心理学院

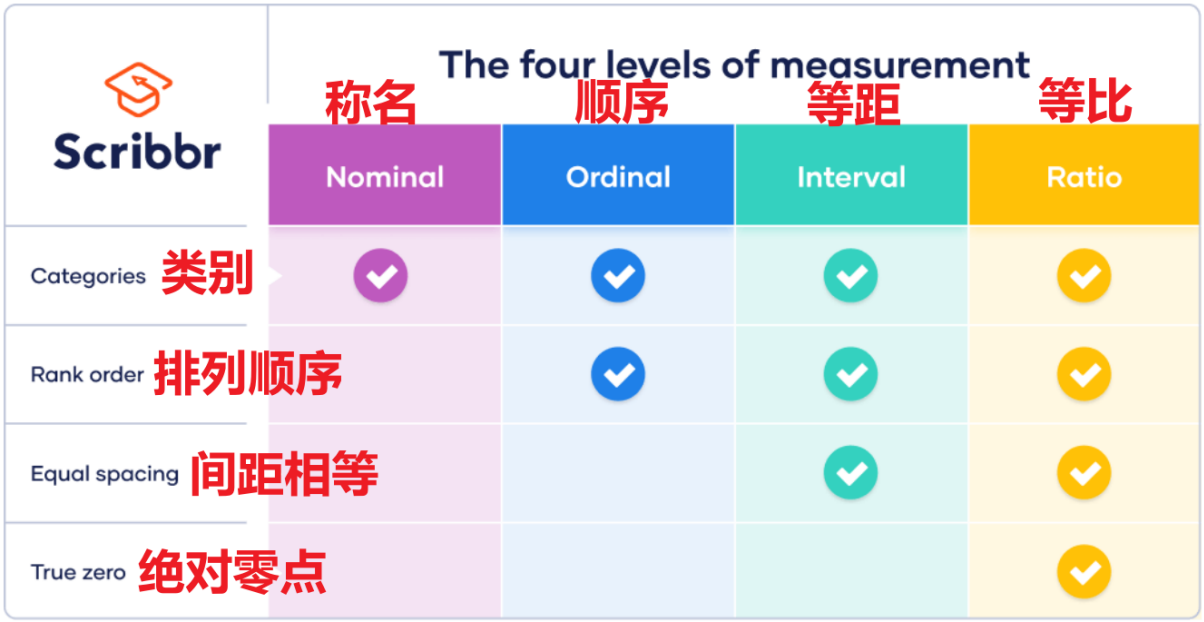
School of Psychology

第2章　数据特征的表图描述法

李喻骏

复习：基本概念

1. 变量：自变量 vs 因变量
2. 变量的操作化定义：测量
3. 数据的含义和类型：



离散	离散或连续
称名	等距
顺序	等比

4. 总体、样本和抽样
5. 描述统计和推断统计

贯穿整个教材的脉络（两条线索）

1. 研究问题从简单到复杂

- 变量数量及角色
- 数据类型
- 研究设计

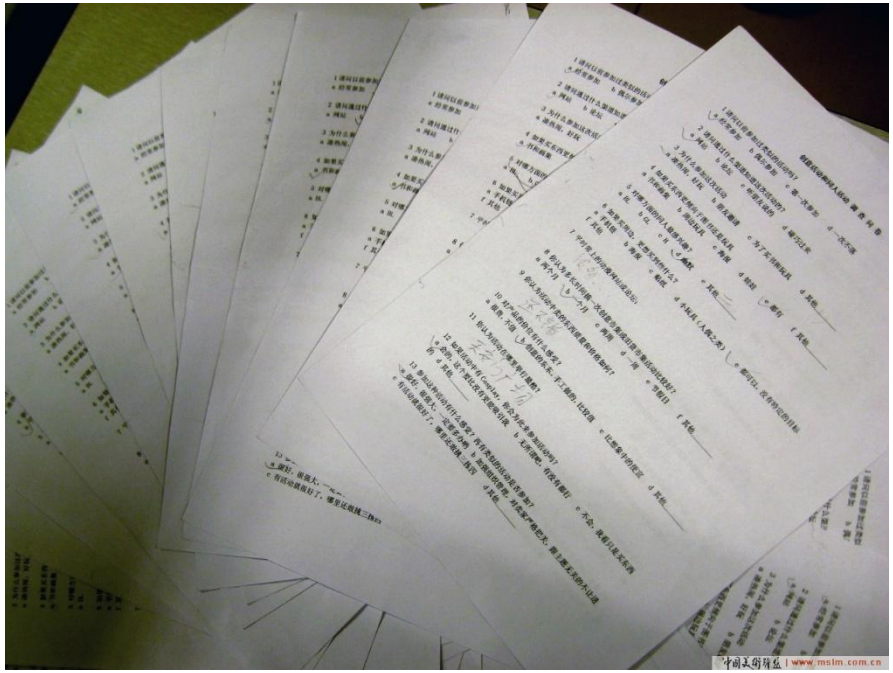
2. 分析流程从简单到复杂

- 先描述统计（了解数据概况）
- 推断统计的前提知识：概率与概率分布，假设检验与参数估计
- 推断统计（从样本提供的信息中推断总体）

第二章 数据特征的表图描述法

问题情境

- 课题组的老师带着几个同学搬来一大堆问卷，放你桌上：“来，把这批数据处理一下，写一个初步的分析报告，介绍下数据的大概情况”。
- 你该怎么完成这个任务？



第二章 数据特征的表图描述法

问题情境（续）

- 在拿到一批数据后，很自然的反应就是：看看数据是什么样的，能不能从中初步看出点什么（趋势、模式等）：

1. 数据的初步整理

1. 录入数据
2. 数据清洗（剔除无效数据、处理缺失值、处理极端值等）

2. 分析数据的概况——频数分析

一、数据的初步整理

1. 录入数据

常用范式：一行一人，一列一量

	A	B	C	D
1	编号	性别	偏好	反应时
2	1	1	2	194
3	2	0	2	190
4	3	0	5	268
5	4	1	3	976
6	5	1	4	860
7	6	1	2	309
8	7	0	3	1952
9	8	1	4	194
10	9	0	3	356
11	10	1	2	1820

一、数据的初步整理（续）

1. 录入数据（续）

问卷类

- 手动录入数据很容易出错，通常需要两个人背对背同时录一批数据，录完了以后对比；
- 在录的过程中，可以制定一定的无效数据筛选标准，不满足条件的数据不录入；
- 通过诸如“问卷星”在线发放问卷，无需录入数据，但同样需要筛选出无效数据。


题型选择
问卷大纲
选择题
单选
多选
下拉框
文件上传
填空题
单项填空
多项填空
矩阵填空
表格填空
多级下拉
签名题
地图
日期
分页说明
分页栏
一题一页
段落说明
分页计时器
矩阵题
矩阵单选
矩阵多选
矩阵量表
矩阵滑动条
表格数值
表格填空
表格下拉框
表格组合
自增表格
评分题
量表题
NPS量表
评分单选
评分多选
矩阵量表
评价题
高级题型
排序
比重题
滑动条
情景随机
商品题型
调研题型
热力图
MaxDiff
知情同意书
个人信息
test
添加问卷说明
[第1页/共1页]
批量添加题目
ID:839107
设计问卷
发送问卷
分析&下载
发布
统计&分析
查看下载答卷
来源分析
以Excel、CSV格式下载答卷数据（可以导入到SPSS分析

一、数据的初步整理（续）

1. 录入数据（续）

行为实验类

- 观察法：行为编码；
- 实验法：软件（如 E-prime）自动采集数据

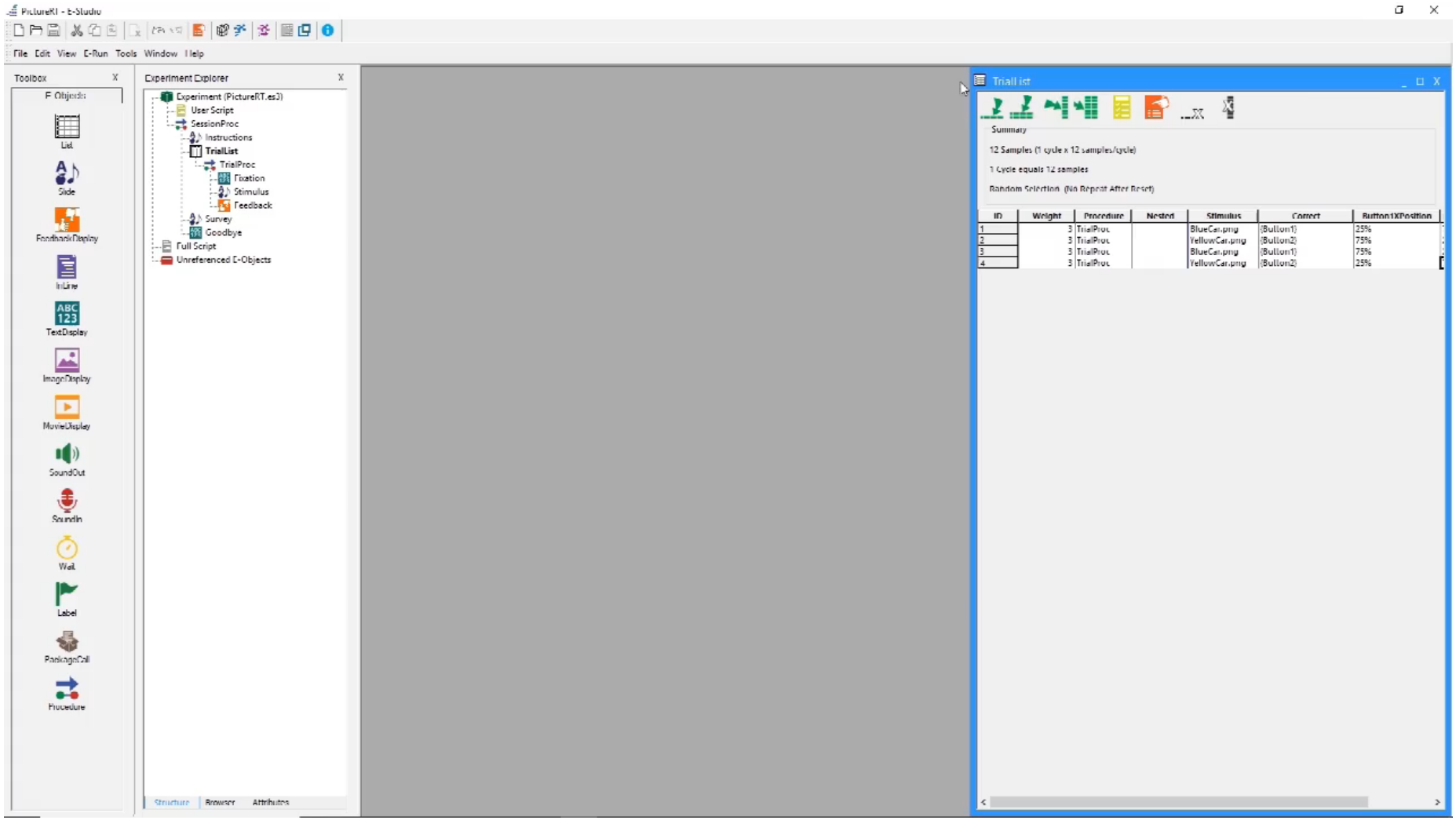
行为编码示例：男女小学数学教师行为差异研究——李姝绮，西南大学，2014

本研究运用视频分析软件（NVivo 10.0）编码分析样本教师的课堂教学视频，但因该研究方法工作量较大，受到人力和时间的限制，仅选择了 24 名教师按照性别分类比较，以探求教师教学行为的性别差异。

表 2-3　男女小学数学教师教学导入方式统计表

男教师			女教师		
编号	教学导入方式	时长(S)	编号	教学导入方式	时长（S）
M_1	复习法	18	F_1	复习法	14
M_2	复习法	21	F_2	阐述	8
M_3	复习法	15	F_3	复习法	19
M_4	设置情景法	8	F_4	设置情景法	94
M_5	设置情景法	13	F_5	复习法	22
M_6	复习法	19	F_6	复习法	15
M_7	复习法	24	F_7	设置情景法	10
M_8	设置情景法	13	F_8	复习法	10
M_9	复习法	5	F_9	设置情景法	13
M_{10}	设置情景法	33	F_{10}	复习法	14
M_{11}	设置情景法	14	F_{11}	设置情景法	28

实验法示例：E-prime



一、数据的初步整理（续）

1. 数据清洗

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. (Wikipedia: Data cleaning)

Data cleaning is time-consuming and cumbersome.

