

# Estadística Básica Edulcorada

*Alejandro Quintela del Rio*

*2019-09-03*



# Índice general

<b>Prólogo</b>	<b>19</b>
<b>1 Introducción</b>	<b>1</b>
1.1 De la aritmética política a la estadística . . . . .	2
1.2 Variabilidad y correlación . . . . .	9
<b>2 Estadística descriptiva</b>	<b>17</b>
2.1 Los censos . . . . .	17
2.2 Proceso científico del análisis estadístico . . . . .	19
2.3 Población . . . . .	19
2.4 Variables y Atributos . . . . .	20
2.5 Variables discretas y continuas . . . . .	21
2.6 Distribuciones de frecuencias . . . . .	22
2.7 Variables continuas . . . . .	28
2.8 Medidas características de una variable . . . . .	45
<b>3 Descriptiva dos-dimensional</b>	<b>89</b>
3.1 Distribución conjunta de caracteres . . . . .	90
3.2 Estructura de la información . . . . .	91
3.3 Representaciones gráficas . . . . .	92
3.4 Dependencia e Independencia estadística. . . . .	93
3.5 Dependencia estadística entre variables . . . . .	93
<b>4 Probabilidad</b>	<b>119</b>
4.1 Introducción histórica . . . . .	119
4.2 La geometría del azar . . . . .	123
4.3 El sentido estadístico-probabilista de la actualidad . . . . .	125
4.4 La familia Bernoulli . . . . .	126

4.5	Laplace . . . . .	129
4.6	desde Poisson hasta Kolmogorov . . . . .	130
4.7	Experimentos aleatorios . . . . .	131
4.8	Definiciones básicas . . . . .	132
4.9	Probabilidad . . . . .	139
4.10	¿Cómo se asignan probabilidades a los sucesos? . . . . .	142
4.11	Propiedades de la probabilidad . . . . .	146
4.12	Probabilidad condicionada . . . . .	150
4.13	Independencia de sucesos . . . . .	154
4.14	Regla de Bayes . . . . .	157
4.15	Teorema de la probabilidad total (Laplace) . . . . .	160
4.16	Teorema de Bayes (Versión de Laplace) . . . . .	162
4.17	Notas históricas. La estadística Bayesiana . . . . .	165
4.18	La falacia del fiscal . . . . .	175
4.19	Sobre la formación correcta de jurados. . . . .	190
4.20	El caso Dreyfuss . . . . .	192
4.21	Problemas de paradojas . . . . .	197
<b>5</b>	<b>Variables aleatorias</b>	<b>205</b>
5.1	Tipos de variables aleatorias . . . . .	207
5.2	Variables aleatorias discretas . . . . .	208
5.3	Variables aleatorias continuas . . . . .	209
5.4	Esperanza Matemática de una variable aleatoria . . . . .	216
5.5	Varianza de una variable aleatoria . . . . .	217
5.6	Mediana y Cuantiles (o percentiles) . . . . .	226
5.7	La moda . . . . .	227
5.8	Variables discretas notables . . . . .	227
5.9	Variables aleatorias continuas notables . . . . .	235
<b>6</b>	<b>La Variable Normal o Gaussiana</b>	<b>251</b>
6.1	Ejemplos de la distribución normal . . . . .	253
6.2	El papel de Quetelet . . . . .	277
6.3	Para pensar un poco: El CI . . . . .	288
6.4	Reproductividad de la variable normal . . . . .	293
6.5	El teorema central del límite . . . . .	297
6.6	Las leyes de los grandes números . . . . .	308
6.7	V.A. obtenidas a partir de la variable normal . . . . .	311

<b>7</b>	<b>Inferencia estadística</b>	<b>319</b>
7.1	Muestreo aleatorio simple: . . . . .	320
7.2	Estimación puntual . . . . .	323
7.3	Ejemplos de interés. . . . .	339
7.4	Intervalos de confianza . . . . .	341
7.5	I.C. para la media (1) . . . . .	350
7.6	I.C. para la media (2) . . . . .	353
7.7	I.C. para una proporción . . . . .	356
7.8	I.C. para la diferencia de proporciones. . . . .	359
7.9	Comparación de variables. . . . .	361
7.10	I.C. para la diferencia de medias . . . . .	361
7.11	I.C. para el ratio de varianzas . . . . .	363
<b>8</b>	<b>Contrastes de hipótesis</b>	<b>365</b>
8.1	De conjeturas y otros . . . . .	365
8.2	Hipótesis estadísticas . . . . .	367
8.3	Test de significación (NHST) . . . . .	369
8.4	La teoría de Neyman-Pearson . . . . .	371
8.5	Contrastes paramétricos y no paramétricos . . . . .	375
8.6	Contrastes de hipótesis paramétricas . . . . .	381
8.7	Contraste para una proporción . . . . .	389
8.8	Contrastes para comparación de poblaciones . . . . .	393
8.9	Para el cociente de varianzas . . . . .	395
8.10	Muestras pareadas o relacionadas . . . . .	399
8.11	Para la diferencia de proporciones . . . . .	401
8.12	Problemas del nivel de significación . . . . .	409



# Índice de tablas

7.4 Table continues below . . . . .	342
-------------------------------------	-----



# Índice de figuras

1	Creative Commons License . . . . .	19
1.1	John Graunt (izquierda) y William Petty (derecha). . . . .	4
1.2	William Farr (izquierda) y Florence Nightingale (derecha). . . . .	6
1.3	Diagrama polar, presentado por Florence Nightingale. . . . .	6
1.4	André Guerry (izquierda) y Adolphe Quetelet (derecha), considerado uno de los padres de la sociología, y creador del índice de masa corporal. . . . .	7
1.5	Sir Francis Galton (izquierda), primo de Charles Darwin, y Pierre Simon Laplace (derecha), dos de las más importantes figuras de la probabilidad y la estadística. . . . .	9
1.6	Dos famosos retratos de Karl Pearson (1857-1936), considerado, junto con Francis Galton, el creador de la estadística moderna. Su hijo Egon Pearson también haría muchos aportes a la estadística. . . . .	10
1.7	Ronald Fisher, considerado el fundador de la ciencia estadística, tal y como la conocemos. . . . .	11
1.8	aaaay. . . . .	14
1.9	En aquella película, Robin Williams sí gana el Oscar como mejor actor. . . . .	15
2.1	Los gráficos: la quintaesencia de la descripción. . . . .	18
2.2	Metodología de un análisis estadístico. . . . .	20
2.3	Resumen de conceptos. . . . .	21
2.4	Resumen de los tipos de variables. . . . .	23
2.5	¿Por qué no termina bien? . . . . .	24
2.6	Ejemplo de diagrama de barras. . . . .	26
2.7	Ejemplo de diagrama de sectores. . . . .	27

2.8	Ejemplo de diagrama de barras con las librerías mosaic y lattice.	27
2.9	Diagrama de barras distinguiendo sexos.	28
2.10	Clasificación de datos en intervalos.	29
2.11	Histograma de la variable Edad	33
2.12	Izquierda: magnitudes de terremotos en Galicia. Fuente: IGN. Derecha: salarios brutos anuales de asalariados en 2012. Fuente: Ministerio.	33
2.13	Histogramas variando el número de intervalos.	35
2.14	Captura de pantalla de histograma.	36
2.15	Un gráfico generado mediante ggvis.	37
2.16	Construcción de un polígono de frecuencias.	38
2.17	Polígono e Histograma de la variable Edad.	39
2.18	Comparativa de un histograma y un estimador núcleo de la densidad.	40
2.19	Histograma y estimador de la densidad	42
2.20	Estimaciones de la densidad de la variable Estatura	43
2.21	Estimaciones de la densidad de la variable Peso	44
2.22	Captura de pantalla de uno de los gráficos generados mediante la librería ggvis.	45
2.23	Esta propiedad la conocen perfectamente los chavales.	47
2.24	Datos=Pesos. La media aritmética es el centro de gravedad.	48
2.25	Blancanieves, la reina, el cazador y los enanitos.	52
2.26	Población de g.. (gente bien vestida): valores más frecuentes de abrigos, solapas y sombreros.	53
2.27	Mostramos de nuevo el gráfico de los salarios Piénsese en la diferencia entre el sueldo moda, mediana y media aritmética.	54
2.28	Sir Winston Churchill, presidente inglés durante la II guerra mundial.	55
2.29	Percentil $p$ en una variable continua.	55
2.30	Posible uso de interés de los percentiles.	56
2.31	Tabla de percentiles de 2 a 20 años, por sexo, peso y estatura.	57
2.32	Cuartiles: Percentil 25, Mediana y Percentil 75.	57
2.33	Percentiles 5 y 95 de la variables Peso.	58
2.34	Cuartiles variable Colesterol LDL	59
2.35	Dos variables con la misma media aritmética y diferente dispersión.	60
2.36	Índice de masa corporal según dolor de espalda. Variabilidad diferente según cada grupo.	61

2.37	Distancias de los puntos a la media aritmética. . . . .	62
2.38	Distribución asimétrica a la derecha. . . . .	67
2.39	Pelo y Simetria. . . . .	69
2.40	Estimación de la densidad de la variable estatura. . . . .	70
2.41	Estimación de la densidad de la variable peso. . . . .	70
2.42	Curva Normal o Campana de Gauss. . . . .	72
2.43	Distribuciones según la curtosis. . . . .	75
2.44	Dibujo original hecho por Student (Gosset) para acordarse de los valores de la curtosis. . . . .	77
2.45	Características de un Boxplot. . . . .	79
2.46	Datos atípicos, extremos y anómalos. . . . .	80
2.47	El inefable John Tukey. . . . .	81
2.48	Boxplot básico . . . . .	83
3.1	Muertes por arma de fuego en USA en 2010. . . . .	89
3.2	Individuo de una población: Ejemplo de caracteres que puede interesar medir. . . . .	90
3.3	Tabla de doble entrada para una variable bidimensional. . . . .	91
3.4	Diagrama de barras para la Tabla de contingencia de los datos del Titanic . . . . .	93
3.5	Datos originales estudiados por Galton. . . . .	94
3.6	Diagrama de dispersión mediante plot. . . . .	95
3.7	Datos relacionando dureza del agua y mortalidad. . . . .	96
3.8	Cuadrantes relativos al centro de gravedad de la nube de puntos. . . . .	98
3.9	Ejemplo de diagrama de dispersión relacionando el índice de masa corporal y la grasa corporal. . . . .	99
3.10	Posibles valores para el coeficiente de correlación. . . . .	101
3.11	Correlación cero y relación matemática perfecta. . . . .	103
3.12	Correlación lineal cero para un donuts y un dinosaurio. . . . .	104
3.13	Ejemplo de una relación de tipo logarítmico entre variables. . . . .	104
3.14	Ejemplo de una relación de tipo exponencial. . . . .	105
3.15	Datos relacionando dureza del agua y mortalidad. . . . .	106
3.16	Línea de regresión para los datos del ejemplo anterior. . . . .	107
3.17	Tiempo realizado frente al año de la carrera, junto con la recta de mínimos cuadrados. . . . .	108
3.18	Francis Galton y una de sus muchas frases célebres. . . . .	109
3.19	Recta de regresión de la estatura de los hijos en función de la media de los padres. . . . .	112

3.20	Gráfico original de Galton. . . . .	113
3.21	Los Bernoulli, una familia de genios de la ciencia. Aquí en España tenemos a los Pelayo y a los del Río (no son primos míos, ojo) . . . . .	114
3.22	Dakota Johnson (50 sombras de Grey), hija de Melanie Griffith y Don Johnson, famosos actores de los años 80. . . . .	115
3.23	Otra hija de Melanie Griffith (de su matrimonio con Antonio Banderas). Además, mucha gente desconoce que su abuela (la madre de Melanie) fue la actriz protagonista de ‘Los Pajaros’. . . . .	115
4.1	El marqués de Laplace. . . . .	120
4.2	Izquierda: Blaise Pascal. Derecha: Pierre Fermat . . . . .	124
4.3	*Abraham de Moivre (1667-1754), matemático francés que predijo el día de su muerte a través de un cálculo matemático (parece que se equivocó en 73 días). Según se dice, contaba el tiempo que dormía, y veía que cada día dormía algo menos que el anterior, por lo cual pudo deducir que llegaría un día en que no dormiría y eso sería cuando muriese - caso para Iker Jiménez.*. . . . .	126
4.4	Carl Gauss. . . . .	129
4.5	Simeón Poisson. . . . .	131
4.6	Andréi Kolmogórov. . . . .	132
4.7	Representación de un suceso contenido en un espacio muestral. . . . .	134
4.8	Diagrama de Venn de la unión de conjuntos. . . . .	135
4.9	Suceso A: doble grado en español y gallego. Suceso B: doble grado en inglés y español. Intersección: grado en español. . . . .	136
4.10	”La ciencia es un juicio verdadero acompañado de razón”, Platón. . . . .	137
4.11	Algunos ejemplos de sucesos incompatibles. . . . .	137
4.12	Diferencia de sucesos. . . . .	137
4.13	Suceso complementario. . . . .	138
4.14	Diagrama de Venn con animales. . . . .	139
4.15	Diagrama de Venn de familias Cristianas. . . . .	139
4.16	A algunos es muy difícil ponerlos de acuerdo hasta con un diagrama de Venn. . . . .	140
4.17	Para entender bien como se componen las islas británicas. . . . .	140
4.18	Venga, más... . . . .	141
4.19	Cocinero experto en probabilidades. . . . .	143

4.20 Diagrama de Venn de la unión e intersección de sucesos: la probabilidad de la unión es la suma de la probabilidad de  $A$  y la probabilidad de  $B$ , pero hay que restar la probabilidad de la intersección, puesto que, en otro caso, la contaríamos dos veces. . . . . 148

4.21 Diagrama de Venn de la diferencia de sucesos. . . . . 149

4.22 Tabla de contingencia menopausia - osteoporosis. . . . . 151

4.23 Diagrama de venn Fumadores (1) . . . . . 152

4.24 Diagrama de venn Fumadores (2) . . . . . 152

4.25 Diagrama de venn Fumadores (3) . . . . . 153

4.26 Mayor dependencia, imposible. . . . . 156

4.27 Thomas Bayes, reverendo presbiteriano inglés (1702-1761). . . 158

4.28 Izquierda: Cartel de la película Rush (2013). Derecha: Los verdaderos Niki Lauda y James Hunt. . . . . 159

4.29 Hasta en la tele. . . . . 160

4.30 Partición: sucesos disjuntos que recubren el total. . . . . 161

4.31 El filósofo, economista, sociólogo e historiador David Hume. . 167

4.32 Un caso real de sueño profético [fuente: digital]. . . . . 170

4.33 Fraga saludando a los cientos de estadísticos bayesianos que se agolparon en la playa, y que la censura franquista impidió ver. . . . . 174

4.34 El estadístico italiano Bruno de Finetti. . . . . 175

4.35 Izquierda: Benedict Cumberbatch en el papel de Alan Touring en la película "The imitation game" (2015). Derecha: el verdadero Alan Touring. . . . . 176

4.36 El implacable fiscal. . . . . 177

4.37 Juicio Real y Serie de TV. . . . . 181

4.38 Mujeres asesinadas dentro del colectivo de mujeres maltratadas: únicamente 1 de cada 2500. . . . . 182

4.39 Imagen de la serie de Tv "El caso O.J. Simpson" (2016) en la famosa escena del guante. . . . . 183

4.40 Recorte de periódico inglés. . . . . 187

4.41 Ibáñez nunca se equivocó. . . . . 190

4.42 Portada de un periódico francés original. . . . . 192

4.43 Henri Poincaré, uno de los matemáticos más prestigiosos de la historia. . . . . 193

4.44 Izquierda: Émile Zola. Derecha: Paul Muni, en la película "La vida de Emile Zola" (1937), que le valió al actor una nominación al Oscar. . . . .	193
4.45 Portada del diario original con la famosa carta de Zola. . . . .	194
4.46 Gregory Perelman. . . . .	196
4.47 El presentador de televisión Monty Hall y su amiga. . . . .	197
4.48 Casos posibles del problema. . . . .	198
5.1 Proceso científico del análisis estadístico. . . . .	206
5.2 Esquema general de relación entre variables aleatorias y estadísticas. . . . .	207
5.3 Ejemplo de una función de densidad bien simple. . . . .	211
5.4 Ejemplo de dos variables discretas con prácticamente la misma media y diferente varianza (dispersión). . . . .	219
5.5 La ruleta francesa: números del 1 al 36, y además el cero. . . . .	219
5.6 Cuantil o percentil. . . . .	226
5.7 Distribución con varias modas. . . . .	227
5.8 Nespresso, what else?. . . . .	228
5.9 Nespresso, binomial. . . . .	229
5.10 Café de Poison. . . . .	231
5.11 Y aún da para más... . . . .	233
5.12 Última. . . . .	234
5.13 función de densidad de una variable aleatoria uniforme entre a y b. . . . .	236
5.14 Intervalos con la misma longitud tienen la misma probabilidad (área). . . . .	236
5.15 La probabilidad es el área del rectángulo que va desde 105 a 120. . . . .	237
5.16 Gráficas de la densidad de la variable exponencial. . . . .	240
5.17 Gráfica del INE de los salarios en 2015 en España. Puede observarse una forma de tipo exponencial o ley de potencias a partir del salario medio, aproximadamente. . . . .	244
5.18 Distribución exponencial, representada mediante su función de densidad, y mediante un gráfico log-log. . . . .	245
5.19 Figura 8 del trabajo de Clauzet, Shalizi y Newman (2009). . . . .	248
5.20 Figura 9 del trabajo de Clauzet, Shalizi y Newman (2009). . . . .	249

6.1	La campana de Gauss presenta simetría y la misma distribución de áreas, independientemente de cuáles sean sus parámetros media y desviación típica. . . . .	253
6.2	Distribución del movimiento de las moléculas de los gases. . .	255
6.3	Gráfica del trabajo de Ritchie et al. (2017)., donde se observa que las características antropométricas del cerebro humano también siguen la distribución normal (algo que, evidentemente, Quetelet no llegó a comprobar) . . . . .	263
6.4	Densidad de los datos de Inestabilidad emocional y Extraversión, por sexos. . . . .	265
6.5	Densidad de los datos de la compulsividad sexual, por sexos. .	265
6.6	Densidad de los datos de dos variables del estudio de Hansenne (2001), por sexos. . . . .	266
6.7	Densidad de los datos de otras dos variables del estudio de Hansenne (2001), por sexos. . . . .	267
6.8	. . . . .	268
6.9	Áreas en la campana de Gauss. . . . .	269
6.10	Densidad del conjunto de datos correspondiente a la velocidad de 82 galaxias. . . . .	272
6.11	Izquierda: Figura original del artículo de Bregman. Proyección de casos hasta el 94. Derecha: datos recientes. . . . .	273
6.12	Número de Bacon de Carmen de Mairena. . . . .	274
6.13	Estudio de los usuarios de Facebook. . . . .	276
6.14	Gustav Fechner, a quien se le atribuye la introducción del concepto de mediana. . . . .	276
6.15	Charles Sander Pierce. . . . .	277
6.16	Adolphe Quetelet. . . . .	278
6.17	Tabla de estaturas de 26.000 soldados americanos del ejército del Norte durante la guerra civil (*Phys. Soc,* i., p. 131; *Anthropom.,* p. 259). . . . .	281
6.18	Tabla de medidas alrededor del pecho, de los soldados del ejército del Potomac (*Phys. Soc,* ii., 59; *Anthropom.,* p. 289). . . . .	281
6.19	Tabla de las medidas de fuerza en hombres, estimadas mediante un dinamómetro (*Anthropom.,* p. 365). . . . .	282
6.20	Tabla original del trabajo de Quetelet. . . . .	284
6.21	Tabla original del trabajo de Quetelet. . . . .	285
6.22	Tabla original del trabajo de Quetelet. . . . .	286

6.23	Tabla original del trabajo de Quetelet. Observemos que los números, año a año, son casi similares, lo cual provocó el asombro de Quetelet (y de cualquiera). . . . .	286
6.24	Cyril Burt, ¿uno de los mayores troleros de la historia? . . . . .	293
6.25	Densidad de la suma de cinco variables aleatorias. . . . .	296
6.26	G. H. Hardy, que fue interpretado por Jeremy Irons en la película *El hombre que conocía el infinito*. . . . .	307
6.27	Simeón Poisson, del que ya hablamos en el capítulo de probabilidades. . . . .	309
6.28	Varias distribuciones Chi-cuadrado para distintos grados de libertad. . . . .	313
6.29	William Gosset, el inventor. . . . .	313
6.30	Varias distribuciones t para distintos grados de libertad. . . . .	315
6.31	Representación de la función de densidad de la distribución F para algunos parámetros. . . . .	316
7.1	Gráfico del proceso científico del estudio estadístico. . . . .	319
7.2	Encuestador y encuestada. . . . .	325
7.3	Histograma y estimación de la densidad de 100 estaturas. . . . .	326
7.4	Histograma y estimación de la densidad de las 250 MEDIAS de todas las muestras. . . . .	327
7.5	Seres extraños de otra galaxia. . . . .	328
7.6	Histograma y estimación de la densidad de los valores obtenidos en cada una de las 250 muestras. . . . .	330
7.7	Tasas de cáncer de riñón en USA. . . . .	334
7.8	Tanque Panzer alemán, de la II guerra mundial. . . . .	339
7.9	Fijémonos que hay intervalos que no contienen al parámetro verdadero. . . . .	347
7.10	Gráfico llamado Forest-Plot del meta-análisis para los medicamentos antidepresivos. . . . .	349
7.11	Niveles de significación en una normal estandarizada . . . . .	351
7.12	J. Bardem y J. Dean Morgan (o al revés). . . . .	361
8.1	Una famosa frase de Ronald Fisher. . . . .	366
8.2	. . . . .	367
8.3	Otra conocida frase de Fisher acerca de la inferencia científica. . . . .	368
8.4	. . . . .	371
8.5	Posibles opciones en un test de hipótesis. . . . .	372

8.6	Ejemplos errores tipo I y II. . . . .	373
8.7	Tres posibles histogramas para una distribución de notas. . . . .	379
8.8	Imagen del cómic Rue del Pércebe, 13, F. Ibáñez. . . . .	381
8.9	Resumen de tipos de contrastes y cálculo del $p$ -valor. . . . .	390
8.10	Siempre de buen humor, qué gran tipo. . . . .	392
8.11	Densidades de las variables $x$ e $y$ . . . . .	397
8.12	Densidad de la estatura de hombres y mujeres. . . . .	398
8.13	$p$ -valor para un test $t$ bilateral. . . . .	411
8.14	Área fuera del intervalo $(-2,2)$ . . . . .	415



# Prólogo

**Advertencia:** Libro en fase de elaboración. No se recomienda copiar trozos, puesto que después podría haber lloros si hay acusaciones de plagio.

La estadística para gente inteligente.

Este libro está bajo licencia [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

Los paquetes o librerías que se utilizan en este libro son las que siguen. Para ejecutar trozos de código particulares no habría que instalar todas, pero, si en algún momento la ejecución de algún ejemplo da error, podría deberse a que no se ha(n) instalado la(s) librería(s) necesaria(s).

```
library(mosaic)
library(dplyr)
library(ggplot2)
library(DT)
library(plotly)
library(pander)
library(xtable)
library(manipulate)
library(HistData)
library(ggthemes)
```



Figura 1: Creative Commons License

```
library(gridExtra)
library(printr)
library(readxl)
library(rcompanion)
library(HSAUR)
library(kableExtra)
library(UsingR)
library(readr)
library(grid)
library(gmodels)
library(nortest)
options(knitr.table.format = "latex")
```

Los ficheros de datos que se utilizan a lo largo del libro se han obtenido de fuentes públicas (generalmente de paquetes de R), pero pueden obtenerse en este [enlace](#) (por si desaparecen del original).

En este libro se usa R y RStudio (entorno gráfico para utilizar R). Existen muchos manuales de R, entre los que podemos citar:

- Para quien no lo haya usado nunca, puede comenzar mirando: *R para Principiantes*, versión española de *R for Beginners*, traducido por [Jorge A. Ahumada](#),
- *Metodos Estadisticos con R y R Commander* by [Antonio Jose Saez Castillo](#).
- En la página <http://wpd.ugr.es/~bioestad/> existen guias de R, Rstudio y RCommander, estructuradas por capítulos.
- Una página web de la Universidad de Las Palmas de Gran Canaria, muy buena, al estilo de la anterior: <http://www.dma.ulpgc.es/profesores/personal/stat/cursoR4ULPGC/index.html>
- Un libro sencillo para empezar a usar Rstudio, la esencia de los datos en R y las gráficas con la librería *ggplot2*: *Ciencia de datos para curiosos* de [Martín Montané](#).
- Un libro de estadística con R muy básico: *Métodos Cuantitativos*, de [Aleksander Dietrichson](#).
- Un libro más técnico (también más completo): *R para profesionales de los datos: una introducción*, de [Carlos Gil Bellosta](#).

# Capítulo 1

## Introducción

El siguiente video realizado por la Universidad Carlos III recoge el sentir mayoritario de la población sobre qué es la estadística, algo un poco alejado de la realidad:

---

“Data, data, give me data. I can’t make bricks without straw  
(Datos, datos, denme datos. No sé hacer ladrillos sin paja)”

— Sherlock Holmes

“With data collection, ‘the sooner the better’ is always the best  
answer.”

— Marissa Mayer, former president and CEO of Yahoo

“Denme menos datos, pero más inteligentes.”

— Lenin

“Un hombre muerto es una desgracia. Un millón de muertos es  
una estadística.”

— Stalin

“Solo creo en las estadísticas que he manipulado.”

— Winston Churchill.

“Cuando las estadísticas no tienen sentido, en general me parece más sabio preferir el sentido común a las estadísticas .”

— John M. Keynes.

Estamos en la época del *Big Data*. Está claro que podemos trabajar con muchos datos, pero también que debemos realizar un análisis *inteligente* de los mismos. Cantidad y calidad nunca han sido sinónimos. En este caso, se trata de extraer, de la cantidad, calidad. Para ello son necesarios algunos conocimientos de estadística.

## 1.1 De la aritmética política a la estadística

La mayoría de las ideas y notas que se ofrecen en esta introducción han sido tomadas de dos libros: *Los orígenes de la estadística: de investigación socio-política empírica a conjunto de técnicas para el análisis de datos*, de Juan I. Piovani (Piovani, 2007) y *Masa Crítica: cambio, caos y complejidad*, de Philip Ball (Ball, 2004), cuya lectura recomiendo para el que le gusten estos temas.

“La concepción estadístico-probabilística de la realidad ha pasado a ser uno de esos sistemas de ideas que el hombre contemporáneo ha naturalizado. Concebir los hechos de la vida —incluso aquellos de la vida cotidiana— desde una perspectiva estadística se encuentra tan difundido que frecuentemente ni siquiera el hombre de ciencia se detiene a pensar que se trata en realidad de una construcción socio-histórica, un producto del pensamiento humano; no de una característica natural intrínseca a la realidad misma”.

— J.I. Piovani

Es muy común entender como estadística cualquier estudio (más o menos breve) que consista en una descripción gráfica y/o numérica de un conjunto de datos. Sin embargo, lo que se conoce como *teoría estadística* o *estadística matemática* corresponde, desde hace ya tiempo, a la **disciplina científica dedicada a los instrumentos conceptuales y operativos para el análisis cuantitativo de datos**.

La estadística como ciencia (de los datos y las probabilidades) surge como

un proceso de transformación de tres disciplinas que se iniciaron a mediados del siglo XVII: la **Political Arithmetic** inglesa, la **Statistik** alemana y la **Théorie mathématique des probabilités** francesa. En un principio, la estadística se dedicaba exclusivamente a la recolección de información a través de datos que realizaban entes oficiales. Posteriormente, este tipo de estudios se enlazaron con la teoría matemática de las probabilidades, para converger en la actual *ciencia de los datos*.

A continuación, detallamos un poco más los comienzos de esta disciplina:

Desde muy antiguo, los distintos gobiernos realizaron, continua o discontinuamente, los llamados censos de población, principalmente encaminados al control de las cifras de habitantes y al pago de los impuestos. Se conocen censos en fechas muy antiguas en China, en Egipto, en el imperio Romano...

A partir del siglo XVII algunos investigadores comenzaron a pensar que los datos numéricos obtenidos de estos censos tal vez podían ser utilizados para interpretar el comportamiento de la sociedad. A través de la observación de los datos a lo largo del tiempo se comprobó que existían ciertas pautas que se mantenían invariablemente, o bien cambiaban de alguna forma que permitía intuir el modo y la magnitud de dicho cambio. Los que primero observaron esta cualidad de los datos fueron los ingleses. El primero en utilizar el análisis estadístico para estudiar los problemas sociales fue **John Graunt** (considerado el primer demógrafo, fundador de la bioestadística y precursor de la epidemiología; 1620-1674), mediante su trabajo *Natural and Political Observations on the Bills of Mortality*.

Las *London Bills of Mortality* fueron la primera recopilación de datos demográficos de Gran Bretaña. En ellas se recogieron tablas de mortalidad y también las causas de los mismos. Dicha recopilación de datos se llevó a cabo por organismos gubernativos, como se haría también en otros estados europeos, fundamentalmente en Francia y Prusia. El trabajo de John Graunt fue precursor en el tipo de análisis, puesto que no se limitó a describir las cifras, sino en darles contenido y sentar las bases de los trabajos que más tarde serían necesarios para la introducción de cualquier reforma social.

Más tarde, **William Petty** (filósofo, médico y economista inglés, 1623-1687), introdujo la expresión *political arithmethic* (Cullen, 1975), que más tarde derivaría en *estadística social*. Esta se definió como una *ciencia de la sociedad cuyas conclusiones dependen de números y medidas*.

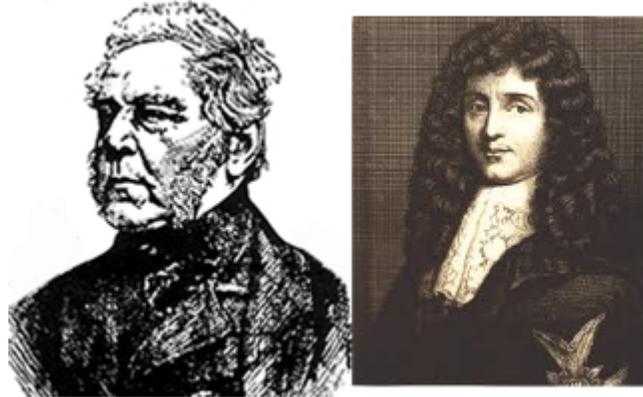


Figura 1.1: John Graunt (izquierda) y William Petty (derecha).

Entre 1780 y 1830 comenzaron a manifestarse algunas señales de expansión y diversificación de las técnicas y usos de las estadísticas sociales en Gran Bretaña: se instituyó un censo nacional, se reforzó el desarrollo de las estadísticas médicas, se propusieron nuevas técnicas de análisis y se aplicaron técnicas ya conocidas a problemas hasta entonces no cuantificados (Cullen, 1975). En este periodo apareció el término **statistics** en la lengua inglesa, como una traducción que Hooper (1770) hizo del vocablo alemán **Statistik**. Siguiendo la tradición alemana, el neologismo inglés fue definido como la ciencia que *enseña el ordenamiento político de los estados modernos conocidos en el mundo*.

En Alemania, sin embargo, la palabra **Statistik** tenía un significado diferente, definiéndose como la *ciencia que se ocupa de los estados naturales de la sociedad*. Uno de los precursores de la disciplina, **J.P. Süßmilch** (presbítero, estadístico y demógrafo alemán; 1707-1767), por ejemplo, sostenía en su libro *El orden divino* que las diferencias de las tasas de natalidad y de mortalidad de niños y niñas se equilibraban perfectamente a fin de que todos los ciudadanos tuvieran perspectivas de matrimonio. De esta forma, empezó a considerarse la idea de que la sociedad mantenía reglas que ningún gobernante decretaba. El mismo **Immanuel Kant** habló de “leyes universales”. En 1752, **Gottfried Achenwall**, (economista y profesor en Göttingen; 1719-1772) presentó un trabajo sobre las constituciones de varios estados europeos, describiendo, de una forma similar a lo que hoy llamaríamos estadística, detalles sobre su agricultura, industria y comercio.

Uno de los primeros trabajos de estadística, en este mismo sentido, apareció por primera vez en lengua inglesa en 1791, con la publicación del primero de los 21 volúmenes de *Statistical Account of Scotland*. En ellos, John Sinclair (1754-1835) describió a la **investigación estadística** como aquella que hace referencia a la población, las circunstancias políticas, la producción y otros aspectos del Estado. Mientras las investigaciones alemanas llamadas *statistik* se interesaban por las cuestiones del Estado, aquellas propuestas por Sinclair buscaban determinar la forma del progreso para el pueblo. Fue la primera persona que usó el término “estadística” en el lenguaje inglés.

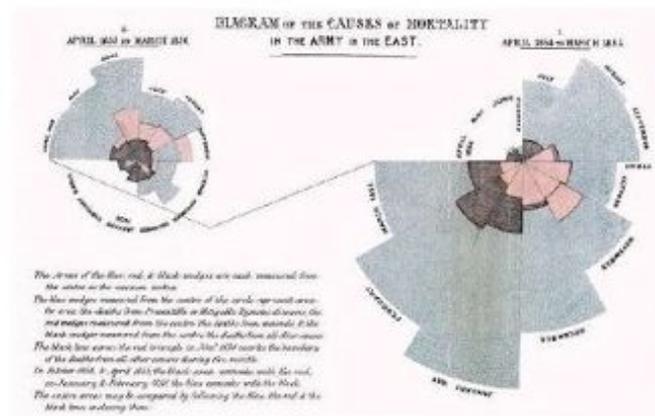
El desarrollo de las estadísticas numéricas comenzó en Gran Bretaña en la primera mitad del siglo XIX. La recolección y compilación de datos numéricos estaba ya bien establecida en esas fechas, pero todavía no se realizaba ningún tratamiento matemático avanzado de los mismos. Según la dinámica habitual, los datos “crudos”, sin “opiniones”, constituían la característica esencial de la ciencia y el fundamento de la estadística. Así, la primera regla de la *Royal Statistical Society* (fundada en 1834 como *Statistical Society* de Londres) fue excluir los pareceres. **William Farr** (epidemiólogo inglés; 1807-1883), en nombre de esta Sociedad, escribió a **Florence Nightingale** (enfermera, escritora y estadística británica; 1820-1910): “no queremos impresiones, sólo hechos”, refiriéndose a los números, puesto que Florence Nightingale se preocupaba no solo de analizar y describir estos hechos (o números), sino de interpretar los mismos, a la manera en que actualmente se realizan los análisis estadísticos de datos. Más adelante, sin embargo, la opinión de Farr cambiaría, cuando F. Nightingale le ayudó a analizar los datos de la guerra de Crimea (1853-1856), revelándose que la causa de 16.000 de las 18.000 muertes de soldados británicos no fueron heridas sufridas en batallas, sino enfermedades provocadas por falta de higiene.

En 1832 se fundó el Departamento de Estadística del Ministerio de Comercio (Board of Trade), una oficina centralizada capaz de generar estadísticas para la ayuda de las reformas promovidas por importantes actores políticos y sociales de la época. Los estudios estadísticos de aquellos años se encaminaban, fundamentalmente, a intentar establecer las causas de problemas sociales, como el aumento de los crímenes, la inestabilidad social, etc. Los promotores de estas investigaciones solían ser, en general, contrarios a la intervención del gobierno en la economía, pero sí partidarios de la intervención en sanidad y educación. Progresivamente, fueron incorporándose técnicas analíticas de contenido matemático más avanzado -ya muy habituales en astronomía- y,



Figura 1.2: William Farr (izquierda) y Florence Nightingale (derecha).

Recolección de datos y registro estadístico con el sistema:  
Diagrama de área polar



Formato Estadístico aprobado por el Congreso Internacional de Estadística en Londres en 1860

Figura 1.3: Diagrama polar, presentado por Florence Nightingale.

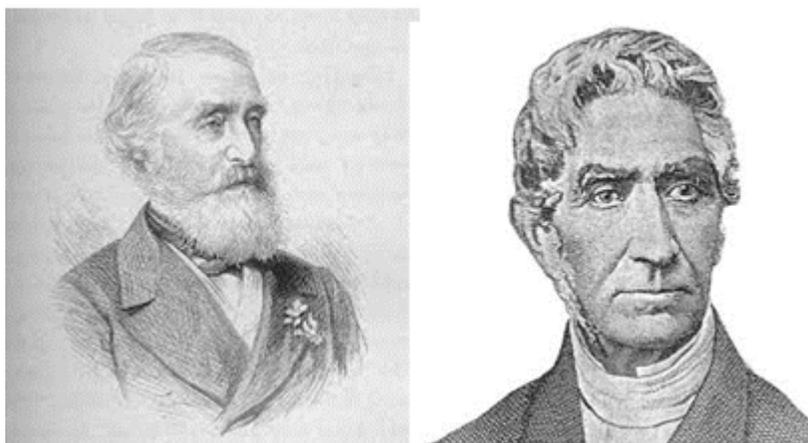


Figura 1.4: André Guerry (izquierda) y Adolphe Quetelet (derecha), considerado uno de los padres de la sociología, y creador del índice de masa corporal.

de esta forma, comenzaron a identificarse muchas regularidades, lo que a su vez se reforzó con el uso de las probabilidades (Porter, 1986).

Paralelamente, en Francia surgirían dos figuras que pasarían a la posteridad, al igual que sus homólogos ingleses: el abogado **André-Michel Guerry** (1802-1866) y el astrónomo **Adolph Quetelet** (Bélgica, 1796-1874)

“La estadística se puede aplicar a la observación de los actos criminales de forma idéntica a cualquier otro acto. Cuando uno sabe cómo tratar hechos establecidos, los datos se agrupan de tal manera que se separen las circunstancias meramente accidentales. Los resultados que se presentan entonces aparecen con una regularidad tal que es imposible atribuirlos al azar. Cada año se registran el mismo número de delitos del mismo tipo. Cada clase de delito tiene su peculiar distribución por sexo, por edad, por estación... Estamos obligados a reconocer que muchos aspectos de las estadísticas judiciales representan una certeza absoluta. Nos vemos obligados a reconocer que los hechos del orden moral están sujetos, lo mismo que los del orden físico, a leyes invariables.”

— Carta de Guerry a Quetelet; julio de 1931.

Guerry había constatado estos hechos a partir de los datos de la publicación anual *Recherches statistiques sur la ville de Paris et le département de la Seine*,

cuyo director era el matemático **Joseph Fourier** (1768-1830). De los mismos, se deducían impresionantes estabilidades sobre los delitos, los crímenes y los suicidios.

En Francia, el suicidio más común era mediante asfixia por carbón de leña y por arrojarse al agua. En Londres, por ahorcamiento o por arma de fuego. No solo se estimaba que los procedimientos suicidas eran regulares, sino que los suicidios variaban según las estaciones. Igual que ahora, los habitantes de Inglaterra y Gales son los menos suicidas de Europa (excluyendo a los irlandeses). De la misma forma, los europeos de cualquier nación eran más suicidas durante el verano que durante el invierno.

Guerry había sido contratado en 1827 por el Ministerio de Justicia francés para estudiar los datos sobre los delitos en París. En 1829 publicó, junto con el geógrafo veneciano Adriano Balbi, una serie de mapas de Francia analizando y comparando los delitos contra la propiedad y contra las personas, en los distintos departamentos. De este modo, no logró encontrarse una relación directa entre educación y criminalidad. El trabajo de Guerry fue el precursor de los mapas geográficos criminológicos. Como puede comprobarse en el trabajo de M. Friendly ([Friendly, 2007](#)), la obra de Guerry es “una oportunidad” para aprender y/o enseñar análisis multivariante con datos de corte histórico (disponibles en el paquete de R *Guerry*).

En 1864, Guerry publicó *Statistique morale de l'Angleterre comparée avec la statistique morale de la France*. En esta obra, analizó 226.000 casos de delitos contra las personas en Francia e Inglaterra, a lo largo de 25 años, y más de 85.000 registros de suicidios, indicando su supuesta causa.

El avance fundamental, sin embargo, que acabaría derivado en la teoría estadística moderna, no surgiría hasta finales del siglo XIX, cuando se empezó a estudiar la variabilidad de los procesos que regían no sólo la sociedad, sino los comportamientos y características individuales. Francis Galton (1822-1911) fue el máximo responsable de lo que sería un paso de gigante en el avance de la disciplina que nos compete.

La teoría darwinista de la evolución llevó a Galton a interesarse por el problema de la herencia de los diferentes atributos humanos. Para ello, estudió la teoría de los errores (la curva Gaussiana o Normal) que ya Quetelet había utilizado para estudiar las características físicas y psicológicas humanas. Profundamente influenciado por los trabajos de Laplace (1749-1827), Quetelet

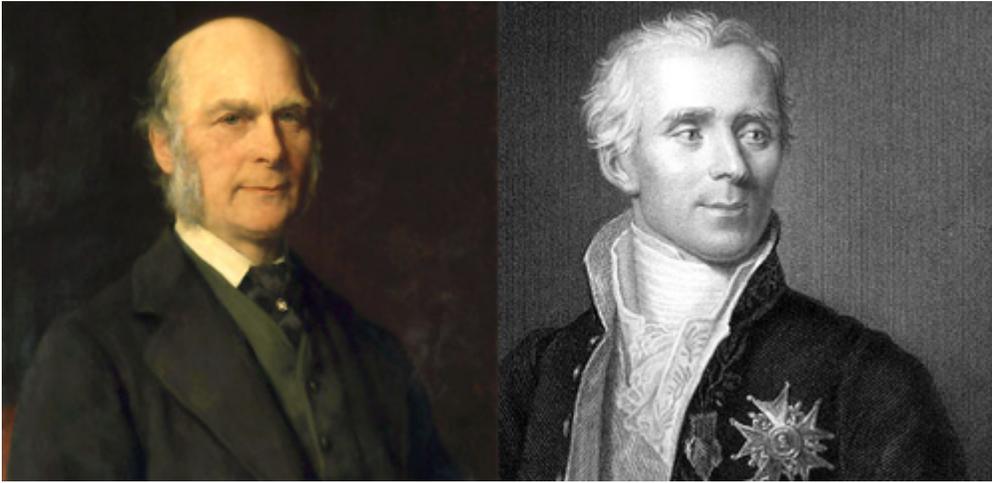


Figura 1.5: Sir Francis Galton (izquierda), primo de Charles Darwin, y Pierre Simon Laplace (derecha), dos de las más importantes figuras de la probabilidad y la estadística.

se convenció, hacia 1830, de la posibilidad de tratar las cuestiones humanas aplicando los métodos estadísticos que se usaban ya en astronomía, yendo más allá de la mera recolección y clasificación de los datos.

## 1.2 Variabilidad y correlación

Quetelet utilizó el término **hombre medio**, como sinónimo del “hombre perfecto”, del cual se producían desviaciones que eran producto del azar y que, cuanto más se diferenciaban de la media, más imperfectas serían. Sin embargo, Galton consideró la hipótesis contraria: las desviaciones en torno a la media no eran necesariamente un defecto. Como reformista social y fundador del movimiento *eugenésico*, le interesaba mejorar la raza, y por lo tanto, el hombre medio no era para nada el modelo ideal; era un mediocre cuyos caracteres debían ser mejorados en la descendencia. Por eso, Galton necesitaba estudiar la forma de poder variar ese hombre medio hacia los valores más extremos (que constituyesen el mejor de los valores posibles, como la altura más elevada o la inteligencia más prodigiosa). Galton introdujo los conceptos de regresión y correlación para analizar la variabilidad de los fenómenos y sus relaciones. La ciencia, como pretendía Galileo, seguiría orientándose a la

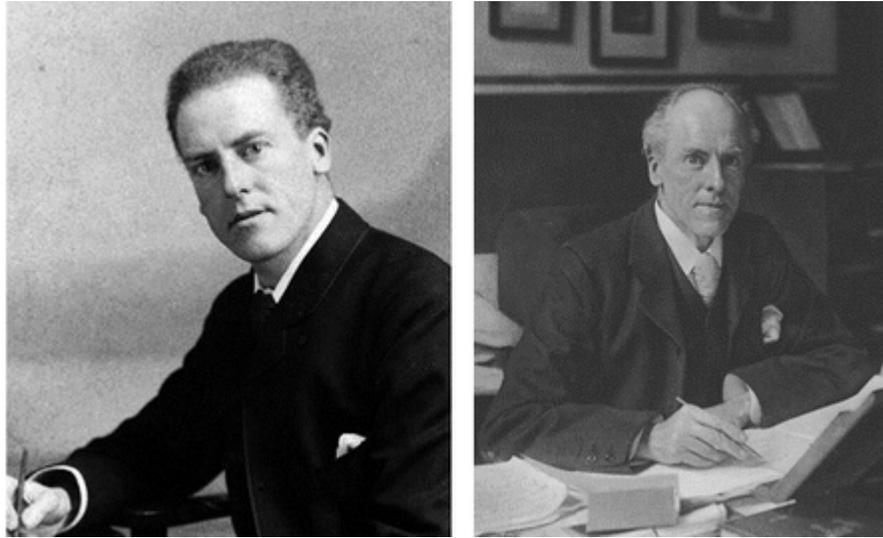


Figura 1.6: Dos famosos retratos de Karl Pearson (1857-1936), considerado, junto con Francis Galton, el creador de la estadística moderna. Su hijo Egon Pearson también haría muchos aportes a la estadística.

búsqueda de relaciones, pero éstas ya no serían consideradas necesarias ni inmutables.

La aparición del libro *Natural Inheritance* de Galton y del análisis de la correlación marcan el inicio del período moderno de la estadística. La difusión de estos conceptos, en torno a 1890, comenzó a llamar la atención de notables estadísticos, como **Francis Edgeworth** (1845-1926), **Frank Weldon** (1860-1906) y **Karl Pearson** (1857-1936).

Hacia 1900, el carácter intelectual de la estadística saldría completamente transformado gracias al trabajo de estos tres últimos científicos. En sus primeros cursos de estadística avanzada, impartidos en la universidad de Londres a finales del siglo XIX, Karl Pearson enseñaba que “estadística es [...] un término utilizado para mediciones agregadas de cualquier hecho, sea social, físico o biológico. La teoría pura de la estadística es una rama de la matemática que se ocupa de la compilación, representación y tratamiento de estos agregados numéricos -independientemente de los hechos que tales números representen. La estadística aplicada es el uso de los métodos de la estadística pura a clases especiales de hechos — observaciones biológicas, físicas o



Figura 1.7: Ronald Fisher, considerado el fundador de la ciencia estadística, tal y como la conocemos.

políticas.” (Yule, 1938).

De su colaboración con Frank R Weldon y Francis Galton surgió la Escuela Biométrica. Desde 1890 hasta los años veinte del siglo XX —cuando **Ronald Fisher** (1890-1962) estableció un centro de investigación en la estación experimental de Rothamstead— la Escuela Biométrica fue la única institución británica dedicada a la teoría estadística. Ya en la primera década de su creación, alrededor de la mitad de todos los artículos ingleses sobre teoría estadística se producían en este ambiente institucional.

Pearson dominó los avances de la teoría científica hasta la primera guerra mundial. Desde los años 20, la figura de Ronald Fisher tomó su testigo, constituyendo hoy en día, junto con Galton y Pearson, los cimientos humanos fundamentales de la disciplina estadística como tal.

Sir Ronald Aylmer Fisher fue un estadístico y biólogo inglés que utilizó las matemáticas para combinar la genética mendeliana y la selección natural. Richard Dawkins dijo de él que fue “el mejor biólogo desde Darwin”. Lo que nadie pone en duda es que proporcionó a los investigadores en biología y medicina sus herramientas de trabajo más importantes.

Fisher obtuvo una beca para estudiar Matemáticas en la Universidad de Cambridge en 1909, obteniendo su primer título en Astronomía en 1912. En 1915 publicó un artículo sobre la evolución de las preferencias sexuales y la elección de pareja. En 1918 publicó el trabajo *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*, en el que introdujo el término **varianza** y propuso su análisis formal. Presentó un modelo conceptual de genética que mostraba que la variación continua entre los rasgos fenotípicos podía producirse por la acción combinada de muchos genes discretos y, por lo tanto, ser el resultado de la herencia mendeliana. Este fue el primer paso hacia el establecimiento de la genética poblacional y la genética cuantitativa, que demostró que la selección natural podía cambiar las frecuencias alélicas de una población, combinando su naturaleza discontinua con la evolución gradual.

En 1919 comenzó a trabajar en el laboratorio de investigación de Rothamsted (Hertfordshire, Inglaterra). En 1925 publicó el famoso libro *Métodos estadísticos para investigadores* (Fisher, 1949), una de las monografías más influyentes del siglo XX en esta ciencia. Este libro también popularizó el **nivel crítico** o **p-valor**. Posteriormente, en 1935, publicaría otro libro fundamental: *El diseño de experimentos*.

Pero, además de por su extraordinaria inteligencia y sus aportaciones a la ciencia, Fisher también dejaría para la posteridad bastantes anécdotas que, en muchas ocasiones, acompañan a la figura del que es catalogado como un genio. En su libro *La teoría que nunca murió*, (McGrayne, 2012) Sharon B. McGraney comenta que Fisher, además de tener una miopía enorme, solía ir tan enfrascado en sus pensamientos y problemas matemáticos que, habitualmente, tenía que ser apartado de sufrir un atropello por parte de los autobuses al cruzar la calle. Además, era de un temperamento arrogante y tomaba muchas preguntas como un ataque personal, de modo que hasta él mismo reconocía que su exaltado temperamento le amargaba la existencia. William Kruskal dijo que la vida de Fisher no había sido más que “una interminable sucesión de altercados científicos —que en ocasiones llegaban incluso a simultanearse—, ya fuera en las reuniones de la profesión o en los artículos propios del oficio” (Kruskal, 1980). Y, en el retrato fundamentalmente amable de la carrera de Fisher hecho por otro famoso estadístico, Jimmie Savage, puede leerse: “a veces decía unos insultos que sólo un santo podría perdonar enteramente [...]” (Savage, 1976).

Fue también famosa la enemistad que mantuvo Fisher con otro de los padres de la estadística, Karl Pearson, a raíz de las críticas de este último a un artículo enviado por Fisher a la revista *Biometrika*, creada por Pearson, Galton y Weldon. Esta enemistad se trasladaría al hijo de Pearson, **Egon Pearson** (1895-1980), que junto con **Jerzy Neyman** (1894-1981) crearon la teoría de los test de hipótesis que lleva sus nombres (Neymann-Pearson). Además, Fisher mantuvo siempre una animadversión severa hacia los estadísticos llamados *bayesianos* (los que utilizan un modo de pensamiento centrado en las probabilidades de tipo subjetivo), lo que seguramente motivo que la estadística bayesiana tardara más tiempo del deseado en tomar arraigo entre los investigadores de esta ciencia.

Es también muy conocido el siguiente hecho: En 1950 se pronunció públicamente en contra de un estudio que demostraba la correlación entre fumar y el cáncer de pulmón; incluso llegó a argumentar que era posible que la existencia de un cáncer motivara la posibilidad de adicción hacia el tabaco. Si bien fue cierto que Fisher había sido asesor de empresas tabacaleras, parece estar demostrado que sus motivaciones para dudar de la relación tabaco - cáncer fueron estrictamente personales. Fisher era un fumador empedernido, capaz de bañarse en la piscina sin dejar de fumar en pipa, siempre había mantenido aversión hacia las tendencias puritanas de todo tipo y había encontrado gran consuelo personal en el tabaco. Años después publicaría una retractación de sus opiniones en este aspecto.

En la siguiente entrada del blog [simplystatistics](#), el autor considera que Fisher es el científico más influyente de la historia, en base a las citas de sus trabajos en Google Scholar. En esta misma página, se indica que varios de los conceptos introducidos por Fisher son de la mayor importancia en la estadística. Por ejemplo:

- el  $p$ -valor.
- el análisis de la varianza o ANOVA.
- el análisis discriminante lineal.
- la matriz de información.

Cualquier estudiante de un curso de estadística, o cualquier científico, ingeniero, psicólogo o profesional de muchas disciplinas ha tenido que oír hablar, casi necesariamente, del  $p$ -valor. Pensemos que, si los conceptos científicos



Los del Río: hicieron la Macarena y se forraron.  
 Mark Zuckerberg: una aplicación web y recontramillonario  
 Fisher: millones de citas de sus trabajos y ni un duro. Aaaaaaaaay

Figura 1.8: aaaay.

podiesen ser registrados igual que un software, o simplemente como una canción en la sociedad general de autores (SGAE), lo bien que vivirían los nietos de Fisher sin tener que dar palo al agua.

### 1.2.1 Curiosidad para cinéfilos

El matemático Jerzy Neyman fue protagonista de la siguiente anécdota histórica: cuando daba clase en la universidad de Berkeley, en 1935, solía dejar en algunos encerados de las paredes laterales algunos problemas de matemáticas que no se habían logrado resolver. Un estudiante que llegó tarde a una clase pensó que uno de ellos era un problema que Neyman había dejado como ejercicio y se lo entregó correctamente resuelto al día siguiente. Neyman se presentó el domingo a primera hora de la mañana en la residencia del estudiante para felicitarlo por semejante proeza. El estudiante era **George B. Dantzig**, famoso matemático que inventó el método del **Simplex**, considerado uno de los algoritmos más importantes del siglo XX -por ejemplo, entre los 10 primeros según la SIAM (Society for Industrial and Applied Mathematics).



Figura 1.9: En aquella película, Robin Williams sí gano el Oscar como mejor actor.

En el periódico [ABC](#) cuentan con detalle esta anécdota, y también cómo fue tomada “prestada” por Matt Damon y Ben Affleck para el guión de la película *El indomable Will Hunting* (1997), que les valió el Oscar al mejor guión original de aquel año (en la que los dos trabajaron como actores).



# Capítulo 2

## Estadística descriptiva

### 2.1 Los censos

Un censo es una recopilación de datos elaborada por un gobierno para recabar información variada de los habitantes de una región o de un país. Los censos son los precedentes de la institucionalización de la recogida de datos demográficos y económicos por los estados modernos.

Se conocen censos tan antiguos como el realizado En Egipto en la Dinastía I, en el año 3050 a.C., o en China, en el año 2238 a.C. el emperador Yao mandó elaborar un censo general que recogía datos sobre la actividad agrícola, industrial y comercial. En antiguas civilizaciones como Babilonia, Egipto, China, Roma etc. era normal que se elaboraran recuentos de la población. La estadística aparece en los textos de varias religiones. Por ejemplo en la Biblia, en el libro de los Números, se menciona la elaboración de un censo de población, en el que se anotarían específicamente los varones mayores de 20 años (aptos para ir a la guerra) [INE: Historia de la Estadística](#) :

“Tomad el encabezamiento de toda la congregación de los hijos de Israel por sus familias, por las casas de sus padres, con la cuenta de los nombres, todos los varones por sus cabezas: 3. De veinte años arriba, todos los que pueden salir á la guerra en Israel, los contaréis tú y Aarón por sus cuadrillas.”

— libro de los Números, cap. 1

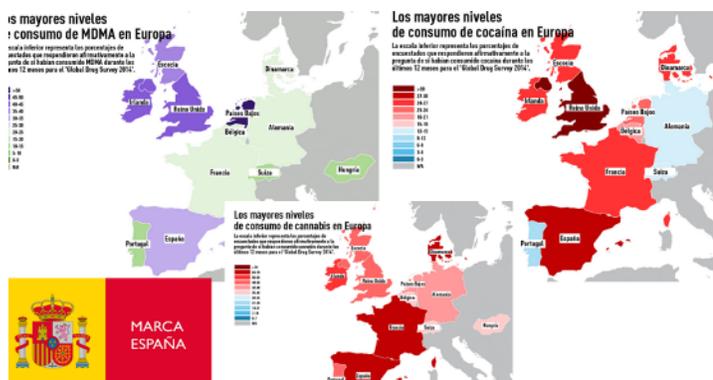


Figura 2.1: Los gráficos: la quintaesencia de la descripción.

Al principio de la película [Ben-hur](#) (justo después de los títulos de crédito) se recoge la escena de como los romanos organizaron un censo en la provincia (ocupada) de Judea, coincidiendo con el nacimiento de Jesús de Nazaret.

La elaboración de censos y publicación de todo tipo de informes estadísticos suelen realizarla los institutos u oficinas de estadística. La creación de estos institutos ha sido habitual en la creación de los estados modernos, desde finales del siglo XVIII hasta nuestros días. Javier Segura del Pozo, en su [blog](#) comenta la importancia que ha tenido y sigue teniendo la Agencia Central de Estadística de Palestina en el intento de consolidación de un estado palestino. Y cómo es uno de los objetivos militares israelitas prioritarios en los ataques a las instalaciones de la Autoridad Palestina en Cisjordania. El estado de Israel es consciente de que las estadísticas sobre población, indicadores económicos, de salud, etc. pueden proporcionar luz sobre muchos elementos del conflicto y exponerlos a debate público en la población mundial.

También en España, paralelamente a la creación de las comunidades autónomas tras la aprobación de la constitución de 1978, surgieron institutos regionales de estadística, para proporcionar información precisa del territorio concreto.

El censo de Quintanilla o censo de 1482, según la fecha que le asignó Tomás González al dar la primera noticia de él en 1829, sería el primer censo de población español del que se tiene constancia, y habría sido realizado con fines militares por Alonso de Quintanilla, contador mayor de cuentas de los Reyes Católicos. En realidad, lo único que se conoce de dicho censo es lo

que consta en un documento de cuatro folios conservado en el Archivo de Simancas, el llamado Apuntamiento de Quintanilla, escrito de su puño y letra y sin fecha, pero probablemente redactado en 1493, tras la toma de Granada [Wikipedia].

El uso de las tablas cruzadas en la Alemania dividida del siglo XVIII sirvió para “cubrir de un vistazo” la multiplicidad de situaciones y puntos de vista, e ir describiendo el estado a partir de un lenguaje común, para así ayudar a construir la unidad de Alemania en el siglo XIX. La urgencia creada por situaciones de hambruna, epidemia o guerra o los objetivos fiscales, fueron el origen de las encuestas sobre la población y las subsistencias en los siglos XVII y XVIII.

En Estados Unidos, la propia constitución, en su primer artículo, recoge la obligación de realizar un censo de población cada diez años, para poder conocer con exactitud el número de representantes legislativos de cada estado.

## 2.2 Proceso científico del análisis estadístico

La estadística se ocupa del estudio de **poblaciones** (entendiendo este estudio a partir de las características de las mismas).

En general, estas poblaciones van a ser muy grandes (en teoría, de infinitos elementos). Por la imposibilidad de examinar todos los elementos de la población, uno a uno, se elegirá un conjunto representativo de esa población, llamada **muestra**. Esa muestra se analizará convenientemente mediante la *estadística descriptiva*, y, a continuación, utilizando herramientas de la teoría de la probabilidad, se intentarán obtener conclusiones de la población en general.

## 2.3 Población

Entendemos genéricamente por **población** el conjunto sobre el cual recaen las observaciones o los datos que analizamos (puede ser una población de personas, de monos, de libros, de radiadores... ). Si tomamos un subconjunto de elementos de la población decimos que hemos tomado una **muestra**. Cada uno de los elementos de la población reciben el nombre de **unidades estadísticas** o **individuos**.



Figura 2.2: Metodología de un análisis estadístico.

Ej.

Un **ejemplo** de población podría ser una biblioteca, y un conjunto de libros de la misma sería una muestra.

La estadística **descriptiva** como su nombre indica, se ocupa de la descripción de una población, lo cual se realiza a través de la descripción de las diferentes variables que interesa conocer o medir en la misma. La descripción de una variable se puede realizar a través de tres conceptos generales:

- su **valor medio**.
- su **variabilidad**.
- su **forma**.

## 2.4 Variables y Atributos

Una primera clasificación del tipo de datos procede del hecho de que las observaciones sean de tipo *cuantitativo* (indican una cantidad) (también se llaman *categorías*) o *cuantitativo* (miden una cantidad).

En el primero de los casos se tiene un **atributo**, y en el segundo una **variable**. Para hacer referencia genéricamente a una variable o a un atributo se utiliza



Figura 2.3: Resumen de conceptos.

el término *carácter*.

Como ejemplos de atributos tenemos el color del pelo, raza, idioma de un grupo de personas, y como variables su estatura, peso, dinero que lleven en el bolsillo, etc.

Si es necesario operar con un atributo, se le asignará a cada una de sus clases (**categorías**) un valor numérico, con lo que se transforma en una variable. Esta asignación se hará de forma que los resultados que se obtengan al final del estudio sean fácilmente interpretables. Por ejemplo, si hay 3 idiomas posibles (inglés, francés y español) podemos usar 1, 2 y 3.



Las variables estadísticas suelen representarse con letras mayúsculas:  $X, Y, Z, T \dots$

## 2.5 Variables discretas y continuas

Si las observaciones corresponden a cantidades, las variables pueden distinguirse entre **discretas** y **continuas**. Se dice que una variable es **discreta** cuando no puede tomar ningún valor entre dos consecutivos, y que es **continua** cuando puede tomar cualquier valor dentro de un intervalo.

Ej.

**Ejemplos** de variable discreta: número de empleados de una fábrica; número de hijos; número de cuentas ocultas en Suiza.

Ej.

**Ejemplos** de variable continua: temperaturas registradas en un observatorio; tiempo en recorrer una distancia en una carrera; contenido de alcohol en un cuba-libre; estatura; tiempo de discurso de un político en las cortes insultando a los del partido contrario.

En la práctica, todas las variables son discretas debido a la limitación de los aparatos de medida. Pensemos en el ejemplo de la estatura: quizá podría detectarse una diferencia de una cienmilésima de metro, o de una millonésima pero, dados dos individuos que se diferencien en una millonésima, seguramente ya no existe otro que tenga una estatura intermedia. De la misma forma, por insignificante que sea la diferencia entre la llegada de dos corredores olímpicos a una meta, la limitación de la precisión en la medida siempre puede acabar produciendo un posible (aunque improbable) empate. Obviando este tipo de limitaciones, las variables continuas se elegirán, desde un punto de vista teórico, con toda la precisión que deseemos (decimales), de manera que siempre podamos escribir un valor que esté entre cualesquiera otros dos.

Por ejemplo, hasta hace unos años, cuando uno se subía a una báscula, lo máximo solían ser dos o tres decimales: (50.35: 50 kilos y 350 gramos...). Realmente, usar más de dos o tres decimales para el peso en kilos de una persona no ofrece ningún interés. Sin embargo, está claro que podrían construirse básculas de baño con mucha mayor precisión, ofreciendo un número mayor de decimales (otra cosa es que se vendieran).

De manera general, consideraremos que *una variable continua toma valores en un intervalo* (por ejemplo: el peso de un grupo de personas varía entre 40 y 100 kilos), y un dato podrá ser cualquier número entre esos dos.

## 2.6 Distribuciones de frecuencias

La organización de los datos constituye la primera etapa de su tratamiento, puesto que facilita los cálculos posteriores y evita posibles confusiones. Realmente, la organización de la información tiene una raíz histórica y, ac-

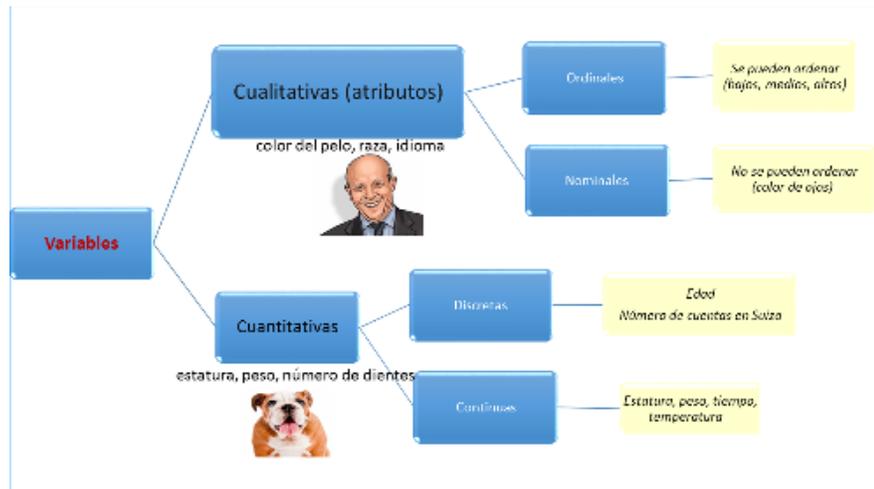


Figura 2.4: Resumen de los tipos de variables.

tualmente, con el desarrollo de los medios informáticos, tiene menos importancia desde un punto de vista aplicado. Cuando no existían ordenadores, o ni siquiera calculadoras, si se disponía de un conjunto de datos, era necesario dotarlos de alguna estructura que permitiera resumirlos y comprenderlos de una forma más o menos sencilla.

La organización va a depender del número de observaciones distintas que se tengan y de las veces que se repitan cada una de ellas. En base a lo anterior, se pueden estructurar los datos de maneras diferentes.

Cuando se tiene un gran número de observaciones, pero muy pocas distintas, se pueden organizar en una **tabla de frecuencias**, es decir, cada uno de los valores acompañado de la frecuencia (también llamada frecuencia absoluta) con la que aparece. Este es el tipo de tabla que acompaña a una variable discreta.

Ej.

La siguiente tabla indica que los valores 1 y 3 se repiten 12 veces, el valor 5 se repite 3 veces, etc....



Figura 2.5: ¿Por qué no termina bien?

$x_i$	$n_i$
1	12
3	12
5	3
6	45
8	72

Ej.

Este es también el formato con que suele representarse también una variable *cuantitativa* o *categorica*, como por ejemplo la *distribución del color del cabello de veinte personas*:

Color del pelo	Número de personas
Rubio	2
Moreno	6
Castaño	5
Verde	7

Ej.

El 10 de abril de 1912, el Titanic zarpaba con 1317 pasajeros a bordo, ante la admiración de una muchedumbre de curiosos que contemplaban atónitos como aquella mole de acero se alejaba majestuosamente del puerto. Cinco días después los medios de comunicación de todo el mundo se hicieron eco de la increíble noticia: el barco más grande jamás construido yacía a casi cuatro mil metros de profundidad.

Esta historia la conocemos todos desde 1997 (cuando se estrenó Titanic).

Desde estas fechas (o quizá de antes) se trabaja muy a menudo con el fichero de datos que contiene la lista de pasajeros, junto con la distribución de los mismos según edad, sexo, supervivencia y clase social.

Este fichero lo utilizaremos en más de una ocasión a lo largo de estos apuntes. En el siguiente trozo de código (*chunk*), los datos, al estar en formato excel, los leemos mediante la librería *readxl*.

```
library(readxl)
Datos_Titanic <-read_excel("Data/PasajerosTitanic.xlsx")
```

### 2.6.1 Gráficos para variables discretas o categóricas

Como una imagen vale más que mil palabras, con las representaciones gráficas se puede obtener una idea del contenido de una variable (su forma, cómo se distribuye, qué valores aparecen más y cuáles menos...). Para una variable discreta (con pocos valores, en general) o categórica existen dos gráficos fundamentales: los **diagramas de barra** y los de **sectores**.

En R los comandos básicos para hacer estos diagramas son *barplot* y *pie*. Para hacer un diagrama de barras:

```
# Recordemos que, en R, para escoger una
# variable de un fichero, se usa Nombre_fichero$Variable
y=Datos_Titanic$class
# El comando table crea una tabla de frecuencias
# es necesario crear esta tabla para hacer los gráficos
x=table(y)
# diagrama de barras
barplot(x, main="Titulo principal",
        xlab="Titulo del eje horizontal")
```

Y uno de sectores:

```
y=Datos_Titanic$class
x=table(y)
# diagrama de sectores
```

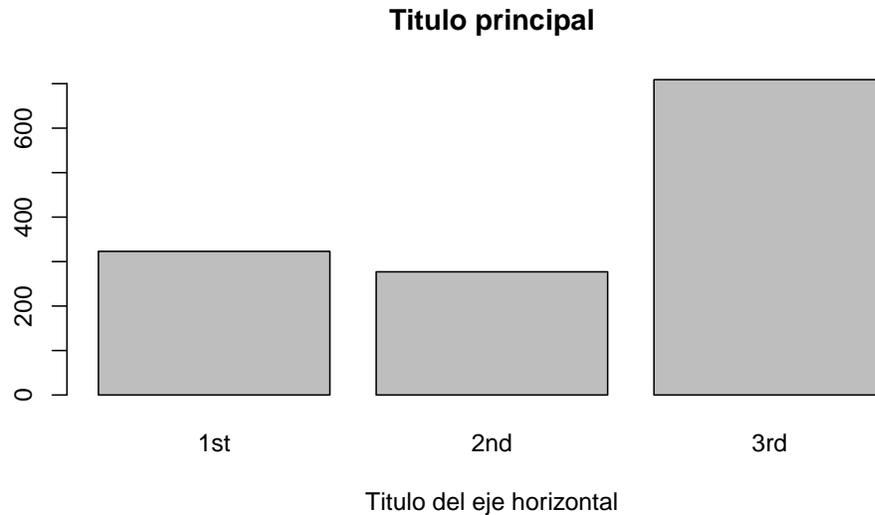


Figura 2.6: Ejemplo de diagrama de barras.

```
pie(x, main="Diagrama de Sectores",
    xlab="Clase del pasajero")
```

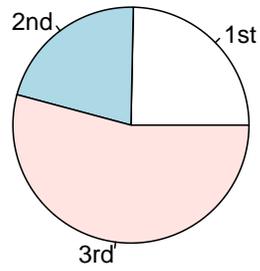
R permite hacer gráficos mucho más profesionales (y escalables). A lo largo de este manual usaremos varias veces la librería **mosaic** (que precisa de la librería **lattice** a su vez), por ser cómoda de utilizar. Por ejemplo, un gráfico de barras se hace (Figura 2.8):

```
library(lattice)
library(mosaic)
#diagrama de barras
bargraph( ~ clase, data = Datos_Titanic,
          main = "Gráfico de Barras")
```

En principio, el gráfico no varía demasiado (tampoco la sintaxis). Con la librería **mosaic**, la sintaxis general es

```
comando( ~ variable, data = fchero de datos)
```

**Diagrama de Sectores**



Clase del pasajero

Figura 2.7: Ejemplo de diagrama de sectores.

**Gráfico de Barras**

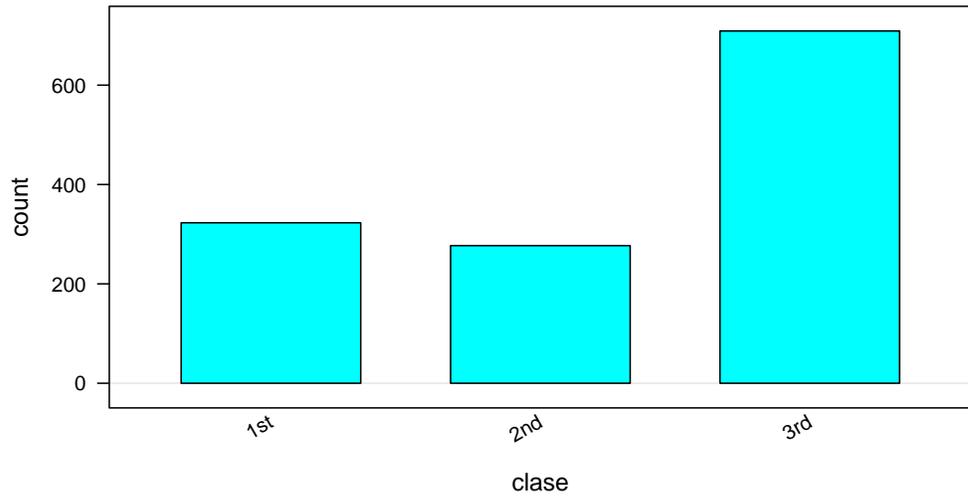


Figura 2.8: Ejemplo de diagrama de barras con las librerías mosaic y lattice.

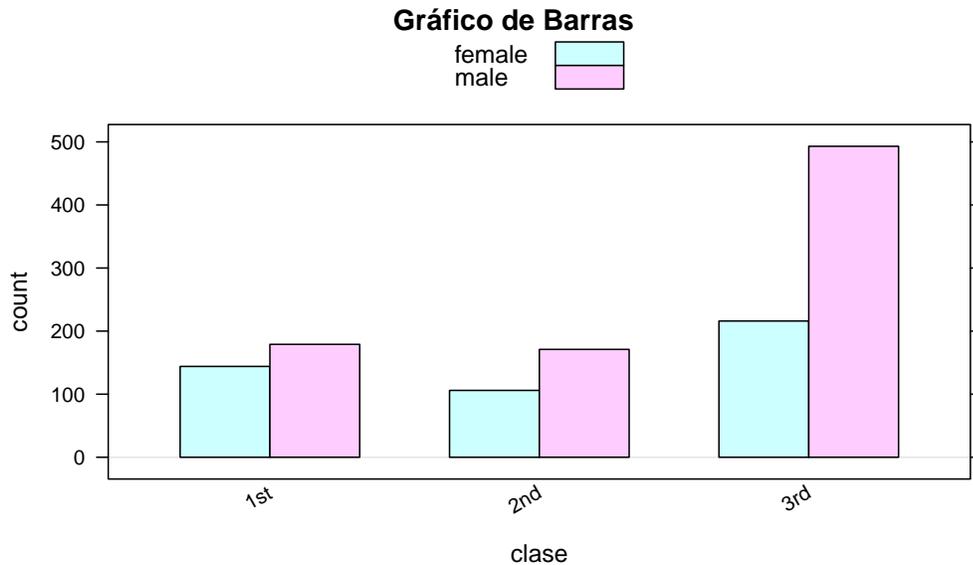


Figura 2.9: Diagrama de barras distinguiendo sexos.

(la ~ es el apóstrofe de la ñ que está al lado del 4 en el teclado). Se escribe `comando( ~ variable)` cuando hacemos un análisis de una sola variable, y

`comando( ~ variable_1, groups= variable_2 )`

si queremos dividir la *variable\_1* por grupos, según la *variable\_2*. Por ejemplo (Figura 2.9):

```
# diagrama de barras relacionando 2 variables
library(lattice)
library(mosaic)
bargraph(~ clase, groups=sexo, data = Datos_Titanic,
         main = "Gráfico de Barras")
```

## 2.7 Variables continuas

Analicemos ahora el caso de variables continuas. En general, el análisis para variables continuas coincide con el análisis para variables discretas cuando existen muchas observaciones, **la mayoría de ellas distintas**. Por ejemplo,

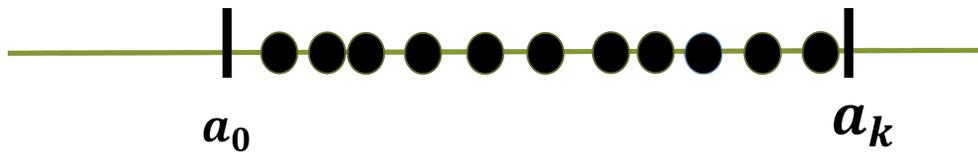


Figura 2.10: Clasificación de datos en intervalos.

la variable edad (en años) se consideraría, en principio, una variable discreta, puesto que no se utilizan valores decimales (tipo 1.5 para año y medio, etc.). Si disponemos, supongamos, de datos de edad correspondientes a muchas personas, lo más probable será que tengamos muchos valores diferentes. Realizar un diagrama de barras o uno de sectores puede no ser una buena idea, ya que ofrecerían poca información.

En este supuesto (variables continuas, o, en general, variables con muchos valores diferentes), los datos pueden disponerse **agrupándolos** o **clasificándolos** en intervalos, e indicando el número de observaciones que caen dentro de cada intervalo.

Para ello se elige un número  $a_0 \leq \min(X)$ , y otro  $a_k \geq \max(X)$ , y se divide el intervalo  $[a_0, a_k]$  en  $k$  intervalos.

Ej.

Supongamos que tenemos los siguientes datos de una variable  $X$ :

2, 3, 4, 4.5, 4.5, 5.6, 5.7, 5.8, 6, 6.1, 6.5, 7, 7, 7,  
7.5, 7.5, 7.5, 8.3, 9, 10.2, 10.4, 11, 11.1, 11.5, 12, 13

Una posible representación o clasificación en intervalos es la siguiente:

```
X=c(2,3,4,4.5,4.5,5.6,5.7,5.8,6,6.1,6.5,7,7,7,
    7.5,7.5,7.5,8.3,9,10.2,10.4,11,11.1,11.5,12,13)
div<-table(cut(X,breaks=7))
library(pander)
mat <-data.frame(div)
names(mat) <-c("Intervalos", "Frecuencias ($n_i$)" )
x<-xtable(mat)
pander(x)
```

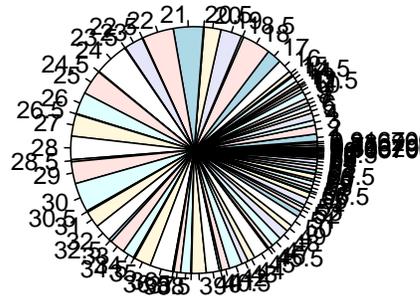
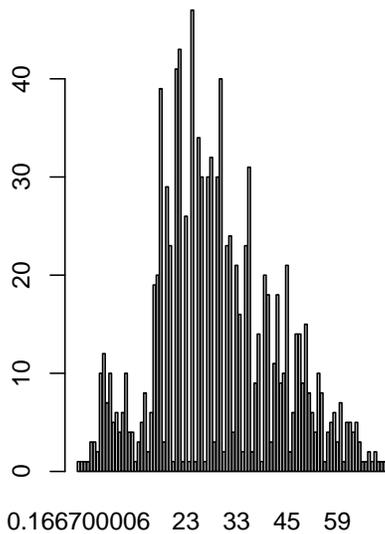
Intervalos	Frecuencias ( $n_i$ )
(1.99,3.57]	2
(3.57,5.14]	3
(5.14,6.71]	6
(6.71,8.29]	6
(8.29,9.86]	2
(9.86,11.4]	4
(11.4,13]	3

Esta clasificación nos dice el número de datos que hay en cada intervalo. El indicar los intervalos de la forma  $(a, b]$  indica que el dato  $a$  no se cuenta en este intervalo, y sí se cuenta el dato  $b$ .

Ej.

Datos de la variable *Edad* de los pasajeros del Titanic. Supongamos, inicialmente, dado que es una variable discreta, que realizamos un diagrama de barras o un diagrama de sectores.

```
x=Datos_Titanic$edad
# clasificamos los datos
y=table(x)
# 2 gráficos en 1 fila, 2 columnas
op <-par(mfrow = c(1,2))
barplot(y)
pie(y)
```



```
# dejamos los gráficos en formato 1 x 1
par(op)
```

Como vemos, el gráfico de barras no es útil porque se “agolpa” la información, y el gráfico de sectores menos. Por este tipo de cosas, es más conveniente *clasificar* los datos, considerándolos como datos procedentes de una variable continua.

Para clasificar los datos en, por ejemplo, seis intervalos, hacemos:

```
X=Datos_Titanic$edad
table(cut(x, breaks=6))
```

(0.0869,13.5]	(13.5,26.8]	(26.8,40.1]	(40.1,53.4]	(53.4,66.7]	(66.7,80.1]
99	375	345	150	68	9

La forma general de una **tabla de frecuencias** es la siguiente:

Intervalo (o dato)	Frecuencia absoluta	Frecuencia relativa
$x_i$ o $[a_0, a_1)$	$n_i$	$f_i$
$x_1$	$n_1$	$f_1 = n_1/n$
$x_2$	$n_2$	$f_2 = n_2/n$
$\vdots$	$\vdots$	$\vdots$
$x_r$	$n_r$	$f_r = n_r/n$

### 2.7.1 Histograma de Frecuencias.

Es uno de los gráficos más antiguos (y de los más utilizados) para representar una variable continua. Una vez que se tienen los intervalos, sobre cada uno de ellos se levanta un rectángulo de área o altura la frecuencia (absoluta o relativa), de manera totalmente equivalente al diagrama de barras: cada intervalo es como un valor de una variable discreta. Cuantos más datos haya en un intervalo, este será más alto, y cuantos menos datos, más bajo.

Su inventor fue Karl Pearson, y el nombre viene de su primera utilización para representar datos históricos.

Ej.

Histograma de la variable *Edad* de los datos del Titanic, realizada por medio de los gráficos básicos de R.

```
x=Datos_Titanic$edad
hist(x, main="Histograma de la edad de los pasajeros del Titanic", xlab="Años")
```

Como comprobamos, con el comando *hist* no es necesario decirle nada sobre la clasificación de datos de la variable; lo hace automáticamente.

Ej.

Histograma de la variable *Magnitud* correspondiente a los terremotos registrados en Galicia por el Instituto Geográfico Nacional (IGN) hasta mayo de 2008, y Salarios en España en 2012. Figura 2.12:

#### 2.7.1.1 El número de intervalos.

El histograma nos sirve para ver el comportamiento de los datos, desde el mínimo al máximo, advirtiéndolo donde se concentran más datos y donde me-

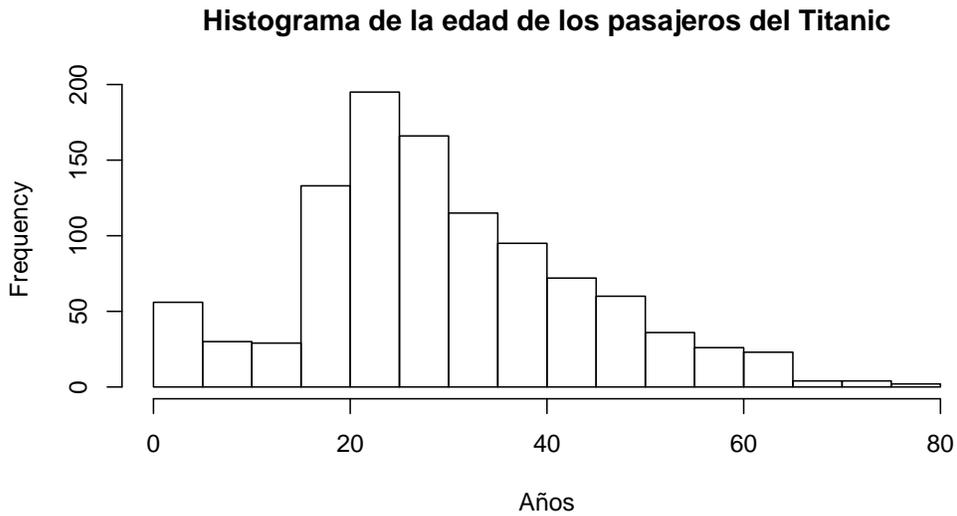


Figura 2.11: Histograma de la variable Edad

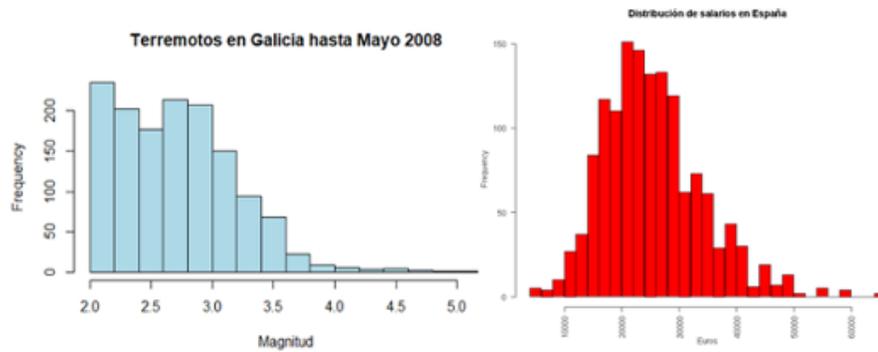


Figura 2.12: Izquierda: magnitudes de terremotos en Galicia. Fuente: IGN. Derecha: salarios brutos anuales de asalariados en 2012. Fuente: Ministerio.

nos. El aspecto de un histograma (y por tanto el de la distribución de los datos) puede cambiar mucho dependiendo del número de intervalos que se utilice. Los programas estadísticos suelen utilizar alguna fórmula que depende del número  $n$  de datos. Por ejemplo, por defecto, el software R o el SPSS utilizan la llamada fórmula de Sturges, que considera el número de intervalos como  $\log_2(n)$ . Otros paquetes utilizan  $\sqrt{n}$ . La selección de un número u otro es un problema matemático con relativa complejidad.

El comando básico `hist` de R puede calcular el número de intervalos mediante otros métodos alternativos al de Sturges, como es el de Scott o el de Friedman-Diaconis (FD) (ver (Wand, 1997) para un estudio sobre la selección del número óptimo de intervalos).

Observemos como cambia la forma o aspecto de un histograma según el número de intervalos. En el comando básico de R `hist` esto se controla escribiendo `breaks=`, en donde `breaks` es el número de intervalos.

Vamos a considerar un ejemplo con el fichero de datos que corresponde a algunas variables de la encuesta nacional de salud en Estados Unidos durante los años 1959 a 1962 (NHES1) (Se puede acceder a estos datos y otros del libro mediante el enlace que está en la página principal). En este fichero se dispone del peso y estatura de 6673 hombres y mujeres (raza blanca y negra). Con los datos del peso de los hombres de raza blanca realizamos 4 histogramas con distintos intervalos cada uno (Figura 2.13), para ver como se producen diferencias en el dibujo y, por lo tanto, en la **forma** de la distribución.

```
NHANES1 <-read.csv("Data/du1003.csv", header=TRUE, sep=";")
library(dplyr)
# filtramos para quedarnos con hombres de raza blanca (código 1)
Datos=NHANES1 %>%
filter(v==1)

# VARIABLE ESTATURA
#
x=Datos$peso

# ponemos gráficos en formato 2 x 2
op <-par(mfrow = c(2, 2))
```

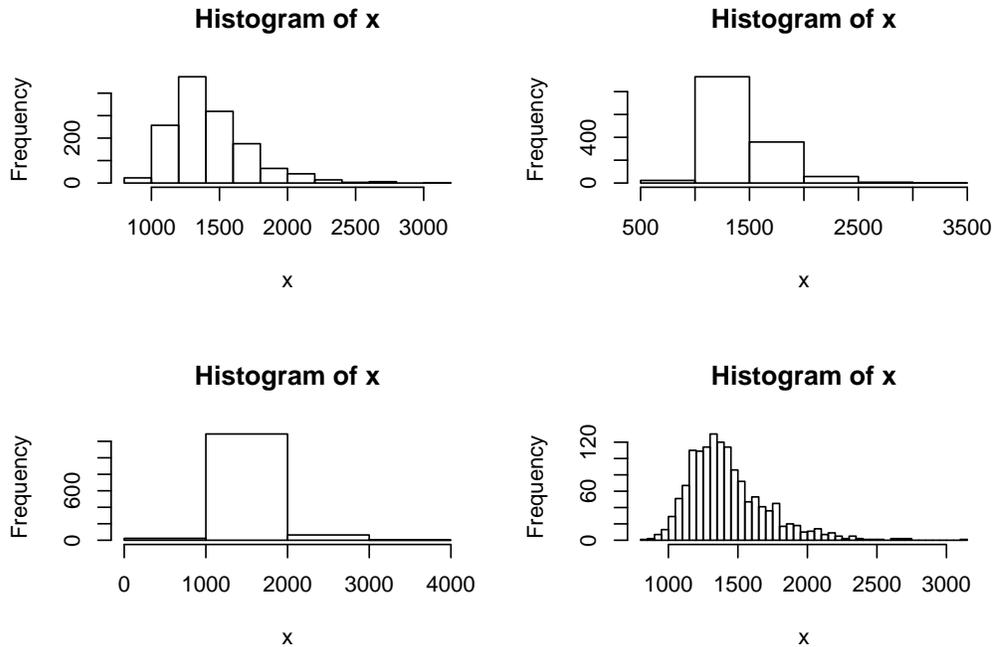


Figura 2.13: Histogramas variando el número de intervalos.

```
# número de intervalos según la fórmula de Sturges
hist(x, breaks="Sturges")
# número de intervalos igual a 5
hist(x, breaks=5)
hist(x, breaks=2)
hist(x, breaks=0)
```

```
# vuelve a poner gráficos de 1 en 1
par(op)
```

Si se quiere probar, de forma interactiva, como cambia un gráfico de histograma según se cambia el número de intervalos (o el ancho de los mismos) se puede probar, dentro de Rstudio, el siguiente trozo de código:

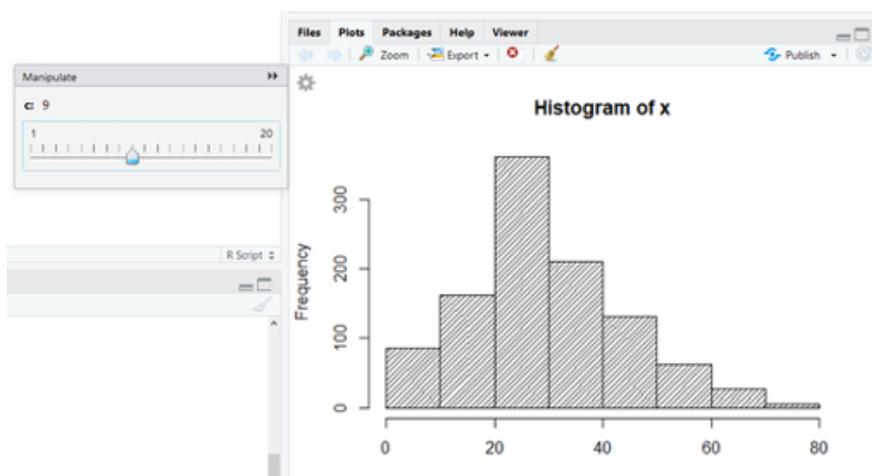


Figura 2.14: Captura de pantalla de histograma.

```
library(manipulate)
x=Datos_Titanic$edad
manipulate( hist(x, breaks=c, density = 25), c=slider(1,11) )
```

Mostramos una gráfica generada por el método anterior (Figura 2.14):

Para probar que, usando otras librerías (o paquetes) se pueden conseguir gráficos con mayor número de prestaciones, ponemos otro ejemplo de gráfico interactivo; en este caso utilizando la librería *ggvis*. Copiando el trozo de código que viene a continuación (hay que instalar primero la librería) obtendremos otra forma de generar histogramas donde podemos variar a mano el número de intervalos.

```
# generamos datos simulados
x1=rnorm(1000,0,3)
# llamamos a la librería ggvis
library(ggvis)
datox=data.frame(x1)
datox %>%
  ggvis(~x1,
        fl1 :="salmon")
```

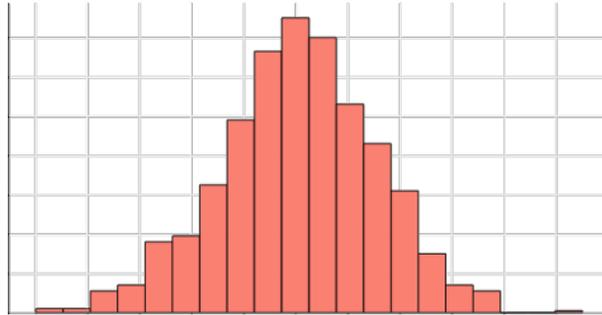


Figura 2.15: Un gráfico generado mediante ggvis.

```

    )%>%
layer_histograms(width = input_slider(0.1, 2))%>%
add_axis("x",title="Histograma de la variable ")%>%
add_axis("y",title="")

```

### 2.7.1.2 Polígono de frecuencias

Consiste en unir los puntos medios de los rectángulos superiores en un histograma. El polígono (Figura 2.16) parte del eje X y regresa al eje X, simplemente marcando como origen y final una distancia de los extremos igual a la longitud de un intervalo dividida entre 2

El área encerrada entre el histograma y el eje horizontal será  $n$ , número total de datos (simplemente sumamos el área de todos los rectángulos). Si en vez de utilizar la frecuencia absoluta usamos la relativa, el área de cada rectángulo será el porcentaje de datos que hay en el mismo. El área encerrada entonces por todo el histograma y el eje horizontal será igual a 1 (es el cien por cien de los datos). Análogamente, puede comprobarse que el área encerrada entre el polígono de frecuencias y el eje horizontal también vale 1 (sólo hay que pensar que, en el polígono de frecuencias, a cada rectángulo le restamos y sumamos el área de dos triángulos, que se van compensando a lo largo de la figura).

Cargando la librería *usingR* se dibuja un polígono de frecuencias sobre un histograma con la orden `simple.feqpoly()` (Figura 2.17).

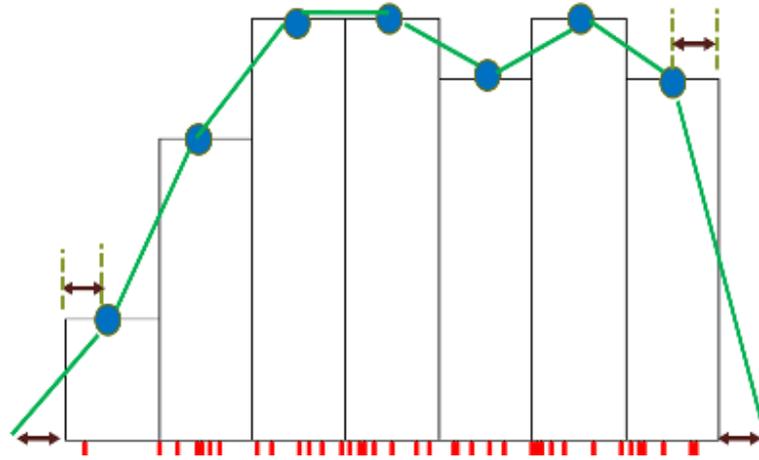


Figura 2.16: Construcción de un polígono de frecuencias.

```
# library(UsingR)
x=Datos_Titanic$edad
simple.freqpoly(x)
```

## 2.7.2 Estimación tipo núcleo de la función de densidad

El histograma y el polígono de frecuencias no dejan de ser más que estimaciones de la **función de densidad** de una **variable aleatoria continua**. Los conceptos de variable aleatoria y de función de densidad (capítulo 5) se definen a partir de la noción de probabilidad. Como una mera aproximación, diremos que la función de densidad sería el polígono de frecuencias que se construiría si dispusiésemos de un conjunto infinito de datos. En este caso, el polígono tendría la forma de una función matemática continua y derivable (no estaría formada por uniones de segmentos).

Matemáticamente, a partir de un conjunto de datos  $x_1, \dots, x_n$ , un estimador no paramétrico tipo núcleo de la función de densidad  $f$ , evaluado en un punto cualquiera (no tiene por qué ser un dato)  $x_0$  es el definido como (Parzen, 1962),

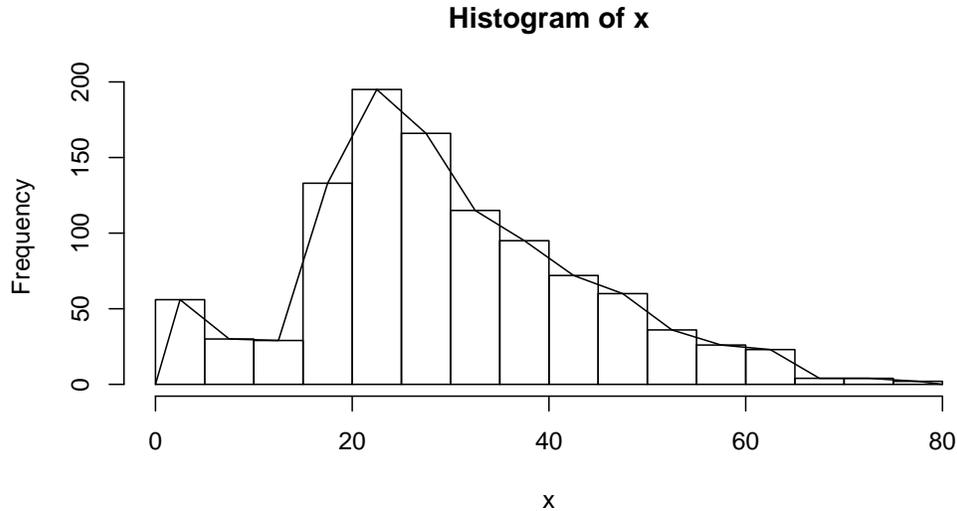


Figura 2.17: Polígono e Histograma de la variable Edad.

$$\hat{f}_h(x_0) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right). \quad (2.1)$$

$K$  es una función (llamada *núcleo* o **kernel**) continua y derivable, y  $h$  es un parámetro llamado *ventana* (**bandwidth**), que ejerce un papel equivalente al del ancho de los intervalos en el histograma (o, equivalentemente, el número de intervalos).

Veamos con un gráfico cómo funciona el estimador núcleo de la densidad. Para ello supongamos que tenemos un conjunto de 6 datos, los que aparecen representados en la gráfica 2.18:

Hemos visto que el histograma consiste en, una vez que los datos están clasificados en intervalos, se levanta un rectángulo proporcional al número de datos. Rectángulos más altos significa que hay más datos en ese intervalo (más bajos menos datos).

Dibujar el estimador núcleo de la densidad de unos datos  $x_1, \dots, x_n$  es como dibujar una función. Si queremos, por ejemplo, dibujar la función  $f(x) = x^2$  entre los valores  $-5$  y  $5$ , seleccionamos una serie de puntos (les llamamos  $x_0$ )

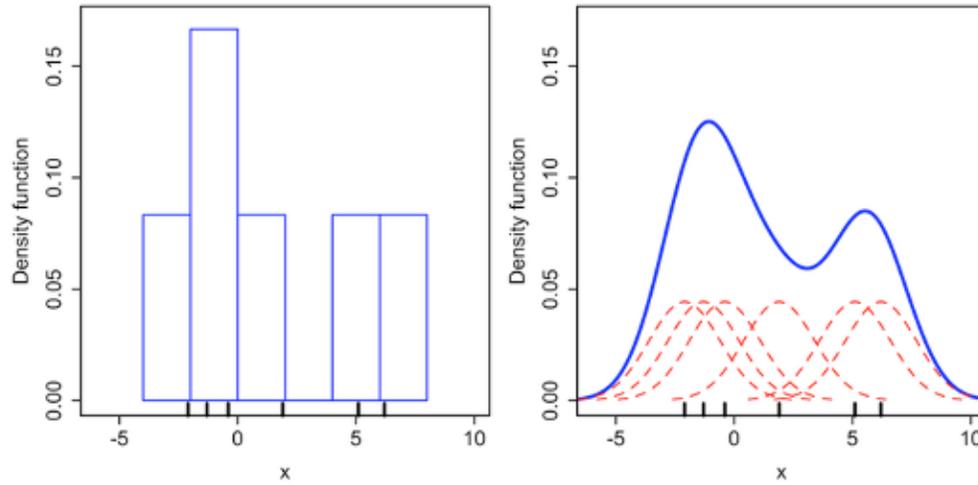


Figura 2.18: Comparativa de un histograma y un estimador núcleo de la densidad.

entre  $-5$  y  $5$ , y marcamos en el plano los puntos de coordenadas  $(x_0, f(x_0))$ . Luego únimos todos esos puntos y tenemos la gráfica. Cuántos más puntos  $x_0$ elijamos, mayor precisión tendrá el dibujo.

La forma de dibujar el estimador núcleo de la densidad es, una vez que seleccionamos esos puntos  $x_0$ , marcar los puntos de la forma  $(x_0, \hat{f}_h(x_0))$ , donde la función depende de la cantidad de datos  $x_1, \dots, x_n$  que haya alrededor del valor  $x_0$ . Cuántos más datos haya “cerca” de  $x_0$ ,  $\hat{f}_h(x_0)$  toma un valor más alto. Cuántos menos datos haya cerca de  $x_0$ ,  $\hat{f}_h(x_0)$  toma un valor más pequeño.

Fijémonos en la gráfica 2.18, donde aparece, sobre cada dato, un dibujo de una curva con forma de campana. Esa es la  $K$  que aparece en la fórmula del estimador  $\hat{f}_h$ , y es la forma en que se tiene en cuenta cada punto que hay “cerca” de  $x_0$ . Cada dato  $x_i$  cerca de  $x_0$  se pondera mediante  $K((x_0 - x_i)/h)$ .

Ese parámetro  $h$  o “ventana” juega un papel similar al del ancho o longitud de los intervalos del histograma. Lo que ocurre es que, ahora, es como si esos intervalos se movieran: para cada valor  $x_0$  donde queremos calcular  $\hat{f}_h(x_0)$  “abrimos” un intervalo de longitud  $h$  centrado en  $x_0$ . Cuántos más datos hay en ese intervalo,  $\hat{f}_h(x_0)$  es más alto, pero no es un valor directamente proporcional al número de datos de ese intervalo, sino que se ponderan en

función de la distancia, mediante esa función  $K((x_0 - x_i)/h)$ . El formato matemático de la construcción de  $\hat{f}_h(x_0)$  garantiza que la curva final que se dibuja va a ser una curva continua digamos “suave”, como la que aparece en la imagen de arriba a la derecha (color azul).

La interpretación gráfica representa la “densidad” o “distribución” del conjunto de datos. Donde hay más datos la curva crece, donde hay menos datos, la curva decrece. Es muy similar al polígono de frecuencias, pero no está formado por segmentos, sino por una línea continua.

Como función  $K$  puede elegirse una función continua cumpliendo condiciones sencillas de regularidad, por ejemplo la curva “normal” o curva de Gauss, o una función polinómica (Silverman, 1986).

Ej.

En la Figura 2.19 vemos el estimador de la función de densidad de la variable *edad de los pasajeros del Titanic*, junto con el histograma, realizado mediante los comandos básicos de R.

```
# na.omit es para no considerar valores en blanco:
X= na.omit(Datos_Titanic$edad)
hist(X, prob=TRUE)
lines(density(X), lty="dotted", lwd=2, col="red")
```

Igual que sucede en el histograma, la forma del estimador de la densidad varía en función del parámetro ventana  $h$  que se utilice. El efecto es el mismo que en el caso del histograma. Cuanto más pequeño es el valor de  $h$ , la forma del estimador de la densidad es más variable. Cuando  $h$  toma un valor muy grande, el estimador resultante adopta una forma muy suave y homogénea, pero que tampoco reflejará la realidad de los datos.

Volvemos ahora a utilizar el fichero de datos de la encuesta nacional de salud en Estados Unidos durante los años 1959 a 1962 (NHES1). Con los datos del peso y la estatura de los hombres de raza blanca realizamos estimaciones de la densidad, cambiando los valores del parámetro  $h$ , para ver como se producen diferencias en el dibujo (Figuras 2.20 y 2.21).

