# 生物统计学 R 上机实验讲义

华东师范大学 2021 暑期课程

## 目录

# 1 假设检验：单样本推断

## 1.1 正态分布均值的单样本检验

### 1.1.1 方差未知时：$t$ 检验

待检验的假设为:

$$H_0 : \mu = \mu_0 \quad \longleftrightarrow \quad H_1 : \mu \neq (\text{或} >, <)\ \mu_0$$

检验统计量为:

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$$

命令使用格式:

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

需要输入的参数为:

- `x`: 所需要进行统计推断的数据
- `alternative`: 备择假设的类型, 从`"two.sided"`, `"less"`, `"greater"` 中三选一.
- `mu`: 即要进行比较的常数 $\mu_0$
- `conf.level`: 置信水平 $1-\alpha$, 这个用来求置信区间.
- 其余参数不必进行声明, 请保持默认值. (下一章两样本检验会用到)

**例 3:** 一种在多州之间传播的大型食源性疾病的罪魁祸首是肠炎沙门氏菌. 流行病学家推定疾病的源头是冰激凌. 他们从某冰激凌生产厂家中采集了九个产品的样本, 检测其中的肠炎沙门氏菌水平, 单位为 MPN/g (每克最可能的数量), 数值如下所示:

```
x <- c(0.593,0.142,0.329,0.691,0.231,0.793,0.519,0.392,0.418)
```

基于此数据, 判断这些冰激凌中肠炎沙门氏菌的平均水平是否高于危险水平 0.3MPN/g. 设 $\alpha = 0.05$.

**解答:**

```
t.test(x, mu=0.3, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  x
## t = 2.2051, df = 8, p-value = 0.02927
## alternative hypothesis: true mean is greater than 0.3
## 95 percent confidence interval:
##  0.3245133       Inf
```

```
## sample estimates:
## mean of x
## 0.4564444
```

**补充练习**: 利用公式手动计算例 3

```
x  <- c(0.593,0.142,0.329,0.691,0.231,0.793,0.519,0.392,0.418)
x.bar <- mean(x); x.sd <- sd(x); n <- length(x)
( t <- (x.bar-0.3)/(x.sd/sqrt(n)) )
```

```
## [1] 2.205059
```

```
( cri.value <- qt(0.95, n-1) )
```

```
## [1] 1.859548
```

```
( p.value <- 1 - pt(t, n-1) )
```

```
## [1] 0.02926516
```

```
( CI.lower <- x.bar - qt(0.95, n-1)*x.sd/sqrt(n) )
```

```
## [1] 0.3245133
```

### 1.1.2 方差已知时: $Z$ 检验

待检验的假设不变, 检验统计量换为

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

命令使用格式:

```
library(BSDA) # 需要额外加载这个包

z.test(x, y = NULL,
       alternative = "two.sided",
       mu = 0, sigma.x = 1, sigma.y = NULL,
       conf.level = 0.95)
```

需要输入的参数为:

- `x`: 所需要进行统计推断的数据
- `alternative`: 备择假设的类型, 从`"two.sided"`, `"less"`, `"greater"` 中三选一.
- `mu`: 即要进行比较的常数 $\mu_0$
- `sigma.x`: 即已知的总体标准差 $\sigma$.
- `conf.level`: 置信水平 $1-\alpha$, 这个用来求置信区间.
- 其余参数不必进行声明, 请保持默认值. (下一章两样本检验会用到)

**例 4:** 某项研究的对象为 25 名接受了充血性心衰新疗法的成年男性病人. 研究的一项指标为一个疗程 (4 周) 内运动耐力的提升量 (单位: 分钟). 数值如下所示

```
increase <- c(2.10,2.86,2.25,0.83,1.87,3.53,2.69, 2.86,1.63,2.23,
              2.82,1.66,1.26,1.15,3.93,2.83,2.07,-0.07,3.55,2.94,
              1.05,3.50,1.83,2.15,0.69)
```

现在研究人员想要评估这种新疗法能不能够提升病人的平均运动耐力. 从前一次大规模的研究可以得到 $\sigma = 1.05$. 设 $\alpha = 0.05$, 能得出什么结论?

**解答:**

```
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
z.test(increase, alternative = "greater",
       mu = 0, sigma.x = 1.05,
       conf.level = 0.95)
```

```
##
##  One-sample z-Test
##
## data:  increase
## z = 10.326, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.822981       NA
## sample estimates:
## mean of x
##    2.1684
```

### 1.1.3 检验的功效

单侧检验:

$$\text{Power}\,(\mu_1) = \Phi\left(z_\alpha + \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right)$$

双侧检验:

$$\text{Power}\,(\mu_1) = \text{P}\left(Z \le -z_{1-\alpha/2} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) + \text{P}\left(Z \le -z_{1-\alpha/2} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

代码:

```
library(asbio) # 需要额外加载这个包

power.z.test(sigma = 1, n = 5, power = NULL, alpha = .05, effect = NULL,
             test = c("two.tail", "one.tail"), strict = FALSE)
```

需要输入的参数为:

- sigma: 即已知的总体标准差 $\sigma$.
- n: 样本量
- power: 功效
  - 注: 此命令中 n 与 power 只需指定其中之一 (然后计算的是另一个), 这里要指定 n.
- alpha: 给定的显著性水平
- effect: 即 $\Delta = \mu_1 - \mu_0$
- test: 有两种选择, 单边为"one.tail", 双边为"two.tail"
- strict: 是否使用近似计算公式, 若是, 则指定 strict = FALSE

**例 5** 一种钙离子阻滞剂新药用于治疗不稳定型心绞痛 (一种严重的类型), 其对心率的影响仍未知. 假设有 20 个病人参与研究, 用药 48 小时后, 心率的标准差为每分钟 10 次. 若假设从基准心率到 48 小时后心率的真实平均改变为每分钟 $\pm 5$ 次, 那么在用药 48 小时后检测到心率有显著性改变的功效是多少? (设 $\alpha = 0.05$)

**解答:**

严格计算

```
power.z.test(sigma = 10, n = 20, power = NULL,
             alpha = .05, effect = 5,
             test = "two.tail", strict = T)$power
```

```
## [1] 0.6087795
```

使用近似公式:

```
power.z.test(sigma = 10, n = 20, power = NULL,
             alpha = .05, effect = 5,
             test = "two.tail", strict = F)$power
```

```
## [1] 0.6087659
```

**补充练习**: 利用公式手动计算例 5

```
delta = 5; alpha=0.05; sigma = 10; n = 20;

# 精确计算
power.exact <- pnorm(-qnorm(1-alpha/2)+delta/(sigma/sqrt(n))) +
                pnorm(-qnorm(1-alpha/2)-delta/(sigma/sqrt(n)))
power.exact
```

```
## [1] 0.6087795
```

```
# 近似公式
power.approx <- pnorm(-qnorm(1-alpha/2)+abs(delta)/(sigma/sqrt(n)))
power.approx
```

```
## [1] 0.6087659
```

### 1.1.4 样本量的确定

单侧检验:

$$n = \frac{\sigma^2 \left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\Delta^2}$$

双侧检验:

$$n = \frac{\sigma^2 \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\Delta^2}$$