一、数据的初步整理（续）

1. 数据清洗（续）

DATA CLEANSING:
WHAT IS IT AND WHY
IS IT IMPORTANT?

MODEL CALCULATIONS

"Garbage In-garbage Out" Paradigm

一、数据的初步整理（续）

1. 数据清洗（续）

（人格）问卷类

- 剔除无效数据
  - 连续重复作答：33333333333333
  - 连续规律作答：34545634545634545456
  - 大面积空白
  - 反应时太长、太短
  - 测谎题：
    - 绝对的情境：我从来不撒谎
    - 相同但相隔很远的题目
    - 规定必须选某项
- 处理缺失值
  - 对删（pairwise deletion）
  - 列删（listwise deletion）
  - 填补（imputation）
- 处理极端值
- 合并数据
- 等等

第二章 数据特征的表图描述法

1. 数据的初步整理——录入数据、剔除无效数据、处理缺失值、处理极端值

2. 分析数据的概况——频数分析：表和图

- 从表到图
- 从单一分组标志到多个分组标志（分组表，复合表）
- 从离散到连续

二、分析数据的概况：频数分析——表

能够得出的主要结论：**谁多谁少**

频数统计表格就是**将原始数据计数**后得到的表格。

称名、称名、顺序（等距）、等比

	A	B	C	D
1	编号	性别	偏好	反应时
2	1	1	2	194
3	2	0	2	190
4	3	0	5	268
5	4	1	3	976
6	5	1	4	860
7	6	1	2	309
8	7	0	3	1952
9	8	1	4	194
10	9	0	3	356
11	10	1	2	1820

二、分析数据的概况：频数分析——表（续）

单一分组标志：统计一个变量的频数

- **离散数据**

表1 不同性别被试人数和百分比

性别	人数	百分比（%）
男	14	46.7
女	16	53.3
合计	30	100.0

表2 不同偏好程度的被试人数和百分比

程度	人数	百分比（%）
1	2	6.7
2	7	23.3
3	8	26.7
4	5	16.7
5	8	26.7
合计	30	100.0

二、分析数据的概况：频数分析——表（续）

单一分组标志（续）

- 离散数据
- 分组标志为时间时，除了能看出谁多谁少的结论，还可以看出涨跌趋势。

表 2 江西师范大学心理学院2017-2025年研究生（全日制）报考人数

年份	人数
2017	291
2018	398
2019	400
2020	490
2021	1318
2022	1396
2023	1311
2024	785
2025	584

二、分析数据的概况：频数分析——表（续）

单一分组标志（续）

- 连续数据：通过设定区间的方式将原始数据分组，使连续数据离散化，绘制次数分布表。

全距：全部数据中的最大值与最小值之差。

组距：每一组的起点数值和终点数值之差。不同组的组距可以相同也可以不同。

表3 年收入分布表

年收入	人数	百分比（%）
0-11,999	44	17.3
12,000-35,999	132	51.8
36,000及以上	79	31.0
合计	255	100.0

二、分析数据的概况：频数分析——表（续）

单一分组标志（续）：累计频数

- 针对有顺序意义的类别，还可以统计累积的频数，能够得出的结论：**有多少在以上/以下**

表4 心理系某年级英语考试成绩分布表

分组	人数	累计人数	百分比（%）	累计百分比（%）
65~69	2	2	2.5	2.5
70~74	6	8	7.5	10.0
75~79	18	26	22.5	32.5
80~84	26	52	32.5	65.0
85~89	16	68	20.0	85.0
90~94	8	76	10.0	95.0
95~99	4	80	5.0	100.0
合计		80		100.0

二、分析数据的概况：频数分析——表（续）

多个分组标志：想要统计两个或两个以上变量的频数

- 离散数据

例如，性别和不同偏好程度

表1不同性别被试人数和百分比

性别	人数	百分比（%）
男	14	46.7
女	16	53.3
合计	30	100.0

表2 不同偏好程度的被试人数和百分比

程度	人数	百分比（%）
1	2	6.7
2	7	23.3
3	8	26.7
4	5	16.7
5	8	26.7
合计	30	100.0

二、分析数据的概况：频数分析——表（续）

多个分组标志（续）

- 离散数据：方式 1，强调性别的对比

表5 不同性别选择不同偏好程度的人数和百分比

偏好	性别	人数	百分比（%）
1	男	1	3.3
	女	1	3.3
2	男	3	10.0
	女	4	13.3
3	男	5	16.7
	女	3	10.0
4	男	2	6.7
	女	3	10.0
5	男	3	10.0
	女	5	16.7

第一层

第二层

二、分析数据的概况：频数分析——表（续）

多个分组标志（续）

- 离散数据：方式 2，强调不同偏好程度的对比

表5 不同性别选择不同偏好程度的人数和百分比

性别	偏好	人数	百分比（%）
男	1	1	3.3
	2	3	10.0
	3	5	16.7
	4	2	6.7
	5	3	10.0
女	1	1	3.3
	2	4	13.3
	3	3	10.0
	4	3	10.0
	5	5	16.7

二、分析数据的概况：频数分析——表（续）

多个分组标志（续）：想要统计两个或两个以上变量的频数。放在最内层的变量通常是想重点比较的变量。

表 4　T1、T2、T3 各 LPA 模型的模型拟合指标

时间点	类别数目	AIC	BIC	aBIC	Entropy	LMRT (*p*)	BLRT (*p*)	所占比例(%)
T1	2	16547.51	16618.36	16577.06	0.90	< 0.01	< 0.01	71.3 / 28.7
	3	**15381.52**	**15479.63**	**15422.44**	**0.86**	**< 0.01**	**< 0.01**	**49.2 / 35.2 / 15.6**
	4	14698.64	14824.00	14750.93	0.88	< 0.01	< 0.01	45.8 / 32.2 / 4.4 / 17.6
	5	14480.28	14632.88	14543.93	0.83	>0.05	< 0.01	38.4 / 28.1 / 14.7 / 4.3 / 14.5
T2	2	16291.11	16361.96	16320.66	0.92	< 0.01	< 0.01	71.3 / 28.7
	3	**15070.49**	**15168.59**	**15111.41**	**0.88**	**< 0.01**	**< 0.01**	**54.2 / 32.8 / 13.0**
	4	14336.19	14461.55	14388.48	0.90	< 0.01	< 0.01	49.3 / 30.3 / 2.7 / 17.7
	5	14015.38	14167.98	14079.03	0.90	>0.05	< 0.01	46.6 / 9.6 / 14.1 / 2.2 / 27.5
T3	2	15736.23	15807.08	15765.78	0.94	< 0.01	< 0.01	76.3 / 23.7
	3	**14499.57**	**14597.67**	**14540.49**	**0.93**	**< 0.05**	**< 0.01**	**68.1 / 24.7 / 7.2**
	4	13692.59	13817.94	13744.87	0.90	< 0.05	< 0.01	53.3 / 27.6 / 15.2 / 3.9
	5	13374.71	13527.32	13438.36	0.88	>0.05	< 0.01	48.9 / 9.2 / 14.3 / 25.4 / 2.2

小学儿童数学焦虑的潜在类别转变及其父母教育卷入效应：3年纵向考察

司继伟, 郭凯玥, 赵晓萌, 张明亮, 李红霞, 黄碧娟, 徐艳丽

2022, 54 (4): 355-370. doi: 10.3724/SP.J.1041.2022.00355

三线表（three-line table）

- 心理学研究中的表格通常是采用三线表的形式，是标准形式。
- 三线表通常只有三条线：顶线、栏目线和底线。

Table 1 | Rates of subtractive changes by condition for experiments 1 to 8

Experiment[a]	Rates of subtraction by condition		Test statistic	*P* value (two-tailed)
	Control	Subtraction cue		
1	41% (40/98)	61% (60/99)	$\chi^2 = 7.72$	0.005
2, 3	28% (47/166)	43% (63/146)	$z = 2.73$	0.006
4 (improve)	21% (17/80)	48% (44/91)	$\chi^2 = 13.63$	<0.001
4 (worsen)	28% (26/92)	50% (53/106)	$\chi^2 = 9.71$	0.002
	Control	Repeated search		
5	49% (74/152)	63% (93/147)	$\chi^2 = 5.87$	0.015
	Higher cognitive load	Lower cognitive load		
6 to 8	2.45/4 ($n = 572$)	2.76/4 ($n = 581$)	$z = 2.97$	0.003

三线表不是只有三根线

表 B.1 基于原始分数的47个题目的难度和区分度的度量值

进阶层次	题目	难度	区分度		进阶层次	题目	难度	区分度	
			鉴别指数	点二列相关				鉴别指数	点二列相关
	1	0	0	0		1	0	0	0
	2	0	0	0		2	0	0	0
L1	3	0	0	0	L2	3	0	0	0
	4	0	0	0		4	0	0	0
	5	0	0	0		5	0	0	0

辅助线

二、分析数据的概况：频数分析——表（续）

总结：

1. 频数分析表是原始数据经过计数处理后得到的；
2. 频数分析表可以得出的结论是“谁多谁少”或“有多少在以下/以上”（累计频数）或涨跌幅（以时间为分组标志）；
3. 当一个表中想要反映多个变量的频数信息，需要先确定变量的层级，想重点比较的变量放在最内层；
4. 用三线表。

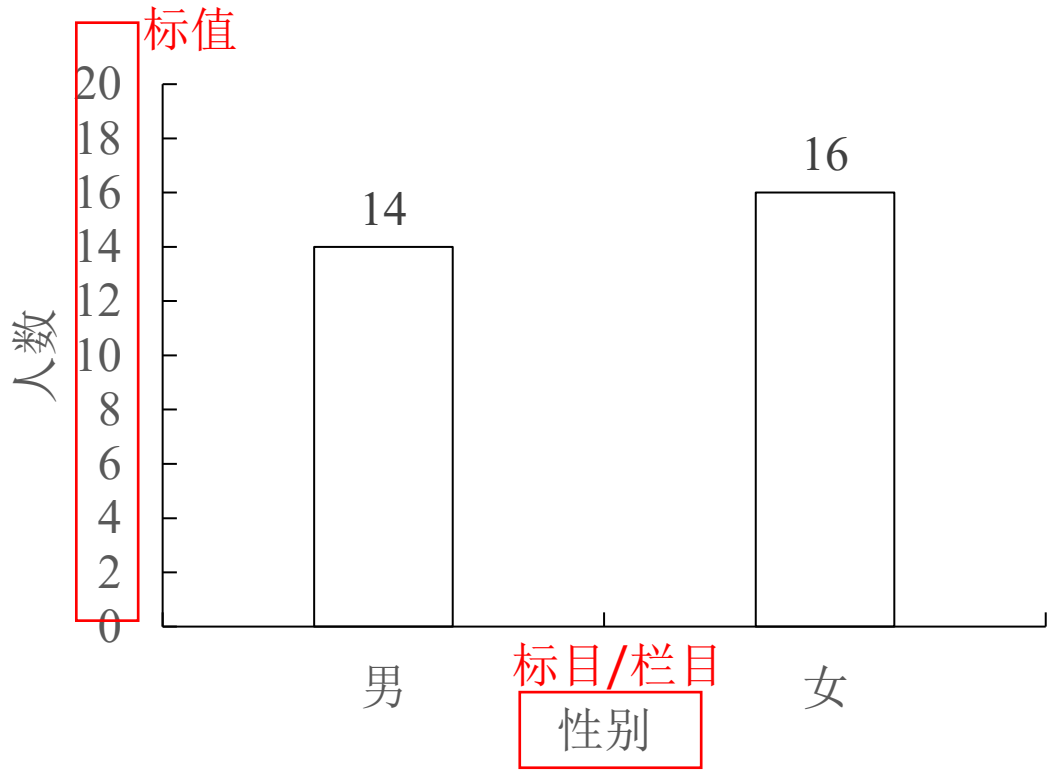
二、分析数据的概况：频数分析——图

- 常用统计图形包括：
  - 条形/柱状图（bar chart）
  - 饼图（pie chart）
  - 直方图（histogram）
- **上述三种图就是将频数表中的数据用图展示出来而已。**

