统计学方法及其应用

Statistical Methods with Applications

Rui Jiang, PhD

Associate Professor

Ministry of Education Key Laboratory of Bioinformatics
Bioinformatics Division, TNLIST/Department of Automation
Tsinghua University, Beijing 100084, China

About me

- 江瑞
- 办公室
  - FIT 1-107
- 联系方式
  - 电子邮件：ruijiang@tsinghua.edu.cn
  - 办公电话：62795578-828

What is statistics?

统计学方法及其应用 概论 统计学是什么？

"You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."

Sherlock Holmes

高血压 (Hypertension)

无声杀手

高血压通常没有明显的症状，但长期血压过高，可引发心肌梗死、脑卒中、肾功能衰竭等严重的并发症。

中国成人高血压患者已超过三亿

目前中国已有3亿左右的高血压患者，每年新增高血压病例达1000万。而中国每年死亡的300万心血管病患者中，50%都与高血压有关。可怕的是，50%的高血压患者并不知道自己平日的血压水平。

吴兆苏，中国高血压防治概况（2015年9月18日）

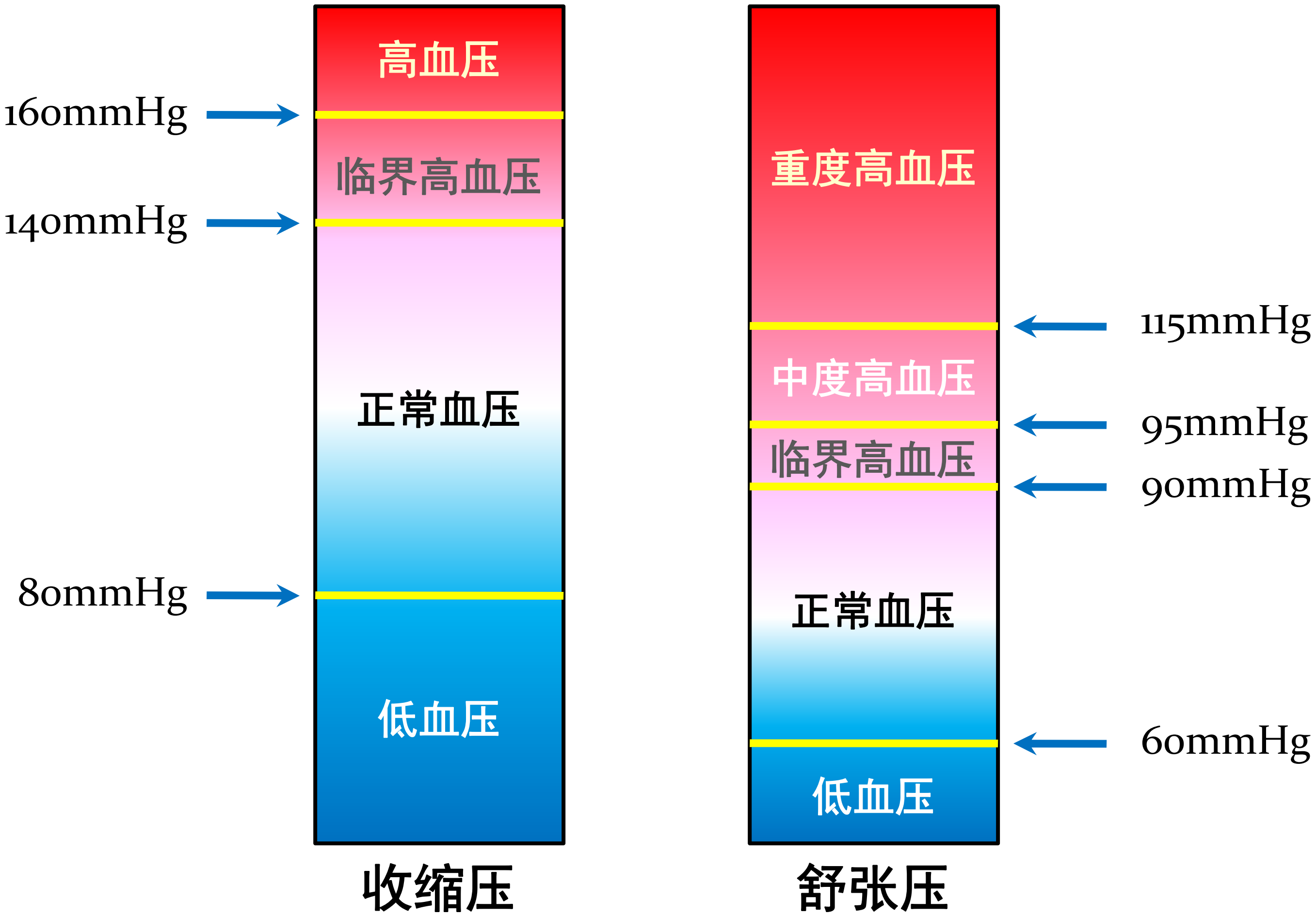
健康生活，绿色饮食

肥胖、缺乏体力活动，吸烟，钠摄入过度和饮酒过量等是导致高血压的重要原因。低盐，低脂，高纤维，新鲜蔬菜和水果等健康膳食可预防高血压。

积极降压

积极降压达标，并选择对心脑肾具有保护作用的药物，全面控制危险因素是降压治疗的重要策略。

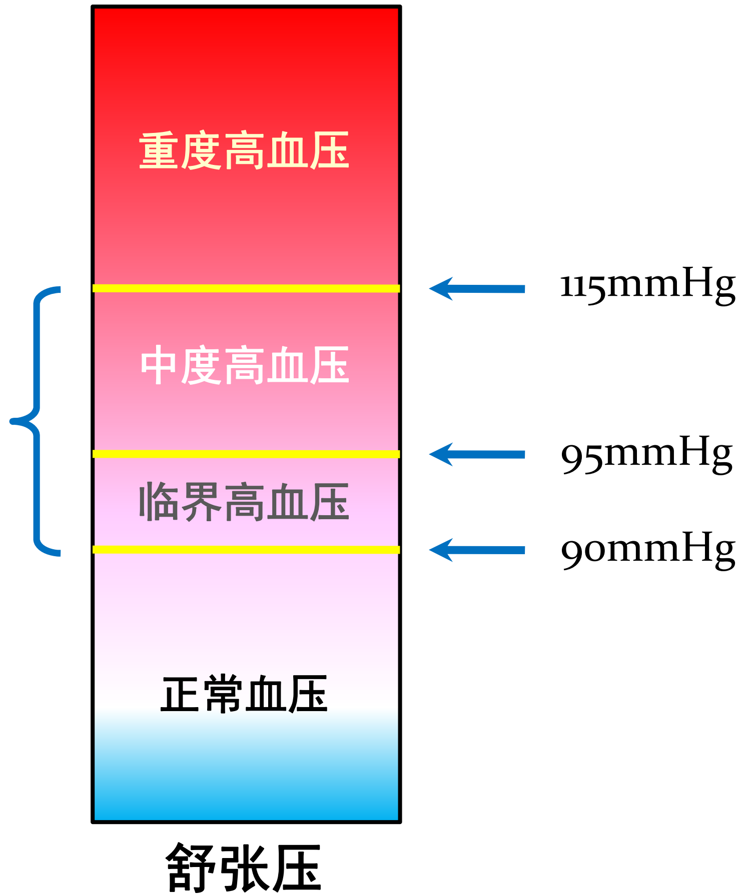
High blood pressure



收缩压（高压）：当心脏收缩时，从心室摄入的血液对血管壁产生的侧压力。
舒张压（低压）：心脏舒张末期，已流入动脉的血液对血管壁产生的侧压力。

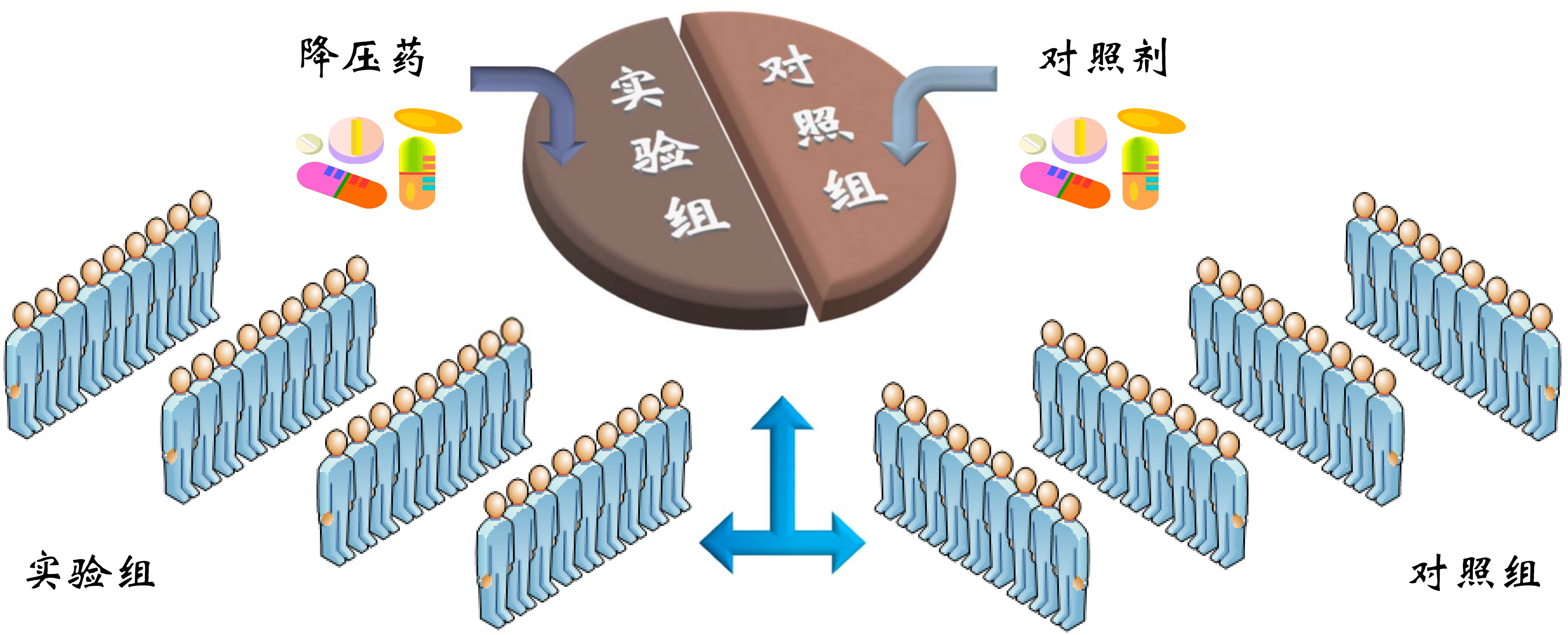
降压药对中度及临界高血压患者的作用

- 某制药公司研制一种降压药物
- 对于重度高血压患者
  - 该药物的降压作用已经清楚了
- 对于中度和临界高血压患者
  - 该药物的降压作用还不清楚
- 通过实验研究降压药物对于中度和临界高血压患者(90-115 mmHg)的降压作用



实验设计

- 从**中度和临界高血压患者**中**随机抽取389**名，分为两组
- 实验组**193**名患者，服用降压药
- 对照组**196**名患者，服用对照剂



数据收集

编号	患者姓名	2000年1月1日	2000年1月8日	2000年1月15日	...	2000年12月31日
1	张三	100	98	99	...	77
2	李四	99	98	99	...	88
3	王五	107	110	108	...	100
4	杨一	96	94	95	...	83
5	丁二	104	104	103	...	108
6	周七	96	96	93	...	64
...						
193	吴九	110	108	109	...	103

- 测量初始血压
- 定期测量血压
- 获得原始数据

原始数据 — 实验组初始血压

100	99	107	96	104	96	94	105	97	91
105	114	93	98	102	98	104	102	100	93
113	99	99	97	102	102	105	114	110	109
115	91	94	112	95	102	95	91	104	111
101	107	114	101	102	99	108	91	94	110
106	114	93	95	95	90	94	96	94	112
110	96	114	97	99	115	106	103	103	106
94	93	92	90	104	113	102	93	95	92
114	105	97	93	95	95	102	115	104	104
108	90	94	109	109	95	95	105	111	109
90	114	94	96	98	105	114	100	113	115
115	92	99	103	93	99	102	114	102	96
98	109	96	112	115	98	109	96	105	106
92	93	93	91	100	114	106	115	96	95
113	99	110	110	104	114	102	92	92	95
108	110	101	99	113	111	111	103	100	91
96	94	91	108	102	93	90	93	109	108
108	114	111	90	104	100	90	95	109	101
98	113	103	96	110	96	97	96	94	91
97	113	110							