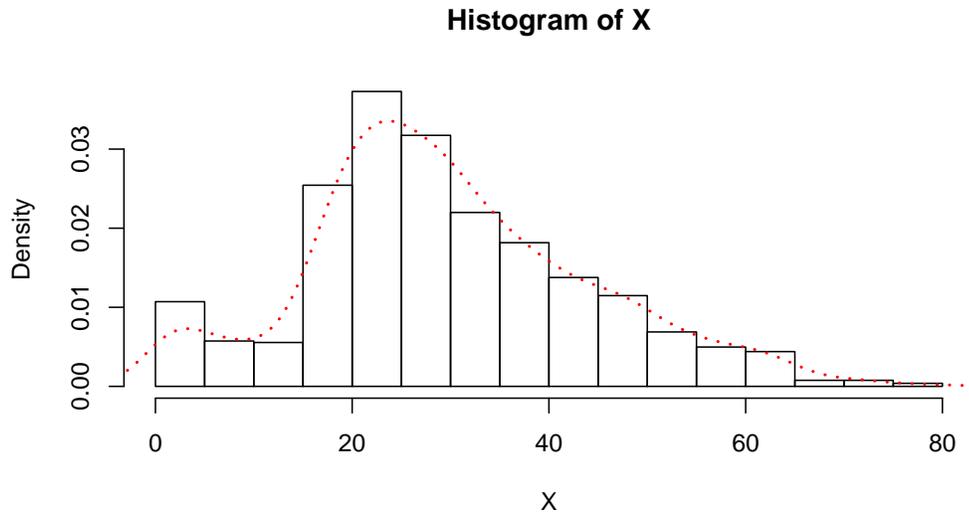


Figura 2.19: Histograma y estimador de la densidad

```
NHANES1 <-read.csv("Data/du1003.csv", header=TRUE, sep=";")
# filtramos para quedarnos con hombres de raza blanca (código 1)
Datos=NHANES1 %>%
  filter(v==1)
# VARIABLE ESTATURA
x=Datos$estatura * 0.254 # pasamos a centímetros
plot(density(x), "Estatura hombres raza blanca")
rug(x) # dibujo de puntos de la variable
lines(density(x, bw = 4), col = 2)
lines(density(x, bw = 10), col = 3)
lines(density(x, bw = 20), col = 4)
lines(density(x, bw = 40), col = 5)
legend(13.5, 0.04,
      legend = c("h=13.3", "h=8", "h=10", "h=15", "h=20"),
      col = 1:5, lty = 1)
```

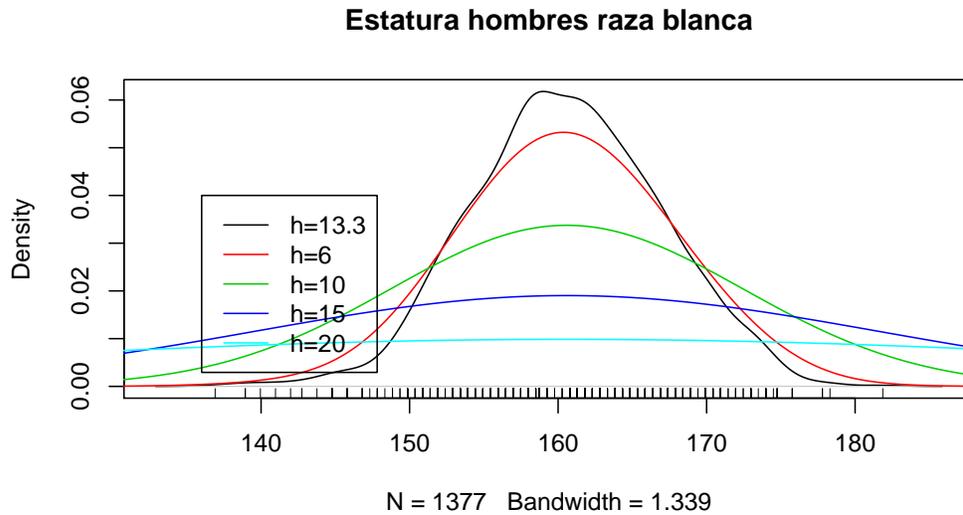


Figura 2.20: Estimaciones de la densidad de la variable Estatura

```
# VARIABLE PESO
# pasamos a kg
x=Datos$peso * 0.0453
x=na.omit(x)
plot(density(x), "Peso hombres raza blanca")
rug(x)
lines(density(x, bw = 10), col = 2)
lines(density(x, bw = 20), col = 3)
lines(density(x, bw = 40), col = 4)
lines(density(x, bw = 50), col = 5)
legend(120, 0.025,
      legend = c("h=23.2", "h=10", "h=20", "h=40", "h=50"),
      col = 1:5, lty = 1)
```

Como podemos comprobar en las dos gráficas de 2.20, el valor de  $h$  influye sustancialmente en la forma que ofrece la curva sobre los datos. De manera general, el valor que el comando `density` de R toma automáticamente suele ser el más adecuado a los datos concretos. En el primer gráfico (estatura),

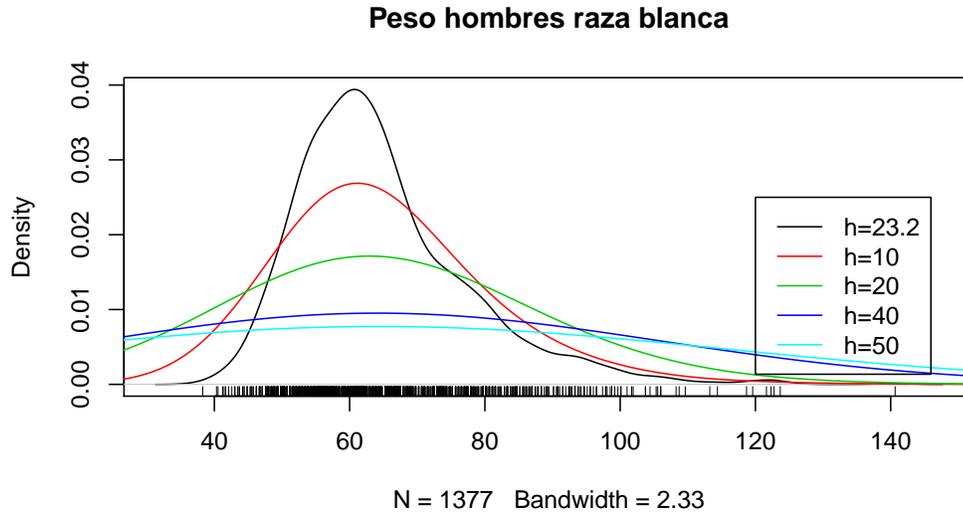


Figura 2.21: Estimaciones de la densidad de la variable Peso

los datos tienen una forma de “campana”, simétrica alrededor de la estatura media, y presentando mucha menor concentración a medida que vamos hacia valores muy bajos o muy altos (hay poca gente con estatura muy alta o estatura muy baja). En cambio, para los datos del peso, esa simetría se pierde, puesto que hay una mayor concentración de datos a la derecha (más gente con peso alto que con peso bajo). Si queremos saber el valor “óptimo” del parámetro  $h$  que nos proporciona el programa escribimos `density(x)$bw`, que, en este último caso, es 2.3299. Este valor óptimo se calcula mediante una fórmula matemática que suele proporcionar buenos resultados prácticos; es decir, habitualmente, con este valor y un tamaño aceptable de datos (50-100 o más datos), no hay que preocuparse en tomar otro, o en hacer gráficos alternativos con otros valores.

Igual que con el histograma, mediante la librería `ggvis` puede uno divertirse haciendo gráficos, cambiando automáticamente el valor del parámetro ventana (Figura 2.22), a través del siguiente código:

```
library(ggvis)
library(dplyr)
```

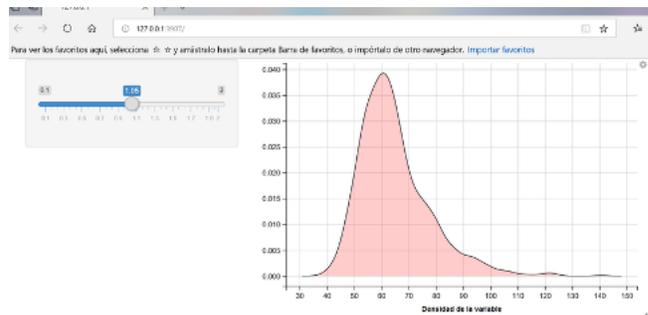


Figura 2.22: Captura de pantalla de uno de los gráficos generados mediante la librería ggvis.

```
NHANES1 <-read.csv("Data/du1003.csv", header=TRUE, sep=";")
# filtramos para quedarnos con hombres de raza blanca (código 1)
Datos=NHANES1 %>%
  filter(v==1)
x=Datos$peso * 0.0453 # peso en kg
x=na.omit(x)
datox=data.frame(x)
datox %>%
  ggvis(~x,
        fill := "red"
  )%>%
  layer_densities(adjust = input_slider(0.1, 2))%>%
  add_axis("x",title="Densidad de la variable Peso ")%>%
  add_axis("y",title="")
```

## 2.8 Medidas características de una variable

Una vez organizados los datos en su correspondiente distribución de frecuencias, procedemos a dar una serie de medidas que resuman, de la mejor forma posible, la información existente en los mismos y que, *de alguna manera*, representen a la distribución en su conjunto. El interés se centra en proporcionar un número reducido de valores que caracterizen bien, o lo mejor posible, el conjunto de datos, por grande que este sea. Hay tres tipos fundamenta-

les: medidas de **posición** (*centro*), medidas de **dispersión** (*variabilidad*) y medidas de **forma**.



A partir de ahora trabajaremos con una variable  $X$ , con valores  $x_1, \dots, x_n$ , ordenados de menor a mayor, que pueden repetirse o no, y así incluimos todos los casos descritos (variable discreta o variable continua).

El objetivo, a partir de un conjunto de datos, que puede ser muy grande, es obtener un conjunto pequeño de números que **resuman** bien el conjunto de datos. No deja de resultar curioso que, para describir la información, reduzcamos la misma. Pero veremos que, en efecto, con un conjunto pequeño de valores podemos resumir bastante bien a un conjunto muy grande.

## 2.8.1 Medidas de posición o de tendencia central

De alguna manera, estas medidas **centralizan** la información, y por ello se llaman **de tendencia central** o **promedios**. Con ellas, se pretende también facilitar la comparación entre distintas variables.

### 2.8.1.1 La media (media aritmética)

La *media* es una medida de representación o de tendencia central que se define de la siguiente manera:

$$x = \frac{x_1 + x_2 + \dots + x_n}{n}.$$



La nota media de los exámenes de una asignatura, el tiempo medio de realización de los mismos, la estatura media, la ganancia media en comisiones ilegales, etc.

En R, para calcular la media de una variable se utiliza `mean(variable)`.

```
X=c(2,3,4,4.5,4.5,5.65,7,5.8661,65)
mean(X)
```



Figura 2.23: Esta propiedad la conocen perfectamente los chavales.

## [1] 4.8

### 2.8.1.1.1 Características esenciales de la media aritmética

1. Si a partir de una variable  $X$  creamos otra  $Y = aX + b$ , entonces la media de la nueva variable es  $\bar{y} = a\bar{x} + b$

Ej.

Supongamos que han salido las notas de una asignatura, y la nota media es 2.5. Para no tener que rellenar cien mil papeles explicando por qué los alumnos son tan desastrosos, el profesor decide subir la nota a todo el mundo en 3 puntos.

Obviamente, si la nota más alta era, por ejemplo, un 5, ahora será 8. Si la nota más baja era un cero, ahora será 3, y parece lógico que la nueva media sea 5.5.

Si el profesor hubiera considerado, en vez de sumar 3 puntos, multiplicar cada nota por 3, ahora la nota más alta sería 15, y la nota media 7.5.

2. **La media aritmética se ve muy alterada por valores extremos de la variable.** Supongamos, por ejemplo, que una variable toma los valores  $X = 1, 4, 12$ . Su media es  $\bar{X} = 5.66$ . Si añadimos un nuevo valor, por ejemplo 100, ahora la media es  $\bar{X} = 29.25$ . Es decir, valores grandes de  $X$  desplazan la media hacia la derecha. Lo mismo ocurre con valores pequeños de  $X$ , que desplazan la media hacia la izquierda (Figura 2.23).

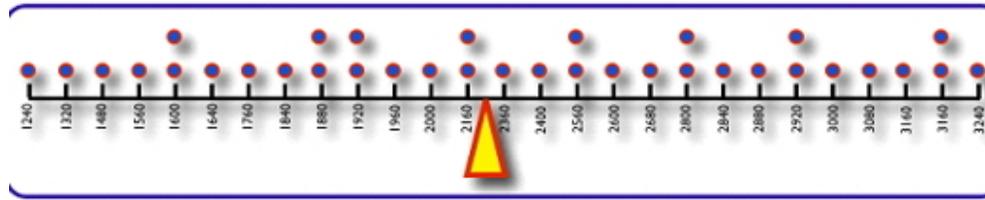


Figura 2.24: Datos=Pesos. La media aritmética es el centro de gravedad.

Esta afectación de la media aritmética por valores extremos se explica por el significado que tiene la misma desde el punto de vista de la física: la media aritmética representa el centro de gravedad de la distribución de los datos. Si los datos son pesos, la media aritmética deja igual peso a un lado que al otro, por eso se desplazaría hacia los lados con valores extremos (Figura 2.24),

### 2.8.1.2 Media recortada

La **Media recortada** (*media  $\alpha$ -trimmed*) a un nivel  $\alpha$  es la media aritmética, calculada quitando el  $\alpha$  por ciento de los datos inferiores y superiores.

```
# Para calcular la media recortada de una variable
# se utiliza mean(Variable, trim= 0 a 0.5)
X <-c(0:10, 50)
mean(X, trim = 0.10)
```

```
## [1] 5.5
```

```
#frente a
mean(X)
```

```
## [1] 875
```

### 2.8.1.3 La media ponderada

Consiste en asignar a cada valor  $x_i$  un peso  $w_i$ , que dependerá de la importancia relativa de dicho valor, bajo algún criterio. Su expresión responde a:

$$x_p = \frac{x_1w_1 + x_2w_2 + \dots + x_nw_n}{w_1 + w_2 + \dots + w_n}.$$

```
wt <-c(5, 5, 4, 1)/15
x <-c(3.7,3.3,3.5,2.8)
xm <-weighted.mean(x, wt)
xm
```

```
## [1] 3.453
```

#### 2.8.1.4 Otras medias

Son, por ejemplo, la **media geométrica** o la **media armónica**, esta última interesante puesto que es la media entre velocidades (en general, es la media resultante cuando las unidades son proporciones, como la velocidad que se mide en km/hora o metros/segundo). Ver, por ejemplo [como consumir menos](#)

#### 2.8.1.5 El concepto histórico de media

##### 2.8.1.5.1 Nominalismo, realismo, y la intervención de Occam para apoyar a los franciscanos contra el papa (siglo XIV)

La media aritmética, desde un punto de vista matemático, existe desde la época de la escuela pitagórica. Sin embargo, no tuvo ningún significado de tipo estadístico hasta el siglo XVII, cuando los astrónomos empezaron a utilizarla para dar una medida única de datos de observaciones que presentaban discrepancias. Al realizar mediciones, por ejemplo de distancias de la tierra a estrellas o entre estas, siempre solían aparecer diferencias (los aparatos de medición no eran, obviamente, los mismos que hoy en día), y se decidió, *realmente sin saberse muy bien por qué*, elegir como medida representativa de estas observaciones a la media aritmética.

Desde el punto de vista histórico, sin embargo, la consideración de la media aritmética como valor representativo de un conjunto tardó mucho en aparecer. Sobre este hecho, conviene destacar la posición nominalista de Guillermo de Occam (fraile franciscano, filósofo y lógico escolástico inglés; 1285-1347), relativo al voto de pobreza pronunciado en el siglo XIII por San Francisco de Asís y asignado como regla a la orden de los franciscanos.

En poco tiempo, los franciscanos prosperarían tanto que pronto comenzaron a dirigir numerosos monasterios y propiedades agrícolas. Sin embargo, para permitirle respetar, al menos nominalmente, su voto de pobreza, el Papa aceptó tomar a su cargo la propiedad de estos bienes, aunque cediéndoles el usufructo. Sin embargo, en el siglo XIV, el Papa, fatigado por la carga administrativa decide entregárselos a la orden, lo que obviamente enriquecería a esta última, pero avivaría las críticas formuladas, por una corriente opositora que reclamaba el retorno a la posición inicial del voto de pobreza de San Francisco.

En esta situación, Occam interviene para defender la posición de los franciscanos frente al papa. Argumenta que no es posible entregar los bienes a la orden considerada como un todo, ya que esto no es más que un nombre que designa a individuos franciscanos. Niega así la posibilidad de que existan personas colectivas distintas de las personas singulares, cuestión que traerá muchas consecuencias en el futuro. Así, al individualismo lógico del nominalismo se le asocia un individualismo moral, a su vez ligado a una concepción de la libertad del individuo, solo frente al Creador. (Desrosières, 2004).

Esta (mas que) anécdota histórica puede servir de argumento a la imposibilidad de contemplar un único valor como representativo de un conjunto, puesto que sólo podían contemplarse características individuales, nunca en grupo. Tendremos que esperar hasta varios siglos después, cuando la ley de los grandes números de Poisson (siglo XIX) permite dar un paso importante a Quetelet (Perrot, 1992). Este último, por encargo del gobierno de Francia, se ocupa de los cálculos necesarios para justificar un nuevo impuesto nacional: *el diezmo real*. Quetelet, para ello, necesita estimaciones variadas: la superficie, los rendimientos agrícolas, las cargas fiscales... En ciertos casos, dispone de varias estimaciones de una magnitud desconocida (la superficie total del país), de la que extrae una *media proporcional*. En otros, en cambio, utiliza informaciones -por ejemplo, sobre los rendimientos agrícolas-, referidas a parroquias o a distintos años. Efectúa entonces un cálculo análogo al precedente, pero no le da el nombre de media, sino de **valor común**. La operación de adición suprime las singularidades locales y permite que surja un objeto nuevo de orden más general, eliminando las contingencias no esenciales.

### 2.8.1.6 La mediana

La **mediana** es un valor que, previa ordenación, deja la mitad de las observaciones a su izquierda y la otra mitad a su derecha. Es decir, el 50% de los datos son menores o iguales que la mediana, y el otro 50% mayores o iguales a ésta.

Para su cálculo, y suponiendo que los valores están ordenados, se procede de la siguiente manera:

- Si hay un número impar de datos, la mediana es el elemento que se encuentra en el centro, es decir, el valor que ocupa el lugar  $\left(\frac{n+1}{2}\right)$ .
- Si el número de datos fuera par habría dos elementos centrales, y la mediana se obtendría como la media de ambos.

**Ej.**

Sea  $X$  la variable que toma los valores 1, 2, 3, 4, 5 (hay  $n = 5$  datos). La mediana es  $Me = 3$  (deja 2 valores a la izquierda y 2 a la derecha).

Si  $X$ , en cambio, toma los valores 1, 2, 2, 3, 4, 5, ahora hay un número par de valores. En el medio tenemos los valores 2 y 3. La mediana es  $Me = (2 + 3) / 2 = 2.5$ .

Veamos ahora un ejemplo de película. Hablamos de [Blancanieves y la leyenda del cazador \(2012\)](#), en la cual no respetan demasiado el cuento, y, como vemos en la Ilustración 2.25, salen 8 enanos en vez de 7 (parece ser que el octavo es el padre de los otros).

Consideremos que todos los enanitos miden 1 metro. La mediana de los 8 datos sería 1 m. Si añadimos a Blancanieves al grupo (1.65 m), la mediana sigue siendo 1. Si añadimos a la madrastra, sigue siendo 1. Y aún añadiendo al cazador, que es un dato bastante separado de los otros (1.92), tenemos en el conjunto total 11 datos. Por lo tanto, la mediana ocupará el lugar sexto (deja 5 datos a cada lado), y sigue siendo 1. Esta propiedad que verifica la mediana, de no dejarse afectar por datos extremos, se llama **robustez**.

A la hora de hablar, por ejemplo, del sueldo promedio o renta **media** de un país, resulta evidente que debería indicarse la medida que se utiliza. Así, un sueldo medio dado por la **mediana** sería aquel tal que el 50 por ciento de la población tendría sueldo más bajo que la mediana, y el otro 50 por ciento



Figura 2.25: Blancanieves, la reina, el cazador y los enanitos.

un sueldo más alto que la mediana. En cambio, el sueldo **media aritmética** es el valor correspondiente a sumar todos los sueldos y dividir por el número de personas. Si existe poca gente con sueldos muy altos, el sueldo media aritmética puede ser alto, pero no será representativo del conjunto total de la población. Es otro ejemplo de la **robustez** de la mediana frente a la **no robustez** de la media aritmética: si hay un porcentaje de personas no muy grande con sueldo muy elevado, provoca que la media aritmética sea también elevada.

```
x=c(2,3,1,1,0,5,5,612,3,4,5,5,4,7)
```

```
Me <median(x)
```

```
Me
```

```
## [1] 4
```

### Nota histórica:

La idea de un medio o mediano (lo cual es una descripción) que sea excelente (lo cual es una evaluación) es una de las enseñanzas más familiares de Aristóteles. “La virtud es un medio entre dos vicios, uno de exceso y otro de deficiencia”. Aristóteles explícitamente restringió la aplicación del concepto de medio porque se trata de una excelencia que contrasta con el exceso o la deficiencia. “No todos los puntos medios son medios. El rencor y el adulterio, son en sí mismo bajos, y no son bajos a causa de exceso y deficiencia” (Hacking and Bixio, 1995).



Figura 2.26: Población de g.. (gente bien vestida): valores más frecuentes de abrigos, solapas y sombreros.

El concepto de mediana apareció con posterioridad al de media aritmética. Quien primero lo utilizó fue Galton, y la generalización al concepto de percentil fue hecha por Pearson.

### 2.8.1.7 La moda

La **moda** (absoluta) de una distribución es el valor que más veces se repite (el valor con mayor frecuencia o más frecuente). Además de la moda absoluta, aquellos valores que tengan frecuencia mayor a la de los valores adyacentes serán modas relativas. Por ejemplo, si tenemos la variable  $X$  que toma los valores 2, 3, 3, 4, 6, 7, 7, 7, 10, la moda absoluta es 7, puesto que es el valor que se repite más veces. Además, el valor 3 es una moda relativa, puesto que su frecuencia es 2, superior a la de los valores 2 y 4, ambas iguales a 1.

En el caso de una variable continua se habla de intervalo modal, que a su vez puede ser absoluto o relativo.

La moda es un valor que se ve directamente al observar el diagrama de barras si la variable es discreta, o el histograma si es continua.

Ej.

Sobre el salario bruto anual en España ((Figura 2.27): El salario más frecuente, o salario moda, es todavía más bajo que el salario mediana. De nuevo hay que incidir en que, cuando se habla del salario medio, debe precisarse qué medida estadística se está utilizando. Quizá la medida más representativa en este caso es la moda (sueldo más frecuente), incluso preferible a la mediana.

Pensemos ahora en la siguiente frase, dicha en su momento por un presidente del gobierno:



Figura 2.27: Mostramos de nuevo el gráfico de los salarios Piénsese en la diferencia entre el sueldo moda, mediana y media aritmética.

“En cuestión de financiación autonómica, todas las comunidades autónomas quedarán por encima de la media”.

— José Luis Rodríguez Zapatero

(Sí, en efecto, sé lo que estás pensando: sea cual sea la media, es imposible que todos los datos estén por encima de ella. Siempre, se trate de la medida que se trate, **la media estará entre los datos**, más al medio o menos al medio, pero nunca dejará todos los datos detrás o después, sino vaya porquería de media sería. Así que, por favor, no cometa usted errores como los de ciertos señores, que convierten en veraces frases como “cuanto más inútil se es, más alto se llega”).

Otra frase famosa:

“El mayor argumento en contra de la democracia son cinco minutos de conversación con el votante medio.”

— Winston Churchill, uno de los mejores políticos del siglo XX

y fiel ejemplo de que el hombre no procede del mono, sino del perro, en este



Figura 2.28: Sir Winston Churchill, presidente inglés durante la II guerra mundial.

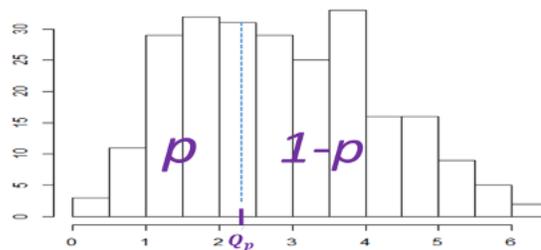


Figura 2.29: Percentil  $p$  en una variable continua.

caso del bull-dog (inglés, of course).

### 2.8.1.8 Cuantiles o percentiles

Se llama **cuantil** o **percentil**  $p$  (o de orden  $p$ ) ( $0 < p < 100$ ) a aquel valor que divide a la variable en dos partes, dejando a su izquierda (o inferiores a él) el  $p$  por ciento de los datos (a su derecha el  $100 - p$  por ciento). Por ejemplo, si  $p = 50$ , el percentil de orden 50 corresponde a la mediana.

Generalmente, mientras los percentiles van de 1 a 100, los cuantiles se toman de 0 a 1, y es entonces lo mismo el percentil 12, por ejemplo, que el cuantil 0.12.

Ej.

Los pediatras disponen de tablas de percentiles (Figura 2.31) para los niños/niñas en ciertos rangos de edad, de manera que, midiendo y pesando a un niño que acuda a su consulta, pueda saber rápidamente

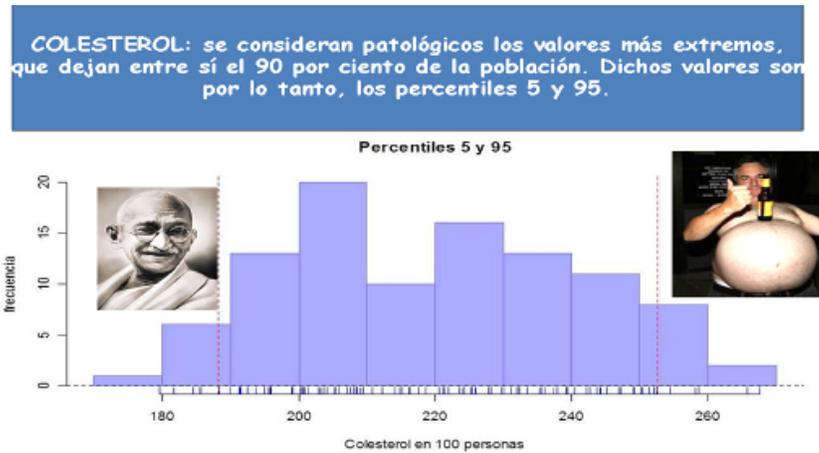


Figura 2.30: Posible uso de interes de los percentiles.

en qué percentil se encuentra, por si tuviese exceso de peso (o pesase excesivamente poco). Igualmente ocurre con la altura: se puede saber la “posición” (clasificar en percentiles no deja de ser una ordenación de los datos) del niño o niña en el conjunto de datos de su edad.

#### 2.8.1.8.1 Cuartiles

Si consideramos los percentiles 25, 50 y 75, estos 3 valores dividen a las observaciones en cuatro partes iguales, y por eso se llaman **cuartil primero**, **cuartil segundo** y **cuartil tercero**. Suelen representarse por  $Q_1$ ,  $Q_2$  y  $Q_3$  (Figura 2.32).

#### 2.8.1.8.2 Deciles

Igual que el caso anterior, si consideramos los percentiles 10, 20, 30... hasta 90 tenemos 9 valores que dividen a las observaciones en 10 partes iguales, y esos valores se llaman **deciles**. Se representan como  $D_1, D_2, \dots, D_9$ .

De forma similar puede considerarse cualquier reparto de los porcentajes, pudiendo así hablarse de *terciles*, *quintiles*, etc.



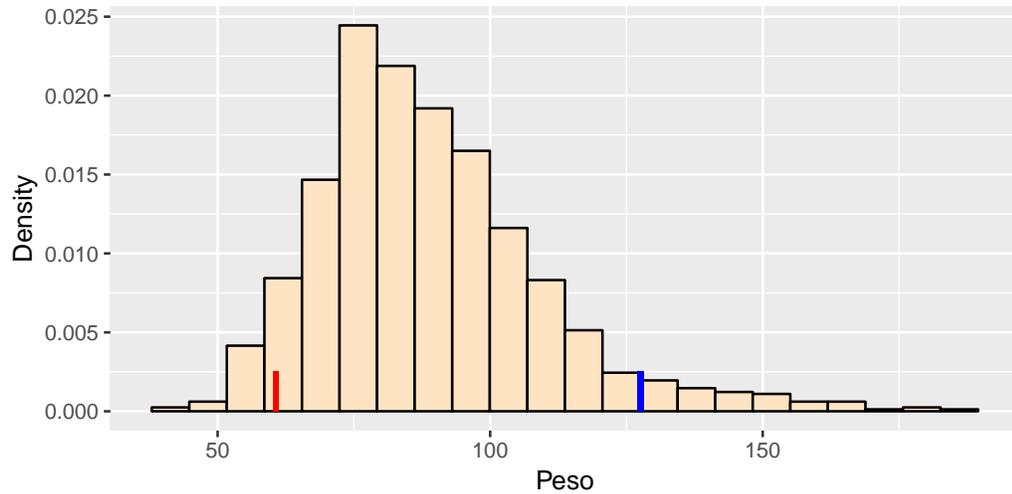


Figura 2.33: Percentiles 5 y 95 de la variables Peso.

Ej.

Vamos a hacer un par de ejemplos considerando los datos de la encuesta nacional de salud (NHANES) de Estados Unidos para el periodo 2013-14.

Primero seleccionamos la variable peso, de los hombres blancos (entre 17 y 50 años). Vemos un sumario básico de los datos, y representamos en el histograma los cuantiles 5 y 95 (el primero deja el 5 por ciento de datos a su izquierda, y el segundo a su derecha).

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
40.4	74.6	85.6	88.88	99.4	187.8

Este gráfico, como vemos, difiere en su aspecto con los tradicionales. Lo hemos realizado usando la librería `ggplot2`, que puede crear gráficos muy bonitos, pero su creación requiere un poco de cuidado. Se puede ver una guía en el libro online (gratuito) de Kieran Healy [Data visualization](#), en la página del grupo [RNA-Seq](#), en la página de [STHDA](#) y, lógicamente, en el libro del autor del paquete `ggplot2`, Hadley Wickham ([Wickham, 2016](#)).

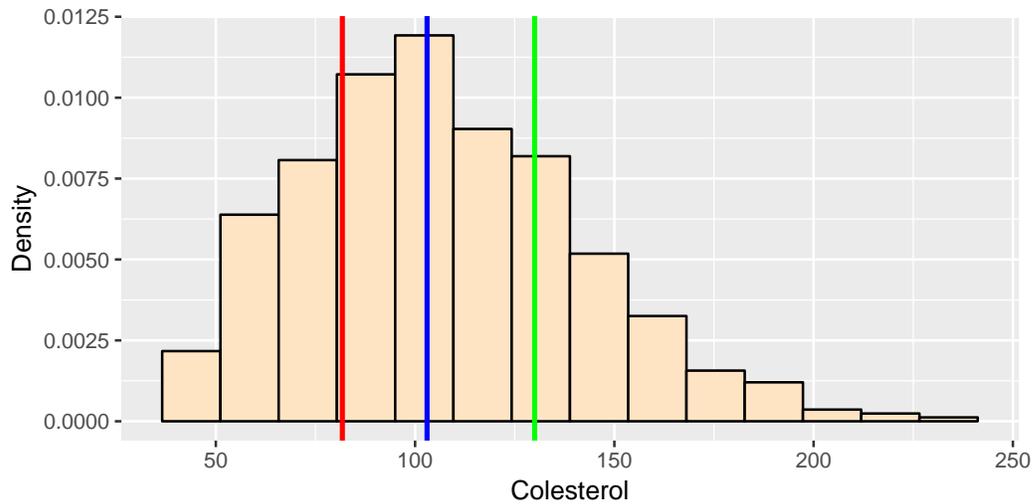


Figura 2.34: Cuartiles variable Colesterol LDL

Ahora, consideramos la variable que mide el colesterol (LDL), seleccionada en el mismo grupo de personas. Representamos sobre el histograma los cuartiles (dividen la distribución en cuatro partes iguales).

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
39	81.75	103	106.7	130	240

```
## 25%
```

```
## 81.75
```

```
## 50%
```

```
## 103
```

```
## 75%
```

```
## 130
```

## 2.8.2 Medidas de dispersión

Las medidas de tendencia central reducen la información de una muestra a un único valor, pero, en algunos casos, este valor estará más próximo a la realidad de las observaciones que en otros.

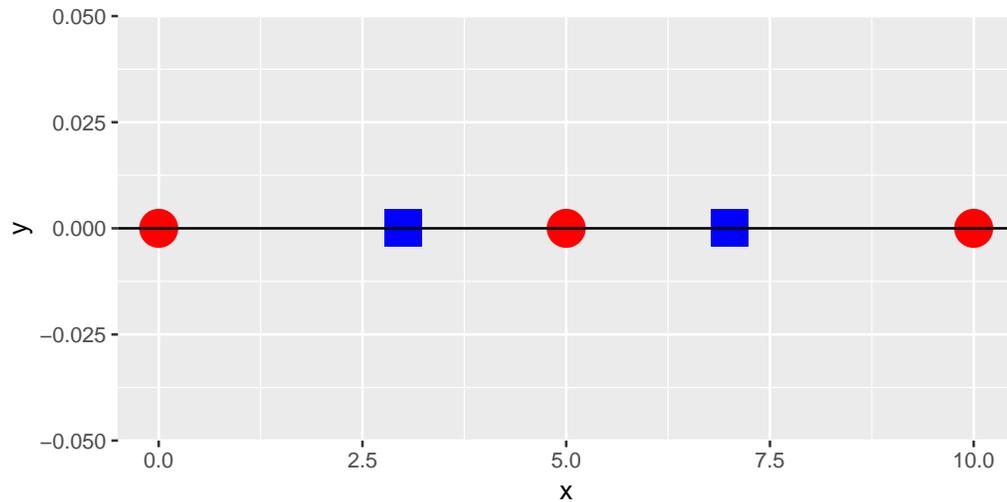


Figura 2.35: Dos variables con la misma media aritmética y diferente dispersión.

Por ejemplo, consideremos la variable  $X = 0, 5, 10$  y la variable  $Y = 3, 7$ . Enseguida podemos ver que las medias aritméticas de ambas variables son iguales ( $\bar{X} = \bar{Y} = 5$ ), pero también que la variable  $X$  está más dispersa (o menos concentrada) que la variable  $Y$ , de manera que la representatividad de  $\bar{Y}$  es mayor que la de  $\bar{X}$  (Gráfica 2.35).

A continuación se estudian una serie de medidas que, por una parte, indicarán el nivel de concentración de los datos que se están analizando y, por otra, informarán sobre la bondad de los promedios calculados como representativos del conjunto de datos. Recordemos que el objetivo de la estadística descriptiva es describir lo más sucintamente posible un conjunto de datos. A través de sus valores medios, podemos tener una gran cantidad de información simplemente con unos pocos números. Lo procedente es saber qué fiabilidad nos ofrecen esas pocas cantidades o números, es decir, cuánta **variabilidad** existe en el conjunto de datos. Si hay poca variabilidad, la información de los valores medios será muy precisa. Si, en cambio, existe mucha variabilidad, la información será menos precisa. Esta fue, concretamente, una de las muchas aportaciones a la estadística que proporcionó Francis Galton. Ya era conocido, en su época, la aportación numérica de un valor medio (en concreto la media aritmética) a la información de un grupo de datos. Galton dijo que el paso siguiente era completar esa información cuantificando la variabilidad.

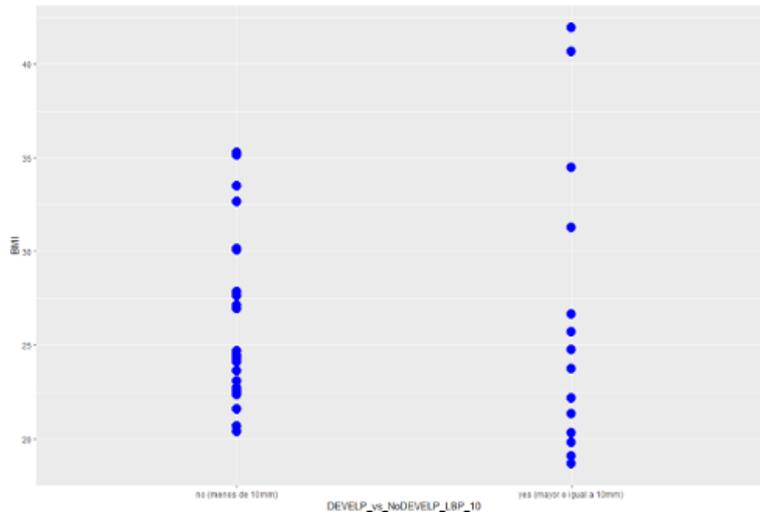


Figura 2.36: Índice de masa corporal según dolor de espalda. Variabilidad diferente según cada grupo.

### 2.8.3 La Varianza y la desviación típica

La **varianza** y su raíz cuadrada positiva, la **desviación típica**, son las medidas de dispersión más importantes, estando íntimamente ligadas a la media como medida de representación de ésta. La varianza viene dada por la expresión:

$$S^2 = \sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Se utiliza esta fórmula por ser la media aritmética de la variable cuyos valores son  $(x_i - \bar{x})^2$ .

Es decir, estamos considerando las distancias entre los datos y la media aritmética, y las promediamos.

Supongamos que tenemos los siguientes datos:  $X = -4, -2, 5, 9, 10, 11, 14$ . Calculamos su media

```
X=c(-4,-2,5,9,10,11,14)
media=mean(X)
# vamos a calcular las distancias
```

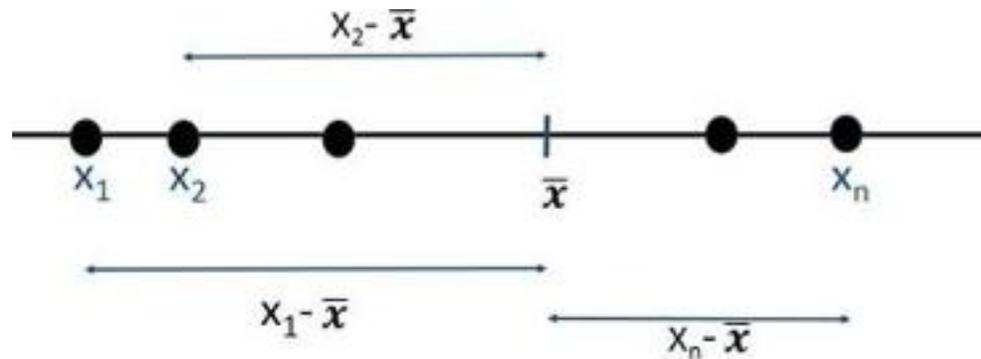


Figura 2.37: Distancias de los puntos a la media aritmética.

```
# de cada dato a la media
```

```
D=(X-media)^2
```

```
D
```

```
## [1] 102.88 6306 1.306 818 14.88 23.592
```

```
## [7] 6.735
```

```
# La varianza es la media de esta nueva variable
```

```
mean(D)
```

```
## [1] 39.8
```

```
# y la desviación típica es la raíz cuadrada
```

```
sqrt(mean(D))
```

```
## [1] 6.312
```

Si las distancias entre los datos y la media, en general, son grandes, la media de estas distancias también lo será.

Si las distancias entre los datos y la media, en general, son pequeñas, la media de las distancias también lo será.

Ahora bien, las distancias  $x_i - \bar{x}$  las elevamos al cuadrado para evitar que se compensen las distancias positivas y negativas (según que los datos  $x_i$  estén

a la izquierda de la media o a la derecha).

Realmente, puede demostrarse que, si no lo hacemos, y considerásemos el promedio  $\frac{1}{n} \sum_{i=1}^n (x_i - x)$  como la varianza, esto no serviría para nada, ya que ese promedio es siempre cero.

```
# Consideremos los mismos datos de antes
X=c(-4,-2,5,9,10,11,14)
media=mean(X)
# vamos a calcular las distancias
# sin elevar al cuadrado
D=X-media
# Si ahora calculamos la media de
# esta variable nos dará cero
mean(D)
```

```
## [1] 1.27e16
```

La desviación típica es la raíz cuadrada, con signo positivo, de la varianza.

$$\sigma = +\sqrt{\sigma^2}$$

que mide la dispersión en las mismas unidades de la variable  $X$ , puesto que la varianza mide la dispersión en las unidades de la variable, pero elevadas al cuadrado (si  $X$  son, por ejemplo, metros, la varianza está en metros al cuadrado, y la desviación típica de nuevo en metros).



Si la varianza (o desviación típica) es grande, la dispersión es grande. Análogamente, si la varianza es pequeña, la dispersión es pequeña.

La forma de calcular la varianza y la desviación típica en R es con

```
X=c(-4,-2,5,9,10,11,14)
var(X)
```

```
## [1] 4648
```

`sd(x)`

## [1] 687

Como vemos, no da exactamente lo mismo que antes. Esto es porque R considera lo que se llama **cuasi-varianza**

$$S_{n-1}^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

El valor que se calcula con **sd** es la raíz cuadrada de  $S_{n-1}^2$ . El motivo es de tipo teórico, puesto que si los datos son observaciones de una variable aleatoria (a definir en capítulos posteriores), este último valor representa mejor a la varianza teórica de la variable (no os preocupéis por entender este detalle en este preciso momento). En cualquier caso, si  $n$  es un número relativamente grande, los valores que se obtienen dividiendo entre  $n$  o  $n - 1$  son prácticamente iguales.

En las calculadoras que realizan cálculos estadísticos suele existir un botón con el símbolo  $x\sigma n$  (o algo parecido), y otro botón con el símbolo  $x\sigma (n - 1)$ , indicando que el primero calcula la varianza (o desviación típica) dividiendo por  $n$ , y el segundo dividiendo por  $n - 1$ .



Tanto la varianza como la desviación típica son siempre positivas, y valen cero sólo en el caso de que todos los valores coincidan con la media (representatividad absoluta de la media).



Desde el punto de vista físico, así como la media aritmética representa el centro de gravedad de una distribución, la varianza mide el momento de inercia alrededor de un eje que sería la media aritmética. Cuanto mayor es la varianza, mayor el momento de inercia. El ejemplo clásico de los patinadores que extienden los brazos para frenarse, aumentando el momento de inercia, sería un ejemplo de poca dispersión alrededor del eje. Los patinadores que encogen los brazos, en cambio, pueden girar mucho más rápido al ser la dispersión mucho menor (el que no sepa esto, o no lo entienda, que pruebe a patinar sobre el hielo y hacer giros, que es muy divertido).

### 2.8.4 Recorrido

Se define como la diferencia entre el mayor y el menor de los valores. Obviamente, es una medida de dispersión, y bien sencilla, además. Tiene la ventaja de que la calcula cualquiera, aunque cuando hay valores aislados en las puntas o extremos de la distribución, da una visión distorsionada de la dispersión de ésta.

$$\text{Recorrido} = \text{Max}(X) - \text{Min}(X).$$

### 2.8.5 Recorrido intercuartílico.

Viene dado por:

$$R_I = Q_3 - Q_1.$$

donde  $Q_3$  y  $Q_1$  son el tercer y primer cuartil, respectivamente.

Es una medida adecuada para el caso en que se desee que determinadas observaciones extremas no intervengan. Es, como vemos, el recorrido, pero teniendo solo en cuenta los valores “centrales” de la distribución.

Las expresiones que se acaban de ver miden la dispersión de la distribución en *términos absolutos* (vienen expresadas en las unidades de la variable, sean kilos, euros, metros cúbicos...). Por eso, se llaman **medidas de dispersión absolutas**. Se precisa definir, a partir de ellas, otras que hagan posible la comparación entre diferentes variables, y que tengan en cuenta el tamaño de las observaciones. Estas últimas se llamarán **medidas de dispersión relativas**.

Por ejemplo, si deseamos comparar los sueldos entre dos países, no solo compararemos el salario medio (incluso refiriéndonos al salario mediana o moda). También es interesante comparar la dispersión, si existe mucha variabilidad entre los salarios, o si hay mucho recorrido (diferencias máximo salario y mínimo salario). Podríamos comparar euros con dolares realizando una conversión, obviamente, para tener las dos variables en la misma unidad de medida. Pero tampoco va a ser necesario, como veremos a continuación.

### 2.8.6 Coeficiente de variación

Si una variable  $X$  viene dada en metros, su desviación típica viene también dada en metros. Si quisieramos comparar su dispersión con la de una variable  $Y$  expresada en centímetros, podríamos convertir la variable  $X$  a centímetros. Pero, si no existe manera de realizar esa conversión (supongamos una variable en metros y otra en kilogramos), no podríamos, a priori, comparar las dispersiones.

El **coeficiente de variación**, también llamado coeficiente de variación de Pearson (debido a su creador [Karl Pearson](#)) se define como el cociente entre la desviación típica y el valor absoluto de la media:

$$CV = \frac{S}{|\bar{X}|}.$$

Se trata de una medida adimensional (no tiene unidades), y permite comparar la dispersión de varias distribuciones. A mayor valor de  $CV$ , menor representatividad de la media aritmética, y viceversa. En general, se suele convenir en que valores de  $CV$  menores a 0.1 indican una alta concentración, entre 0.1 y 0.5 una concentración media, y valores superiores a 0.5 (o 1 según algunos libros) una alta dispersión y una media poco o nada representativa.

**Ejercicio 2.1** (Cálculo del coeficiente de variación). Supongamos dos variables  $X$  e  $Y$  que toman los siguientes valores:

```
X=c(0.1,0.2,0.3,0.4,0.5)
Y=c(1000.1,1000.2,1000.3,1000.4,1000.5)
# ambas variables tienen la misma
# desviación típica,
sd(X);sd(Y)
```

```
## [1] 0.158
```

```
## [1] 0.158
```

```
# sin embargo, los coeficientes
# de variación son muy diferentes
sd(X)/mean(X); sd(Y)/mean(Y)
```

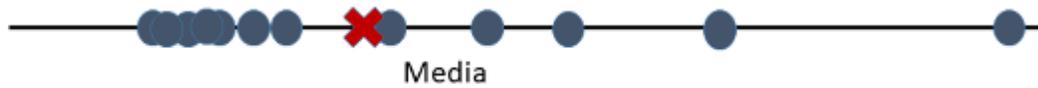


Figura 2.38: Distribución asimétrica a la derecha.

```
## [1] 0.527
```

```
## [1] 0.000158
```

```
# la media de la segunda variable
# es mucho más representativa.
```



El coeficiente de variación de Pearson es el que debe usarse para comparar la dispersión entre diferentes variables. Aquella variable con mayor coeficiente tiene mayor dispersión.

### 2.8.7 Simetría

Diremos que una distribución es **simétrica** respecto a un parámetro cuando los valores de la variable equidistantes de dicho parámetro tienen la misma frecuencia. La simetría suele referirse a la simetría respecto de la media aritmética, o respecto de la mediana.

Una distribución o variable es simétrica si, gráficamente, levantamos un eje o línea vertical sobre la media (o mediana, según el caso) y el dibujo a ambos lados de dicho eje es idéntico. Tengamos en cuenta que, si una distribución es simétrica, la media aritmética y la mediana **van a coincidir**.

Estudiar la simetría de una distribución es una manera de estudiar **la forma** de una distribución. Hemos dicho que la estadística se ocupa del estudio de poblaciones, que a su vez están compuestas de variables. La manera de estudiar las mismas es conocer sus valores medios (medidas de posición), su variabilidad (dispersión) y su forma. Dos variables (por ejemplo las estaturas de los hombres de dos naciones), cuanto más se parezcan en estos 3 conceptos, más similares serán. Como veremos en capítulos posteriores, la simetría es una propiedad que aparece en más ocasiones de lo que quizá pudiera suponerse, a priori, en muchas variables o distribuciones.

Si una distribución no es simétrica, entonces es **asimétrica**, y la asimetría puede presentarse:

- *a la derecha* (*asimetría positiva*: cola de la distribución más larga a la derecha)
- *a la izquierda* (*asimetría negativa*: cola de la distribución más larga a la izquierda).

Los *coeficientes de simetría* son valores numéricos que indican si la distribución es simétrica y, caso de no serlo, la tendencia o signo de su asimetría. Uno de los coeficientes de simetría más utilizados es el llamado **primer coeficiente de Fisher**:

$$g_1 = \frac{m_3}{S^3}$$

siendo  $m_3$  el momento respecto a la media de orden 3, es decir

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

y  $S$  la desviación típica. Como vemos, es una medida adimensional (tanto en el numerador como en el denominador las unidades de la variable aparecen elevadas al cubo, por lo que al efectuar la división no hay unidades), y esto nos permite comparar simetrías de distintas variables.

Si una distribución es simétrica,  $g_1 = 0$ .

Si  $g_1 < 0$  entonces la distribución es asimétrica negativa.

Si  $g_1 > 0$  entonces es asimétrica positiva.

Cuando la distribución es simétrica, coinciden la media y la mediana.

Al igual que ocurría con la varianza, la fórmula del coeficiente de asimetría puede variar, por motivos técnicos, dependiendo del programa estadístico que se utilice. Conviene siempre mirar el manual para tener clara la fórmula. En todo caso, los valores deben ser parecidos, y lo importante es el signo (positivo para asimetría a la derecha y negativo al contrario), que no debe depender del programa utilizado. Esto debe servir como regla para cualquier otro coeficiente de simetría que encontremos (en un libro, en google o en una papelera de algún matemático loco): el coeficiente de simetría positivo denotará asimetría a la derecha. Coeficiente negativo, asimetría a la izquierda. Un coeficiente cero (en realidad, próximo a cero puesto que la realidad muchas veces es más tozuda que la teoría) implicará simetría.

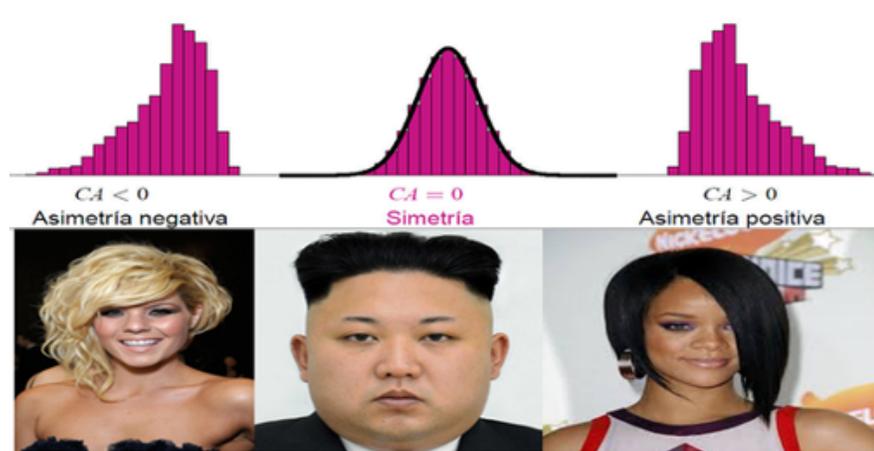


Figura 2.39: Pelo y Simetría.

Ej.

Los salarios en España, que vimos anteriormente, representan una distribución con una asimetría positiva o a la derecha.

Ej.

Con los datos de la encuesta nacional de salud americana NHES1 usados anteriormente, vamos a comparar la distribución de la estatura (Figura @ref(fig: densesth1)) y el peso (Figura 2.41) en varones (mayores de 18 años) de raza blanca.

```
# estatura en centimetros
x=Datos$estatura * 0.254
x=na.omit(x)
plot(density(x), "Estatura hombres raza blanca")
```

```
# peso en kg
x=Datos$peso * 0.0453
x=na.omit(x)
plot(density(x), "Peso hombres raza blanca")
```

Distinguimos claramente la asimetría del Peso frente a la simetría de la Es-

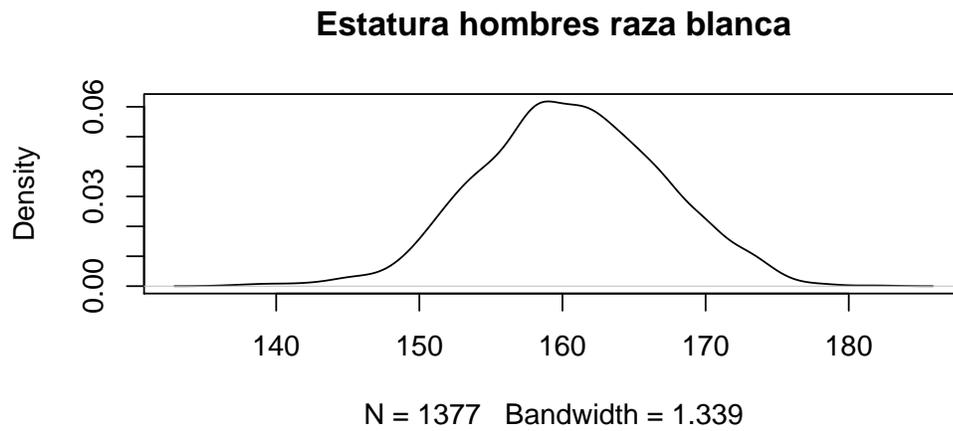


Figura 2.40: Estimación de la densidad de la variable estatura.

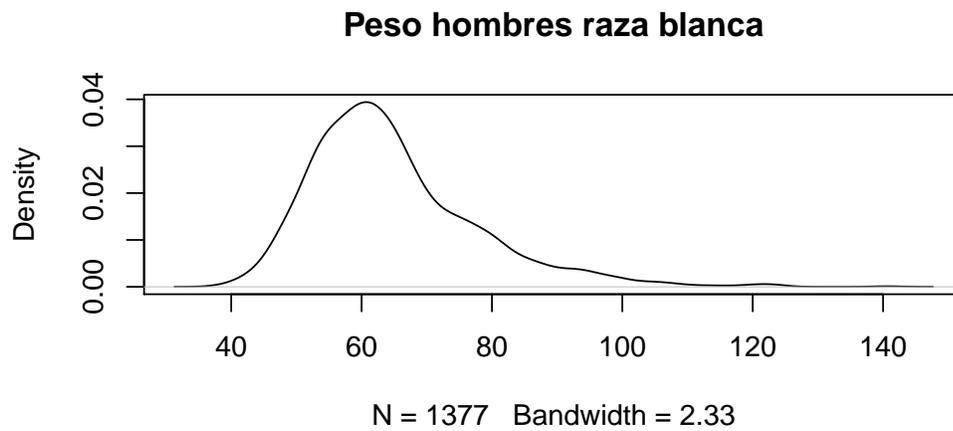


Figura 2.41: Estimación de la densidad de la variable peso.

tatura. La distribución del Peso es asimétrica a la derecha. Tengamos en cuenta que la altura es una variable antropométrica que no es susceptible de ser modificada como el peso (porque los bollitos de pan con chocolate están muy buenos, pero no aumentan la estatura).



La fórmula exacta del coeficiente de asimetría puede variar según el programa estadístico que se utilice (SPSS,R,...). En cualquier caso, el valor en sí mismo no importa demasiado, sino el signo.

### 2.8.8 Curtosis

Otra manera de estudiar la forma de una distribución es mediante la concentración existente en su “zona central” (alrededor de la media o mediana, esto es, considerando distribuciones simétricas o próximas a la simetría). La mayor o menor concentración de frecuencias alrededor de la media, en este caso, dará lugar a una distribución más o menos apuntada. El grado de apuntamiento de una distribución se calcula a través del coeficiente de **apuntamiento** o de **curtosis**, para lo cual se compara con la llamada distribución Normal o Gaussiana.

Supongamos una variable  $X$  cuya media es  $\bar{x}$  y su desviación típica es  $\sigma_x$ . La siguiente función matemática (Figura 2.42) recibe el nombre de función gaussiana (en honor a Karl Gauss): que, obviamente, es una función harto rara, pero que su dibujo es el de una campana, con eje de simetría en el valor medio  $\bar{x}$

Para estudiar el apuntamiento de una variable se puede dibujar el histograma de la misma, junto con la función 2.42. Por ejemplo, consideremos los datos de peso considerados hace un momento

```
# pasamos a kg
x=Datos$peso * 0.0453
x=na.omit(x)
# calculamos la media y desviación típica
media=mean(x)
des = sd(x)
#dibujamos el histograma
```

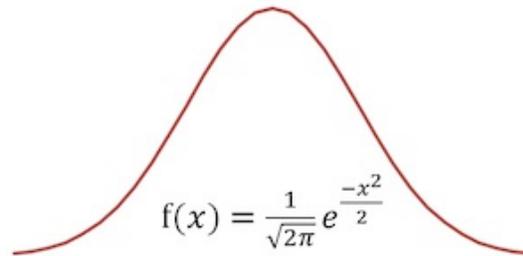
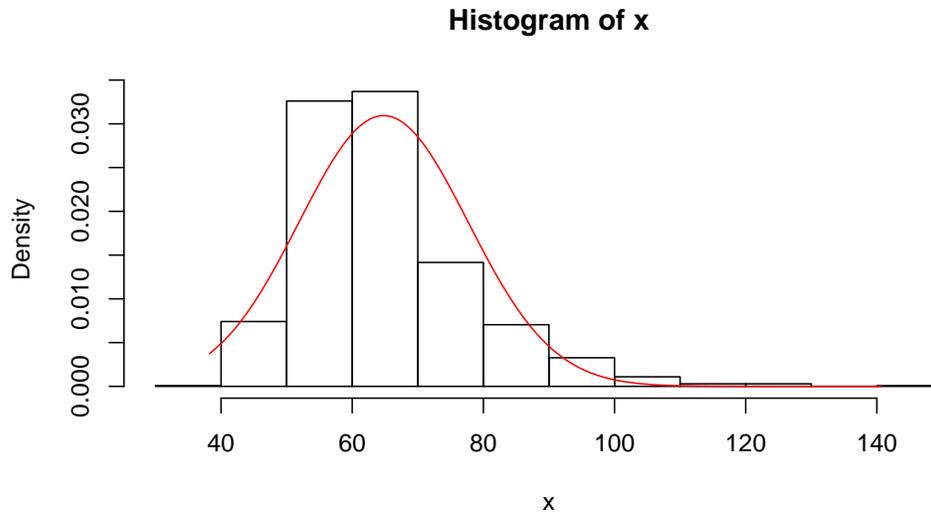
**Normal Distribution**

Figura 2.42: Curva Normal o Campana de Gauss.

```
hist(x, probability = TRUE)
# ordenamos los datos x
x0=sort(x)
# calculamos la función f en los puntos x0
y0=dnorm(x0, mean=media, sd=des)
# dibujamos la función f
lines(x0,y0, col="red")
```



Ante todo, observamos que la comparación del apuntamiento, de manera visual, no va a ser perfecta. El apuntamiento se fija en la concentración alrededor de la zona central. Como ya habíamos visto, la variable *peso* no es simétrica. Cuando dibujamos la función de densidad 2.42 utilizando la media y la desviación típica calculada a partir de los datos, vemos que la correspondiente *campana de Gauss* está “desplazada” con respecto al histograma de la distribución.

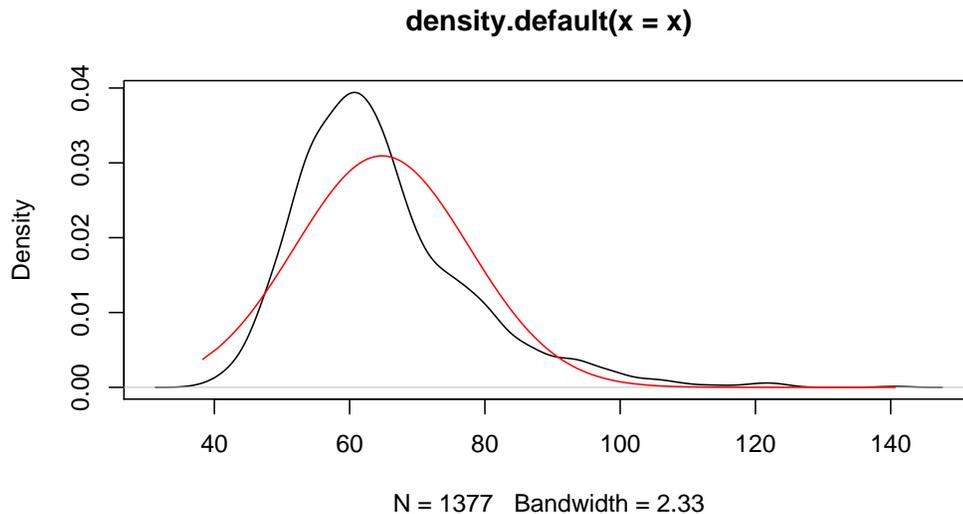
Quizá esta comparación sea más clara si realizamos una estimación de la función de densidad de los datos de la misma variable *peso*.

```
# pasamos a kg
x=Datos$peso * 0.0453
x=na.omit(x)
media=mean(x); des=sd(x)
# dibujamos la estimación de la
# función de densidad del peso
plot(density(x))
# ordenamos los datos x
x0=sort(x)
# calculamos la función f en los puntos x0
```

```

y0=dnorm(x0, mean=media, sd=des)
# dibujamos la función f
lines(x0,y0, col="red")

```



De cualquier modo, a simple vista, diríamos que esta variable tiene mayor apuntamiento o curtosis que la distribución normal, que es la que se establece de referencia. Así, se dice que la distribución es:

- *mesocúrtica* (o que la variable tiene el mismo apuntamiento que la normal),
- *platicúrtica* (la variable es menos apuntada que la normal)
- *leptocúrtica* (la variable es más apuntada que la normal).

En el caso anterior, la variable *peso* sería leptocúrtica.

De todas formas, al igual que en el caso de la simetría, es posible definir coeficientes numéricos para medir la curtosis. El llamado *coeficiente de curtosis* (o también *segundo coeficiente de Fisher*) toma la expresión

$$g_2 = \frac{m_4}{s^4} - 3,$$

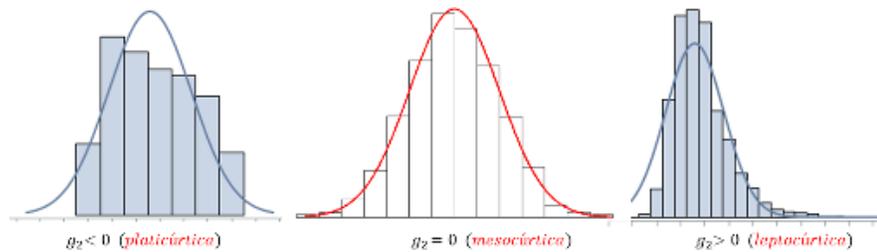


Figura 2.43: Distribuciones según la curtosis.

siendo  $s$  la desviación típica y

$$m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4.$$

Como vemos, este coeficiente  $g_2$  es adimensional, con lo cual sirve para comparar la curtosis de diferentes variables. Cuando dicho coeficiente vale 0, coincide con el de la campana de Gauss,

Básicamente, el cálculo de la curtosis de una variable se utiliza para establecer una comparación con la variable normal que tenga la misma media y desviación típica. El objetivo es analizar si podemos considerar que la variable en estudio es “aproximadamente normal”. Repetimos que la curtosis sólo tiene interés medirla en distribuciones simétricas o ligeramente asimétricas, que “puedan parecerse” a la curva Normal o campana de Gauss.



La fórmula exacta del coeficiente de curtosis puede variar según el programa estadístico que se utilice (SPSS,R,...), y puede tratarse de alguna variación del coeficiente aquí definido, por motivos esencialmente teóricos. Igual que ocurría con el caso de la simetría, más que el valor en sí mismo interesa el signo.



**Nota histórica:** la palabra KURTOSIS fue utilizada por primera vez por Karl Pearson en (Pearson, 1905). Este término está basado en el griego kyrtos o kurtos (curvado o arqueado).

Pearson introdujo los términos leptocúrtico, platicúrtico y mesocúrtico, escribiendo en

(Pearl, 1905): “Given two frequency distributions which have the same variability as measured by the standard deviation, they may be relatively more or less flat-topped than the normal curve. If more flat-topped I term them platykurtic, if less flat-topped leptokurtic, and if equally flat-topped mesokurtic”.

Otros autores, como por ejemplo (Balanda and MacGillivray, 1988), criticarían posteriormente esta definición de la curtosis de Pearson, diciendo que es “una pobre medida de la curtosis, apuntamiento o peso en las colas”, así que propusieron “definir curtosis de alguna forma como el movimiento de la masa de probabilidad desde los hombros de la distribución hasta su centro y sus colas”.



Desde un punto de vista físico, hemos visto que:

- la media aritmética representa el centro de gravedad.
- la varianza representa el momento de inercia.

La curtosis podría entenderse como una especie de “varianza de la varianza”. La curtosis viene a ser una medida de la dispersión de la variable, pero alrededor de los valores media - desviación típica y media + desviación típica.

Los Valores altos para la curtosis se dan en 2 circunstancias: cuando la masa de probabilidad se concentra alrededor de la media, o cuando la masa de probabilidad se concentra en las colas.

La definición de Pearson se utiliza en física, como un indicador de intermitencias en turbulencias.

### 2.8.9 Tipificación de una variable

Dada una variable  $X$  con media  $\bar{X}$  y desviación típica  $S_X$ , la tipificación consiste en realizar la siguiente transformación:

$$Z = \frac{X - \bar{X}}{S_X}$$

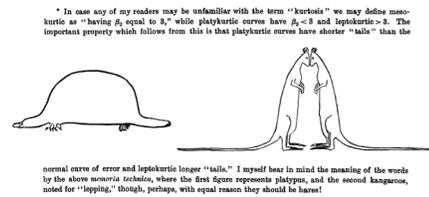


Figura 2.44: Dibujo original hecho por Student (Gosset) para acordarse de los valores de la curtosis.

A la nueva variable  $Z$  se le llama **variable estandarizada** o **tipificada**, y tiene media 0 y desviación típica 1.

Aunque un ejemplo no demuestra nada, es más fácil de entender:

```
# consideramos un conjunto cualquiera de datos
X= c(3,91,1.1, -3.5, -6 15)
#calculamos su media y desviación típica
media=mean(X)
des = sd(X)
# creamos una nueva variable donde a cada dato
# le restamos la media y dividimos
# por la desviación típica
Z= (X-media)/des
# calculamos la media y desviación típica de Z
mean(Z)
```

```
## [1] 5.8e17
```

```
sd(Z)
```

```
## [1] 1
```

Restar la media a cada dato  $x_i$  es trasladar los datos, **centrarlos**, puesto que ahora el centro de los datos es cero.

Dividir por la desviación típica es hacer un cambio de escala. Ahora la escala va a ser una unidad.

Los valores tipificados se convierten en datos adimensionales, centrados en el

cero y escala uno. Por todo lo anterior, la tipificación tiene la propiedad de hacer comparables valores individuales que pertenecen a distintas distribuciones, aún en el caso de que éstas vinieran expresadas en diferentes unidades.

**Ej.**

Izán se ha ido de Erasmus a Andorra, donde ha obtenido una nota de 25 en estadística, mientras que Yonathan ha estudiado en Corea del Norte, obteniendo una nota de 740. Para poder comparar las notas de Izán y de Yonathan, hay que saber que las notas de la clase de Izán tienen media de 20 y desviación típica de 4, mientras que en Corea del Norte la nota media es de 666 con desviación típica de 66.

Así, las puntuaciones tipificadas fueron  $\frac{25-20}{4} = 1.25$  y  $\frac{740-666}{66} = 1.12$

Una vez estandarizadas, observamos que la nota de Izan es superior a la de Yonathan.

Esta operación es la única forma que se tiene de comparar valores individuales de dos medidas diferentes. Estandarizar es una palabra muy utilizada, que significa “ajustar a un estándar” o patrón de uso común. Si hablamos de que un determinado sitio “queda muy lejos”, obviamente no sería lo mismo si nos referimos a llegar a dicho sitio en coche que andando. Para poder realizar una comparación habrá siempre que tener algún valor de referencia que nos permita realizar esa comparación. Es algo equivalente a cuando realizamos un porcentaje. Si decimos que el precio de un producto ha subido el 20 por ciento, frente al precio de otro que ha subido sólo un 3 por ciento, entendemos la diferencia, aún cuando estemos hablando de productos muy distintos entre sí.

### 2.8.10 Tamaño del efecto

El llamado **tamaño del efecto** (effect size) en una prueba estadística corresponde a una magnitud estandarizada. Si, por ejemplo, se afirma que un tratamiento disminuye el peso en 10, no sería lo mismo decir 10 kilos que 10 libras, y lo lógico sería indicar un 10 por ciento. Si se desean comparar dos tratamientos, la forma más clara de hacerlo sería a través de puntuaciones estandarizadas (como en el ejemplo que hemos puesto arriba: si comparamos dos técnicas de estudio, una podría mejorar el rendimiento más que otra, y lo lógico es indicar cual tiene un “tamaño” mayor).

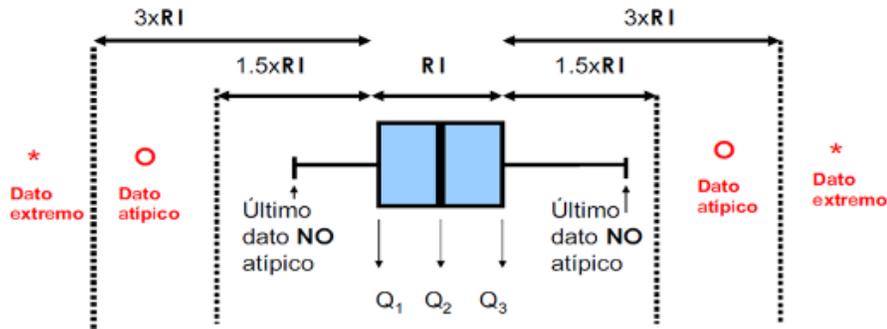


Figura 2.45: Características de un Boxplot.

### 2.8.11 Diagrama de caja (Box-Plot)

Se trata de una representación gráfica sencilla que no necesita un número elevado de valores para su construcción. Sirve para visualizar tanto la dispersión como la forma de una variable. Asimismo, es especialmente útil para comparar diferentes distribuciones de manera simultánea.

Como dice su nombre, consta de una caja, donde la misma representa el cincuenta por ciento central de la distribución (va de  $Q_1$  o primer cuartil a  $Q_3$  o tercer cuartil), y la línea situada en el interior de la caja es la mediana.

En este gráfico,  $Q_1$  recibe el nombre de *bisagra inferior* y  $Q_3$  *bisagra superior*. Los extremos inferiores y superiores de los segmentos (también llamados bigotes) delimitan lo que se denomina como valores *normales* y coinciden, respectivamente, con el mínimo y el máximo de los valores una vez excluidos los candidatos a valores **anómalos**. Los candidatos a valores anómalos se etiquetan como atípicos y coinciden con aquellas observaciones que se encuentran fuera del intervalo  $(LI, LS)$ , donde:

$$LI = Q_1 - 1.5R_I, \quad LS = Q_3 + 1.5R_I,$$

es decir, a una distancia de  $Q_1$ , por la izquierda, o de  $Q_3$ , por la derecha, superior a una vez y media el recorrido intercuartílico ( $R_I = Q_3 - Q_1$ ). En este caso se llaman atípicos de primer nivel. Cuando la distancia, por uno de los dos lados, es superior a tres recorridos intercuartílicos, el valor atípico se denomina de segundo nivel, o dato **extremo**.

Los valores atípicos de primer y segundo nivel quedan normalmente identifica-



Figura 2.46: Datos atípicos, extremos y anómalos.

dos en el diagrama de cajas por símbolos diferenciados, debiendo considerarse la revisión de los mismos (pueden corresponder a mediciones mal efectuadas), puesto que podrían corresponder a mediciones mal efectuadas, o a datos no pertenecientes a la variable que se está observando.

El diagrama de cajas revela rápidamente la simetría o asimetría de la distribución, pues será asimétrica a la derecha si desde la mediana la caja y los bigotes son más largos a la derecha que a la izquierda (será asimétrica a la izquierda si ocurre lo contrario).



El diagrama de caja y bigotes (box-plot and whisker) fue desarrollado por John Tukey (1915-2000), excelente estadístico americano famoso, entre otras cosas, por el desarrollo del *análisis exploratorio de datos*, una forma de tratamiento de datos que surgió a partir de su libro del mismo nombre (Tukey, 1977). Inventor de otro gráfico muy conocido: *el diagrama de tallo y hojas*, Tukey adquirió pronto fama de genio. El gráfico de caja y bigotes fue bautizado así, por lo visto, porque las líneas que sobresalen de la caja hacia los lados le recordaron los bigotes de un gato (recordemos que “whisker” significa bigotes de gato). A día de hoy, el boxplot, que es un gráfico muy sencillo de realizar y muy

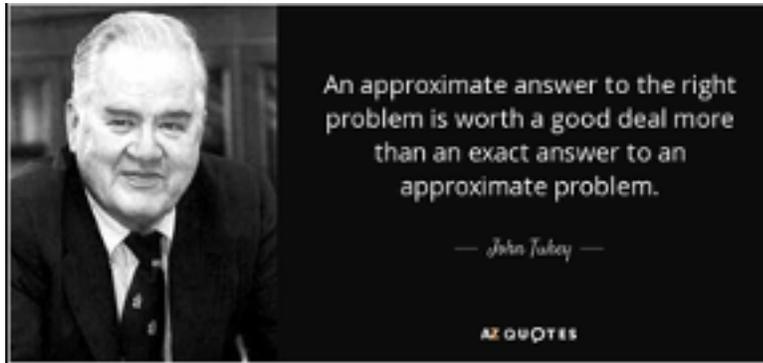


Figura 2.47: El inefable John Tukey.

práctico para la comparación rápida de múltiples variables, es uno de los diagramas más utilizados en estadística. Curiosamente, el término que se generalizó fue el de boxplot; el de whisker ni siquiera aparece en muchos textos.

John Tukey fue también famoso por su cordialidad y frases ocurrentes. En la siguiente imagen aparece una de las más conocidas:

---

**Ejercicio 2.2** (Edades en el cine). vamos a analizar las variables *Edad que tenía un actor/actriz al recibir el Oscar al mejor actor/actriz* (datos actualizados hasta 2017).

Tenemos 4 variables con edades (oscar a mejor actor/actriz principal/secundario). Para los oscars a mejor actriz y mejor actor principal tenemos 89 datos (un dato por año desde 1929 hasta 2017). Los oscar a mejor actor o actriz de reparto comenzaron a darse varios años más tarde.

Primero hacemos un sumario de las medidas estadísticas básicas:

```
premios_oscar <-read_excel("Data/premiososcar.xlsx")
#cargamos los datos y los preparamos
X<-(premios_oscar$actor_principal)
```

```

Y<-(premios_oscar$actriz_principal)
Z<-(premios_oscar$actor_secundario)
W<-(premios_oscar$actriz_secundaria)
pander(summary(X))

```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
29	38	42	44.07	49	76

```
pander(summary(Y))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
21	28	33	36.03	41	80

```
pander(summary(Z))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
21	41	47	50.68	60	83	8

```
pander(summary(W))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
11	31	39	40.69	47	78	8

y, a continuación, un boxplot de las 4 variables (Figura 2.48).

```
boxplot(X, Y, Z, W)
```

Si queremos hacer un boxplot bastante “profesional”, podemos usar la libre-

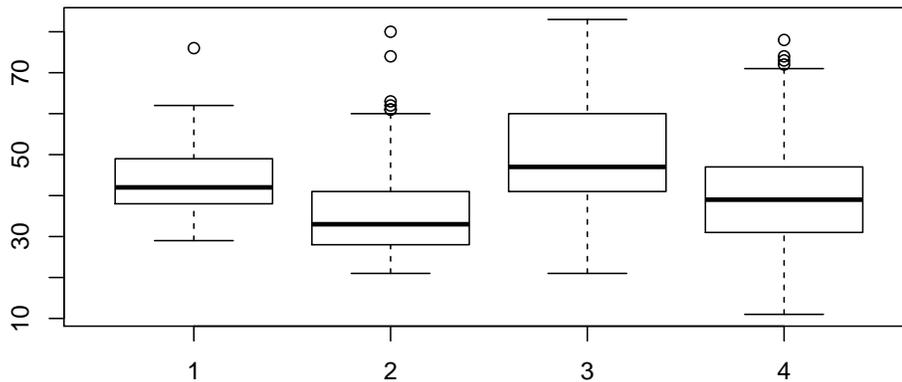


Figura 2.48: Boxplot básico

ria `plotly`, que, básicamente, nos permite crear un gráfico interactivo y que permite hacer zoom, mediante la instrucción `plotly(g)` donde `g` es un gráfico que podemos crear mediante la librería `ggplot2`. De todas formas, está claro que la interactividad solo se puede ver en el formato de página web de este libro. Además, tengase en cuenta que utiliza bastante memoria.

```

# actor principal
# actriz principal
# actor secundario
# actriz secundaria

# vector con todas las edades
edad <- c(X, Y, Z, W)

# creamos una estructura
# de dos vectores, edad y
# tipo (actor principal, actriz principal...)
dt <- data.frame(edad, #c)

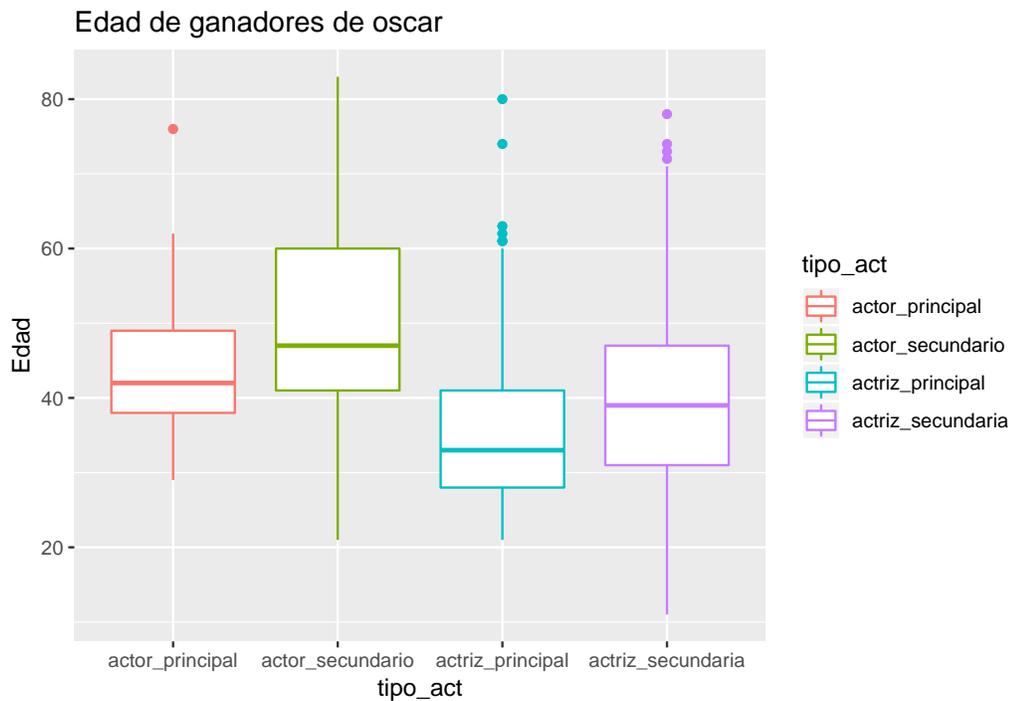
tipo_act <- c(rep("actor_principal", length(X)),
             rep("actriz_principal", length(Y)),

```

```

    rep("actor_secundario", length(Z)),
    rep("actriz_secundaria", length(W))
edad<-c(X,Y,Z,W)
dt<data.frame(edad,tipo_act)
library(plotly)
library(ggplot2)
p <-ggplot()+
  geom_boxplot(data=dt,
               aes(x=tipo_act, y=edad, color=tipo_act))+
  ggtitle("Edad de ganadores de oscar")+
  labs(y="Edad")
p

```



Vemos que, para la variable *Edad que tenía el actor que ganó el oscar al mejor actor principal* la caja se mueve entre los 38 y los 50 años. La mediana es 42 años. La edad media está alrededor de los 44 años (aunque este valor

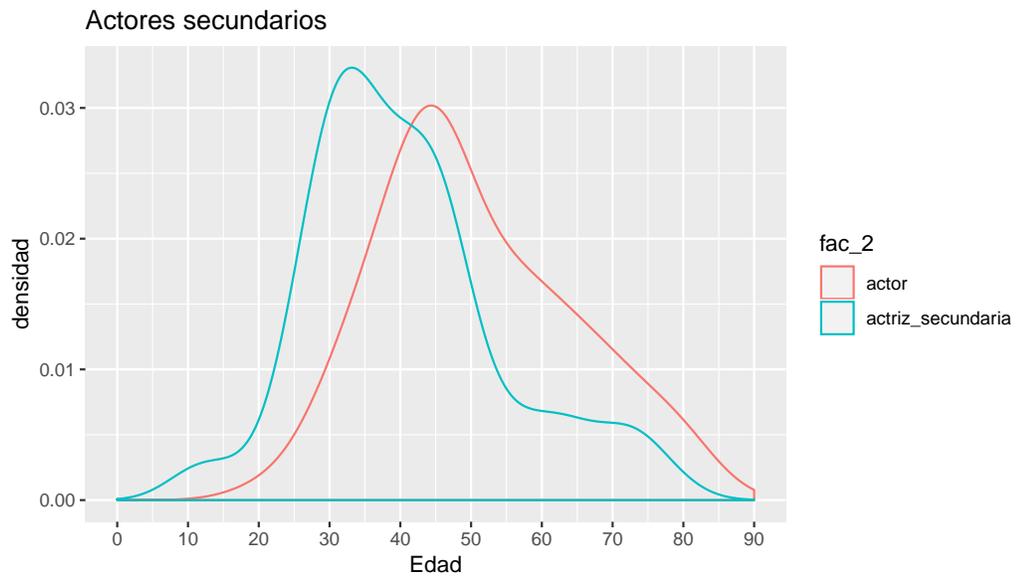
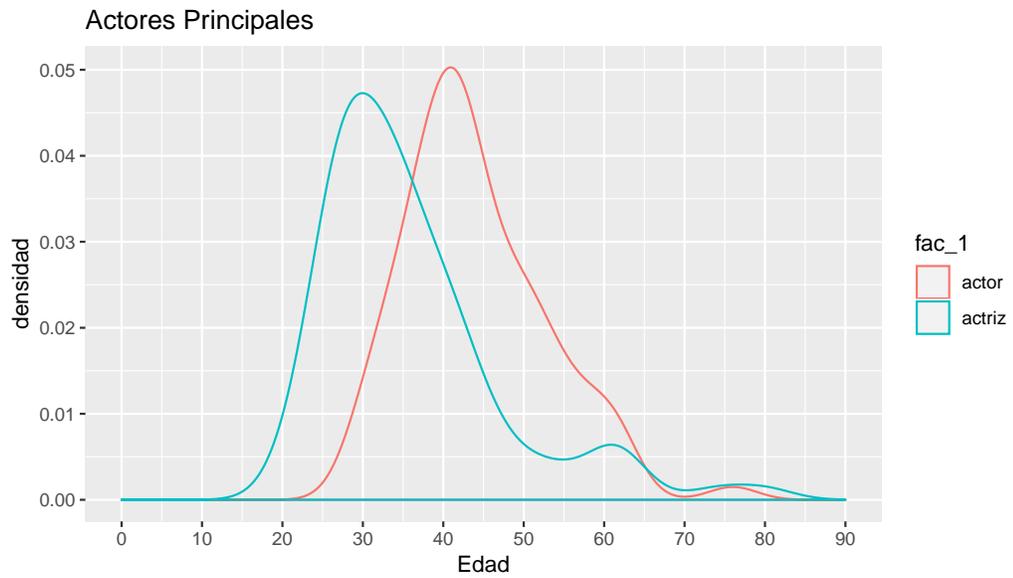
no aparece en el diagrama).

Las edades varían desde un mínimo de 29 años (Adrien Brody en “El pianista” en el 2002) y un máximo de 76 años (Henry Fonda en 1982). Con respecto a las mujeres que ganaron el Oscar a la mejor actriz principal, la caja del gráfico correspondiente (el tercero, de color azul) nos indica que el 50 por ciento central de las edades es bastante menor. Esto quiere decir que las mujeres ganadoras del oscar, por lo general, lo ganan siendo más jóvenes (comparadas con los hombres). La mediana de las edades es 33 y la media aritmética ronda los 36 años. Si bien los extremos son 21 años (Marlee Matlin en 1987) y 80 (Jessica Tandy en 1990), vemos que en esta segunda variable hay unos cuantos valores atípicos (esos puntos aislados que se separan de la línea del gráfico), que significa que están alejados de la tónica general de los datos.

Así, de un solo vistazo, podemos comparar estas dos variables y también las correspondientes a las edades de ganadores de oscar a mejor actor y actriz secundario/a, donde comprobamos que la tendencia de mayor edad (en general) en los hombres frente a las mujeres se mantiene.

El diagrama de cajas es muy útil también para estudiar la asimetría de una variable y poder comparar entre varias. En este ejemplo, vemos que la distribución que parece más simétrica es la correspondiente a las edades de los actores secundarios (la longitud desde la mediana hacia abajo y hacia arriba son similares). Mientras tanto, las otras tres variables presentan asimetría positiva (son más largas desde la mediana hacia la derecha).

Presentamos ahora en dos gráficos la estimación de la densidad de las variables edad actor/actriz principal y edad actor/actriz secundario. Estas gráficas nos permiten comparar la distribución de las edades con mucha claridad.



Cuanto más alta es la densidad en un cierto rango de edad, significa que hay más datos en el mismo. Observamos que, en ambos casos, las edades de las mujeres van “retrasadas” con respecto a los hombres, es decir, que los hombres tienden a ganar el Oscar con mayor edad que las mujeres. Estos

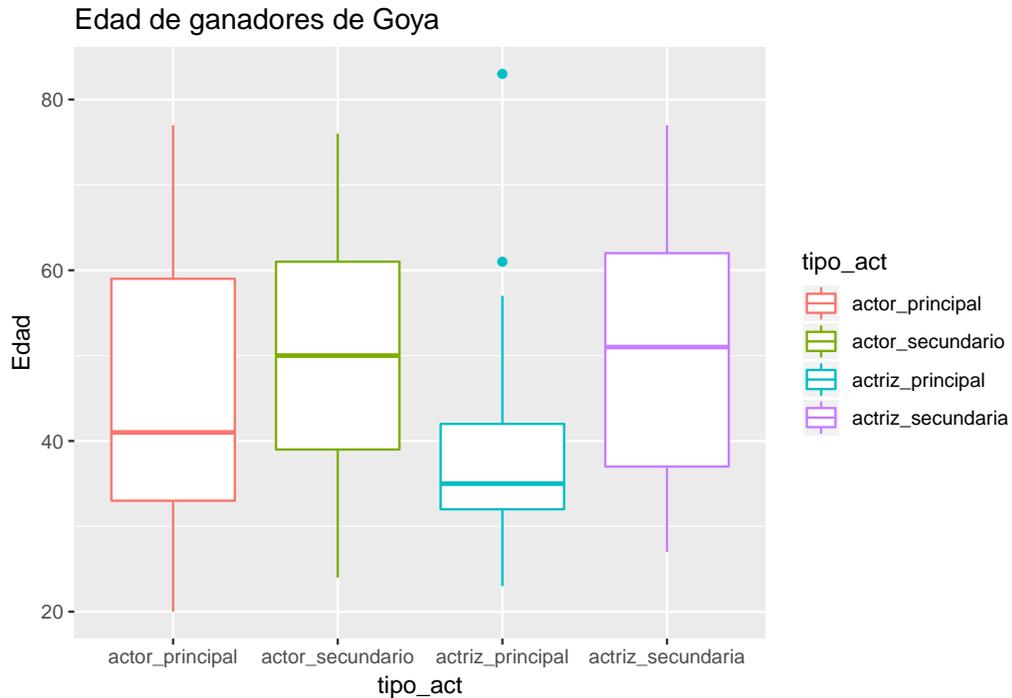
resultados pueden servir para corroborar que los papeles más interesantes para las mujeres (y que les permiten ser nominadas y por tanto ganar un Oscar) se le ofrecen a las mujeres más jóvenes, mientras que en los hombres parece que existen papeles atractivos en todas las edades.

---

*Estudiemos ahora la situación en España*

Aquí no tenemos Oscar, tenemos los Goya, pero desde hace menos tiempo. En concreto, desde 1986. Vamos a realizar exactamente el mismo análisis de antes. El gráfico de box-plot es

```
# Leemos los datos directamente de excel
# con la libreria readxl
library(readxl)
premios_goya <- read_excel("Data/premiosgoya.xlsx")
#cargamos los datos y los preparamos
X<na.omit(premios_goya$edad_actor_principal)
Y<na.omit(premios_goya$edad_actriz_principal)
Z<na.omit(premios_goya$edad_actor_reparto)
W<na.omit(premios_goya$edad_actriz_reparto)
tipo_act<e(rep("actor_principal", length(X)),
           rep("actriz_principal", length(Y)),
           rep("actor_secundario", length(Z)),
           rep("actriz_secundaria", length(W)))
edad<e(X,Y,Z,W)
dt<data.frame(edad, tipo_act)
library(plotly)
library(ggplot2)
ggplot()+geom_boxplot(data=dt,
                      aes(x=tipo_act, y=edad, color=tipo_act))+
  ggtitle("Edad de ganadores de Goya")+
  labs(y="Edad")
```



Fijémonos que la diferencia de gráficos es acusada entre actores y actrices cuando nos fijamos en el premio a actriz/actor principal, pero no en cambio al comparar edades en premios a actor/actriz de reparto. En el caso de las edades de las actrices que ganaron el Goya a la mejor actriz principal, ya aparecen como puntos atípicos (aislados del resto) una actriz de 61 años (Amparo Rivelles, que ganó en la primera ceremonia de los Goya), y Rafaela Aparicio que tenía 83 años cuando lo ganó.

Las gráficas corroboran la impresión del diagrama de cajas. Las distribuciones son muy parecidas en el caso de las edades de los premios secundarios, pero muy diferente en el caso de los premios a papeles principales (resaltemos como hay mucha concentración de datos en el rango de 20 a 40 años).

La diferencia de nuestro cine con el de Hollywood se centra, como vemos, en lo que respecto a los papeles no principales. Sin embargo, con respecto a los papeles principales, las actrices españolas se quejan de igual forma de que no existen tantos buenos papeles principales para actrices “maduras”, como sí existen para sus compañeros masculinos.

# Capítulo 3

## Descriptiva dos-dimensional

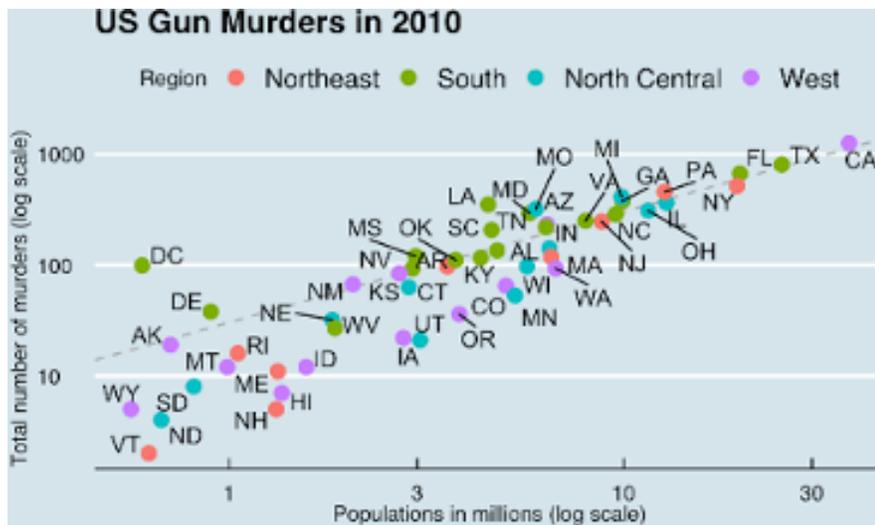


Figura 3.1: Muertes por arma de fuego en USA en 2010.

En el capítulo anterior se ha considerado el estudio de un único carácter o variable. Sin embargo, es frecuente estudiar conjuntamente varios caracteres y preguntarse si existe o no algún tipo de relación entre ellos. Comenzamos con la organización y resumen de la información, siguiendo un esquema análogo al establecido en el capítulo anterior, para ver posteriormente como puede estudiarse la relación, si existe, entre los dos caracteres o variables.



Figura 3.2: Individuo de una población: Ejemplo de caracteres que puede interesar medir.

Habitualmente, cuando se tiene una muestra de una población, si es posible y no resulta especialmente costoso, se observan varias variables y/o atributos (cuanta más información se pueda obtener, siempre es mejor). Por ejemplo, en un análisis de sangre, se miden varios niveles; si se elige una muestra de personas para una encuesta, normalmente se le realizan varias preguntas, etc. A continuación, resulta lógico cuestionarse sobre qué grado de relación o dependencia existe entre las variables analizadas: ¿hay relación entre la estatura y el peso?, ¿y entre el nivel de renta y el consumo?; ¿entre la estadística y las mentiras?, ¿la política y la corrupción?, ¿la ruina y los impuestos?, ¿la pesadez de una película y el número de premios en festivales europeos?...

### 3.1 Distribución conjunta de caracteres

En general, un investigador está interesado en más de una característica de los individuos de una población.

Si el investigador recoge datos de  $d$  características, tiene  $d$  variables unidimensionales. En muchos casos, le interesará estudiar varias de ellas conjuntamente para ver las posibles relaciones de interdependencia entre unas y otras. En el caso particular de dos variables, al par  $(X, Y)$  se le llama *variable estadística bidimensional*.

A partir de ahora, trabajaremos con un par de variables  $(X, Y)$  de interés, medidas en  $n$  individuos.

$X \ Y$	$y_1$	$\dots$	$y_j$	$\dots$	$y_s$	
$x_1$	$n_{11}$	$\dots$	$n_{1j}$	$\dots$	$n_{1s}$	$n_{1\cdot}$
$\vdots$	$\vdots$	$\backslash$	$\vdots$	$\backslash$	$\vdots$	$\vdots$
$x_i$	$n_{i1}$	$\dots$	$n_{ij}$	$\dots$	$n_{is}$	$n_{i\cdot}$
$\vdots$	$\vdots$	$\backslash$	$\vdots$	$\backslash$	$\vdots$	$\vdots$
$x_r$	$n_{r1}$	$\dots$	$n_{rj}$	$\dots$	$n_{rs}$	$n_{r\cdot}$
	$n_{\cdot 1}$	$\dots$	$n_{\cdot j}$		$n_{\cdot s}$	$n$

Figura 3.3: Tabla de doble entrada para una variable bidimensional.

## 3.2 Estructura de la información

En variables unidimensionales, los datos podían organizarse en tablas de frecuencias (bien de valores, en el caso de variables **discretas**, o bien de intervalos, en el caso de variables **continuas**)

Cuando leemos un fichero de datos, como por ejemplo el del *Titanic*

```
Titanic <- read_excel("Data/PasajerosTitanic.xlsx")
```

disponemos de una serie de variables ordenadas en columnas. A partir de una tabla de datos podemos seleccionar diferentes variables unidimensionales (**discretas**, **continuas** o **atributos (categorías)**). Si se seleccionan dos variables discretas (o atributos), en ocasiones se presenta la información construyendo una tabla de doble entrada con los valores de cada variable y las frecuencias ( $n_{(i,j)}$  es la frecuencia absoluta del valor  $(x_i, y_j)$ , o número de veces que se repite dicho par de valores).

En la última columna y en la última fila de la tabla suelen presentarse las sumas por fila y columna, respectivamente, siendo el total de datos  $n$ .

Si  $X$  e  $Y$  son variables numéricas, la tabla suele llamarse **tabla de correlación**.

Si  $X$  e  $Y$  son atributos o categorías, la tabla suele llamarse **tabla de contingencia**.

Ej.

Tabla de contingencia de supervivientes del fichero de datos *Titanic*. La variable “X” sería un atributo (sobrevivió o no) y la  $Y$  otro (clase en la que viajaba).

```
t1<-table(Titanic$sobrevivio, Titanic$clase)
t2<addmargins(t1)
pander(t2)
```

	1st	2nd	3rd	Sum
no	123	158	528	809
yes	200	119	181	500
Sum	323	277	709	1309

### 3.3 Representaciones gráficas

Si ambas variables  $X$  e  $Y$  son cuantitativas discretas o atributos, se puede realizar un diagrama de barras simultaneo de ambas variables.

Ej.

Gráfico de barras a partir de la tabla de contingencia anterior.

```
dt<data.frame(Titanic)
ggplot(dt, aes(x=clase))+
  geom_bar(aes(fill= sobrevivio), position="dodge")
```

En el caso de que las variables  $X$  e  $Y$  sean continuas, el gráfico más utilizado, sobre todo por su sencilla interpretación, es el **diagrama de dispersión**, que comentamos más adelante.

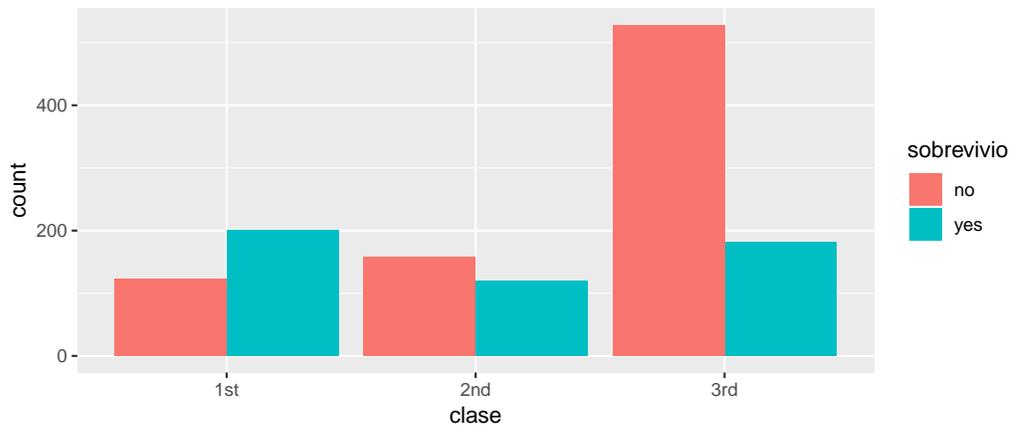


Figura 3.4: Diagrama de barras para la Tabla de contingencia de los datos del Titanic

### 3.4 Dependencia e Independencia estadística.

Entre dos variables o atributos puede ocurrir que exista *independencia*, *dependencia funcional* (matemática) o *dependencia estadística*. Por ejemplo:

- *Altura y renta mensual* de los trabajadores de Mercamona (independencia)
- *Minutos de llamadas y factura mensual* (dependencia funcional o matemática) de un número de teléfono.
- *Política y Corrupción, Peso y Altura* de un grupo de hombres de una misma población (dependencia estadística).

Intuitivamente, dos variables (atributos) son independientes si los valores que toma una variable no están influenciados por los de la otra.

### 3.5 Dependencia estadística entre variables

Cuando consideramos dos variables  $X$  e  $Y$  midiendo dos características diferentes en una misma población, nos interesa estudiar si existe relación entre las mismas. Una primera forma de verlo es mediante el **diagrama de dispersión** o **nube de puntos**.

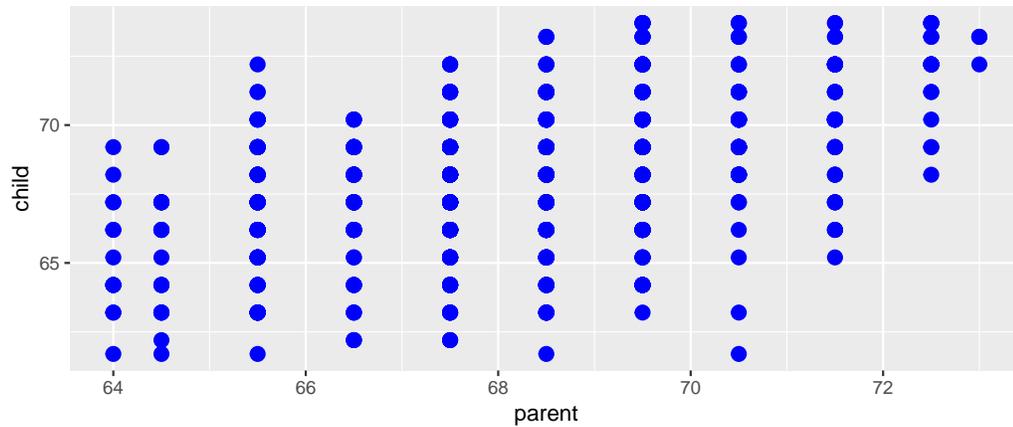


Figura 3.5: Datos originales estudiados por Galton.

### 3.5.1 Diagrama de dispersión.

Consiste en representar, en un plano, las coordenadas  $(x_i, y_j)$  correspondiente a los valores de  $X$  e  $Y$ , respectivamente.

Ej.

Dentro del paquete *HistData* tenemos el conjunto de datos *galton*. Este conjunto de datos (tomados en 1886 por Francis Galton) corresponde a las estaturas de 928 adultos varones (variable  $Y$ ) y la estatura media del padre y la madre (variable  $X$ ).

