统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2021-05-08

第三章 数据的图表展示

3.1 数据的预处理

3.2 品质数据的整理与显示

3.3 数值型数据的整理与显示

3.4 合理使用图表

3.1 数据的预处理

数据清洗（data cleaning）

数据变换（data transformation）

数据子集(data subset)

数据预处理的主要内容

- 数据审核：检查数据中的错误
- 数据筛选：找出符合条件的数据
- 数据排序：升序和降序；寻找数据的基本特征
- 数据透视：按需要汇总

数据预处理：数据审核

对于**原始数据(raw data)**需要进行：

完整性审核：

- 应调查的单位或个体是否有遗漏
- 所有的调查项目或变量是否填写齐全

准确性审核：

- 数据是否真实反映实际情况，内容是否符合实际
- 数据是否有错误，计算是否正确等

数据预处理：数据审核

对于**二手数据(second hand data)**需要进行：

适用性审核：

- 弄清楚数据的来源、数据的口径以及有关的背景材料
- 确定数据是否符合自己分析研究的需要

时效性审核：

- 尽可能使用最新的数据
- 确认是否有必要做进一步的加工整理

数据预处理：数据筛选

数据预处理：当数据中的错误不能予以纠正，或者有些数据不符合调查的要求而又无法弥补时，需要对数据进行筛选。

数据筛选的内容：

- 将某些不符合要求的数据或有明显错误的数据予以剔除
- 将符合某种特定条件的数据筛选出来，而不符合特定条件的数据予以剔除

数据预处理：数据排序

数据排序的作用：

- 按一定顺序将数据排列，以发现一些明显的特征或趋势，找到解决问题的线索
- 排序有助于对数据检查纠错，以及为重新归类或分组等提供依据
- 在某些场合，排序本身就是分析的目的之一
- 排序可借助于计算机完成

数据预处理：数据透视表

数据透视表（pivot table）的作用：

- 可以从复杂的数据中提取有用的信息
- 可以对数据表的重要信息按使用者的习惯或分析要求进行汇总和作图
- 形成一个符合需要的交叉表(列联表)
- 在利用数据透视表时，数据源表中的首行必须有列标题

数据预处理：Excel中的数据透视表

利用Excel软件创建数据透视表的主要步骤：

- 第1步：在Excel工作表中建立数据清单
- 第2步：选中数据清单中的任意单元格，并选择【数据】菜单中的【数据透视表和数据透视图】
- 第3步：确定数据源区域
- 第4步：在【向导—3步骤之3】中选择数据透视表的输出位置。然后选择【布局】
- 第5步：在【向导—布局】对话框中，依次将”分类变量“拖至左边的“行”区域，上边的“列”区域，将需要汇总的“变量”拖至“数据区域”
- 第6步：然后单击【确定】，自动返回【向导—3步骤之3】对话框。然后单击【完成】，即可输出数据透视表

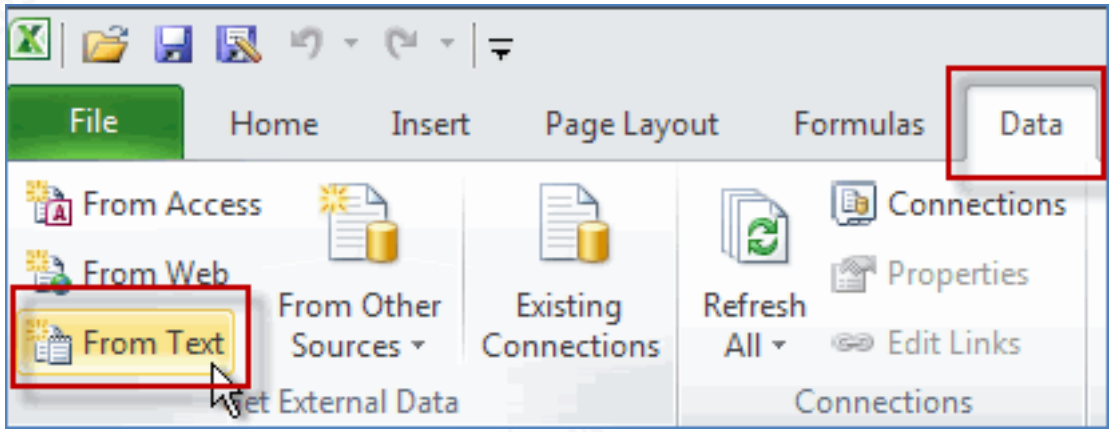
（演示）Excel数据预处理：数据准备

利用Excel进行数据准备，主要工作包括：

- 读取或导入其他数据格式（例如`.txt`或`.csv`格式）
- 删除说明性内容
- 数据表形式（long data VS wide data）
- 数值表达方式（labels VS values）
- 备注重要信息

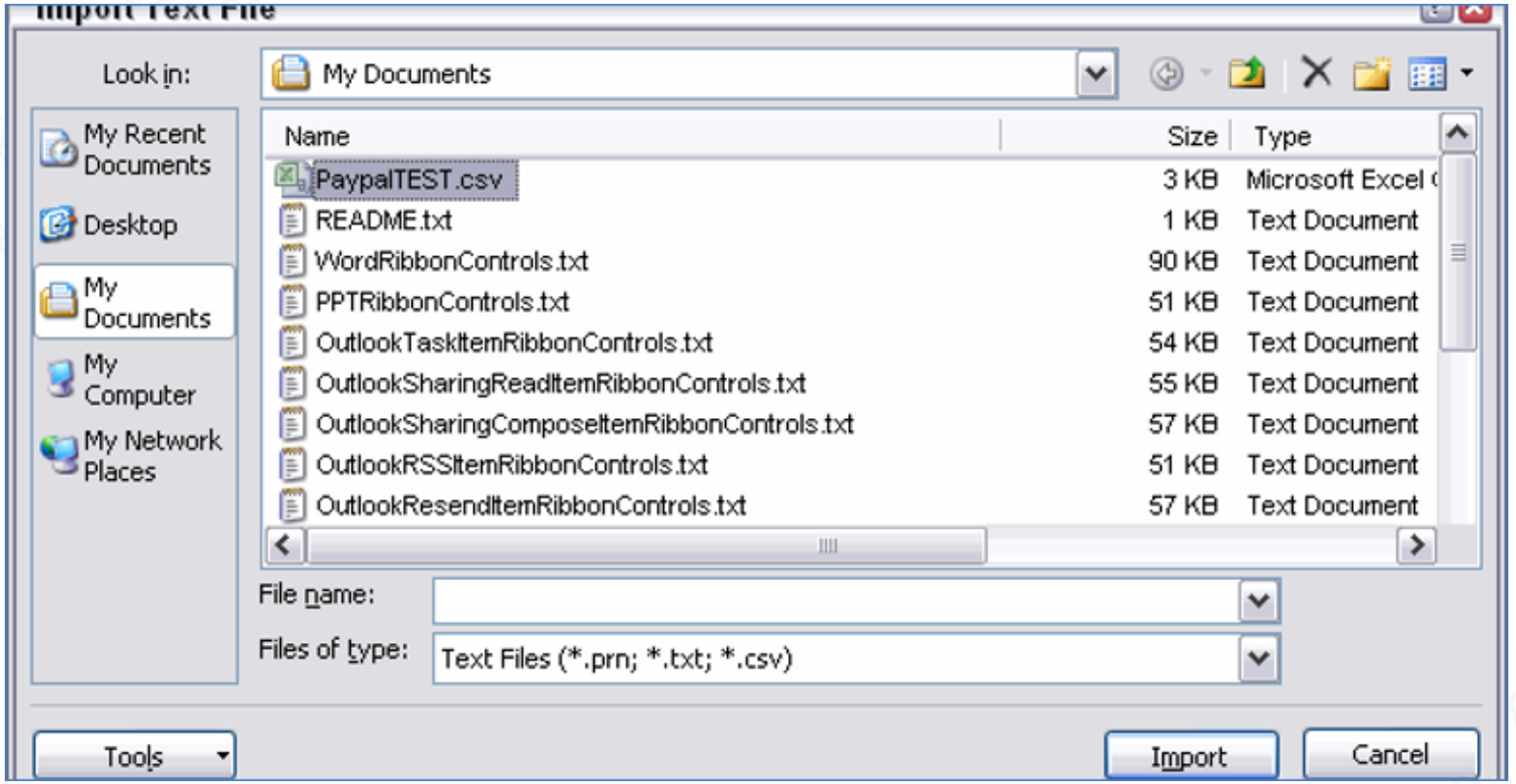
（演示）导入其他数据格式：Excel操作

a.数据菜单　b.选择文件　c.数据引导



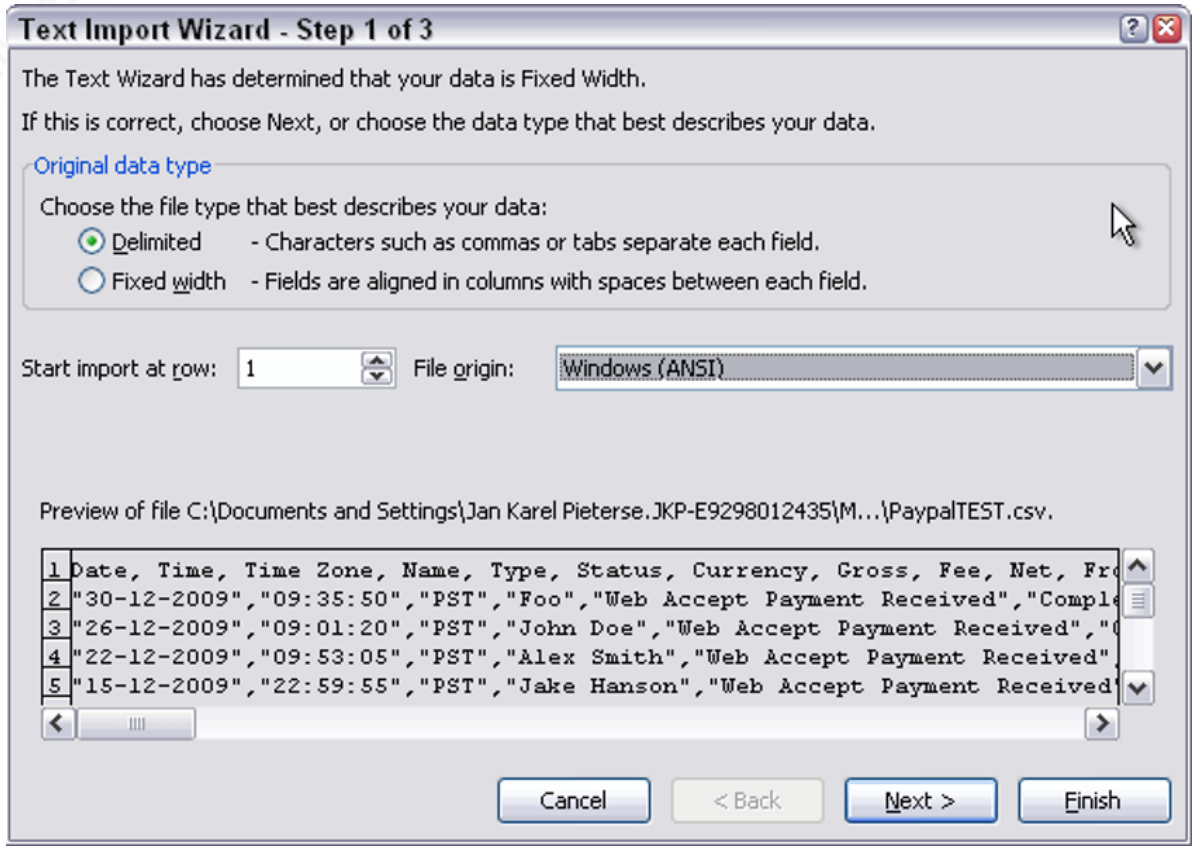
（演示）导入其他数据格式：Excel操作

a.数据菜单　b.选择文件　c.数据引导



（演示）导入其他数据格式：Excel操作

a.数据菜单　　b.选择文件　　c.数据引导



（演示）删除说明性内容：Excel操作

a.删除前

F12

	A	B	C	D	E
1	Report date:	04/05/2007		Manager name:	john Smith
2				Division:	Fruits
3					
4			Category	Date	Sales
5			Apples	01/05/2007	62
6			Pears	01/05/2007	42
7			Melons	01/05/2007	32
8			Apples	02/05/2007	45
9			Pears	02/05/2007	34
10			Melons	02/05/2007	20
11					

（演示）删除说明性内容：Excel操作

a.删除前

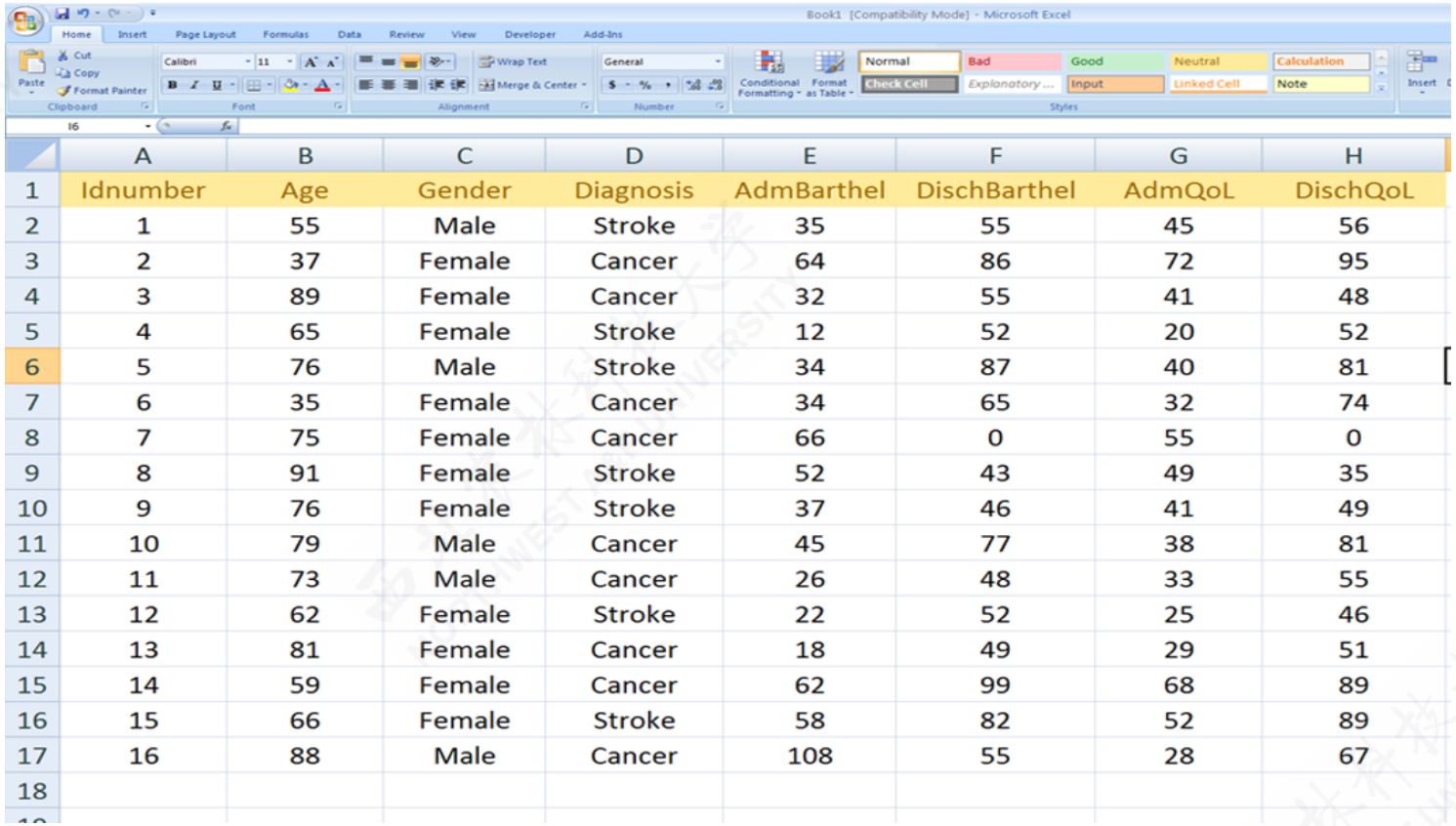
b.删除后

H7

	A	B	C	D	E	F
1						
2						
3						
4			Category	Date	Sales	
5			Apples	01/05/2007	62	
6			Pears	01/05/2007	42	
7			Melons	01/05/2007	32	
8			Apples	02/05/2007	45	
9			Pears	02/05/2007	34	
10			Melons	02/05/2007	20	
11						

（演示）数据表现形式：Excel操作

1）扁平形式



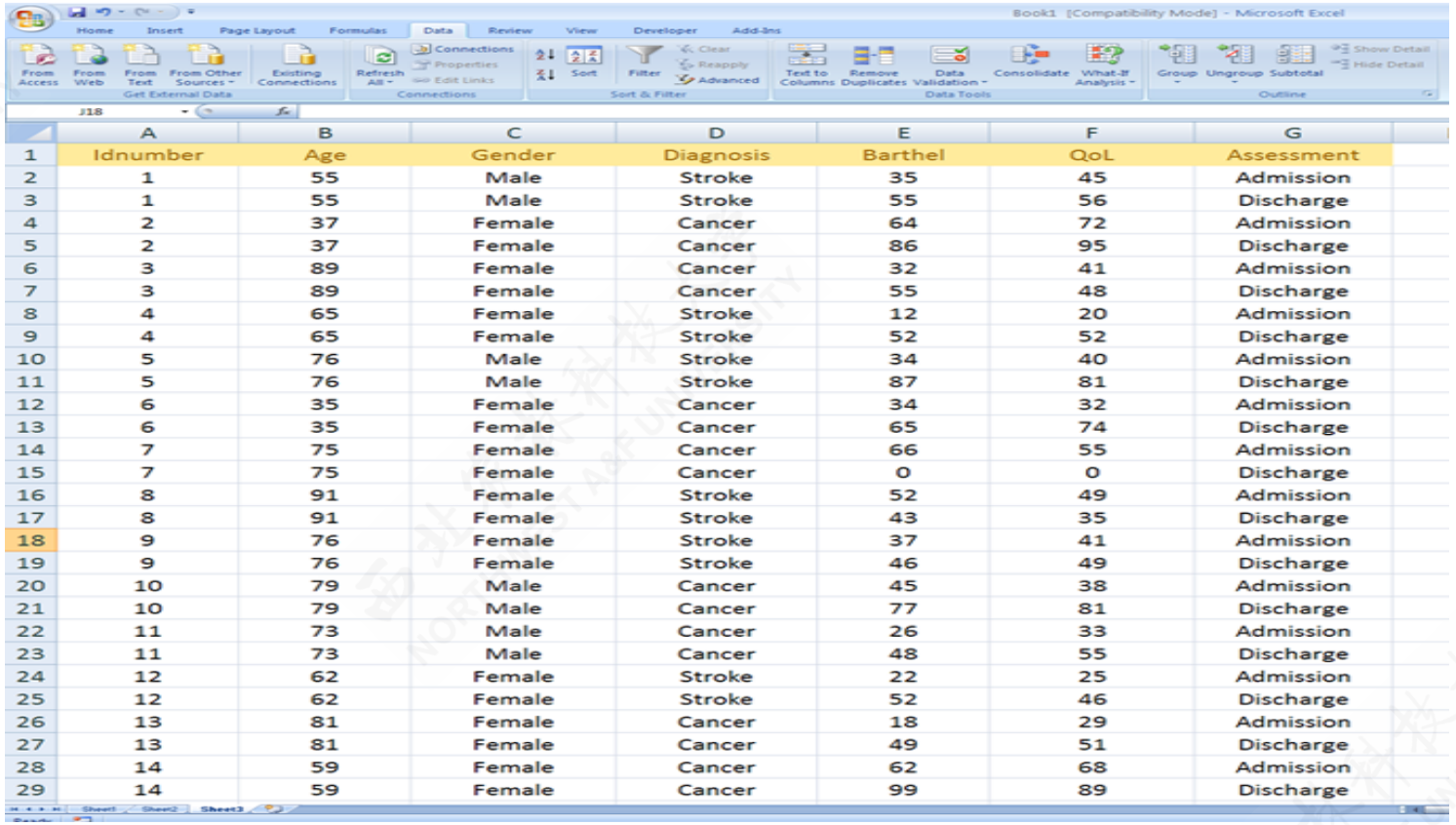
	A	B	C	D	E	F	G	H
1	Idnumber	Age	Gender	Diagnosis	AdmBarthel	DischBarthel	AdmQoL	DischQoL
2	1	55	Male	Stroke	35	55	45	56
3	2	37	Female	Cancer	64	86	72	95
4	3	89	Female	Cancer	32	55	41	48
5	4	65	Female	Stroke	12	52	20	52
6	5	76	Male	Stroke	34	87	40	81
7	6	35	Female	Cancer	34	65	32	74
8	7	75	Female	Cancer	66	0	55	0
9	8	91	Female	Stroke	52	43	49	35
10	9	76	Female	Stroke	37	46	41	49
11	10	79	Male	Cancer	45	77	38	81
12	11	73	Male	Cancer	26	48	33	55
13	12	62	Female	Stroke	22	52	25	46
14	13	81	Female	Cancer	18	49	29	51
15	14	59	Female	Cancer	62	99	68	89
16	15	66	Female	Stroke	58	82	52	89
17	16	88	Male	Cancer	108	55	28	67
18								

- 一个病人的数据，由一行就能够完全进行表达。

（演示）数据表现形式：Excel操作

1）扁平形式

2）窄长形式



	A	B	C	D	E	F	G
1	Idnumber	Age	Gender	Diagnosis	Barthel	QoL	Assessment
2	1	55	Male	Stroke	35	45	Admission
3	1	55	Male	Stroke	55	56	Discharge
4	2	37	Female	Cancer	64	72	Admission
5	2	37	Female	Cancer	86	95	Discharge
6	3	89	Female	Cancer	32	41	Admission
7	3	89	Female	Cancer	55	48	Discharge
8	4	65	Female	Stroke	12	20	Admission
9	4	65	Female	Stroke	52	52	Discharge
10	5	76	Male	Stroke	34	40	Admission
11	5	76	Male	Stroke	87	81	Discharge
12	6	35	Female	Cancer	34	32	Admission
13	6	35	Female	Cancer	65	74	Discharge
14	7	75	Female	Cancer	66	55	Admission
15	7	75	Female	Cancer	0	0	Discharge
16	8	91	Female	Stroke	52	49	Admission
17	8	91	Female	Stroke	43	35	Discharge
18	9	76	Female	Stroke	37	41	Admission
19	9	76	Female	Stroke	46	49	Discharge
20	10	79	Male	Cancer	45	38	Admission
21	10	79	Male	Cancer	77	81	Discharge
22	11	73	Male	Cancer	26	33	Admission
23	11	73	Male	Cancer	48	55	Discharge
24	12	62	Female	Stroke	22	25	Admission
25	12	62	Female	Stroke	52	46	Discharge
26	13	81	Female	Cancer	18	29	Admission
27	13	81	Female	Cancer	49	51	Discharge
28	14	59	Female	Cancer	62	68	Admission
29	14	59	Female	Cancer	99	89	Discharge

- 一个病人的数据，需要多行才能进行完整表达。

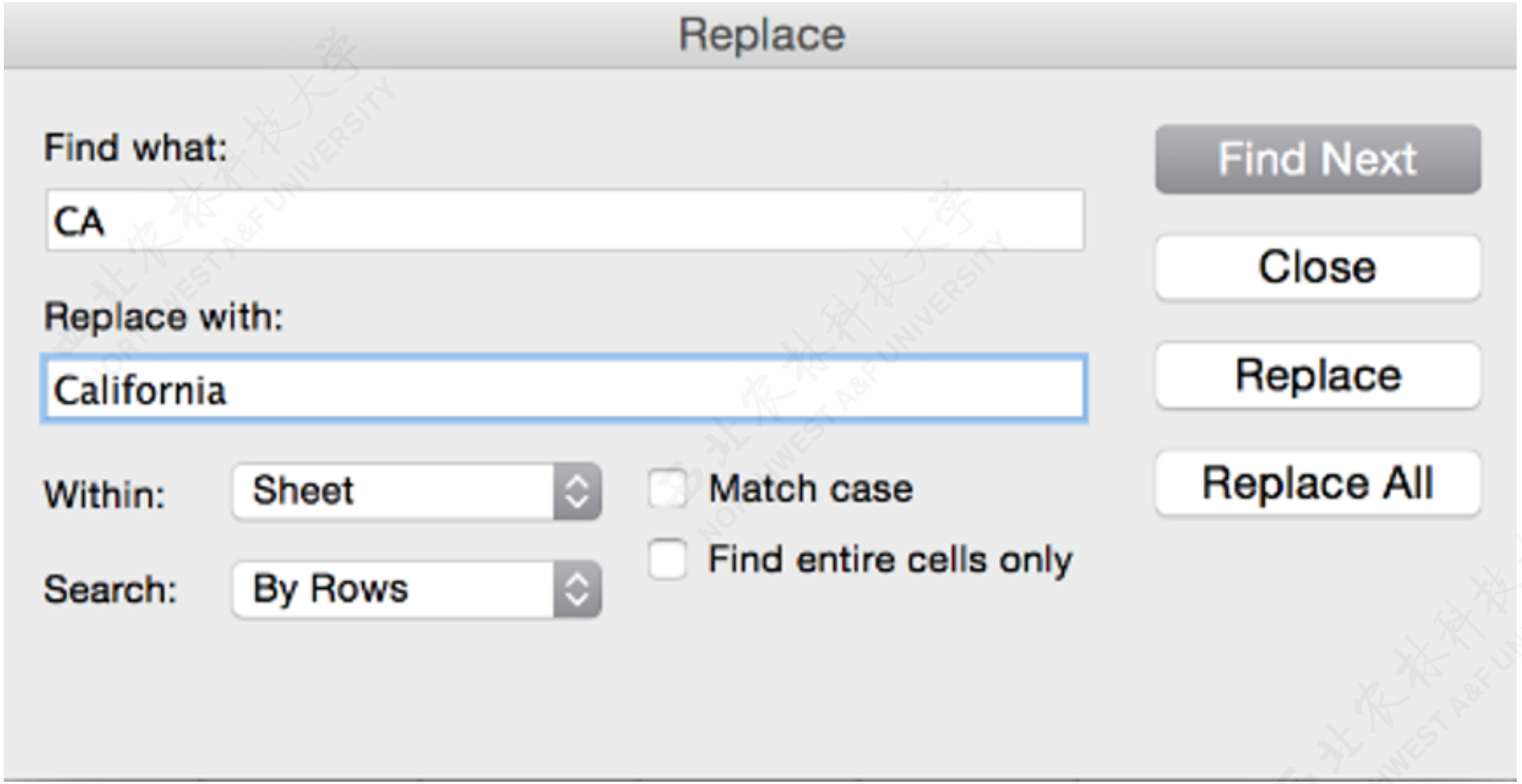
数据清洗：Excel常用操作1

- 查找/替换：
- 提取文本字符：
  - 从左侧：`=LEFT(text, [num_chars])`
  - 从右侧：`=RIGHT(text, [num_chars])`
  - 从指定位置：`=MID(text,start_num,num_chars)`
- 正确大小写：
  - 全部小写：`=LOWER(text)`
  - 全部大写：`=UPPER(text)`
  - 首字母大写：`=PROPER(text)`
  - 自定义大小写：`=UPPER(LEFT(A2,1)&LOWER(MID(A2,2,60)`

数据清洗：Excel常用操作2

- 删除重复值：
- 合并文本内容：`=[Cell 1]&[Cell 2]`
- 清除空格：`=TRIM( text )`
- 清除非打印字符：`=CLEAN( text )`

（演示）Excel数据清洗：查找/替换



（演示）Excel数据清洗：提取文本字符

	A	B
1	Zip code (9 digits)	Zip code (5 digit)
2	98765-1234	98765
3	98765-1234	=LEFT(A3,5)

MID =MID(A3,6,3)

	A	B	C	D
1	Phone number	Area code		
2	001 (206) 123 4567	206		
3	001 (206) 123 4567	=MID(A3,6,3)		
4				

（演示）Excel数据清洗：自定义大小写

MID

=UPPER(LEFT(A5,1))&LOWER(MID(A5,2,60))

	A	B	C	D
1	City Name	LOWER	UPPER	UPPER and LOWER
2	boston	boston	BOSTON	Boston
3	BOSTON	boston	BOSTON	Boston
4	BostoN	boston	BOSTON	Boston
5	bOSTON	boston	BOSTON	=UPPER(LEFT(A5,1))&LOWER(MID(A5,2,60))
6				
7				
8				

（演示）Excel数据清洗：删除重复值

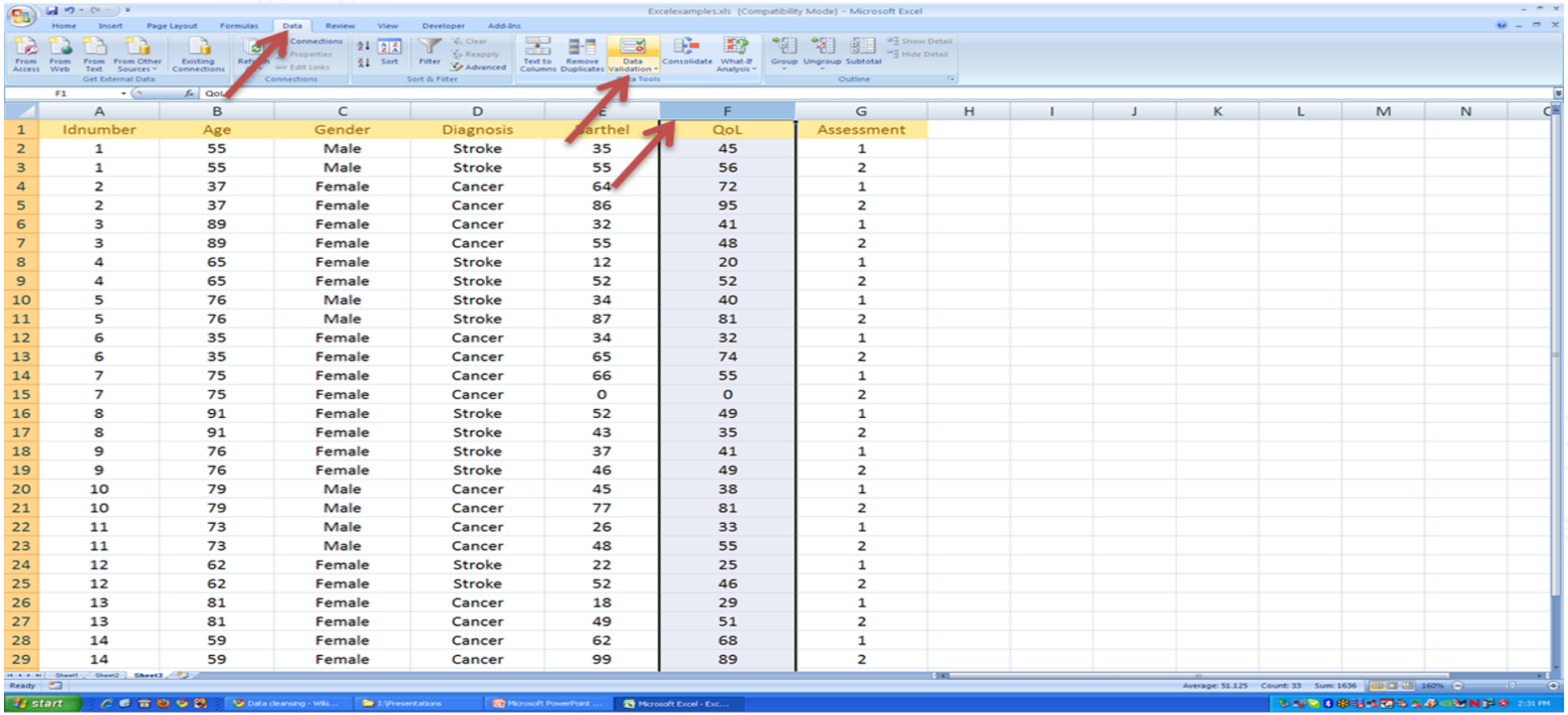
IF | =IF(C6=C5,"Duplicate","")

	A	B	C	D	E	F
1	First Name	Last Name	First & Last Name	Check for duplicates		
2	Jane	Doe	Jane Doe			
3	John	Doe	John Doe			
4	Georg	Lucas	Georg Lucas			
5	Georg	Lucas	Georg Lucas	Duplicate		
6	Steven	Spielberg	Steven Spielberg	=IF(C6=C5,"Duplicate","")		
7						
8						
9						
10						

IF(logical_test, [value_if_true], [value_if_false])

（演示）Excel数据审核：

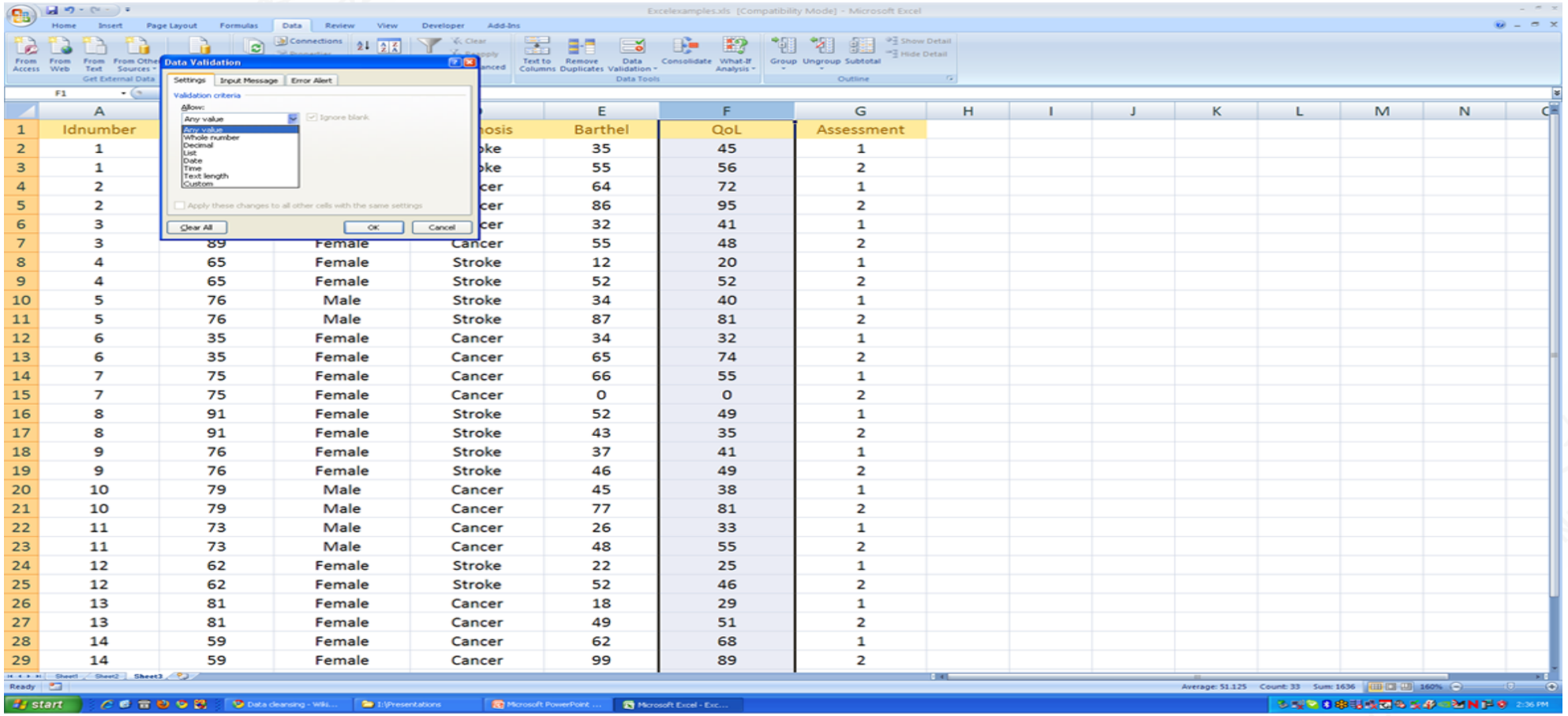
a.选择区域　b.设定类型　c.设置范围　d.设置提示



	A	B	C	D	E	F	G
1	Idnumber	Age	Gender	Diagnosis	Barthel	QoL	Assessment
2	1	55	Male	Stroke	35	45	1
3	1	55	Male	Stroke	55	56	2
4	2	37	Female	Cancer	64	72	1
5	2	37	Female	Cancer	86	95	2
6	3	89	Female	Cancer	32	41	1
7	3	89	Female	Cancer	55	48	2
8	4	65	Female	Stroke	12	20	1
9	4	65	Female	Stroke	52	52	2
10	5	76	Male	Stroke	34	40	1
11	5	76	Male	Stroke	87	81	2
12	6	35	Female	Cancer	34	32	1
13	6	35	Female	Cancer	65	74	2
14	7	75	Female	Cancer	66	55	1
15	7	75	Female	Cancer	0	0	2
16	8	91	Female	Stroke	52	49	1
17	8	91	Female	Stroke	43	35	2
18	9	76	Female	Stroke	37	41	1
19	9	76	Female	Stroke	46	49	2
20	10	79	Male	Cancer	45	38	1
21	10	79	Male	Cancer	77	81	2
22	11	73	Male	Cancer	26	33	1
23	11	73	Male	Cancer	48	55	2
24	12	62	Female	Stroke	22	25	1
25	12	62	Female	Stroke	52	46	2
26	13	81	Female	Cancer	18	29	1
27	13	81	Female	Cancer	49	51	2
28	14	59	Female	Cancer	62	68	1
29	14	59	Female	Cancer	99	89	2

（演示）Excel数据审核：

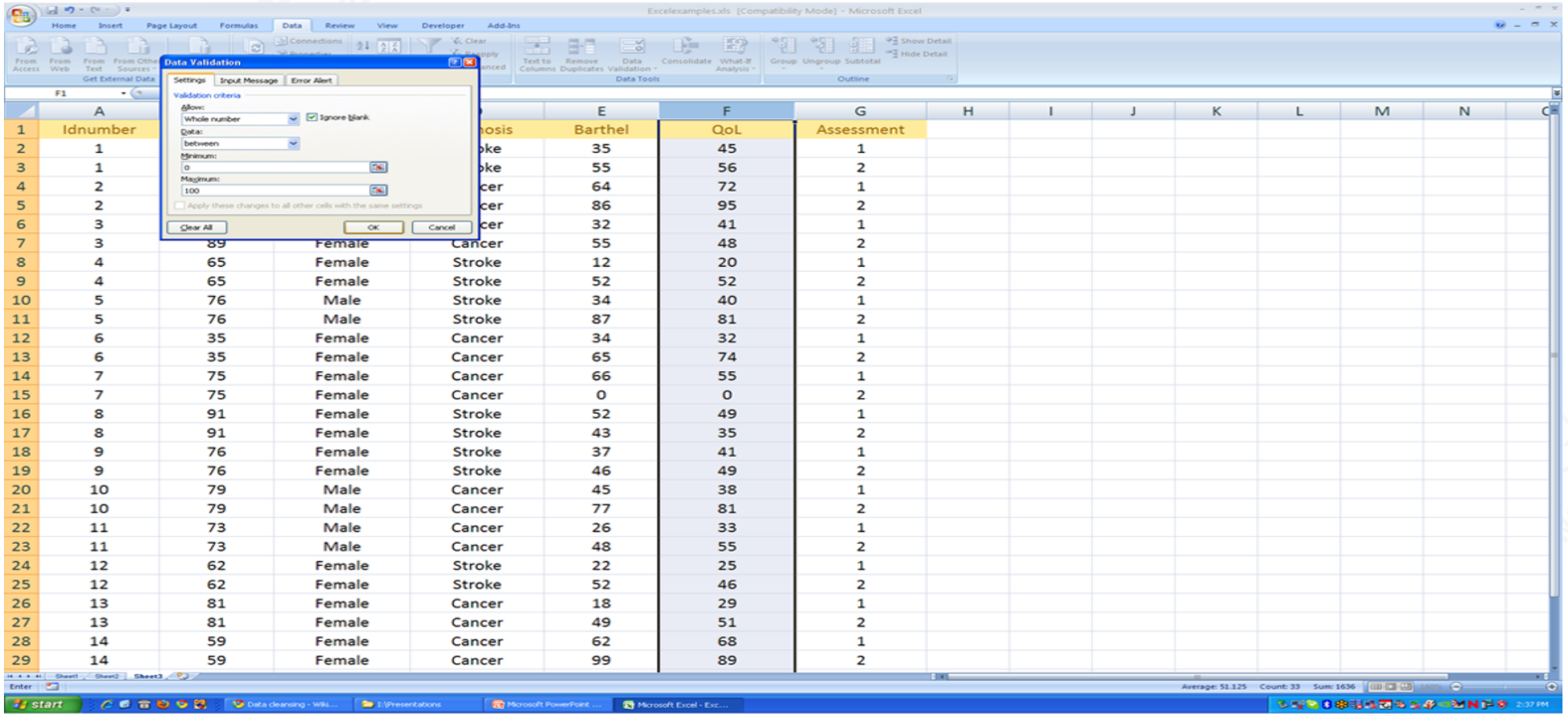
a.选择区域　b.设定类型　c.设置范围　d.设置提示



	A	B	C	D	E	F	G
1	Idnumber				Barthel	QoL	Assessment
2	1				35	45	1
3	1				55	56	2
4	2				64	72	1
5	2				86	95	2
6	3				32	41	1
7	3	89	Female	Cancer	55	48	2
8	4	65	Female	Stroke	12	20	1
9	4	65	Female	Stroke	52	52	2
10	5	76	Male	Stroke	34	40	1
11	5	76	Male	Stroke	87	81	2
12	6	35	Female	Cancer	34	32	1
13	6	35	Female	Cancer	65	74	2
14	7	75	Female	Cancer	66	55	1
15	7	75	Female	Cancer	0	0	2
16	8	91	Female	Stroke	52	49	1
17	8	91	Female	Stroke	43	35	2
18	9	76	Female	Stroke	37	41	1
19	9	76	Female	Stroke	46	49	2
20	10	79	Male	Cancer	45	38	1
21	10	79	Male	Cancer	77	81	2
22	11	73	Male	Cancer	26	33	1
23	11	73	Male	Cancer	48	55	2
24	12	62	Female	Stroke	22	25	1
25	12	62	Female	Stroke	52	46	2
26	13	81	Female	Cancer	18	29	1
27	13	81	Female	Cancer	49	51	2
28	14	59	Female	Cancer	62	68	1
29	14	59	Female	Cancer	99	89	2

（演示）Excel数据审核：

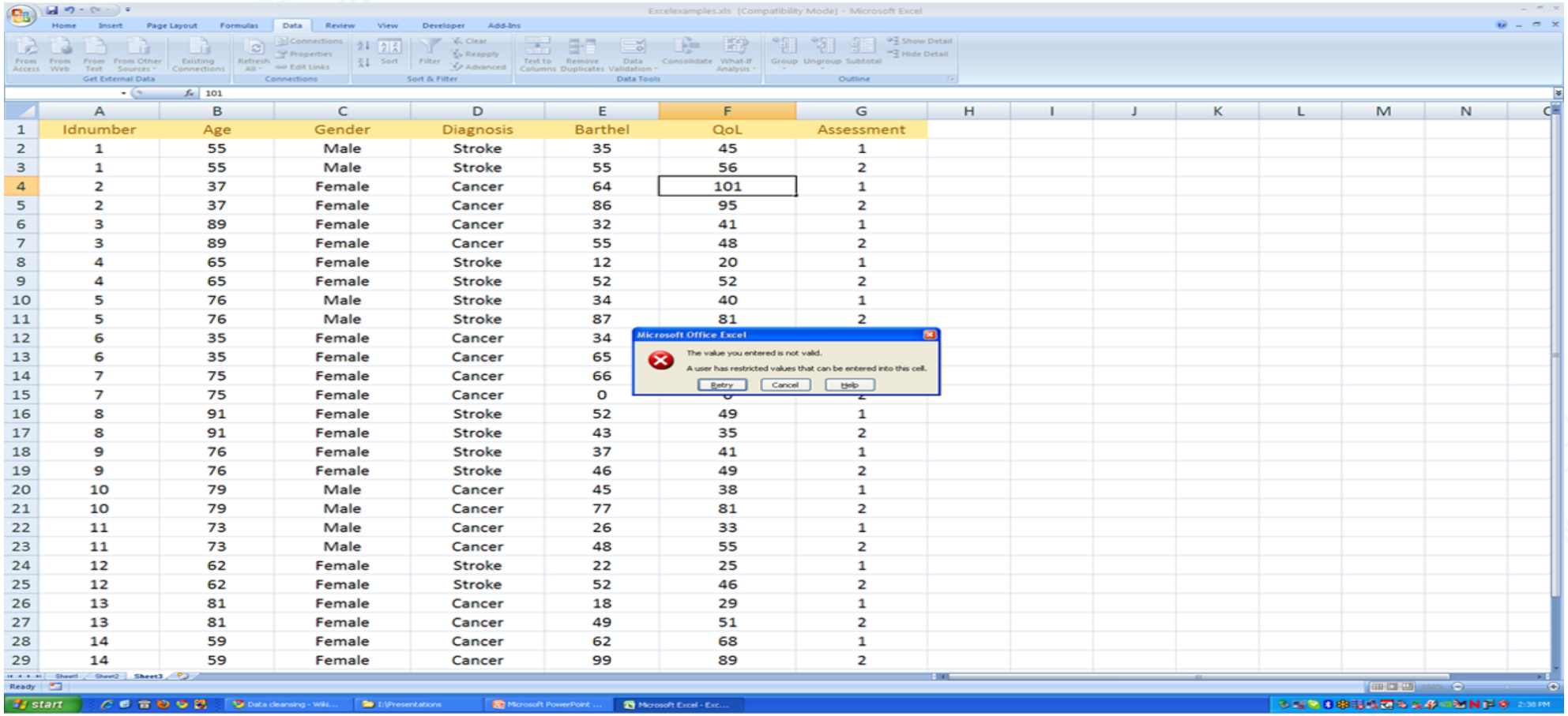
a.选择区域　b.设定类型　c.设置范围　d.设置提示



	A	B	C	D	E	F	G
1	Idnumber				Barthel	QoL	Assessment
2	1				35	45	1
3	1				55	56	2
4	2				64	72	1
5	2				86	95	2
6	3				32	41	1
7	3	89	Female	Cancer	55	48	2
8	4	65	Female	Stroke	12	20	1
9	4	65	Female	Stroke	52	52	2
10	5	76	Male	Stroke	34	40	1
11	5	76	Male	Stroke	87	81	2
12	6	35	Female	Cancer	34	32	1
13	6	35	Female	Cancer	65	74	2
14	7	75	Female	Cancer	66	55	1
15	7	75	Female	Cancer	0	0	2
16	8	91	Female	Stroke	52	49	1
17	8	91	Female	Stroke	43	35	2
18	9	76	Female	Stroke	37	41	1
19	9	76	Female	Stroke	46	49	2
20	10	79	Male	Cancer	45	38	1
21	10	79	Male	Cancer	77	81	2
22	11	73	Male	Cancer	26	33	1
23	11	73	Male	Cancer	48	55	2
24	12	62	Female	Stroke	22	25	1
25	12	62	Female	Stroke	52	46	2
26	13	81	Female	Cancer	18	29	1
27	13	81	Female	Cancer	49	51	2
28	14	59	Female	Cancer	62	68	1
29	14	59	Female	Cancer	99	89	2

（演示）Excel数据审核：

a.选择区域　　b.设定类型　　c.设置范围　　**d.设置提示**



	A	B	C	D	E	F	G
1	Idnumber	Age	Gender	Diagnosis	Barthel	QoL	Assessment
2	1	55	Male	Stroke	35	45	1
3	1	55	Male	Stroke	55	56	2
4	2	37	Female	Cancer	64	101	1
5	2	37	Female	Cancer	86	95	2
6	3	89	Female	Cancer	32	41	1
7	3	89	Female	Cancer	55	48	2
8	4	65	Female	Stroke	12	20	1
9	4	65	Female	Stroke	52	52	2
10	5	76	Male	Stroke	34	40	1
11	5	76	Male	Stroke	87	81	2
12	6	35	Female	Cancer	34		
13	6	35	Female	Cancer	65		
14	7	75	Female	Cancer	66		
15	7	75	Female	Cancer	0		
16	8	91	Female	Stroke	52	49	1
17	8	91	Female	Stroke	43	35	2
18	9	76	Female	Stroke	37	41	1
19	9	76	Female	Stroke	46	49	2
20	10	79	Male	Cancer	45	38	1
21	10	79	Male	Cancer	77	81	2
22	11	73	Male	Cancer	26	33	1
23	11	73	Male	Cancer	48	55	2
24	12	62	Female	Stroke	22	25	1
25	12	62	Female	Stroke	52	46	2
26	13	81	Female	Cancer	18	29	1
27	13	81	Female	Cancer	49	51	2
28	14	59	Female	Cancer	62	68	1
29	14	59	Female	Cancer	99	89	2

案例：家庭税收情况

案例说明：一项家庭税收情况调查，一共收集了样本数 n = 73262个家庭在13个变量上的基本情况。

```
Rows: 73,262
Columns: 13
$ id             <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...
$ custid         <chr> "000006646_03", "000007827_01", "000008359...
$ sex            <fct> Male, Female, Female, Female, Male, Male, ...
$ is_employed    <lgl> TRUE, NA, TRUE, NA, TRUE, NA, TRUE, NA, TR...
$ income         <dbl> 22000, 23200, 21000, 37770, 39000, 11100, ...
$ marital_status <fct> Never married, Divorced/Separated, Never m...
$ health_ins     <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE,...
$ housing_type   <fct> Homeowner free and clear, Rented, Homeowne...
$ recent_move    <lgl> FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, F...
$ num_vehicles   <dbl> 0, 0, 2, 1, 2, 2, 2, 2, 5, 3, 2, 2, 5, 1, ...
$ age            <dbl> 24, 82, 31, 93, 67, 76, 26, 73, 27, 54, 61...
$ state_of_res   <fct> Alabama, Alabama, Alabama, Alabama, Alabam...
$ gas_usage      <dbl> 210, 3, 40, 120, 3, 200, 3, 50, 3, 20, 3, ...
```

（税收案例）变量视图

序号	变量名	变量含义
1	id	样本编号
2	custid	客户编号
3	sex	性别
4	is_employed	雇佣情况
5	income	收入
6	marital_status	婚姻状况
7	health_ins	医保情况
8	housing_type	住房情况
9	recent_move	迁徙情况
10	num_vehicles	机动车数量

Showing 1 to 10 of 13 entries

Previous 1 2 Next

（税收案例）数据视图

数据包含的样本数 $n = 73262$，下表展示了前500行。

id	sex	is_employed	income	marital_status	health_ins	recent_move
1	Male	true	22000	Never married	true	false
2	Female		23200	Divorced/Separated	true	true
3	Female	true	21000	Never married	true	false
4	Female		37770	Widowed	true	false
5	Male	true	39000	Divorced/Separated	true	false
6	Male		11100	Married	true	false
7	Female	true	25800	Married	false	false
8	Female		34600	Married	true	false

Showing 1 to 8 of 500 entries

Previous 1 2 3 4 5 … 63 Next

（税收案例）数据清洗问题——存在混合编码

a.变量取值　b.编码问题　c.处理办法

gas_usage变量表示用油支出，样本数据中的分布情况如下：

1	2	3	4	10	20	30	40	50	60	70
2389	6609	24984	455	1376	4163	5149	4242	4118	2585	1909
80	90	100	110	120	130	140	150	160	170	180
2330	1331	2651	636	836	700	446	1049	299	238	331
190	200	210	220	230	240	250	260	270	280	290
175	823	113	88	111	89	256	54	40	61	33
300	310	320	330	340	350	360	370	380	390	400
255	13	26	20	36	53	5	39	19	9	79
410	420	430	440	450	460	470	480	490	510	520
37	15	1	3	48	43	39	72	35	3	3
540	570	<NA>								
9	11	1720								

（税收案例）数据清洗问题——存在混合编码

a.变量取值　b.编码问题　c.处理办法

根据“数据集说明”，我们容易发现`gas_usage`变量数据存在混合编码问题：

- 混合了数值（number）和字符（string）：
  - 1表示包含在电子支付中；2表示包含在出租或分户中；3表示没有用油；
  - 004-999表示用油支出数（美元）。
- 有缺失值：
  - `NA`表示缺失。

（税收案例）数据清洗问题——存在混合编码

a.变量取值　b.编码问题　c.处理办法

我们需要把原来的变量`gass_usage`提取构建为4个新变量：`gas_usage_new`、`gas_with_rent`、`gas_with_electricity`、`no_gas_bill`。

id	gas_usage	gas_usage_new	gas_with_rent	gas_with_electricity	no_gas_bill
1	210	210	false	false	false
2	3		false	false	true
3	40	40	false	false	false
4	120	120	false	false	false
5	3		false	false	true

Showing 1 to 5 of 500 entries　Previous 1 2 3 4 5 … 100 Next

（税收案例）数据清洗问题——数据范围失常

a.年龄范围　b.收入范围　c.处理办法

我们容易发现，年龄变量age存在取值为0，或大于100的情形。

```
   0   21   22   23   24   25   26   27   28   29   30   31   32   33
  77 1365 1312 1348 1302 1424 1489 1407 1364 1284 1444 1303 1351 1327
  34   35   36   37   38   39   40   41   42   43   44   45   46   47
1285 1367 1284 1216 1226 1190 1327 1202 1215 1203 1262 1449 1392 1217
  48   49   50   51   52   53   54   55   56   57   58   59   60   61
1244 1197 1355 1352 1349 1397 1429 1292 1392 1326 1290 1307 1333 1281
  62   63   64   65   66   67   68   69   70   71   72   73   74   75
1220 1175 1124 1113 1082 1062  999 1026  792  696  800  757  645  549
  76   77   78   79   80   81   82   83   84   85   86   87   88   89
 519  530  517  431  419  424  358  314  301  317  267  251  202  162
  90   91   92   93   94   95   96  100  110  114  120 <NA>
 130   57   83  124  277   91    6   72   66   62   66    0
```

a.年龄范围　b.收入范围　c.处理办法

此外，我们还发现收入变量income存在取值小于0的情形。

```
-6900   -6800   -6700   -6600   -6100   -6000   -5900   -5800
    1       2       2       1       1       5       1       1
-5700   -5500   -5400   -5300   -5200   -5000   -4900   -4700
    1       2       2       1       2       2       1       2
-4520   -4500   -4200   -3500   -3100   -2800   -2700   -2400
    1       2       1       1       1       1       1       1
-1800   -1700   -1500    -800    -700    -630    -400    -160
    1       1       2       1       1       1       1       1
    0       1       4      10      20      30      40      50
 6811      18      18       8      10      16       4      10
   60      70      80      90     100     110     120     130
   10       6       7       3      39       4       9       6
  140     150     160     170     180     190     200     210
    7       8       4       5       3       5      56       3
  220     230     240     250     260     270     280     300
```