原始数据 — 实验组12个月后血压

77	88	100	83	108	64	74	87	104	59
91	96	99	108	98	83	83	92	110	73
98	89	88	87	102	89	91	108	119	90
105	72	88	117	78	96	80	97	96	92
85	109	108	87	80	88	85	85	82	98
106	114	90	81	61	84	64	79	77	111
80	65	97	89	107	92	94	112	70	110
76	77	64	80	96	93	94	73	87	111
102	94	78	81	85	75	91	113	87	80
90	77	119	113	96	94	88	103	89	84
71	106	77	86	71	104	108	81	117	99
86	96	91	87	77	100	95	83	93	96
94	93	91	106	101	95	109	127	95	83
78	69	82	89	120	109	85	114	94	92
92	94	100	98	104	109	112	66	83	85
124	95	75	100	95	92	95	113	100	95
87	90	92	101	97	82	82	88	90	103
106	114	93	57	76	77	72	82	101	98
96	114	109	76	99	90	95	64	67	69
80	101	103							

原始数据 —— 对照组初始血压

97	105	110	103	90	94	115	111	114	99
105	113	110	103	90	106	93	93	91	113
113	91	100	99	104	96	114	98	101	92
106	106	95	94	98	98	109	93	112	104
105	91	113	111	115	109	98	108	114	115
103	102	113	113	104	110	112	97	112	98
103	99	100	104	104	115	99	103	113	107
97	96	107	115	114	102	103	96	93	94
101	90	91	107	100	109	92	90	112	98
99	108	97	97	113	106	91	96	91	100
110	109	105	96	115	113	107	109	96	102
92	96	113	113	112	100	104	97	101	115
110	109	103	115	94	102	94	94	94	111
99	110	112	109	95	98	107	93	111	96
105	114	99	91	111	102	105	91	104	111
113	92	102	91	112	114	101	107	112	94
95	110	105	97	91	106	112	94	99	110
93	91	110	101	109	115	114	108	111	94
109	97	112	115	113	110	105	114	115	90
92	104	109	104	115	90				

原始数据 —— 对照组12个月后血压

82	103	116	94	87	93	124	126	131	102
115	103	92	105	105	103	92	103	96	133
99	85	103	109	101	97	130	98	101	87
112	92	96	102	89	108	115	83	116	101
93	96	130	113	135	112	90	92	102	102
97	107	130	121	99	102	103	109	105	77
93	97	96	86	110	107	91	113	133	112
86	77	94	134	108	92	101	104	95	81
112	98	91	90	100	93	69	110	91	92
103	103	85	80	93	100	93	91	96	102
110	124	106	100	133	128	126	92	91	92
78	104	117	133	111	110	116	92	106	110
130	116	110	111	94	100	95	94	95	111
99	110	102	116	99	98	107	67	113	102
125	137	97	102	107	95	125	95	107	131
136	90	113	87	105	134	105	110	132	109
97	100	107	81	90	100	115	75	106	116
100	93	132	105	103	135	79	105	134	87
106	102	122	130	105	142	132	136	132	102
99	106	106	96	102	72				

实验组 VS. 对照组

1

98	113	103	96	110	96	97	96	94	91
97	113	110							

2

77	88	100	83	108	64	74	87	104	59
91	96	99	108	98	83	83	92	110	73
98	89	88	87	102	89	91	108	119	90
105	72	88	117	78	96	80	97	96	92
85	109	108	87	80	88	85	85	82	98
106	114	90	81	61	84	64	79	77	111
80	65	97	89	107	92	94	112	70	110
76	77	64	80	96	93	94	73	87	111
102	94	78	81	85	75	91	113	87	80
90	77	119	113	96	94	88	103	89	84
71	106	77	86	71	104	108	81	117	99
86	96	91	87	77	100	95	83	93	96
94	93	91	106	101	95	109	127	95	83
78	69	82	89	120	109	85	114	94	92
92	94	100	98	104	109	112	66	83	85
124	95	75	100	95	92	95	113	100	95
87	90	92	101	97	82	82	88	90	103
106	114	93	57	76	77	72	82	101	98
96	114	109	76	99	90	95	64	67	69
80	101	103							

3

109	97	112	115	113	110	105	114	115	90
92	104	109	104	115	90				

4

82	103	116	94	87	93	124	126	131	102
115	103	92	105	105	103	92	103	96	133
99	85	103	109	101	97	130	98	101	87
112	92	96	102	89	108	115	83	116	101
93	96	130	113	135	112	90	92	102	102
97	107	130	121	99	102	103	109	105	77
93	97	96	86	110	107	91	113	133	112
86	77	94	134	108	92	101	104	95	81
112	98	91	90	100	93	69	110	91	92
103	103	85	80	93	100	93	91	96	102
110	124	106	100	133	128	126	92	91	92
78	104	117	133	111	110	116	92	106	110
130	116	110	111	94	100	95	94	95	111
99	110	102	116	99	98	107	67	113	102
125	137	97	102	107	95	125	95	107	131
136	90	113	87	105	134	105	110	132	109
97	100	107	81	90	100	115	75	106	116
100	93	132	105	103	135	79	105	134	87
106	102	122	130	105	142	132	136	132	102
99	106	106	96	102	72				

数据的组织 — 实验组血压变化量

-23	-11	-7	-13	4	-32	-20	-18	7	-32
-14	-18	6	10	-4	-15	-21	-10	10	-20
-15	-10	-11	-10	0	-13	-14	-6	9	-19
-10	-19	-6	5	-17	-6	-15	6	-8	-19
-16	2	-6	-14	-22	-11	-23	-6	-12	-12
0	0	-3	-14	-34	-6	-30	-17	-17	-1
-30	-31	-17	-8	8	-23	-12	9	-33	4
-18	-16	-28	-10	-8	-20	-8	-20	-8	19
-12	-11	-19	-12	-10	-20	-11	-2	-17	-24
-18	-13	25	4	-13	-1	-7	-2	-22	-25
-19	-8	-17	-10	-27	-1	-6	-19	4	-16
-29	4	-8	-16	-16	1	-7	-31	-9	0
-4	-16	-5	-6	-14	-3	0	31	-10	-23
-14	-24	-11	-2	20	-5	-21	-1	-2	-3
-21	-5	-10	-12	0	-5	10	-26	-9	-10
16	-15	-26	1	-18	-19	-16	10	0	4
-9	-4	1	-7	-5	-11	-8	-5	-19	-5
-2	0	-18	-33	-28	-23	-18	-13	-8	-3
-2	1	6	-20	-11	-6	-2	-32	-27	-22
-17	-12	-7							

数据的组织 — 对照组血压变化量

-15	-2	6	-9	-3	-1	9	15	17	3
10	-10	-18	2	-18	5	-3	-1	10	5
20	-14	-6	3	10	-3	1	16	0	0
-5	6	-14	1	8	-9	10	6	-10	4
-3	-12	5	17	2	20	3	-8	-16	-12
-13	-6	5	17	8	-5	-8	-9	12	-7
-21	-10	-2	-4	-18	6	-8	-8	10	20
5	-11	-19	-13	19	-6	-10	-2	8	2
-13	11	8	0	-17	0	-16	-23	20	-21
-6	4	-5	-12	-17	-20	-6	2	-5	5
2	0	15	1	4	18	15	19	-17	-5
-10	-14	8	4	20	-1	10	12	-5	5
-5	20	7	7	-4	0	-2	1	0	1
0	0	0	-10	7	4	0	0	-26	2
6	20	23	-2	11	-4	-7	20	4	3
20	23	-2	11	-4	-7	20	4	3	20
15	2	-10	2	-16	-1	-6	3	-19	7
6	7	2	22	4	-6	20	-35	-3	23
-7	-3	5	10	15	-8	32	27	22	17
12	7	2	-3	-8	-13				

数据的组织

1

98	113	103	96	110	96	97	96	94	91
97	113	110							

2

96	114	109	76	99	90	95	64	67	69
80	101	103							

3

109	97	112	115	113	110	105	114	115	90
92	104	109	104	115	90				

4

82	103	116	94	87	93	124	126	131	102
115	103	92	105	105	103	92	103	96	133
99	85	103	109	101	97	130	98	101	87
112	92	96	102	89	108	115	83	116	101
93	96	130	113	135	112	90	92	102	102
97	107	130	121	99	102	103	109	105	77
93	97	96	86	110	107	91	113	133	112
86	77	94	134	108	92	101	104	95	81
112	98	91	90	100	93	69	110	91	92
103	103	85	80	93	100	93	91	96	102
110	124	106	100	133	128	126	92	91	92
78	104	117	133	111	110	116	92	106	110
130	116	110	111	94	100	95	94	95	111
99	110	102	116	99	98	107	67	113	102
125	137	97	102	107	95	125	95	107	131
136	90	113	87	105	134	105	110	132	109
97	100	107	81	90	100	115	75	106	116
100	93	132	105	103	135	79	105	134	87
106	102	122	130	105	142	132	136	132	102
99	106	106	96	102	72				

1

-2	1	6	-20	-11	-6	-2	-32	-27	-22
-17	-12	-7							

2

-15	-2	6	-9	-3	-1	9	15	17	3
10	-10	-18	2	-18	5	-3	-1	10	5
20	-14	-6	3	10	-3	1	16	0	0
-5	6	-14	1	8	-9	10	6	-10	4
-3	-12	5	17	2	20	3	-8	-16	-12
-13	-6	5	17	8	-5	-8	-9	12	-7
-21	-10	-2	-4	-18	6	-8	-8	10	20
5	-11	-19	-13	19	-6	-10	-2	8	2
-13	11	8	0	-17	0	-16	-23	20	-21
-6	4	-5	-12	-17	-20	-6	2	-5	5
2	0	15	1	4	18	15	19	-17	-5
-10	-14	8	4	20	-1	10	12	-5	5
-5	20	7	7	-4	0	-2	1	0	1
0	0	0	-10	7	4	0	0	-26	2
6	20	23	-2	11	-4	-7	20	4	3
20	23	-2	11	-4	-7	20	4	3	20
15	2	-10	2	-16	-1	-6	3	-19	7
6	7	2	22	4	-6	20	-35	-3	23
-7	-3	5	10	15	-8	32	27	22	17
12	7	2	-3	-8	-13				

数据的表述 — 实验组血压变化量的符号

1

-23	-11	-7	-13	4	-32	-20	-18	7	-32
-14	-18	6	10	-4	-15	-21	-10	10	-20
-15	-10	-11	-10	0	-13	-14	-6	9	-19
-10	-19	-6	5	-17	-6	-15	6	-8	-19
-16	2	-6	-14	-22	-11	-23	-6	-12	-12
0	0	-3	-14	-34	-6	-30	-17	-17	-1
-30	-31	-17	-8	8	-23	-12	9	-33	4
-18	-16	-28	-10	-8	-20	-8	-20	-8	19
-12	-11	-19	-12	-10	-20	-11	-2	-17	-24
-18	-13	25	4	-13	-1	-7	-2	-22	-25
-19	-8	-17	-10	-27	-1	-6	-19	4	-16
-29	4	-8	-16	-16	1	-7	-31	-9	0
-4	-16	-5	-6	-14	-3	0	31	-10	-23
-14	-24	-11	-2	20	-5	-21	-1	-2	-3
-21	-5	-10	-12	0	-5	10	-26	-9	-10
16	-15	-26	1	-18	-19	-16	10	0	4
-9	-4	1	-7	-5	-11	-8	-5	-19	-5
-2	0	-18	-33	-28	-23	-18	-13	-8	-3
-2	1	6	-20	-11	-6	-2	-32	-27	-22
-17	-12	-7							

157名患者的血压值降低了，36名患者的血压值不变或升高了

数据的表述 —— 对照组血压变化量的符号

2

-15	-2	6	-9	-3	-1	9	15	17	3
10	-10	-18	2	-18	5	-3	-1	10	5
20	-14	-6	3	10	-3	1	16	0	0
-5	6	-14	1	8	-9	10	6	-10	4
-3	-12	5	17	2	20	3	-8	-16	-12
-13	-6	5	17	8	-5	-8	-9	12	-7
-21	-10	-2	-4	-18	6	-8	-8	10	20
5	-11	-19	-13	19	-6	-10	-2	8	2
-13	11	8	0	-17	0	-16	-23	20	-21
-6	4	-5	-12	-17	-20	-6	2	-5	5
2	0	15	1	4	18	15	19	-17	-5
-10	-14	8	4	20	-1	10	12	-5	5
-5	20	7	7	-4	0	-2	1	0	1
0	0	0	-10	7	4	0	0	-26	2
6	20	23	-2	11	-4	-7	20	4	3
20	23	-2	11	-4	-7	20	4	3	20
15	2	-10	2	-16	-1	-6	3	-19	7
6	7	2	22	4	-6	20	-35	-3	23
-7	-3	5	10	15	-8	32	27	22	17
12	7	2	-3	-8	-13				