```
data(Galton)
G1<data.frame(Galton)
ggplot()+
  geom_point(data=G1,
             aes(x=parent, y=child),
             size=3, color="blue")
```

Mediante el gráfico 3.5, podemos ver que, a medida que el padre y la madre tienen estatura mayor (eje  $X$ : media de las alturas) los hijos tienen, de manera general, también mayor estatura. Se observa, por lo tanto, un tipo de relación *lineal creciente*. Este gráfico se ha realizado mediante la librería *ggplot2* que, obviamente, hay que llamarlo primero, mediante *library(ggplot2)*.

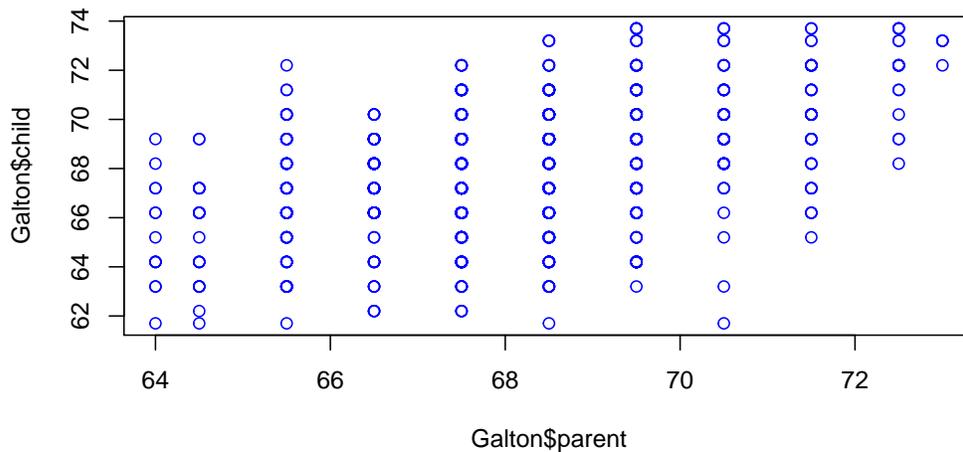


Figura 3.6: Diagrama de dispersión mediante plot.

La resolución gráfica es mucho mejor que con los comandos básicos de R, que serían, simplemente:

```
data(Galton)
plot(Galton$parent, Galton$child, col="blue")
```

Como vemos, el gráfico 3.6 es igual que el anterior, pero más cutrillo.

Ej.

Dentro de la librería *HSAUR* podemos encontrar el conjunto de datos *water*, correspondiente a la mortalidad y dureza del agua en 61 ciudades de Inglaterra y Gales durante los años 1958 a 1964. Se trata de 61 observaciones de 4 variables: *location* (un factor con niveles *North* y *South*), *town* con el nombre de la ciudad, *X* correspondiente a la **mortalidad** (mortalidad anual media por cada 100.000 varones) e *Y* correspondiente a la **dureza del agua** (concentración de calcio en partes por millón) (Hand et al., 1993).

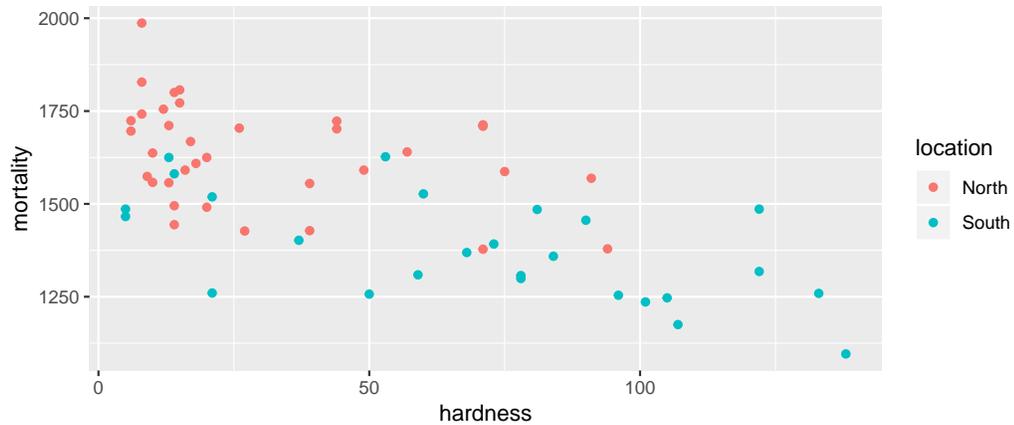


Figura 3.7: Datos relacionando dureza del agua y mortalidad.

```
library(HSAUR)
data("water", package = "HSAUR")
mort<data.frame(water)
ggplot()+
  geom_point( data=mort,
              aes(x=hardness, y=mortality, colour=location))
```

La estructura de la nube de puntos deja entrever que, a mayor dureza del agua (mayor concentración de calcio), la mortalidad disminuye. Es, al contrario que antes, una relación *lineal decreciente*.

### 3.5.2 Covarianza. Correlación lineal.

Sir Francis Galton fue geógrafo, meteorólogo, antropólogo y estadístico, e introdujo por primera vez el término **correlación**, de la siguiente forma:

“La longitud del cúbito [el antebrazo] está correlacionada con la estatura, ya que un cúbito largo implica en general un hombre alto. Si la correlación entre ellas es muy próxima, un cubito muy largo implicaría una gran estatura; en cambio, si no lo es tanto, un cúbito muy largo estaría asociado en promedio con una estatura simplemente alta, pero no muy alta; mientras que, si la

correlación fuese nula, un cubito muy largo no estaría asociado con ninguna estatura en particular y, por consiguiente, en promedio, con la mediocridad.”

Este discurso original se enmarcó dentro de su trabajo como científico, interesado en estimar la estatura de las poblaciones, a partir de los huesos encontrados en excavaciones. Antes de definir el concepto matemático de correlación, necesitamos estudiar la llamada **covarianza**, o medida de variabilidad conjunta entre dos variables.

La covarianza de una variable bidimensional  $(X, Y)$  que toma valores  $\{(x_i, y_i)\}_{i=1}^n$  viene dada por la expresión

$$S_{XY} = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

$S_{XY}$  es una medida simétrica (porque es igual a  $S_{YX}$ ) y se puede leer como la suma de los productos de las desviaciones de  $X$  por las desviaciones de  $Y$  con respecto a sus respectivas medias. Fijémonos en la gráfica siguiente, donde se considera como centro el punto de coordenadas las medias aritméticas,  $(\bar{x}, \bar{y})$ , que se llama **centro de gravedad de la nube de puntos**. Alrededor de él consideramos cuatro cuadrantes:

- por primer cuadrante entendemos los puntos  $(x_i, y_i)$  donde  $x_i > \bar{x}$  e  $y_i > \bar{y}$ .
- Por tercer cuadrante los puntos  $(x_i, y_i)$  donde  $x_i < \bar{x}$  e  $y_i < \bar{y}$ .
- Por segundo cuadrante los puntos  $(x_i, y_i)$  donde  $x_i > \bar{x}$  e  $y_i < \bar{y}$ .
- por cuarto cuadrante los puntos  $(x_i, y_i)$  donde  $x_i < \bar{x}$  e  $y_i > \bar{y}$ .

Si el signo de la desviación de  $X$  coincide con la de  $Y$ , como ocurre en el primer y tercer cuadrante, se genera un sumando positivo en la fórmula de la covarianza; y cuando el signo es distinto -segundo y cuarto cuadrante- la aportación a la covarianza es negativa.

Entonces, en el primer y tercer cuadrante  $(x_i - \bar{x}) \cdot (y_i - \bar{y})$  siempre es un número mayor o igual a cero. En cambio, en el segundo y cuarto cuadrante, el producto  $(x_i - \bar{x}) \cdot (y_i - \bar{y})$  siempre es un número menor o igual a cero.

El punto de coordenadas

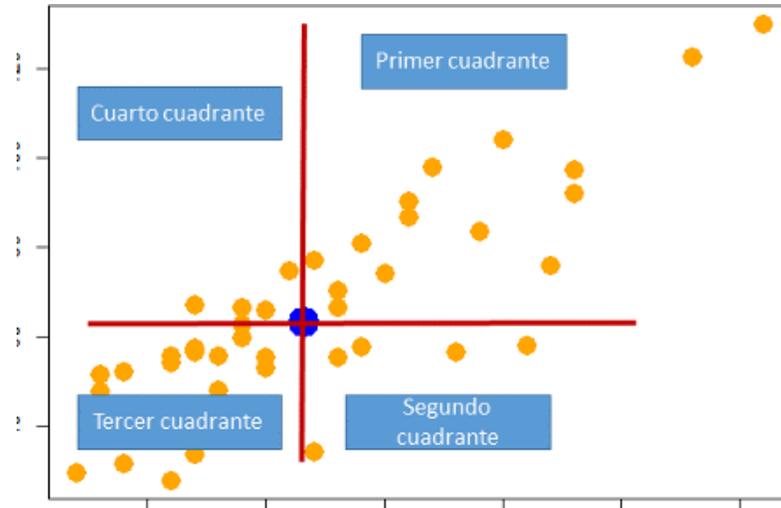


Figura 3.8: Cuadrantes relativos al centro de gravedad de la nube de puntos.

Por lo tanto, la concentración de valores en los distintos cuadrantes determina el signo y el valor de  $S_{XY}$ . La covarianza mide, pues, la cantidad de relación lineal entre las variables y el sentido de esta, de la forma:

- $S_{XY} > 0$ , relación lineal positiva (si crece una variable, la otra también)
- $S_{XY} < 0$ , relación lineal negativa (si crece una variable, la otra decrece).
- $S_{XY} = 0$ , no hay relación lineal entre las variables.

De la simple observación de la nube de puntos podemos deducir que existe una relación lineal positiva entre las dos variables.

Calculamos ahora la covarianza entre los dos conjuntos de datos anteriores.

```
data(Galton)
Gl<data.frame(Galton)
cov(Gl$parent,Gl$child)
```

```
## [1] 2.06
```

```
data("water", package = "HSAUR")
mort<data.frame(water)
```

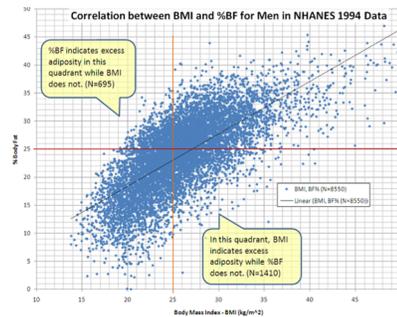


Figura 3.9: Ejemplo de diagrama de dispersión relacionando el índice de masa corporal y la grasa corporal.

```
cov(mort$hardness, mort$mortality)
```

```
## [1] 40
```

Observamos que la covarianza en el primer conjunto de datos (Galton) es positiva (relación lineal creciente, puesto que ambas variables crecen en el mismo sentido), y en el segundo grupo de datos es decreciente (una variable crece, la otra decrece).

De todas formas, el número que resulte es de difícil interpretación, puesto que depende de las unidades en que vengan expresadas las variables. Es por ello que, en vez de trabajar con la covarianza, se trabaja con el llamado **coeficiente de correlación lineal** (o coeficiente de correlación lineal de Pearson).

### 3.5.3 Coeficiente de Correlación lineal (Pearson)

El coeficiente de correlación lineal o coeficiente de correlación de Pearson viene dado por

$$\rho = r = \frac{S_{XY}}{S_X \cdot S_Y},$$

y verifica

- es una medida adimensional,
- siempre toma valores en el intervalo  $[-1, 1]$  y

- tiene el signo de  $S_{XY}$ .

Veamos cuánto vale la correlación para cada uno de los conjuntos de datos anteriores:

```
data(Galton)
Gl<data.frame(Galton)
cor(Gl$parent,Gl$child)
```

```
## [1] 0.458
```

```
data("water", package = "HSAUR")
mort<data.frame(water)
cor(mort$hardness, mort$mortality)
```

```
## [1] 0.648
```

**De manera general, tenemos:**

- cuando la relación lineal entre  $X$  e  $Y$  es exacta y directa, es decir, todos los puntos se encuentran sobre una recta con pendiente positiva,  $r$  vale 1.
- cuando la relación lineal es exacta e inversa, es decir, todos los puntos se encuentran sobre una recta con pendiente negativa,  $r$  vale -1.
- los valores intermedios ( $0 < r < 1$  o  $-1 < r < 0$ ) darán lugar a que los puntos se aproximen más o menos a una recta que pasa por el medio de los mismos.
- cuando no hay relación lineal,  $r$  vale 0. Este último caso se llama **inco-relación**, y se dice que las variables están incorreladas.

En la gráfica que sigue aparecen representados diferentes casos:

El coeficiente de correlación lineal  $r$  también verifica que es invariante (salvo el signo) ante cambios de escala y origen, es decir, si construimos 2 nuevas variables  $Z = a + bX$ ,  $W = c + dY$ , entonces:  $r_{(Z,W)} = r_{(X,Y)}$ .

Tengamos en cuenta que la correlación mide la variación conjunta de las variables  $X$  e  $Y$ . Si a una o a las dos variables les sumamos una constante,

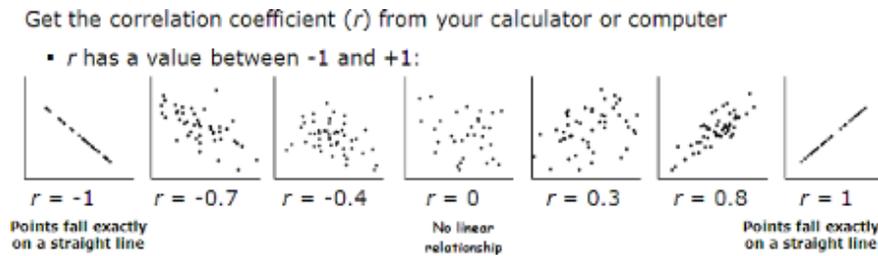


Figura 3.10: Posibles valores para el coeficiente de correlación.

la variación conjunta entre las nuevas variables es la misma de antes. Si multiplicamos alguna (o las dos) por una constante, la variabilidad (varianza o desviación típica) se verá multiplicada por esa constante (salvo el signo).

Al cuadrado de  $r$  se le llama **coeficiente de determinación**, y se le denota por  $R^2$ . Lógicamente, se verifica

$$0 \leq R^2 \leq 1,$$

y, cuánto más próximo esté  $R^2$  a 1, mayor es la relación lineal existente entre las variables, y menor cuanto más próximo esté  $R^2$  a cero. Muchas veces se multiplica esta medida por cien, y se habla entonces de un valor de  $R^2$  de, por ejemplo, el 80 por ciento, indicando la cantidad de relación lineal entre las variables. La relación lineal perfecta será, por lo tanto, del cien por cien. La relación será menor cuanto más se aproxime al cero por cien.

Calculemos la correlación en los dos conjuntos de datos usados anteriormente (*Galton* y *water*), y la elevamos al cuadrado para obtener el coeficiente de determinación:

```
data(Galton)
Gl<-data.frame(Galton)
cor(Gl$parent,Gl$child)^2
```

```
## [1] 0.2105
```

```
data("water", package = "HSAUR")
mort<data.frame(water)
cor(mort$hardness, mort$mortality)^2
```

```
## [1] 0.428
```

Así, podemos decir que la relación entre la altura de los hijos y la altura media de sus padres es lineal (positiva o creciente) en un 21.04 por ciento, y que la relación entre la mortalidad y la dureza del agua es lineal (negativa o decreciente) en un 42.88 por ciento.

Se concluye este apartado indicando que la independencia implica incorrelación, pero el recíproco no siempre es cierto (recordemos que la incorrelación se refiere a ausencia de relación **lineal**. Dos variables pueden estar relacionadas muy fuertemente mediante una función diferente a una lineal, y la incorrelación puede ser cero). Un ejemplo lo tenemos en la gráfica 3.11, donde las variables X e Y están relacionadas totalmente por una función matemática, pero si se calcula el valor de la correlación lineal da cero.

```
# 100 puntos desde -1 a 1
X=seq(-1,1,length=100)
# ecuación de una semi-circunferencia
Y=sqrt(1-X^2)
xy=data.frame(X,Y)
ggplot(data=xy, aes(x=X,y=Y))+
  geom_point(color="darkorange", size=3)
```

```
cor(X,Y)
```

```
## [1] 3.18e-16
```

Y en la gráfica 3.12 dibujamos otros dos ejemplos de sendos conjuntos de datos, en los que observamos que la correlación puede tomar valores muy próximos a cero y las nubes de puntos indican, en cambio, una fuerte relación entre las variables (y demostrar además que hay gente que no parece tener mucho qué hacer, pero, al menos, no se divierte haciendo maldades) (Matejka and Fitzmaurice, 2017)

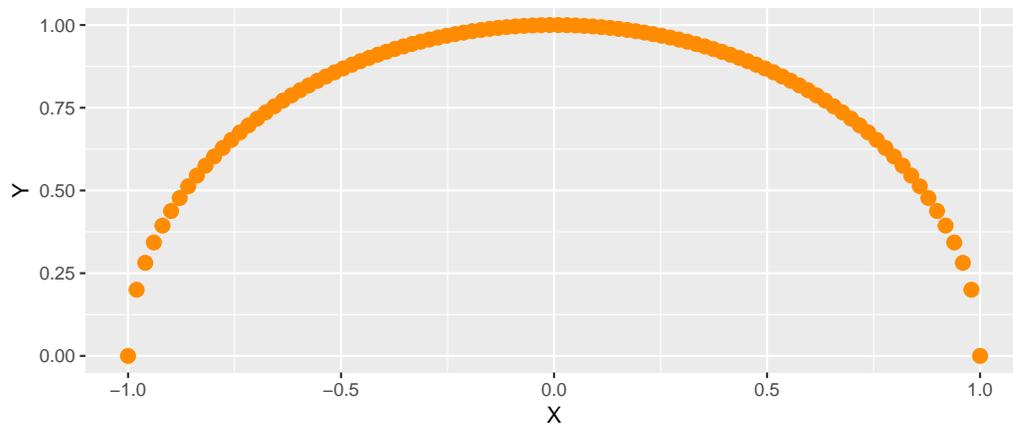


Figura 3.11: Correlación cero y relación matemática perfecta.

```
## [1] 0.089
```

```
## [1] 0.0647
```

### 3.5.4 Ajuste y regresión bidimensional

Considerada una serie estadística  $(x_1, y_1), \dots, (x_n, y_n)$ , procedente de una distribución  $(X, Y)$ , el problema que se denomina *ajuste de una nube de puntos* o *regresión bidimensional* consiste en encontrar alguna relación que exprese los valores de una variable en función de los de la otra. La cuestión será elegir la mejor función, y determinar los parámetros (fórmula) de la misma. Esta relación podrá ser utilizada, posteriormente, para hacer predicciones aproximadas; por ejemplo, para hacer previsiones de ventas a corto o medio plazo, estimar el volumen de cosecha en función de la lluvia caída, la estatura de los hijos en función de la de los padres, etc...

La elección de esa función particular que mejor se adapte a las variables es el primer problema que habrá que solventar. En un principio, la observación de la nube de puntos puede dar una idea de la evolución de los valores de la variable dependiente (a partir de ahora  $Y$ ) en función de los de la independiente ( $X$ ).

A través del dibujo de la nube de puntos podemos, en muchas ocasiones, intuir que existirán mejores funciones que la línea recta (que es la más sencilla de

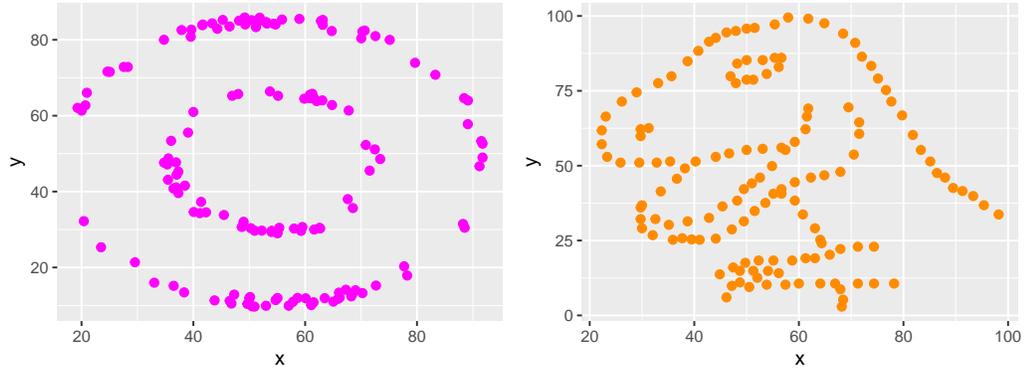


Figura 3.12: Correlación lineal cero para un donuts y un dinosaurio.

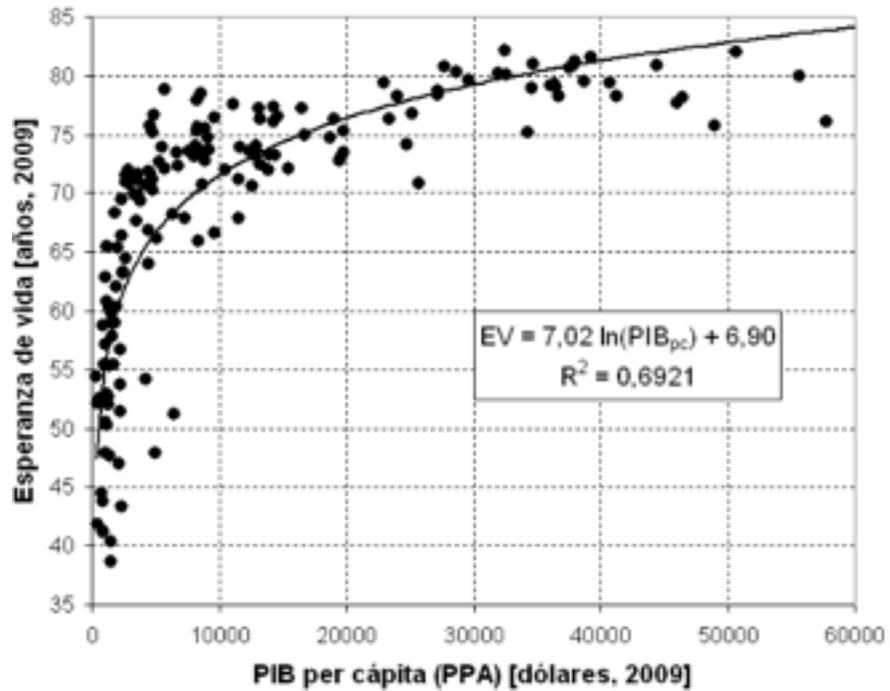


Figura 3.13: Ejemplo de una relación de tipo logarítmico entre variables.

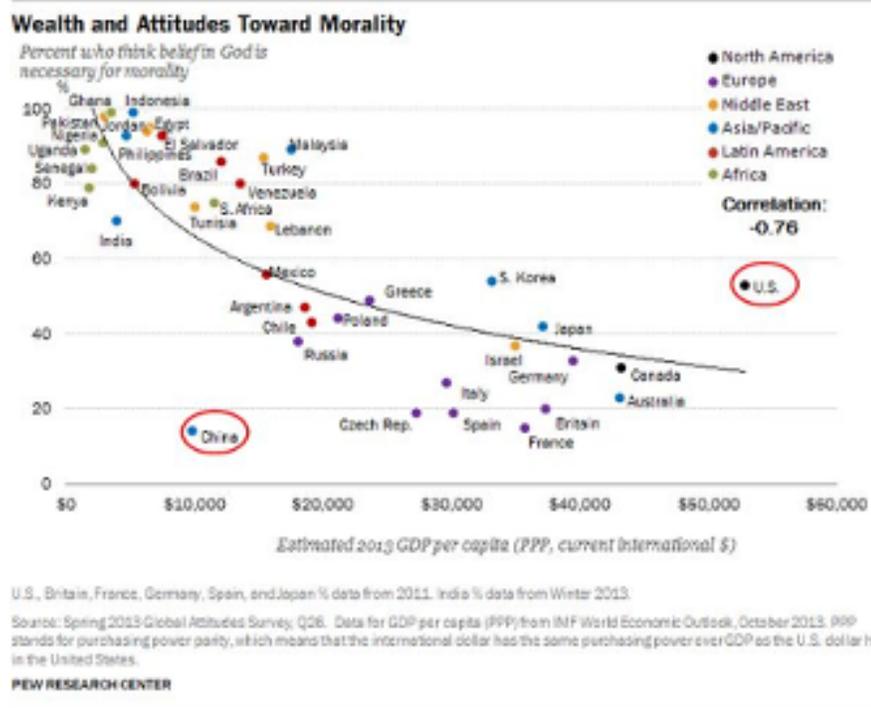


Figura 3.14: Ejemplo de una relación de tipo exponencial.

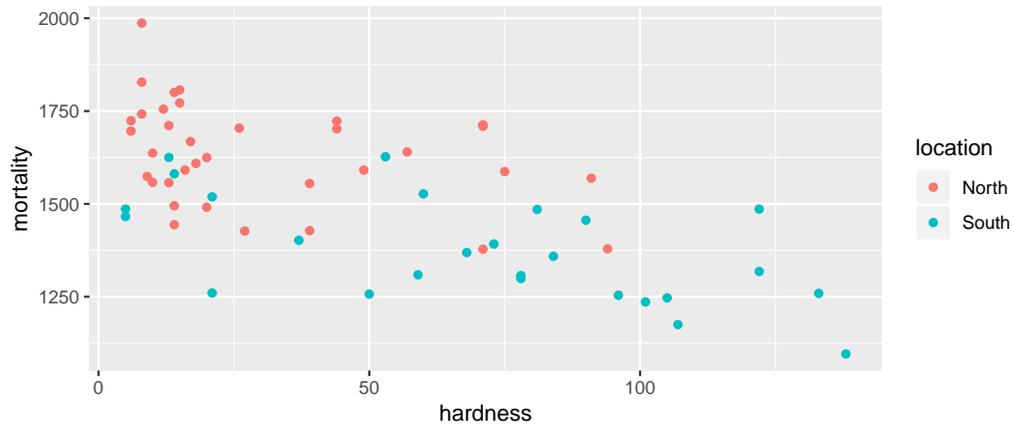


Figura 3.15: Datos relacionando dureza del agua y mortalidad.

todas) para explicar la variable  $Y$  en función de la variable  $X$ . El proceso de elegir la mejor función no tiene por qué ser sencillo ni simple. Debemos tener también en cuenta que quizá no haya una única variable  $X$  influyendo en la variable  $Y$ , sino que pueden existir diferentes variables explicativas  $X_1, X_2, \dots, X_k$  que sean necesarias para poder establecer predicciones de la variable  $Y$  de interés.

### 3.5.5 El caso lineal

Como hemos visto en los ejemplos utilizados antes, observando el dibujo de la nube de puntos, tendremos que existe una relación de tipo lineal entre las variables si los puntos “forman” alguna disposición que se pueda concentrar alrededor de una línea recta (Figura 3.15: relación lineal decreciente).

Matemáticamente, la variable  $Y$  puede expresarse en función de  $X$  como una línea recta a través de una función del tipo

$$Y = a + bX$$

Esta recta se llama recta de regresión, y sirve para *predecir* el valor de  $Y$  para un valor nuevo de la variable  $X$ .

La forma de calcular los *mejores* valores para  $a$  y  $b$  (es decir, aquellos valores que consigan que la recta se “aproxime” lo más posible a los datos y, por consiguiente, luego permita obtener mejores predicciones) se llama *método de*

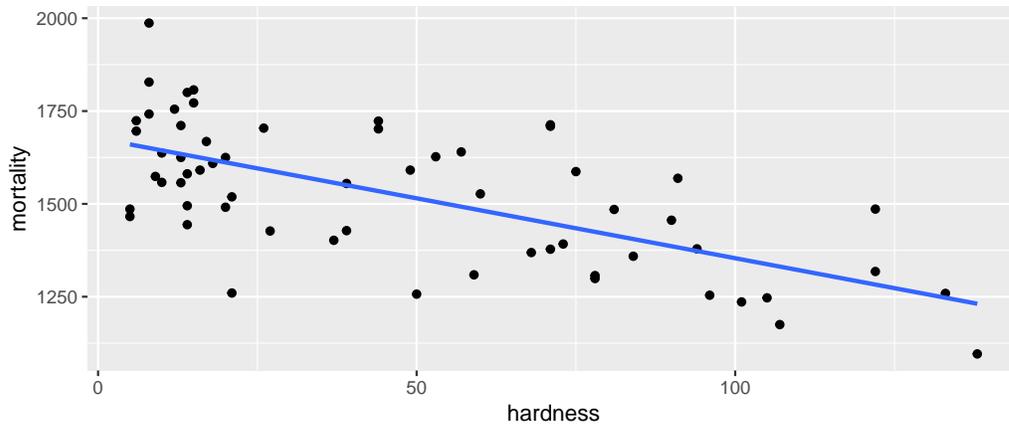


Figura 3.16: Línea de regresión para los datos del ejemplo anterior.

los *mínimos cuadrados* (ver el capítulo de probabilidades si se tiene interés por el desarrollo histórico de este método).

Los valores de los parámetros  $a$  y  $b$ , utilizando este método, se obtienen en función de los datos como:

$$b = \frac{S_{XY}}{S_X^2}$$

$$a = \bar{y} - b\bar{x}$$

```
.
##
## Call:
## lm(formula = mortality ~ hardness, data = mort)
##
## Coefficients:
## (Intercept)      hardness
##    16636         3.23
```

Si ahora quisiéramos utilizar la recta para hacer alguna predicción, escribiríamos  $a = 1676.356$  y  $b = -3.226$  y calcularíamos  $y = a + bx$ .



La predicción usando una recta de regresión tiene claros problemas. Uno es el de la **extrapolación** (salirnos de los límites del rango de

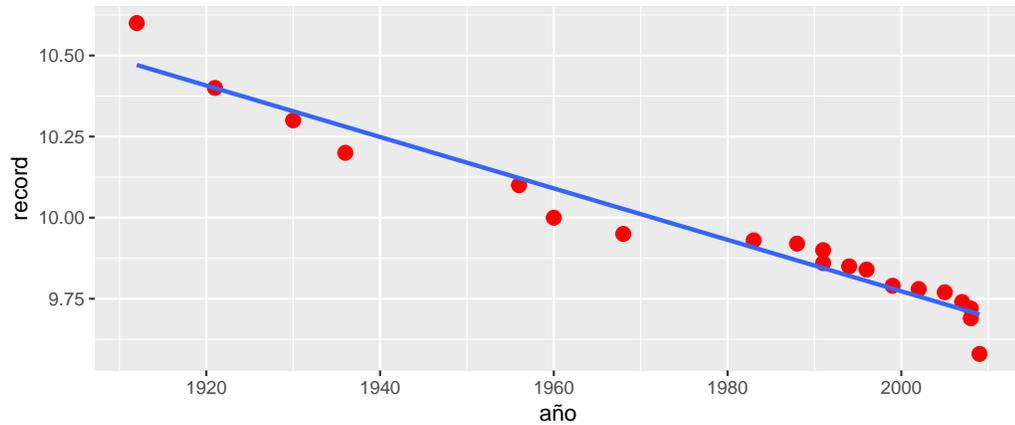


Figura 3.17: Tiempo realizado frente al año de la carrera, junto con la recta de mínimos cuadrados.

valores analizado). Un ejemplo clásico es el de los récords de los cien metros.

Como sabemos, la marca mundial de los 100 metros lisos ha ido disminuyendo con el paso de los años. Si calculamos la recta de regresión que relacione  $X$ ="año" e  $Y$ ="tiempo récord para recorrer los 100 metros", podríamos predecir cual sería el año en que se llegaría a un tiempo de 0 segundos, o incluso un tiempo negativo.

```
record_100_m <- read_excel("Data/record100m.xlsx")
dt = data.frame(x=record_100_m[,1], y=record_100_m[,2])
names(dt) <- c("año", "record")
p <- ggplot(data=dt,
            aes(x=año, y=record), color="red", size=3) +
  geom_point(color="red", size=3) +
  geom_smooth(method = "lm", se=FALSE)
p
```

El valor que se obtiene para  $R^2$  con estos datos es 0.9484.

Es evidente que este tipo de previsiones no tiene sentido, puesto que los

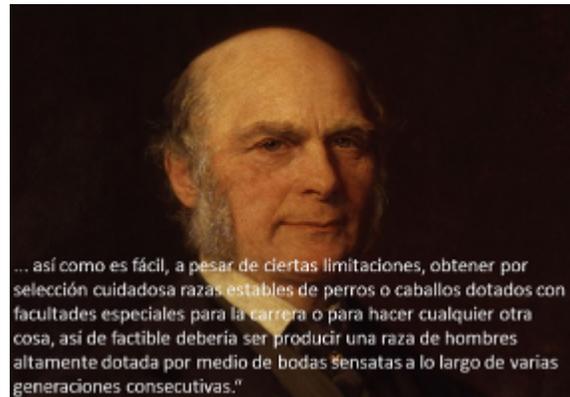


Figura 3.18: Francis Galton y una de sus muchas frases célebres.

valores mínimos de  $Y$  para los años actuales parece muy difícil rebajarlos. En Internet pueden encontrarse estudios de cuál es la función más adecuada para el ajuste de estas variables, con el fin de obtener una posible predicción dentro de límites razonables.



A veces una variable depende de otra, pero no a la inversa. Por ejemplo, un descenso de temperatura puede influir en un aumento del consumo eléctrico (por las estufas), pero un aumento del consumo eléctrico no influirá en el descenso de temperatura. De hecho, el aumento del consumo eléctrico podrá relacionarse tanto con un descenso de temperatura como con un aumento (si hace más calor, aumentará el uso del aire acondicionado).

### 3.5.6 El origen del término “regresión”

Fue Francis Galton el creador de este término. Su fama histórica procede, entre otras cosas, por ser el creador de la psicología diferencial, del mapa del tiempo, del saco de dormir, del silbato para perros, ser pionero en la clasificación de huellas dactilares, y por la creación del término “eugenesia”. En 1884 fundó el primer laboratorio de Biometría, y calculó que la probabilidad de que haya dos huellas dactilares iguales es prácticamente nula.

### 3.5.6.1 Galton y la eugenesia

La selección artificial de seres humanos fue sugerida desde muy antiguo, al menos desde Platón, quien creía que la reproducción humana debía ser controlada por el gobierno. Platón registró estos puntos de vista en *La República*: «que los mejores cohabiten con las mejores tantas veces como sea posible y los peores con las peores al contrario». Platón proponía que el proceso se ocultase al público mediante una especie de lotería. Otros ejemplos antiguos incluyen la supuesta práctica de las polis de Esparta de abandonar a los bebés fuera de los límites de la ciudad durante un periodo de tiempo, considerándose más fuertes a los supervivientes.

Durante los años 1860 y 1870, Galton sistematizó estas ideas y costumbres de acuerdo al nuevo conocimiento sobre la evolución del hombre y los animales provisto por la teoría de su primo Charles Darwin. Tras leer *El origen de las especies*, Galton observó una interpretación de la obra de Darwin a través de la cual los mecanismos de la selección natural eran potencialmente frustrados por la civilización humana. Galton razonó que, dado que muchas sociedades humanas buscaban proteger a los desfavorecidos y los débiles, dichas sociedades estaban reñidas con la selección natural responsable de la extinción de los más débiles.

Galton esbozó por vez primera su teoría en el artículo de 1865 *Talento y personalidad hereditarios*, explicándola luego más detalladamente en su libro de 1869 *El genio hereditario*. Galton comenzó estudiando la forma en la que los rasgos humanos intelectuales, morales y de personalidad tendían a presentarse en las familias. Su argumento básico era que el «genio» y el «talento» eran rasgos hereditarios en los humanos (aunque ni él ni Darwin tenían aún un modelo de trabajo para este tipo de herencia). Galton concluyó que, puesto que puede usarse la selección artificial para exagerar rasgos en otros animales, podían esperarse resultados similares al aplicar estas prácticas en humanos. Como escribió en la introducción de *El genio hereditario*:

“Me propongo mostrar en este libro que las habilidades naturales del hombre se derivan de la herencia, bajo exactamente las mismas limitaciones en que lo son las características físicas de todo el mundo orgánico. Consecuentemente, como es fácil, a pesar de estas limitaciones, lograr mediante la cuidadosa selección una raza permanente de perros o caballos dotada de especiales facultades

para correr o hacer cualquier otra cosa, de la misma forma sería bastante factible producir una raza de hombre altamente dotada, mediante matrimonios sensatos durante varias generaciones consecutivas.”

Según Galton, la sociedad ya fomentaba las enfermedades disgenéticas, afirmando que los menos inteligentes se reproducían más que los más inteligentes. Galton no propuso sistema de selección alguno, sino que esperaba que se hallaría una solución cambiando las buenas costumbres sociales de forma que animasen a la gente a ver la importancia de la reproducción.

Galton usó por primera vez la palabra **eugenesia** en su libro de 1883 *Investigaciones sobre las facultades humanas y su desarrollo (Inquiries into Human Faculty and Its Development)*, en el que quiso “mencionar los diversos tópicos más o menos relacionados con el cultivo de la raza o, como podríamos llamarlo, con las cuestiones eugenésicas”. En 1904, Galton aclaró su definición de eugenesia como “la ciencia que trata sobre todas las influencias que mejoran las cualidades innatas de una raza, y también con aquellas que las desarrollan hasta la mayor ventaja”.

La formulación de Galton de la eugenesia estaba basada en un fuerte enfoque estadístico, fuertemente influenciado por la “física social” de Adolphe Quetelet. Sin embargo, a diferencia de éste, Galton no exaltaba al “hombre medio”, sino que lo despreciaba por mediocre. Galton y su heredero estadístico Karl Pearson desarrollaron lo que se llamó el enfoque biométrico de la eugenesia, que desarrolló nuevos y complejos modelos estadísticos (más tarde exportados a campos completamente diferentes) para describir la herencia de los rasgos.

La eugenesia terminó aludiendo a la reproducción humana selectiva como intento de obtener niños con rasgos deseables, generalmente mediante el enfoque de influir sobre las tasas de natalidad diferenciales. Estas políticas se clasificaban en su mayoría en dos categorías: eugenesia positiva, la mayor reproducción de los que se consideraba que contaban con rasgos hereditarios ventajosos, y la eugenesia negativa, la disuasión de la reproducción de los que tenían rasgos hereditarios considerados malos. En el pasado, las políticas eugenésicas negativas han ido de intentos de segregación a esterilizaciones e incluso genocidio. Las políticas eugenésicas positivas han tomado típicamente la forma de premios o bonificaciones para los padres “aptos” que tenían otro hijo.

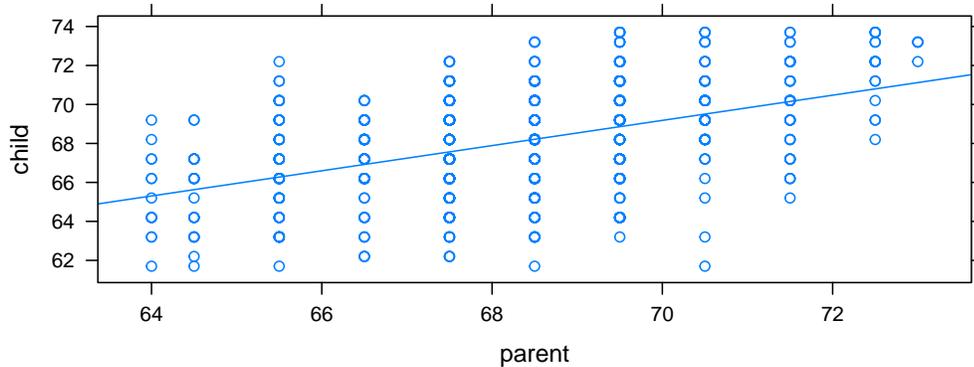


Figura 3.19: Recta de regresión de la estatura de los hijos en función de la media de los padres.

El ejemplo que motivo a Galton: la altura de los padres correlaciona linealmente con la estatura de los hijos (Figura 3.19), pero la estatura de los hijos tiende a la media. Es decir, el hijo de unos padres muy altos es muy probable que sea más bajo que ellos. Significa que existe una **regresión a la media**, o lo que Galton llamó **regresión hacia la mediocridad**.

```
data(Galton)
Gl<data.frame(Galton)
# dibujo de los datos y la línea de regresión
xyplot(child ~ parent, type = c("p", "r"), data = Gl)
```

```
# cálculo de los coeficientes de la recta
linea<-lm(child~ parent, data = Gl)
linea
```

```
##
## Call:
## lm(formula = child ~ parent, data = Gl)
##
## Coeficients:
## (Intercept)      parent
```

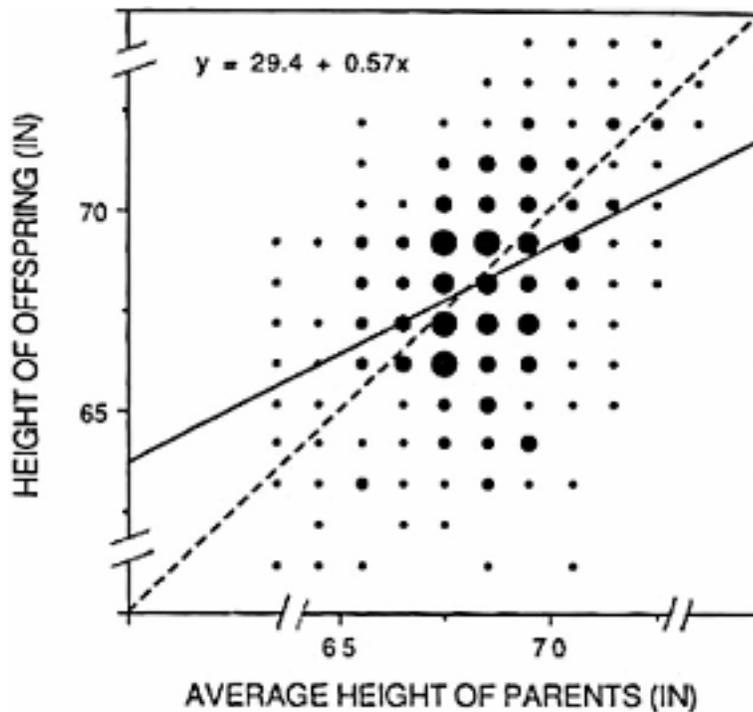


Figura 3.20: Gráfico original de Galton.

```
##      23.942      0.66
```

En el dibujo original que hizo en su día Galton (Figura 3.20), la línea de regresión o de mínimos cuadrados aparece comparada con la diagonal del gráfico (sería la recta  $Y = X$ ). De esta manera se puede comprender perfectamente la regresión hacia la media. A partir del centro de gravedad, la línea de regresión está más baja que la diagonal, lo que significa que, en media, los valores altos de la  $Y$  “descienden” o “regresan” hacia la media. Si no existiera esa regresión, el crecimiento sería continuo de generación en generación (se iría creciendo continuadamente).

```
knitr::include_graphics('Figure/Galtonheightregress.jpg')
```

Pero no sólo la estatura es una característica, por así decirlo *heredable*, es decir que parte de la altura de una persona depende de la de sus padres. Otras características, como la inteligencia, la valentía, la capacidad de pro-

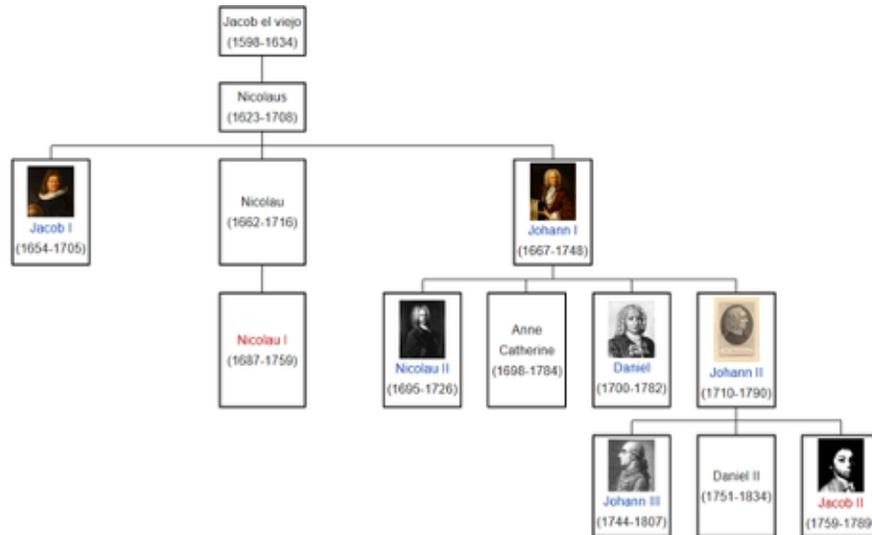


Figura 3.21: Los Bernoulli, una familia de genios de la ciencia. Aquí en España tenemos a los Pelayo y a los del Río (no son primos míos, ojo) .

gresar en una asignatura, o en un deporte, presentan propiedades similares. Observemos:

Familias de genios: <https://hipertextual.com/2018/03/nobel-padres-hijos>

Un caso particular en matemáticas: los Bernoulli

Fijémonos ahora en otra hija de Melanie Griffith (de su matrimonio con Antonio Banderas). Además, mucha gente joven no sabe que su abuela (la madre de Melanie) fue la actriz protagonista de *Los Pajaros*.

### 3.5.6.2 La importancia de la regresión a la media en entornos educativos

El premio nobel de economía Daniel Kahneman, en su libro *Pensar rápido, pensar despacio* (Kahneman and Egan, 2011), expone:

“La regresión a la media se da en todas las circunstancias donde se dan medidas extremas y tiene consecuencias predictivas de mucho interés, por ejemplo en los rendimientos escolares, deportivos o en cualquier otra situación donde los resultados pueden cuantificarse. Por ejemplo, si en un primer examen un niño saca



Figura 3.22: Dakota Johnson (50 sombras de Grey), hija de Melanie Griffith y Don Johnson, famosos actores de los años 80.



Figura 3.23: Otra hija de Melanie Griffith (de su matrimonio con Antonio Banderas). Además, mucha gente desconoce que su abuela (la madre de Melanie) fue la actriz protagonista de ‘Los Pajaros’.

un 10 y otro saca un 2, siendo la nota media de la clase un 5, es bastante posible que en un segundo examen el que sacó un 10 disminuya su nota mientras que el que sacó un 2 la aumente, pues las puntuaciones tienden a igualarse buscando la media.

Pero lo más interesante de esta cuestión es que esta regresión va a darse igual, tanto si animamos o reprendemos al suspendido como si nos dedicamos a lisonjear al sobresaliente. Algo que va en contra de nuestras intuiciones pedagógicas. En ellas tendemos a “echar la bronca” a los que quedan por debajo de la media y a alabar los buenos rendimientos de los que se salen por arriba, sin caer en la cuenta de que nuestros esfuerzos no sirven de nada cuando estamos enfrentando puntuaciones extremas. Lo cual no quiere decir que la motivación, el apoyo o el empoderamiento no sean buenas estrategias pedagógicas, sino que solo sirven en las medianías y no en los casos extremos.

Y como tenemos tendencia a premiar a los buenos y a castigar a los malos, estamos condenados a cosechar decepciones por parte de los buenos (y con los que somos mas agradables) y sorpresas agradables con aquellos con los que somos desagradables.”

Kahneman expone estas conclusiones en su libro, explicando a su vez una experiencia vivida durante una etapa de colaboración con pilotos de aviones del ejercito israelí, comprobando que las felicitaciones tras un buen vuelo (refuerzo positivo) se revelaban como mejor estrategia en contra de una reconvencción (refuerzo negativo), puesto que durante una serie de entrenamientos, todos los pilotos tienden a hacer unos ejercicios mejores que otros, por el efecto de regresión a la media.

### 3.5.6.3 Correlación espuria

Algunos ejemplos tomados de la literatura estadística (resumidos en la wikipedia) dan idea de errores comunes y detalles que cualquier investigador debe considerar siempre que relacione dos variables. Los siguientes son ejemplos de conclusiones erróneas obtenidas de una mala interpretación de la correlación entre dos variables (llamada correlación **espuria**).

- *Dormir sin quitarse los zapatos tiene una alta correlación con despertarse con dolor de cabeza. Por lo tanto, el dormir con los zapatos puestos*

*ocasiona levantarse con dolor de cabeza.*

Este resultado mezcla los conceptos de correlación y causalidad, porque concluye que dormir con los zapatos puestos provoca dolor de cabeza al levantarse. Hay un tercer factor que no se ha tenido en cuenta, que es que irse borracho a la cama provoca ambos efectos: no poder ni quitarse los zapatos y levantarse mareado.

- *Los niños pequeños que duermen con la luz encendida son mucho más propensos a desarrollar miopía en la edad adulta.* Esta fue la conclusión de un estudio de la Universidad de Pennsylvania, publicado en 1999 en la revista Nature. Un estudio posterior de la Ohio State University refutó esta teoría, y encontró una fuerte relación entre la miopía de los padres y el desarrollo de la miopía infantil, advirtiendo que los padres miopes tenían más probabilidades de dejar una luz encendida en el dormitorio de sus hijos. De nuevo una tercera variable no tenida en cuenta (la miopía de los padres), causaba la alta correlación entre las otras dos variables relacionadas.
- *Cuando aumentan las ventas de helado, la tasa de muertes por ahogamiento también aumenta. Por lo tanto, el consumo de helado provoca ahogamiento.* El helado se vende durante los meses de verano a un ritmo mucho mayor que en épocas más frías, y es durante estos meses de verano que las personas son más propensas a participar en actividades relacionadas con el agua, como la natación. El mayor porcentaje de muertes por ahogamiento es causado por una mayor exposición a las actividades acuáticas.
- *Desde 1950, tanto el nivel de dióxido de carbono (CO<sub>2</sub>) en la atmósfera como los niveles de obesidad han aumentado considerablemente. Por lo tanto, el CO<sub>2</sub> atmosférico provoca la obesidad.* Obviamente, en las ciudades que han mejorado su nivel de vida a partir de los años 50, se come más y se expulsa más CO<sub>2</sub> a la atmósfera (coches, fábricas... ).
- *Con una disminución en el uso de sombreros, ha habido un aumento en el calentamiento global durante el mismo período.* Por lo tanto, el calentamiento global es causado por personas que abandonan la práctica de usar sombreros. La explicación de este efecto sería muy parecida a la del ejemplo anterior.
- Por último, una página donde se pueden encontrar montones de gráficos

con correlación espuria:

<http://tylervigen.com/spurious-correlations>

# Capítulo 4

## Probabilidad

Estudio de la teoría matemática de la probabilidad

### 4.1 Introducción histórica

“Se ve que, en el fondo, la teoría de las probabilidades no es más que el buen sentido reducido a cálculo. Hace apreciar con exactitud lo que los espíritus justos sienten por una especie de instinto, sin que a menudo puedan darse cuenta de ello. No deja nada de arbitrario en la elección de las opiniones y del partido a tomar, siempre que se pueda mediante ella determinar la elección más ventajosa. Por ello, viene a ser el suplemento más dichoso a la ignorancia y debilidad del espíritu humano. Si se consideran los métodos analíticos a los que esta teoría ha dado nacimiento, la veracidad de las teorías que le sirven de base, la lógica fina y delicada que exige su empleo en la solución de problemas, las instituciones públicas que se apoyan en ella, y la extensión que ha alcanzado y puede alcanzar todavía, por su aplicación a las cuestiones más importantes de la filosofía natural y de las ciencias morales; si se observa luego que incluso en las cosas que no pueden someterse al cálculo, da la apreciación más segura que pueda guiarnos en nuestros juicios, y que enseña a protegernos de las ilusiones que a menudo nos extravían, se vería que no hay ciencia más digna de nuestras reflexiones, ni que sea más digna



Figura 4.1: El marqués de Laplace.

de hacerse entrar en el sistema de instrucción pública.”

— Pierre-Simon Laplace (1749-1827, matemático francés que llegó a ser ministro del interior con Napoleón).



la mayor parte de lo que se escribe en esta introducción está basada en (Camacho, 2002), (del Cerro, 2002) y (Secades, 2002).

El divulgador científico Ian Hacking afirma que “el evento conceptual más decisivo del siglo veinte fue el descubrimiento de que el mundo no es determinístico. La **causalidad**, por mucho tiempo bastión de la metafísica, fue desacreditada, y paulatinamente emergió una nueva forma de “conocimiento objetivo” — la estadística (Hacking and Bixio, 1995).

**Causalidad** y **Azar** son conceptos opuestos, en el sentido de que si todos los procesos que rigen el universo (y la vida) son causales, el azar no existe. Por otro lado, la existencia del azar lleva implícito la posibilidad del “libre albedrío”. Un nombre equivalente a “teoría de la probabilidad” es, precisamente, “ley del azar”, porque mientras sea posible “medir” la certeza o incertidumbre de los sucesos que se presentan ante nosotros, ese azar deja de ser desconocido.

La Teoría de la Probabilidad (ciencia del azar o de la incertidumbre) es, históricamente, la unión de la disciplina matemática conocida como “geometría del azar” y, por otro lado, la probabilidad heredada del pensamiento clásico y cristiano. En efecto, en ambas materias se estudian cuestiones acerca de las cuales el hombre no posee la certeza absoluta sobre su realización o existencia.

El concepto de probabilidad vigente hasta mediados del siglo XVII, fecha de la creación definitiva del **Cálculo de Probabilidades**, tuvo un carácter filosófico y teológico. La primera aparición del término **probable** se remonta a **Aristóteles** (384-322 a.c.). En los *Tópicos* aparece la siguiente definición:

“El razonamiento probable es aquel que parte de las cosas plausibles. Son cosas plausibles las que parecen bien a todos, o a su mayoría, o a los más conocidos y reputados”.

El concepto de probabilidad es el concepto *opinión certificada por la autoridad*. Una cosa es probable si quien la sostiene es alguien con reconocimiento social (no sería lo mismo que un médico, por ejemplo, dijera que una enfermedad probablemente se curase, a que lo dijera alguien que no tuviera estudios).

Más adelante, Carnéades de Cirene (Libia, 214-129 a.C.), director de la Academia platónica, defendió un cierto libre albedrío y voluntariedad en la explicación de la conducta humana, algo que chocaba completamente con la concepción predominante del determinismo. Empezó a abrirse un conflicto que adquiriría especial trascendencia en los siglos XVI y XVII, entre el determinismo propio de un destino voluntad de Dios, frente al libre albedrío, que supone la libertad del hombre de elegir su propio camino.

La siguiente reseña histórica referida a la probabilidad aparece en Marco Tulio Cicerón (106-43 a.C.):

“No somos de los que niegan la existencia de la verdad, sino de los que sostienen que la verdad y la falsedad andan tan hermanadas y mezcladas, que en ellas no hay ninguna señal cierta para discernirlas y prestarles asentimiento, de donde resulta que el sabio debe regir su vida según la probabilidad”.

Según los historiadores, para Cicerón lo más probable es aquella que sucede la mayoría de las veces, como enuncia del siguiente modo:

“Y no basta tener esa fortaleza en teoría, si no se practica. Así

como puede ciertamente tenerse la teoría de una ciencia aunque no se practique, la virtud de la fortaleza consiste enteramente en la práctica”.

Asímismo, es ampliamente conocido que los conceptos de fortuna y azar eran perfectamente conocidos en Roma. Suetonio en su obra “Vida de los doce Césares”, nos relata como ya en aquella época era habitual jugar a los dados. El mismo Emperador Claudio (10 a.C. - 54 d.C.) realizó un tratado sobre el juego de los dados (recomendamos la visión de la serie de la BBC “Yo, Claudio”).

Sin embargo, durante los primeros siglos del cristianismo, el concepto de probabilidad desaparece, por cuanto se cree que todo lo que ocurre es designio divino. Para San Agustín (354-430), la verdad era sólo conocida por Dios, y no es posible acercarse a ella por medio de la probabilidad.

“Sólo algún divino numen puede manifestar al hombre lo que es verdad”.

“Cuando tú (se refiere a Carnéades) dices que no conoces ninguna verdad, ¿cómo puedes abrazar lo que se asemeja a ella?”.

San Agustín asegura que, del empleo de la probabilidad para la formación de la conciencia, *nada bueno puede derivarse*:

“Si la probabilidad, término tan clásico para la formación de la conciencia, ha de ser norma de la acción, se abre una ancha brecha a la moral para los homicidios, parricidios, sacrilegios... etc”.

Como curiosidad, señalamos también esta otra frase suya bastante conocida:

“El buen cristiano deberá guardarse de los matemáticos y de todos aquellos que practican la predicción sacrílega, particularmente cuando proclaman la verdad. Porque existe el peligro de que esta gente, aliada con el diablo, pueda cegar las almas de los hombres y atraparlos en las redes del infierno” (afirmación que se refería no a lo que hoy entendemos por matemático, sino a los científicos en general).

No sería hasta San Alberto Magno (1200-1280) y Santo Tomás (1225-1274) que se retomarían los conceptos de azar, probabilidad y su estudio. Santo Tomás llevó a cabo una destacable recuperación del pensamiento Aristotélico.

Aún así, no sería hasta el siglo XVII, y debido, entre otras cosas, a la reforma protestante del siglo XVI y a la revolución científica, cuando la certeza dejó de basarse en la autoridad, y pasó a fundarse en criterios que se consideraban objetivos, esto es, que tenían que ver con la experiencia. La objetividad vino a sustituir así a la autoridad, y donde antes era el sujeto el que legitimaba la objetividad de nuestro conocimiento, ahora será el objeto de ese conocimiento el que legitime la autoridad del sujeto.

En el siglo XVII se culminó también un proceso de cambio donde se empezaba a aplicar el lenguaje matemático a los fenómenos naturales. Aristóteles había descrito que la “naturaleza” de las cosas era un principio interior que garantizaba su modo de actuar. La acción humana tenía también su propia esencia o razón de ser, y esa esencia convertía a la acción humana en una acción diferente a la acción mecánica, fruto sólo de causas eficientes. La finalidad de la actuación humana no era otra que la felicidad eterna, de la que no se podía excluir necesariamente la felicidad terrena, dentro de este mundo. La revolución científica de ese siglo destruyó la destrucción de la visión aristotélica de la naturaleza y la sustituyó por una visión mecanicista (explicar los fenómenos mediante las leyes mecánicas del movimiento).

## 4.2 La geometría del azar

El matemático italiano del siglo XVI Gerolamo Cardano definió la suerte de un evento como la relación entre los resultados favorables y los desfavorables (lo que implica que la probabilidad de un evento viene dada por la relación entre los resultados favorables y el número total de resultados posibles). Christian Huygens (1657) redactó el primer tratado matemático sobre los juegos de azar. Sin embargo, la mayor parte de los autores señalan como fecha clave de su creación la segunda mitad del siglo XVII, cuando los matemáticos Blaise Pascal (1623-1662) y Pierre de Fermat (1607-1665) tratan de resolver ciertos problemas relativos a los juegos de azar planteados por el Caballero de Mére, célebre jugador de la corte francesa de aquel entonces, a su amigo Pascal.

El matemático Blaise Pascal advirtió que la mayor parte de los fenómenos naturales se presentan con mayor o menor grado de incertidumbre, sin alcanzar el grado de certeza absoluta. Es por ello por lo que Pascal optó por una postura diferente a las que, en aquel momento predominaban, que eran



Figura 4.2: Izquierda: Blaise Pascal. Derecha: Pierre Fermat

el pirronismo y el dogmatismo. Uno de los elementos fundamentales que contribuyeron a la conexión definitiva del cálculo de lo que Pascal denominó geometría del azar, con la conceptualización filosófica y teológica de la probabilidad, fue *la Lógica de Port-Royal*, cuyos autores -Antoine Arnauld y Pierre Nicole- tuvieron una estrecha relación con Pascal. Además, esta obra ejerció sobre el pensamiento posterior una notoria influencia, de la que destacamos la que asimiló Jacques Bernoulli (1655-1705) en su *Ars Conjectandi* (Arte de las conjeturas).

La correspondencia mantenida por Pascal y Fermat se ocupó fundamentalmente del llamado problema de la división de las apuestas, que consistía básicamente en establecer una regla fija que permitiera dividir entre los jugadores el montante de las apuestas de un juego cuando este, por la razón que sea, se interrumpe y no puede terminarse. Esta correspondencia sentó las bases de los criterios analíticos que permitirían medir matemáticamente la probabilidad.

Posteriormente a *Ars Conjectandi* de Bernoulli (1713, póstumo, y que hizo que 2013 fuese considerado año internacional de la Estadística, al cumplirse 300 años de su publicación), surge *La doctrina del azar* de Abraham de Moivre, trabajo donde se relaciona la matemática y la probabilidad. Más tarde, Laplace desarrollaría su famoso *Ensayo filosófico sobre las probabilidades*, y su famosa regla de los casos favorables entre los casos posibles.

### 4.3 El sentido estadístico-probabilista de la actualidad

Hacking (Hacking and Bixio, 1995) describe la concepción estadístico-probabilística de la realidad como un nuevo “estilo de razonamiento” (Piovani, 2007). Hacking también afirma que **el evento conceptual más decisivo del siglo veinte fue el descubrimiento de que el mundo no es determinístico**. Otros prefieren recurrir al concepto de paradigma, que, desde su introducción en los años sesenta, ha pasado a ocupar un lugar fundamental, cuando se trata de entender que las ideas científicas compartidas en un momento determinado —y dadas por descontadas por las mayor parte de los científicos— son en realidad complejas construcciones sociales.

Muy habitualmente se atribuye la creación de las herramientas matemático-probabilísticas al interés de obtener mejores apuestas en los juegos de azar. Muchos textos de estadística y probabilidades comienzan hablando que los juegos de azar existen desde tiempos ancestrales, y de que se conocen ejemplos de pueblos que utilizaban el hueso astrágalo de animales para realizar una especie de antecedente del dado. Es muy probable, por lo tanto, que cualquier estadístico sepa pronunciar bien *astrágalo* por raro que pueda parecer. A continuación, se habla de que Pascal y Fermat fueron básicamente los precursores de la teoría de la probabilidad, al intentar dar respuesta a las preguntas de un noble y conocido jugador francés, Antoine Gombaud, llamado *Chevalier de Méré*. Aunque este hecho es rigurosamente cierto, sería incorrecto decir que el desarrollo del cálculo de probabilidades se motivó o surgió paralelamente a la creación de juegos de azar y/o de apuestas. En realidad, la teoría de la probabilidad nació a partir de la consideración de problemas científicos concretos, uno de ellos el de la medición en astronomía. Este problema favoreció el perfeccionamiento de los instrumentos conceptuales y heurísticos en los que se funda la teoría de los errores, que se “desarrolló en los siglos XVII y XVIII como un apéndice [...] de la astronomía” (MacKenzie, 1981).

Debido a la lógica complejidad de las mediciones en astronomía, y considerando las herramientas disponibles en los siglos XVI-XVII, los astrónomos “debieron reconocer que era imposible medir algo con completa exactitud”. Sin embargo, el investigador podía aprovechar la posibilidad de repetir sus mediciones a los efectos de acrecentar la precisión. El objetivo de la teoría

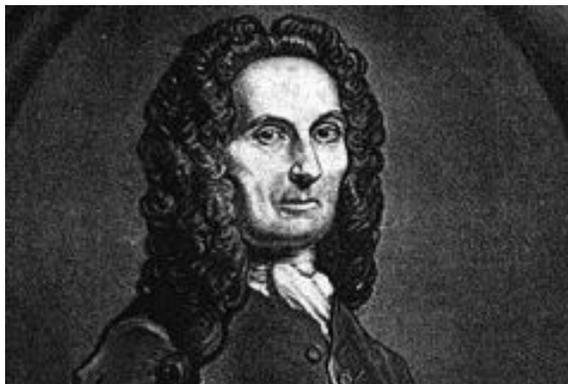


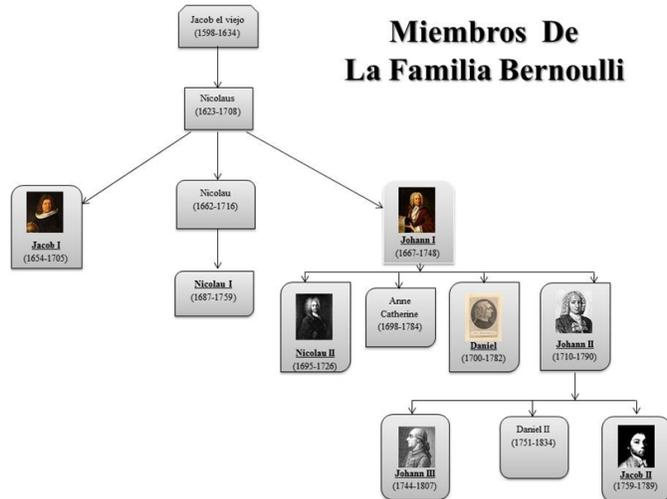
Figura 4.3: \*Abraham de Moivre (1667-1754), matemático francés que predijo el día de su muerte a través de un cálculo matemático (parece que se equivocó en 73 días). Según se dice, contaba el tiempo que dormía, y veía que cada día dormía algo menos que el anterior, por lo cual pudo deducir que llegaría un día en que no dormiría y eso sería cuando muriese - caso para Iker Jiménez.\*.

de los errores era justamente este: **reducir al máximo posible el error en una cantidad dada, y dar una estimación confiable de su valor**. Las aplicaciones de la teoría de los errores en astronomía se basaban en dos instrumentos de excepcional importancia, todavía hoy fundamentales en la investigación científica: la distribución normal y el método de los mínimos cuadrados.

En 1733, **Abraham De Moivre** (1667-1754) introdujo en la teoría de las probabilidades la función de la curva normal, entonces conocida como ley de los errores. Por primera vez era posible aplicar la teoría de las probabilidades a un número indefinidamente grande de sucesos independientes ([Porter, 1986](#)).

## 4.4 La familia Bernoulli

Si hay algún ejemplo de que la genética influye en el talento, ese es el de la familia Bernoulli, original de Amberes, por entonces en los Países Bajos españoles (hablamos de la guerra de independencia de los Países Bajos, que duró 80 años), y que emigró para escapar de la persecución de los hugonotes (protestantes). Tras un breve período en Frankfurt se establecieron en Basilea,



Suiza.

Jacob Bernoulli (1654 - 1705), también conocido como Jacob, Jacques o James Bernoulli, era el hermano mayor de Johann Bernoulli. En 1690 se convirtió en la primera persona en desarrollar la técnica para resolver ecuaciones diferenciales separables. Se familiarizó con el cálculo mediante su correspondencia con Gottfried Leibniz, y colaboró con su hermano Johann en varias aplicaciones

Su obra maestra fue *Ars Conjectandi* (el Arte de la conjetura), un trabajo pionero en la teoría de la probabilidad. La publicó su sobrino Nicholas en 1713, ocho años después de su muerte por tuberculosis. Los términos *ensayo de Bernoulli* y *números de Bernoulli* son resultado de su trabajo. También existe un cráter en la Luna bautizado cráter Bernoulli en honor suyo y de su hermano Johann.

El físico y matemático Daniel Bernoulli (1700-1782) fue el más famoso de la familia. Nació en Groningen el 8 de febrero de 1700. Pasó los primeros 5 años de su vida en esta ciudad, donde su padre trabajaba como catedrático . Matemático y físico, también se doctoró en medicina en 1721 con una tesis sobre la respiración. En 1724, Catalina I de Rusia le propuso junto a su hermano Nicolás para la recién fundada Academia de Ciencias de San Petersburgo. Un año después, ambos se instalaron en la ciudad rusa, donde Nicolás trabajaría como profesor de mecánica, y Daniel como profesor de Fisiología.

Nicolás fallecería poco después a causa de unas fiebres. Daniel permaneció en la Academia donde fundó una escuela rusa de fisiología mecánico-matemática a la que invitó a su amigo y compañero de estudios en Basilea, Leonhard Euler. En San Petersburgo participó en los cálculos necesarios para la construcción de más de 80 fuentes, estanques y un acueducto en el Palacio de Verano, residencia de los zares. Fue famoso por desarrollar la teoría de la mecánica de fluidos.

Pero en su vertiente de investigador médico también desarrolló un resultado fundamental. Hasta aquel momento lo habitual era dejar sangrar a los enfermos como remedio de casi todo. Los sangrados se basaban en las investigaciones de Hipócrates en el siglo V a.C. Había descubierto que los procesos inflamatorios se debían a una acumulación de sangre y que tan sólo había que disminuir la cantidad de ésta en el cuerpo humano para curar al paciente. El método se hizo tan popular que se convirtió en la técnica para tratar casi cualquier enfermedad. El problema era que ningún médico sabía hasta dónde llegar, y el ojo clínico era la única herramienta para parar a tiempo el drenaje de fluido.

Tanto Aristóteles como Hipócrates habían otorgado al corazón la responsabilidad de producir calor para el cuerpo. Tuvieron que pasar 23 siglos para que William Harvey afirmara que el corazón es como una bomba, y nuestros vasos sanguíneos constituían una red de canales. El único método para conocer la presión de un fluido lo descubrió Edme Mariotte, y consistía en dejar que el líquido saliera libremente por la tubería para que golpeará contra un balancín vertical con una pesa en el otro extremo. Bernoulli, con su experiencia en la teoría de fluidos, pensó que, si la presión es la misma en todas las direcciones, dará igual seccionar el conducto que pincharlo en un solo punto. De esta forma un pequeño chorro emana verticalmente llegando a una altura determinada. Si la altura era grande, la presión sería alta; si la altura era pequeña, la presión sería baja. Se le ocurrió poner un pequeño capilar de vidrio en el lugar del pinchazo. El líquido subía por el capilar hasta una altura proporcional a la presión, pero sin que se derramara un chorro constante de líquido. El sistema funcionó igualmente con una arteria. Sólo había que repetir la metodología y observar la altura de la columna de sangre. Pronto todos los médicos de Europa adoptaron esta técnica antes de dejar sangrar a un paciente.



Figura 4.4: Carl Gauss.

## 4.5 Laplace

Pierre-Simon Laplace (1749-1827) fue un astrónomo, físico y matemático francés. Continuador de la mecánica newtoniana, descubrió y desarrolló la transformada de Laplace y la ecuación de Laplace. Como estadístico, sentó las bases de la teoría analítica de la probabilidad; y como astrónomo planteó la teoría nebular sobre la formación del sistema solar. Él vio en la curva normal un instrumento excelente cuya aplicabilidad extendió a la probabilidad (Simon, 1951). Sin embargo, Laplace estaba convencido de que las más importantes cuestiones de la vida eran problemas de probabilidades, y por consiguiente susceptibles de la aplicación de sus teorías.

En 1795, Laplace empezó a publicar el primero de los cinco volúmenes que constituirían su famosa *Mecánica celeste*. En 1799 fue nombrado ministro del Interior por Napoleón, aunque no estuvo en el cargo más que seis semanas. En 1812 publicó su *Teoría analítica de las probabilidades* y en 1814 el *Ensayo filosófico sobre la probabilidad*. Prácticamente todo el mundo conoce por su nombre la famosa regla de los *casos favorables entre los casos posibles*.

Intimamente ligado con el problema del error de medición, el **método de los mínimos cuadrados** tiene una historia más reciente. En 1805 el matemático Adrien-Marie Legendre (1752-1833) anunció un método general para reducir las múltiples observaciones de un objeto —como una estrella o un planeta—, aunque lo presentó sin una justificación probabilística. Esta fue la primera comunicación efectiva del método, ya que posiblemente Carl F. Gauss (1777-1855) —como él mismo declarara en varias ocasiones— lo usaba desde hacía casi diez años, sin haber tenido la oportunidad de hacerlo público.

En 1810, Laplace desarrolló una derivación alternativa del método de los mínimos cuadrados, y estableció que los errores en astronomía, como en los estudios poblacionales, deberían distribuirse según la ley de los errores (o ley normal). Como afirma MacKenzie (MacKenzie, 1981):

“Los teóricos de los errores mostraron que la mejor estimación de una cantidad [...] era generalmente la media (aritmética) de las diversas mediciones, y que estas mediciones seguían típicamente la distribución matemática que llamaban “ley de la frecuencia de los errores“.

Los astrónomos y matemáticos del siglo XIX produjeron una enorme cantidad de artículos sobre el método de los mínimos cuadrados. A la muerte de Laplace, en 1827, la teoría de las probabilidades había alcanzado un alto nivel. Hacia mediados de siglo, el conocimiento de las técnicas matemáticas de las probabilidades estaba muy difundido en la astronomía y las matemáticas, pero no en otros campos. Fue otro astrónomo -**Adolphe Quetelet** (Bélgica, 1796-1874)— quien impulsaría la aplicación de la ley de los errores (curva de Gauss) a los fenómenos sociales. Otro francés, \*Simeón Denis Poisson\*\* (1781-1840) promovería también esa aplicación en aquellos campos en donde la regularidad podía empezar a entenderse matemáticamente.

## 4.6 desde Poisson hasta Kolmogorov

Con la revolución francesa y la independencia de las colonias americanas, se instituyó la figura del jurado popular (la justicia emana del pueblo y para el pueblo) en la constitución francesa y americana. Poisson realizó estudios sobre el número de personas que debían constituir un jurado para emitir un juicio más acertado, igual que la mayoría necesaria para dicho fin.

**Augustus de Morgan** (matemático británico nacido en la India, 1806-1871) fue quien presentó las teorías de Laplace en Inglaterra en dos trabajos fundamentales: *Theory of Probabilities*, publicado en 1838 en la Encyclopaedia Metropolitana, y un ensayo sobre las probabilidades y su aplicación a los seguros, aparecido en la Cabinet Cyclopaedia.

En 1850, **William Herschel** (astrónomo germano-británico, descubridor del planeta Urano; 1738-1822) llamó la atención de los científicos ingleses con relación a las ideas de Quetelet sobre la aplicación de las probabilidades a los



Figura 4.5: Simeón Poisson.

datos sociales. Por otra parte, Quetelet ya era conocido en los círculos académicos de Gran Bretaña. En los años 30 había jugado un papel decisivo en la creación de la sección de estadística de la Association for the Advancement of Science, y en 1844 había presentado, en un encuentro de estadísticos llevado a cabo en Plymouth, un largo listado de temas (meteorología, física, química, botánica, agricultura, zoología, además de las cuestiones humanas) que a su juicio podían ser estudiados con los nuevos métodos estadísticos.

Hemos hablado ya de los considerados *padres* de la estadística moderna, que vendrían a continuación: Francis Galton, Karl Pearson y, ya en el siglo XX, Ronald Fisher. La moderna teoría de la probabilidad sería completada por el desarrollo de la **Axiomática de Kolmogorov**, realizada por el matemático ruso Andrei Kolmogorov (1903-1987), quien, por decirlo vulgarmente, “ordenó” las propiedades que existían hasta ese momento, y sentó las bases de la teoría matemática de la probabilidad tal como la conocemos en la actualidad.

## 4.7 Experimentos aleatorios

“El desarrollo de la mecánica cuántica —la teoría del mundo subatómico— en el siglo XX ha demostrado que las expectativas de un universo totalmente determinista pecan de exceso de optimismo. De hecho, la física moderna ha demostrado que no es posible predecir el resultado de todos los experimentos, ni siquiera



Figura 4.6: Andréi Kolmogórov.

en principio. La teoría puede únicamente predecir las probabilidades de distintos resultados. En las ciencias sociales, la situación es aún más compleja debido a la multiplicidad de elementos interrelacionados, muchos de los cuales son, como mínimo, inciertos. Los investigadores sociales del siglo XVII pronto se dieron cuenta de que su búsqueda de principios universales del tipo de la ley de gravitación de Newton estaba condenada al fracaso. Durante un tiempo parecía que, al introducir las complejidades de la naturaleza humana en la ecuación, es virtualmente imposible llegar a predicción segura alguna. La situación aún parecía más desesperada si se tomaba en cuenta el pensamiento de toda una población. Sin embargo, en lugar de desesperar, algunos astutos pensadores desarrollaron un innovador arsenal de herramientas matemáticas: la estadística y la teoría de probabilidades.” (Ball, 2004)

## 4.8 Definiciones básicas

Para introducir detalladamente la noción de probabilidad, vamos a hablar de dos tipos de experimentos. Consideraremos que un experimento es **un proceso por medio del cual se obtiene una observación**. Bajo este

enfoque, podemos distinguir entre experimentos *deterministas* y *aleatorios*.

Los primeros son aquellos que siempre que se repitan bajo condiciones análogas llevan al mismo resultado, por tanto este resultado puede predecirse.

Ej.

Una reacción química en condiciones prefijadas de antemano; cualquier experimento físico que se realice en las mismas condiciones.

Por el contrario, un *experimento aleatorio* es el que puede dar lugar a varios resultados conocidos previamente, sin que sea posible saber de antemano cuál de ellos se va a producir.

Ej.

El lanzamiento de una moneda; el lanzamiento de un dado; la observación de un paciente para saber si tiene o no una enfermedad.

### Espacio muestral El conjunto de los posibles resultados de un experimento aleatorio se llama espacio muestral. Se representa con la letra griega  $\Omega$ .

#### 4.8.1 Suceso elemental.

Se llama suceso elemental a cada uno de los posibles resultados de un experimento aleatorio. En el ejemplo del lanzamiento de un dado, cualquiera de los números del 1 al 6 es un suceso elemental.

#### 4.8.2 Suceso.

Es un subconjunto del espacio muestral,  $A \subset \Omega$ . Se dice que ocurre un suceso si ocurre alguno de los sucesos elementales que lo componen.

Ej.

en el lanzamiento de un dado, el suceso  $A = \text{salir número par} = \{2, 4, 6\}$ .

#### 4.8.3 Diagramas de Venn

Son las representaciones gráficas (habitualmente con círculos) de sucesos, y las posibles operaciones o relaciones entre ellos. Nos van a resultar muy útiles para entender las propiedades de las probabilidades de los sucesos.

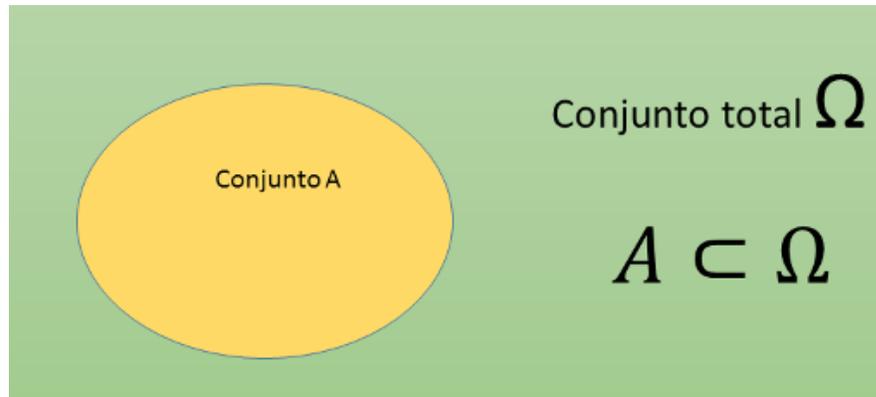


Figura 4.7: Representación de un suceso contenido en un espacio muestral.

Se dice de John Venn (lógico matemático, 1834-1923) que fue quién inventó la noción de frecuencia.

“La probabilidad de un suceso es su frecuencia relativa dentro de la serie” (*La lógica del azar*; 1866)

#### 4.8.4 Suceso imposible.

Es el suceso que no puede ocurrir nunca. Como conjunto, es el conjunto vacío  $\emptyset$ .

Ej.

Al lanzar un dado,  $\emptyset$  = “el resultado es 7, 8, o el número pi”.

Ej.

La policía observa a un político corrupto, y este es cogido infraganti mientras entrega todo lo robado a Cáritas, o la mitad de lo robado, o la mitad de la mitad, o la mitad de...

#### 4.8.5 Unión de sucesos

La Unión (de conjuntos o sucesos)  $A \cup B$  es el conjunto o suceso formado por los elementos que están en  $A$  o están en  $B$ .

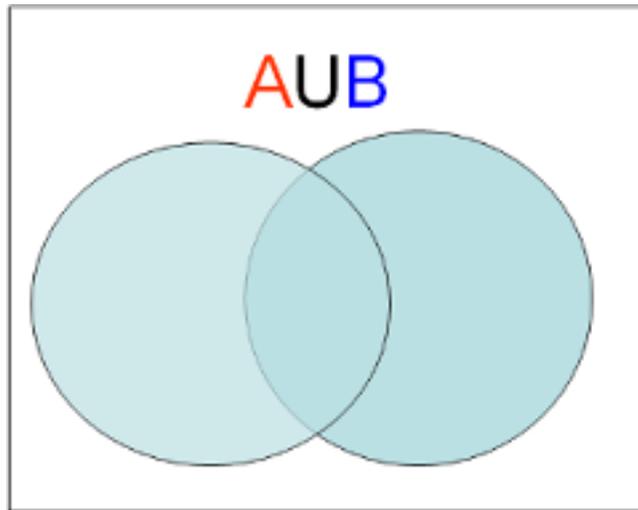


Figura 4.8: Diagrama de Venn de la unión de conjuntos.

Ej.

Supongamos la observación de un enfermo. Llamemos suceso  $A$  tener un hueso roto, y el suceso  $B$  estar resfriado. El suceso  $A \cup B$  consiste en que el enfermo puede tener un hueso roto, o estar resfriado, o tener ambas cosas a la vez.

Se puede generalizar a varios sucesos:  $A_1$  tener la enfermedad 1,  $A_2$  tener la enfermedad 2 y  $A_n$  tener la enfermedad  $n$ . El suceso unión  $A_1 \cup A_2 \cup \dots \cup A_n$  sería tener una, dos, o varias enfermedades simultáneamente (pudiendo tenerlas todas).

#### 4.8.6 Intersección de sucesos

$A \cap B$  está formado por los elementos que están simultáneamente en  $A$  y en  $B$ .

Ej.

En el ejemplo anterior, la intersección de  $A_1 \cap A_2 \cap \dots \cap A_n$  sería tener simultáneamente **TODAS** las enfermedades (el enfermo sería un gafe, pero eso no es cosa nuestra).

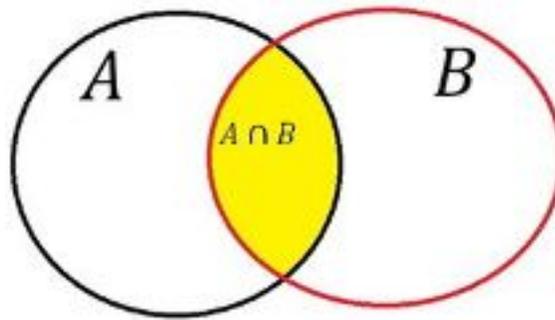


Figura 4.9: Suceso A: doble grado en español y gallego. Suceso B: doble grado en inglés y español. Intersección: grado en español.

Ej.

Un estudiante acaba de aprobar la selectividad. Mirando la página web de la UDC, decide que quiere hacer dos dobles grados: el de español y gallego y el de inglés y español. La administrativa de la facultad de Filología hace un diagrama de Venn y ve que con matricularse en español, gallego e inglés por separado le llega.

#### 4.8.7 Sucesos incompatibles

Dos sucesos se dicen **incompatibles** si no pueden ocurrir simultáneamente (al lanzar un dado no puede salir un número par e impar a la vez). Dos sucesos son incompatibles si como conjuntos son disjuntos ( $A \cap B = \emptyset$ ) (esta definición se extiende a más de dos sucesos o conjuntos).

#### 4.8.8 Diferencia de sucesos

$A - B$  está formado por los elementos de  $A$  que no están en  $B$ .

Ej.

Si  $A$  es *tener un hueso roto*, y  $B$  *estar resfiado*,  $A - B$  es *tener un hueso roto y NO estar resfiado*.  $B - A$  sería *estar resfiado y no tener un hueso roto*.

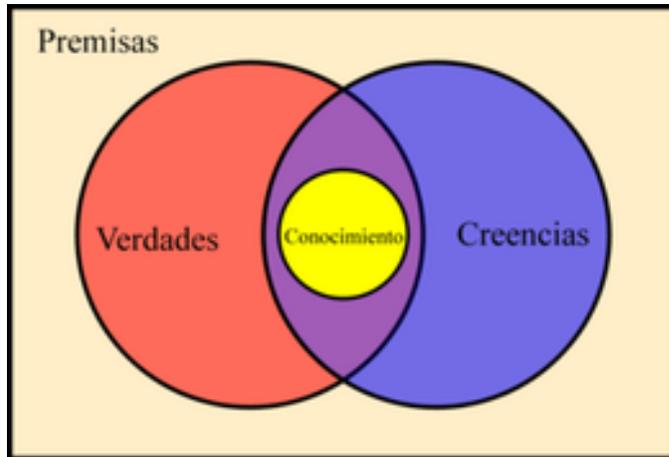


Figura 4.10: "La ciencia es un juicio verdadero acompañado de razón", Platón.



**Mensaje dirigido a los políticos actuales, exponiéndoles otro ejemplo de sucesos incompatibles**

Sin embargo, los políticos no parecen entenderlo.

Está claro que "inteligencia y política" también son incompatibles,

Figura 4.11: Algunos ejemplos de sucesos incompatibles.

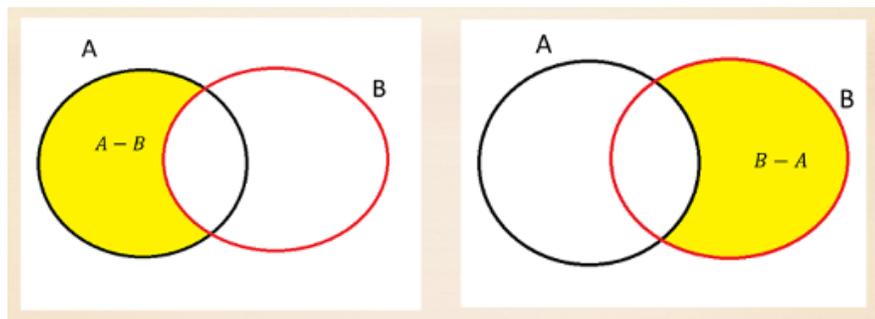


Figura 4.12: Diferencia de sucesos.

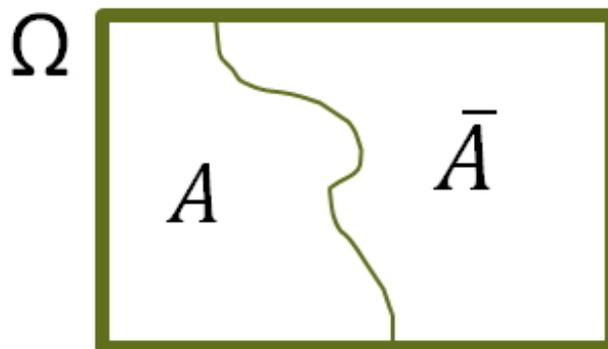


Figura 4.13: Suceso complementario.

Ej.

Supongamos que buscamos en Google información sobre el nuevo modelo de teléfono móvil de Chamchung: el **Chamchung-Boom2**. Una búsqueda genérica podría ser escribir: *ChamchungBoom2* a secas.

Si queremos garantizar que salga información sobre el precio pondríamos *ChamchungBoom2 +precio* que sería equivalente a la intersección de **Chamchung-Boom2** y **precio**.

Si escribimos *Chamchung Boom2 -Ifone* le estamos indicando a Google que no aparezca en los resultados la palabra *Ifone*, puesto que muchos resultados van a ofrecer comparativas con el *Ifone*, y puede que eso no nos interese. Esto valdría como ejemplo de diferencia de sucesos.

#### 4.8.9 Complementario de un suceso

El complementario de un suceso  $A$  se escribe  $A^c$  o  $\bar{A}$ . Corresponde al suceso formado por todos los elementos del espacio total  $\Omega$  que no están en  $A$ .

Se verifica que  $A \cap \bar{A} = \emptyset$  y  $A \cup \bar{A} = \Omega$ .

Ej.

Al lanzar un dado, si el suceso  $A$  es *que salga un número par*, el suceso  $B$  es *que salga número impar*.

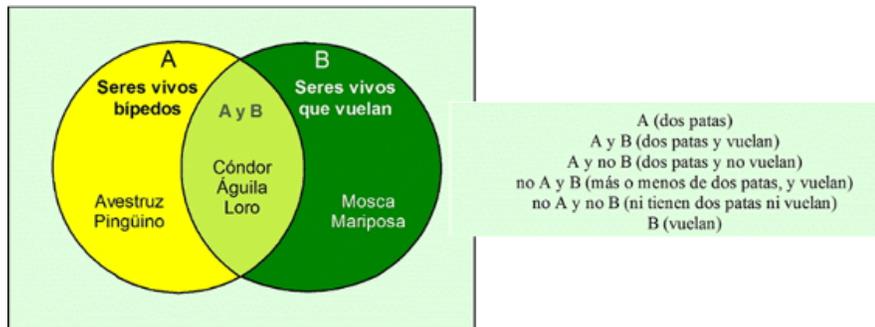


Figura 4.14: Diagrama de Venn con animales.

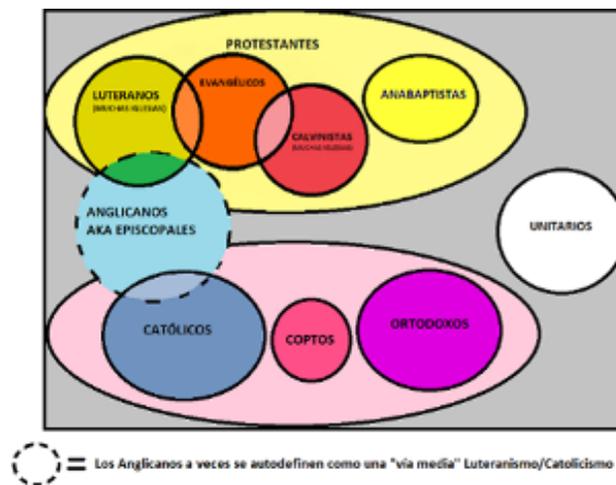


Figura 4.15: Diagrama de Venn de familias Cristianas.

**Pequeño resumen:**

Y más ejemplos:

## 4.9 Probabilidad

Si consideramos un experimento aleatorio que da lugar a un espacio muestral  $\Omega$ , la probabilidad es una función que asigna a los sucesos  $A \subset \Omega$  un número entre 0 y 1, verificando además que



Figura 4.16: A algunos es muy difícil ponerlos de acuerdo hasta con un diagrama de Venn.

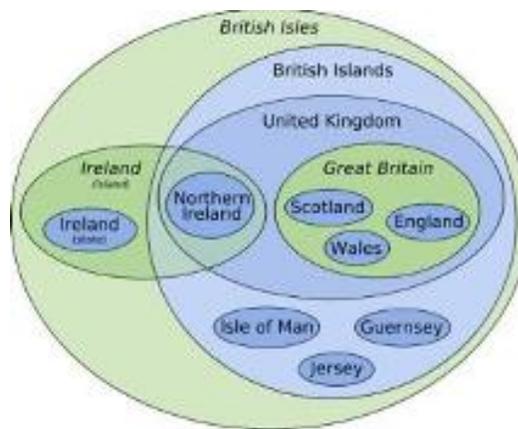


Figura 4.17: Para entender bien como se componen las islas británicas.

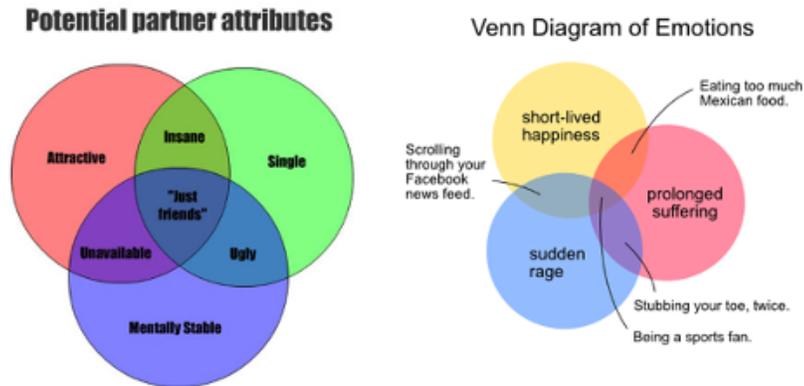


Figura 4.18: Venga, más...

A.1

$P(\Omega) = 1$  (la probabilidad del suceso seguro o total es 1)

A.2

$P(\emptyset) = 0$  (la probabilidad del suceso imposible es cero)

A.3

Si  $A$  y  $B$  son dos sucesos incompatibles ( $A \cap B = \emptyset$ ), entonces

$$P(A \cup B) = P(A) + P(B)$$

(de manera general, si  $n$  sucesos son incompatibles dos a dos, entonces:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Estas tres propiedades se llaman *Axiomas de Kolmogorov*.

Ej.

Consideremos el experimento aleatorio consistente en la observación de un paciente.

La probabilidad de los sucesos de la forma  $A = \text{sufir una enfermedad}$  varía entre 0 y 1 para cualquier posible enfermedad.

Si consideramos los sucesos:  $A_1 = \text{paciente no tiene fiebre}$ ,  $A_2 = \text{paciente}$

*tiene poca fiebre(36.537.5),  $A_3$  =paciente tiene fiebre alta (37.89), tenemos un ejemplo de sucesos incompatibles. La probabilidad de la unión es la suma de las probabilidades.*

Ej.

Se registra la edad de los pacientes de fisioterapia en el hospital de Valdecilla.

$A_1$  = paciente entre 18 y 30 años,  $A_2$  = paciente entre 31 y 40 años,  $A_3$  = paciente entre 41 y 60.

La probabilidad de que el próximo paciente que ingrese tenga entre 18 y 60 años es la suma de las probabilidades de cada uno de los sucesos.

## 4.10 ¿Cómo se asignan probabilidades a los sucesos?

### 4.10.1 Asignación equiprobable

Cuando un experimento aleatorio da lugar a un espacio muestral con finitos elementos equiprobables, entonces se aplica la famosa **Regla de Laplace**:

*Casos favorables entre casos posibles*

Ej.

Lanzamos un dado de seis caras. El experimento aleatorio da lugar a un espacio muestral con seis posibles resultados (todos igualmente probables).

La probabilidad de que salga un número par es 3 casos favorables entre 6 posibles, es decir 0.5

Ej.

Nos “encontramos” una tarjeta de débito al lado de un cajero automático. Si la introducimos en el cajero, disponemos de tres intentos para acertar con la clave de 4 dígitos. ¿Cuál es la probabilidad de acertar?

#### 4.10. ¿CÓMO SE ASIGNAN PROBABILIDADES A LOS SUCESOS? 143



Figura 4.19: Cocinero experto en probabilidades.

### Solución

Con la regla de Laplace el número de casos favorables es 3. El número de casos posibles es la cantidad de posibles números de cuatro dígitos que existen si utilizamos los números del 0 al 9: 0000, 0001, ..., 9999 que son diez mil posibilidades. Por lo tanto, la probabilidad de acertar es  $3/10000$ .

### 4.10.2 Asignación frecuentista

Cuando se repite muchas veces un mismo experimento, las frecuencias relativas de sus posibles resultados tienden a estabilizarse en torno a unos valores (números) concretos.

Esto se conoce como **ley de estabilidad de las frecuencias** (Jacob Bernoulli, 1713, uno de los resultados conocidos como *leyes de los grandes números*).

Es un hecho experimental que, cuando se lanza una moneda al aire un número alto de veces, la mitad de las veces, aproximadamente, aparece cada cara. Análogamente, si se lanza repetidamente un dado de seis caras, cada una de las caras sale aproximadamente la sexta parte de las veces.

Si la moneda, o el dado, se lanza un número bajo de veces, digamos 10 o 15 veces, puede suceder que la indicada aproximación a la mitad de las veces

para cada cara de la moneda, o a la sexta parte de las veces para cada cara del dado no sea correcta. No sería impensable que, de 10 lanzamientos, en nueve aparezca cara y solo aparezca cruz una vez, en la moneda; o bien, que al lanzar el dado 10 o 15 veces, no aparezca el 3 en ninguna ocasión.

Pero ocurre que, cuanto mayor es el número de veces que se lanza el dado, o que se lanza la moneda, mayor es la aproximación de la frecuencia relativa (el número de veces que aparece el suceso, dividido por el número total de veces que se realizó el experimento) a la probabilidad del suceso.

**Ej.**

La observación de un número elevado de pacientes con unos síntomas determinados nos permite conocer la probabilidad de que se tenga una determinada enfermedad. Cada vez que analizamos un paciente, tendrá la enfermedad o no. Anotando el número de pacientes que tiene la enfermedad ( $k$ ) de un total de  $n$  pacientes, la frecuencia relativa de dicha enfermedad es  $k/n$ , que, a medida que el valor  $n$  sea más elevado, se irá aproximando a un valor numérico concreto, que será su probabilidad.

**Ej.**

Las compañías de seguros elevan o disminuyen el precio de los seguros (por ejemplo de accidente) en función de la probabilidad de los mismos. Por ejemplo, el seguro de hombres jóvenes es más elevado que el de mujeres porque el número de accidentes es mayor.

El precio de un seguro de accidente en avión es bajo porque la probabilidad de que se registre un accidente de avión también lo es. Esta probabilidad está calculada tras haberse contabilizado el número de accidentes tras un número muy elevado de vuelos.

**Ej.**

*El Lanzamiento de una moneda.*

Un experimento muy simple para comprobar la ley de Bernoulli de estabilidad de las frecuencias consiste en lanzar una moneda, anotar si sale cara o cruz, y repetirlo un montón de veces. A medida que el número de repeticiones del experimento (lanzamientos) crezca, la frecuencia de aparición de cara (o de cruz) irá aproximándose a 0.5.

El naturalista francés George Lecler (conde de Buffon) (1707-1788) lanzó al

#### 4.10. ¿CÓMO SE ASIGNAN PROBABILIDADES A LOS SUCESOS? 145

aire una moneda 4.040 veces. El resultado: 2.048 caras, o una proporción de caras de  $2048/4040 = 0.5069$ . Cerca del año 1900, Karl Pearson lanzó al aire una moneda 24.000 veces. El resultado: 12.012 caras, una proporción de 0.5005.

Mientras estuvo preso por los alemanes durante la Segunda Guerra Mundial, el matemático sudafricano John Kerrich lanzó 10.000 veces una moneda al aire. El resultado: 5.067 caras, una proporción de 0.5067.

Esto con R (o cualquier lenguaje de programación) puede hacerse muy fácilmente mediante el comando `sample` que arroja números de forma aleatoria. `sample(a:b, N, replace=T)` escribe **N** números enteros entre **a** y **b**, y si queremos que se puedan repetir, escribimos `replace=T`. Por ejemplo, ahora vamos a “simular” el lanzamiento de 1000 monedas (el número 1 es cara, 2 es cruz), y los resultados los visualizamos en una tabla.

```
#libreria para que la tabla de resultados salga más bonita
library(pander)
# esto genera 1000 números aleatorios con valor de 1 a 2
x=sample(1:2,1000, replace=T)
pander(table(x))
```

1	2
494	506

Ej.

El Lanzamiento de un dado.

Ahora hacemos lo mismo simulando el lanzamiento de un dado.

