

实用统计分析与R入门

参数估计与假设检验

参数估计

- 总体是由总体分布来刻画的，在实际问题中，我们根据实际问题本身的专业知识或者经验或使用适当的统计方法，有时可以判断总体分布的类型。但总体分布的参数还是未知的，需要通过样本来估计。
- 根据样本来估计总体分布中的未知参数，叫做参数估计。

点估计与区间估计

- 点估计
 - 用一个统计量来估计一个未知参数
 - 能够明确告诉我们“未知参数大概是多少”
 - 不能反映出估计的可靠程度
- 区间估计
 - 用两个统计量构成的区间来估计一个未知参数，并同时指明此区间可以覆盖住这个参数的可靠程度（置信度）。
 - 不能直接告诉我们“未知参数是多少”

假设检验

- 假设检验：利用样本数据对某个事先做出的统计假设，依照某种设计好的方法进行检验，判断此假设是否正确
- 基本思想：
 - 为了检验一个假设是否成立，先假定这个假设是成立的。
 - 如果这个假设导致一个不合理的现象出现，那么就表明原先的假设不正确，因此，我们就拒绝原假设。
 - 判断一个现象是否合理的原则：小概率事件在一次观察中不会发生

两类错误

- 第一类错误

- 原假设正确，但拒绝原假设
- 犯第一类错误的概率通常用 α 来表示，也就是显著性水平

$$\alpha = P\{\text{拒绝}H_0|H_0\text{为真}\}$$

- 第二类错误

- 原假设不正确，但接受原假设
- 犯第二类错误的概率通常用 β 来表示

$$\beta = P\{\text{接受}H_0|H_0\text{为假}\}$$

- 统计功效

- 原假设错误，拒绝原假设的概率

$$\pi = 1 - \beta = P\{\text{拒绝}H_0|H_0\text{为假}\}$$

假设检验的基本步骤

- 根据问题的需要，对待检验的未知参数 θ ，作出一个零假设(H_0)，进而确定备择假设(H_1)
- 选定一个显著性水平 α ，常用的为0.05、0.01、0.001
- 选定统计方法，根据样本观察值，计算出相应的统计量(如 t 、 F 、 χ^2 等)
- 根据统计量的大小及其分布，计算 p 值(p value)的大小，并判定零假设是否成立

读入数据

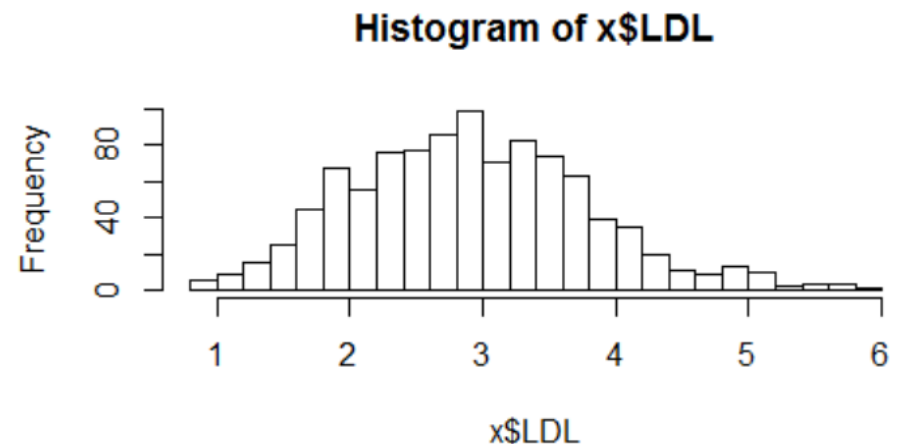
- `x <- read.csv('pheno.csv', head=T)`
- `dim(x)`
- `head(x)`

```
> x <- read.csv('pheno.csv', head=T)
```

```
> dim(x)
[1] 995  7
```

```
> head(x)
  id LDL CAD AGE SEX
1 p0001 2.7 N  55  M
2 p0002 3.4 N  64  M
3 p0003 4.2 N  49  M
4 p0004 3.5 N  43  F
5 p0005 2.5 Y  66  M
6 p0006 2.5 N  54  F
```

```
> hist(x$LDL, breaks=20)
```