85名患者的血压值降低了，111名患者的血压值不变或升高了

数据的表述 —— 血压变化量符号的对比

1

-23	-11	-7	-13	4	-32	-20	-18	7	-32
-14	-18	6	10	-4	-15	-21	-10	10	-20
-15	-10	-11	-10	0	-13	-14	-6	9	-19
-10	-19	-6	5	-17	-6	-15	6	-8	-19
-16	2	-6	-14	-22	-11	-23	-6	-12	-12
0	0	-3	-14	-34	-6	-30	-17	-17	-1
-30	-31	-17	-8	8	-23	-12	9	-33	4
-18	-16	-28	-10	-8	-20	-8	-20	-8	19
-12	-11	-19	-12	-10	-20	-11	-2	-17	-24
-18	-13	25	4	-13	-1	-7	-2	-22	-25
-19	-8	-17	-10	-27	-1	-6	-19	4	-16
-29	4	-8	-16	-16	1	-7	-31	-9	0
-4	-16	-5	-6	-14	-3	0	31	-10	-23
-14	-24	-11	-2	20	-5	-21	-1	-2	-3
-21	-5	-10	-12	0	-5	10	-26	-9	-10
16	-15	-26	1	-18	-19	-16	10	0	4
-9	-4	1	-7	-5	-11	-8	-5	-19	-5
-2	0	-18	-33	-28	-23	-18	-13	-8	-3
-2	1	6	-20	-11	-6	-2	-32	-27	-22
-17	-12	-7							

2

-15	-2	6	-9	-3	-1	9	15	17	3
10	-10	-18	2	-18	5	-3	-1	10	5
20	-14	-6	3	10	-3	1	16	0	0
-5	6	-14	1	8	-9	10	6	-10	4
-3	-12	5	17	2	20	3	-8	-16	-12
-13	-6	5	17	8	-5	-8	-9	12	-7
-21	-10	-2	-4	-18	6	-8	-8	10	20
5	-11	-19	-13	19	-6	-10	-2	8	2
-13	11	8	0	-17	0	-16	-23	20	-21
-6	4	-5	-12	-17	-20	-6	2	-5	5
2	0	15	1	4	18	15	19	-17	-5
-10	-14	8	4	20	-1	10	12	-5	5
-5	20	7	7	-4	0	-2	1	0	1
0	0	0	-10	7	4	0	0	-26	2
6	20	23	-2	11	-4	-7	20	4	3
20	23	-2	11	-4	-7	20	4	3	20
15	2	-10	2	-16	-1	-6	3	-19	7
6	7	2	22	4	-6	20	-35	-3	23
-7	-3	5	10	15	-8	32	27	22	17
12	7	2	-3	-8	-13				

实验组
157/193 （81.3%）

对照组
85/196 （43.4%）

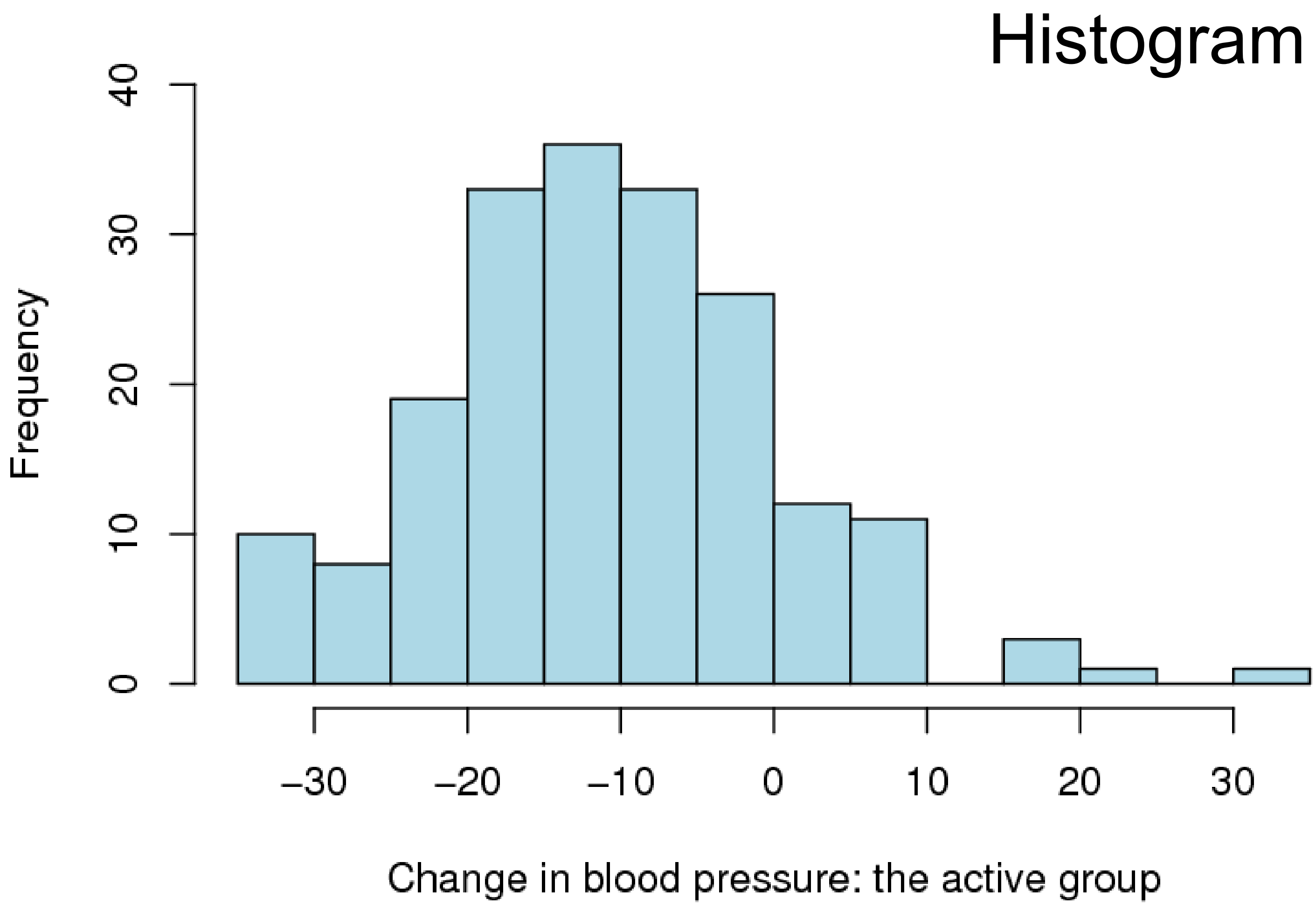
数据的表述 —— 实验组血压变化量的频数

Bin Limits		Bin Boundaries		Bin	Bin	Relative Bin	Percentage of
Lower	Upper	Lower	Upper	Mark	Frequency	Frequency	Observations
-35	-31	-35.5	-30.5	-33	10	0.0518	5.18
-30	-26	-30.5	-25.5	-28	8	0.0415	4.15
-25	-21	-25.5	-20.5	-23	19	0.0984	9.84
-20	-16	-20.5	-15.5	-18	33	0.1710	17.10
-15	-11	-15.5	-10.5	-13	36	0.1865	18.65
-10	-6	-10.5	-5.5	-8	33	0.1710	17.10
-5	-1	-5.5	-0.5	-3	26	0.1350	13.50
0	4	-0.5	4.5	2	12	0.0622	6.22
5	9	4.5	9.5	7	11	0.0570	5.70
10	14	9.5	14.5	12	0	0.0000	0.00
15	19	14.5	19.5	17	3	0.0155	1.55
20	24	19.5	24.5	22	1	0.0052	0.52
25	29	24.5	29.5	27	0	0.0000	0.00
30	34	29.5	34.5	32	1	0.0052	0.52

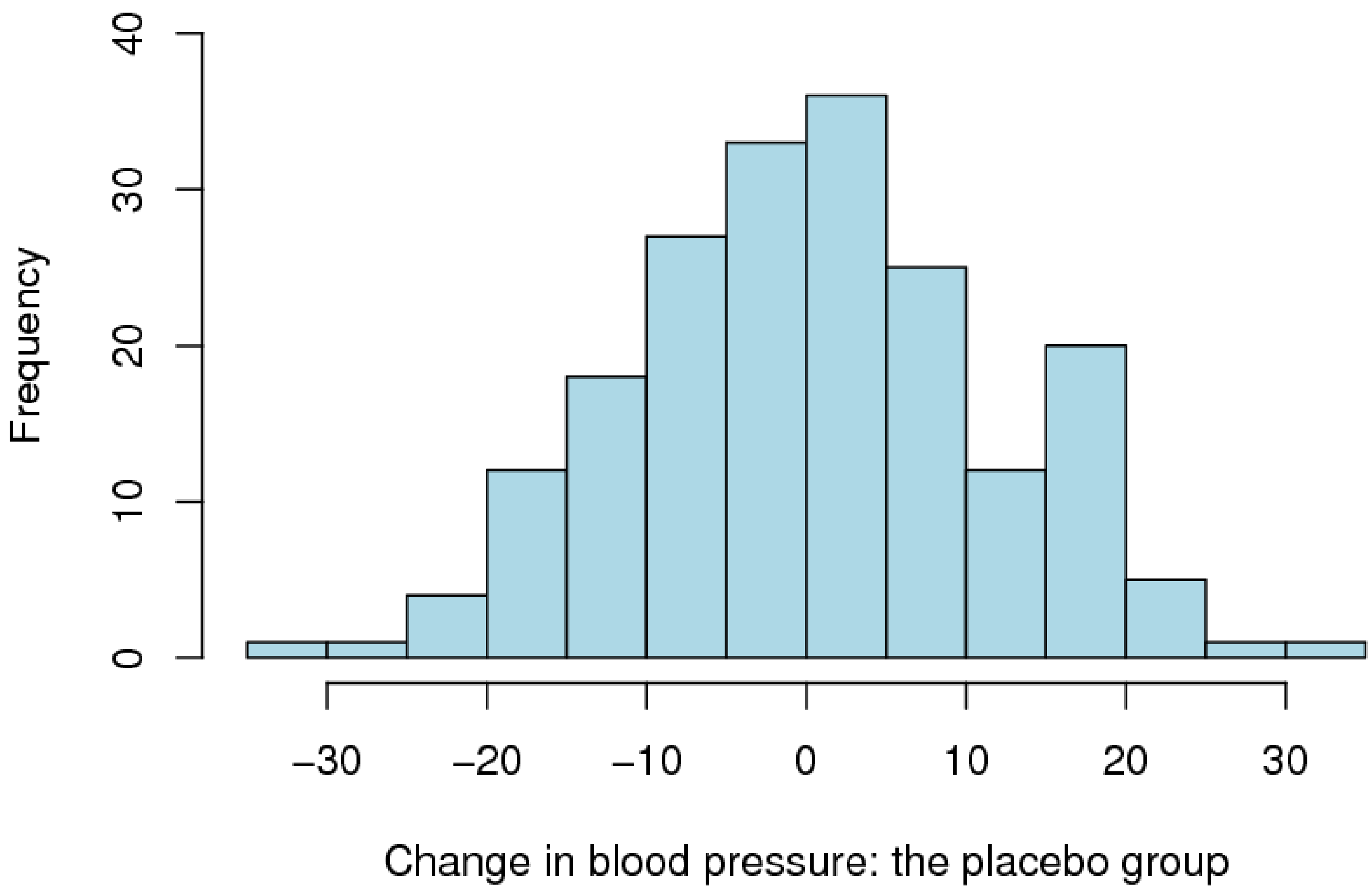
数据的表述 —— 对照组血压变化量的频数

Bin Limits Lower	Bin Limits Upper	Bin Boundaries Lower	Bin Boundaries Upper	Bin Mark	Bin Frequency	Relative Bin Frequency	Percentage of Observations
-35	-31	-35.5	-30.5	-33	1	0.0051	0.51
-30	-26	-30.5	-25.5	-28	1	0.0051	0.51
-25	-21	-25.5	-20.5	-23	4	0.0204	2.04
-20	-16	-20.5	-15.5	-18	12	0.0612	6.12
-15	-11	-15.5	-10.5	-13	18	0.0918	9.18
-10	-6	-10.5	-5.5	-8	27	0.1378	13.78
-5	-1	-5.5	-0.5	-3	33	0.1684	16.84
0	4	-0.5	4.5	2	36	0.1837	18.37
5	9	4.5	9.5	7	25	0.1276	12.76
10	14	9.5	14.5	12	12	0.0612	6.12
15	19	14.5	19.5	17	20	0.102	10.2
20	24	19.5	24.5	22	5	0.0255	2.55
25	29	24.5	29.5	27	1	0.0051	0.51
30	34	29.5	34.5	32	1	0.0051	0.51