（税收案例）数据清洗问题——数据范围失常

a.年龄范围　　b.收入范围　　c.处理办法

- age取值等于0，则转换为缺失值NA
- income取值小于0，则转换为缺失值NA

id	age	age_new	income	income_new
1274	0		0	NA
9494	0		0	NA
9533	0		0	NA
11423	0		0	NA
26439	0		0	NA

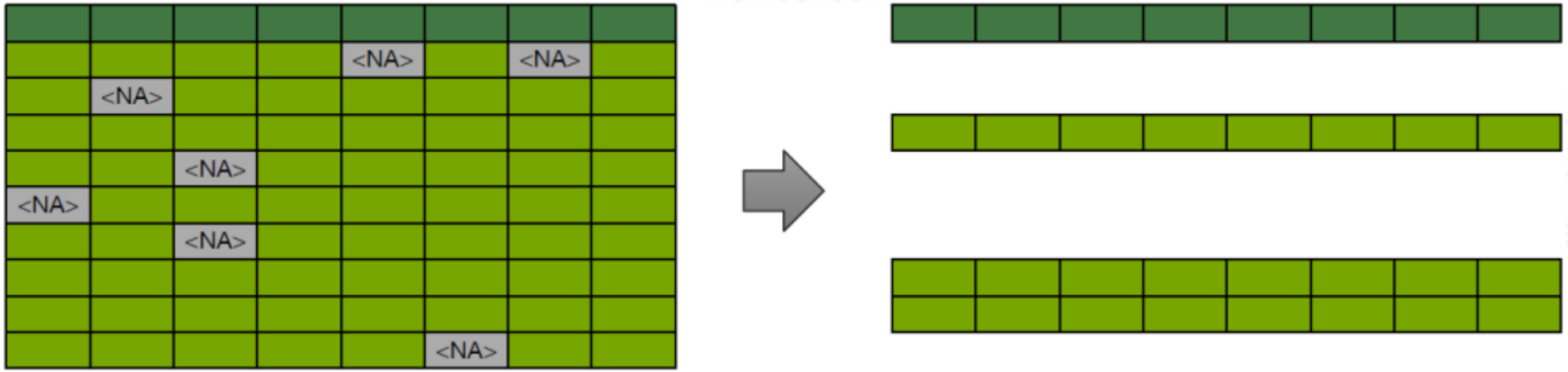
Showing 1 to 5 of 10 entries　　Previous　1　2　Next

（税收案例）数据清洗问题——存在缺失值

a.简单删除法　b.填加标签法　c.数值替换法

处理缺失值最简单粗暴的办法就是直接删除行或列，但是这样也会直接去掉很多样本信息：

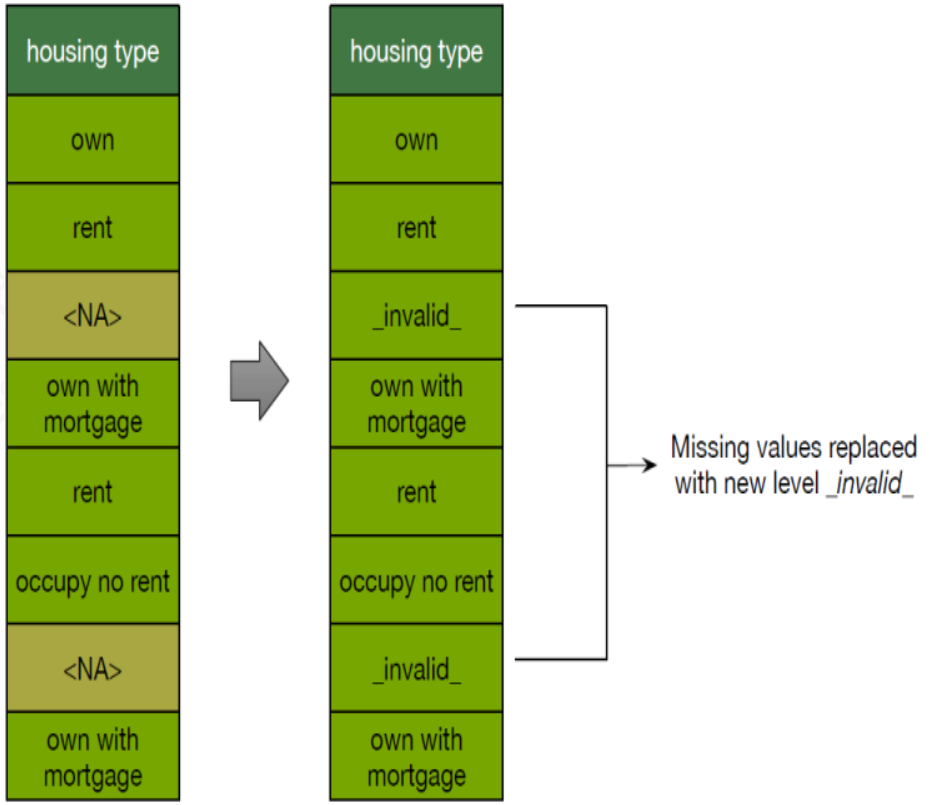
- 行删除(row delete)
- 列删除(column delete)



（税收案例）数据清洗问题——存在缺失值

a.简单删除法　b.填加标签法　c.数值替换法

对于**分类变量**的缺失值，可以直接加一个特定标签（level）。



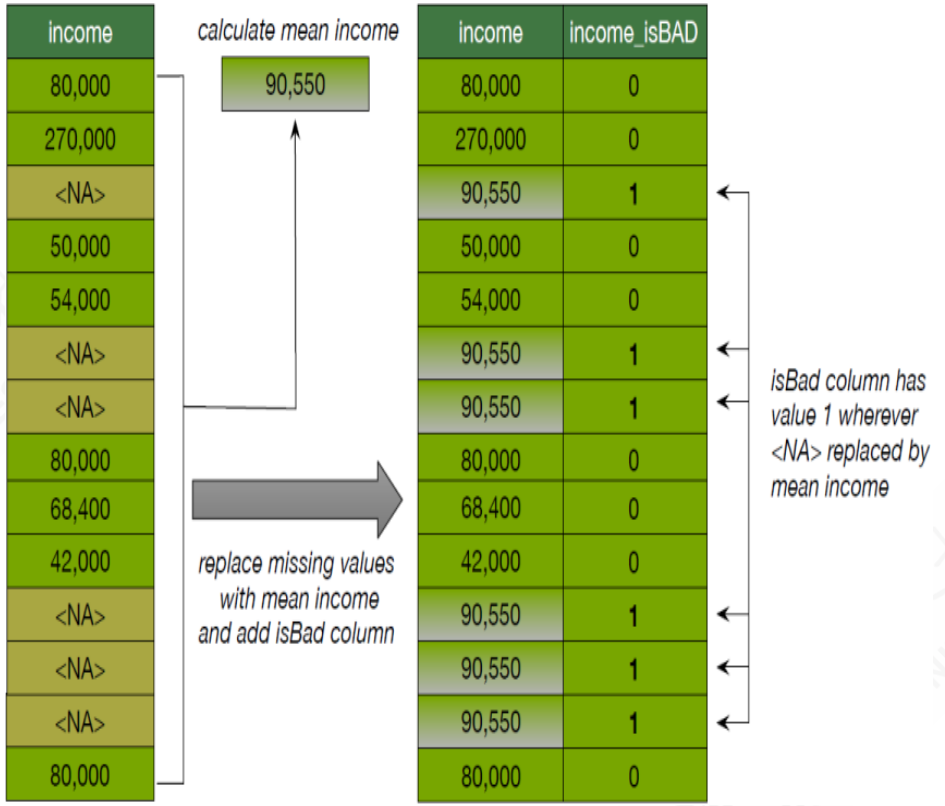
（税收案例）数据清洗问题——存在缺失值

a.简单删除法　b.填加标签法　**c.数值替换法**

对于**数值型变量**的缺失值，可以给缺失值进行**补值**：

- 用均值代替
- 使用各种插值办法

补值后，一定要记得新增加1个变量，指明哪些样本进行了补值操作！



a.处理前变量　b.处理后变量　c.处理前数据　d.处理后数据

处理前数据集customer_data的变量：

```
 [1] "id"                   "custid"
 [3] "sex"                  "is_employed"
 [5] "income"               "marital_status"
 [7] "health_ins"           "housing_type"
 [9] "recent_move"          "num_vehicles"
[11] "age"                  "state_of_res"
[13] "gas_usage"            "gas_with_rent"
[15] "gas_with_electricity" "no_gas_bill"
[17] "gas_usage_new"        "age_new"
[19] "income_new"
```

a.处理前变量　b.处理后变量　c.处理前数据　d.处理后数据

采用R包vtreat进行自动补值后新数据集training_prepared的变量：

```
 [1] "custid"                        "health_ins"
 [3] "id"                            "sex"
 [5] "is_employed"                   "is_employed_isBAD"
 [7] "income"                        "marital_status"
 [9] "housing_type"                  "recent_move"
[11] "recent_move_isBAD"             "num_vehicles"
[13] "num_vehicles_isBAD"            "age"
[15] "state_of_res"                  "gas_usage"
[17] "gas_usage_isBAD"               "gas_with_rent"
[19] "gas_with_rent_isBAD"           "gas_with_electricity"
[21] "gas_with_electricity_isBAD"    "no_gas_bill"
[23] "no_gas_bill_isBAD"             "gas_usage_new"
[25] "gas_usage_new_isBAD"           "age_new"
[27] "age_new_isBAD"                 "income_new"
```

（税收案例）数据清洗问题——R补值工具

a.处理前变量　b.处理后变量　c.处理前数据　d.处理后数据

处理前的数据缺失情况（4个变量，前6行）：

	id	is_employed	num_vehicles	housing_type	health_ins
55		true			false
65		true			true
162					false
207					false
219					true
294					true

（税收案例）数据清洗问题——R补值工具

a.处理前变量　b.处理后变量　c.处理前数据　**d.处理后数据**

补值工具**处理后**的数据补齐情况（前6行）：

id	is_employed	is_employed_isBAD	num_vehicles	num_vehicles_isBAD	hou
55	1.00	0	2.07	1	_i
65	1.00	0	2.07	1	_i
162	0.95	1	2.07	1	_i
207	0.95	1	2.07	1	_i
219	0.95	1	2.07	1	_i
294	0.95	1	2.07	1	_i

（税收案例）数据变换

数据变换主要目的是为让数据能更适合于建模分析（modeling）。

主要变换操作包括：

- 中位数（median）变换
- 均值（mean）变换
- 对数化（log）变换 $ln(X_i)$：
- 标准化（sd）变换 $\frac{(X_i-\bar{X})}{S_X}$
- 把**连续变量**转换为**离散变量**

（税收案例）数据变换：原始收入变量

假设我们还有51个州的收入的中位数数据集（`median_income_table`）：

state_of_res	median_income
Alabama	21100
Alaska	32050
Arizona	26000
Arkansas	22900
California	25000
Colorado	32000
Connecticut	36000
Delaware	29400

Showing 1 to 8 of 51 entries

Previous 1 2 3 4 5 6 7 Next

（税收案例）数据变换：收入变量的中位数变换

a.变换公式　b.R计算过程　c.变换结果

现在我们可以，通过如下步骤进行收入的**中位数**变换：

- 把收入的州数据集（`median_income_table`）与前面的案例数据集（`training_prepared`）匹配起来
- 对收入进行中位数变换 $income\_normalized = \frac{income_i}{median\_income}$。

（税收案例）数据变换：收入变量的中位数变换

a.变换公式　b.R计算过程　c.变换结果

下面展示的是用R软件进行前述的匹配和计算过程：

```
mean_income <- round(mean(training_prepared$income, na.rm = T),2)
sd_income <- round(sd(training_prepared$income, na.rm = T),2)

training_prepared <-  training_prepared %>%
  left_join(., median_income_table, by="state_of_res") %>%
  mutate(income_byMedian = income/median_income,
         income_byMean = income/mean_income,
         income_bySd = (income- mean_income)/sd_income,
         income_byLog10 =log10(income))
```

（税收案例）数据变换：收入变量的中位数变换

a.变换公式　b.R计算过程　c.变换结果

最终得到收入的中位数变换结果：

id	income	median_income	income_byMedian
1	22000	21100	1.0427
2	23200	21100	1.0995
3	21000	21100	0.9953
4	37770	21100	1.7900
5	39000	21100	1.8483

Showing 1 to 5 of 500 entries　Previous 1 2 3 4 5 … 100 Next

（税收案例）数据变换：收入变量的均值变换

a.变换公式　b.R计算过程　c.变换结果

现在我们可以，通过如下步骤进行收入的均值变换：

- 对收入进行中位数变换 $income\_byMean = \frac{income_i}{\overline{income}}$。

（税收案例）数据变换：收入变量的均值变换

a.变换公式　b.R计算过程　c.变换结果

下面展示的是用R计算过程：

```
mean_income <- round(mean(training_prepared$income, na.rm = T),2)
sd_income <- round(sd(training_prepared$income, na.rm = T),2)

training_prepared <-  training_prepared %>%
  left_join(., median_income_table, by="state_of_res") %>%
  mutate(income_byMedian = income/median_income,
         income_byMean = income/mean_income,
         income_bySd = (income- mean_income)/sd_income,
         income_byLog10 =log10(income))
```

（税收案例）数据变换：收入变量的均值变换

a.变换公式　b.R计算过程　**c.变换结果**

最终得到收入的**均值变换**结果：

id	income	income_byMedian	income_byMean
1	22000	1.0427	0.5268
2	23200	1.0995	0.5555
3	21000	0.9953	0.5028
4	37770	1.7900	0.9044
5	39000	1.8483	0.9338

Showing 1 to 5 of 500 entries　Previous　1　2　3　4　5　…　100　Next

（税收案例）数据变换：收入变量的标准化变换

a.变换公式　b.R计算过程　c.变换结果

现在我们可以，通过如下步骤进行收入的**标准化**变换：

- 对收入进行标准化变换 $income\_bySd = \frac{income_i - \overline{income}}{sd\_income}$。

（税收案例）数据变换：收入变量的标准化变换

a.变换公式　b.R计算过程　c.变换结果

下面展示的是用R计算过程：

```
mean_income <- round(mean(training_prepared$income, na.rm = T),2)
sd_income <- round(sd(training_prepared$income, na.rm = T),2)

training_prepared <-  training_prepared %>%
  left_join(., median_income_table, by="state_of_res") %>%
  mutate(income_byMedian = income/median_income,
         income_byMean = income/mean_income,
         income_bySd = (income- mean_income)/sd_income,
         income_byLog10 =log10(income))
```

（税收案例）数据变换：收入变量的标准化变换

a.变换公式　b.R计算过程　c.变换结果

最终得到收入的**标准化**变换结果：

id	income	income_byMedian	income_byMean	income_bySd
1	22000	1.0427	0.5268	-0.3401
2	23200	1.0995	0.5555	-0.3194
3	21000	0.9953	0.5028	-0.3573
4	37770	1.7900	0.9044	-0.0687
5	39000	1.8483	0.9338	-0.0476

Showing 1 to 5 of 500 entries　Previous 1 2 3 4 5 … 100 Next

（税收案例）数据变换：收入变量的对数化变换

a.变换公式　b.R计算过程　c.变换结果

现在我们可以，通过如下步骤进行收入的**对数化**变换：

- 对收入进行**对数化**变换 $income\_byLog10 = log_{10}^{(income_i)}$。

（税收案例）数据变换：收入变量的对数化变换

a.变换公式　b.R计算过程　c.变换结果

下面展示的是用R计算过程：

```
mean_income <- round(mean(training_prepared$income, na.rm = T),2)
sd_income <- round(sd(training_prepared$income, na.rm = T),2)

training_prepared <-  training_prepared %>%
  left_join(., median_income_table, by="state_of_res") %>%
  mutate(income_byMedian = income/median_income,
         income_byMean = income/mean_income,
         income_bySd = (income- mean_income)/sd_income,
         income_byLog10 =log10(income))
```

（税收案例）数据变换：收入变量的对数化变换

a.变换公式　b.R计算过程　c.变换结果

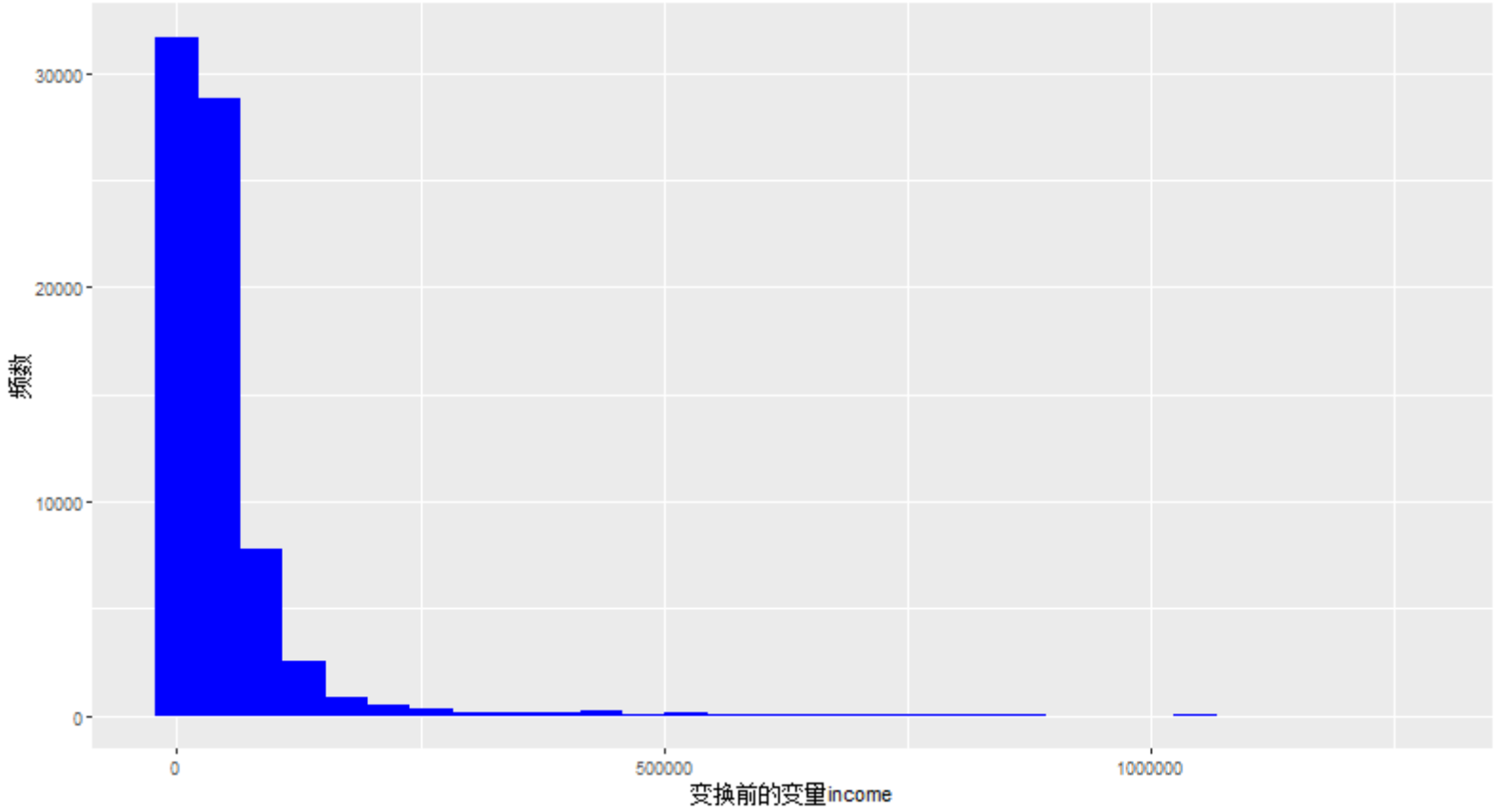
最终得到收入的对数化变换结果：

id	income	income_byMedian	income_byMean	income_bySd	income_byLog10
1	22000	1.0427	0.5268	-0.3401	4.3424
2	23200	1.0995	0.5555	-0.3194	4.3655
3	21000	0.9953	0.5028	-0.3573	4.3222
4	37770	1.7900	0.9044	-0.0687	4.5771
5	39000	1.8483	0.9338	-0.0476	4.5911

Showing 1 to 5 of 500 entries　Previous 1 2 3 4 5 … 100 Next

（税收案例）数据变换：收入变换前后分布的比较

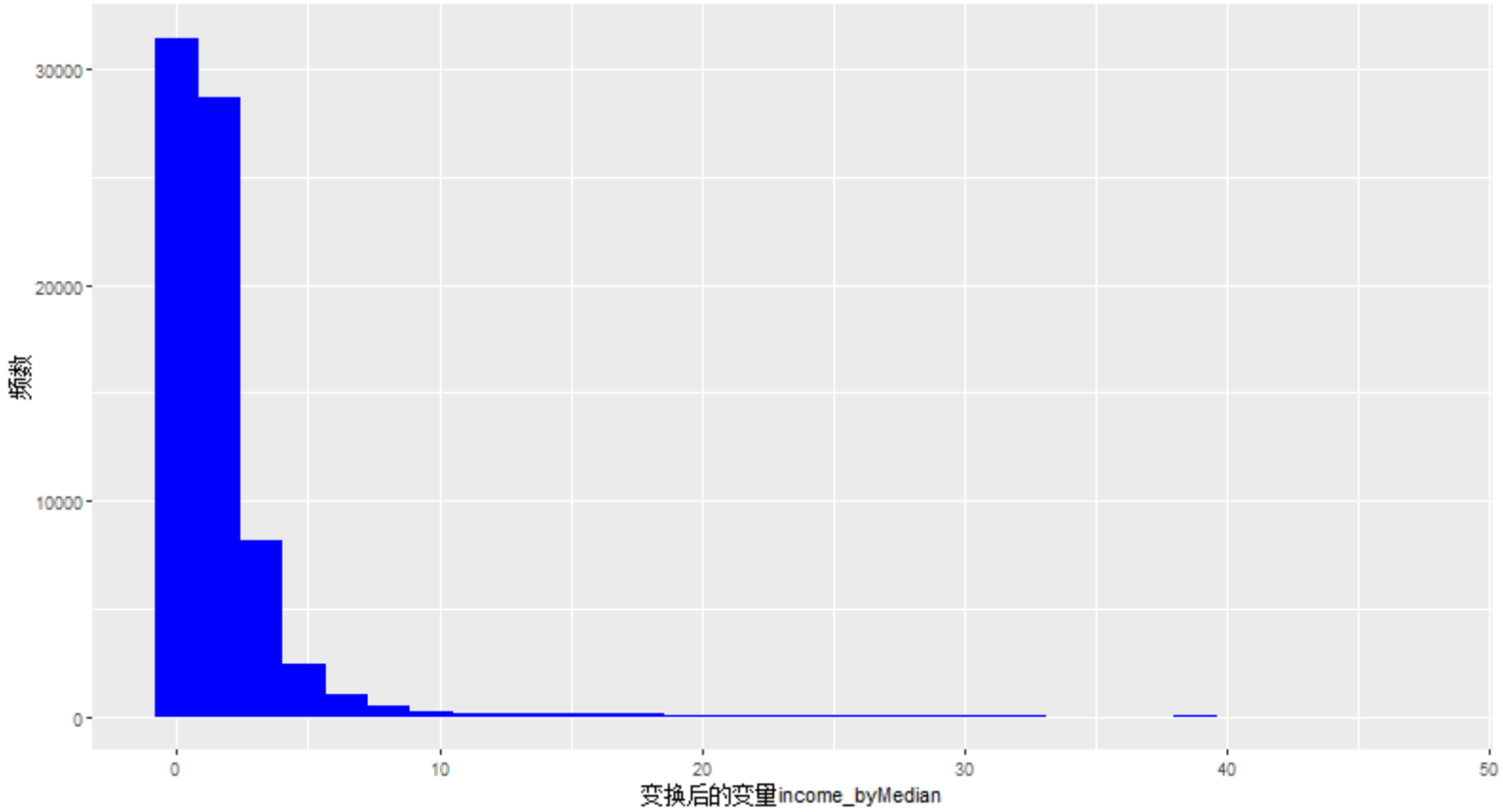
a.原变量　b.中位数变换　c.均值变换　d.标准化　e.对数化



年收入的直方图

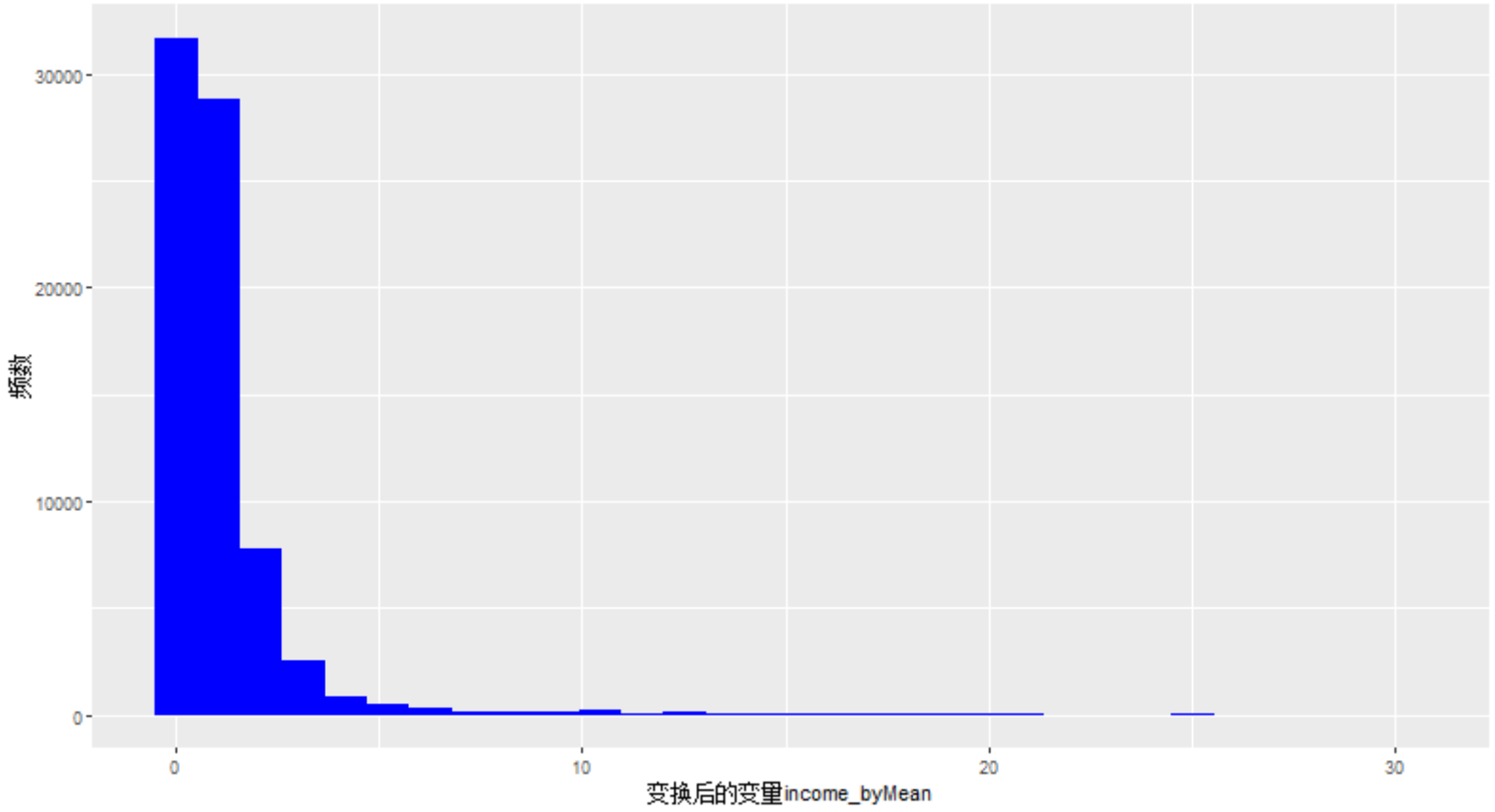
（税收案例）数据变换：收入变换前后分布的比较

a.原变量　b.中位数变换　c.均值变换　d.标准化　e.对数化



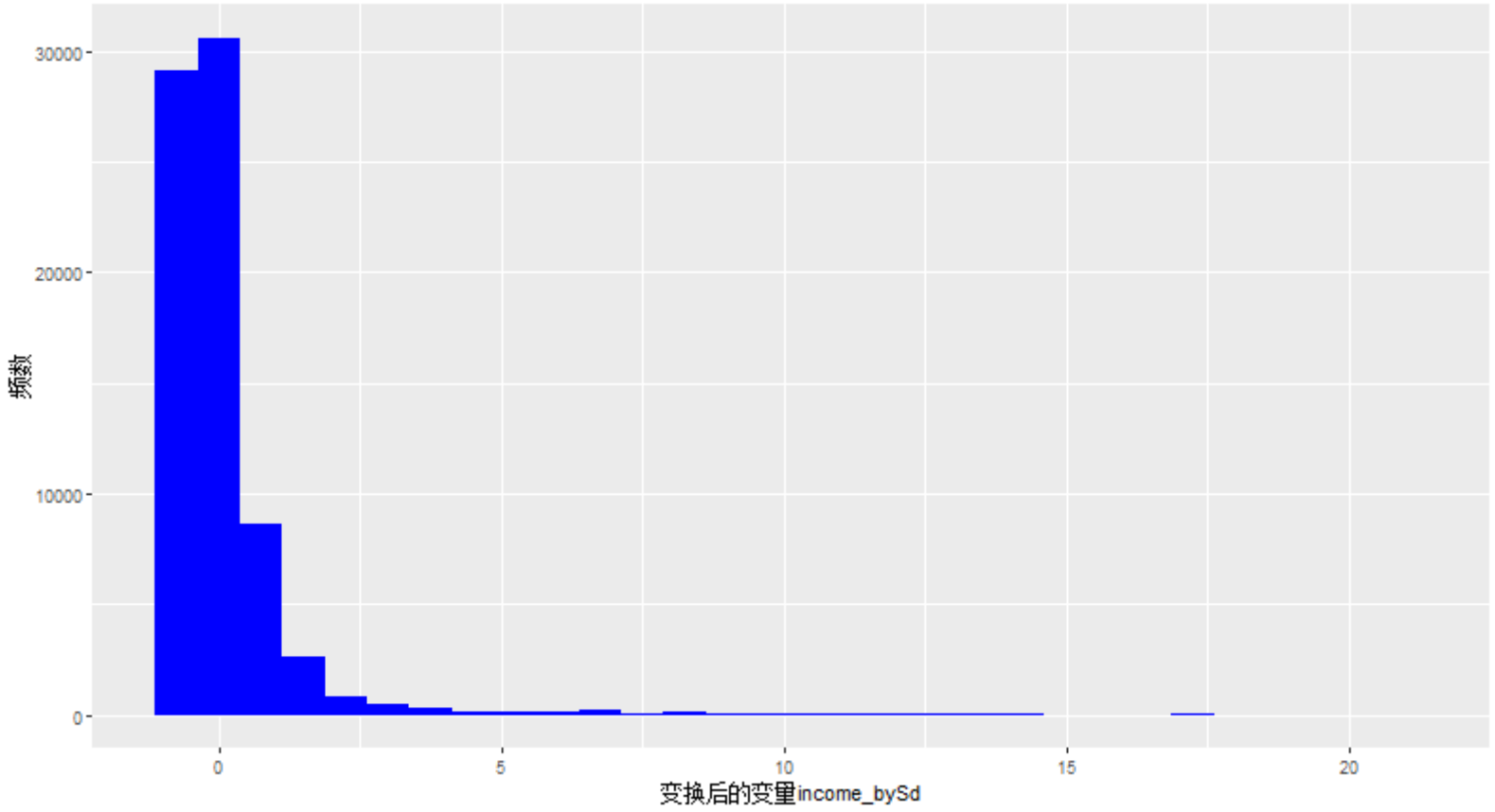
（税收案例）数据变换：收入变换前后分布的比较

a.原变量　b.中位数变换　**c.均值变换**　d.标准化　e.对数化



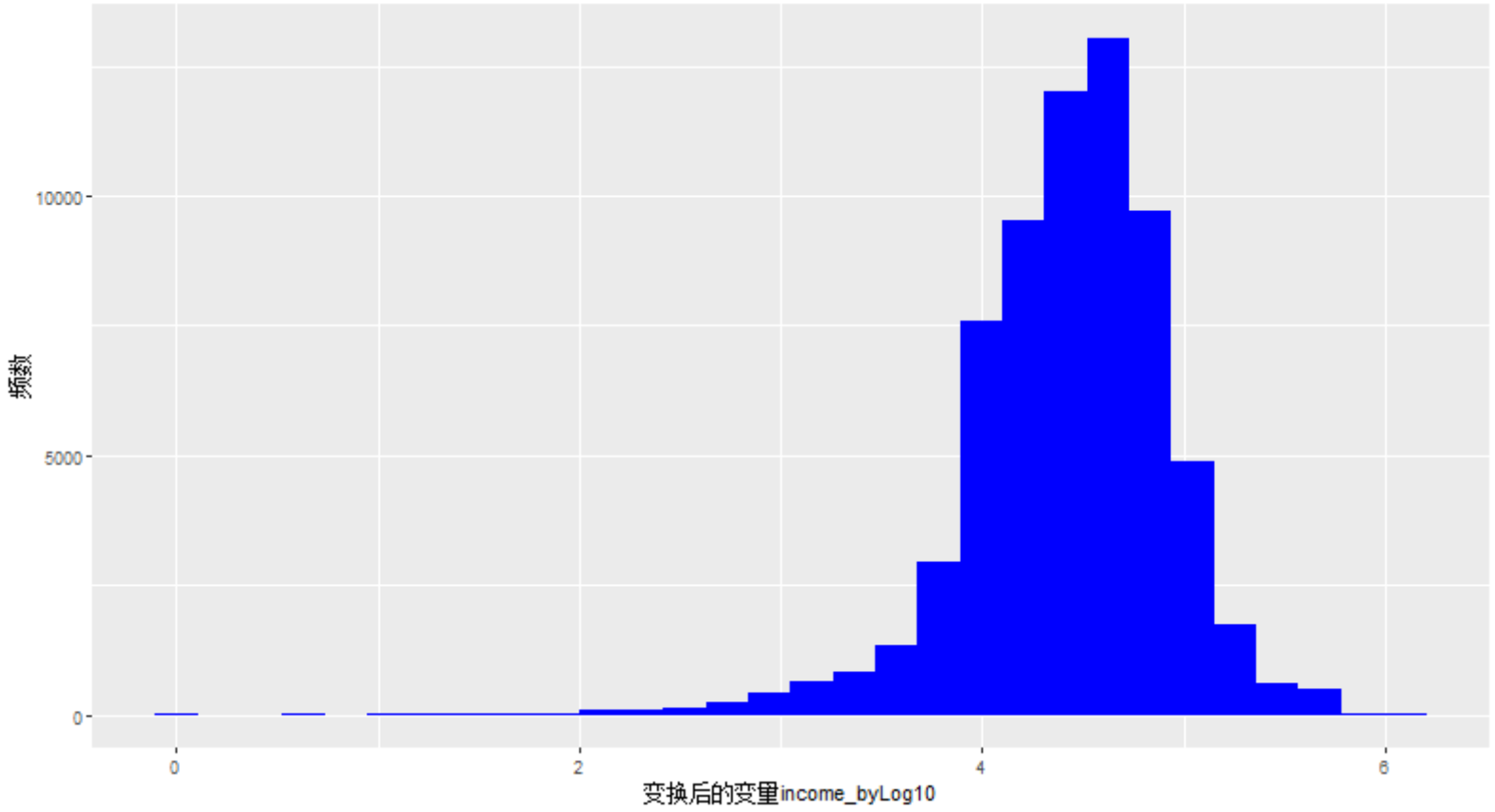
（税收案例）数据变换：收入变换前后分布的比较

a.原变量　b.中位数变换　c.均值变换　**d.标准化**　e.对数化



（税收案例）数据变换：收入变换前后分布的比较

a.原变量　b.中位数变换　c.均值变换　d.标准化　**e.对数化**



（税收案例）数据变换：批量标准化处理

a.编写R代码　b.原数据集　c.批处理后数据集

对于数据集的数值型变量（age,income,num_veicles,gas_usage），我们可以同时进行批量**标准化**变换。从而为下一步建模分析做准备。

如下是利用R软件函数scale()进行批量标准化处理的代码：

```
dataf <- training_prepared[, c("age", "income", "num_vehicles", "gas_usage")]
dataf_scaled <- scale(dataf, center=TRUE, scale=TRUE)
```

（税收案例）数据变换：批量标准化处理

a.编写R代码　b.原数据集　c.批处理后数据集

id	age	income	num_vehicles	gas_usage
1	24	22000	0	210
2	82	23200	0	3
3	31	21000	2	40
4	93	37770	1	120
5	67	39000	2	3

Showing 1 to 5 of 500 entries

Previous 1 2 3 4 5 … 100 Next

（税收案例）数据变换：批量标准化处理

a.编写R代码　b.原数据集　c.批处理后数据集

id	age	income	num_vehicles	gas_usage
1	-1.39	-0.34	-1.79	2.71
2	1.82	-0.32	-1.79	-0.61
3	-1.00	-0.36	-0.06	-0.02
4	2.42	-0.07	-0.92	1.27
5	0.99	-0.05	-0.06	-0.61

Showing 1 to 5 of 500 entries

Previous 1 2 3 4 5 … 100 Next

（税收案例）数据变换：批量变换为分类变量

a.编写R代码　b.变换结果

我们还可以把**连续变量**年龄(age)，变换为**分类变量**(age_range和age_cat)。其中分割依据为brks <- c(0, 18, 45, 65, Inf)。

```
brks <- c(0, 18,45, 65, Inf)

training_prepared <- training_prepared %>%
  select(id, age) %>%
  filter(!is.na(age)) %>%
  mutate(age_range = cut(age, breaks = brks,
                         include.lowest = T),
         age_cat = cut(age, breaks = brks,
                       include.lowest = T, labels = FALSE))
```

（税收案例）数据变换：批量变换为分类变量

a.编写R代码　b.变换结果

id	age	age_range	age_cat
1	24	(18,45]	2
2	82	(65,Inf]	4
3	31	(18,45]	2
4	93	(65,Inf]	4
5	67	(65,Inf]	4

Showing 1 to 5 of 500 entries

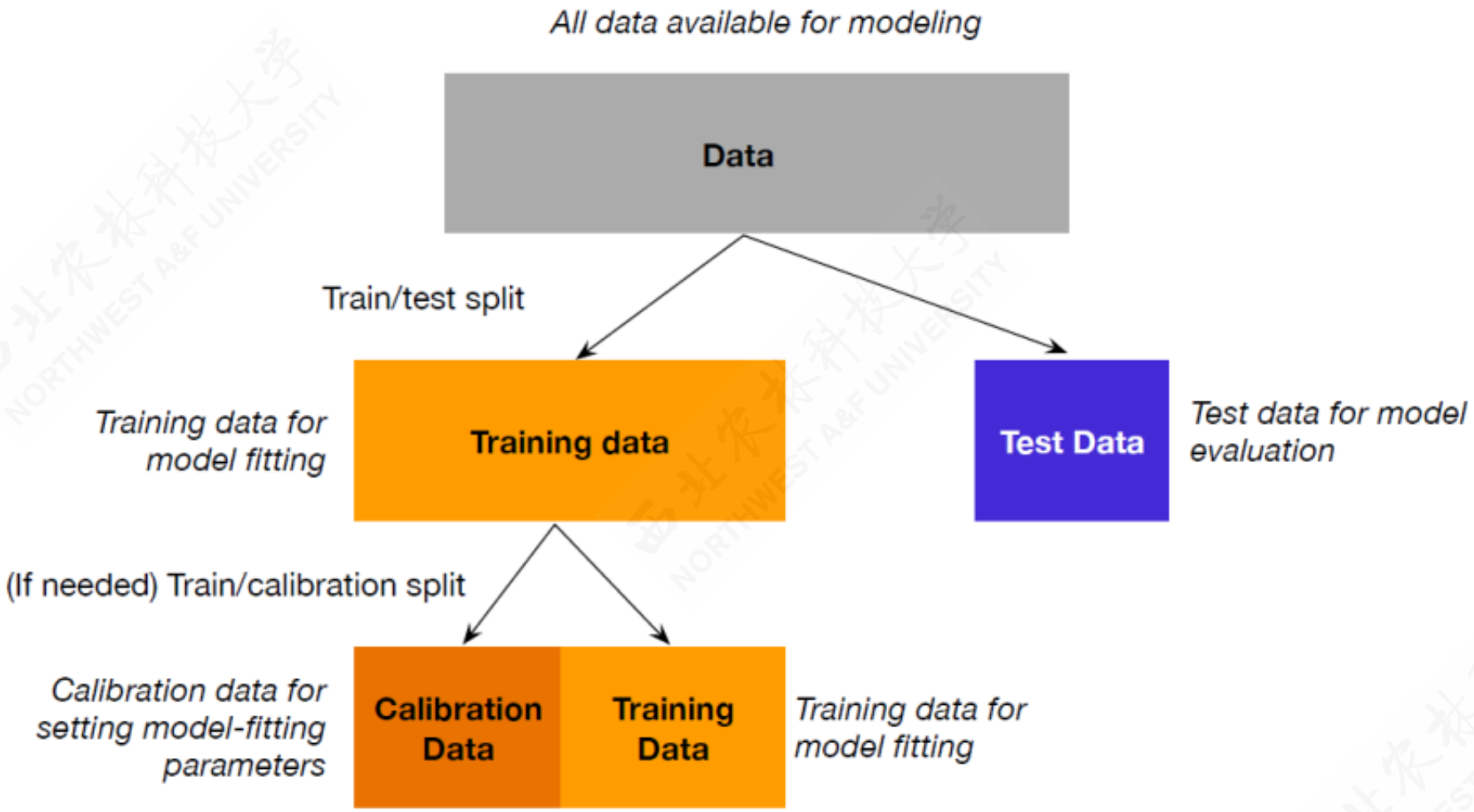
Previous 1 2 3 4 5 … 100 Next

数据子集：概念

数据子集（data subset）是对数据集进行抽样（sampling）的过程，形成的数据子集一般用于后续的建模分析。

- 训练集（training subset）：主要用于参数估计，得到一个最终估计模型。
- 测试集（test subset）：主要用于对估计模型的预测准确性进行评估。

数据子集：训练集和测试集



（税收案例）数据子集：设置子集样本数量

a.子集样本数　b.编写R代码

对全部数据集`custom_data`（样本数n=73262），我们可以生成名为`gp`的新列，采用[0,1]的**均匀随机分布**（uniform distribution），对应的R函数为`runif(n)`

假定我们希望数据子集的容量分别为：

- **训练集**样本数 n_{train} = 90% X 73262 =65421个。
- **测试集**样本数 n_{test} = 10% X 73262 =7841个。

（税收案例）数据子集：设置子集样本数量

a.子集样本数　b.编写R代码

```
set.seed(25643)
customer_data <- customer_data %>%
  mutate(gp =round(runif(nrow(.)),2))
pct <- 0.1

customer_test <- subset(customer_data, gp <= pct)
customer_train <- subset(customer_data, gp > pct)

n_all <- nrow(customer_data)
n_test <- nrow(customer_test)

n_train <- nrow(customer_train)
```

（税收案例）数据子集：比较三个数据集

a.全部数据集　b.训练数据集　c.测试数据集

全部数据样本量n=73262

id	gp	age	income	num_vehicles	gas_usage
1	0.13	24	22000	0	210
2	0.64	82	23200	0	3
3	0.92	31	21000	2	40
4	0.59	93	37770	1	120
5	0.11	67	39000	2	3

Showing 1 to 5 of 500 entries

Previous 1 2 3 4 5 … 100 Next

（税收案例）数据子集：比较三个数据集

a.全部数据集　b.训练数据集　c.测试数据集

训练集-样本量n=65421

id	gp	age	income	num_vehicles	gas_usage
1	0.13	24	22000	0	210
2	0.64	82	23200	0	3
3	0.92	31	21000	2	40
4	0.59	93	37770	1	120
5	0.11	67	39000	2	3

Showing 1 to 5 of 500 entries

Previous 1 2 3 4 5 … 100 Next

（税收案例）数据子集：比较三个数据集

a.全部数据集　b.训练数据集　c.测试数据集

测试集-样本量n=7841

id	gp	age	income	num_vehicles	gas_usage
8	0.05	73	34600	2	50
10	0.03	54	31200	3	20
28	0.03	62	19100	1	210
49	0.04	40	45000	2	3
56	0.03	79	12000	3	3

Showing 1 to 5 of 500 entries　Previous 1 2 3 4 5 … 100 Next

（税收案例）数据子集：等比例随机抽取

a.抽取规则　b.抽取结果　c.按州抽取前　d.按州抽取后

下面我们随机抽取数据集的10%作为子集，其中要求按所在州（state_of_res）来等比例分配：

```
vars_sel <- c("id","state_of_res","age", "sex","income")

spl_state <- customer_data %>%
  select(one_of(vars_sel)) %>%
  group_by(state_of_res) %>%
  sample_frac(0.1) %>%
  arrange(id)
```

（税收案例）数据子集：等比例随机抽取

a.抽取规则　b.抽取结果　c.按州抽取前　d.按州抽取后

数据样本量n=7325

id	state_of_res	age	sex	income
5	Alabama	67	Male	39000
12	Alabama	64	Male	40000
13	Alabama	57	Female	0
17	Alabama	27	Female	20200
60	Alabama	22	Male	7800

Showing 1 to 5 of 7,325 entries

Previous 1 2 3 4 5 … 1465 Next

（税收案例）数据子集：等比例随机抽取

a.抽取规则　b.抽取结果　c.按州抽取前　d.按州抽取后

抽取前各州样本数量n=73262

index	state_of_res	n
1	Alabama	1047
2	Alaska	162
3	Arizona	1534
4	Arkansas	653
5	California	8962

Showing 1 to 5 of 51 entries

Previous 1 2 3 4 5 … 11 Next

（税收案例）数据子集：等比例随机抽取

a.抽取规则　b.抽取结果　c.按州抽取前　d.按州抽取后

抽取后各州样本数量n=7325

index	state_of_res	n
1	Alabama	105
2	Alaska	16
3	Arizona	153
4	Arkansas	65
5	California	896

Showing 1 to 5 of 51 entries

Previous 1 2 3 4 5 … 11 Next

3.2 品质数据的整理与展示

分类数据的整理与图示

顺序数据的整理与图示

分类数据：整理与图示的基本过程

- 列出各类别
- 分类数据的整理与图示
- 制作频数分布表
- 用图形显示数据

分类数据：计算统计量并制表

- 频数(frequency)：落在各类别中的数据个数
- 比例(proportion)：某一类别数据个数占全部数据个数的比值
- 百分比(percentage)：将对比的基数作为100而计算的比值
- 比率(ratio)：不同类别数值个数的比值

（案例）饮料销售

a.案例说明　b.原始数据

案例：一家市场调查公司为研究不同品牌饮料的市场占有率，对随机抽取的一家超市进行了调查。

调查员在某天对50名顾客购买饮料的品牌进行了记录，如果一个顾客购买某一品牌的饮料，就将这一饮料的品牌名字记录一次。

（案例）饮科销售

a.案例说明　b.原始数据

下边就是记录的原始数据：

index	gender	brand
1	女	碳酸饮料
2	男	绿茶
3	男	矿泉水
4	女	矿泉水
5	男	碳酸饮料
6	男	矿泉水

Showing 1 to 6 of 50 entries　Previous　1　2　3　4　5　…　9　Next

（案例）饮料销售：单变量制表（饮科类别）

a.类别分组　b.频次表　c.频率表

根据原始数据，我们可以整理并列出所有饮料类别：

brand
果汁
矿泉水
绿茶
其他
碳酸饮料

（案例）饮料销售：单变量制表（饮料类别）

a.类别分组　b.频次表　c.频率表

统计得到各饮料类别的购买人数（频次）：

brand	n
果汁	6
矿泉水	10
绿茶	11
其他	8
碳酸饮料	15
Total	50

（案例）饮料销售：单变量制表（饮料类别）

a.类别分组　b.频次表　**c.频率表**

进一步统计得到各饮料类别的购买人数占比（频率）：

brand	n	percent
果汁	6	12.0%
矿泉水	10	20.0%
绿茶	11	22.0%
其他	8	16.0%
碳酸饮料	15	30.0%
Total	50	100.0%

（案例）饮料销售：双变量制表（饮料类别VS性别）

a.交叉分组　b.频次表　c1.列频率表　c2.行频率表　d.复合表

根据原始数据，我们可以对饮料类别和性别进行交叉分组：

brand	男	女
果汁		
矿泉水		
绿茶		
其他		
碳酸饮料		

（案例）饮料销售：双变量制表（饮料类别VS性别）

a.交叉分组　b.频次表　c1.列频率表　c2.行频率表　d.复合表

统计得到交叉分组下的购买人数（频次）：

brand	男	女	Total
果汁	1	5	6
矿泉水	6	4	10
绿茶	7	4	11
其他	2	6	8
碳酸饮料	6	9	15
Total	22	28	50

（案例）饮料销售：双变量制表（饮料类别VS性别）

a.交叉分组　b.频次表　**c1.列频率表**　c2.行频率表　d.复合表

进一步统计得到交叉分组下购买人数占比（频率）及其**列合计**：

brand	男	女	Total
果汁	4.5%	17.9%	12.0%
矿泉水	27.3%	14.3%	20.0%
绿茶	31.8%	14.3%	22.0%
其他	9.1%	21.4%	16.0%
碳酸饮料	27.3%	32.1%	30.0%
Total	100.0%	100.0%	100.0%

（案例）饮料销售：双变量制表（饮料类别VS性别）

a.交叉分组　b.频次表　c1.列频率表　**c2.行频率表**　d.复合表

同时，也可统计得到交叉分组下购买人数占比（频率）及其行合计：

brand	男	女	Total
果汁	16.7%	83.3%	100.0%
矿泉水	60.0%	40.0%	100.0%
绿茶	63.6%	36.4%	100.0%
其他	25.0%	75.0%	100.0%
碳酸饮料	40.0%	60.0%	100.0%
Total	44.0%	56.0%	100.0%

（案例）饮料销售：双变量制表（饮料类别VS性别）

a.交叉分组　b.频次表　c1.列频率表　c2.行频率表　**d.复合表**

最后，还可以同时统计交叉分组的人数和占比（列合计）：

brand	男	女	Total
果汁	4.5% (1)	17.9% (5)	12.0% (6)
矿泉水	27.3% (6)	14.3% (4)	20.0% (10)
绿茶	31.8% (7)	14.3% (4)	22.0% (11)
其他	9.1% (2)	21.4% (6)	16.0% (8)
碳酸饮料	27.3% (6)	32.1% (9)	30.0% (15)
Total	100.0% (22)	100.0% (28)	100.0% (50)

分类数据：统计制图1（条形/柱状图）

条形/柱状图：用宽度相同的条形的高度或长短来表示各类别数据的图形。

- 各类别可以放在纵轴，称为**条形图**(bar Chart)
- 各类别也可以放在横轴，称为**柱形图**(column chart)

