非参数方法
Nonparametric Methods

什么是非参数方法

参数方法 vs 非参数方法

- 总体的分布形式已知, 而其中的某些参数未知, 我们可以通过从总体中随机抽取样本, 根据样本信息对总体参数进行估计和假设检验, 这就是一般所说的参数方法

- 总体的分布未知, 或虽已知却不能用有限个参数刻画, 这时要对总体的某些性质进行估计或假设检验, 就要使用非参数方法

参数方法 vs 非参数方法

- 多数情况下, 非参数问题与参数问题界线清晰; 少数情况下, 会因为各人出发点不同而有不同看法; 非参数方法并非绝对只能解决非参数问题, 有些也适用于典型的参数问题

非参数方法的特点

- 在利用样本信息对总体性质进行估计或检验时, 不依赖总体的分布形式, 构造的统计量通常与总体分布无关, 因此, 非参数方法也称为“自由分布(distribution-free)”方法

- 非参数方法对变量的量化要求很低, 不论是分类变量还是数值变量, 都可以采用非参数方法进行估计或检验

非参数方法的优点

- 不受总体分布的限制, 应用范围广泛; 它仅应用样本观察值中一些非常直观(例如次序/秩)的信息
- 基本上每一种参数方法都有相对应的非参数方法; 非参数方法对数据的要求不像参数方法那么严格
- 非参数方法常具较好的稳健性(比如不受数据中极端值的影响)

非参数方法的缺点

- 非参数方法需要考虑在约束条件十分宽松的情况下使用，有可能导致效率的下降

- 对符合使用参数方法条件的数据，使用非参数方法时，犯第二类错误的概率β比参数方法要大，即统计功效要小；若要达到相同的统计功效，非参数方法比参数方法所需要的样本容量要大

什么时候选择非参数方法

- 如果中位数(而不是均值)可以更好地描述数据的集中趋势
- 如果处理对象为有序变量或者有不能移除的极端值
- 对符合使用参数方法条件的数据，首选参数方法；当不满足使用参数方法的条件时，才选用非参数方法
- 如果样本不能很好的代表总体，任何检验方法都是无效的

非参数方法举例

One sample/Paired t-test ⟷ Wilcoxon signed-rank test

Two sample t-test ⟷ Wilcoxon rank-sum test

Pearson Correlation test ⟷ Spearman Correlation test

Confidence interval ⟵ Bootstrap

Significance ⟵ Permutation

Paired t-test ⟷ Wilcoxon signed-rank test

运动是否影响睡眠?

Before	After
7.2	6
6.2	8.6
7	8.3
6	5.3
7.5	7.5
5.5	6.5
7.1	8.9
4.8	5.6
5.2	7.2
8	7.4

→

Diff(A-B)	Abs	Sign	Rank	Signed-rank
-1.2	1.2	-	5	-5
2.4	2.4	+	9	9
1.3	1.3	+	6	6
-0.7	0.7	-	2	-2
0				
1.0	1.0	+	4	4
1.8	1.8	+	7	7
0.8	0.8	+	3	3
2	2	+	8	8
-0.6	0.6	-	1	-1

Paired t-test ⟷ Wilcoxon signed-rank test

运动是否影响睡眠?

Diff(A-B)	Abs	Sign	Rank	Signed-rank
-1.2	1.2	-	5	-5
2.4	2.4	+	9	9
1.3	1.3	+	6	6
-0.7	0.7	-	2	-2
0				
1.0	1.0	+	4	4
1.8	1.8	+	7	7
0.8	0.8	+	3	3
2	2	+	8	8
-0.6	0.6	-	1	-1

Sum of positive signed-rank

V or W = 9+6+4+7+3+8 = 37

$$Z = \frac{\sum_{i=1}^{n} SR_i}{\sqrt{\sum_{i=1}^{n} SR_i^2}}$$

= 29/16.88 = 1.72

p = 2x0.043 = 0.086

Two sample t-test ⟷ Wilcoxon rank-sum test

两个总体均值相等

两个总体中位数相等

Sleep_male	Sleep_female
7.2	6
6.2	8.6
7	8.3
6	5.3
7.5	7.5
5.5	6.5
7.1	8.9
4.8	5.6
5.2	7.2
8	7.4