```
x=sample(1:6,1000, replace=T)
pander(table(x))
```

1	2	3	4	5	6
148	165	176	176	169	166

## 4.11 Propiedades de la probabilidad

A partir de las tres primeras propiedades de la probabilidad (axiomas de Kolmogorov), pueden deducirse fácilmente las siguientes:



La probabilidad del suceso complementario vale 1 menos la del suceso:

$$P(\bar{A}) = 1 - P(A)$$



Si un suceso  $A$  está contenido en un suceso  $B$  (siempre que ocurre  $A$  ocurre  $B$ ), entonces  $P(A) \leq P(B)$



*Linda (Menos es más)*

El siguiente ejemplo corresponde a una prueba realizada por Kahneman (Kahneman and Egan, 2011) durante los años 70:

Se contaba a la gente que iba a realizar la prueba la siguiente historia sobre una mujer llamada Linda:

Linda tiene treinta y un años, es soltera, franca y muy inteligente. Se graduó en filosofía. De estudiante le preocupaban mucho los asuntos de discriminación y justicia social, y también participó en manifestaciones antinucleares.

Ahora, a los participantes de la prueba se les proponía elegir, de entre las siguientes afirmaciones, aquella que pareciera más probable sobre la descripción dada.

*Linda es profesora de primaria.*

*Linda trabaja en una librería y recibe clases de yoga.*

*Linda milita en el movimiento feminista.*

*Linda presta asistencia social en psiquiatría.*

*Linda es reparadora de calderas.*

*Linda es cajera de un banco.*

*Linda es corredora de seguros.*

*Linda es cajera y activista del movimiento feminista.*

Linda encaja muy bien en el tipo de la feminista activa, bastante bien en el de alguien que trabaja en una librería y recibe clases de yoga, y muy poco en el de la cajera de un banco o en el de la corredora de seguros.

Centrémonos ahora en los ítems críticos de la lista: ¿parece Linda más una cajera de banco o una cajera de banco que milita en el movimiento feminista? Cualquiera estará de acuerdo en que Linda se ajusta más a la idea de una «cajera de banco feminista» que al estereotipo de las cajeras de banco. La cajera de banco, a priori, no es una militante feminista, y añadir este detalle a la descripción hace que la historia sea más coherente.

Piénsese en términos de los diagramas de Venn. El conjunto de las cajeras de banco feministas está enteramente incluido en el conjunto de las cajeras de banco, pues cada cajera feminista es una cajera más. Por consiguiente, la probabilidad de que Linda sea una cajera feminista tiene que ser más baja que la de que sea una cajera de banco. Si especificamos un posible suceso con mayor detalle, solo podemos bajar su probabilidad. El problema crea así un conflicto entre la intuición y la lógica de la probabilidad.

Ej.

Nadal va a jugar un partido. Decir cuál es el suceso más probable

- A. Nadal perderá el primer set.
- B. Nadal perderá el primer set, pero ganará el partido.
- C. Nadal ganará el primer set, pero perderá el partido.

De nuevo, a primera vista, podríamos tender a afirmar que el suceso más probable es el B, pero B está contenido en el suceso A (A es B con una condición más). Así que la probabilidad de B es menor o igual que la de A (la opción C, que Nadal pierda, no se contempla, desde luego).

P<sub>3</sub>

Si  $A$  y  $B$  son sucesos cualesquiera

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

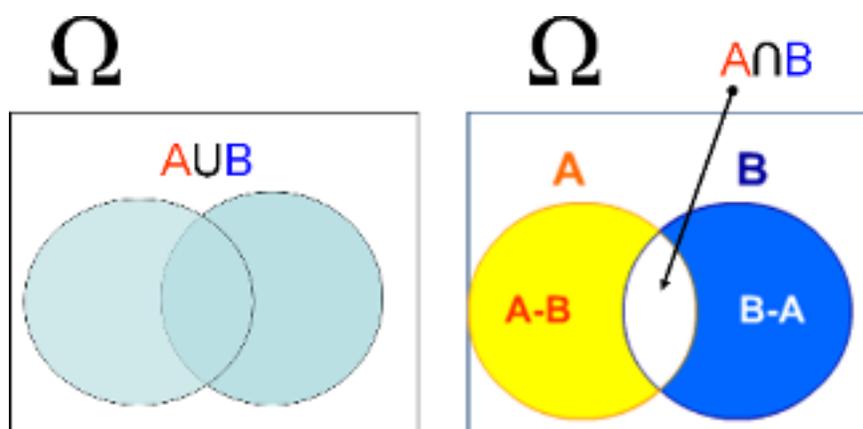


Figura 4.20: Diagrama de Venn de la unión e intersección de sucesos: la probabilidad de la unión es la suma de la probabilidad de  $A$  y la probabilidad de  $B$ , pero hay que restar la probabilidad de la intersección, puesto que, en otro caso, la contaríamos dos veces.

Ej.

Si la probabilidad de que llueva el sábado es 0.6, y la de que llueva el domingo es 0.7, ¿cuál es la probabilidad de que llueva el fin de semana?

La probabilidad de que llueva el fin de semana es que llueva el sábado o el domingo, o que llueva los dos días:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.7 - P(A \cap B).$$

Como vemos, si sólo consideramos el sábado y el domingo aisladamente, la probabilidad daría más que uno, algo imposible. La posibilidad de que lluevan los dos días y no uno solo hace que debamos restar la probabilidad de la intersección, que por ahora no sabemos calcular, y que veremos en breve.

““

Ej.

Un medicamento  $A$  para el dolor de estómago ofrece un 70 por ciento de probabilidades de eliminar el dolor con la toma de una pastilla. Otro medicamento  $B$  ofrece un 85 por ciento de éxito con la toma de una pastilla. Uno que se ha hinchado a comer y le dan retortijones decide

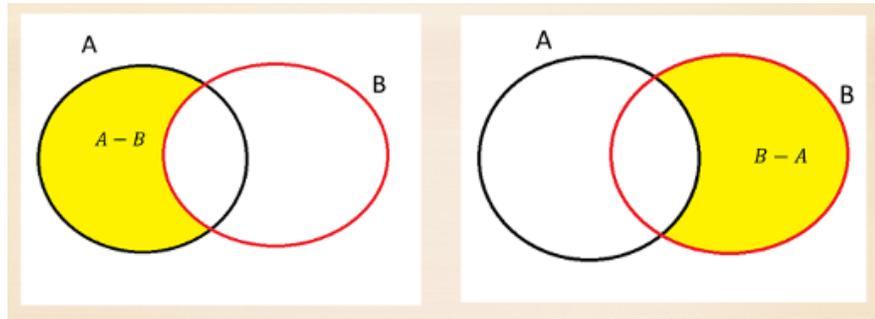


Figura 4.21: Diagrama de Venn de la diferencia de sucesos.

tomarse una pastilla de cada (los medicamentos no son incompatibles).  
¿Cuál es la probabilidad de que se le quite el dolor de estómago?



La probabilidad de la diferencia es

$$P(A - B) = P(A) - P(A \cap B)$$



En el último trimestre, se ha realizado la siguiente clasificación de pacientes de una clínica en función de tener o no artrosis.

	Artrosis	No Artrosis	Totales
Hombres	45	30	75
Mujeres	55	15	70
Totales	100	45	145

La probabilidad de que un paciente tenga artrosis es:  $P(A) = 100/145$ , y de que no la tenga es  $P(\bar{A}) = 1 - (100/145)$

La probabilidad de que un paciente tenga artrosis o sea una mujer es

$$P(A \cup M) = P(A) + P(M) - P(A \cap M) = \frac{100}{145} + \frac{70}{145} - \frac{55}{145}$$

Probabilidad de que un paciente sea hombre y no tenga artrosis:

$$P(H - A) = P(H) - P(H \cap A) = \frac{75}{145} - \frac{45}{145} = \frac{30}{145} = P(H \cap \bar{A})$$

## 4.12 Probabilidad condicionada

En la incertidumbre asociada a los experimentos aleatorios puede existir un cierto dinamismo, ya que la información adicional que vayamos obteniendo sobre un proceso puede modificar las probabilidades de los sucesos.

Retomemos este ejemplo donde considerábamos la clasificación de hombres y mujeres en función de padecer o no artrosis:

	Artrosis	No Artrosis	Totales
Hombres	45	30	75
Mujeres	55	15	70
Totales	100	45	145

Entra en la consulta una mujer. ¿Cuál es la probabilidad de que tenga artrosis?

Esto se escribe  $P(A | M)$  y se calcula como siempre, como casos favorables entre posibles (55 mujeres con artrosis, entre 70 mujeres).

$$P(A | M) = 55/70$$

A partir de la tabla, podemos ver que la probabilidad del suceso intersección es

$$P(A \cap M) = 55/145$$

y que la probabilidad de ser mujer

$$P(M) = 55/145$$

Se ve rápidamente la relación:

$$P(A | M) = \frac{P(A \cap M)}{P(M)}$$

En general, **la probabilidad de un suceso  $A$  condicionado a otro suceso  $B$**  es



$$P(A | B) = \frac{P(A \cap B)}{P(B)},$$

Recuento		MENOPAUSIA		Total
		NO	SI	
CLASIFICACION OMS	NORMAL	189	280	469
	OSTEOPENIA	108	359	467
	OSTEOPOROSIS	6	58	64
Total		303	697	1000

Figura 4.22: Tabla de contingencia menopausia - osteoporosis.

donde se supone que la probabilidad del suceso  $B$  no puede ser cero.

De esta definición se deduce la forma general de calcular la **probabilidad de una intersección**:



$$P(A \cap B) = P(A) \cdot P(B | A) = P(B) \cdot P(A | B)$$

**Ejercicio 4.1.** De 1000 mujeres mayores de 40 años que acudieron a un centro de salud, se realizó la siguiente tabla para el estudio de la relación entre la presencia o no de osteoporosis y haber tenido la menopausia:

La probabilidad de tener osteoporosis es  $64/1000 = 0.064$ .

Ahora bien, es conocido que la probabilidad de **tener osteoporosis** es mayor si se **ha tenido la menopausia**. Entonces podríamos calcular

$$P(\text{Osteoporosis} | \text{Menopausia}) = 58/697 = 0.098$$

Si calculamos en cambio la probabilidad de **Haber tenido la Menopausia** condicionado a **Tener osteoporosis**:

$$P(\text{Menopausia} | \text{Osteoporosis}) = 58/64$$

Comprobamos lo que sucede de manera general:  $P(A | B)$  no coincide con  $P(B | A)$ .

Aunque a veces pueda ser fácil liarse. Veamos otro ejemplo:



Figura 4.23: Diagrama de venn Fumadores (1)

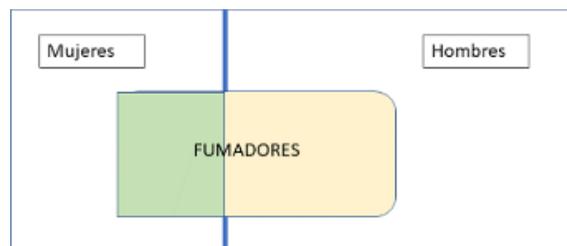


Figura 4.24: Diagrama de venn Fumadores (2)

**Ejercicio 4.2.** Un médico observa que el 30% de sus pacientes son mujeres y que el 40% de sus pacientes fuma. De los fumadores, el 75% son hombres. Calcular el porcentaje de mujeres que fuman entre los pacientes del doctor, y el porcentaje de fumadoras entre las mujeres.

**Solución:**

Primero; en caso de duda, diagrama de Venn:

Nos piden, primero, porcentaje de mujeres que fuman:  $P(M \cap F)$

Lo que nos piden es el área representada por el rectángulo verde (mujeres y a la vez fumadores)

Ahora

$$P(M \cap F) = P(F) \cdot P(M | F) = 0.4 \cdot (1 - 0.75) = 0.4 \cdot 0.25 = 0.1,$$

es decir el 10 por ciento.

El porcentaje de fumadoras entre las mujeres será el mismo rectángulo, pero considerado sólo en el grupo de mujeres.



Figura 4.25: Diagrama de venn Fumadores (3)

$$P(F | M) = \frac{P(F \cap M)}{P(M)} = \frac{0.1}{0.3} = 0.33.$$

Vemos que la probabilidad condicionada y la probabilidad de la intersección no son lo mismo.

Por otro lado,

$$P(F | M) \neq P(M | F)$$

La primera probabilidad nos da el porcentaje de mujeres que fuman (en el grupo de mujeres como TOTAL) y la segunda nos da el porcentaje de mujeres que hay en el grupo de fumadores, y sería:

$$P(M | F) = \frac{P(M \cap F)}{P(F)} = \frac{0.1}{0.4} = 0.25.$$



Puede demostrarse que la probabilidad condicionada cumple todas las propiedades vistas de la probabilidad. Esto es, si  $E$  es cualquier suceso con probabilidad mayor que cero:

- $P(\bar{A} | E) = 1 - P(A | E)$
- $P(A \cup B | E) = P(A | E) + P(B | E) - P(A \cap B | E)$
- Si  $A \subset B$ , entonces  $P(A | E) \leq P(B | E)$

$$\bullet P((A - B) | E) = P(A | E) - P((A \cap B) | E)$$

**Ejercicio 4.3.** En las bodas de Caná, el 70% de los presentes eran mujeres. De ellas el 10% eran católicas. De los hombres, eran católicos el 20%. a) ¿Qué porcentaje de católicos había? b) Se elije a una persona al azar y es católica. ¿Cuál es la probabilidad de que sea un hombre?

**Solución:**

Con los datos, creamos una tabla de doble entrada:

	Católicos	No Católicos	Totales
Hombres	$0.2 \cdot 0.3$		
Mujeres	$0.7 \cdot 0.10$		0.7
Totales			1

Rellenamos los huecos:

	Católicos	No Católicos	Totales
Hombres	$0.2 \cdot 0.3$	0.24	0.3
Mujeres	$0.7 \cdot 0.10$	0.63	0.7
Totales	0.13	0.87	1

Mirando a la tabla respondemos

1.  $P(C) = 0.13$
2.  $P(H | C) = P(H \cap C)/P(C)$  y estas son  $0.06/0.13 = 0.46$

## 4.13 Independencia de sucesos

Dos sucesos son **independientes** si la ocurrencia de uno de ellos no modifica la probabilidad del otro. Matemáticamente  $P(A | B) = P(A)$  y  $P(B | A) = P(B)$ .



En consecuencia, si dos sucesos son independientes

$$P(A \cap B) = P(A) \cdot P(B)$$

Ej.

Con la tabla de clasificación de osteoporosis y menopausia utilizada anteriormente, ¿son independientes los sucesos “Menopausia” y “Osteoporosis”?

Podemos comprobar que la probabilidad de la intersección no coincide con el producto de las probabilidades.

$$P(A \cap B) = \frac{58}{1000} = 0.058,$$

mientras que

$$P(A) \cdot P(B) = \frac{697}{1000} \cdot \frac{64}{1000} = 0.045$$

### 4.13.1 Regla del producto

Se conoce como **regla del producto** a la fórmula que establece la probabilidad de la intersección de, en general,  $n$  sucesos:

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_n) &= \\ &= P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot P(A_4 | A_1 \cap A_2 \cap A_3) \cdot \dots \\ &\quad \dots \cdot P(A_n | A_1 \cap A_2 \cap \dots \cap A_{n-1}), \end{aligned}$$

que, como vemos, consiste en que la probabilidad de la intersección de  $n$  sucesos se calcula multiplicando las probabilidades sucesivas, pero condicionando cada suceso a que ocurran todos los inmediatamente anteriores.

Si los sucesos son **independientes**, la ocurrencia de cualquier suceso no influye en la probabilidad de ocurrencia del resto de sucesos, de manera que la fórmula es mucho más fácil de recordar:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_n)$$



Figura 4.26: Mayor dependencia, imposible.

**Ejercicio 4.4.** Mientras los adolescentes españoles se divierten los fines de semana practicando el noble arte del botellón, los rusos juegan a la ruleta que lleva su nombre (ruleta rusa). Por si existe alguien que no lo sepa, el juego consiste en meter una bala en un revólver, girar el tambor a lo loco y apuntarse a la cabeza. Cada vez que se dispara (y no hay bala), el tambor gira una posición, de manera que la bala se irá acercando más a la posición de disparo. Gana el que se muere (es que los rusos son todos igual de valientes que Putin).

Supongamos que nos toca jugar con otros cinco jugadores (los revólveres que se usan suelen tener seis balas). ¿En qué posición debemos colocarnos para jugar y tener la mayor (o menor) probabilidad de que salga la bala cuando nos toque dispararnos?

Independientemente del cálculo de las probabilidades de supervivencia de cada uno de los jugadores, podemos realizar un experimento de simulación:

Cada uno de los jugadores, desde el primero al último en disparar, tiene asignado un número del 1 al 6. Cuando se introduce la bala en la recámara del revólver y se hace girar el tambor, la bala queda colocada en la posición 1,2,3,4,5 o 6. Como se va a ir disparando sucesivamente (hasta que salga la bala), el jugador cuya posición coincida con la posición de la bala va a ser el que reciba el disparo.

Así pues, en vez de realizar un experimento con jugadores vivos y balas (que además de necesitar muchas personas y ataúdes, sería ilegal; no sé en Rusia pero aquí sí), podemos hacerlo con R. Cada juego de la ruleta rusa es tan

sencillo como lanzar un dado, y el número del 1 al 6 que salga es la posición en que queda la bala, y por tanto el número del jugador que va a recibir el disparo. Por lo tanto, podemos replicar el juego 1000 veces, por ejemplo, igual que hemos hecho antes con el lanzamiento de un dado.

```
x=sample(1:6,1000, replace=T)
y=table(x)
pander(y)
```

1	2	3	4	5	6
154	161	176	180	164	165

Como vemos, todos los jugadores, a la larga, tienen la misma probabilidad de recibir un disparo

### Solución numérica

Llamemos  $M_i$ =el jugador en posición  $i$  recibe el disparo (muere) y  $A_i$ =la bala sale en el disparo  $i$ .

$$P(\text{el primero muere}) = P(M_1) = P(A_1) = 1/6.$$

$$P(\text{el segundo muere}) = P(M_2) = P(\overline{A_1} \cap A_2) = P(\overline{A_1}) \cdot P(A_2 | \overline{A_1}) = \frac{5}{6} \cdot \frac{1}{5} = \frac{1}{6}.$$

$$P(M_3) = P(\overline{A_1} \cap \overline{A_2} \cap A_3) = P(\overline{A_1}) \cdot P(\overline{A_2} | \overline{A_1}) \cdot P(A_3 | \overline{A_1} \cap \overline{A_2}) = \frac{5}{6} \cdot \frac{4}{5} \cdot \frac{1}{4} = \frac{1}{6}.$$

De igual manera se calculan las otras probabilidades, y todas son iguales a  $1/6$ .

---

## 4.14 Regla de Bayes

Vimos antes que  $P(A | B)$  no coincide con  $P(B | A)$ . La regla de Bayes, también llamada *regla de la probabilidad inversa* establece la relación entre



Figura 4.27: Thomas Bayes, reverendo presbiteriano inglés (1702-1761).

estas probabilidades.

$$P(B | A) = \frac{P(A | B) \cdot P(B)}{P(A)}$$

La probabilidad  $P(B)$  se conoce como probabilidad *a priori* (ex ante) y la probabilidad  $P(B/A)$  se llama probabilidad *a posteriori* (ex post). En una relación causa-efecto, se trata de determinar la probabilidad de la causa, cuando se ha producido un determinado efecto.

Ej.

Los síntomas de un paciente (cara amarilla) pueden ser provocados por diferentes causas (anemia, virus, fiebre... ). Cuando se presentan los síntomas (efecto), interesa saber la probabilidad de que haya sido por una u otra causa.

Ej.

El mal funcionamiento de un ordenador puede deberse a diferentes causas: virus, fallos físicos, incompatibilidades de software... Las causas (virus...) ocasionan el efecto del mal funcionamiento. A partir de que se observa este efecto, se pretende conocer la probabilidad de que lo haya originado una causa concreta (sería calcular la probabilidad inversa de la causa, dado “o condicionado” al efecto).

Ej.

La película Rush (2013) nos narra la rivalidad existente en los años 70 entre los pilotos de carreras Niki Lauda y James Hunt.



Figura 4.28: Izquierda: Cartel de la película Rush (2013). Derecha: Los verdaderos Niki Lauda y James Hunt.

Supongamos que, en 4 carreras de Fórmula 1 entre Niki Lauda y James Hunt, Niki ganó 3 veces frente a 1 que ganó James Hunt. Se van a enfrentar en una próxima carrera. ¿Cuál es la probabilidad de ganar de Hunt?

*Claramente, esta probabilidad es  $1/4=0.25$ .*

Supongamos ahora que, una vez que ganó Niki Lauda, estaba lloviendo, y la vez que ganó Hunt también llovía. El pronóstico del tiempo dice que la probabilidad de que llueva en la próxima carrera es del 50 por ciento. ¿Cambia la probabilidad de ganar de James Hunt?

Sea  $H$ ="gana Hunt".  $A$ ="llueva".

$$P(H | A) = \frac{P(A | H) \cdot P(H)}{P(A)} = \frac{1 \cdot \frac{1}{4}}{\frac{1}{2}} = 0.5$$

Como vemos, al existir una probabilidad a priori, la probabilidad a posteriori (también llamada **probabilidad inversa**) varía, pasa a ser el doble.

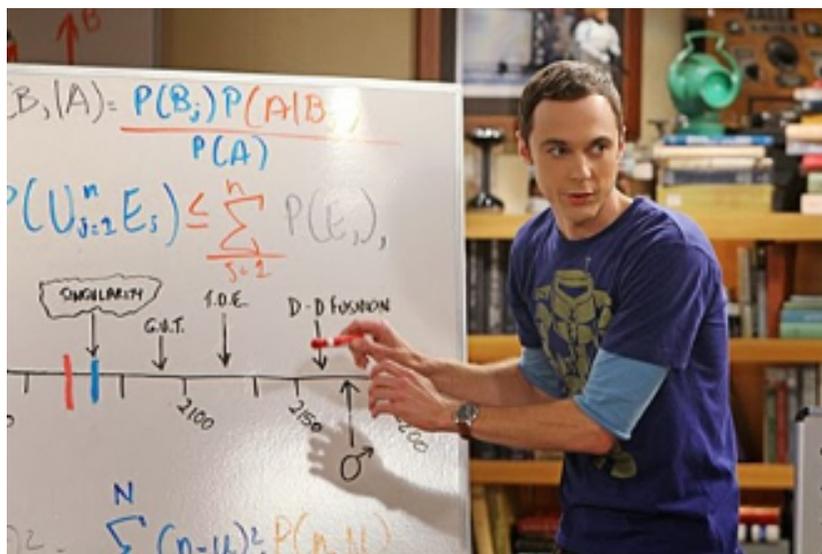


Figura 4.29: Hasta en la tele.

## 4.15 Teorema de la probabilidad total (Laplace)

Sean los sucesos  $A_1, A_2, \dots, A_n$  una partición del espacio muestral  $\Omega$ , y sea  $B$  un suceso cualquiera.

Se cumple que la probabilidad del suceso  $B$  puede expresarse en función de los sucesos  $A_i$  de la siguiente manera (fórmula de la probabilidad total):

$$P(B) = P(B | A_1) \cdot P(A_1) + P(B | A_2) \cdot P(A_2) + \dots + P(B | A_n) \cdot P(A_n)$$

**Ejemplo 4.1.** La tienda online favorita del 33 por ciento de los socios de un foro es Pccomponentas, un 8 por ciento prefiere Medianamart, el 2 por ciento prefiere Evoy y el resto prefieren comprar en Amazonas. La probabilidad de que el pedido se pierda y no llegue al destinatario, según la casa que lo envíe, es 0.8, 0.9, 0.7 y 0.6 respectivamente. Pepe, forero del 2003, ha pedido el último modelo de linterna. Si acaba de entrar al foro y el primer hilo que abre ya es para insultar, ¿crees que está cabreado porque no ha recibido el envío?

**Solución**

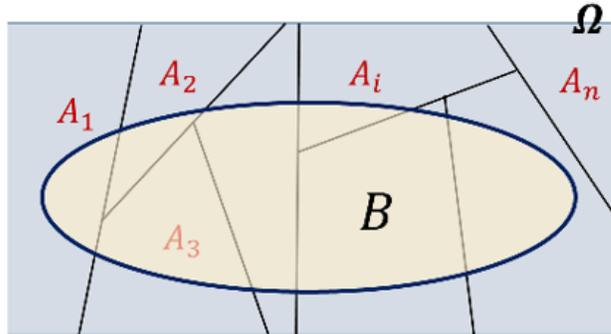


Figura 4.30: Partición: sucesos disjuntos que recubren el total.

Definimos los sucesos:  $A_1$ ="pedido a Pccomponendas",

$A_2$ ="pedido a Medianamart",

$A_3$ ="pedido a Evoy" ,

$A_4$ ="pedido a Amazonas",

y  $B$ ="el pedido no llega a tiempo".

Se tiene que

$$P(A_1) = 0.33, P(A_2) = 0.08, P(A_3) = 0.02, P(A_4) = 0.57.$$

Como vemos, los sucesos  $A_1, A_2, A_3$  y  $A_4$  son incompatibles y sus probabilidades suman 1, por lo que cumplen las hipótesis del teorema de las probabilidades totales.

Nos dicen, además, que  $P(B | A_1) = 0.8, P(B | A_2) = 0.9, P(B | A_3) = 0.7, P(B | A_4) = 0.6$ .

Por el teorema de las probabilidades totales, la probabilidad de que el pedido no se ha recibido es  $P(B) =$

$$\begin{aligned} P(B | A_1) \cdot P(A_1) + P(B | A_2) \cdot P(A_2) + P(B | A_3) \cdot P(A_3) + P(B | A_4) \cdot P(A_4) = \\ = 0.8 \cdot 0.33 + 0.9 \cdot 0.08 + 0.7 \cdot 0.02 + 0.6 \cdot 0.57 = 0.692. \end{aligned}$$

Vemos que la probabilidad de que Pepe no haya recibido el envío es más alta que la probabilidad de que sí lo haya recibido, por lo tanto comprendemos su enfado.

## 4.16 Teorema de Bayes (Versión de Laplace)

Disponemos de una partición  $A_1, A_2, \dots, A_n$  y de las probabilidades  $P(A_i)$ , conocidas como **probabilidades a priori**.

El objetivo es calcular en qué medida el conocimiento de un suceso  $B$  modifica las probabilidades a priori, dando lugar a las probabilidades a posteriori  $P(A_i | B)$ . Reciben este nombre porque se calculan una vez obtenida la evidencia empírica  $B$ .

Sean los sucesos  $A_1, A_2, \dots, A_n$  una partición de  $\Omega$ , y sea  $B$  un suceso cualquiera (las mismas hipótesis del teorema de la probabilidad total). Se cumple

$$\begin{aligned} P(A_i | B) &= \frac{P(B | A_i) \cdot P(A_i)}{P(B)} = \\ &= \frac{P(B | A_i) \cdot P(A_i)}{P(B | A_1) \cdot P(A_1) + \dots + P(B | A_n) \cdot P(A_n)} \end{aligned}$$

Como vemos, se trata de aplicar la fórmula de la probabilidad condicionada y, a continuación, aplicar en el denominador la fórmula de la probabilidad total.

En el año 1774, Laplace desarrolló (de forma independiente de Bayes) este teorema, que en todos los textos aparece como teorema de Bayes, en honor a Thomas Bayes. En base a este teorema, Laplace llegaría a concluir que un dato bien contrastado —el de que nacen más niños que niñas— tenía que emanar, casi con toda certeza, de una ley natural. En la gran mayoría de textos de enseñanza de la estadística, se llama a este teorema el teorema de Bayes, aunque en realidad sea de Laplace.

**Ejemplo 4.2.** En una clase el 70% de los alumnos son mujeres. De ellas, el 10% son fans de Pitingo. De los varones, son fans de Pitingo el 20%. ¿Qué porcentaje de fans de Pitingo hay en total?

**Solución:**

Empezamos construyendo una tabla con los datos que nos dan:

	Fan	No Fan	Totales
Hombres	$0.2 \times 30 = 6$		
Mujeres	$0.10 \times 70 = 7$		70
Totales			100

Completamos:

	Fan	No Fan	Totales
Hombres	6	24	30
Mujeres	7	63	70
Totales	13	87	100

$$P(F) = 0.7 \cdot 0.1 + 0.3 \cdot 0.2 = 0.13$$

Ahora se elije a un alumno/a al azar y va escuchando a Pitingo. ¿Probabilidad de que sea un hombre?

$$P(H/F) = \frac{P(F/H) \cdot P(H)}{P(F)} = \frac{0.2 \cdot 0.3}{0.13} = 0.46$$

**Ejercicio 4.5.** Un estudiante, después de una semana terrible de exámenes, noches sin dormir, cafés, red-bull, alguna que otra sustancia no muy legal y muchos whatsapps de antiguos amantes, se encuentra con graves desvaríos mentales. Decide ir a un psiquiatra, quien, por sus síntomas, cree que puede tener esquizofrenia, a pesar de no tener ningún antecedente familiar de tal enfermedad.

Para comprobarlo, decide aplicarle un nuevo test con una alta eficacia: si una persona es esquizofrénica, lo detecta con probabilidad 0.99. Si no lo es, lo dice con probabilidad 0.98. El psiquiatra realiza el test y da positivo. Sabiendo que la prevalencia de la enfermedad en la población es del uno por ciento, ¿debería el estudiante preocuparse o no preocuparse demasiado?

**Solución:**

Llamemos  $E$ ="ser esquizofrénico". Nos dicen que la proporción de esquizofrénicos en la población es del 1 por ciento, es decir,  $P(E) = 0.01$ .

Por otro lado, el test realizado verifica que:  $P(+ | E) = 0.99$ , y  $P(- | \bar{E}) = 0.98$

Por el teorema de Bayes:

$$P(E | +) = \frac{P(+ | E) \cdot P(E)}{P(+)},$$

y el denominador lo calculamos mediante el teorema de la probabilidad total:

$$P(+) = P(+ | E) \cdot P(E) + P(+ | \bar{E}) \cdot P(\bar{E})$$

Entonces (usamos que  $P(+ | \bar{E}) = 1 - P(- | \bar{E})$ )

$$P(E | +) = \frac{0.99 \cdot 0.01}{0.99 \cdot 0.01 + (1 - 0.98) \cdot 0.99} = \frac{0.0099}{0.0099 + 0.0198} = \frac{0.0099}{0.0297} = 0.33.$$

Podemos plantearlo con una tabla:

	Esquizofrenia	No Esquizofrenia	Totales
Test Positivo	$0.99 \cdot 0.01$		
Test Negativo		$0.98 \cdot 0.99$	
Totales	0.01		1

Completamos la tabla:

	Esquizofrenia	No Esquizofrenia	Totales
Test Positivo	0.0099	0.0198	0.0297
Test Negativo		0.9702	
Totales	0.15	0.99	1

Como vemos, en realidad, la probabilidad de que tenga esquizofrenia, aún habiendo dado positivo en el test, es solo de 0.33. lo tanto, es mucho más probable que no la tenga (1-0.33). Esto se debe a que la probabilidad a priori (en este caso es la prevalencia de la enfermedad) es muy baja.

Observamos dos resultados importantes:

- Por un lado, resaltar que no es lo mismo  $P(A/B)$  que  $P(B/A)$ , por ejemplo comparemos  $P(+ | E)$  con  $P(E | +)$ .
- En un test o prueba diagnóstica, se conoce como **sensibilidad** a la probabilidad de clasificar correctamente a un individuo enfermo  $P(+ | E)$ . Se llama **especificidad** a la probabilidad de clasificar correctamente a un individuo sano  $P(- | \bar{E})$ .
- Independientemente de lo bueno que sea el test, si tiene un margen de error pequeño (como en el ejemplo que hemos visto), la probabilidad REAL de que el individuo tenga o no la enfermedad está directamente

relacionada con la prevalencia de la enfermedad en la población. Así pues, cuando alguien se hace un test que dice que puede tener cáncer u otra enfermedad grave, la preocupación puede dejarse en un segundo plano hasta el momento en que se haga una segunda prueba que diga definitivamente si la persona está enferma o no (biopsia, etc.)

**Ejercicio 4.6.** Una señora de 65 años, a la que le han dado vez en el seguro (para hacerse una resonancia) para dentro de 3 años, decide acudir a fisioterapia.

La señora tiene dificultad y dolor al mover una pierna. La fisioterapeuta, después de un análisis concienzudo según aprendió en la UDC, supone que el problema puede deberse a:

Artritis ( $A_1$ ), ó sobrecarga por bailar zumba el sábado noche ( $A_2$ ), con probabilidades 0.7 y 0.3 respectivamente.

La fisioterapeuta ve que lo mejor es aplicar un tratamiento estándar, que tiene probabilidad de mejora:  $P(M | A_1) = 0.9$  y  $P(M | A_2) = 0.7$ .

Después del tiempo establecido, la paciente ha mejorado. ¿Cuál es la probabilidad de que tuviese artritis? Si no mejoró, ¿probabilidad de que el problema no fuese por artritis?

---

## 4.17 Notas históricas. La estadística Bayesiana

Sacado del libro *La teoría que nunca murió* (McGrayne, 2012).

La regla de Bayes aparece escrita en el ensayo *An Essay Towards Solving a Problem in the Doctrine of Chances* (*Un ensayo hacia la solución de problemas en la disciplina del azar*). Este ensayo puede leerse al español traducido por M. Villegas (Villegas, 2001).

Este trabajo lo escribió Thomas Bayes en la década de 1740, en medio de una polémica religiosa marcada por la siguiente cuestión: ¿es posible establecer conclusiones racionales relativas a la existencia de Dios, sobre la base de las pruebas que nos proporciona el mundo a nuestro alrededor? Concretamente,

Bayes se plantéo si es posible conciliar la presencia del mal con la presunta bondad divina. En 1731, Bayes escribió su ensayo, en el que dice (entre otras muchas cosas) que Dios otorga a las personas «la máxima felicidad que les es dado alcanzar». Resulta curioso, primero, que haya sido un sacerdote quien creara una de las reglas del azar más utilizada a lo largo de la historia. Segundo, contrasta los sucesos de la existencia de Dios con la evidencia empírica que podría negar su existencia.

Cuando el reverendo Bayes falleció, sus parientes solicitaron a Richard Price, buen amigo del difunto, que revisara sus trabajos matemáticos. Price, que también era ministro de la Iglesia presbiteriana y matemático, alcanzaría posteriormente una notable fama como abogado de las libertades civiles y defensor de las revoluciones estadounidense y francesa. Personajes tan ilustres como Benjamin Franklin, John Adams y Thomas Jefferson (segundo y tercer presidente de Estados Unidos, respectivamente); la feminista Mary Wollstonecraft, que acostumbraba a acudir a su iglesia; el reformador del sistema penitenciario John Howard, que era su mejor amigo; y Joseph Priestley, el descubridor del oxígeno, fueron admiradores de su personalidad y conocimientos científicos.

En 1781, la Universidad de Yale concedió un doctorado honoris causa a Price, simultáneamente que a George Washington. Sin embargo, el nombre de Price es conocido, fundamentalmente, por sacar a la luz la regla de Bayes. Al empezar a estudiar los papeles de Bayes que le había dejado su familia, Price se encontró frente a *una imperfecta solución a uno de los problemas más difíciles de la doctrina de las probabilidades*. Se refería al ensayo que Bayes había compuesto a fin de indagar en la probabilidad de las causas, y, a partir de las observaciones relativas al mundo real, tratar de encontrar su causa más probable.

Price comprendió que el ensayo era la respuesta al ataque que David Hume (Edimburgo; 1711- 1776) había dirigido contra la *teoría de la causalidad*. De manera muy breve, podemos decir que toda predicción – consciente o inconsciente, científica o propia de la vida cotidiana – se funda en un razonamiento acerca de cuestiones de hecho, y este razonamiento depende por entero de la fe en el principio de causalidad. Cuando comemos un trozo de pan, no nos paramos a pensar en que el pan podría estar envenenado si no lo hemos sospechado, o que al caminar bajando unos escalones la tierra vaya a abrirse bajo nuestros pies. La sucesión natural de hechos a la que, cotidianamente,



Figura 4.31: El filósofo, economista, sociólogo e historiador David Hume.

nos vamos acostumbrando (a una cosa le sigue otra, a la noche siempre le sucede el día porque nunca hubo -que se sepa- un día en que no saliera el sol) es la base de la experiencia que, a la vez, condiciona a la razón.

Y la cuestión fundamental que nos afecta era que, bajo dicha teoría, se mantenía la existencia de Dios como algo necesario, puesto que era la causa de, por ejemplo, la ocurrencia de los milagros históricamente señalados como ciertos por el cristianismo. El más destacado, la resurrección de Jesús. Hume (en innegable atrevimiento para la época), en su trabajo “On Miracles”, afirmaba que no debía darse validez a los reportes sobre milagros, puesto que no existía evidencia científica sobre los mismos y, por lo tanto, no debía tenerse por cierta la resurrección de Jesús, por mucho que existieran testimonios que lo aseverasen. En ningún caso debía irse más allá de la experiencia, y así, los conceptos metafísicos (Dios, mundo, alma) no tenían por qué ser inexorablemente ciertos.

En un trabajo dirigido a la Real Sociedad de Londres para el avance de la Ciencia, Price señaló que la regla de Bayes intentaba demostrar, matemáticamente, que «el mundo debía ser necesariamente una consecuencia de la sabiduría y el poder de una causa inteligente, de modo que con él se viene a confirmar [...] partiendo de las causas finales [...] la existencia de un Ser Divino». El trabajo se tituló: “A method of calculating the exact probability

of all conclusions founded on induction”.

Hay que tener en cuenta que el ensayo de Bayes fue escrito bastantes años antes de que Laplace escribiera su famosa regla de los casos favorables entre casos posibles. La fórmula de la probabilidad total (y el teorema posterior) también son posteriores a la regla de Bayes. Es decir, Bayes dedujo su regla sin conocer no solo esto, sino la axiomática de Kolmogorov y las propiedades que se deducen de ella (que es como nosotros hemos calculado la probabilidad condicionada). En particular, Hume hacía referencia a los testimonios históricos existentes que apoyaban la resurrección de Cristo (los cuales existen y pueden buscarse en internet).

Utilizando el resultado formulado por Bayes, y con argumentos matemáticos del estilo de los utilizados por Bayes para su demostración, Richard Price concluyó lo siguiente: supongamos que  $p$  sea la probabilidad de que un hecho suceda en cada uno de  $n$  ensayos independientes ( $p$  podría ser la probabilidad de que ocurra un milagro en  $n$  días consecutivos). Supongamos que este hecho ha ocurrido  $x$  veces. La distribución a posteriori para  $p$  bajo el supuesto de que todos los valores de  $p$ , a priori, son igualmente probables, permite aceptar que, aunque **el número  $x$  sea cero**, la probabilidad  $p$  puede suponerse más grande que *uno entre un millón seiscientos mil* con probabilidad 0.533.

Este resultado no es fácil de obtenerse (por ello no lo exponemos) pero matemáticamente es correcto. Viene a decirnos que la probabilidad de un milagro en un día concreto (aunque nunca se haya observado ninguno) es más grande que ese número (1/1.600.000) con probabilidad algo mayor que 1/2. Permite establecer una base matemática para la probabilidad de ocurrencia de un hecho muy anormal en la naturaleza:

Si  $p$  es la probabilidad de que un suceso (“milagro”) ocurra (tal como que mañana no salga el sol, en palabras de Price, o la resurrección de una persona), admitamos que  $p$  pueda ser 1/1600000. Esto es, admitimos que un milagro puede suceder, aunque con una probabilidad muy pequeña. **Ahora nos planteamos: ¿Cuál es la probabilidad de que ocurra al menos un suceso de este tipo en un millón de ensayos independientes? (en el próximo millón de días, ¿cuál es la probabilidad de que ocurra algún milagro? -uno, al menos).**

La probabilidad de que ocurra algún suceso puede calcularse como 1 menos

la probabilidad del complementario (“no ocurra ningún suceso”), es decir

$$1 - \left(1 - \frac{1}{1.600.000}\right)^{1.000.000} = 1 - \left(\frac{1.599.999}{1.600.000}\right)^{1.000.000}$$

y este número es, aproximadamente, 0.46, lo cual no es tan pequeño.



Lo que se ha probado es que la probabilidad de que ocurra un suceso muy raro (probabilidad muy pequeña) es bastante alta (próxima a un medio, en este caso) siempre que se espere el tiempo suficiente. Obviamente, esto también sirve para argumentos que favorecen las supersticiones o le dan credibilidad a programas de televisión donde relatan sucesos paranormales.



(extraído del libro “el hombre anúmerico” de John Paulos (Paulos and Llosa, 1990)) Supongamos que la probabilidad de un sueño profético sea de  $1/10.000$  (muy poco frecuente; cambiese este número por otro más pequeño si se quiere). *Entendemos por sueño profético aquel donde alguien sueña con su tía y la ve muriendo en un accidente de coche y, efectivamente, muere en accidente de coche, etc.* En una población con 10 millones de personas adultas, ¿cuántas tendrán al menos un sueño profético al año?

### Solución:

Según suponemos en el enunciado, la probabilidad de que un sueño NO sea profético es muy grande:  $9.999/10.000$ .

Por consiguiente, la probabilidad de que una persona tenga al menos un sueño profético al cabo de un año es

$$\begin{aligned} 1 - P(\text{no tenga ninguno}) &= 1 - \left(\frac{9999}{10000}\right)^{365} = \\ &= 1 - (0.9999)^{365} = 1 - 0.9641 = 0.0359. \end{aligned}$$

Supongamos una población con 10 millones de adultos. Cada uno de ellos tiene probabilidad 0.0359 de tener al menos un sueño profético al año. Es



Figura 4.32: Un caso real de sueño profético [fuente: digital].

decir, el 3.59 por ciento de los adultos (que son 3.590.000 adultos) tendrán al menos un sueño profético al año (lo cual, ciertamente, es un número muy alto).

Si se cambia la probabilidad inicial  $1/10.000$  por un número más pequeño, el porcentaje de personas con, al menos, un sueño profético al año, evidentemente disminuirá, pero seguirá siendo muy alto. Es por eso que en la prensa se recogen situaciones (verídicas) de gente que ha soñado con algo que se ha cumplido (accidentes de avión, etc.). Es evidente que somos muchas personas a soñar y se pueden tener muchos sueños diferentes. La probabilidad de que se cumpla algún sueño profético es alta, pero eso no significa que la persona que ha soñado algo que se ha cumplido tenga propiedades extrasensoriales.

**Caso real:**

#### 4.17.1 Aprendiendo de la experiencia: uso recursivo de la fórmula de Bayes

El teorema de Bayes y la fórmula de la probabilidad inversa no es sólo una fórmula. El uso de esta fórmula de una manera encadenada permite contemplar la probabilidad como una forma de conocimiento en función de la experiencia que se va adquiriendo.

Ej.

Tras una relación sexual, una mujer sospecha que puede estar embarazada. Para estar segura de su estado compra un test del cual se conoce que tiene una eficacia del 90% en detectar embarazos. La mujer se realiza el test y obtiene un resultado positivo. ¿Cuál es la probabilidad de que dicha mujer esté embarazada? (suponemos que el test da falsos positivos el 50 por ciento de las veces, y que la probabilidad de quedarse embarazada tras una relación sexual es del 15 por ciento)

**Solución:**

Llamemos  $E$ ="estar embarazada". Nos dicen que

$$P(E) = 0.15, P(+ | E) = 0.9, P(+ | \bar{E}) = 0.5$$

Por el teorema de Bayes:

$$P(E | +) = \frac{P(+ | E) \cdot P(E)}{P(+)},$$

y el denominador lo calculamos mediante el teorema de la probabilidad total:

$$P(+)=P(+|E) \cdot P(E)+P(+|\bar{E}) \cdot P(\bar{E})$$

Entonces

$$P(E | +) = \frac{0.9 \cdot 0.15}{0.9 \cdot 0.15 + 0.15 \cdot 0.85} = \frac{0.135}{0.135 + 0.425} = \frac{0.135}{0.56} = 0.241.$$

Como en otras ocasiones, quizá sea posible entenderlo mejor mediante una tabla

	Embarazo	No Embarazo	Totales
Test Positivo	$0.9 \cdot 0.15$	$0.5 \cdot 0.15$	
Test Negativo			
Totales	0.15		1

Completamos la tabla:

	Embarazo	No Embarazo	Totales
Test Positivo	0.135	0.075	0.21
Test Negativo	0.015	0.775	0.79
Totales	0.15	0.85	1

- Ahora supongamos que la mujer, para confirmar su estado, se realiza un nuevo test de embarazo y obtiene nuevamente un resultado positivo. Con esta información adicional, ¿cómo cambian las conclusiones?

Calculemos ahora la probabilidad de estar embarazada, sabiendo que el nuevo test ha dado positivo (la probabilidad de embarazo “a priori” ha cambiado a 0.241):

$$\begin{aligned} P(E | +) &= \frac{P(+ | E) \cdot P(E)}{P(+ | E) \cdot P(E) + P(+ | \bar{E}) \cdot P(\bar{E})} = \\ &= \frac{0.9 \cdot 0.241}{0.9 \cdot 0.241 + 0.15 \cdot 0.759} = 0.364 \end{aligned}$$

Si se repite sucesivamente el test de embarazo, obteniendo resultados positivos, la probabilidad de embarazo sería:  $Test3 = 0.507$ ,  $Test4 = 0.649$ ,  $Test5 = 0.769$ ,  $Test6 = 0.857$ ,  $Test7 = 0.915 \cdots Test10 = 0.984$ .

#### 4.17.2 Ejemplos importantes de aplicación de la estadística Bayesiana

En el mencionado libro *La teoría que nunca murió* (McGrayne, 2012) se describen con detalle ejemplos reales de aplicación de métodos estadísticos basados en la regla de Bayes o, mejor dicho, en la Estadística Bayesiana. El nombre de estadística Bayesiana surgió como una especie de alternativa a la estadística llamada **frecuentista**, la cual, básicamente, se formó a partir de los trabajos de Ronald Fisher de inferencia matemática. La estadística Bayesiana parte del principio de que los parámetros que describen una población no son valores fijos e inmutables, sino que son variables aleatorias con una distribución *a priori* determinada. Asimismo, considera que las probabilidades de los sucesos pueden ser subjetivas, e irse modificando a medida que se van conociendo datos del problema en el que se trabaje (como en el ejemplo del test de embarazo del apartado anterior).

Fue precisamente Ronald Fisher uno de los mayores objetores a este tipo de pensamiento, por cuanto consideraba que las probabilidades deben calcularse, bien a través de la frecuencia relativa, bien por la regla de Laplace, y que los parámetros de las poblaciones no podían considerarse variables, sino valores fijos que había precisamente que estimar. Este hecho fue quizá lo que motivó que las técnicas basadas en el pensamiento bayesiano fuesen menos

difundidas e investigadas durante varias décadas. Además, la derivación de muchas técnicas de tipo bayesiano y la resolución de problemas asociados no pudieron llevarse a cabo hasta la existencia de ordenadores con capacidad de cálculo adecuada, por cuanto, en muchas ocasiones, conllevan aparejados problemas de tipo numérico importantes.

En realidad, la estadística bayesiana y la estadística frecuentista, si bien en un principio (primeras décadas del siglo XX) parecían formas de pensar contrapuestas, no lo son en absoluto. Hoy en día existe un campo de trabajo y de aplicación de la estadística bayesiana muy amplio, tanto a nivel teórico como aplicado, del cual puede uno hacerse idea sin más que teclear los términos “bayesian inference” en google.

A continuación, citamos algunas aplicaciones de la estadística bayesiana que adquirieron popularidad:

- El índice de paternidad: La fórmula descrita por Essen-Moller es una consecuencia del teorema de Bayes. Puede verse, por ejemplo, en ([Mickey et al., 1986](#)).
- En problemas legales (juicios) existen numerosos ejemplos de utilización del teorema de Bayes. Uno de los más conocidos es la llamada *filicia del fiscal*. [The Guardian: A formula for justice](#). Asimismo, existen numerosos ejemplos de aplicación del teorema de Bayes para la resolución de casos judiciales. Muchos de ellos pueden consultarse en el libro de Gabriel Ruiz Garzón *Condenados por la estadística* ([Ruiz-Garzón, 2015](#)).
- Búsqueda de barcos perdidos: [Búsqueda Bayesiana](#). Un ejemplo de búsqueda de este tipo fue la realizada para localizar cuatro bombas atómicas que el ejército americano perdió en el mar en Palomares (Almería), en 1966, tras el accidente aéreo de un B-52. El suceso terminó con el famoso baño del ministro Fraga y el embajador americano en la playa, para demostrar que no existía radioactividad alguna.

En el paquete de R **rSARP** está implementada la búsqueda bayesiana. El lector interesado encontrará en la viñeta (pequeño manual) del paquete anotaciones muy interesantes.

- Los filtros de spam del correo electrónico son un ejemplo sencillo de lo que se conoce como filtros bayesianos. Un gestor de correo electrónico



Figura 4.33: Fraga saludando a los cientos de estadísticos bayesianos que se agolparon en la playa, y que la censura franquista impidió ver.

debe decidir, a partir de un efecto (el mensaje de correo que contiene cosas raras, como caracteres extraños, mensajes en lengua extranjera, declaraciones de amor de desconocidos...) si la causa es un mensaje o no de spam (y filtrar o no dicho correo). Para ello, el gestor irá “aprendiendo” de los datos que el usuario le va enseñando paulatinamente, con lo cual la probabilidad de acierto crece con el tiempo.

También, las **redes bayesianas** son un ejemplo de utilización del resultado de Bayes en la inteligencia artificial ([Castillo et al., 1997](#));([García, 2012](#)).

- Los métodos actuariales proceden de la aplicación del teorema de Bayes a los seguros, a través de los trabajos del italiano Bruno de Finetti (1906-1985) ([Bernardo, 1998](#)).
- El famoso caso Dreyfuss. [wikipedia](#) (1894–1906), referente a un militar judío francés injustamente acusado de espionaje y encarcelado en la isla del diablo. En una revisión de las pruebas presentadas contra el militar (entre cuyos testigos favorables destacó el matemático Henri Poincaré) se utilizó la regla de Bayes en varias ocasiones ([Ruiz-Garzón, 2015](#)). En la sección [El caso Dreyfuss](#) detallamos este caso.
- La descodificación de los códigos generados por la máquina alemana



Figura 4.34: El estadístico italiano Bruno de Finetti.

Enigma en la II guerra mundial: en Blentchley Park, un equipo de científicos dirigido por el matemático Alan Turing (considerado el padre de la informática) utilizó la regla de Bayes de manera continuada, parece que sin haberla conocido previamente. [BlogElPais](#) ([Mardia and Cooper, 2016](#)) ([Good, 1979](#))

- La invasión alemana de Rusia en la II guerra mundial: como Kolmogorov usó la regla de Bayes para conseguir mejorar la precisión de la artillería rusa contra el ejército alemán que estuvo a punto de entrar en Moscú ([Kolmogorov and Hewitt, 1948](#)).
- Cálculo del riesgo de explosión de una bomba nuclear por accidente durante la guerra fría. ([Barrett et al., 2013](#))

Estos y más ejemplos son citados y ampliamente comentados y documentados en el libro de Sharon McGrayne ([McGrayne, 2012](#)).

## 4.18 La falacia del fiscal

Siempre nos ha parecido que la estadística y las probabilidades, o las matemáticas en general, no guardan ninguna relación con el derecho y los tribunales

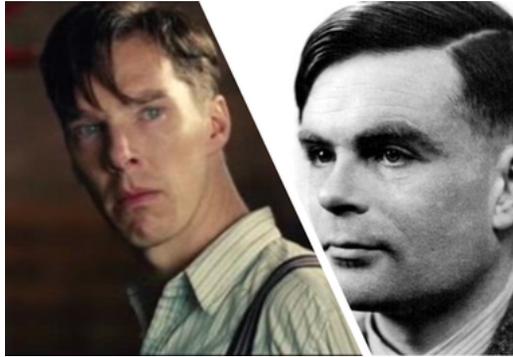


Figura 4.35: Izquierda: Benedict Cumberbatch en el papel de Alan Turing en la película "The imitation game" (2015). Derecha: el verdadero Alan Turing.

de justicia. Precisamente, el concepto de probabilidad condicionada es algo que, por lo que veremos, los expertos en justicia deberían conocer en profundidad, para que no se den casos de condenados injustamente (o también injustamente puestos en libertad). El citado caso Dreyfuss es uno de ellos. Veremos algunos más que han sido muy destacados en la historia de los errores judiciales, y como las probabilidades tuvieron mucho que ver en el desarrollo de los correspondientes procesos.

Supongamos que se ha cometido un asesinato y que el autor ha dejado algún tipo de evidencia en la escena del crimen como, por ejemplo, una mancha de sangre en la alfombra.

Supongamos que, atendiendo a ciertos marcadores bioquímicos, la sangre encontrada en la escena del crimen es de un tipo tal que sólo la sangre de una de cada 1000 personas coincide con ella.

Tenemos un sospechoso (persona con antecedentes policiales, que se encontraba cerca del lugar del suceso el día de autos) cuya sangre coincide con la encontrada en la escena del crimen, que es acusado del asesinato y llevado a juicio. El fiscal, durante el juicio, asegura lo siguiente:

“La probabilidad de que la sangre de un inocente coincida con la de la escena del crimen es de 1 entre 1000. La sangre del acusado coincide con la de la escena del crimen. Entonces, la probabilidad de que sea inocente es 0.001, es decir, es culpable con probabilidad 0.999”.



Figura 4.36: El implacable fiscal.

Esta aseveración, que puede sonar convincente e influir decisivamente en un juez o un jurado es, sencillamente, falsa.

Veamos por qué. Imaginemos que la población de posibles autores del crimen es de 100.000 personas, y que hay por tanto 100 personas cuya sangre coincide con la de la escena del crimen, uno de ellos el asesino.

Denotemos por  $S$  el suceso “poseer el tipo de sangre del autor del crimen”, y construyamos, con los datos que tenemos, la siguiente tabla de la población:

	C (Culpable)	No Culpable	Totales
S	1	99	$100000 \cdot \frac{1}{1000} = 100$
No S	0	99.900	99.900
Totales	1	99.999	100.000

$$P(\text{Culpable} \mid \text{datos}) = P(C \mid S) = \frac{P(C \cap S)}{P(S)} = \frac{1/100000}{100/100000} = \frac{1}{100}$$

Luego

$$P(\text{Inocente} \mid \text{datos}) = 1 - 0.01 = 0.99$$

Como vemos, la probabilidad de ser inocente no es directamente 1 sobre 1000, que podría pensarse a la luz de que ese es el porcentaje del tipo de sangre en la población general. Lo que hay que hacer es ver cuál es realmente ese porcentaje **dentro de la población de posibles sospechosos**. En este caso hemos considerado una ciudad de 100.000 personas, con lo cual la población

de posibles sospechosos (gente en esa ciudad con ese tipo de sangre) tiene  $100.000 \cdot \frac{1}{1000} = 100$  elementos, es decir la probabilidad de ser culpable en base a ese marcador en la sangre es de una entre cien. Pensemos que, si la ciudad tuviese un millón de habitantes, esa cantidad, en vez de ser 100, sería de 1000, con lo que la probabilidad de ser culpable (en base a la sangre) variaría a una entre mil.

La cuestión fundamental estriba en que, en un caso judicial, no se puede considerar la probabilidad “a secas” de ser culpable o inocente. Esa probabilidad tiene que venir condicionada por la evidencia existente, es decir, las pruebas o datos. Cuando se detiene a una persona y se le lleva a juicio, tiene que haber unas evidencias en su contra lo suficientemente consistentes para que, precisamente, el juicio se lleve a cabo. Así pues, la probabilidad que se debe calcular es una probabilidad condicionada, es decir, la **probabilidad de ser culpable en función de los datos que existan** (o uno menos la probabilidad de ser inocente condicionado a los datos que existan). Y no sólo habría de tenerse en cuenta, en un ejemplo como el anterior, el tener el mismo tipo de sangre que el del autor del crimen, sino otras evidencias (animadversión contra el fallecido, amenazas, etc...). Ahora bien, está claro que un tribunal popular es lego en probabilidades (y también la casi totalidad de jueces), con lo cual el argumento dado al principio: inocente solo con probabilidad una entre mil, puede resultar convincente, y no tiene por qué darse cuenta nadie (y muchas veces, por desgracia, ocurre) que ha de considerarse también el tamaño de la población.

Trabajemos ahora en otro ejemplo muy similar:

Ej.

Supongamos que en cierta ciudad se ha cometido un crimen. Hay 10.000 hombres en esa ciudad que podrían haberlo cometido, de los que 200 trabajan en un pozo minero. En la escena del delito se ha encontrado cierta evidencia que determina que el criminal ha de ser uno de los 200 mineros, se trata de restos de mineral que sólo pueden provenir del pozo minero. Se ha identificado a un sospechoso y en sus ropas se han encontrado restos de mineral similares a los encontrados en la escena del delito. ¿Cómo podría evaluarse esta evidencia?

Representemos la evidencia por  $E$ , el suceso *se han encontrado restos de mineral en la ropa del sospechoso* que son similares a los restos de mineral

encontrados en la escena del delito. Denotemos la hipótesis de que *el sospechoso es culpable* mediante  $C$ , y la de que es inocente mediante  $\bar{C}$ .

Parece razonable suponer que todos los trabajadores del pozo minero tienen en alguna parte de sus ropas restos de mineral similares a los encontrados en la escena del delito. En cualquier caso, la probabilidad de encontrar la evidencia en una persona inocente puede calcularse de la forma siguiente: hay 9999 hombres inocentes en la ciudad, de los que 199 trabajan en la mina. Esos 199 hombres, por la suposición inicial, tendrán la evidencia en sus ropas debido a su trabajo. Así pues

$$P(E | \bar{C}) = \frac{199}{9999} = 0.019$$

Una confusión en la interpretación de esta probabilidad puede tener graves consecuencias para el presunto culpable. En efecto, si a la hora de evaluar la evidencia permutamos las posiciones de  $E$  y  $\bar{C}$  en la anterior expresión, estaremos diciendo que una persona a la que se encuentra la evidencia es inocente con una probabilidad de aproximadamente 0.02 (**por lo tanto culpable con probabilidad 0.98**). El paso siguiente por parte del fiscal será reclamar la culpabilidad del acusado.

Pero, en realidad, igual que en el ejemplo anterior, lo que debemos calcular es la probabilidad de ser culpable (o inocente) en función (condicionado a) los datos existentes. En la ciudad hay 200 hombres con la evidencia  $E$ , de los que 199 son inocentes. Por lo tanto,

$$P(\bar{C} | E) = \frac{199}{200} = 0.995$$

y, entonces,

$$P(C | E) = 1 - 0.995 = 0.005$$

es decir, la probabilidad de ser culpable tan sólo 0.005.

La utilización de  $P(E | \bar{C})$  en lugar de  $P(\bar{C} | E)$ , se conoce como la *falacia de la condicional transpuesta* o **falacia del fiscal** y, desgraciadamente, puede ocurrir y ocurre con más frecuencia de la deseable.

En resumen, si llamamos  $H$  a la hipótesis de ser culpable -o inocente, según interese, porque son complementarias-,  $E$  a las evidencias o pruebas, podemos tener una probabilidad a priori  $P(H)$ , y la fórmula de Bayes nos permite

calcular la probabilidad a posteriori dado que se ha presentado una evidencia,  $P(H | E)$

a partir de la probabilidad a priori y de una probabilidad que, normalmente, es más fácil conocer, que es  $P(E | H)$ .

$$P(H | E) = \frac{P(E | H) \cdot P(H)}{P(E)}$$

$P(E | H)$  es la probabilidad de la evidencia  $E$  si la hipótesis  $H$  de la inocencia del acusado es cierta. Esta probabilidad se conoce como verosimilitud, ya que representa lo verosímil o creíble que sería la evidencia  $E$  que hemos observado, si la hipótesis  $H$  fuese cierta.

Dicho de otro modo, si alguien es culpable, es lógico que todas las pruebas apunten contra él ( $P(E | C)$  es alta). Ahora bien, si todas las pruebas apuntan contra él, no quiere decir que sea culpable ( $P(C | E)$  alta?). Por la regla de Bayes, como hemos visto

$$P(C | E) = \frac{P(E | C) \cdot P(C)}{P(E)}$$

Un ejemplo que puede entenderse fácilmente es el de que alguien vaya a cobrar un décimo de la lotería premiado. La probabilidad de que, comprando un décimo, te toque un premio muy alto (el gordo, por ejemplo) es muy pequeña. Si alguien te acusa de haber robado el décimo, automáticamente serías culpable precisamente por ese motivo.

Veamos ahora algunos casos reales muy conocidos de la falacia del fiscal, ampliamente documentados y de los que se puede encontrar mucha información en internet.

#### 4.18.0.1 El caso O.J. Simpson o Estadisticidio

Un caso muy conocido de la falacia del fiscal es el de O.J. Simpson (famoso jugador de rugby y posteriormente actor secundario en muchas películas taquilleras de Hollywood, como “El coloso en llamas”, 1974), acusado del asesinato de su ex mujer, Nicole Brown y su amante. Fue uno de los mayores acontecimientos periodísticos en 1995 en Estados Unidos, puesto que la



Figura 4.37: Juicio Real y Serie de TV.

persecución policial de Simpson -que huyó en su coche del lugar del crimen- fue retransmitida en directo por la televisión.

La policía tenía multitud de pruebas contra Simpson (antecedentes de violencia de Simpson contra su mujer, sangre de Simpson en el lugar del crimen, sangre de la víctima en el coche de Simpson...). Poco podía hacer la defensa, aparte de criticar al Departamento de Policía de Los Ángeles por racismo, así como por irregularidades en la obtención y autenticidad de las pruebas.

Pero claro, poderoso caballero es don dinero. Simpson se gastó sobre cuatro millones de dólares en un equipo de nueve abogados, entre los que se encontraban algunos tan famosos en EEUU como Johnny Cochran (acostumbrado a ganar juicios contra abusos de derechos civiles contra ciudadanos negros) o Robert Shapiro (el número uno de los abogados criminalistas de los Ángeles). Casualmente, hasta Robert Kardashian (padre de las posteriormente televisivas hermanas) también participó en la defensa de Simpson, puesto que sus familias eran amigas.

La fiscalía se encargó de dar cuenta de los reiterados abusos físicos de Simpson contra su mujer Nicole. Sin embargo, la defensa adujo que las pruebas previas no significaban nada. Según las estadísticas del año 1993, si bien 4 millones de mujeres eran maltratadas anualmente por maridos y novios en EEUU, solo 1432 mujeres (o sea, solo una entre 2500, aproximadamente), fueron asesinadas por estos. Por lo tanto, "pocos hombres que abofetean o golpean a sus compañeras domésticas continúan hasta matarlas".

Sin embargo, este es un nuevo ejemplo de la falacia del fiscal. Este argumento

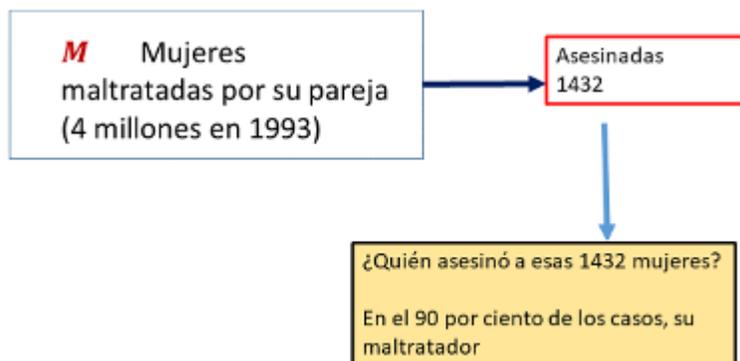


Figura 4.38: Mujeres asesinadas dentro del colectivo de mujeres maltratadas: únicamente 1 de cada 2500.

puede convencer, pero no es el realmente importante. El número relevante no es la probabilidad de que un hombre que maltrate a una mujer acabe matándola (1 entre 2.500), sino la probabilidad de que una mujer maltratada sea asesinada por su maltratador. Según el *Uniform Crime Reports for the United States and its Possessions* de 1993, se obtuvo que: de todas las mujeres maltratadas asesinadas en 1993, el 90% fueron asesinadas por su maltratador. Y esta estadística no fue citada en el juicio. Obviamente, el equipo defensor no iba a hacerlo aunque lo supiera, pero el equipo acusador seguramente desconocía todo lo referente a probabilidades condicionadas.

Para que podamos comprender mejor lo sucedido vamos a realizar un esquema. Llamemos

$M = \text{mujeres maltratadas por su marido o pareja}$

Ahora llamemos  $A = \text{mujeres asesinadas por su pareja}$ . El dato que dio la defensa de Simpson fue  $P(A | M) = \frac{1432}{4 \cdot 10^6} \approx \frac{1}{2500}$  (la probabilidad de que una mujer maltratada también sea asesinada es muy pequeña).

Descartando cualquier otro tipo de consideraciones, Simpson solo tendría un 10 por ciento de probabilidades de ser inocente, no 1 entre 2.500.

Consideraciones y argumentos de los abogados hubo muchos más, por supuesto, que para eso eran tan buenos y caros. Como puede leerse en [El PAIS](#):

En el lugar del crimen se encontró un guante, usado por el asesino



Figura 4.39: Imagen de la serie de Tv "El caso O.J. Simpson" (2016) en la famosa escena del guante.

de Nicole para no dejar huellas; la policía encontró el otro en la casa del deportista. Johnny Cochran defendió la tesis de que el segundo guante había sido colocado "por un policía racista" para incriminar a su defendido, e hizo que O. J. Simpson se los probara frente al jurado, de manera muy aparatosa, para demostrar que no eran de su talla. En su intervención final, el abogado acuñó, en rima, la expresión que le hizo famoso: "If it doesn't fit, you must acquit" (Si no se los puede poner, tienen que absolver).

"El desenlace del caso O. J. Simpson, uno de los juicios del siglo que hay cada dos o tres años en EE UU, fue polémico; la fiscalía aseguró tener una montaña de pruebas en contra del acusado, entre ellas, su rastro de sangre en el lugar del crimen y la sangre de la víctima en el famoso segundo guante. Pero Cochran jugó la carta del complót racista y pidió a un jurado en el que los negros eran mayoría que recordara los graves incidentes de 1992 -los desórdenes en Los Ángeles después de la absolución de los policías blancos que dieron una tremenda paliza a Rodney King- y que asestara con su veredicto un golpe a la corrupción policial."

#### 4.18.0.2 El pueblo contra Collins (1968)

Una mujer mayor, mientras caminaba en el área de San Pedro en los Ángeles, fue asaltada por detrás para robarle el bolso. La víctima dijo que le pareció

reconocer a una mujer joven y rubia, que salió corriendo. Otro testigo dijo ver corriendo a una mujer con el pelo rubio y coleta, que se introdujo en un automóvil amarillo conducido por un hombre de raza negra que tenía barba y bigote.

Unos pocos días más tarde, la policía arrestó a una pareja que cumplía las descripciones: Malcolm y Paula Collins. Cuando la policía llegó a casa de los Collins para arrestarlos, Malcolm salió huyendo por la parte trasera de la casa. La policía lo detuvo y encontró en el bolsillo de Malcolm dos recibos de sendos pagos por importe total igual al dinero robado. Preguntados Malcolm y Jane por el dinero robado, ambos ofrecen versiones contradictorias.

El fiscal no tenía evidencias tangibles ni testigos fiables contra los sospechosos y construyó su caso sobre lo improbable que resultaba que la Sta. Collins y su amigo tuvieran todas estas características y no fueran culpables. Para ello asignó probabilidades a las citadas características, basadas en la incidencia de las mismas en la población de Los Angeles:

Características	Probabilidad
Automóvil amarillo	1/10
Varón con bigote	1/4
Mujer con coleta	1/10
Mujer rubia	1/3
Varón negro con barba	1/10
Pareja interracial en coche	1/1000

El fiscal argumentó que la probabilidad de que todas estas características se dieran conjuntamente, admitiendo la hipótesis de independencia entre ellas, venía dada por el producto de sus respectivas probabilidades (probabilidad de la intersección) y que dicho producto, como fácilmente puede comprobarse, era una entre doce millones.

$$P(A_1 \cap A_2 \cap \dots \cap A_6) = \frac{1}{10} \cdot \frac{1}{4} \cdot \frac{1}{10} \cdot \frac{1}{3} \cdot \frac{1}{10} \cdot \frac{1}{1000} = \frac{1}{12 \cdot 10^6}.$$

Lo que significaba que era tan improbable encontrar una pareja que se ajustara a todas las características que, verificándolas Janet Collins y su compañero, la única decisión razonable, según el fiscal, era proclamarlos culpables, como efectivamente ocurrió.

El abogado de la Sta. Collins apeló a la Corte Suprema de California argumentando que el razonamiento probabilístico era incorrecto y engañoso. El defensor sostuvo que era posible aproximarse a los datos desde una perspectiva diferente, que mantenía la duda razonable sobre la culpabilidad de sus clientes.

En efecto, el razonamiento alternativo comenzaba suponiendo que había  $n$  parejas en el área geográfica donde ocurrieron los hechos y que existía una probabilidad  $p$  de que cualquiera de estas parejas compartiera las seis características introducidas por el fiscal como evidencias. De acuerdo con lo anterior  $p = 1/12.000.000$ . El defensor centró su atención en los sucesos  $A$ ="entre las  $n$  parejas existen al menos 2 con iguales características" y  $B$ ="entre las  $n$  parejas existe al menos 1 con iguales características", y más concretamente en el cociente de sus probabilidades.

Entonces

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}$$

ya que  $A \subset B$ .

Dicho en otros términos de mayor interés para la defensa, se trata de la probabilidad de que al menos otra pareja hubiera podido cometer la acción criminal. Si este cociente no fuera muy pequeño, habría que admitir la posibilidad de que la Sta. Collins y su amigo tenían competidores que podrían ser los culpables.

Tenemos que  $\bar{B}$  es el suceso de que ninguna pareja de las  $n$  posee las seis características mencionadas. Para una sola de estas parejas, la probabilidad de no poseerlas es  $(1 - p)$ , y como las  $n$  parejas podemos suponerlas independientes,

$$P(B) = 1 - P(\bar{B}) = 1 - (1 - p)^n$$

Para calcular la probabilidad de  $A$  también consideramos

$$P(A) = 1 - P(\bar{A}) = 1 - P(\text{ninguna pareja as o una pareja as})$$

Puede probarse que esto es igual a

$$= 1 - (1 - p)^n - n \cdot p \cdot (1 - p)^{n-1}$$

De manera que

$$P(A | B) = \frac{P(A)}{P(B)} = \frac{1 - (1 - p)^n - n \cdot p \cdot (1 - p)^{n-1}}{1 - (1 - p)^n}$$

En la siguiente tabla vemos como varía la probabilidad en función del número  $n$  de parejas.

<b>n</b>	<b>P(A B)</b>
<b>1.000.000</b>	0.042
<b>2.000.000</b>	0.0786
<b>5.000.000</b>	0.1875
<b>10.000.000</b>	0.3479

En una ciudad de las dimensiones de Los Angeles, con unos 2.000.000 de parejas, no era tan improbable, Como vemos en la tabla, el valor sale 0.0786, que viene a ser una entre doce o trece, bastante lejos de una entre doce millones.

#### 4.18.0.3 El caso de Sally Clark

El primer hijo de una mujer llamada Sally Clark murió a los 11 meses de vida (en 1996). Se informó de que su muerte se debía al SMLS (siglas en inglés de Síndrome de Muerte Súbita del Lactante). O sea, la autopsia no reveló la causa. La senora Clark quedó embarazada por segunda vez (1997), y tuvo su segundo hijo, que murió a las ocho semanas, otra vez por SMLS. En ambos casos, la senora Clark estaba sola en casa con sus bebés. Fue detenida y acusada de asfixiar a sus dos hijos.

En el juicio, la acusación llamo a un experto pediatra, sir Roy Meadow, quien declaró basándose en la rareza del SMLS, que las probabilidades de que un niño muriera de SMLS era de una entre 8.543. Como ambos hijos murieron de esa forma, y las muertes eran independientes, la probabilidad de que dos hubieran muerto de esta manera es

Siendo  $A_i$ ="el niño  $i$  muere"

$$P(A_1 \cap A_2) = \frac{1}{8543} \cdot \frac{1}{8543} = \frac{1}{73 \cdot 10^6}$$

Otro experto, el profesor Berry, sugirió que habría que tener en cuenta posibles antecedentes familiares, y si hay una muerte súbita en una familia, no



Figura 4.40: Recorte de periódico inglés.

quiere decir que no pueda ocurrir otra.

Aparte de la prueba estadística, no había ninguna prueba material; ninguna prueba física. A pesar de ello, el jurado condeno a la acusada por 10 votos contra 2, a cadena perpetua (1999).

Semanas después, la revista *British Medical Journal* publicó que la probabilidad de que ambos hermanos hubieran muerto de SMLS debía estimarse en una entre 2,75 millones. Aun así, la probabilidad era muy baja.

Otra vez tenemos la falacia del fiscal. No había que considerar la probabilidad de que dos niños murieran por SMLS, sino que, dadas las muertes de dos niños, ¿qué es más probable? ¿Qué hayan muerto por SMLS o que hayan sido asesinados por su madre? Dos años después de que la señora Clark fuera encarcelada, la mismísima *Royal Statistical Society* intervino en un comunicado de prensa:

La decisión del jurado está basada en un serio error de conocimiento lógico conocido como la falacia del fiscal. El jurado necesita sopesar dos explicaciones contradictorias sobre las muertes de los bebés: SMLS o asesinato. Dos muertes a causa del SMLS o dos asesinatos son bastante improbables, pero aparentemente uno de ellos ha sucedido en este caso. Lo que importa es la probabilidad relativa de las muertes... no solamente lo improbable que es. Lo que se debió buscar no era la probabilidad de que hubiera dos muertes en la misma familia, sino la probabilidad de que una madre cometa un doble asesinato.

Después del juicio, Ray Hill, de la universidad de Salford, analizó los datos y

estimó que la probabilidad de una segunda muerte súbita estaba entre  $1/60$  y  $1/130$ .

Si elegimos, por ejemplo,  $1/100$ , tendríamos  $P(A_1 \cap A_2) = \frac{1}{8543} \cdot \frac{1}{100} = \frac{1}{854.300}$  que, obviamente, es un número pequeño, pero no tanto como la primera probabilidad sugerida.

En una revisión del caso, el veredicto fue puesto en duda ya que se descubrió que el patólogo que había trabajado para la acusación había ocultado que el segundo niño había sufrido una infección bacteriana (*Staphylococcus aureus*), la cual podía haber sido la causa de la muerte. Además, se encontró que, de 325 familias donde se había producido una muerte súbita, 5 habían tenido un fallecimiento anterior.

Analizando pormenorizadamente las estadísticas, se obtuvo que, en Inglaterra y Gales, se dan 30 casos de homicidio de un hijo por cada 650.000 nacimientos, y la probabilidad de que una madre cometa un doble homicidio sobre sus hijos es, aproximadamente, de una entre mil:

$$P(A_2 | A_1) = \frac{1}{1000}$$

Con lo que

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1) = \frac{30}{650000} \cdot \frac{1}{1000}$$

La conclusión es que era 9 veces más probable que los niños hubieran muerto de SMLS que de asesinato.

Basándose en estas y otras pruebas, y después de tres años y medio (en 2003), Sally Clark fue liberada. Murió en 2007.

#### 4.18.0.4 Ejercicio: padre a cara o cruz

(extraído de (Montes, 2003)): Un hombre es acusado en un caso de paternidad sobre la base de un marcador genético cuya frecuencia en la población adulta es del 1%, y que se transmite con probabilidad 1 de padres a hijos. Tanto el presunto padre como el niño causante del litigio poseían el citado marcador, por lo que el fiscal del caso planteo la conveniencia de obtener la probabilidad de que el acusado fuera el padre dado que tenía el marcador.

Si  $A$ =el acusado es el padre y  $B$ =el niño tiene el marcador, la probabilidad se obtiene aplicando Bayes:

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B | A) \cdot P(A) + P(B | \bar{A}) \cdot P(\bar{A})}$$

De acuerdo con los datos,  $P(B | A) = 1$  y  $P(B | \bar{A}) = 0.01$ . En cuanto a  $P(A)$  y  $P(\bar{A})$  se estimó conveniente que ambas eran iguales a 0.5, valor que trataba de reflejar el desconocimiento acerca de la paternidad, asignando igual probabilidad a ambos supuestos.

El resultado: el hombre es el padre, puesto que  $P(A | B)$  es aproximadamente 0.99.

El defensor recurrió, basándose en la asignación de probabilidades a  $A$  y a su complementario. Llevada a sus últimas consecuencias, dijo el abogado, semejante asignación de probabilidades equivalía a declarar padre a cualquier adulto por el procedimiento a cara o cruz. Se estaba confundiendo ignorancia con equiprobabilidad.

En la tabla siguiente se muestran valores de  $P(A | B)$  en función de  $P(A)$ , evidenciándose la importancia de la elección de esta última probabilidad, puesto que valores bajos de  $P(A)$  (entre 0 y 0.1) dan lugar a valores bajos de  $P(A | B)$ , que difícilmente conseguirían una condena.

<b>P(A)</b>	<b>P(A B)</b>
<b>0.01</b>	0.5025
<b>0.03</b>	0.7557
<b>0.05</b>	0.8403
<b>0.07</b>	0.8827
<b>0.09</b>	0.9082
<b>0.1</b>	0.9174
<b>0.3</b>	0.9772
<b>0.5</b>	0.9901
<b>0.7</b>	0.9957
<b>0.9</b>	0.9989



Figura 4.41: Ibáñez nunca se equivocó.

## 4.19 Sobre la formación correcta de jurados.

Con la revolución francesa, la figura del jurado popular se convirtió en una conquista social: el pueblo llano pasaría, a partir de ese momento, a participar en la impartición de la justicia. El primer matemático que adoptó una postura, basada en sus conocimientos acerca de esta institución, fue Marie-Jean-Antoine Nicolas de Caritat (1743-1794), más conocido por su marquesado, el de Condorcet. Fue un destacado matemático y filósofo, alumno de d'Alembert, participó en la redacción de la Enciclopedia y pidió el voto para la mujer, en idénticas condiciones a las del hombre, en un artículo del Journal de la Société de julio de 1789.

Condorcet establece, como garantía de éxito en la decisión de un jurado, una probabilidad de acierto de  $144.767/144.768$  para que el error de la decisión sea despreciable, y añade que el objeto de tal exigencia “no es sólo evitar que el inocente sea condenado, (...) si no evitar al mismo tiempo el riesgo de absolver a un culpable cuando el crimen esté realmente probado, es decir que este riesgo debe ser lo suficientemente pequeño para poder ser despreciado”.

Condorcet obtuvo esta cifra a partir de los siguientes cálculos: en una asamblea de 61 electores, en la que se exija una mayoría de 9 votos, se daría este caso, supuesto que la probabilidad de acierto de cada uno sea de  $4/5$ , es decir, que cada uno no se equivoque más que una de cada cinco veces. Si, además, se exige que el error no se dé más que una de cada diez veces, entonces bastaría con exigir una mayoría de seis votos en una asamblea de 44 Electores.

Laplace se ocupó también del cálculo del número de jueces y del número de votos que deben tomarse para garantizar la presunción de inocencia del reo, así como la defensa de los ciudadanos ante la posible absolución de un culpable. Según Laplace, la probabilidad de que la decisión de cada juez sea justa es la pieza clave de todo el proceso, afirmando que la probabilidad de que un juez emita una decisión acertada debería variar desde un mínimo de  $1/2$  a la casi certeza (menor probabilidad daría que fuese más fiable lanzar una moneda al aire que fiarse del juez).

En cuanto al número de jueces que deben componer un tribunal, planteó que el objetivo debía ser minimizar la probabilidad de error en la decisión. Y comparó la situación de una mayoría exigida de dos votos en los casos de composiciones de 8 o de 6 miembros. La probabilidad de error en la condena superaría el 25% en el primer caso, mientras que en el segundo quedaría por debajo, lo que supondría, con la aplicación de la misma mayoría simple, una ventaja para el acusado.

Tras analizar distintas opciones, concluye que, *para garantizar convenientemente la inocencia del acusado, debería exigirse al menos una mayoría de nueve miembros en un tribunal formado por doce*. Planteó que una posible solución de compromiso podría ser la de un jurado de nueve miembros<sup>1</sup>, exigiendo la unanimidad.

En 1837, Siméon-Denis Poisson publica su *Investigación sobre la probabilidad de los juicios*, una obra de más de 400 páginas. Poisson utilizaría su *ley de los grandes números* (capítulo 6) para su análisis y posteriores recomendaciones. Así, exige tener presentes dos cantidades previas deducidas de la observación:

- las relativas a que “un jurado, escogido al azar, no equivoque su voto”, establecida en algo más de  $2/3$  para los crímenes contras las personas, y en alrededor de  $13/17$  en el caso de los juicios por crímenes contra las propiedades, y en  $3/4$  si no se distingue entre unos juicios y otros;
- la probabilidad a priori (*avant l'ouverture des débats*) de que el acusado sea culpable, fijada entre 0.53 y 0.54 en el primer caso, y alrededor de  $2/3$  en el segundo, siendo de 0.64 si no se efectúa la distinción y se evalúa de forma global.

Para el cálculo de estos números, Poisson empleó las cifras existentes de con-

---

<sup>1</sup>este es el número de miembros de un jurado en España

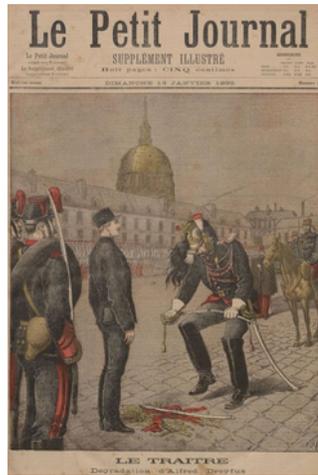


Figura 4.42: Portada de un periódico frances original.

denados por, al menos, siete votos contra cinco, y entre ellos los condenados exactamente por tal mayoría, divididos ambos por el número total de acusados; analizando por separado los juicios por crímenes contra las personas y por atentados contra las propiedades.

Poisson señaló también la necesidad de que este estudio fuese aplicado a todo juicio en el que existiese la posibilidad de estudiar un número suficientemente grande de casos previos para obtener las cifras necesarias. Y hace referencia a los tribunales de excepción de la Revolución (35 años antes de la publicación de su obra), justificando el alto número de condenas en la parte pasional de la acusación y del jurado, “muy alejada de la necesaria calma de espíritu que debe guiar la aplicación de la justicia” (Poisson, 1837) (Zafra and de Paz Cobo, 2012), (Hacking and Bixio, 1995).

## 4.20 El caso Dreyfuss

Para terminar este apartado de derecho y probabilidades, no podemos dejar de citar este caso, por cuanto en él se vio involucrado Jules Henri Poincaré (1854-1912); físico, científico teórico y filósofo de la ciencia, primo del presidente de Francia Raymond Poincaré.

No sería el único personaje célebre que se involucró en la defensa de una



Figura 4.43: Henri Poincaré, uno de los matemáticos más prestigiosos de la historia.



Figura 4.44: Izquierda: Émile Zola. Derecha: Paul Muni, en la película "La vida de Emile Zola" (1937), que le valió al actor una nominación al Oscar.

acusación injusta. El caso Dreyfus es un caso histórico por sus connotaciones racistas y políticas, que causaría un vuelco en la historia de Francia. Émile Zola (1840-1902), el famoso escritor, se implicaría tanto en este caso que debería exiliarse en Londres para no ir a la cárcel.

Este caso apareció reflejado en la película "La vida de Émile Zola" (1937), y sería también inspiración de diferentes películas y documentales. Aunque puede consultarse con detalle en la wikipedia ([https://es.wikipedia.org/wiki/Caso\\_Dreyfus](https://es.wikipedia.org/wiki/Caso_Dreyfus)), expondremos un breve resumen y la importancia de las pruebas periciales y las probabilidades en este asunto.

Alfred Dreyfus (1859-1935) era un capitán del ejército de origen judío que fue falsamente condenado por pasar secretos militares a los alemanes. En 1894, después de que un espía francés de la embajada alemana en París descubriera una carta, arrojada en una papelería, con una escritura que se parecía a la de



Figura 4.45: Portada del diario original con la famosa carta de Zola.

Dreyfus, fue juzgado en consejo de guerra, declarado culpable de traición y condenado a cadena perpetua en la Isla del Diablo, frente a las costas de la Guayana Francesa.

En 1896, el nuevo jefe de la unidad de inteligencia del ejército, Georges Picquart, descubrió pruebas que señalaban a otro oficial militar francés, el mayor Ferdinand Esterhazy, como el verdadero traidor. Sin embargo, cuando Picquart les dijo a sus jefes lo que había descubierto, fue transferido al norte de África y luego encarcelado. Sin embargo, empezaron a circular noticias sobre la posible culpabilidad de Esterhazy. En 1898 fue juzgado en consejo de guerra pero, rápidamente, fue declarado inocente; más tarde, huyó del país.

Tras la absolución de Esterhazy, un periódico francés publicó una carta abierta en primera página titulada “J'Accuse...” de Emile Zola, en la que defendía a Dreyfus y acusaba a los militares de un encubrimiento importante en el caso. Ya antes de la intervención de Zola, el caso Dreyfus había generado una división en la opinión pública francesa entre partidarios y defensores de este. Después se irían sucediendo revisiones de la condena, hasta conseguirse una amnistía del condenado en 1899, y la absolución final en 1906.

Como resultado de su famosa carta “Yo acuso”, Zola fue condenado por difamación, aunque escapó a Inglaterra para no ir a la cárcel, volviendo más tarde a Francia.

La acusación contra Dreyfus, en el juicio inicial que lo condenó a prisión,

estuvo marcada por las pruebas periciales realizadas por Alphonse Bertillon (policía francés, 1853-1914), hijo de Louis-Adolphe Bertillon (médico, antropólogo y estadístico al igual que el hermano de Alphonse, Jacques Bertillon, que también fue médico y estadístico). Se le considera el creador de la **antropometría**, en su caso la recopilación de una serie de medidas físicas para poder reconocer a los delincuentes. Esta técnica, inicialmente conocida como *Bertillonage*, tenía el problema de que las medidas recogidas de las personas “fichadas” iban cambiando con la edad, algo que no sucede con las huellas dactilares. Fue precisamente Francis Galton quien le daría el impulso definitivo a esta técnica de identificación de criminales mediante el uso de estas últimas, calculando además la probabilidad de que dos personas en el mundo puedan tener las mismas huellas, que es prácticamente nula.

Bertillon fue perito de la acusación contra Dreyfus. Argumentó que la letra encontrada en la carta que demostraba espionaje alemán contra Francia era de Dreyfus, y que la había alterado para esconderla. Calculó la probabilidad de que la letra hubiera sido falsificada para acusar a Dreyfus en un número muy pequeño.

Pueden leerse los detalles con cierta profusión en el libro, ya comentado, de Garzón (Ruiz-Garzón, 2015), donde se advierte que el error cometido por Bertillon es un ejemplo más de la *falacia del fiscal*. En 1899, Paul Painlevé, matemático que llegó a primer ministro de Francia, recibió una carta de Poincaré, donde este criticaba abiertamente los cálculos realizados por Bertillon. Se abriría posteriormente una revisión del caso, donde el tribunal encargaría un informe pericial a Gaston Darboux (secretario perpetuo de la Academia de Ciencias), Paul Apell (decano de la facultad de ciencias de París) y al mismo Poincaré, sobre el tema de la falsificación. Este informe supondría un espaldarazo a la inocencia de Dreyfus, a la par de una desacreditación para Bertillon.

### 4.20.1 Curiosidad histórica

Poincaré es conocido como el último *universalista* (especialista en más de una rama científica), y fue autor de importantes logros reconocidos en muchas de sus áreas de conocimiento. Escribió grandes avances sobre la teoría de la luz y desarrolló avances en la teoría de la relatividad. Como matemático, dejó escrito una de los famosos [siete problemas del milenio](#): la llamada *Conjetura de Poincaré*, un problema topológico que fue resuelto por el matemático ruso

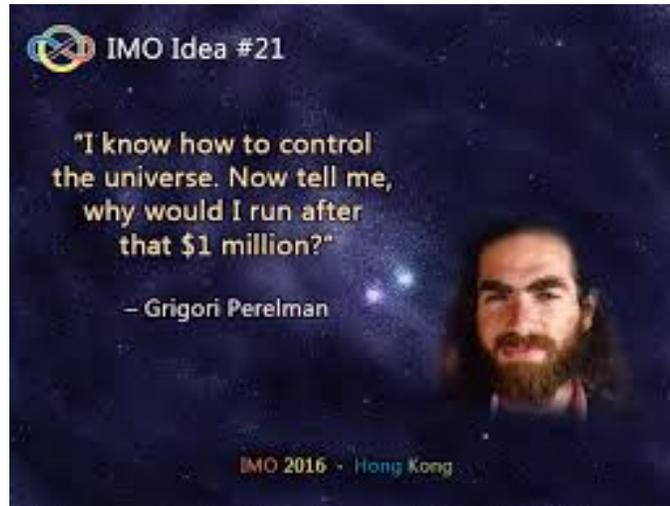


Figura 4.46: Gregory Perelman.

Gregory Perelman en 1906. A este último le sirvió para obtener la medalla Fields (equivalente al premio Nobel en matemáticas) y el premio del milenio otorgado por el Clay Mathematical Institute, dotado con un millón de dolares. Sin embargo, Perelman no fue a recoger el premio porque, según parece ser que dijo que *estaba ocupado cuidando a su madre, a quien no podía dejar sola en su pequeño apartamento de Moscú*. Según aparece en la Wikipedia, renunció al premio diciendo:

“No quiero estar expuesto como un animal en el zoológico. No soy un héroe de las matemáticas. Ni siquiera soy tan exitoso. Por eso no quiero que todo el mundo me esté mirando.”

Buscando en internet, me he encontrado, según la fuente que se mire, que renunció a un premio diciendo una cosa, o al otro diciendo la contraria. Yo mismo recuerdo escuchar en el telediario, en el año 2006, que Perelman no había ido a recoger la medalla Fields, y según comentaban por la razón anteriormente expuesta. Lo que sí parece ser verídico es que debe ser tan buen matemático como persona honrada.



Figura 4.47: El presentador de televisión Monty Hall y su amiga.

## 4.21 Problemas de paradojas

### 4.21.1 Problema de Monty Hall

La dinámica de un concurso de televisión es la siguiente: hay tres puertas y detrás de una de ellas está el premio; el concursante elige una de las puertas pero, antes de abrirla, el presentador, que sabe donde está el premio, abre una de las otras dos puertas, muestra que en ella no se encuentra el premio y, acto seguido, le ofrece al concursante una última oportunidad de cambiar la puerta elegida. ¿Qué debe hacer el concursante?

El Problema de Monty Hall es un problema de probabilidad que está inspirado por el concurso televisivo estadounidense *Let's Make a Deal* (Hagamos un trato), famoso entre 1963 y 1986. Su nombre proviene del presentador, Monty Hall.

En este concurso, el concursante escoge una puerta entre tres, y su premio consiste en lo que se encuentra detrás. Una de ellas oculta un coche, y tras las otras dos hay una cabra. Sin embargo, antes de abrirla, el presentador, que sabe dónde está el premio, abre una de las otras dos puertas y muestra que detrás de ella hay una cabra. Ahora tiene el concursante una última oportunidad de cambiar la puerta escogida. ¿Debe el concursante mantener su elección original o escoger la otra puerta? ¿Hay alguna diferencia?

A primera vista, parece que da igual. La intuición nos dice que ahora, quitando una puerta sin premio, la puerta que nosotros escogimos tiene un 50%



Figura 4.48: Casos posibles del problema.

de tener una cabra y por tanto da igual cambiar que no hacerlo. Desarrollamos todas las posibilidades:

Si miramos las posibilidades de éxito de cambiar o no cambiar, vemos que si no cambiamos tenemos  $1/3$  y si cambiamos tenemos  $2/3$ .

Veamos cómo se hace con probabilidades condicionadas. Definimos los siguientes sucesos:  $A$ ="El jugador selecciona la puerta que contiene el coche en su selección inicial".  $B$ ="El jugador selecciona una puerta que contiene una cabra en su selección inicial".  $G$ ="El jugador gana el coche".

Para calcular  $P(G)$ , basta con notar que  $G = (G \cap A) \cup (G \cap B)$ , ya que  $A \cap B = \emptyset$  y  $A \cup B = \Omega$  (esto es equivalente a decir que  $A, B$  es una partición de  $\Omega$ )

$$\begin{aligned} P(G) &= P((G \cap A) \cup (G \cap B)) = \\ &= P(G \cap A) + P(G \cap B) = P(G | A) \cdot P(A) + P(G | B) \cdot P(B) \end{aligned}$$

Tenemos que  $P(A) = 1/3$  y  $P(B) = 2/3$  pues hay un coche y dos cabras. Ahora calculamos la probabilidad de ganar el coche,  $P(G)$ , en función de que el jugador acepte o no la oportunidad de cambiar de puerta elegida.

#### Jugador que no se cambia.

En este caso  $P(G | A) = 1$  y  $P(G | B) = 0$  pues el jugador se queda con su selección inicial, es decir  $P(G) = 1/3$ .

$$P(\text{ganar}) = P(\text{coche} \mid (\text{eligiendo bien al principio})) \cdot P(\text{elegir bien al principio}) + \\ + P(\text{coche} \mid (\text{elegir mal al principio})) \cdot P(\text{elegir mal al principio}) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{3}$$

**Jugador que cambia.**

En este caso  $P(G \mid A) = 0$  y  $P(G \mid B) = 1$  pues el jugador se cambia a la única puerta cerrada que queda (y sabemos que como el presentador sabe dónde está el coche, siempre mostrará una cabra).

$$P(\text{ganar}) = P(\text{coche} \mid (\text{eligiendo bien al principio})) \cdot P(\text{elegir bien al principio}) + \\ + P(\text{coche} \mid (\text{elegir mal al principio})) \cdot P(\text{elegir mal al principio}) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{3}$$

Claramente la mejor estrategia es cambiar siempre, pues la probabilidad efectiva de ganar es el doble de la correspondiente al jugador que no cambia nunca.

**4.21.2 Paradoja de los hijos**

El matrimonio Smith camina por la calle y se encuentra a don Pepito. Don Pepito reconoce a su amigo del colegio José Smith, así que se detienen a saludarse: “Hola, don Pepito”, “Hola, don José”. “¿Recuerda usted a mi esposa?” -dice don José, señalando a su pareja. Don Pepito asiente-. ¿Y a mi hijo Joseíto? -dice, señalando a un chico que va a su lado.

Sabiendo que el matrimonio Smith tiene dos hijos, ¿cuál es la probabilidad de que el otro hijo sea también varón? (Suponemos que la probabilidad de nacer hombre o mujer es 0.5).

A primera vista, da la impresión de que la probabilidad es 0.5. Lo que ocurre es que, aunque no lo parezca, tenemos una información que modifica esa probabilidad. Si un matrimonio tiene dos hijos, el espacio muestral de los posibles resultados (en cuanto al sexo) de los hijos es

$$\Omega = \{H_1H_2, H_1M_2, M_1H_2, M_1M_2\},$$

donde estamos utilizando la notación  $H_1H_2$  indicando: hijo menor hombre, hijo mayor hombre... Pero, ya que sabemos que uno de los hijos es hombre (el que acompaña a sus padres), tenemos que el suceso  $M_1M_2$  no puede darse, luego el espacio muestral se reduce a

$$\Omega = \{H_1H_2, H_1M_2, M_1H_2\}$$

con lo que vemos que la probabilidad de que el otro hijo sea hombre es la probabilidad de que los dos hijos sean hombres ( $H_1H_2$ ) que es  $1/3$ .

**Resolver el problema anterior (calcular la probabilidad de que el otro hijo del matrimonio sea varón), sabiendo que el hijo que está paseando con el matrimonio nació un martes (suponemos que todos los días son equiprobables para nacer).**

Ahora sí que parece imposible que pueda servir de algo que nos digan que el hijo (Joseíto) que pasea con el matrimonio nació en martes. Pues en realidad sí que es una información que condiciona la probabilidad que queremos obtener. Llamemos  $H$ ="hombre",  $M$ ="mujer",  $A$ ="un hijo varón nació en martes", y utilizaremos subíndices 1 y 2 para indicar hija(o) menor o hija(o) mayor. La probabilidad que nos piden es

$$P(H_1H_2 | A) = \frac{P(H_1H_2 \cap A)}{P(A)}$$

En el numerador tenemos que calcular la probabilidad de que, teniendo dos hijos, los dos sean varones y uno nazca en martes. Puede ocurrir:

- El primero nace en martes y el segundo no.
- El primero y el segundo nacen en martes.
- El primero no nace en martes y el segundo tampoco.

Los casos posibles son Lunes y Lunes, Lunes y Martes,..., Martes y Lunes,...; hasta Domingo y Domingo. Total 49 casos. Los casos favorables son Martes y Lunes, Martes y Martes,... Martes y Domingo (7) y luego Lunes y Martes, Lunes y Miércoles, ..., Lunes y Domingo (6: Martes y Martes no podemos considerarlo dos veces). Por lo tanto, el numerador es  $13/49$ .

Vamos con el denominador. Los casos posibles son 49, igual que antes. Los casos favorables son los días a considerar según las posibilidades

$H_1H_2, M_1H_2, H_1M_2$  (primer hijo hombre y segundo hombre, primer hijo hombre y segundo mujer.... ).

- En el primer caso, las posibilidades pueden ser 13 (es lo ocurrido con el numerador).
- En el segundo caso serían Martes y Lunes, Martes y Martes, ..., Martes y Domingo. Total 7.
- En el tercer caso serían Lunes y Martes, Martes y Martes, ..., Domingo y Martes. Total 7.

Por consiguiente, el denominador es  $13 + 7 + 7 = 27$  dividido entre 49. Finalmente, la probabilidad es  $(13/49)/(27/49) = 13/27$

### 4.21.3 Problema del prisionero (Hardin, 1968).

En una cárcel, tres prisioneros de historiales similares, solicitan el indulto a un tribunal. Poco después se sabe que el indulto ha sido concedido a dos de los tres presos. Uno de los prisioneros conoce a uno de los miembros del tribunal y sabe que si le pregunta podrá obtener algo de información. Podrá preguntarle por el nombre de uno de los indultados, pero no podrá preguntar si él es uno de ellos.

Reflexionando, concluye que, si no pregunta, la probabilidad de ser uno de los indultados es  $2/3$ ; mientras que, si pregunta, obtendrá una respuesta, y entonces la probabilidad de ser él otro de los indultados es  $1/2$ . Por lo tanto, concluye que será mejor no preguntar, puesto que eso solo le servirá para disminuir su probabilidad de ser uno de los indultados.

**¿Es correcto el razonamiento del preso?**

En este caso, el conocimiento de la ocurrencia de un suceso (nombre de un indultado) no condiciona la probabilidad de que el preso que pregunta sea o no indultado, puesto que eso ya ha sucedido, y no hay ninguna acción (cambio de puerta) que vaya a modificar la probabilidad. La probabilidad es la misma a priori que a posteriori.

- **Consideremos ahora el siguiente enunciado:** Tres prisioneros esperan encarcelados su juicio sabiendo que sólo uno de ellos morirá. El

juez le dice al primer preso que el tercero se salva y le pregunta si quiere intercambiar su suerte con el segundo. ¿Qué debe hacer el primer prisionero?

Como vamos a ver, este problema es equivalente al problema de Monty Hall. Llamamos  $S$ ="salvarse",  $A$ ="tener suerte al principio",  $B$ ="tener mala suerte al principio"

$$P(S) = P(S | A) \cdot P(A) + P(S | B) \cdot P(B)$$

#### Prisionero que no se cambia.

En este caso  $P(S | A) = 1$  y  $P(S | B) = 0$  pues el prisionero no quiere cambiar.

Por lo tanto  $P(S) = 1/3$ . En efecto:

$$\begin{aligned} P(S) &= P(\text{salvarse} | (\text{teniendo suerte al principio})) \cdot P(\text{tener suerte al principio}) + \\ &+ P(\text{salvarse} | (\text{teniendo mala suerte al principio})) \cdot P(\text{tener m.s.a.p.}) = \end{aligned}$$

$$= 1 \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{3}$$

#### Prisionero que cambia

En este caso  $P(S | A) = 0$  y  $P(S | B) = 1$  pues el prisionero se cambia. Por lo tanto  $P(S) = 2/3$ . En efecto:

$$\begin{aligned} P(S) &= P(\text{salvarse} | (\text{teniendo suerte al principio})) \cdot P(\text{tener suerte al principio}) + \\ &+ P(\text{salvarse} | (\text{teniendo mala suerte al principio})) \cdot P(\text{tener m.s.a.p.}) = \end{aligned}$$

$$= 0 \cdot \frac{2}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}$$

#### 4.21.4 Paradoja de la caja de Bertrand, o Gold-Silver box (Bertrand, 1988).

Tenemos tres cajas, y cada caja tiene dos cajones, con una moneda cada uno: Una caja contiene dos monedas de oro, otra caja dos

monedas de plata, y la caja final una de cada tipo. Después de elegir una caja al azar se toma un cajón al azar, y resulta que contiene una moneda de oro. ¿Cuál es la probabilidad de que la otra también sea de oro?

Muchos seguirán el siguiente razonamiento: Después de elegir una caja al azar y retirar una moneda también al azar, si esta resultase ser una moneda de oro, sólo tenemos dos opciones: (a) que hayamos elegido la caja con dos monedas de oro; o (b) que hayamos elegido la caja con una moneda de oro y otra de plata. Por tanto, la probabilidad de que la otra moneda también sea de oro es igual a  $1/2$ . Esta solución es incorrecta. Veamos: Llamemos  $O_i =$  "moneda de oro en la extracción  $i$ ". Nos referimos al suceso obtener una moneda de oro la primera vez, y la segunda vez. Nos piden  $P(O_2 | O_1)$

Por Bayes:

$$P(O_2 | O_1) = \frac{P(O_1 | O_2) \cdot P(O_2)}{P(O_1)}$$

La probabilidad de sacar una moneda de oro en la primera extracción depende de que cajón se haya seleccionado, es decir

$$P(O_1) = P(O_1 | C_1) \cdot P(C_1) + P(O_1 | C_2) \cdot P(C_2) + P(O_1 | C_3) \cdot P(C_3)$$

**Cajón 1:** dos oros. **Cajón 2:** oro y plata. **Cajón 3:** plata y plata. Luego

$$P(O_1) = 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = \frac{1}{2}$$

Ahora tenemos que darnos cuenta de que, cuando escribimos  $P(O_1 | O_2)$  queremos calcular la probabilidad de que el primero fuera oro, sabiendo que el segundo lo fue, es decir

$$P(O_1 | O_2) = P(O_1 | C_1)$$

ya que la única forma de que hayamos obtenido oro en la segunda extracción es que el primer oro fuese extraído en la caja 1, que es la que tiene dos oros.

En definitiva:

$$P(O_2 | O_1) = \frac{P(O_1 | O_2) \cdot P(O_2)}{P(O_1)} = \frac{P(O_1 | C_1) \cdot P(C_1)}{P(O_1)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{3} \cdot \frac{3}{2}} = \frac{2}{3}$$



# Capítulo 5

## VARIABLES ALEATORIAS

En los temas 1 y 2 hemos visto como analizar un conjunto de datos. La metodología general consiste en, a partir de una muestra de una población, resumir el contenido de esa muestra mediante análisis gráficos y numéricos.

La muestra, habitualmente, consiste en una serie de variables (numéricas o no), que llamamos **variables estadísticas**. Esta variable es la medición u observación de una característica concreta (estatura, peso, color... ) en cada uno de los individuos de la muestra.

Centrémonos ahora en variables numéricas. Si en vez de referirnos a los valores de la variable en una muestra concreta (que es una parte de una población) nos referimos a los valores de la variable en **TODA** la población, tendremos una **variable aleatoria**.

Recordemos, antes de nada, el esquema de trabajo de la estadística matemática (Figura 5.1):

El objetivo es conocer lo mejor posible una población, a partir de una muestra de individuos. Para ello se elige una muestra de la población, lo más representativa posible de la misma, y se analiza, extrayendo toda la información que se pueda.

Para poder extrapolar los resultados de la muestra a la población, necesitaremos extender los conceptos vistos de una variable estadística a la población general.

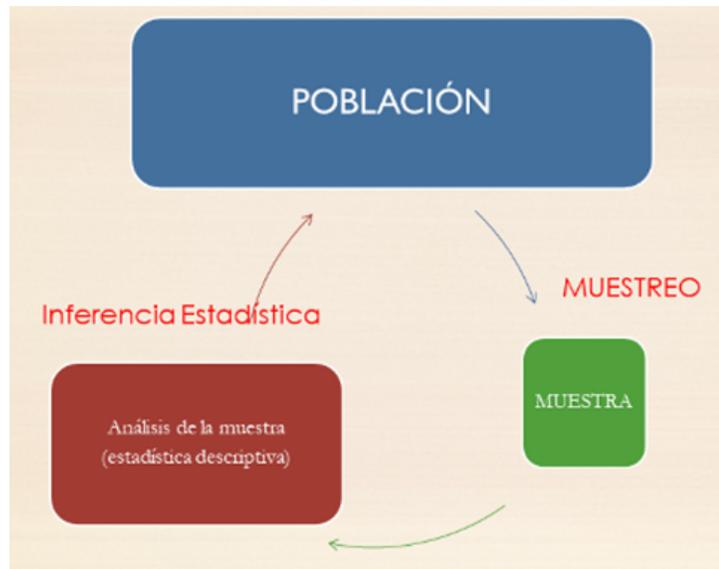


Figura 5.1: Proceso científico del análisis estadístico.

La variable en la población general se llama **variable aleatoria**.

Ej.

Cuando tenemos una base de datos de una población, en general tenemos una muestra de dicha población. Un fichero de datos de pacientes de un hospital concreto es una muestra de la población total de pacientes. En ese fichero de datos podemos tener mediciones de diferentes variables: altura, peso, temperatura corporal, niveles en sangre, si tiene una enfermedad o no (variable que podría estar representada con ceros y unos, cero si no tiene la enfermedad, uno si la tiene)... Las variables que hemos citado: *altura*, *peso*, *temperatura*, etc., son ejemplos de variables aleatorias.

El nombre de *variable aleatoria* hace referencia a que los valores de esta variable dependen del azar, mediante alguna ley concreta. El proceso de comprensión o interpretación matemática de esa aleatoriedad es el que realizaremos en este tema.

Ej.

Cuando lanzamos una moneda, por ejemplo, cien veces, y anotamos

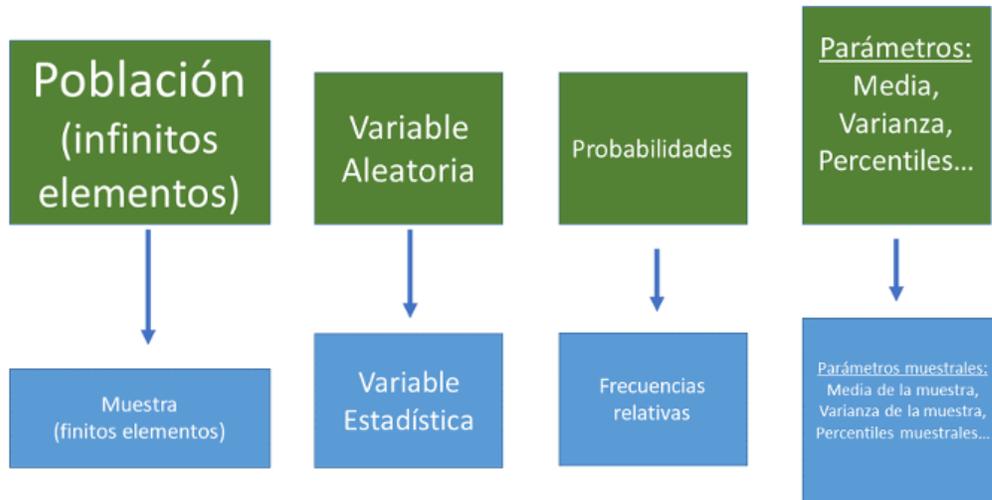


Figura 5.2: Esquema general de relación entre variables aleatorias y estadísticas.

1 si sale cara y 0 si sale cruz, tenemos una variable estadística que toma los valores 0 y 1, con frecuencias de aparición de, por ejemplo, 56 y 44, respectivamente. Cuando consideramos el proceso general de lanzar una moneda, tenemos una variable aleatoria que toma los valores 0 y 1 con probabilidad 0.5, respectivamente.

## 5.1 Tipos de variables aleatorias

Al igual que sucede con las variables estadísticas, las variables aleatorias pueden ser **discretas** o **continuas**, en función de que tomen un número finito (o infinito numerable)<sup>1</sup> de valores, o bien un número infinito no numerable de valores<sup>2</sup>, respectivamente.

<sup>1</sup>infinito, pero puede ordenarse, es decir cada elemento tiene uno anterior y otro posterior

<sup>2</sup>los elementos no tienen uno anterior y uno posterior

Ej.

La variable *número que sale al lanzar un dado con seis caras* es una variable aleatoria discreta (toma los valores del 1 al 6).

Ej.

La variable *número de veces que un pesado le pide una cita a una mujer casada* es una variable que toma los valores  $1, 2, \dots$  y no tiene fin (aunque las probabilidades de los sucesivos valores vayan disminuyendo). Se llama conjunto infinito numerable porque cada elemento  $i$  tiene un anterior ( $i - 1$ ) y un posterior ( $i + 1$ ).

Ej.

La variable *número de coches que pasan por un cruce en un tiempo determinado* es otro ejemplo de variable aleatoria discreta que toma un conjunto infinito pero numerable de valores.

Ej.

La variable *estatura* se mueve en un rango de valores determinado por el mínimo y el máximo de la población. Los valores que puede tomar dependen de la precisión del aparato de medida, con lo cual se trata de un conjunto no numerable de valores, es decir un intervalo  $[a, b]$ . En principio, la variable puede tomar cualquier valor dentro de ese intervalo o rango de valores  $[a, b]$ .

No es numerable, porque un número cualquiera puede tomar tantas cifras decimales como se quiera, y por eso no existe un elemento inmediatamente anterior y un elemento inmediatamente posterior.

## 5.2 Variables aleatorias discretas

Una variable aleatoria discreta  $X$  está definida por los valores que toma y sus probabilidades, las cuales deberán sumar 1.

$X$	$P(X = x_i)$
$x_1$	$p_1$
$x_2$	$p_2$
$\vdots$	$\vdots$
$x_n$	$p_n$

verificando que  $p_1 + \dots + p_n = 1$ . Esta tabla se conoce como ley de probabilidad, distribución de probabilidad, función de probabilidad o función de masa de probabilidad.

**Ej.**

Ante la observación de un paciente, la variable aleatoria que toma los valores 1 y 0 (1 si el paciente tiene una enfermedad, 0 si no la tiene).

$X$	$P(X = x_i)$
0	$1 - p$
1	$p$

recibe el nombre de **variable aleatoria de Bernoulli** de parámetro  $p$ , siendo  $p$  la probabilidad de tener la enfermedad.

**Ej.**

Cuando realizamos el experimento aleatorio “elegir un número al azar entre 1 y  $N$ ”, la variable aleatoria  $X$  = “valor que se observa” se llama **variable uniforme discreta**.

$X$	$P(X = x_i)$
1	$1/N$
2	$1/N$
$\vdots$	$\vdots$
$N$	$1/N$

## 5.3 Variables aleatorias continuas

Una variable aleatoria continua es aquella que puede tomar cualquier valor (al menos teóricamente) entre 2 fijados. Los valores de la variable (al menos teóricamente) no se repiten.

Ej.

“Tiempo observado al recorrer una cierta distancia”, “estatura”, “peso”, “nivel de colesterol en sangre”...

Todas las precisiones realizadas en el capítulo de variables estadísticas son igual de adecuadas en este caso. Cuando observamos valores de una variable aleatoria continua, existe una limitación en cuanto al número de valores que puede tomar la misma. Esto es, en la práctica, la variable no toma infinitos valores. A la hora de medir el peso o la estatura, por ejemplo, se trabaja con un número preciso de decimales (que puede ser grande pero nunca será infinito). Lo que se está haciendo es lo que se llama una discretización a la hora de tomar datos. Sin embargo, desde un punto de vista matemático, consideraremos siempre que una variable continua puede tomar infinitos valores. Esto nos permitirá trabajar con propiedades matemáticas que nos aportarán mucha información de la variable considerada.

### 5.3.1 Función de densidad

Igual que una variable aleatoria discreta viene caracterizada por su función de probabilidad, las variables aleatorias continuas vienen caracterizadas por una función llamada **función de densidad**, que es una generalización de la función de probabilidad.

Matemáticamente, una función  $f$  es una función de densidad si verifica dos propiedades:

- $f(x)$  es mayor o igual que cero en cualquier punto  $x$  (el dibujo de la función debe estar por encima del eje horizontal).
- $\int_{-\infty}^{\infty} f(x)dx = 1$  (el área bajo la curva y el eje horizontal vale uno).

El concepto de función de densidad procede de considerar que tenemos una población con todos sus (infinitos) datos o posibles valores y dibujamos el histograma, polígono de frecuencias o estimación de la densidad.

Supongamos que nos ponemos en medio de la calle y a cada mujer mayor de 18 años le preguntamos su estatura. Hacemos esto hasta tener una muestra de 15 datos y, a continuación, clasificamos los datos en intervalos, construimos el histograma y el polígono de frecuencias.

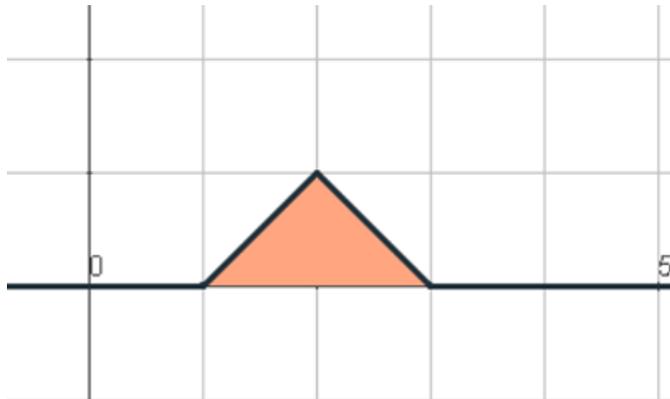
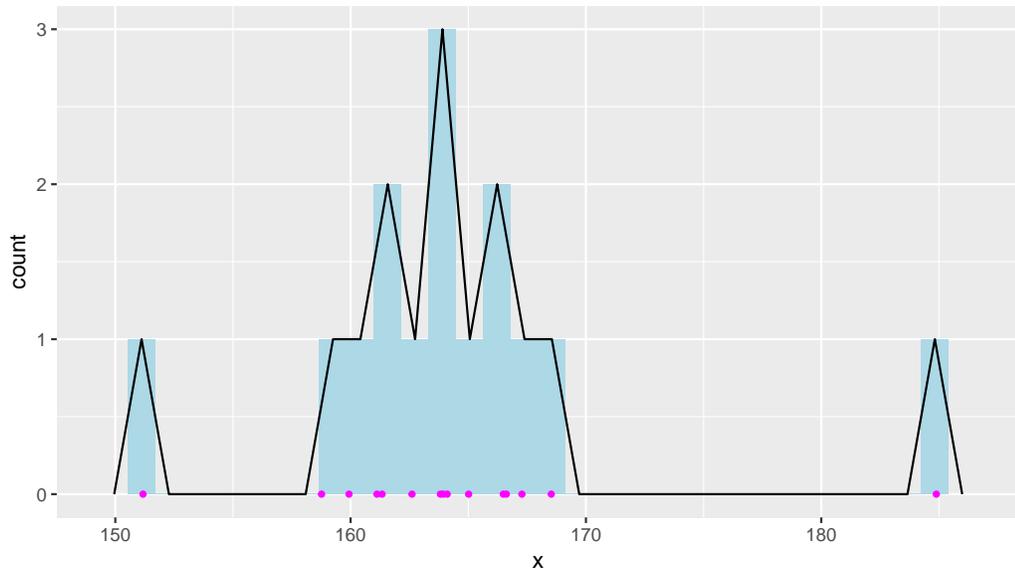


Figura 5.3: Ejemplo de una función de densidad bien simple.

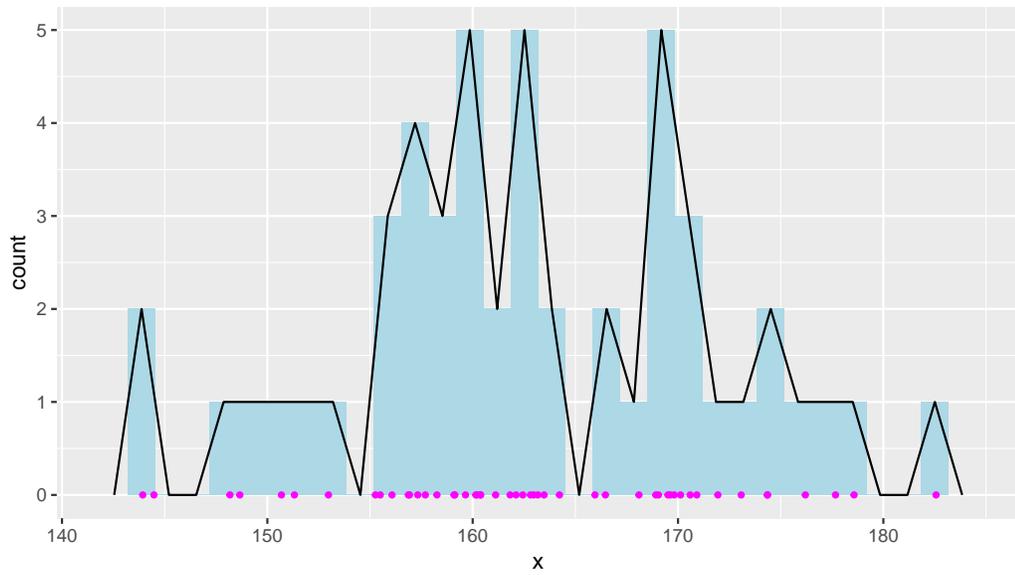
```
x=rnorm(15, 18,9)
print(x)
```