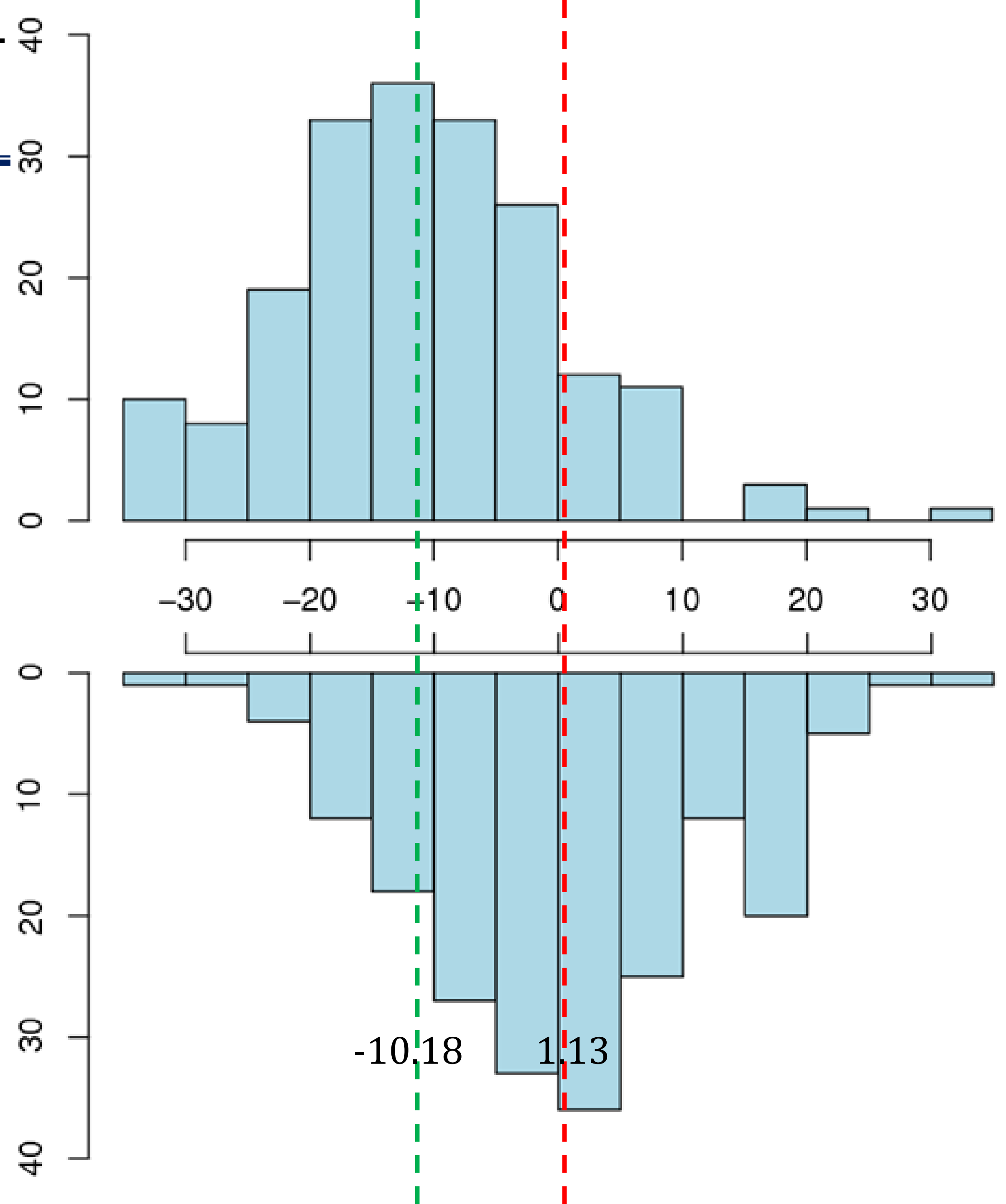
数据的表述 —— 血压变化量的直方图



数据的表述 — 血压变化量的直方图



数据的表述 —— 直方图对比



数据的概括 — 均值 (Mean)

- 数据集聚位置的一种度量
- 一组数据的平均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 实验组血压变化量的均值 -10.18
- 对照组血压变化量的均值 1.13

数据的概括 —— 方差与标准差 (Variance and standard derivation)

- 数据分散程度的度量
- 方差：一组数据相对于均值的均方值

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- 标准差：方差的正平方根
- 实验组血压变化量的标准差 11.16
- 对照组血压变化量的标准差 11.70

数据的分析 —— 提出问题

- 降压药物是否有降压作用？
- 降压药物的降压作用有多大？
- 降压药物是否能够阻止高血压的恶化？
- 降压药物是否有助于缓解高血压症状？
- 降压药物是否有助于缓解冠心病？

数据的分析 —— 是否有降压作用？

- 实验数据

	实验组	对照组
患者数量	193	196
服用药物12个月后的血压变化的平均值	-10.18	1.13
服用药物12个月后的血压变化的标准差	11.16	11.70

- 直观推测
  - 降压药有降压作用，因为 -10.18 比 1.13 小很多
- 统计学方法
  - 进行两正态总体期望值是否相等的**假设检验** —— **不相等**

数据的分析 —— 降压作用有多大？

- 实验数据

	实验组	对照组
患者数量	193	196
服用药物12个月后的血压变化的平均值	-10.18	1.13

- 直观推测
  - 降压药的降压作用为 -10.18
- 统计学方法
  - 进行正态总体均值的**点估计**

数据的分析 —— 降压作用有多大？

- 实验数据

	实验组	对照组
患者数量	193	196
服用药物12个月后的血压变化的平均值	-10.18	1.13
服用药物12个月后的血压变化的标准差	11.16	11.70

- 直观推测
  - 降压作用的真实值被怎样一个区间以较高的可信度覆盖？
- 统计学方法
  - 进行正态总体均值的**区间估计** —— [-11.8, -8.6]

数据的分析 — 是否能够阻止高血压的恶化？

- 实验数据

	实验组	对照组
患者数量	193	196
中度高血压恶化为重度高血压患者的数量	0	24
病情恶化患者的比例	0	0.12

- 直观推测
  - 降压药物能够阻止高血压的恶化
- 统计学方法
  - 两比例是否相等的**假设检验** — **不相等**

数据的分析 — 是否有助于缓解高血压症状？

- 实验数据

	实验组	对照组
患者数量	193	196
出现高血压症状患者的数量	37	89
出现高血压症状患者的比例	0.19	0.45

- 直观推测
  - 降压药物有助于缓解高血压症状
- 统计学方法
  - 两比例是否相等的**假设检验**　　　　　— **不相等**

数据的分析 —是否有助于缓解冠心病？

- 实验数据

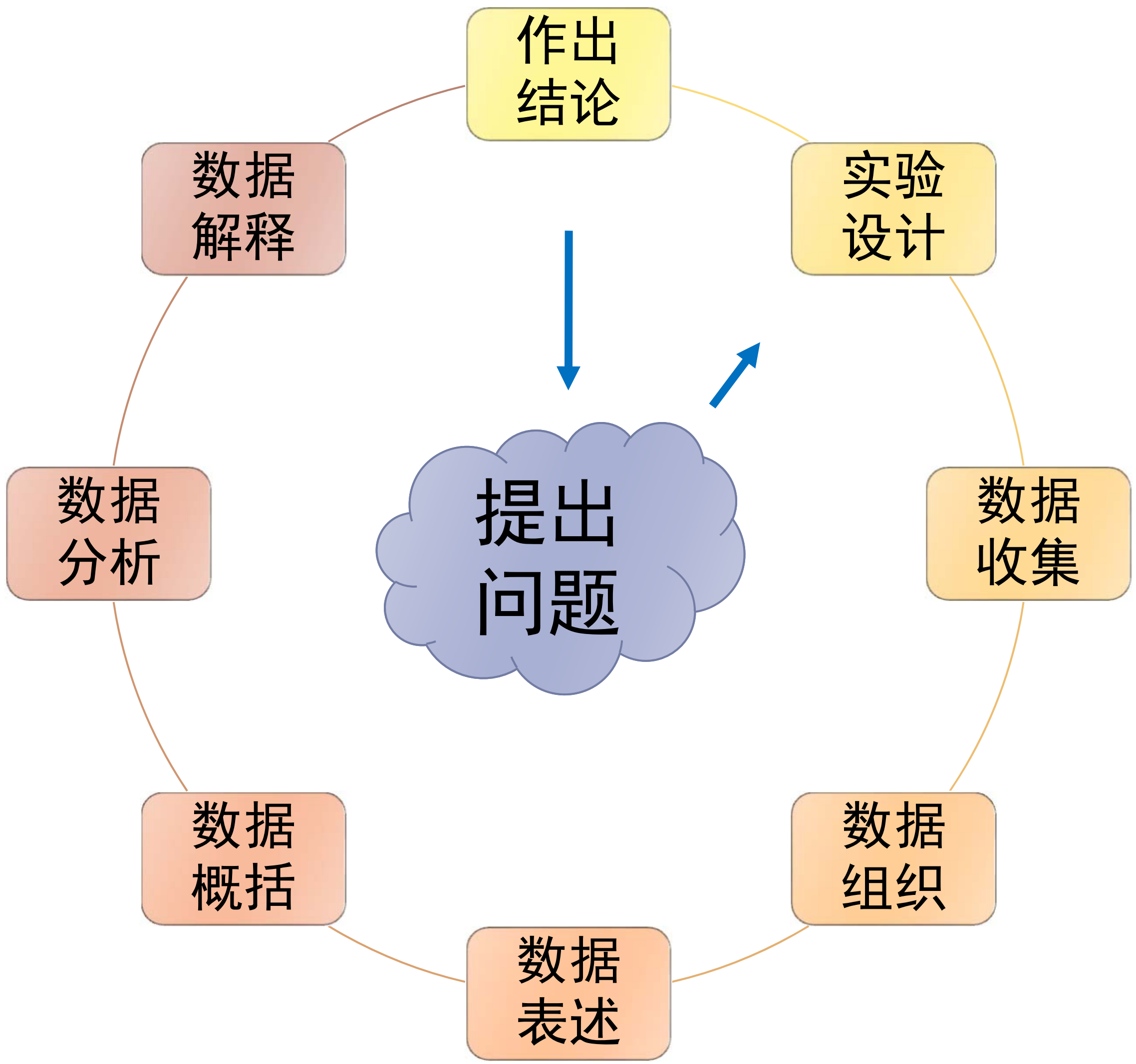
	实验组	对照组
患者数量	193	196
冠心病患者数量	35	38
冠心病患者比例	0.18	0.19

- 直观推测
  - 降压药物对缓解冠心病没有帮助
- 统计学方法
  - 两比例是否相等的**假设检验** — **相等**

研究结论

- 降压药物是否有降压作用？ — 是
- 降压药物的降压作用有多大？ — 10
- 降压药物是否能够阻止高血压的恶化？ — 是
- 降压药物是否有助于缓解高血压症状？ — 是
- 降压药物是否有助于缓解冠心病？ — 否

研究过程总结



统计学的定义

统计学

统计学是一门关于实验设计和数据收集，以及对实验数据进行组织、表述、概括、分析、解释，并最终作出结论的应用科学。

Statistics

- 来源于拉丁语 ***statisticum collegium*** (国会, council of state)和意大利语 ***statista*** (政治家, statesman)
- 德语 ***statistik*** 由 Gottfried Achenwall 于1749年使用
  - 表示对国家的资料进行分析的学问，也就是“研究国家的科学”
- 英语 ***statistics*** 由 Sir John Sinclair 于1791-1799年使用
  - 最初是政府（通常是中央政府）以及管理阶层通对国家资料的收集和分析掌握国家信息的工具
  - 现在已经发展为数学的一个分支，并被广泛应用于自然科学和社会科学的各个领域中

Sir John Sinclair

2011年国务院学位委员会批准统计学为一级学科

Gottfried Achenwall

Contents

Academic calendar

	M	T	W	T	F	S	S	
				1	2	3	4	September
	5	6	7	8	9	10	11	
1	12	13	14	15	16	17	18	
2	19	20	21	**22**	23	24	25	
3	26	27	28	**29**	30	1	2	
4	3	4	5	6	7	**8**	9	October
5	10	11	12	**13**	14	15	16	
6	17	18	19	**20**	21	22	23	
7	24	25	26	**27**	28	29	30	
8	31	1	2	**3**	4	5	6	November
9	7	8	9	**10**	11	12	13	
10	14	15	16	**17**	18	19	20	
11	21	22	23	**24**	25	26	27	
12	28	29	30	**1**	2	3	4	
13	5	6	7	**8**	9	10	11	December
14	12	13	14	**15**	16	17	18	
15	19	20	21	**22**	23	24	25	
16	26	27	28	**29**	30	31	1	

2016 年秋季学期调课示意图

周	月	星期一	星期二	星期三	星期四	星期五	星期六	星期日
1	九月	12	13	14	15 原排课程 停上	16 原排课程 停上	17 原排课程 照常进行	18 原排课程 照常进行
2		19	20	21	22	23	24	25
3		26	27	28	29	30		
	十月						1 原排课程 停上	2 原排课程 停上
4		3 原排课程 停上	4 原排课程 停上	5 原排课程 停上	6 改上 8 日 （周六）课程	7 改上 9 日 （周日）课程	8 改上 6 日 （周四）课程	9 改上 7 日 （周五）课程
5		10	11	12	13	14	15	16