t 检验

- 男性与女性LDL的均值是否相等？

```
x.male <- x[x$SEX == 'M',]
```

```
x.female <- x[x$SEX == 'F',]
```

```
head(x.male)
```

```
head(x.female)
```

```
> x.male <- x[x$SEX == 'M',]
```

```
> x.female <- x[x$SEX == 'F',]
```

```
> head(x.male)
```

	id	LDL	CAD	AGE	SEX
1	p0001	2.7	N	55	M
2	p0002	3.4	N	64	M
3	p0003	4.2	N	49	M
5	p0005	2.5	Y	66	M
7	p0007	2.6	N	65	M
8	p0008	1.7	Y	75	M

```
> head(x.female)
```

	id	LDL	CAD	AGE	SEX
4	p0004	3.5	N	43	F
6	p0006	2.5	N	54	F
9	p0009	1.9	N	67	F
10	p0010	2.7	N	67	F
13	p0013	4.4	N	68	F
15	p0015	2.8	N	51	F

t 检验

- 男性与女性LDL的均值是否相等？
 - H_0 : 男性与女性LDL血脂水平均值相等
 - H_1 : 男性与女性LDL血脂水平均值不相等
- `t.test(x.male$LDL, x.female$LDL)`

```
> t.test(x.male$LDL, x.female$LDL)
```

```
Welch Two Sample t-test
```

```
data: x.male$LDL and x.female$LDL
```

```
t = -6.2186, df = 871.91, p-value = 7.772e-10 P值
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.4579649 -0.2382342 置信区间
```

```
sample estimates:
```

```
mean of x mean of y
```

```
2.757002 3.105102
```

t统计量

**置信水平
(1- α)**

H_1

t 检验

- 男性与女性LDL的均值是否相等？

- H_0 : 男性与女性LDL血脂水平均值相等

- H_1 : 男性与女性LDL血脂水平均值不相等

$\alpha = 0.01$

`t.test(x.male$LDL, x.female$LDL, conf.level = 0.99)`

```
> t.test(x.male$LDL, x.female$LDL, conf.level = 0.99)
```

```
Welch Two Sample t-test
```

```
data: x.male$LDL and x.female$LDL
```

```
t = -6.2186, df = 871.91, p-value = 7.772e-10
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
99 percent confidence interval:
```

```
-0.4926030 -0.2035962 置信区间
```

```
sample estimates:
```

```
mean of x mean of y
```

```
2.757002 3.105102
```

置信水平
($1-\alpha$)

二项分布总体的假设检验

- 二型糖尿病发病率是否为8.5%?

- H_0 : 发病率等于8.5%

- H_1 : 发病率不等于8.5%

`binom.test(sum(x$T2D == 'Y'), sum(x$T2D == 'Y' | x$T2D == 'N'), p = 0.085)`

```
> binom.test(sum(x$T2D == 'Y'), sum(x$T2D == 'Y' | x$T2D == 'N'), p = 0.085)
```

Exact binomial test

data: `sum(x$T2D == "Y")` and `sum(x$T2D == "Y" | x$T2D == "N")`
number of successes = 149, number of trials = 995, p-value =
1.951e-11

alternative hypothesis: true probability of success is not equal to 0.085

95 percent confidence interval:

0.1281370 0.1734556

H_1

sample estimates:

probability of success

0.1497487

二项分布总体的假设检验

- 二型糖尿病发病率是否高于8.5%?