→

Rank(male)	Rank(female)
12.5	6.5
8	19
10	18
6.5	3
15.5	15.5
4	9
11	20
1	5
2	12.5
17	14
Sum = 87.5	Sum = 122.5

Two sample t-test ⟷ Wilcoxon rank-sum test

两个总体均值相等

两个总体中位数相等

Rank(male)	Rank(female)
12.5	6.5
8	19
10	18
6.5	3
15.5	15.5
4	9
11	20
1	5
2	12.5
17	14
Sum = 87.5	Sum = 122.5

$W = Sum\_R_1 - n_1*(n_1+1)/2 = 32.5$

$$Z = \frac{W - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

$= (32.5 - 50) / 13.23 = -1.33$

$p = 2*0.093 = 0.186$

Pearson Correlation test ⟷ Spearman Correlation test

Score	Happy
10	10
20	30
30	20
40	60
50	40
60	50
70	80
80	90
90	90
100	70

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

Pearson Correlation

- H_0: r = 0
- H_A: r ≠ 0

$$\frac{r}{\sqrt{1-r^2}/\sqrt{n-2}} \sim t(n-2)$$

Pearson Correlation test ⟷ Spearman Correlation test

Score	Happy
10	10
20	30
30	20
40	60
50	40
60	50
70	80
80	90
90	90
100	70

→

rank(Score)	rank(Happy)
1	1
2	3
3	2
4	6
5	4
6	5
7	8
8	9.5
9	9.5
10	7

Pearson Correlation test ⟷ Spearman Correlation test

rank(Score)	rank(Happy)
1	1
2	3
3	2
4	6
5	4
6	5
7	8
8	9.5
9	9.5
10	7

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$\rho = \frac{Cov(X,Y)}{S_X S_Y}$$

Spearman Correlation

- H_0: $\rho = 0$
- H_A: $\rho \neq 0$

$$\frac{\rho}{\sqrt{1-\rho^2}/\sqrt{n-2}} \sim t(n-2)$$

$S = (1-\rho)(n^3-n)/6$

再从回归的角度比较常见的参数与非参数方法

参数方法	非参数方法
One sample t-test	Wilcoxon signed-rank test
$y = \beta_0$	$signed\_rank(y) = \beta_0$
Paired t-test	Wilcoxon signed-rank test
$y_1 - y_2 = \beta_0$	$signed\_rank(y_1 - y_2) = \beta_0$

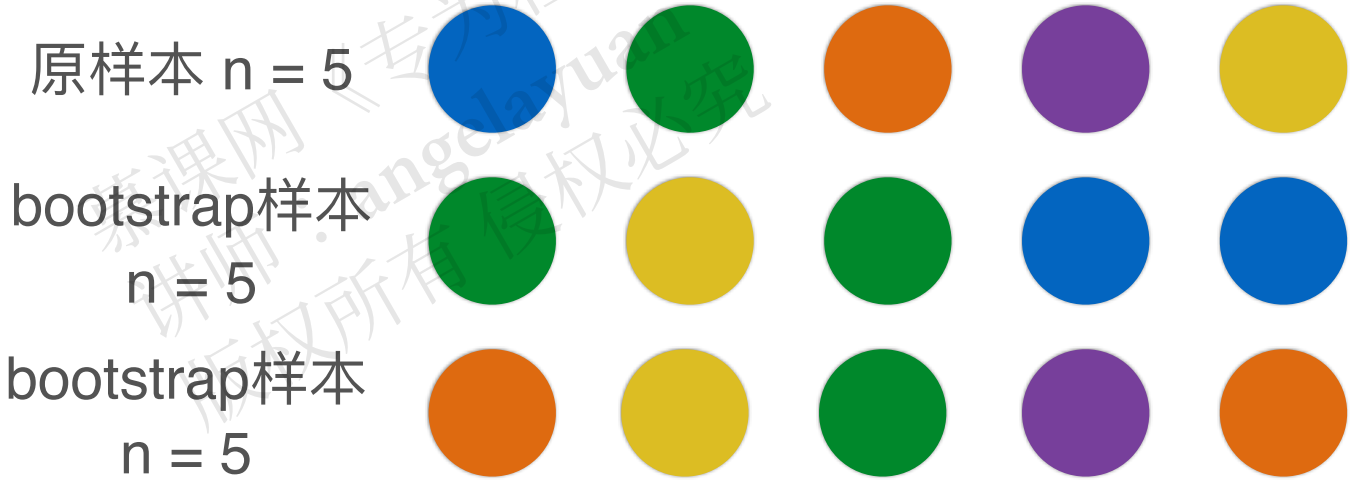
再从回归的角度比较常见的参数与非参数方法

参数方法	非参数方法
Two sample t-test	Wilcoxon rank-sum test
$y = \beta_0 + \beta_1 x$	$rank(y) = \beta_0 + \beta_1 x$
Pearson Correlation test	Spearman Correlation test
$y = \beta_0 + \beta_1 x$	$rank(y) = \beta_0 + \beta_1 rank(x)$