16	十二月	26	27	28	29	30	31 原排课程 照常进行	
	2017 一月							1 原排课程 停上
17		2	3 开始 期末考试	4	5	6	7	8

Fundamentals of statistics

- 概率论基础 Lecture 1
- 随机变量 Lecture 2
- 随机向量 Lecture 3
- 随机抽样 Lecture 4

Descriptive statistics

- 数据的简约
- 数据的概括
- 数据的表述

Lecture 4

Inferential statistics

- 点估计 (Point estimation) Lecture 5,6,7
- 期中考试 (Middle term) Lecture 8
- 假设检验 (Hypothesis testing) Lecture 9,10,11
- 区间估计 (Interval estimation) Lecture 12
- 方差分析 (Analysis of variance) Lecture 13
- 回归分析 (Linear regression) Lecture 14
- 二值分类 (Binary classification) Lecture 15

Contents

方法 应用

方法

EM 算法 | Gibbs采样 | Bayesian | 二值分类 | …

点估计 | 假设检验 | 区间估计 | 方差分析 | 回归分析

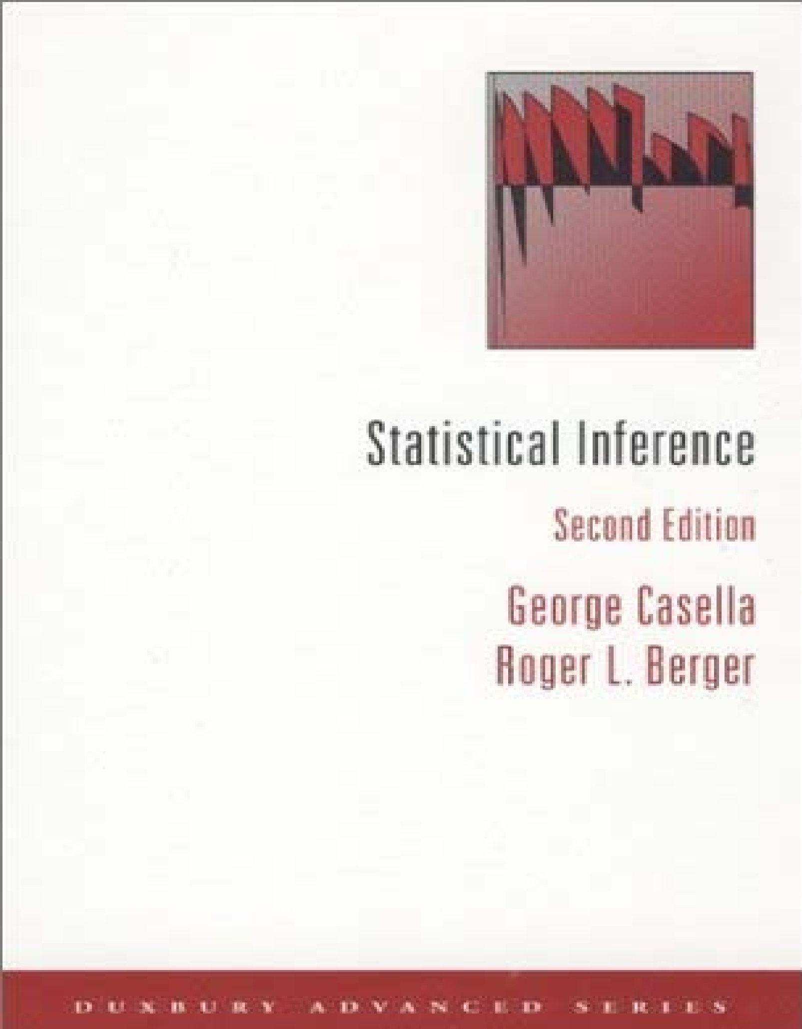
基础

随机变量 | 概率分布 | 随机抽样

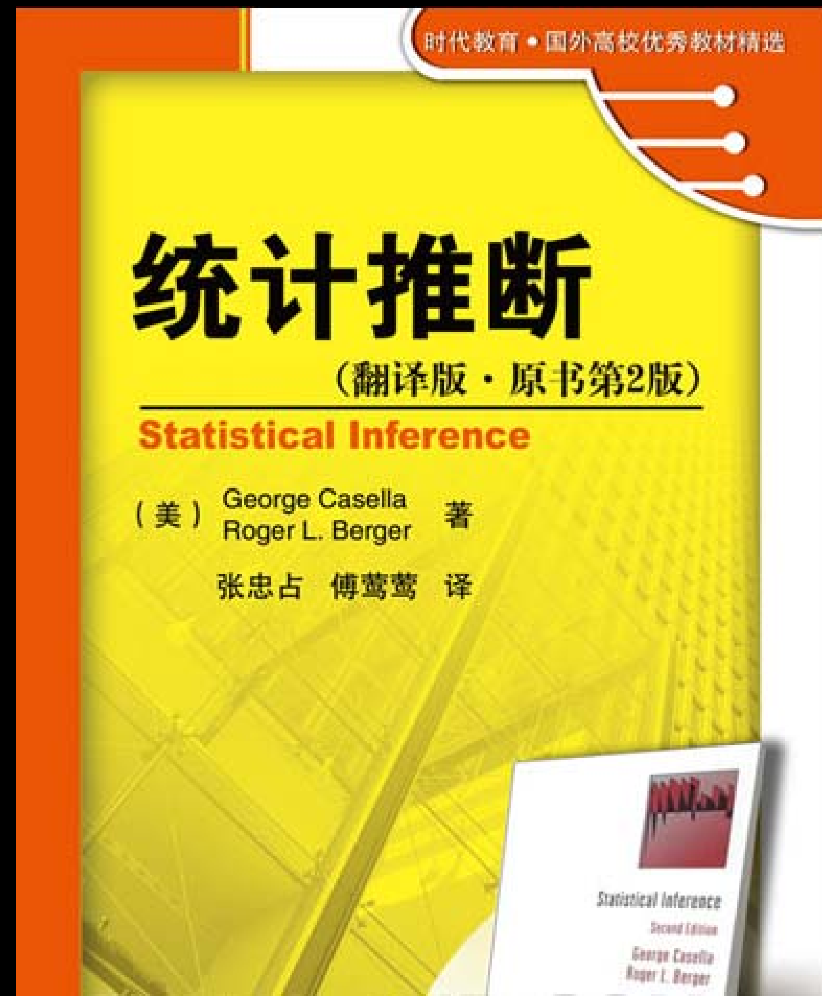
概率论 | 微积分

统计计算软件 R

George Casella and Roger L. Berger,
Statistical inference (*second edition*),
Duxbury Thomson learning, 2002
http://product.china-pub.com/27908



George Casella and Roger L. Berger 著，
张忠占，傅莺莺 译，
统计推断（翻译版，原书第二版）
机械工业出版社
http://product.china-pub.com/196285



薛毅，陈立萍，统计建模与R软件，清华大学出版社，2007

王静龙，梁小筠，非参数统计分析，高等教育出版社，2006

考核方式

- 小作业
- 大作业
- 期中考试
- 期末考试

助教

- 花　奎：　huak14@mails.tsinghua.edu.cn
- 曾婉雯：　zengww14@mails.tsinghua.edu.cn
- 李文然：　llwr15@mails.tsinghua.edu.cn
- 陈风玲：　cfl15@mails.tsinghua.edu.cn
- 奉雨娟：　fyj15@mails.tsinghua.edu.cn
- 崔佳欣：　cjx13@mails.tsinghua.edu.cn
- 黄　浩：　zlzr200599@163.com

课代表

- 刘桥
  - liuqiao@buaa.edu.cn

http://learn.tsinghua.edu.cn

Exercises

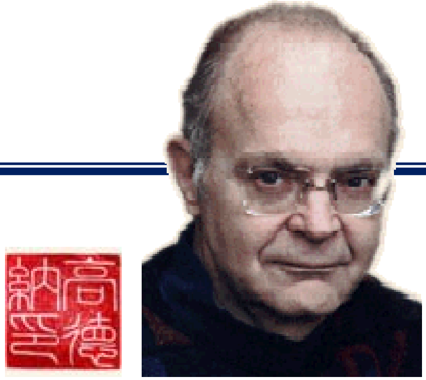
1. Pdf file generated from **latex**. (*****)
2. Pdf file generated from word. (****)
3. Doc file. (***)
4. Pdf file generated from handwriting. (**)
5. Paper. (*)

请不要抄作业！

发现后本门课记零分！

Exercises

LaTeX



Donald E. Knuth
Turing Award, 1974

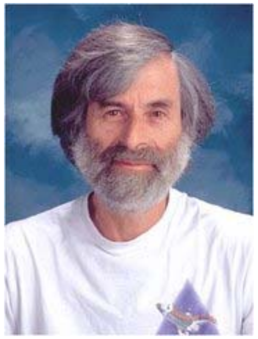
MiKTeX

```
\documentclass{article}
\title{Point estimation}
\author{Rui Jiang}
\date{September 2009}

\begin{document}
\maketitle

Hello world!

\end{document}
```



Leslie Lamport

The R Project for Statistical

统计学方法及其应用 | 概论 | 统计计算软件R

"R is a statistical computing program, make available through the Internet under the General Public License (GPL). It exists for Microsoft Windows (95 or later), for a variety of Unix and Linux platforms, and for the Apple platforms (Mac OS 8.6 or newer)."

The R Core Team

Statistical computing software

	R	S-Plus	SAS	SPSS
价格	免费	1-2万	8万+年费(55%)	1-2万+模块费
版权	GPL	永久版权	一年版权	永久版权
适合领域	自然科学 制造、金融、生物、医药 、…	自然科学 制造、金融、生物、医药 、…	管理科学 企业、资料、财务、会计、经济、…	社会科学 社会、教育、心理、行政、传播、…
产品定位	统计研究应用人员	统计研究应用人员	统计应用人员	统计应用人员
扩展性	具有优秀的扩展性，可自创或扩展新的统计分析方法。	具有优秀的扩展性，可自创或扩展新的统计分析方法	不具有对新方法的集成功能。只能随软件的更新进行扩展	无法编写新算法，只能使用软件提供的固定功能
操作界面	主要为命令行 操作灵活	命令行及图形界面 操作方便灵活	编程界面 操作困难	图形界面 操作简单
适用平台	几乎任何平台	几乎任何平台	Windows,Unix,Linux	Windows, Mac

The beginning of R

Ross Ihaka

The legend of R

R started in **1996** as a project by **Ross Ihaka** and **Robert Gentleman** at the University of Auckland, New Zealand, intended to provide a **statistical computing environment** in their teaching lab.

Robert Gentleman

R: A Language for Data Analysis and Graphics

Ross IHAKA and Robert GENTLEMAN

Ross Ihaka is Senior Lecturer, and Robert Gentleman is Senior Lecturer, Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand; e-mail: ihaka@stat.auckland.ac.nz.

©1996 *American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*

Journal of Computational and Graphical Statistics, Volume 5, Number 3, Pages 299–314

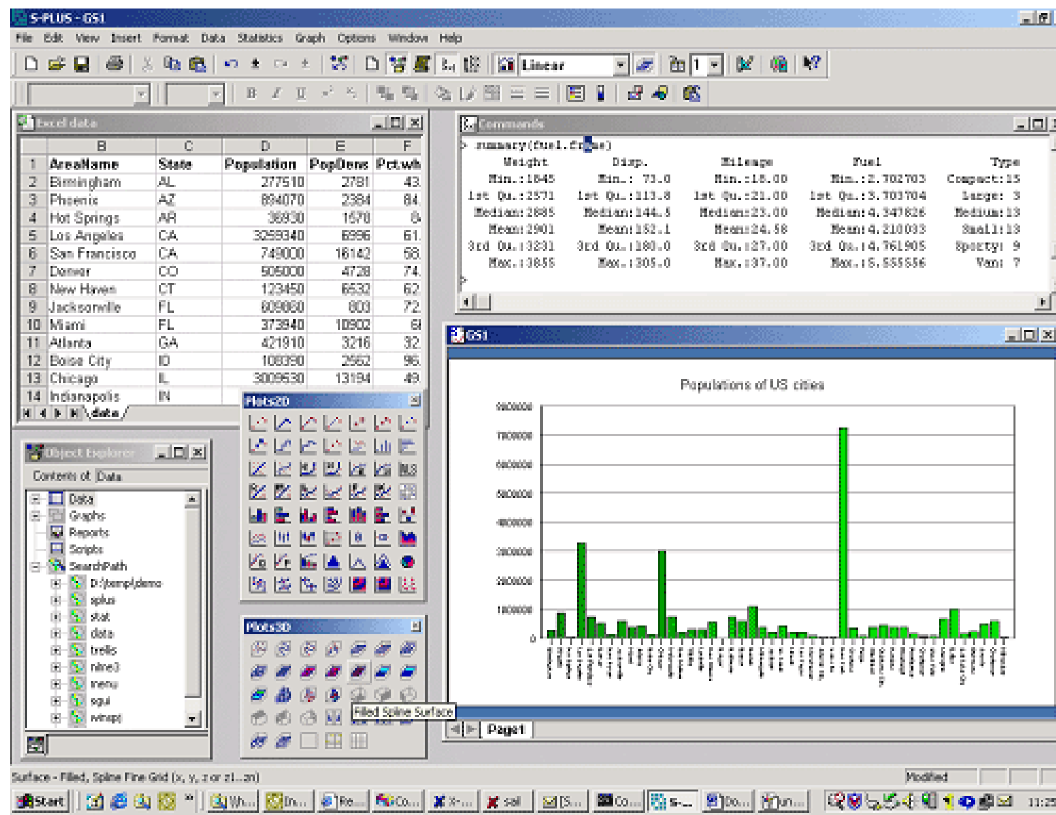
The S language

- An interactive environment for data analysis developed at ***Bell Laboratories*** since 1976
- Exclusively licensed by *AT&T/Lucent* to *Insightful Corporation*, Seattle, WA. Now it becomes the product named **S-PLUS**