代码 (与求 power 的代码类似, 这里指定 `power`):

```
power.z.test(sigma = 1, n = NULL, power = .8, alpha = .05, effect = NULL,
             test = c("two.tail", "one.tail"), strict = FALSE)
```

**例 6** 钙离子阻滞剂例子中如果想要功效达到 80%, 至少需要多少样本量?

**解答:**

```
power.z.test(sigma = 10, n = NULL, power = 0.8,
             alpha = .05, effect = 5,
             test = "two.tail", strict = F)$n
```

```
## [1] 31.39552
```

**补充练习**: 利用公式手动计算例 6

```
delta = 5; alpha=0.05; sigma = 10; power = 0.8;

beta <- 1-power
(n <- sigma^2*(qnorm(1-alpha/2)+qnorm(1-beta))^2/delta^2)
```

```
## [1] 31.39552
```

## 1.2 二项分布比例的单样本检验

### 1.2.1 大样本近似检验

假设为

$$H_0 : p = p_0 \quad \longleftrightarrow \quad H_1 : p \neq (\text{或} >, <)\ p_0$$

检验统计量

$$Z=\frac{\hat{p}-p_0}{\sqrt{p_0\left(1-p_0\right)/n}} \quad \overset{H_0}{\longrightarrow} \quad N(0,1)$$

可直接沿用 $Z$ 检验的手算代码, 或是使用

```
prop.test(x,n,p=p0,
          alternative = c("two.sided", "less", "greater"),correct = T)
```

注意以下两点:

- 这里尽管是 $Z$ 检验, 但因为给的数据通常为 “试验 xx 次, 成功 xx 次” 这样的表述, 所以不适合使用 `z.test` 函数.
- `prop.test` 严格意义上不能算是正态近似, 因为它用的是

$$X^2_{\text{corr}}=\frac{\left(|\hat{p}-p_0|-\frac{1}{2n}\right)^2}{p_0q_0/n} \overset{H_0}{\sim} \chi^2_1$$

  也就是将 $Z$ 检验统计量进行了平方 (回忆数理统计内容, 若 $X\sim N(0,1)$, 则 $X^2\sim\chi^2_1$). 但使用此方法算出的 $p$ 值与使用常规的正态近似 ($Z$ 检验) 所得到的 $p$ 值是相等的.

**例 7** 考察乳腺癌家族史在乳腺癌发病中的影响. 调查了 10,000 名 50-54 岁的女性, 她们的母亲都患有乳腺癌. 在这些被调查者中, 有 400 名女性患有乳腺癌. 现已知该年龄段美国女性的乳腺癌患病率为 2%. 这个样本是否能反映出乳腺癌家族史与乳腺癌发病之间有联系?

**解答:**

使用 `prop.test`

```
prop.test(x=400,n=10000,p=0.02, alternative="two.sided",correct=T)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  400 out of 10000, null probability 0.02
## X-squared = 203.06, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.02
## 95 percent confidence interval:
##  0.03628490 0.04407297
## sample estimates:
##    p
## 0.04
```

手算:

```
p0 <- 0.02; n <- 10000; p <- 400/n; alpha <- 0.05
(Z <- (p-p0-1/(2*n))/sqrt(p0*(1-p0)/n))
```

```
## [1] 14.25
```

```
(cri.value <- qnorm(1-alpha/2))
```

```
## [1] 1.959964
```

```
(p.value <- 2*(1-pnorm(Z)))
```

```
## [1] 0
```

### 1.2.2 精确检验

精确检验利用的是 $p$ 值的定义 (原假设成立的条件下, 出现样本结果或比样本结果更极端的结果的概率) 来直接计算 $p$ 值得到结论.

代码:

```
binom.test(x, n, p = 0.5,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95)
```

需要输入的参数为:

- x: "成功" 次数
- n: 试验总次数
- p: 即 $p_0$, 需要拿来比较的常数
- alternative: 备择假设的类型, 从"two.sided", "less", "greater" 中三选一.
- conf.level: 置信水平 $1-\alpha$, 这个用来求置信区间.

注意: 使用此方法与书上给出的 stata 结果一致, 但与使用公式进行计算的结果有出入.

**例 8** 工作或生活在核电站周围的人的安全性受到广泛讨论, 其中一个可能风险就是暴露在辐射中的人死于癌症的比例更高. 一种研究方法是比例-死亡率研究, 即把在暴露组中由于某特殊原因而造成死亡的比例与总体中对应的比例进行比较. 假设在某核电站工作的 55-60 岁的人中一共有 13 人死亡, 其中 5 人死于癌症. 而根据人口统计报告, 死亡的人中约有 20% 归因于癌症. 由此判断, 这个结果是否显著?

**解答:**

```
binom.test(5, 13, p = 0.2, alternative = "two.sided")
```

```
##
##  Exact binomial test
##
## data:  5 and 13
## number of successes = 5, number of trials = 13, p-value = 0.1541
## alternative hypothesis: true probability of success is not equal to 0.2
## 95 percent confidence interval:
##  0.1385793 0.6842224
## sample estimates:
```

```
## probability of success
##               0.3846154
```

如果使用书上公式:

```
p.value <- 0
for (k in 5:13) {
    p.value <- p.value + dbinom(k,13,0.2)
}
(p.value <- p.value*2)
```

```
## [1] 0.1982612
```

或是使用如下更为简洁的版本 (上下两种代码是等价的):

```
2*sum(dbinom(5:13,13,0.2))
```

```
## [1] 0.1982612
```

### 1.2.3 (大样本近似) 检验的功效

基于大样本近似, 双侧检验下的功效 (单侧检验只需把 $\alpha/2$ 改为 $\alpha$):

$$\mathrm{PWR}\left(p_1\right)=P\left(Z\leq\sqrt{\frac{p_0\left(1-p_0\right)}{p_1\left(1-p_1\right)}}\left(-z_{1-\alpha/2}+\frac{\left|p_0-p_1\right|\sqrt{n}}{\sqrt{p_0\left(1-p_0\right)}}\right)\right)$$

```
power.binom.test<-function(n,p0,p1,sig.level=0.05,side=2)
{
  P0<-sqrt(p0*(1-p0))
  P1<-sqrt(p1*(1-p1))
  if (side==2) power<-pnorm(P0/P1*(qnorm(sig.level/2)
                                   +abs(p0-p1)*sqrt(n)/P0))
  else power<-pnorm(P0/P1*(qnorm(sig.level)+abs(p0-p1)*sqrt(n)/P0))
  return(power)
}
```

**例 9** 我们想要检验的一个假设是, 姐妹之一患过乳腺癌的女性, 其本人患乳腺癌的概率比正常人要高. 假设 50-54 岁女性人群的乳腺癌患病率为 2%. 现在想要调查 500 名 50-54 岁的女性, 其姐妹之一有乳腺癌病史. 如果姐妹有乳腺癌病史的 50-54 岁女性的乳腺癌患病率为 5%, 这个检验的功效是多少?