二、分析数据的概况：频数分析——图（续）

单一分组标志——条形图

- 离散数据



图序 图 1 不同性别被试的人数 图题

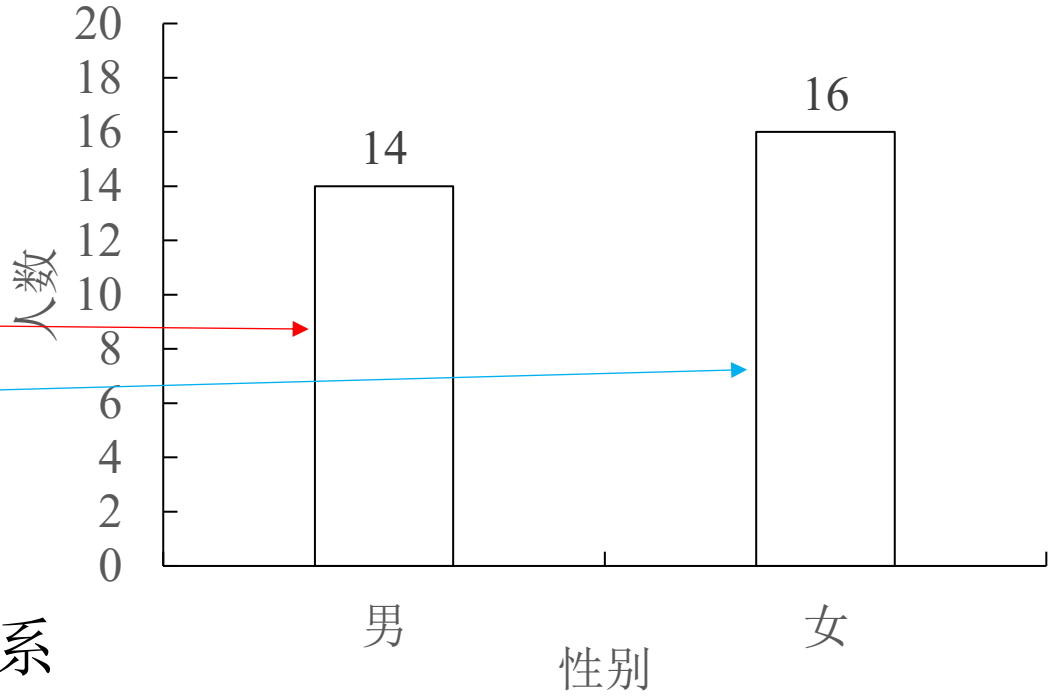
二、分析数据的概况：频数分析——图（续）

单一分组标志——条形图

- 离散数据

表1不同性别被试人数和百分比

性别	人数	百分比（%）
男	14	46.7
女	16	53.3
合计	30	100.0



表和图之间的一一对应关系

图 1 不同性别被试的人数

- 如果条形图呈现的是频数，画图时提供原始数据即可，统计软件会自动帮你统计频数。

	A	B
1	编号	性别
2	1	1
3	2	0
4	3	0
5	4	1
6	5	1
7	6	1
8	7	0
9	8	1
10	9	0
11	10	1

	A	B	C	D	E
1	考生号	姓名	身份证号	专业	年份
2	20xxxxxxx	xxx	xxxxxxxxx	心理学	2017
3	20xxxxxxx	xxx	xxxxxxxxx	教育学	2017
4	20xxxxxxx	xxx	xxxxxxxxx	法学	2017
5	20xxxxxxx	xxx	xxxxxxxxx	心理学	2018
6	20xxxxxxx	xxx	xxxxxxxxx	教育学	2018
7	20xxxxxxx	xxx	xxxxxxxxx	法学	2018
8	20xxxxxxx	xxx	xxxxxxxxx	心理学	2019
9	20xxxxxxx	xxx	xxxxxxxxx	教育学	2019
10	20xxxxxxx	xxx	xxxxxxxxx	法学	2019

二、分析数据的概况：频数分析——图（续）

- 条形图不仅限于频率分析。

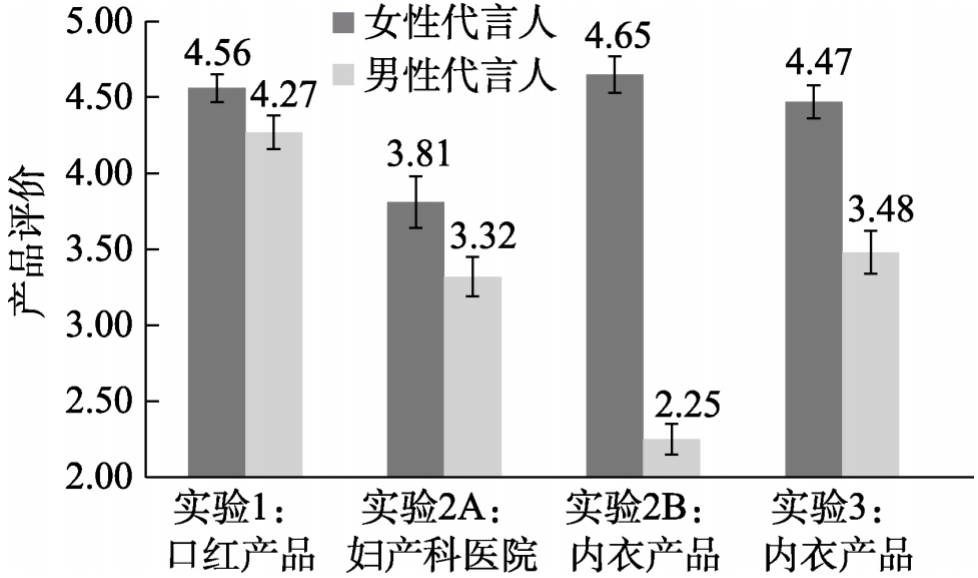


图 1　实验 1~3 的主效应检验

ID	实验	代言人性别	被试评分
1	实验1	男	4.12
…	…	…	…
1	实验2	女	3.64
…	…	…	…
1	实验3	女	4.12
…	…	…	…
1	实验4	男	3.47
…	…	…	…

实验	代言人性别	被试平均评分
实验1	男	4.27
实验1	女	4.56
实验2	男	3.81
实验2	女	3.32
实验3	男	4.65
实验3	女	2.25
实验4	男	4.47
实验4	女	3.48

二、分析数据的概况：频数分析——图（续）

单一分组标志——饼图

- 离散数据

图例，通常对应着数据中的某个分组标志/变量。

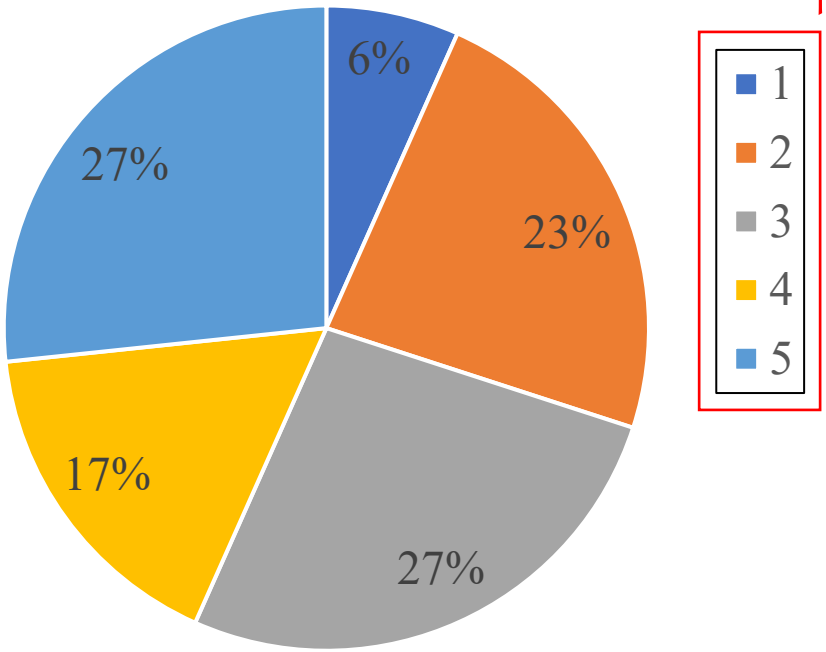


图 2 不同偏好被试的人数占比

二、分析数据的概况：频数分析——图（续）

单一分组标志——直方图

Histogram was first introduced by Karl Pearson, English mathematician and biostatistician. He founded the world's first university statistics department at University College, London in 1911. He and Weldon established Biometrika in 1902. (Wikipedia: Histogram, Karl Pearson)

二、分析数据的概况：频数分析——图（续）

单一分组标志——直方图（续）

- 连续数据：直方图就是将连续数据的频数表格图形化。

表4 心理系某年级英语考试成绩分布表

分组	人数	累计人数	百分比（%）	累计百分比（%）
65~69	2	2	2.5	2.5
70~74	6	8	7.5	10.0
75~79	18	26	22.5	32.5
80~84	26	52	32.5	65.0
85~89	16	68	20.0	85.0
90~94	8	76	10.0	95.0
95~99	4	80	5.0	100.0
合计		80		100.0