作用：主要用于反映分类数据的频数分布。

形式：单式条形图/复式条形图。

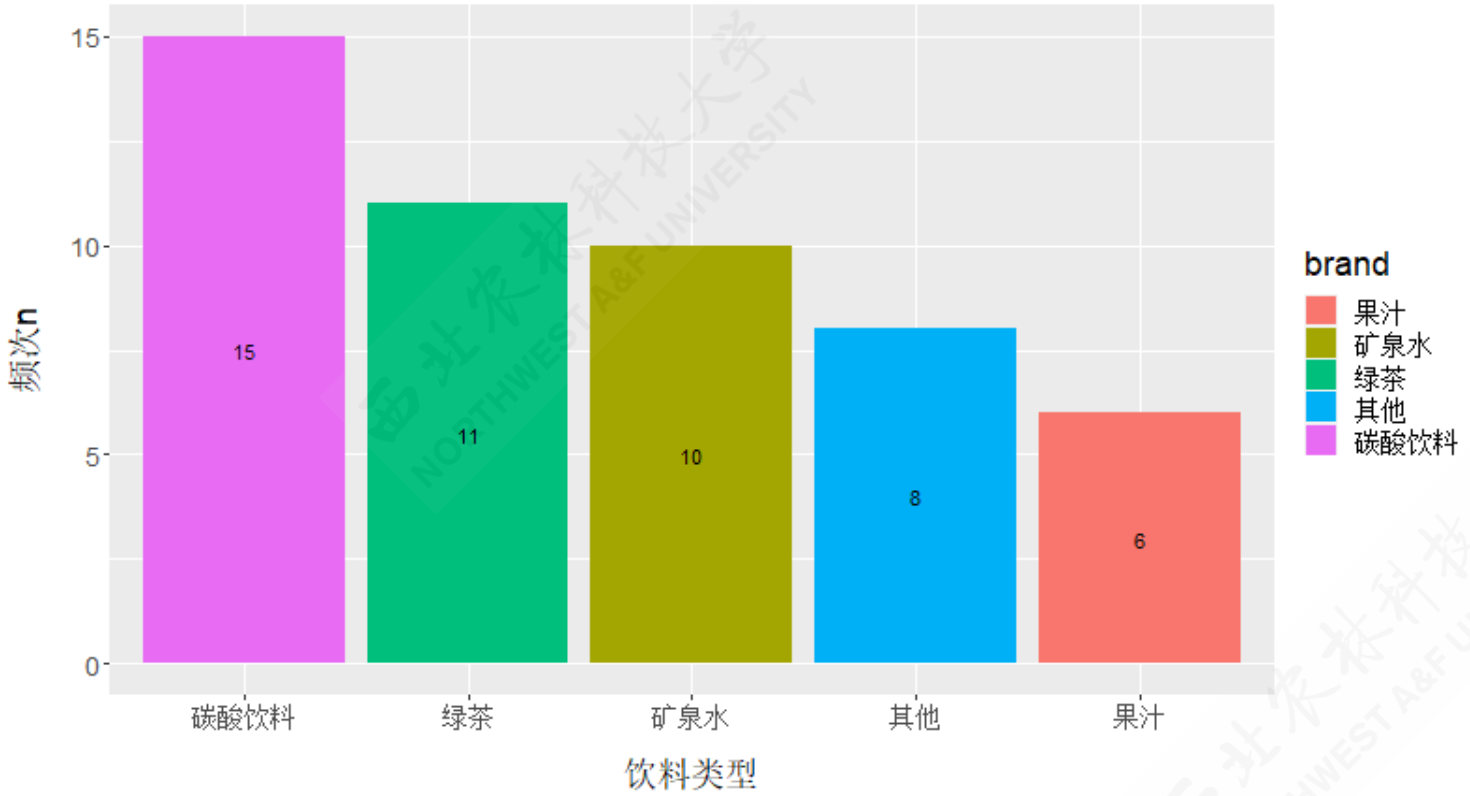
复式条形图主要用于：

- 分类变量在不同时间或不同空间上有多个取值
- 对比分类变量的取值在不同时间或不同空间上的差异或变化趋势

（案例）饮科销售：单变量柱状图/条形图

a.柱状图

根据饮料类型购买次数的数据表，我们可以绘制出如下柱状图：

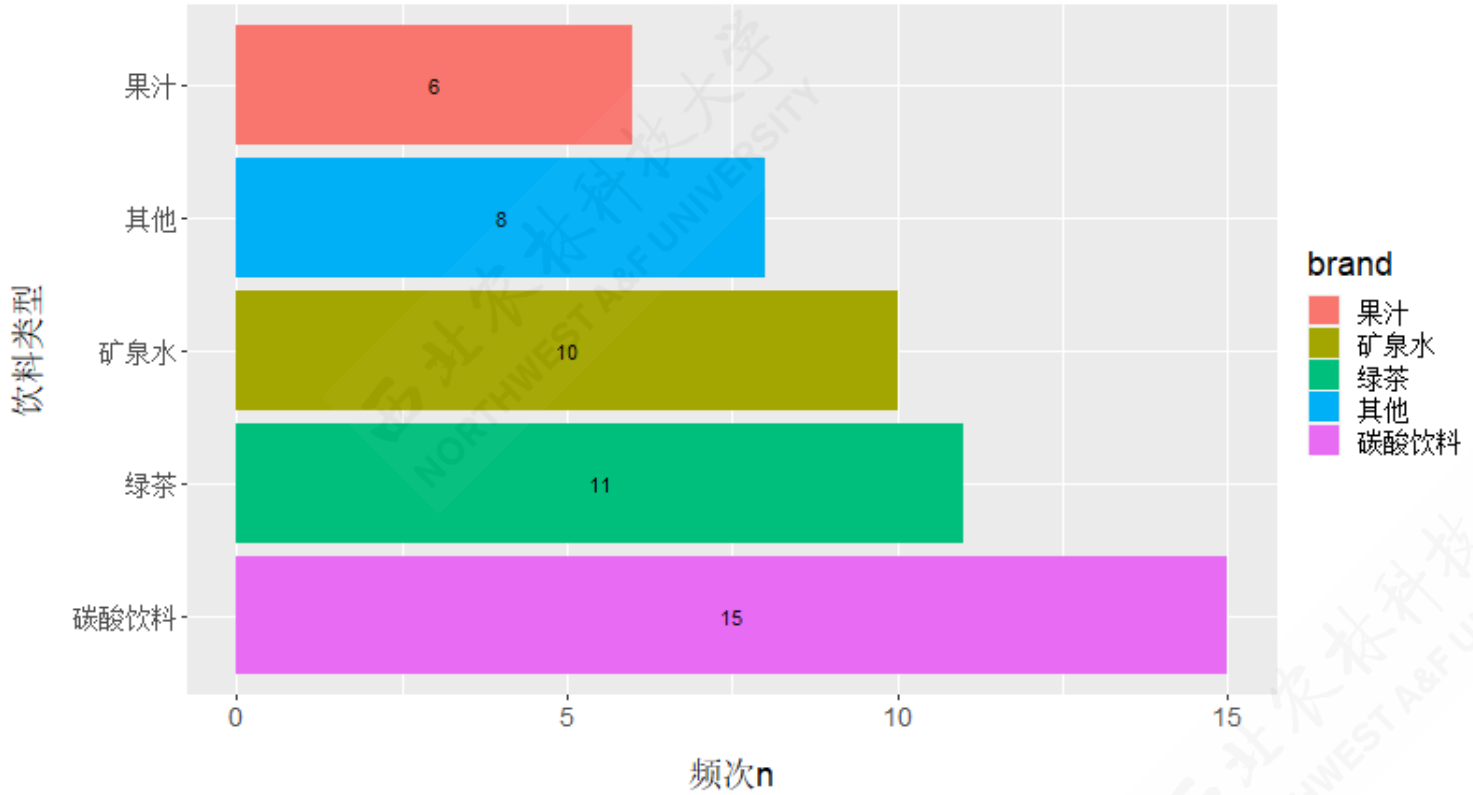


（案例）饮科销售：单变量柱状图/条形图

a.柱状图

b.条形图

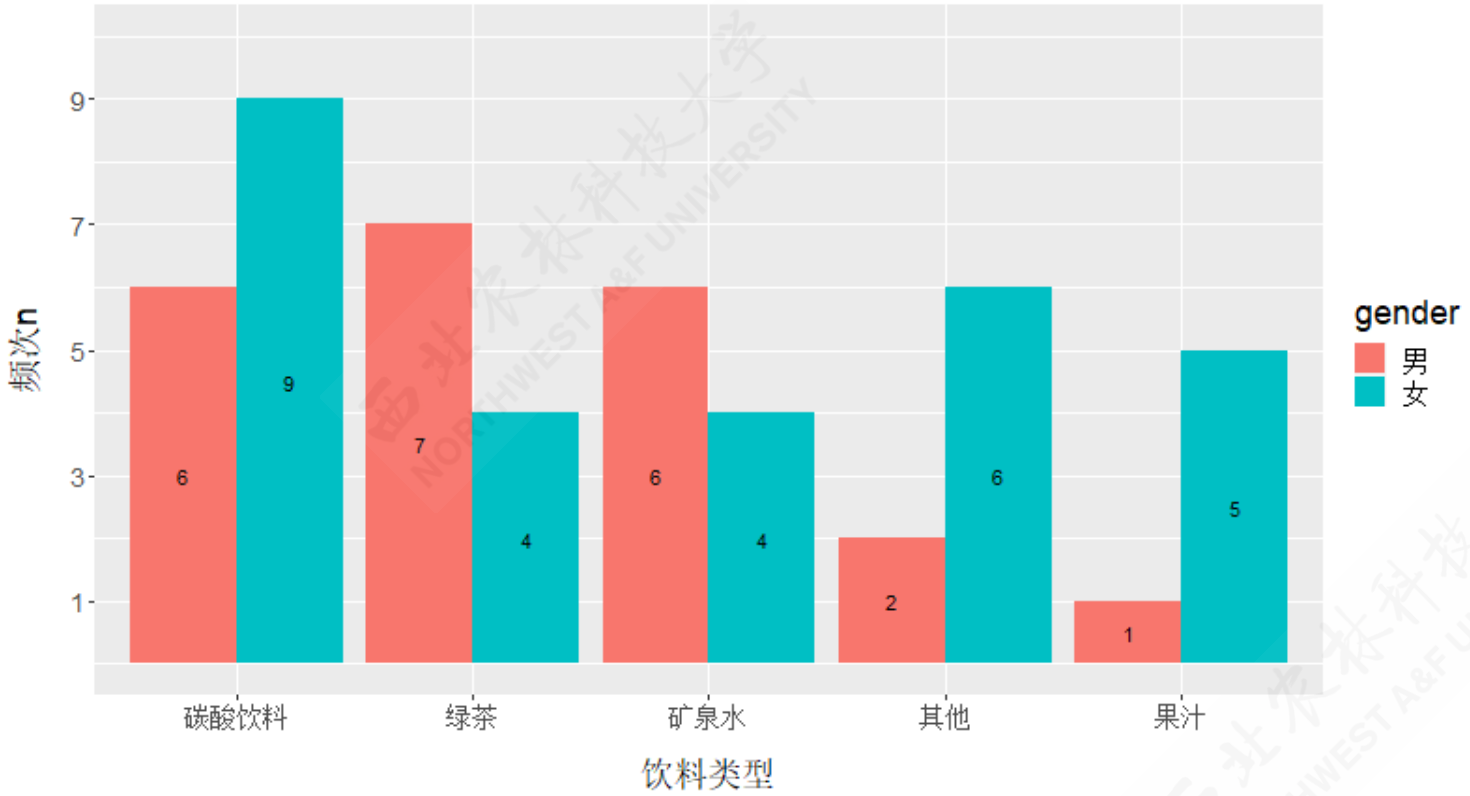
根据饮料类型购买次数的数据表，还可以绘制出如下**条形图**：



（案例）饮料销售：多变量柱状图/条形图

a.双变量柱状图

根据饮料类型（brand）和性别（gender）交叉分组下的购买次数表，可以绘制出如下柱状图：

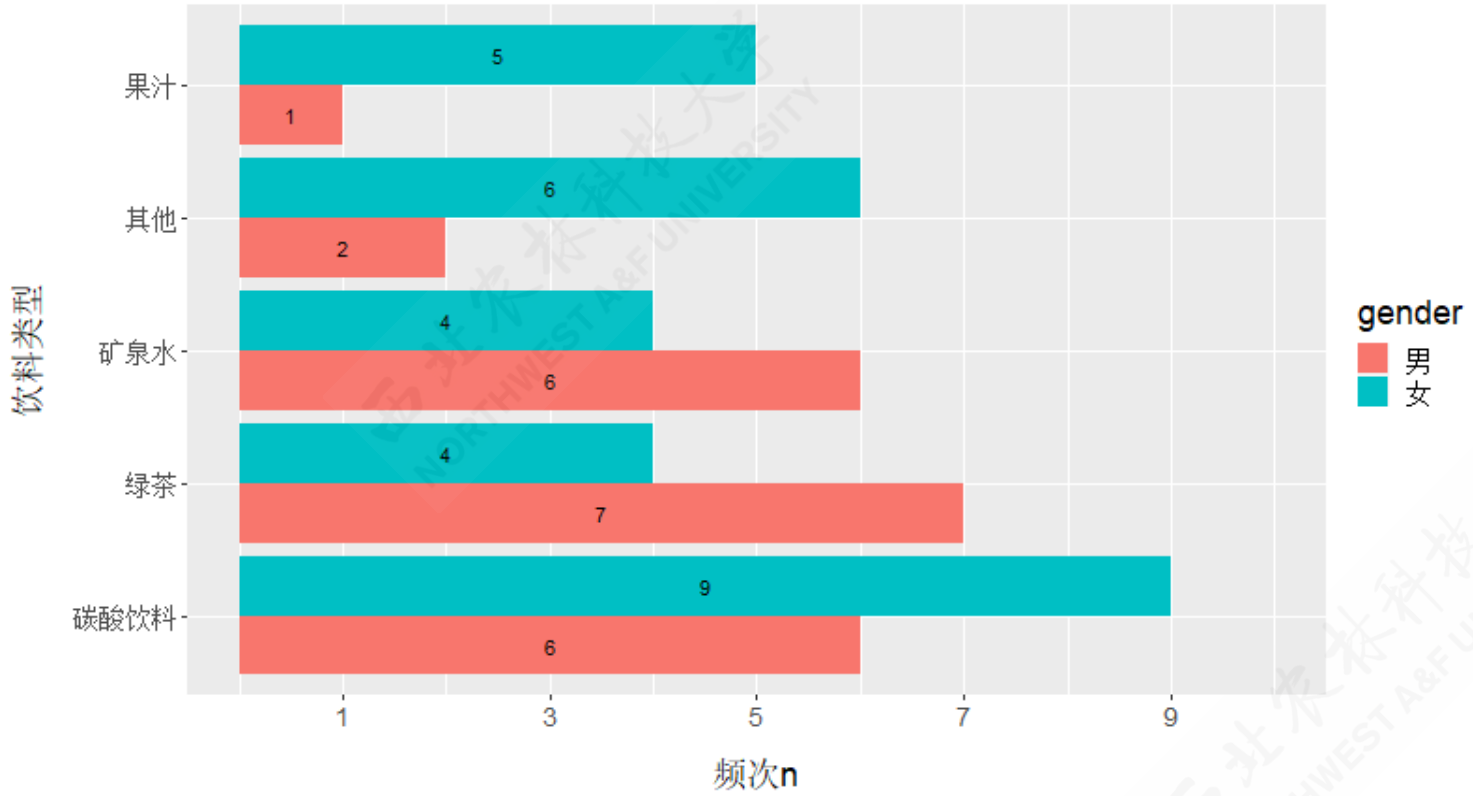


（案例）饮料销售：多变量柱状图/条形图

a.双变量柱状图

b.双变量条形图

根据饮料类型（brand）和性别（gender）交叉分组下的购买次数表，可以绘制出如下条形图：



分类数据：统计制图1（条形/柱状图）

思考：什么时候适合使用柱状图？什么时候适合使用条形图？

待完成：找到新数据，绘制一张条形图，但其不适合制作柱状图。

分类数据：统计制图2（饼图）

饼图（pie Chart）：也称圆形图，是用圆形及圆内扇形的角度来表示数值大小的图形。

用途：用于表示样本或总体中各组成部分所占的比例，用于研究结构性问题。

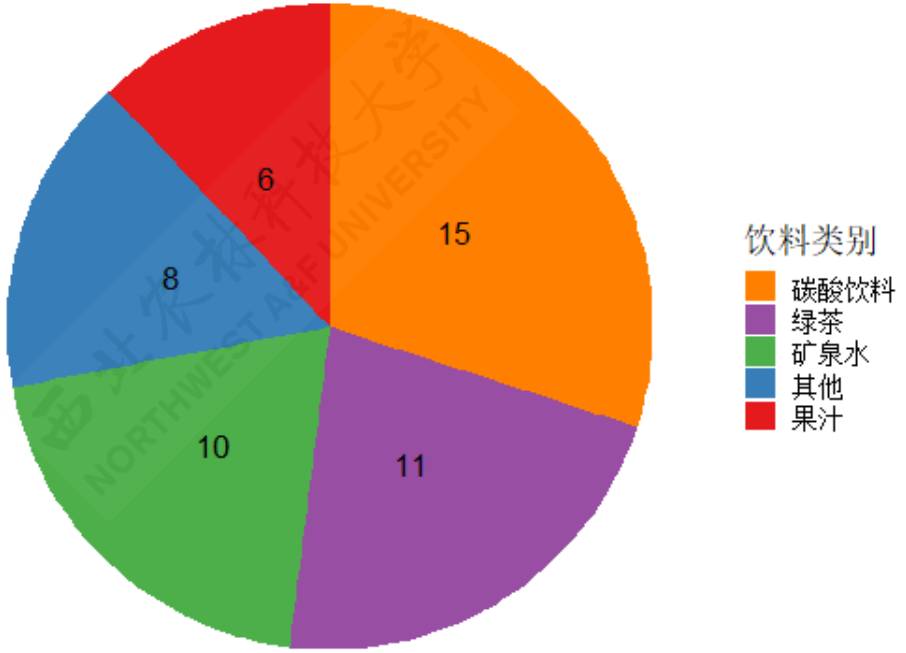
绘制要点：

- 样本或总体中各部分所占的频次/百分比用圆内的各个扇形角度表示。
- 扇形块的中心角度，进行**极坐标变换**（按各部分数据占比乘以3600确定）。
- 排列顺序、标签值显示。

（案例）饮科销售：绘制饼图（频次1）

a.次数饼图1

根据分配数据表，我们可以绘制出如下次数饼图：



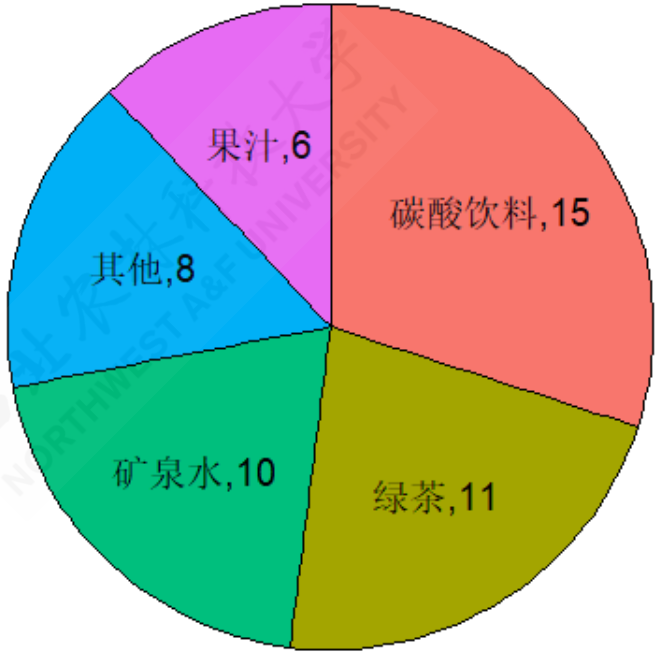
图a.饮科销售量分布（有图例）

（案例）饮科销售：绘制饼图（频次2）

a.次数饼图1

b.次数饼图2

可以进一步调整次数饼图的图例和标签数值显示：



图b.饮科销售量分布（无图例）

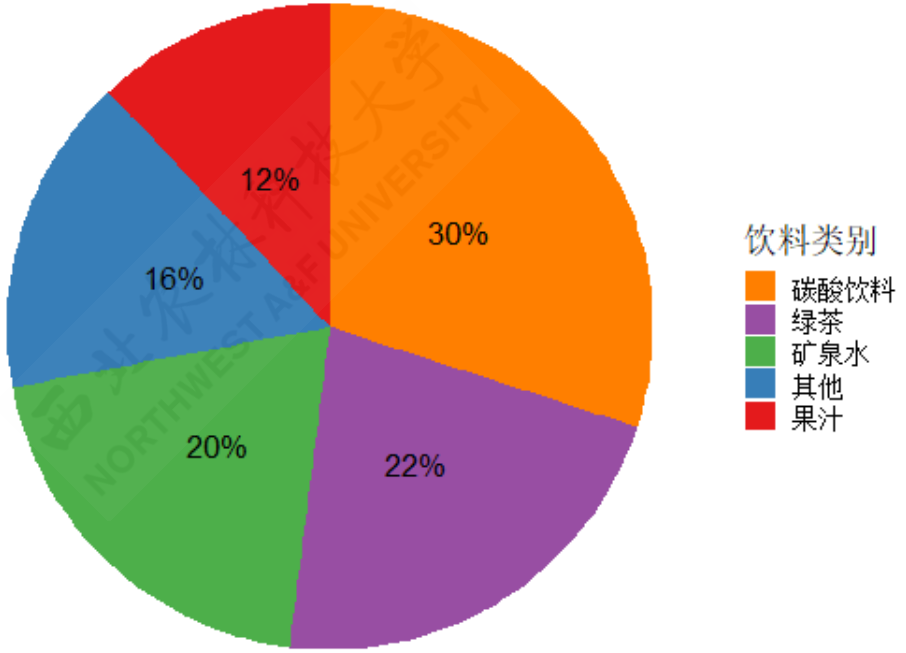
（案例）饮料销售：绘制饼图（占比）

a.次数饼图1

b.次数饼图2

c.占比饼图1

可以进一步调整**占比**饼图的图例和标签数值显示：



图c.饮料销售占比（有图例）

（案例）饮科销售：绘制饼图（占比2）

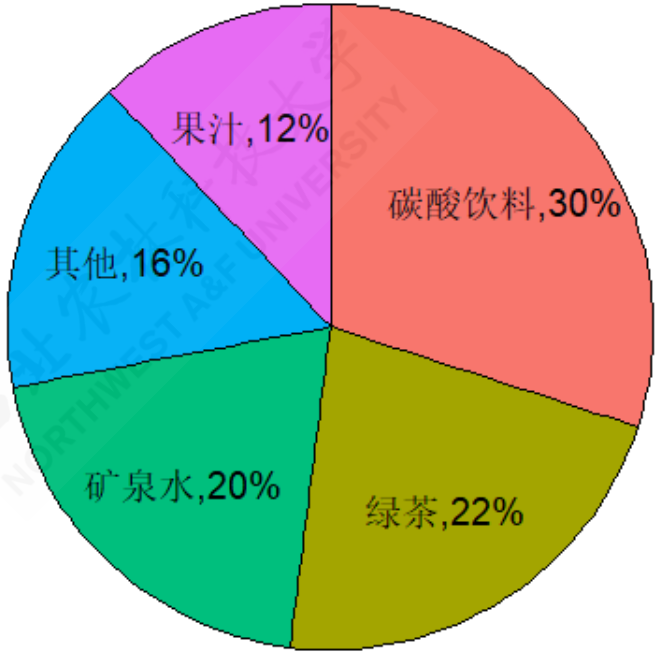
a.次数饼图1

b.次数饼图2

c.占比饼图1

d.占比饼图2

根据分配数据表，我们可以绘制出如下次数饼图：



图d饮科销售占比（无图例）

顺序数据：统计量和图表

对于顺序型分组（上小下大）数据，主要的统计指标包括：

- **累积频数**（cumulative frequencies）：各类别频数的逐级累加。
- **累积频率**（cumulative percentages）：各类别频率(百分比)的逐级累加。

对于以上累积统计指标，又分别包括：

- **较小制累积**（频数/频率）：又称为向上累积或以下累积，本组及以下次数/频率的逐级累加。
- **较大制累积**（频数/频率）：又称为向下累积或以上累积，本组及以上次数/频率的逐级累加。

制表和绘图分别有：

- 累积频数/频率表等
- 累积频数/频率图、环形图等

（案例）住房满意度：案例数据

1)案例说明　2)案例数据表

案例说明：在一项城市住房满意度问题的研究中，研究人员在甲城市抽样调查300家庭户，其中的一个问题是：

> 您对您家庭目前的住房状况是否满意?
>
> 1.非常不满意；2.不满意；3.一般；4.满意；5.非常满意

（案例）住房满意度：案例数据

1)案例说明　2)案例数据表

下边就是收集到的不同满意度评价水平的频次和频率数据表：

groups	satisfication	n	percent
A	非常不满意	24	8.0%
B	不满意	108	36.0%
C	一般	93	31.0%
D	满意	45	15.0%
E	非常满意	30	10.0%
-	Total	300	100.0%

（案例）住房满意度：计算统计量并制表

1)较小制累积表　2)较大制累积表　3)较小制和较大制对比

我们可以计算得到**较小制**下的累积频次和频率，并制表：

groups	satisfication	n	percent	min_cum_n	min_cum_p
A	非常不满意	24	8.0%	24	8.0%
B	不满意	108	36.0%	132	44.0%
C	一般	93	31.0%	225	75.0%
D	满意	45	15.0%	270	90.0%
E	非常满意	30	10.0%	300	100.0%
-	Total	300	100.0%	-	-

（案例）住房满意度：计算统计量并制表

1)较小制累积表　2)较大制累积表　3)较小制和较大制对比

我们也可以计算得到**较大制**下的累积频次和频率，并制表：

groups	satisfication	n	percent	max_cum_n	max_cum_p
A	非常不满意	24	8.0%	300	100.0%
B	不满意	108	36.0%	276	92.0%
C	一般	93	31.0%	168	56.0%
D	满意	45	15.0%	75	25.0%
E	非常满意	30	10.0%	30	10.0%
-	Total	300	100.0%	-	-

（案例）住房满意度：计算统计量并制表

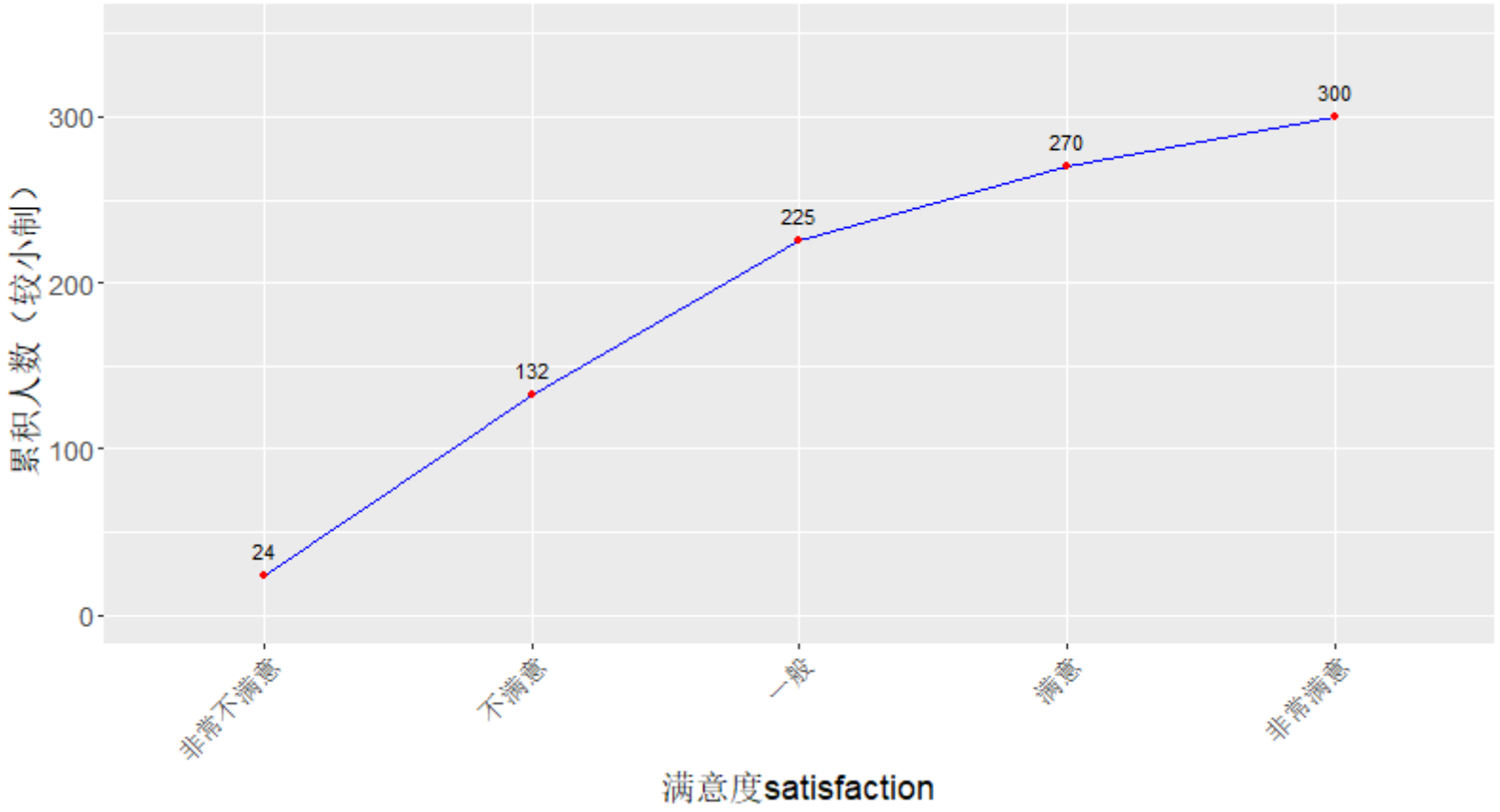
1)较小制累积表　2)较大制累积表　3)较小制和较大制对比

我们可以对比观测**较小制**和**较大制**下的累积频次和频率：

groups	satisfication	n	percent	min_cum_n	min_cum_p	max_cum_n	m
A	非常不满意	24	8.0%	24	8.0%	300	
B	不满意	108	36.0%	132	44.0%	276	
C	一般	93	31.0%	225	75.0%	168	
D	满意	45	15.0%	270	90.0%	75	
E	非常满意	30	10.0%	300	100.0%	30	
-	Total	300	100.0%	-	-	-	

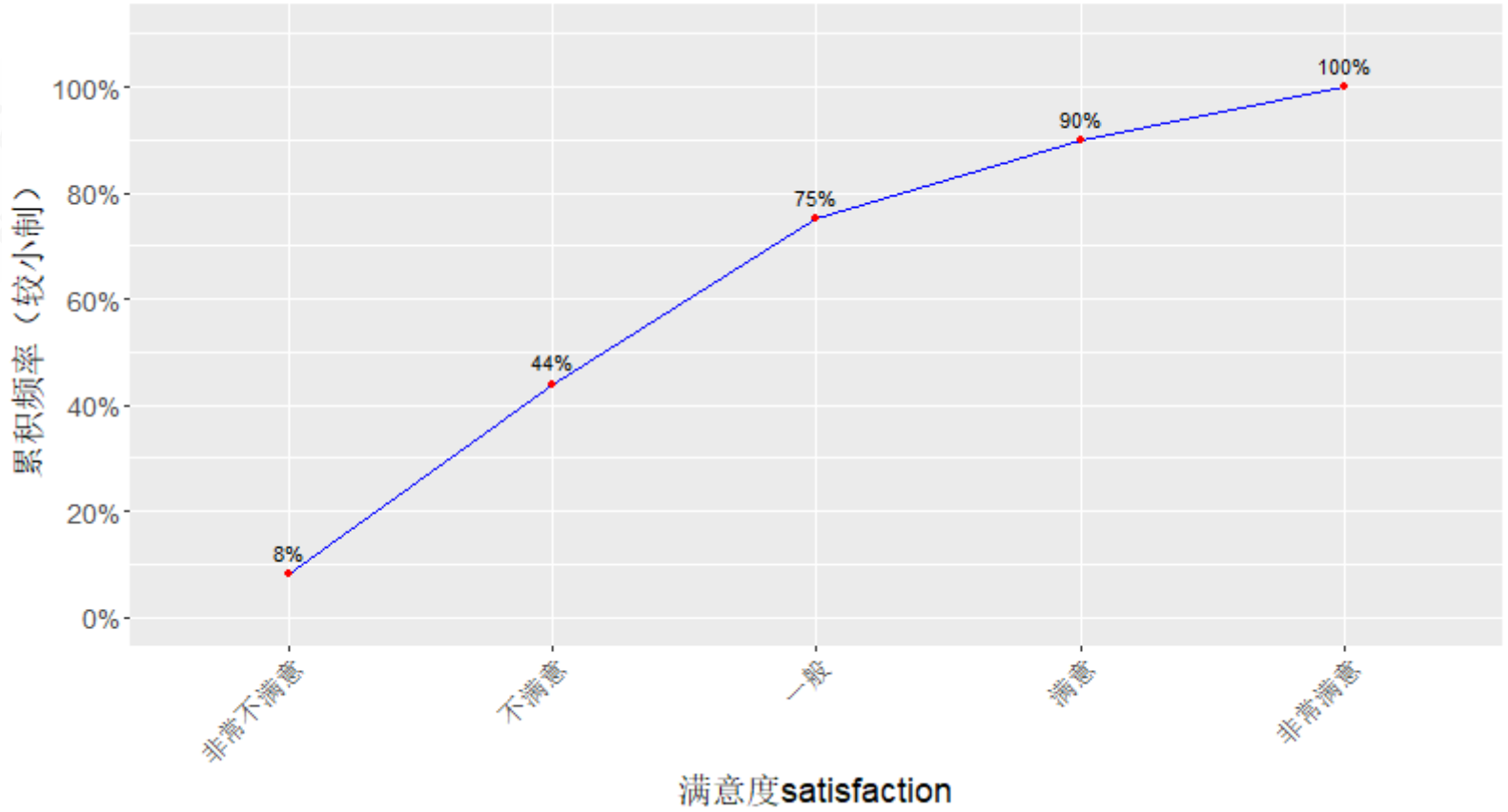
（案例）住房满意度：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　3)较大累积频次　4)较大累积频率



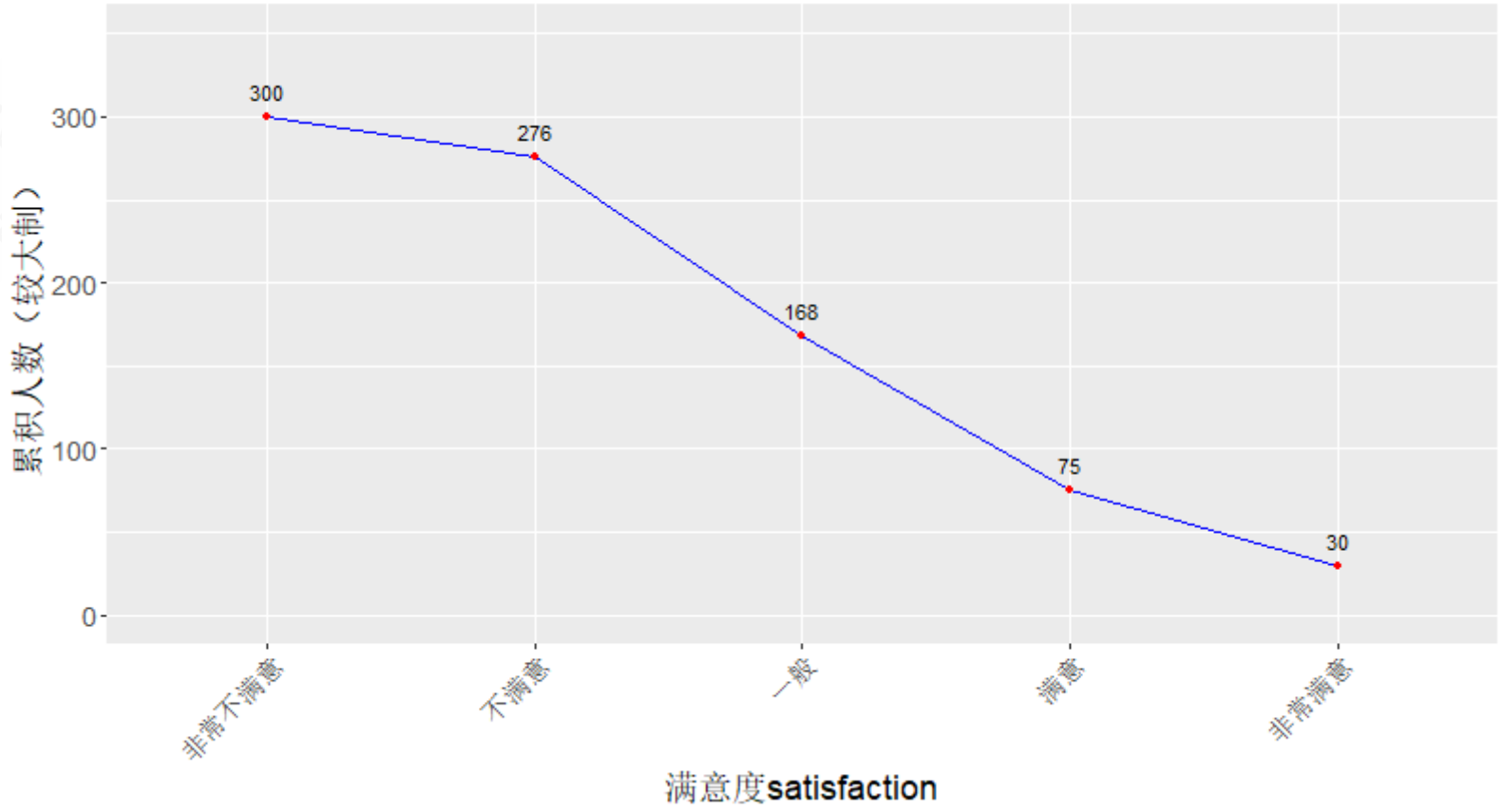
（案例）住房满意度：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　3)较大累积频次　4)较大累积频率



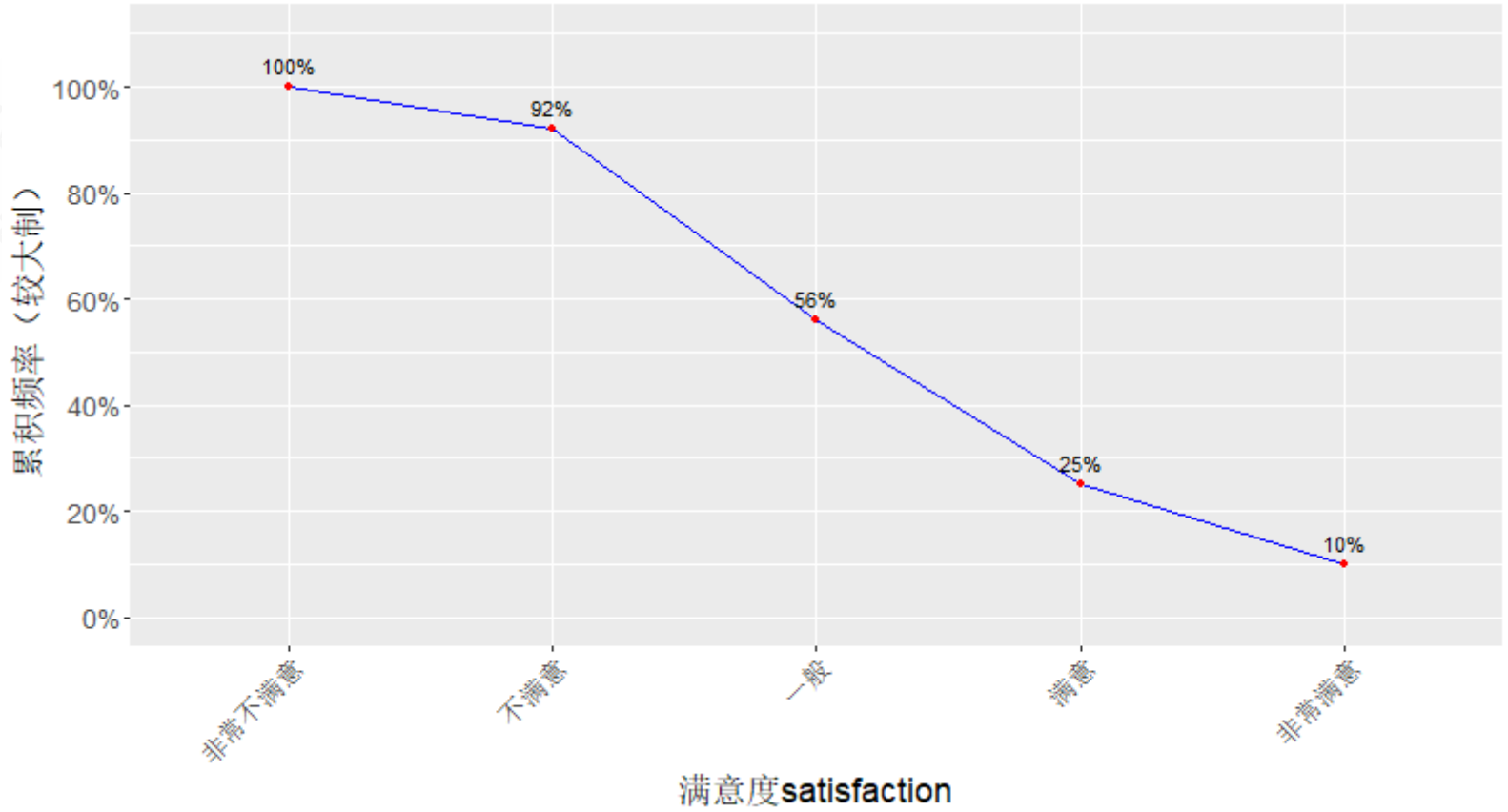
（案例）住房满意度：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　**3)较大累积频次**　4)较大累积频率



（案例）住房满意度：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　3)较大累积频次　**4)较大累积频率**



（案例）住房满意度：绘制环形图

环形图(doughnut chart)：环形图中间有一个“空洞”，样本或总体中的每一部分数据用环中的一段表示。

用途：用于结构比较研究；用于展示分类和顺序数据。

与饼图的联系与区别：

- 饼图只能显示一个总体各部分所占的比例。
- 环形图则可以同时绘制多个样本或总体的数据系列，每一个样本或总体的数据系列为一个环。

（案例）住房满意度：绘制环形图

1)补充数据　2)甲城市　3)乙城市　4)两个城市

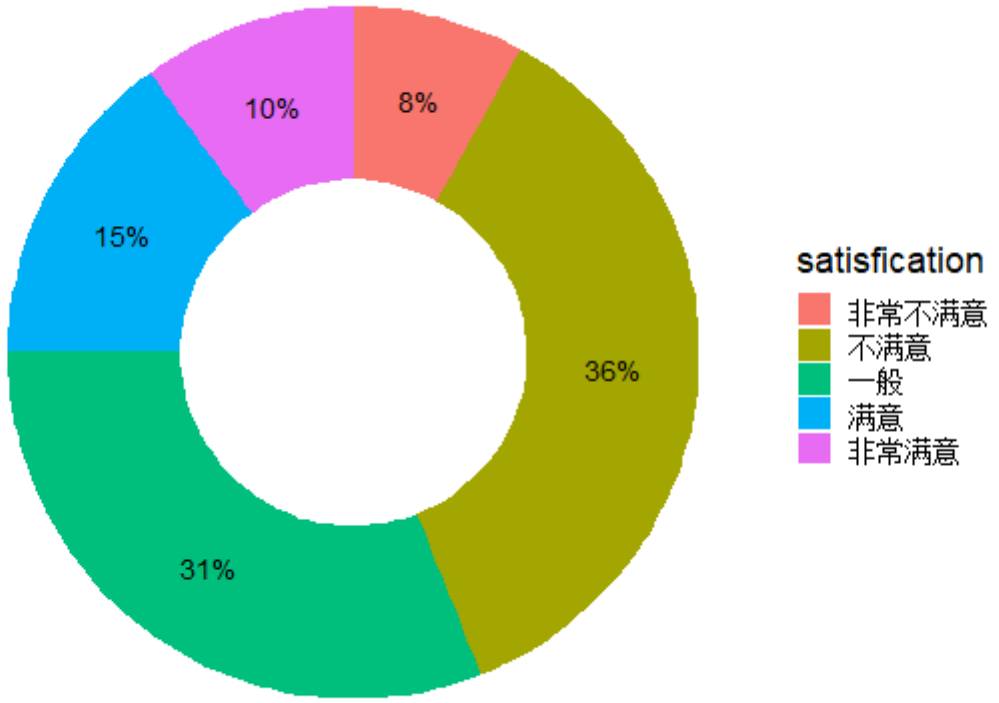
案例说明：继续前述甲城市满意度的研究，为了综合比较城市家庭满意度。研究者继续收集并获得了乙城市家庭的满意度评价数据。甲乙两个城市的家庭住户满意度数据如下表所示：

city	satisfication	n	percent
甲城市	非常不满意	24	8.0%
甲城市	不满意	108	36.0%
甲城市	一般	93	31.0%
甲城市	满意	45	15.0%
甲城市	非常满意	30	10.0%

Showing 1 to 05 of 10 entries　Previous　1　2　Next

（案例）住房满意度：绘制环形图

1)补充数据　2)甲城市　3)乙城市　4)两个城市



图a.甲城市评价分布

（案例）住房满意度：绘制环形图

1)补充数据　2)甲城市　3)乙城市　4)两个城市

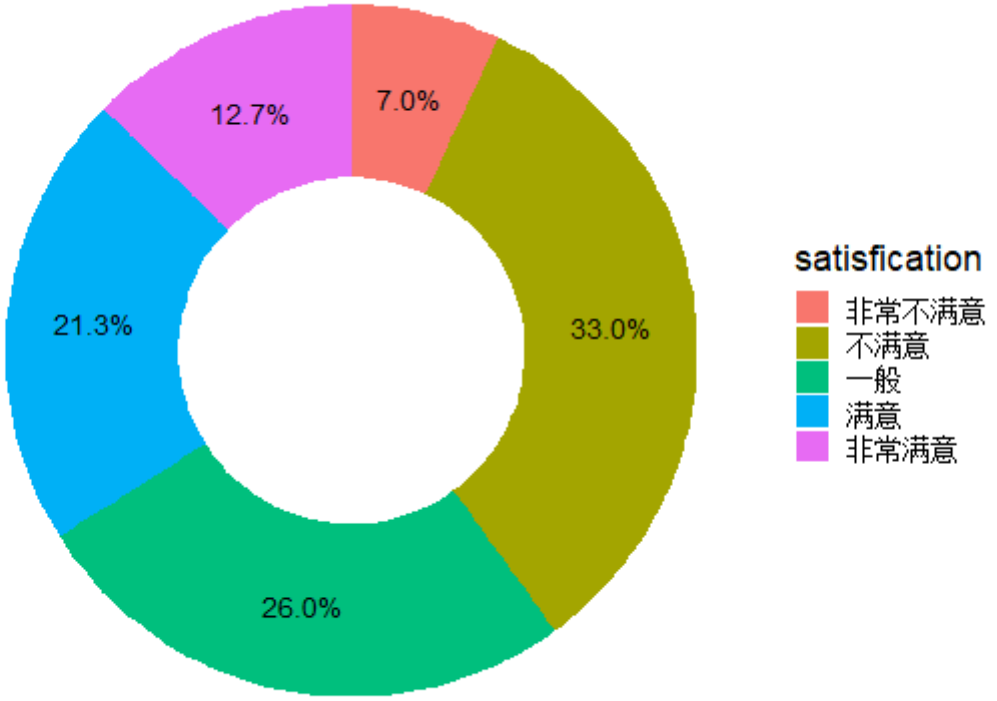
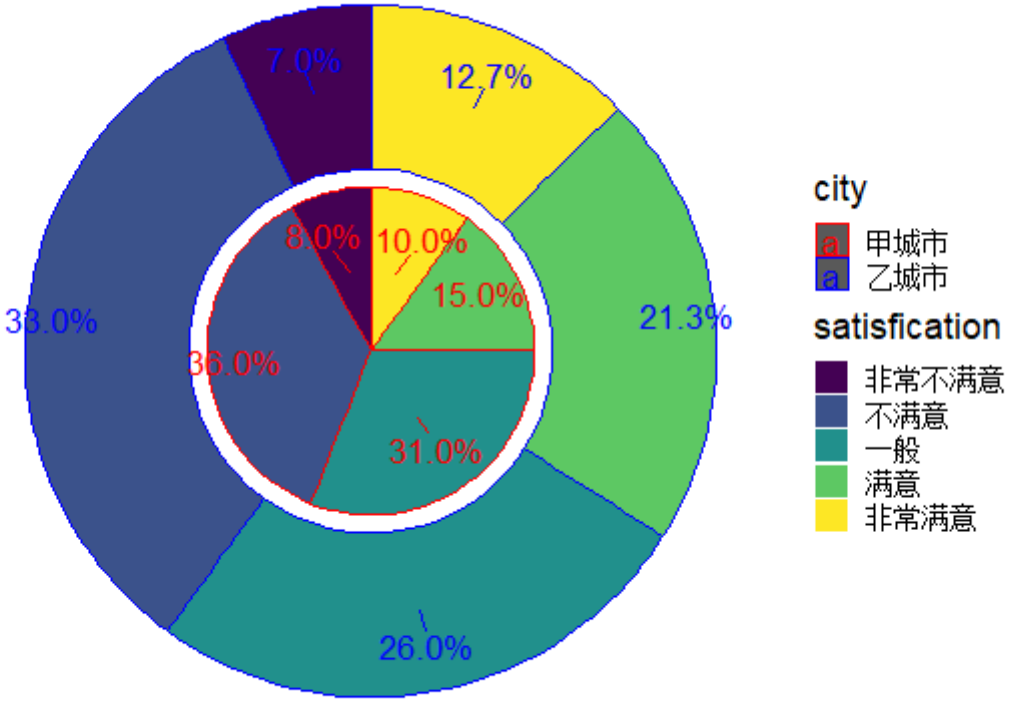


图6.乙城市评价分布

（案例）住房满意度：绘制环形图

1)补充数据　2)甲城市　3)乙城市　4)两个城市



图c.两个城市评价分布对比

3.3 数值型数据的整理与展示

数据分组的图表展示

数据未分组的图表展示

统计报表的设计

数据分组：概念和作用

数据分组：把同质总体中的具有不同特点的单位分开，从而正确地认识事物的本质及其规律性。

数据分组的作用：

- 类型分组：揭露社会经济现象的类型，反映各类型的特点。
- 结构分组：说明社会经济现象的内部结构。
- 分析分组：研究经济现象之间的依存关系。

数据分组：选择分组标志

选择分组标志的**原则**：

- 科学性
- 完备性
- 互斥性

选择分组标志的**方法**：

- 根据研究问题的目的来选择。
- 要选择最能反映被研究现象本质特征的标志。
- 要结合现象所处的具体历史条件或经济条件来选择。

数据分组：数据分组类型

A.按分组标志的特征分：

- 品质标志分组：反映事物属性差异
  - 简单分组：如人口按性别分组。
  - 复杂分组：如人口按职业分组。
- 数量标志分组：反映事物数量差异
  - 单项式数量分组：运用于变量变动幅度小、项目少的分组。
  - 组距式分组：运用于变量变动幅度大、项目多的分组。

B.按总体所选择标志的个数分：

- 单一分组：按一个标志对总体进行分组。
- 复合分组：按两个或两个以上标志对同一总体进行分组。

（示例）数据分组类型：按标志特征分组1

1)品质-简单分组　2)品质-复杂分组

性别	人数
男	
女	

（示例）数据分组类型：按标志特征分组I

1)品质-简单分组　2)品质-复杂分组

职业	人数
急诊科医师	
乡村医师	
急诊护士	
手术室护士	
护理员	
放射线之技术人员	

Showing 1 to 6 of 12 entries

Previous 1 2 Next

（示例）数据分组类型：按标志特征分组2

3)数量-单项式分组　4)数量-组距分组

日产量	职工人数
12	
13	
15	
16	
20	

（示例）数据分组类型：按标志特征分组2

3)数量-单项式分组　4)数量-组距分组

年龄	职工人数
18岁以下	
19-25岁	
26-35岁	
36-45岁	
45岁以上	

（示例）数据分组类型：按标志个数分组

1)单一分组　2)复合分组

性别	人数
男	
女	

--

日产量	职工人数
12	
13	
15	
16	
20	

--

年龄	职工人数
18岁以下	
19-25岁	
26-35岁	
36-45岁	
45岁以上	

（示例）数据分组类型：按标志个数分组

1)单一分组　2)复合分组

学科	学历	性别	人数
理科生	本科生	男生	
理科生	专科生	男生	
理科生	本科生	女生	
理科生	专科生	女生	
文科生	本科生	男生	
文科生	专科生	男生	
文科生	本科生	女生	
文科生	专科生	女生	

分配数列：概念和类型

分配数列：统计总体按照某一标志分组以后，用以反映总体各单位分配情况的统计数列，称为分配数列，又可称次数分配或次数分布。

- **品质型分配数列**。
- **数值型分配数列**。根据分组情况，具体又分为：
  - **单项变量数列**：按每个变量值分别列组编制数列。适用于不连续变量或变量能以整数表示，其变动范围不大时。
  - **组距变量数列**：按组距分组编制数列。适用于连续变量或变量可用小数表示，其变动范围。

（示例）分配数列的类型：数值型分配数列

1)单项式分配数列　2)组距式分配数列

日产量	职工人数
12	109
13	136
15	144
16	107
20	124

（示例）分配数列的类型：数值型分配数列

1)单项式分配数列　2)组距式分配数列

年龄	职工人数
18岁以下	1117
19-25岁	903
26-35岁	1113
36-45岁	883
45岁以上	860

组距式数据分组：类型

根据分组是否开口以及是否等距，组距式分配数列可以分为如下类型：

- 按两端组是否开口分：
  - **开口式分组**：最小组与最大组不封口。
  - **闭口式分组**：所有组都有明确上限和下限。
- 按组距是否为等距分：
  - **等距式分组**：所有分组的组距全部相等。
  - **异距式分组**：各个分组组距不是完全相等。

（示例）组距式分配数列的类型

1)闭口-等距　2)开口-异矩

成绩	人数
50-60	2
60-70	7
70-80	11
80-90	12
90-100	8

（示例）组距式分配数列的类型

1)闭口-等距　2)开口-异矩

年龄	职工人数
18岁以下	1117
19-25岁	903
26-35岁	1113
36-45岁	883
45岁以上	860

组距式数据分组：重要概念

关于组距式分配数列，我们需要掌握如下重要概念：

- **组数（bins）**：数据分组的总组数。
- **组限（limits）**：组距两端的数值。分为上限和下限。

  a. **下限(lower limit)**：一个组的最小值。

  b. **上限(upper limit)**：一个组的最大值。
- **组距（width）**：某组的上限与下限之差。
- **组中值(class midpoint)**：某组的下限与上限之间的中点值
- **全距（range）**：整个分组数列中，最大组**上限**与最小组**下限**之差。
- **最大组/最小组**：整个分组数列中，分组标志数值最大/最小的那一组。

组距式数据分组：分组步骤

组距式数据分组的主要步骤包括：

- 确定组数：组数的确定应以能够显示数据的分布特征和规律为目的。在实际分组时，组数一般为 $5 \leq K \leq 15$。
- 确定组距：组距是一个组的上限与下限之差，可根据全部数据的最大值和最小值及所分的组数来确定。例如，$\text{组距} = \frac{(\text{最大值} - \text{最小值})}{\text{组数}}$
- 确定组限。对于连续变量分组，各组之间的组限也要连续。对于不连续变量分组，组与组之间的组限往往是间断的。此外，在同一个组距数列中，组限标准保持一致。
- 统计出各组的频数并整理成频数分布表。在登记次数时，应遵守：
  - 上组限不在内：适用于越大越好的变量，如产值。
  - 下组限不在内：适用于越小越好的变量，如成本。

组距式数据分组：组中值的计算

组中值的计算，需要考虑分组是否开口：

- 闭口式分组的组中值求法：

$$\begin{aligned}\text{组中值} &= \frac{\text{组上限}+\text{组下限}}{2} = \text{组下限} + \frac{\text{组上限}-\text{组下限}}{2} \\ &= \text{组上限} - \frac{\text{组上限}-\text{组下限}}{2}\end{aligned}$$

- 开口式分组的组中值求法：

$$\begin{aligned}\text{下开口组的组中值} &= \text{组上限} - \frac{\text{邻组组距}}{2} \\ \text{上开口组的组中值} &= \text{组下限} + \frac{\text{邻组组距}}{2}\end{aligned}$$

（案例）学生考试成绩：原始数据

案例说明：某班级共有40名学生，在《统计学原理》课程考试中成绩如下：

```
s01 s02 s03 s04 s05 s06 s07 s08 s09 s10 s11 s12 s13 s14 s15 s16 s17
 63  88  72  69  80  80  61  68  79  81  76  76  77  78  90  89  61
s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29 s30 s31 s32 s33 s34
 84  92  67  65  57  62  60  66  87  80  92  78  86  71  71  63  74
s35 s36 s37 s38 s39 s40
 67  70  91  64  79  65
```