**解答:**

```
power.binom.test(500,0.02,0.05,sig.level=0.05,side=2)
```

```
## [1] 0.9655385
```

### 1.2.4 (大样本近似) 样本量的确定

基于大样本近似, 双侧检验下的样本量 (单侧检验只需把 $\alpha/2$ 改为 $\alpha$):

$$n = \frac{p_0\left(1-p_0\right)\left(z_{1-\alpha/2}+z_{1-\beta}\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}\right)^2}{\left(p_1-p_0\right)^2}$$

```
size.binom.test<-function(p0,p1,power,sig.level=0.05,side=2)
{
    P<-sqrt((p1*(1-p1))/(p0*(1-p0)))
    P2<-(p1-p0)^2
    P3<-p0*(1-p0)/P2
    if (side==2) size<-P3*((qnorm(1-sig.level/2)+qnorm(power)*P)^2)
    else size<-P3*((qnorm(1-sig.level)+qnorm(power)*P)^2)
    return(size)
}
```

**例 10** 在姐妹乳腺癌病史的案例中, 我们假设了 50-54 岁女性人群的乳腺癌患病率为 2%. 如果姐妹有乳腺癌病史的 50-54 岁女性的乳腺癌患病率为 5%, 给定 $\alpha = 0.05$, 为使功效达到 90%, 需要调查多少名女性?

**解答:**

```
size.binom.test(0.02,0.05,0.9,sig.level=0.05,side=2)
```

```
## [1] 340.6518
```

此外还有一些关于求功效以及样本量的方法, 大家可以自行探索:

- 包 `jmuOutlier` 中的 `power.binom.test` 函数.
- test_binomial: Power and Sample Size Analysis for a Binomial Test (https://rdrr.io/github/nutterb/StudyPlanning/man/test_binomial.html)

# 2 假设检验：双样本推断

## 2.1 配对 $t$ 检验

### 2.1.1 方式 1: 求差转化为单样本 $t$ 检验

令 $d_i$ 为第 $i$ 对中两种处理的指标差值, $\mu_d = \mu_1 - \mu_2 = \frac{1}{n}\sum_{i=1}^{n} d_i$.

待检验的假设为

$$H_0 : \mu_d = \delta \quad \longleftrightarrow \quad H_1 : \mu_d \neq \delta$$

我们只讨论 $\delta = 0$ 的情形. 归结为单样本 $t$ 检验.

```
d <- x - y
t.test(d, alternative = c("two.sided", "less", "greater"),
       mu = 0, conf.level = 0.95)
```

需要输入的参数为:

- `d`: 两配对样本作差, 即 `x - y`
- `alternative`: 备择假设的类型, 从`"two.sided"`, `"less"`, `"greater"` 中三选一.
- `mu`: 即要进行比较的常数 $\delta$(常取 $\delta = 0$)
- `conf.level`: 置信水平 $1-\alpha$, 这个用来求置信区间.

### 2.1.2 方式 2: 使用 t.test 直接进行检验

命令使用格式:

```
t.test(x, y,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = TRUE, var.equal = FALSE,
       conf.level = 0.95)
```

需要输入的参数为:

- `x,y`: 两个给定的样本
- `alternative`: 备择假设的类型, 从`"two.sided"`, `"less"`, `"greater"` 中三选一.
- `mu`: 即要进行比较的常数 $\delta$(常取 $\delta = 0$)
- `conf.level`: 置信水平 $1-\alpha$, 这个用来求置信区间.

需要关注的输出的内容:

- `t`: 检验统计量的值
- `df`: 自由度 $(n-1)$
- `xx percent confidence interval`: $\mu_d$ 的置信区间, 置信水平由输入参数 `conf.level` 来定
- `sample estimates`: $\mu_d$ 的点估计

注意: 此方法得出的结果基于 `x - y`, 使用时注意你的备择假设.

**例 1** 高血压相关研究中的一个重要的假设是, 限制钠的摄入可以降低血压. 但长时间保持低钠饮食有一定难度, 因此有时需要专家介入进行饮食咨询. 现有进行低钠饮食的 8 个研究对象, 收集他们在研究开始时以及进行一周饮食咨询后的尿钠水平 (单位: mEq/8hr), 具体数据如下表所示

| Person | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Baseline | 7.85 | 12.03 | 21.84 | 13.94 | 16.68 | 41.78 | 14.97 | 12.072 |
| Week 1 | 9.59 | 34.50 | 4.55 | 20.78 | 11.69 | 32.51 | 5.46 | 12.95 |
| $d_i$ | -1.74 | -22.47 | 17.29 | -6.84 | 4.99 | 9.27 | 9.51 | -0.88 |

提出合理假设并进行检验. 求一周后尿钠水平平均改变量的 95% 置信区间. 正态性假设是否满足?

**解答:**

使用方法 1:

```
baseline <- c(7.85,12.03,21.84,13.94,16.68,41.78,14.97,12.072)
week1 <- c(9.59,34.50,4.55,20.78,11.69,32.51,5.46,12.95)
d <-  baseline - week1
t.test(d,mu=0,paired = F,alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  d
## t = 0.26421, df = 7, p-value = 0.3996
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -7.043798       Inf
## sample estimates:
## mean of x
##    1.1415
```

使用方法 2:

```
baseline <- c(7.85,12.03,21.84,13.94,16.68,41.78,14.97,12.072)
week1 <- c(9.59,34.50,4.55,20.78,11.69,32.51,5.46,12.95)
t.test(baseline,week1,mu=0,paired = T,alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  baseline and week1
## t = 0.26421, df = 7, p-value = 0.3996
## alternative hypothesis: true difference in means is greater than 0
```

```
## 95 percent confidence interval:
##  -7.043798       Inf
## sample estimates:
## mean of the differences
##                  1.1415
```

## 2.2 两独立正态样本均值的 $t$ 检验

待检验的假设为

$$H_0 : \mu_1 - \mu_2 = \delta \quad \longleftrightarrow \quad H_0 : \mu_1 \neq (\text{或 } >, <)\mu_2 + \delta$$

我们只讨论 $\delta = 0$ 的情形.

### 2.2.1 方差相等

检验统计量

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

其中

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

使用

```
t.test(x1, x2,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = F, var.equal = TRUE,
       conf.level = 0.95)
```

把 `var.equal` 设之为 `TRUE`.

需要关注的输出的内容:

- `t`: 检验统计量的值
- `df`: 自由度 $(n_1 + n_2 - 2)$
- `p-value`: $p-$ 值
- `xx percent confidence interval`: $\mu_1$ 的置信区间, 置信水平由输入参数 `conf.level` 来定
- `sample estimates`: $\mu_x$ 与 $\mu_y$ 的点估计 (两个总体各自均值的点估计)

注意: 数据结构不同, 命令格式也会不一样:

- 两个样本分别放在向量 `x1` 与 `x2` 中

```
t.test(x1, x2,
   alternative = c("two.sided", "less", "greater"),
   mu = 0, paired = F, var.equal = TRUE,
   conf.level = 0.95)
```

- 两个样本都放在向量 y 中, 向量 x 用于指示数据属于哪个样本:

```
t.test(y~x,
   alternative = c("two.sided", "less", "greater"),
   mu = 0, paired = F, var.equal = TRUE,
   conf.level = 0.95)
```

比如说

```
set.seed(123)
x1 <- rnorm(10,5,1); x2 <- rnorm(10,4.9,1)
t.test(x1, x2,
   alternative = c("two.sided", "less", "greater"),
   mu = 0, paired = F, var.equal = TRUE,
   conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = -0.076261, df = 18, p-value = 0.9401
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9705694  0.9025768
## sample estimates:
## mean of x mean of y
##  5.074626  5.108622
```