二、分析数据的概况：频数分析——图（续）

单一分组标志——直方图（续）

- 连续数据：直方图就是将连续数据的频数表格图形化。

分组	人数
65~69	2
70~74	6
75~79	18
80~84	26
85~89	16
90~94	8
95~99	4

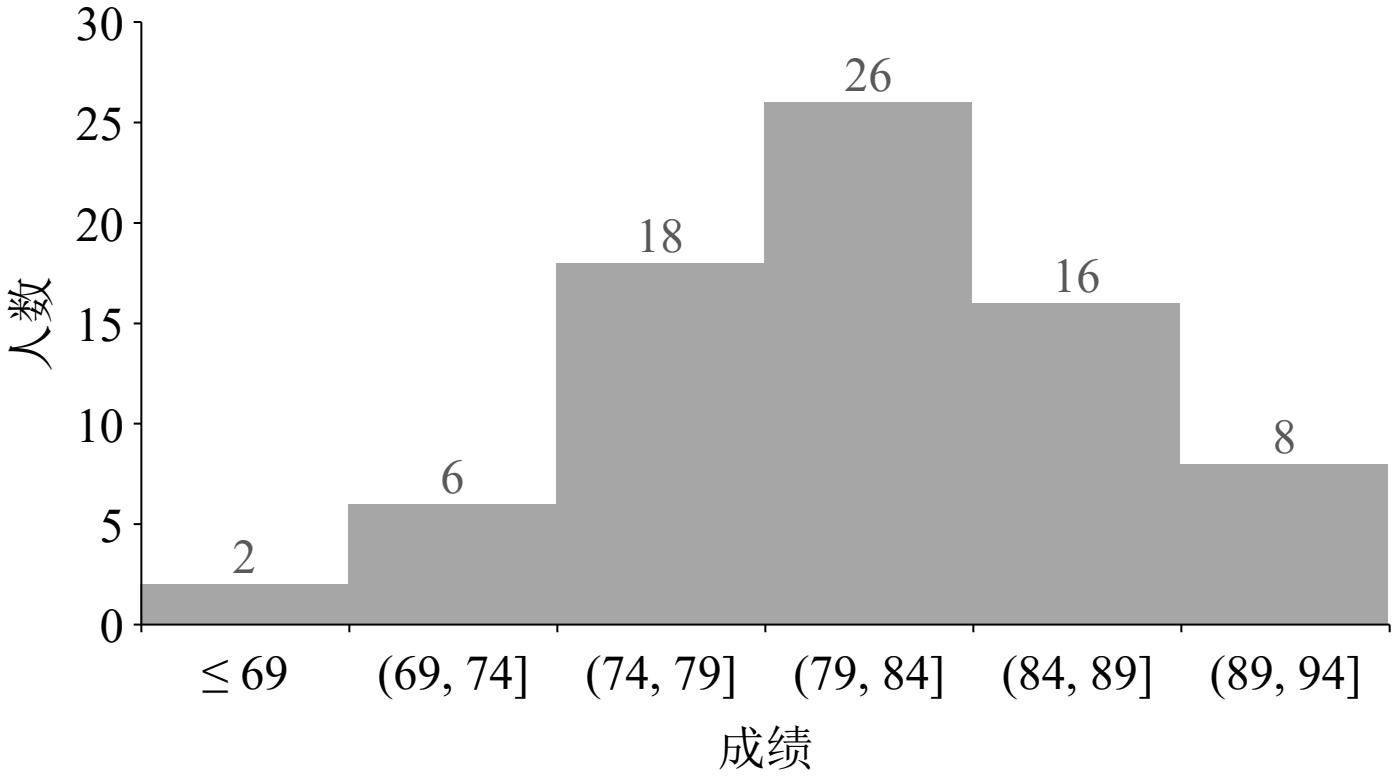


图3 心理系某年级英语考试成绩分布

二、分析数据的概况：频数分析——图（续）

多个分组标志——条形图：

- 离散数据：方式 1，强调性别的对比

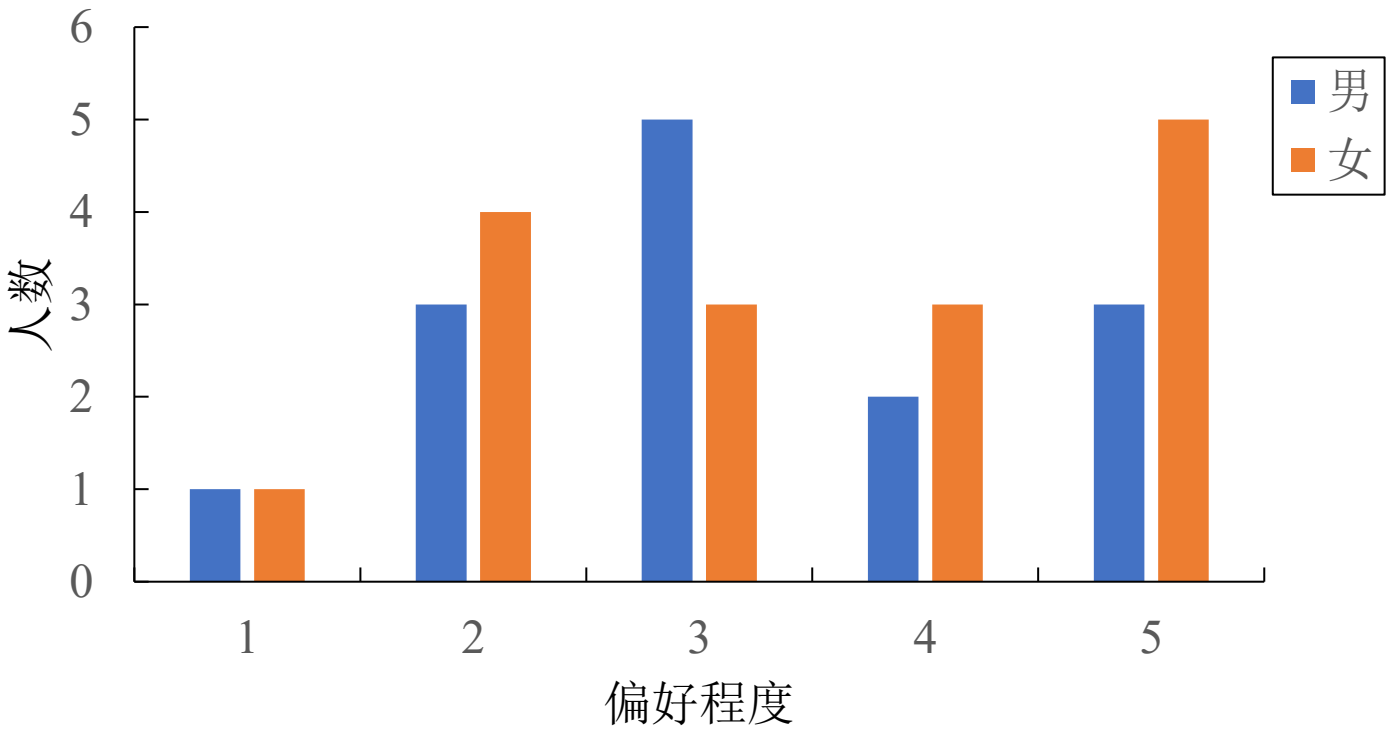


图4 不同性别选择不同偏好程度的人数

二、分析数据的概况：频数分析——图（续）

多个分组标志——条形图（续）

- 离散数据：方式 2，强调不同偏好程度的对比

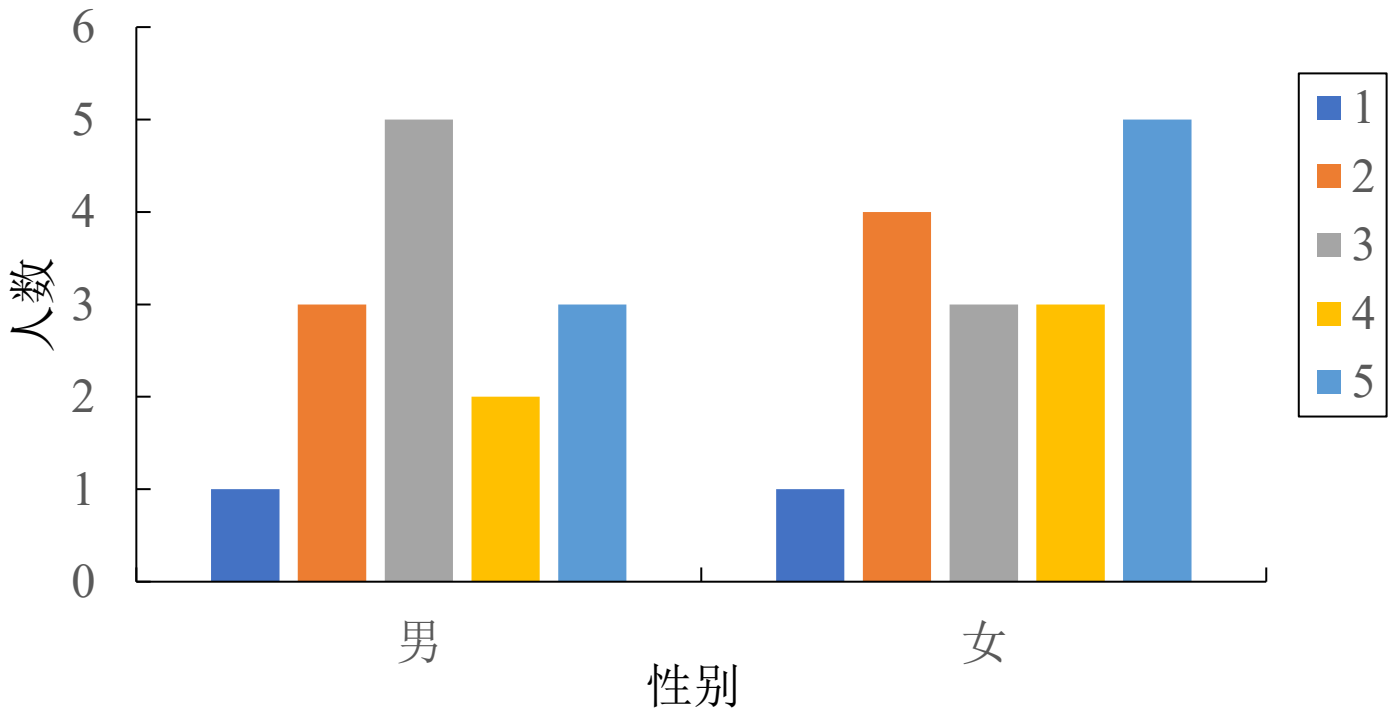
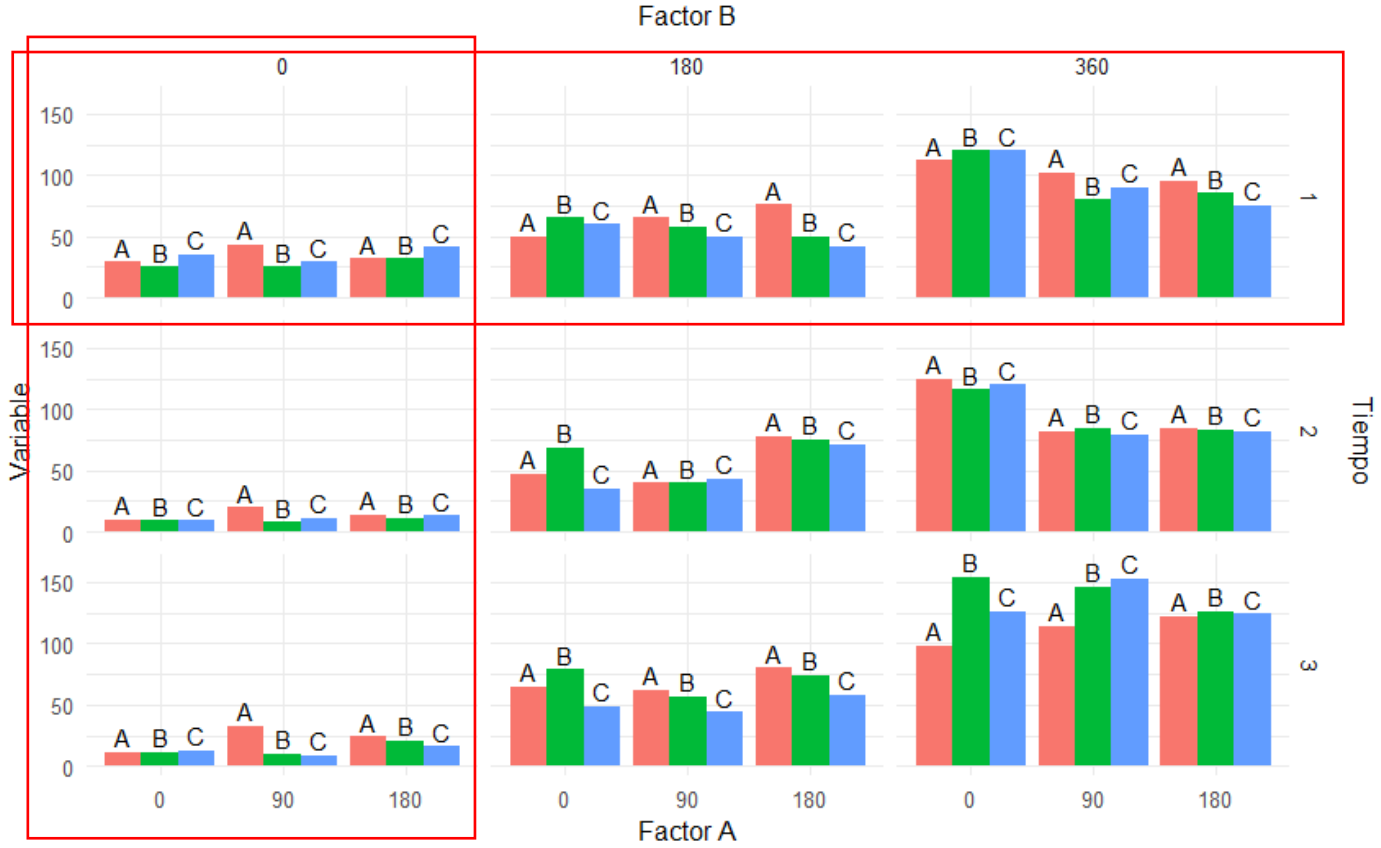