AT&T Bell Laboratories

Bell Laboratories

Insightful the knowledge to act



The Scheme language

- A statically scoped and properly tail-recursive dialect of the **Lisp** programming language invented by **Guy Lewis Steele Jr.** and **Gerald Jay Sussman**

- **Scheme's underlying semantics + S's syntax = R**

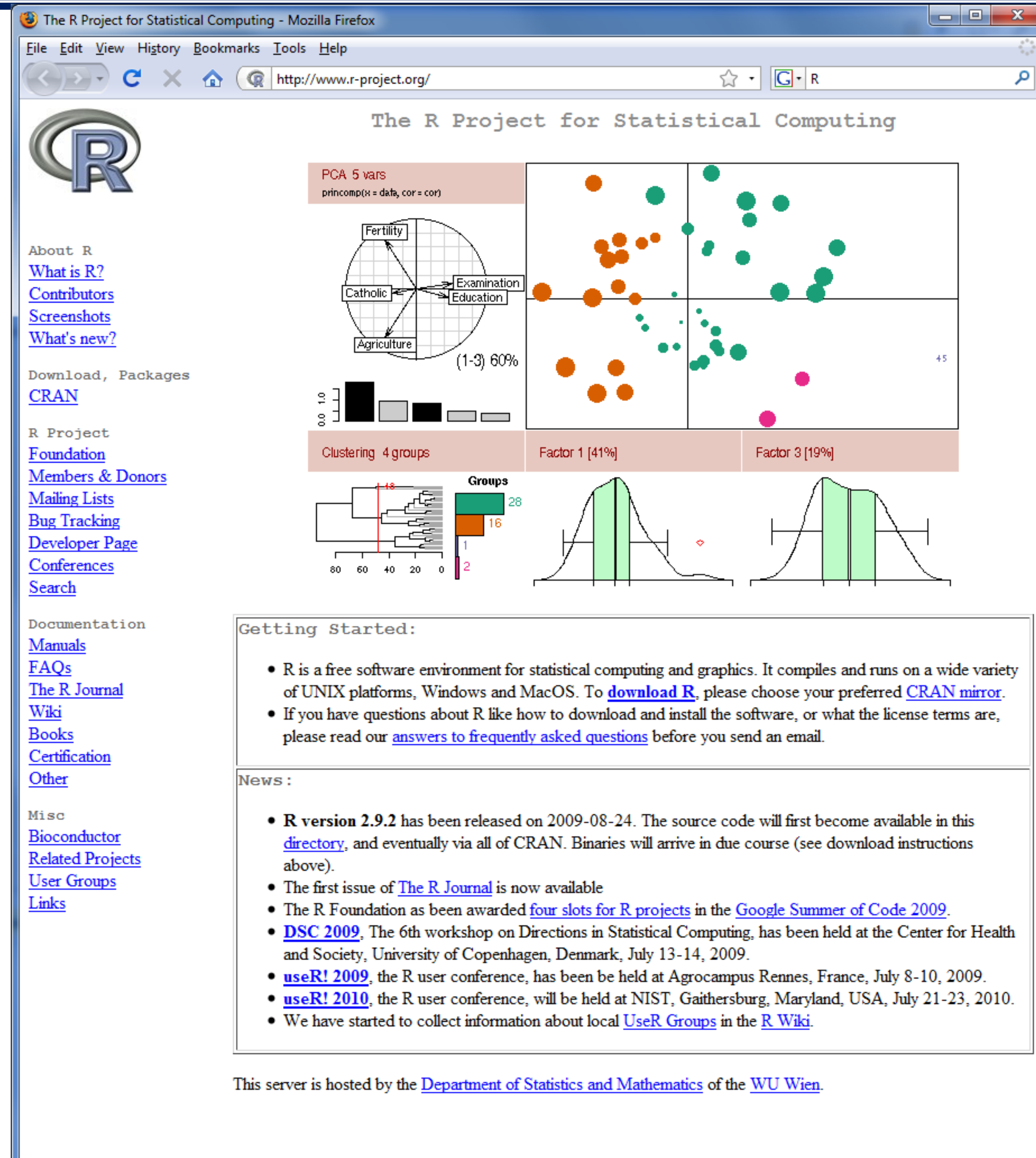
*We have named our language **R** — in part to acknowledge the influence of S and in part to celebrate our own efforts.*

Ross Ihaka
Robert Gentleman

The R project for statistical computing

- Maintained by the **R core team** since 1997, under the **G**eneral **P**ublic **L**icense
- Free
- Open source
- Cross-platform

http://www.r-project.org

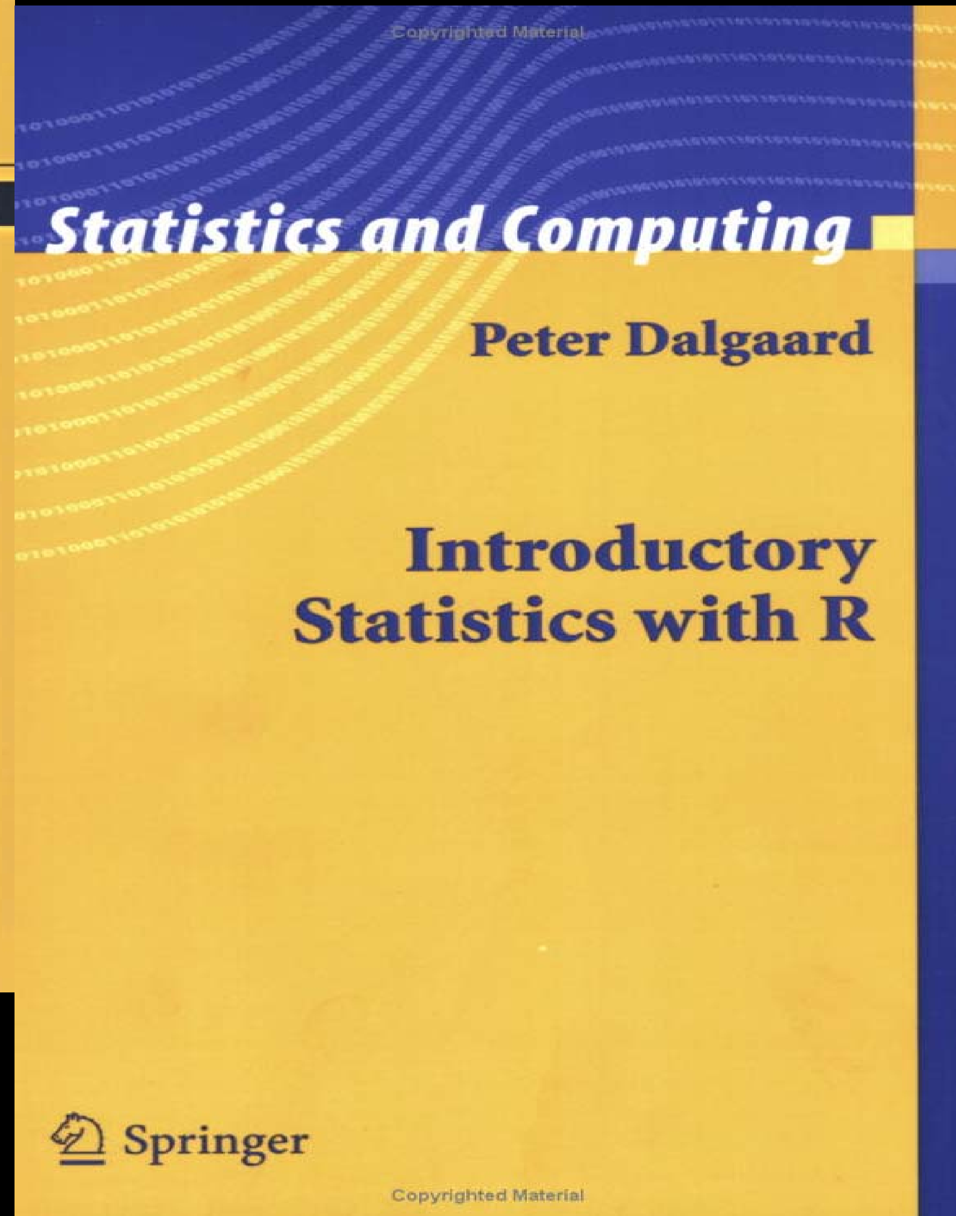


R Programming

统计学方法及其应用 概论 统计计算软件R

结合课件与参考书自学


R统计建模与R软件
薛毅 陈立萍 编著
http://www.tup.com.cn
清华大学出版社


Copyrighted Material
Statistics and Computing
Peter Dalgaard
Introductory
Statistics with R
Springer
Copyrighted Material

Comprehensive books

- 薛毅，陈立萍，统计建模与R软件，清华大学出版社，2007 第二章“R软件的使用”
- Introductory Statistics with R

  by Peter Dalgaard
- Using R for Introductory Statistics

  by John Verzani
- Data Analysis and Graphics Using R

  by John Maindonald and John Braun
- R Graphics
  by Paul Murrell

Tutorial documents

- An introduction to R
  - http://cran.r-project.org/doc/manuals/R-intro.pdf
- The R language definition
  - http://cran.r-project.org/doc/manuals/R-lang.pdf
- R Data Import/Export
  - http://cran.r-project.org/doc/manuals/R-data.pdf
- Writing R Extensions
  - http://cran.r-project.org/doc/manuals/R-exts.pdf
- Other documents
  - http://cran.r-project.org/other-docs.html
  - http://www.biosino.org/R/R-doc/ (中文)

Fundamentals of Statistics

统计学方法及其应用 | 统计学基础 | 概述

Fundamentals of statistics

- 概率论基础
- 随机变量
  - 随机变量的概念
  - 随机变量的变换
  - 随机变量的期望
  - 随机变量的分布
- 多维随机变量
  - 随机向量的概念
  - 随机向量的变换
  - 随机向量的期望
- 随机抽样
  - 随机抽样
  - 抽样分布

Basics of Probability Theory

统计学方法及其应用 | 统计学基础 | 概率论基础

"Probability is the chance that something is likely to happen or be the case. Probability theory is used extensively in areas such as statistics, social science and philosophy to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems."

Probability theory

Probability theory

The subject of probability theory is the foundation upon which all of statistics is built, providing **a means for modeling random phenomena**. Through these models, statisticians are able to draw inferences on the basis of the examination of only a part of the whole.

Classical probabilities

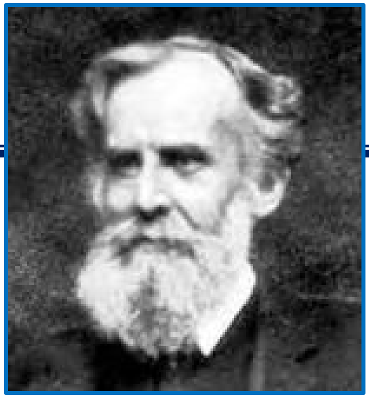
Pierre-Simon Laplace

Théorie analytique des probabilités

The probability of an event is **the ratio of the number of cases favorable to it, to the number of all cases possible** when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, **equally possible**.

http://en.wikipedia.org/wiki/Classical_definition_of_probability

Frequency probabilities



John Venn

Frequency probabilities

Probabilities are related to well-defined **random experiments**. The set of all possible outcomes of a random experiment is called the **sample space** of the experiment. An **event** is defined as a particular subset of the sample space. The relative frequency of occurrence of an event, in a number of repetitions of the experiment, is a measure of the **probability** of that event.

http://en.wikipedia.org/wiki/Frequency_probability

Subjective probabilities

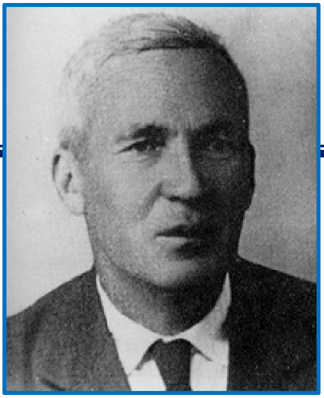
Thomas Bayes

Bayesian probability

Probability is the degree to which a person (or community) believes that a proposition is true, **the degree of belief**.

http://en.wikipedia.org/wiki/Bayesian_probability

Axiomatic definition



Kolmogorov

Probability axioms

The probability P of some event is defined in such a way that P satisfies the **Kolmogorov axioms**.