等价于

```
x <- rep(1:2, each = 10) # 相当于 1...1 2...2
y <- append(x1,x2)   # x1 与 x2 拼在一起
t.test(y~x,
   alternative = c("two.sided", "less", "greater"),
   mu = 0, paired = F, var.equal = TRUE,
   conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  y by x
## t = -0.076261, df = 18, p-value = 0.9401
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9705694  0.9025768
## sample estimates:
```

```
## mean in group 1 mean in group 2
##        5.074626        5.108622
```

**例 2** 一项实验研究某种治疗羊胃中绦虫病方案的效果. 将 24 只年龄与健康状况相近并感染了绦虫的绵羊随机地均分为两组, 其中一组注射药物, 另一组不做任何处理. 6 个月后剖开羊胃并对里面的绦虫进行计数, 所得数值如下所示

```
treatment <- c(18, 43, 28, 50, 16, 32, 13, 35, 38, 33, 6, 7)
control <- c(40, 54, 26, 63, 21, 37, 39, 23, 48, 58, 28, 39)
```

设 $\alpha = 0.05$, 检验注射药物后羊胃中的绦虫平均数是否低于不做处理的羊胃中的绦虫平均数.

**解答:**

```
t.test(treatment,control,alternative="less",var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  treatment and control
## t = -2.2709, df = 22, p-value = 0.01665
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.190165
## sample estimates:
## mean of x mean of y
##  26.58333  39.66667
```

### 2.2.2 方差不等

检验统计量

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \dot{\sim} t_d$$

其中

$$d = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1 - 1) + \left(s_2^2/n_2\right)^2/(n_2 - 1)}$$

命令使用格式:

```
t.test(x, y,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = F, var.equal = FALSE,
       conf.level = 0.95)
```

**例 3** 影响儿童肺功能一个可能的重要环境因素为家庭吸烟量. 选择两组对象进行研究:

- 第一组包括 23 名不吸烟的 5-9 岁儿童, 其父母都吸烟. 可测得这些儿童的用力呼气量 (FEV) 如下所示;

```
x1 <- c(1.59,1.66,0.92,2.69,2.21,1.30,2.98,2.40,1.89,2.73,2.71,
        2.68,2.58,2.49,2.06,1.89,1.83,1.61,1.95,1.21,3.62,2.95,1.31)
```

- 第二组为 20 名同龄的不吸烟儿童, 其父母都不吸烟. 可测得这些儿童的 FEV 如下所示.

```
x2 <- c(2.14,2.11,2.61,2.27,2.40,2.29,2.28,2.85,2.21,2.91,
        1.68,2.53,2.35,2.39,2.45,2.10,2.17,1.89,1.87,2.42)
```

进行假设检验并给出 p 值.

**解答:**

```
t.test(x1, x2,
       alternative = "two.sided",
       mu = 0, paired = F, var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -0.98708, df = 31.425, p-value = 0.3311
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4728212  0.1642994
## sample estimates:
## mean of x mean of y
##  2.141739  2.296000
```

## 2.3 两独立正态样本方差齐性检验

待检验的假设为

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \longleftrightarrow \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

检验统计量为

$$F = \frac{s_1^2}{s_2^2} \overset{H_0}{\sim} F_{n_1-1,n_2-1}$$

命令使用格式:

```
var.test(x, y, ratio = 1,
         alternative = c("two.sided", "less", "greater"),
         conf.level = 0.95)
```

**例 4** 考虑前一章的案例: 心血管风险因素的家族性聚集现象. 现在有两组受试对象:

- 第一组为 100 名 2-14 岁的儿童, 其父亲死于心脏病. 这些儿童的胆固醇平均水平 $x_1$ 为 207.3 mg/dL, 标准差 $s_1$ 为 35.6 mg/dL.
- 第二组为 74 名 2-14 岁的儿童, 其父亲仍健在并且没有患心脏病. 这些儿童的胆固醇平均水平 $x_2$ 为 193.4 mg/dL, 标准差 $s_2$ 为 17.3 mg/dL.

对同方差性进行检验并得出 $p$ 值.

**解答:**

```
# 此处出于练习需要, 采用模拟数据.
# 如果遇到只给描述统计量的情形, 请使用公式算, 不要模拟.
x1 <- rnorm(100,207.3,35.6); x2 <- rnorm(74,193.4,17.3)
var.test(x1, x2, ratio = 1,
         alternative = "two.sided",
         conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 3.3163, num df = 99, denom df = 73, p-value = 2.286e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.140848 5.064974
## sample estimates:
## ratio of variances
##           3.316342
```

## 2.4 离群值的处理

```
library(PMCMRplus)
out <- gesdTest(x, maxr)
```

需要输入的参数:

- x: 样本量
- maxr: 预估的最大离群值数量

**例 5** 判断暴露组的手指-腕轻叩得分数据中是否有离群值.

**解答:**

```
load("C:\\Users\\zjchen\\Desktop\\2021Spring\\2021Summer-BiostatTA\\LAB-slides\\02\\LEAD.DAT.rdata")
library(PMCMRplus)
```

```
## Warning: package 'PMCMRplus' was built under R version 4.0.5
```

```r
# 数据预处理, 不是每个数据集都要这样做
data <- lead[lead$Group==2,]$maxfwt
data <- data[data!=99]
stem(data,scale=1) # 画茎叶图
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 34
##   2 |
##   3 | 45788
##   4 | 001224456889
##   5 | 0122244567789
##   6 | 2
##   7 | 0
##   8 | 3
```

```r
# 寻找离群的值
out <- gesdTest(data,3)
data[out$ix]
```

```
## [1] 83 13 14
```

这里由于数据集删掉了一些等于 99 的值, 因此数据标号会跟讲义不一样.

```r
summary(out)  # 查看 ESD 值, 以及离群值对应的样本序号
```

```
##
##   GESD multiple outlier test
##
## Outliers tested:
##           i        R Pr(>|R|)
## 1        31 2.703855 0.151982
## 2     31 12 2.832990 0.087344 .
## 3  31 12 10 3.223667 0.013816 *
##
## alternative hypothesis:  two.sided
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

这里 R 这一列就代表着 ESD 值.

## 2.5 样本量与功效

### 2.5.1 两独立样本的均值比较

双侧检验下功效的计算公式为:

$$\text{power}(\Delta) = \Phi\left(-z_{1-\alpha/2} + \frac{\Delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right)$$

单侧检验时把 $\alpha/2$ 换为 $\alpha$ 即可.

自己写函数:

```
power.z.twosample <- function(n1,n2,Delta,sigma1,sigma2,alpha=0.05,side=2){
  if (side==2){
    return(pnorm(-qnorm(1-alpha/2)+Delta/sqrt(sigma1^2/n1+sigma2^2/n2)))
  } else {
    return(pnorm(-qnorm(1-alpha)+Delta/sqrt(sigma1^2/n1+sigma2^2/n2)))
  }
}
```

需要输入的参数为:

- `n1,n2`: 两个样本的样本量
- `Delta`: 即 $\Delta = |\mu_1 - \mu_2| = |\mu_d|$, 为两样本之间均值的差异
- `sigma1, sigma2`: 两个样本对应的总体的的标准差
- `alpha`: 显著性水平, 默认为 0.05
- `side`: 双侧检验请用 `side=2`(默认为双侧), 输入其他值时为单侧检验

**例 6** 假设现有服用 OC 者与未服用 OC 者各 100 人, 两组之间实际平均血压差异预期为 $\mu_1 - \mu_2 = 5$ mmHg, 其中服用 OC 的人的平均血压更高. 假设在预先实验中已经知道服用 OC 者与未服用 OC 者的血压标准差估计分别为 15.34 mmHg 与 18.23 mmHg. 这个检验的功效是多少?

**解答:**

```
power.z.twosample(100,100,5,15.34,18.23,alpha=0.05,side=1)
```

```
## [1] 0.6749953
```

假设两个组的样本量比 $n2 : n1 = k : 1$, 即 $n_2 = kn_1$, 则样本量计算公式为

$$n_1 = (\sigma_1^2 + \frac{\sigma_2^2}{k})\frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

自己写函数:

```
size.z.twosample <- function(power,Delta,sigma1,sigma2,alpha=0.05,side=2,k=1){
  beta <- 1-power
  if (side==2){
    n1 <- ceiling((sigma1^2+ sigma2^2/k)*(qnorm(1-alpha/2)+qnorm(1-beta))^2/Delta^2)
    n2 <- ceiling(k*n1)
  } else {
```

```
    n1 <- ceiling((sigma1^2+ sigma2^2/k)*(qnorm(1-alpha)+qnorm(1-beta))^2/Delta^2)
    n2 <- ceiling(k*n1)
  }
  return(list(n1=n1,n2=n2))
}
```

需要输入的参数为:

- power: 想要达到的功效
- Delta: 即 $\Delta = |\mu_1 - \mu_2| = |\mu_d|$, 为两样本之间均值的差异
- sigma1, sigma2: 两个样本对应的总体的的标准差
- alpha: 显著性水平, 默认为 0.05
- side: 双侧检验请用 side=2(默认为双侧), 输入其他值时为单侧检验
- k: 样本量比 $n2 : n1 = k : 1$, 即 $n_2 = kn_1$

**例 7** 现规定不服用 OC 组的人数是服用 OC 组人数的两倍. 在预先实验中已经知道服用 OC 者与未服用 OC 者的血压标准差估计分别为 15.34 mmHg 与 18.23 mmHg. 在假定双侧检验, 显著性水平 $\alpha = 0.05$, 功效为 80%, 平均血压的真实差异为 5 mmHg, 求样本量.

**解答:**

```
size.z.twosample(0.8,5,15.34,18.23,alpha=0.05,side=1,k=2)
```

```
## $n1
## [1] 100
##
## $n2
## [1] 200
```

### 2.5.2 纵向研究

双侧检验下的功效计算公式:

$$\text{power}(\mu_d) \approx P\left(Z \leq -z_{1-\alpha/2} + \frac{|\mu_d|}{\sigma_d/\sqrt{n}}\right)$$

注: 这里方便起见, 假设 $\sigma_d$ 未知而 $\sigma_1, \sigma_2, \rho$ 已知.

```
power.z.long <- function(n,Delta,sigma1,sigma2,rho,alpha=0.05,side=2){
  sigma_d <- sqrt(sigma1^2+sigma2^2-2*rho*sigma1*sigma2)
  if (side==2){
    return(pnorm(-qnorm(1-alpha/2)+abs(Delta)/(sigma_d/sqrt(n))))
  } else {
    return(pnorm(-qnorm(1-alpha)+abs(Delta)/(sigma_d/sqrt(n))))
  }
}
```

需要输入的参数为:

- `n`: 样本量
- `Delta`: 即 $\Delta = |\mu_1 - \mu_2| = |\mu_d|$, 为受试对象在研究开始与随访中的两次指标测量值之差的均值
- `sigma1, sigma2`: 研究开始与随访中的两次指标测量值对应的标准差
- `rho`: 基础水平与一年后随访时水平的相关系数
- `alpha`: 显著性水平, 默认为 0.05
- `side`: 双侧检验请用 `side=2`(默认为双侧), 输入其他值时为单侧检验

**例 8** 假设我们计划在纵向研究中比较处理组与对照组的收缩压 (SBP) 平均变化值. 在研究开始时 SBP 水平与随访时水平的标准差都为 15 mmHg 基础 SBP 水平与一年后随访时水平的相关系数为 0.70. 已知处理组的 SBP 在一年间的变化平均降低量为 8 mmHg, 而对照组的平均降低量为 3mmHg. 若要使用双侧检验, 并且给定 $\alpha = 0.05$, 且现在每组都招募到了 35 名研究对象. 这个研究的功效是多少?

**解答:**

```
power.z.long(35,-5,15,15,0.7,alpha=0.05,side=2)
```

```
## [1] 0.7210325
```

样本量:

$$n = \frac{\sigma_d^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\mu_d^2}$$

```
size.z.long <- function(power,Delta,sigma1,sigma2,rho,alpha=0.05,side=2){
  beta <- 1-power
  sigma_d <- sqrt(sigma1^2+sigma2^2-2*rho*sigma1*sigma2)
  if (side==2){
    return( ceiling(sigma_d^2*(qnorm(1-alpha/2)+qnorm(1-beta))^2/Delta^2 ))
  } else {
    return( ceiling(sigma_d^2*(qnorm(1-alpha)+qnorm(1-beta))^2/Delta^2 ))
  }
}
```

需要输入的参数为:

- `power`: 功效
- `Delta`: 即 $\Delta = |\mu_1 - \mu_2| = |\mu_d|$, 为受试对象在研究开始与随访中的两次指标测量值之差的均值
- `sigma1, sigma2`: 研究开始与随访中的两次指标测量值对应的标准差
- `rho`: 基础水平与一年后随访时水平的相关系数
- `alpha`: 显著性水平, 默认为 0.05
- `side`: 双侧检验请用 `side=2`(默认为双侧), 输入其他值时为单侧检验

**例 9** 假设我们计划在纵向研究中比较处理组与对照组的收缩压 (SBP) 平均变化值. 在研究开始时 SBP 水平与随访时水平的标准差都为 15 mmHg 基础 SBP 水平与一年后随访时水平的相关系数为 0.70. 已知处理组的 SBP 在一年间的变化平均降低量为 8 mmHg, 而对照组的平均降低量为 3mmHg. 若要使用双侧检验, 并且给定 $\alpha = 0.05$, 则应该要抽多少样本才能以 80% 的功效检验出两组之间存在显著差异?

**解答:**

```
size.z.long(0.8,-5,15,15,0.7,alpha=0.05,side=2)
```

```
## [1] 43
```

# 3 非参数检验

## 3.1 配对样本的检验

### 3.1.1 符号检验

注意到我们的假设等价于

$$H_0 : p = \frac{1}{2} \quad \longleftrightarrow \quad H_1 : p \neq \frac{1}{2}$$

其中 $p$ 是差值中正数的比例. 于是这归结到第七章种二项分布比例的单样本检验.

#### 3.1.1.1 大样本: 正态近似法

沿用第七章代码 (这里 p 固定为 p=0.5):