图4 不同性别选择不同偏好程度的人数

二、分析数据的概况：频数分析——图（续）

多个分组标志——条形图（续）

- 离散数据：分组标志更复杂时需分版面



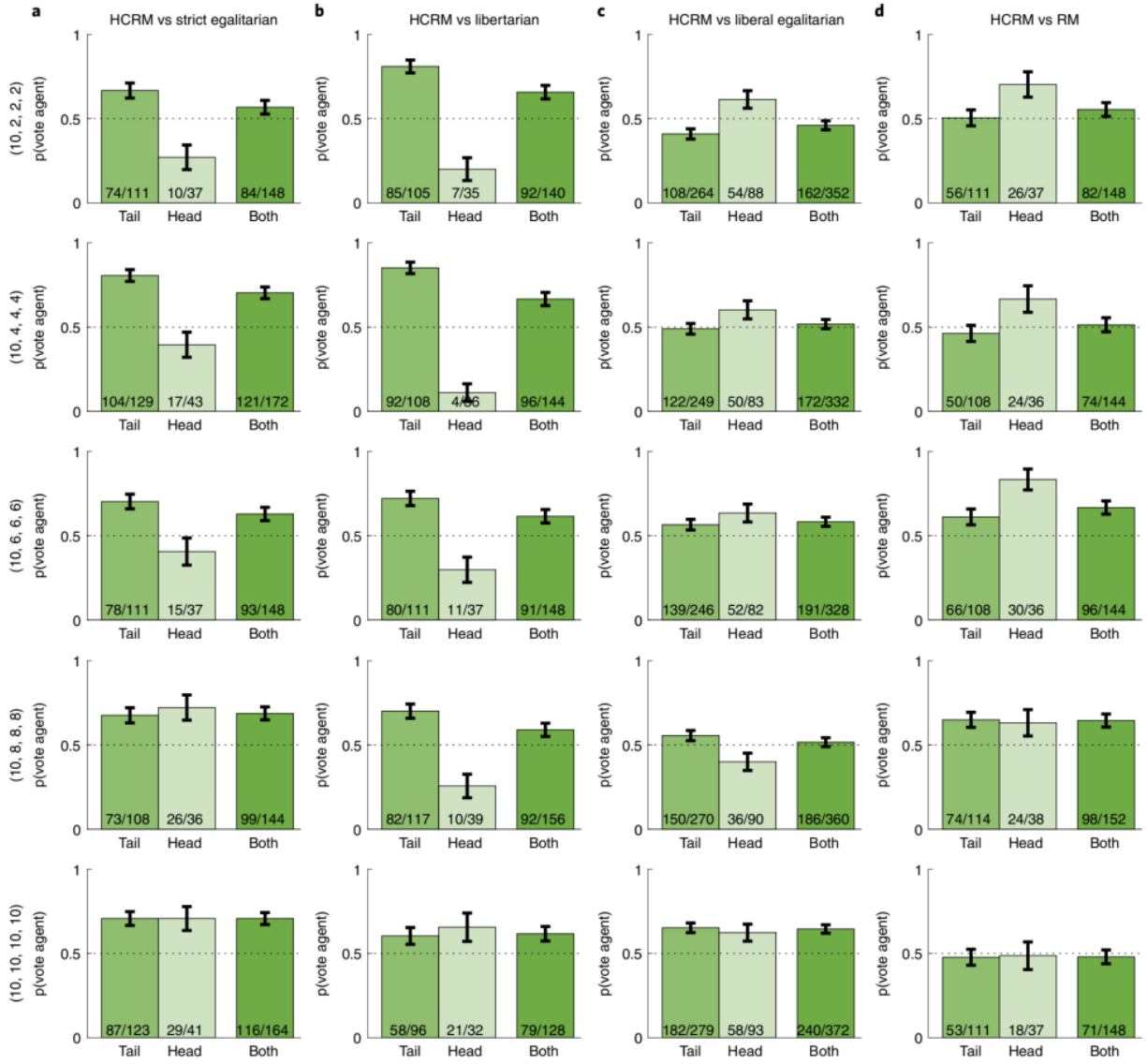
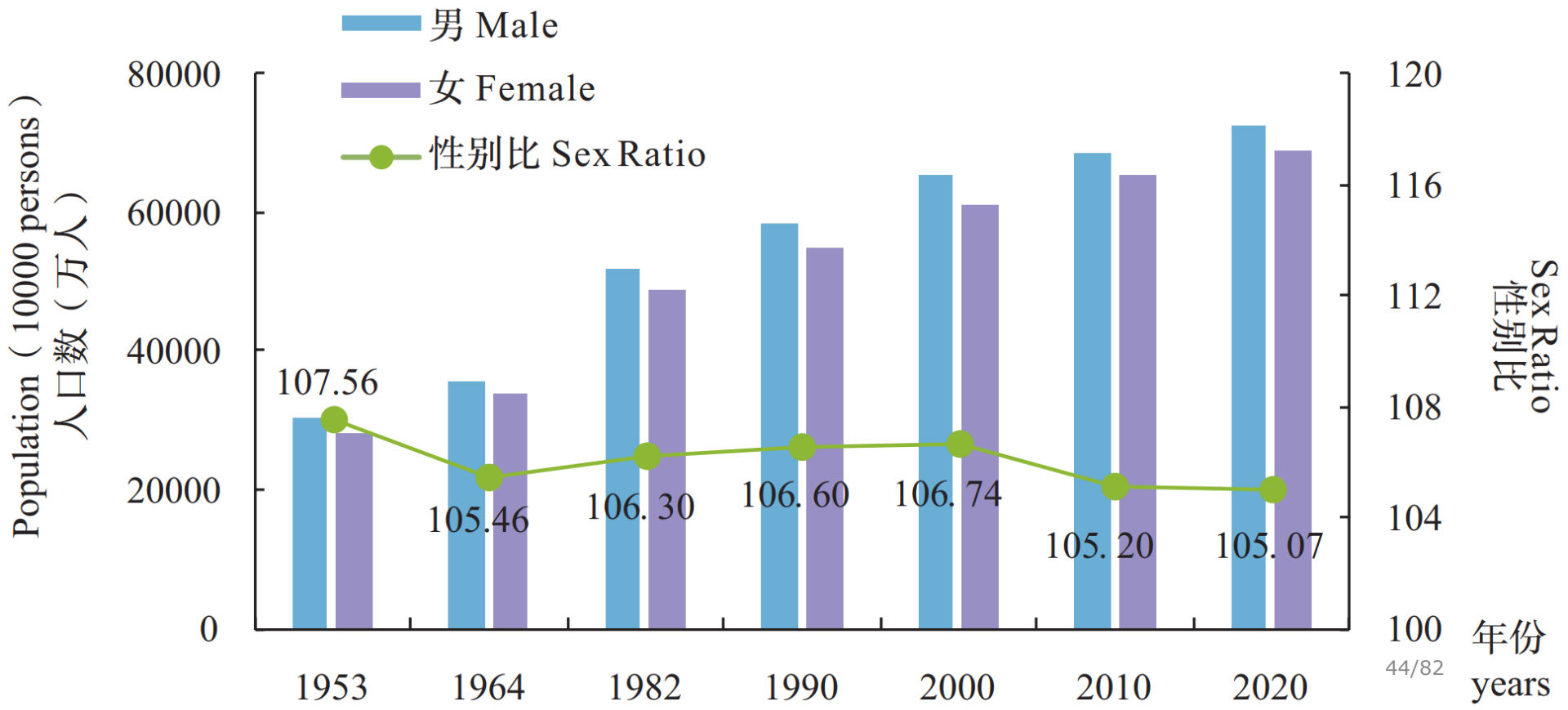


Fig. 2 | Overall vote share by endowment and rival mechanism. a–d, Vote share for the HCRM against the three canonical baselines (**a**–**c**) and the RM (**d**) for each endowment condition. The three bars show the average number of votes for the agent given by the tail players, the head player and all players. In all plots, bars show binomial standard error.