（案例）学生考试成绩：组距式分组操作1

假定采用**等距-闭口**分组方法，且设定分组数量为5，组距为10。

利用原始数据，进行分组得到如下结果：

成绩分组	学生成绩
(50,60]	(s22)57; (s24)60
(60,70]	(s01)63; (s04)69; (s07)61; (s08)68; (s17)61; (s20)67; (s21)65; (s23)62; (s25)66; (s33)63; (s35)67; (s36)70; (s38)64; (s40)65
(70,80]	(s03)72; (s05)80; (s06)80; (s09)79; (s11)76; (s12)76; (s13)77; (s14)78; (s27)80; (s29)78; (s31)71; (s32)71; (s34)74; (s39)79
(80,90]	(s02)88; (s10)81; (s15)90; (s16)89; (s18)84; (s26)87; (s30)86
(90,100]	(s19)92; (s28)92; (s37)91

（案例）学生考试成绩：组距式分组操作2

以前述**考试成绩**案例分组数据为例，相关概念包括（见下表）：

- **组数（bins）**为5
- **组距（width）** 为10
- **全距（range）**为50（注意原始数据的全距）
- 最小组和最大组分别为第1组和第5组
- 下限、上限、组距和组中值分别见右表

分组数据表：

成绩分组	人数	下限	上限	组距	组中组
(50,60]	2	50	60	10	55
(60,70]	14	60	70	10	65
(70,80]	14	70	80	10	75
(80,90]	7	80	90	10	85
(90,100]	3	90	100	10	95

组距式数据分组：异矩情形

在**异矩分组**下，我们需要进一步计算各组的**组密度**以及**标准组距分布**：

$$\text{次数密度} = \frac{\text{各组次数}}{\text{各组组距}}$$

$$\text{频率密度} = \frac{\text{各组频率}}{\text{各组组距}}$$

$$\text{标准组距次数分布} = \frac{\text{各组次数}}{\text{标准化组距}}$$

$$\text{标准组距频率分布} = \frac{\text{各组频率}}{\text{标准化组距}}$$

（案例）学生考试成绩：异矩分组情形

继续前述**考试成绩**案例，如果分组时不小心处理成了如下异矩分组：

成绩分组	人数	下限	上限	组距	组中组
(50,60]	2	50	60	10	55
(60,80]	28	60	80	20	70
(80,90]	7	80	90	10	85
(90,100]	3	90	100	10	95

（案例）学生考试成绩：异矩分组统计量

在异矩分组情形下，**频次密度**和**标准组距频次**计算如下：

	成绩分组	人数	下限	上限	组距	组中组	频次密度	标准组距人数
(50,60]	2	50	60	10	55	0.2	2
(60,80]	28	60	80	20	70	1.4	14
(80,90]	7	80	90	10	85	0.7	7
(90,100]	3	90	100	10	95	0.3	3

数据分组：统计制表（类型）

对分组数据进行统计制表，也即用统计表来表示次数/频率等统计量在各组的分布情况，主要包括。

1)制表类型A　2)制表类型B　3)制表类型C

- **频数表/百分数表**（已经展示，见前面slide）

数据分组：统计制表（类型）

对分组数据进行统计制表，也即用统计表来表示次数/频率等统计量在各组的分布情况，主要包括。

1)制表类型A　2)制表类型B　3)制表类型C

累计次数表/累计百分数表。复习之前的定义，具体为：

- 较小制累计（以下累计、向上累计）：即（上限）以下累计次数，每一组的累计次数表示小于该组上限（变量）值的次数/频率共有多少。
- 较大制累计（以上累计、向下累计）：即（下限）以上累计次数：每一组的累计次数表示大于该组下限（变量）值的次数/频率共有多少。

数据分组：统计制表（类型）

对分组数据进行统计制表，也即用统计表来表示次数/频率等统计量在各组的分布情况，主要包括。

1)制表类型A　2)制表类型B　3)制表类型C

交叉分析表，又称为**交叉列联表**（cross-table），是对复合式分组数列的频数或频率统计，便于对两个或多个分组标志（分组变量）关系的直接观察。

数据交叉形式可以是：

- 品质变量VS品质变量
- 品质变量VS数值变量（较少用*）
- 数值变量VS数值变量（较少用*）

（案例）学生考试成绩：累积次数/频率表

对于等距式分组情形，我们可以分别计算出较小/较大累计次数/频率：

1)较小制累积表　2)较大制累积表　3)较小制和较大制对比

我们可以计算得到**较小制**下的累积频次和频率，并制表：

groups	n	percent	min_cum_n	min_cum_p
(50,60]	2	5.0%	2	5.0%
(60,70]	14	35.0%	16	40.0%
(70,80]	14	35.0%	30	75.0%
(80,90]	7	17.5%	37	92.5%
(90,100]	3	7.5%	40	100.0%

（案例）学生考试成绩：累积次数/频率表

对于等距式分组情形，我们可以分别计算出较小/较大累计次数/频率：

1)较小制累积表　2)较大制累积表　3)较小制和较大制对比

我们也可以计算得到**较大制**下的累积频次和频率，并制表：

groups	n	percent	max_cum_n	max_cum_p
(50,60]	2	5.0%	40	100.0%
(60,70]	14	35.0%	38	95.0%
(70,80]	14	35.0%	24	60.0%
(80,90]	7	17.5%	10	25.0%
(90,100]	3	7.5%	3	7.5%

（案例）学生考试成绩：累积次数/频率表

对于等距式分组情形，我们可以分别计算出较小/较大累计次数/频率：

1)较小制累积表　2)较大制累积表　3)较小制和较大制对比

我们可以对比观测**较小制**和**较大制**下的累积频次和频率：

groups	n	percent	min_cum_n	min_cum_p	max_cum_n	max_cum_p
(50,60]	2	5.0%	2	5.0%	40	100.0%
(60,70]	14	35.0%	16	40.0%	38	95.0%
(70,80]	14	35.0%	30	75.0%	24	60.0%
(80,90]	7	17.5%	37	92.5%	10	25.0%
(90,100]	3	7.5%	40	100.0%	3	7.5%

（案例）学生考试成绩：等距分组的交叉分析表

a.性别信息　b.交叉列表

对于前述**学生考试成绩**案例，研究者还收集了40名学生的性别信息（见下表）。

ID	score	gender	groups
s01	63	女生	(60,70]
s02	88	男生	(80,90]
s03	72	女生	(70,80]
s04	69	女生	(60,70]
s05	80	男生	(70,80]
s06	80	女生	(70,80]

Showing 1 to 06 of 40 entries

Previous 1 2 3 4 5 6 7 Next

（案例）学生考试成绩：等距分组的交叉分析表

a.性别信息　　b.交叉列表

此时，可以根据需要绘制出性别变量与成绩分组的**交叉分析表**（列联表）：

成绩分组	男生	女生	Total
(50,60]	0	2	2
(60,70]	5	9	14
(70,80]	6	8	14
(80,90]	4	3	7
(90,100]	1	2	3
Total	16	24	40

数据分组：统计制图

分组数列的图示方法，也即用统计图形来表示频数/频率在各组的分布情况，主要包括的图形类型有：

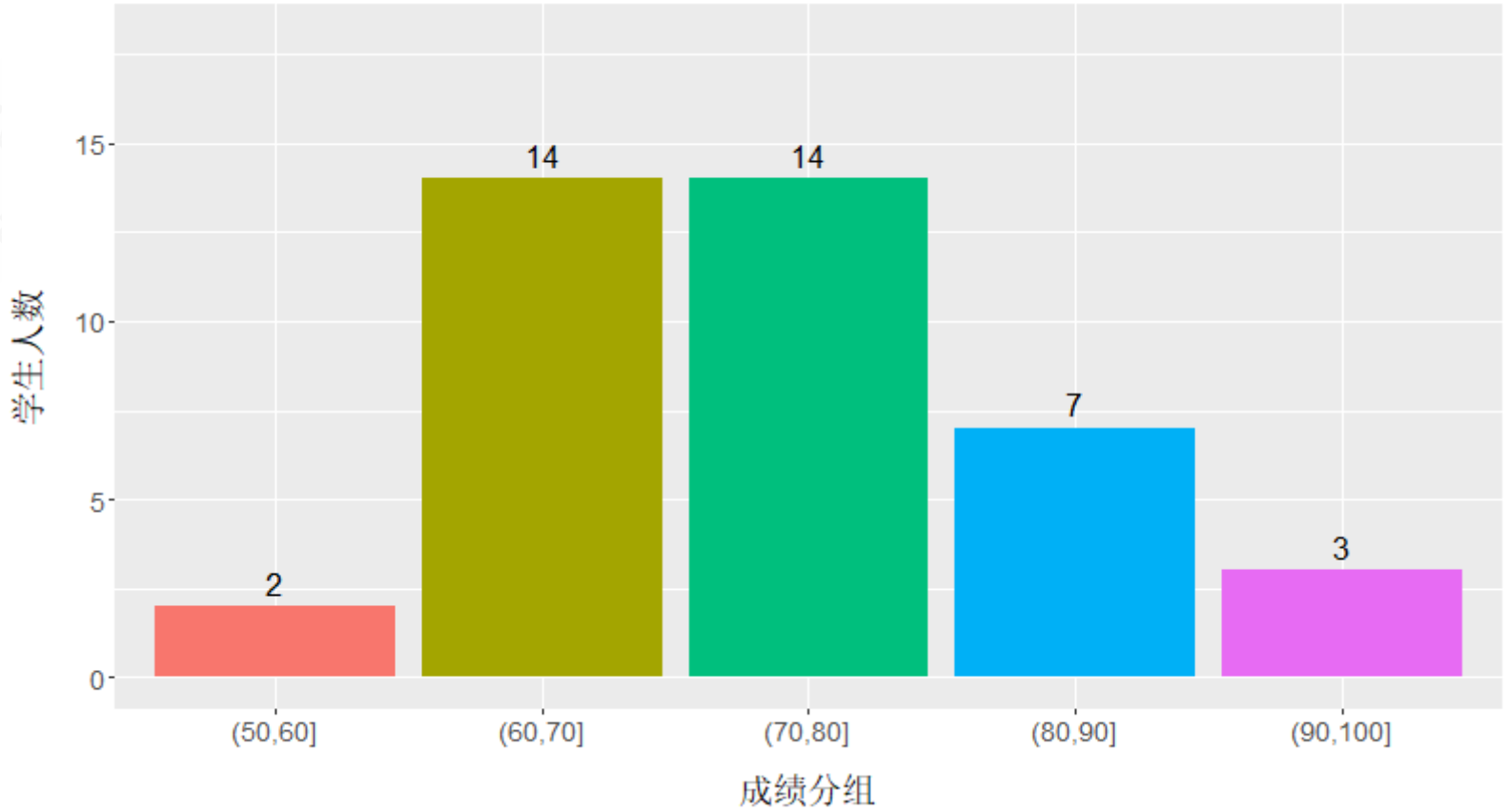
a. **条形/柱状图**[已讲，见前面slide]

b. **折线图**（line chart）

c. **累积频次/频率图**（cumulative chart）

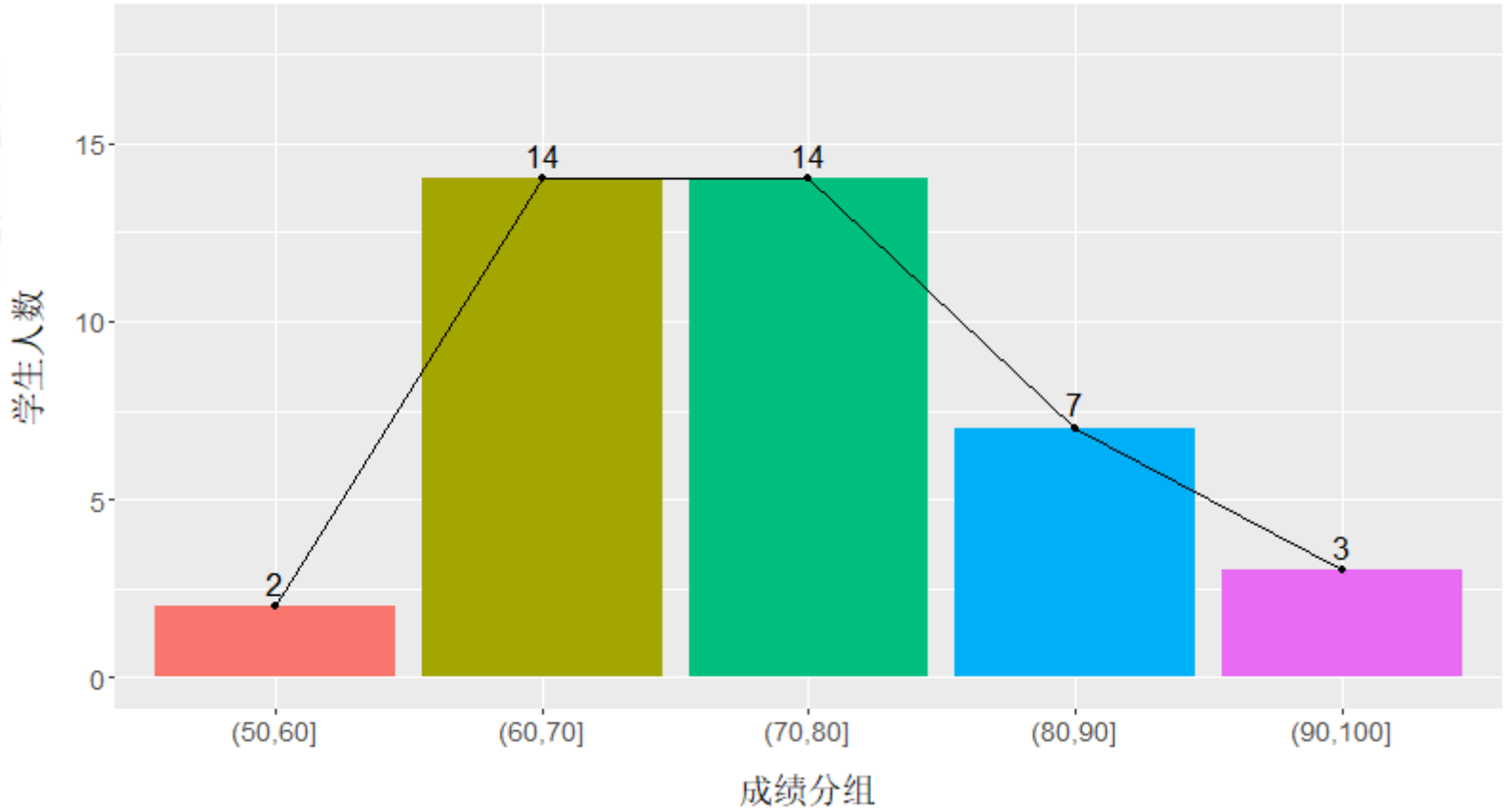
（案例）学生考试成绩：绘制柱状图/折线图

1)柱状图　2)折线图



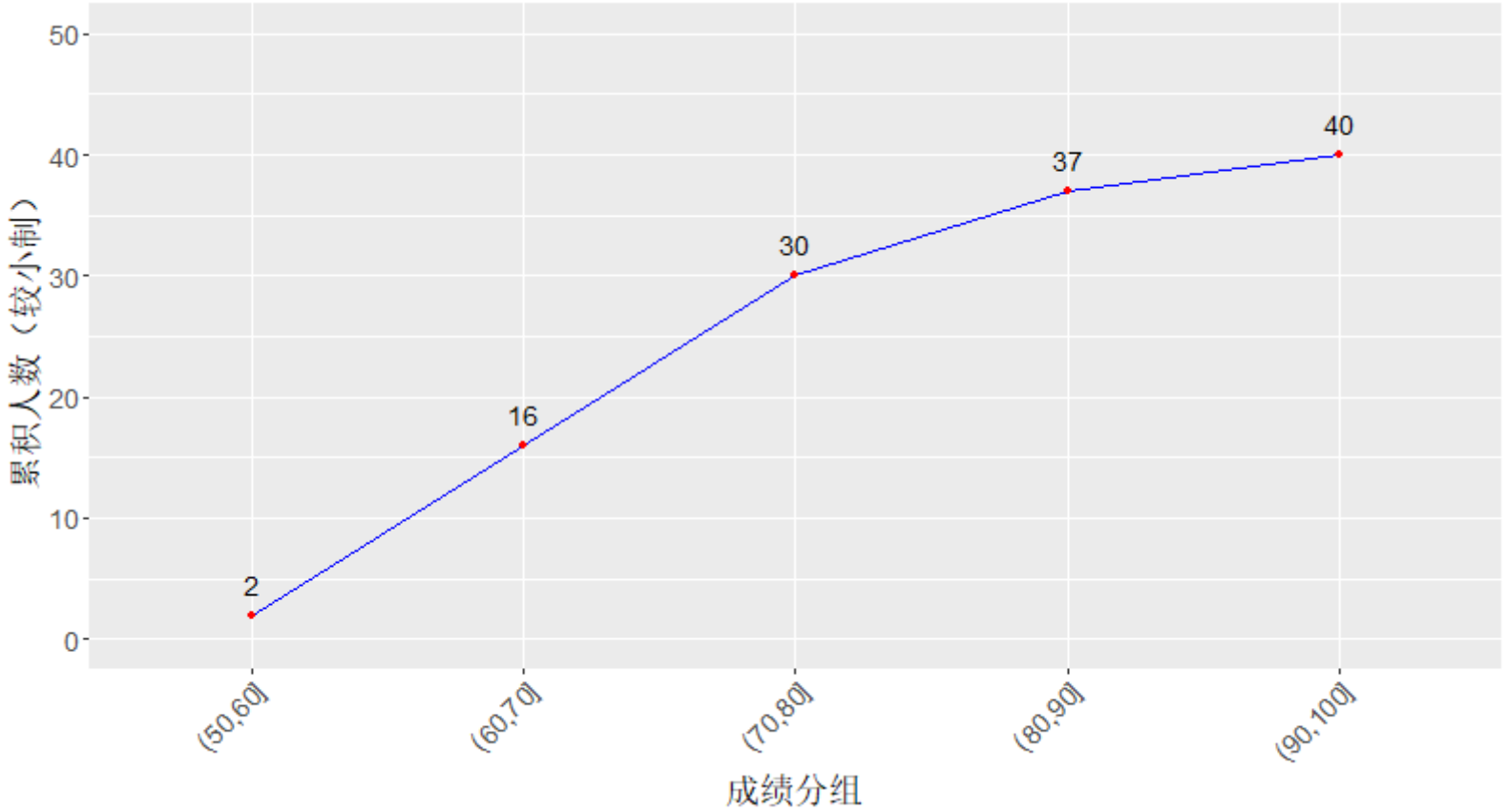
（案例）学生考试成绩：绘制柱状图/折线图

1)柱状图　2)折线图



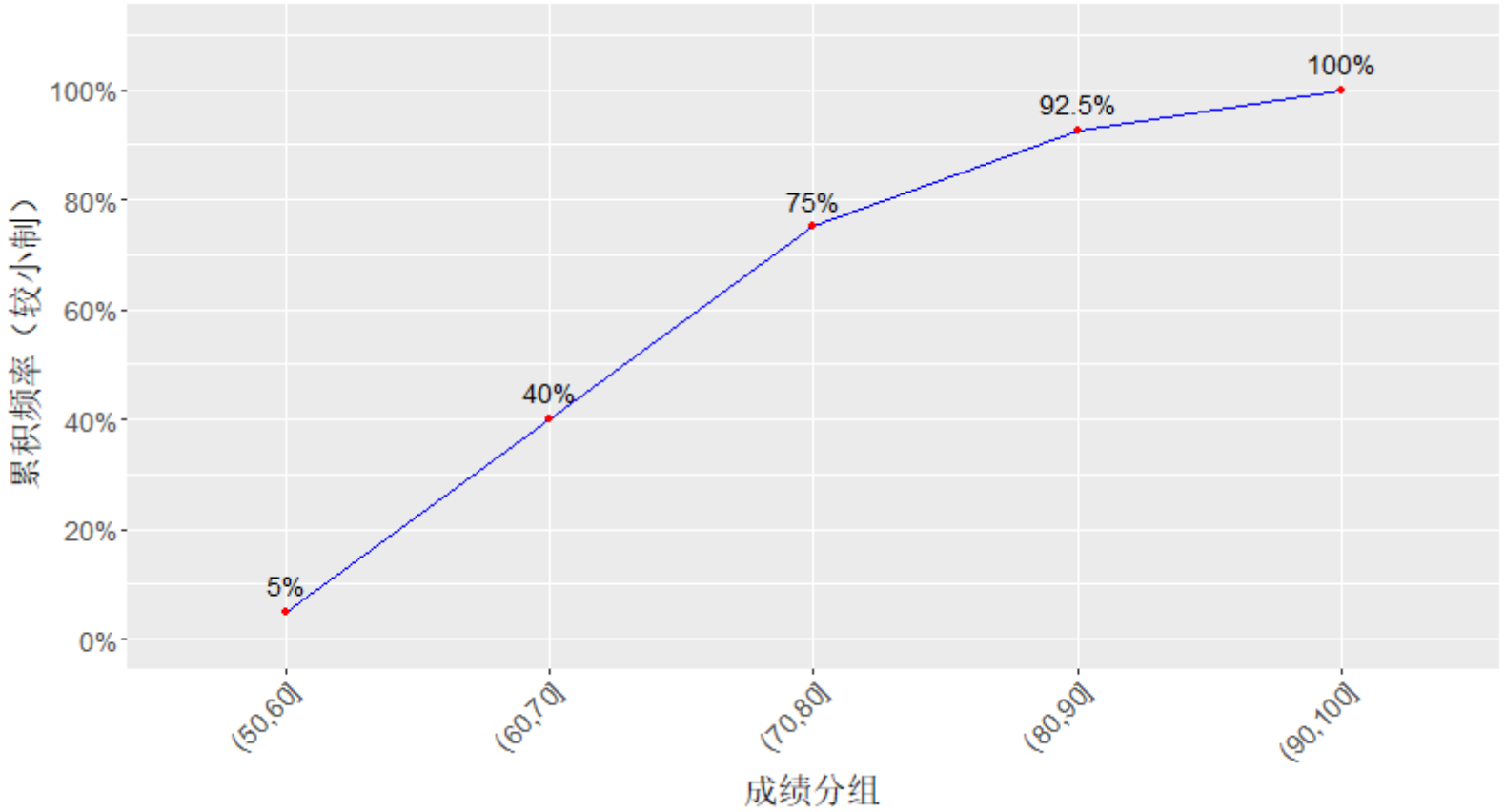
（案例）学生考试成绩：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　3)较大累积频次　4)较大累积频率



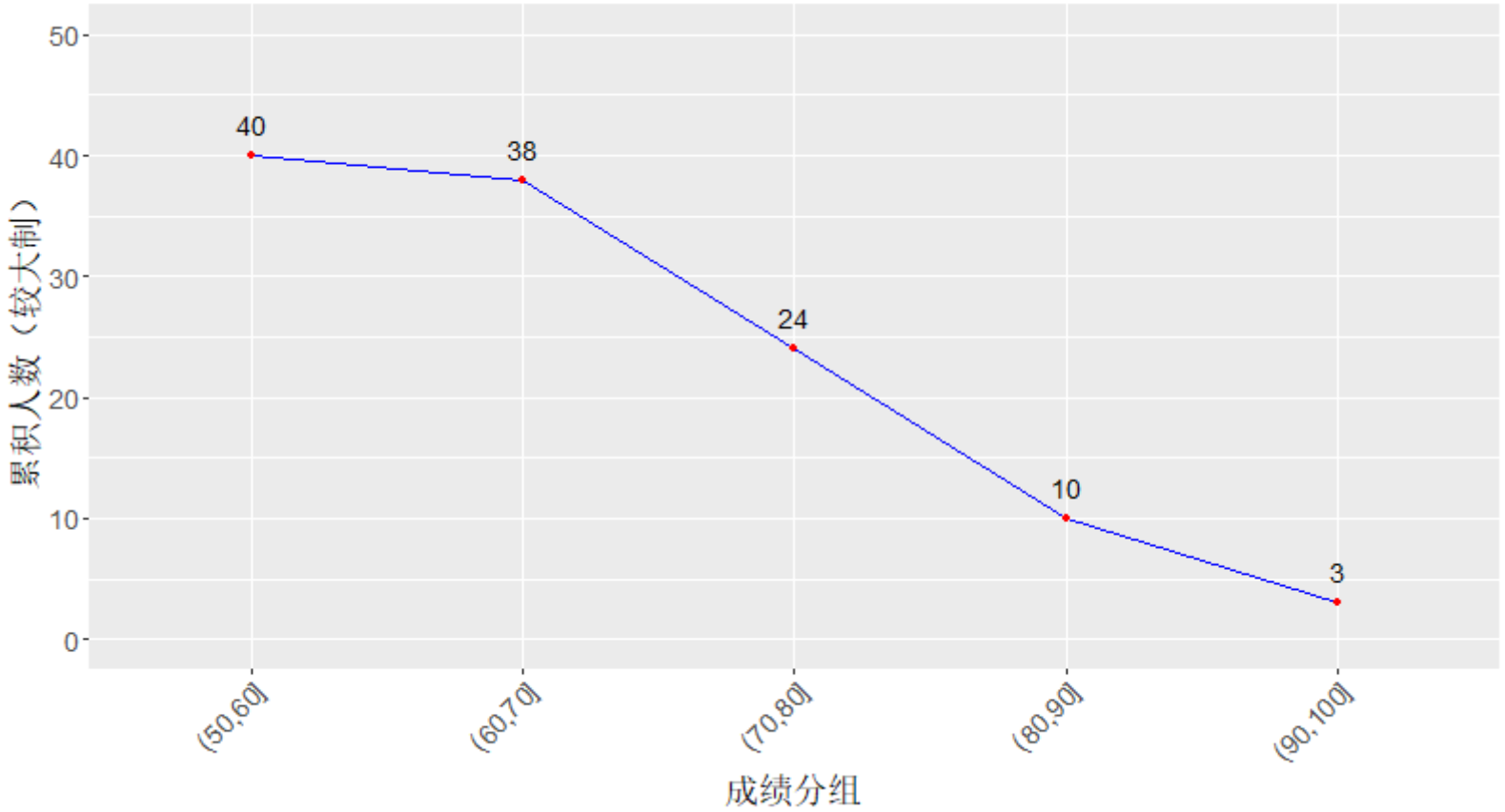
（案例）学生考试成绩：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　3)较大累积频次　4)较大累积频率



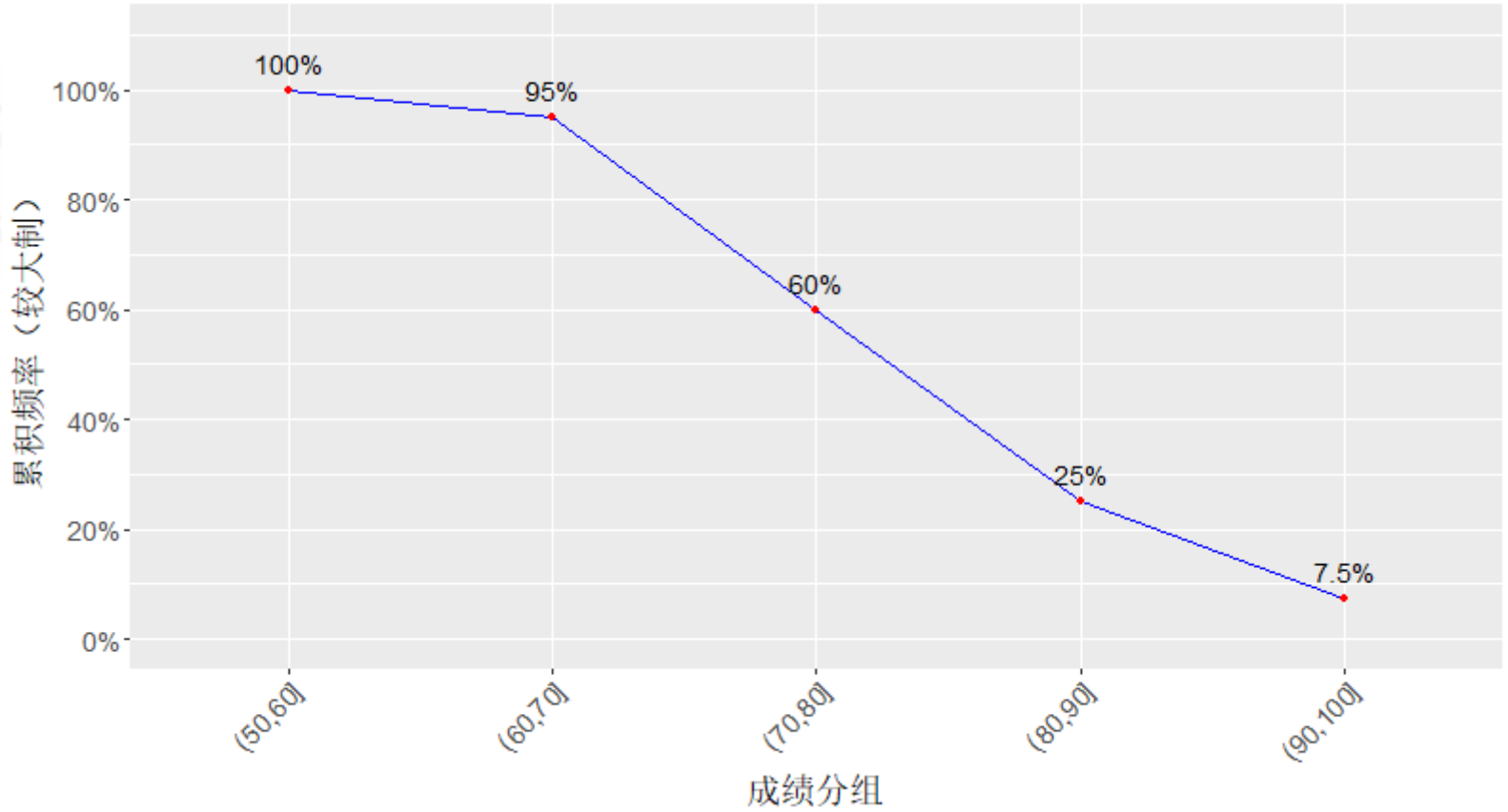
（案例）学生考试成绩：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　**3)较大累积频次**　4)较大累积频率



（案例）学生考试成绩：绘制累计频次/频率图

1)较小累积频次　2)较小累积频率　3)较大累积频次　**4)较大累积频率**



数据未分组：统计制图1（茎叶图）

茎叶图（stem and leaf diagrams）：用于显示未分组的原始数据的分布。由“茎”和“叶”两部分构成，其图形是由数字组成的，以该组数据的高位数值作树茎，低位数字作树叶，树叶上只保留最后一位数字。

茎叶图与**横直方图**的区别*：

- 直方图可观察一组数据的分布状况，但没有给出具体的数值。
- 茎叶图既能给出数据的分布状况，又能给出每一个原始数值，保留了原始数据的信息。
- 直方图适用于大批量数据，茎叶图适用于小批量数据。

注释：* 茎叶图曾经的优势（简单、可手工绘制）在今天这个计算机时代也显得并不突出，因此，除非特殊情况，建议主要使用直方图作为密度函数估计工具。see link

（案例）学生考试成绩：绘制茎叶图

1)原始成绩单　2)按成绩排序　3)制作茎叶图

继续考虑之前的**学生考试成绩案例**。40名学生课程考试成绩如下：

```
s01 s02 s03 s04 s05 s06 s07 s08 s09 s10 s11 s12 s13 s14 s15 s16 s17
 63  88  72  69  80  80  61  68  79  81  76  76  77  78  90  89  61
s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29 s30 s31 s32 s33 s34
 84  92  67  65  57  62  60  66  87  80  92  78  86  71  71  63  74
s35 s36 s37 s38 s39 s40
 67  70  91  64  79  65
```

（案例）学生考试成绩：绘制茎叶图

1)原始成绩单　2)按成绩排序　3)制作茎叶图

我们先按成绩由低到高进行排序：

```
s22 s24 s07 s17 s23 s01 s33 s38 s21 s40 s25 s20 s35 s08 s04 s36 s31
 57  60  61  61  62  63  63  64  65  65  66  67  67  68  69  70  71
s32 s03 s34 s11 s12 s13 s14 s29 s09 s39 s05 s06 s27 s10 s18 s30 s26
 71  72  74  76  76  77  78  78  79  79  80  80  80  81  84  86  87
s02 s16 s15 s37 s19 s28
 88  89  90  91  92  92
```

（案例）学生考试成绩：绘制茎叶图

1)原始成绩单　2)按成绩排序　3)制作茎叶图

```
The decimal point is 1 digit(s) to the right of the |

5 | 7
6 | 01123345567789
7 | 011246678899
8 | 000146789
9 | 0122
```

数据未分组：统计制图2（箱线图）

箱线图（box plot）：用于显示未分组的原始数据的分布。由一组数据的5个特征值绘制而成，它由一个箱子和两条线段组成。箱线图也被称为Median/Quartile/Range箱线图。

绘制方法：

- 首先找出一组数据的5个特征值，即最大值（Max）、最小值（Min）、中位数 M_e和两个四分位数（下四分位数 Q_L和上四分位数 Q_U）。
- 连接两个四分位数画出箱子，再将两个极值点与箱子相连接。

图形分类：

- 单一箱线图
- 多批箱线图

（案例）学生考试成绩：绘制单一箱线图

1)原始成绩单　2)计算统计量　3)单一箱线图

继续考虑之前的**学生考试成绩案例**。40名学生课程考试成绩如下：

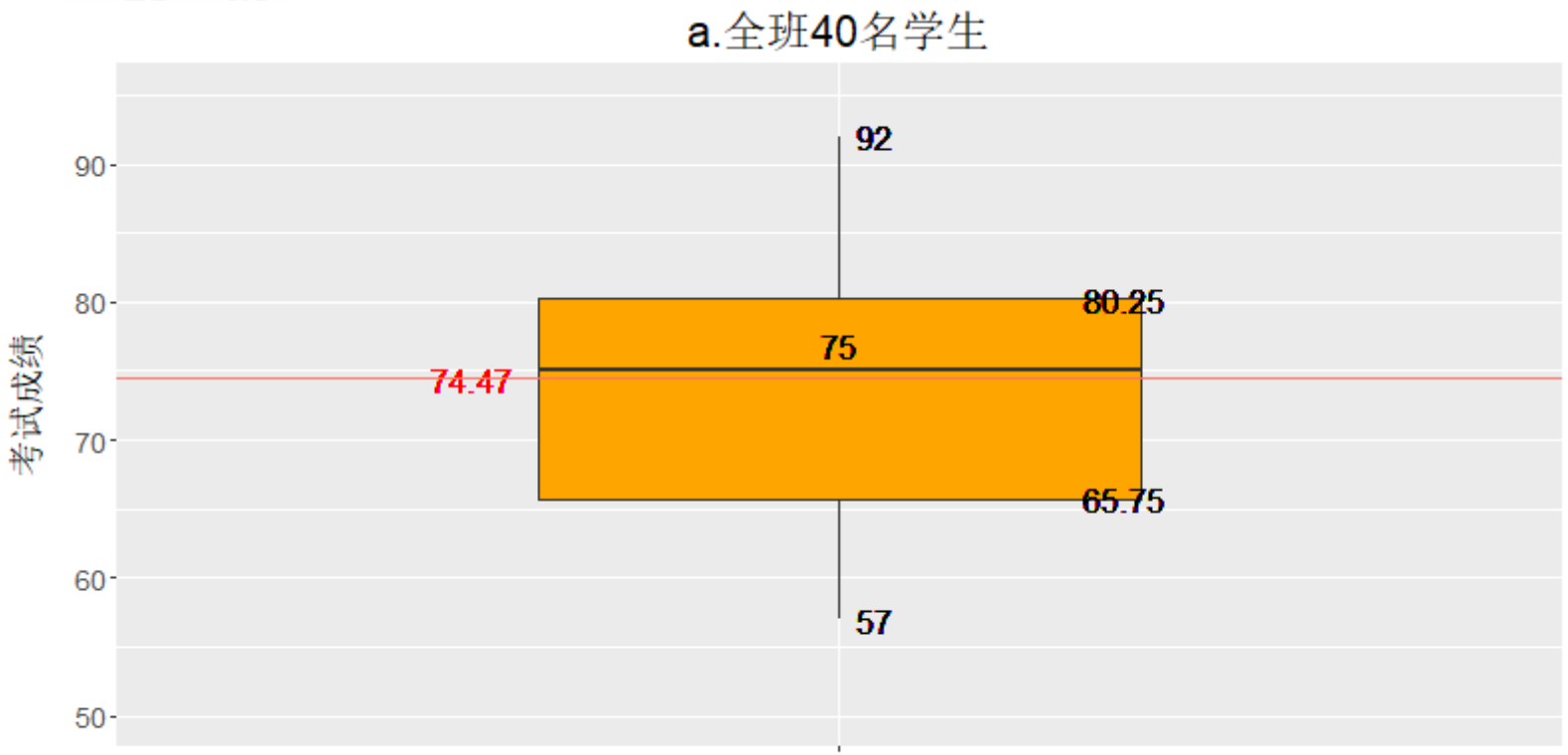
```
s01 s02 s03 s04 s05 s06 s07 s08 s09 s10 s11 s12 s13 s14 s15 s16 s17
 63  88  72  69  80  80  61  68  79  81  76  76  77  78  90  89  61
s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29 s30 s31 s32 s33 s34
 84  92  67  65  57  62  60  66  87  80  92  78  86  71  71  63  74
s35 s36 s37 s38 s39 s40
 67  70  91  64  79  65
```

我们可以先计算出箱线图的五个制表：

- 中位数 $median = 75$
- 极大值 $median = 92$
- 极小值 $median = 57$
- 四分之一位数 $median = 65.75$
- 四分之三位数 $median = 80.25$

（案例）学生考试成绩：绘制单一箱线图

1)原始成绩单　2)计算统计量　**3)单一箱线图**



（案例）学生考试成绩：绘制多批箱线图

1)原始成绩单　2)计算统计量　3)多批箱线图

同时考虑学生性别和**考试成绩**。40名学生课程考试信息如下：

ID	score	gender
s01	63	女生
s02	88	男生
s03	72	女生
s04	69	女生
s05	80	男生

Showing 1 to 05 of 40 entries　Previous 1 2 3 4 5 … 8 Next

（案例）学生考试成绩：绘制多批箱线图

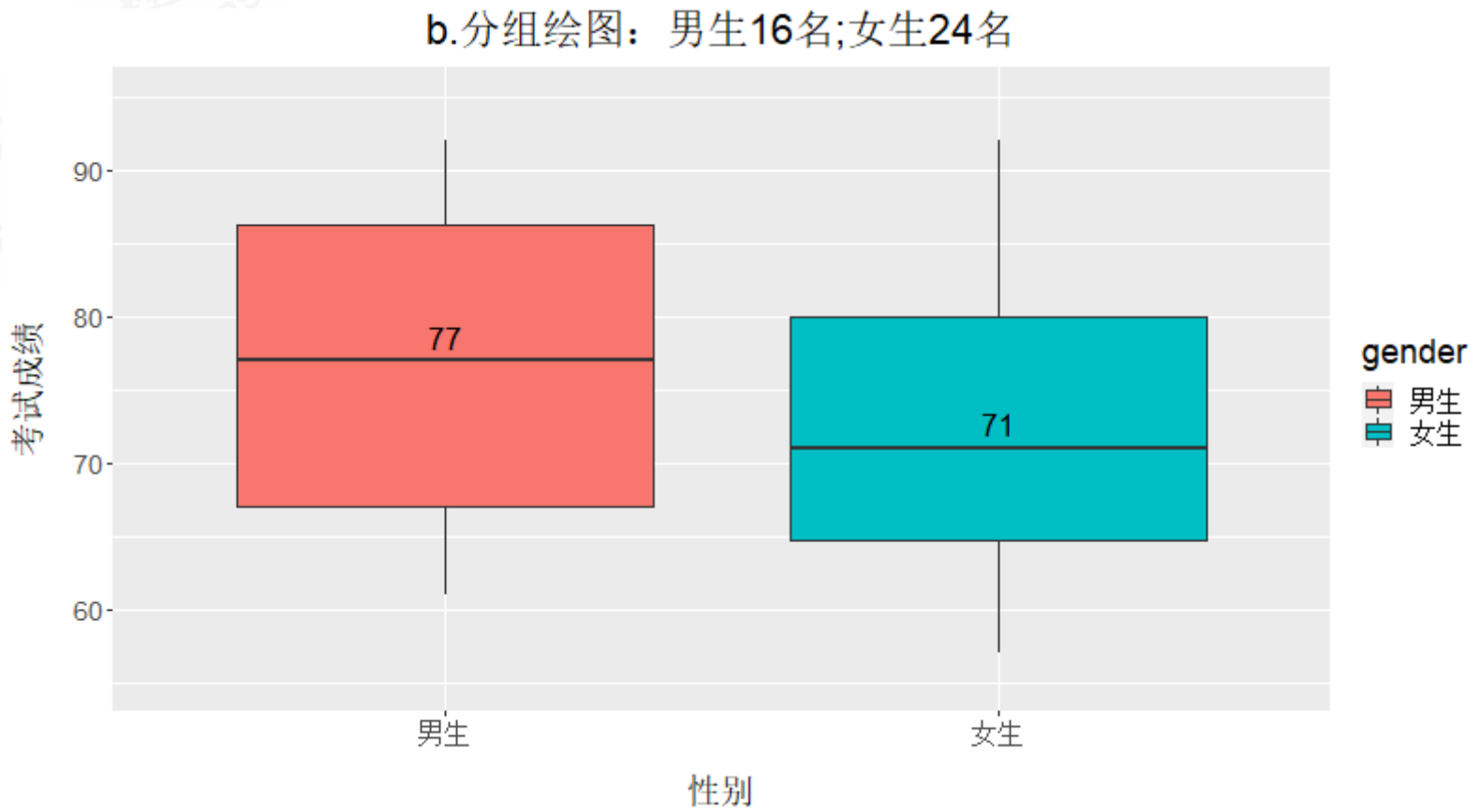
1)原始成绩单　2)计算统计量　3)多批箱线图

我们可以根据性别分组，分别计算出箱线图的五个指标：

gender	med	mean	Q1	Q3	max	min
男生	77	76.75	67	86.25	92	61
女生	71	72.96	64.75	80	92	57

（案例）学生考试成绩：绘制多批箱线图

1)原始成绩单　2)计算统计量　3)多批箱线图



数据未分组：统计制图3（直方图）

直方图（histogram）：用于展示分组数据分布的一种图形，用矩形的宽度和高度来表示频数分布。

- 本质上是用矩形的面积来表示频数分布。
- 在直角坐标中，用横轴表示数据分组，纵轴表示频数或频率，各组与相应的频数就形成了一个矩形。

直方图与柱状图的区别：

- 柱状图是用柱形的高度表示各类别频数的多少，其宽度（表示类别）则是固定的。
- 直方图是用面积表示各组频数的多少，矩形的高度表示每一组的频数或百分比，宽度则表示各组的组距，其高度与宽度均有意义。
- 直方图的各矩形通常是连续排列，柱状图则是分开排列。
- 柱状图主要用于展示分类/分组数据，直方图则主要用于展示数值型数据。

（示例）数据未分组的制图：直方图（histogram）

1)数据表　2)数据概览　3)全体直方图　4)分组直方图

案例说明：某个学院共有2000名学生参加《统计学原理》课程考试，考试成绩和性别信息如下：

ID	gender	score
s0001	女生	84
s0002	女生	87
s0003	女生	82
s0004	女生	81
s0005	男生	81

Showing 1 to 5 of 2,000 entries　Previous 1 2 3 4 5 … 400 Next

（示例）数据未分组的制图：直方图（histogram）

1)数据表　2)数据概览　3)全体直方图　4)分组直方图

全样本概览

variables	n	mean	median	max	min
score	2000	83.46	82	100	69

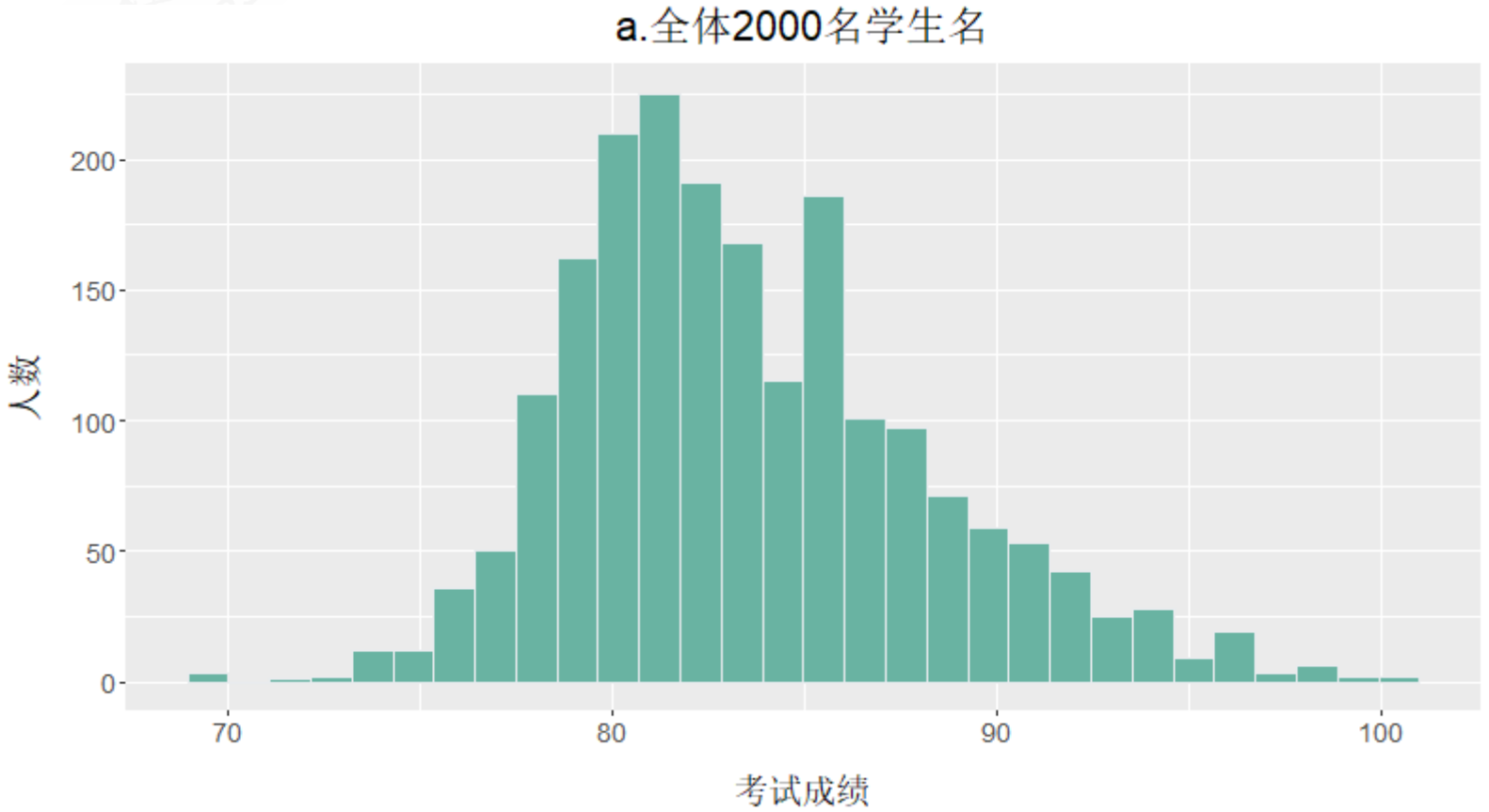
我们再按性别分两类样本数据来看成绩情况：

性别子集概览

set	variables	n	mean	median	max	min
男生组	score	800	85.43	85.5	100	70
女生组	score	1200	82.15	81	98	69

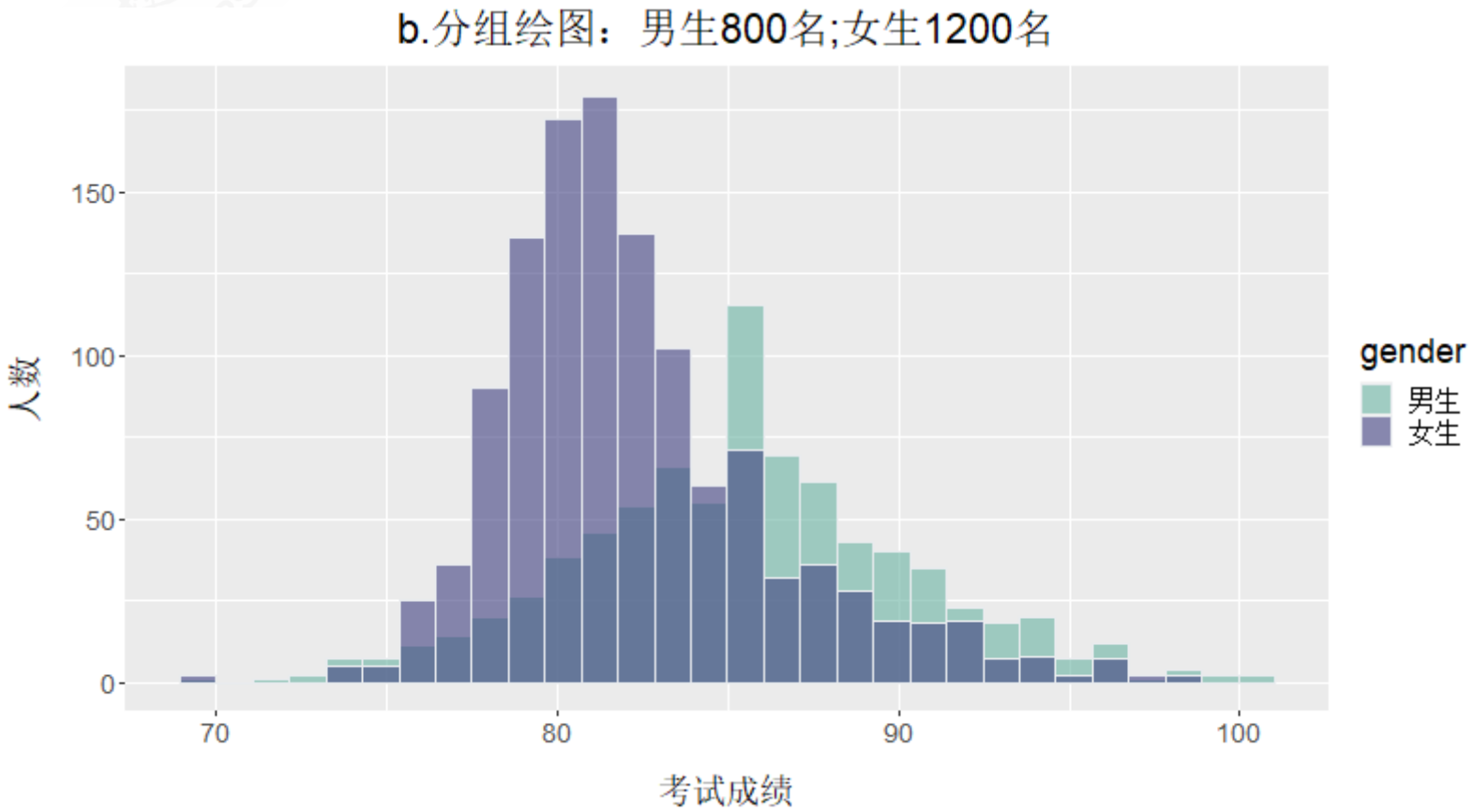
（示例）数据未分组的制图：直方图（histogram）

1)数据表　2)数据概览　3)全体直方图　4)分组直方图



（示例）数据未分组的制图：直方图（histogram）

1)数据表　2)数据概览　3)全体直方图　4)分组直方图



数据未分组：统计制图4（线图）

线图(line plot)：主要用于表示时间序列数据趋势的图形。

绘制要点：

- 时间一般绘在横轴，数据绘在纵轴。
- 图形的长宽比例大致为`10:7`。
- 一般情况下，纵轴数据下端应从“0”开始，以便于比较。数据与“0”之间的间距过大时，可以采取折断的符号将纵轴折断。

（案例）全球新冠疫情：数据说明

案例说明：研究人员收集了8个国家（US、France、Norway、Switzerland、United Kingdom、Germany、Italy、Spain），共3260条新冠疫情数据（见下表）。

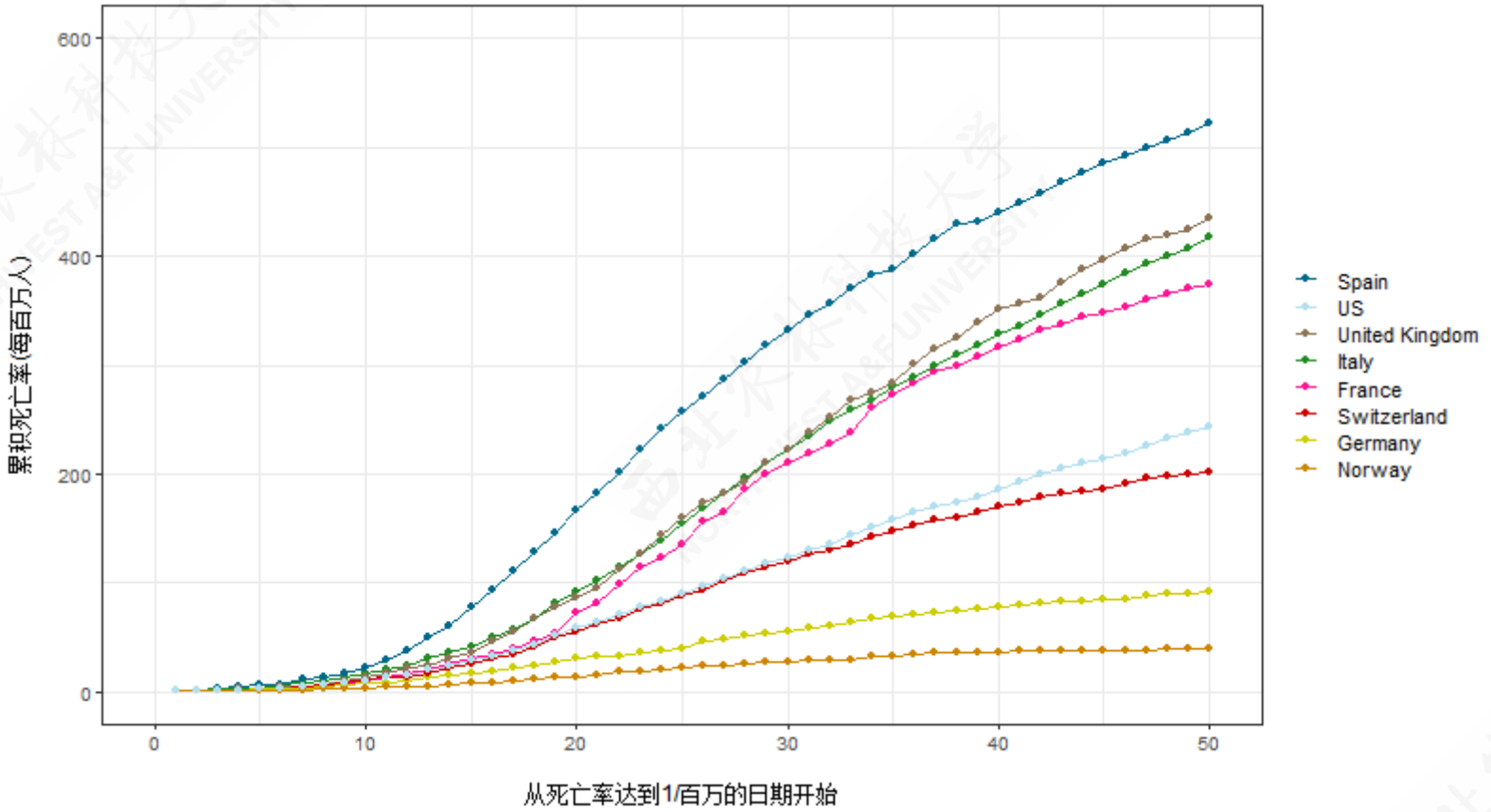
index	country	deaths	population	deathspc	t0	days
1	France	79	65721	1.20	true	1
2	France	91	65721	1.38	true	2
3	France	91	65721	1.38	true	3
4	France	149	65721	2.27	true	4
5	France	149	65721	2.27	true	5
6	France	149	65721	2.27	true	6