Euclidean axioms

公理1：任意一点到另外任意一点可以画直线
公理2：一条有限线段可以继续延长
公理3：以任意点为心及任意的距离可以画圆
公理4：凡直角都彼此相等
公理5：同平面内一条直线和另外两条直线相交，若在某一侧的两个内角和小于二直角的和，则这二直线经无限延长后在这一侧相交。

http://en.wikipedia.org/wiki/Kolmogorov_axioms

Random experiments（随机试验）

Random experiments

A **random experiment** is an experiment for which the outcome cannot be predicted with certainty. The term "random experiment" is often simplified as "experiment."

- 随机试验在相同的条件下可以重复进行
- 随机试验的所有可能结果能够事先明确地指出来
- 某一次随机试验的结果不能在试验进行之前预料到

Sample space（样本空间）

Sample space

The set, $\mathcal{S}$, of all possible outcomes of a particular random experiment is called the **sample space** for the experiment.

- 有限可列 (Finite countable)
- 无限可列 (Infinite countable)
- 无限不可列 (Infinite uncountable)

Examples of random experiments

- 随机试验
  - 掷一只骰子，观察朝上一面的点数
  - 在一批产品中，任取一件，观察是正品还是次品
  - 射击一目标，直到击中为止，记录射击的次数
  - 从一批灯泡中，任取一只，测其寿命

- 样本空间
  - 掷骰子试验（有限可列）： $\mathcal{S}=\{1,2,3,4,5,6\}$
  - 取一件产品（有限可列）： $\mathcal{S}=\{\text{正品，次品}\}$
  - 射击目标试验（无限可列）： $\mathcal{S}=\{1,2,3,\ldots\}$
  - 灯泡寿命试验（无限不可列）： $\mathcal{S}=\{t|t\geq 0\}$

Random event （随机事件）

Event

An (random) **event** is any collection of possible outcomes of an experiment, that is, any subset of $\mathcal{S}$ (including $\mathcal{S}$ itself).

- 基本事件 vs. 复合事件
- 必然事件 vs. 不可能事件

Event operations

- 包含
  Containment
  $$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B$$
  $$A = B \Leftrightarrow A \subset B \text{ and } B \subset A$$
- 合集
  Union
  $$A \cup B = \{x : x \in A \text{ or } x \in B\}$$
- 交集
  Intersection
  $$A \cap B = \{x : x \in A \text{ and } x \in B\}$$
- 补集
  Complementation
  $$A^c = \{x : x \notin A\}$$
- 差集
  Theoretic difference
  $$A - B = \{x : x \in A \text{ and } x \notin B\}$$

Laws of event operations

- 交换律
  Commutativity

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

- 结合律
  Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

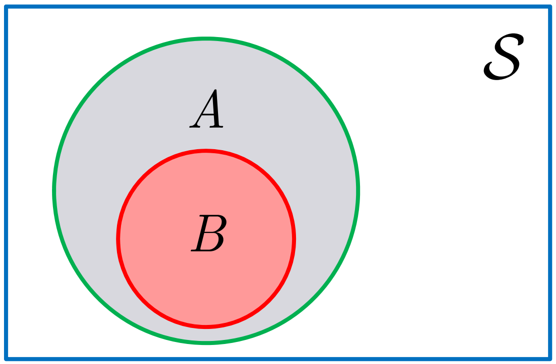
- 分配律
  Distributive Laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

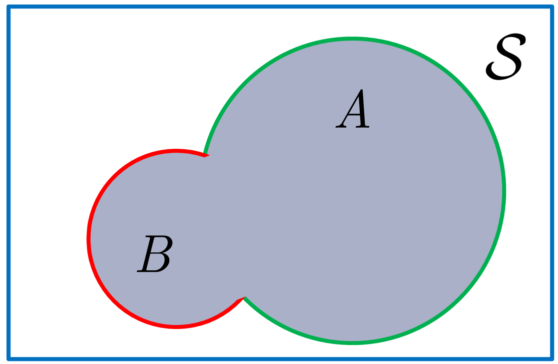
- 摩根律
  DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

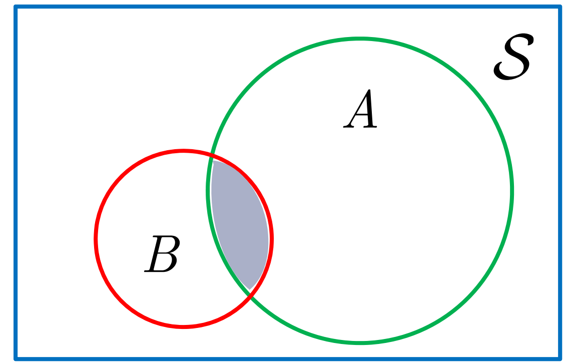
Venn diagrams



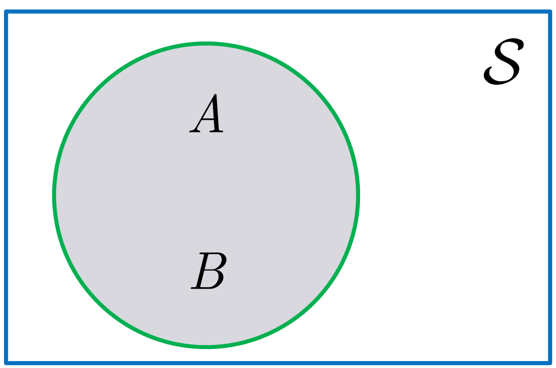
$A \supset B$



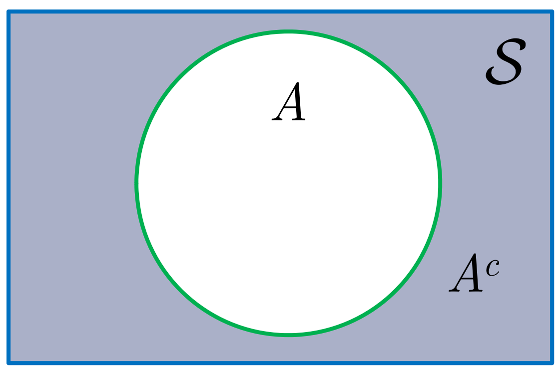
$A \cup B$



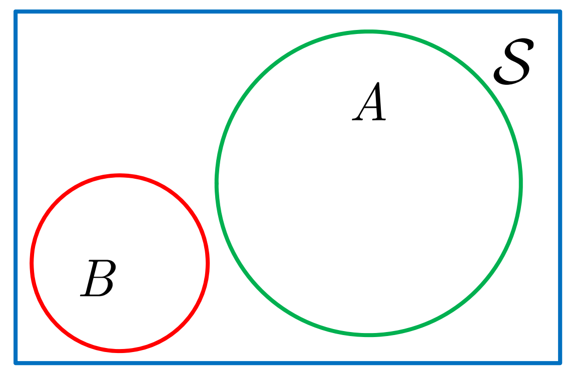
$A \cap B$



$A = B$



A^c



$A \cap B$

Extension of event operations

Countable infinite collection of sets

$$\bigcup_{i=1}^{\infty} A_i = \{x \in \mathcal{S} : x \in A_i \text{ for some } i\}$$

$$\bigcap_{i=1}^{\infty} A_i = \{x \in \mathcal{S} : x \in A_i \text{ for all } i\}$$

Uncountable infinite collection of sets

$$\bigcup_{a \in \Gamma} A_a = \{x \in \mathcal{S} : x \in A_a \text{ for some } a\}$$

$$\bigcap_{a \in \Gamma} A_a = \{x \in \mathcal{S} : x \in A_a \text{ for all } a\}$$

$\Gamma :$ All possible real numbers. $A_a : (0, a]$.

Generalized DeMorgan's Laws

Let $\{A_1, A_2, \dots, A_n\}$ be a finite collection of sets. Then

$$a)\ \left(\bigcup_{i=1}^{n} A_i\right)^c = \bigcap_{i=1}^{n} A_i^c, \quad \text{and} \quad b)\ \left(\bigcap_{i=1}^{n} A_i\right)^c = \bigcup_{i=1}^{n} A_i^c.$$

Let $\{A_1, A_2, \dots, A_\infty\}$ be an infinite countable collection of sets. Then

$$a)\ \left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c, \quad \text{and} \quad b)\ \left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c.$$

Let $\{A_\alpha : \alpha \in \Gamma\}$ be a (possible uncountable) collection of sets. Then

$$a)\ (\cup_\alpha A_\alpha)^c = \cap_\alpha A_\alpha^c, \quad \text{and} \quad b)\ (\cap_\alpha A_\alpha)^c = \cup_\alpha A_\alpha^c.$$

Mutually exclusive（互斥）

Mutually exclusive

Two events A and B are **disjoint** (mutually exclusive) if $A \cap B = \emptyset$.

The events $A_1, A_2, \ldots$ are **pairwise disjoint** (mutually exclusive) if $A_i \cap B_j = \emptyset$ for all $i \neq j$.

$$A_i = \left[i, i+1\right), i = 0, 1, 2, \ldots$$

$$A_i \cap A_j = \emptyset \text{ for all } i \neq j$$

Partition of the sample space

Partition of the sample space

If $A_1, A_2, \ldots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, \ldots$ forms a **partition** of $\mathcal{S}$.

$$\begin{aligned} A_i &= \left[i, i+1\right), i = 0, 1, 2, \ldots \\ A_i \cap A_j &= \emptyset \text{ for all } i \neq j \\ \bigcup_{i=0}^{\infty} A_i &= \left[0, \infty\right) \end{aligned}$$

Sigma algebra

sigma algebra (Borel field)

A collection of subsets of $\mathcal{S}$ is called a **sigma algebra**, denoted by $\mathcal{B}$, if it satisfies the following three properties:

1. $\emptyset \in \mathcal{B}$
   (the empty set is an element of $\mathcal{B}$);
2. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$
   ($\mathcal{B}$ is closed under complementation);
3. If $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$
   ($\mathcal{B}$ is closed under countable unions).

Examples of sigma algebra-I

- $\mathcal{B}_1 = \{\emptyset\, , \mathcal{S}\}$ (the trivial sigma algebra)
  - $\emptyset \in \mathcal{B}_1$
  - $\mathcal{B}_1$ is closed under complementation
  - $\mathcal{B}_1$ is closed under countable unions

Examples of sigma algebra-II

- $\mathcal{B}_2 = \{$all subsets of $\mathcal{S}$, including $\mathcal{S}$ itself$\}$
  - $\emptyset \in \mathcal{B}_2$
  - $\mathcal{B}_2$ is closed under complementation
  - $\mathcal{B}_2$ is closed under countable unions
- Example
  - $\mathcal{B} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$ for $\mathcal{S} = \{1,2,3\}$

Properties of a sigma algebra

- $\emptyset$ is always in a sigma algebra
  - By definition (1)
- $\mathcal{S}$ is always in a sigma algebra
  - By definitions (1) and (2)
- A sigma algebra is also closed under countable intersections
  - By definition (2), (3), and the DeMorgan's law

Kolmogorov axioms

Kolmogorov axioms

Given a sample space $\mathcal{S}$ and an associated sigma algbra $\mathcal{B}$, a **probability function** is a function with domain $\mathcal{B}$ that satisfies

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$;
2. $P(\mathcal{S}) = 1$;
3. If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Defining a probability function

- 定义扔硬币时观测到正面的概率
  - 样本空间 $\mathcal{S} = \{\text{H, T}\}$
  - Sigma algebra $\mathcal{B} = \{\emptyset, \{\text{H}\}, \{\text{T}\}, \{\text{H, T}\}\}$
  - 概率函数

$$P(A) = \begin{cases} 0 & \text{if } A = \emptyset; \\ 1/2 & \text{if } A = \{\text{H}\}; \\ 1/2 & \text{if } A = \{\text{T}\}; \\ 1 & \text{if } A = \{\text{H,T}\}. \end{cases}$$

Defining probability functions

Let $S = \{s_1, \dots, s_n\}$ be a finit set. Let $\mathcal{B}$ be any sigma algbra of subsets of S. Let $p_1, \dots, p_n$ be nonnegative numbers that sum to 1. For any $A \in \mathcal{B}$, define $P(A)$ by

$$P(A) = \sum_{\{i : s_i \in A\}} p_i .$$

Then P is a probability funciton on $\mathcal{B}$. This remains true if $S = \{s_1, s_2, \dots\}$ is a countable set.