- H_0 : 发病率小于或等于8.5%

- H_1 : 发病率大于8.5%

binom.test(sum(x\$T2D == 'Y'), sum(x\$T2D == 'Y' | x\$T2D == 'N'), p = 0.085, alternative = 'greater')

```
> binom.test(sum(x$T2D == 'Y'), sum(x$T2D == 'Y' | x$T2D == 'N'), p = 0.085,  
alternative = 'greater')
```

Exact binomial test

```
data:  sum(x$T2D == "Y") and sum(x$T2D == "Y" | x$T2D == "N")  
number of successes = 149, number of trials = 995, p-value = 1.539e-11  
alternative hypothesis: true probability of success is greater than 0.085  
95 percent confidence interval:  
 0.1314254 1.0000000  
sample estimates:  
probability of success  
 0.1497487
```

H_1

χ^2 检验

- 男性与女性糖尿病发病率是否相同
 - H_0 : 男性与女性糖尿病发病率相同
 - H_1 : 男性与女性糖尿病发病率不同

```
table(x[,c(5,6)])
```

```
chisq.test(table(x[,c(5,6)]))
```

```
> table(x[,c(5,6)])  
      T2D  
SEX      N      Y  
 F    495    93  
 M    329    78
```

```
> chisq.test(table(x[,c(5,6)]))
```

```
Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  table(x[, c(5, 6)])
```

```
X-squared = 1.6666, df = 1, p-value = 0.1967
```

Fisher精确检验

- 60岁以上与60岁以下人群糖尿病发病率是否相同

```
x$AGE.group <- cut(x$AGE, breaks=c(min(x$AGE), 60,  
max(x$AGE)), include.lowest = T)
```

```
table(x[,6:7])
```

```
> x$AGE.group <- cut(x$AGE, breaks=c(min(x$AGE), 60, max(x$AGE)),  
include.lowest = T)
```

```
> head(x)
```

	id	LDL	CAD	AGE	SEX	T2D	AGE.group
1	p0001	2.7	N	55	M	N	[30,60]
2	p0002	3.4	N	64	M	Y	(60,84]
3	p0003	4.2	N	49	M	N	[30,60]
4	p0004	3.5	N	43	F	N	[30,60]
5	p0005	2.5	Y	66	M	N	(60,84]
6	p0006	2.5	N	54	F	N	[30,60]

```
> table(x[,6:7])
```

	AGE.group	
T2D	[30,60]	(60,84]
N	477	369
Y	73	76

Fisher精确检验

- 60岁以上与60岁以下人群糖尿病发病率是否相同
 - H_0 : 60岁以上与60岁以下人群糖尿病发病率相同
 - H_1 : 60岁以上与60岁以下人群糖尿病发病率不同
- `fisher.test(table(x[,6:7]))`

```
> fisher.test(table(x[,6:7]))

Fisher's Exact Test for Count Data

data:  table(x[, 6:7])
p-value = 0.1077
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9348395 1.9374881
sample estimates:
odds ratio
1.345385
```

Fisher精确检验

- 2x2列联表中的odds ratio

	1	2
A	A1	A2
B	B1	B2

$$\text{odds ratio} = \frac{\frac{B2}{A2}}{\frac{B1}{A1}} = \frac{\frac{B2}{A2}}{\frac{B1}{A1}} = \frac{B2 * A1}{B1 * A2}$$

Fisher精确检验

- 60岁以上人群糖尿病发病率是否高于60岁以下人群？
 - H_0 : 60岁以上人群糖尿病发病率低于或者等于60岁以下人群糖尿病发病率
 - H_1 : 60岁以上人群糖尿病发病率高于60岁以下人群糖尿病发病率

```
> table(x[, 6:7])  
      AGE.group  
T2D [30,60] (60,84]  
  N      477 A1    369 A2  
  Y       73 B1     76 B2
```


$$\frac{B2}{A2} > \frac{B1}{A1}$$


$$\text{odds ratio} = \frac{\frac{B2}{A2}}{\frac{B1}{A1}} > 1$$

Fisher精确检验

- 60岁以上人群糖尿病发病率是否高于60岁以下人群？
 - H_0 : 60岁以上人群糖尿病发病率低于或者等于60岁以下人群糖尿病发病率
 - H_1 : 60岁以上人群糖尿病发病率高于60岁以下人群糖尿病发病率

```
> fisher.test(table(x[,6:7]), alternative = 'greater')

Fisher's Exact Test for Count Data

data:  table(x[, 6:7])
p-value = 0.05692
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.9887374      Inf
sample estimates:
odds ratio
 1.345385
```