Showing 1 to 06 of 3,260 entries

Previous 1 2 3 4 5 … 544 Next

（案例）全球新冠疫情：死亡率线图

为了比较各国新冠死亡率的变化趋势，绘制得到如下线图：



数据未分组：统计制图5（散点图）

散点图（Scatter plots）：展示两个变量之间的关系。

用横轴代表变量 X_i，纵轴代表变量 Y_i，每组数据 (X_i, Y_i)在坐标系中用一个点表示，n组数据在坐标系中形成的 n个点称为散点，由坐标及其散点形成的二维数据图。

（案例）汽车油耗：绘制散点图

1)数据表　2)散点图1　3)散点图2

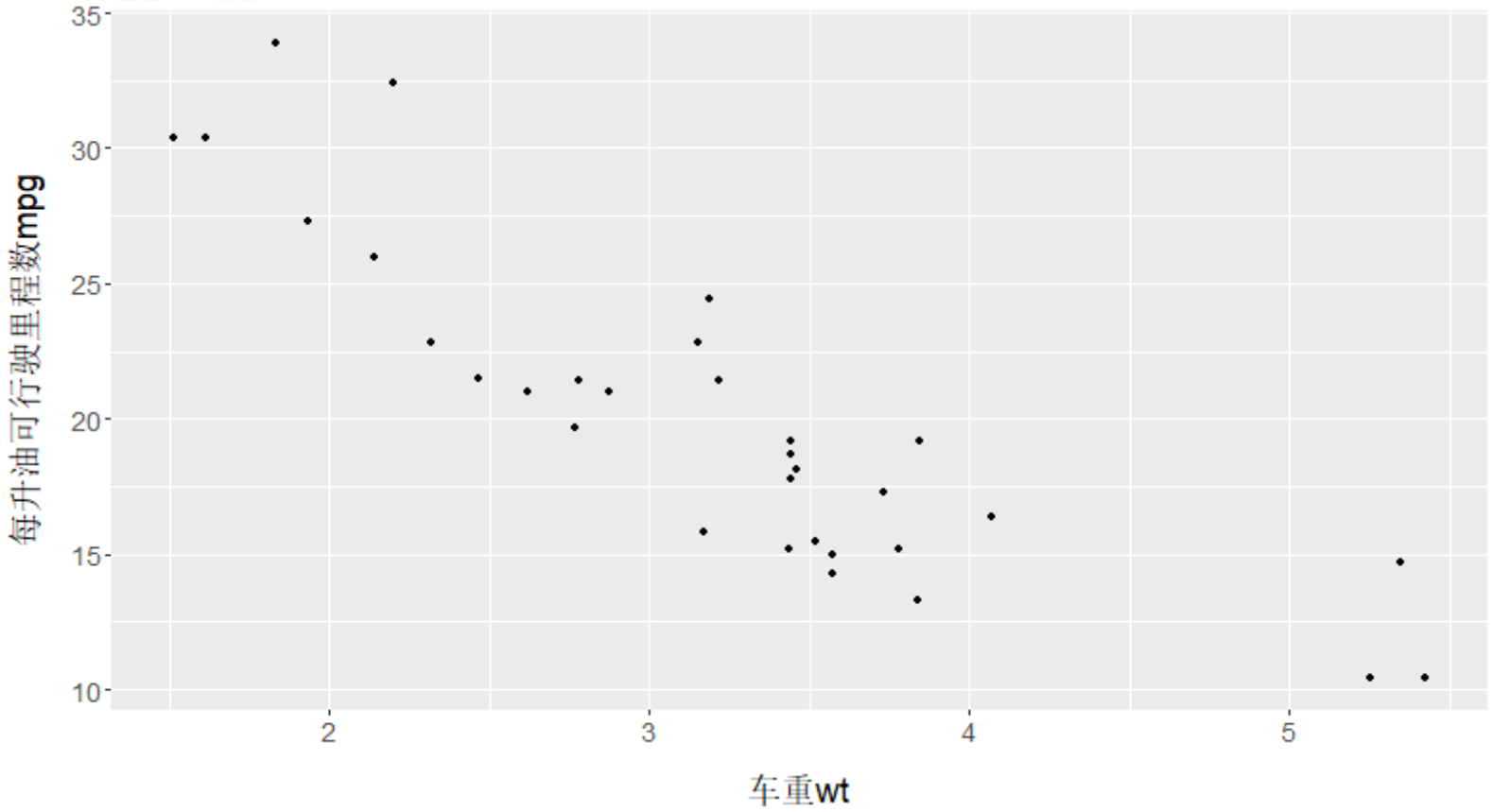
案例说明：研究人员希望了解汽车油耗的情况，收集了汽车车重、轴距、气缸数量等的数据信息（见下表）：

index	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	21	6	160	110	3.9	2.62	16.46	0	1	4	4
2	21	6	160	110	3.9	2.875	17.02	0	1	4	4
3	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
4	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
5	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2

Showing 1 to 5 of 32 entries　Previous 1 2 3 4 5 6 7 Next

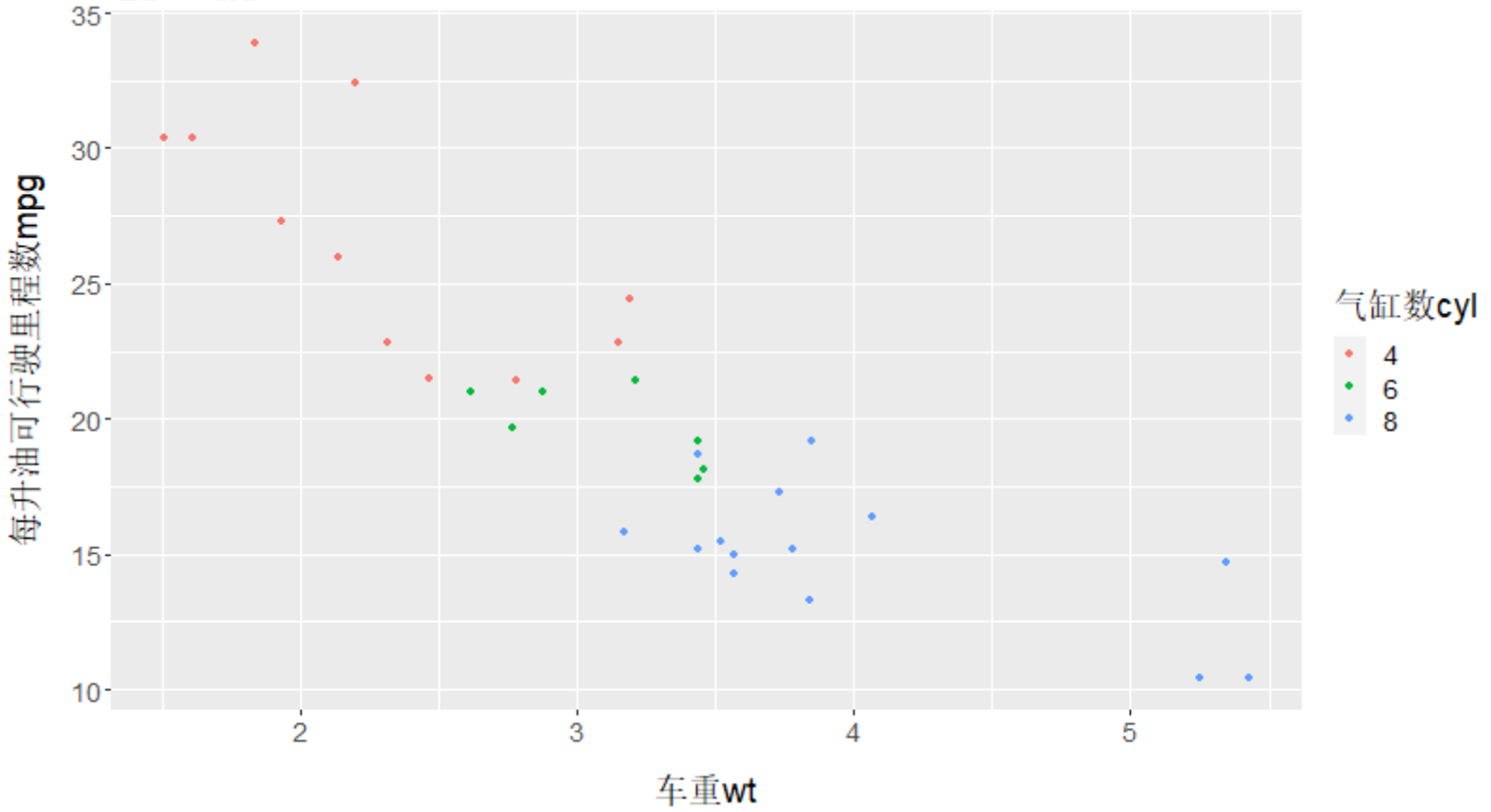
（案例）汽车油耗：绘制散点图

1)数据表　2)散点图1　3)散点图2



（案例）汽车油耗：绘制散点图

1)数据表　2)散点图1　3)散点图2



数据未分组：统计制图6（散点图）

气泡图（bubble chart）：显示**三个变量**之间的关系， 图中数据点的大小依赖于第三个变量。

（案例）人均寿命：绘制气泡图

1)数据表　2)气泡图

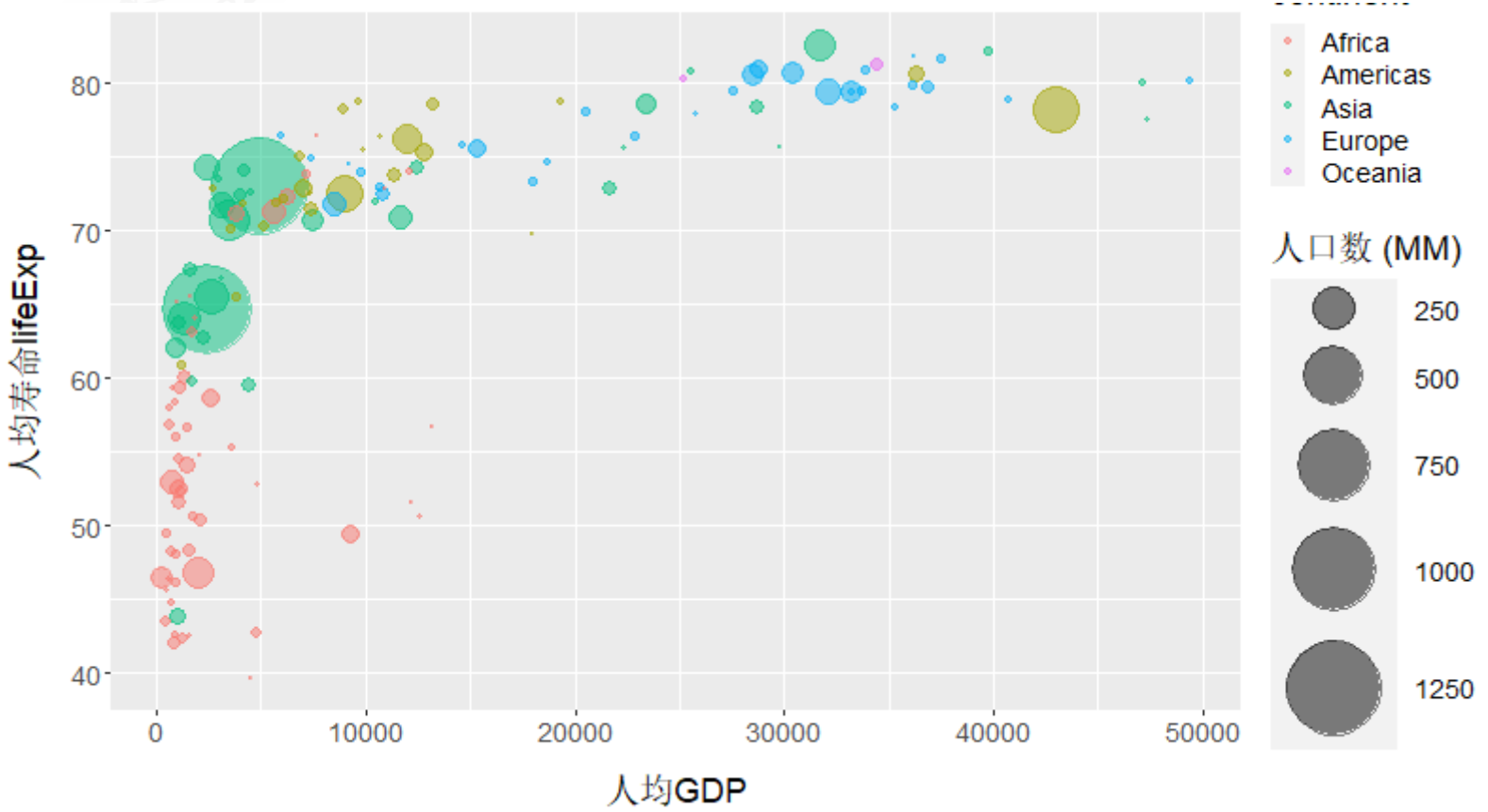
案例说明：下表给出了各个地区在2007年的人均寿命、人均GDP等数据。

index	country	continent	lifeExp	pop	gdpPercap
1	Afghanistan	Asia	43.828	31889923	974.58
2	Albania	Europe	76.423	3600523	5,937.03
3	Algeria	Africa	72.301	33333216	6,223.37
4	Angola	Africa	42.731	12420476	4,797.23
5	Argentina	Americas	75.32	40301927	12,779.38

Showing 1 to 5 of 142 entries　Previous 1 2 3 4 5 … 29 Next

（案例）人均寿命：绘制气泡图

1)数据表　2)气泡图



数据未分组：统计制图7（雷达图）

雷达图（radar chart）：也称为蜘蛛图（spider chart），显示**多个变量**的图示方法。

用途：

- 在显示或对比各变量的数值总和时十分有用。
- 假定各变量的取值具有相同的正负号，总的绝对值与图形所围成的区域成正比。
- 可用于研究多个样本之间的相似程度。

（案例）领域发展评估：绘制雷达图

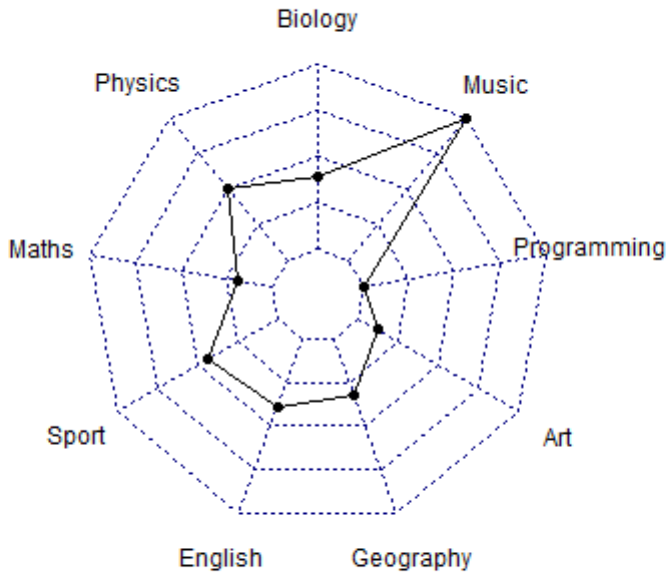
1)数据表　2)雷达图1　3)雷达图2

案例说明：下表给出了三个学生在不同领域发展的评分结果。

students	Biology	Physics	Maths	Sport	English	Geography	Art	Progı
S1	7.9	10	3.7	8.7	7.9	6.4	2.4	
S2	3.9	20	11.5	20	7.2	10.5	0.2	
S3	9.4	0	2.5	4	12.4	6.5	9.8	

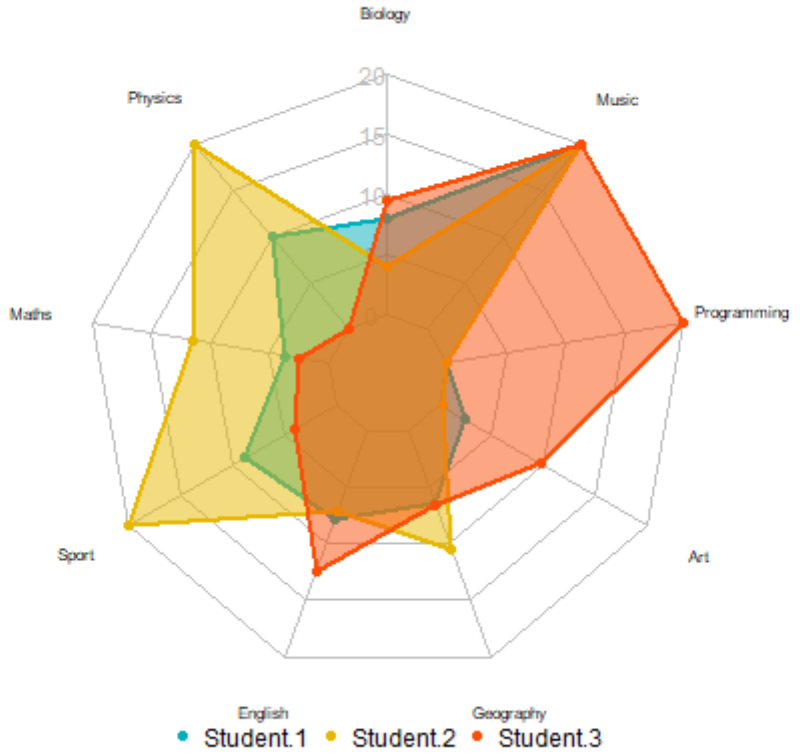
（案例）领域发展评估：绘制雷达图

1)数据表　2)雷达图1　3)雷达图2



（案例）领域发展评估：绘制雷达图

1)数据表　　2)雷达图1　　3)雷达图2



数据分布形态：概述

典型分布形态：

- 正态分布，也称正态曲线
- 偏态分布，或称偏态曲线

其他分布形态：

- U型分布
- 双峰分布
- J型分布
- 反J型分布

待完成：提供示例图形，给出代表分布类型示例。

数据分布形态：偏度系数

偏度（Skewness）系数：分布形态的一个重要衡量标准就是分布偏度。一组数据分布偏度的计算是复杂的，但使用统计软件可以很容易的计算出偏度系数。偏度系数 SK的理论计算公式如下：

$$SK = \frac{n}{(n-1)(n-2)} \sum_{1}^{n} \left(\frac{X_i - \bar{X}}{S_X}\right)^3$$

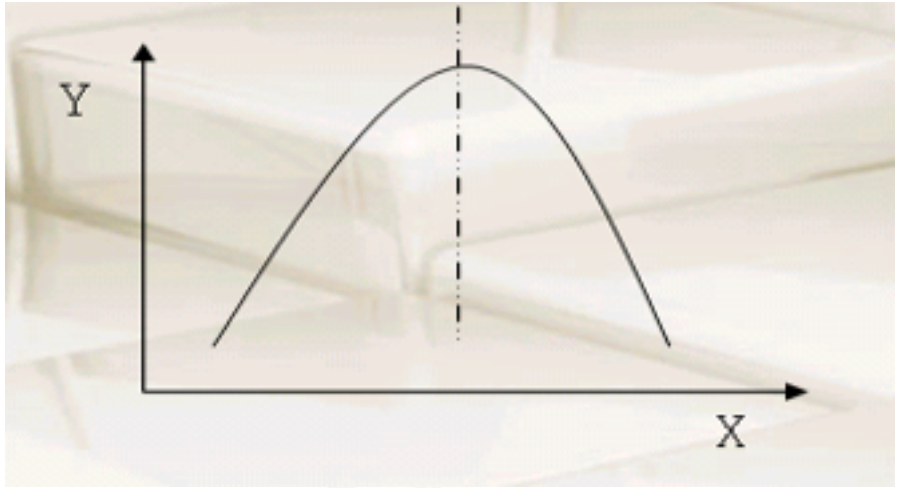
> 其中： n表示样本数； S_X表示样本标准差 $S_X = \sqrt{\frac{\sum_1^n (X_i - \bar{X})^2}{n-1}}$。

数据分布形态：偏度系数

- 若偏度系数 $SK=0$，则数据分布是对称的（无偏的），此时均值和中位数相等，也即：$\bar{X}=M_e$。
- 若偏度系数 $SK<0$，则数据分布是非对称的（左偏的），此时均值小于中位数，也即：$\bar{X}<M_e$。
- 若偏度系数 $SK>0$，则数据分布是非对称的（左偏的），此时均值大于中位数，也即：$\bar{X}>M_e$。

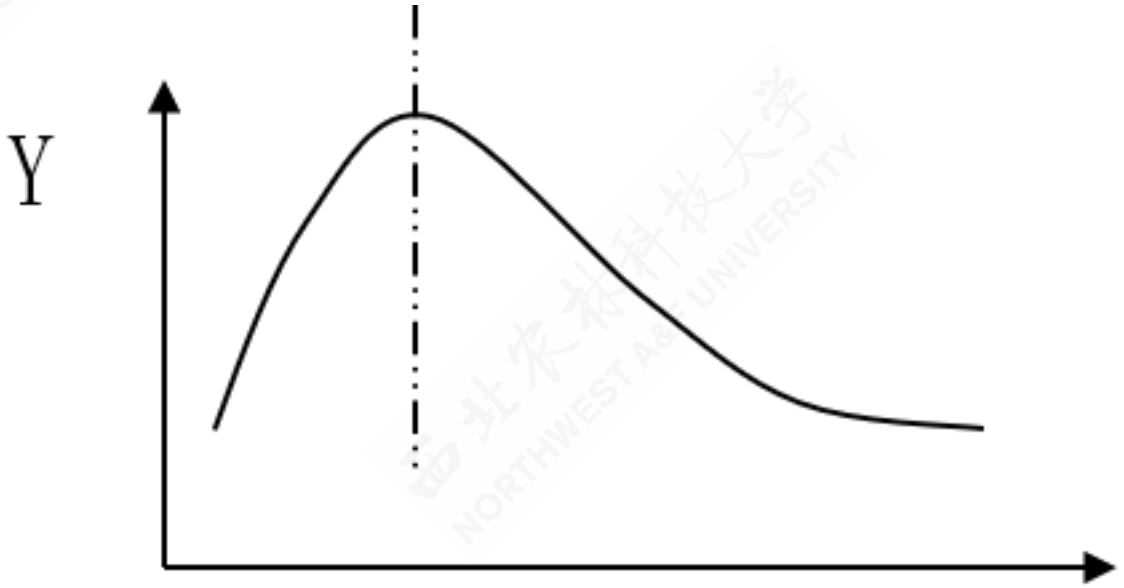
（示例）数据分布形态：常见形态

1)对称分布　2)右偏分布　3)左偏分布



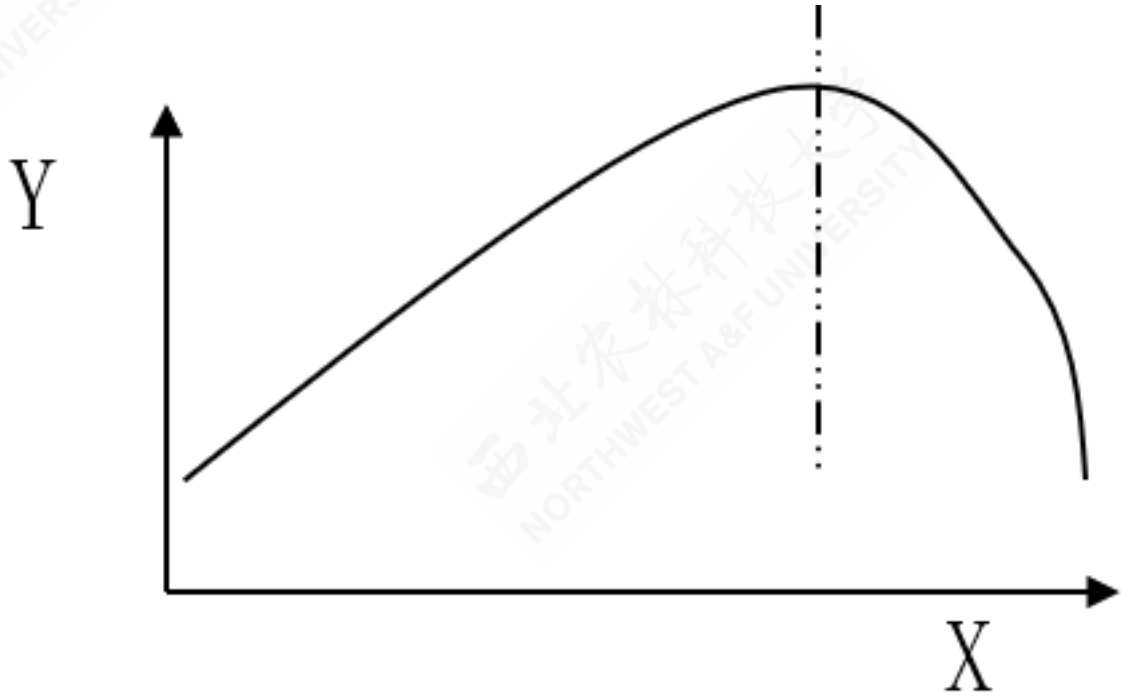
（示例）数据分布形态：常见形态

1)对称分布　2)右偏分布　3)左偏分布



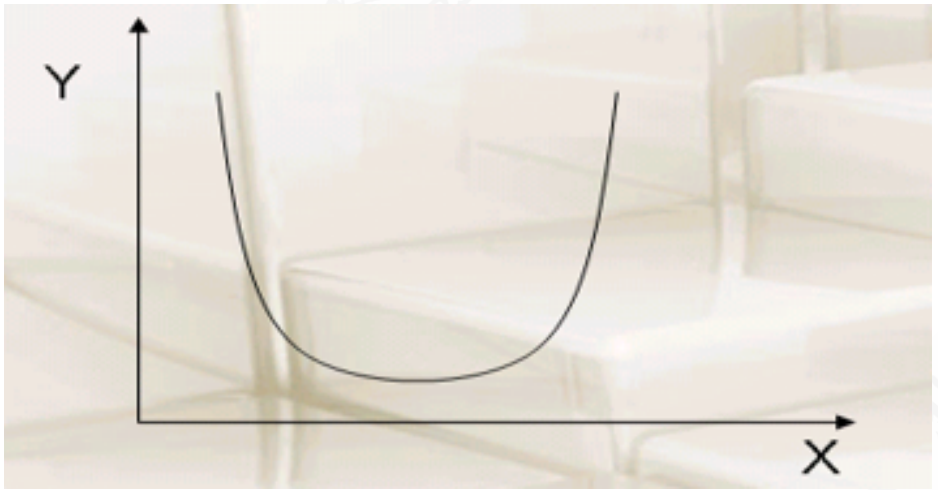
（示例）数据分布形态：常见形态

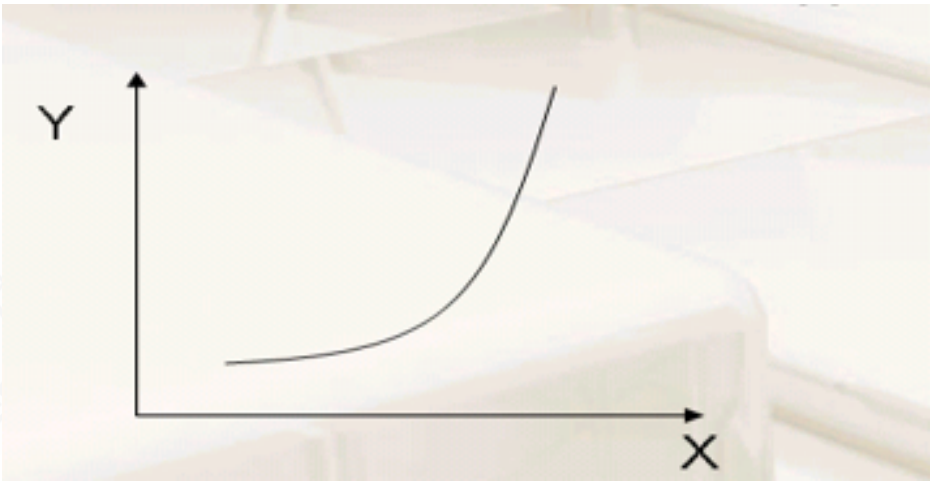
1)对称分布　　2)右偏分布　　3)左偏分布



（示例）数据分布形态：其他形态

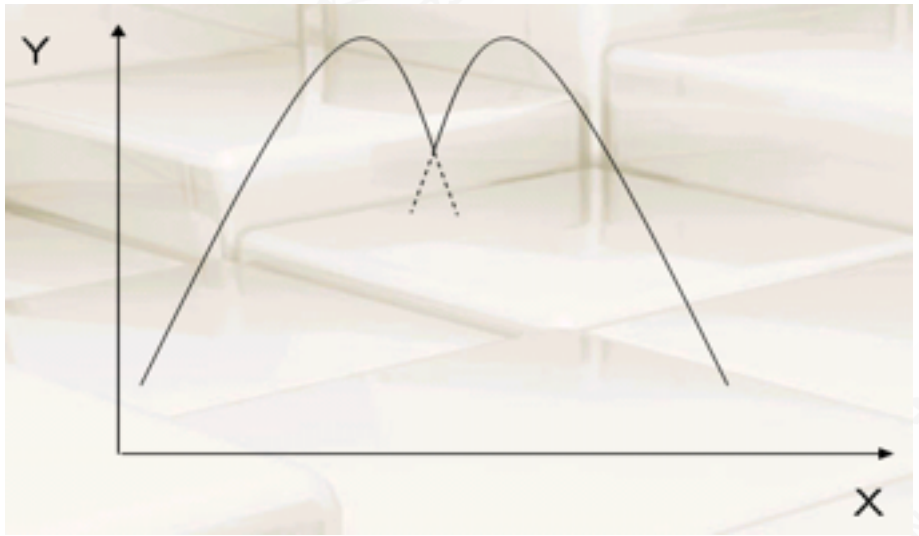
1)U型分布和J型分布　　2)M分布和反J分布

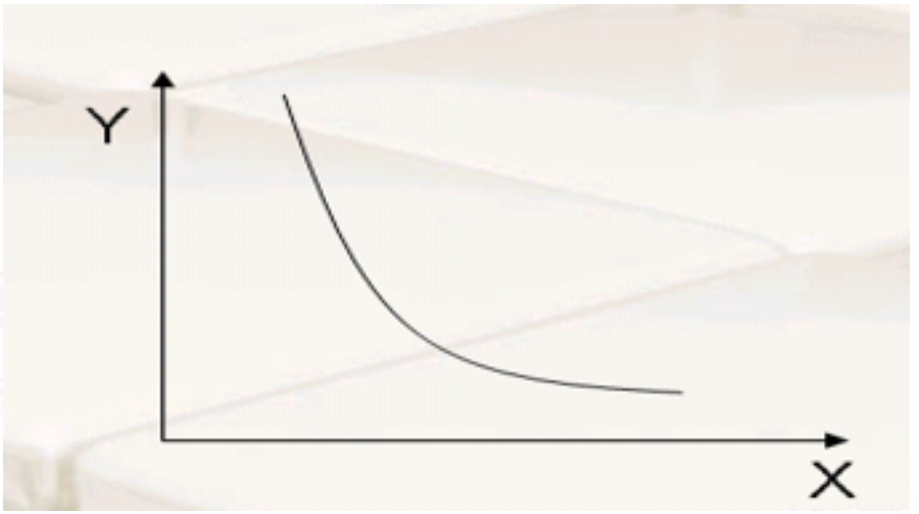




（示例）数据分布形态：其他形态

1)U型分布和J型分布　　2)M分布和反J分布





数据分布形态：峰度系数

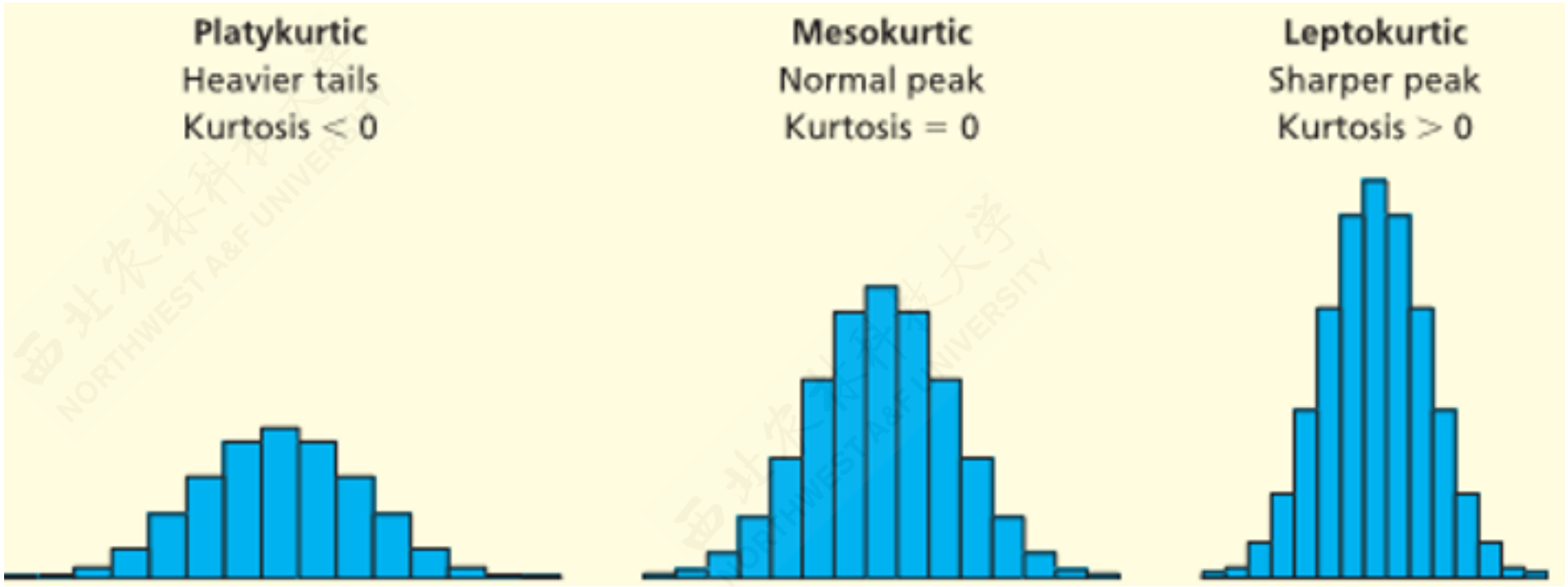
峰度（Kurtosis）系数：峰度刻画数据分布的拖尾长度和集中度。峰度系数 KT的理论计算公式如下：

$$KT = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{S_X^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

> 其中：n表示样本数；S_X表示样本标准差 $S_X = \sqrt{\frac{\sum_1^n (X_i - \bar{X})^2}{n-1}}$。

- 若峰度系数 $KT = 0$，则数据分布是**常峰态的**。
- 若峰度系数 $KT < 0$，则数据分布是**低峰态的**。
- 若峰度系数 $KT > 0$，则数据分布是**尖峰态的**。

（示例）数据分布形态：不同的峰态形状



统计制表：结构

统计表的结构与内容一般为：

从**形式**上看：统计表由总标题、横行标题、纵栏标题、指标数值构成。

从**内容**上看：统计表由主词和宾词两部分构成。

- 主词：说明总体或总体的分组。
- 宾词：用哪些指标数值来说明总体或总体的分组。

（示例）统计指标的形式和规范

某年某月某公司各企业劳动生产率统计表　　总标题

单位____

分组	总产值(万元)	职工人数(人)	劳动生产率(元/人)
P	1	2	3
大型			
中型			
小型			
合计			

纵栏标题

横行标题

数据资料（指标数值）

主词　　宾词

资料来源：《XXX统计摘要》

统计制表：特点

统计表的特点

- 开口式
- 上下有基线
- 编号：主词一般按A、B、C...，宾词按1、2、3...
- 有计量单位
- 表中不允许有空格：若不需要此资料则用“-”;暂缺某资料则用“.....”

统计制表：设计准则

统计表的一般设计准则包括：

- 合理安排统计表的结构
- 总标题内容应满足3W要求
- 数据计量单位相同时，可放在表的右上角标明，不同时应放在每个变量后或单列出一列标明
- 表中的上下两条横线一般用粗线，其他线用细线
- 通常情况下，统计表的左右两边不封口
- 表中的数据一般是右对齐，有小数点时应以小数点对齐，而且小数点的位数应统一
- 对于没有数字的表格单元，一般用“—”表示
- 必要时可在表的下方加上注释