```
prop.test(x, n, p = 0.5,
          alternative = c("two.sided", "less", "greater"),correct = T)
```

输入参数 (与第七章中是一致的, 但这里有特殊含义):

- x: 差值 $d_i$ 中正数的个数 (讲座课课件上用的是 $C$)
- n: 非零差值的个数 (注意! $n$ 在此之前被用于表示样本量.)
- alternative 与 correct 与第七章所述一致.

**例 1** 某研究进行一项口腔教育计划以促进更好的口腔卫生. 在计划实施前以及实施 6 个月后分别对 28 例轻度牙周病的成人患者进行评估. 计划实施六个月后, 有 15 例患者的牙周状况有所改善, 8 例患者牙周状况恶化, 5 例患者状况不变. 试用统计学的方法来评价这项计划的效果.

**解答:**

```
prop.test(x=15,n=23,p=0.5, alternative="greater",correct=T)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  15 out of 23, null probability 0.5
## X-squared = 1.5652, df = 1, p-value = 0.1055
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.4595101 1.0000000
## sample estimates:
##         p
## 0.6521739
```

#### 3.1.1.2 小样本: 精确法

沿用第七章代码 (这里 p 固定为 p=0.5):

```
binom.test(x, n, p = 0.5,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95)
```

类似于 prop.test, 这里的输入参数为

- x: 差值 $d_i$ 中正数的个数 (讲座课课件上用的是 $C$)
- n: 非零差值的个数 (注意! $n$ 在此之前被用于表示样本量.)
- alternative 与 correct 与第七章所述一致.

**例 2** 假设要比较两种不同的眼药水 (A,B) 的效果, 其作用是防止花粉热病人染上红眼病. 眼药水 A 被随机地点往其中一只眼, 眼药水 B 被点往另一只眼. 眼红程度在治疗前 (基准) 已被测定, 且 10 分钟后由一个不知晓用药情况的观察者再次测定. 15 人在治疗前左右眼的眼红程度相同. 10 分钟后:

- 2 人点了 A 药的眼比点了 B 药的眼红色程度更浅;
- 8 人点了 B 药的眼比点了 A 药的眼红色程度更浅;
- 5 人两只眼睛红色程度相同.

问: 两种药的效果有差别吗?

**解答:**

```
binom.test(x=2,n=10,p=0.5,alternative="two.sided")
```

```
##
##  Exact binomial test
##
## data:  2 and 10
## number of successes = 2, number of trials = 10, p-value = 0.1094
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.02521073 0.55609546
## sample estimates:
## probability of success
##                    0.2
```

#### 3.1.1.3 补充内容: 今天课上的小证明

只考虑 $C \neq D$ 的情形, 注意到 $C + D = n$, 当 $C > n/2$ 时, $|C - D| > 0 \Longrightarrow |C - D| = C - D = C - (n - C) = 2C - n$, 所以

$$
\begin{aligned}
p &= 2 \times \left[1 - \Phi\left(\frac{|C - D| - 1}{\sqrt{n}}\right)\right] \\
&= 2 \times \left[1 - \Phi\left(\frac{2C - n - 1}{\sqrt{n}}\right)\right] \\
&= 2 \times \left[1 - \Phi\left(\frac{C - n/2 - 1/2}{\sqrt{n}/2}\right)\right]
\end{aligned}
$$

同理 $C < n/2$ 时, $|C-D|<0 \Longrightarrow |C-D| = D-C = (n-C)-C = n-2C$, 所以

$$
\begin{aligned}
p &= 2\times\left[1-\Phi\left(\frac{|C-D|-1}{\sqrt{n}}\right)\right] \\
&= 2\times\left[1-\Phi\left(\frac{n-2C-1}{\sqrt{n}}\right)\right] \\
&= 2\times\left[1-\Phi\left(\frac{n/2-C-1/2}{\sqrt{n}/2}\right)\right] \\
&= 2\times\Phi\left(\frac{C-n/2-1/2}{\sqrt{n}/2}\right)
\end{aligned}
$$

证毕.

### 3.1.2 Wilcoxon 符号秩检验

```
wilcox.test(x, y= NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = TRUE, exact = NULL,
            correct = TRUE, conf.int = FALSE, conf.level = 0.95)
```

输入参数为:

- `x, y`: 观察数据构成的向量
- `paired`: 说明变量 `x, y` 是否非为成对数据
  - 如果在上述命令中只指定了 `x`, 那么这些值会被自动认作是差值, 此处要用 `paired=F`
  - 如果在上述命令中同时指定了 `x` 与 `y`, 那么这里要用 `paired=T`
- `exact`: 说明是否精确计算 $p$ 值. 若 $n<16$, 设置 `exact=TRUE`, 否则用 `exact=FALSE`.
- `correct`: 说明是否对 $p$ 值采用连续性修正.
- `conf.int`: 说明是否给出相应的置信区间

补充说明: 此函数无法精确计算带结的数据的 $p$ 值 (会出现 warning message: cannot compute exact p-value with ties). 这是这个函数本身的缺陷. 解决方法一是使用 `coin` 包中的 `wilcox_test`(似乎用不了), 二是使用 `CNPS` 包中的 `pairwise_test`.

```
pairwise_test(x, y= rep(0,length(x)),alternative = c("two.sided", "less", "greater"))
```

输入的参数

- `x,y`: 数据向量
  - 如果只知道差值, 那就输入进 `x`, 然后令 `y= rep(0,length(x))`
  - 如果知道两个配对的样本的数据, 则分别输入 `x` 与 `y` 中
- `alternative`: 备择假设, 默认为"greater"

总结: 在 Wilcoxon 符号秩检验中,

- 存在结的时候:
  - 如果要用正态近似, 则请手动按照讲义上的公式进行计算

  - 如果不需要正态近似, 则用 `pairwise_test`
- 不存在结的时候: 直接用 `wilcox.test`

**例 3** 假设我们想比较两种软膏 (A,B) 在治疗因阳光照射而导致的皮肤过度发红症状中的效果. 两种软膏随机地被涂抹在左或右手手臂上. 被试者在涂抹软膏后接受一小时的阳光照射. 用 10 分制来定量地衡量晒红的程度, 其中 10 分表示最严重的晒红, 1 分表示没有被晒红. 打分如下 `x` 所示 (注意: 这里打的分数已经是差值了.)

```
d <- -8:3 # 可以打到的分数
f <- c(1,3,2,2,1,5,4,4,5,10,6,2) # 每个分数对应的频数
x <- rep(d,f)
```

两种药的效果有差别吗?

**解答:**

注意: 这个例子中数据有结, 不适宜使用 `wilcox.test`. 此处仅为练习, 更合适的方法是用书本上的修正后的正态近似公式进行计算.