Bootstrap

Bootstrap

- 总体的分布未知, 但已经有一个容量为n的来自总体的样本, 自这一样本按放回抽样(sampling with replacement)的方法抽取一个容量为n的样本, 这种样本称为bootstrap样本或自助样本



Bootstrap

- 反复地, 独立地从原始样本中抽取很多个bootstrap样本(通常要抽取不少于1000个样本), 利用这些bootstrap样本对总体进行统计推断, 这种方法称为非参数bootstrap方法, 又称自助法

> “Pulling oneself up by one’s bootstraps”: accomplishing an impossible task without any outside help

Bootstrap举例

Sleep_male	Sleep_female
7.2	6
6.2	8.6
7	8.3
6	5.3
7.5	7.5
5.5	6.5
7.1	8.9
4.8	5.6
5.2	7.2
8	7.4

- 原始样本 n = 20
- 以样本中位数作为总体中位数的估计
- 求男女睡眠时间中位数差异的估计的标准误差和置信区间

Bootstrap举例

Sleep_male	Sleep_female
7.2	6
6.2	8.6
7	8.3
6	5.3
7.5	7.5
5.5	6.5
7.1	8.9
4.8	5.6
5.2	7.2
8	7.4

bootstrap样本1

M: 7.2, 7, 6, 5.5, 6, 8, 7.1, 4.8, 5.5, 7.1

F: 6, 8.3, 6.5, 5.6, 8.6, 7.2, 7.4, 6, 5.3, 7.4

求该bootstrap样本中男女各自睡眠的中位数的差: D_1 = median(M) - median(F)

Bootstrap举例

Sleep_male	Sleep_female
7.2	6
6.2	8.6
7	8.3
6	5.3
7.5	7.5
5.5	6.5
7.1	8.9
4.8	5.6
5.2	7.2
8	7.4

bootstrap样本2

M: 6.2, 7.1, 6, 5.2, 6.2, 7.5, 7.2, 4.8, 5.2, 6.2

F: 8.6, 5.6, 6.5, 5.3, 8.9, 7.5, 7.5, 6.5, 6, 7.4

求该bootstrap样本中男女各自睡眠的中位数的差: D_2 = median(M) - median(F)

Bootstrap举例

Sleep_male	Sleep_female
7.2	6
6.2	8.6
7	8.3
6	5.3
7.5	7.5
5.5	6.5
7.1	8.9
4.8	5.6
5.2	7.2
8	7.4

bootstrap样本1000

M: 8, 4.8, 7.5, 6, 6.2, 6.2, 8, 5.5, 5.2, 7.2

F: 5.3, 8.9, 6, 7.4, 6, 7.5, 7.2, 8.6, 8.3, 7.2

求该bootstrap样本中男女各自睡眠的中位数的差: D_{1000} = median(M) - median(F)