3.4 合理使用图表

图表体系和要素

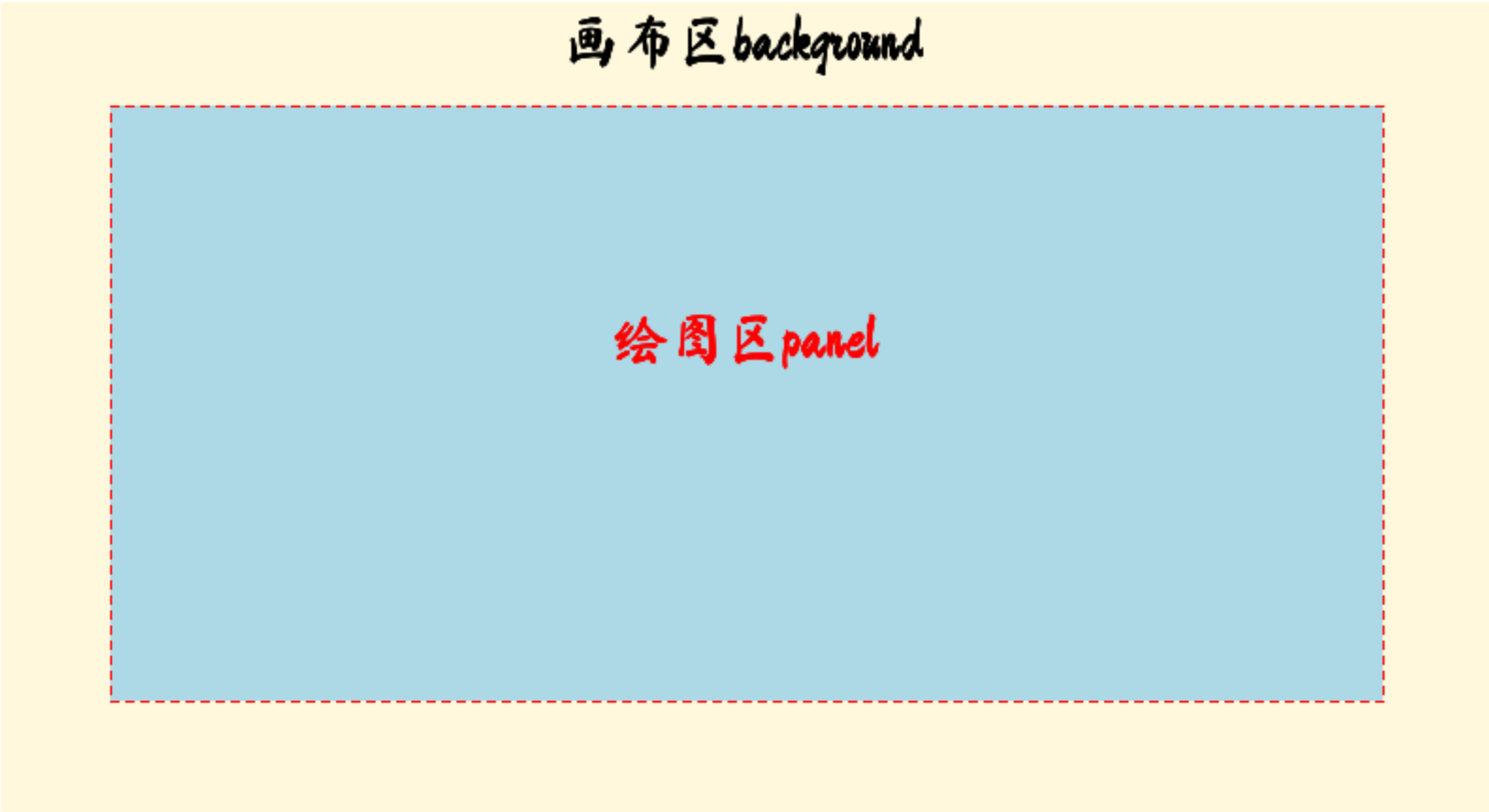
鉴别图形优劣的准则

制图体系：绘图区 panel

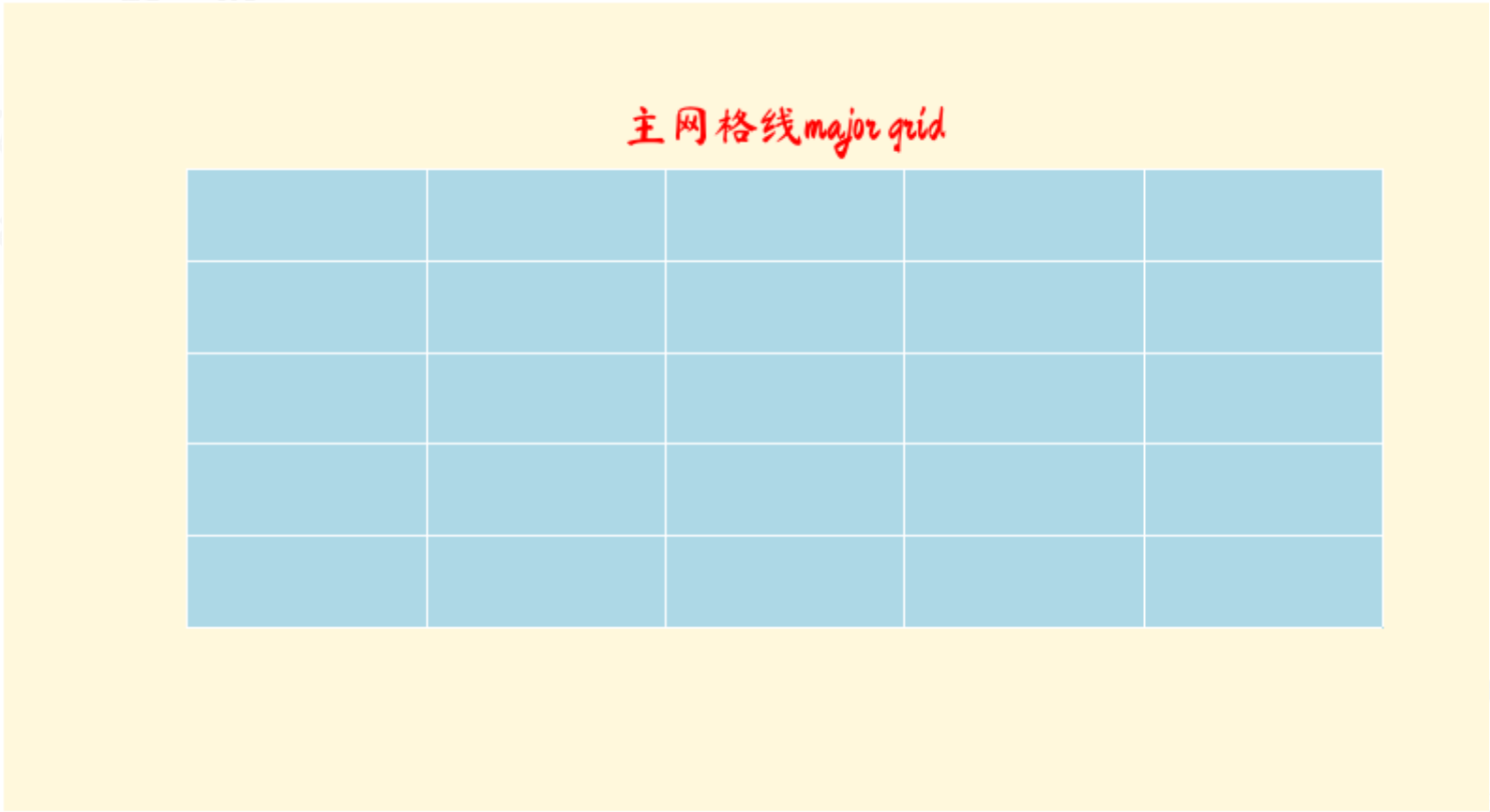
a.画布区　b.绘图区　c.主网格　d.次网格

制图体系：绘图区 panel

a.画布区　b.绘图区　c.主网格　d.次网格



a.画布区　b.绘图区　c.主网格　d.次网格



a.画布区　　b.绘图区　　c.主网格　　d.次网格

次网格线 minor grid

制图体系：坐标轴 | axis

a.下横轴X　b.上横轴x　c.左纵轴Y　d.右纵轴Y

制图体系：坐标轴 axis

a.下横轴X　　b.上横轴x　　c.左纵轴Y　　d.右纵轴Y

制图体系：坐标轴 axis

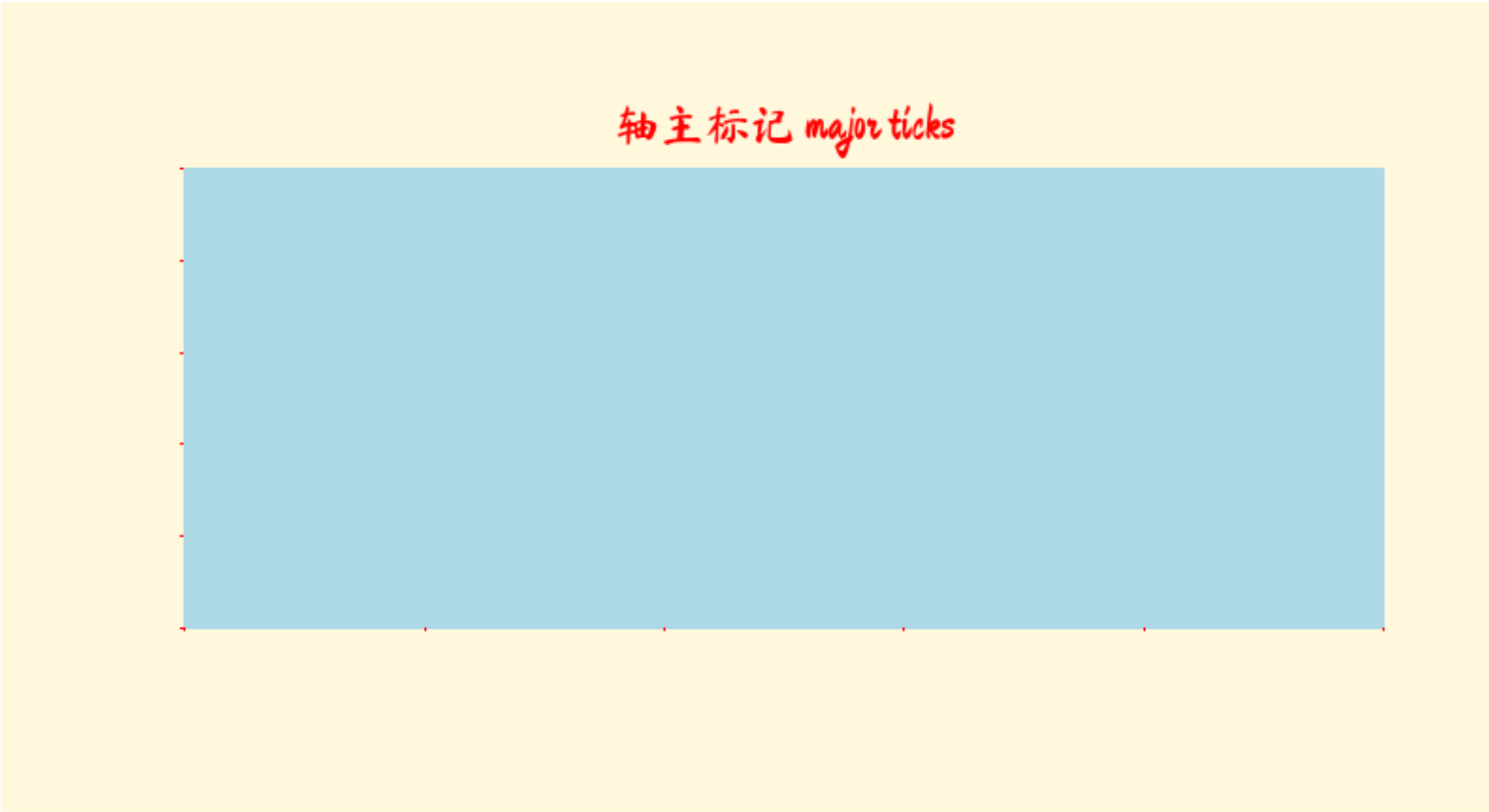
a.下横轴X　b.上横轴x　**c.左纵轴Y**　d.右纵轴Y

制图体系：坐标轴 axis

a.下横轴X　　b.上横轴x　　c.左纵轴Y　　**d.右纵轴Y**

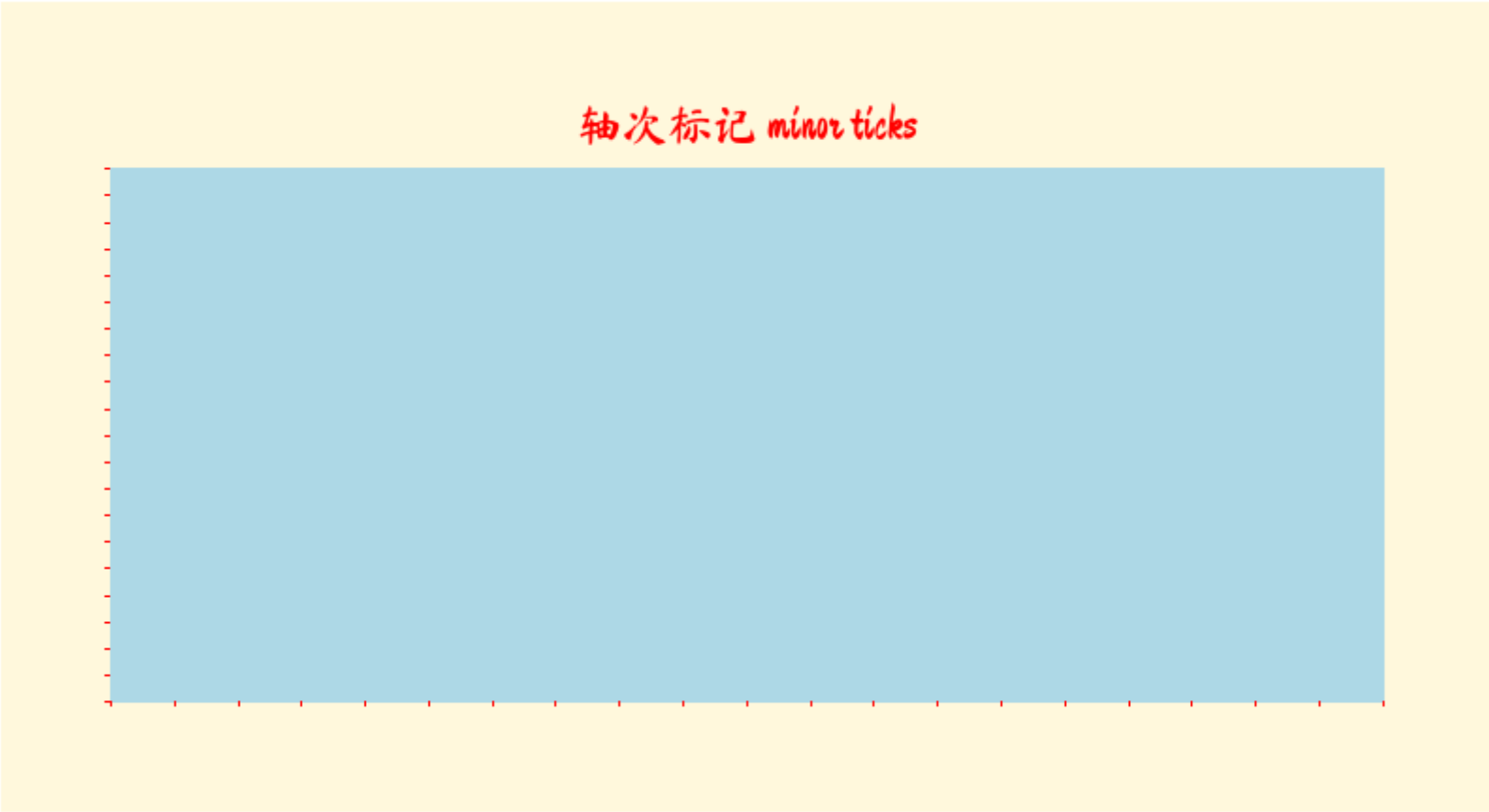
制图体系：坐标轴2 axis

d.轴主标记　e.轴次标记　f.轴标签值1　g.轴标签值2



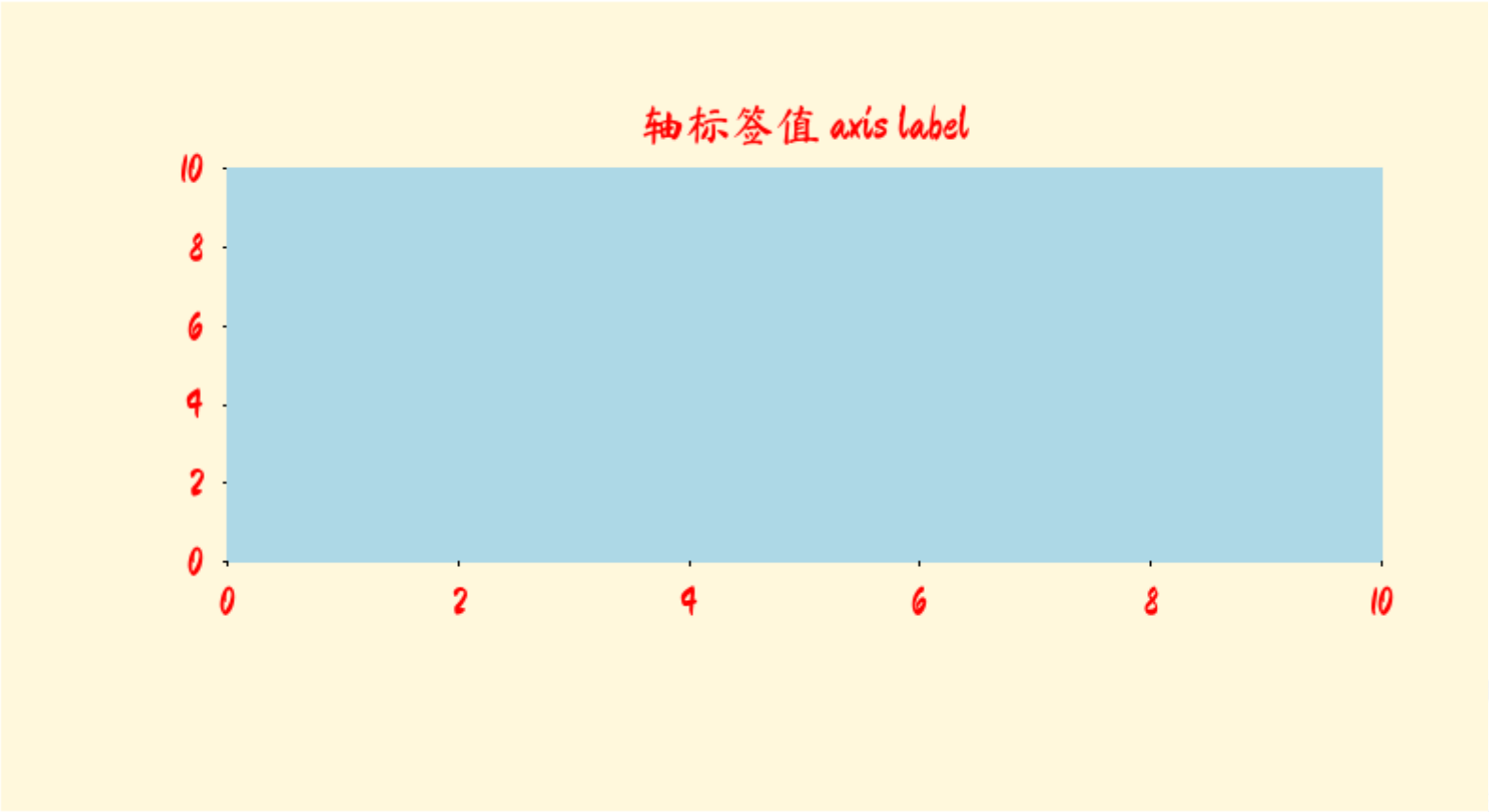
制图体系：坐标轴2 axis

d.轴主标记　e.轴次标记　f.轴标签值1　g.轴标签值2



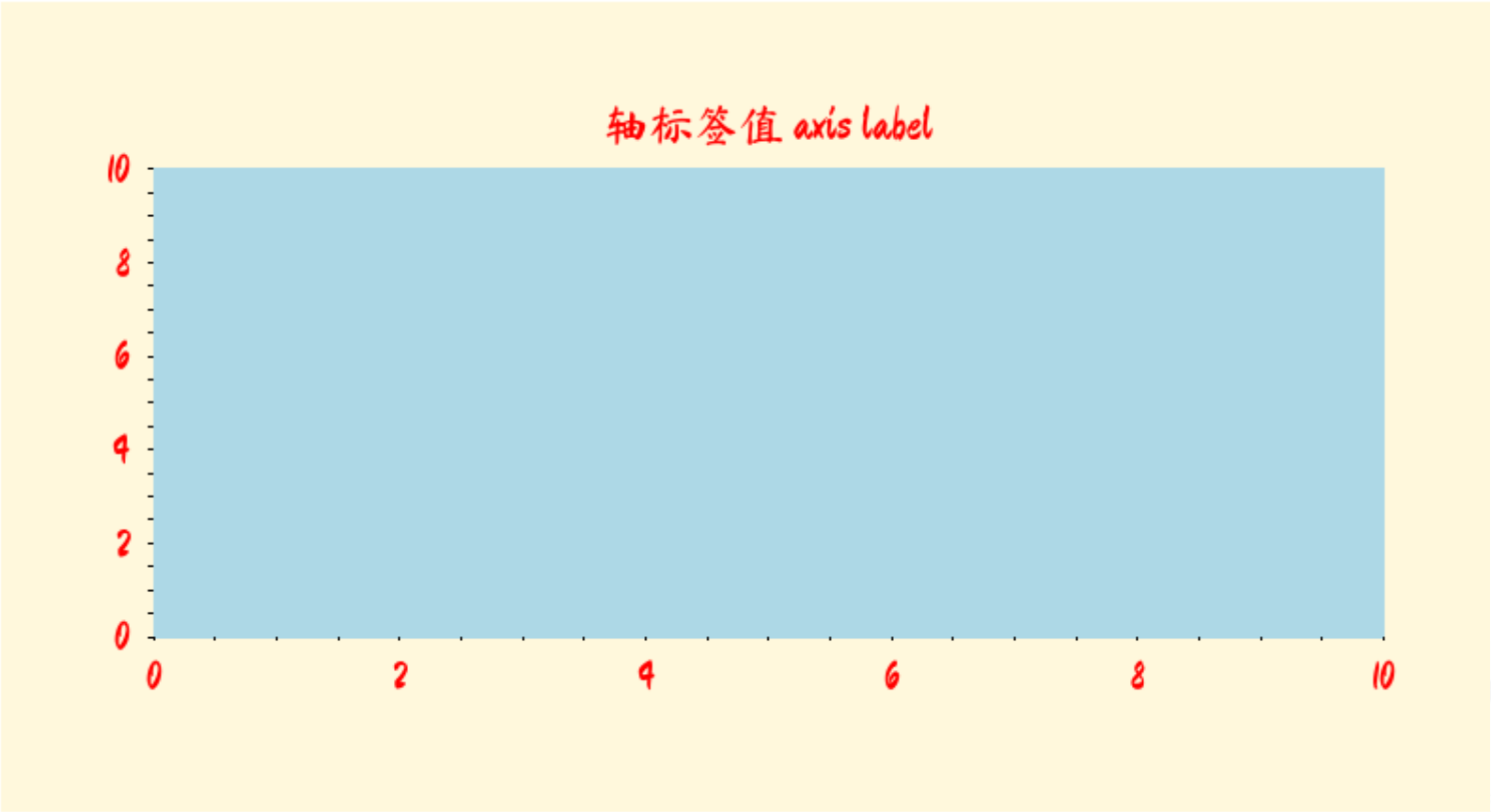
制图体系：坐标轴2 axis

d.轴主标记　e.轴次标记　f.轴标签值1　g.轴标签值2

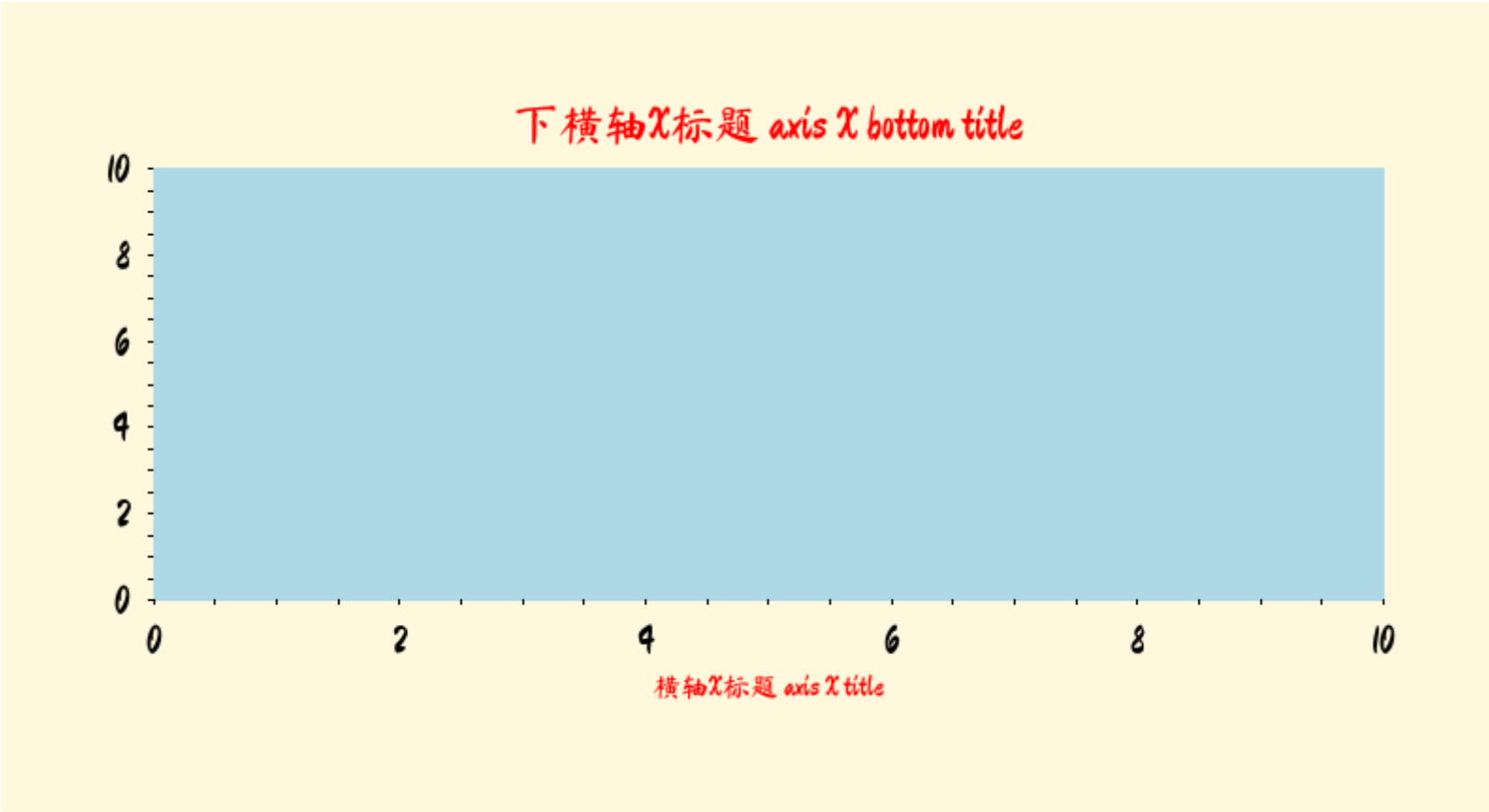


制图体系：坐标轴2 axis

d.轴主标记　e.轴次标记　f.轴标签值1　**g.轴标签值2**

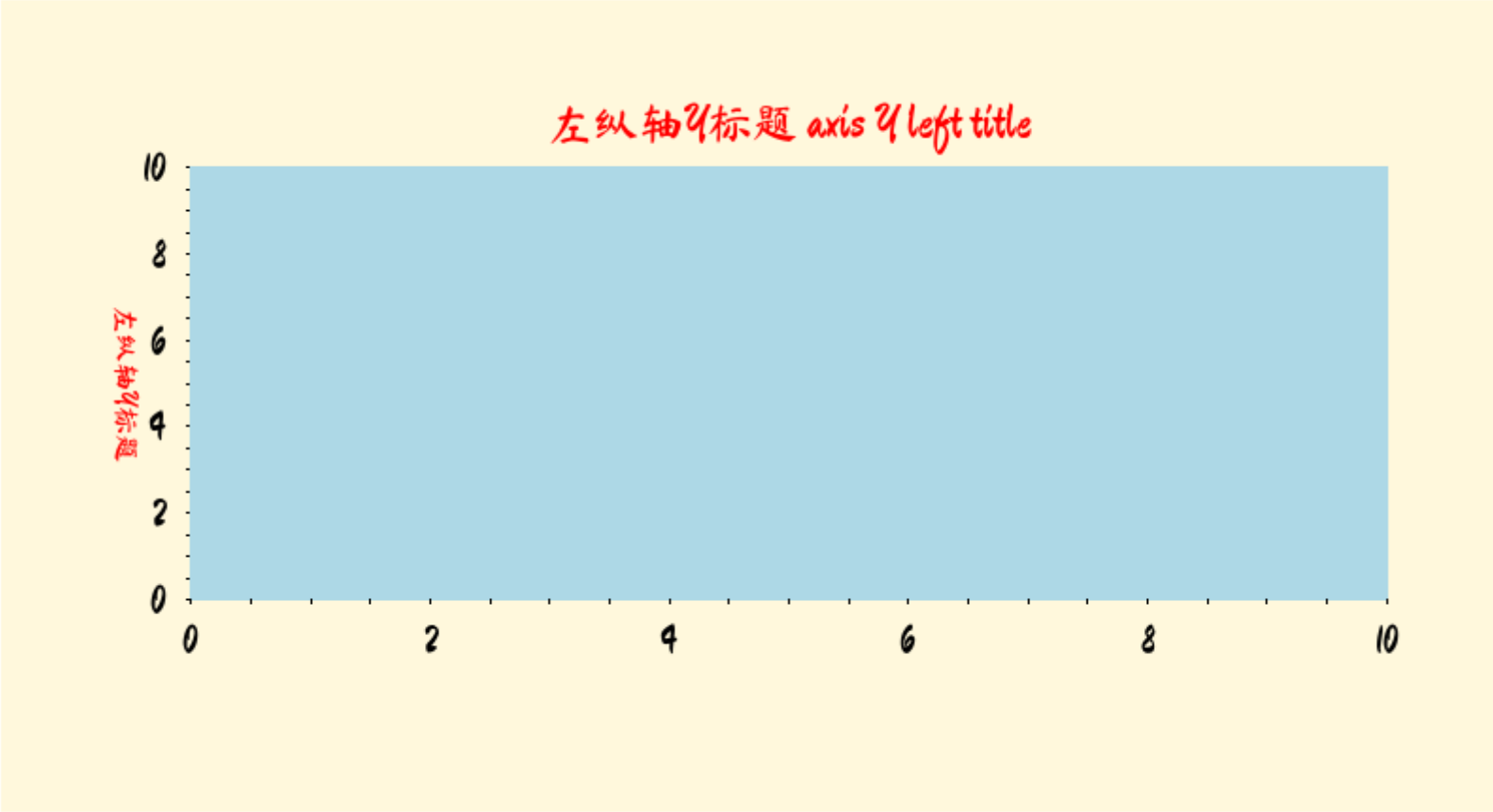


h.横标题　h.左纵轴标题　h.右纵轴标题

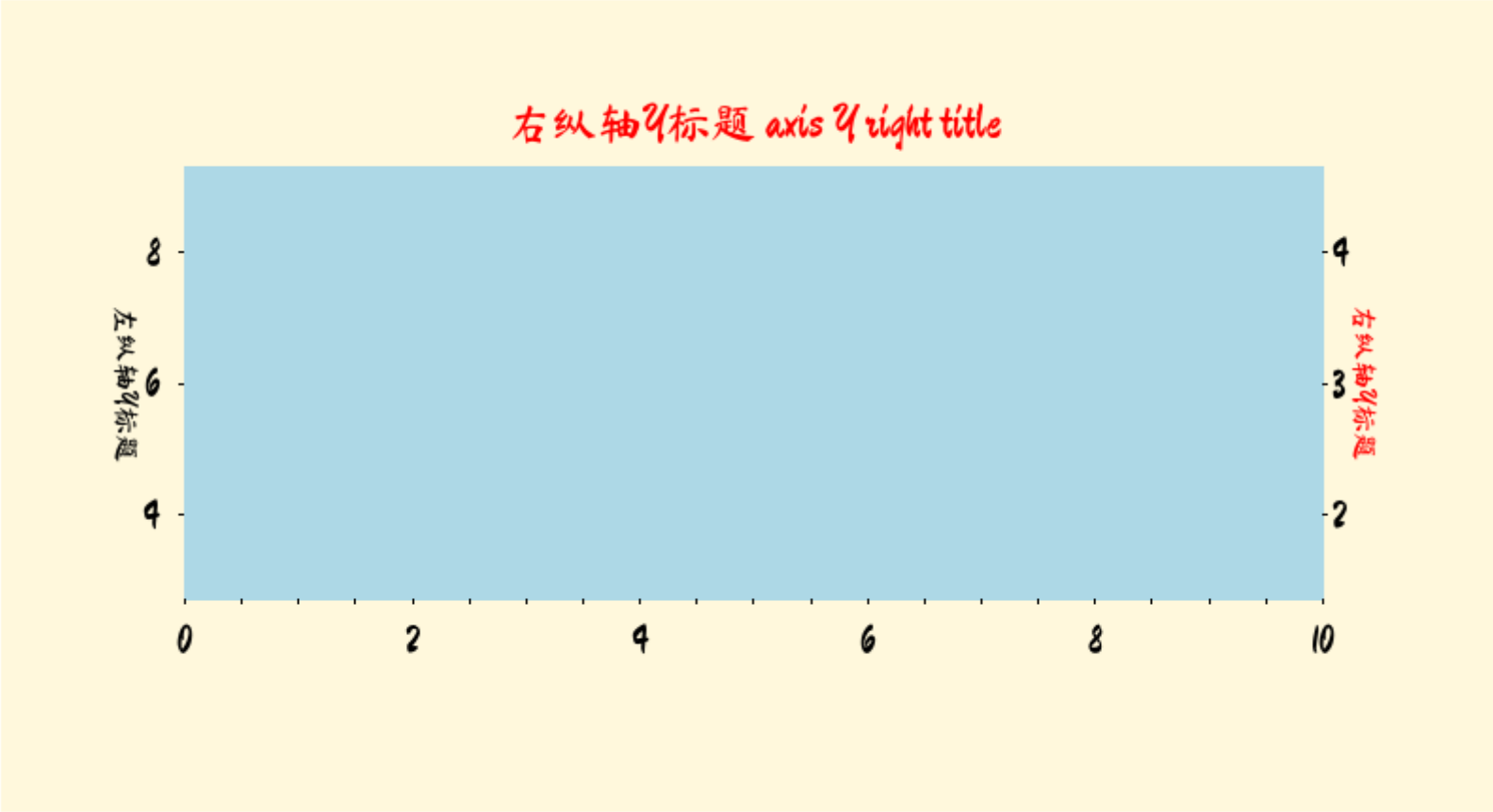


制图体系：坐标轴3 axis

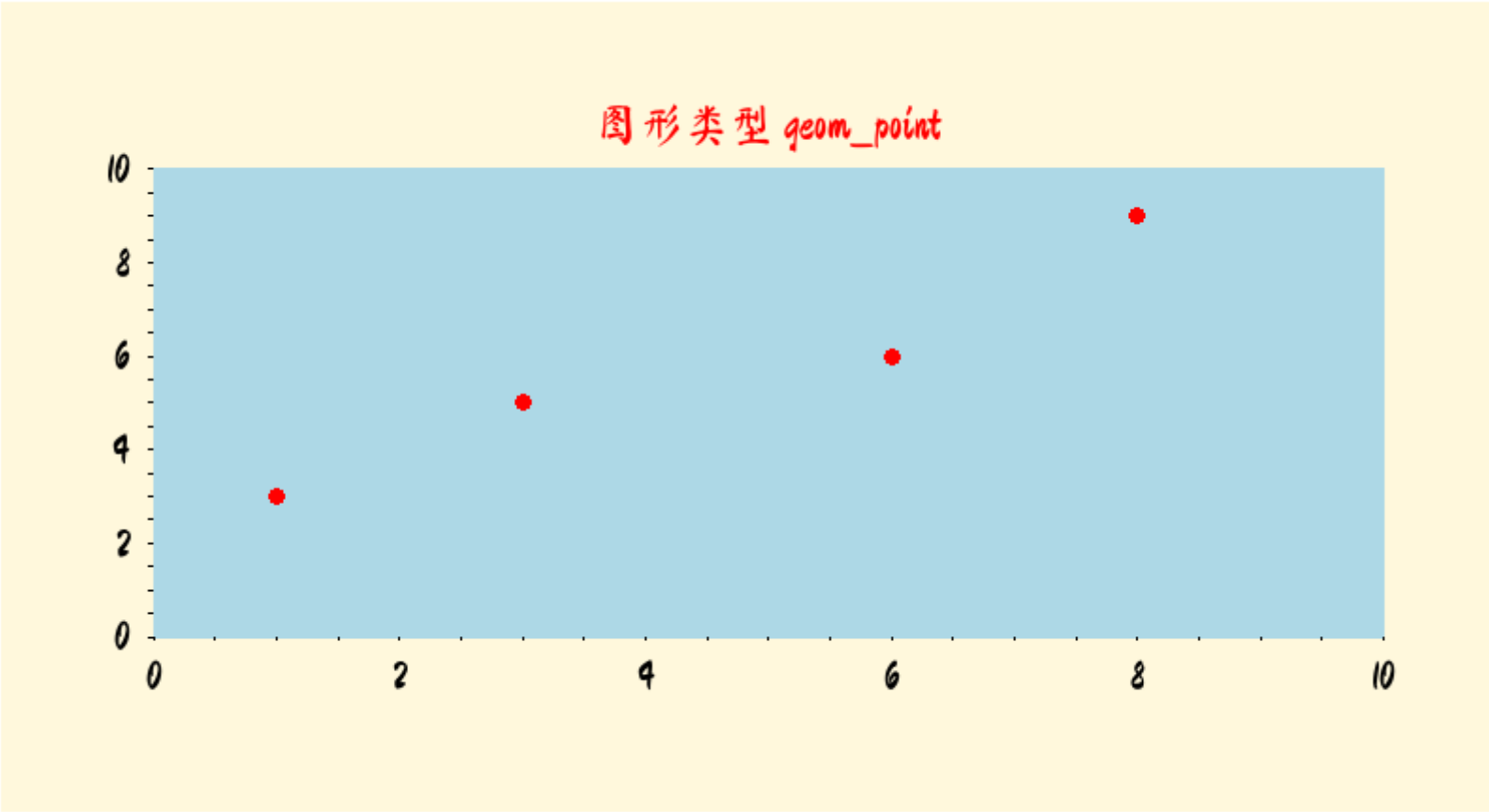
h.横标题　　h.左纵轴标题　　h.右纵轴标题



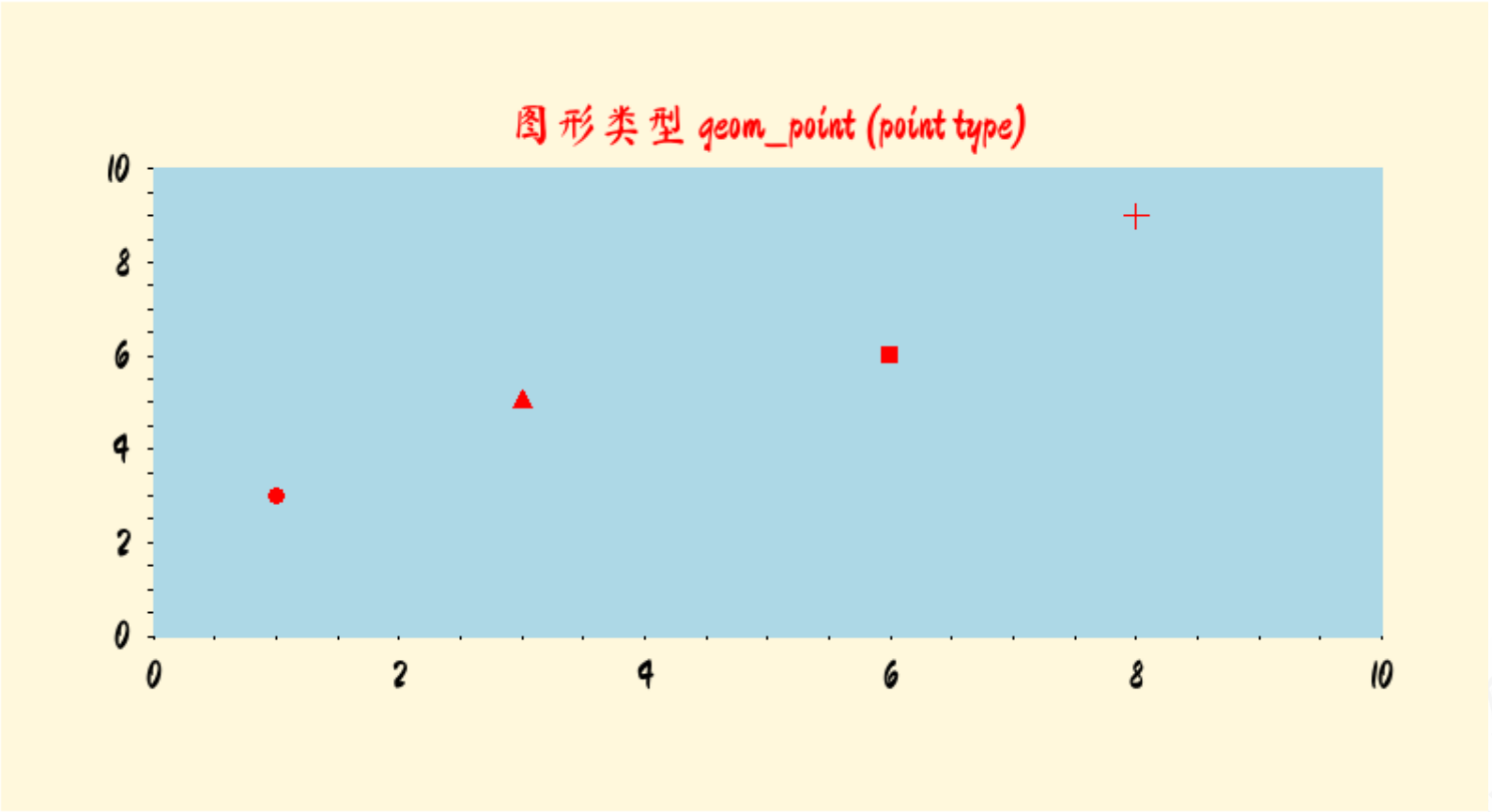
h.横标题　　h.左纵轴标题　　h.右纵轴标题



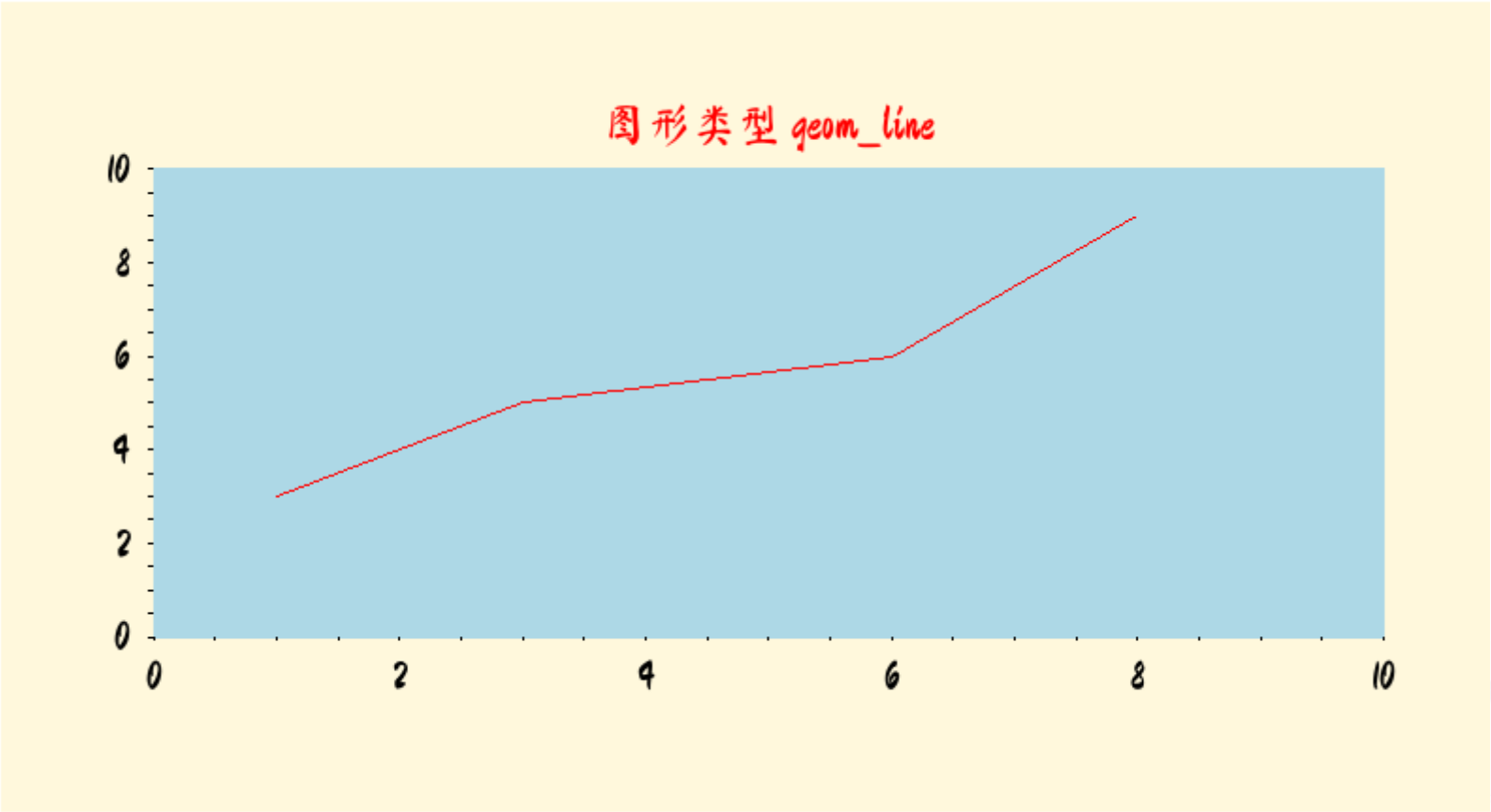
a.点图1　b.点图2　c.线图1　d.线图2



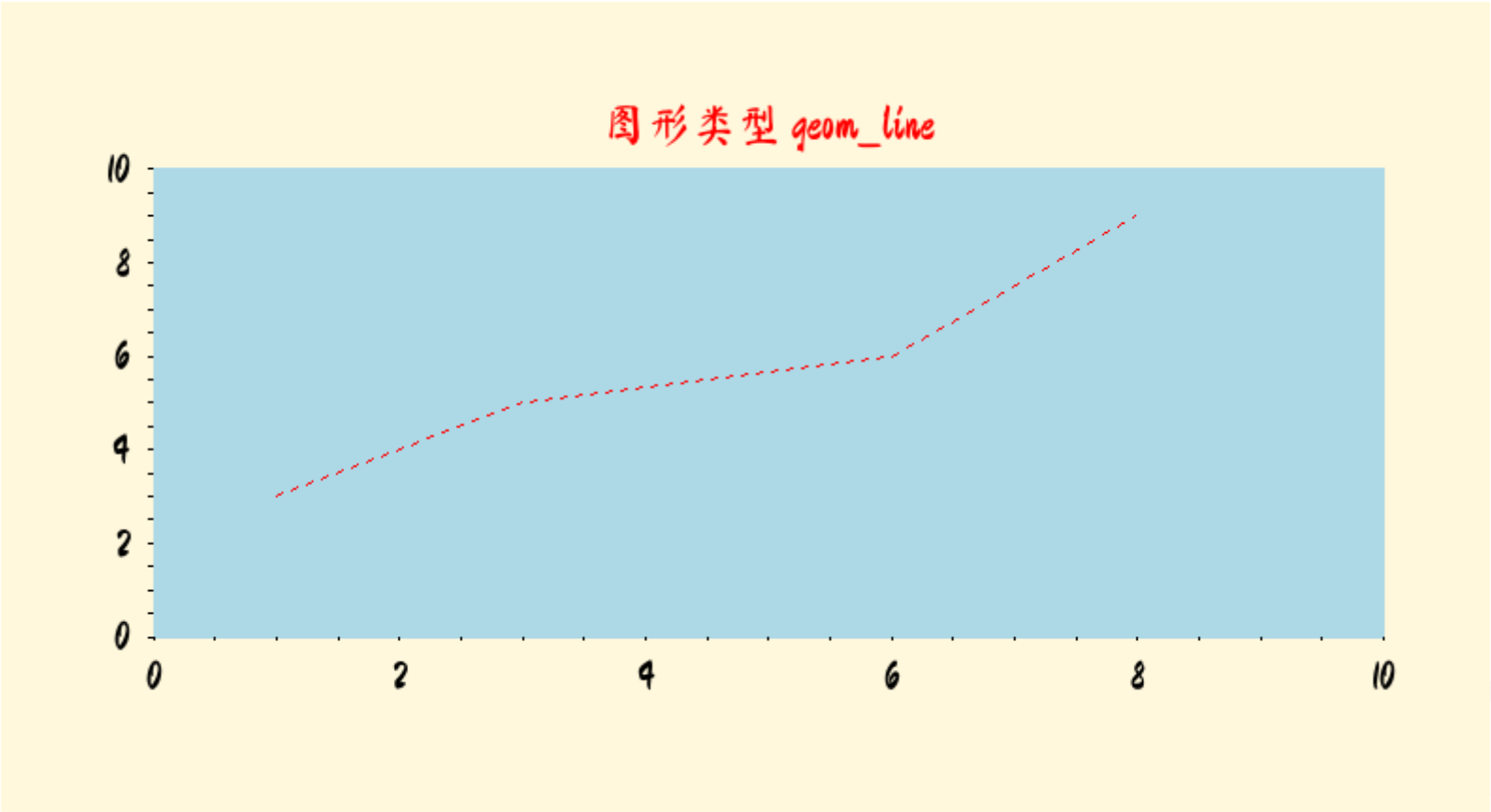
a.点图1　**b.点图2**　c.线图1　d.线图2



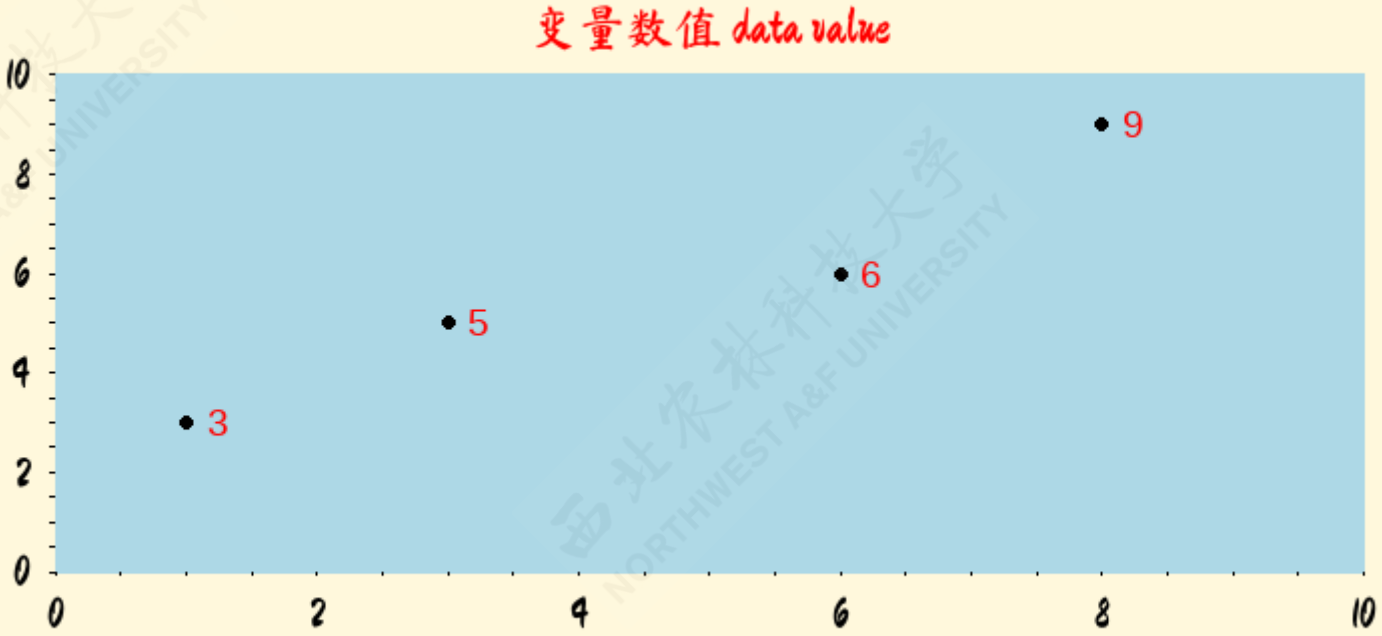
a.点图1　b.点图2　**c.线图1**　d.线图2



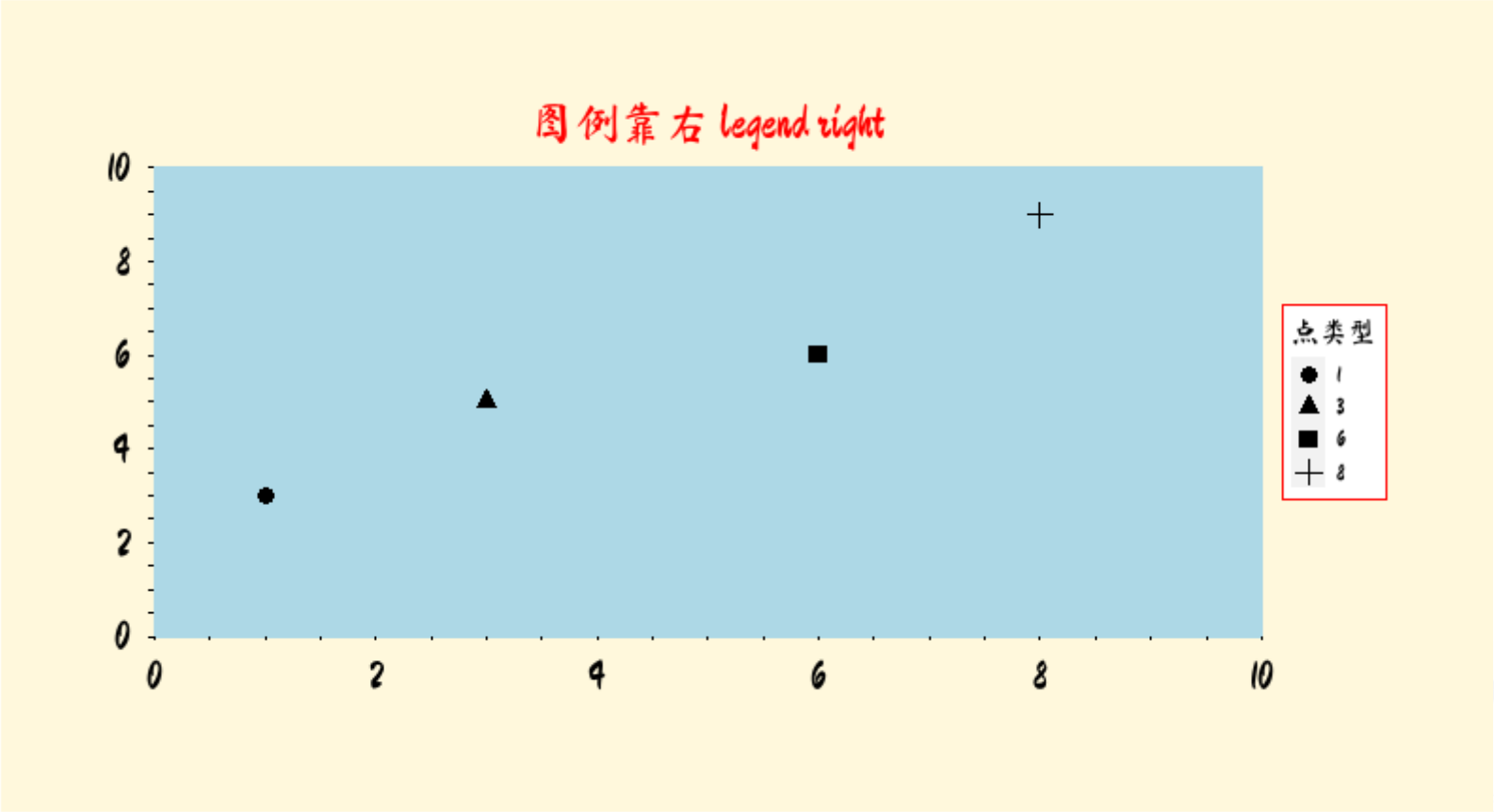
a.点图1　b.点图2　c.线图1　**d.线图2**



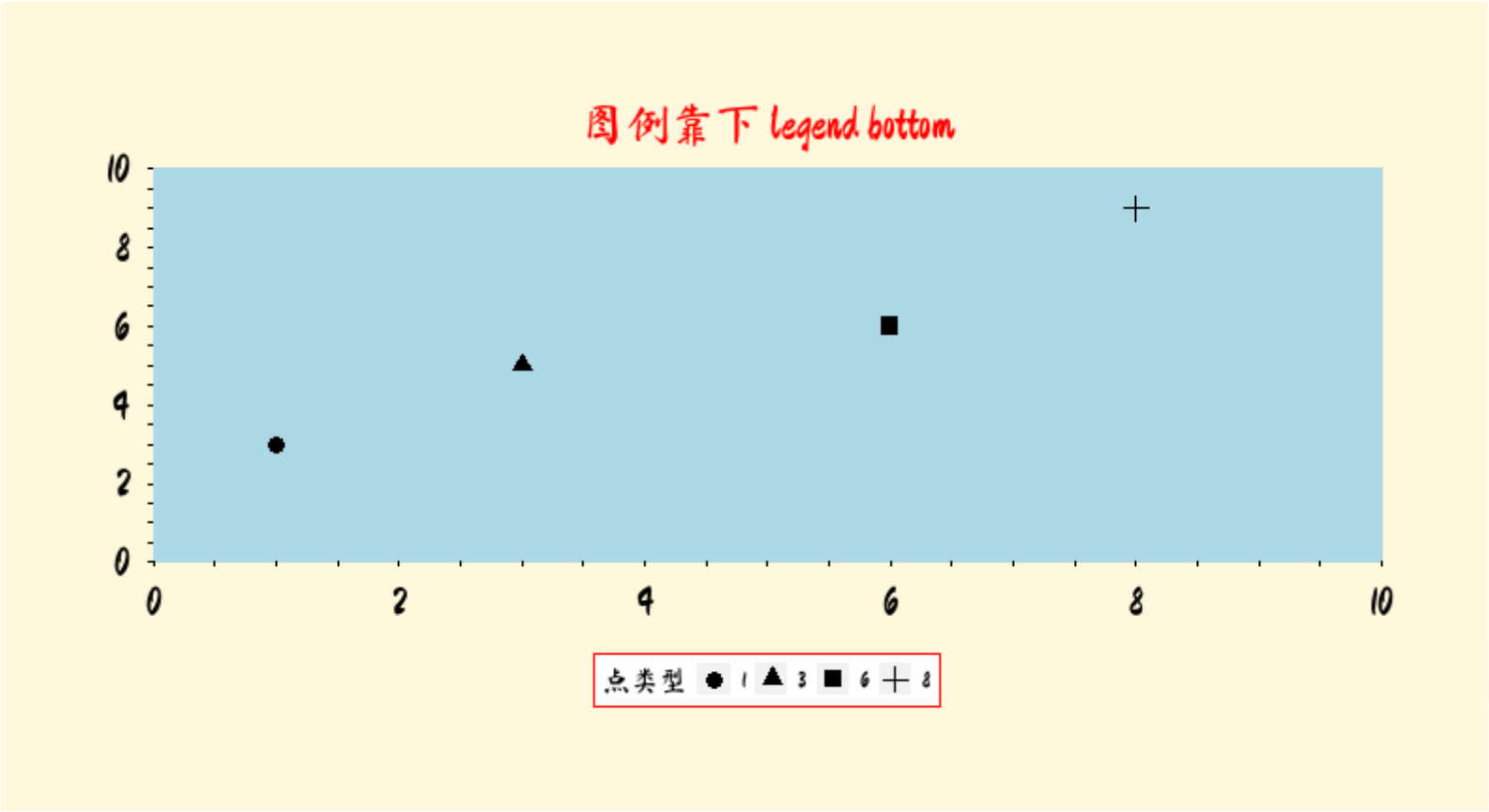
制图体系：变量数值 data value



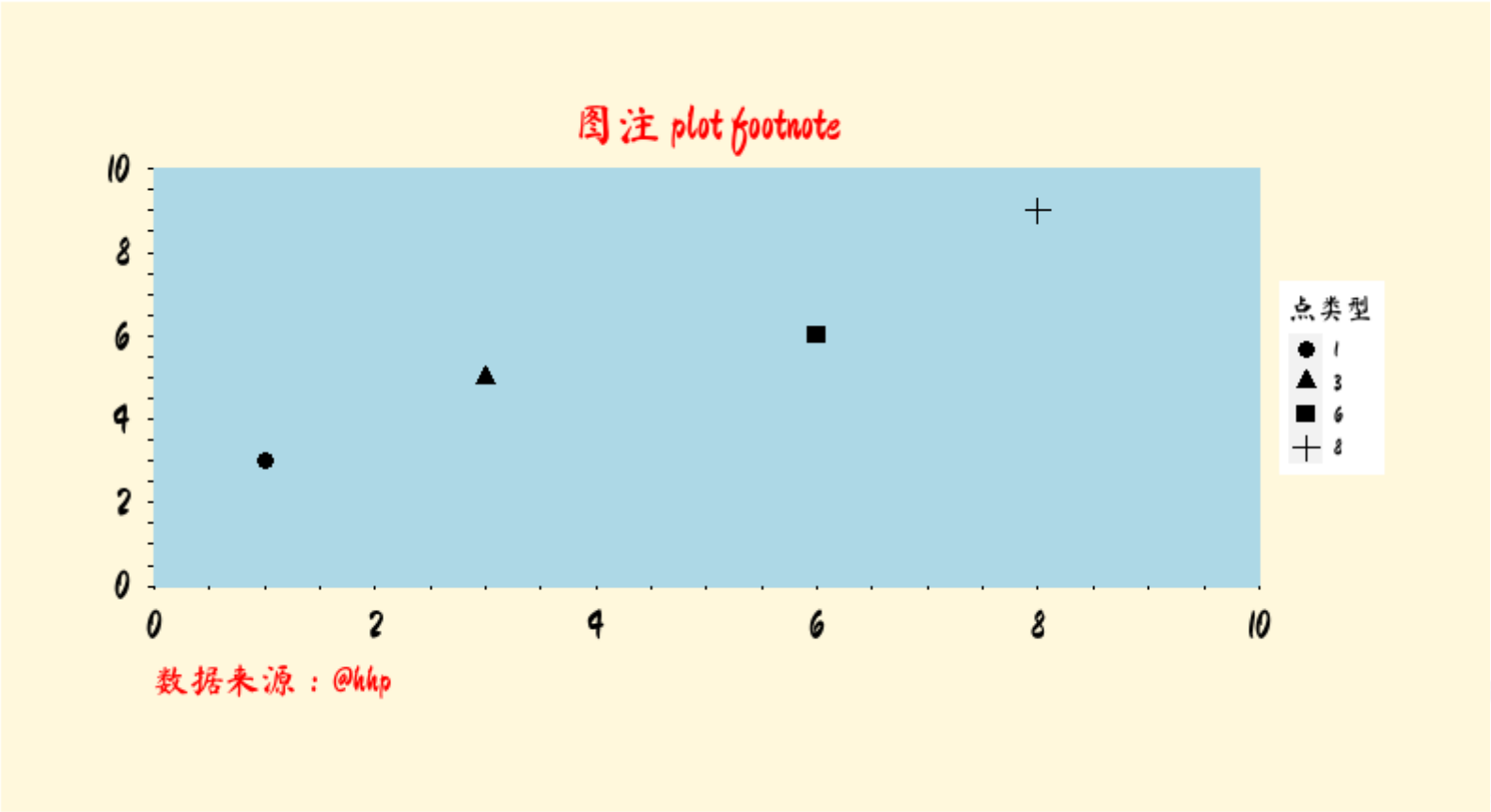
a.图例靠右　　b.图例靠下



a.图例靠右　　b.图例靠下



a.图注　b.图题



a.图注　b.图题

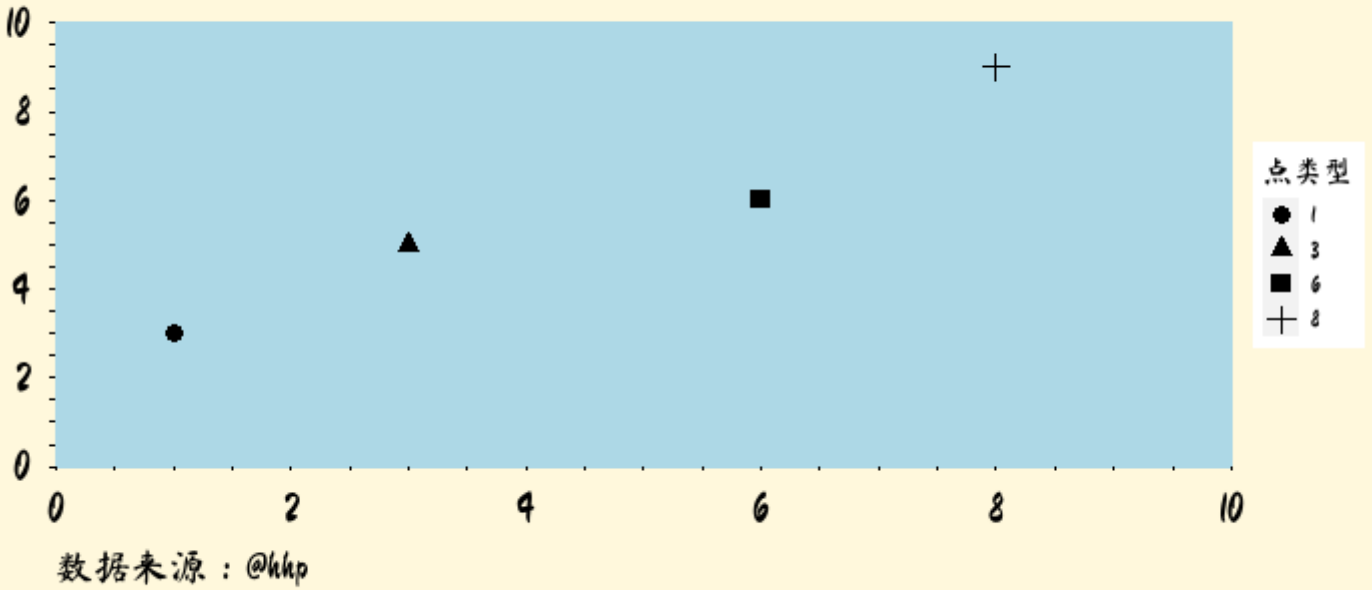


图3-1:plot caption演示

表格要素包括：表序号（numbering）、表题（title）、表头（header）、主体（body）和表注（footer）等部分构成。

表3-2：iris dataset数据集

Sepal.Length[a]	Sepal.Width[b]	Petal.Length[c]	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

[a]This is footnote one; [b]This is footnote two; [c]This is footnote three

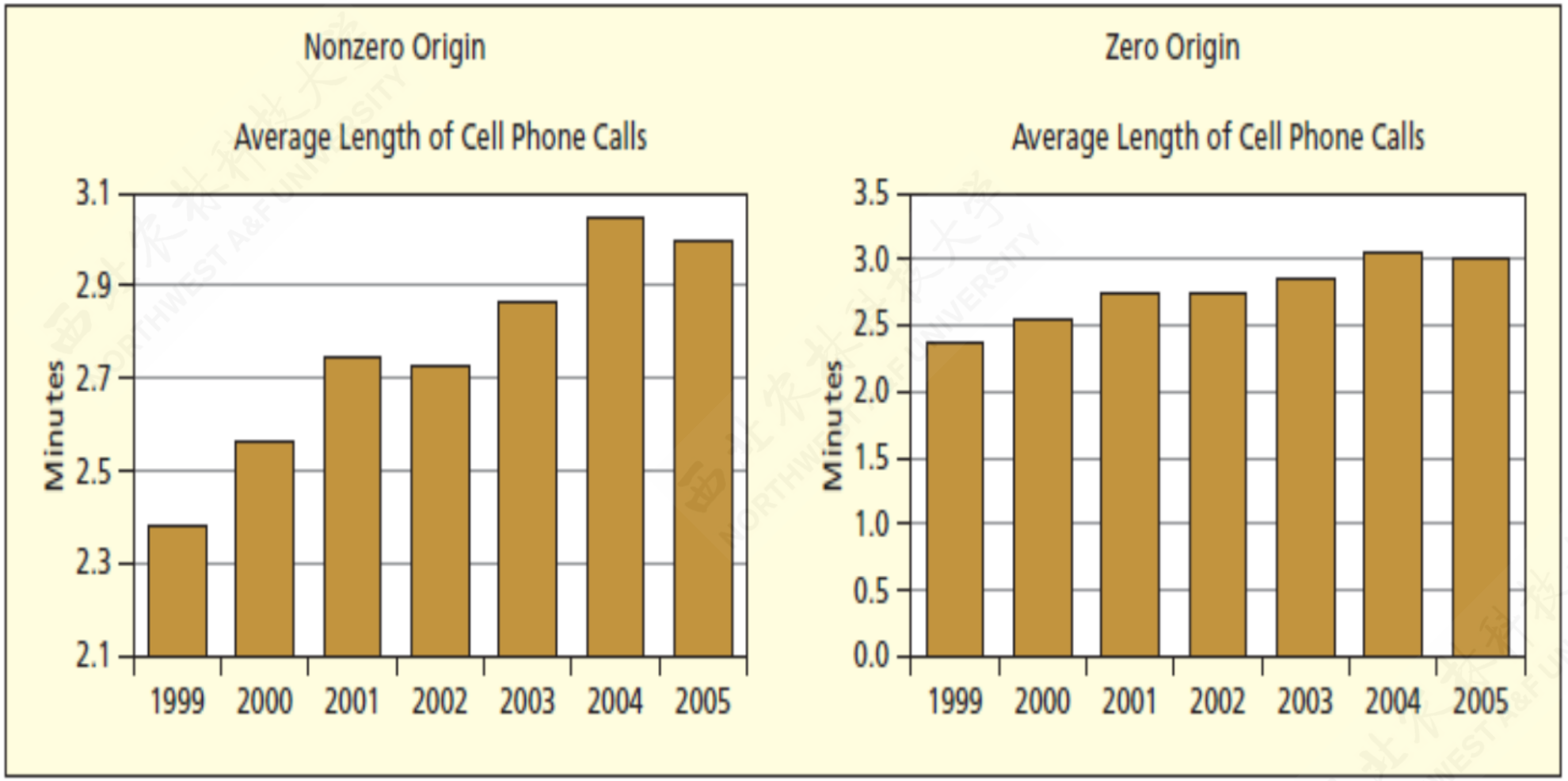
良好图表应具备的基本特征

- 服务于一个明确的目的
- 显示数据
- 强调数据之间的比较
- 有对图表的统计描述和文字说明
- 让读者把注意力集中在图表的内容上，而不是制作图表的程序上
- 避免歪曲

鉴别图表优劣的准则

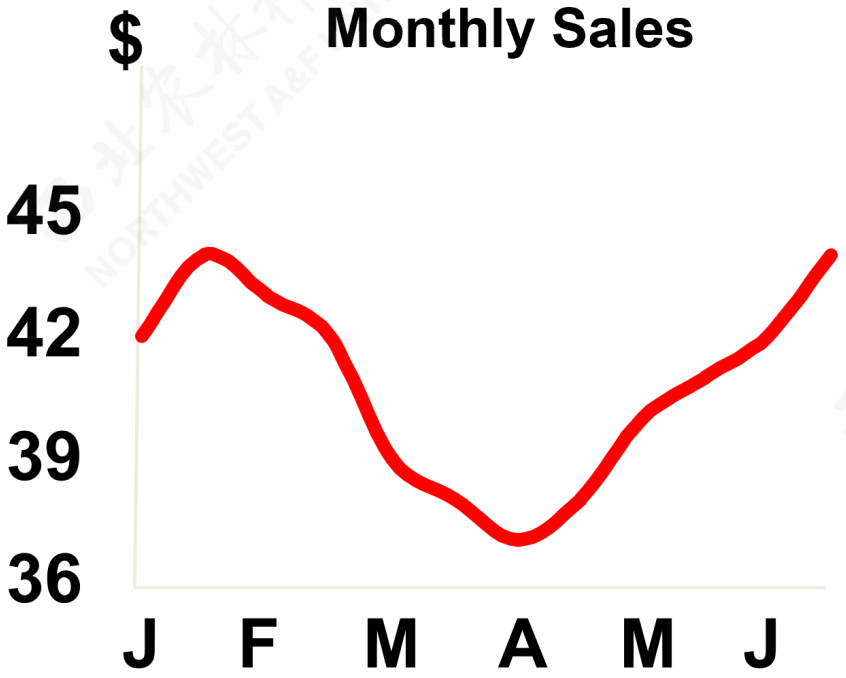
- 表述数据的真实情况
- 使复杂的观点得到简明、确切、高效的阐述
- 精心设计、有助于洞察问题的实质
- 能在最短的时间内以最少的笔墨给读者提供最大量的信息
- 多维度地对问题进行客观反映

制图常见误区：非零起始点！

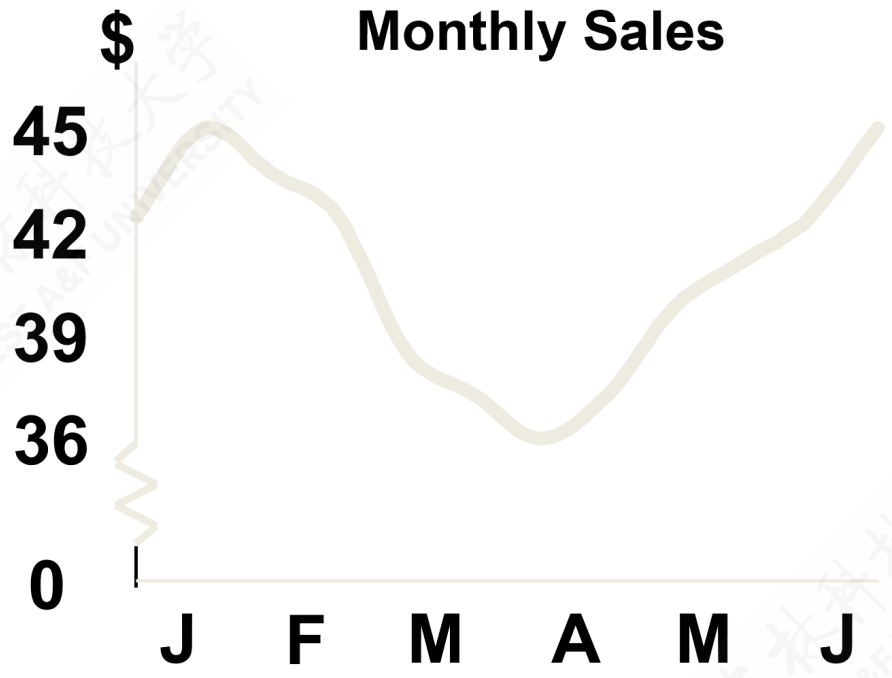


制图常见误区：非零起始点?