Classical probabilities

- Sample space

  $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ A finite countable sample space

- Define probability

  $P(s_i) = p_i = 1/n$ Equal probability

- Probability function

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{n} = \frac{\#\{\text{elements in } A\}}{\#\{\text{elements in } \mathcal{S}\}}$$

where $A \in \mathcal{B} = \{\text{all subsets of } \mathcal{S}, \text{ including } \mathcal{S} \text{ itself}\}$

Dice

- Sample space

  $$\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$$

- Define probability

  $$P(s_i) = 1/6$$

- Probability function

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{n} = \frac{\#\{\text{elements in } A\}}{\#\{\text{elements in } \mathcal{S}\}}$$

- Calculation
  - P(观测到 3 的概率) $= 1/6$
  - P(观测到奇数点的概率) $= 3/6 = 1/2$
  - P(观测到大于等于 3 的概率) $= 4/6 = 2/3$

Poker

- Sample space

$$n = \#\{\text{elements in } \mathcal{S}\} = \binom{52}{5} = 2{,}598{,}960$$

- Define probability

$$P(s_i) = 1/2{,}598{,}960$$

- Probability function

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} \frac{1}{n} = \frac{\#\{\text{elements in } A\}}{\#\{\text{elements in } \mathcal{S}\}}$$

- Calculation
  - P(抽到同花顺的概率) $= 10\times 4/2{,}598{,}960 \approx 0.0015\%$
  - P(抽到四张A的概率) $= 48/2{,}598{,}960 \approx 0.0018\%$
  - P(抽到同花的概率) $= 4\times 1{,}287/2{,}598{,}960 \approx 0.1981\%$
  - P(抽到顺子的概率) $= 10\times 4^5/2{,}598{,}960 \approx 0.3940\%$
  - P(抽到一个对子的概率) $= 13\times 6\times 220\times 4^3/2{,}598{,}960 \approx 42.26\%$

Basic principles for counting

- 加法原理

Addition principle

If A and B are disjoint events and their are n_1 possible outcomes for event A and n_2 possible outcomes for event B, then there are $n_1 + n_2$ possible outcomes for event A or B.

- 乘法原理

Multiplication principle

If there is a sequence of k events with $n_1, n_2, \dots, n_k$ possible outcomes, then the total number of outcomes for the sequence of k events is $n_1 \times n_2 \times \cdots \times n_k$.

Four methods of counting

Select k objects from n	Without replacement	With replacement
Ordered		
Unordered		

Ordered without replacement

- 运用乘法原理
  - 选第一个对象有 n 种可能
  - 选第二个对象有 $n-1$ 种可能
  - $\ldots$
  - 选第 k 个对象有 $n-k+1$ 种可能

$$
\begin{aligned}
& n \times (n-1) \times \ldots \times (n-k+1) \\
&= \frac{n \times (n-1) \times \ldots \times (n-k+1) \times (n-k) \times (n-k-1) \times \ldots \times 1}{(n-k) \times (n-k-1) \times \ldots \times 1} \\
&= \frac{n!}{(n-k)!}
\end{aligned}
$$

Unordered without replacement

- 假设考虑顺序

$$\frac{n!}{(n-k)!}$$

- 除以重复计数的次数 $k!$

$$\frac{n!}{k!(n-k)!} = \binom{n}{k}$$

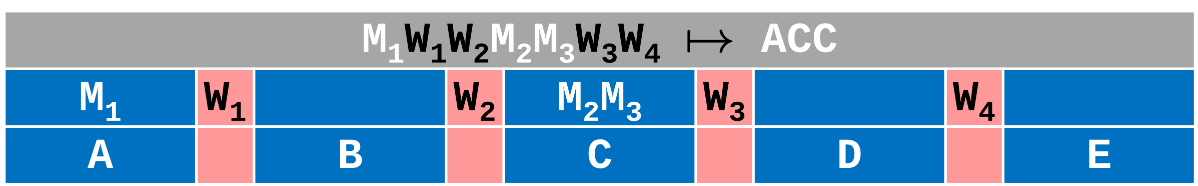
Ordered with replacement

- 运用乘法原理
  - 选第一个对象有 n 种可能
  - 选第二个对象有 n 种可能
  - $\dots$
  - 选第 k 个对象有 n 种可能

$$\underbrace{n \times n \times \ldots \times n}_{k \text{ times}} = n^k$$

Unordered with replacement

- 等同于模型
  - n 个对象固定，用 $n-1$ 块隔板隔开
  - 用 k 个标记来标记 n 个对象
- 考虑顺序 $(n+k-1)\times(n+k-2)\times\ldots\times 1$
- 除以重复计数 $$\frac{(n+k-1)\times(n+k-2)\times\ldots\times 1}{k!(n-1)!}=\binom{n+k-1}{k}$$
- 举例



Four methods of counting

- 选择的四种情况

Select k objects from n	Without replacement	With replacement
Ordered	$\frac{n!}{(n-k)!}$	n^k
Unordered	$\binom{n}{k}$	$\binom{n+k-1}{k}$

- 阶乘 $n! = n\times(n-1)\times\cdots\times 2\times 1$ factorial(x) lfactorial(x)
- 排列 $P(n,k) = n!/(n-k)!$
- 组合 $C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$ choose$(n,\ k)$ lchoose$(n,\ k)$

The calculus of probabilities

For a single event

If P is a probability function and A is any set in $\mathcal{B}$, then

1. $P(\emptyset) = 0$;
2. $P(A) \leq 1$;
3. $P(A^c) = 1 - P(A)$.

The calculus of probabilities

For two events

If P is a probability function and A and B are any two sets in $\mathcal{B}$, then

1. $P(A) = P(A \cap B) + P(A \cap B^c)$;
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
3. $P(A \cup B) \leq P(A) + P(B)$;
4. $P(A \cap B) \geq P(A) + P(B) - 1$ (Bonferroni's inequality);
5. If $A \subset B$, then $P(A) \leq P(B)$.

The calculus of probabilities

For countable events

If P is a probability function, then

1. $P(A)=\sum_{i=1}^{\infty} P(A \cap C_i)$ for any partition $C_1, C_2, \ldots$;
2. $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$ for any sets $A_1, A_2, \ldots$.

Conditional probability

Conditional probability

if A and B are events in $\mathcal{S}$, and $P(B) > 0$, then the **conditional probability** of A given B, written $P(A \mid B)$, is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

- 样本空间由 $\mathcal{S}$ 变为 B

Defining conditional probability

- 掷骰子试验
  - $A = \{\text{观测到 } 3\}$
  - $B = \{\text{观测到奇数点}\}$
  - $P(\text{观测到 } 3 \mid \text{观测到奇数点})$
    $= P(\text{观测到 } 3 \textbf{ 并且 } \text{观测到奇数点}) / P(\text{观测到奇数点})$
    $= P(\text{观测到 } 3) / P(\text{观测到奇数点})$
    $= (1/6)/(1/2)$
    $= 1/3$

Defining conditional probability

- 从一幅扑克牌中随机抽出五张
  - $A = \{$抽到一对K$\}$
  - $\mathrm{B} = \{$抽到一个对子$\}$
  - $P($抽到一对K $|$ 抽到一个对子$)$

    $= P($抽到一对K **并且** 抽到一个对子$)/P($抽到一个对子$)$

    $= P($抽到一对K$)/P($抽到一个对子$)$

    $= (6 \times 220 \times 4^3/2{,}598{,}960)/(13 \times 6 \times 220 \times 4^3/2{,}598{,}960)$

    $= 1/13$

Statistically independent

Statistically independent

Two events, A and B, are said to be **statistically independent** if

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

A collection of events, $A_1, A_2, \ldots A_n$ are **mutually independent** if for **any subcollection** $A_{i_1}, A_{i_2}, \ldots A_{i_k}$

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Tossing a fair coin three times

Sample space: {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Define H_i = the i-th toss is a head ($i = 1,2,3$). We have

$$P(H_1) = P(H_2) = P(H_3) = \frac{4}{8} = \frac{1}{2}$$

Now,

$$P(H_1 \cap H_2) = P(\{\text{HHH, HHT}\}) = \frac{2}{8} = \frac{1}{4} = P(H_1)P(H_2)$$

$$P(H_2 \cap H_3) = P(\{\text{HHH, THH}\}) = \frac{2}{8} = \frac{1}{4} = P(H_2)P(H_3)$$

$$P(H_3 \cap H_1) = P(\{\text{HHH, HTH}\}) = \frac{2}{8} = \frac{1}{4} = P(H_3)P(H_1)$$

$$P(H_1 \cap H_2 \cap H_3) = P(\{\text{HHH}\}) = \frac{1}{8} = P(H_1)P(H_2)P(H_3)$$

Therefore,

$H_1, H_2,$ and H_3 are mutually independent.

Mutually independent $\neq$ pairwise independent

A B | **B C** | **C A** |

$A =$ the card has letter A; $B =$ the card has letter B; $C =$ the card has letter C

$$P(A) = P(B) = P(C) = \frac{2}{4} = \frac{1}{2}$$

Now,

$$P(A \cap B) = \frac{1}{4} = P(A)P(B)$$
$$P(B \cap C) = \frac{1}{4} = P(B)P(C)$$
$$P(C \cap A) = \frac{1}{4} = P(C)P(A)$$
$$P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C)$$

Independence of complements

If A and B are independent, then the following pairs are also independent

$a)$ A and B^c,

b) A^c and B,

$c)$ A^c and B^c

Multiplication rule

Multiplication rule

Let A and B be two events in $\mathcal{S}$. If $P(A) > 0$, then

$$P(A \cap B) = P(A)P(B \mid A);$$

if $P(B) > 0$, then

$$P(A \cap B) = P(B)P(A \mid B).$$

$$P(AB) = P(A)P(B \mid A)$$

$$P(AB) = P(B)P(A \mid B)$$

Multiplication rule

Multiplication rule

Let A, B, and C be three events in $\mathcal{S}$, then

$$\begin{aligned} P(A \cap B \cap C) &= P((A \cap B) \cap C) \\ &= P(A \cap B)P(C \mid A \cap B) \\ &= P(A)P(B \mid A)P(C \mid A \cap B) \end{aligned}$$

$$P(ABC) = P(A)P(B \mid A)P(C \mid AB)$$

Chain rule

Chain rule

Let $A_1, \ldots, A_k$ be k events in $\mathcal{S}$, then

$$
\begin{aligned}
P\left(\bigcap_{i=1}^{k} A_i\right) &= P\left(\bigcap_{i=1}^{k-1} A_i\right) P\left(A_k \,\middle|\, \bigcap_{i=1}^{k-1} A_i\right) \\
&= P\left(\bigcap_{i=1}^{k-2} A_i\right) P\left(A_{k-1} \,\middle|\, \bigcap_{i=1}^{k-2} A_i\right) P\left(A_k \,\middle|\, \bigcap_{i=1}^{k-1} A_i\right) \\
&= \cdots \\
&= P(A_1 \cap A_2 \cap A_3) P(A_4 \mid A_1 \cap A_2 \cap A_3) \cdots P\left(A_k \,\middle|\, \bigcap_{i=1}^{k-1} A_i\right) \\
&= P(A_1 \cap A_2) P(A_3 \mid A_1 \cap A_2) P(A_4 \mid A_1 \cap A_2 \cap A_3) \cdots P\left(A_k \,\middle|\, \bigcap_{i=1}^{k-1} A_i\right) \\
&= P(A_1) p(A_2 \mid A_1) P(A_3 \mid A_1 \cap A_2) P(A_4 \mid A_1 \cap A_2 \cap A_3) \cdots P\left(A_k \,\middle|\, \bigcap_{i=1}^{k-1} A_i\right)
\end{aligned}
$$

Law of total probability

Law of total probability

Let $A_1, A_2, \ldots$ be a partition of the sample space $\mathcal{S}$, then for any event B,

$$P(B) = \sum_{i=1}^{\infty} P(B \mid A_i) P(A_i).$$

$$P(B) = \sum_{i=1}^{\infty} P(B \cap A_i), \text{ and } P(B \cap A_i) = P(B \mid A_i) P(A_i)$$

Bayes' Rule

Bayes' rule

Let $A_1, A_2, \ldots$ be a partition of the sample space $\mathcal{S}$, and let B be any set. Then, for each $i = 1, 2, \ldots,$

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B \mid A_j)P(A_j)}.$$

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)},$$

$$P(A_i \cap B) = P(B \mid A_i)P(A_i), \text{ and}$$

$$P(B) = \sum_{i=1}^{\infty} P(B \mid A_i)P(A_i)$$

Applications of the Baye's rule

An investigation has shown that 5% of men and 0.25% of women are color-blind. A person is chosen at random and that person is color-blind. What is the probability that the person is female.

Define events

M=The person is a man

F=The person is a woman

C=The person is color blind

We like to calculate $P(F \mid C)$. Using Baye's rule:

$$\begin{aligned} P(F \mid C) &= \frac{P(C \mid F)P(F)}{P(C \mid F)P(F) + P(C \mid M)P(M)} \\ &= \frac{0.25\% \times 50\%}{0.25\% \times 50\% + 5\% \times 50\%} \\ &\approx 4.76\% \end{aligned}$$

Thank you very much