二、分析数据的概况：频数分析——图（续）

- 多分组标志，且时间为分组标志之一



二、分析数据的概况：频数分析——图（续）

- 多分组标志，且时间为分组标志之一

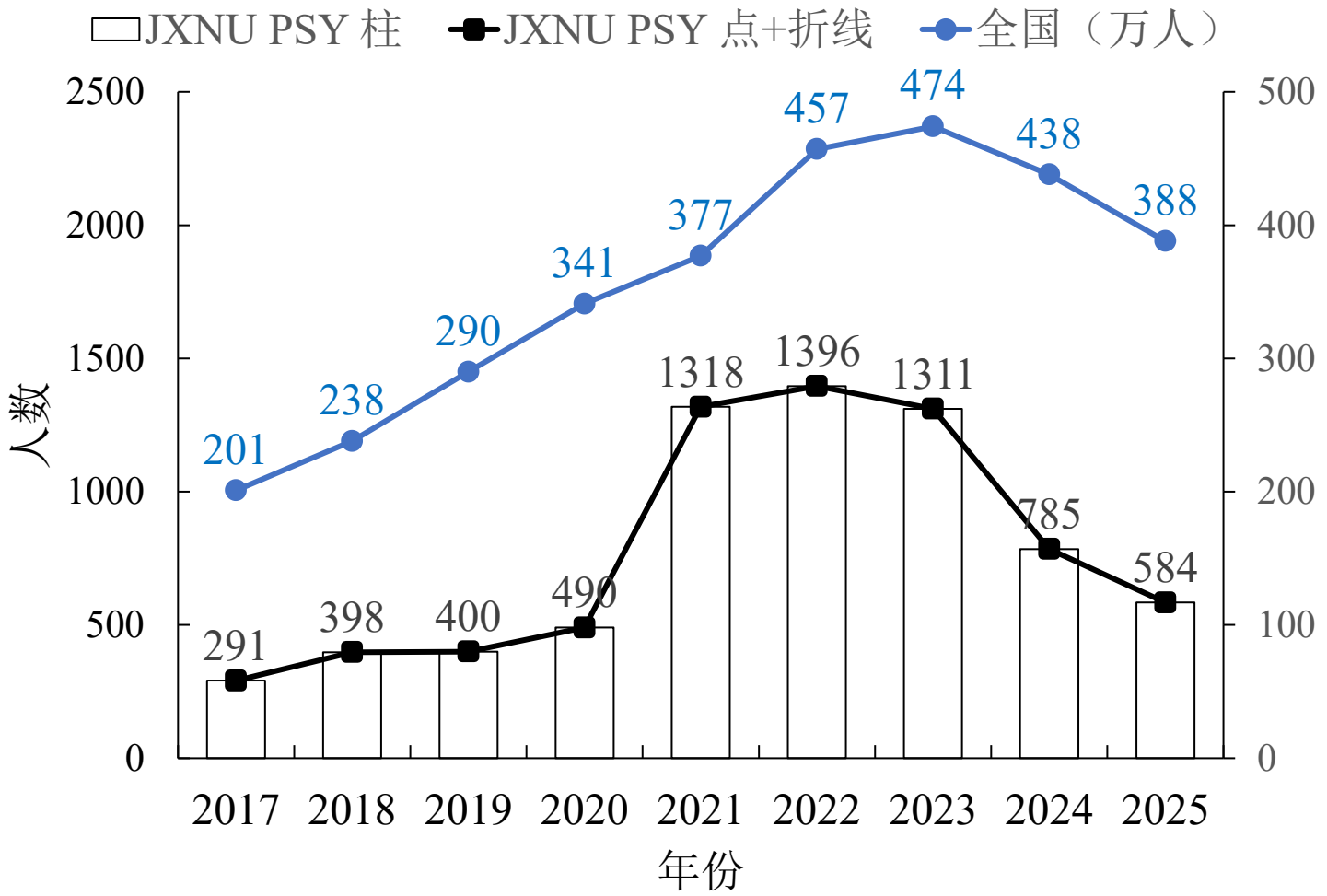
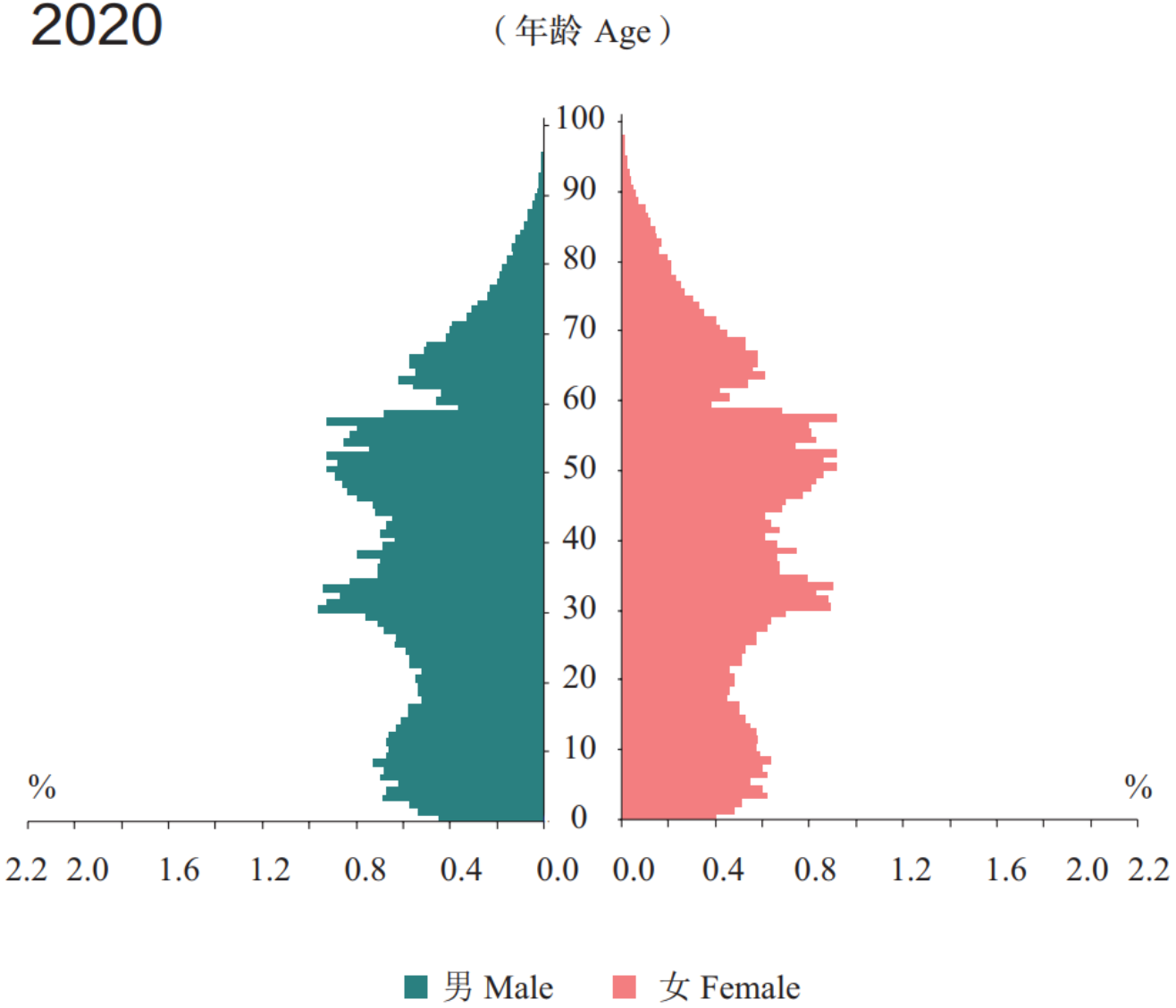
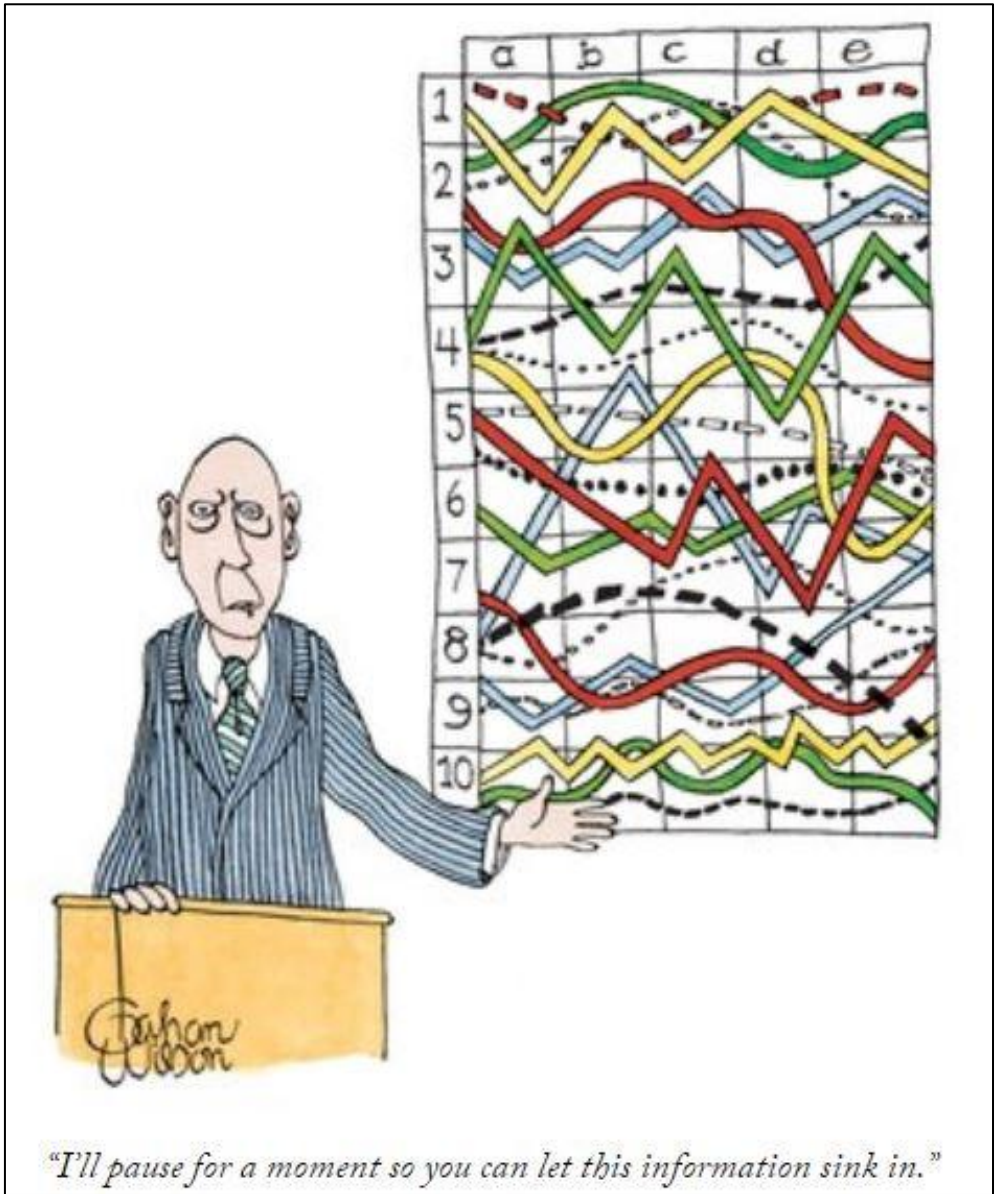


图1 2017-2025年研究生（全日制）报考人数变化图

二、分析数据的概况：频数分析——图（续）



Stats joke



"I'll pause for a moment so you can let this information sink in."

二、分析数据的概况：频数分析——图（续）

- 总结：

1. 统计图实际上就是频数表中的数据用图展示出来而已；
2. 图相对于表更直观；
3. 统计图形的选择一定要服务于信息展示的目标，能够清晰地呈现出数据中趋势、现象的图才是好图；
4. 只要看到图例就知道图中含有至少一个分组标志；
5. 当要在图形中考虑的分组变量变多时，图会不可避免地变得更加复杂，使得呈现信息越来越多的同时，趋势、现象也会越来越难总结。

条形图、饼图、折线图、直方图例子

- 爬取BAT官网及各大招聘网站后，我们找到了心理学的17135 种可能
- 以及就业指南：北京师范大学心理学部 生涯规划

散点图：无需任何分析就能画的图

- 频数分析回答的是有多少的问题，频数分析的所有表和图都需要对原始数据进行统计频数的处理，不论原始数据是离散还是连续，最终在表格和图中呈现的都是离散的形式（计数）。
- 如果，想要**了解两个变量之间的关系**，并且恰巧这两个**变量的数据都是等距或等比**时，就可以用散点图来呈现数据的概况。
- 散点图能得出的结论：
  - **A是否会随着B变化；**
  - **数据是否集中；**
  - **以及其他和数据内容高度相关的结论。**

散点图：无需任何分析就能画的图（续）

原始数据

表6 30名学生的语、数、英成绩

性别	年龄	语文	数学	英语
1	16	87	82	80
0	16	78	90	82
1	17	76	95	66
1	15	83	77	88
1	15	75	53	68
0	15	78	74	84
1	16	81	78	77
0	17	81	75	64
…	…	…	…	…