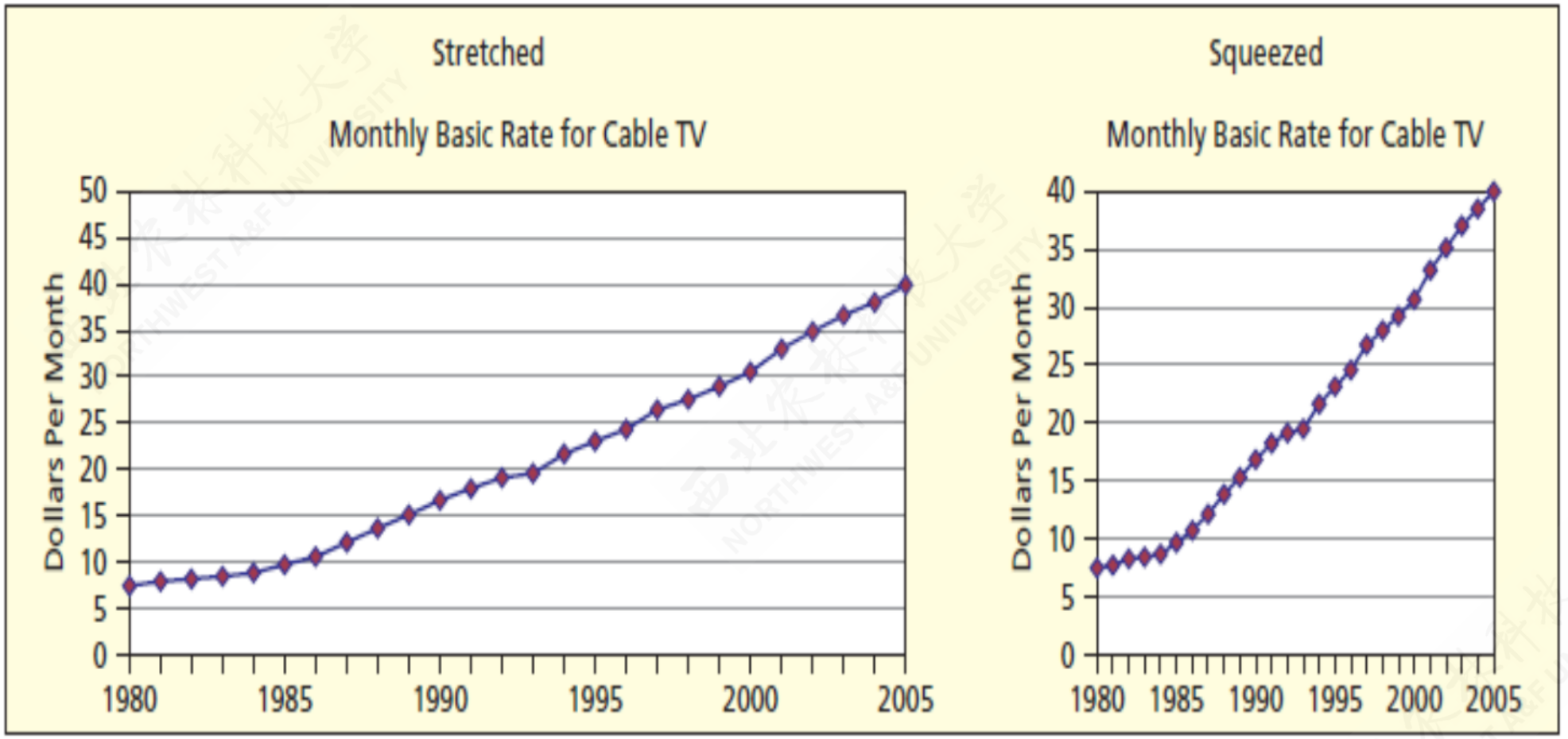
Bad Presentation



Good Presentations



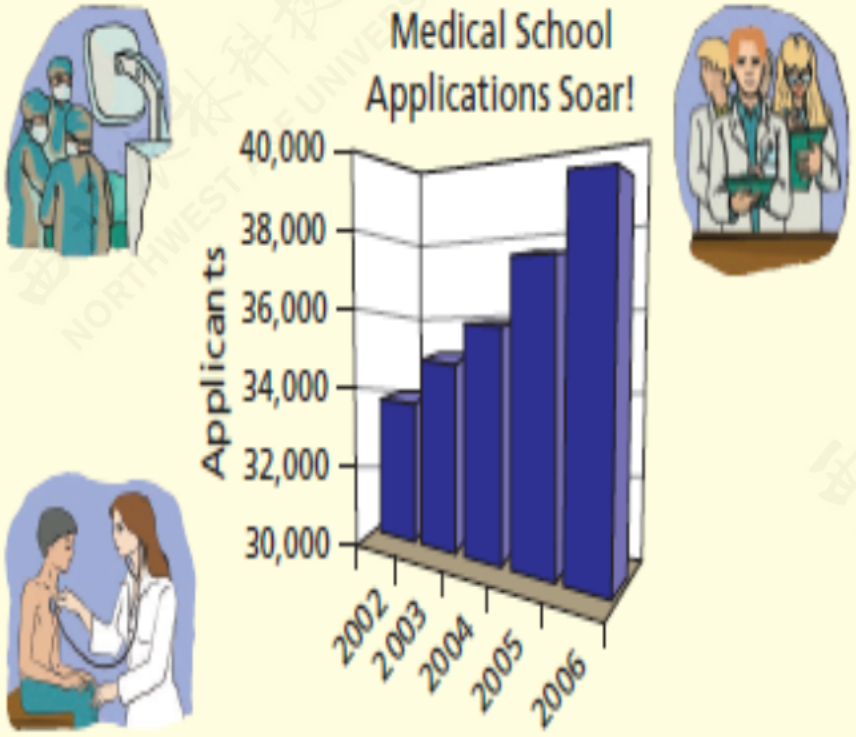
制图常见误区：图片比例拉伸失调



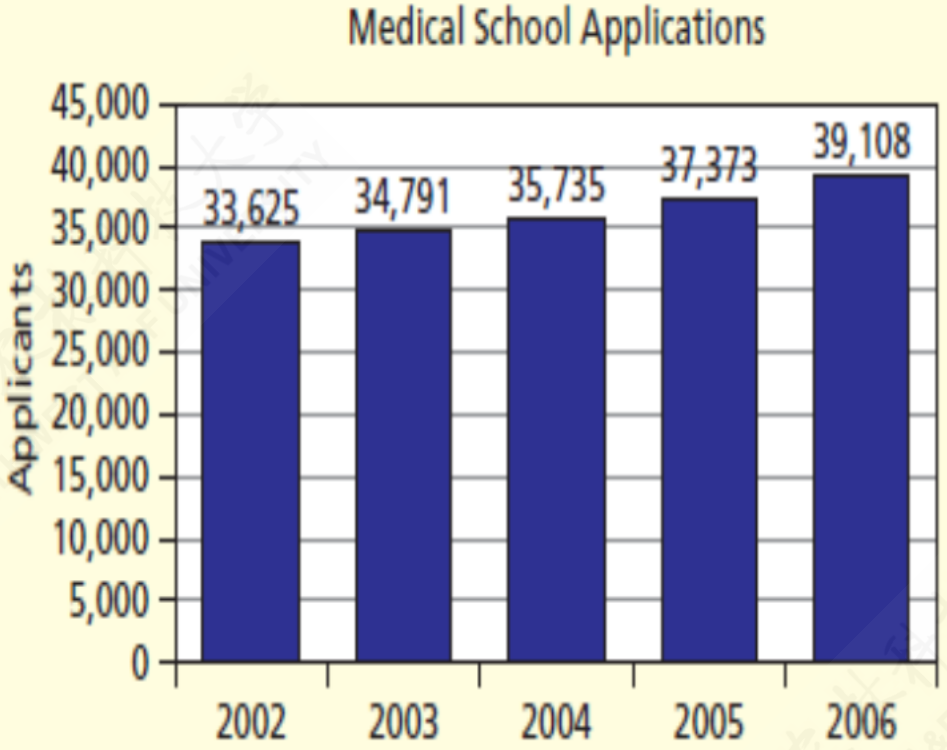
Source: *Statistical Abstract of the United States, 2007*, p. 717.

制图常见误区：炫技乱人心

Deceptive

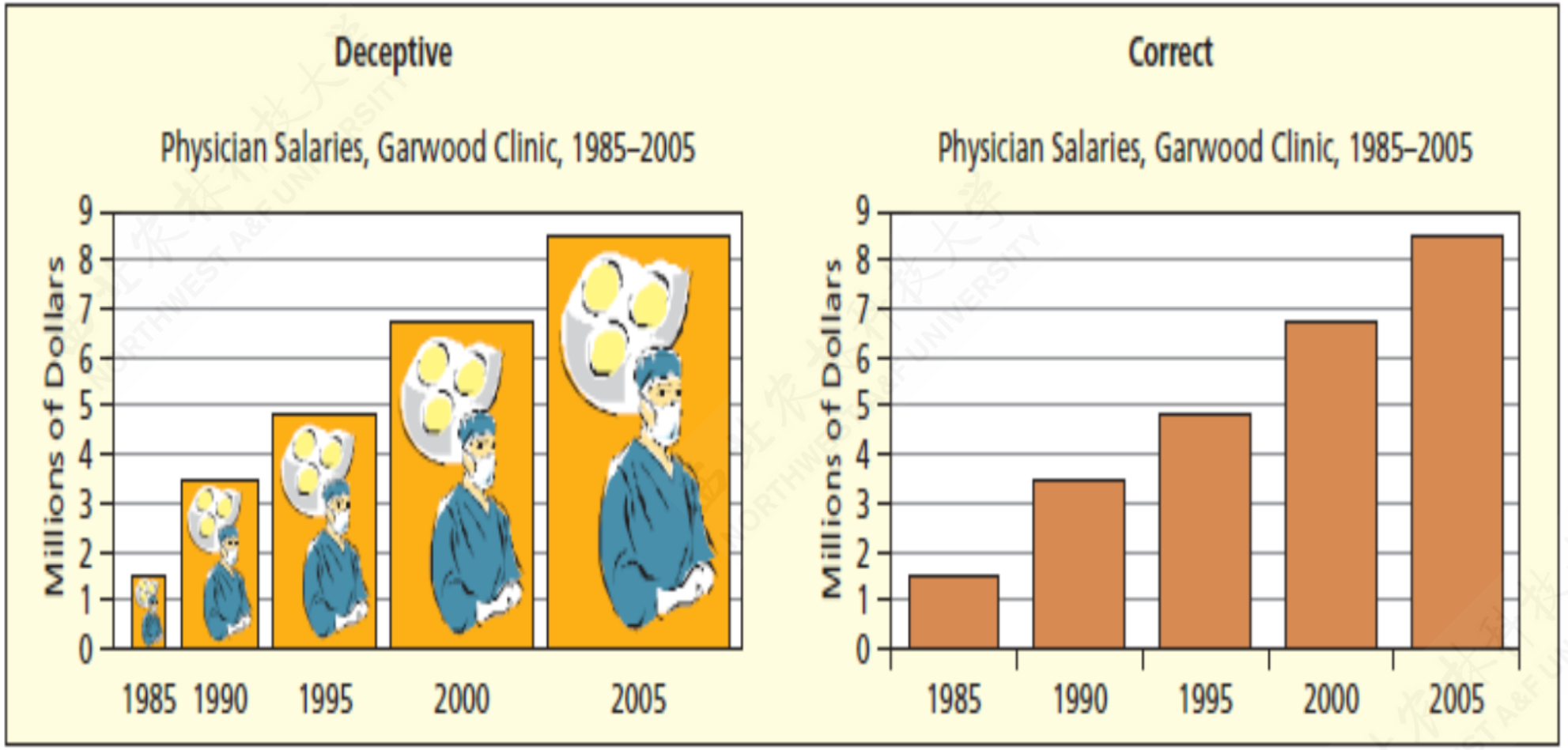


Correct



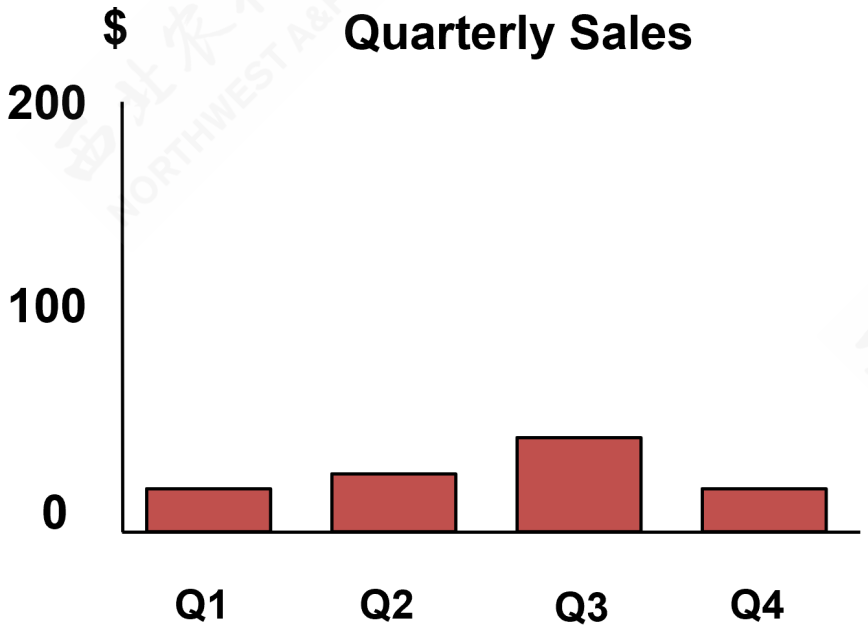
Source: www.aamc.org.

制图常见误区：视觉误导1

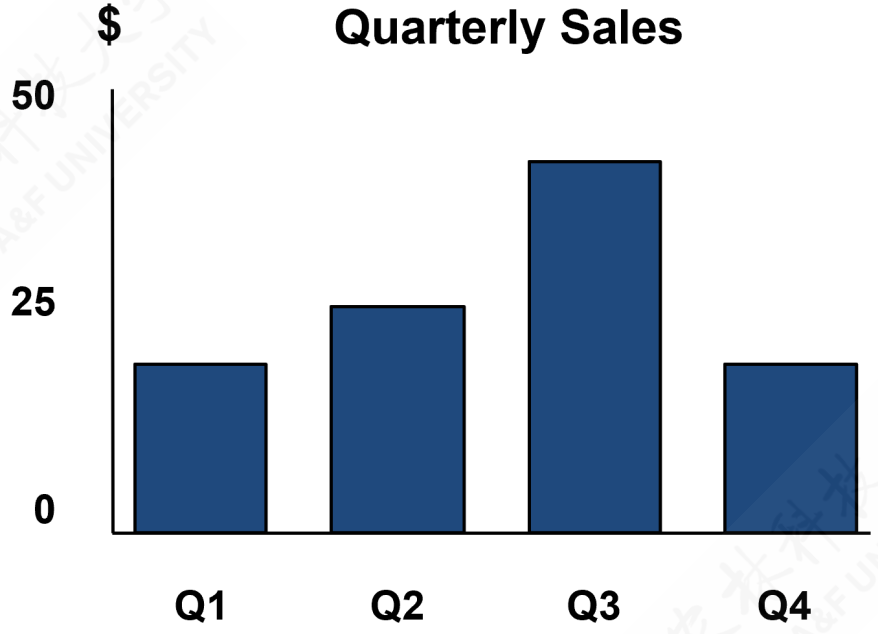


制图常见误区：视觉误导2

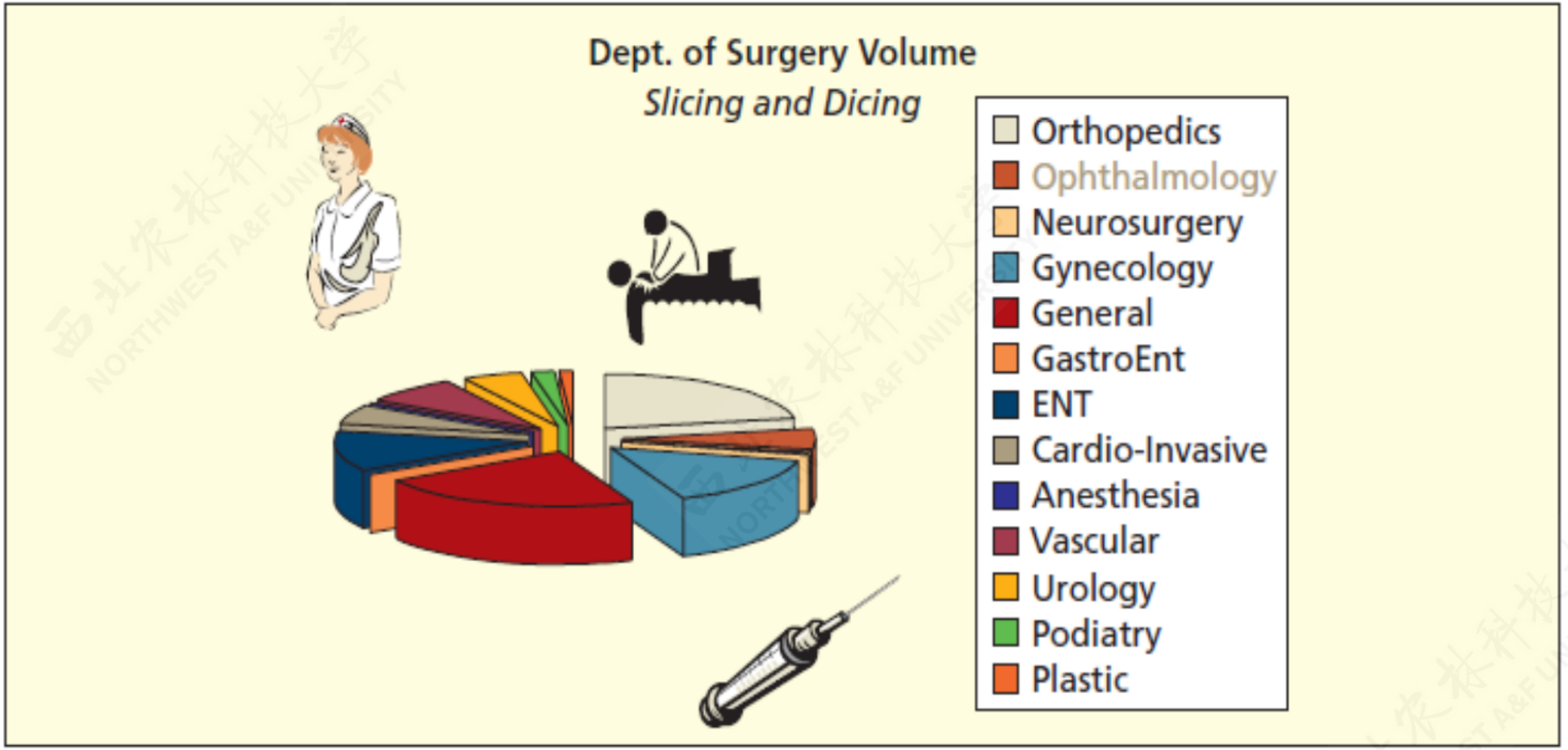
Bad Presentation



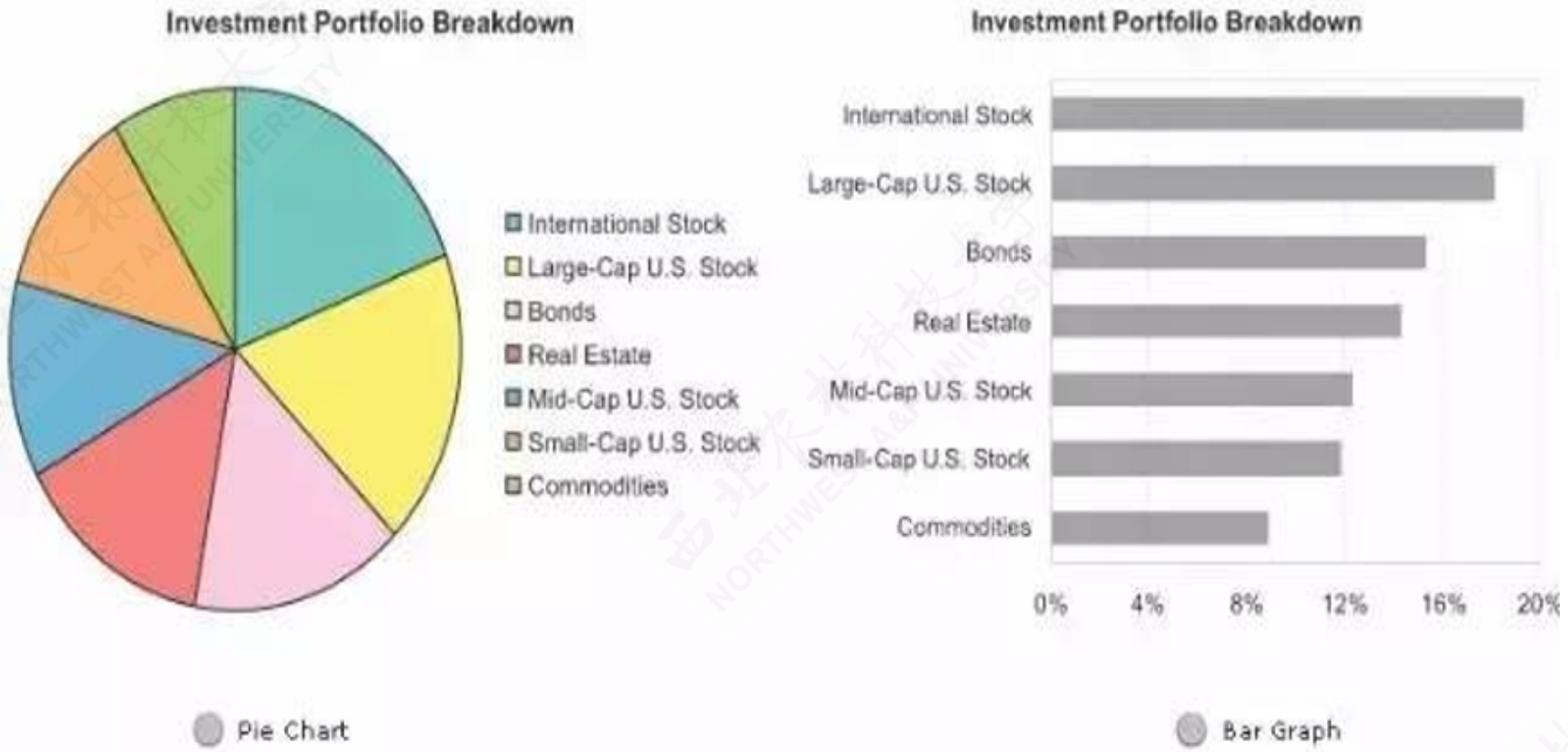
Good Presentation



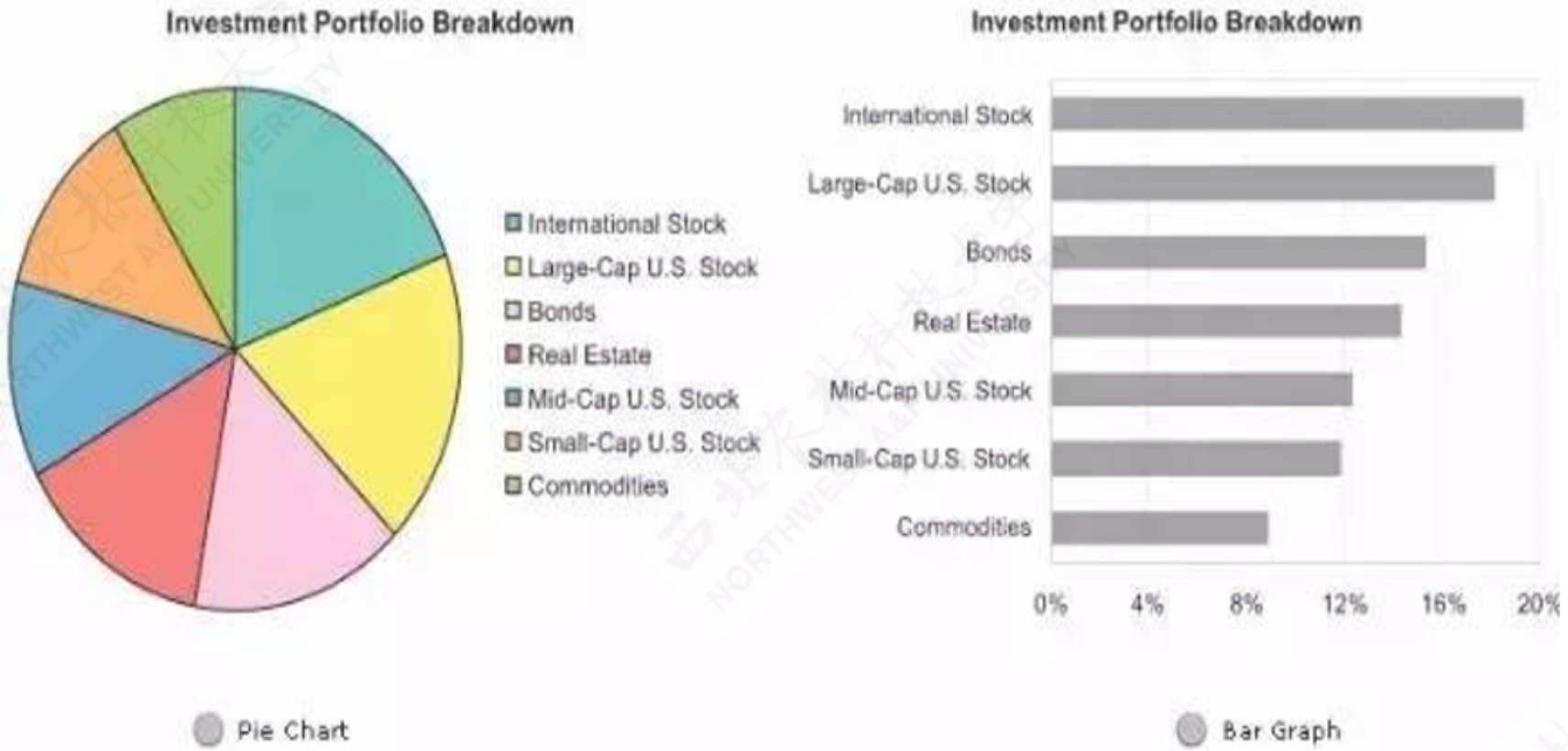
制图常见误区：目标不明确



图设计要点：饼图VS条形图？

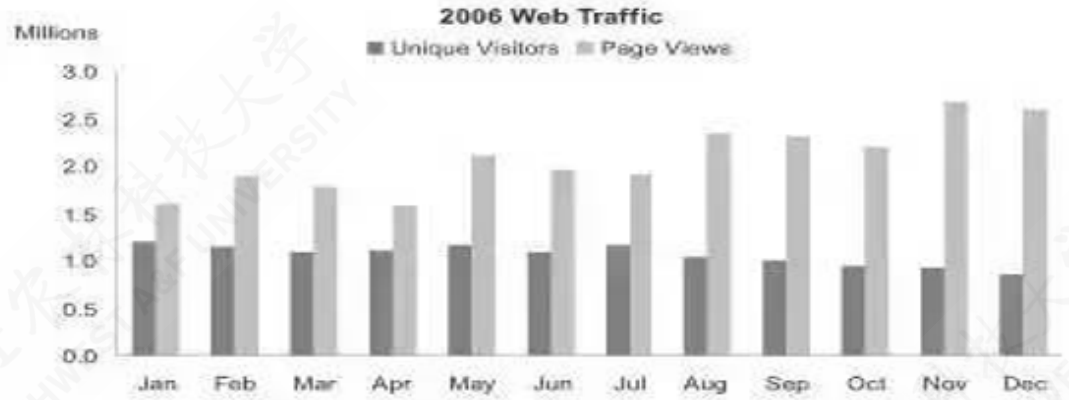


图设计要点：饼图VS条形图？

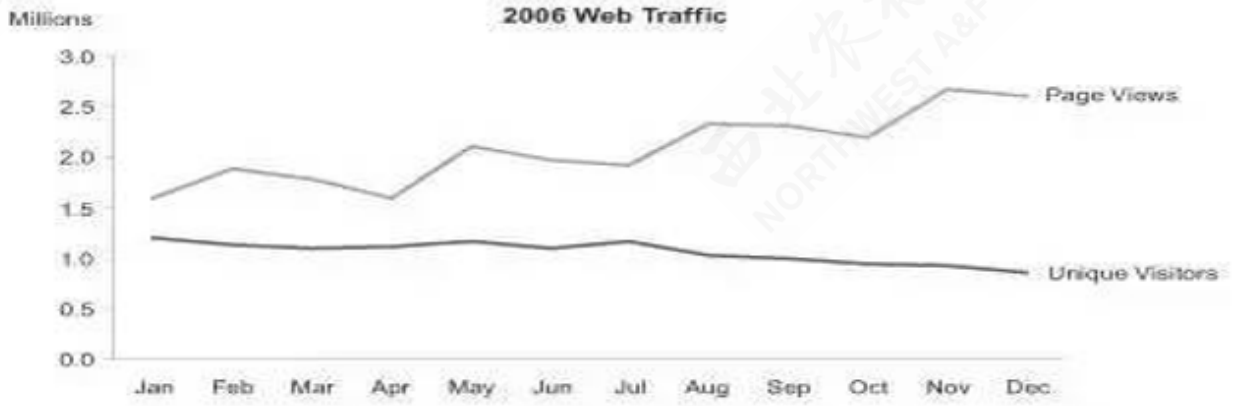


点评：一般来讲表述市场份额是100%，所以大家习惯用饼图表述，研究表明人们更习惯从条形图来比较大小，更醒目的看到差异！

图设计要点：线形图VS柱状图？

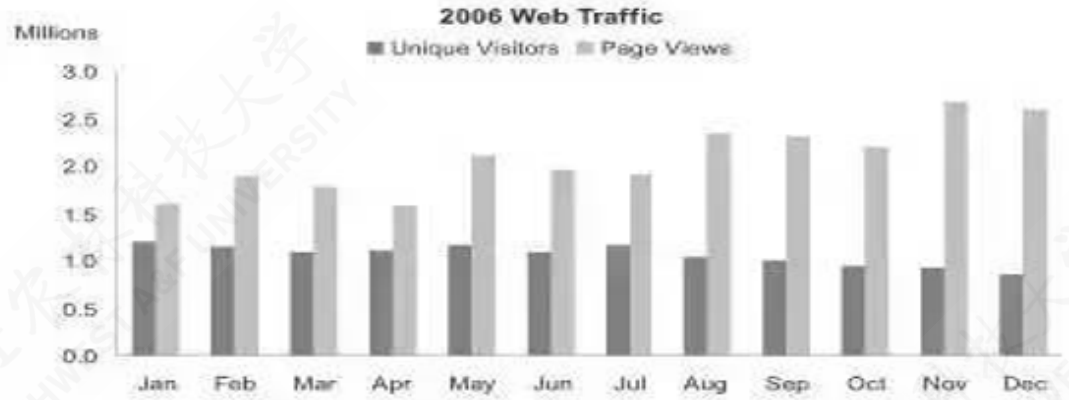


Bar Graph

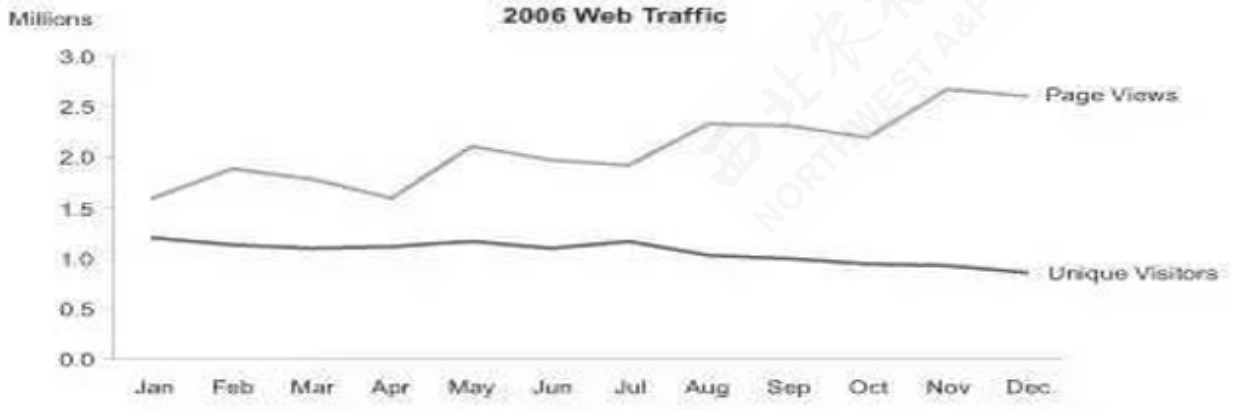


Line Graph

图设计要点：线形图VS柱状图？



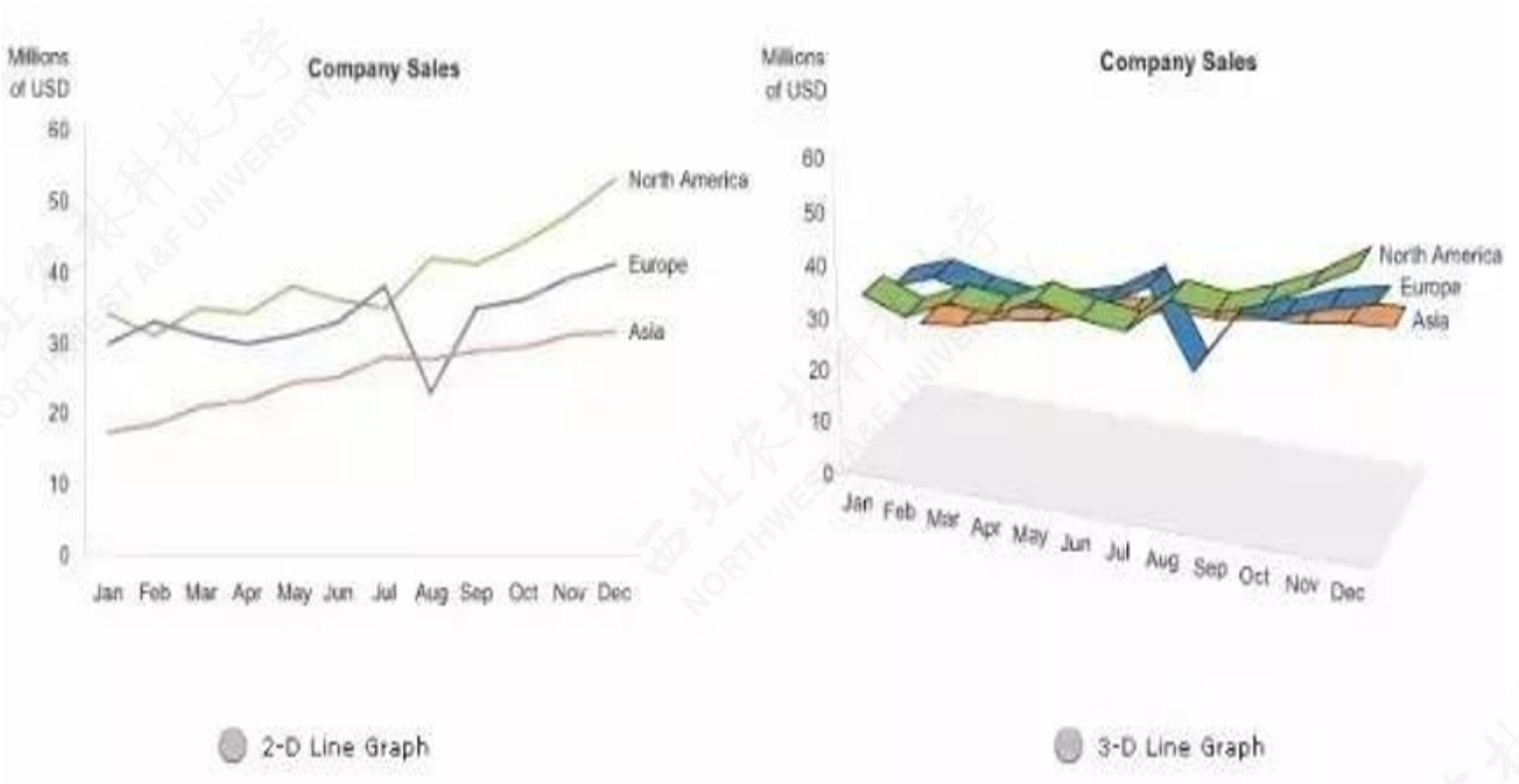
Bar Graph



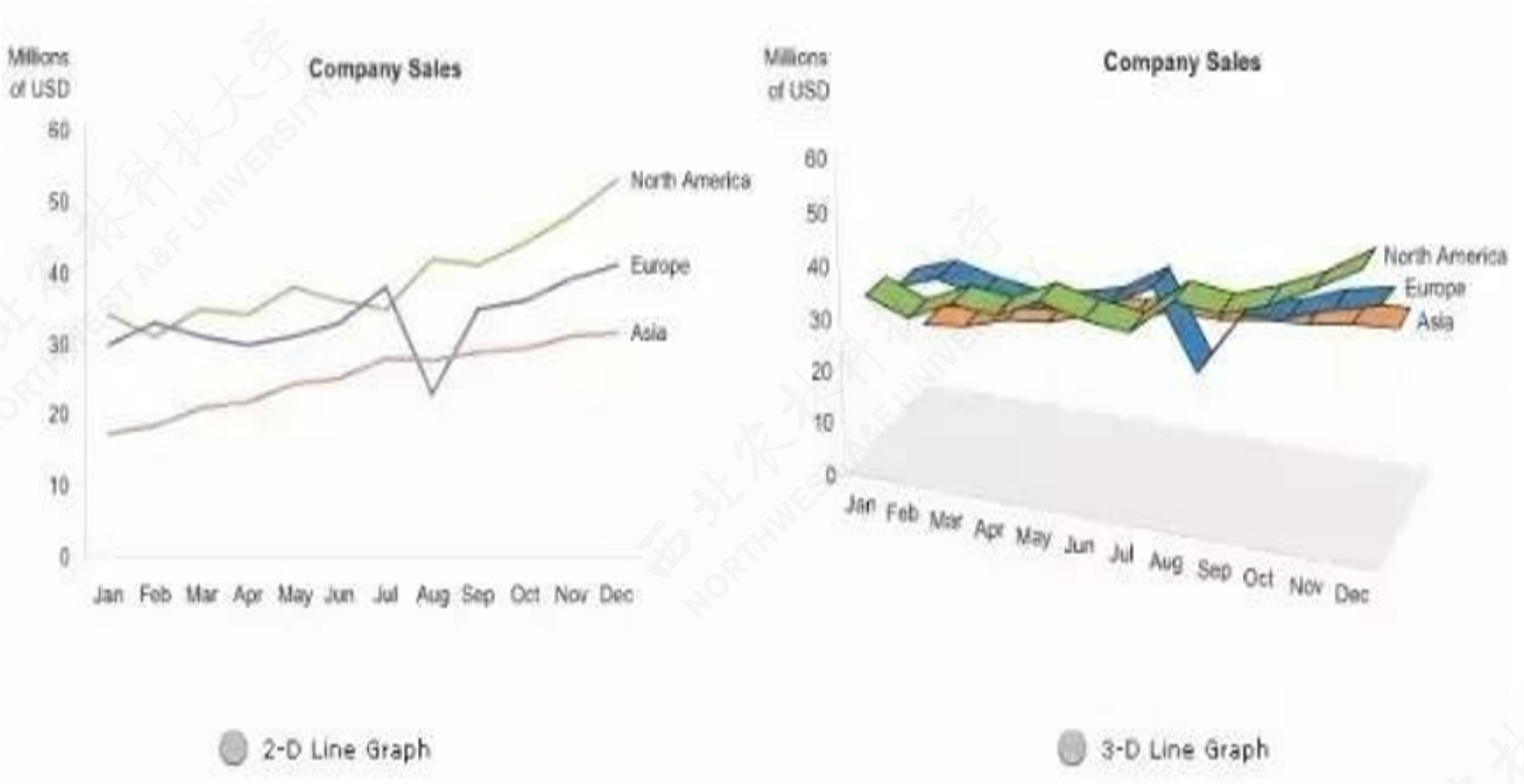
Line Graph

点评：X轴是时间，是时间序列数据，所以折线图更能够感知的趋势、模式的变化！当然如果你表现的是不同品牌的市场份额，柱状图也是可以的！

图设计要点：平面图VS立体图？

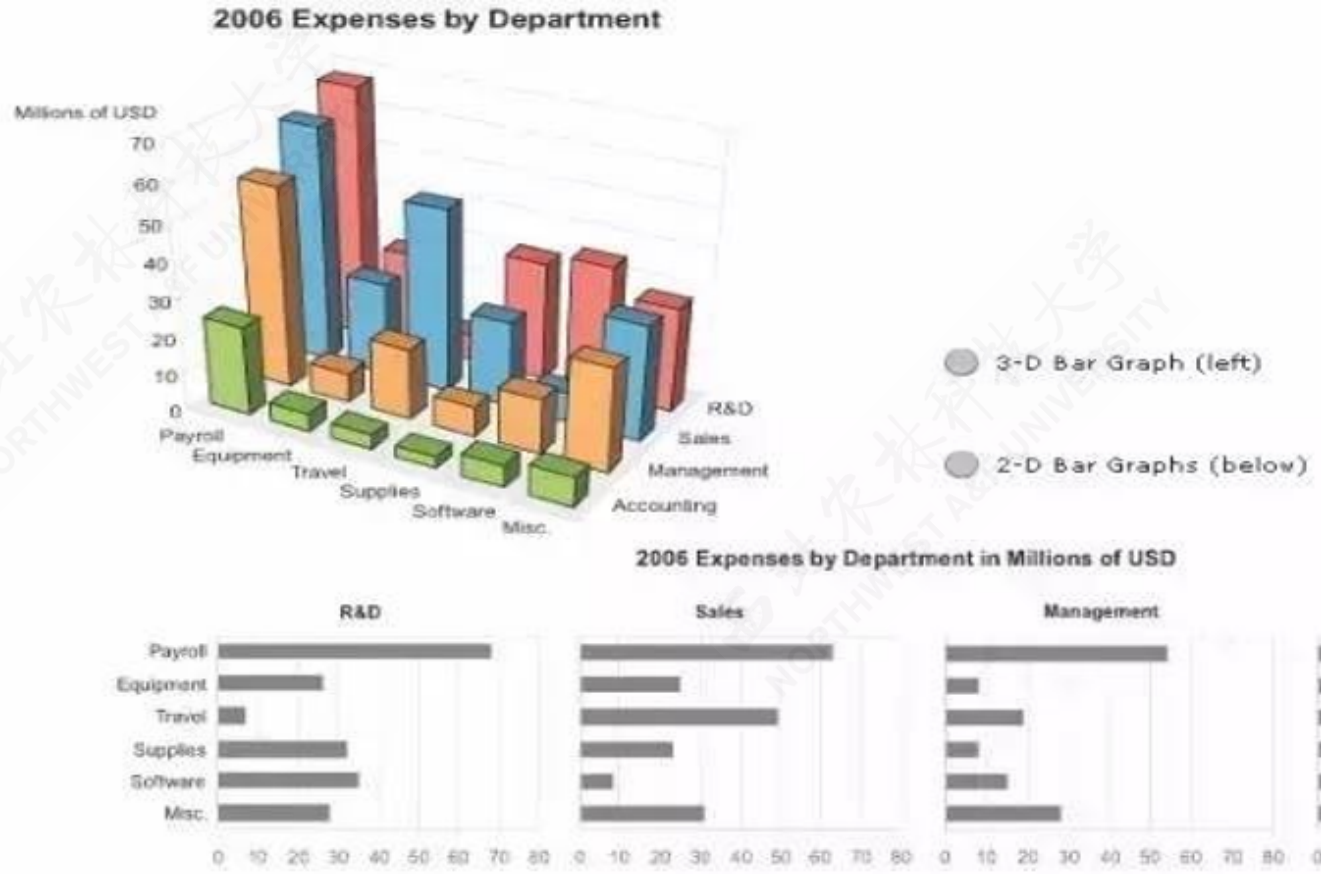


图设计要点：平面图VS立体图？

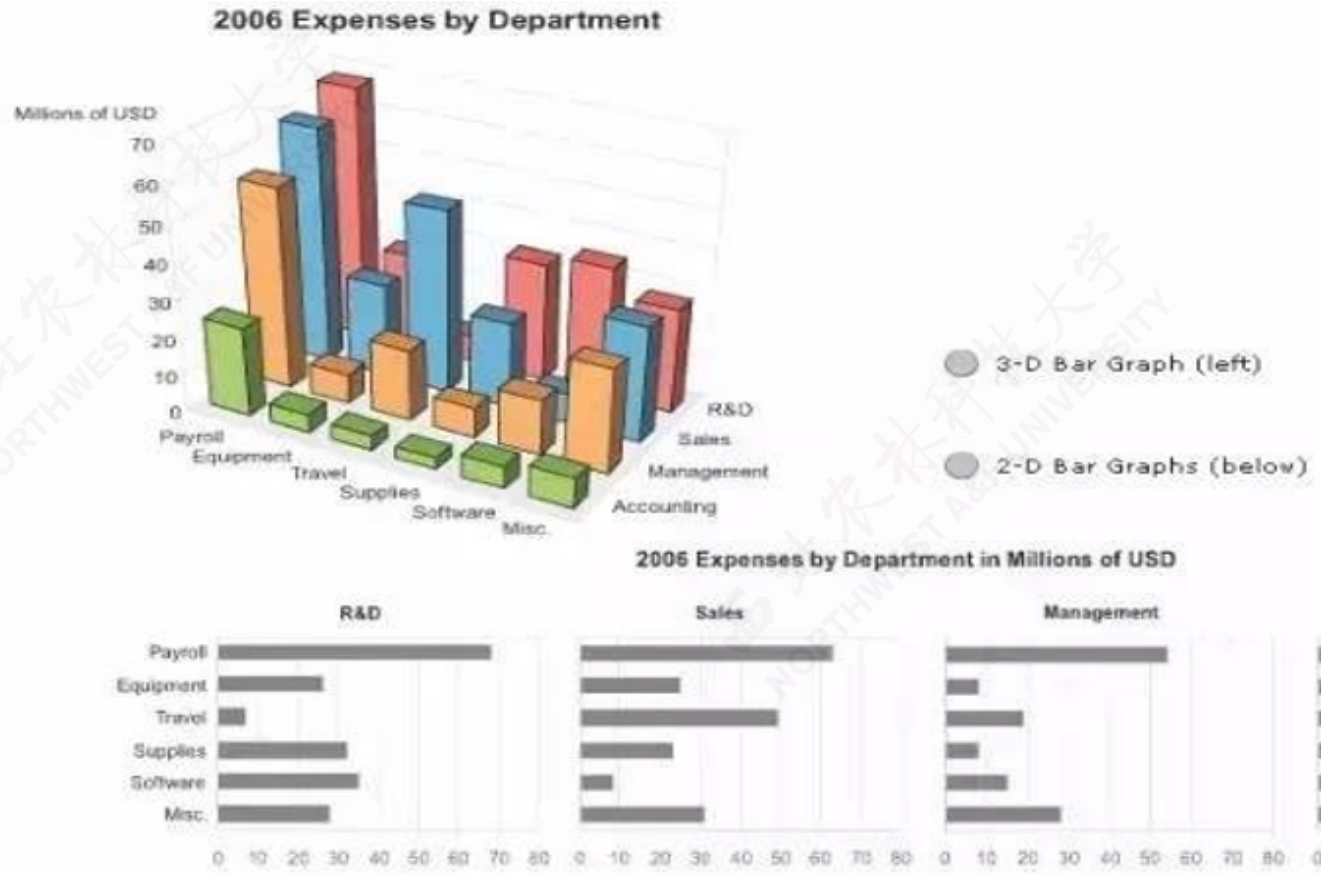


点评：二维图更适合人们观察，三维并不适合观察，毕竟人类视觉空间最低维度是二维！另时序数据应该采用折线图表述趋势！

图设计要点：3D VS 2D？



图设计要点：3D VS 2D？



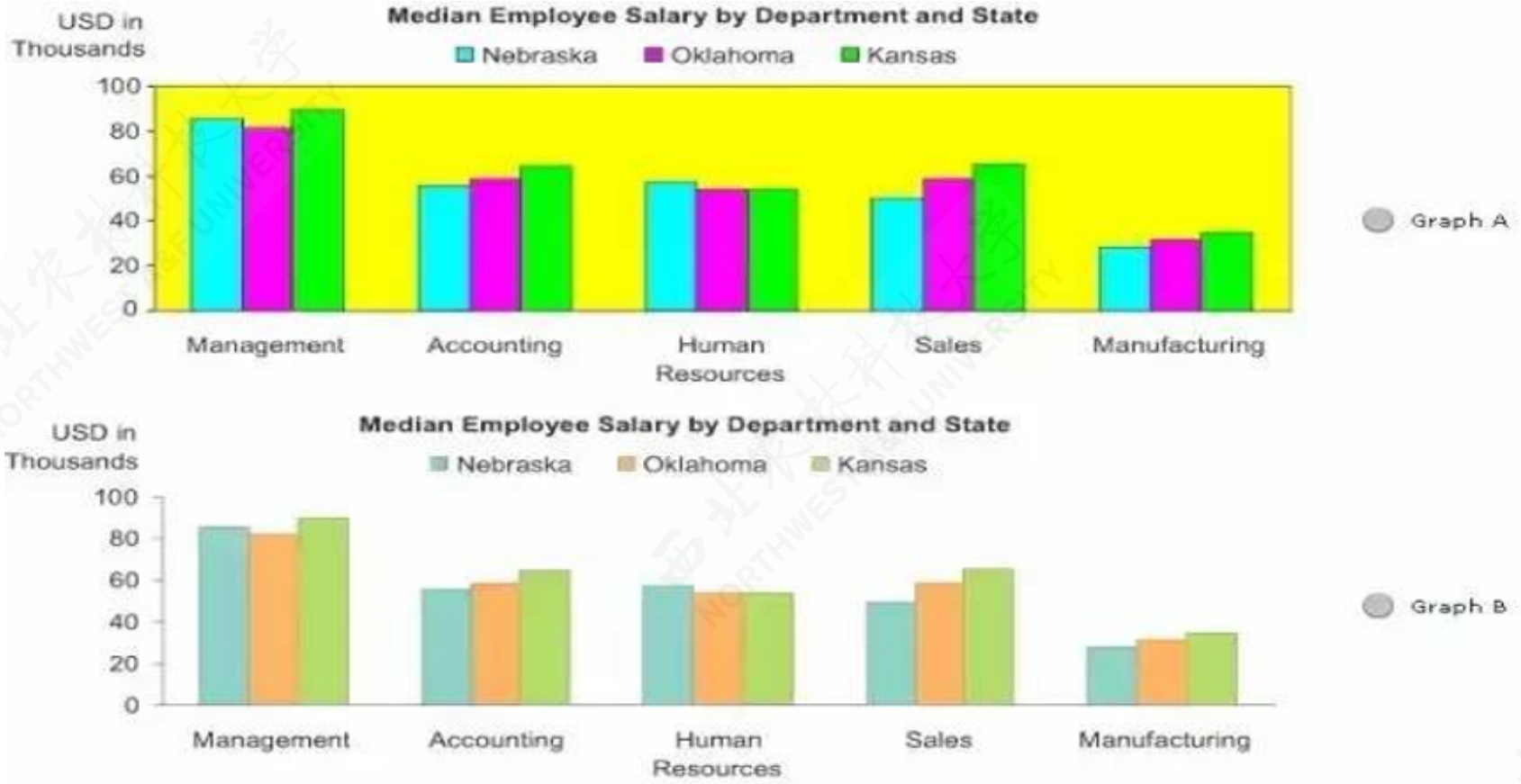
点评：尽量不用用三维图，但是大家是不是会把信息按某个维度作出分散的二维图呢？一定注意要用统一的纵坐标，否则是四张图，要贴成一张图！