Bootstrap举例

bootstrap样本1 → D_1

bootstrap样本2 → D_2

bootstrap样本3 → D_3

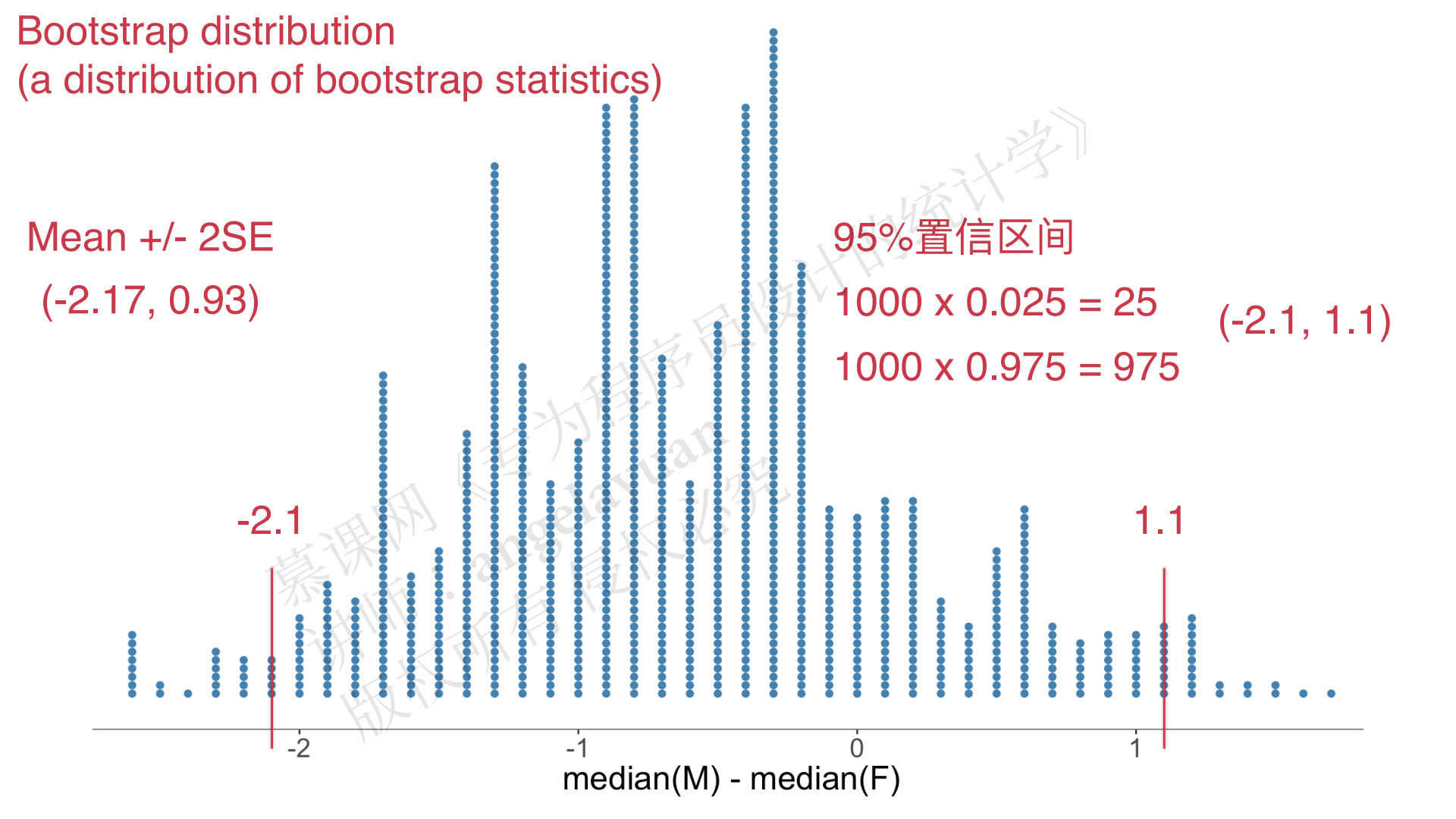
bootstrap样本4 → D_4

⋮ ⋮

bootstrap样本1000 → D_{1000}

- 求男女睡眠时间中位数差异的估计的标准误差和置信区间

$$SE = \sqrt{\frac{\sum_{i=1}^{B}(D_i - \overline{D})^2}{B-1}}$$


Bootstrap distribution
(a distribution of bootstrap statistics)
Mean +/- 2SE
(-2.17, 0.93)
95%置信区间
1000 x 0.025 = 25
1000 x 0.975 = 975
(-2.1, 1.1)
-2.1
1.1
-2
-1
0
1
median(M) - median(F)

Permutation

Permutation

- 假设检验: 构造一个统计量, 得到这个统计量在零假设为真时的抽样分布, 然后把样本统计量的观测值与抽样分布进行比较, 进行统计推断(计算在该抽样分布下得到该观测值或更极端值的概率)