```
[1] 184.9 161.1 161.3 162.6 159.9 158.8 163.8 164.1 [9] 168.5 167.3 151.2 165.0
163.9 166.5 166.6
```

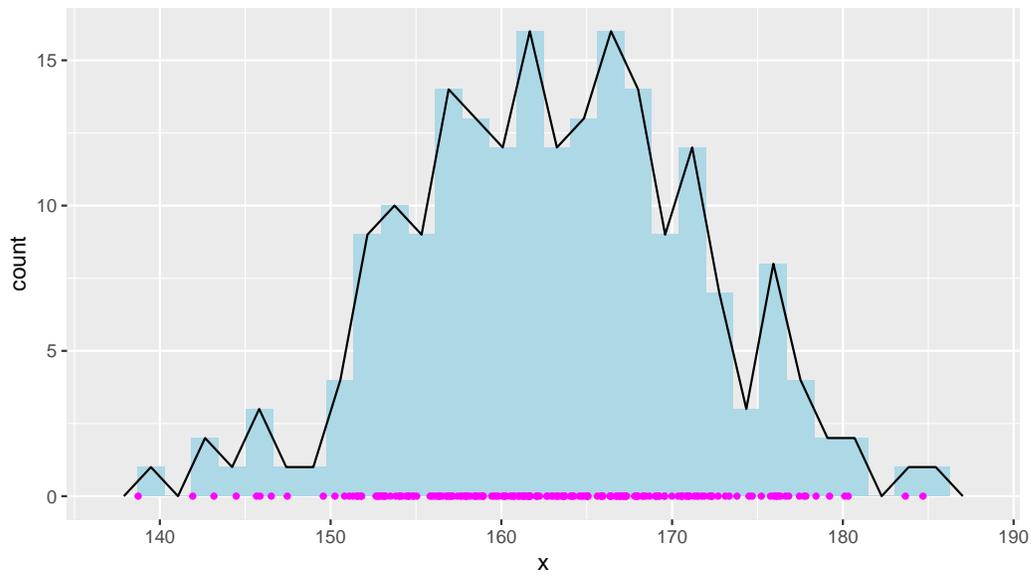
```
df<-data.frame(x)
N=length(x)
y=rep(0,N)
d2<-data.frame(x=x,y=y)
ggplot(df)+
  geom_histogram(aes(x=x), fill="lightblue")+
  geom_freqpoly(aes(x=x))+
  geom_point(data=d2, aes(x=x,y=y),
            size=1, col="magenta")
```



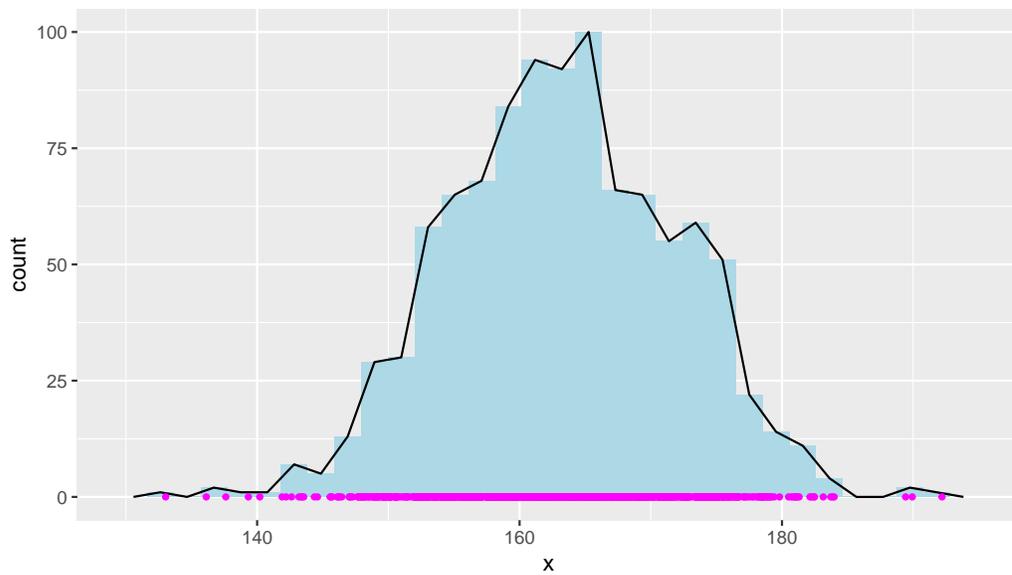
Ahora lo hacemos 50 veces



Ahora 200 veces



Ahora 1000 veces



Observamos como el polígono de frecuencias se va “perfilando” y suavizando, hasta construir el dibujo de una curva, a medida que tenemos más datos.

*Obviamente, no estamos tomando datos en la calle, sino “simulándolos” en el ordenador. Mediante la función de R `rnorm` estamos generando números aleatoriamente, pero que corresponden a medidas reales de mujeres en España mayores de 18 (la media es, aproximadamente, 163 cm y la desviación típica 9). En el tema siguiente, en el que se habla de la distribución normal, se entenderá mejor todo esto. Ahora, lo único con que tenemos que quedarnos es que, paulatinamente, vamos añadiendo más datos de estaturas en nuestra muestra.*

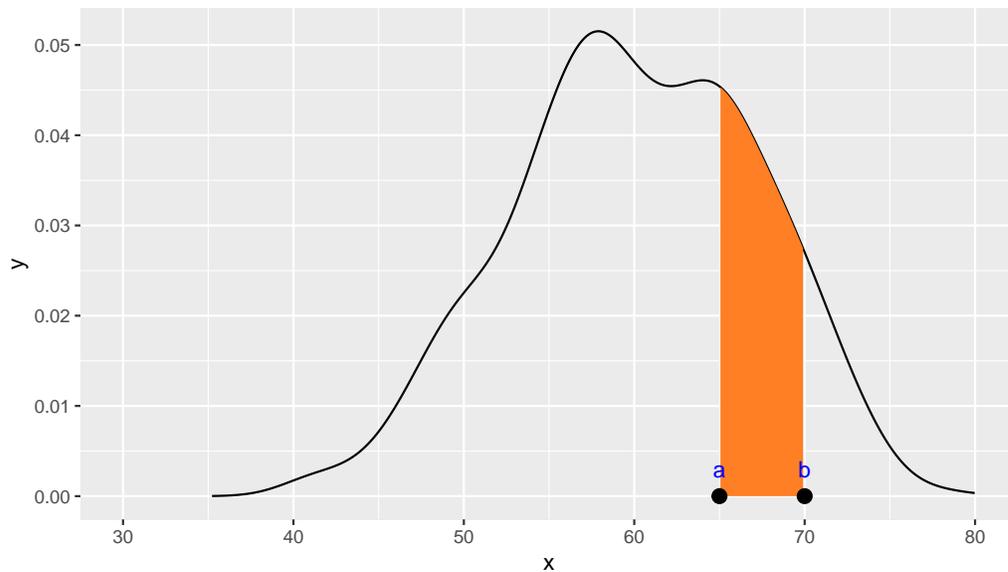
Con cada gráfica, el polígono de frecuencias acaba convirtiéndose en una curva que verifica las dos propiedades de la función de densidad (es una función no negativa y el área bajo la curva es uno, puesto que es el área bajo el polígono de frecuencias. Puede demostrarse geoméricamente que el área bajo un polígono de frecuencias coincide con el área existente bajo un histograma de frecuencias, y el área total del histograma corresponde al cien por cien de los datos).

La función de densidad corresponde, desde un punto de vista teórico, al polígono de frecuencias cuando tenemos todos los datos de la población (en teoría, infinitos).

Una vez expuesto que, en una variable aleatoria continua, las propiedades de la misma vendrán descritas por la función de densidad, indiquemos que las probabilidades se calcularán como una integral definida:

$$P(a < X < b) = \int_a^b f(x)dx$$

es decir, la probabilidad de que la variable aleatoria  $X$  tome valores entre dos números  $a$  y  $b$  corresponde al área bajo la curva  $f$ , el eje  $X$  y los puntos  $a$  y  $b$ .



En el caso de una variable aleatoria continua, la probabilidad de cualquier punto concreto  $a$  es cero, porque no hay área bajo la curva:

$$P(a < X < a) = \int_a^a f(x)dx = 0.$$

Esto puede sonar un poco raro, al principio. Si hablamos, por ejemplo, de la variable altura, nos podemos preguntar:

¿cuál es la probabilidad de medir 1.72?

Según lo que acabamos de decir, la probabilidad de un punto es cero. ¿Qué sucede? Pues que, como se comentó al principio del tema, en la práctica realizamos una discretización de la variable continua altura.

La pregunta matemáticamente correcta sería:

¿Cuál es la probabilidad de tener una estatura mayor que 1.72?

o

¿Cuál es la probabilidad de tener una estatura entre 2 valores  $a$  y  $b$ ?

Lo que sí se podría calcular sería algo como

Probabilidad de medir entre  $1.72 - a$  y  $1.72 + a$ , siendo  $a$  cualquier número, aunque sea muy pequeño.

La probabilidad de un intervalo (por pequeño que sea), siempre será un número más grande que cero.

Precisamente por este hecho, cuando calculemos la probabilidad de que una variable continua tome valores entre dos números  $a$  y  $b$ , podemos tener en cuenta que

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X \leq b) = P(a \leq X < b),$$

o sea, todas esas probabilidades dan lo mismo, porque considerar un punto más (o dos) no cuenta (al tener probabilidad cero).

“En algunos años, todas las grandes constantes de la física habrán sido estimadas y la única ocupación que quedará a los hombres de ciencia será la de refinar estas medidas al siguiente decimal”.

— Albert Michelson (1852-1931), famoso por su trabajo en la medición de la velocidad de la luz.

## 5.4 Esperanza Matemática de una variable aleatoria

Es la generalización de la media aritmética a toda la población, es decir, es la media de la variable aleatoria. También se llama **valor medio**, **valor esperado** o **esperanza matemática**, y se representa por la letra griega  $\mu$ .

Si  $X$  es una variable aleatoria discreta (representada, de manera general, por una tabla de valores  $x_i$  y probabilidades  $p_i = P(X = x_i)$ ),

$X$	$P(X = x_i)$
$x_1$	$p_1$
$x_2$	$p_2$
$\vdots$	$\vdots$
$x_n$	$p_n$

la esperanza se calcula como la media aritmética de los valores, es decir la suma de los valores por sus probabilidades (las probabilidades serían las

frecuencias relativas).

$$\mu = E(X) = \sum_{i=1}^k x_i \cdot p_i.$$

Recordemos que la media aritmética de una variable estadística se definió como

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n},$$

que, obviamente, sería equivalente a escribir

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i \cdot \frac{1}{n},$$

es decir, sería la esperanza de una variable cuyos valores aparecen todos con la misma probabilidad  $p_i = 1/n$ .

Si a una variable estadística la representamos por sus valores  $x_i$ , y sus frecuencias relativas son  $f_i = n_i/n$ , entonces la media aritmética se puede escribir como

$$\bar{x} = \sum_{i=1}^n x_i \cdot f_i,$$

esto es, suma de valores por frecuencias. En el caso de una variable aleatoria, las frecuencias se transforman en probabilidades (de ocurrencia). Por eso la esperanza es un *valor medio esperado*.

Si  $X$  es una variable aleatoria continua, la variable toma infinitos valores. El equivalente continuo de la suma es la integral. La fórmula matemática incluye en este caso a la función de densidad:

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx.$$

## 5.5 Varianza de una variable aleatoria

Se representa  $\sigma^2 = Var(X)$ , y la **desviación típica**  $\sigma$  es la raíz cuadrada (con signo positivo) de la varianza.

Igual que en el caso de variables estadísticas, mide la dispersión de la variable, y se calcula como la media de las desviaciones (elevadas al cuadrado) de los valores a su media:

$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2].$$

También puede calcularse como

$$\sigma^2 = E[X^2] - \mu^2$$

Si  $X$  es una variable discreta, la forma de hacer los cálculos será

$$\sigma^2 = \sum_{i=1}^k (x_i - \mu)^2 p_i = \left( \sum_{i=1}^k x_i^2 p_i \right) - \mu^2.$$

Si  $X$  es una variable continua,

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx,$$

y que también puede calcularse como

$$\sigma^2 = \left( \int_{-\infty}^{\infty} x^2 f(x) dx \right) - \mu^2.$$

**Ej.**

Consideremos la variable aleatoria que representa el número que puede salir en una ruleta:

La ley de probabilidad de esta variable es la de la tabla siguiente:

$X$	$P(X = x_i)$
0	1/37
1	1/37
⋮	⋮
36	1/37

Mediante la fórmula,

$$E(X) = \sum_{i=1}^k x_i \cdot p_i = \sum_{i=0}^{36} i \cdot \frac{1}{37} = \frac{1}{37} \sum_{i=0}^{36} i = \frac{1}{37} \cdot 666 = 18,$$

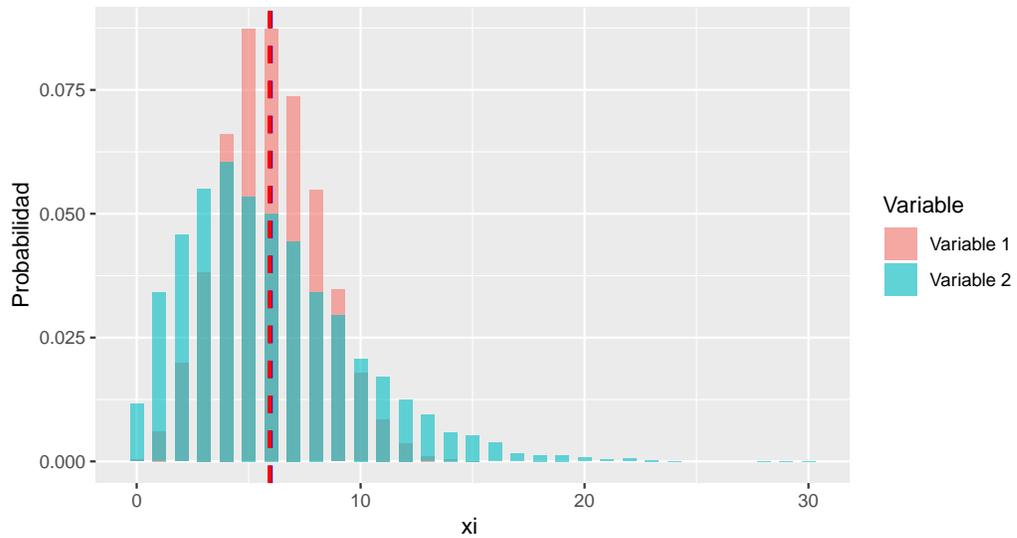


Figura 5.4: Ejemplo de dos variables discretas con prácticamente la misma media y diferente varianza (dispersión).

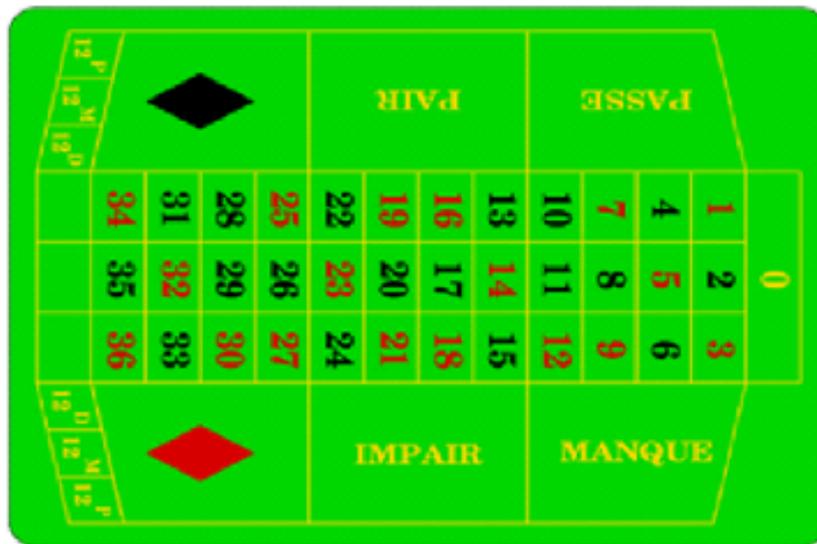


Figura 5.5: La ruleta francesa: números del 1 al 36, y además el cero.

lo que quiere decir que, si jugásemos a la ruleta infinitas veces, y fuésemos anotando el número que sale, la media aritmética de esos infinitos números daría 18.

Por medio de una simulación, podemos comprobarlo: vamos a ver qué sucede si jugamos en la ruleta y anotamos los números durante un gran número de jugadas. Esto juego lo “simulamos” en R generando números aleatorios que tomen valores enteros entre 0 y 36, lo cual se hace con el comando `runif` del siguiente modo:

```
x=runif(1000,0,36 # 1000 números
x=round(x) # les quitamos los decimales
mean(x)
```

```
## [1] 1821
```

Si, en vez de hacerlo 1000 veces, lo hacemos, por ejemplo, 10.000 veces:

```
x=runif(10000,0,36
x=round(x)
mean(x)
```

```
## [1] 1816
```

Observamos que, en efecto, la media de los números obtenidos se aproxima al valor medio o esperado que es 18.

Vamos a considerar ahora la variable  $X$ =*ganancia al apostar a un número concreto*. Las reglas de la ruleta francesa (la que tiene un cero) son: cuando apostamos una cantidad a un número concreto entre 1 y 36 (al cero no podemos apostar) y no acertamos, perdemos la cantidad apostada; si acertamos, recibimos 35 veces la cantidad apostada.

De esta forma, si llamamos  $c$  a la cantidad apostada, vemos que la variable  $X$  toma sólo dos valores:  $-c$  si perdemos (sale cualquier número, incluido el cero, excepto el que hemos apostado), y  $c \cdot 35$  si ganamos (sale el número apostado).

La variable  $X$  tiene la siguiente ley de probabilidad:

$X$	$P(X = x_i)$
$-c$	$36/37$
$35 \cdot c$	$1/37$

Esto es, de cada 37 jugadas, teóricamente 1 vez ganamos y el resto perdemos, y la variable  $X$  mide la cantidad que recibimos.

La esperanza o valor esperado de esta variable es

$$E(X) = -c \cdot \frac{36}{37} + 35 \cdot c \cdot \frac{1}{37} = -c \cdot \frac{1}{37}$$

Como vemos, para un jugador, la media o valor esperado siempre es negativo. ¿Qué significa este valor? Que, si se juega a la ruleta infinitas veces, el valor medio que se espera ganar es negativo (para el jugador; para la banca es positivo). En la práctica, si  $N$  es un número muy grande de jugadas, el total de dinero que ganará el jugador va a ser  $N \cdot E(X) = -N \cdot c \cdot \frac{1}{37}$ .

Si consideramos un juego tan simple como lanzar una moneda, apostar 1, ganar 1 si sale cara y perder 1 si sale cruz, la variable aleatoria es

$X$	$P(X = x_i)$
$-1$	$1/2$
$1$	$1/2$

La esperanza es  $E(X) = 1 \cdot 0.5 - 1 \cdot 0.5 = 0$ . Esto se llama *juego de suma nula*, que significa que, en teoría, si juegan 2 jugadores, los dos pierden tantas veces como ganan. Vemos que no es lo que ocurre en el caso de la ruleta. La ruleta está pensada para que, a la larga, gane el casino.

**Ej.**

Vamos a jugar a la ruleta 100 veces y ver cuanto dinero podríamos ganar:

```
x=round(runif(100,0,36))
# Vamos a suponer que siempre apostamos 1 euro al número 12
premio=ifelse(x==12, 35, -1)
sum(premio) # contamos la cantidad que ganamos (o perdemos)
```

```
## [1] 44
```

Si, en vez de jugar 100 veces, jugamos 100.000 veces, veamos qué podría ocurrir.

```
X=round(runif(10000,0,35))
# Vamos a suponer que siempre apostamos 1 euro al número 12
premio=ifelse(X==12, 35, -1)
sum(premio)
```

```
## [1] 100
```

Estos dos ejemplos son simulaciones. En algunos casos ganaremos y otras perderemos, pero nunca van a ser cantidades desorbitadas, y habrá que jugar muchas horas para llegar a 100.000 apuestas

### 5.5.1 La falacia del jugador

En el siguiente trozo de programación construimos una variable  $X$  que va desde 10 al valor *tope* (número que elijamos). Para cada valor de  $X$  tiramos una moneda y contamos el número de caras y de cruces y los vamos anotando. Es decir, si  $X = 12$ , lanzamos la moneda 12 veces. Si  $X = 23$ , lanzamos la moneda 23 veces, y así sucesivamente...

Una vez realizados los lanzamientos, en la tabla de resultados guardamos los resultados obtenidos: número de lanzamientos de la moneda, número de caras, número de cruces, diferencia entre número de caras y número de cruces y proporción entra ambos números.

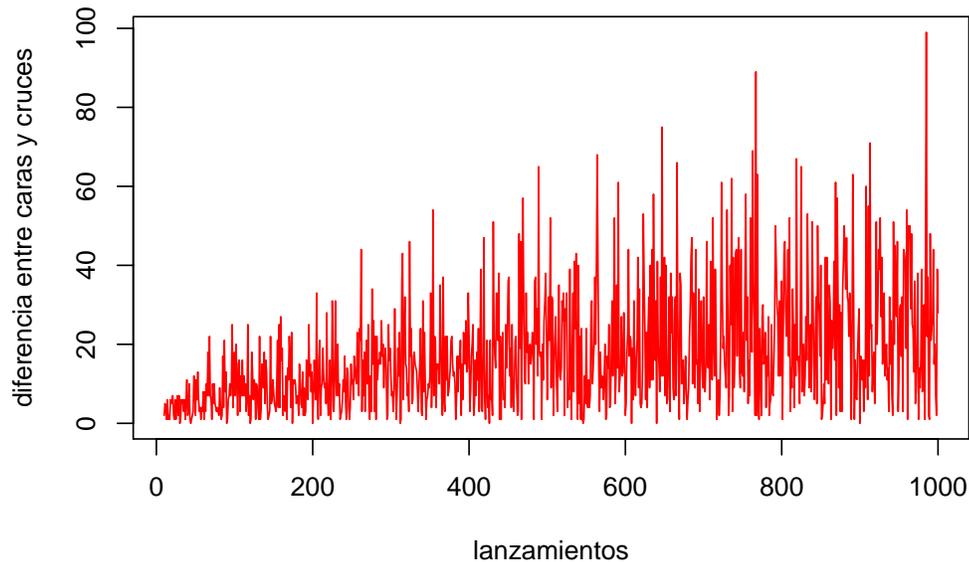
Empezamos lanzando la moneda 10 veces porque así nos “aseguramos” de que el número de cruces no sea cero (evidentemente puede ocurrir, pero sería raro), y así la proporción no nos dé infinito al dividir por cero.

```
tope=1000
X<seq(from=10, to=tope, by=1 )
N=length(X)
n_caras<e(0)
n_cruces<e(0)
A<e(0)
for (i in 1:N)
```

```
{
  A=runif(X[i])
  a=length(A[A<0.5])
  n_caras[i]<-a
  n_cruces[i]<-X[i]-a
}
dif=n_caras-n_cruces
prop<n_caras/n_cruces
resultados<data.frame(lanzamientos=X,
                      caras=n_caras,
                      cruces=n_cruces,
                      diferencia=dif
                      proporcion=prop)
```

Fijémonos que ocurre si dibujamos el resultado anterior en una gráfica que nos muestre las diferencias entre caras y cruces según que lanzamos la moneda  $N$  veces, con  $N$  de 10 a 1000.

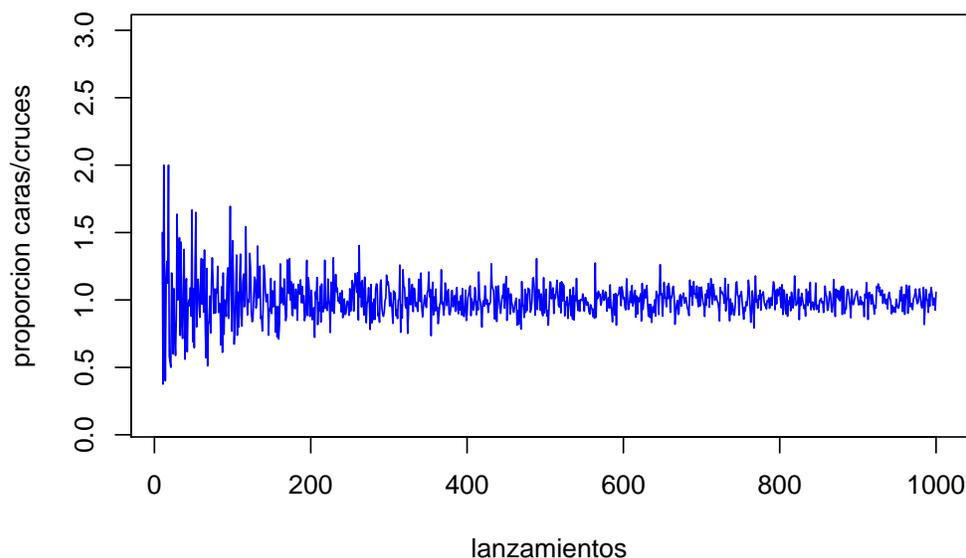
```
plot(X, abs(dif),
     col="red",
     type="l",
     #ylim=c(0.7,1.4),
     xlab="lanzamientos",
     ylab="diferencia entre caras y cruces")
```



Según podemos observar, la diferencia se agranda cuanto mayor sea  $N$ . A la derecha del gráfico vemos que, lanzando la moneda cerca de 1000 veces, hay diferencia de 80 resultados entre caras y cruces. Este hecho puede resultar un poco extraño, a primera vista: cuantas más veces lancemos la moneda, el número de caras y el de cruces debería tender a acercarse, puesto que cara y cruz son igual de probables.

Realmente es lo que sucede, pero para verlo mejor hay que considerar no la diferencia, sino las proporciones:

```
plot(X, prop,
     col="blue",
     type="l",
     ylim=c(0.1,3),
     xlab="lanzamientos",
     ylab="proporcion caras/cruces")
```



La ley de estabilidad de las frecuencias (que la frecuencia converge a la probabilidad) nos garantiza que la proporción entre el número de caras y de cruces tiende a ser 1 cuantas más veces se arroje la moneda. La diferencia entre el número de caras y cruces también tiende a cero, pero hay que ver que una diferencia de 80, en 1000 lanzamientos, es un número pequeño, proporcionalmente.

La conocida como *falacia del jugador* consiste en creerse que, porque hayan salido de forma continuada un número de caras relativamente grande, en la siguiente jugada deberá salir cruz puesto que los resultados deberán compensarse. Así, en una ruleta, si han salido 3 o 4 veces seguidas números de color rojo, pensar que en el siguiente movimiento de la ruleta es más probable que salga negro es una falacia. Cada jugada es independiente de la anterior. La probabilidad de que salga negro o rojo va a ser  $1/2$ , independientemente de que hayan salido 2, 3 o 40 negros seguidos.

En particular, está documentado que, en el Casino de Monte Carlo, el 18 de agosto de 1913, la bola cayó 26 veces consecutivas en números de color negro. Por eso, a esta falacia también se le conoce como “falacia de Monte Carlo”.

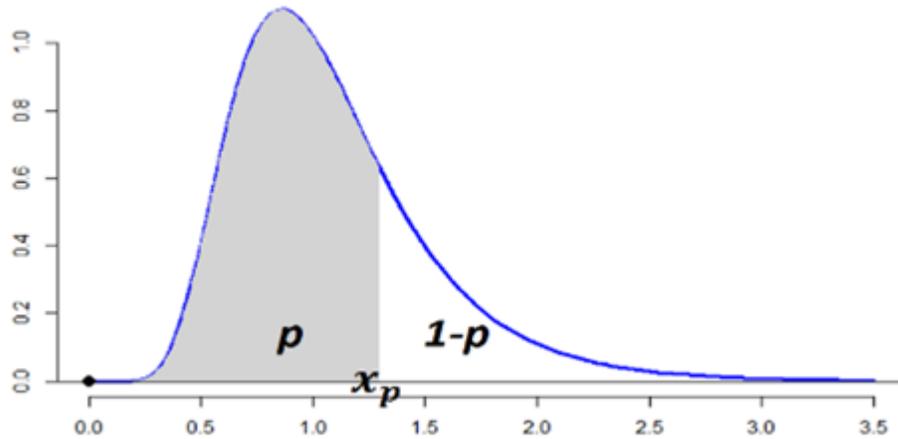


Figura 5.6: Cuantil o percentil.

Otros planteamientos (incorrectos) equivalentes son: «Un resultado aleatorio tiene más probabilidades de ocurrir, si no ha ocurrido durante cierto periodo de tiempo»; o «Un resultado tiene menos probabilidades de ocurrir, si no ha ocurrido durante cierto periodo de tiempo».

## 5.6 Mediana y Cuantiles (o percentiles)

La definición y significado de estos conceptos son el mismo que el que se usó en el tema de variables estadísticas.

Así, por ejemplo, la **mediana** es el valor que deja a un lado y a otro el 50 por ciento de los valores, o de la frecuencia (en este caso de la probabilidad):

$$P(X < Me) = 0.5$$

Dado un valor  $0 < p < 1$ , el cuantil o percentil  $p$  es el valor  $x_p$  que deja a su izquierda una probabilidad  $p$ .

$$P(X < x_p) = p$$

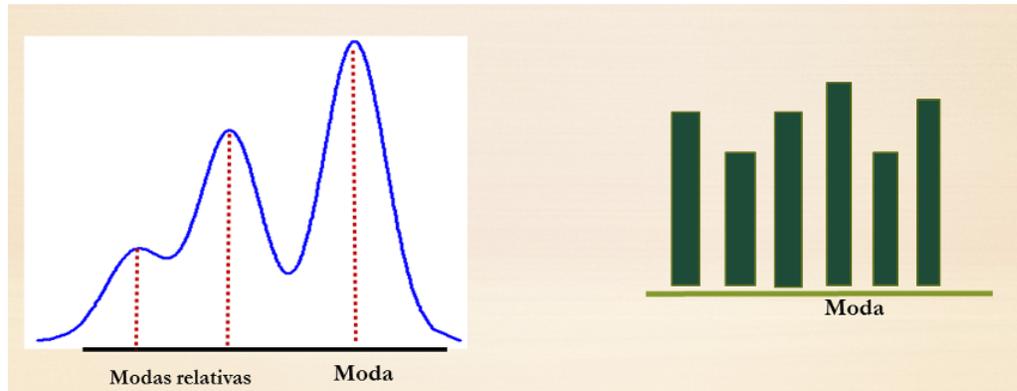


Figura 5.7: Distribución con varias modas.

## 5.7 La moda

La moda  $M_o$  es el valor más probable o frecuente, es decir, el valor con mayor probabilidad, si la variable es discreta, o el valor que maximiza la función de densidad si es continua.

La moda no tiene por qué ser única. En el caso de una única moda se dice que la distribución es unimodal, y con más modas multimodal.

## 5.8 Variables discretas notables

En esta sección hablaremos de las variables aleatorias discretas *con nombre propio*, es decir aquellas cuya forma o ley de probabilidad son utilizadas muy a menudo. La descripción de cada variable se hará de forma breve. Quien desee ampliar conocimientos, puede consultar (del Río, 2018). Seguramente hay textos mucho mejores para ello, pero este es mío.

### 5.8.1 La variable de Bernoulli

Supongamos un experimento aleatorio con sólo dos posibles resultados: un suceso  $A$  que llamaremos *éxito* y un suceso  $\bar{A}$  que llamaremos *fracaso*. Supondremos que las probabilidades de ocurrencia son  $p = P(A)$  y  $q = 1 - p = P(\bar{A})$ .

A la variable aleatoria que toma los valores 1 (si ocurre el suceso éxito) y 0 (si ocurre el suceso fracaso) se le llama *variable aleatoria de Bernoulli* de



Figura 5.8: Nespresso, what else?.

parámetro  $p$ .

Se escribe  $X \in \text{Bernoulli}(p)$ .

Ej.

Lanzar una moneda y anotar 1 si sale cara y 0 si sale cruz es un ejemplo de variable de Bernoulli.

Ej.

Observar un paciente para ver si tiene una enfermedad concreta cuya probabilidad es  $p$ . La variable de Bernoulli asociada a este experimento cuenta 1 si el paciente tiene la enfermedad, y 0 si no la tiene.

### 5.8.2 Variable Binomial

Supongamos que se realizan  $n$  experimentos de Bernoulli de manera sucesiva, siendo cada experimento o prueba independiente del anterior.

La Variable  $X$  = número de veces que ocurre el suceso  $A$  en las  $n$  pruebas o número de éxitos en las  $n$  pruebas, recibe el nombre de variable binomial de parámetros  $n$  y  $p$  ( $p = P(A) = p(\text{éxito en 1 prueba})$ ).

Se escribe  $X \in B(n, p)$ .

La variable  $X$  puede tomar los valores  $0, 1, 2, \dots, n$ , siendo la probabilidad con que los toma:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \text{ donde } \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

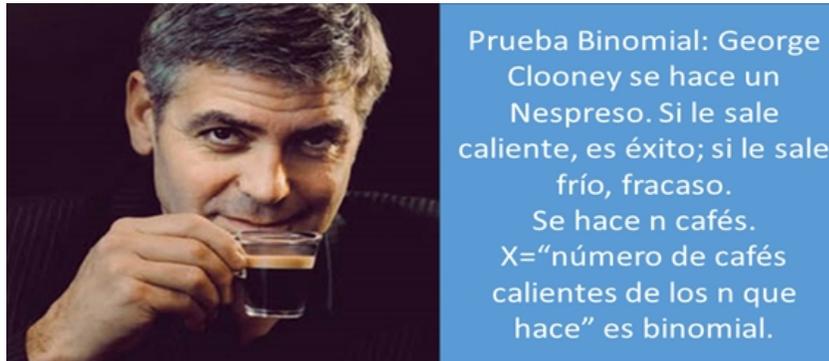


Figura 5.9: Nespresso, binomial.

La media y la varianza (los cálculos no son sencillos) son:

$$E(X) = np \text{ y } Var(X) = npq.$$

Ej.

Un acusado va a ser declarado inocente o culpable por un jurado popular. Para ser condenado es necesario que al menos 7 personas de las 10 del jurado voten culpable. Dado que en los programas de televisión ya han dado muchos detalles del caso, los miembros del jurado están atendiendo al twitter o leyendo el periódico en vez de escuchar al fiscal y al abogado, porque van a decidir tirando una moneda al aire. ¿Cuál es la probabilidad de que el acusado sea declarado inocente?

**Solución.**

Definimos  $A$ ="éxito"="inocente".  $p = P(A) = 0.5$

$X$ ="número de éxitos en 10 pruebas"  $\in B(10, 0.5)$ .

La probabilidad de ser declarado inocente es  $P(X \geq 4)$ .

$$P(X \geq 4) = \sum_{k=4}^{10} \binom{10}{k} 0.5^k 0.5^{10-k} = 0.82.$$

Con R se calcula de la siguiente forma:

```
1-pbinom(3,10,0.5)
```

```
## [1] 0.28
```

`pbinom(3,10,0.5)` es la propiedad acumulada hasta 3, en una distribución binomial de parámetros  $n = 10$  y  $p = 0.5$ . Esto es:  $P(X \leq 3)$ . Como el complementario del suceso  $\{X \geq 4\}$  es  $\{X < 4\} = \{X \leq 3\}$ , calculamos la probabilidad de ser declarado inocente de esta forma sencilla.

### 5.8.2.1 Propiedad aditiva

La Variable binomial es reproductiva respecto al parámetro  $n$ . Si  $X \in B(n_1, p)$  e  $Y \in B(n_2, p)$  son 2 variables independientes, la suma  $X + Y \in B(n_1 + n_2, p)$  (esta propiedad es generalizable a un número finito de variables).

De hecho, esto ocurre puesto que la variable Binomial  $B(n, p)$  es la suma de  $n$  variables aleatorias independientes de Bernoulli de parámetro  $p$ . Obviamente, el número de éxitos en  $n$  pruebas =  $1 + 0 + 1 + \dots + =$ , es decir, es la suma de unos y ceros, según haya éxito o fracaso en cada prueba.

### 5.8.3 Variable de Poisson

Un proceso de Poisson es un experimento aleatorio donde se observa la aparición de un suceso concreto (éxito) sobre un soporte continuo (generalmente el tiempo). Además, debe cumplirse que los sucesos ocurren de forma independiente y con media estable (el número medio de sucesos por unidad de medida es constante).

Ejemplos interesantes de procesos de Poisson son: clientes que acuden a un mostrador por unidad de tiempo, llamadas por unidad de tiempo a una centralita, defectos por metro de cable, baches por kilometro de autopista...

En un proceso de Poisson, la variable  $X = \text{número de éxitos en un intervalo}$  se dice que sigue una distribución de Poisson de parámetro  $\lambda$ . Se escribe  $X \in Pois(\lambda)$ .

Su distribución de probabilidad es

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$



Figura 5.10: Café de Poison.

Se verifica que

$$E(X) = Var(X) = \lambda,$$

es decir, que el parámetro  $\lambda$  es precisamente el número medio de sucesos que estamos contando, y, en este caso, también coincide con la varianza de la variable.

La variable de Poisson es una generalización de la variable binomial. Supongamos que un experimento de Bernoulli tiene un suceso éxito  $A$  con probabilidad  $p = P(A)$  muy pequeña (por ejemplo, una enfermedad rara, de baja probabilidad). Puede demostrarse que, si  $X$  es la variable  $Bi(n, p)$ , que mide el número de sucesos  $A$  en  $n$  pruebas,  $X$  puede aproximarse por una variable de Poisson de parámetro  $\lambda = np$ .

Por este motivo, la distribución de Poisson es conocida como distribución de los **sucesos raros**.

$$P(X = k) = \binom{n}{k} p^k q^{n-k} \rightarrow e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{con } \lambda = np$$

En la práctica, esta aproximación funciona si  $n > 30$  y  $p < 0.1$ .

**Ej.**

Supóngase que en un hotel donde descansan sufridos cazadores de elefantes ocurren de manera aleatoria e independiente dos accidentes de caídas con rompimiento de cadera por semana. Determinar la probabilidad de que ocurra un accidente en una semana.

**Solución**

$X$ ="número de accidentes por semana" sigue una distribución de Poisson de media  $\lambda = 2$ .

$$P(X = 1) = e^{-2} \frac{2^1}{1!} = 0.270.$$

O sea, existe un 27 por ciento de probabilidades de que se te fastidie la semana de caza por caerte y romperte la cadera.

**5.8.3.1 Propiedad aditiva**

La Variable de Poisson es reproductiva respecto al parámetro  $\lambda$ .

**5.8.4 Variable Binomial negativa**

Supongamos que se realiza un experimento de Bernoulli hasta que se obtiene el éxito número  $r$ . Definamos la variable  $X$ =*número de fracasos hasta obtener el éxito*  $r$ .  $X$  se dice que sigue una distribución Binomial Negativa de parámetros  $r$  y  $p$ . Se escribe  $X \in BN(r, p)$ .

Su ley de probabilidad es

$$P(X = k) = \binom{r+k-1}{k} p^r q^k, \quad k = 0, 1, 2, \dots$$

Se obtiene que

$$E(X) = \frac{rq}{p} \quad \text{y} \quad \text{Var}(X) = \frac{rq}{p^2}.$$

**Ej.**

Al llamar al servicio de atención al cliente de una compañía de teléfonos, la probabilidad de que se consiga resolver el problema es 0.1. Pepe necesita que le arreglen la conexión a internet y, por otro lado, quiere que le devuelvan 2 euros que le cobraron mal hace dos décadas. ¿Cuál es la probabilidad de que tenga que llamar más de 5 veces?

**Solución**

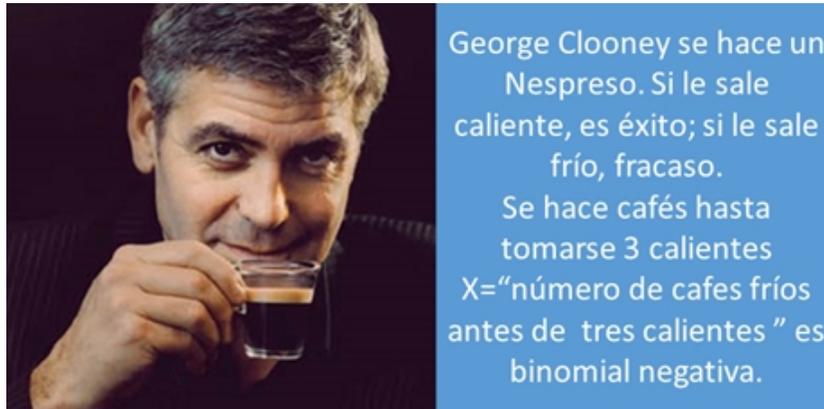


Figura 5.11: Y aún da para más...

$X$ ="número de fracasos hasta obtener el éxito 2"  $\in BN(2, 0.1)$ .

$$P(X > 5) = 1 - P(X \leq 4) = 1 - \sum_{k=0}^4 \binom{2+k-1}{k} 0.1^2 0.9^k = 0.885.$$

### 5.8.5 Variable Hipergeométrica

Supongamos que tenemos una población de  $N$  elementos, que se divide en dos clases:  $A$  y  $\bar{A}$ . El número de elementos de cada clase los denotamos como  $n_A$  y  $n_{\bar{A}}$ . Lógicamente  $n_A + n_{\bar{A}} = N$ .

Supongamos que se extrae una muestra de tamaño  $n$  de la población, sin reemplazamiento. La variable  $X$  = número de elementos de la clase  $A$  en la muestra se dice que sigue una distribución hipergeométrica de parámetros  $N, n_A$  y  $n$ . Se escribe  $X \in H(N, n_A, n)$ .

Su ley de probabilidad es

$$P(X = k) = \frac{\binom{n_A}{k} \cdot \binom{n_{\bar{A}}}{n-k}}{\binom{N}{n}}, \quad k = \max\{0, n + n_A - N\}, \dots, \min\{n_A, n\}.$$

Sus parámetros media y varianza:

$$E(X) = \frac{n \cdot n_A}{N}, \quad Var(X) = \frac{N-n}{N-1} \cdot \frac{n \cdot n_A}{N} \cdot \left(1 - \frac{n_A}{N}\right).$$

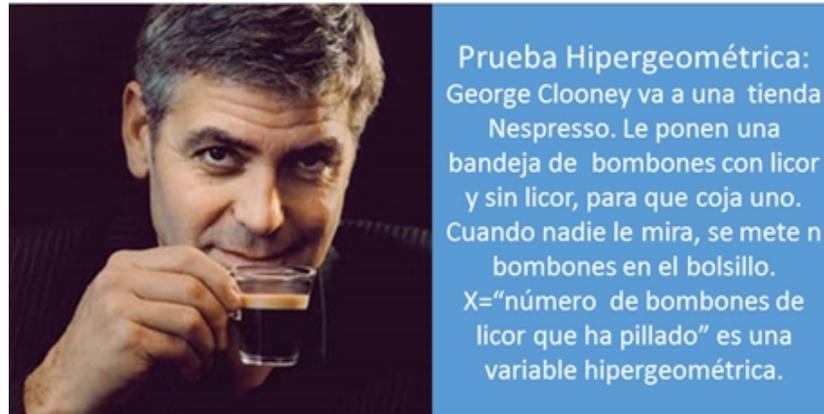


Figura 5.12: Última.

Si se escribe  $p = \frac{n_A}{N}$ ,  $q = 1 - p$ , se obtiene:

$$E(X) = np, \quad \text{Var}(X) = npq \frac{N-n}{N-1}.$$

Ej.

Un opositor a registrador de la propiedad tiene que preparar cien temas. En el examen se sacan tres a sorteo, de los cuales deberá exponer uno. El opositor decide estudiar solamente la mitad y probar suerte. ¿Cuál es la probabilidad de que apruebe?

**Solución:**

Dividimos la población, de tamaño  $N = 100$ , en 2 clases: los temas que ha estudiado el opositor ( $A$ , con  $n_A = 50$ ) y los que no ha estudiado ( $\bar{A}$ , con  $n_{\bar{A}} = 50$ ).

La variable  $X$ ="número de temas que el opositor conoce, en la muestra de tamaño 3" sigue una distribución hipergeométrica de parámetros  $N = 100$ ,  $n_A = 50$  y  $n = 3$ .

Con que se sepa uno de los temas, el opositor aprobará. Tenemos que calcular, entonces,

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) =$$

$$= 1 - \frac{\binom{50}{0} \cdot \binom{50}{3}}{\binom{100}{3}} = 1 - 0.1212 = 0.8788.$$

Como vemos, la probabilidad de aprobar es alta estudiando solo la mitad.

## 5.9 Variables aleatorias continuas notables

Como en el caso de las variables discretas, particularizamos ahora algunas variables continuas que se emplean habitualmente más que otras, y por eso también poseen nombre propio. La variable aleatoria continua más conocida y utilizada es la variable normal, que, por sus peculiaridades, merece un capítulo aparte (el siguiente). De forma análoga al caso anterior, ahora únicamente comentamos un par de ejemplos (distribución uniforme y distribución exponencial), y remitimos al querido lector a cualquier otro manual de estadística (incluido el mío, que ya no pongo cita porque a estas alturas se sabrá a cuál me refiero) para profundizar en estas variables y en otras (variable beta, gamma, Weibul...). En los casos particulares de las leyes de potencias y de la distribución normal, creo, sin embargo, que se ha efectuado aquí un resumen exhaustivo bastante amplio (en el caso de la distribución normal dudo que se encuentre otro parecido, y no por la calidad -que se da por supuesta-, sino por lo extenso).

### 5.9.1 Variable uniforme continua

Una variable aleatoria continua  $X$  se dice que sigue una distribución uniforme entre dos valores  $a$  y  $b$  (se representa  $X \in U(a, b)$ ) si su función de densidad tiene la siguiente expresión:

$$f(x) = \frac{1}{b-a}, \quad x \in [a, b]$$

y vale 0 en cualquier otro caso.

Esta variable es la generalización, al caso continuo, de la variable uniforme discreta (la que da a todos los valores la misma probabilidad). La variable uniforme reparte de manera continua y equivalente la probabilidad, es decir, intervalos de igual longitud (dentro de  $[a, b]$ ) tienen igual probabilidad.

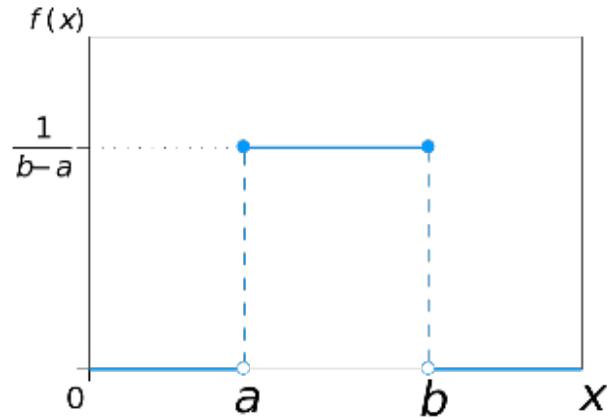


Figura 5.13: función de densidad de una variable aleatoria uniforme entre  $a$  y  $b$ .

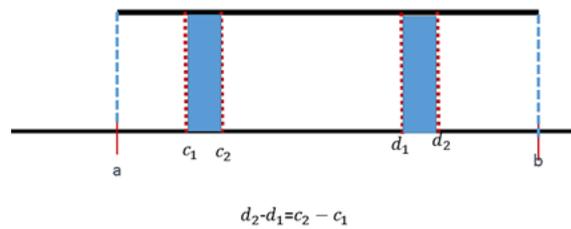


Figura 5.14: Intervalos con la misma longitud tienen la misma probabilidad (área).

Recuérdese que, en las variables continuas, la probabilidad entre 2 puntos  $v_1$  y  $v_2$  es el área bajo la función de densidad. En este caso, el área sería el área de un rectángulo, es decir:

$$P(v_1 < X < v_2) = \int_{v_1}^{v_2} f(t) dt = (v_2 - v_1) \cdot \frac{1}{b - a}$$

Gráficamente, está claro que intervalos de igual longitud tienen el mismo área y, por lo tanto, igual probabilidad. Matemáticamente, también es muy simple:

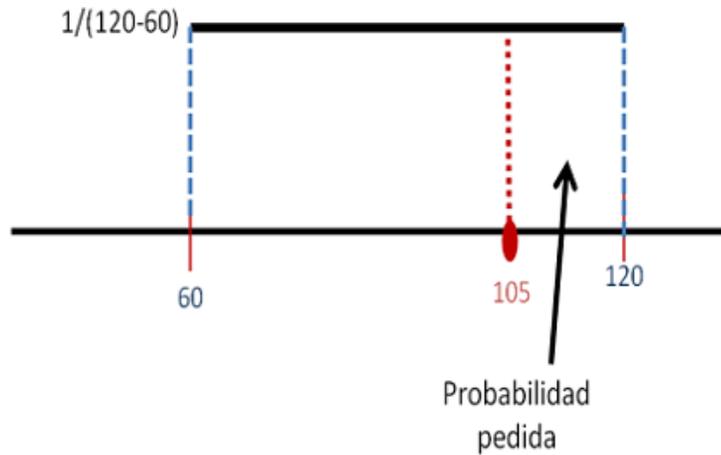


Figura 5.15: La probabilidad es el área del rectángulo que va desde 105 a 120.

$$\int_{c_1}^{c_2} f(t)dt = (c_2 - c_1) \cdot \frac{1}{b - a} = \int_{d_1}^{d_2} f(t)dt = (d_2 - d_1) \cdot \frac{1}{b - a}$$

### 5.9.1.1 Media y Varianza

Puede comprobarse que  $E(X) = \frac{a + b}{2}$  y  $Var(X) = \frac{(b - a)^2}{12}$ . Es decir, el valor medio es el punto medio del segmento o intervalo  $(a, b)$ , lo cual coincide con lo que nos diría la intuición: un reparto uniforme de probabilidad supone que la media esté en el punto medio.

**Ej.**

Un estudiante llega a la parada del bus justo cuando este acaba de marcharse. El siguiente tardará en llegar como mínimo una hora, y puede llegar en cualquier momento en el transcurso de la hora siguiente. ¿Cuál es la probabilidad de que el estudiante tenga que esperar más de una hora y cuarenta y cinco minutos?

### Solución

El tiempo que tarda en llegar el bus es una variable uniforme  $X \in [60, 120]$ . La probabilidad de que tarde más de 45 minutos en llegar (pasada la primera hora) es  $P(X > 60 + 45) = 0.25$ .

Cuando en R generamos números aleatorios entre 0 y 1 (o entre 2 números cualesquiera), estamos trabajando con una variable aleatoria uniforme.

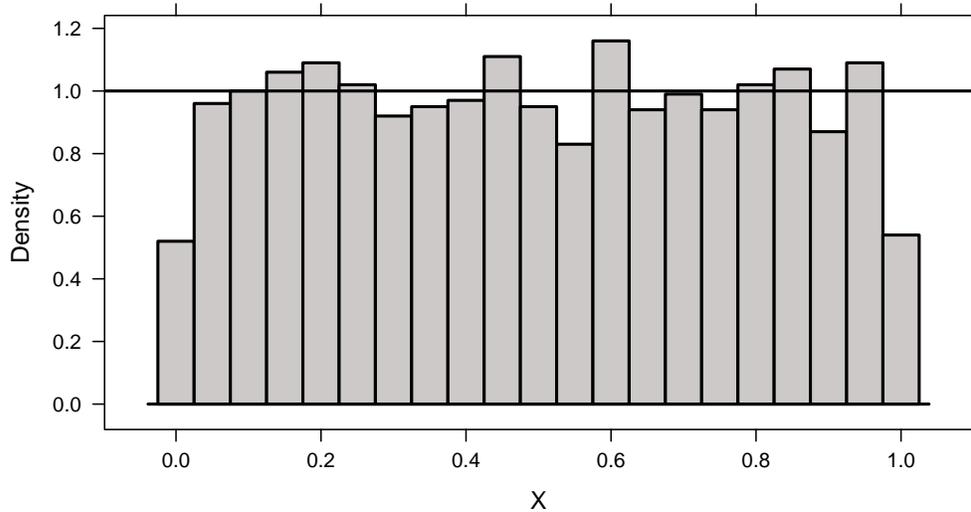
La siguiente línea de código, por ejemplo, genera 10 números aleatorios con distribución uniforme entre 0 y 1.

```
X=runif(10,0,1)
X
```

```
## [1] 0.0780 0.34320 0.41522 0.01053 0.0307 0.22457
## [7] 0.17260 0.8132 0.44757 0.74377
```

Vemos que nos aparecen, en efecto, 10 números entre 0 y 1. Si en vez de 10 números generamos unos cuantos más, por ejemplo, 2000, y hacemos un histograma o una estimación de la densidad, veremos que se parece a la densidad de la distribución uniforme.

```
X=runif(2000,0,1)
datos=data.frame(X)
histogram(~X, data=datos,
          col="snow3", dcol="mediumblue",
          h=1, type='density',
          width=0.05, lwd=2 )
```



### 5.9.2 Variable exponencial

Una variable continua  $X$  se dice que sigue una distribución exponencial de parámetro  $\lambda$  (siendo  $\lambda$  cualquier número real mayor que cero) si su función de densidad es:

$$f(x) = \lambda \cdot e^{-\lambda x}, \quad x > 0$$

y vale cero en cualquier otro caso (es decir, la variable exponencial solo toma valores positivos).

Se escribe  $X \in \text{Exp}(\lambda)$ .

Puede comprobarse que, independientemente de cuánto valga  $\lambda$ , la integral  $\int_0^{\infty} \lambda \cdot e^{-\lambda x} dx = 1$ .

En la siguiente gráfica dibujamos la forma de la función de densidad para 3 valores diferentes de  $\lambda$ :

```
### Exponencial
curve(dexp(x, rate = 0.5), xlim = c(0, 4), ylim = c(0, 2),
      xlab = "x", ylab = "Función de densidad")
curve(dexp(x, rate = 1), col = "red", lty = 3, add = T)
```

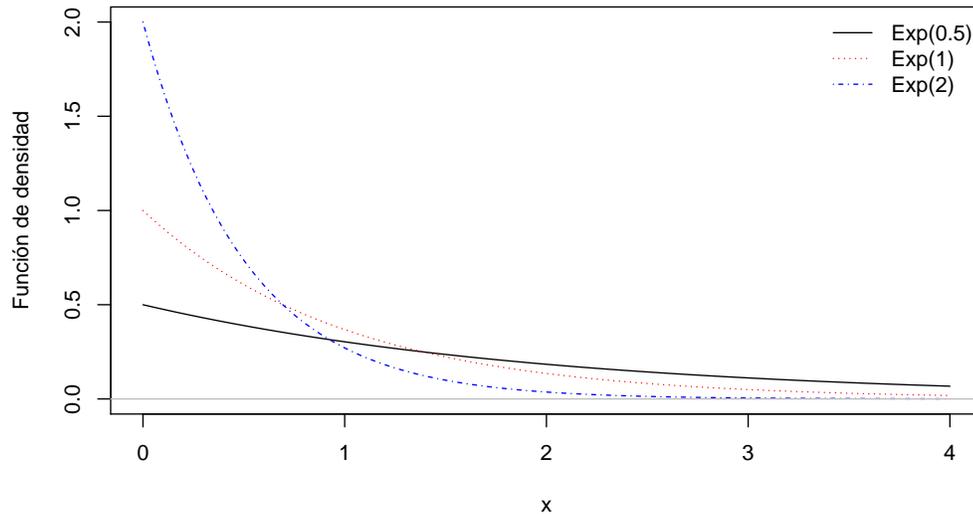


Figura 5.16: Gráficas de la densidad de la variable exponencial.

```

curve(dexp(x, rate = 2), col = "blue", lty = 4, add = T)
abline(h = 0, col = "gray")
legend("topright", c("Exp(0.5)", "Exp(1)", "Exp(2)"), col = c("black",
  "red", "blue"), lty = c(1, 3, 4), bty = "n")

```

Existen numerosos ejemplos de variables que siguen esta ley. La duración de componentes electrónicos, baterías, células en enfermedades, tienen este tipo de comportamiento: los valores más altos son mucho menos probables que los valores más bajos.

El tiempo de espera, en muchas ocasiones, también sigue una distribución exponencial: tiempo en ser atendidos en una cola en una ventanilla, tiempo entre la llegada de dos taxis a una parada...

La magnitud de los terremotos que se producen en una determinada región sigue, por regla general, una distribución de este tipo. En la terminología de sismos, se le conoce como ley de Gutenberg-Richter. También es una ley de este tipo la que rige la velocidad de los vientos que atraviesan una determi-

nada región. La idea básica es que terremotos (o huracanes) de magnitudes bajas hay muchos más (por suerte) que terremotos con magnitud alta.

También se verifica que la duración de las relaciones entre parejas sigue una distribución exponencial: hay muchas parejas que duran poco y pocas que duran mucho (Newman, 2005).

El físico (además de matemático y pacifista) británico Lewis Fry Richardson (1881-1953) acumuló datos de las estadísticas de “disputas mortales”, en las que incluyó la guerra entre los tipos de asesinatos. Observó que existía una especie de ley de Gutenberg-Richter de los conflictos en la que todos los acontecimientos, desde el asesinato de una persona (“conflictos de magnitud 0”) a las dos guerras mundiales (“magnitud 7”), obedecían a una distribución de probabilidades también de este tipo.

### 5.9.2.1 Media y Varianza

Puede comprobarse que  $E(X) = \frac{1}{\lambda}$  y  $Var(X) = \frac{1}{\lambda^2}$ .



La duración **media** de una prótesis de cadera se estima en 20 años. La duración se corresponde con una variable aleatoria de tipo exponencial.

En este caso, al ser la media 20, tenemos que

$$E(X) = \frac{1}{\lambda} = 20 \longrightarrow \lambda = \frac{1}{20} = 0.05$$

De acuerdo con esto, podemos calcular probabilidades del tipo *probabilidad de que la prótesis dure como mínimo  $r$  años* como

$$P(X > r) = 1 - P(X \leq r) = 1 - (1 - e^{-\lambda r}) = e^{-\lambda r}$$

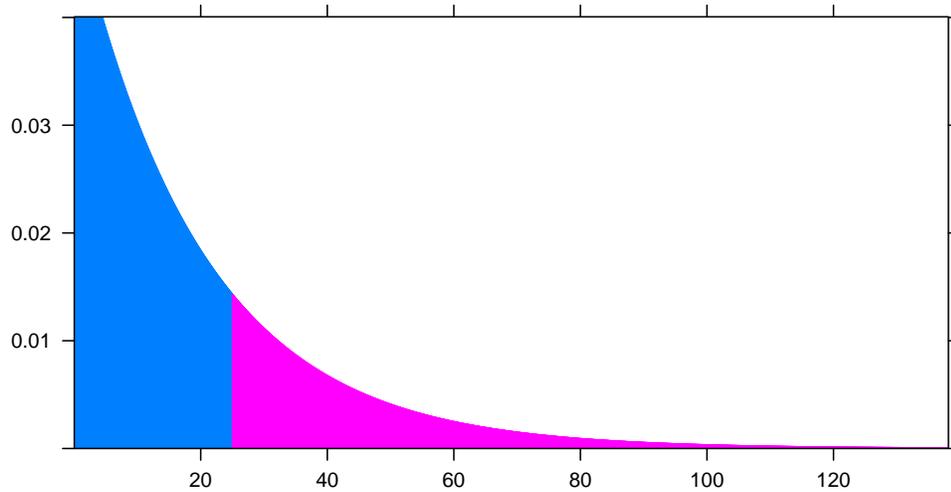
Aquí hemos utilizado que la probabilidad acumulada

$$P(X \leq r) = \int_0^r \lambda \cdot e^{-\lambda t} dt = 1 - e^{-\lambda r}$$

mediante integración por partes.

Si, por ejemplo,  $r = 25$ , la probabilidad de que una prótesis dure más de 25 años será

$$P(X > 25) = e^{-0.05 \cdot 25} = 0.28$$



Todo esto último son muchas matemáticas. Obviamente, con R es mucho más simple:

```
pexp(25, rate=0.05)
```

```
## [1] 0.7135
```

nos da la probabilidad acumulada en el valor 25 para una distribución exponencial de parámetro  $\lambda$  (*rate*) igual a 0.05.

Comparemos la media de una variable exponencial con la media de una variable uniforme. En la variable uniforme, la media es el punto medio. Sin embargo, en el caso que estamos considerando, la vida media de una prótesis de cadera es 20 años, pero este valor no deja igual probabilidad a cada lado (si lo hiciera, la media coincidiría con la mediana). Veamos qué área o probabilidad deja a la derecha el valor 20.

$$P(X > 20) = e^{-0.05 \cdot 20} = 0.3678,$$

lo que significa que deja 36.78 por ciento del área a la derecha ( $1 - 36.78 = 63.22$  a su izquierda). Por lo tanto, la media es el percentil 63, aproximadamente.

### 5.9.3 Las leyes de potencias (power law)

Sobre este apartado, gran parte de la información se ha extraído del trabajo de (Clauset et al., 2009).

La distribución exponencial es un caso particular de lo que se conoce como leyes de potencias. Matemáticamente, una variable  $X$  sigue una ley de potencias si su función de densidad es de la forma

$$f(x) \propto x^{-\alpha},$$

donde  $\alpha$  es el parámetro de escala, que habitualmente cae entre 1 y 3, aunque puede haber excepciones. El símbolo  $\propto$  significa *proporcional* (por ejemplo, la densidad exponencial  $e^{-x}$  es de este tipo, pues  $e^{-x} \propto x^{-2}$ ) (de manera aproximada, podemos decir que los dibujos de ambas funciones son prácticamente similares).

En la práctica, pocos fenómenos empíricos obedecen las leyes de potencias para todos los valores de  $x$ . En general, la ley de potencias se aplica a partir de un valor mínimo  $x_{min}$ . En tal caso, se dice que la cola de la distribución sigue una ley de potencias. Estas distribuciones se llaman *heavytail distributions* (colas pesadas). Significa que su cola (habitualmente a la derecha) es descendente, pero de forma muy pausada, es decir que pueden aparecer valores extremadamente grandes y muy alejados del valor modal o del valor mediana.

La probabilidad de que  $X$  sea mayor que un número  $x$  viene dada por:

$$Pr(X > x) = \begin{cases} \left(\frac{x_{min}}{x}\right)^\alpha & \text{si } x \geq x_{min}, \\ 1 & \text{si } x < x_{min}. \end{cases}$$

#### 5.9.3.1 La ley de Pareto y la regla 80-20

Uno de los ejemplos más conocidos es el relativo a la distribución de la riqueza de un país. En este caso, el parámetro  $\alpha$  es conocido como índice de Pareto. Esto se debe a que, en 1897, Vilfredo Pareto (1848-1923) afirmó que las rentas hacia el extremo más rico del espectro social están distribuidas de acuerdo a una ley de potencias. Esto implica que una gran parte de la riqueza de la nación está en manos de unos pocos individuos. En general, los porcentajes son del tipo 80 – 20, es decir el 20 por ciento de unos pocos poseen el 80 por ciento de la riqueza, mientras que el 80 por ciento restante solo tiene el

### ■ Distribución de la ganancia bruta anual

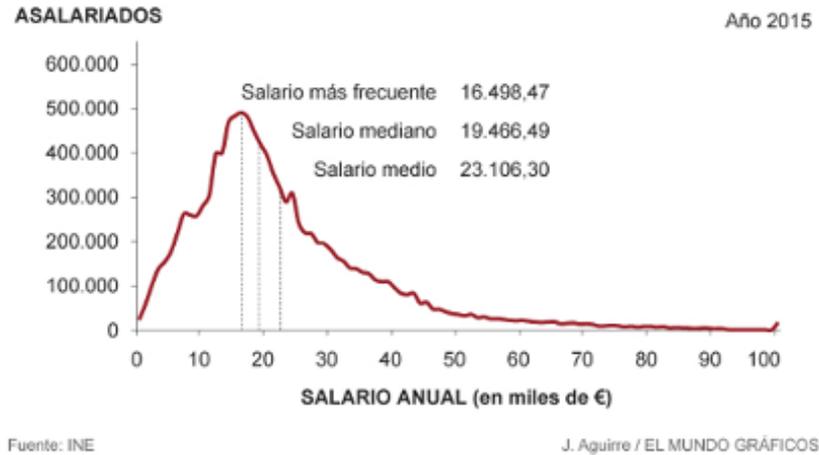


Figura 5.17: Gráfica del INE de los salarios en 2015 en España. Puede observarse una forma de tipo exponencial o ley de potencias a partir del salario medio, aproximadamente.

20 por ciento. Este tipo de regla 80 – 20 también se da en otras situaciones, como vamos a ver (buscad en google “regla 80 20” y encontraréis mucha más información y curiosidades).

En la gráfica de los salarios de España en 2015 se ve perfectamente esta forma. El valor modal del salario (el más frecuente) es 16.498 euros. El valor mediana es 19.466. Vemos que a partir de algún valor cercano a 30.000 o 40.000 euros brutos anuales la cola desciende pero lo hace muy pausadamente, puesto que hay personas que pueden ganar mucho más que la mayoría, pero son pocas. El cálculo del valor exacto  $x_{min}$  a partir del cual la cola de la distribución sigue una ley de potencias es un problema matemático con una cierta complejidad (Resnick, 2007).

#### 5.9.3.2 Otras leyes de potencias

Una variable aleatoria que siga una ley de potencias puede ser tanto continua como discreta. Hemos visto cual es su función de densidad si la variable es continua. Si es discreta, su ley de probabilidad es  $P(X = x) = C \cdot x^{-\alpha}$  para  $x > x_{min}$ .

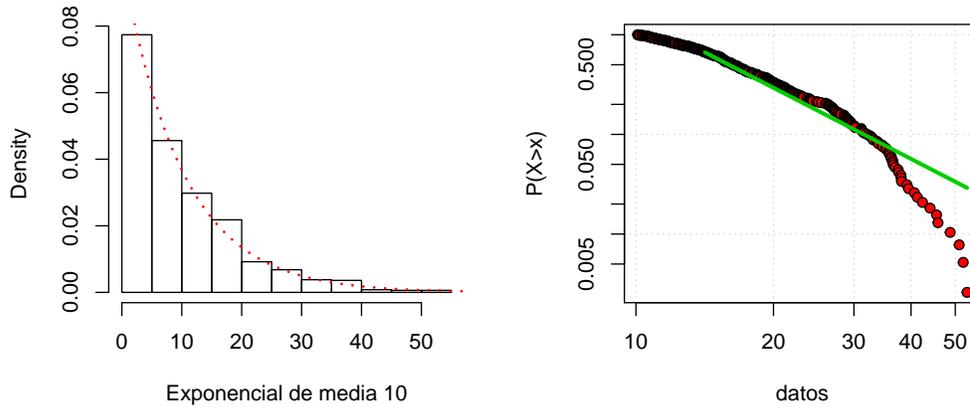


Figura 5.18: Distribución exponencial, representada mediante su función de densidad, y mediante un gráfico log-log.

Seguramente todos hemos oído hablar de leyes de este tipo en algún momento. El tamaño de los grupos de amigos, por ejemplo. Habitualmente, hay mucha gente que tiene pocos amigos y poca gente que tiene muchos. Muchos novelistas que tienen pocos lectores y pocos novelistas que tienen muchos. Muchos trabajos científicos son poco o escasamente leídos, y unos pocos son leídos o consultados por muchos...

Una representación muy habitual de las leyes de potencias es un diagrama llamado *log-log plot*. Es un gráfico en el que, en el eje horizontal, se representa la variable  $X$  en escala logarítmica, y en el eje vertical el logaritmo de  $1 - F(x)$ , que corresponde a la probabilidad  $P(X > x)$ . De esta forma, se obtiene una función descendente. En cierto momento, ese descenso se convierte en una línea recta, cuya pendiente (salvo el signo) corresponde al índice  $\alpha$  de la ley de potencias.

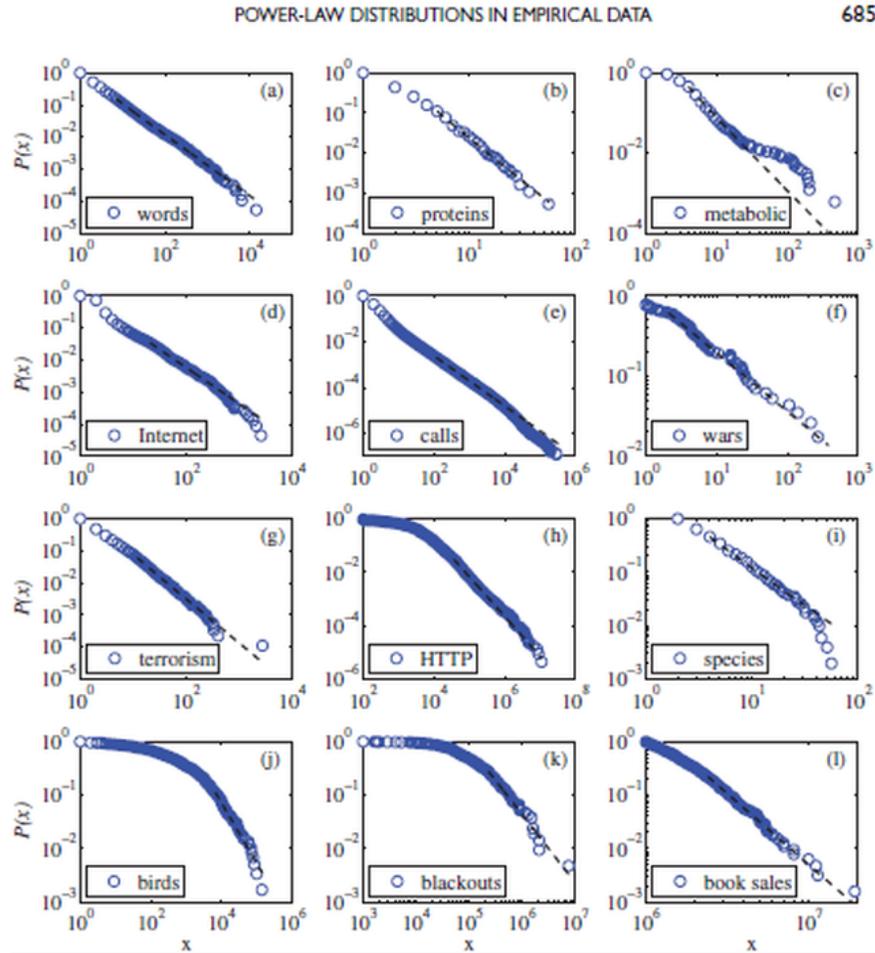
Veamos un ejemplo con una variable exponencial (las unidades del gráfico son logarítmicas, pero aparecen las originales para facilitar la interpretación del mismo).

En (Clauset et al., 2009) se detallan los siguientes ejemplos (ver a continuación las Figuras 8 y 9 del artículo) de casos empíricos reales que se ajustan

a una ley de potencia:

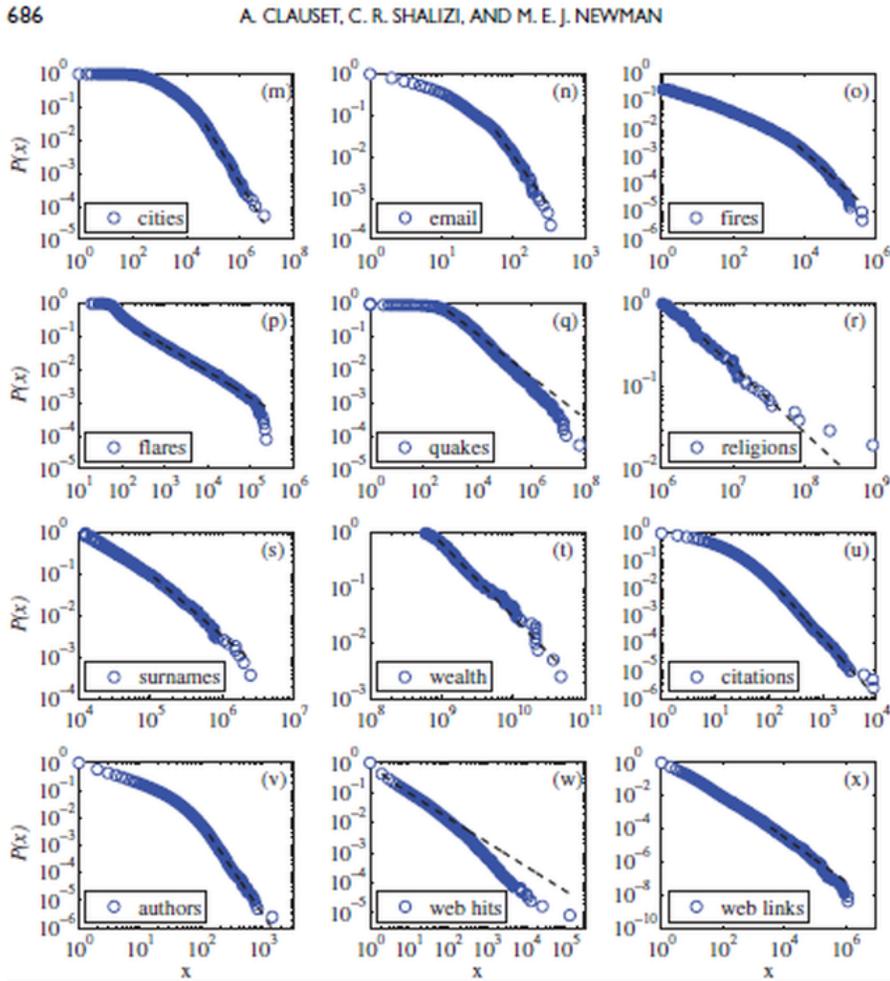
- a) La frecuencia de aparición de palabras únicas en la novela *Moby Dick* (Herman Melville).
- b) Los grados (es decir, el número de compañeros de interacción distintos) de las proteínas, en la red de interacción proteica de la levadura *Saccharomyces cerevisiae*.
- c) Los grados de metabolitos en la red metabólica de la bacteria *Escherichia coli*.
- d) Los grados de nodos en la representación de la red de Internet en mayo de 2006.
- e) El número de llamadas recibidas por los clientes de la compañía telefónica AT&T, en los Estados Unidos, durante un solo día.
- f) La intensidad de las guerras entre 1816 y 1980, medida como el número de muertos en combate por cada 10.000 habitantes de las naciones en guerra.
- g) La gravedad de los atentados terroristas perpetrados en todo el mundo, entre febrero de 1968 y junio de 2006, medida como el número de muertes directamente resultantes.
- h) El número de bytes de datos recibidos -como resultado del comando `http-` en un gran laboratorio de investigación, durante las 24 horas del día en junio de 1996. En términos generales, esta distribución representa la distribución del tamaño de archivos web transmitidos a través de Internet.
- i) El número de especies por género de mamíferos. Este conjunto está compuesto principalmente de especies vivas hoy en día, pero también incluye algunas especies recientemente extinguidas, donde “reciente” en este contexto significa “en los últimos miles de años”.
- j) El número de avistamientos de aves de diferentes especies en América del Norte en el año 2003.
- k) El número de clientes afectados por cortes de electricidad en los Estados Unidos entre 1984 y 2002.
- l) El número de copias de los libros más vendidos en los Estados Unidos durante el período de 1895 a 1965.
- m) La población de las ciudades en el Censo de los Estados Unidos del año 2000.
- n) El tamaño de las libretas de direcciones de correo electrónico de los usuarios de una universidad.

- o) El tamaño en acres de los incendios forestales que ocurrieron en tierras federales de EE.UU. entre 1986 y 1996.
- p) Intensidad máxima de rayos gamma de las erupciones solares entre 1980 y 1989.
- q) La intensidad de los terremotos que ocurrieron en California entre 1910 y 1992.
- r) El número de seguidores de religiones y sectas, según consta en el sitio web adherents.com.
- s) La frecuencia de ocurrencia de apellidos en los Estados Unidos en el censo de 1990.
- t) El patrimonio neto agregado en dólares de los individuos más ricos de los Estados Unidos en octubre de 2003.
- u) El número de citas recibidas por los artículos científicos del Science Citation Index entre 1981 y 1997.
- v) El número de trabajos académicos escritos o co-escritos por matemáticos en la base de datos MathSciNet de la Sociedad Matemática Americana.
- w) El número de “hits” recibidos por los sitios web de AOL en un sólo día.
- x) El número de enlaces a sitios web, encontrados en un rastreo de la Web en 1997, de unos 200 millones de páginas web.



**Fig. 8** The CDFs  $P(x)$  and their maximum likelihood power-law fits for the first 12 of our 24 empirical data sets. (a) The frequency of occurrence of unique words in the novel *Moby Dick* by Herman Melville. (b) The degree distribution of proteins in the protein interaction network of the yeast *S. cerevisiae*. (c) The degree distribution of metabolites in the metabolic network of the bacterium *E. coli*. (d) The degree distribution of autonomous systems (groups of computers under single administrative control) on the Internet. (e) The number of calls received by U.S. customers of the long-distance telephone carrier AT&T. (f) The intensity of wars from 1816–1980 measured as the number of battle deaths per 10000 of the combined populations of the warring nations. (g) The severity of terrorist attacks worldwide from February 1968 to June 2006, measured by number of deaths. (h) The number of bytes of data received in response to HTTP (web) requests from computers at a large research laboratory. (i) The number of species per genus of mammals during the late Quaternary period. (j) The frequency of sightings of bird species in the United States. (k) The number of customers affected by electrical blackouts in the United States. (l) The sales volume of bestselling books in the United States.

Figura 5.19: Figura 8 del trabajo de Clauzet, Shalizi y Newman (2009).



**Fig. 9** The CDFs  $P(x)$  and their maximum likelihood power-law fits for the second 12 of our 24 empirical data sets. (m) The populations of cities in the United States. (n) The sizes of email address books at a university. (o) The number of acres burned in California forest fires. (p) The intensities of solar flares. (q) The intensities of earthquakes. (r) The numbers of adherents of religious sects. (s) The frequencies of surnames in the United States. (t) The net worth in U.S. dollars of the richest people in the United States. (u) The numbers of citations received by published academic papers. (v) The numbers of papers authored by mathematicians. (w) The numbers of hits on web sites from AOL users. (x) The numbers of hyperlinks to web sites.

Figura 5.20: Figura 9 del trabajo de Clauzet, Shalizi y Newman (2009).



# Capítulo 6

## La Variable Normal o Gaussiana

Karl F. Gauss no fue el inventor de la campana (función de densidad de la variable normal), pero esta lleva su nombre en su honor. El concepto *normal* es sinónimo aquí de *común*, de general, como opuesto a lo atípico o poco probable. Esta concepción tiene a su vez un desarrollo histórico no sin interés. Sin embargo, para la introducción de la variable, el nombre resulta más que adecuado, a partir de la gran cantidad de ejemplos que veremos a continuación. La ley normal, también llamada **ley de los errores**, fue utilizada por primera vez en 1733, cuando de Moivre la propuso como aproximación de la distribución binomial. En 1783, Laplace la propuso para describir la distribución de los errores accidentales en la medición de una magnitud física. El término “ley normal” no aparece hasta 1894, de la mano, precisamente, de Karl Pearson (Porter, 1986).

“No se me ocurre nada tan propenso a impresionar la imaginación como la maravillosa forma de orden cósmico expresada por la ley del error (la curva de Gauss). Si un salvaje la hubiera comprendido, le habría rendido culto como a una divinidad. Cuanto más grande es la multitud y mayor la anarquía aparente, mas perfecto es su dominio. Es la ley suprema de la sinrazón: siempre que una muestra grande de elementos caóticos juegan un papel similar en magnitud, emerge una forma insospechada y bella de regularidad, presente en estado latente desde el principio”.

— Sir Francis Galton

Unos años después de escribir este párrafo tan rimbombante (*¿Quién dijo que la mejor lírica y las matemáticas son incompatibles?*), Galton escribió una versión 2.0 de la misma, quizá más políticamente correcta. Sustituyó “Si un salvaje [...] divinidad” por “Los griegos, de haberla conocido, la habrían divinizado”.

De manera general, una variable aleatoria continua  $X$  se dice que sigue una distribución **normal** o **gaussiana** de parámetros  $\mu$  y  $\sigma$  si su función de densidad es de la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

Se verifica que

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \mu, \quad Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \sigma^2$$

esto es,



la **media** o esperanza es el parámetro  $\mu$  y la **varianza** es el parámetro  $\sigma$ .

La función de densidad es simétrica respecto de la media  $\mu$ , es decir, áreas a la derecha y a la izquierda (probabilidades) coinciden. Las áreas entre valores de  $\mu - k\sigma$  y  $\mu + k\sigma$  ( $k = 1, 2, 3$ ) pueden verse en la siguiente gráfica, llamada la campana de Gauss.



Esta variable, cuando se consideran los valores  $\mu = 0$  y  $\sigma = 1$ , se llama *Normal estándar* o *Normal tipificada*.



Si tenemos una variable  $X$  con media  $\mu$  y desviación típica  $\sigma$ , a partir de ella se puede construir lo que se conoce como variable tipificada o estandarizada  $Z = \frac{X - \mu}{\sigma}$ , verificándose que esta variable  $Z$  tiene media cero y desviación típica uno.

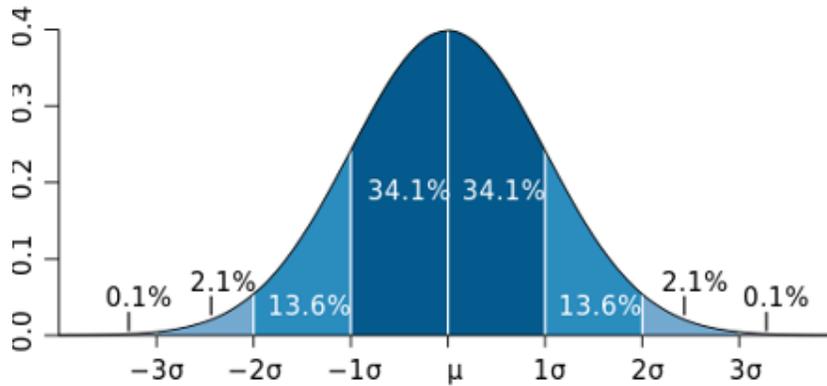


Figura 6.1: La campana de Gauss presenta simetría y la misma distribución de áreas, independientemente de cuáles sean sus parámetros media y desviación típica.

Esta propiedad resultaba de mucha importancia hasta hace unos años, cuando había que calcular probabilidades de esta variable “a mano”, y se utilizaba siempre la llamada *tabla de la distribución normal*, que es una tabla con probabilidades (áreas) de la distribución normal estándar, de media cero y desviación típica 1. Gracias a la propiedad de estandarización de cualquier variable normal, se pueden calcular las probabilidades asociadas a dicha variable tipificando la misma.

## 6.1 Ejemplos de la distribución normal

“muchos sucesos que, por su naturaleza, parecen caprichosos e inciertos, y para los cuales en ningún caso individual existe grado obtenible de conocimientos que nos permita preverlos, ocurren con un grado de regularidad que raya lo matemático, si se tienen en cuenta cifras considerables”

— John Stuart Mill (*Un sistema de la lógica*, 1862)

Así como hemos visto que hay otros casos de variables muy representativas de un conjunto de situaciones, como las leyes de potencias, la distribución normal o Gaussiana aparece en muchas más ocasiones en la práctica, debido,

entre otros motivos, al teorema central del límite (al final de este tema). Veamos ahora unos cuantos ejemplos de variables que siguen esta ley.

### 6.1.1 La mecánica de los gases de Maxwell

Fue la obra del astrónomo Quetelet (más abajo) quien puso a James Clerk Maxwell (1831-1879) en el camino de la mecánica estadística: las moléculas de un gas son como los individuos de una población, ya que el desorden a escala individual se transforma en orden a escala poblacional.

Lo que importa de los movimientos de las partículas de gas, afirmó Maxwell, son dos cosas: la velocidad media a la que se mueve cada partícula —la cual determina su energía cinética media— y cuánto se aparta de esa media por arriba o por abajo. Maxwell intuyó que la distribución de velocidades semeja el tipo de curva en forma de campana. La curva de Maxwell indica cuántas partículas de gas se mueven a una velocidad determinada y se eleva suavemente desde velocidades bajas, alcanza el pico en la media, y vuelve a descender suavemente hacia las velocidades altas. Esta distribución muestra que muy pocas partículas alcanzan velocidades mucho más altas que el promedio. Maxwell dijo que lo importante no es la trayectoria precisa de todas las partículas del gas, sino su comportamiento medio. Maxwell imaginó un enjambre de abejas: las abejas zumban furiosamente en todas direcciones, pero el propio enjambre se mantiene estacionario, porque, teniendo en cuenta el promedio de sus vuelos, no es más probable que las abejas vuelen en una dirección que en otra. (Ball, 2004)

La curva de Maxwell indica cuántas partículas de gas se mueven a una velocidad determinada y se eleva suavemente desde velocidades bajas, alcanza el pico en la media, y vuelve a descender suavemente hacia las velocidades altas. Esta distribución muestra que la mayoría de las partículas se mueven a una velocidad alrededor del promedio.

### 6.1.2 Los datos antropométricos en los seres humanos

En la película “El protegido”, Samuel L. Jackson es un hombre con osteogénesis imperfecta, al cual le apodaron de pequeño “Don cristal”. Es una persona

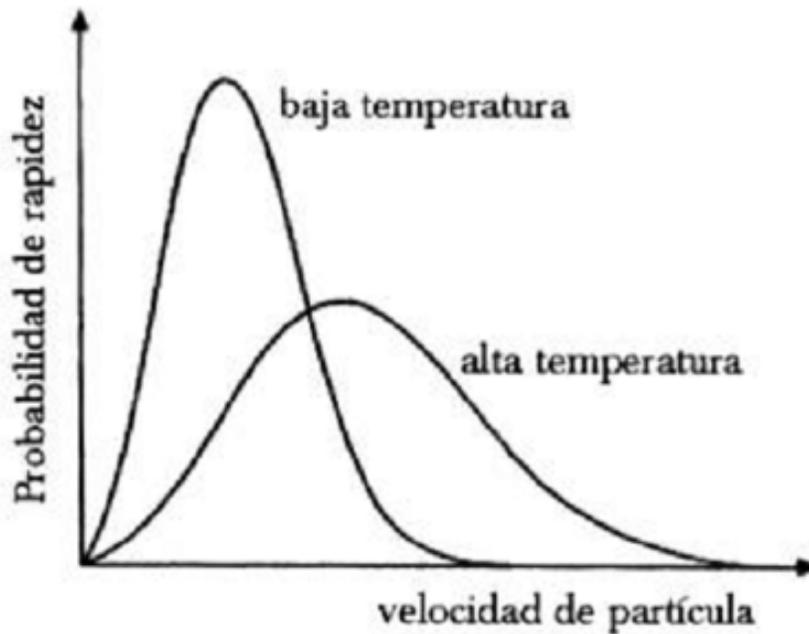


Figura 6.2: Distribución del movimiento de las moléculas de los gases.

que ha tenido múltiples fracturas, muy débil. Piensa que él se encuentra a un extremo (de un 'espectro', le llama. Sería una forma particular de llamar a una distribución), y, al otro lado de un punto medio (un eje de referencia) podría existir otra persona que nunca se lesione, que no enferme... Bruce Willis permanece callado, porque al principio de la película hemos visto como iba en un tren que ha descarrilado y matado a todos sus ocupantes, excepto a él, que no tenía un solo rasguño.

¿Podemos suponer que, si existe una persona con unas determinadas características -llamemos 'extremas'-, existirá otro con las mismas características 'opuestas'? Por ejemplo, si existe un 'genio del mal', ¿existe un 'genio del bien'? ¿Si existen personas muy altas en una población, necesariamente tienen que existir personas muy bajas? E incluso más: ¿el número de personas muy altas coincide con el número de personas muy bajas? ¿El número de personas extremadamente inteligentes coincide con el número de personas extremadamente poco inteligentes?...

El primero que respondió a este tipo de preguntas fue Adolphe Quetelet. Fue el primero en advertir que los datos antropométricos de los individuos de una

misma población, raza, sexo y edad, constituyen una variable normal.

Para verlo con un ejemplo, elegimos el fichero de datos “body\_dat.csv”, cogido de <http://ww2.amstat.org/publications/jse/v11n2/datasets.heinz.html>

El fichero consta de la medición de 25 Variables en 507 individuos (247 hombres y 260 mujeres, principalmente de entre 20 y 30 años de edad, todos haciendo ejercicio varias horas a la semana). Las mediciones fueron tomadas inicialmente for Grete Heinz y Louis J. Peterson en la Universidad Estatal de San José (USA) y en la Escuela de Postgrado Nval de Monterrey, California. Más tarde, se tomaron mediciones en centros de salud y acondicionamiento físico de California. La siguiente es la descripción del fichero por columnas:

```
# Columnas Variable
#
# Medidas esqueléticas:
#
# 1 - 4 Diámetro biacromial
# 6 - 9 Diámetro biológico, o "anchura pélvica"
# 11 - 14 Diámetro bitrocantéreo
# 16 - 19 Profundidad del pecho entre la
#         columna vertebral
#         y esternón a la altura de los pezones
# 21 - 24 Diámetro del pecho a nivel del pezón
# 26 - 29 Diámetro del codo,
#         suma de dos codos
# 31 - 34 Diámetro de muñeca,
#         suma de dos muñecas
# 36 - 39 Diámetro de rodilla,
#         suma de dos rodillas
# 41 - 44 Diámetro del tobillo,
#         suma de dos tobillos
#
# Medidas de circunferencia:
#
# 46 - 50 Circunferencia del hombro
#         sobre los músculos deltoides
# 52 - 56 Perímetro torácico, línea de los
```

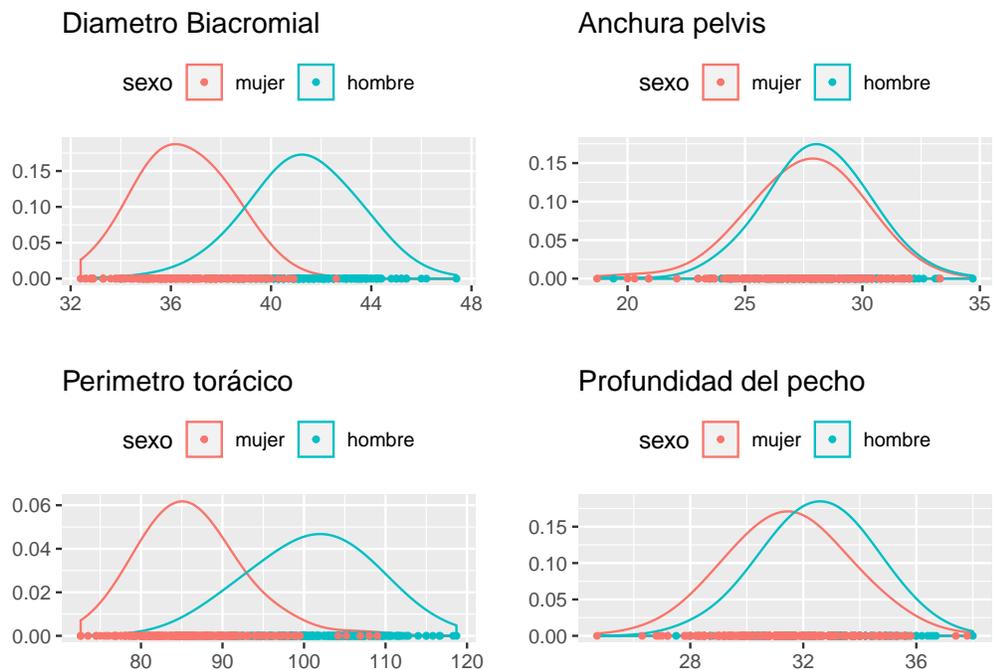
```
#           pezones en los hombros y
#           justo por encima del pecho
# 58 - 62   Cincha de la cintura, parte
#           más estrecha del torso por
#           debajo de la caja torácica
# 64 - 68   Circunferencia del ombligo
#           (o "abdominal")
#           en el ombligo y la
#           cresta ilíaca
# 70 - 74   Cincha de la cadera al nivel
#           del diámetro bitrocantéreo
# 76 - 79   Circunferencia del muslo
#           por debajo del pliegue glúteo,
#           promedio de dcha. e izda.
# 81 - 84   Cincha del bíceps, flexionada,
#           promedio de cinchas
#           derecha e izquierda
# 86 - 89   Cincha del antebrazo, extendida,
#           palma hacia arriba, promedio
#           de circunferencia derecha
#           e izquierda
# 91 - 94   Cincha de la rodilla sobre
#           la rótula, posición
#           ligeramente flexionada, promedio
# 96 - 99   circunferencia máxima de la
#           pantorrilla, promedio de
#           circunferencias derecha e izquierda
# 101 -104  Circunferencia mínima del tobillo,
#           media de cir. derecha e izda.
# 106 -109  circunferencia mínima de muñeca,
#           promedio de cir. derecha e izda.
#
# Otras medidas:
#
# 111-114   Edad (años)
# 116-120   Peso (kg)
```

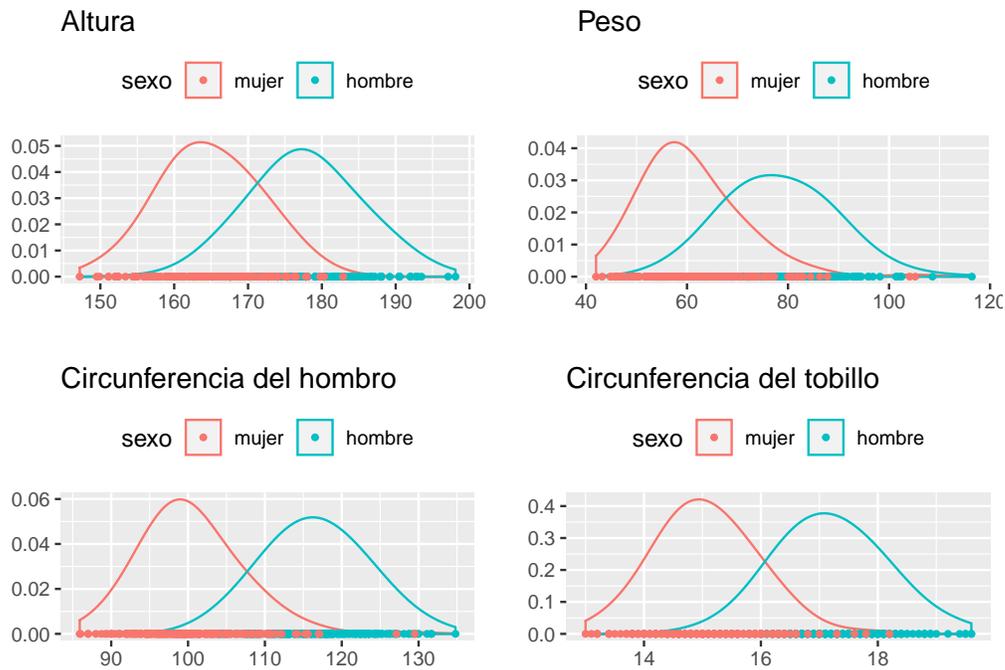
```

# 122-126      Altura (cm)
# 128 Género  (1 - hombre, 0 - mujer)
#
# Las primeras 21 variables
# se miden en centímetros (cm).
#
# Los valores están separados por
# espacios en blanco. No faltan valores.

```

A continuación, seleccionamos 8 variables y dibujamos la estimación de la densidad, diferenciándolas por sexos. Compruébese que siempre sale la campana de Gauss.





Ej.

**Ejercicio:** En España, la estatura media de los hombres mayores de 18 años es de 177.7 cm y desviación típica de 5.9, mientras que en las mujeres la media es 164.7 y desviación típica 5.4. Calcular:

Probabilidad de que tu vecino tenga un hijo que sea igual o más alto que Pau Gasol (213cm)

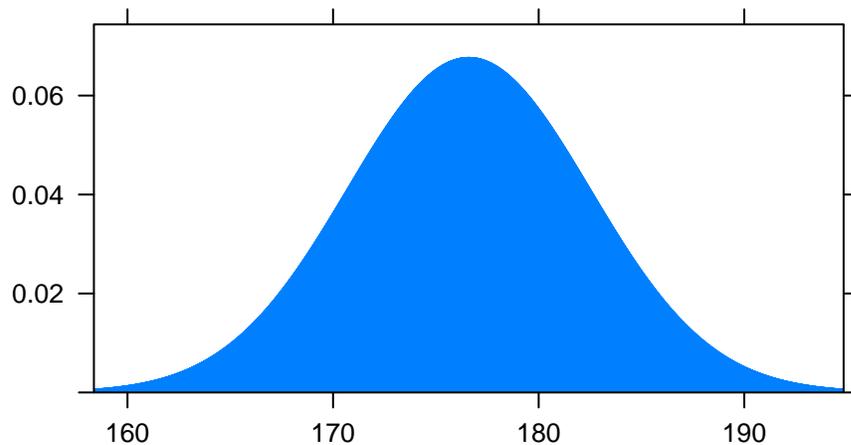
Probabilidad de que su hermana sea igual o más alta que Nicole Kidman (180cm).

Si una mujer está en el percentil 85 de estatura, ¿cuánto mide?

Un hombre mide 190 cm y una mujer 178. ¿Cuál de los dos ocuparía la mayor posición dentro de su grupo, si clasificamos las alturas de menor a mayor?

1 y 2. Dibujamos y hacemos las cuentas con R:

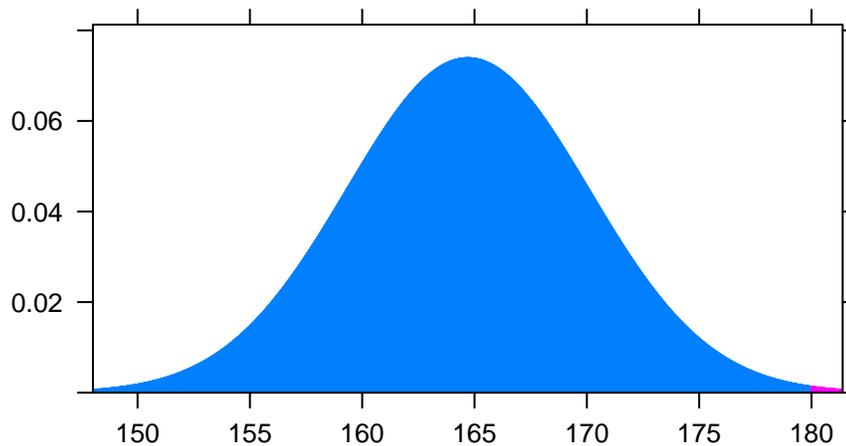
```
library(mosaic)
plotDist("norm", mean = 176.6, sd = 5.9, groups = x > 213,
         type = "h")
```



```
p1 = (1 - pnorm(200, mean = 176.6, sd = 5.9))
```

La probabilidad de ser más alto (o igual) que Pau Gasol es el área a la derecha del valor 213cm. Gráficamente, no se ve destacada área ninguna, aunque la hay. Su valor es lo que llamamos  $p1$  y que calculamos directamente con R mediante `1-pnorm(200,mean=176.6,sd=5.9)`.