```
wilcox.test(x, y= NULL, alternative = "two.sided",
            paired = FALSE, exact = F, correct = TRUE, conf.int = FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x
## V = 248, p-value = 0.02869
## alternative hypothesis: true location is not equal to 0
```

## 3.2 独立样本的检验

### 3.2.1 Wilcoxon 秩和检验

```
wilcox.test(x, y,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = FALSE,
            correct = TRUE, conf.int = FALSE, conf.level = 0.95)
```

在这里, 你要:

- 同时输入 `x,y`
- `paired=FALSE`
- 若 $\min\{n_1, n_2\} < 10$, 则使用 `exact = TRUE`

其余参数的同上.

注意: 这里输出的检验统计量不是我们上课定义的正态近似下的 $T$ 统计量, 而是 W 统计量 (也称为 Mann-

Whitney 统计量):

$$W = R_1 - \frac{n(n+1)}{2}$$

补充说明: 这里仍然避免不了无法处理有结的问题. 但仍可以使用包 `CNPS` 中的 `twosample_test` 函数解决:

```
library(CNPS)
twosample_test(x,y,alternative="two.sided", method_p=c("sampling","asymptotic","exact"))
```

现在要用到的输入参数的设置:

- `alternative`: 备择假设, 默认是 `greater`
- `method_p`: p-值的计算方法
  - `sampling`: 置换检验 + 抽样
  - `asymptotic`: 大样本正态近似 (对应的检验统计量与讲座课课件以及书本一致), 但计算代码还是有误, 就不要用了.
  - `exact`: 置换检验 (遍历所有情形)

这里我们只关注函数输出的两个内容:

- `Dobs` 这是修正后的 `x` 的秩和
  - 为什么这个能处理有结的情形呢? 因为这里面调用了 `R` 中的 `rank` 函数, 能够自动处理有结的秩分配.
- `p-value`: p-值

包中函数具体的计算原理以及更多的参数设置可以等到大三第二学期《非参数统计》课程 (Lecturer: 项冬冬) 再去了解.

总结: 在 Wilcoxon 秩和检验中,

- 存在结的时候:
  - 如果要用正态近似, 则请手动按照讲义上的公式进行计算
  - 如果不需要正态近似, 则可在 `twosample_test` 中调用 `method_p="sampling"` 或 `method_p="exact"` 进行计算
- 不存在结的时候: 直接用 `wilcox.test`

**例 4** 通常认为视网膜色素变性 (RP) 的不同基因型对应着不同的疾病进展速度: 显性进展最慢, 隐性次之, 伴性形式进展最快. 假设现在有 25 名 10-19 岁的显性患者, 30 名伴性患者. 在配镜矫正后, 这些患者较好视力眼的视敏度如下表所示

| 视敏度 | 20-20 | 20-25 | 20-30 | 20-40 | 20-50 | 20-60 | 20-70 | 20-80 |
|---|---|---|---|---|---|---|---|---|
| 显性患者 | 5 | 9 | 6 | 3 | 2 | 0 | 0 | 0 |
| 伴性患者 | 1 | 5 | 4 | 4 | 8 | 5 | 2 | 1 |

检验这两组患者的视敏度的分布是否存在差异.

**解答:**

```r
Ratings <- 8:1
fx <- c(5,9,6,3,2,0,0,0)
fy <- c(1,5,4,4,8,5,2,1)
x <- rep(Ratings,fx)
y <- rep(Ratings,fy)

wilcox.test(x, y, alternative = "two.sided",
            paired = FALSE, exaxt=FALSE,
            correct = TRUE, conf.int = FALSE)
```

```
## Warning in wilcox.test.default(x, y, alternative = "two.sided", paired =
## FALSE, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 596, p-value = 0.0001513
## alternative hypothesis: true location shift is not equal to 0
```

```r
library(CNPS)
```

```
## Warning: package 'CNPS' was built under R version 4.0.5
```

```r
twosample_test(x,y,alternative="two.sided",method_p = "sampling")
```

```
##
##      Two sample test using wilcoxon scoring
## Dobs = 921
## sampling with replacement method to calculate :
##  p-value =  < 2.2e-16
## alternative hypothesis:
##  The mean of the first is not equal to the mean of the second
## 95% confidence interval of p-value :
##  [ 0 , 0 ]
##
##  The Hodges-Lehmann statistic = 2
## The 95 % CI for mean difference is [ 1 , 3 ]
```

# 4 假设检验: 分类数据 (属性数据)

## 4.1 二项分布比例的两样本检验

假设

$$x_1 \sim \text{bin}(p_1, n_1), \quad x_2 \sim \text{bin}(p_2, n_2)$$

我们想要检验的假设为

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_a : p_1 \neq p_2$$

### 4.1.1 (独立, 大样本) 正态近似

根据

$$\hat{p}_1 - \hat{p}_2 \overset{H_0}{\sim} N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

构造如下检验统计量:

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

使用的命令为:

```
prop.test(x, n, alternative = "two.sided",
          conf.level = 0.95, correct = TRUE)
```

输入的参数:

- x: 向量 $(n_{11}, n_{21})$, 其中 $n_{i1} = p_i n_i$
- n: 向量 $(n_1, n_2)$

**例 1** 要研究导致乳腺癌 (BCa) 的风险因素, 病例组为选定的医院里的女性乳腺癌患者, 对照组为同一时间同一间医院里的非乳腺癌且年龄相近的女性. 数据如下所示

| | 首次生育年龄 | | |
|---|---|---|---|
| **BC 状态** | $\geq 30$ | $\leq 29$ | 总数 |
| 病例组 | 683 | 2,537 | 3,220 |
| 对照组 | 1,498 | 8,747 | 10,245 |

问: 乳腺癌与首次生育年龄有关吗? (即, 病例组与对照组中首次生育年龄 $\geq 30$ 的比例是否相同?)

**解答:**

```
x=c(683,1498); n=c(3220, 10245)
prop.test(x, n, alternative = "two.sided",
          conf.level = 0.95, correct = TRUE)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 77.885, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04999981 0.08178846
## sample estimates:
##    prop 1    prop 2
## 0.2121118 0.1462177
```

### 4.1.2 (独立, 大样本) 卡方检验

$$X^2 = \sum_{i,j=1,2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{H_0}{\sim} \chi_1^2$$

构造列联表: 使用 `matrix` 函数

```
t22 <- matrix(c(a11,a12
                a21,a22),
              nrow=2, ncol=2, byrow=TRUE)
```

检验的时候用 `chisq.test`

```
chisq.test(t22)
```

**例 2** 某项研究关注口服避孕药 (OC) 对 40 至 44 岁女性的心脏疾病的影响, 数据如下表所示.

| | 三年内患心肌梗塞 | | |
|---|---|---|---|
| **OC 服用情况** | 是 | 否 | 总数 |
| 服用 | 13 | 4,987 | 5,000 |
| 未服用 | 7 | 9,993 | 10,000 |
| 总数 | 20 | 14,980 | 15,000 |

问: 三年内患心肌埂塞与是否服用口服避孕药有关系吗?

**解答:**

```
t22<-matrix(c(13,4987,7,9993),nrow = 2, ncol = 2,byrow = TRUE)
t22
```

```
##      [,1] [,2]
## [1,]   13 4987
## [2,]    7 9993
```

```
chisq.test(t22)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t22
## X-squared = 7.6665, df = 1, p-value = 0.005626
```

### 4.1.3 (独立, 小样本) Fisher 精确检验

对列联表进行重排, 使得 $n_1 \le n_2$ 且 $m_1 \le m_2$. 重排后 $p_1$ 对应于样本量较小的那个群体. 令 $k = \min(m_1, n_1)$.

- 对形如 $H_a : p_1 < p_2$ 的备择假设,

$$p-\text{value} = P(X \le a) = P(X = 0) + P(X = 1) + \cdots + P(X = a)$$

- 对形如 $H_a : p_1 > p_2$ 的备择假设

$$p-\text{value} = P(X \ge a) = P(X = a) + P(X = a + 1) + \cdots + P(X = k)$$

- 双边检验 $H_a : p_1 \ne p_2$

$$p-\text{value} = 2 \times \min[P(X \le a), P(X \ge a), 0.5].$$

构造列联表的方式同上, 检验的时候使用 `fisher.test`