秩统计量

- 设 X_1 ，为一组样本，将 X_1 从小到大排成一列，用 R_1 记为 X_1 在上述排列中的位置序号， $i = 1, 2, 3, \dots, n$ 。称 R_1 为样本 X_1 产生的秩序统计量。
- 秩统计量的一个重要特征是分布无关性

秩和检验

- 男性与女性LDL的均值是否相等？
 - H_0 : 男性与女性LDL血脂水平均值相等
 - H_1 : 男性与女性LDL血脂水平均值不相等

`wilcox.test(x.male$LDL, x.female$LDL)`

```
> wilcox.test(x.male$LDL, x.female$LDL)

      Wilcoxon rank sum test with continuity correction

data:  x.male$LDL and x.female$LDL
W = 91848, p-value = 4.272e-10
alternative hypothesis: true location shift is not equal to 0
```

多重假设检验校正

- 各个年龄段男性与女性血脂水平是否有差异
 - (40, 45]
 - (45, 50]
 - ...
 - (70, 75]
 - (80, 85]

```
res <- data.frame()
for (i in seq(40, 80, 5)) {
  ldl.male <- x.male[x.male$AGE > i & x.male$AGE <= i + 5, 2]
  ldl.female <- x.female[x.female$AGE > i & x.male$AGE <= i + 5, 2]
  ldl.test <- wilcox.test(ldl.male, ldl.female)
  res <- rbind(res, c(i, i+5, ldl.test$p.value))
}
```

多重假设检验校正

- 各个年龄段男性与女性血脂水平是否有差异

```
colnames(res) <- c('AGE1', 'AGE2', 'P')
```

```
> res
  X40 X45 X0.00694590059157067
1  40  45      0.006945901
2  45  50      0.133278783
3  50  55      0.154470954
4  55  60      0.182092585
5  60  65      0.034658789
6  65  70      0.131205501
7  70  75      0.001626903
8  75  80      0.002615996
9  80  85      0.440786714
```

```
> colnames(res) <- c('AGE1', 'AGE2', 'P')
> res
  AGE1 AGE2      P
1   40   45 0.006945901
2   45   50 0.133278783
3   50   55 0.154470954
4   55   60 0.182092585
5   60   65 0.034658789
6   65   70 0.131205501
7   70   75 0.001626903
8   75   80 0.002615996
9   80   85 0.440786714
```

多重假设检验校正

- 在进行多重假设检验时，每个单独的假设都具有其本身的I型错误。在这种情况下，如果不进行任何的控制，犯I-型错误的概率会随着假设检验的个数而迅速增加
- 多重假设检验校正的方法
 - 调整显著性水平 α
 - 对p值进行校正

多重假设检验校正

- 在进行多重假设检验时，如果不对显著性水平 α 做出调整，那么一般需要对p值进行校正

```
res$P.BF <- p.adjust(res$P, method='bonferroni')
```

```
res$P.FDR <- p.adjust(res$P, method='fdr')
```

```
> res$P.BF <- p.adjust(res$P, method='bonferroni')
> res$P.FDR <- p.adjust(res$P, method='fdr')
> res
```

	AGE1	AGE2	P	P.BF	P.FDR
1	40	45	0.006945901	0.06251311	0.02083770
2	45	50	0.133278783	1.00000000	0.19860551
3	50	55	0.154470954	1.00000000	0.19860551
4	55	60	0.182092585	1.00000000	0.20485416
5	60	65	0.034658789	0.31192910	0.07798228
6	65	70	0.131205501	1.00000000	0.19860551
7	70	75	0.001626903	0.01464213	0.01177198
8	75	80	0.002615996	0.02354396	0.01177198
9	80	85	0.440786714	1.00000000	0.44078671