```
library(mosaic)
plotDist("norm", mean = 16.7, sd = 5.4, groups = x > 10,
         type = "h")
```



```
p2 = 1 - pnorm(18, mean = 16.7, sd = 5.4)
```

La probabilidad de ser más alta que Nicole Kidman es el área (en rosa) a la derecha de 180cm, cuyo valor le llamamos  $p2$ .

1. y 2. Calculando con R, los valores son 0.00003652 y 0.0023033.

Aproximadamente, en España hay 46 millones de habitantes, de los cuales un 85 por ciento son mayores de 18 años. Supongamos que la mitad son hombres y la mitad son mujeres. Se obtiene una cantidad aproximada de 19.550.000 hombres o mujeres. Multiplicando por las probabilidades obtenidas, estimamos que el número de hombres con estatura mayor o igual a Pau Gasol es 714.13, y el número de mujeres con estatura mayor o igual que la de Nicole Kidman es 45028.8, más o menos.

3. Se calcula como

```
p8 = qnorm(0.8, mean = 16.7, sd = 5.4)
```

El valor es 170.2967.

4. Para realizar una comparación entre dos valores que pertenecen a distribuciones normales diferentes, debemos **tipificar** ambos valores.

Si un hombre mide 190 cm, su puntuación tipificada es  $\frac{190 - 177.7}{5.9} = 2.084$ ,  
y si una mujer mide 178, su puntuación tipificada es  $\frac{178 - 164.7}{5.4} = 2.4629$ .

Directamente, observamos que la mujer ocupa una mejor posición dentro del grupo “conjunto”, esto es una distribución normal con media cero y desviación típica 1.

### 6.1.3 La morfología del cerebro

En la página web [cultura colectiva](#) podemos leer lo siguiente:

“Una de las preguntas más comunes con respecto a las particularidades del cerebro radica en la diferencia que existe entre sexos. Tradicionalmente, distintos estudios científicos consideran que existen desigualdades insalvables entre los cerebros de hombres y mujeres. Tales diferencias se han utilizado a través de la historia para justificar el comportamiento de uno y otro sexo, estableciendo estereotipos entre la mentalidad masculina y femenina.

A pesar de que la ciencia médica consideraba en el pasado que no existían diferencias significativas entre ambos cerebros, el estudio más ambicioso jamás realizado para descubrir distinciones en la anatomía cerebral entre hombre y mujer arrojó resultados que podrían cambiar la forma en que comprendemos a cada sexo a partir de su pensamiento. El doctor Stuart Ritchie, especialista en Psicología de la Universidad de Edimburgo, se valió de más de 500 mil datos disponibles en el UK Biobank (una iniciativa científica del Reino Unido para la recopilación de información relativa a los órganos del cuerpo humano) para escanear cerebros y descifrar si la información arrojada a través de resonancias magnéticas era suficiente para descubrir si se trataba de un hombre o una mujer.

La amígdala, el hipocampo, el cuerpo estriado y el tálamo de los hombres son visiblemente más grandes que las mismas regiones en las mujeres. No obstante, la anatomía del sistema nervioso central femenino posee una corteza cerebral más gruesa que la masculina, un elemento que parece decisivo en la obtención de

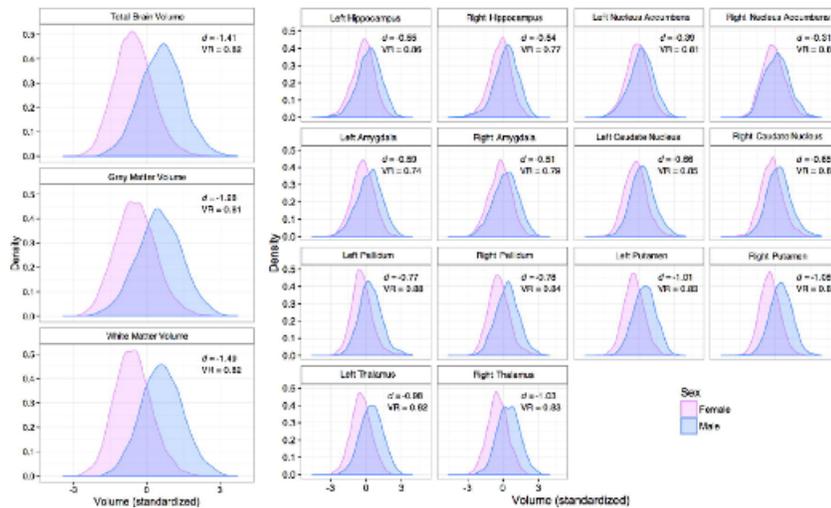


Figura 6.3: Gráfica del trabajo de Ritchie et al. (2017), donde se observa que las características antropométricas del cerebro humano también siguen la distribución normal (algo que, evidentemente, Quetelet no llegó a comprobar)

un resultado más alto en las pruebas de habilidades cognitivas e inteligencia.

El resultado fue estadísticamente notorio: en cerca del 77% de los casos fue posible identificar correctamente el sexo del cerebro en cuestión. Según el estudio (Ritchie et al., 2017), la diferencia más significativa entre el sistema nervioso central de los sexos de nuestra especie está en el tamaño: los cerebros masculinos poseen un volumen mayor que los de su contraparte femenina, no sólo en el total de materia gris, también en áreas específicas del mismo.”

#### 6.1.4 Las características psico-sociales

Adolphe Quetelet fue también el primero en descubrir que, además de los caracteres antropométricos, los caracteres psicológicos y sociales de los seres humanos tienden a distribuirse como campanas de Gauss. La propensión a la violencia, a la aficción, la empatía hacia los demás, el sentido del humor... Existe una mayoría de personas con características de este tipo en la media, disminuyendo la proporción a medida que vamos hacia los extremos.

Veamos algunos ejemplos con datos concretos procedentes de estudios científicos.

En R, en el paquete *Car*, tenemos el archivo de datos *cowles* que consta de 1421 filas y 4 columnas, con datos de interés psicológico para el voluntariado social. Las gráficas por sexo aparecen en la Figura 6.4.

```
require(car)
data(Cowles)

Cowles$sexo=Cowles$sex
N=nrow(Cowles)
y=rep(0,N)
y1<data.frame(y)
names(y1)<"ceros"
Cowles2 <bind_cols(Cowles, y1)

p1<ggplot(Cowles2)+
  geom_density(aes(x= neuroticism), adjust=2) +
  aes(colour = sexo) +
  labs(title="Inestabilidad Emocional", y=" ", x=" ") +
  theme(legend.position = "top") +
  geom_point(aes(x=neuroticism, y=ceros), size=1)

p2<ggplot(Cowles2)+
  geom_density(aes(x= extraversion), adjust=2) +
  aes(colour = sexo) +
  labs(title="Extraversion", y=" ", x=" ") +
  theme(legend.position = "top") +
  geom_point(aes(x=extraversion, y=ceros), size=1)

grid.arrange(p1,p2,ncol=2)
```

Los siguientes datos corresponden a un estudio descrito en (Kalichman and Rompa, 1995): Sexual Compulsivity Scale. En los siguientes enlaces puede leerse la descripción de las preguntas y ver los datos. [Descripcion](#) y [Datos](#). Las gráficas por sexo aparecen en la figura 6.5.

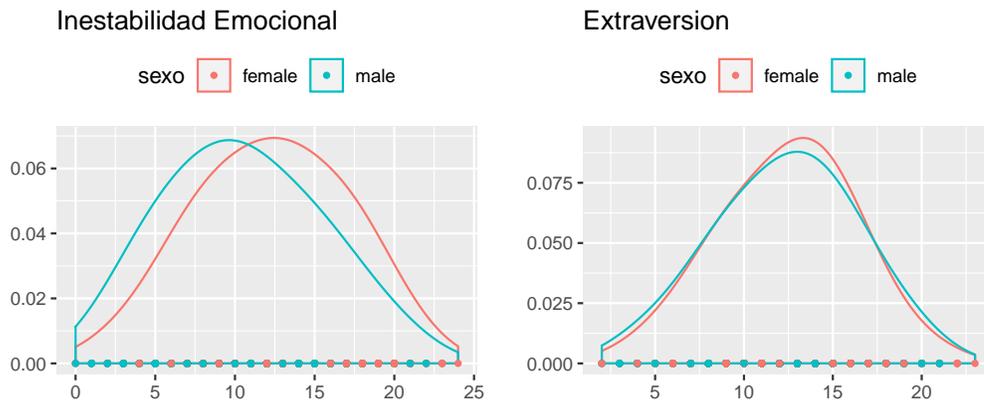


Figura 6.4: Densidad de los datos de Inestabilidad emocional y Extraversión, por sexos.

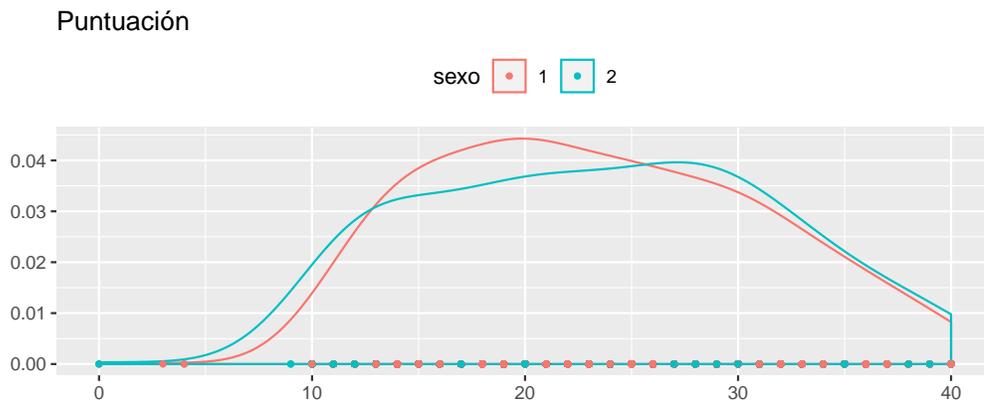


Figura 6.5: Densidad de los datos de la compulsividad sexual, por sexos.

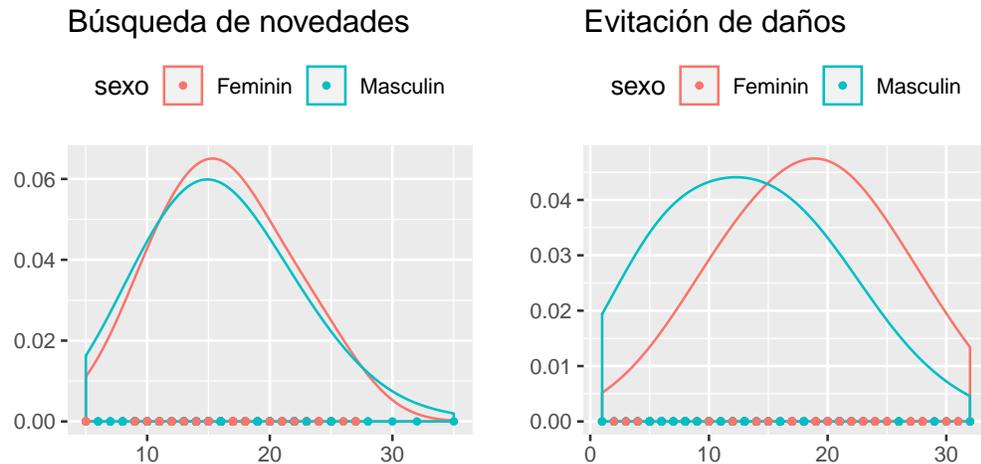


Figura 6.6: Densidad de los datos de dos variables del estudio de Hansenne (2001), por sexos.

De nuevo, campanas de Gauss para describir la distribución de los datos.

Otro ejemplo: el trabajo de (Hansenne et al., 2001), donde se detalla un cuestionario de 226 ítems del *Inventario de Temperamento y Carácter* (TCI, por sus siglas en inglés) para evaluar las siete dimensiones de la personalidad descritas por Cloninger (Cloninger et al., 1994). El cuestionario fue respondido por 322 ciudadanos Belgas. En este enlace: [Datos\\_y\\_trabajo](#) pueden descargarse el trabajo y los datos.

Análizamos 4 variables del estudio: Novelty Seeking (búsqueda de novedad) (NS), Harm Avoidance (evitación de daños) (HA), Reward dependence (búsqueda de recompensas) (RD) y Sentimentality (sentimentalismo) (RD1): gráficas 6.6 y 6.7.

De estos (y de otros muchos estudios publicados) parece concluirse también que las características psico-sociales de los seres humanos, correspondientes a una misma raza, cultura, población y sexo parecen venir descritas mediante leyes Gaussianas.

Retomemos ahora la pregunta que nos planteábamos tras ver la escena de “El protegido”: *¿es posible la existencia de personas opuestas, en cuanto a las características de "fagilidad", como expone el personaje interpretado por*

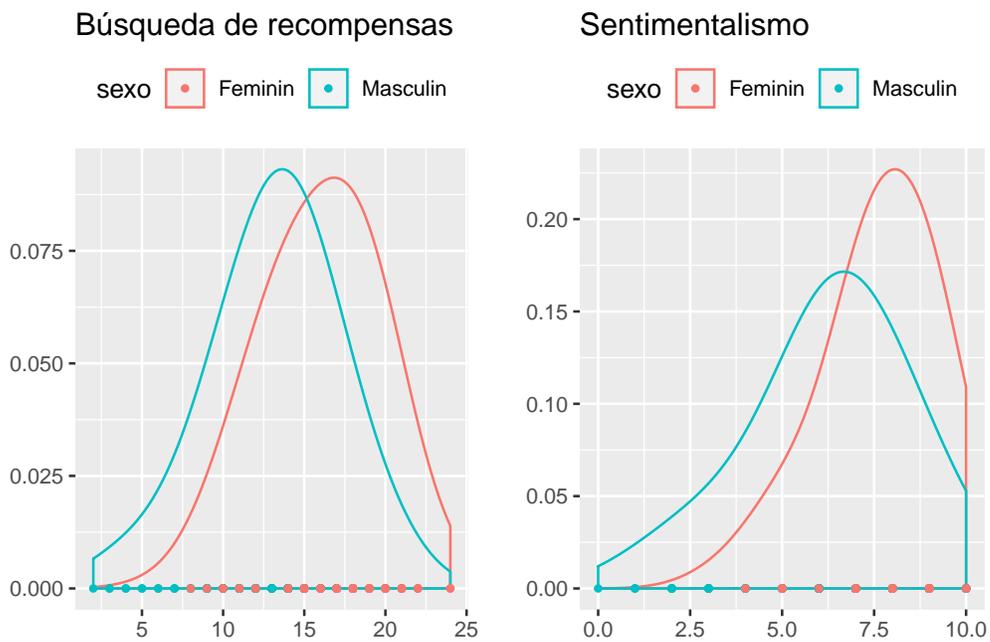


Figura 6.7: Densidad de los datos de otras dos variables del estudio de Hansenne (2001), por sexos.

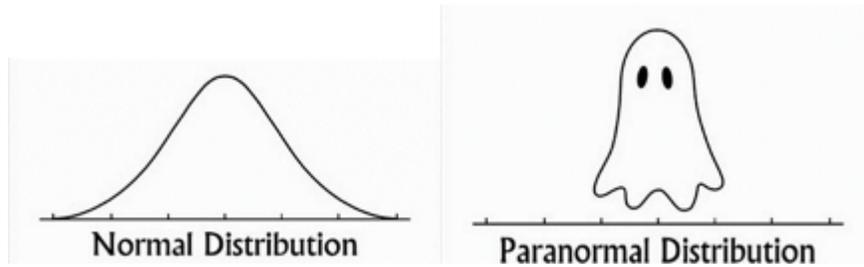


Figura 6.8:

*Samuel L. Jackson?* Él es el hombre que ‘se rompe constantemente’. ¿Existe uno que ‘prácticamente nunca se rompa’, que sería Bruce Willis?

Tenemos que darnos cuenta de que falla una hipótesis. Es la de encontrarnos en lo que, precisamente, se llama **población normal**, como sinónimo de común. El personaje de ‘Don Cristal’ no pertenece a la población más común, puesto que es una persona con una enfermedad genética, de la que tendríamos que estudiar sus características (por ejemplo, si esa población sigue o no una distribución normal). Pero, sin embargo, esta escena puede hacernos pensar en personas que a menudo conocemos. Por un lado, gente muy propensa a enfermedades, o a lesionarse, y a la vez gente que rara vez enferma, que se lesiona poco. El azar asociado a la campana de Gauss provoca que los extremos (a ambos lados) sean los más raros (los de menos probabilidad), frente al término medio, que es lo más común.

Por último, recordemos que, en el documental de la 2 que veíamos al principio de este capítulo, nos contaban que Gauss era un genio, y que pertenecía al dos por ciento de personas con cociente intelectual, digamos, “altísimo”. Los genios son aquellas personas con cociente intelectual extremo (al este, o sea a la derecha). En una distribución normal, los datos extremos, o las colas de la distribución, son los datos que están alejados  $2\sigma$  (dos desviaciones típicas) de la media. Tengamos en cuenta que, en una distribución normal cualquiera  $X \in N(\mu, \sigma)$ , la probabilidad en los extremos o colas de la distribución es  $(1 - p)/2$ , siendo  $p$

$$p = P(\mu - 2\sigma < X < \mu + 2\sigma)$$

*Recordemos las áreas en una distribución normal:*

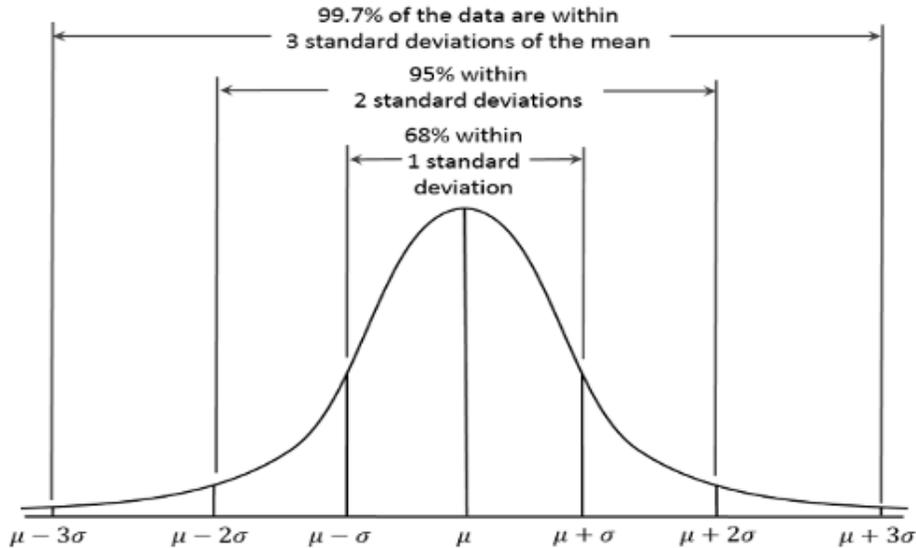


Figura 6.9: Áreas en la campana de Gauss.

Al “este” de la campana de Gauss están, como vemos, el dos por ciento: los superdotados intelectualmente, los más altos, los más destacados por alguna característica física o psico-social humana...

### 6.1.5 El consumo de petróleo, gas, electricidad, de una ciudad, un país, en un determinado periodo de tiempo

El consumo de petróleo (o gas, o electricidad, propano...) de una ciudad (ayuntamiento, urbanización, etc.) es la suma de los consumos individuales de las familias o particulares. Por ello, y debido al teorema central del límite (del que hablamos más adelante), la distribución de esta variable (consumo) va a seguir una distribución normal.

### 6.1.6 Los errores de medición

“Imaginemos, por ejemplo, que debe medir con mucha precisión la temperatura de un líquido en un recipiente. Puede utilizar un termómetro de alta precisión y tomar mil medidas a lo largo de un período de una hora. Debido a errores aleatorios y posible-

mente a fluctuaciones en la temperatura, hallará que no todas las mediciones dan exactamente el mismo valor, sino que tienden a agruparse alrededor de un valor central; algunas mediciones dan un valor superior y otras, uno inferior. Si representa el número de veces que aparece cada medida en función de la temperatura, obtendrá el mismo tipo de curva en forma de campana que Quetelet halló para las características humanas. De hecho, cuanto mayor sea el número de mediciones efectuadas de cualquier magnitud física, más se aproximará la distribución de frecuencias a la curva normal. La influencia inmediata de este hecho en la cuestión de por qué las matemáticas son tan extraordinariamente eficaces es bastante espectacular: ¡incluso los errores humanos obedecen leyes matemáticas estrictas!” (Ball, 2004)

El problema de los errores de medición era algo que preocupaba a los astrónomos desde siempre: al realizarse mediciones de, por ejemplo, distancias de la tierra a otros puntos del mapa celeste, estas no coincidían (las mediciones se realizaban desde diferentes puntos de la tierra en diferentes momentos pero, obviamente, al realizar las transformaciones matemáticas pertinentes, las mediciones deberían ser iguales, cosa que no ocurría). La teoría de los errores de medida fue iniciada por Galileo (1564-1642), que advirtió que estos errores eran simétricos, y que los errores pequeños se producen con más frecuencia que los grandes. Esta idea fue apoyada por otros muchos científicos, en su mayoría astrónomos; como, por ejemplo, Ticho Brahe (1546–1601), quien encontró que cada medida tiene un posible error. Brahe fue quien intuía que, debido a estos errores, la medida se podía precisar realizando varias observaciones y calculando la media aritmética.

R. Cotes (1682–1716), T. Simpson (1710–1761) y Daniel Bernoulli fueron los primeros en tratar matemáticamente la teoría de la medida de errores. Cotes opinaba que los errores se distribuían uniformemente, es decir, que se cometían tantos errores pequeños como grandes. Más adelante, Laplace afirmó que los errores de medida observados eran la suma de una gran cantidad de pequeños errores; si estos errores tenían una distribución normal, su suma también debería tenerla. Como estimación del valor desconocido del error, Laplace sugirió tomar el valor que minimiza la cantidad que es igual a la mediana de las observaciones realizadas. Sin embargo, su trabajo no alcanzó mucha difusión, debido a las aportaciones de Gauss y Legendre (1752–1833), que propusieron y desarrollaron el método de mínimos cuadrados. Gauss de-

mostró que, bajo ciertas condiciones generales, la función de densidad de los errores de medida tiene la forma de la distribución normal. (Salinero, 2006)

### 6.1.7 Duración de un embarazo

Se sabe que es una variable aproximadamente normal, con una media de 266 días y una desviación típica de 16 (Moore and Kirkland, 2007).

### 6.1.8 Velocidad de las Galaxias

En el trabajo de (Roeder, 1990) se puede ver la distribución que sigue el conjunto de velocidades de 82 galaxias de la Corona Boreal.

En el paquete *MASS* de R se encuentra este conjunto de datos. Se trata de la velocidad en km/seg de 82 galaxias de seis secciones cónicas bien separadas, de un estudio de la región de Corona Borealis. Como se ve, no hay una única distribución normal, sino lo que se conoce como una *mixtura* de distribuciones, donde cada una de las distribuciones que se “mezclan” aparece representada por la moda (máximo). Según explican en el trabajo de (Roeder, 1990), esa multimodalidad es evidencia de vacíos y superclusters en el universo lejano (obvio para cualquiera).

```
library(MASS)
gal <-galaxies/1000
# c(width.SJ(gal, method = "dpi"), width.SJ(gal))
plot(x = c(5, 40), y = c(0, 0.2), type = "n", bty = "l",
      xlab = "velocidad (1000km/s)", ylab = "densidad")
rug(gal)
lines(density(gal, width = 3.25, n = 200), lty = 1)
lines(density(gal, width = 2.56, n = 200), lty = 3)
```

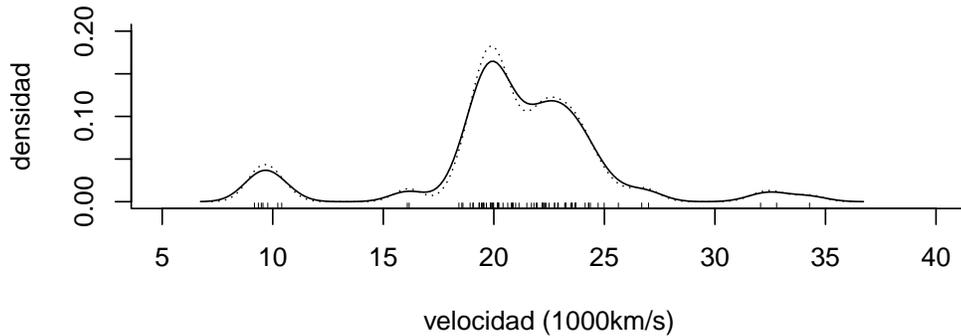


Figura 6.10: Densidad del conjunto de datos correspondiente a la velocidad de 82 galaxias.

### 6.1.9 La ley de Farr de las epidemias

(Brownlee, 1915)

William Farr, el famoso epidemiólogo inglés (1807–1883), postuló que las epidemias tienden a crecer y caer con un patrón aproximadamente simétrico, del tipo Campana de Gauss. La idea fue resucitada por Brownlee a principios del siglo XX. Un artículo famoso sobre la epidemia del sida (Bregman et al., 1990) analizó 200.000 casos desde 1982 hasta 1988, estimando que la enfermedad tendería a desaparecer hacia 1994. Sin embargo, artículos posteriores advirtieron de la no consideración del periodo de incubación de la enfermedad, lo que debería hacer construir una curva desde diez años más atrás y hasta después del año 2000.

### 6.1.10 Crecimiento de las plantas

Las reglas matemáticas que rigen el crecimiento de las plantas son similares a cómo brotan conexiones en las células cerebrales, según descubrieron científicos del Salk Institute, según puede leerse en el trabajo de (Conn et al., 2017).

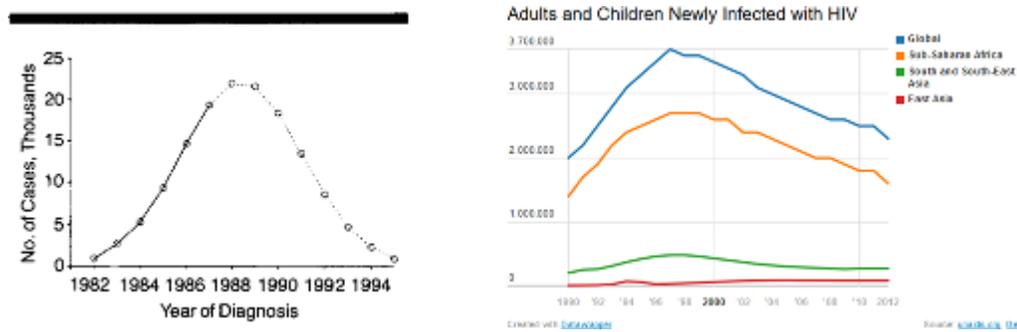


Figura 6.11: Izquierda: Figura original del artículo de Bregman. Proyección de casos hasta el 94. Derecha: datos recientes.

### 6.1.11 Votos en las elecciones: Putin contra Gauss

Gracias a la campana de Gauss (entre otros detalles), se sabe desde hace tiempo que Putin gana las elecciones haciendo trampas: la curva de distribución de la variable *mesas electorales según su participación* no sigue una curva gaussiana. Existe una amplia cantidad de artículos científicos donde se demuestran las constantes irregularidades en los procesos electorales en Rusia. Ver, por ejemplo, (Klimek et al., 2012)

Con detalle aquí

### 6.1.12 Los seis grados de separación

Esta teoría la propuso el escritor Frigyes Karinthy en 1930, y viene a decir que se puede acceder a cualquier persona del planeta en sólo seis (o menos) pasos o conexiones, o dicho de otra manera:

Cualquier persona del mundo estaría unida a nosotros a través de una cadena de conocidos de no más de cinco intermediarios o intermediarias, conectándonos con sólo seis enlaces, pasos o saltos.

Existe una película sobre este tema, llamada precisamente *Seis grados de Separación*

Esta teoría hizo famoso en los años 90 el llamado “juego de Kevin Bacon”, que se convirtió en un fenómeno popular entre los amantes del cine, y predijo la web social de conexiones en línea que hoy encontramos en Facebook o Twitter.

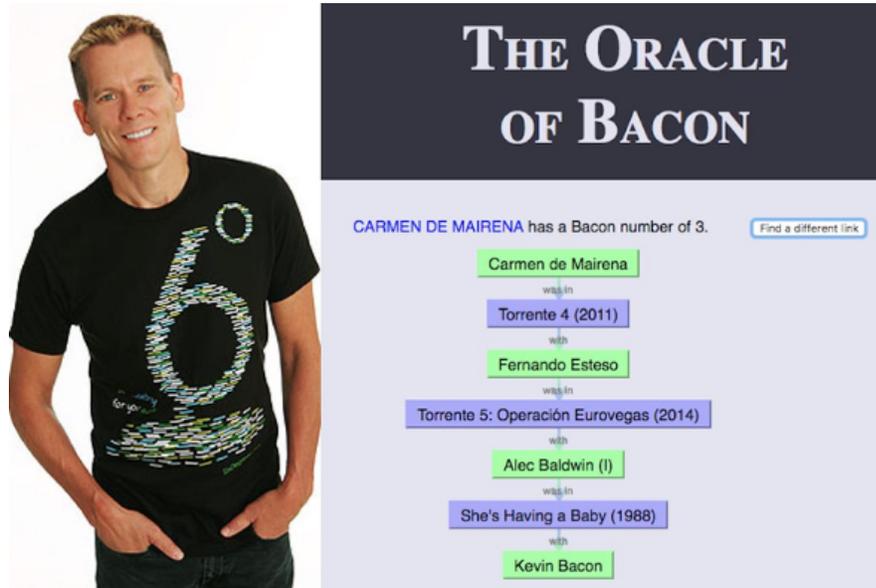


Figura 6.12: Número de Bacon de Carmen de Mairena.

El juego requiere que quienes participan en él vinculen a celebridades con Bacon, en la menor cantidad de pasos posible, a través de las películas que tienen en común. Mientras más extraña o aleatoria sea la celebridad, mejor. Por ejemplo, O.J. Simpson estuvo en *The Naked Gun 33 1/3* con Olympia Dukakis, quien estuvo en *Picture Perfect* con Kevin Bacon.

Búsquese usted mismo: [en esta página](#)

En 1967, el psicólogo norteamericano Stanley Milgram realizó el denominado experimento del ‘mundo pequeño’ para intentar demostrar la teoría.

El experimento consistió en seleccionar al azar a varias personas del Medio Oeste para que enviaran un paquete a un extraño situado en Massachussetts, a varios miles de kilómetros de distancia. Las personas remitentes sabían el nombre y la dirección aproximada, y tenían que enviar el paquete a una persona que ellos conocieran directamente y que pensasen que fuese la que más probabilidades tenía, de entre todo su círculo de amistades, de conocer directamente al destinatario. Esta persona tendría que hacer lo mismo, y así sucesivamente, hasta que el paquete fuera entregado personalmente a su destinatario final en la Costa Este.

Para sorpresa de todos, Milgram comprobó que hacía falta, en promedio, entre cinco y siete intermediarios o intermediarias para hacer llegar el paquete a esa persona desconocida. De nuevo aparece la curva normal. La media o mediana es tres, y hay tanta probabilidad de que haya un grado o dos más como un grado o dos menos.

En la reciente película [Experimenter: la historia de Stanley Milgram](#) se habla de este experimento, y muchos otros que hizo famosos el doctor Milgram, de los cuales tal vez el más conocido sea el de la capacidad de obediencia del ser humano (aunque sea para ejercer dolor contra otro ser humano).

En 2003, el sociólogo Duncan J. Watts recogió la teoría actualizada en su libro *Seis grados: la ciencia de las redes en la era conectada del acceso*. Watts intentó probar la teoría de nuevo, esta vez a través del correo electrónico con personas de todo el mundo. El resultado promedió de nuevo los famosos seis grados.

El estudio más extenso hasta la fecha sobre la teoría de los seis grados de separación lo realizó Facebook en 2011 (“Anatomy of Facebook”): el estudio se realizó con todas las personas usuarias activas de su página en esa fecha (que rondaban los 720 millones, alrededor de un 10% de la población mundial. Consistió en analizar el conjunto de amigos o amigas en común de las personas usuarias de la página, para promediar cuántos eslabones hay entre dos usuarios cualesquiera de la página. El estudio mostró que un 99’6% de pares de personas usuarias estuvieron conectados por cinco grados de separación (4.75 eslabones de promedio).

### 6.1.13 La psicofísica

La psicofísica fue introducida en 1850 por Gustav Fechner (Polonia, 1801 - Alemania, 1887), quien se preguntaba hasta qué punto una persona puede distinguir objetos de pesos ligeramente diferentes.

Posteriormente, Charles S. Peirce (filósofo, matemático y padre de la semiótica; 1839-1914) y un alumno suyo, Joseph Jastrow (1863-1944), hicieron experimentos “ciegos” (el sujeto no sabía si se le daba una caja más pesada o liviana). Es importante destacar que este fue el primer experimento en que la serie de pruebas fue determinada por un casualizador artificial. ([Peirce and Jastrow, 1884](#))

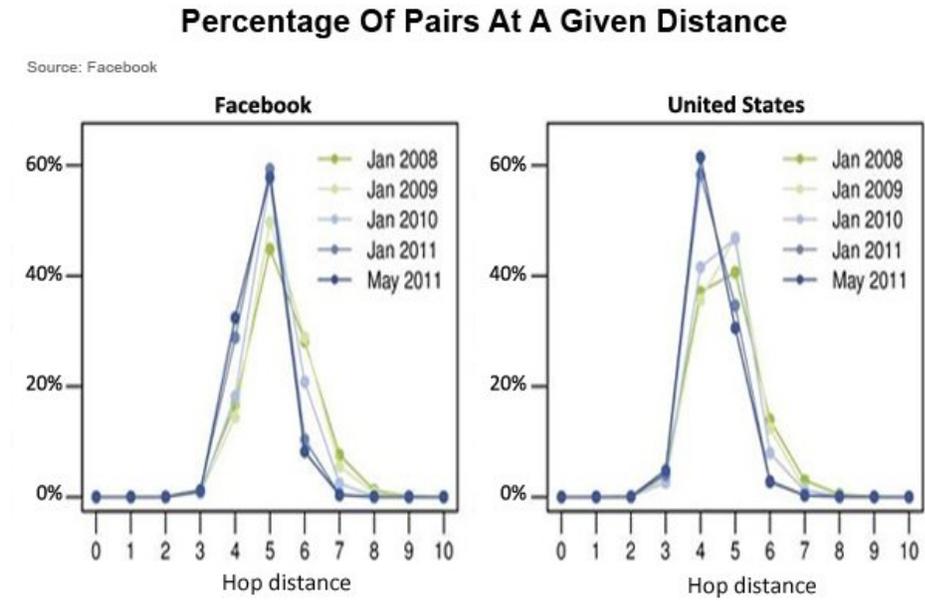


Figura 6.13: Estudio de los usuarios de Facebook.



Figura 6.14: Gustav Fechner, a quien se le atribuye la introducción del concepto de mediana.



Figura 6.15: Charles Sander Pierce.

Peirce y Jastrow se preguntaban: *¿existe, en una persona, una ley general para discriminar?*.

Sí, la curva de gauss. La variación mide la sensibilidad de un individuo. Desde el punto de vista histórico, esta comprobación significó una mayor autonomía a las leyes estadísticas. Estas son también capaces de medir una realidad psicológica de la cual ni siquiera tenemos conciencia. La sensibilidad alrededor del peso real sigue la curva de Gauss. Tanta gente se equivoca “por abajo” como “por arriba”, y menos cuanto más se aleja del peso real.

Peirce pensó que su descubrimiento de que no existe un umbral mínimo podía explicar la intuición femenina y la telepatía (tenues sensaciones de las que no tenemos plena conciencia). Posteriormente, se fundó en Londres, en 1882, la sociedad de investigación psíquica. Sus miembros deseaban reemplazar por un estudio científico el entusiasmo vulgar por la figura del medium: en lugar de suponer que había comunicación con los muertos, se suponía que podía existir transferencia de pensamiento entre personas vivas.

## 6.2 El papel de Quetelet

Adolphe Quetelet nació el 2 de febrero de 1796 en la ciudad belga de Gante. Su padre murió cuando él tenía siete años. Obligado a buscarse el sustento, Quetelet empezó a enseñar matemáticas a los diecisiete años. También compuso poesía, escribió el libreto de una ópera, fue coautor de dos obras

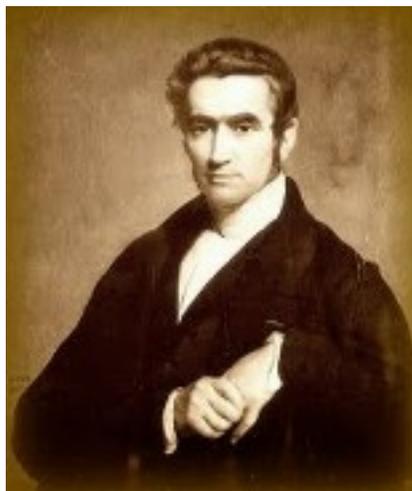


Figura 6.16: Adolphe Quetelet.

de teatro y tradujo diversas obras literarias. Sin embargo, su tema favorito siempre fueron las matemáticas, y fue la primera persona que obtuvo el grado de Doctor en Ciencias por la Universidad de Gante. En 1820, fue elegido miembro de la Real Academia de Ciencias de Bruselas.

Los años posteriores los dedicó especialmente a la enseñanza y a la publicación de diversos tratados de matemáticas, física y astronomía.

Quetelet solía empezar su curso de historia de la ciencia con la siguiente observación:

“Cuanto más avanzan las ciencias, más invaden el dominio de la matemática, que actúa como una especie de punto de convergencia. Podemos juzgar el grado de perfección al que ha llegado una ciencia por la mayor o menor facilidad con la que se le pueden aplicar cálculos”.

En diciembre de 1823, Quetelet viajó a París, enviado por el gobierno belga, con el fin de estudiar técnicas de observación en astronomía. Sin embargo, esta visita de tres meses sirvió para que Quetelet conociera a Laplace, y fijase su atención en la teoría de las probabilidades. Más adelante, Quetelet hablaría de este modo de su experiencia con la estadística y la probabilidad:

“El azar, ese misterioso vocablo del que tanto se ha abusado,

se debe considerar nada más que como un velo para nuestra ignorancia; es un espectro que domina de forma absoluta la mente común, acostumbrada a considerar los acontecimientos de un modo aislado, pero que queda reducido a nada ante el filósofo, cuyo ojo abarca largas series de eventos y cuya lucidez no se extravía en variaciones, que desaparecen cuando adquiere una perspectiva suficiente para aprehender las leyes de la naturaleza.”

En esencia, Quetelet negaba el papel del azar y lo sustituía por la idea de que incluso los fenómenos sociales poseen causas, y que las regularidades que presentan los resultados estadísticos se pueden emplear para desentrañar las reglas que subyacen al orden social.

Con la intención de probar la validez de su punto de vista estadístico, Quetelet puso en marcha un ambicioso proyecto de recopilación de miles de medidas relacionadas con el cuerpo humano. Estudió, por ejemplo, la distribución de medidas del pecho de 5.738 soldados escoceses, y de la altura de 100.000 reclutas franceses, y representó gráficamente la frecuencia de aparición de cada rasgo humano.

Luego construyó curvas similares incluso para aquellos rasgos «morales» (según él los denominaba) de los que poseía suficientes datos. Entre estas cualidades se hallaba la propensión al comportamiento criminal, los suicidios y los matrimonios. Para su sorpresa, Quetelet descubrió que todas las características humanas siguen lo que ahora se denomina una distribución normal.

Ya se tratase de alturas, pesos, longitudes de extremidades o incluso cualidades intelectuales determinadas a través de los primeros tipos de tests psicológicos, una y otra vez aparecía el mismo tipo de curva. Hasta el momento, solía ser nombrada como curva de error, porque solía aparecer en cualquier tipo de errores de medida.

Quetelet consideró que el hecho de que las características humanas siguiesen la curva de los errores era indicativo de que el **hombre medio** (*l'homme moyen*) era lo que la naturaleza estaba tratando de generar. Según Quetelet, de igual modo que los errores de fabricación crearían una distribución de longitudes alrededor de la longitud promedio (correcta) de un clavo, los errores de la naturaleza estaban distribuidos alrededor de un tipo biológico preferible. Así, afirmó que las personas de una nación estaban agrupadas alrededor de su promedio, “de igual modo que los resultados de mediciones efectuadas

sobre una misma persona, pero con instrumentos imprecisos que justifican el tamaño de la variación”.

Puesto que, evidentemente, es deseable que la sociedad “exista y se conserve”, de ello se deduce que el comportamiento promedio es el comportamiento correcto. De este modo, la física social de Quetelet se fundó en el concepto de hombre medio:

“un individuo que, en un momento dado, es el epítome de todas las cualidades del hombre promedio, representaría toda la grandeza, belleza y bondad del ser”.

Esta inquietante veneración de la uniformidad tiene su corolario en el aborrecimiento de todas las singularidades:

“Las desviaciones más o menos pronunciadas del promedio han constituido [...] la fealdad en el cuerpo como el vicio en la moral, y un estado enfermizo de la constitución general.”

La idea de que la perfección moral y física de la humanidad queda reflejada en la conformidad a unos datos matemáticos se remonta al Renacimiento, cuando, además, llegaron a definirse los instrumentos para cimentar la perfección.

Desde nuestra limitada perspectiva actual, el hecho sorprendente consiste en que prácticamente todos los detalles medibles de los seres humanos (de una raza determinada) están distribuidos según un solo tipo de función matemática.

En un ejemplo descrito en un documento (*Letters*) en 1846, Quetelet comprobó que las alturas registradas de cien mil reclutas del ejército francés no se ajustaban con precisión a la distribución normal. Según los documentos oficiales, había 28.620 hombres sobre cien mil que se encontraban por debajo de 1.57 metros. Concretamente, existían excesivos individuos en las clases (intervalos) correspondientes a las alturas entre 1.5 y 1.57, y demasiados entre 1.57 y 1.597 metros. Precisamente, 1.57 era la talla que servía para excluir del servicio militar. Con los cálculos correspondientes, se comprobó que 2275 individuos habían sido excluidos ilegalmente (habrían pagado a quien midiera para que les anotara una estatura más baja).

Pero la regularidad en la distribución de los datos no sólo se adaptaba a las características humanas. De los amplios estudios que Quetelet realizó sobre

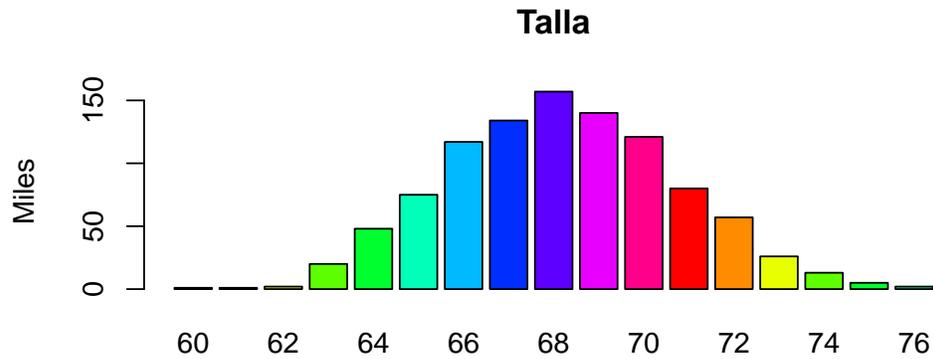


Figura 6.17: Tabla de estaturas de 26.000 soldados americanos del ejercito del Norte durante la guerra civil (\*Phys. Soc,\* i., p. 131; \*Anthropom,\* p. 259).

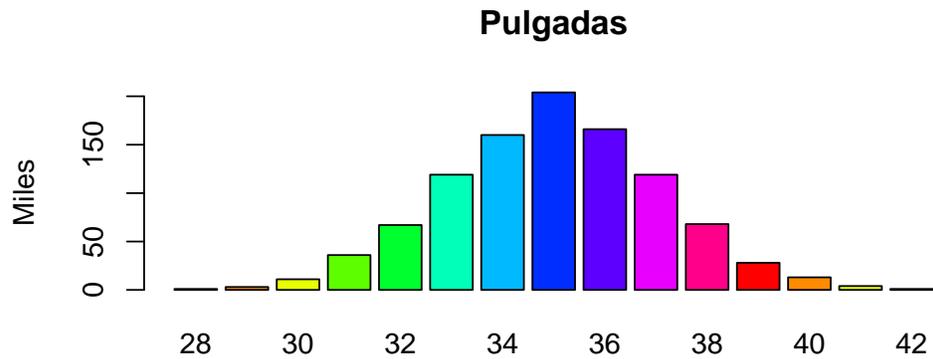


Figura 6.18: Tabla de medidas alrededor del pecho, de los soldados del ejercito del Potomac (\*Phys. Soc,\* ii., 59; \*Anthropom,\* p. 289).

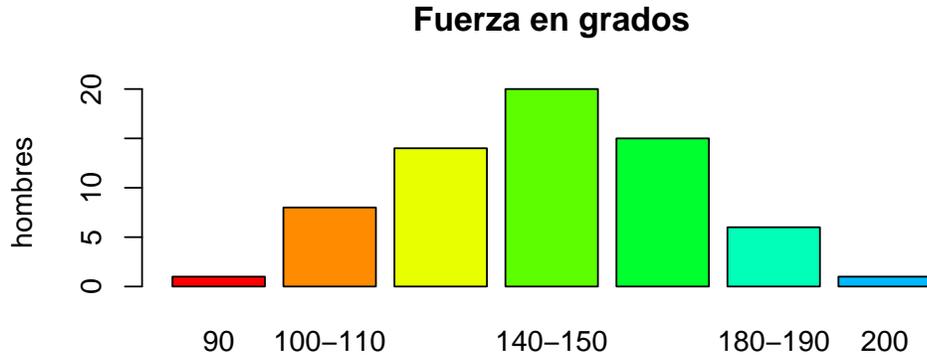


Figura 6.19: Tabla de las medidas de fuerza en hombres, estimadas mediante un dinamómetro (\*Anthropom.,\* p. 365).

las estadísticas disponibles, observó que existían leyes que se adaptaban muy bien a otro tipo de situaciones, como el número de crímenes, de personas con enfermedad mental... En (Caponi, 2013) podemos leer:

“Quetelet concluye, a partir de estudios cuantitativos y de la observación de datos empíricos, que cada año se repite, de manera exacta, el mismo número de crímenes, de suicidios, de matrimonios y de nacimientos en una población determinada. Analizando las estadísticas del ejército concluirá que los soldados presentaban, año tras año, las mismas medidas de peso, altura, tamaño de tórax. Observa que el número de alienados, internados en los asilos, se mantiene asombrosamente constante y que el tipo de crímenes y las penas aplicadas permanecen regulares según los datos suministrados por los registros civiles y los psiquiátricos. El único modo de poder explicar esas constantes que tanto seducían a Quetelet era analizar cada uno de estos hechos desde una perspectiva, ya no individual, sino poblacional. No se trataba de entender por qué razón ocurría cierto tipo de crimen y no otro, o por qué motivo un individuo presentaba una altura o peso determinados. Se trataba de explicar la repetición de fenómenos

que, por su constancia, parecían indicar alguna fuerza común a todos ellos, una causa común capaz de mantener ese equilibrio, la acción de leyes tan regulares como las que rigen los astros o la caída de los cuerpos.”

Ej.

Tenemos disponible, entre los datos de este libro, la Tabla de muertos en Francia desde 1827 a 1831: (Ficheros *CrímenesFrancia@quetelet.xlsx* y *duelos-quetelet.xlsx*).

```
## # A tibble: 12 x 7
##   Muerte Por`186187`188189`180`
##   <chr>          <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Crímenes en~    241   234   227   231   205
## 2 pistola         56    6    6    6    57
## 3 espada, sab~   15     7     8     7    12
## 4 cuchillo       33    40    34    46    44
## 5 bastones       23    28    31    24    12
## 6 piedras        20    20    21    21    11
## 7 instrumento~  35    40    42    45    46
## 8 estrangulam~   2     5     2     2     2
## 9 precipitaci~   6    16     6     1     4
## 10 patadas y g~  28    12    21    23    17
## 11 fuego         0     1     0     1     0
## 12 desconocidas  17     1     2     0     2
## # ... with 1 more variable: 181`<dbl>

## # A tibble: 5 x 5
##   año muertes accide~ suicidios duelos con mue~
##   <dbl>          <dbl>    <dbl>          <dbl>
## 1 187            4744    1542            19
## 2 188            485     1754            29
## 3 189            5048    1904            13
## 4 80             4478    1756            20
## 5 181            5045    208             23
## # ... with 1 more variable: duelos sin muerte`<dbl>
```

“Hoy en día, la idea de los promedios se da por sentada. Forman parte del zumbido de los medios de comunicación diarios. Mien-

\* The following is the result of the reports of criminal justice in France, &c. :—

	1826.	1827.	1828.	1829.	1830.	1831.
Murders in general, -	241	234	227	231	205	206
Gun and pistol, - -	56	64	60	61	57	68
Sabre, sword, stiletto, poniard, dagger, &c.,	15	7	8	7	12	30
Knife, - - - - -	39	40	34	46	44	34
Cudgels, cane, &c., -	23	28	31	24	12	21
Stones, - - - - -	20	20	21	21	11	9
Cutting, stabbing, and bruising instruments,	35	40	42	45	46	49
Strangulations, - - -	2	5	2	2	2	4
By precipitating and drowning, - - - -	6	16	6	1	4	3
Kicks and blows with the fist, - - - - -	28	12	21	23	17	26
Fire, - - - - -	..	1	..	1	..	..
Unknown, - - - - -	17	1	2	..	2	2

Figura 6.20: Tabla original del trabajo de Quetelet.

\* The following is the result of the reports of criminal justice in France, &c. :—

	1826.	1827.	1828.	1829.	1830.	1831.
Murders in general, -	241	234	227	231	205	206
Gun and pistol, - -	56	64	60	61	57	68
Sabre, sword, stiletto, poniard, dagger, &c.,	15	7	8	7	12	30
Knife, - - - - -	39	40	34	36	44	34
Cudgels, cane, &c., -	23	28	31	24	12	21
Stones, - - - - -	20	20	21	21	11	9
Cutting, stabbing, and bruising instruments,	35	40	42	45	46	49
Strangulations, - - -	2	5	2	2	2	4
By precipitating and drowning, - - - -	6	16	6	1	4	3
Kicks and blows with the fist, - - - - -	23	12	21	23	17	26
Fire, - - - - -	..	1	..	1	..	..
Unknown, - - - - -	17	1	2	..	2	9

Figura 6.21: Tabla original del trabajo de Quetelet.

Years.	Accidental Deaths.	Suicides.	Duels.	
			followed by Death.	not followed by Death.
1827, - -	4,744	1542	19	51
1828, - -	4,855	1754	29	57
1829, - -	5,048	1904	13	40
1830, - -	4,478	1756	20	21
1831, - -	5,045	2084	25	36
Total, -	24,170	9040	106	205

Figura 6.22: Tabla original del trabajo de Quetelet.

**2. Of Suicides and Duels.**

The destruction of man by his own hands, although generally repugnant to the notions of modern society, has nevertheless found panegyrista, and those who have proclaimed its advantages. Suicide, among some nations, continues to be branded with infamy by the public. The ancients were not entirely of this opinion: it was often practised by the most illustrious men, and has been mentioned with admiration by their gravest historians. We are naturally excited by the death of Cato, who wished not to survive the liberty of his country; by the death of Lucretia, who wished not to survive her dishonour; or even by the death of the criminal, who seeks to spare his family the shame of seeing his head fall on the scaffold.

The destruction of one man by another, excites horror; yet this dreadful crime may also, in our manners and modern institutions, present the appearance of a virtue under certain circumstances. We can only comprehend these apparent contradictions, by admitting that the crime consists, not in the action, but in the intention of him who commits it; so that, if the intention was noble or generous, the action may also be considered of the same character. This is the

number of duels, it may be supposed that the values in the table are too low.

A very great number of suicides takes place in the department of the Seine. They have been committed in the following manner, during the years from 1817 to 1825 inclusive:—

Years.	Total.	Submerston.	Fire-arms.	Asphyxia.	Voluntary Fall.	Strangulation.	Cutting Instruments.	Poisoning.
1817, - -	322	50	46	35	39	36	21	15
1818, - -	370	131	45	35	45	37	27	20
1819, - -	378	149	59	46	30	44	39	20
1820, - -	325	129	46	30	37	33	28	14
1821, - -	348	127	60	42	33	36	25	23
1822, - -	307	130	48	45	33	21	31	15
1823, - -	360	114	35	61	43	48	47	21
1824, - -	371	115	43	61	47	36	40	23
1825, - -	336	134	36	56	40	40	30	20
Total, -	3265	1178	401	437	323	234	260	175

The average number of suicides, therefore, in the department of the Seine, annually reaches 356: which,

Figura 6.23: Tabla original del trabajo de Quetelet. Observemos que los números, año a año, son casi similares, lo cual provocó el asombro de Quetelet (y de cualquiera).

tras escribo esto, el New York Times del día reporta la cantidad promedio de deuda estudiantil, el número promedio de televidentes de la televisión en horario de máxima audiencia y el salario promedio de los médicos. Pero cada vez que Quetelet revelaba un nuevo promedio, el público se asombraba. Por ejemplo, Quetelet mostró que la tasa promedio de suicidio era relativamente estable de año en año. Aunque esto no sería una noticia sorprendente en estos días, en la década de 1830 el suicidio fue visto como una decisión privada altamente irracional que no podía ajustarse a un patrón más profundo. En cambio, Quetelet mostró que los suicidios ocurrían con regularidad confiable y consistente. Y no sólo eso: Afirmó que la estabilidad de los hechos indicaba que todo el mundo tiene una propensión promedio al suicidio.”

— Todd Rose, 2016. [The Atlantic](#)

Académicos y pensadores en todos los campos aclamaron a Quetelet como un genio para descubrir las leyes ocultas que gobiernan la sociedad. **Florence Nightingale** adoptó sus ideas en enfermería, declarando que el hombre promedio encarnaba la “Voluntad de Dios”. **Karl Marx** se basó también en las ideas de Quetelet, afirmando que “el hombre común” demostraba la existencia del determinismo histórico. El físico **James Maxwell** se inspiró en las matemáticas de Quetelet para formular la teoría clásica de la mecánica del gas. El médico **John Snow** utilizó las ideas de Quetelet para luchar contra el cólera en Londres, marcando el inicio del campo de la salud pública. **Wilhelm Wundt**, padre de la psicología experimental, leyó a Quetelet y proclamó: “Se puede afirmar sin exagerar que de los promedios estadísticos se puede aprender más psicología que de todos los filósofos, excepto de Aristóteles”.

Según Quetelet, el hombre medio está impuesto de tal forma que no es posible realizar modificaciones. No solo las características fisiológicas como la cantidad de respiraciones o de pulsaciones por minuto, la altura o la fuerza que podemos alcanzar se definen en relación a la constancia representada por el hombre medio, sino que también hechos sociales tan variados como la cantidad de crímenes, matrimonios, casos de alienación mental, suicidios, ya están preestablecidos, de modo tal que escapa de nuestras manos cualquier modificación.

“Podemos enumerar anticipadamente cuántos individuos man-

charán sus manos con la sangre de sus semejantes, cuántos serán falsificadores, cuántos envenenadores, con tanta precisión como podemos enumerar la cantidad de nacimientos y muertes que ocurrirán en una sociedad. La sociedad contiene en ella los gérmenes de todos los crímenes que se cometerán, al mismo tiempo que las condiciones para que ellos ocurran. Es ella quien prepara sus crímenes, y el culpable no es más que el instrumento que los ejecuta.

— Quetelet, 1848, p.315.

De esta forma, el ‘libre albedrío’ no es más que un obstáculo para la comprensión de las leyes que rigen las constantes biológicas y sociales. Quetelet no creía que nuestros actos estuviesen predeterminados, sino que, a pesar de la libertad del individuo, las acciones de unos y otros se mezclarían y se verían compensadas, encadenándose una permanencia de los hechos sociales año tras año. De nada serviría que nos neguemos a provocar un crimen; este acto de libertad será compensado con la acción contraria, posibilitando que se realice la natural tendencia a la constancia y a la repetición.

### 6.3 Para pensar un poco: El CI

La inteligencia de una persona se compone de múltiples factores (razonamiento aritmético, manejo del lenguaje, memoria, capacidad de anticipación...). En realidad, no existe una manera “real” de medir la inteligencia de nadie. Lo único que se puede hacer es “estimarla” de alguna forma, de manera que siempre podrá existir alguna imprecisión en la medida que se dé sobre la inteligencia de alguien.

Lo que se conoce como cociente de inteligencia (muchas veces mal llamado coeficiente) es el resultado de un test que mide y califica los diferentes factores que forman la misma. El primer test fue inventado en 1905 por Alfred Binet, psicólogo francés, con la intención de identificar a escolares que necesitasen atención especial.

Antes que Binet, Francis Galton diseñó una serie de cuestionarios para medir los rasgos y las características de grupos de población que consideraba relevantes, viendo que las personas de mejor posición social y económica tendían a dar mayores signos de inteligencia que el resto. Estos estudios también le

permitieron ver que la inteligencia, al igual que las características físicas, se comporta estadísticamente mediante una distribución normal: la gran mayoría de las personas tiene un nivel de inteligencia muy cercano a la media, mientras que las personas con valores extremos (por su inteligencia muy baja o muy alta) son siempre claras minorías.

Al ver que la estadística podría ser muy útil para conocer las características mentales de nuestra especie y el modo en el que se expresan las diferencias individuales en ella, Galton decidió utilizarla para comprobar la validez de sus hipótesis sobre la inteligencia. Había llegado a la conclusión de que las personas más inteligentes eran una minoría y que esta coincidía con la minoría más acomodada, pero... ¿era esto un signo de que la educación cara favorecía el desarrollo de grandes intelectos, o es que la herencia biológica de las familias ricas tendía a generar individuos inteligentes?

Para responder a la pregunta anterior, Galton decidió buscar casos en los que se pudiese descartar la influencia de la herencia innata, lo cual permitiría ver los efectos del aprendizaje. Para ello recurrió al estudio de gemelos monocigóticos. Estudiando las diferencias en cuanto a características mentales de estos gemelos a lo largo de varios años, observó algo curioso: podían ser muy diferentes o muy semejantes, pero este patrón raramente cambiaba con el tiempo. Es decir, los gemelos que eran muy parecidos al nacer seguían pareciéndose mucho años más tarde, y los que eran muy distintos desde sus primeros años seguían siéndolo en etapas posteriores.

Este descubrimiento hizo que Francis Galton, aún reconociendo la influencia del aprendizaje y del entorno sobre el individuo, terminase por darle más importancia a lo innato y a la herencia recibida por los padres y madres: a fin de cuentas, los efectos de un ambiente que cambia constantemente no parecía ser muy significativo en los rasgos psicológicos de los gemelos, que se mantenían más o menos iguales con el paso del tiempo.

Posteriormente al desarrollo del test de Binet en 1905, el psicólogo Lewis M. Terman (1877-1956) realizaría una revisión de la escala de aquel, que recibiría el nombre de escala Stanford-Binet. En dicha escala incluiría la medición del Cociente de Inteligencia desarrollado por William Stern (1871-1938), multiplicándolo por cien con el fin de eliminar fracciones. Crearía pues el Cociente Intelectual que se conoce hoy en día, permitiendo una medición más precisa del nivel de inteligencia.

Uno de los interrogantes que más controversia ha causado a lo largo del último siglo es si la inteligencia es mayoritariamente heredada. De ser así, este planteamiento serviría de base a los partidarios de la eugenesia (como Galton) para el desarrollo de medidas activas que favoreciesen a los más privilegiados (en este caso por su capacidad intelectual).

Curiosamente, uno de los factores más influyentes en esta controversia constituye también una de los más audaces historias de falsificación científica de la ciencia reciente. Y cuyo protagonista fue ni más ni menos que **Cyril Burt** (1883-1971), que había sido médico de cabecera de Francis Galton. Galton creó por primera vez en Inglaterra una cátedra de psicología de la que su discípulo se convirtió en titular en 1907. Su mayor tarea fue continuar el camino trazado por Galton y Charles Spearman, el primer gran psicólogo inglés y creador del análisis factorial. Esta historia puede leerse con más detalle en el libro de Federico di Trocchio ([Trocchio, 1977](#)).

En sus numerosas publicaciones, Burt utilizó ampliamente las pruebas de inteligencia **con el objeto de demostrar sus hipótesis de que la inteligencia está determinada por factores hereditarios**. Los estudios más importantes fueron los de gemelos idénticos separados; es decir, los denominados “gemelos verdaderos”, que por diferentes razones habían sido educados por familias distintas. El motivo del interés de Burt por estos individuos era evidente. Los gemelos idénticos poseían idéntico patrimonio genético. Si hubiera podido demostrar que, aunque crecieran en familias diferentes, mantenían el mismo cociente de inteligencia, habría tenido entre manos la prueba más evidente del carácter hereditario de la inteligencia. Habría demostrado que las costumbres y las capacidades adquiridas en el ambiente cultural en el que crecemos no mejoran los potenciales intelectuales innatos.

Para demostrar matemáticamente su teoría, Burt trabajó con el coeficiente de correlación de Pearson. En el caso de los gemelos idénticos, el razonamiento de Burt fue el siguiente: este tipo de gemelos posee el mismo patrimonio hereditario. Si los sometemos a algunas pruebas de inteligencia, evaluando en forma numérica sus respuestas y comparando luego las puntuaciones de uno y otro, el coeficiente de correlación entre las dos medidas nos permitirá evaluar la influencia de la herencia en la inteligencia. En particular, si las dos medidas resultan correladas positivamente, el índice de correlación obtenido puede considerarse una medida exacta de la incidencia efectiva de la herencia en la inteligencia.

Burt obtuvo una **correlación positiva** de  $0.771$  para los gemelos idénticos criados en familias separadas, y una aún más elevada de  $0.944$  para los gemelos criados por la misma familia. Creía haber demostrado de forma rigurosamente matemática que la inteligencia es una cualidad que se hereda de los padres y no puede adquirirse, sino en una mínima parte, a través de la educación.

En aquella época nadie se atrevió a discutir los fundamentos de los estudios y la exactitud de las conclusiones de Burt, pero muchos años después se reveló que en aquellos investigaciones podían plantearse varios interrogantes:

- Ante todo, los gemelos eran demasiados. Los primeros datos presentados por Burt en 1955 se referían a 21 pares de gemelos idénticos. En 1958 los pares estudiados ascendieron a 30, y en su último artículo de 1966 se convirtieron en 53.
- Los gemelos idénticos son, de hecho, poco frecuentes, y lo son mucho menos los criados separadamente.
- A pesar de que el número de gemelos aumentaba con el tiempo, los coeficientes de correlación seguían siendo los mismos, es decir  $0.771$  y  $0.944$ . Desde el punto de vista estadístico esto parecía imposible, dado que, a medida que aumentaba el número de gemelos, esos coeficientes deberían variar, por lo menos, en alguna cifra decimal. Sin embargo, durante muchos años, nadie dio importancia a esta anomalía.

Mientras tanto, las ideas de Burt se afirmaron en el mundo científico e influyeron en el sistema educativo, tanto en Inglaterra como en Estados Unidos. El gobierno inglés, por ejemplo, adoptó inmediatamente después de la Segunda Guerra Mundial una prueba elaborada sobre las bases de las ideas de Burt, a la que se sometían todos los niños ingleses a los 11 años, a fin de establecer qué tipo de educación era la más adecuada para ellos, si la superior o la inferior. El examen fue abolido en 1969 aunque las críticas habían comenzado a surgir a partir de 1950, el año en que Burt, que entonces tenía 68 años, fue nombrado profesor honorario y se jubiló.

El efecto más sorprendente de las ideas de Burt se presentó en septiembre de 1971 cuando Richard Herrnstein, profesor de Harvard, publicó un artículo en el que sostenía que la clase social a la que pertenece un individuo está determinada en gran parte por las diferencias hereditarias del cociente de inteligencia. Es decir, que los pobres son tal porque son hijos de pobres y

estúpidos y, recíprocamente, los ricos son así porque son hijos de padres que su mayoría son ricos e inteligentes. Posteriormente, Hernstein publicó un libro con otro autor, llamado “The Bell Curve” (Murray and Herrnstein, 1994) que redundaría más en la polémica y del que hablaremos un poco más abajo.

En 1954, cuando se le solicitaron a Cyril Burt los datos de sus experimentos, respondió que un compañero suyo iba a publicarlos, pero nunca los dio a conocer. En octubre de 1976, un periodista del Sunday Times descubrió que otras dos presuntas colaboradoras de Burt, Margareth Howard y Jane Conway, resultaban por completo desconocidas en la Universidad de Londres. Howard y Conway aparecían además como autoras de numerosos artículos y reseñas publicadas en el *Journal of Statistical Psychology* que, casualmente, estaba dirigido por el mismo Burt. Todas eran reseñas muy favorables a las aportaciones de Burt, reivindicando sus prioridades científicas e incluyendo duros ataques contra aquellos que no compartían sus ideas.

Finalmente, en 1979, Leslie Hearnshaw, un psicólogo que ocupaba en Liverpool la cátedra que había sido de Burt, fue contratado por la hermana de este para escribir una biografía. Hearnshaw encontró datos relativos a otros 15 pares de gemelos y comprobó que los datos de los otros 38 eran totalmente inventados. También verifico que Burt había escrito con nombres falsos un total de más de 20 cartas y reseñas con el fin de poder citar una y otra vez sus trabajos. De la biografía escrita por Hearnshaw surgió el retrato de un hombre muy inteligente pero con graves problemas de carácter.

En su libro, Federico diTroccio comenta que Burt ha sido el más afortunado de los investigadores acusados y encontrados culpables de fraude científico, ya que, gracias a algunos otros científicos que trataron de apoyar sus teorías y trabajos, sus investigaciones aun son consideradas por buenas en muchos casos. Por ejemplo, el psicólogo Robert B. Joyson y el sociólogo Ronald Fletcher publicaron sendos libros en los que intentaron rehabilitar a Burt. El más importante es *The Burt affair* (Joyson, 1989). El objetivo fue refutar las acusaciones de Hearnshaw y demostrar que las teorías de Burt se apoyaban en investigaciones efectivamente realizadas, no en datos falsificados.

A día de hoy, no se conoce explícitamente cuánta parte de la inteligencia es heredada y cuánta adquirida. Quien quiera leer algo más sobre el asunto (y sobre estudios en gemelos) puede leer este artículo de [Xataka](#).



Figura 6.24: Cyril Burt, ¿uno de los mayores troleros de la historia?

## 6.4 Reproductividad de la variable normal

Igual que sucede con otras variables, como la de Bernoulli, la binomial o la de Poisson, la variable aleatoria normal es reproductiva; esto quiere decir que, si sumamos variables aleatorias normales, la variable suma también es otra variable normal.

Supongamos que juntamos a 5 personas. La altura, por ejemplo, sigue una distribución normal, según hemos visto anteriormente. Sin embargo, si de esas 5 personas 2 son hombres y 3 mujeres, de entrada ya tendremos que los parámetros media y desviación típica de hombres y mujeres diferirán. Si, además, los 2 hombres son de raza diferente, también tendrán parámetros diferentes entre sí.

La altura suma de las alturas de las cinco personas, a pesar de que tengan parámetros diferentes, será también una variable normal. De manera que, para un grupo de 5 personas determinado cumpliendo esas características (2 hombres de distinta raza, 3 mujeres de la misma), la altura total será un valor, llamémosle  $y_1$ .

Si juntamos otro grupo de 5 personas con esas mismas características, la

altura suma será otro valor  $y_2$ .

Si realizamos esto (juntar grupos de 5 personas verificando las mismas características)  $n$  veces, tendremos  $n$  valores  $y_1, y_2, \dots, y_n$ .

La reproductividad de la variable normal nos asegura que la densidad de estos valores (o un histograma) tendrá el aspecto de la curva de Gauss. La forma de saber los parámetros concretos (media y desviación típica) de la variable suma nos la da el siguiente resultado teórico:

La suma de variables aleatorias normales independientes es otra variable aleatoria normal, con media la suma de las medias y varianza la suma de varianzas.

Sean  $X_i \in N(\mu_i, \sigma_i)$ ,  $i = 1, \dots, n$ . Se verifica que la suma

$$Y = X_1 + X_2 + \dots + X_n \in N\left(\mu_1 + \mu_2 + \dots + \mu_n, \sqrt{\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2}\right).$$

Vamos a realizar una simulación con R de lo que acabamos de comentar. Supongamos que juntamos 2 hombres (un español y un ciudadano de Timor Oriental) (mayores de 18 años) y 3 mujeres españolas, también mayores de 18 años. La altura de los hombres españoles mayores de 18 años sigue una distribución normal de media 177.7 y desviación típica 5.9. Las mujeres siguen una distribución normal de media 164.7 y desviación típica 5.4. La altura de los hombres de Timor Oriental es normal de media 159.8cm, y vamos a suponer que su desviación típica es 4.6.

```
x1=rnorm(1, mean=177.7, sd=5.9)
x2=rnorm(1, mean=159.8, sd=4.6)
x3=rnorm(1, mean=164.7, sd=5.4)
x4=rnorm(1, mean=164.7, sd=5.4)
x5=rnorm(1, mean=164.7, sd=5.4)
y=x1+x2+x3+x4+x5
x1;x2;x3;x4;x5;y
```

```
## [1] 1783
```

```
## [1] 153.3
```

```
## [1] 172.1
```

```
## [1] 173.1
```

```
## [1] 16.7
```

```
## [1] 80.4
```

Con esto hemos simulado la medición de las estaturas de esas 5 personas, y hemos sumado las mismas. Nos da 840.4248.

Si realizamos este proceso, por ejemplo, 1000 veces, los 1000 valores que obtengamos de  $y$  deberán dibujar la densidad de una variable normal, con media la suma de las medias, es decir  $157.7 + 159.8 + 3 * 164.7 = 811.6$  y desviación típica la raíz cuadrada de las varianzas, es decir

$$\sigma = \sqrt{5.9^2 + 4.6^2 + 3 * 5.4^2} = 11.97.$$

Veamos como se puede simular este proceso en R, y la densidad de la suma ( Figura 6.25).

```
y<-c()
for (i in 1:1000) {
x1=rnorm(1, mean=177.7, sd=5.9)
x2=rnorm(1, mean=159.8, sd=4.6)
x3=rnorm(1, mean=16.7, sd=5.4)
x4=rnorm(1, mean=16.7, sd=5.4)
x5=rnorm(1, mean=16.7, sd=5.4)
y[i]=x1+x2+x3+x4+x5
}
mean(y)
```

```
## [1] 81
```

```
sd(y)
```

```
## [1] 11.6
```

```
plot(density(y))
```

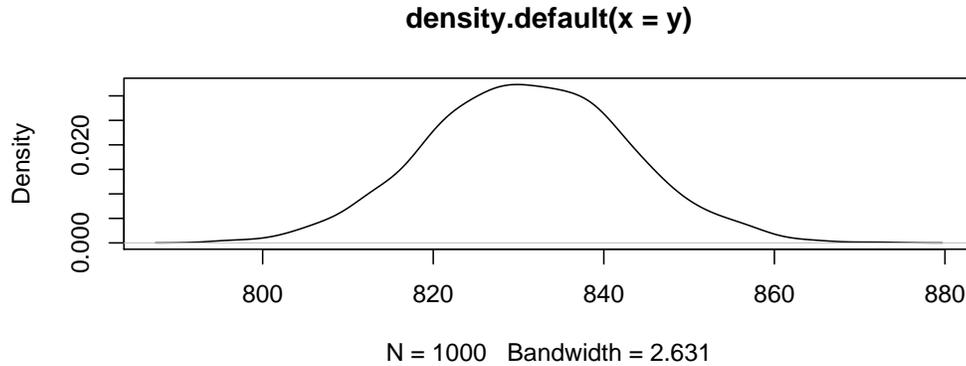


Figura 6.25: Densidad de la suma de cinco variables aleatorias.

Comprobamos que la media de los 1000 valores  $y$  obtenidos es un número muy próximo a 811.6, y la desviación típica también se aproxima mucho a 11.97. Además, vemos que la estimación de la densidad es una campana de Gauss.

De esta forma, hemos simulado el proceso de medir 1000 grupos de personas con las características citadas, cada vez sumando las estaturas. Obtenemos 1000 sumas, que, por ser cada una de ellas una suma de números procedentes de variables normales, el número resultante también pertenece a una distribución normal.

Este resultado todavía se puede generalizar más, si, en vez de considerar una simple suma, se considera una combinación lineal de las variables de partida  $X_i$ .

Sean  $X_i \in N(\mu_i, \sigma_i)$ ,  $c_i$  cualesquiera números reales,  $i = 1, \dots, n$ . Se verifica:

$$Y = \sum_{i=1}^n c_i X_i \in N \left( \sum_{i=1}^n c_i \mu_i, \sqrt{\sum_{i=1}^n c_i^2 \sigma_i^2} \right).$$

**Ej.**

El indicador de una carretilla elevadora dice “Máximo 1200 kg”. Un repartidor de mercancía introduce 10 cajas grandes, cuyo peso se distribuye según una distribución normal de media 140 kg y desviación típica

10 kg. ¿Cuál es la probabilidad de que meta 10 cajas en la carretilla y esta se quede bloqueada?

**Solución:**

La variable  $Y = \text{Peso de 10 cajas}$  corresponde a la suma de 10 variables normales  $X_i \in N(140, 10)$ .

$$Y = X_1 + \dots + X_{10},$$

Por lo tanto, por la fórmula anterior (serían todos los números  $c_i = 1$ ), tenemos que

$$Y \in N(140 \cdot 10, \sqrt{10 \cdot 10^2}) \equiv N(1400, 31.623).$$

Entonces

$$\begin{aligned} P(\text{la máquina se bloquee}) &= P(Y > 1200) = \\ &= P\left(Z > \frac{1200 - 1400}{31.623}\right) = P(Z > -6.32) \simeq 1. \end{aligned}$$

Por lo tanto, la máquina se bloquea seguro.

## 6.5 El teorema central del límite

En el resultado anterior, veíamos que la suma de variables aleatorias normales es otra variable aleatoria normal. Sin embargo, la normalidad de una suma de variables no se limita solo a las variables normales. El teorema central del límite es un resultado matemático que garantiza que, si sumamos variables cualesquiera (no necesariamente normales), la variable suma también seguirá una distribución normal (esto siempre que se cumplan algunas condiciones básicas).

Así, cuando un dato o resultado es la suma de contribuciones independientes, de igual magnitud y “con un tamaño típico”, este resultado corresponderá

a una distribución Gaussiana siempre que el número de contribuciones (el número de sumandos) sea un número considerable (no pequeño).

Con un tamaño típico se quiere garantizar que las contribuciones tienen que “estar controladas”, esto es, las contribuciones extremas tienen que estar controladas por una probabilidad muy pequeña (En jerga matemática las contribuciones tienen que tener varianza finita).

Este teorema asegura, de manera esquemática, que, cuando sumamos un número grande de variables, la variable resultante sigue una distribución normal.

De manera general, si  $X_1, X_2, \dots, X_n$  son variables de media o esperanza  $\mu_i = E(X_i)$  y varianza  $\sigma_i^2 = \text{Var}(X_i)$ ,  $i = 1, \dots, n$ , se verifica que la variable suma  $Y = X_1 + X_2 + \dots + X_n$  (si  $n$  es un número tendiendo a infinito) se puede aproximar por una variable normal, de media la suma de las medias y varianza la suma de varianzas (desviación típica = raíz de la suma de varianzas), es decir

$$Y = X_1 + X_2 + \dots + X_n \approx N \left( \sum_{i=1}^n \mu_i, \sqrt{\sum_{i=1}^n \sigma_i^2} \right).$$



En el caso de sumar variables aleatorias normales, la aproximación anterior no es tal, sino que es una distribución exacta, como hemos visto anteriormente.



Si, en vez de sumar variables, realizamos la media aritmética de las mismas, también podemos utilizar el teorema central del límite (puesto que la media aritmética es sumar y luego dividir por una constante).

Este teorema (del que damos únicamente una idea general, sin establecer las hipótesis matemáticas reales) establece la importancia de la distribución normal. Su resultado es que, cuando se suma un número grande de variables aleatorias, la variable resultante es una variable con distribución aproximadamente igual a la distribución normal. Incluso, el término *número grande* (porque matemáticamente el teorema se establece cuando  $n$  tiende a infinito) no lo es tanto, porque, en la práctica, con tener que  $n$  sea un número mayor o igual a 30, la aproximación ya proporciona buenos resultados.

Además, el teorema es cierto independientemente de la distribución que sigan las variables que se sumen (no importa si son exponenciales, binomiales, etc.). Lo único que se necesita es saber su media y su varianza.

Ej.

El *consumo de petróleo (gas, electricidad...)* de una ciudad es la suma de los consumos individuales de las familias o particulares. Por ello la distribución de esta variable (consumo) va a seguir una distribución normal.

Ej.

El *tiempo de realización de un proyecto complejo* (como construir una casa, un submarino, un avión, una red de carreteras, un oleoducto...) es la suma de los tiempos de las distintas tareas que componen el proyecto. A pesar de que habrá tareas que tendrán un tiempo fijo, la mayoría serán variables con diferente tiempo medio y diferente variación. Pero la suma de los tiempos seguirá una distribución normal, y se podrán calcular probabilidades de finalización en un tiempo determinado (y a su vez el coste de este tiempo).

Ej.

La proporción de una característica  $A$  en una muestra sigue una distribución normal. Comprobémoslo.

La proporción muestral de una característica  $A$  es el número de veces que dicha característica  $A$  aparece en una muestra. Por ejemplo, si  $A$  representa tener una enfermedad cualquiera,  $p = P(A)$  es la probabilidad de que una persona tenga la enfermedad.

Si se seleccionan, de manera independiente,  $n$  personas, tenemos una muestra de  $n$  individuos de esa población, y la proporción muestral es:

$$\hat{p} = \frac{\text{número de individuos en la muestra con esa enfermedad}}{n}$$

En vez de tener una enfermedad,  $A$  puede representar *estar de acuerdo o no con algo, tener trabajo o no*, etc (cualquier cosa que admita solo 2 posibilidades complementarias).

Cada vez que consideramos una persona, podemos considerar la variable de

Bernoulli  $X$  = tiene la enfermedad (o característica)  $A$ . Esta variable toma los valores 1 con probabilidad  $p$  y 0 con probabilidad  $1 - p$ .

De esta manera, la proporción muestral que acabamos de definir se puede considerar como

$$\hat{p} = \frac{X_1 + X_2 + \dots + X_n}{n},$$

donde  $X_1$  es la variable  $X$  en el individuo 1, ...,  $X_n$  es la variable  $X$  en el individuo  $n$ , es decir vale 1 o 0 en cada individuo, según tenga la característica  $A$  o no la tenga.

De manera que, si  $n$  es grande, por el teorema central del límite, la variable suma  $X_1 + X_2 + \dots + X_n$  se aproximará mediante una distribución normal, de media la suma de las medias (cada variable de Bernoulli tiene de media  $p$ ) y de desviación típica la raíz cuadrada de la suma de varianzas (y cada variable de Bernoulli tiene de varianza  $p \cdot (1 - p)$ ). En consecuencia, la variable suma  $Y$  verificará:

$$Y = X_1 + X_2 + \dots + X_n \approx N(\mu, \sigma)$$

donde  $\mu = p + p + \dots + p = np$  y  $\sigma = \sqrt{p \cdot (1 - p) + \dots + p \cdot (1 - p)} = \sqrt{np(1 - p)}$ .

Supongamos ahora que lanzamos una moneda. La variable  $X$  que vale 1 si sale cara y 0 si sale cruz es una variable de Bernoulli. Si lanzamos una moneda, por ejemplo, 200 veces, la variable que mide el número de caras que salen es una suma de 200 variables ( $X_i$  cuenta 1 o 0 si sale cara en el lanzamiento  $i$ ).

Supongamos que repetimos esta operación 10 veces (cada operación es lanzar la moneda 200 veces). La primera vez pueden salir 115 caras, la segunda 94, etc. Se supone que el número de caras andará cerca de 100 (es la media,  $200 \cdot 0.5$ ).

Podemos simular el experimento con R:

```
y=rbinom(10,200,0.5)
```

```
y
```

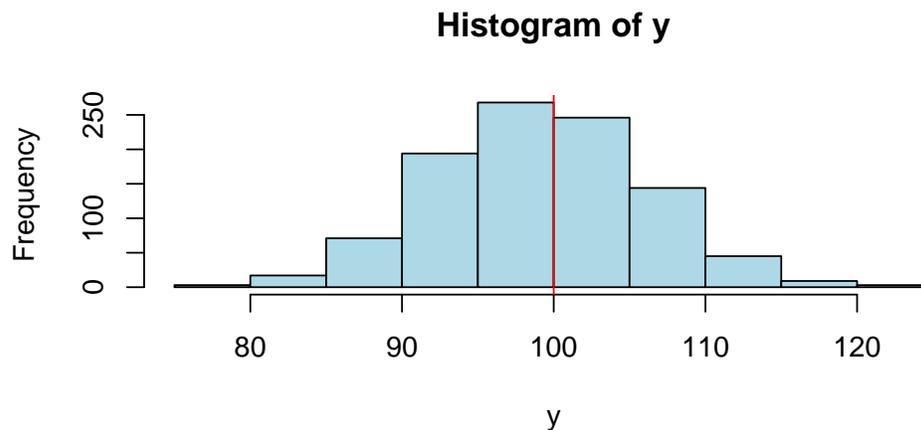
```
## [1] 115 90 105 90 99 91 98 100 95 95
```

Si en vez de repetirlo 10 veces, lo hacemos 1000, tenemos mil valores de la variable. Dibujamos su histograma:

```

y=rbinom(1000,200,0.5)
hist(y, col="lightblue")
abline(v=100, col="red")

```



Como vemos, se parece mucho a la campana de Gauss, con media 100 y desviación típica  $\sqrt{200 \cdot 0.5 \cdot 0.5} = 7.071$ .

Consideremos de nuevo una proporción. Según acabamos de ver, la proporción muestral es

$$\hat{p} = \frac{X_1 + X_2 + \dots + X_n}{n},$$

y, como la suma de arriba es aproximadamente una distribución normal, de parámetros media  $np$  y varianza  $np(1-p)$ , la proporción muestral también sigue aproximadamente una distribución normal.

$$\hat{p} = \frac{X_1 + X_2 + \dots + X_n}{n} \approx N\left(\frac{np}{n}, \sqrt{\frac{np(1-p)}{n^2}}\right) = N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

**Ej.**

Supongamos que entrevistamos a 50 rusos. Parece ser que la probabilidad de que un ruso no crea que el hombre llegó a la luna es 0.57. [fuente](#)

La proporción de gente en la muestra que dirá “NO”, cuando le preguntemos si cree que el hombre llegó a la luna, será un número más o menos cercano a 0.57 (quizá no demasiado, puesto que 50 son pocas personas).

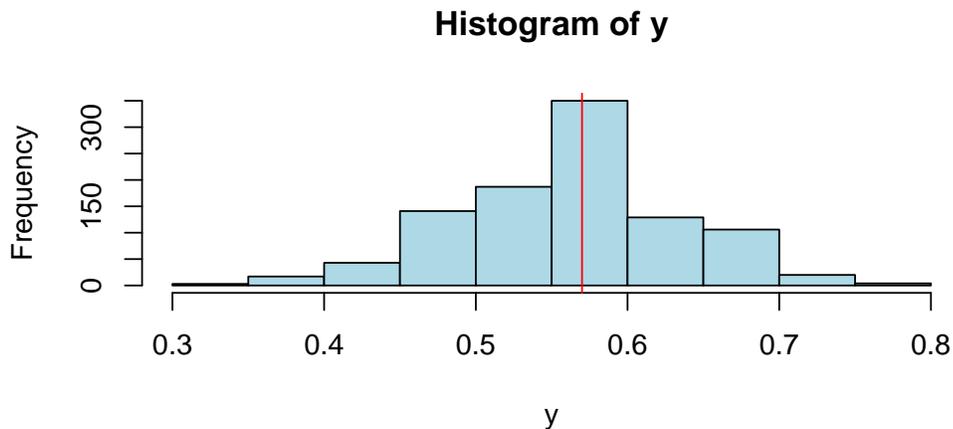
Podemos simular en R el proceso de preguntar a 50 personas hoy, mañana, pasado y así, por ejemplo, durante 20 días:

```
y=rbinom(20,50,0.57)
y/50 # proporción en cada muestra de 50 personas
```

```
## [1] 0.0 0.0 0.50 0.54 0.58 0.56 0.0 0.0 0.56 0.56
## [11] 0.54 0.52 0.0 0.0 0.0 0.0 0.0 0.0 0.52 0.54 0.0
```

Si en vez de repetirlo 20 días, lo hacemos 1000, tenemos mil valores de la proporción muestral, y, como antes, podemos dibujar su histograma:

```
y=rbinom(1000,50,0.57)/50
hist(y, col="lightblue")
abline(v=0.57, col="red")
```



Vemos, por lo tanto, que la proporción muestral toma valores cuya distribución es aproximadamente normal, de media la verdadera proporción  $p$  y desviación típica  $\sqrt{p(1-p)/n}$ , en este caso  $\sqrt{0.57(1-0.57)/1000} = 0.015$ .

Son, precisamente, este tipo de resultados un fuerte apoyo de la teoría del muestreo, en la cual se fundamentan las encuestas de opinión o electorales. Si las muestras de la población en la que se realiza un sondeo son relativamente grandes, se puede precisar con bastante fiabilidad la opinión de una población a través de la muestra, y la variabilidad existente en esta “opinión muestral”.

Ej.

Zenón de Citium era un famoso filósofo que tardaba en comer un jabalí de media 12 minutos y desviación típica 3. ¿Cuál era la probabilidad de que tardase menos de 9 horas y media en comer 50 jabalies?

### Solución

La variable  $T$ =tiempo que tarda Zenón en comer un jabalí tiene media 12 y varianza 9.

Por lo tanto, la variable que mide el tiempo en comer 50 jabalies sigue, aproximadamente, una distribución normal de media  $50 \cdot 12$  y varianza  $50 \cdot 9$ . Es, por lo tanto, aproximadamente  $N(600, \sqrt{450}) = N(600, 21.21)$ .

Así, la probabilidad pedida es

$$P(\text{tiempo total} < 570) = P\left(Z < \frac{570 - 600}{21.21}\right) = P(Z < -1.41) = 0.0786.$$

### 6.5.1 Para pensar un poco: genética.

Galton se dio cuenta que en variables como la estatura aparecen dos efectos principales que hay que tener en cuenta: Los efectos genéticos (la estatura de los padres) y los que no lo son. Ahora bien, ya hemos hablado de que los padres altos tienden a tener descendencia alta, aunque exista una regresión a la media. Pero, curiosamente, debería existir una cierta clase dentro de la población total (los altos) cuya descendencia debería ser más bien alta; al igual que los padres bajos deberían tener descendencia más bien baja.

Esta situación puede expresarse de la siguiente manera: supongamos que elegimos dos números reales positivos concretos  $x_0$  e  $y_0$ , y buscamos todos los pares de padres verificando que uno de ellos mide  $x_0$  unidades de altura y el otro  $y_0$  unidades de altura. A continuación, buscamos en todos los descendientes de estos pares de padres. Se puede suponer que existe una cierta

función  $f$  que denota el efecto genético de las alturas de los padres en la altura de la descendencia, y una variable aleatoria  $W$  que denote los efectos aleatorios en las alturas de la descendencia. Luego, para un conjunto dado de alturas de padres  $x_0, y_0$ , la variable aleatoria que representa las alturas de la descendencia está dada por

$$H = f(x_0, y_0) + W.$$

Si suponemos que el efecto de  $f$  es grande en comparación con el efecto de  $W$ , entonces la varianza (variabilidad) de  $W$  será pequeña. Pero, como  $f$  es determinista, la varianza de  $H$  es igual a la varianza de  $W$ , por lo que la varianza de  $H$  también será pequeña. Sin embargo, Galton observó de sus datos que la varianza de las alturas de la descendencia de un determinado par de la altura de los padres no es pequeña. Esto parece implicar que la herencia juega un papel pequeño en la determinación de la altura de un individuo.

Ahora consideremos la explicación moderna de por qué ciertos rasgos, tales como las alturas, se distribuyen normalmente. Para hacerlo, necesitamos introducir alguna terminología del campo de la genética. Las células en un organismo vivo que no están directamente involucradas en la transmisión de material genético a la descendencia se llaman células somáticas, y las células restantes se llaman células germinales. Los organismos de una especie dada tiene su información genética codificada en conjuntos de entidades físicas, llamados cromosomas. Los cromosomas están emparejados en cada célula somática.

Por ejemplo, los seres humanos tienen 23 pares de cromosomas en cada célula somática. Las células sexuales contienen un cromosoma de cada par. En la reproducción sexual, dos células sexuales, una de cada padre, contribuye con sus cromosomas para crear el conjunto de cromosomas para la descendencia. Los cromosomas contienen muchas subunidades, llamadas genes. Los genes consisten en moléculas de ADN, y un gen tiene, codificado en su ADN, información que conduce al registro de proteínas. En el presente contexto, consideraremos aquellos genes que contienen información que tiene un efecto sobre algún rasgo físico, como la altura. El emparejamiento de los cromosomas da lugar a un emparejamiento de los genes.

En una especie dada, cada gen puede ser de varias formas. Estas formas se llaman alelos. Los diferentes alelos producen diferentes efectos en un rasgo

físico en cuestión. De los dos alelos que se encuentran en un par de genes dado en un organismo, uno de los alelos proviene del padre y el otro alelo de la madre. Los posibles tipos de pares de alelos (sin tener en cuenta el orden) se llaman genotipos.

Si suponemos que la altura de un ser humano está en gran medida controlada por un gen, estamos asumiendo que cada padre tiene un par de alelos que controla en gran medida su altura. Debido a que cada padre contribuye con un alelo de este par de genes a cada uno de sus descendientes, hay cuatro posibles pares de alelos para la descendencia en esta ubicación del gen. La suposición es que estos pares de alelos controlan en gran medida la altura de la descendencia, y también suponemos que los factores genéticos superan a los factores no genéticos. Se sigue que entre la descendencia deberíamos ver varios modos en la distribución de la altura; un modo correspondiente a cada posible par de alelos.

Una hipótesis alternativa que explica la observación de los alturas en la descendencia de un sexo dado, es la hipótesis de múltiples genes. Debajo esta hipótesis, suponemos que hay muchos genes que afectan a la altura de un individuo. Estos genes pueden diferir en la cantidad de sus efectos. Por lo tanto, podemos representar cada par de genes por una variable aleatoria  $X_i$ , donde el valor de cada variable es el efecto del par de alelos en la altura del individuo. Por lo tanto, por ejemplo, si cada padre tiene dos alelos diferentes en el par de genes en consideración, entonces la descendencia tiene uno de cuatro posibles pares de alelos en esta ubicación del gen. Ahora, la altura de la descendencia es una variable aleatoria, que se puede expresar como

$$H = X_1 + X_2 + \dots + X_n + W,$$

si hay  $n$  genes que afectan a la altura (aquí, como antes, la variable aleatoria  $W$  denota los efectos no genéticos). Aunque  $n$  es fijo, si es bastante grande, entonces el teorema central del límite garantiza que  $X_1 + X_2 + \dots + X_n$  tiene una distribución aproximadamente normal. Ahora, si suponemos que la suma tiene un efecto acumulativo significativamente mayor que  $W$ , entonces  $H$  tiene una distribución aproximadamente normal.

Otra característica observada de la distribución de las alturas de los adultos de un sexo en una población es que la varianza no parece aumentar o disminuir de una generación a la siguiente. Esto era conocido en el momento de

Galton, y sus intentos de explicar esto lo llevó a la idea de la regresión a la media.

Usando la hipótesis de múltiples genes, es fácil explicar por qué la varianza debería ser constante de generación en generación. Comenzaremos por suponer que, para una ubicación específica de un gen, hay  $k$  alelos que denotaremos por  $A_1, A_2, \dots, A_k$ . Supongamos que la descendencia se produce por apareamiento aleatorio. Consideramos el conjunto  $S$  de todos los alelos (en un gen concreto) en todas las células germinales de todos los individuos en la generación principal. En términos del conjunto  $S$ , por apareamiento aleatorio queremos decir que cada par de alelos en  $S$  es igualmente probable que resida en cualquier descendencia en particular. Para  $1 \leq i \leq k$ , sea  $p_i$  la proporción de alelos en la población de padres que son de tipo  $A_i$ . Está claro que esto es lo mismo que la proporción de alelos en las células germinales de la población de origen, suponiendo que cada padre produce aproximadamente la misma cantidad de células germinales.

Consideremos la distribución de alelos en la descendencia. Como cada célula germinal tiene la misma probabilidad de ser elegida para cualquier descendencia en particular, la distribución de alelos en la descendencia es la misma que en los padres. A continuación, consideramos la distribución de genotipos en las dos generaciones.

La distribución de genotipos en la generación de descendientes depende solo de la distribución de los alelos en la generación principal (en particular, no depende de la distribución de genotipos en la generación principal). Las frecuencias de los genotipos dependerá únicamente de las frecuencias de los alelos en la generación principal.

Esto significa que si comenzamos con cierta generación y cierta distribución de alelos, en todas las generaciones posteriores a una considerada, tanto el alelo como la distribución del genotipo serán corregidas. Esta última declaración es conocida como la Ley Hardy-Weinberg.

La ley de Hardy-Weinberg -llamada así por el matemático G. H. Hardy (1877-1947) y el médico alemán Wilhelm Weinberg (1862-1937)-, es un principio fundamental que los genetistas utilizan en el estudio de la evolución de las poblaciones. En términos sencillos, la ley de Hardy-Weinberg afirma que, si una gran población se aparea de forma totalmente aleatoria (y no sufre los efectos de mutaciones, migraciones o selecciones), la constitución genética

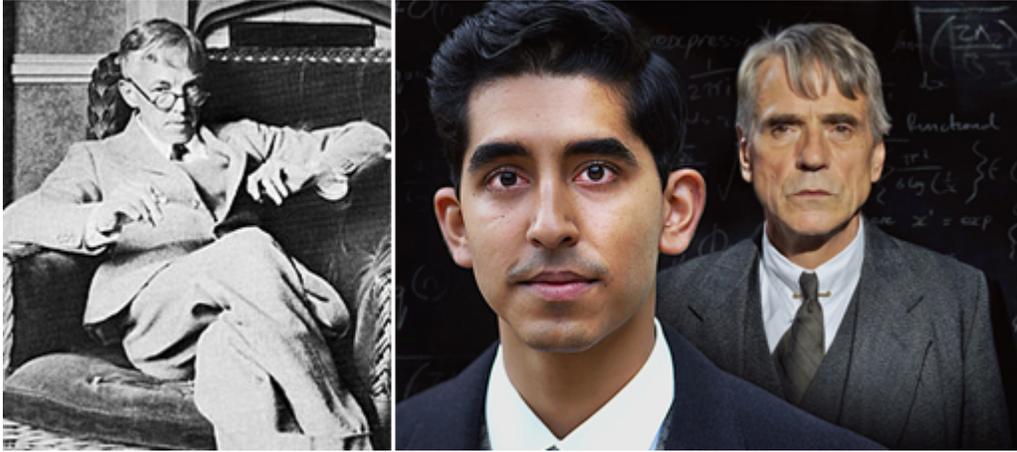


Figura 6.26: G. H. Hardy, que fue interpretado por Jeremy Irons en la película \*El hombre que conocía el infinito\*.

permanece constante de una generación a la siguiente.

Podemos describir las consecuencias de esta ley para la distribución de alturas entre adultos de un sexo en una población. Recordamos que la altura de una descendencia viene dada por una variable aleatoria  $H$ , donde

$$H = X_1 + X_2 + \dots + X_n + W,$$

con las  $X_i$  correspondientes a los genes que afectan a la altura, y la variable aleatoria  $W$  que denota efectos no genéticos.

La ley de Hardy-Weinberg establece que, para cada  $X$ , la distribución en la generación descendiente es la misma que la distribución en la generación de padres. Por lo tanto, si suponemos que la distribución de  $W$  es aproximadamente la misma de generación en generación (o si suponemos que sus efectos son pequeños), entonces la distribución de  $H$  es la misma de generación en generación. De hecho, los efectos de la nutrición forman parte de  $W$ , y está claro que en muchas poblaciones humanas, las dietas han cambiado bastante de una generación a otra en los últimos tiempos. Se cree que este cambio es una de las razones por las cuales los humanos, en promedio, estamos creciendo. Asimismo, también se cree que los efectos de  $W$  son pequeños en relación con la efectos de los padres (la genética).

Obviamente, este razonamiento puede realizarse no solo con la estatura, sino

con cualquier otra variable característica de los seres humanos, para entender la importancia de la herencia en el desarrollo de dichas características en generaciones posteriores.

## 6.6 Las leyes de los grandes números

“grandes números de individuos, actuando independientemente en un sistema, producen regularidades que no dependen de su coordinación mutua, de manera que es posible razonar sobre la colectividad sin ningún conocimiento detallado de los individuos”

— ‘Simeon Denis Poisson’

En resumen: **“No se puede predecir el comportamiento individual, pero si el comportamiento promedio”**.

En 1837 Siméon-Denis Poisson publica su *Investigación sobre la probabilidad de los juicios* (Poisson, 1837), una obra de más de 400 páginas, donde abunda en la temática contemplada por su maestro Laplace acerca de la composición más equitativa e imparcial para los jurados populares. Laplace había realizado cálculos y determinado, según ciertas condiciones, lo que él entendía como la composición más justa y la mayoría necesaria para los mismos. Sin embargo, no se había adentrado en las características o formación que debería exigírsele a los miembros de un jurado, algo que para Poisson resultaba imprescindible. Por otro lado, también expone que debe tenerse en cuenta la proporcionalidad del daño de una condena errónea, en el sentido de que la probabilidad del error debe ser tal que sea más peligroso para la seguridad de la sociedad la absolución de un culpable que la condena de un inocente.

En esta obra, Poisson explica la ley de los grandes números (“la base de todas las aplicaciones del cálculo de las probabilidades”, según él mismo dice) a través de diversos ejemplos de su aplicación. Muchos tienen que ver con la extraña regularidad de multitud de fenómenos físicos: los golpes de azar en juegos, regularidad en las mareas, la vida media de las personas. Fenómenos que, en principio, podrían parecer independientes, como los accidentes de un barco, que podrían depender del navío, del mar, de su país de procedencia (que influirían en la mejor o peor construcción) acaban presentando patrones de comportamiento en el tiempo que permiten precisamente aproximar la probabilidad de los mismos. La regularidad en los golpes de azar en los juegos



Figura 6.27: Simeón Poisson, del que ya hablamos en el capítulo de probabilidades.

de cartas, las mareas, los índices de mortalidad, los fallos condenatorios, los tipos de crímenes, son otros hechos en donde la regularidad se exhibe de forma similar a lo que Bernoulli había pronosticado en su ley de estabilidad de las frecuencias.

Pero no sólo en hechos de índoles física, sino también *moral* existe este mismo tipo de regularidad. Así, al igual que Bernoulli había pronosticado que la repetición de un experimento sirve para calcular una probabilidad de un suceso determinado, Poisson establece que las tasas de ocurrencia de sucesos de esta índole también aventuran las probabilidades. La comparación de los cocientes de acusados frente a juzgados en Francia y en Bélgica, bajo un sistema judicial similar, resultan casi idénticas. A lo largo de los años, se mantienen las diferencias entre los tipos de delitos, de las diferencias en las condenas de hombres a mujeres... De esta forma, por ejemplo, la proporción de condenados anualmente permitirá conocer de manera bastante exacta la probabilidad de ser condenado y bajo qué acusación. Por lo tanto, bajo la misma jurisprudencia, podrá también calcularse la proporción de condenas incorrectas, y esta jurisprudencia podrá alterarse paulatinamente para estar más acorde al desarrollo de la sociedad.

Poisson se muestra tan defensor de la ley de los grandes números que afirma **que no sólo no hay que preocuparse por tal regularidad ni buscar la acción de una mano oculta**, sino que más bien habrá que hacerlo cuando

esa regularidad no se produzca.

Según una antigua tradición relacionada con discípulos de Newton, como por ejemplo De Moivre, la estabilidad de las frecuencias relativas era un signo de la acción de la Divina Providencia. Poisson pensaba que su teorema dejaba saldada la cuestión:

“Podría uno sentirse tentado a atribuir [la estabilidad estadística] a la intervención de un poder oculto, diferente de las causas físicas o morales de los sucesos, y que obrara de alguna manera para mantener el orden; pero la teoría muestra que esa permanencia se da necesariamente, mientras no cambie la ley de la probabilidad de las causas relativas a cada clase de sucesos”. (Poisson, 1837), pag. 144.

### 6.6.1 Las matemáticas

La ley de los grandes números viene a decir que (bajo ciertas condiciones generales) la media de  $n$  variables aleatorias  $X_1, X_2, \dots, X_n$  se aproxima a la media de las  $n$  medias  $\mu_1, \mu_2, \dots, \mu_n$  (donde  $\mu_i = E(X_i)$ ).

$$\frac{X_1 + X_2 + \dots + X_n}{n} \longrightarrow \frac{\mu_1 + \mu_2 + \dots + \mu_n}{n}$$

Si todas las variables tienen la misma media  $\mu$ , entonces la media aritmética de las variables se aproxima al mismo valor.

Un caso particular de esta ley es el principio de estabilidad de las frecuencias, o teorema de Bernoulli, que ya hemos visto. Efectivamente, recordemos que una variable de Bernoulli es aquella que toma solo el valor 0 o 1 cuando no ocurre (u ocurre, respectivamente) un suceso  $A$  con probabilidades  $1 - p$  y  $p$ . Sumar  $n$  variables de Bernoulli es contar el número de veces que se repite el suceso  $A$  en  $n$  pruebas.

Una variable de Bernoulli tiene media  $p$  (cálculo muy sencillo). Luego la media de  $n$  medias será también  $p$ .

La ley de los grandes números generaliza este resultado a experimentos donde no necesariamente repetimos siempre la misma prueba (como en el caso anterior).  $X_1$  podría contar si ocurre un suceso  $A_1$  (de probabilidad  $p_1$ ),  $X_2$

si ocurre un suceso  $A_2$  (de probabilidad  $p_2$ ), etc... con diferentes probabilidades cada uno. La ley de los grandes números establecerá la regularidad por cuanto la suma de frecuencias de ocurrencia de los sucesos tenderá a la media de las probabilidades  $(p_1, p_2, \dots)$

El mismo Poisson lo explica sencillamente en su obra citada anteriormente: “Supongamos que lanzamos al aire una moneda de cinco francos, y observamos que, en 2.000 tiradas, la moneda sale cara 1.100 veces. Entendemos que hay una frecuencia o probabilidad constante de que la moneda salga cara, esto es,  $11/20$ . Esta constante es la consecuencia de una causa común, de la manera en que está hecha la moneda y de la manera de arrojarla. Pero supongamos ahora que tiramos 2.000 monedas diferentes y obtenemos 1.100 caras. No podemos imaginar que las monedas tengan constituciones idénticas. Las causas y, por lo tanto, las probabilidades de salir cara, variarán de un caso a otro.”

Muchos sucesos legales, sociales, de la moral y de las ciencias naturales son como el caso de las múltiples monedas. Cada viaje por mar es diferente. Poisson indicaba que un barco es atacado por un tifón, otro no, otro tiene un piloto incompetente y otro es atacado por piratas. No hay una causa constante que obre sobre los marinos, pero, sin embargo, existía un efecto constante, una proporción constante y demostrada de naufragios. Lo mismo ocurría con los jurados cuyos miembros varían en cuanto a sabiduría y prejuicios, pero que manifiestan un efecto general estable en las tabulaciones del Ministerio de justicia en cuanto a resultados prácticamente invariables de año a año.

## 6.7 V.A. obtenidas a partir de la variable normal

Definimos, a continuación, tres variables aleatorias muy utilizadas en la inferencia estadística, y que se construyen a partir de la variable aleatoria normal.

### 6.7.1 Variable Chi-cuadrado (Pearson)

Se escribe variable  $\chi^2$ . El hecho de que su función de densidad dependa de un número entero positivo llamado *grados de libertad* hace que se hable de la distribución  $\chi_k^2$  con  $k$  grados de libertad. Así, existe una variable para cada

valor de  $k$  mayor o igual a 1. Esta variable aparece cuando se suman  $k$  variables aleatorias independientes con distribución  $N(0, 1)$ , elevadas al cuadrado.

$$\chi_k^2 = X_1^2 + X_2^2 + \dots + X_k^2, \text{ con } X_i \in N(0, 1).$$

Esta distribución es necesaria para la construcción de intervalos de confianza y la realización de muchos contrastes de hipótesis.

En la Figura 6.28 dibujamos con R las gráficas de las funciones de densidad de la variable Chi-cuadrado con 1, 2, 4, 10 y 20 grados de libertad, respectivamente. Fijémonos que, comenzando con uno y dos grados de libertad ( $df$  en la gráfica, de *degree freedom*), donde la función de densidad es parecida a la función de densidad de una variable exponencial, la curva se vuelve asimétrica a la derecha y, a medida que el número de grados de libertad se hace más grande (20 en la gráfica) la curva se va pareciendo a la campana de Gauss. Esto sucede por la definición de la variable: al ser una suma de variables, por el teorema central del límite, la densidad de la suma se va aproximando a la densidad de la variable normal.