散点图：无需任何分析就能画的图（续）

散点图可以大致呈现出两个变量之间的关系，包括集中或变化趋势。

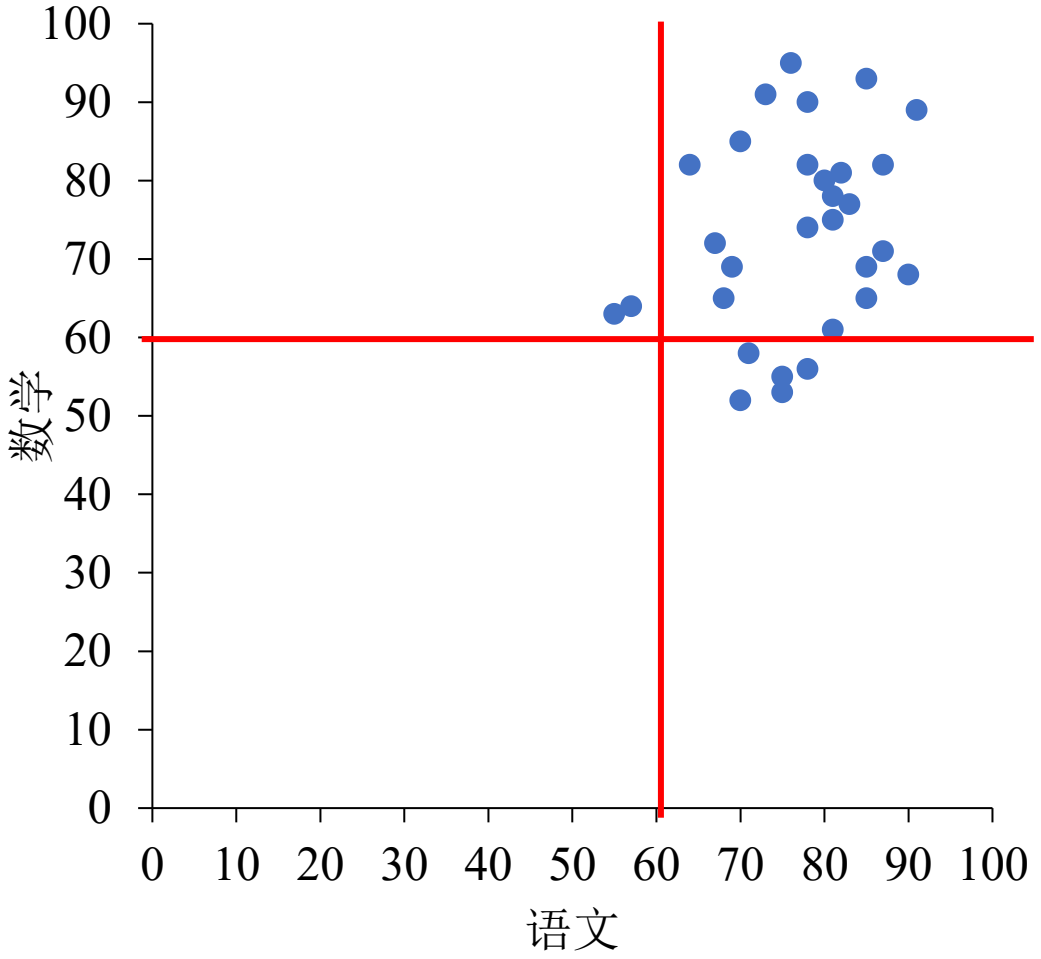


图 5 30名学生的语文与数学

散点图：无需任何分析就能画的图（续）

不同的变量间关系会有所不同。

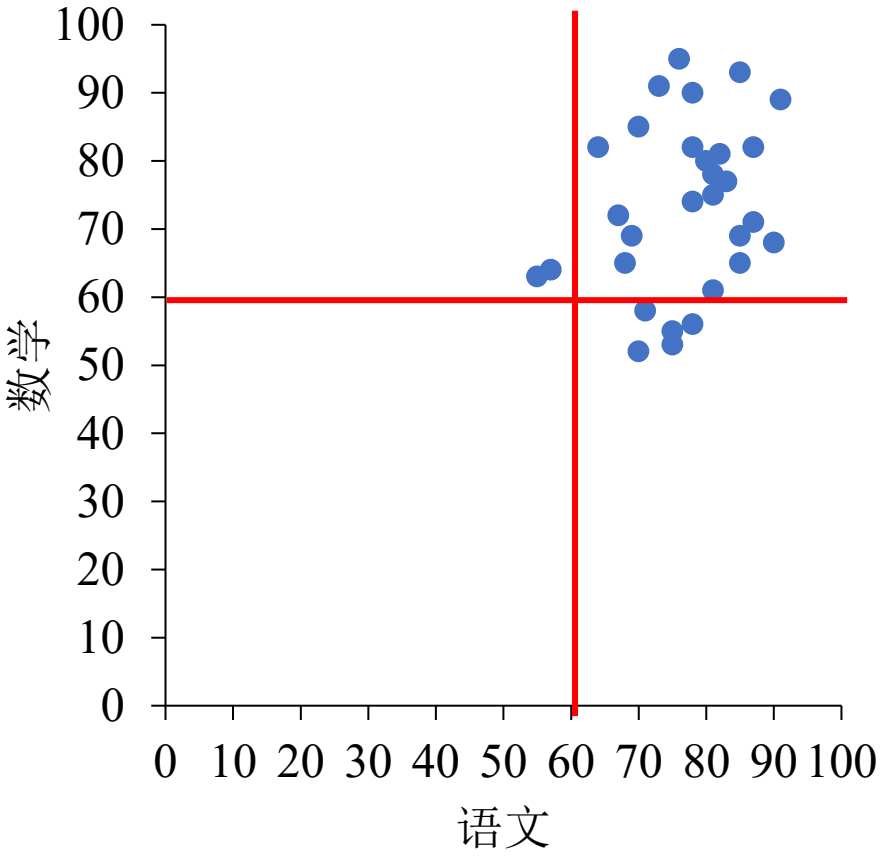


图 5 30名学生的语文与数学

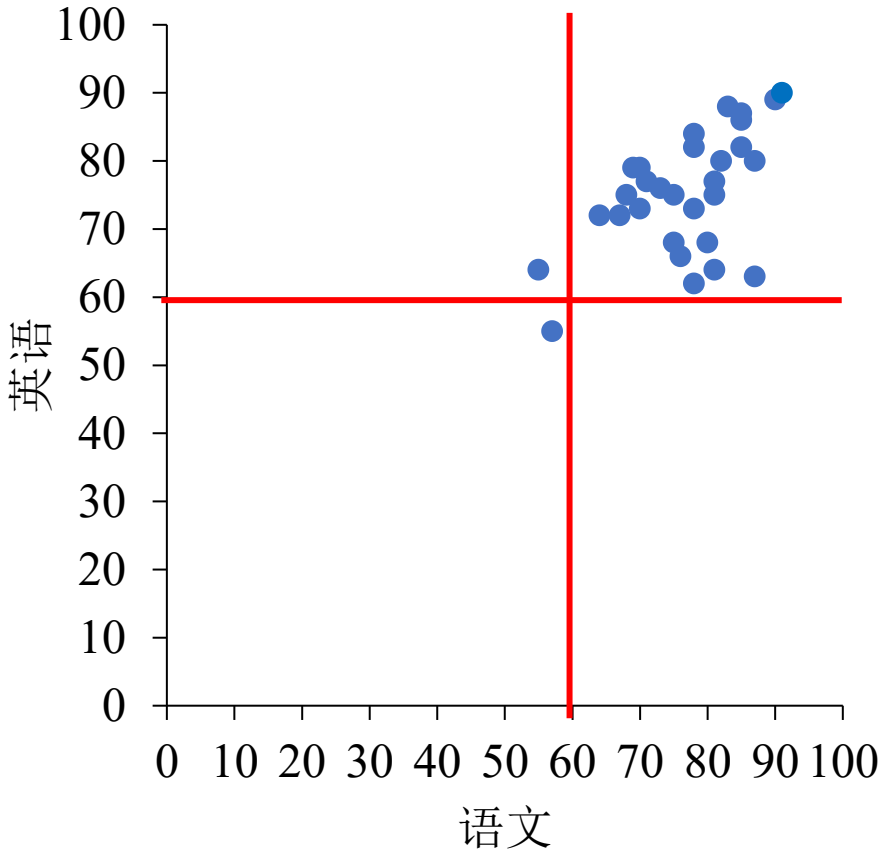


图 6 30名学生的语文与英语

散点图：无需任何分析就能画的图（续）

- 散点图同样也可以通过设置颜色的方式添加分组标签（分组变量）

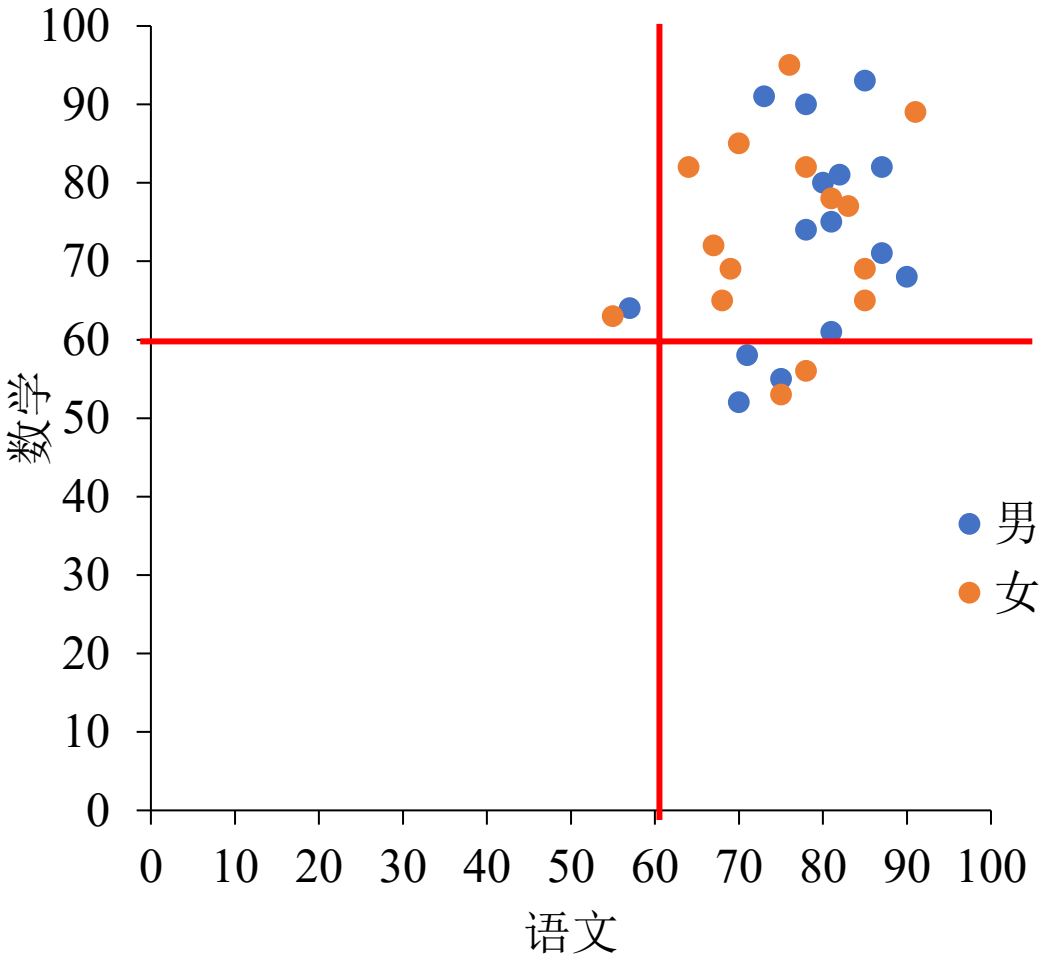


图 5 30名学生的语文与数学

散点图：无需任何分析就能画的图（续）

- 散点图同样也可以通过设置颜色的方式添加分组标签（分组变量）

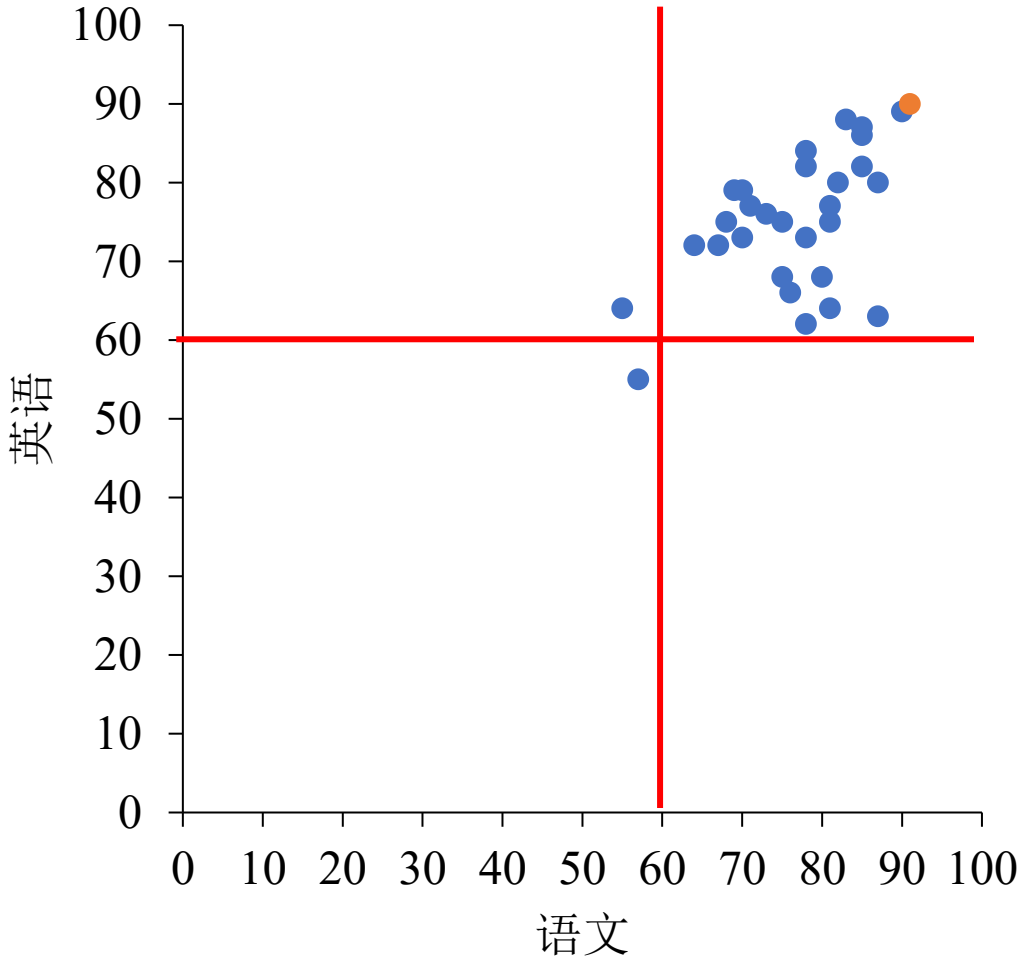
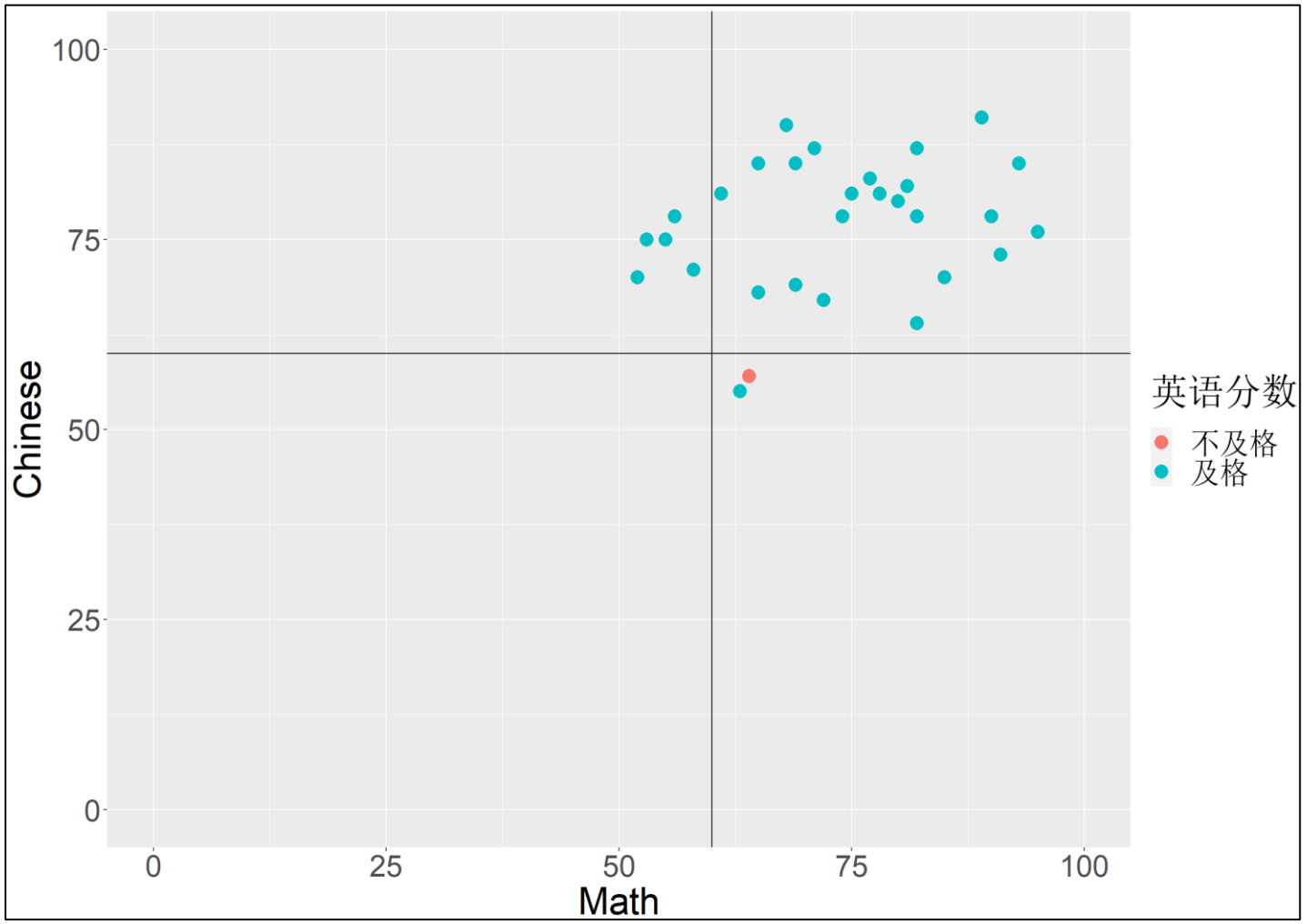
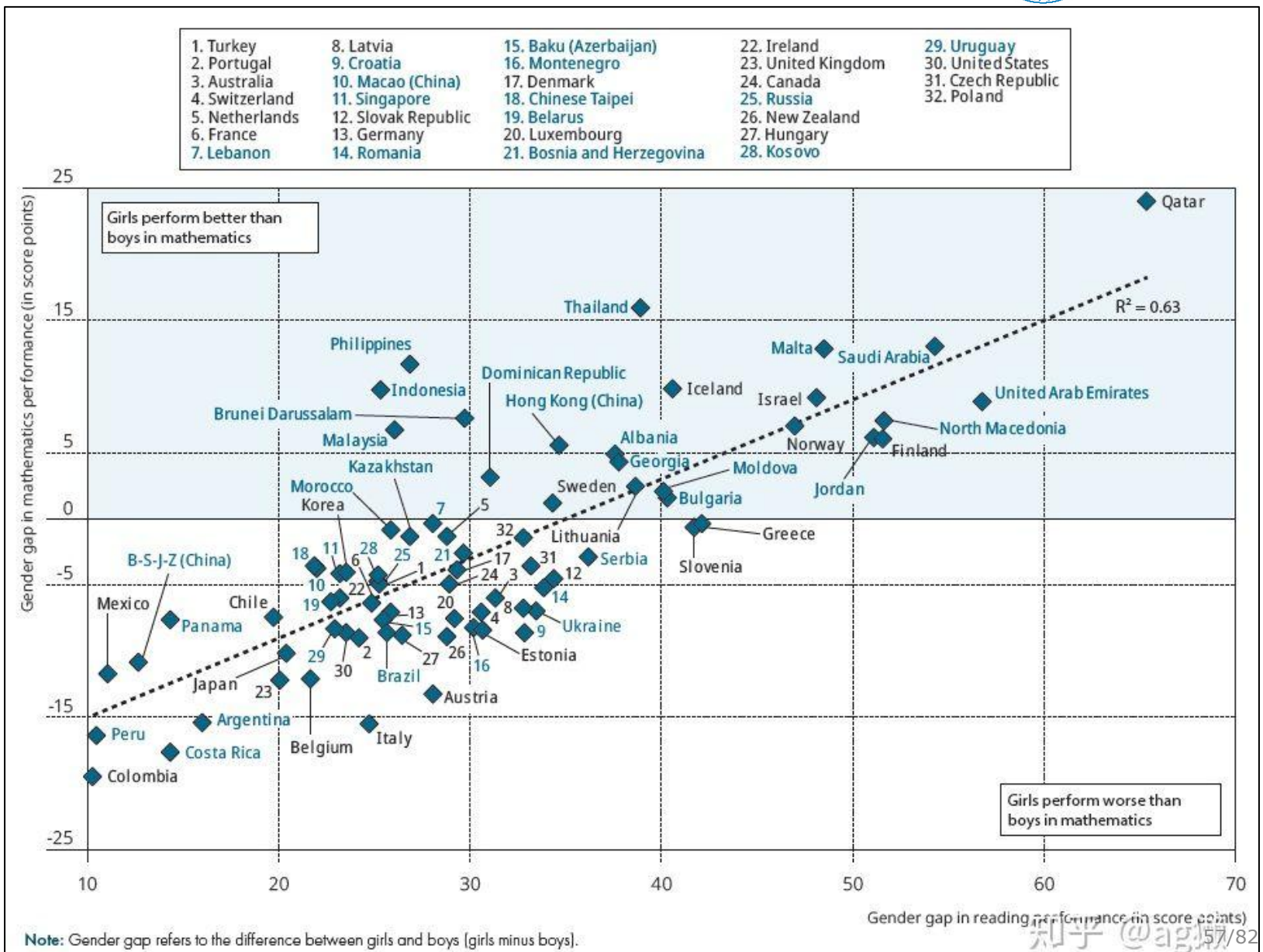