图设计要点：颜色VS数据？

Graph A

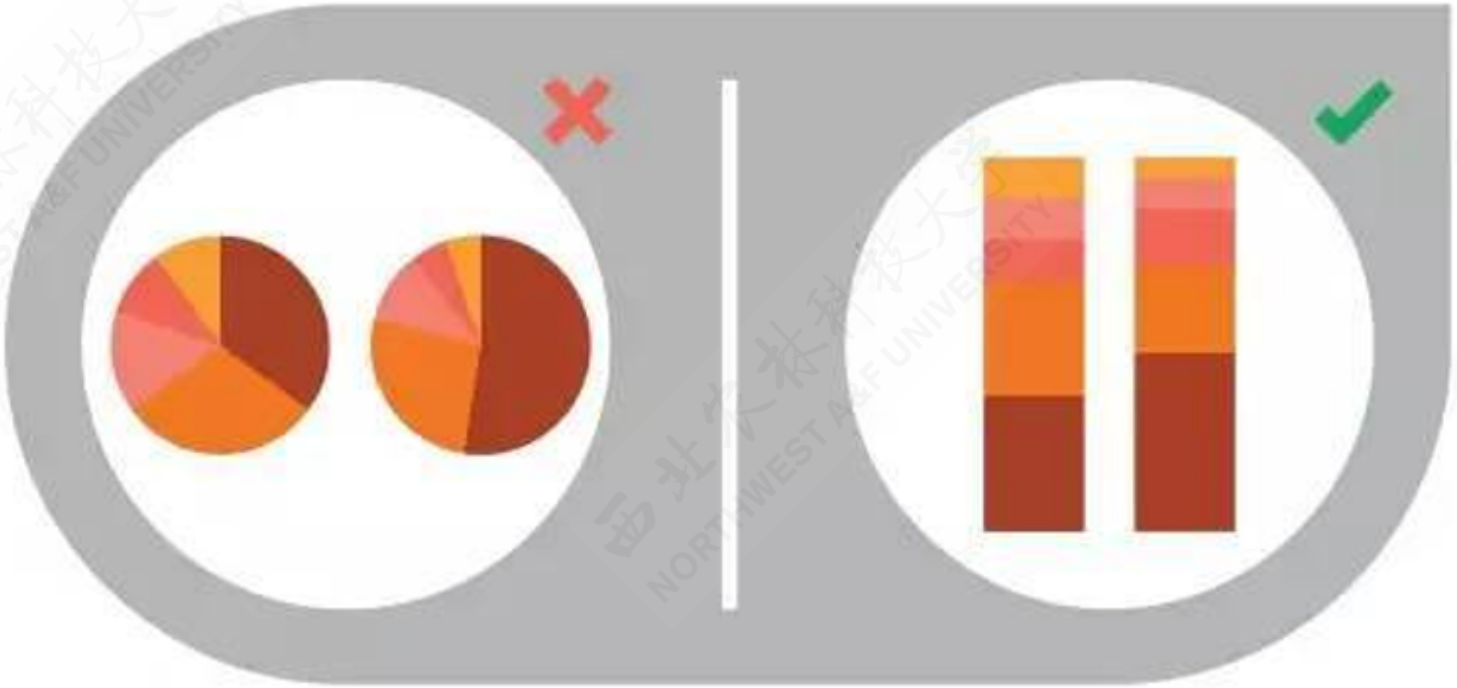
Graph B

图设计要点：颜色VS数据？

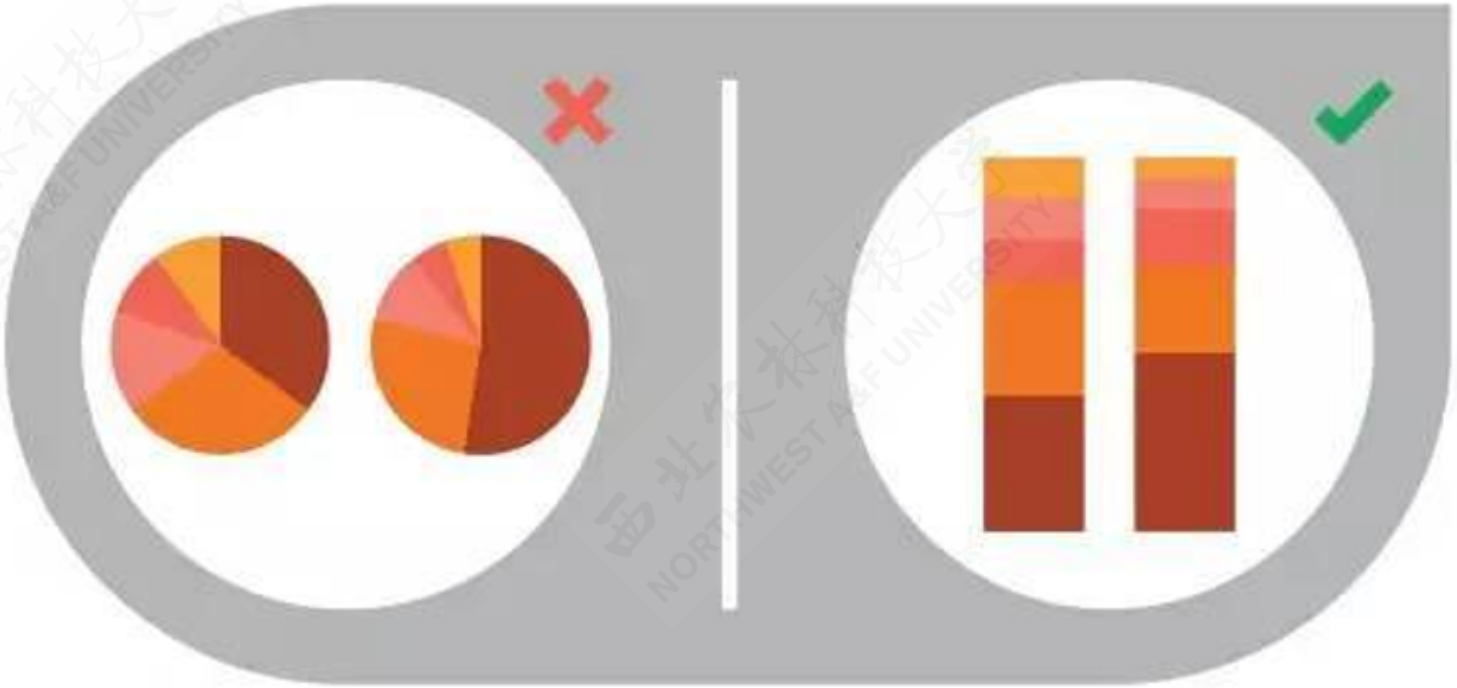


点评：如果选上面的打印都费墨，坚持简单是最好的。当然如果有艺术细胞的话，背景也是可以更为生动些，但是更多是考虑展示结果！

图设计要点：选色VS选图？

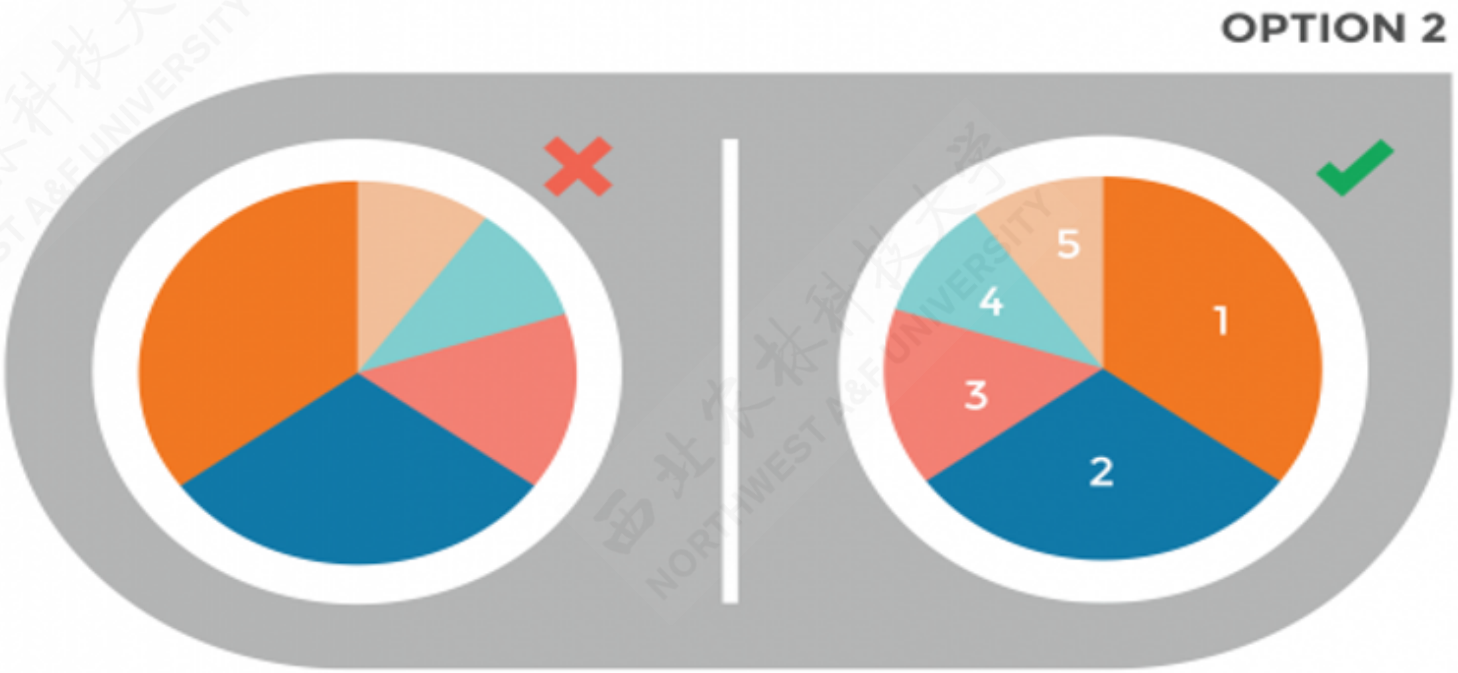


图设计要点：选色VS选图？

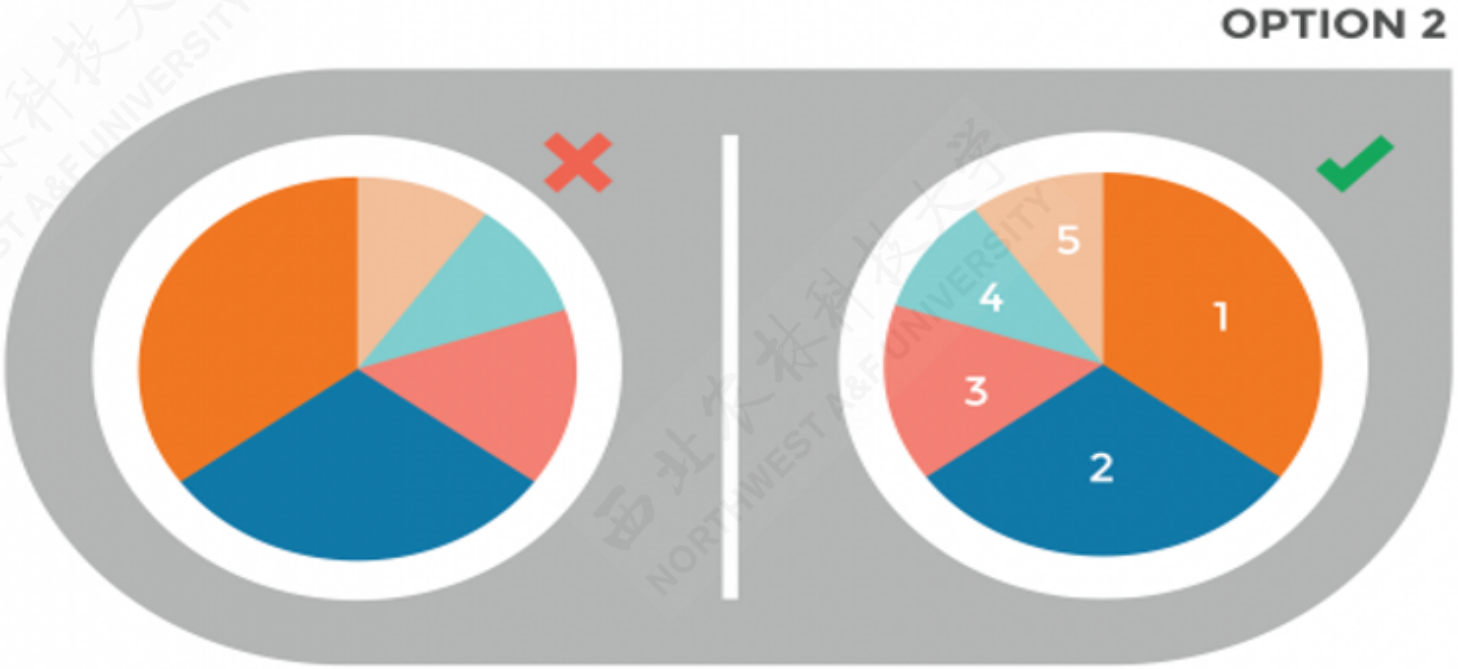


点评：比较是展示数据差异的好法子，但是如果读者不容易看出差别的话，那么比较就毫无意义。确保选择最合适的比较方法。

图设计要点：饼图-排序VS分块？

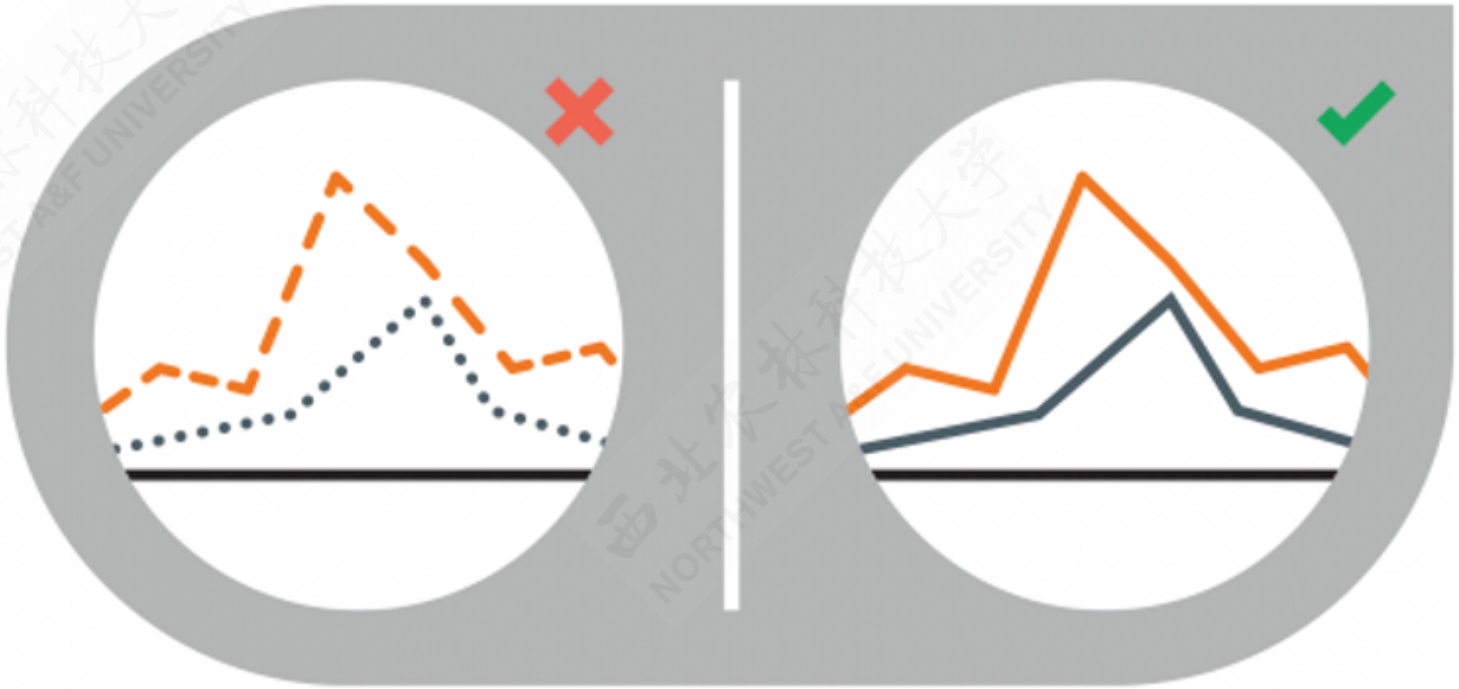


图设计要点：饼图-排序VS分块？

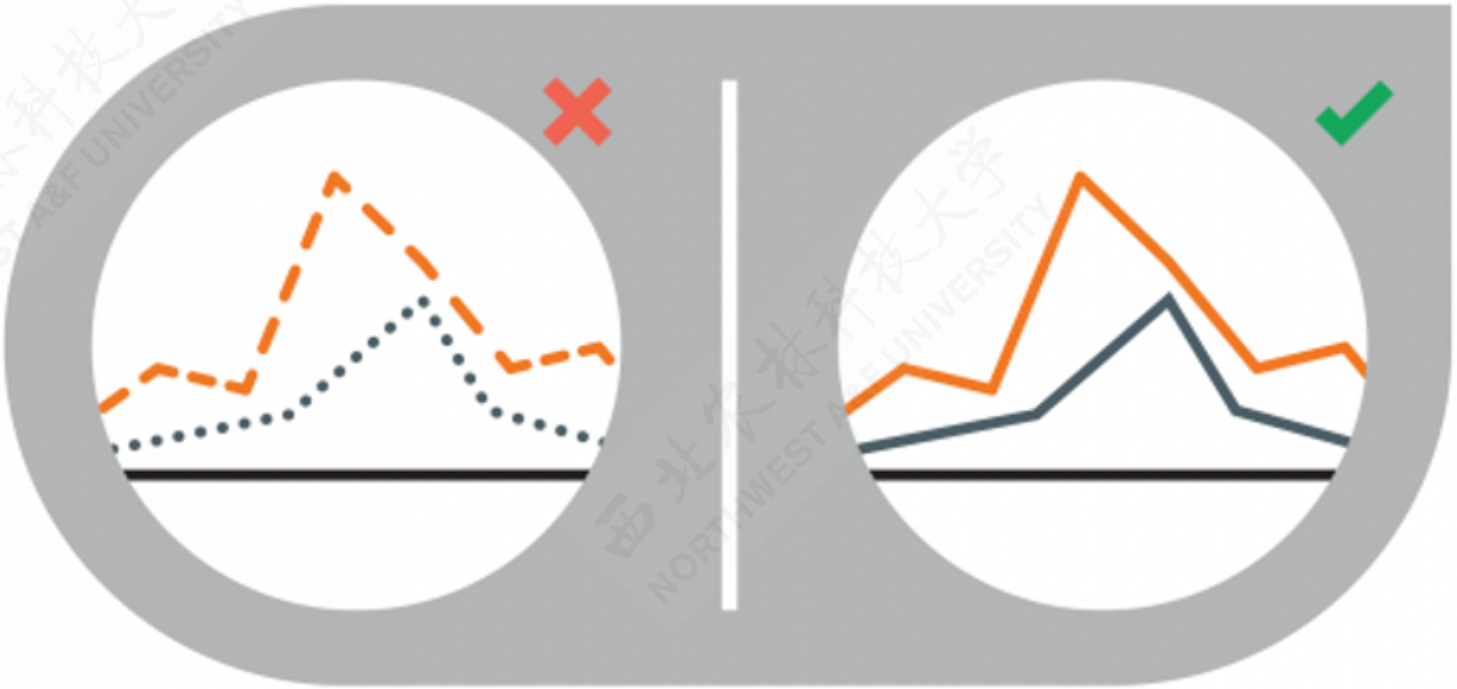


点评：理论上，一个饼图不应该分割超过5块。注意排序，最大一块12点钟开始，顺时针方向旋转。剩余部分再降序排列，顺时针。

图设计要点：线图-实线VS虚线？

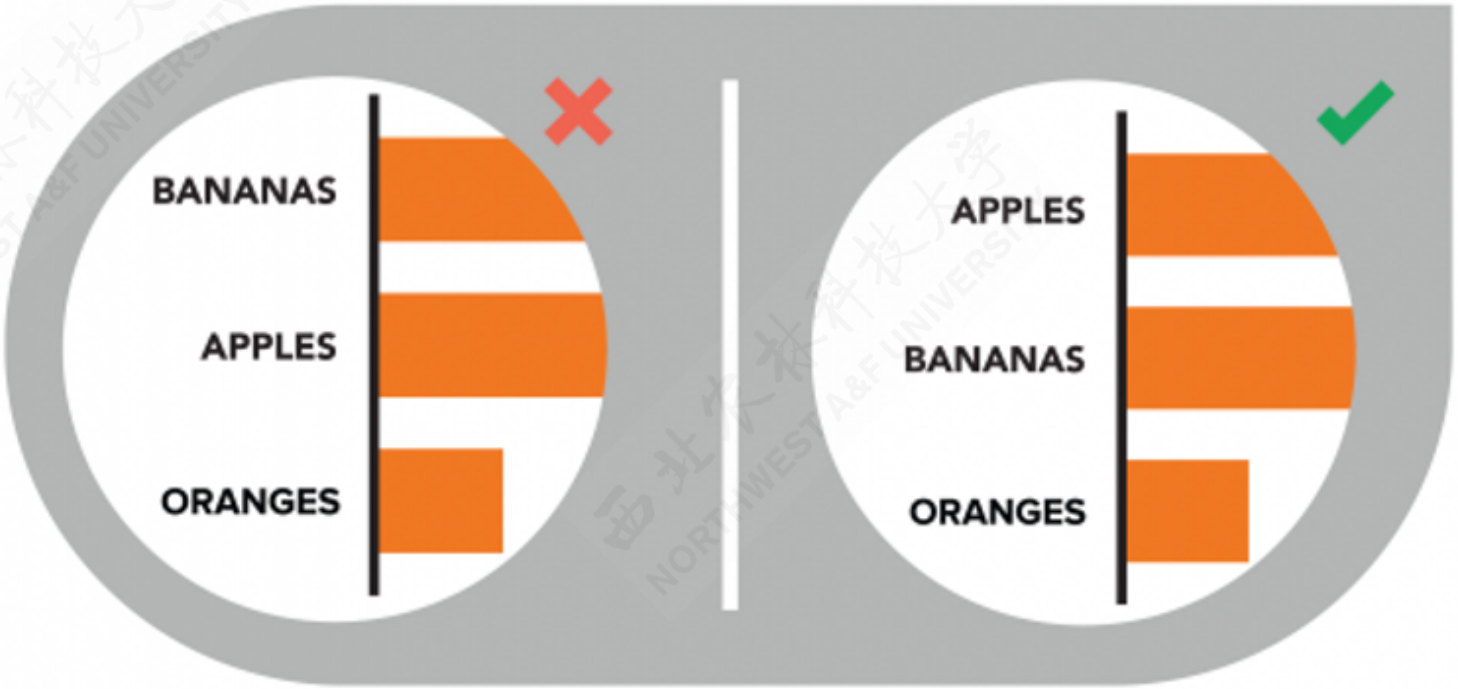


图设计要点：线图-实线VS虚线？

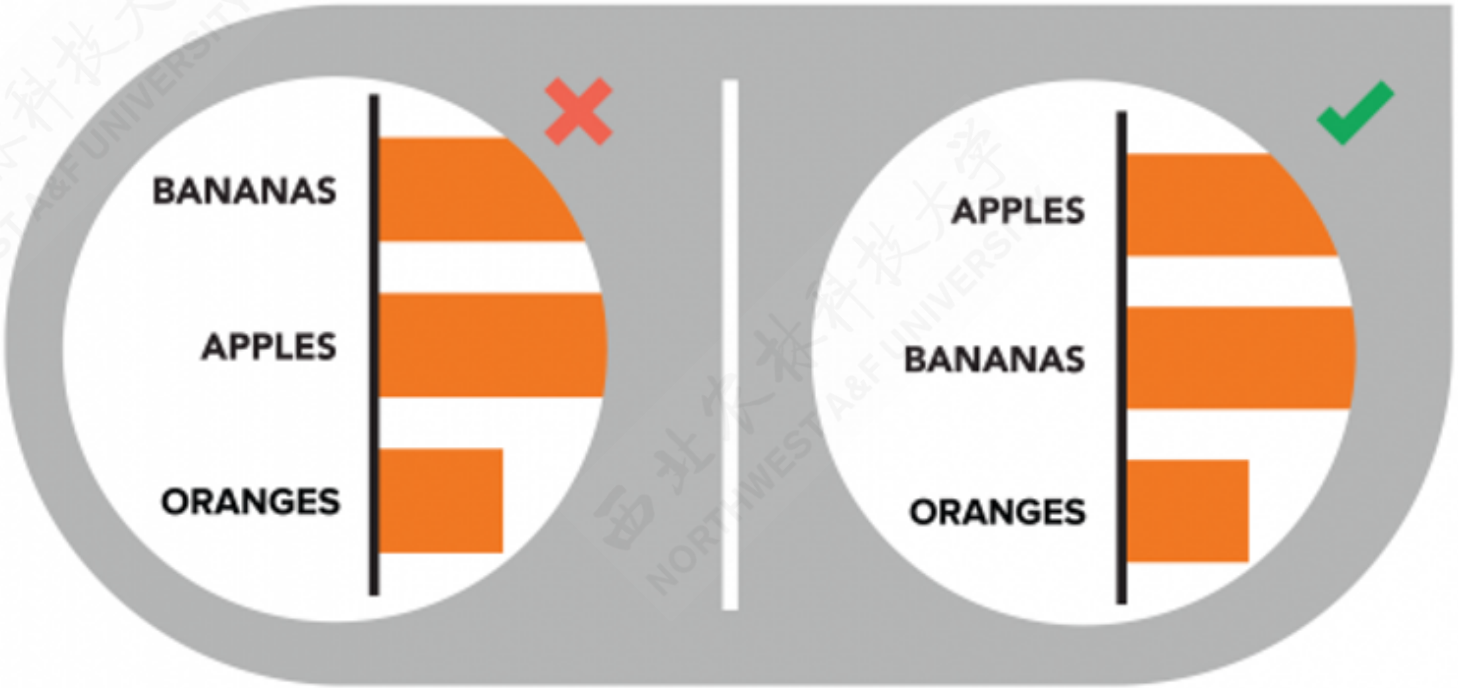


点评：虚线容易分散注意力。相反,使用实线和颜色,反而容易区分彼此的区别。

图设计要点：条形图-排序数据VS名称？

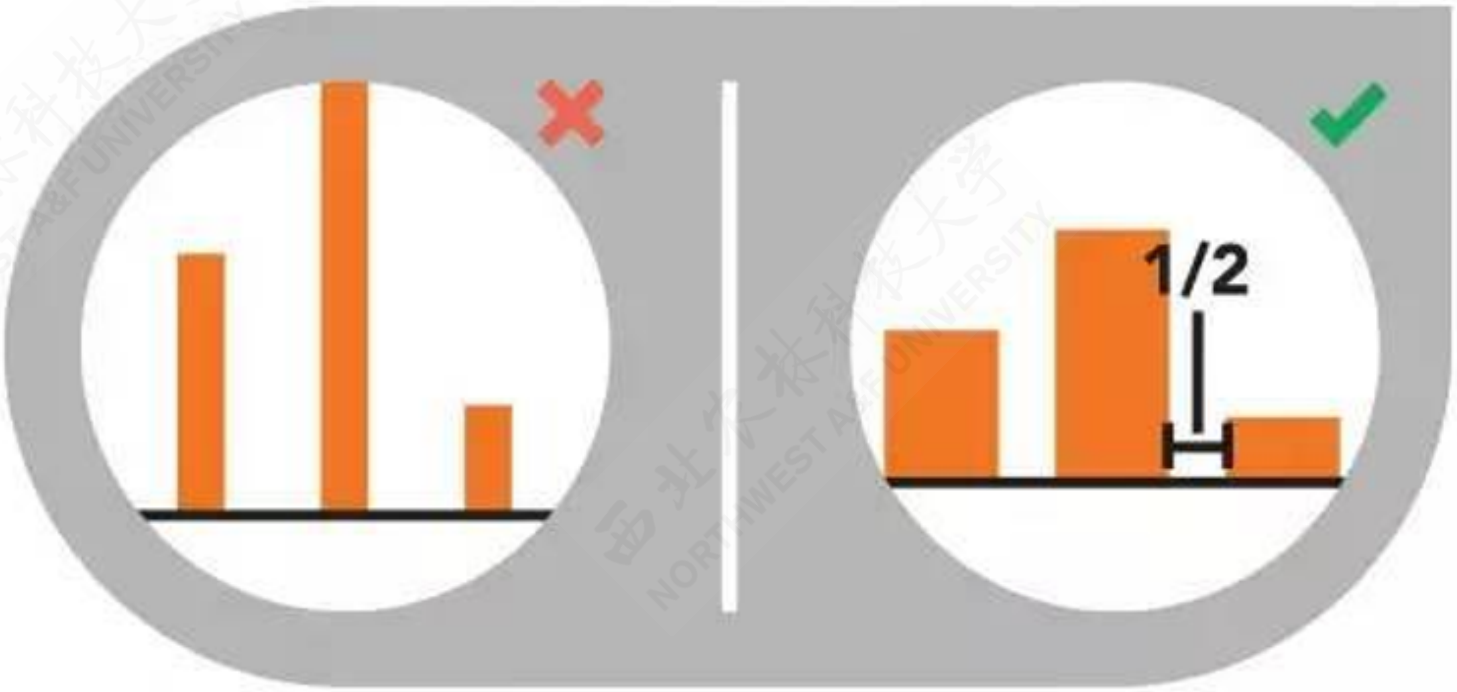


图设计要点：条形图-排序数据VS名称？

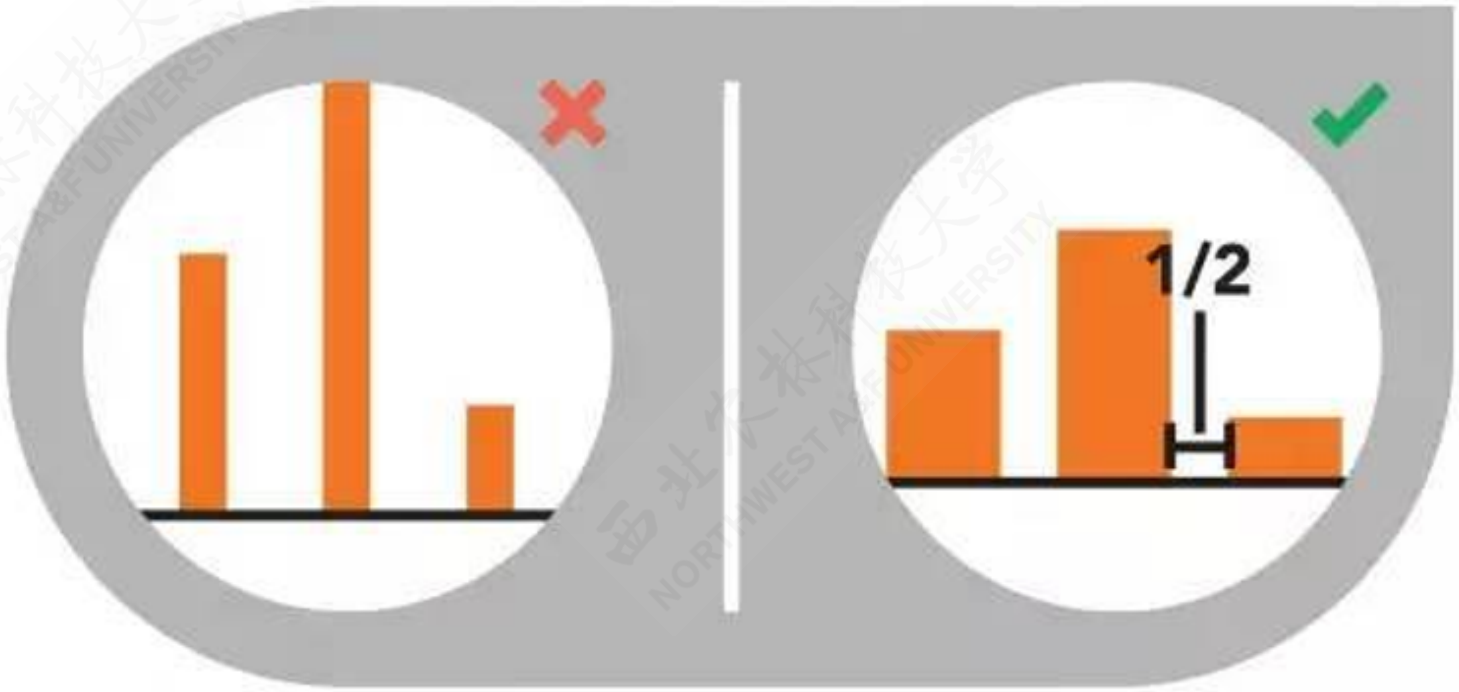


点评：你的内容应该以一种合乎逻辑的和直观的方式来引导读者了解数据。所以，记得将数据类别按字母顺序、大小顺序、或数据值进行排序。

图设计要点：柱状图-宽度VS间距？



图设计要点：柱状图-宽度VS间距？



点评：或许你的报告很有创意，非常精彩，但是记得图表设计水平也要跟上。条形图之间的间隔应该是1/2栏宽度。

图设计要点：面积图-堆叠VS透明度？

图设计要点：面积图-堆叠VS透明度？

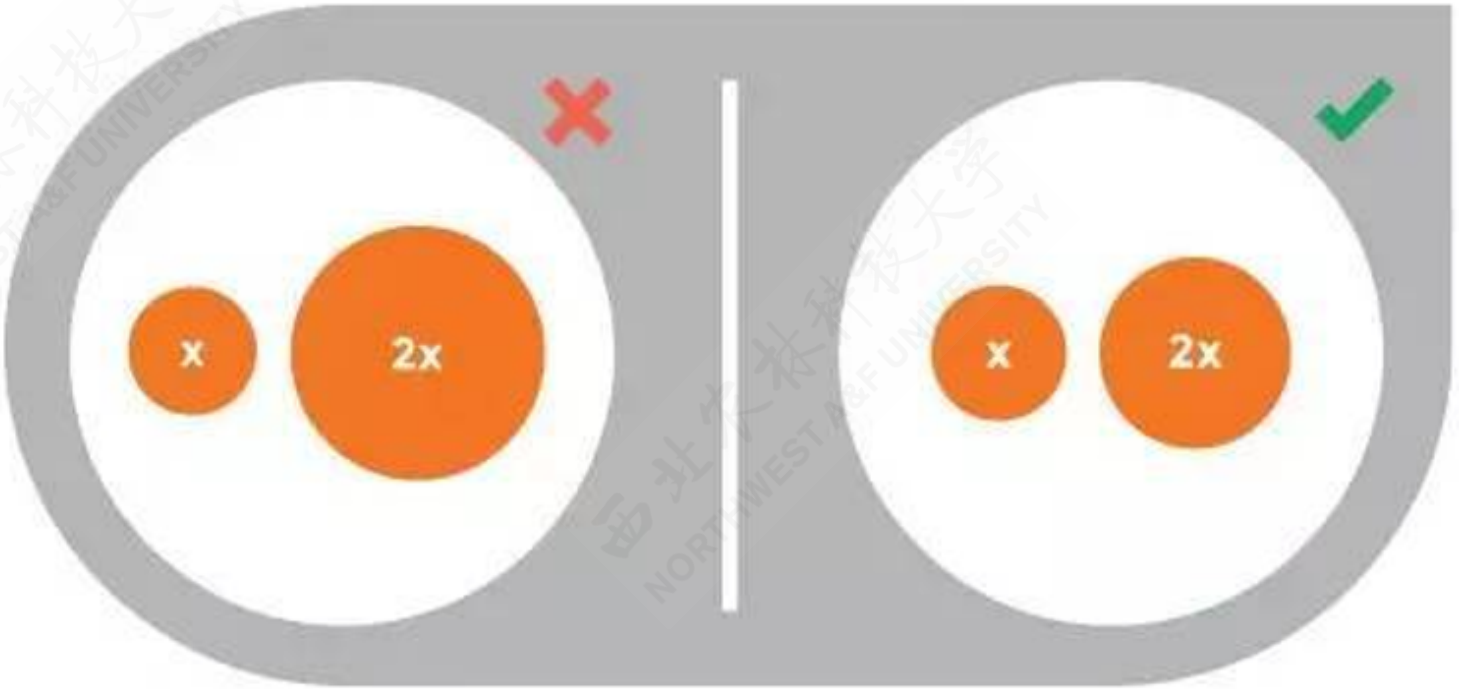
点评：确保没有数据丢失或被设计修改。例如，使用标准的面积图时，可以添加透明度，确保读者可以看到所有数据。

图设计要点：散点图-原始VS趋势？

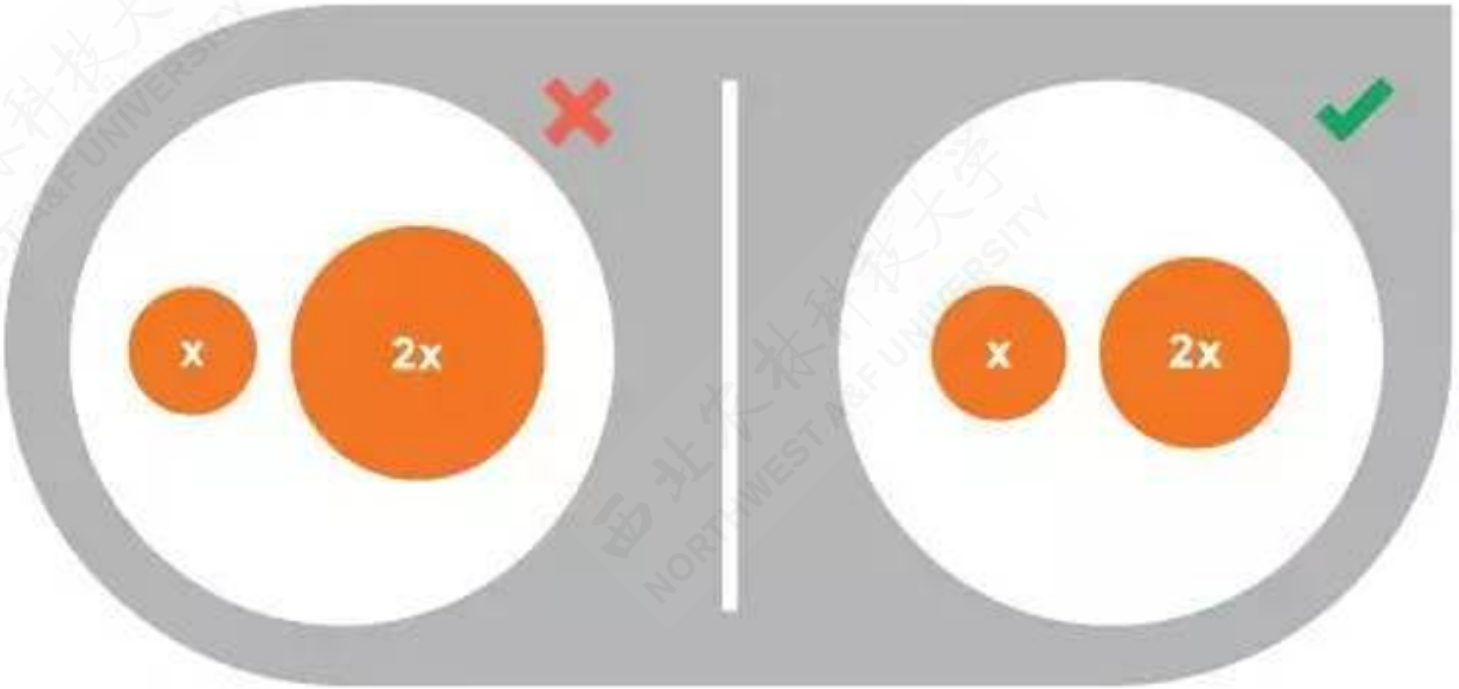
图设计要点：散点图-原始VS趋势？

点评：应该使图表尽可能轻松地帮助读者理解数据。例如，在散点图中添加趋势线来强调的趋势。

图设计要点：气泡图-形状VS数值？

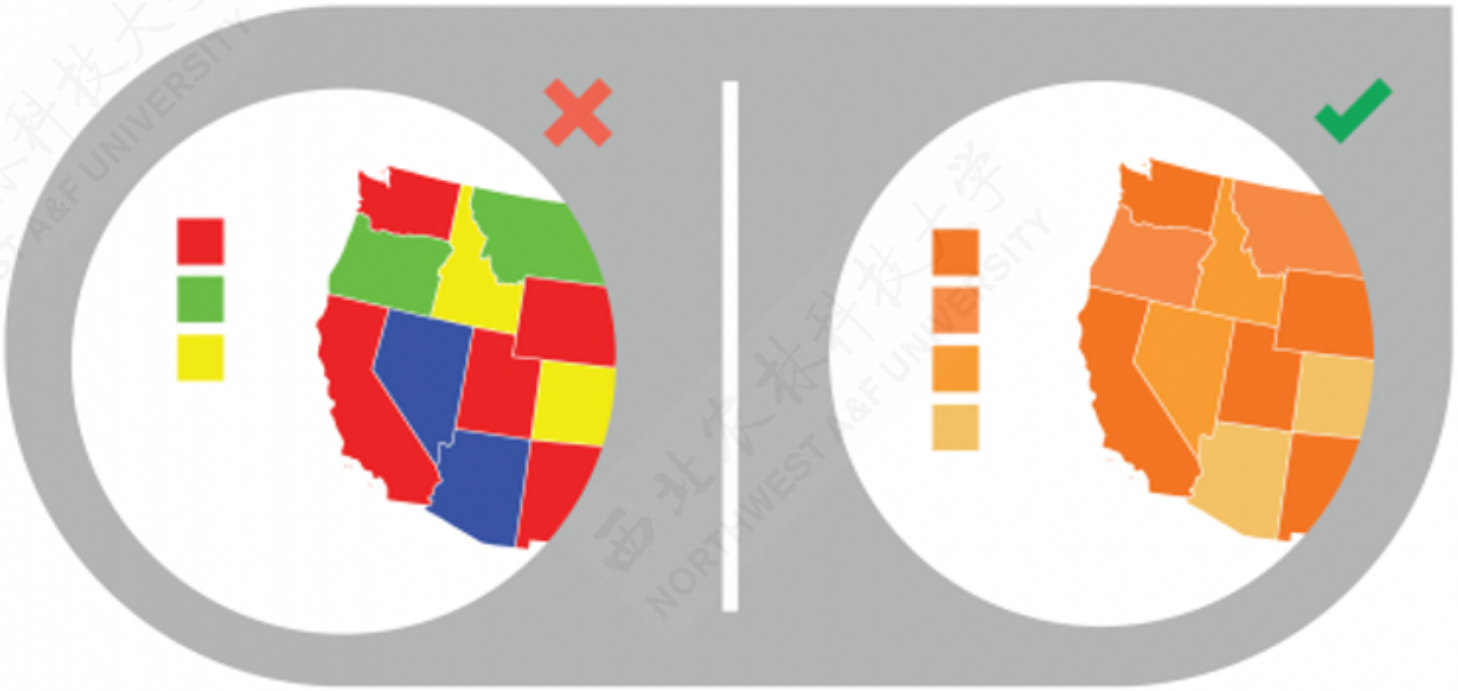


图设计要点：气泡图-形状VS数值？

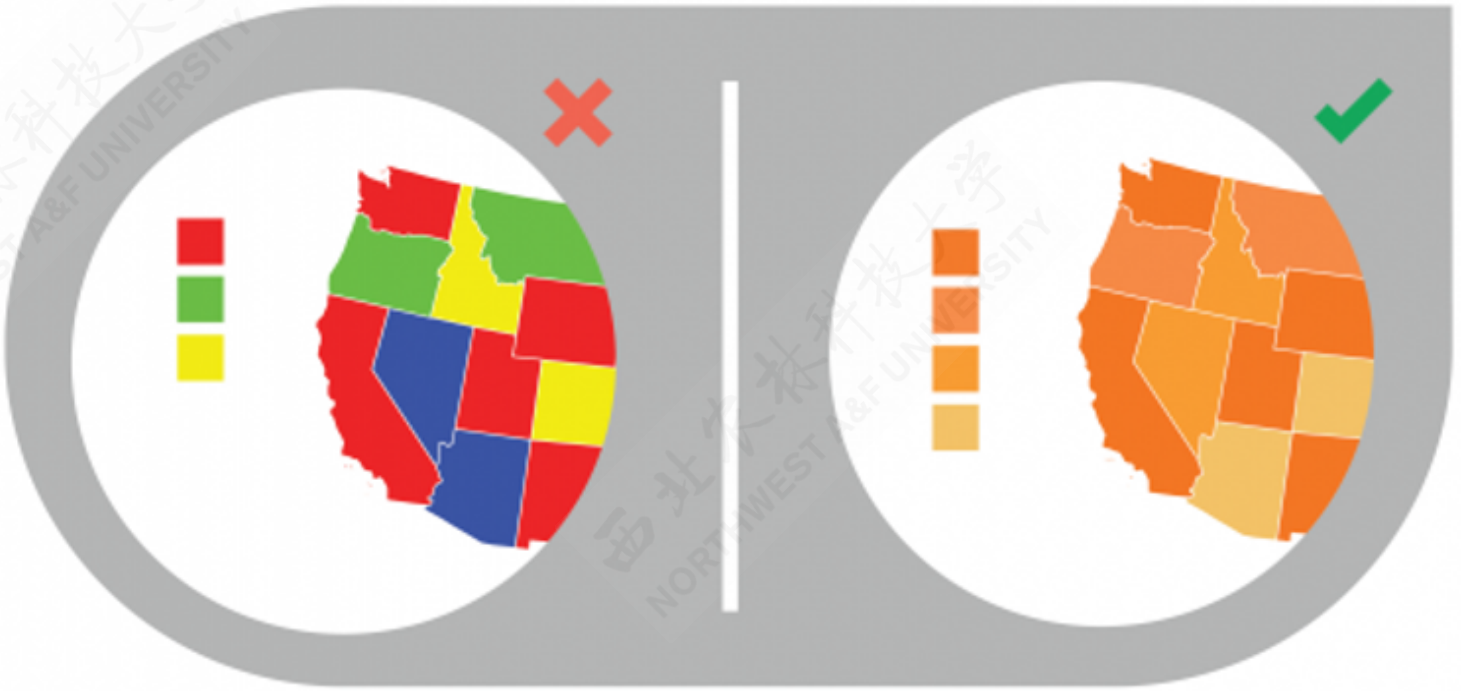


点评：确保所有可视化方式是准确的。例如，气泡图大小应该根据区域扩展，而不是直径。

图设计要点：热力图-颜色堆VS颜色系？



图设计要点：热力图-颜色堆VS颜色系？



点评：颜色用得太花，会给数据增加不可承受之重，相反，应该采用同一色系，或者类比色。

表设计要点：交叉表-封闭VS开放？

Top Table

Sales Summary by Region
1st Quarter, 2007
Regions are Sorted by Revenue

Region	Revenue	% of Total Revenue	Expenses	Profit	% of Total Profit
Europe	$75,904,604.00	31.06%	$40,988,486.16	$34,916,117.84	22.31%
Canada	$51,572,694.00	21.10%	$17,534,715.96	$34,037,978.04	21.75%
Western U.S.	$42,660,178.00	17.46%	$11,944,849.84	$30,715,328.16	19.63%
Eastern U.S.	$33,977,385.00	13.90%	$7,135,250.85	$26,842,134.15	17.15%
Central U.S.	$26,139,598.00	10.70%	$3,920,939.70	$22,218,658.30	14.20%
Asia	$14,135,278.00	5.78%	$6,360,875.10	$7,774,402.90	4.97%
Total (or Avg)	$244,389,737.00	100.00%	$87,885,117.61	$156,504,619.39	100.00%

Bottom Table

Sales Summary by Region (USD)
1st Quarter, 2007
Regions are Sorted by Revenue

Region	Revenue	% of Total Revenue	Expenses	Profit	% of Total Profit
Europe	75,904,604	31.1%	40,988,486	34,916,118	22.3%
Canada	51,572,694	21.1%	17,534,716	34,037,978	21.7%
Western U.S.	42,660,178	17.5%	11,944,850	30,715,328	19.6%
Eastern U.S.	33,977,385	13.9%	7,135,251	26,842,134	17.2%
Central U.S.	26,139,598	10.7%	3,920,940	22,218,658	14.2%
Asia	14,135,278	5.8%	6,360,875	7,774,403	5.0%
Total (or Avg)	$244,389,737	100.0%	$87,885,118	$156,504,619	100.0%

表设计要点：交叉表-封闭VS开放？

Top Table

Sales Summary by Region
1st Quarter, 2007
Regions are Sorted by Revenue

Region	Revenue	% of Total Revenue	Expenses	Profit	% of Total Profit
Europe	$75,904,604.00	31.06%	$40,988,486.16	$34,916,117.84	22.31%
Canada	$51,572,694.00	21.10%	$17,534,715.96	$34,037,978.04	21.75%
Western U.S.	$42,660,178.00	17.46%	$11,944,849.84	$30,715,328.16	19.63%
Eastern U.S.	$33,977,385.00	13.90%	$7,135,250.85	$26,842,134.15	17.15%
Central U.S.	$26,139,598.00	10.70%	$3,920,939.70	$22,218,658.30	14.20%
Asia	$14,135,278.00	5.78%	$6,360,875.10	$7,774,402.90	4.97%
Total (or Avg)	$244,389,737.00	100.00%	$87,885,117.61	$156,504,619.39	100.00%

Bottom Table

Sales Summary by Region (USD)
1st Quarter, 2007
Regions are Sorted by Revenue

Region	Revenue	% of Total Revenue	Expenses	Profit	% of Total Profit
Europe	75,904,604	31.1%	40,988,486	34,916,118	22.3%
Canada	51,572,694	21.1%	17,534,716	34,037,978	21.7%
Western U.S.	42,660,178	17.5%	11,944,850	30,715,328	19.6%
Eastern U.S.	33,977,385	13.9%	7,135,251	26,842,134	17.2%
Central U.S.	26,139,598	10.7%	3,920,940	22,218,658	14.2%
Asia	14,135,278	5.8%	6,360,875	7,774,403	5.0%
Total (or Avg)	$244,389,737	100.0%	$87,885,118	$156,504,619	100.0%

点评：下表更加清晰明快！我们习惯采用中国式上表，一般都是封闭边框线的。但是我们如果经常看英文的论文，你会发现很多论文都是下面开放式三线表！

表设计要点：交叉表-颜色VS线条？

2006 Key Metrics

Region	Overall	Revenue	Expenses	Profit	Avg. Order Size
East	Good	$4,652,462	$2,682,765	$1,969,697	$6,845
West	Fair	3,705,426	2,211,773	1,493,653	4,266
North	Fair	3,215,789	2,712,984	502,805	4,568
South	Poor	2,215,752	1,562,735	653,017	1,358
Overall	**Fair**	**$13,789,429**	**$9,170,257**	**$4,619,172**	**$4,259**

Table A

2006 Key Metrics

Region	Overall	Revenue	Expenses	Profit	Avg. Order Size
East	Good	$4,652,462	$2,682,765	$1,969,697	$6,845
West	Fair	3,705,426	2,211,773	1,493,653	4,266
North	Fair	3,215,789	2,712,984	502,805	4,568
South	Poor	2,215,752	1,562,735	653,017	1,358
Overall	**Fair**	**$13,789,429**	**$9,170,257**	**$4,619,172**	**$4,259**

Table B

表设计要点：交叉表-颜色VS线条？

2006 Key Metrics

Region	Overall	Revenue	Expenses	Profit	Avg. Order Size
East	Good	$4,652,462	$2,682,765	$1,969,697	$6,845
West	Fair	3,705,426	2,211,773	1,493,653	4,266
North	Fair	3,215,789	2,712,984	502,805	4,568
South	Poor	2,215,752	1,562,735	653,017	1,358
Overall	**Fair**	**$13,789,429**	**$9,170,257**	**$4,619,172**	**$4,259**

Table A

2006 Key Metrics

Region	Overall	Revenue	Expenses	Profit	Avg. Order Size
East	Good	$4,652,462	$2,682,765	$1,969,697	$6,845
West	Fair	3,705,426	2,211,773	1,493,653	4,266
North	Fair	3,215,789	2,712,984	502,805	4,568
South	Poor	2,215,752	1,562,735	653,017	1,358
Overall	**Fair**	**$13,789,429**	**$9,170,257**	**$4,619,172**	**$4,259**

Table B

点评：应选择上表。网格线会让我们看不清晰。当然颜色也是非常重要的，要学会使用条件格式规则调色，还有如字体、数值、对齐等格式问题！

本章结束