```
curve(dchisq(x, df= 1), xlim = c(0, 20), ylim = c(0, 0.2),
      xlab = " ", ylab = "Función de densidad")
curve(dchisq(x, df= 2), col = "red", lty = 2, add = T)
curve(dchisq(x, df= 4), col = "blue", lty = 3, add = T)
curve(dchisq(x, df= 10), col = "green", lty = 4, add = T)
curve(dchisq(x, df= 20), col = "magenta", lty = 5, add = T)
abline(h = 0, col = "gray")
legend("topright", c("1", "2", "4", "10", "20"), col = c("black",
  "red", "blue", "green", "magenta"), lty = c(1, 2, 3,
  4, 5), bty = "n")
```

### 6.7.2 Variable t de Student

Aparece a partir de la distribución normal y la Chi-cuadrado, puesto que si  $Z$  es una variable  $N(0, 1)$ , e  $Y$  es una variable aleatoria independiente de  $Z$ , con distribución Chi-cuadrado con  $n$  grados de libertad, entonces la variable

$$t_n = \frac{Z}{\sqrt{Y/n}}$$

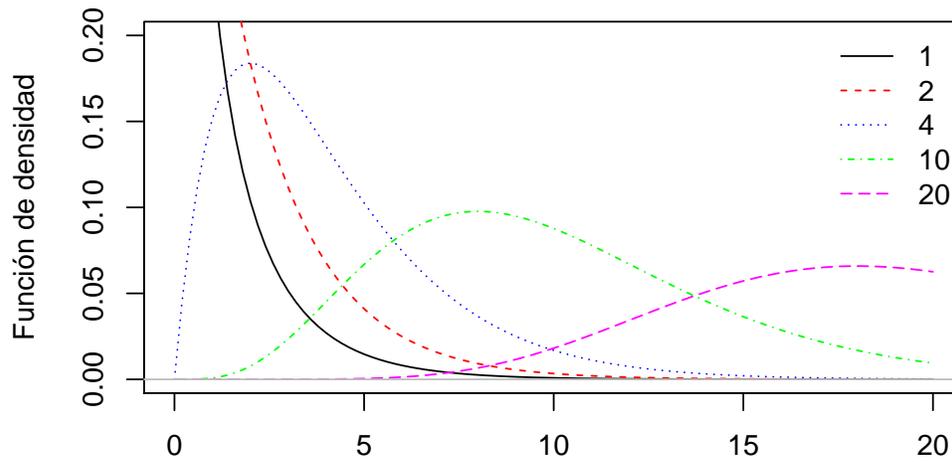


Figura 6.28: Varias distribuciones Chi-cuadrado para distintos grados de libertad.



Figura 6.29: William Gosset, el inventor.

es una variable con distribución  $t$  con  $n$  grados de libertad. La distribución  $t$  de Student (con  $n$  *grados de libertad*) es una variable aleatoria cuya función de densidad también tiene forma de campana y es simétrica (Figura 6.30). Es, por lo tanto, muy parecida a la densidad de la variable normal (de hecho, la media o esperanza de cualquier variable  $t$  de Student es cero). Sin embargo, tiene colas más pesadas que la campana de Gauss, lo que significa que el área o probabilidad en los extremos izquierdo y derecho de la curva es mayor que en el caso de la distribución normal. Esto pueden visualizarse en la gráfica siguiente, donde se representan diferentes valores del parámetro  $n$ .

Hay que reseñar que, a medida que el número de grados de libertad aumenta, la curva se parece cada vez más a la campana de Gauss, lo cual sucede a partir de valores como  $n = 40$ , siendo la coincidencia total para  $n = \infty$ .

```
curve(dt(x,d#1),xlim=c(-3,3),ylim=c(0,0.4),
      xlab=' ',ylab='Función de densidad')
curve(dt(x,d#2),col='red', lty=2,add=T)
curve(dt(x,d#4),col='blue',lty=3, add=T)
curve(dt(x,d#10),col='green',lty=4, add=T)
curve(dt(x,d#20),col='magenta',lty=5, add=T)
abline(h=0, col="gray")
legend("topright",c("1","2","4","10"),
col=c("black","red","blue", "green","magenta"),
lty=c(1,2,3,4,5),bty="n")
```

El interés de esta variable es, igual que en el caso de la variable Chi-cuadrado, su aparición en la construcción de intervalos de confianza y realización de contrastes de hipótesis.

La fórmula de la función de densidad de la variable  $t$  fue publicada en 1908 por William Sealy Gosset (1876-1937), mientras trabajaba en la fábrica de cervezas Guinness, en Dublin. Existe la anécdota de que el origen del seudónimo *student*, utilizado por Gosset para escribir el artículo de investigación matemático donde define la variable, vino motivado porque la dirección de la fabrica impedía a los empleados la publicación de trabajos científicos.

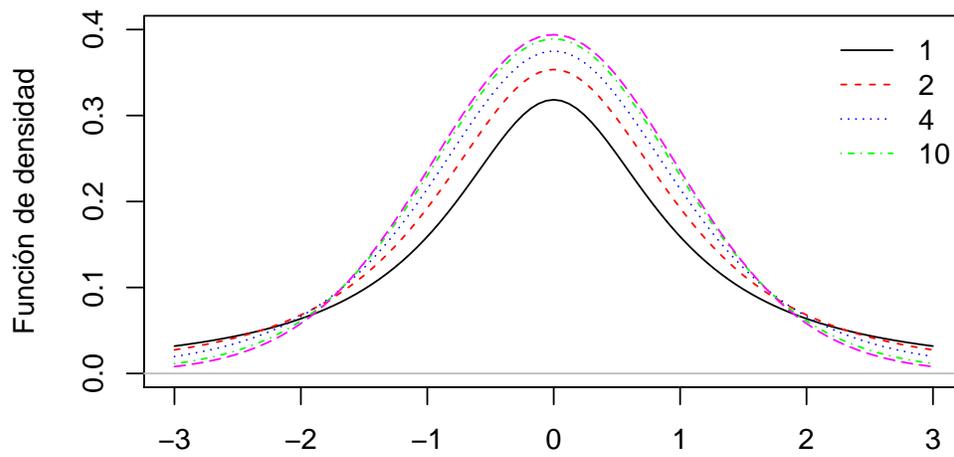


Figura 6.30: Varias distribuciones t para distintos grados de libertad.

### 6.7.3 Variable F de Fisher-Snedecor

La distribución  $F$  es conocida habitualmente como la distribución  $F$  de Snedecor, o distribución  $F$  de Fisher-Snedecor, en honor a R. Fisher y George W. Snedecor (1881-1974). Su función de densidad es bastante complicada, y depende de dos parámetros  $n_1$  y  $n_2$  que son sus grados de libertad.

Esta distribución aparece a través de la distribución  $\chi^2$ , puesto que si  $X$  sigue una distribución  $\chi_{n_1}^2$ , e  $Y$  es otra variable (independiente de  $X$ ) con distribución  $\chi_{n_2}^2$ , entonces la variable  $W$  definida como

$$W = \frac{X/n_1}{Y/n_2}$$

sigue una distribución  $F$  con  $n_1$  y  $n_2$  grados de libertad. En la Figura 6.31 vemos algunos casos concretos para la densidad (fuente: wikipedia).

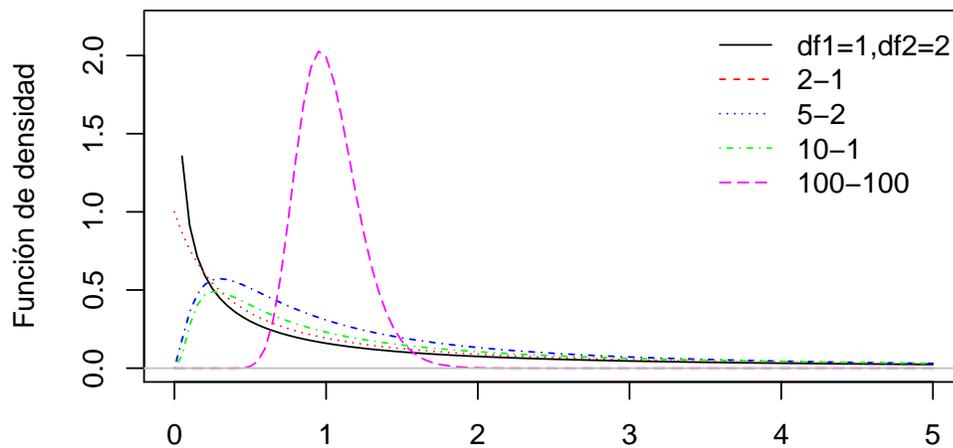


Figura 6.31: Representación de la función de densidad de la distribución F para algunos parámetros.

Odio la Estadística 



El interés de su estudio corresponde a ser una variable fundamental en la teoría de los contrastes de hipótesis (Anovas).



# Capítulo 7

## Inferencia estadística

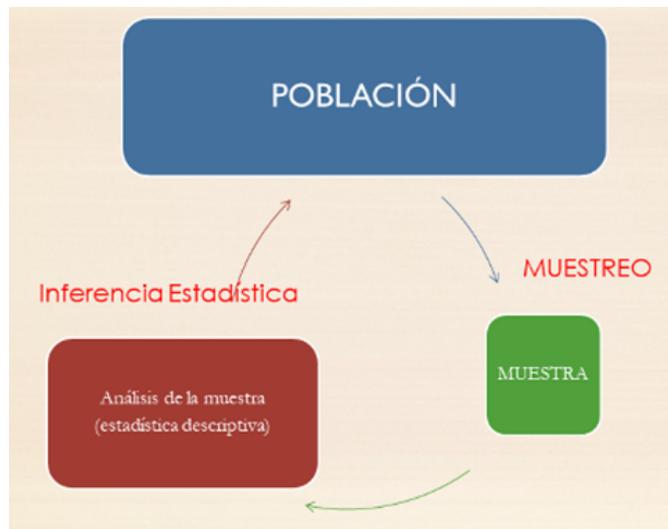


Figura 7.1: Gráfico del proceso científico del estudio estadístico.

Los dos objetivos principales de la Estadística son:

- Describir una muestra o subconjunto de una población (Estadística descriptiva).
- Extraer de la muestra información acerca de la población.

La Inferencia Estadística se utiliza cuando no se puede observar toda la

población (principalmente por motivos económicos). A partir de la muestra, se intenta obtener información que sirva para caracterizar toda la población (Figura 7.1).

Hay varios procedimientos para obtener una muestra de una población. El objetivo de estos procedimientos es que la muestra represente lo mejor posible a dicha población.

Nosotros nos centraremos en el conocido como *muestreo aleatorio simple*.

## 7.1 Muestreo aleatorio simple:

Se llama así al procedimiento de seleccionar una muestra cumpliendo dos propiedades fundamentales:

1. Todos los individuos de la población tienen la misma probabilidad de ser elegidos.
2. Todas las muestras del mismo tamaño son igualmente probables.

Desde un punto de vista matemático, el muestreo aleatorio simple se realiza suponiendo que la población tiene infinitos elementos. En la práctica, suele ocurrir que las poblaciones a considerar tienen un número muy grande de elementos, con lo cual puede equipararse a una situación con infinitos elementos.

Supongamos, por ejemplo, que nos interesa realizar una estimación del sueldo medio de la población trabajadora de un país. Desde un punto de vista meramente intuitivo, una forma de realizar una **inferencia** consistiría en seleccionar una muestra de individuos de la población y preguntarles el sueldo, para posteriormente calcular la media (aritmética, por ejemplo).

La población de trabajadores de un país no tiene infinitos elementos, pero suele ser muy grande. Cuando seleccionamos una muestra (aleatoria simple) de elementos de la población, estamos considerando que todos los individuos se seleccionan de manera independiente; esto es, no hay sesgo de ninguna clase a la hora de distinguir un individuo de otro para preguntarle su sueldo.

A la hora de seleccionar una muestra aleatoria simple, se supondría que todos los miembros a los que se puede encuestar están numerados del 1 al  $N$ . Se realizaría un sorteo en el que se sacarían  $n$  números, y esas serían las personas

seleccionadas.

El motivo de trabajar con un muestreo aleatorio con reemplazamiento es de naturaleza matemática, ya que permite estudiar propiedades de los distintos procesos de inferencia (decir si lo que se hace está bien o está mal, si una estimación es mejor que otra, etc.).

Una muestra aleatoria simple de tamaño  $n$  de una variable aleatoria  $X$  (relativa a una población) viene dada por las variables aleatorias  $X_1, \dots, X_n$ , que serán independientes e idénticamente distribuidas a  $X$ .

En la práctica, una muestra aleatoria simple de tamaño  $n$  de una variable aleatoria  $X$  es un conjunto de datos  $(x_1, x_2, \dots, x_n)$ , obtenidos al observar la variable  $X$  en  $n$  individuos. Por ejemplo,  $n$  datos de la variable  $X = \text{estatura}$ .

Debe tenerse en cuenta que, si la población se compone de grupos importantes que deben considerarse por separado (por sexo, grupos de edad, nivel educativo...), ha de realizarse un **muestreo aleatorio estratificado**. Este consistirá en tener en cuenta el tamaño de cada estrato dentro de la población (por ejemplo, 70 por ciento hombres, 30 por ciento mujeres) y, a la hora de seleccionar la muestra total, mantener el porcentaje de cada estrato dentro de la misma.

Nuestro objetivo es obtener información sobre un parámetro o característica desconocida de la población a partir de una muestra. Para ello podemos:

- Aproximar el parámetro por un valor. Esto se conoce como **Estimación puntual**.
- Construir un intervalo que, con cierta seguridad, contenga el verdadero valor del parámetro. Esto se llama un **Intervalo de confianza** o estimador por intervalo.
- Verificar si cierta hipótesis sobre el parámetro es coherente con los datos observados. **Contraste de hipótesis**.

Dos son los resultados fundamentales sobre los que se plantea gran parte de la introducción a la teoría de la inferencia estadística: el **teorema central del límite** y la **ley de los grandes números** de Poisson. Hemos hablado de ellos anteriormente, pero no está de más recordarlos, desde el punto de vista intuitivo.

El teorema central del límite nos permite considerar que una variable sigue

una distribución aproximadamente normal, cuando sea suma de un conjunto relativamente grande de variables (también si es la media aritmética de un conjunto relativamente grande de variables).

La ley de los grandes números dice que “grandes números de individuos, actuando independientemente en un sistema, producen regularidades que no dependen de su coordinación mutua, de manera que es posible razonar sobre la colectividad sin ningún conocimiento detallado de los individuos”. No se puede predecir el comportamiento individual, pero si el comportamiento promedio. Significa que elaborar predicciones sobre la media de una variable es matemáticamente viable (y, entre comillas, sencillo). A través de los datos de una muestra podremos, por ejemplo, predecir con bastante fiabilidad una estatura media, el tiempo medio para la realización de una actividad, anticipar la efectividad media de un medicamento. No va a ser posible, en general, predecir el resultado de una variable para un caso particular (para un paciente determinado, qué grado de efectividad tendrá el medicamento, o el tiempo de curación para ese paciente, o la estatura que tendrá esa persona cuando alcance su madurez).

Ej.

La variable estatura  $X$  en una población humana es una variable aleatoria con distribución normal. Por lo tanto, depende de dos **parámetros**, que son su media y su desviación típica.

Si tenemos la oportunidad de medir la estatura de todos los miembros de una población concreta (por ejemplo los varones mayores de 18 años), obviamente podemos obtener su media y su desviación típica. Pero si podemos disponer de todos los miembros de la población, no necesitamos la estadística para nada. Ahora bien, lo habitual no es poder disponer de todos los miembros de la población, con lo cual lo que se hace es **estimar** esos valores teóricos a través de la muestra.

Ej.

Si queremos saber la proporción de gente mayor de edad que cree en la existencia de Dios, podría organizarse un referendun y que todo el mundo votara. En este caso, estamos hablando de observar la variable  $X$  = “opinión sobre la existencia de Dios”, que toma dos posibles valores (SI o NO). Se trata de una variable aleatoria de Benoulli de parámetro  $p$ . Si se realiza un referendun, el verdadero valor de  $p$  se puede cono-

cer, porque entonces se estudia a TODA la población. Si no se puede estudiar a toda la población, podemos realizar una estimación de ese valor de  $p$  por medio de una muestra.

## 7.2 Estimación puntual

El objetivo de la estimación puntual es aproximar el valor del parámetro desconocido (tiempo medio de ejecución de un algoritmo, altura media de las mujeres de una población, diferencia del resultado medio entre dos tratamientos médicos, proporción de gente que mejora con un tratamiento médico...)

Para ello se utiliza la información de la muestra  $(x_1, x_2, \dots, x_n)$ , a través de un **estimador**.

Algunos estimadores frecuentes son:

- **Media muestral**, para estimar la media teórica de una variable  $X$ .

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

- **Proporción muestral**, para estimar una proporción  $p$ :

$\hat{p} = \frac{x_1 + \dots + x_n}{n}$ , siendo  $x_1, \dots, x_n$  una muestra aleatoria simple de la variable  $X \in B(1, p)$ , es decir, son unos o ceros.

- **Varianza muestral**: para estimar la varianza teórica de una población, se puede usar la varianza de una muestra:

$$S^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n},$$

y también la llamada

- **Cuasi-varianza muestral**:

$$S_{n-1}^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1},$$

que corresponde a la varianza de la muestra, pero dividiendo por  $n - 1$ , en lugar de dividir por  $n$ . En el capítulo de estadística descriptiva, ya comentamos

que el R, por defecto, al calcular la desviación típica de una muestra, mediante el comando `sd`, calcula directamente la cuasi-varianza y luego obtiene la raíz cuadrada.

La evaluación del estimador sobre la muestra fija da lugar a una **estimación puntual**.

Ej.

Cálculo de la media muestral tomando la muestra fija  $(x_1, x_2, x_3) = (2, 7, 1)$ .

$$\bar{x} = \frac{2 + 7 + 1}{3} = \frac{10}{3}$$

## 7.2.1 Propiedades de los estimadores

Estamos diciendo que un estimador es una aproximación de un parámetro teórico o desconocido de una población. Para estimar la media de la altura de una población, podemos seleccionar una muestra y calcular la media aritmética de la muestra. Ahora bien, también tendría sentido usar como estimador el siguiente:

$$\frac{\min(x_1, x_2, \dots, x_n) + \max(x_1, x_2, \dots, x_n)}{2}$$

¿Cuál de los dos se aproxima más al verdadero valor desconocido? En principio, no habría manera de saberlo, puesto que deberíamos conocer el valor teórico (el desconocido). Por eso, interesa estudiar propiedades de los estimadores, que nos permitan decidir entre usar unos u otros para los casos concretos.

### 7.2.1.1 Estimadores insesgados

Una primera propiedad deseable para un estimador es que el centro de la distribución de los valores que puede tomar coincida con el valor del parámetro que queremos aproximar.

A esta propiedad se le llama *insesgades*. Así, un estimador **insesgado** es aquel cuya media coincide con el valor del parámetro a estimar.

Veámoslo con un ejemplo para entenderlo mejor: supongamos que deseamos tener una estimación de la estatura media de los hombres mayores de 18 en



Figura 7.2: Encuestador y encuestada.

una población. Podríamos ponernos en medio de la calle y seleccionar aleatoriamente una muestra de  $n$  hombres, medir su estatura (o preguntársela) y calcular después la media aritmética de los datos obtenidos. Esa sería una estimación puntual; llamémosla  $\bar{x}_1$ .

Por medio de R podemos hacer una **simulación** de este proceso. En vez de bajar a la calle, parar a la gente y preguntarle lo que mide, simulamos cien datos correspondientes a 100 estaturas de varones mayores de 18. En este caso, tenemos que “simular” que medimos a cien personas, de una población de varones españoles mayores de 18.

```
# Consideremos n =100 personas
set.seed(1)
n=100
# así se simulan n datos que siguen
# una distribución normal de
# media 177.7 y desviación típica 5.9 :
X1=rnorm(n,177.7,5.9)
# dibujamos el histograma:
hist(X1, probability = TRUE, col = 'lightblue',
main="100 estaturas de varones mayores de 18")
# dibujamos los puntos:
rug(X1)
# dibujamos la estimación de la densidad:
lines(density(X1), col="red",lwd=2)
```

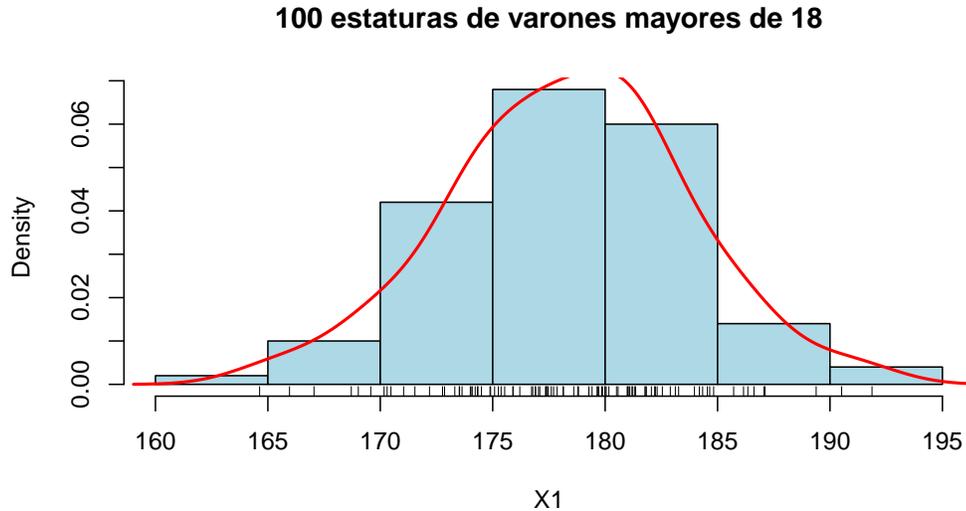


Figura 7.3: Histograma y estimación de la densidad de 100 estaturas.

La media muestral de esos 100 valores es  $\bar{x}_1 = 178.3424$ .

Si vamos al día siguiente a la misma calle y seleccionamos aleatoriamente otra muestra del mismo número  $n$  de personas, medimos su estatura y calculamos la media aritmética, tenemos otra estimación puntual ( $\bar{x}_2$ ).

```
# Otras 100 personas
n=100
X2=rnorm(n,177.7,5.9)
```

La media es  $\bar{x}_2 = 177.4769$ .

Obviamente, estos valores  $\bar{x}_1$  y  $\bar{x}_2$  no coinciden, y no tienen por qué coincidir. En cada caso, hemos seleccionado 100 personas aleatoriamente, hemos medido su estatura y hemos calculado la media muestral. Los datos no van a ser los mismos, y por lo tanto las medias muestrales tampoco. Cada vez que seleccionemos otra muestra, el estimador media muestral da un valor diferente. Esto es, **la media muestral es una variable aleatoria**.

Vamos ahora a suponer que realizamos este proceso un número grande  $B$  de

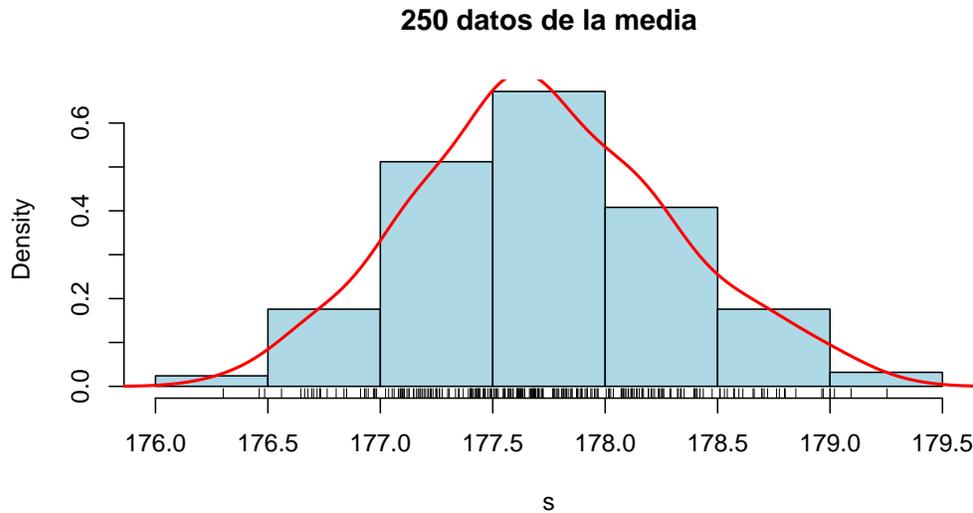


Figura 7.4: Histograma y estimación de la densidad de las 250 MEDIAS de todas las muestras.

veces; es decir, salimos a la calle, medimos a 100 personas, y calculamos la media muestral; al día siguiente volvemos a hacer lo mismo, y así sucesivamente, hasta  $B = 250$  veces, por ejemplo. Mediante el siguiente procedimiento en R, simulamos este procedimiento y hacemos una gráfica (Figura 7.4) de la distribución de los 250 valores obtenidos.

```
n=100;B=250
s<0
for (i in 1:B) s[i]=mean(rnorm(n,177.7,5.9))
hist(s, probability = TRUE, col = 'lightblue',
main="250 datos de la media")
rug(s)
lines(density(s), col="red",lwd=2)
```

La media de estos 250 valores es 177.7205 que es muy próxima al verdadero valor 177.7

De esta forma, comprobamos que la media (de las diferentes medias) se apro-

xima al verdadero valor 177.7. Matemáticamente, se puede demostrar que siempre ocurre así; es decir, que la media muestral es un estimador insesgado.

---

Veamos otro ejemplo:



Figura 7.5: Seres extraños de otra galaxia.

Supongamos que preguntamos en la calle si la gente cree o no en los extraterrestres. En este caso, las respuestas van a ser “Sí” o “No”, que anotaremos como 1 o 0, es decir valores de una variable aleatoria de Bernoulli de parámetro  $p$ , siendo  $p = \text{proporción de gente que cree en los extraterrestres}$ , que es el 61 por ciento ([sacado de aquí](#)). Hacemos como en el caso anterior: cogemos una muestra de tamaño 100 (o cualquier otro número un poco grandecito), preguntamos y anotamos la respuesta. Pero, en vez de hacerlo realmente, lo simulamos con el ordenador, de esta forma:

```
# Consideremos n =100 personas
set.seed(1)
n=100
# así se simulan n datos que siguen
# una distribución de Bernoulli de parámetro 0.61
Y1=rbinom(n,1,0.6)
pander(table(Y1))
```

0	1
42	58

El número de unos en esta encuesta (gente que dijo que creía en los extraterrestres) es de 58, con lo que la proporción muestral es  $\hat{p}_1 = 0.58$ .

Si vamos al día siguiente a la misma calle y seleccionamos aleatoriamente otra muestra del mismo número  $n$  de personas, les preguntamos lo mismo, tendremos otra estimación puntual del número de gente que cree en los extraterrestres ( $\hat{p}_2$ ).

```
n=100
Y2=rbinom(n,1,0.6)
pander(table(Y2))
```

0	1
38	62

El número de unos en esta encuesta (gente que dijo que creía en los extraterrestres) es de 62, con lo que la proporción muestral es  $\hat{p}_2 = 0.62$ .

Vamos ahora a suponer que realizamos este proceso un número grande  $B$  de veces; es decir, salimos a la calle, le preguntamos a 100 personas, calculamos la proporción muestral; al día siguiente volvemos a hacer lo mismo, y así sucesivamente, hasta  $B=250$  veces, por ejemplo. Mediante el siguiente procedimiento en R, simulamos este procedimiento y hacemos una gráfica (Figura 7.6) de la distribución de los 250 valores obtenidos.

```
n=100;B=250
x<0
for (i in 1:B) x[i]=sum(rbinom(n,1,0.6))/n
hist(x, probability = TRUE,
col = 'lightblue', main="250 encuestas")
rug(x)
lines(density(x), col="red",lwd=2)
```

La media de estos 250 valores es 0.6119 que es muy próxima al verdadero valor 0.61

Para que se pueda entender este ejemplo mejor, supongamos que las muestras

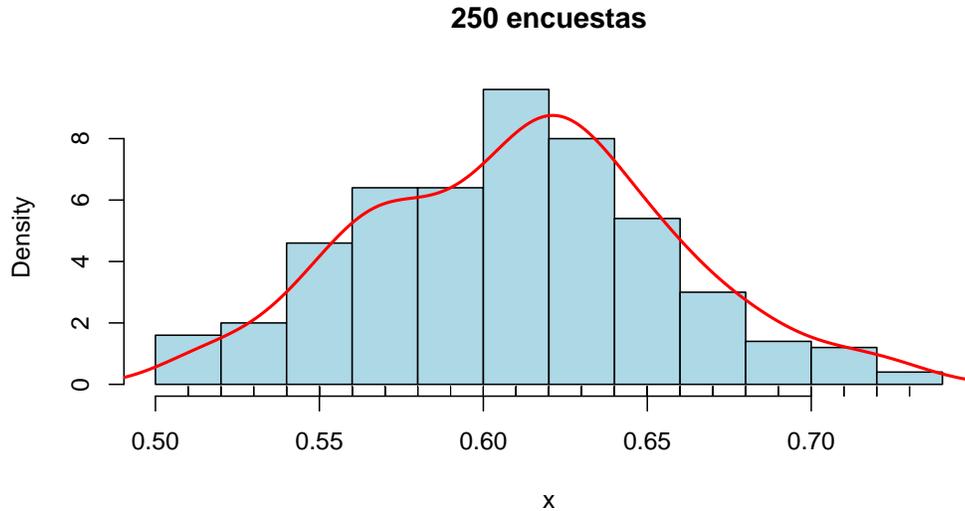


Figura 7.6: Histograma y estimación de la densidad de los valores obtenidos en cada una de las 250 muestras.

seleccionadas son de tamaño 7, y realizamos el proceso 15 veces. A continuación realizamos una simulación donde podemos observar todos los detalles: para cada muestra, los valores 1 o 0 denotan si la persona en cuestión ha contestado “Sí” o “No”, y, a la derecha, calculamos la proporción muestral de “Síes”, es decir el número de ellos dividido por (en este caso) 7.

```
n=7
B=15
muestras <- as.data.frame(matrix(rbinom(n*B,
                                     1, 0.6), ncol=n))
rownames(muestras) <- paste("muestra", 1:B, sep="")
muestras$mean <- rowMeans(muestras[,1:n])
ptilde <- muestras$mean
colnames(muestras) <- c(paste(" obs", 1:n,
                              sep=""), " ptilde")
# muestras:
pander(muestras)
```

	obs1	obs2	obs3	obs4	obs5	obs6	obs7	ptilde
<b>muestra1</b>	0	1	1	0	0	0	0	0.2857
<b>muestra2</b>	0	1	1	1	0	1	1	0.7143
<b>muestra3</b>	0	1	1	1	0	0	1	0.5714
<b>muestra4</b>	0	1	0	0	1	0	1	0.4286
<b>muestra5</b>	1	1	1	0	0	1	0	0.5714
<b>muestra6</b>	0	1	0	1	1	0	0	0.4286
<b>muestra7</b>	0	0	1	1	1	1	0	0.5714
<b>muestra8</b>	1	0	1	1	1	0	1	0.7143
<b>muestra9</b>	1	0	1	0	1	0	1	0.5714
<b>muestra10</b>	1	1	1	1	1	1	0	0.8571
<b>muestra11</b>	0	1	1	1	1	1	1	0.8571
<b>muestra12</b>	1	1	0	0	1	0	1	0.5714
<b>muestra13</b>	1	0	0	0	1	0	1	0.4286
<b>muestra14</b>	1	1	1	1	1	0	1	0.8571
<b>muestra15</b>	1	0	0	1	1	1	1	0.7143

```
mean(ptilde)
```

```
## [1] 0.695
```

Observamos que la proporción muestral también es un estimador insesgado.

### 7.2.1.2 Estimadores consistentes

Un estimador insesgado es **consistente** cuando su varianza tiende a 0 si  $n$  crece hacia infinito.

Recordemos que la varianza mide la dispersión. A mayor varianza, mayor dispersión entre los valores de la variable, y a menor varianza menor dispersión. La propiedad de consistencia indica que, si tomamos muestras muy grandes ( $n$  muy grande, creciendo hacia infinito), la varianza se hará próxima a cero, es decir que obtendremos siempre valores muy próximos entre sí.

Pensemos en los ejemplos que hemos estado viendo donde realizábamos encuestas en la calle (o medíamos la estatura). Lo hacíamos en muestras de tamaño 100, y obteníamos valores diferentes para nuestro estimador (o bien

la media muestral, o bien la proporción muestral). Parece lógico pensar que, si en vez de tomar muestras de tamaño 100, tomamos muestras muy grandes, los valores que obtengamos de cada vez se parezcan mucho entre si (por lo tanto la dispersión o varianza tiende a cero), puesto que es como si estuviésemos de cada vez midiendo a (casi) toda la población.

### 7.2.2 Propiedades de la media muestral

La media muestral  $\bar{X} = \frac{X_1 + \dots + X_n}{n}$ :

- Es un estimador de la media poblacional  $\mu$ .
- Es insesgado.
- Es consistente.
- Si  $X \in N(\mu, \sigma)$  entonces  $\bar{X} \in N(\mu, \sigma\sqrt{n})$ .

### 7.2.3 El error estándar de la media (muestral).

Esta última propiedad nos dice que, si consideramos muestras de una variable normal (como la estatura), la media muestral (la media de las muestras, que es otra variable aleatoria, como hemos visto), sigue también una distribución normal (recordemos que, en el ejemplo de arriba, donde simulamos 250 muestras de cien estaturas, el histograma de las medias muestrales tenía la forma de la campana de Gauss).

Si  $X$  no sigue una distribución normal, pero tiene una media  $\mu$  y una desviación típica  $\sigma$  finitas, entonces, por el teorema central del límite  $\bar{X} \approx N(\mu, \sigma\sqrt{n})$ .

Esta propiedad es casi como la inmediatamente anterior. Si las variables que consideremos no siguen una distribución normal, pues no hay excesivo problema, puesto que, por el teorema central del límite, la media muestral seguirá aproximadamente una distribución normal (siempre que se promedien bastantes variables; en la práctica, más de 30).

El resultado nos dice que  $\bar{X} \in N(\mu, \sigma\sqrt{n})$  (exacta o aproximadamente, según acabamos de comentar), es decir que el parámetro media de la variable  $\bar{X}$  es, precisamente, la media teórica (la misma de la variable  $X$ ), y la desviación típica es la misma que la teórica, pero **dividida por**  $\sqrt{n}$ . Esto se conoce

como el **error estándar** de la media muestral. Por ejemplo, en muestras de tamaño 100, la desviación típica o error estándar es la de la variable  $X$  dividida por 10. En muestras de tamaño 10000,  $\sigma$  aparece dividida solo por 100. Esto nos mide, en cierta manera, la “velocidad” en que la dispersión se va acercando a cero. Y vemos que con tamaños de muestra, por ejemplo, de un millón (que ya es una señora muestra), la desviación típica solo aparece dividida por mil.

En el capítulo 10 del libro “Pensar rápido, pensar despacio” (Kahneman, 2014) Daniel Kahneman explica que la mente humana tiende automáticamente a sacar conclusiones y explicaciones causales de resultados que, procedentes de muestras pequeñas y poco representativas, son meras ilusiones estadísticas, carentes de significado. Kahneman llama a ese frecuente error la “ley de los pequeños números”.

Ilustra el fenómeno con los resultados de un estudio sobre la distribución geográfica del porcentaje de cáncer de riñón entre los más de 3.000 condados de Estados Unidos. Se observó que los porcentajes más bajos se daban en condados rurales poco poblados del Oeste, Medio Oeste y Sur de los Estados Unidos, de esos (añade Kahneman irónicamente) donde predominan los votantes del Partido Republicano.

Como no parece lógico que la intención de voto proteja contra el cáncer, surgen otras explicaciones mucho más sensatas: la vida rural es más sana, con menos estrés, mejor alimentación, menos contaminación, etc.

Sin embargo, al observar, en el mismo estudio, los lugares con mayor incidencia del cáncer de riñón, la localización geográfica era **la misma**. Si no se ha leído el párrafo anterior, uno podría plantearse que en el medio rural hay peor alimentación (rica en solo ciertos contenidos y pobre en otros), peor acceso a la sanidad, etc.

El gráfico 7.7 muestra los resultados citados: en naranja, los condados con porcentajes más altos de cáncer de riñón; y, en verde, aquellos con porcentajes más bajos. En general, unos están pegados a los otros.

La explicación la ofreció el estadístico americano Howard Wainer en el artículo “The most dangerous equation” (Wainer, 2007) (de donde procede el gráfico): las llamativas diferencias en la incidencia del cáncer entre condados obedecen al azar.

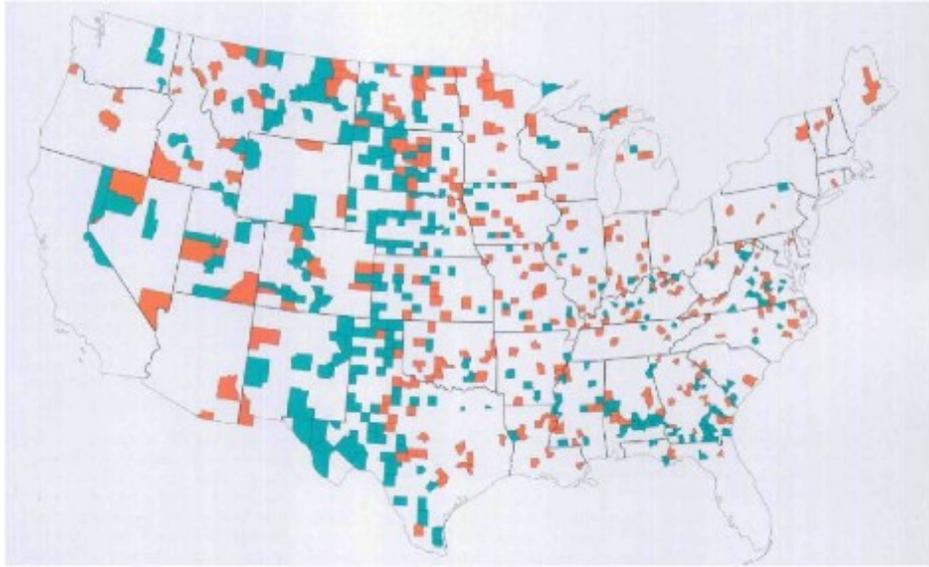


Figura 7.7: Tasas de cáncer de riñón en USA.

En efecto, si en Estados Unidos se dan, en promedio, 5 casos de cáncer de riñón por cada 100.000 habitantes, la “ley de los grandes números” hará que la incidencia de la enfermedad esté muy próxima a ese valor en los condados muy poblados (como en Los Angeles). Pero, en los que tengan poca población, ese promedio oscilará mucho, pues la variabilidad (desviación estándar) de la media aritmética de la variable analizada -en nuestro caso, el número de casos de cáncer de riñón por habitante- guarda una relación inversa con el tamaño de la muestra usada para calcularla (o, para ser exactos, de su raíz cuadrada).

Por eso, según Wainer, la gran diferencia estadística entre condados en la incidencia de la enfermedad no es un “hecho” (fact) genuino, sino un “artefacto” (arti-fact), es decir un resultado artificial nacido de la interacción entre el azar y el tamaño de las “muestras” utilizadas para calcular la incidencia media.

Así, cuando el tamaño de las muestras ( $n$ ) es muy grande -como ocurre en condados con mucha población-, la variabilidad de los valores medios que salgan será muy pequeña: tales valores medios nos saldrán muy parecidos al valor medio de la variable en el conjunto de la población (en nuestro ejemplo, el 5 por 100.000 habitantes con cáncer de riñón que se da en Estados Unidos,

considerado como un todo).

Wainer explica que la gran variabilidad de los valores medios en muestras pequeñas se da también en las notas medias de los alumnos de una escuela o clase: cuanto más pequeña sea la escuela o clase, más frecuente será que la nota media de sus alumnos tome valores extremos, altos o bajos. Así, que muchas de las mejores notas medias se alcancen en escuelas pequeñas no obedecerá necesariamente a que sean mejores, sino al puro azar.

Wainer señala que la Fundación Gates ignoró esa relación cuando en los años 90 dedicó muchos millones de dólares a fomentar las escuelas pequeñas, a la vista de que un elevado porcentaje de las escuelas con mejores resultados académicos tenían pocos alumnos. Pero abandonaron la idea cuando advirtieron que ese fenómeno estadístico podía obedecer al azar: también las escuelas pequeñas estaban sobre-representadas entre las que obtenían los peores resultados. O sea, las escuelas pequeñas, en efecto, contaban con los mejores alumnos, pero también contaban con los peores.

Esta argumentación sirve para responder al siguiente problema, planteado en [naukas](#) por Pablo Rodríguez en 2014:

“Un acertijo: bebés y probabilidad.

En una ciudad hay dos hospitales. Uno de ellos es mucho más grande que el otro, y por lo tanto tiene capacidad para más pacientes. El mes pasado sucedió una cosa curiosa. La sala de maternidad de uno de los hospitales registró que la proporción de bebés niña había sido ese mes muy superior a la de bebés niño (pongamos 75% de niñas, 25% de niños). ¿En cuál de los dos hospitales es más probable que sucediese esta rareza, en el hospital grande o en el hospital pequeño?

Para los más exigentes en cuánto a detalles, dejo las siguientes aclaraciones: Consideramos que las probabilidades en cada parto son 50% niño 50% niña. No tiene importancia sobre el resultado final, pero pongamos que ese mes no hubo partos múltiples de gemelos o mellizos. La única diferencia relevante entre el hospital grande y el pequeño es que el hospital grande atendió a muchas más madres que el pequeño.

En el propio blog de Pablo Rodríguez da tres explicaciones, que resumo (ir al original para ampliar las ideas):

- Explicación intuitiva:

En donde la muestra es más pequeña (hospital pequeño), es más fácil que una rareza (en el sentido probabilístico, algo de poca probabilidad) destaque; puesto que, en donde la muestra es más grande (hospital grande), las frecuencias van a compensarse. De hecho, la ley de los grandes números nos recuerda que la frecuencia relativa de ocurrencia de un suceso tenderá a su probabilidad cuando el número de repeticiones sea grande.

- Explicación con teoría:

El nacimiento de un bebé se puede modelar como una variable de Bernoulli: dos posibles sucesos, complementarios uno del otro. Los distintos nacimientos en un hospital van a constituir, por lo tanto, una variable Binomial. Llamemos, por ejemplo,  $X$ =*número de niñas nacidas en un hospital, tras  $n$  partos*. Esta variable es una variable Binomial de parámetros  $n$  y  $p = 0.5$ .

Podemos calcular la probabilidad de que el número de niñas nacidas sea mayor que el 75% del total como  $P(X > 0.75 \cdot n)$ . Cambiando el valor de  $n$  podemos ir viendo los resultados. En R, la probabilidad  $P(X \leq c)$  en una binomial  $Bi(n, p)$  se calcula con `pbinom(c, n, p)`; por lo tanto,  $P(X > 0.75 \cdot n)$  será  $1 - P(X \leq 0.75 \cdot n)$ .

Para  $n = 10$ ,

```
n=10
c=0.75*n
p=0.5
1-pbinom(c, n, p)
```

```
## [1] 0.0546875
```

Vemos que, para 10 camas, solo ocurriría esto en un 5.46% de los casos.

Para  $n = 30$ ,

```
n=30
c=0.75*n
p=0.5
1-pbinom(c, n, p)
```

## [1] 0.00261

esto solo ocurriría en un 0.26% de los casos.

Estas (y algunas otras) curiosidades estadísticas también han sido analizadas por Manuel Conthe (licenciado en Derecho, Economista del Estado y ex-presidente de la Comisión Nacional del Mercado de Valores, entre otros méritos) en [expansion](#). Son destacables también los siguientes párrafos:

“Pero creo que también acertó el presidente Rajoy cuando en un acto público en Nueva York manifestó: Permítanme que haga un reconocimiento a la mayoría de españoles que no se manifiestan, que no salen en las portadas de la prensa y que no abren los telediarios. No se les ve, pero están ahí”.

“La prensa es como el rayo de una linterna que se mueve sin cesar y saca de la oscuridad un suceso tras otro. Las noticias y la verdad no son lo mismo, y deben distinguirse”.

“Nuestra mente, por desgracia, no sólo tiende a identificarlas, sino que, como enseña Kahneman, tiende a deducir verdaderas generales de artefactos nacidos del azar. Por eso, cuando los medios eligen sus encuadres no iluminan: construyen su realidad.”

### 7.2.4 Propiedades de la cuasi-varianza (muestral)

A estas alturas ya no nos acordamos de lo que es. Refresquemos:

La cuasi-varianza muestral  $S_{n-1}^2 = \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1}$ :

- Es un estimador de la varianza poblacional  $\sigma^2$ .
- Es **insesgado** (el motivo de dividir por  $n - 1$  es que, de esta forma, el estimador cumple esta propiedad. Si se divide por  $n$ , esta propiedad no se verifica).
- Es consistente.
- Además,  $S_{n-1}^2 = \frac{n}{n-1} S^2$  (la relación entre la varianza muestral y la cuasi-varianza es muy simple).

### 7.2.5 Propiedades de la proporción muestral

La proporción muestral  $\hat{p} = \frac{X_1 + \cdots + X_n}{n}$ :

- Es un estimador de la proporción poblacional  $p$ .
- Es insesgado.
- Es consistente.
- Para  $n$  grande ( $n > 30$ ), por el Teorema Central del Límite, se tiene que  $\hat{p} \approx N(p, \sqrt{p(1-p)/n})$

Ej.

En una clínica de fisioterapia desean saber el número medio de grados que puede doblar una rodilla destrozada, tras un tratamiento de diez sesiones de rehabilitación (el seguro no paga más) con un método nuevo recién salido al mercado. Se seleccionan diez pacientes y, tras dos semanas de tratamiento, se mide el número de grados que dobla la rodilla hasta que los gritos son inaguantables:

41.60, 41.48, 42.34, 41.95, 41.86, 42.41, 41.72, 42.26, 41.81, 42.04.

Para estimar la media poblacional (parámetro desconocido en este caso) podemos considerar la media muestral:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{10} x_i = \frac{419.47}{10} = 41.947$$

Si nos pidiesen estimar una proporción; por ejemplo, proporción de pacientes que doblarán más de 42 grados, ( $p = P(X > 42)$ ) podríamos utilizar la proporción muestral:

$$\hat{p} = \frac{1}{n} \sum_{i/x_i > 42} 1 = \frac{4}{10} = 0.4$$

## 7.3 Ejemplos de interés.

### 7.3.1 El problema de los tanques alemanes

Recomiendo la lectura del trabajo de (Luque, 2013) sobre este famoso ejemplo.

La tabla siguiente está recogida en el artículo de (Ruggles and Brodie, 1947). En ella se ven las estimaciones sobre el número de tanques que poseía el ejército alemán (II guerra mundial) en tres fechas distintas. Las estimaciones fueron realizadas por el departamento de estadística y por el servicio de inteligencia militar y, como vemos, son muy diferentes. En la columna de la derecha está anotado el número verdadero. Como podemos comprobar, los estadísticos lograron acercarse a la cifra real de tanques mucho más que los servicios de inteligencia.

Fecha	Est. estadística	Est. Militares	Valor real
<b>Junio de 1940</b>	169	1000	122
<b>Junio de 1941</b>	244	1550	271
<b>Agosto de 1942</b>	327	1550	342



Figura 7.8: Tanque Panzer alemán, de la II guerra mundial.

La estimación de estas cifras se corresponde con el resultado del llamado *problema de estimación del número de taxis de una ciudad*: si en una ciudad hay  $N$  taxis en servicio. Si una persona acaba de llegar a la ciudad, ¿cómo puede obtener un número aproximado a  $N$ , es decir, una estimación?

Habitualmente, los taxis llevan el número en la puerta. A partir de la observación de una muestra de  $n$  taxis, ¿cómo podría estimarse el total ( $N$ )? El problema de los tanques alemanes en la segunda guerra mundial se planteó

de la misma forma, al darse cuenta los aliados de que los tanques alemanes recogidos tras una batalla tenían números de serie consecutivos.

Supongamos que el número de taxis en una ciudad es  $N = 50$ , y que se observa una muestra de  $n$  taxis. Una estimación lógica podría ser, si conocemos el número  $m$  medio, considerar que hay  $(m - 1)$  taxis a cada lado, es decir,  $\hat{N} = (m - 1) + 1 + (m - 1) = 2m - 1$ .

Pongamos un caso práctico: observamos 5 taxis con los números 2, 10, 15, 25, 40. Una estimación de la media podría ser la media muestral o la mediana muestral. La media muestral es  $\bar{x} = 18.4$ , y la mediana muestral es 15. Las estimaciones de  $N$  son, por lo tanto,  $\bar{N} = 2 \cdot 18.4 - 1 = 36.8$  (36 o 37 taxis) o  $\bar{N} = 2 \cdot 15 - 1 = 29$ .

Claramente, el número obtenido infraestima el valor verdadero de 50. Además, está claro que cualquiera de los dos números obtenidos no serían buenos estimadores del tamaño real, puesto que 40 es un dato de la muestra, es decir hay 40 o más taxis seguro.

Una posibilidad podría ser considerar un estimador de la siguiente manera:  $\hat{N}_1 = \max\{x_1, x_2, \dots, x_n\}$ .

Otra posibilidad sería considerar que hay, por simetría, tantos datos a la derecha de  $x_n$  como antes de  $x_1$ , es decir  $N - x_n = x_1 - 1$ , de donde obtenemos

$$\hat{N}_2 = x_n - x_1 + 1$$

En nuestro ejemplo sale  $\hat{N}_2 = 40 - 2 + 1 = 39$ .

Otra aproximación es considerar que el número de taxis a la derecha de  $x_n$  es aproximadamente igual al promedio de las distancias entre cada dos valores de la muestra:

$$N - x_n = \frac{(x_1 - 1) + (x_2 - x_1 - 1) + \dots + (x_n - x_{n-1} - 1)}{n} = \frac{x_n}{n} - 1$$

Así tenemos otro estimador:

$$\hat{N}_3 = x_n + \frac{x_n}{n} - 1$$

En nuestro ejemplo sale  $\hat{N}_3 = 47$ , y es de las mejores opciones que pueden darse, junto con una estimación de tipo bayesiano (no es posible dar una

descripción intuitiva de este):

$$\hat{N} = (x_n - 1) \cdot \frac{n - 1}{n - 2} \pm \sqrt{\frac{(n - 1) \cdot (x_n - 1) \cdot (x_n - n + 1)}{(n - 3) \cdot (n - 2)^2}}$$

## 7.4 Intervalos de confianza

La estimación puntual aproxima mediante un número el valor de una característica poblacional o parámetro desconocido (la altura media de los españoles, la intención de voto a un partido en las próximas elecciones generales, el tiempo medio de ejecución de un algoritmo, el número de taxis...) pero no nos indica el error que se comete en dicha estimación.

Lo razonable, en la práctica, es adjuntar, junto a la estimación puntual del parámetro, un intervalo que mida el margen de error de la estimación. La construcción de dicho intervalo es el objetivo de la **estimación por intervalos de confianza**.

Un **intervalo de confianza** para un parámetro con un **nivel de confianza**  $1 - \alpha$  ( $0 < \alpha < 1$ ), es un intervalo de extremos aleatorios  $(L, U)$  que, con probabilidad  $1 - \alpha$ , contiene al parámetro en cuestión.

$$P(\text{parámetro} \in (L, U)) = 1 - \alpha.$$

Los valores más habituales del nivel de confianza  $1 - \alpha$  son 0.9, 0.95 o 0.99 (la confianza es del 90%, 95% o 99%). En ocasiones también se emplea la terminología **nivel de significación** para el valor  $\alpha$ .

En la estimación por intervalos de confianza partimos de una muestra  $x_1, \dots, x_n$ . A partir de estos valores obtenemos un intervalo numérico. Por ejemplo, podríamos hablar de que, con una confianza del 99 por ciento, la proporción de voto al partido político “Unidas Ciudadanas” está entre el 29 y el 31 por ciento. O que, con una confianza del 90 por ciento, la estatura media está entre 1.80 y 1.84.

### 7.4.1 Interpretación

Igual que vimos antes con las encuestas de las estaturas, o de la proporción de gente que cree en los extraterrestres, con cada muestra obteníamos  $n$

datos diferentes, y valores diferentes (de la media muestral o de la proporción muestral).

De cada muestra también puede obtenerse un intervalo de confianza. Entonces, con cada muestra diferente, obtendremos un intervalo también diferente. A medida que aumenta la cantidad de intervalos que hemos construido, el porcentaje de intervalos que contienen el verdadero valor del parámetro se aproximará al  $100(1 - \alpha)\%$ .

Así, por ejemplo, un intervalo de confianza al 95% garantiza que, si tomamos 100 muestras, el verdadero valor del parámetro estará dentro del intervalo en **aproximadamente** el 95 de los intervalos construidos.

Veamos un ejemplo mediante simulación. Vamos a simular que realizamos encuestas, en este caso preguntando a  $n = 12$  personas (de nuevo si creen en los extraterrestres). De cada muestra podemos obtener una estimación puntual (calculada mediante la proporción en la muestra), y también un intervalo de confianza (que más adelante veremos cómo se calcula). Haremos este proceso  $B = 20$  veces.

```
set.seed(1)
n<12
B<20
muestras <-as.data.frame(matrix(rbinom(n*B,
                                     1, 0.6), ncol=n))
rownames(muestras) <-paste("muestra", 1:B, sep="")
colnames(muestras) <-paste(" obs ", 1:n , sep="")
#muestras
pander(muestras)
```

Tabla 7.4: Table continues below

	obs 1	obs 2	obs 3	obs 4	obs 5	obs 6	obs 7	obs 8
<b>muestra1</b>	1	0	0	0	1	0	0	0
<b>muestra2</b>	1	1	0	1	0	1	1	1
<b>muestra3</b>	1	0	0	1	1	1	1	1
<b>muestra4</b>	0	1	1	1	1	0	1	1
<b>muestra5</b>	1	1	1	0	0	0	0	0
<b>muestra6</b>	0	1	0	1	1	1	1	1

	obs 1	obs 2	obs 3	obs 4	obs 5	obs 6	obs 7	obs 8
<b>muestra7</b>	0	1	1	1	0	1	1	1
<b>muestra8</b>	0	1	1	0	1	1	1	0
<b>muestra9</b>	0	0	0	1	1	0	1	1
<b>muestra10</b>	1	1	0	0	1	1	1	0
<b>muestra11</b>	1	1	1	1	1	0	1	0
<b>muestra12</b>	1	1	0	0	1	0	1	1
<b>muestra13</b>	0	1	1	1	0	1	1	1
<b>muestra14</b>	1	1	1	1	0	1	0	1
<b>muestra15</b>	0	0	1	1	0	1	0	1
<b>muestra16</b>	1	0	1	0	0	1	1	1
<b>muestra17</b>	0	0	1	0	1	0	1	1
<b>muestra18</b>	0	1	1	1	1	1	1	1
<b>muestra19</b>	1	0	0	0	0	1	0	1
<b>muestra20</b>	0	1	1	0	1	0	1	1

	obs 9	obs 10	obs 11	obs 12
<b>muestra1</b>	1	1	1	1
<b>muestra2</b>	0	1	1	1
<b>muestra3</b>	1	0	1	1
<b>muestra4</b>	0	1	1	1
<b>muestra5</b>	0	0	1	0
<b>muestra6</b>	1	1	1	1
<b>muestra7</b>	1	0	1	1
<b>muestra8</b>	1	0	1	1
<b>muestra9</b>	0	0	1	1
<b>muestra10</b>	1	1	0	0
<b>muestra11</b>	0	0	0	1
<b>muestra12</b>	0	1	1	1
<b>muestra13</b>	0	1	0	1
<b>muestra14</b>	1	0	0	0
<b>muestra15</b>	1	1	0	1
<b>muestra16</b>	0	1	1	0
<b>muestra17</b>	0	1	0	0
<b>muestra18</b>	0	0	0	1
<b>muestra19</b>	1	1	0	1

	obs 9	obs 10	obs 11	obs 12
<b>muestra20</b>	0	0	1	1

```

muestras$mean <-rowMeans(muestras[,1:n])

alpha <-0.05
z <-qnorm(1 - alpha/2)
p <muestras$mean
muestras$ici <-p - z*sqrt(p*(1-p)/n)
muestras$ics <-p + z*sqrt(p*(1-p)/n)
esta<theta
b=ifelse(0.6<=muestras$ics & 0.6 >= muestras$ici,
          "SI", "NO")
Resultado<data.frame(muestras$mean, muestras$ici ,
                     muestras$ics, b)
colnames(Resultado)<e("ptilde", "L", "U",
                     " 0.6 está o no?")
pander(Resultado)

```

ptilde	L	U	0.61 está o no?
0.5	0.2171	0.7829	SI
0.75	0.505	0.995	SI
0.75	0.505	0.995	SI
0.75	0.505	0.995	SI
0.3333	0.06662	0.6001	NO
0.8333	0.6225	1.044	NO
0.75	0.505	0.995	SI
0.6667	0.3999	0.9334	SI
0.5	0.2171	0.7829	SI
0.5833	0.3044	0.8623	SI
0.5833	0.3044	0.8623	SI
0.6667	0.3999	0.9334	SI
0.6667	0.3999	0.9334	SI
0.5833	0.3044	0.8623	SI
0.5833	0.3044	0.8623	SI

ptilde	L	U	0.61 está o no?
0.5833	0.3044	0.8623	SI
0.4167	0.1377	0.6956	SI
0.6667	0.3999	0.9334	SI
0.5	0.2171	0.7829	SI
0.5833	0.3044	0.8623	SI

Vemos que, en la mayoría de las ocasiones, el intervalo de confianza contiene al verdadero valor del parámetro (0.61 en este caso), pero en otras (las menos) no. En este ejemplo hemos utilizado un tamaño de muestra pequeño y hemos repetido el proceso pocas veces, pero ahí está el código para replicar el procedimiento con otros valores. Se puede comprobar que, con tamaños de muestra un poco más grandes y repetido el proceso un número grande de veces, 95 de cada 100 veces (aquí hemos usado un nivel de confianza del 95%) los intervalos contendrán el valor 0.61, y aproximadamente 5 de cada 100 veces no lo contendrán.

Observemos el mismo proceso, ahora para 500 muestras de tamaño 10, de estaturas (de hombres mayores de 18 años en España). Simulamos que seleccionamos aleatoriamente a 10 hombres, les medimos la estatura, calculamos la estatura media de la muestra y el intervalo de confianza para dicha estatura media, con confianza del 95%. Realizamos este proceso 500 veces.

Ahora, además, hacemos un dibujo de los 50 primeros intervalos, para comprobar cuáles contienen al verdadero valor buscado (la estatura media real o teórica, en este caso, que es 177.7).

```
set.seed(54321)
nsim <-500
nx <-10

# Valores teóricos:
mux <-177.7
sdx <-5.9
#
# Simulación de las muestras
muestras <-as.data.frame(matrix(rnorm(nsim*nx,
```

```

      mean=mux, sd=sdx), ncol=nx))
rownames(muestras) <-paste("muestra", 1:nsim, sep="")
colnames(muestras) <-paste("obs", 1:nx, sep="")

# Estimaciones
muestras$mean <-rowMeans(muestras[,1:nx])
muestras$sd <-apply(muestras[,1:nx], 1, sd)

alfa <-0.05
z <-qnorm(1 - alfa/2)
muestras$ici <-muestras$mean - z*sdx/sqrt(nx)
muestras$ics <-muestras$mean + z*sdx/sqrt(nx)

# Cobertura de las estimaciones por IC:
muestras$cob <-(muestras$ici < mux) & (mux < muestras$ics)
ncob <-sum(muestras$cob)

# N° de intervalos que contienen la verdadera media
ncob

```

```
[1] 480
```

```

# Proporción de intervalos
100*ncob/nsim

```

```
[1] 96
```

Como vemos, de los intervalos calculados, el 96 por ciento de los mismos contiene el verdadero valor 177.7. Este valor es casi el 95%, que es el nivel utilizado.

Hacemos ahora el gráfico de los primeros 50 intervalos (Figura 7.9:). Los intervalos de color rojo no contienen al verdadero valor del parámetro; los otros sí.

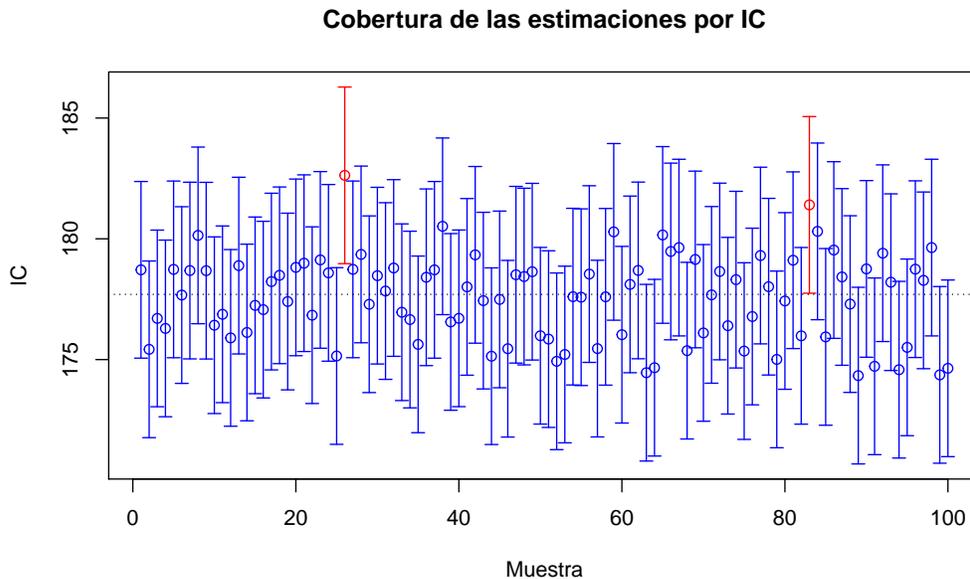


Figura 7.9: Fijémonos que hay intervalos que no contienen al parámetro verdadero.

```

m <-100
tmp <-muestras[1:m,]
attach(tmp)
color <-ifelse(cob,"blue","red")
plot(1:m, mean, col = color,
     ylim = c(min(ici),max(ics)),
     main = "Cobertura de las estimaciones por IC",
     xlab = "Muestra", ylab = "IC")
arrows(1:m, ici, 1:m, ics, angle = 90,
       length = 0.05, code = 3, col = color)
abline(h = mux, lty = 3)

```

```
detach(tmp)
```

### 7.4.2 Ejemplo: Meta-análisis de eficacia de Antidepresivos

Se llama meta-análisis a un procedimiento basado en técnicas estadísticas, mediante el que se analizan datos de diferentes estudios realizados sobre el mismo tema (revisión sistemática). El esquema básico consiste en elegir un grupo de artículos científicos en donde se haya medido el mismo parámetro (efecto medio, proporción... ) y estudiar la variabilidad existente entre unos resultados y otros.

En el gráfico 7.10: se considera un meta-análisis del efecto de los antidepresivos más comunes, siendo el parámetro considerado el llamado *oddsratio* entre la efectividad de cada antidepresivo contra placebo.

Si  $p_1$  es la probabilidad de que el antidepresivo sea efectivo, y  $p_2$  la del placebo, el odds-ratio -“razón de momios” o “razón de posibilidades”- es:

$$\frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}$$

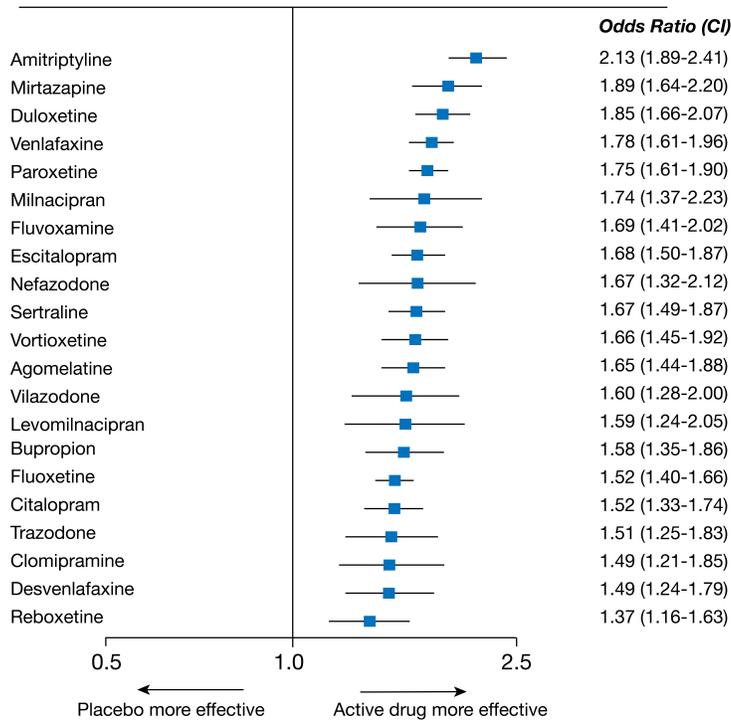
De manera general, el odds-ratio entre dos sucesos expresa las posibilidades de uno frente al otro. Si, por ejemplo, las odds de que llueva son de dos a uno,  $(2/1) = 2$ , quiere decir que es dos veces más probable que llueva que no llueva; con lo cual, la probabilidad de lluvia sería  $2/3$  y de no lluvia  $1/3$ . Si, de un grupo de 100 personas, 85 presentan síntomas de una enfermedad, la probabilidad de enfermedad será  $85/100$  mientras que las odds serán de 85 a 15, es decir 5.7. Las odds exceden de la unidad siempre que haya más posibilidades del suceso de “arriba” (numerador) frente al de “abajo” (denominador). Con respecto al intervalo de confianza, es importante ver si incluye o no el valor 1. Si lo incluye, significa que la asociación no es estadísticamente significativa, y que los resultados sólo pueden deberse a la casualidad. Si no lo incluye, indicaría mayor probabilidad de ocurrencia de un suceso frente al otro. El 1 equivale a la misma probabilidad (llover o no llover, o, en este caso, efecto de un antidepresivo frente al efecto de un placebo).

Como observamos en el gráfico 7.10, ninguno de los intervalos de confianza para ninguno de los antidepresivos considerados contiene al 1 (lo que significaría que su eficacia sería equivalente a un placebo). En cambio, todos los intervalos de confianza están situados a la derecha, lo que significa que, con

## Are antidepressant drugs effective?



The effectiveness (here measured as the efficacy) of antidepressant drugs measured relative to placebos from meta-analysis of 522 trials comprising 116,477 participants. Effectiveness was measured as the odds-ratio (OR) of a  $\geq 50\%$  reduction of the total score on a standardised observer-rating scale for depression, relative to a placebo. Values greater than 1.0 indicate the antidepressant drug was more effective than the placebo. All results were statistically significant.



Source: Cipriani et al. (2019). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder. The data visualization is available at [OurWorldinData.org](https://ourworldindata.org). There you will find more on this topic. Licensed under CC-BY-SA by the authors Hannah Ritchie & Max Roser.

Figura 7.10: Gráfico llamado Forest-Plot del meta-análisis para los medicamentos antidepresivos.

un 95% de confianza (este nivel es el más utilizado), el efecto de un antidepresivo es siempre mayor (y tanto más cuanto más grande el odds-ratio, o la raya este más a la derecha). En este meta-análisis, se observó que el antidepresivo de mayor eficacia (comparándolo con un placebo) es la Amitriptilina.

Fuente: <https://ourworldindata.org/happiness-and-life-satisfaction>

## 7.5 I.C. para la media (1)

Nos centramos en la estimación de la media  $\mu$  de una población o variable **Normal** (altura media, peso medio, tiempo medio haciendo gimnasia...) . Inicialmente, consideramos que la desviación típica de la variable es **conocida** (es un número que sabemos).

Por una vez, y sin que sirva de precedente, vamos a ver cómo es la construcción matemática del intervalo de confianza. Consideremos la variable  $X \in N(\mu, \sigma)$ , que representa a la característica que estamos midiendo (altura, peso...). Supongamos que  $\sigma$  es conocida.

Consideramos una muestra aleatoria simple  $X_1, \dots, X_n$  de la variable  $X$ . Dado el nivel de confianza  $1 - \alpha$ , elegimos el llamado **estadístico pivote**

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Un **estadístico** es una función de variables aleatorias y es también otra variable aleatoria. En este caso, vamos a ver que distribución sigue esta variable  $T$  que acabamos de definir (el término pivote es una nomenclatura utilizada en los test de hipótesis).

Como vimos anteriormente, la media muestral verifica

$$\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Por lo tanto, si tipificamos la variable (restamos la media y dividimos por la desviación típica), obtenemos la variable  $T$ , lo que quiere decir que esta variable sigue una distribución normal estándar ( $N(0, 1)$ ).

Teniendo en cuenta que  $\frac{\alpha}{2} = P(Z \geq z_{\alpha/2})$  (Figura 7.11), sabemos que

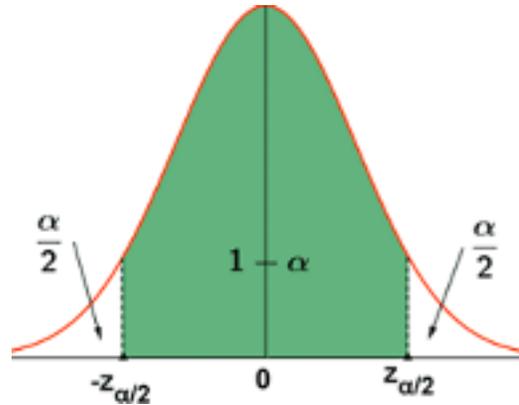


Figura 7.11: Niveles de significación en una normal estandarizada

$$1 - \alpha = P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right)$$

Despejando el parámetro  $\mu$  obtenemos

$$1 - \alpha = P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

Por tanto, el I.C. para  $\mu$  al nivel de confianza  $1 - \alpha$  es

$$(L, U) = \left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

El procedimiento teórico para llegar a esta fórmula es simple, aunque difícil de seguir para cualquiera con pocos conocimientos matemáticos. En todo caso, lo importante es que la fórmula del intervalo no tiene excesiva dificultad. El intervalo está centrado en el estimador media muestral, y los extremos consisten en restar y sumar la misma cantidad: un valor que depende del nivel de confianza utilizado, multiplicado por el error muestral de la media.

**Ej.**

\*Retomamos el Ejercicio anterior.\*\*

En una clínica de fisioterapia se quiere saber el número de grados que acaba doblando una rodilla después de dos semanas de tratamiento.

Las medidas de 10 pacientes fueron

41.60, 41.48, 42.34, 41.95, 41.86, 42.41, 41.72, 42.26, 41.81, 42.04.

Aceptando que la variable aleatoria  $X$  = “grados que dobla la rodilla” sigue una distribución normal, y suponiendo que  $\sigma = 0.30$  grados,

- Obtener un intervalo de confianza para la temperatura media al nivel del 90%.
- Deduce el tamaño muestral necesario para conseguir un intervalo de confianza al 99%, con un error menor o igual que 0.05.

### Solución

- Sabemos que  $\sigma = 0.3$  y  $n = 10$

La media muestral es  $\bar{x} = \frac{1}{n} \sum_{i=1}^{10} x_i = \frac{419.47}{10} = 41.947$

El I.C. para  $\mu$  al nivel de confianza  $1 - \alpha$  es:

$$\left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = \left( 41.947 \pm z_{\alpha/2} \frac{0.3}{\sqrt{10}} \right)$$

donde el valor  $z_{\alpha/2} = 1.645$  se puede obtener como

```
qnorm(0.1/2)
```

```
## [1] 1.65
```

es decir, calculamos el cuantil de una normal (por defecto, los parámetros 0 y 1 no hace falta escribirlos) mediante `qnorm`.

El I.C. para  $\mu$  al 95% es, entonces:

$$\left( 41.947 \pm 1.96 \cdot \frac{0.3}{\sqrt{10}} \right) = (41.947 \pm 0.186) = (41.761, 42.133).$$

b. Escribimos de nuevo la fórmula del intervalo de confianza:

$$\left( \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),$$

para ver que, con una probabilidad  $1 - \alpha$  el parámetro verdadero ( $\mu$ ) está dentro de ese intervalo; es decir, que la distancia entre  $\mu$  y  $\bar{x}$  es, como mucho,  $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ .

Esto es, el error de estimación está acotado:

$$|\bar{x} - \mu| \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Si queremos calcular el tamaño muestral necesario para que el error sea menor o igual a una cantidad  $e$  (0.05 en este caso), hacemos

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq e \iff n \geq \frac{z_{\alpha/2}^2 \cdot \sigma^2}{e^2} = \left( \frac{1.96 \cdot 0.3}{0.05} \right)^2 = 138.298.$$

Hay que tomar entonces  $n = 139$  mediciones.

Fijémonos en que, si quisiésemos -con la misma confianza- obtener un error la mitad de pequeño ( $e/2$ ), la fórmula que obtenemos es

$$n \geq \frac{z_{\alpha/2}^2 \cdot \sigma^2}{(e/2)^2} = 4 \times \frac{z_{\alpha/2}^2 \cdot \sigma^2}{e^2},$$

es decir, habría que tomar una muestra 4 veces más grande.



Regla de la raíz de  $n$ : “si quieres multiplicar la exactitud de una investigación, no basta con duplicar el esfuerzo, debes multiplicarlo por 4”.

‘S. Stigler (Stigler, 1986)’

## 7.6 I.C. para la media (2)

En la práctica, no es habitual conocer la desviación típica, así que esta debe estimarse a partir de la muestra, igual que se estima la media. El intervalo

de confianza para la media de una variable aleatoria normal, con desviación típica desconocida, tiene la siguiente forma:

$$\left( \bar{x} \pm t_{n-1, \alpha/2} \frac{\hat{S}_{n-1}}{\sqrt{n}} \right),$$

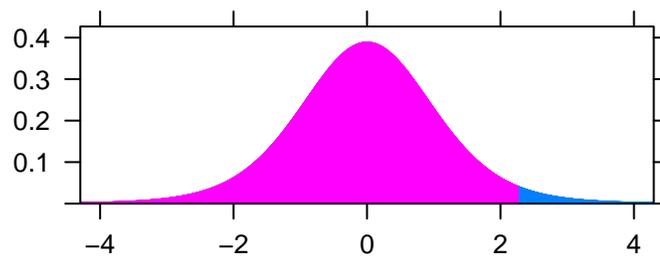
siendo  $t_{n-1, \alpha/2}$  el valor de una  $t$  de Student con  $n - 1$  grados de libertad que deja a la derecha  $\alpha/2$  de área (mismo significado que en el caso anterior, pero debemos buscar dicho valor en la densidad  $t$  con  $n - 1$  grados de libertad).  $\hat{S}_{n-1}$  es la cuasi-desviación típica muestral, es decir la raíz cuadrada de la cuasi-varianza muestral.

Por ejemplo, supongamos  $\alpha = 0.05$  y el tamaño de muestra es 10. El valor del que hablamos es

```
qt(0.05/2, df=9)
```

```
## [1] 2.26
```

```
library(ggplot2)
library(mosaic)
plotDist("t", df= 9, groups = x < 2.26 type = "h")
```



Con los datos del ejemplo anterior, vamos a calcular rápidamente el intervalo de confianza. En R, debemos utilizar la orden `t.test` que, en realidad, es para hacer un contraste de hipótesis. Sin embargo, a mayores, obtenemos el intervalo de confianza.