图 6 30名学生的语文与英语

散点图：无需任何分析就能画的图（续）

想要在一张散点图里描述三个变量之间的关系怎么办？



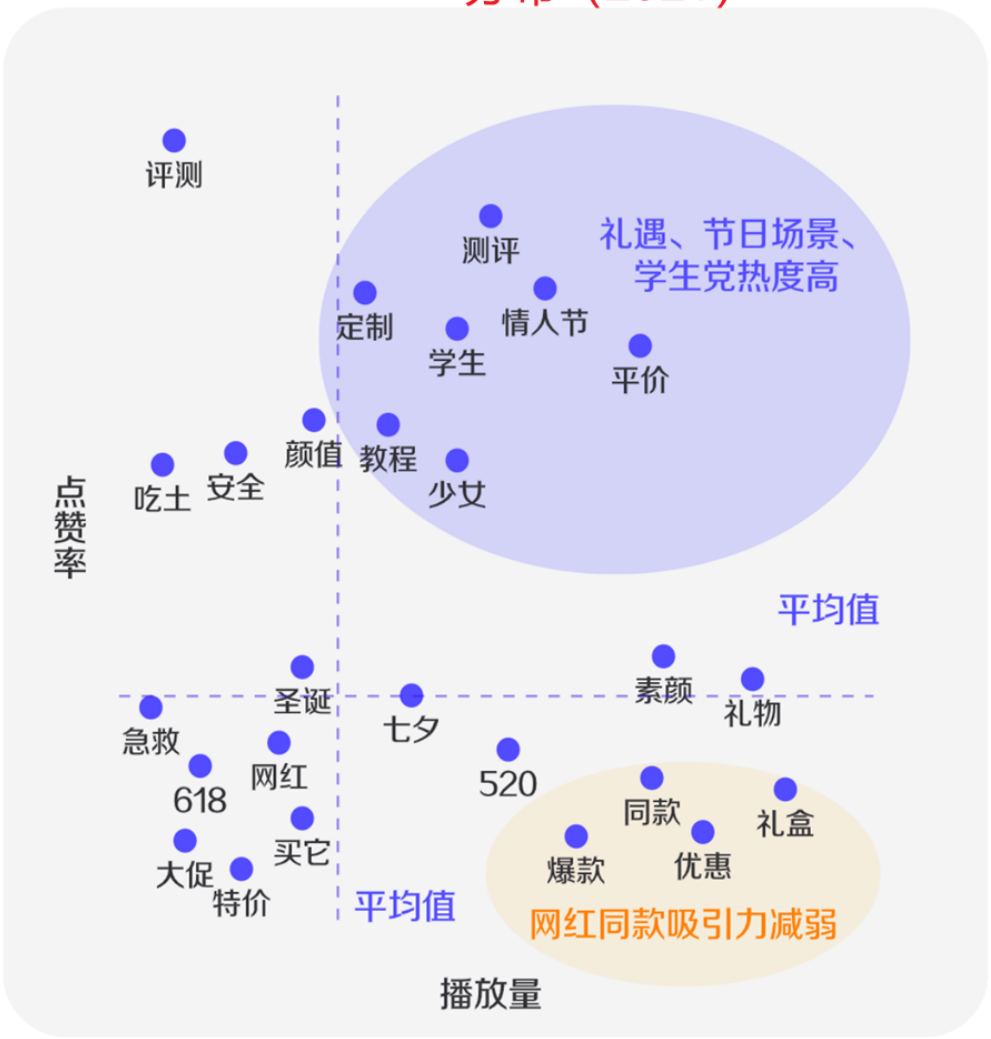


Note: Gender gap refers to the difference between girls and boys (girls minus boys).

Source: OECD, PISA 2018 Database, Tables II.B1.7.1 and II.B1.7.3; Figure II.7.3.

口红

不同美妆品类的
热点词播放量和点赞率
分布（2021）



第二章总结

1. 数据的初步整理

1. 录入数据
2. 数据清洗（剔除无效数据、处理缺失值、处理极端值等）

2. 分析数据的概况——频数分析

- 通过频数分析，可以得到**谁多谁少 、有多少在以上/以下**（累计）、**涨跌幅**（分组标志为时间）的结论。
- 表格要使用三线表
- 表题在上，图题在下
- 有分组标志、图中有颜色区别时，一般会有图例
- **精确选表，直观选图**
- 为更好地展示数据和描述趋势，可以在条形图、饼图、折线图、直方图、散点图的基础上对图形进行巧妙地变换。