- Permutation: 使我们可以得到对任意统计量在零假设为真时的抽样分布, 从而进行统计推断

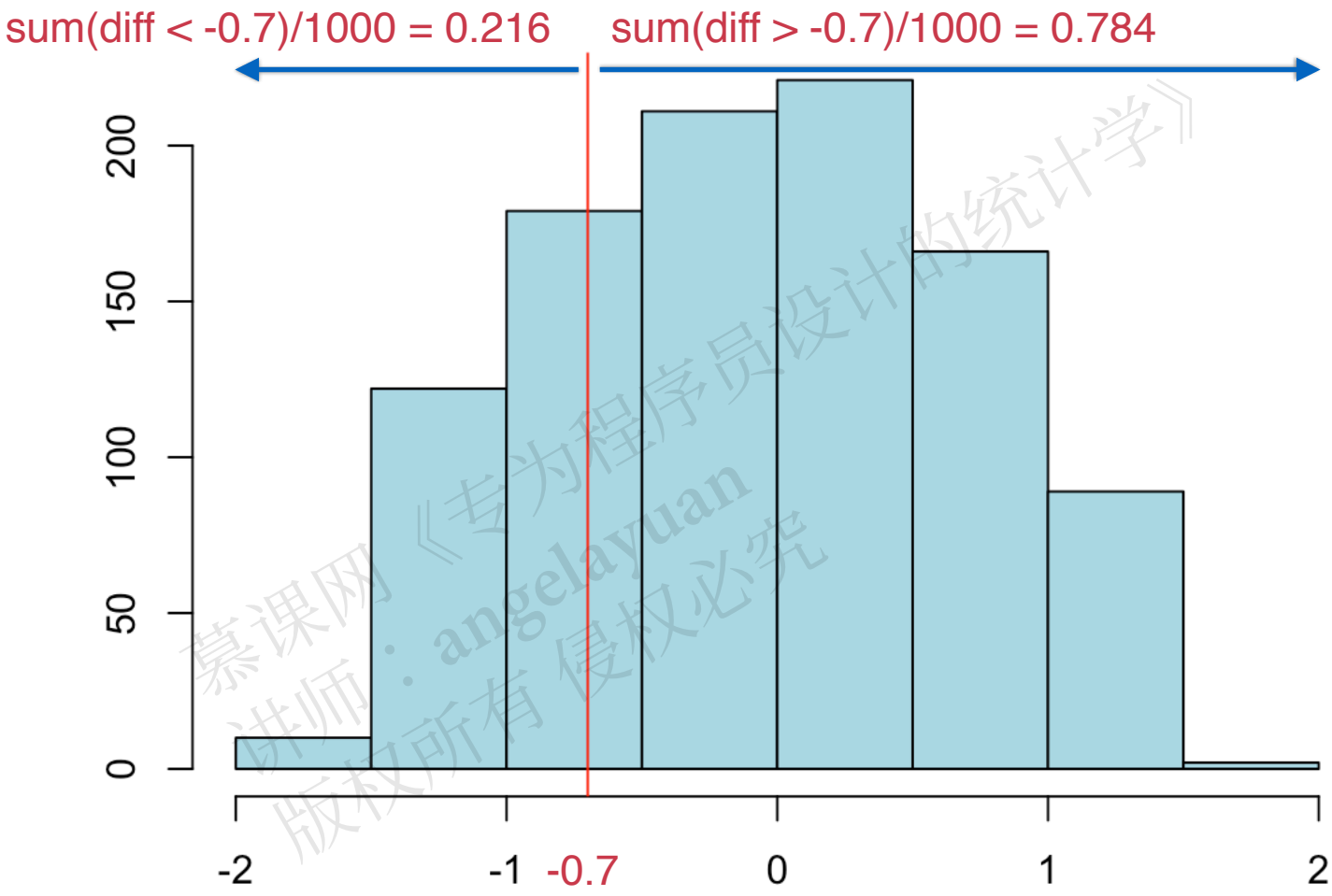
Permutation

Sleep_male	Sleep_female
7.2	6
6.2	8.6
7	8.3
6	5.3
7.5	7.5
5.5	6.5
7.1	8.9
4.8	5.6
5.2	7.2
8	7.4
6.6	7.3

H_0: 男性睡眠时长的中位数与女性睡眠时长的中位数没有差异

对下列过程重复1000次

- 把20个数字随机分配给sleep_male (n=10) 和sleep_female (n=10)
- 计算sleep_male与sleep_female的中位数
- 计算并保存 median(M) - median(F)



本章小结