```
x=c(41.0, 41.48 42.34, 41.95, 41.6
42.41, 41.72, 42.26 41.8, 42.04)
y=t.test(x, mu=0, conflevel=0.9)
y$confint
```

```
## [1] 41.7642.13
## attr("conflevel")
## [1] 0.9
```

El intervalo es (41.76, 42.12). Lo hemos calculado para un 90 (*conflevel*) por ciento de confianza.

Si lo calculamos para un 95 y un 99 por ciento de confianza, respectivamente, obtenemos:

```
x=c(41.0, 41.48 42.34, 41.95, 41.6
42.41, 41.72, 42.26 41.8, 42.04)
y=t.test(x, mu=0, conflevel=0.95)
y$confint
```

```
## [1] 41.72 42.17
## attr("conflevel")
## [1] 0.95
```

```
y=t.test(x, mu=0, conflevel=0.99)
y$confint
```

```
## [1] 41.62 42.27
## attr("conflevel")
## [1] 0.99
```

Vemos que los intervalos son (41.72, 42.17) (al 95 por ciento), que es un poco más largo que el anterior; y, a la vez, el intervalo al 99 por ciento es (41.62, 42.27) que vuelve a ser más grande.

Como parece lógico, si se desea una mayor confianza de que el parámetro buscado esté dentro del intervalo, el intervalo va a salir más grande. La única manera de obtener intervalos más pequeños sería aumentar el tamaño

muestral (recordemos que la forma del intervalo es  $(\bar{x} \pm t_{n-1, \alpha/2} \frac{\hat{S}_{n-1}}{\sqrt{n}})$ , donde la longitud es inversamente proporcional a  $\sqrt{n}$ ).

Como hemos dicho anteriormente, en la práctica, si de una población no se conoce la media (y por eso intentamos estimarla), es raro conocer la desviación típica. Por ello el intervalo de confianza para la media que calculan, en general, los paquetes estadísticos (como R) es este último.

## 7.7 I.C. para una proporción

El objetivo es construir un intervalo de confianza para la proporción de elementos ( $p$ ) de una población que poseen una determinada característica (votantes de un partido político, alumnos que usan una determinada red social, elementos defectuosos...) a partir de una muestra aleatoria simple de la población.

De esta forma, consideramos la variable

$X$	$P(X = x_i)$
0	$1 - p$
1	$p$

Es decir, la variable aleatoria que toma los valores 1 y 0 (1 si tiene la característica, con probabilidad  $p$ , 0 si no la tiene)

Tomamos entonces una muestra aleatoria simple  $X_1, \dots, X_n$  de la variable  $X \in B(1, p)$

Dado el nivel de confianza  $1 - \alpha$ ,

El intervalo es:

$$\left( \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right).$$

Para calcular este intervalo de confianza con R, debemos utilizar la siguiente función:

```

z.test <-function(x,n,p=NULL,conflevel=0.95,
                 alternative="two.sided") {
  ts.z <-NULL
  cint <-NULL
  p.val <-NULL
  phat <-x/n
  qhat <-1 - phat
  if(length(p) > 0) {
    q <-1-p
    SE.phat <-sqrt((p*q)/n)
    ts.z <- (phat - p)/SE.phat
    if(ts.z<0) p.val <-2*pnorm(ts.z)
    else p.val<2*(1-pnorm(ts.z))
    if(alternative=="less") {
      p.val <-pnorm(ts.z)
    }
    if(alternative=="greater") {
      p.val <-1 - pnorm(ts.z)
    }
  } else {
    SE.phat <-sqrt((phat*qhat)/n)
  }
  cint <-phat + c(
    -1*((qnorm(((1 - conflevel)/2) +
              conflevel))*SE.phat),
    ((qnorm(((1 - conflevel)/2) +
              conflevel))*SE.phat) )
  return(list(estimate=phat,ts.z=ts.z,
             p.val=p.val,cint=cint))
}

```

Esta función no solo calcula el intervalo de confianza para una proporción, sino que realiza un test de hipótesis sobre dicho parámetro, por lo que la utilizaremos en el capítulo siguiente.

Ej.

Al finalizar un ensayo clínico para un trabajo fin de master, la profesora Chifuentes dispuso de una m.a.s. de 100 pacientes, de las cuales 7 dijeron no haber notado mejoría. Determinar un intervalo de confianza (con un nivel de confianza aproximado del 99%) para  $p$ , proporción de pacientes que notan mejoría en la población en general.

¿Qué tamaño muestral  $n$  recomendarías para que la longitud del intervalo de confianza (con un nivel de confianza aproximado del 95%) para la proporción  $p$  fuera inferior a 0.01 unidades?

**Solución:**

Tenemos que  $n = 100$  pacientes y que la proporción muestral es  $\hat{p} = \frac{93}{100} = 0.93$

El intervalo de confianza es:

$$\left( 0.93 \pm 2.575 \sqrt{\frac{0.93 \times 0.07}{100}} \right) = (0.93 \pm 0.0657) = (0.8643, 0.9957)$$

Calculado con la función de R:

```
y=z.test(93,100, conflevel=0.99)
y$scint
```

```
## [1] 0.863 0.9957
```

Respondemos ahora a la segunda parte del ejercicio:

*Cálculo del tamaño muestral necesario para obtener una cierta precisión.*

El intervalo de confianza está centrado en  $\hat{p}$ . Quiere esto decir que la longitud máxima del intervalo (llamémosle  $L$ )

$$2 \cdot z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq L \iff n \geq \frac{4\hat{p}(1-\hat{p})z_{\alpha/2}^2}{L^2} = 4\hat{p}(1-\hat{p}) \left( \frac{1.96}{0.01} \right)^2$$

Como  $\hat{p} = 0.93$ , obtenemos que  $n \geq 10003.53$ . Hay que preguntarle a 10004 pacientes .

## 7.8 I.C. para la diferencia de proporciones.

Ahora suponemos dos poblaciones en donde se considera la misma característica  $A$ .  $p_1$  es la proporción de elementos con dicha característica en la primera población, y  $p_2$  es la proporción en la segunda población. Se toma una muestra de tamaño  $n_1$  de la primera población y otra de tamaño  $n_2$  en la segunda, y se calculan las respectivas proporciones muestrales  $\hat{p}_1$  y  $\hat{p}_2$ . El intervalo de confianza para la diferencia  $p_1 - p_2$  es

$$\left( (\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right).$$

Para calcularlo con R, introducimos la siguiente función:

```
z.test.2 <-function(x1,n1,x2,n2,p=0,
  conflevel=0.95,alternative="two.sided")
  {
  ts.z <-NULL
  cint <-NULL
  p.val <-NULL
  phat <-NULL
  p1 <-x1/n1
  p2 <-x2/n2
  q1 <-1 - p1
  q2 <-1 - p2
  phat <- p1 - p2
  SE.phat <-sqrt( ((p1*q1)/n1) + ((p2*q2)/n2) )
  ts.z <-(phat - p)/SE.phat
  p.val <-pnorm(ts.z)
  ts.z <- (phat - p )/SE.phat
  if(ts.z<0) p.val <-2*pnorm(ts.z)
  else p.val<2*(1-pnorm(ts.z))
  if(alternative=="less") {
    p.val <-pnorm(ts.z)
  }
  if(alternative=="greater") {
    p.val <-1 - pnorm(ts.z)
  }
  }
```

```

}

cint <-phat + c(
  -1*((qnorm(((1 - conflevel)/2) +
    conflevel))*SE.phat),
  ((qnorm(((1 - conflevel)/2) +
    conflevel))*SE.phat) )
return(list(estimate=phat,ts.z=ts.z,
  p.val=p.val,cint=cint))
}

```

Al igual que la función inmediatamente anterior, esta calcula el intervalo de confianza para la diferencia de proporciones, y también realiza un test de hipótesis para contrastar si las proporciones teóricas de las dos poblaciones pueden considerarse iguales.

**Ej.**

En Hollywood existe bastante confusión entre los actores Javier Bardem y Jeffrey Dean Morgan. Leer, por ejemplo [esto](#) o [aquello](#).

Se elige por la calle aleatoriamente a 100 personas, y se les enseña una foto de Bardem, pidiéndoles que marquen una casilla según quién crean que es (Bardem o el actor americano). 52 personas aciertan. Se hace lo mismo con otras 100 personas a las que se les enseña una foto de Jeffrey Dean Morgan, y 53 personas aciertan. ¿Crees que a lo mejor son la misma persona?

Calculamos el intervalo usando la función anterior. Únicamente es necesario, para calcular el intervalo de confianza, indicar el nivel de confianza que queremos (aunque, en esta ocasión, vamos a utilizar 0.95, y tampoco haría falta escribirlo, puesto que es el que se usa por defecto:

```
z.test.2(52,100,53,100, conflevel=0.95 )$cint
```

```
## [1] 0.148 0.128
```

Como vemos, la proporción de gente que reconoce correctamente a Bardem puede coincidir con la proporción de gente que reconoce correctamente a J.D.



Figura 7.12: J. Bardem y J. Dean Morgan (o al revés).

Morgan (puesto que el intervalo de confianza contiene al cero). O, lo que es lo mismo, la proporción de gente que se equivoca (creyéndose que es la otra persona) pueden coincidir.

## 7.9 Comparación de variables.

Ahora consideramos dos variables normales (por ejemplo, la estatura de dos poblaciones diferentes; la efectividad de un tratamiento 1 y un tratamiento 2...)

Se parte de dos muestras aleatorias simples  $(x_1, x_2, \dots, x_n)$  e  $(y_1, y_2, \dots, y_m)$ , de las variables  $X \in N(\mu_1, \sigma_1)$  e  $Y \in N(\mu_2, \sigma_2)$ , respectivamente (los tamaños de la muestra no tienen por qué ser iguales).

Llamamos  $\bar{x}$  a la media de la muestra de la primera variable ( $X$ ) e  $\bar{y}$  a la media de la muestra de la otra variable;  $\hat{S}_{n-1}$  es la cuasi-desviación típica de la primera muestra y  $\hat{S}_{m-1}$  la de la segunda muestra.

### 7.10 I.C. para la diferencia de medias

El intervalo de confianza para el parámetro diferencia de medias  $\mu_1 - \mu_2$  es:

$$\left( (\bar{x} - \bar{y}) \pm t_{n+m-2-\Delta, \alpha/2} \sqrt{\frac{\hat{S}_{n-1}^2}{n} + \frac{\hat{S}_{m-1}^2}{m}} \right),$$

siendo  $\Delta$  el entero más próximo a (Corrección de Welch)

$$\frac{\left( (m-1) \frac{\hat{S}_{n-1}^2}{n} - (n-1) \frac{\hat{S}_{m-1}^2}{m} \right)^2}{(m-1) \left( \frac{\hat{S}_{n-1}^2}{n} \right)^2 + (n-1) \left( \frac{\hat{S}_{m-1}^2}{m} \right)^2}.$$

**Ej.**

Uno de los dilemas que más trae de cabeza a los historiadores y antropólogos es si existen diferencias físicas y psicológicas (inteligencia, fuerza, carácter... ) entre la gente de derechas y de izquierdas. Hoy vamos a traer algo de luz a este tema, comparando las estaturas de famosos personajes históricos de izquierdas y de derechas.

Calcular un intervalo de confianza, al 95 por ciento, para la diferencia de estaturas medias, y razonar si alguno de los grupos puede considerarse más alto que el otro.

De derechas	Estatura	De izquierdas	Estatura
Adolf Hitler	175	Boris Yeltsin	187
Rudolf Hess	175	Josif Stalin	168
Francisco Franco	163	Fidel Castro	191
Reinhard Heydrich	191	Nicolas Maduro	190
Benito Mussolini	169	Che Guevara	175
Donald Trump	188	Hugo Chavez	173
Joseph Goebbels	165	Lenin	165
Hermann Goering	178	Nikita Khrushchev	160
Heinrich Himmler	174	Leonid Brezhnev	173
Jordi Pujol	165	Gorbachov	175
Jose Maria Aznar	171	Pablo Iglesias	176
Silvio Berlusconi	171	Evo Morales	174

**Solución:**

Calcular el intervalo de confianza a mano es bastante tedioso (lo dice uno que lo tuvo que hacer muchas veces), por culpa del valor  $\Delta$  que aparece en la fórmula. Los afortunados hijos de la era digital tenéis la suerte de contar con R o herramientas similares. Tan sencillo como poner los datos y una orden:

```
e1=c(175,175,18,191,18,1816,178174,16,171,171)
e2=c(18,18191, 190,175,173,16,18,173,175,178174)
t.test(e1,e2, conflevel=0.95)$confint
```

```
## [1] 9.56 5.89
## attr("conflevel")
## [1] 0.95
```

Este es el procedimiento para realizar un test de comparación de medias. Ahora mismo nos interesa el intervalo de confianza, que es  $(-9.56, 5.89)$ . Como podemos ver, el intervalo de confianza contiene al cero, o sea que hay poca diferencia entre la estatura media de un grupo y otro. De todas formas, para tomar una decisión más acertada, desde el punto de vista estadístico, es mejor realizar un contraste de hipótesis.

## 7.11 I.C. para el ratio de varianzas

Si queremos comparar, en vez de las medias, las desviaciones típicas (para ver si una población tiene mayor variabilidad que otra), utilizaremos el cociente de varianzas  $\sigma_2^2/\sigma_1^2$ . Una estimación puntual de este cociente es el que se obtiene dividiendo las cuasi-varianzas muestrales, y el intervalo de confianza para el cociente de las varianzas es:

$$\left( F_{n-1,m-1,1-\alpha/2} \frac{\hat{S}_{m-1}^2}{\hat{S}_{n-1}^2}, F_{n-1,m-1,\alpha/2} \frac{\hat{S}_{m-1}^2}{\hat{S}_{n-1}^2} \right),$$

siendo  $F_{n-1,m-1,\alpha/2}$  el valor de una F de Snedecor con  $n-1$  y  $m-1$  grados de libertad que deja a la derecha  $\alpha/2$  de área.



# Capítulo 8

## Contrastes de hipótesis

Antes de pasar a la matemática, dediquemos un poco de tiempo a hablar de filosofía y ciencia.

### 8.1 De conjeturas y otros

La ciencia se basa en la toma de decisiones. Estas decisiones a veces son acertadas y a veces no, y en ese sucesivo esfuerzo de “ensayo y error” el conocimiento avanza. Cualquier investigador está continuamente elaborando conjeturas y comprobando (o intentando comprobar) si son ciertas o no.

La ciencia plantea y desarrolla teorías orientadas a entender el mundo, desentrañando los mecanismos que determinan que las cosas funcionen como funcionan. Las teorías, no obstante, son sólo un conjunto de hipótesis, con un cierto grado de confirmación y, en un primer momento, solo simples conjeturas.

Demostrar la validez absoluta de una teoría es imposible, ya que no es admisible verificar cada una de las consecuencias derivadas de una teoría dada y con ello la teoría en su totalidad. Lo único que es factible es *refutarla* (*desmentir una hipótesis o una teoría mediante pruebas o experimentos*), a partir de un experimento u observación cuyo resultado contradiga lo predicho por la misma, o bien incrementar su grado de corroboración, a partir de la acumulación de verificaciones positivas de consecuencias observables anticipadas por la teoría.

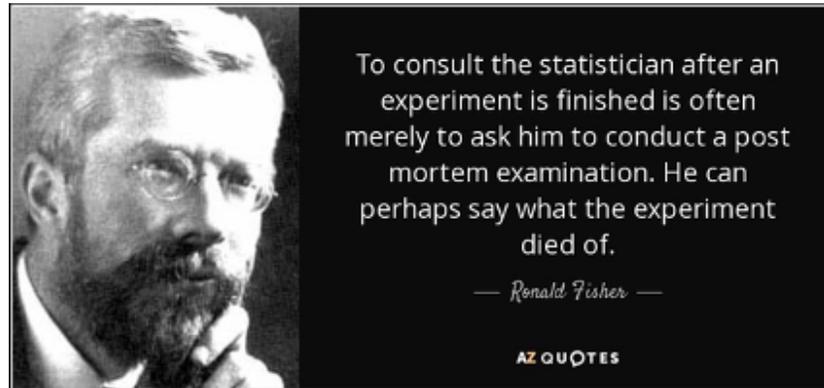


Figura 8.1: Una famosa frase de Ronald Fisher.

Las teorías están sujetas a un proceso de selección a cargo de la comunidad científica, teniendo siempre como juez de última instancia a la experiencia. A la ciencia se la puede tildar de dogmática en el sentido de defender en un momento determinado (en escalas que van de varias décadas hasta siglos) un paradigma aceptado. Pero la ciencia, a diferencia de otros ámbitos humanos, se caracteriza por tener una profunda actitud crítica. Lo que hace que, llegado un momento, después de la acumulación de anomalías y de la incapacidad por parte de una teoría para explicar nuevos fenómenos, existiendo a su vez una teoría en competencia “superadora”, la ciencia sea capaz de abandonar el viejo paradigma para abrazar uno nuevo.

El “falsacionismo” fue propuesto por el epistemólogo austriaco Karl R. Popper (1902-1994) en 1934, como criterio de demarcación entre lo científico y lo metafísico. Si una hipótesis no es potencialmente falsable y no admite por naturaleza, llegado el caso, ser refutada a partir de su confrontación con la experiencia, entonces, por definición, no es científica.

Un ejemplo muy conocido es el de la teoría de Newton, que llegado un momento fue desplazada por la teoría de la relatividad, en la cual se plantearon ideas nuevas (en su momento revolucionarias) con respecto al espacio, el tiempo y la energía.

Otro ejemplo es el de los cisnes negros. Antes del siglo XVII, se asumía como imposible la existencia de los mismos. Si tras cientos de años de observaciones de cisnes, nunca se había visto uno, la conclusión más razonable sería que no existían. Hasta que apareció uno (Taleb and Mosquera, 2011).

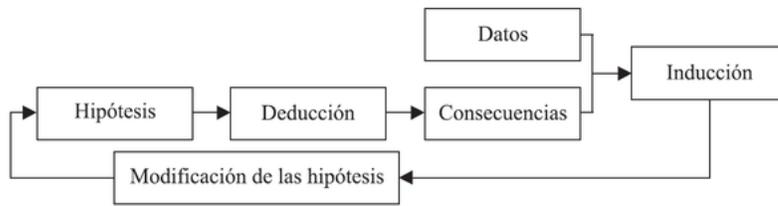


Figura 8.2:

El término *cisne negro* fue una expresión latina, cuya referencia conocida más antigua proviene de la descripción de algo que hizo el poeta Juvenal (poeta romano, año 50 a. de C.): *rara avis in terris nigroque simillima cygno*, cuya traducción en español significa un ave rara en la tierra, y muy parecida a un cisne negro. La importancia del símil radica en su analogía con la fragilidad de cualquier sistema de pensamiento. La frase de Juvenal era una expresión común en el Londres del siglo XVI como una declaración de imposibilidad. Todos los cisnes debían de ser blancos puesto que todos los registros históricos conocidos decían que tenían plumas blancas. Sin embargo, después de que una expedición holandesa, dirigida por el explorador Willem de Vlamingh en el río Swan en 1697, descubrió cisnes negros en Australia Occidental, lo que parecía imposible se convirtió, de repente, en cierto.

Hoy en día, el término *cisne negro* se refiere, a un suceso de probabilidad muy pequeña que, si ocurre, produce unos resultados extraordinariamente impactantes. Se utiliza fundamentalmente en economía, y corresponden a sucesos como las caídas bruscas de la bolsa (la depresión de 1929, o el estallido de la burbuja inmobiliaria). La teoría desarrollada por Nassim Taleb ([Taleb and Mosquera, 2011](#)) es que un suceso altamente improbable acaba siendo racionalizado por retrospectión. Nadie lo esperaba, pero posteriormente se analiza retrospectivamente, y se racionaliza (los datos estaban ahí, pero no se supieron ver).

## 8.2 Hipótesis estadísticas

Un principio general de la investigación científica es escoger siempre la hipótesis más simple capaz de explicar la realidad observada. La razón es que una hipótesis simple es más fácil de contrastar empíricamente y descubrir sus deficiencias, lo que permite aprender de los datos con mayor rapidez y

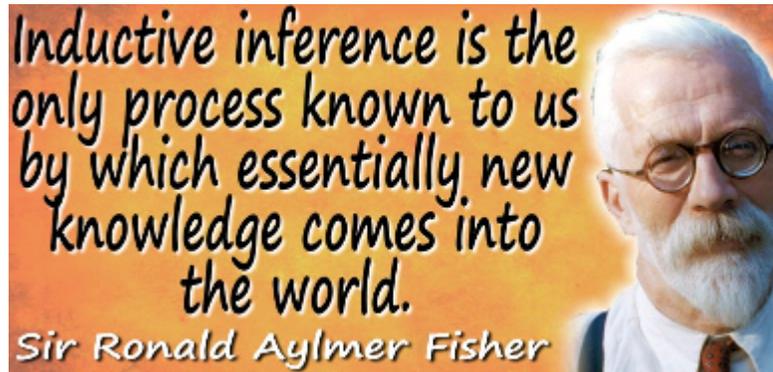


Figura 8.3: Otra conocida frase de Fisher acerca de la inferencia científica.

seguridad.

Este principio justifica que muchas investigaciones estadísticas tengan por objeto contrastar una hipótesis simplificadora del tipo: una población es idéntica a otra de referencia (el efecto de un medicamento nuevo es igual a otro existente...); dos o más poblaciones son iguales entre sí...

Una hipótesis se contrasta comparando sus predicciones con la realidad: si coinciden, dentro del margen de error admisible, mantendremos la hipótesis; en caso contrario, la rechazaremos.

Una **hipótesis estadística** es una afirmación que se hace sobre una o más características de una población (decir que la vida media de una batería son  $x$  horas, que un tratamiento reduce el dolor, que un determinado producto hace adelgazar, que a la gente le gusta mayoritariamente el morbo...). Un **test** o **contraste de hipótesis** es algún procedimiento para aceptar o rebatir dicha hipótesis o afirmación.

La teoría de los test de hipótesis surge a partir de 1925, cuando Ronald Fisher publicó el libro *Métodos estadísticos para investigadores*. En él, Fisher definía los **contrastos de significación**, cuyo funcionamiento básico sigue el esquema del falsacionismo, como veremos.

## 8.3 Test de significación (NHST)

Una prueba de significación consta de una hipótesis  $H_0$  (hipótesis nula) que consiste en establecer que el valor de un parámetro (llamémosle  $\theta$ ) es un número concreto (llamémosle  $\theta_0$ ):

$$H_0 : \theta = \theta_0$$

La hipótesis de partida del investigador fue bautizada con este nombre por Fisher porque significaba que no había cambio alguno con el uso de un nuevo fertilizante, que su efecto era nulo. Es decir, para poder demostrar que un nuevo fertilizante es efectivo, supondremos que no lo es, que tiene efecto nulo, y trataremos de **falsar** esta afirmación. El acrónimo *NHST* viene de “Null hypothesis signification testing”.

Ej.

Se conoce que la vida media de los elementos resultantes de un proceso de fabricación (por ejemplo bombillas) es 5.000 horas. Se introducen cambios en el proceso y se desea contrastar que la vida media no ha variado.

Ej.

Se desea saber si un tratamiento es efectivo para reducir el dolor en una lesión determinada. El tratamiento se realiza a una serie de pacientes y se comprueba si ha existido reducción del dolor (por término medio; unos pacientes reducirán más que otros, incluso alguno puede no ver reducido el dolor y alguno hasta podría aumentarlo) o no lo ha habido.

Entonces, una vez que se delimita la hipótesis nula que se desea poner a prueba, la forma general de proceder es elegir una muestra de esa población, y ver si los resultados de esa muestra son coherentes con la afirmación que se está realizando. Básicamente, que la *evidencia* que nos proporcione la muestra sea lo suficientemente fuerte para poder tomar una decisión.

Para comprobar la *coherencia* de los resultados, o la *fuerza de la evidencia*, se tratará de ver la diferencia entre lo observado en la muestra y lo que dice la hipótesis nula. Para ello se elige el llamado **estadístico  $T$**  (*estadístico pivote*) del test y se calcula su valor sobre los datos de una muestra observada  $(x_1, x_2, \dots, x_n)$ , lo que se denota como  $T(x_1, x_2, \dots, x_n)$ . Dado que la distribución en el muestreo del estadístico  $T$  ha de ser conocida, se determina la

probabilidad de que el estadístico tome un valor igual o más extremo que el valor observado  $(x_1, x_2, \dots, x_n)$ , bajo la suposición de que la hipótesis nula es cierta.

Se escribe simbólicamente

$$P(T \geq T(x_1, x_2, \dots, x_n)/H_0),$$

y este número se denomina *p-valor*.

Entonces, si el *p-valor* es excesivamente pequeño -en general, por debajo de 0.05-, el resultado del test se dice que es significativo, porque permite rechazar la hipótesis nula  $H_0$ . En otro caso, el resultado del test no es significativo y no puede rechazarse esta hipótesis.

“un fenómeno es demostrable experimentalmente cuando se conoce como conducir un experimento que raramente falla para darnos un resultado estadísticamente significativo”

— Sir Ronald Fisher.

De acuerdo con esta lógica se entiende que, con un nivel de significación del 5%, en promedio cinco de cada 100 veces que la hipótesis nula sea cierta se rechazará por azar. Es decir, 5 de cada 100 veces que rechazamos la hipótesis nos estaremos equivocando, pues se asume que es cierta y la prueba no demuestra ni su falsedad ni su certeza.

La hipótesis nula sólo se rechaza si la probabilidad de observar una muestra como la dada es demasiado baja. Esto es, la hipótesis nula se rechaza *si la muestra es demasiado rara* cuando la hipótesis nula es cierta. El razonamiento estadístico se basa en la siguiente disyunción lógica: “o bien ha ocurrido un suceso excepcional (muy improbable) o bien la hipótesis nula no es correcta.”

El *p-valor* o probabilidad de significación funciona como una suerte de la evidencia en contra de la hipótesis nula: cuanto menor es, más evidencia en contra. Un valor demasiado pequeño indica que la muestra observada se separa de lo esperado mucho más de lo que sería achacable al azar, a las circunstancias del muestreo aleatorio. Por lo tanto, el investigador se encuentra ante una hipótesis nula inverosímil, descartable.

Fisher describía los test de significación como un procedimiento para rechazar la hipótesis nula, que en ningún caso podía ser probada o establecida

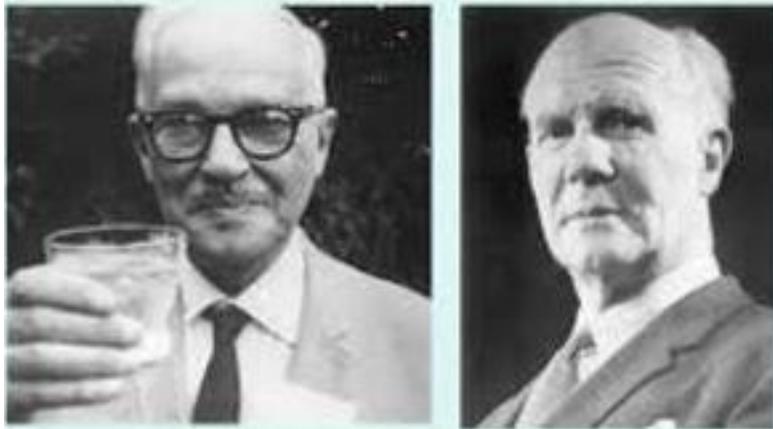


Figura 8.4:

definitivamente. Este planteamiento refutacionista era coherente con la corriente falsacionista. La propuesta metodológica de Fisher era una especie de falsacionismo aplicado a la estadística: se trata de rechazar aquellas hipótesis para las cuales las observaciones sean relativamente inverosímiles.

## 8.4 La teoría de Neyman-Pearson

Buscando fortalecer las bases lógicas de los test de significación de Fisher, Egon Pearson (1895-1980) (hijo de Karl Pearson) y Jerzy Neyman (1894-1981) idearon varias mejoras. El eje principal de su investigación era el siguiente interrogante: ¿qué hacer si se obtiene un resultado significativo en un test estadístico? Se rechaza la hipótesis nula, pero los test de significación no arrojan ninguna pista sobre qué hipótesis elegir a cambio.

La teoría de Neyman-Pearson utilizó el NHST de Fisher y el  $p$ -valor como parte de un **proceso formal de decisión**. Así, plantearon una elección real entre dos hipótesis rivales. El contraste de hipótesis quedó convertido en un método para discernir entre dos hipótesis: la hipótesis nula y la hipótesis **alternativa**  $H_1$ .

Todo contraste de hipótesis conduce pues, a aceptar o rechazar la hipótesis nula planteada (aceptando, en este último caso, la hipótesis alternativa). Ahora bien, pueden darse las siguientes situaciones.

CONDICIÓN REAL	DECISIÓN	
	Rechazar $H_0$	No Rechazar $H_0$
$H_0$ cierta	<b><i>Error (Tipo I)</i></b>	<b><i>Acierto</i></b>
$H_0$ falsa	<b><i>Acierto</i></b>	<b><i>Error (Tipo II)</i></b>

Figura 8.5: Posibles opciones en un test de hipótesis.

- Se acepta la hipótesis nula siendo verdadera. Esta es una decisión correcta.
- Se rechaza la hipótesis nula siendo falsa. Esta es otra situación correcta.
- Se rechaza la hipótesis nula siendo verdadera. Estamos cometiendo un error, que se llama *error de tipo uno*. La probabilidad de cometer este error viene dada por el nivel de significación  $\alpha$ , fijado de antemano.
- Se acepta la hipótesis nula siendo falsa. También cometemos un error, que se llama *error de tipo II*. La probabilidad de cometer este error se representa por  $\beta$ , y la probabilidad  $1 - \beta$  se llama *potencia del contraste*, que cuantifica la probabilidad de rechazar la hipótesis nula cuando es falsa.

Veamos. Si tenemos una hipótesis nula:  $H_0$ : *un tratamiento nuevo no es efectivo*, frente a  $H_1$ : *el tratamiento sí es efectivo* siempre es posible construir más de un test de hipótesis para contrastar la hipótesis nula frente a la alternativa.

Por ejemplo, tiramos una moneda al aire. Si sale cara, aceptamos  $H_0$ . Si sale cruz, rechazamos  $H_1$ .

La probabilidad de cometer un error de tipo I es 0.5, igual que la probabilidad de cometer un error de tipo II.

Si en vez de tirar una moneda tiramos un dado y decidimos mediante la regla: “aceptamos la hipótesis nula si sale un 1, la rechazamos si sale cualquier otro número”, la probabilidad de error de tipo I es 5/6 y la de error de tipo II es 1/6.

Obviamente ambos test son bastante absurdos, pero nos sirve para ver que siempre existen test con sus correspondientes errores.

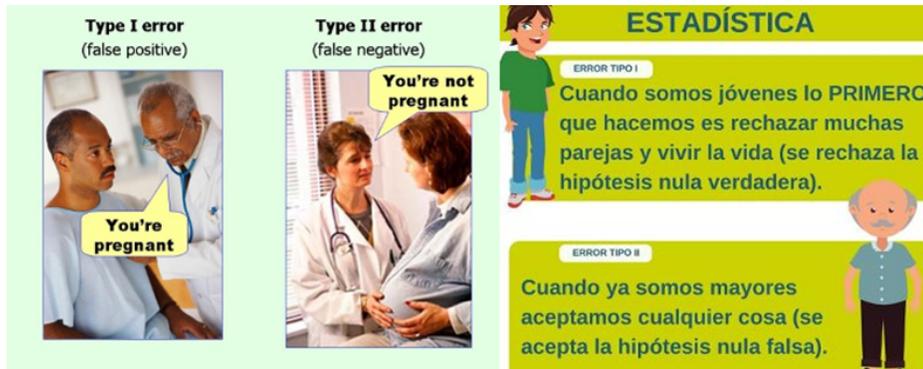


Figura 8.6: Ejemplos errores tipo I y II.

Ej.

Un acusado ante un tribunal:

 $H_0 : \textit{inocente}$  $H_1 : \textit{culpable}$ El error de tipo I es *rechazar que es inocente, siéndolo*.El error de tipo II es *rechazar que es culpable, cuando es inocente*.Si se ponen las hipótesis al revés:  $H_0 : \textit{culpable}$  frente a  $H_1 : \textit{inocente}$  se comprueba enseguida que los errores de tipo I y tipo II se permutan.

Ej.

Una alarma de incendio. Cuando suena una alarma, ante un exceso de calor, o bien que un gamberro ha acercado un mechero al sensor (se ve en las películas), la alarma puede sonar y no haber fuego.

 $H_0 : \textit{fuego}$  $H_1 : \textit{nofuego}$ El error de tipo I es *rechazar que hay fuego, cuando en realidad lo hay*. El error de tipo II es *aceptar que hay fuego, cuando en realidad no lo hay*.

Con estos dos ejemplos, podemos ver que no es posible disminuir simultáneamente la probabilidad de error de tipo I y la probabilidad de error de tipo II: una opción para no cometer errores de tipo I en el caso de un juicio sería

declarar inocente a casi todo el mundo, lo cual conlleva a cometer muchos errores de tipo II. En el caso de la alarma, puede hacerse que el aparato no tenga demasiada sensibilidad, para que no haya falsas alarmas, pero esto puede hacer peligrar el hecho de que, ante un incendio de verdad, la alarma no se active.

Neyman y Pearson demostraron que, en bastantes circunstancias, una vez fijada la probabilidad  $\alpha$  de error de tipo I -esto es, una vez acotado el porcentaje de veces que tomaremos una decisión equivocada al rechazar la hipótesis nula cuándo es verdadera- es posible construir y utilizar contrastes de **máxima potencia**, es decir, contrastes que minimizan la probabilidad  $\beta$ , o de error de tipo II (o sea, maximizan la llamada **potencia** del test: su sensibilidad o capacidad para detectar que la hipótesis nula es falsa).

Este último párrafo seguro que apesta a matemáticas. Por si alguien no se había dado cuenta, para decidir entre dos hipótesis (la nula y la alternativa) podemos, habitualmente, realizar más de un test o proceso de decisión.

Supongamos que queremos elegir entre:

$H_0$  : *inocente* frente a  $H_1$  : *culpable*

podemos revisar las pruebas, interrogar a los testigos, etc. y tomar una decisión. Habrá una probabilidad de error de tipo I y una probabilidad de error de tipo II (1- potencia).

Pero se podría decidir simplemente tirando una moneda al aire. La probabilidad de cometer un error de tipo I es 0.5. Igual que la potencia del test (1- probabilidad de error de tipo II).

Ahora, en vez de tirar una moneda tiramos un dado. Si sale el 1, decidimos que el acusado es culpable, y si no es inocente. La probabilidad de cometer un error de tipo I es 1/6. La potencia es la probabilidad de rechazar la hipótesis nula siendo falsa, es decir 5/6 (el que el acusado sea inocente o culpable no va a influir en el resultado del lanzamiento de la moneda; son sucesos independientes).

Vemos que reglas de decisión diferentes ocasionan probabilidades de error diferentes. En una situación “seria”, las hipótesis a elegir son de tipo estadístico. La forma de decidir será alguna función también de tipo estadístico (va a ser una variable aleatoria). A partir de ella calcularemos las probabilidades de error.

Supongamos que hay 2 posibles formas de decidir (2 variables aleatorias). Neymann y Pearson dijeron que la mejor manera de decidir entre ambas es mantener para ambas la misma probabilidad de error de tipo I, y luego elegir la que dé mayor potencia (o menor probabilidad de error de tipo II) (Mismo valor de  $\alpha$ , máxima potencia).

En un célebre resultado publicado en 1933, Neyman y Pearson probaron que en el caso de hipótesis rivales simples (que asignan valores específicos al parámetro desconocido) existe automáticamente una clase de test óptimos, de bajo tamaño y máxima potencia: los basados en la razón de verosimilitudes. Los contrastes de hipótesis que se utilizan en los casos prácticos, cuando se quiere saber, por ejemplo, si un tratamiento es efectivo, o si una proporción en una muestra aproxima una proporción real (caso de una encuesta electoral) son los propuestos mediante la teoría de Neyman y Pearson.

Quiere esto decir que son los que, desde un punto de vista matemático, son mejores para decidir entre la hipótesis nula y alternativa. Ya hemos visto antes que podemos construir muchos procedimientos de decisión (por ejemplo, tirando una moneda al aire). Aunque se puedan pensar procedimientos matemáticos más correctos, los test de Neyman y Pearson son óptimos desde el punto de vista antes comentado.

## 8.5 Contrastes paramétricos y no paramétricos

Los contrastes pueden ser de tipo **paramétrico** o **no paramétrico**, según se refieran o no a parámetros de una población (a la media, a la varianza, a una proporción...).

Una **hipótesis paramétrica** es una afirmación sobre una o más características (parámetros) de una población. Si dicha hipótesis especifica un único valor para el parámetro la llamaremos **hipótesis simple**.

Ej.

La estatura media de los varones españoles mayores de 18 años es 1.77 m ( $\mu = 1.77$ )

Si se especifica más de un valor para el parámetro la llamaremos **hipótesis compuesta**.

Ej.

La temperatura media de una caldera nunca alcanza un valor superior a los 90 grados ( $\mu \geq 90$ ).

Ej.

En un supermercado venden dos clases de naranjas, A y B. Las naranjas difieren en el diámetro, siendo en ambos casos de medias 25 y 30 cm, respectivamente. El diámetro sigue una distribución normal que, en ambos casos, tiene una desviación típica de 2 cm. Al llegar a casa, un señor ve que no le han puesto etiqueta a las bolsas, por lo que, en principio, no sabe cuál es cual. Teniendo en cuenta que las naranjas de la clase A son para él y las de la clase B para el vecino, necesita poder diferenciarlas.

Después de pensar un rato, el señor decide realizar el siguiente proceso para clasificar las naranjas: coge las de una bolsa (que son 20), y las mide. Aceptará que son del tipo normal si la media de las longitudes no supera los 28 cm.

Vamos a calcular las probabilidades de los errores que es posible cometer.

Se plantea el contraste  $H_0 : \mu = 30$  frente a  $H_1 : \mu = 25$ . La regla de decisión es aceptar  $H_0$  si  $\bar{x} > 28$ .

Recordemos que la variable normal es reproductiva, es decir, la suma de variables aleatorias normales sigue también una distribución normal:

si  $X_1, X_2, \dots, X_n$  son variables normales de media o esperanza  $\mu_i = E(X_i)$  y varianza  $\sigma_i^2 = Var(X_i), i = 1, \dots, n$ , la variable suma  $Y = X_1 + X_2 + \dots + X_n$  es también una variable normal, de media la suma de las medias, y varianza la suma de varianzas (desviación típica = raíz de la suma de varianzas), es decir

$$Y = X_1 + X_2 + \dots + X_n \in N \left( \sum_{i=1}^n \mu_i, \sqrt{\sum_{i=1}^n \sigma_i^2} \right).$$

$X \in N(\mu, \sigma)$  entonces, dada una muestra de tamaño  $n$ , la media muestral  $\bar{x} \in N(\mu, \frac{\sigma}{\sqrt{n}})$ .

$P(\text{Error tipo I}) = P(\text{Rechazar } H_0 \text{ siendo cierta}) = P(\bar{x} \leq 28 / H_0 \text{ es cierta}).$

$P(\text{Error tipo II}) = P(\text{Aceptar } H_0 \text{ siendo falsa}) = P(\bar{x} > 28/H_0 \text{ es falsa}).$

Si  $H_0$  es cierta,  $\mu = 30$ , luego  $\bar{x} \in \left(30, \frac{2}{\sqrt{20}}\right)$ .

$$\begin{aligned} P(\bar{x} \leq 28/H_0 \text{ es cierta}) &= P(\bar{x} \leq 28/\mu = 30) = \\ &= P\left(Z \leq \frac{28 - 30}{\frac{2}{\sqrt{20}}}\right) = P(Z \leq -4.472) \cong 0. \end{aligned}$$

Si  $H_0$  es falsa,  $\mu = 25$ , por lo tanto  $\bar{x} \in N\left(25, \frac{2}{\sqrt{20}}\right)$ .

$$\begin{aligned} P(\bar{x} > 28/H_0 \text{ es falsa}) &= P(\bar{x} > 28/\mu = 25) = \\ &= P\left(Z > \frac{28 - 25}{\frac{2}{\sqrt{20}}}\right) = P(Z > 6.7) \cong 0 \end{aligned}$$

Comprobamos que las probabilidades de cometer los errores de tipo I y II son prácticamente cero, por lo que el señor ha inventado una regla de decisión estupenda.

### Ejemplos de hipótesis no paramétricas:

**Ej.**

Las notas de una asignatura son normales (la variable  $X = \text{'nota'}$  sigue una distribución normal).

**Ej.**

Los resultados de un sorteo no son aleatorios (la variable  $X = \text{'resultados'}$  no sigue una distribución uniforme).

Como dijimos antes, la manera habitual de realizar un contraste o test de hipótesis es tomar una muestra, y ver si los resultados son coherentes o no con lo que se afirma en la hipótesis nula.

Supongamos que un investigador desarrolla un método o tratamiento para reducir el dolor. Para comprobar que el método es efectivo, habrá que observar una muestra de pacientes con dolor y, tras aplicarles el tratamiento, ver si el dolor ha disminuido.

Si la muestra consta, en general, de  $n$  pacientes, se observa en ellos la variable

$X = \text{dolor antes del tratamiento}$

Tendremos  $n$  datos  $x_1, x_2, \dots, x_n$ .

Una vez realizado el tratamiento, volvemos a observar la variable

$Y = \text{dolor tras el tratamiento}$

Tendremos otros  $n$  datos  $y_1, y_2, \dots, y_n$ . A continuación, restamos el dolor antes y el dolor después  $Z = X - Y$  y tenemos los datos  $z_1, z_2, \dots, z_n$  donde  $z_i = x_i - y_i$ .

De manera general, si el tratamiento es efectivo, los valores de  $X$  serán mayores que los de  $Y$  (el dolor antes era más alto que el dolor después). Pero, en general, el descenso de dolor no va a ser igual en todos los pacientes. En unos se reducirá más, en otros menos. Por ello lo que interesa comprobar es si la reducción de dolor, **en media**, es grande o no lo es (porque si la reducción de dolor es pequeña, o casi nula, el tratamiento a lo mejor ni interesa por ser caro).

La hipótesis nula a comprobar, entonces, será de la forma:

$$H_0 : \mu = 0$$

donde  $\mu$  representa el nivel medio de disminución del dolor. Esta es la hipótesis que planteamos, porque es la que se pretende “falsar”, es decir, encontrar alguna prueba de que no es cierta (cuando usamos el tratamiento). Por lo tanto, lo interesante será (para demostrar que el tratamiento es efectivo) que la media de la muestra ( $z_1, z_2, \dots, z_n$ ) no sea un número próximo a cero.

**Ej.**

una variable sigue una distribución normal. *Por ejemplo, las notas de una asignatura.*

Una posible forma de realizar un contraste sería por medio de un histograma. Fijémonos en la Figura siguiente. Tenemos 3 muestras o 3 grupos de notas

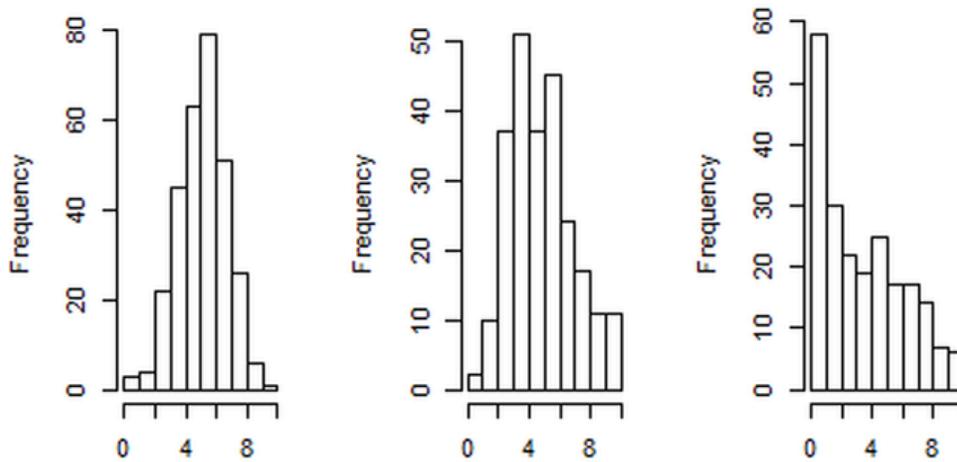


Figura 8.7: Tres posibles histogramas para una distribución de notas.

de 250 alumnos. Claramente, el primero de los histogramas nos diría que la variable  $X$  es normal, puesto que su forma se parece a la campana de Gauss. Mientras, el segundo histograma nos diría que existe una asimetría a la derecha, y con el tercer histograma claramente rechazaríamos la hipótesis nula.

Este ejemplo sería un contraste de tipo no paramétrico, puesto que no se realiza ninguna afirmación sobre parámetros de la variable. Intentamos saber si  $X$  es una variable normal o no.

Ej.

Veamos ahora un caso paramétrico. Nos basamos en la película de James Bond “Octopussy” (1982); en concreto, en el siguiente fragmento, cuando Kamal Khan, el malo malísimo de la película, está jugando a los dados contra un millonario inglés en un casino de la India. Kamal apuesta a que es capaz de sacar un seis doble al lanzar dos dados. Dice que siempre lo logra, que todo es “cuestión de muñeca”.

<https://player.vimeo.com/video/242488892>”

Desde un punto de vista estadístico, ¿cómo se debería proceder para saber si Kamal está haciendo trampa?.

Veamos: Si los dados están cargados en el 6, la probabilidad de salir un seis doble será mayor de lo que le correspondería, es decir:  $p(6 \cap 6) = p_0 = 1/36$ .

Como en principio, todo malo, por muy malo que sea, tiene derecho a la presunción de inocencia, la hipótesis nula es cierto salvo que los resultados demuestren lo contrario. Vamos a plantear entonces las hipótesis nula y alternativa como

$H_0$  : Los dados no están cargados, es decir

$H_0 : p = p_0 = 1/36$  frente a

$H_1$  : Los dados sí están trucados ( $p > 1/36$ ).

Lo procedente, desde el punto de vista estadístico, es observar una serie larga de tiradas del dado por parte de Kamal, para ver si tiene más suerte de la que le correspondería por azar. Como hemos visto en el capítulo anterior, un estimador natural del parámetro  $p$  es la proporción muestral  $\hat{p}$

Vamos a suponer que James Bond observa 30 tiradas de los dos dados, y que en esas 30 tiradas el seis doble sale 3 veces. Tenemos entonces una muestra en la cual En esta muestra,  $\hat{p} = 3/30 = 0.1$  que es distinta y más grande que  $p_0 = 1/36 = 0.027$ .

Como vemos, existe una discrepancia. Ahora bien, ¿es porque realmente los dados están trucados, o el resultado es fruto de la casualidad en la muestra elegida? Dicho de otra manera: ¿qué probabilidad existe de que el dado no esté preparado ( $H_0$  es cierta) y que la muestra arroje una proporción muestral de 0.1. Dicho de otro modo: ¿Qué probabilidad existe de que el dado no este trucado ( $H_0$  es cierta) y que exista esa diferencia entre lo observado (0.1) y lo teórico (0.027)?

La discrepancia obtenida entre lo que dice la hipótesis nula ( $p = 0.027$ ) discrepa de lo observado en la realidad ( $p = 0.1$ ). ¿Esta discrepancia es grande o pequeña? A simple vista parece grande, pero ¿podemos medirla de alguna forma?

En la terminología de Fisher:

$T =$  medida de la discrepancia

$T(x_1, x_2, \dots, x_n) = 0.1 - 0.027$

y tenemos que calcular el  $p$ -valor  $P(T \geq (x_1, x_2, \dots, x_n) / H_0)$ ,



Figura 8.8: Imagen del cómic Rue del Pércibe, 13, F. Ibáñez.

El  $p$ -valor es la **probabilidad de que los datos (si la hipótesis nula es cierta) se hayan obtenido por casualidad.**



La forma “general” de medir la discrepancia entre un valor teórico y uno estimado es mediante la diferencia de ambos, dividida por la desviación típica.

Se utiliza la misma idea para medir la diferencia entre dos grupos, o dos tratamientos. Se comparan las medias dividiendo por la desviación típica.

En este caso, El  $p$ -valor (más adelante vemos como se calcula) es 0.0073, por lo tanto, muy pequeño. Resulta así, muy difícil creer en que los datos no estén trucados.

## 8.6 Contrastes de hipótesis paramétricas

### 8.6.1 Tipos de contrastes: bilaterales y unilaterales

Un contraste es bilateral cuando tiene la forma

$H_0 : \theta = \theta_0$  (Por  $\theta$  nos referimos a un parámetro teórico y por  $\theta_0$  a un valor constante, un número) frente a

$H_1 : \theta \neq \theta_0$

Un contraste **unilateral** es de la forma:  $H_0 : \theta = \theta_0$  frente a

$H_1 : \theta = \theta_0$  o bien  $H_1 : \theta = \theta_0$

Ej.

Con el mayor consumo de chucherías y comida basura, parece que el peso medio de los niños de 12 años ha aumentado.

Contraste unilateral:

$H_0 : \mu = 26kg$  frente a  $H_1 : \mu > 26$

Ej.

El nuevo virus zombi ha provocado una alteración en el peso de los adultos.

Contraste bilateral:

$H_0 : \mu = 60kg$  frente a  $H_1 : \mu \neq 60$

### 8.6.2 Pasos a seguir al realizar un contraste de hipótesis

- 1.- Especificar las hipótesis nula y alternativa.
- 2.- Elegir un estadístico  $T$  para el contraste (para medir la discrepancia entre lo observado y lo teórico). *Este estadístico tendrá una función de densidad determinada que nos servirá para calcular el  $p$ -valor.*
- 3.- Tomar una muestra  $(x_1, x_2, \dots, x_n)$  y evaluar el estadístico del contraste  $T(x_1, x_2, \dots, x_n)$ .
- 4.- Calcular el  $p$ -valor

$$P(T \geq (x_1, x_2, \dots, x_n) / H_0),$$

que viene a ser la probabilidad de obtener esos datos, si  $H_0$  es cierta. Si ese valor es muy pequeño, significa que esos datos son muy improbables bajo la hipótesis nula, con lo cual tenderemos a pensar que no es cierta.

Una especie de nivel crítico es el valor 0.1, de forma que

Si el  $p$ -valor es más pequeño que 0.1, tenderemos a no creer en  $H_0$ , y, en cambio, si es más grande, tenderemos a creer en  $H_0$ .

En muchas ocasiones (muchos libros, artículos de investigación), se trabaja fijando un **nivel de significación**  $\alpha$  (error de tipo I) y realizando la siguiente comparación:

- Si  $p < \alpha$  se rechaza  $H_0$
- si  $p \geq \alpha$  se acepta  $H_0$ .

Los valores con los que se suele trabajar son  $\alpha = 0.1, 0.05$  o  $0.01$ . El más habitual es  $\alpha = 0.05$

Para rechazar la hipótesis nula se requiere que la probabilidad del fenómeno que se produce por casualidad sea inferior al cinco por ciento (este es el origen de la anécdota del estadístico que presencié la decapitación de 25 vacas, advirtió que una sobrevivía y desechó el fenómeno por no significativo).

Ver [Problemas del nivel de significación](#)

No rechazar una hipótesis no prueba que sea totalmente cierta. Podemos cometer un error de tipo II.

A continuación, indicamos los estadísticos que se utilizan para los principales contrastes de tipo paramétrico, y la distribución que siguen cuando la hipótesis nula es cierta.

### 8.6.3 Para la media de una variable normal

Supongamos que estamos trabajando con datos de una variable aleatoria  $X$  **que sigue una distribución normal**. Supongamos que establecemos la hipótesis de que la media (teórica, que es desconocida) es igual a un cierto valor numérico:

$$H_0 : \mu = \mu_0.$$

$\mu_0$  es un número concreto. Por ejemplo, se ha realizado un tratamiento a un conjunto de pacientes con dolor lumbar. A cada uno de ellos se le ha pedido que diga, en una escala de 0 a 10, cuánto dolor le ha reducido el tratamiento (0 si nada, 10 si se ha quedado sin dolor alguno).

Suponemos que la variable  $X$  = *reducción del dolor con el tratamiento* sigue una distribución normal. Para saber si el tratamiento es efectivo, la hipótesis nula que plantearíamos es:

$$H_0 : \mu = 0.$$

De lo que se trata es de *afisar* esta hipótesis, es decir, encontrar una diferencia estadísticamente significativa entre lo que se obtenga en la muestra y lo que dice la hipótesis nula. Esta última, al considerar que la media es cero, implica que el tratamiento no tiene efectividad, puesto que, en media, no reduce el dolor.

Para realizar el contraste de hipótesis, inicialmente consideramos dos posibilidades: que conozcamos la desviación típica de la variable, o que no la conozcamos. Esta última opción es la más habitual, y la más lógica, puesto que, si no tenemos información sobre la media, es raro tenerla de la desviación típica. En el ejemplo que hemos puesto, si no se conoce la reducción media de dolor con un tratamiento (hablamos habitualmente de un tratamiento en fase experimental) pues será raro conocer su variabilidad.

En todo caso, en la mayoría de los textos realizan la distinción entre los casos de desviación típica conocida o no. Consideraremos aquí también los dos casos, igual que se hizo en el capítulo anterior de intervalos de confianza.

#### 8.6.4 Si se conoce la desviación típica

Si tenemos una muestra  $(x_1, \dots, x_n)$  de  $n$  datos de una variable aleatoria normal, de parámetros  $\mu$  y  $\sigma$ , la media muestral verifica

$$\bar{x} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Por lo tanto, si tipificamos la variable (restamos la media y dividimos por la desviación típica), obtenemos la variable  $T$ , lo que quiere decir que esta variable sigue una distribución normal estándar  $(N(0, 1))$ .

$$T = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \in N(0, 1)$$

Este estadístico o variable aleatoria  $T$  se llama estadístico “pivote” o estadístico del contraste. Es con el que se realiza el contraste y se toma una decisión, según el valor que tome con los datos de la muestra.

El  $p$ -valor se calcula en función de la distribución que sigue el estadístico “pivote” del contraste, y de que el contraste sea bilateral o unilateral.

**Ej.**

Los catedráticos de la universidad se quejan al rector de que los precios del menú del día de las cafeterías universitarias han subido espectacularmente. Para verificarlo, se envía a un vicerrector a comer a las cafeterías de las distintas facultades y se anota el precio del menú:

6, 6.6, 6.5, 5.8, 7, 6.3, 6.2, 7.2, 5.7, 6.4, 6.5, 6.2, 6, 6.5, 7.2, 7.3, 7.6, 6.8, 6

El curso anterior, el precio medio del menú era de 6.8 euros, y la desviación típica de 0.7. ¿Podrá el rector decirle a sus profesores que los precios no han subido?

**Solución:**

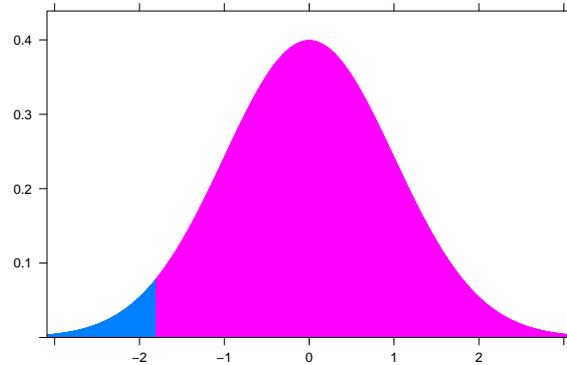
El test de hipótesis que debemos plantear es  $H_0 : \mu = 6.8$  frente a  $H_1 : \mu < 6.8$ .

A partir de la muestra, calculamos la media muestral, que es  $\bar{x} = 6.51$ . El valor del estadístico es, entonces:

$$T = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{6.51 - 6.8}{0.7/\sqrt{19}} = -1.8$$

Al ser un test unilateral, donde la hipótesis alternativa  $H_1$  es de la forma  $H_1 : <$ , el  $p$ -valor se calcula como el área a la izquierda de  $-1.8$  (en este caso, de una variable  $N(0, 1)$  que es la distribución del estadístico  $T$ ).

```
library(mosaic)
plotDist("norm", groups = x >-1.8 , type="h")
```



```
pnorm(-1.5)
```

```
## [1] 0.03593
```

El  $p$ -valor es 0.03593. Si lo comparamos con el valor más utilizado para  $\alpha = 0.05$ , se rechazaría que la media es 6.8 y aceptaríamos que el precio medio ha bajado.

### 8.6.5 La prueba $t$

En la práctica, es bastante extraño conocer la desviación típica (puesto que precisamente estamos en una situación de incertidumbre de la variable. Lo habitual es no conocer ni la media ni la desviación típica). En este caso, lo que se hace es estimar la desviación típica a partir de la muestra, utilizando para ello la cuasi-desviación típica muestral  $\hat{S}_{n-1}$ . En este caso, el estadístico que se utiliza cambia, y también la distribución del mismo. Tenemos que usar el estimador

$$T = \frac{\bar{x} - \mu_0}{\hat{S}_{n-1}/\sqrt{n}} \in t_{n-1}.$$

Este estimador sigue una distribución  $t$  de Student, con  $n - 1$  grados de libertad. A este contraste de hipótesis se le llama **prueba  $t$** .

**Ej.**

Vamos a realizar el ejercicio anterior, pero suponiendo que desconocemos la desviación típica teórica de la población. En este caso, calculamos  $\hat{S}_{n-1}$  con los datos de la muestra.

Recordemos que la desviación típica se calcula con `sd`

```
x=c(66.65, 5.87, 63, 62, 7.2, 5.7, 64, 65,
    62, 66.5, 7.2, 7.3, 7.668)
mean(x)
```

```
## [1] 65.16
```

```
sd(x)
```

```
## [1] 0.5419
```

El valor del estadístico es, por lo tanto:

$$T = \frac{\bar{x} - \mu_0}{\hat{S}_{n-1}/\sqrt{n}} = \frac{6.51 - 6.8}{0.54/\sqrt{19}} = -2.3.$$

Y ahora el  $p$ -valor lo buscamos a partir de la distribución  $t$ , en este caso con  $19 - 1$  grados de libertad.

```
pt(-2.3, df=18) #df son grados de libertad
```

```
## [1] 0.016
```

El  $p$ -valor en este caso es 0.016.

Este test puede realizarse de manera directa en R, de la forma:

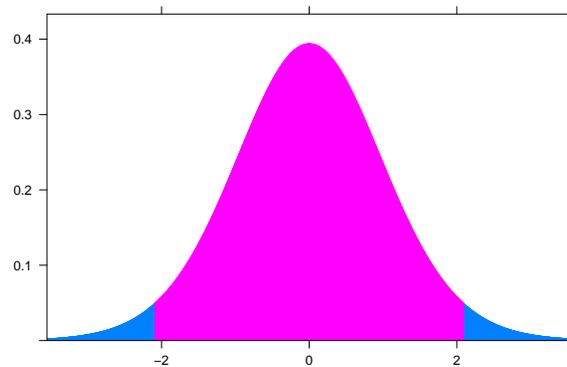
```
x=c(66.65, 5.87, 63, 62, 7.2, 5.7, 64,
    65, 62, 66.5, 7.2, 7.3, 7.668)
t.test(x, mu=68, alternative="less")
```

```
##
## One Sample t-test
##
## data: x
## t = 2.3, df= 18, pvalue = 0.02
## alternative hypothesis: true mean is less than 68
```

```
## 95 percent confidence interval:
##      Inf6731
## sample estimates:
## mean of x
##      6516
```

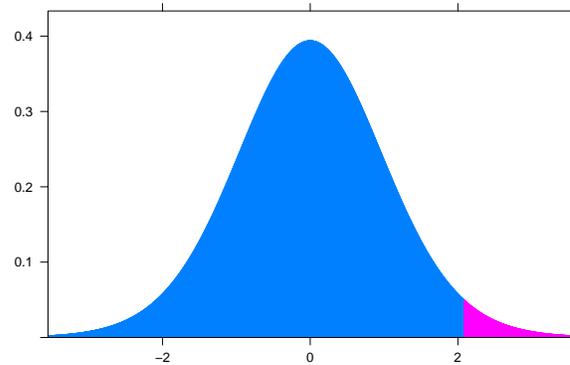
Supongamos, por ejemplo, que la hipótesis alternativa  $H_1$  es de la forma  $H_1 : \mu \neq \mu_0$ , entonces el **nivel crítico** o **p-valor** es 2 veces el área a la derecha del valor absoluto del estadístico del contraste  $\hat{w}$ , que se calcula como  $qt(0.975, df=20)$  (recordemos que el p-valores es la probabilidad  $P(T \geq (x_1, x_2, \dots, x_n)/H_0)$ ).

```
plotDist("t", df= 20, groups = x > -2.08 & x < 2.08,
         type = "h")
```



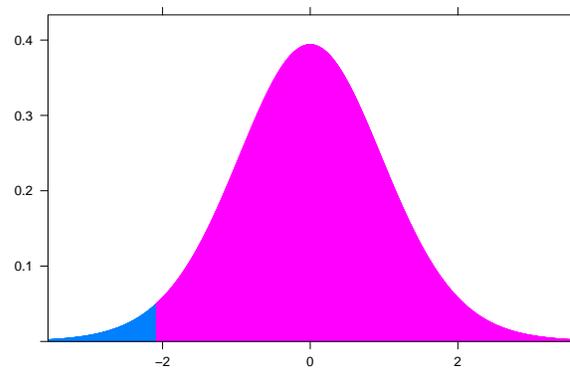
Si  $H_1$  es de la forma  $H_1 : \mu > \mu_0$ , el **p-valor** es el área a la derecha del estadístico del contraste.

```
plotDist("t", df=20, groups = x > 2.08 , type="h")
```



Si  $H_1$  es de la forma  $H_1 : \mu < \mu_0$ , el **p-valor** es el área a la izquierda del estadístico del contraste.

```
plotDist("t", df=20, groups = x >-2.08 , type="h")
```



## 8.7 Contraste para una proporción

Ahora, consideramos  $p$  la proporción teórica de ocurrencia de un suceso en una población (proporción de voto a un partido, proporción de gente enferma de amor...). Establecemos la siguiente hipótesis nula, de que  $p$  es igual a un número concreto:

$$H_0 : p = p_0$$

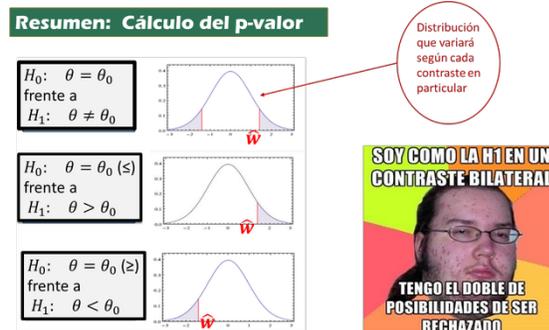


Figura 8.9: Resumen de tipos de contrastes y cálculo del  $p$ -valor.

El estadístico “pivote” para este contraste es:

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}} \approx N(0, 1).$$

**Ej.**

El portavoz del gobierno de BestKorea ha dicho que más de la mitad de la población está de acuerdo con la aplicación del artículo 155 en una república autónoma rebelde. Una televisión independiente (que no independentista) decide realizar una encuesta. De 288 personas encuestadas, 155 son favorables a la aplicación del artículo 155. ¿Ponen en duda estos resultados la publicidad del gobierno?

• **Solución:**

Llamamos  $p$  = proporción de personas a favor del 155.

El gobierno afirma que  $p > 0.5$ , luego, al no llevar el signo =, esta opción irá en la hipótesis alternativa. Así, el test será:

$$H_0 : p \leq 0.5 \text{ frente a } H_1 : p > 0.5.$$

La forma de tomar la decisión de rechazar la hipótesis nula o no es ver si hay mucha diferencia entre lo que dice la muestra y lo que dice la hipótesis nula.

La hipótesis nula dice que la proporción teórica  $p_0$  es 0.5. La muestra nos da una proporción muestral  $\hat{p} = 155/288 = 0.538$ . Obviamente, este valor difiere

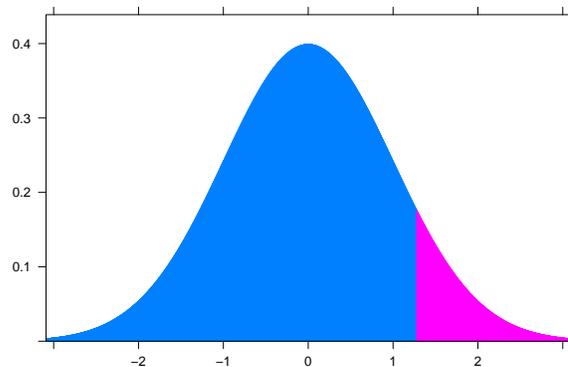
de 0.5, pero ¿mucho, poco, regular? Es el mismo caso que se nos planteaba arriba en el ejemplo de la película de James Bond.

La forma de “medir” la diferencia es mediante el estadístico  $T$  que, en este caso, es

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}} = \frac{0.538 - 0.5}{\sqrt{\frac{0.5 \cdot (1-0.5)}{n}}} = \frac{0.038}{0.029} = 1.28$$

El estadístico que se utiliza sigue una distribución aproximadamente normal, quiere esto decir que el  $p$ -valor no será tan exacto, sino que hay una cierta diferencia, pero nada para asustar.

```
plotDist("norm", groups = x > 1.28 , type="h")
```



```
1-pnorm(1.28)
```

```
## [1] 0.1003
```

Observamos que el  $p$ -valor es, aproximadamente, el área a la derecha de 1.28, que es 0.1. En función de ese valor se decide.

Para realizar este test en R, en el paquete básico existe el procedimiento siguiente:

```
prop.test(155, 280, 0.5, alternative="greater")
```

```
##  
## 1sample proportions test with continuity  
## correction  
##  
## data: 155 out of 28  
## Xsquared = 1.5, df= 1, pvalue = 0.1  
## alternative hypothesis: true p is greater than 0.5  
## 95 percent confidence interval:  
## 0.481 0.000  
## sample estimates:  
##      p  
## 0.532
```

Observamos que no sale exactamente lo mismo que lo que hemos realizado nosotros. La cuestión es que el procedimiento que usa R utiliza un estadístico diferente (compara la proporción muestral con la proporción teórica mediante otra fórmula distinta). Al utilizar otra fórmula, la distribución que sigue ese estadístico no es una distribución normal. En este caso, es una distribución Chi-cuadrado. De todos modos, lo que importa es que el  $p$ -valor va a salir igual o muy parecido, que es lo que ocurre.

Tenemos la opción de usar la función `z.test` que se definió en el capítulo anterior:

```
y= z.test(155,28 p=0.5,alternative="greater")
```

El  $p$ -valor del test es 0.0974, que es mayor que el nivel habitual de 0.05, luego no se podría rechazar la hipótesis nula.



Figura 8.10: Siempre de buen humor, qué gran tipo.

## 8.8 Contrastes para comparación de poblaciones

En muchas ocasiones interesa contrastar si los parámetros de dos poblaciones son iguales o no lo son. Si queremos estudiar si hay diferencias entre la estatura de los hombres (o mujeres) de una población (España, por ejemplo) con otra (Francia), teniendo en cuenta que la estatura es una variable aleatoria que sigue una distribución normal, lo que tendremos que comparar son sus parámetros media y desviación típica

Si se desea comparar las diferencias que puedan existir entre dos tratamientos que se apliquen para una enfermedad, consideraremos dos muestras de pacientes a los que, de forma aleatoria, les aplicaremos uno u otro tratamiento, y después consideraremos la efectividad de cada una de los tratamientos.

De manera formal, tendremos dos variables  $X = \text{efecto del tratamiento 1}$  e  $Y = \text{efecto del tratamiento 2}$ . Este efecto podrá ser, por ejemplo, la disminución de dolor, disminución de temperatura, aumento de movilidad, etc.

Para contrastar cual de los dos es mejor, puede realizarse un contraste de igualdad de medias, de la forma.  $H_0 : \mu_1 = \mu_2$  frente a  $H_1 : \mu_1 \neq \mu_2$ .

El supuesto más habitual es considerar que las variables aleatorias  $X$  e  $Y$  son variables aleatorias que siguen una distribución normal. Esto es

$$X \in N(\mu_1, \sigma_1), \quad Y \in N(\mu_2, \sigma_2)$$

Un contraste de diferencia de medias nos sirve, entonces, para comprobar si hay diferencia o no entre el efecto medio de los tratamientos. Esta igualdad de medias (o no) es equivalente a considerar si la diferencia entre las medias es igual (o no) a cero.

$$H_0 : \mu_1 = \mu_2 \quad \text{o} \quad \mu_1 - \mu_2 = 0$$

El estadístico que se utiliza es el siguiente:

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_{n-1}^2}{n} + \frac{\hat{S}_{m-1}^2}{m}}} \in t_{n+m-2-\Delta},$$

siendo  $\Delta$  el entero más próximo a

$$\frac{\left( (m-1) \frac{\hat{S}_{n-1}^2}{n} - (n-1) \frac{\hat{S}_{m-1}^2}{m} \right)^2}{(m-1) \left( \frac{\hat{S}_{n-1}^2}{n} \right)^2 + (n-1) \left( \frac{\hat{S}_{m-1}^2}{m} \right)^2}$$

Ej.

Se quiere estudiar qué aumenta más la presión sanguínea: (1) resistir las tonterías del cuñado en la comida de navidad, o (2) estar esperando ansioso por un paquete pedido por Internet y, luego de estar toda la tarde en casa, descubrir que el repartidor ha pegado un papel en tu buzón donde dice “Ausente en el momento del reparto”.

Para ello se seleccionan dos grupos, se les somete a la tortura explicada, y luego se obtienen las presiones sistólicas en el momento de finalizar la sesión:

**Grupo 1:** 104,88,100, 98,102,92,96,100,96,96

**Grupo 2:** 100,102,96,106,110,110,120,112,112,90

¿Puede considerarse que las presiones medias son iguales en ambos casos?

En R, así de sencillo:

```
x=c(104,88,100,98,102,92,96,100,96,96)
y=c(100,102,96,106,110,110,120,112,112,90)
t.test(x,y)

##
## Welch Two Sample ttest
##
## data: x and y
## t = 2.7, df = 14, pvalue = 0.02
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  15.429  1.771
## sample estimates:
```

```
## mean of x mean of y
##      97.2      105.8
```

Como el  $p$ -valor es más pequeño que 0.05, que es el que se usa habitualmente para decidir, diríamos que las presiones medias no pueden considerarse iguales.

## 8.9 Para el cociente de varianzas

Si se quiere contrastar que la proporción de las varianzas de las 2 variables  $X$  e  $Y$  es igual a un número concreto (una constante):

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = cte,$$

se utiliza el siguiente estimador:

$$T = \frac{\hat{S}_{n-1}^2 \sigma_2^2}{\hat{S}_{m-1}^2 \sigma_1^2} \in F_{n-1, m-1}.$$

Fijémonos que, si lo que interesa es contrastar si las 2 variables tienen la misma varianza, entonces esa constante es igual a 1, de manera que el estadístico será:

$$T = \frac{\hat{S}_{n-1}^2}{\hat{S}_{m-1}^2} \in F_{n-1, m-1},$$

esto es, hacer el cociente entre las cuasi-varianzas muestrales. Ese estadístico (si la hipótesis nula de que las varianzas teóricas son iguales es cierta) sigue una distribución  $F$  de Fisher-Snedecor de  $n - 1$  y  $m - 1$  grados de libertad.

**Ej.**

Con los datos del ejercicio anterior, contrastar si las varianzas de las variables originales pueden considerarse iguales o no.

```
x=c(104, 8100, 98102, 92, 96100, 9698)
y=c(100, 102, 96106110, 110, 120, 112, 112, 90)
var.test(x, y)
```

```
##
## F test to compare two variances
##
## data: x and y
## F = 0.28 num df= 9, denom df= 9, pvalue =
## 0.08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.07077 1.14703
## sample estimates:
## ratio of variances
## 0.289
```

En este caso, el  $p$ -valor del test es 0.075. Atendiendo al valor estándar 0.05 de decisión, no podríamos decir que las varianzas son diferentes.

### 8.9.1 El poder de los gráficos

Hoy en día, con la potencia gráfica de que disponemos con R y programas similares, es absurdo no acompañar este tipo de análisis con una comparativa gráfica que nos ayude en la toma de decisiones.

Para ver si las dos variables que generan las muestras que tenemos tienen medias y/o varianzas similares podemos hacer, simplemente, una comparación de las estimaciones de la densidad:

```
x=c(104, 100, 98, 102, 92, 96, 100, 96, 98)
y=c(100, 102, 96, 106, 110, 110, 120, 112, 112, 90)
m1=length(x) # tamaño muestra x
m2=length(y) # tamaño muestra y
m=c(rep(1, m1), rep(2, m2))
m=as.factor(m)
datos<-data.frame(m, c(x, y))
names(datos)<-c("muestra", "variable")
library(ggplot2)
ggplot(datos) + geom_density(aes(x=variable), adjust=2) +
  aes(color = muestra)
```

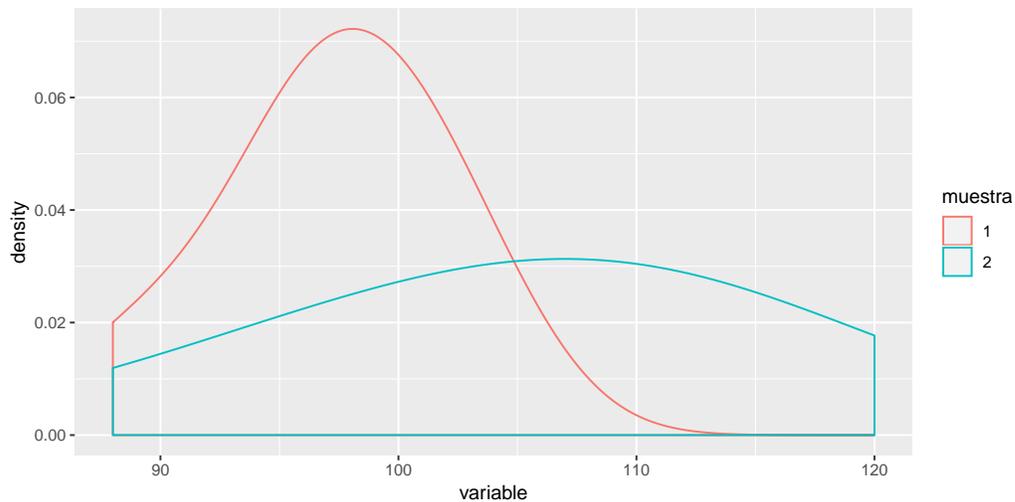


Figura 8.11: Densidades de las variables  $x$  e  $y$ .

Como observamos en la gráfica (Figura 8.11), tanto las medias como las varianzas (variabilidad) son diferentes.

Hagamos ahora un ejemplo con un fichero de datos real, que ya hemos utilizado anteriormente. El fichero “body\_dat.csv”, que utilizamos en el capítulo de la distribución normal, contiene datos antropométricos de hombres y mujeres. Vamos a realizar una comparación de la variable estatura (height),

```
body <-read_csv("Data/body_dat.csv")

body <-within(body, {
  sexo <-factor(sexo, labels=c('mujer', 'hombre'))
})

N=nrow(body)
y=rep(0, N)
y1<data.frame(y)
names(y1)<"ceros"
body2 <bind_cols(body, y1)

ggplot(body2)+
```

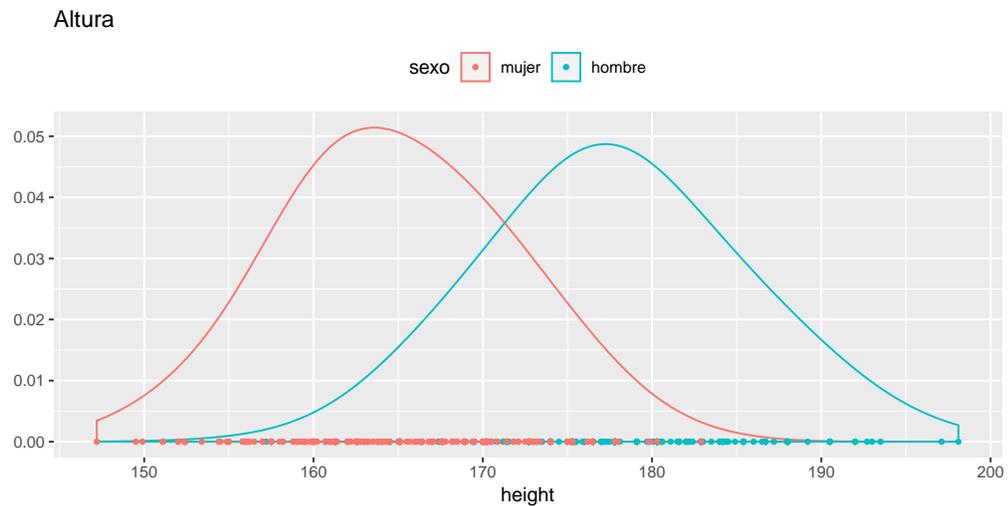


Figura 8.12: Densidad de la estatura de hombres y mujeres.

```
geom_density(aes(x=height), adjust=2) +
aes(colour = sexo) + labs(title="Altura", y=" ") +
theme(legend.position = "top") +
geom_point(aes(x=height, y=ceros), size=1)
```

Gráficamente (Figura 8.12), observamos que las medias difieren, pero las varianzas no lo parecen.

```
x1=body2$height[body2$sexo=="mujer"]
x2=body2$height[body2$sexo=="hombre"]
t.test(x1,x2)$p.value
```

```
## [1] 9.217e-71
```

```
var.test(x1,x2)$p.value
```

```
## [1] 0.139
```

## 8.10 Muestras pareadas o relacionadas

Si suponemos 2 variables o poblaciones  $X$  e  $Y$ , pero dependientes, estaremos en el caso de muestras o variables **apareadas**. Este caso recoge el ejemplo del estudio del efecto de un tratamiento: para saber si un nuevo tratamiento es efectivo sobre un cierto factor (dolor, temperatura, movilidad), se prueba en un grupo de personas y se miden los efectos *antes* y *después* del mismo. La variable  $X$  representa la medición del factor de interés en una muestra de  $n$  pacientes antes del tratamiento, y la variable  $Y$  representa la medición del mismo factor después de realizado el tratamiento. La diferencia entre las mediciones *después* y *antes*,  $D = Y - X$  es una nueva variable medida en una muestra de tamaño  $n$ :

$$(d_1, d_2, \dots, d_n) = (y_1 - x_1, y_2 - x_2, \dots, y_n - x_n)$$

Con lo cual, para saber si el tratamiento ha sido efectivo, plantearemos la hipótesis nula de que el efecto medio es cero (el tratamiento no sirve):

$$H_0 : \mu_d = 0 \text{ frente a } H_1 : \mu_d < 0 \text{ o } H_1 : \mu_d > 0,$$

según sea la dirección que nos interesa. Por ejemplo, si tenemos interés en saber si el tratamiento reduce el dolor (hemos medido el dolor antes y después), la hipótesis alternativa será  $H_1 : \mu_d < 0$ .

Si estamos interesados en saber si, por ejemplo, un programa de ejercicio físico aumenta la resistencia, la hipótesis alternativa será  $H_1 : \mu_d > 0$  (la resistencia después del tratamiento, en media, es mayor que antes del tratamiento).

**Ej.**

Unos investigadores están tratando de descubrir la fórmula de la poción mágica que utilizaba el druida Panorámix de los libros de Astérix el galo, a partir de unas pócimas encontradas a través de internet. Para probar si es eficaz o no, plantean el siguiente experimento: contratan a 8 grandes luchadores de artes marciales mixtas, y controlan el tiempo que tardan en levantarse después de recibir una patada giratoria de Chuck Norris. Al cabo de un mes de la experiencia, se toman la pócima y se repite la prueba. El número de horas que tardaron en despertarse la primera vez, para los luchadores numerados del 1 al 8, respectivamente, fueron:

sin poción: 38 32 41 35 42 32 45 37

y, para cada luchador, respectivamente, tras tomar la supuesta poción:

con poción: 30 32 34 37 35 26 38 32

¿Es efectiva la nueva poción mágica?

### Solución:

Se quiere saber si el número medio de horas que tardaron en levantarse fue inferior cuando los luchadores tomaron la supuesta poción mágica, es decir, hay que plantear el contraste

$$H_0 : \mu_X = \mu_Y \text{ frente a } H_1 : \mu_X > \mu_Y.$$

( $X$  mide las horas durmiendo sin poción,  $Y$  con poción).

En este caso, las muestras son dependientes (muestras relacionadas o apareadas), porque son los mismos luchadores. Hay que calcular las diferencias entre los datos de una muestra y la otra:

$$d = (d_1 = x_1 - y_1, d_n = x_n - y_n) = (8, 0, 7, -2, 7, 6, 7, 5),$$

y ahora el contraste a realizar es

$$H_0 : \mu_d = 0 \text{ frente a } H_1 : \mu_d > 0.$$

Lo que es, simplemente, un test para la media de una muestra.

```
x=c(8,0,7,-2,7,6,7,5)
y=t.test(x, mu=0, alternative="greater")
```

El  $p$ -valor del contraste es 0.0042. Así, no aceptaríamos la hipótesis nula, o sea que los investigadores van bien encaminados con su fórmula.

## 8.11 Para la diferencia de proporciones

Ahora consideramos la hipótesis nula de que la proporción  $p_1$ , que mide la proporción de una característica  $A$  en una población, es igual a  $p_2$ , que es la proporción de la misma característica en otra población.

$$H_0 : p_1 = p_2 \text{ o } p_1 - p_2 = 0$$

El estadístico para este caso se escribe, a partir de las proporciones muestrales  $\hat{p}_1$  y  $\hat{p}_2$  en dos muestras, como:

$$T = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \approx N(0, 1).$$

**Ej.**

Para comprobar si interceder ante el altísimo (rezar) por una persona enferma influye o no en una más rápida curación, se llevó a cabo el siguiente experimento: los pacientes que iban a ser operados de cirugía de injerto de derivación arterial coronaria (CABG) serían informados de que habría una serie de voluntarios rezando por una pronta recuperación, de la siguiente forma: 1201 pacientes fueron informados de que podrían recibir o no oraciones intercediendo por ellos. Estos pacientes se dividieron en 2 grupos A y B de 604 y 597 personas, que recibieron ayuda en forma de rezos (grupo A) y no la recibieron (grupo B) (Es decir, cada persona de cada grupo no sabía si recibía esa ayuda o no).

Un tercer grupo de 601 pacientes fueron informados de que recibirían oraciones intercediendo por ellos, y todos recibieron ayuda de esa manera. Las oraciones duraron 14 días, empezando a la noche antes de la intervención.

Al cabo de 30 días de la intervención se controló la presencia de complicaciones. Los resultados finalmente fueron

Complicaciones en el grupo A = 315.

Complicaciones en el grupo B = 304.

Complicaciones en el grupo C = 352.

¿Qué conclusiones podemos sacar de estos resultados?



Este estudio puede leerse con detalle en: “Study of the Therapeutic Effects of Intercessory Prayer (STEP) in cardiac bypass patients: A multicenter randomized trial of uncertainty and certainty of receiving intercessory prayer”. *American Heart Journal*, 2006 151(4):934-42.

Empecemos comparando los que recibieron ayuda con los que no (grupos A y B, respectivamente) y ellos no lo sabían. Las proporciones muestrales son  $\hat{p}_1 = 315/604 = 0.52$ ,  $\hat{p}_2 = 304/597 = 0.50$ . El contraste que se plantea es  $H_0 : p_1 = p_2$  frente a  $H_1 : p_1 \neq p_2$ .

Para realizar el contraste usamos la función `z.test.2` que se definió en el capítulo anterior:

```
z.test.2(315,604,304,597,alternative="two.sided")
```

```
## $estimate
## [1] 0.01231
##
## $ts.z
## [1] 0.428
##
## $p.val
## [1] 0.05
##
## $cint
## [1] 0.04422 0.06
```

Comprobamos que el  $p$ -valor para este contraste es `z.test.2$p.val`, lo cual significa que no podemos rechazar la hipótesis nula. Esto quiere decir que rezar (o no rezar) por una tercera persona es indiferente (si esa tercera persona no lo sabe).

Ahora comparemos los grupos A y B (no saber si rezan por ti) en conjunto, con el C (sabes que rezan por ti).

Las proporciones muestrales son:

$$\hat{p}_1 = \frac{315 + 304}{604 + 597} = \frac{619}{1201} = 0.51 \text{ y } \hat{p}_2 = \frac{352}{601} = 0.58.$$

De nuevo, planteamos el contraste  $H_0 : p_1 = p_2$  frente a  $H_1 : p_1 \neq p_2$ .

Ahora:

```
z.test.2(69,1201,352,61,alternative="two.sided")
```

```
## $estimate
## [1] 0.07029
##
## $ts.z
## [1] 2.82
##
## $p.val
## [1] 0.00446
##
## $cint
## [1] 0.11860.0218
```

Se obtiene que el  $p$ -valor es prácticamente cero, es decir, que se puede rechazar la hipótesis de que las proporciones no son iguales. De hecho, si planteamos como hipótesis alternativa la unilateral:  $H_1 : p_1 < p_2$ , el  $p$ -valor sigue siendo muy pequeño, con lo que convendríamos que *es mejor que sepas que rezan por tí, que no sepas si lo hacen o no.* ## Algunos contrastes no paramétricos

### 8.11.1 Contrastes de normalidad

Si se dispone de una muestra  $(x_1, \dots, x_n)$  aleatoria de una variable  $X$  y se quiere comprobar que esta variable sigue una distribución normal, es decir se plantea el contraste:

$H_0 : X$  sigue una distribución normal, frente a  $H_1 : X$  no sigue una distribución normal,

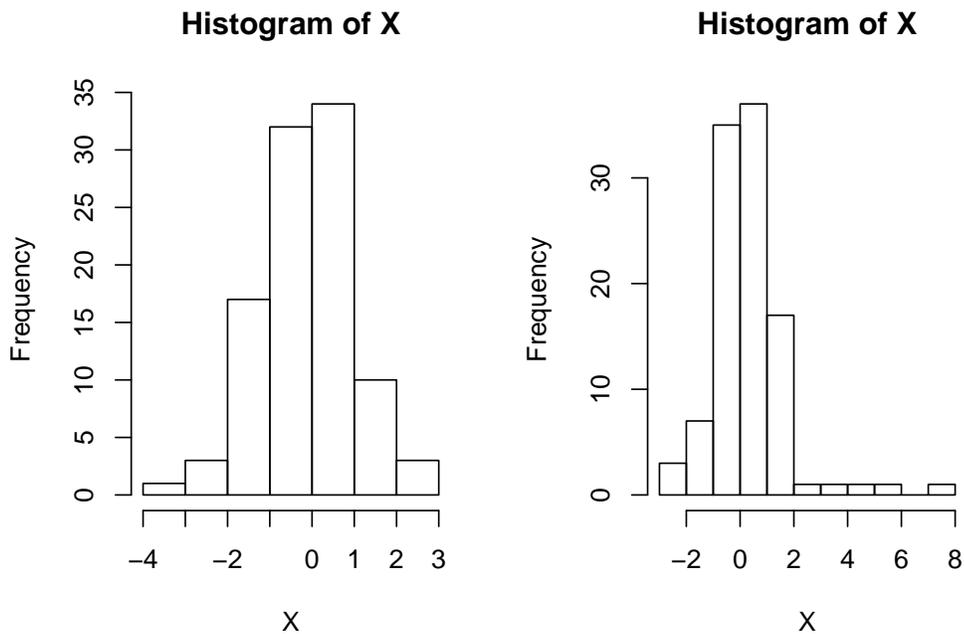
lo más fácil que puede hacerse es un histograma o un gráfico de la estimación de la densidad, y ver si se parece a la campana de Gauss.

Ejemplo:

```

X=rnorm(100)
op<-par(mfow=c(1,2))
hist(X)
X=c(rnorm(100),4,5,6,8)
hist(X)

```



```
par(op)
```

Primero hemos simulado 100 datos de una distribución normal de media cero y desviación típica 1, y el histograma realizado se parece bastante a la campana de Gauss. Sin embargo, a continuación incluimos en el conjunto X los datos 4, 5, 6 y 8, que serían datos atípicos o anómalos, y el histograma deja de parecerse a la campana de Gauss.

Sin embargo, también existen procedimientos de cálculo para llevar a cabo un contraste de normalidad, como son el contraste de Kolmogorov-Smirnov (KS test), el de Anderson-Darling (AD test) o el de Shapiro-Wils (shapiro test).

Este último test se encuentra directamente en el paquete base, los otros dos los podemos encontrar en el paquete **nortest**.

```
x=rnorm(100)
lillie.test(x )
```

```
##
##  Lilliefors (KolmogorovSmirnov) normality test
##
## data:  x
## D = 0.06, pvalue = 0.4
```

```
ad.test(x)
```

```
##
##  AndersonDarling normality test
##
## data:  x
## A = 0.32, pvalue = 0.5
```

```
shapiro.test(x)
```

```
##
##  ShapiroWilk normality test
##
## data:  x
## W = 0.99, pvalue = 0.7
```

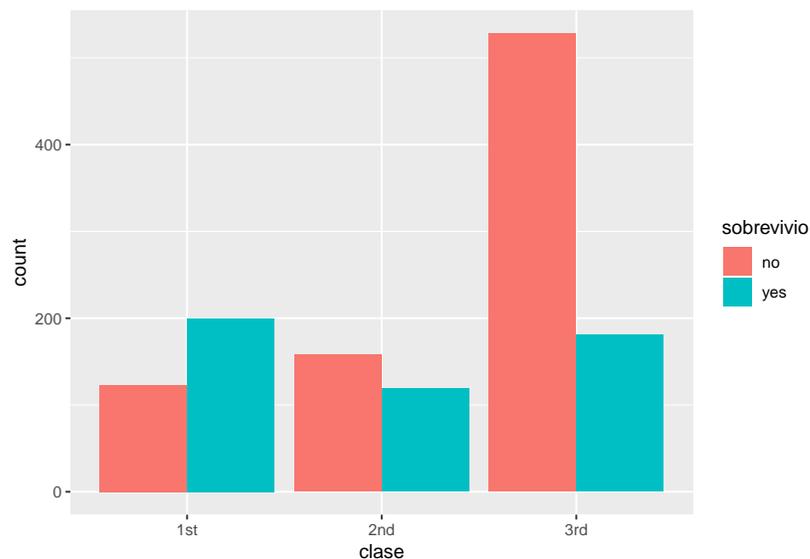
### 8.11.2 Contrastes de independencia entre caracteres

En el capítulo de estadística bidimensional se vieron las tablas de contingencia para atributos o caracteres. Vamos a recordar el ejemplo en el que usamos el conjunto de datos de pasajeros del Titanic, y comparabamos los supervivientes con la clase en que viajaban

```
Titanic <- read_excel("Data/PasajerosTitanic.xlsx")
t1 <- table(Titanic$sobrevivio, Titanic$clase)
t2 <- addmargins(t1)
pander(t2)
```

	1st	2nd	3rd	Sum
no	123	158	528	809
yes	200	119	181	500
Sum	323	277	709	1309

```
dt <- data.frame(Titanic)
ggplot(dt, aes(x=clase)) +
  geom_bar(aes(fill=sobrevivio), position="dodge")
```



Gráficamente, se intuye que las variables “sobrevivir” y “clase en la que viajaba” están relacionadas, pues no hay independencia entre las mismas.

De manera general, podemos plantear, para 2 variables categóricas  $X$  e  $Y$  de un fichero de datos, el test de hipótesis:

$H_0$  :  $X$  e  $Y$  son independientes, frente a  $H_1$  : no lo son.

Esto puede realizarse mediante el llamado test  $\chi^2$  de Pearson o el test exacto de Fisher;

```
t2=table(Titanic$sobrevivio, Titanic$clase)
pander(t2)
```

	1st	2nd	3rd
no	123	158	528
yes	200	119	181

```
chisq.test(t2)
```

```
##
## Pearson's Chi-squared test
##
## data:  t2
## X-squared = 130, df= 2, p-value <2e-16
```

```
fisher.test(t2)
```

```
##
## Fisher's Exact Test for Count Data
##
## data:  t2
## p-value <2e-16
## alternative hypothesis: two.sided
```

Ej.

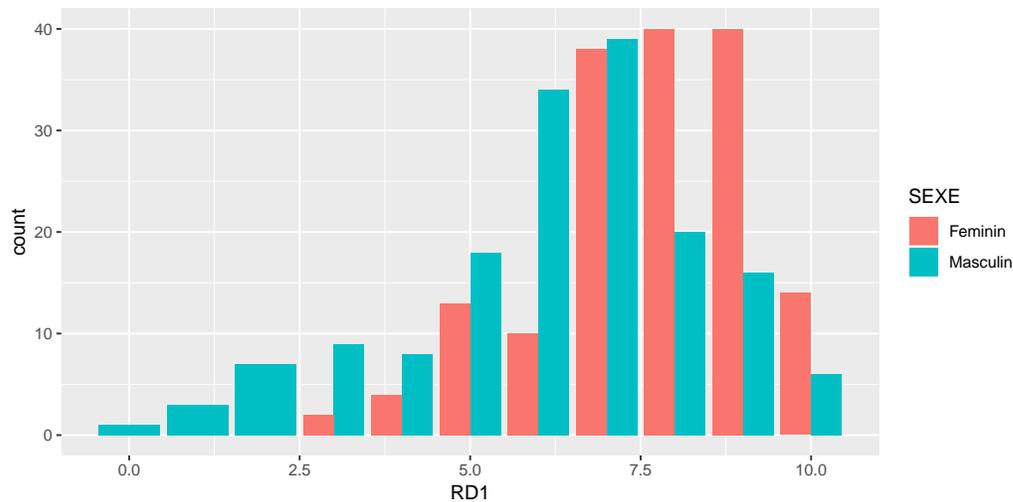
En el capítulo de la variable aleatoria normal describimos el trabajo de (Hansenne et al., 2001), donde se detalla un cuestionario de 226 items de el *Inventario de Temperamento y Carácter* para evaluar las siete dimensiones de la personalidad descritas por Cloninger (Cloninger et al., 1994). El cuestionario fue respondido por 322 ciudadanos Belgas. La variable

RD1 mide la Sentimentalidad. Vamos a ver si hay relación entre esta variable y el Sexo.

```
library(readxl)
hansenne <- read_excel("Data/20011701_hansenne/data.xls")
t1 <- table(hansenne$RD1, hansenne$SEXE)
t2 <- addmargins(t1)
pander(t2)
```

	Feminin	Masculin	Sum
<b>0</b>	0	1	1
<b>1</b>	0	3	3
<b>2</b>	0	7	7
<b>3</b>	2	9	11
<b>4</b>	4	8	12
<b>5</b>	13	18	31
<b>6</b>	10	34	44
<b>7</b>	38	39	77
<b>8</b>	40	20	60
<b>9</b>	40	16	56
<b>10</b>	14	6	20
<b>Sum</b>	161	161	322

```
dt <- data.frame(hansenne)
ggplot(dt, aes(x=RD1)) +
  geom_bar(aes(fill=SEXE), position="dodge")
```



```
chisq.test(t2)
```

```
##
## Pearson's Chi-squared test
##
## data:  t2
## X-squared = 51, df= 22, p-value = 4e-04
```

Gráficamente podemos intuir que existen diferencias por sexo. El contraste Chi-cuadrado da un  $p$ -valor muy próximo a cero, con lo que la independencia entre las variables Sexo y RD1 se rechazarían.

## 8.12 Problemas del nivel de significación

En el siguiente artículo del periódico digital [El Confidencial](#) podemos leer

“Los investigadores no saben estadística (y eso perjudica a la ciencia)”

En el interior de la noticia, podemos leer que

“la estadística es la herramienta más importante de que dispone la ciencia para confirmar o refutar hipótesis. Ahora, investigadores denuncian que su uso incorrecto, por mala fe o ignorancia,

amenaza la calidad de la investigación. El culpable de todos los males es el  $p$ -valor, un término estadístico cuya popularidad ha aumentado con los años al mismo ritmo que su mal uso.”

Más adelante, comenta

“Un estudio publicado esta semana en la revista ‘JAMA’ ha revisado” millones” de estudios biomédicos fechados entre 1990 y 2015 para concluir que la ‘mala estadística’ es cada vez más empleada. “El  $p$ -valor ya es una técnica subóptima, y si encima se usa de una forma sesgada puede ser muy confusa”, asegura el director del Centro de Investigación Preventiva de la Universidad de Stanford y autor principal del estudio, John Ioannidis.”

Cualquier interesado, a estas alturas, ya ha acudido a la fuente original y leído el artículo entero. Vamos a tratar en las siguientes líneas de aclarar algunas de los problemas que puede presentar el  $p$ -valor (si no se utiliza correctamente).

Hemos visto que el test  $t$  para la media de una muestra utiliza el estadístico

$$T = \frac{\bar{x} - \mu_o}{\hat{S}_{n-1}/\sqrt{n}} \in t_{n-1}.$$

A partir de una muestra, calculamos la media muestral  $\bar{x}$ , la cuasi-desviación típica muestral  $\hat{S}_{n-1}$  y hacemos la cuenta:

$$T = \frac{\bar{x} - \mu_o}{\hat{S}_{n-1}/\sqrt{n}} = \sqrt{n} \cdot \frac{\bar{x} - \mu_o}{\hat{S}_{n-1}}$$

En muchas ocasiones, si se pretende probar, por ejemplo, si un nuevo tratamiento es eficaz, el contraste es de la forma  $H_0 : \mu = 0$  frente a  $H_1 : \mu \neq 0$ , ya que obtendremos una muestra de pacientes y mediremos una variable antes de efectuar el tratamiento (dolor, temperatura, horas de sueño... ) y después de realizar el tratamiento. Si el tratamiento es efectivo las diferencias de los valores antes y después serán diferentes a cero (el dolor ha disminuido, o la temperatura, o las horas de sueño han aumentado...).

En este caso el estadístico  $T$  es

$$T = \frac{\bar{x} - 0}{\hat{S}_{n-1}/\sqrt{n}} = \sqrt{n} \cdot \frac{\bar{x}}{\hat{S}_{n-1}}$$

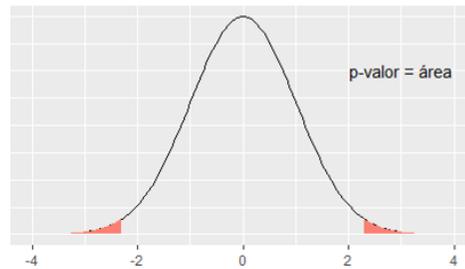


Figura 8.13:  $p$ -valor para un test  $t$  bilateral.

Si el tamaño de la muestra es grande, este valor también lo va a ser (puesto que se supone que, tomando muestras diferentes, la media y la cuasi-desviación típica muestral nos darán siempre valores parecidos). Es decir, el valor de  $T$  aumenta con el tamaño de la muestra, y recordemos que el  $p$ -valor es el área a la derecha del valor  $T$  (o dos veces ese valor). Esto significa que el  $p$ -valor va a ser próximo a cero, y la hipótesis nula siempre se va a rechazar.

La conclusión es que, para demostrar que un tratamiento es efectivo, no es necesario que lo sea, sino simplemente hay que hacer un ensayo clínico con un tamaño de muestra lo suficientemente grande.

Esto ocurre porque el test de hipótesis mide la diferencia entre la media muestral  $\bar{x}$  y el valor cero. Cuando el tamaño de la muestra es muy grande, es como si dispusiésemos de toda la población, con lo cual cualquier diferencia de  $\bar{x}$  con el valor cero se hace significativa (tenemos toda la población y queremos saber si la media es cero. Si obtenemos un valor distinto de cero, aunque sea muy poco distinto, rechazaremos que sea cero).

En su libro *Métodos estadísticos para investigadores* (Fisher, 1992), Ronald Fisher escribió:

“Personally, the writer prefers to set a low standard of significance at the 5 percent point... A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance”

“Un hecho científico se considerará probado experimentalmente sólo si un experimento correctamente diseñado raramente falla en dar este nivel de significación. Un cinco por ciento significaría que el experimento debería repetirse y, de cada cien repeticiones,

debería arrojar valores significativos en, al menos, 95 de cada 100 veces”. De esta manera quedaría demostrado que los efectos observados no serían a causa del azar. En definitiva, un  $p$ -valor más pequeño de 0.05 haría al experimento **digno de atención**, y debería repetirse para tener más certeza sobre el mismo.

El problema surge cuando un experimento es costoso o involucra seres humanos, con todos los problemas añadidos (posibles efectos secundarios, dificultad de repetición... ). Paulatinamente, se realizaron experimentos donde se obtenía un  $p$ -valor. Y había que tomar una decisión. Estamos hablando de la primera mitad del siglo XX, cuando no había ordenadores ni programas que calcularan nada. Lo que había eran tablas estadísticas para poder discernir si el  $p$ -valor era mayor o menor a algún nivel determinado. Se fijaron tres valores fundamentales:  $\alpha = 0.1, 0.05$  y  $0.01$ . Por aquello de estar en el medio, el nivel  $\alpha = 0.05$  fue el más utilizado. De manera que, si el  $p$ -valor era menor que 0.05 se rechazaba la hipótesis nula, y el tratamiento es efectivo. Si el  $p$ -valor era mayor o igual, no se podía rechazar la hipótesis nula, y por lo tanto el tratamiento no podía aceptarse como efectivo.

Esta claro que si, por ejemplo, se pone el límite de significación en 0.05, estamos afirmando que, sobre la base de que 95 veces sobre 100 esperaríamos encontrar que el tratamiento es efectivo. Pero si obtenemos  $p = 0.06$  entonces ocurre que 94 de cada cien veces esperaremos encontrar que el tratamiento es efectivo. ¿Es tanta la diferencia?

Lo que ha sucedido, históricamente, es lo que se llama el *sesgo de publicación*. Si un tratamiento resulta efectivo, es mucho más probable que sea merecedor de ser publicado en una revista científica que si no resulta efectivo. De ahí que se hayan visto todo tipo de artífugos para conseguir que un  $p$ -valor consiga que un tratamiento sea efectivo. Si en el experimento se obtuvo  $p = 0.049$  se dice  $p < 0.05$  y punto. Si no se obtuvo un valor menor que 0.05, se intenta conseguir un tamaño de muestra lo suficientemente grande para que cualquier diferencia resulte significativa (y digna de mención en alguna revista científica).

Ya en 1994, el famoso psicólogo y estadístico Jacob Cohen (1923-1998), conocido por sus estudios del tamaño del efecto y otros trabajos que ayudaron a sentar las bases para los metaanálisis, publicó un artículo enormemente citado: *The earth is round ( $p < .05$ )*, cuyo abstract, traducido, viene a decir:

“Tras cuatro décadas de severa crítica, el ritual del contraste de hipótesis (NHST) —decisiones mecánicas y dicotómicas alrededor del sagrado criterio del 0.05— todavía perdura. Este artículo repasa los problemas derivados de esta práctica, incluyendo la casi universal malinterpretación del  $p$ -valor como la probabilidad de que  $H_0$  sea falsa, la malinterpretación de su complementario como la probabilidad de una réplica exitosa y la falsa premisa de que rechazar  $H_0$  valida la teoría que condujo a la prueba. Como alternativa, se recomiendan el análisis exploratorio de datos y los métodos gráficos, la mejora y la estandarización progresiva de las medidas, el énfasis en la estimación de los tamaños de los efectos usando intervalos de confianza y el uso adecuado de los métodos estadísticos disponibles. Para garantizar la generalización, los psicólogos deben apoyarse, como ocurre en el resto de las ciencias, en la replicación.”

En la decimotercera edición de *Métodos estadísticos para investigadores*, Fisher realizó la siguiente aclaración sobre los  $p$ -valores:

“el  $p$ -valor indica la fuerza de la evidencia contra la hipótesis nula... y los contrastes de significación deben utilizarse como ayuda para el juicio, y no deben confundirse con pruebas de aceptación automática, o funciones de decisión.”

En 2016, la American Statistical Association (ASA) publicó una declaración sobre los  $p$ -valores, elaborada por un grupo de más de dos docenas de expertos (Wasserstein et al., 2016). Aunque hubo discusiones controvertidas sobre muchos temas, el informe de consenso de la ASA incluye la siguiente declaración: “El uso generalizado de la ‘significación estadística’ (generalmente interpretada como  $p < 0.05$ ) como una licencia para hacer una afirmación de un hallazgo científico (o verdad implícita) conduce a una considerable distorsión del proceso científico”

Además, un grupo de siete estadísticos de la ASA publicó en *European Journal of Epidemiology* (Greenland et al., 2016) una extensa revisión de 25 malas interpretaciones de los  $p$ -valores, los intervalos de confianza y la potencia estadística, cerrando con las palabras: “Nos unimos a otros para señalar la degradación de los  $p$ -valores en significativos y no significativos como una práctica estadística especialmente perniciosa”.

Algunos de los comentarios más relevantes inciden en que la significación estadística no puede tomarse como evidencia de que la hipótesis de investigación sea cierta; ni proporciona la probabilidad de la hipótesis, por lo que no hay base para estudiar la replicación y tampoco nos proporciona evidencias verificables de replicación.

Para evitar malas interpretaciones y ayudar en la toma de decisiones, muchos estadísticos sugieren utilizar el enfoque bayesiano, tanto en la realización de test estadísticos (test bayesianos) como en el cálculo de intervalos de confianza y el uso del factor de Bayes (Gelman et al., 2013), (Casella and Berger, 1987).

### 8.12.1 Evidencia y descubrimientos en física

Como detalle de interés, resaltamos que, desde los años 90, los experimentos de física de partículas utilizan como criterio el correspondiente a un  $p$ -valor de  $2.87 \cdot 10^{-7}$ , para refutar con seguridad una hipótesis nula. Solo en ese caso se habla de *observación o descubrimiento*. Este criterio se llama  $5\sigma$ , El criterio llamado  $3\sigma$ , correspondiente a un  $p$ -valor de  $1.35 \cdot 10^{-3}$ , se utiliza para hablar de *evidencia*. Fuente: [trabajo](#).

Hablando de forma más mundana, un nivel de significación  $\alpha = 0.05$ , que es de los más habituales en los artículos científicos, supone un criterio de  $2\sigma$ , que, recordemos que en una distribución normal (de media genérica  $\mu$  y desviación típica  $\sigma$ ), fuera del intervalo  $(\mu - 2 \cdot \sigma, \mu + 2 \cdot \sigma)$  se encuentran los valores «más raros», que vienen a ser tan sólo un 5 por ciento.

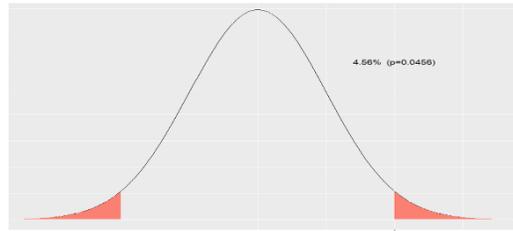
Realmente, no es un 5 por ciento, sino un 4.55 por ciento. Lo calculamos en una  $N(0, 1)$ :

```
2*(1-pnorm(2))
```

```
## [1] 0.0455
```

```
knitr::include_graphics('Figure/pvalue3.png')
```

Fuera del intervalo  $(\mu - 3 \cdot \sigma, \mu + 3 \cdot \sigma)$  el área es

Figura 8.14: Área fuera del intervalo  $(-2, 2)$ .

```
2*(1-pnorm(3))
```

```
## [1] 0.0027
```

Si lo dividimos por 2, nos da el valor  $1.35 \cdot 10^{-3}$ , que es el que citan en el trabajo (es decir, que consideran el área a la derecha de  $\mu + 3\sigma$ ).

Fuera del intervalo  $(\mu - 4 \cdot \sigma, \mu + 4 \cdot \sigma)$

```
2*(1-pnorm(4))
```

```
## [1] 6.334e-05
```

y fuera de  $(\mu - 5 \cdot \sigma, \mu + 5 \cdot \sigma)$

```
2*(1-pnorm(5))
```

```
## [1] 5.733e-07
```

que, al dividirlo por 2, nos da, aproximadamente,  $2.87 \cdot 10^{-7}$  (área a la derecha de  $\mu + 5\sigma$ ).



# Bibliografía

- Balanda, K. P. and MacGillivray, H. (1988). Kurtosis: a critical review. *The American Statistician*, 42(2):111–119.
- Ball, P. (2004). *Masa crítica. Cambio, caos y complejidad*. Turner Publicaciones.
- Barrett, A. M., Baum, S. D., and Hostetler, K. (2013). Analyzing and reducing the risks of inadvertent nuclear war between the united states and russia. *Science & Global Security*, 21(2):106–133.
- Bernardo, J. M. (1998). Bruno de finetti en la estadística contemporánea. *Historia de la Matemática en el siglo XX*, S. Rios (ed.), Real Academia de Ciencias, Madrid, pages 63–80.
- Bregman, D. J., Langmuir, A. D., et al. (1990). Farr’s law applied to aids projections. *Jama*, 263(11):1522–5.
- Brownlee, J. (1915). Historical note on farr’s theory of the epidemic. *British Medical Journal*, 2(2850):250.
- Camacho, F. G. (2002). Probabilismo y toma de decisiones en la escolástica española. In *Historia de la probabilidad y de la estadística/AHEPE*, pages 81–102.
- Caponi, S. (2013). Quetelet, el hombre medio y el saber médico. *História, Ciências, Saúde-Manguinhos*, 20(3).
- Casella, G. and Berger, R. L. (1987). Reconciling bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82(397):106–111.

- Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1997). Sistemas expertos y modelos de redes probabilísticas. *Academia de Ingeniería*.
- Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.
- Cloninger, C. R., Przybeck, T. R., Svrakic, D. M., and Wetzell, R. D. (1994). The temperament and character inventory (tc): A guide to its development and use.
- Conn, A., Pedmale, U. V., Chory, J., Stevens, C. F., and Navlakha, S. (2017). A statistical description of plant shoot architecture. *Current Biology*, 27(14):2078–2088.
- Cullen, M. J. (1975). *The statistical movement in early Victorian Britain: The foundations of empirical social research*. Harvester Press.
- del Cerro, J. S. (2002). Probabilismo moral y probabilidad. In *Historia de la probabilidad y de la estadística/AHEPE*, pages 103–118.
- del Río, A. Q. (2018). *PEPE (Problemas Estimulantes de Probabilidad y Estadística)*. Editorial CreateSpace.
- Desrosières, A. (2004). La política de los grandes números. *Historia de la razón estadística. Barcelona: Melusina*.
- Fisher, R. A. (1949). Métodos estadísticos para investigadores. Technical report.
- Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer.
- Friendly, M. (2007). A.-m. guerry’s” moral statistics of france”: Challenges for multivariable spatial analysis. *Statistical Science*, pages 368–399.
- García, A. (2012). *Inteligencia artificial: fundamentos, práctica y aplicaciones*. Rc Libros.
- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

- Good, I. J. (1979). Studies in the history of probability and statistics. xxxvii  
an turing's statistical work in world war ii. *Biometrika*, pages 393–396.
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman,  
S. N., and Altman, D. G. (2016). Statistical tests, p values, confidence  
intervals, and power: a guide to misinterpretations. *European Journal of  
Epidemiology*, 31(4):337–350.
- Hacking, I. and Bixio, A. L. (1995). *La domesticación del azar: la erosión del  
determinismo y el nacimiento de las ciencias del caos*. Gedisa Editorial.
- Hand, D., Daly, F., McConway, K., Lunn, D., and Ostrowski, E. (1993). *A  
Handbook of Small Data Sets*. Number v. 1 in A Handbook of Small Data  
Sets. Taylor & Francis.
- Hansenne, M., Le Bon, O., Gauthier, A., and Ansseau, M. (2001). Belgian  
normative data of the temperament and character inventory. *European  
Journal of Psychological Assessment*, 17(1):56.
- Joynton, R. B. (1989). *The Burt affair*. Taylor & Francis/Routledge.
- Kahneman, D. (2014). *Pensar rápido, pensar despacio / Thinking, Fast and  
Slow*. Debolsillo Mexico.
- Kahneman, D. and Egan, P. (2011). *Thinking, fast and slow*, volume 1. Farrar,  
Straus and Giroux New York.
- Kalichman, S. C. and Rompa, D. (1995). Sexual sensation seeking and sexual  
compulsivity scales: Validity, and predicting hiv risk behavior. *Journal of  
personality assessment*, 65(3):586–601.
- Klimek, P., Yegorov, Y., Hanel, R., and Thurner, S. (2012). Statistical de-  
tection of systematic election irregularities. *Proceedings of the National  
Academy of Sciences*, 109(41):16469–16473.
- Kolmogorov, A. N. and Hewitt, E. (1948). Collection of articles on the theory  
of firing.
- Kruskal, W. (1980). The significance of fisher: a review of ra fisher: the life of  
a scientist. *Journal of the American Statistical Association*, 75(372):1019–  
1030.

- Luque, B. (2013). El problema de los tanques alemanes. *Investigación y Ciencia*.
- MacKenzie, D. A. (1981). *Statistics in Britain: 1865-1930; the social construction of scientific knowledge*. Edinburgh University Press.
- Mardia, K. V. and Cooper, S. B. (2016). Alan turing and enigmatic statistics.
- Matejka, J. and Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1290–1294, New York, NY, USA. ACM.
- McGrayne, S. (2012). *La teoría que nunca murió*. Crítica.
- Mickey, M., Gjertson, D., and Terasaki, P. (1986). Empirical validation of the essen-möller probability of paternity. *American journal of human genetics*, 39(1):123.
- Montes, F. (2003). Ley y probabilidad.
- Moore, D. S. and Kirkland, S. (2007). *The basic practice of statistics*, volume 2. WH Freeman New York.
- Murray, C. and Herrnstein, R. (1994). The bell curve. *Intelligence and Class Structure in American Life*, New York.
- Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.
- Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.
- Paulos, J. and Llosa, J. (1990). *El hombre anumérico: el analfabetismo matemático y sus consecuencias*. Matatemas (Tusquets Editores). Tusquets.
- Pearl, R. (1905). Biometrical studies on man: I. variation and correlation in brain-weight. *Biometrika*, 4(1/2):13–104.
- Pearson, K. (1905). Das fehlergesetz und seine verallgemeinerungen durch fechner und pearson. a rejoinder. *Biometrika*, 4(1-2):169–212.

- Peirce, C. S. and Jastrow, J. (1884). On small differences in sensation.
- Perrot, J.-C. (1992). Une histoire intellectuelle de l'économie politique(xviie-xviiiè siècle). *Civilisations et sociétés*.
- Piovani, J. I. (2007). Los orígenes de la estadística: de investigación sociopolítica empírica a conjunto de técnicas para el análisis de datos. *Revista de Ciencia Política y Relaciones Internacionales*, 1(1):25–44.
- Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile précédées des règles générales du calcul des probabilités par SD Poisson*. Bachelier.
- Porter, T. M. (1986). *The rise of statistical thinking, 1820-1900*. Princeton University Press.
- Resnick, S. I. (2007). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.
- Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., Harris, M. A., Alderson, H., Hunter, S., Neilson, E., et al. (2017). Sex differences in the adult human brain: Evidence from 5,216 uk biobank participants. *bioRxiv*, page 123729.
- Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624.
- Ruggles, R. and Brodie, H. (1947). An empirical approach to economic intelligence in world war ii. *Journal of the American Statistical Association*, 42(237):72–91.
- Ruiz-Garzón, G. (2015). *Condenados por la estadística*. Servicio de Publicaciones de la Universidad de Cadiz.
- Salinero, P. (2006). Historia de la teoría de la probabilidad. Ver [www.uam.es/personal\\_pdi/ciencias/ezuazua/informweb/trabajosdehistoria/saline-ro\\_probabilidad.pdf](http://www.uam.es/personal_pdi/ciencias/ezuazua/informweb/trabajosdehistoria/saline-ro_probabilidad.pdf). Consultado el, 20:1–21.
- Savage, L. J. (1976). On rereading ra fisher. *The Annals of Statistics*, pages 441–500.

- Secades, M. G. (2002). Antecedentes de la concepción subjetivista de la probabilidad. In *Historia de la probabilidad y de la estadística/AHEPE*, pages 119–132.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.
- Simon, P. (1951). *A philosophical essay on probabilities*.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.
- Taleb, N. and Mosquera, A. (2011). *El cisne negro: el impacto de lo altamente improbable*. Paidós transiciones. Ediciones Paidós Ibérica, S.A.
- Trocchio, F. d. (1977). *Las mentiras de la ciencia*. Madrid: Alianza editorial.
- Tukey, J. W. (1977). *Exploratory data analysis*, volume 2. Reading, Mass.
- Villegas, M. A. G. (2001). El ensayo encaminado a resolver un problema en la doctrina del azar. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales*, 95(1):81–85.
- Wainer, H. (2007). The most dangerous equation. *American Scientist*, 95(3):249.
- Wand, M. (1997). Data-based choice of histogram bin width. *The American Statistician*, 51(1):59–64.
- Wasserstein, R. L., Lazar, N. A., et al. (2016). The ASA’s statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133.
- Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.
- Yule, G. U. (1938). Notes of Karl Pearson’s lectures on the theory of statistics, 1884–96. *Biometrika*, 30(1/2):198–203.
- Zafra, J. M. L. and de Paz Cobo, S. (2012). 7. justicia y probabilidad en la Francia de la revolución: las posturas de Condorcet, Laplace y Poisson. In *Historia de la probabilidad y de la estadística VI*, pages 159–172. Universidad Nacional de Educación a Distancia, UNED.