```
fisher.test(t22)
```

**例 3** 某项回顾性研究关注饮食习惯与心血管病之间的关系, 研究对象为某地区一个月内已故的 50-54 岁男性.

| 死因 | 饮食类型 | | |
|---|---|---|---|
| | 高盐 | 低盐 | 总数 |
| 非心血管病 | 2 | 23 | 25 |
| 心血管病 | 5 | 30 | 35 |
| 总数 | 7 | 53 | 60 |

问: 饮食习惯与心血管病之间有关系吗?

**解答:**

```
t22<-matrix(c(2,23,5,30),nrow = 2, ncol = 2, byrow = TRUE)
t22
```

```
##      [,1] [,2]
## [1,]    2   23
## [2,]    5   30
```

```
fisher.test(t22)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t22
## p-value = 0.6882
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.04625243 3.58478157
## sample estimates:
## odds ratio
##   0.527113
```

#### 4.1.4 (成对) McNemar 检验

设 $p = \mathbb{P}$(不一致对为 A 型), 如果两种处理效果相同, 那么 A 型与 B 型不一致对的数量相等. 因此我们想要检验的假设为

$$H_0 : p = \frac{1}{2} \quad \text{vs} \quad H_a : p \neq \frac{1}{2}.$$

**4.1.4.1 方法 1: 直接使用函数对列联表进行检验** 列联表的构造同上, 检验的时候使用:

```
mcnemar.test(t22)
```

**4.1.4.2 方法 2: 当作符号检验进行处理**

- 大样本下, 使用 `prop.test`

  ```
  prop.test(x=n_A,n=n_D, p=0.5,correct=TRUE,
          alternative="two.sided")
  ```

  其中 `n_A`, `n_D` 分别为 A 型不一致对与不一致对的总数, 下同.

- 小样本下, 使用 `binom.test`

  ```
  binom.test(x=n_A,n=n_D,p=0.5, alternative="two.sided")
  ```

**例 4** 比较乳房切除术后两种化疗方式对乳腺癌的治疗效果. 将病人按年龄 (差别在五岁以内) 及临床状况进行配对, 并随访五年.

| | 疗法 B | | |
|---|---|---|---|
| **疗法 A** | 存活 | 死亡 | 总数 |
| 存活 | 510 | 16 | 526 |
| 死亡 | 5 | 90 | 95 |
| 总数 | 515 | 106 | 621 (对) |

问: 两种治疗方法没有显著差异?

**解答:**

方法 1: 直接使用函数对列联表进行检验

```
t22<-matrix(c(510,16,5,90),nrow = 2, ncol = 2, byrow = TRUE)
t22
```

```
##      [,1] [,2]
## [1,]  510   16
## [2,]    5   90
```

```
mcnemar.test(t22)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  t22
## McNemar's chi-squared = 4.7619, df = 1, p-value = 0.0291
```

方法 2: 当作符号检验进行处理

```
prop.test(x=16,n=16+5, p=0.5,correct=TRUE,
          alternative="two.sided")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  16 out of 16 + 5, null probability 0.5
## X-squared = 4.7619, df = 1, p-value = 0.0291
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5245026 0.9088286
## sample estimates:
##         p
## 0.7619048
```

## 4.2 $R \times C$ 列联表中的检验

### 4.2.1 关联性检验

仍然使用 `chisq.test`.

**例 5** 一项关于乳腺癌的病例对照研究的数据如下所示:

| 状态 | 首次生育年龄 | | | | | 总数 |
|---|---|---|---|---|---|---|
| | $< 20$ | $20-24$ | $25-29$ | $30-34$ | $\geq 35$ | |
| 病例组 | 320 | 1206 | 1011 | 463 | 220 | 3220 |
| 对照组 | 1422 | 4432 | 2893 | 1092 | 406 | 10245 |
| 总数 | 1742 | 5638 | 3904 | 1555 | 626 | 13465 |

问: 是否患病与首次生育年龄有关吗?

**解答:**

```
freq<-matrix(c( 320,1206,1011, 463,220,
                1422,4432,2893,1092,406),nrow = 2, ncol = 5, byrow = TRUE)
freq
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  320 1206 1011  463  220
## [2,] 1422 4432 2893 1092  406
```

```
chisq.test(freq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  freq
## X-squared = 130.34, df = 4, p-value < 2.2e-16
```

### 4.2.2 趋势的检验

使用的命令:

```
prop.trend.test(x, n, score)
```

输入的参数为:

- `x` 为向量, 各分量为每组成功的数量.
- `n` 为向量, 各分量为每组样本量.
- `score` 为向量, 各分量为每组的分数. 关于 score 的设置:
  - 默认情况下组中样本量作为分数
  - 可根据实际情况赋予分数 (比如说, 行指标为年龄, 则可用组距的 1/2 作为分数, 见 Example 10.42, P418)
  - 简便起见, 可以设分数为 `1:` 总的组数
- 如果有 $k$ 组, 那么 `x,n,score` 都应该为 $k$ 维向量.

**例 6** 前面的结果揭示了乳腺癌与首次生育年龄之间存在某种关联, 但并没有告诉我们具体存在何种关联. 我们可以看出女性患乳腺癌的比例有随着表格列数 (即年龄) 增加而增加的趋势. 问: 诸 $p_i$ 之间是否存在趋势?

**解答:**

```
x<- c(320, 1206, 1011, 463, 220)
n<- c(1742, 5638, 3904, 1555, 626)
score<- c(1:5)
prop.trend.test(x,n,score)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x out of n ,
##  using scores: 1 2 3 4 5
## X-squared = 129.01, df = 1, p-value < 2.2e-16
```

## 4.3 卡方拟合优度检验

因为这里可以用很多分布进行拟合, 每个分布的命令都不一样, 所以就不写成一个大的函数的形式, 用的时候直接修改下面给出的代码即可.

比如说我们想要检验正态分布对于数据的拟合性.

方法 1:

```
L <- # 这里构造各个组的下限
U <- # 这里构造各个组的上限
prob <- pnorm(U, x.bar , x.sd)-pnorm(L, x.bar , x.sd) # 计算想要的分布在各组的概率
observed <- c() # 各组的观测频数
n <- sum(observed)
expected <- n*prob # 计算各组的期望频数
X <- sum((observed- expected)^2/expected) # 检验统计量
df <- length(observed) - 1 - 2 # 自由度
p_value <- 1 - pchisq(X,df) #p 值
```

方法 2:

```
observed <- c() # 各组的观测频数
p <- pnorm(c(seq(下限, 上限, 组距)), x.bar , x.sd)
p <- c(p[1], p[2]-p[1], p[3]-p[2], ... , p[组数]-p[6], 1-p[7])
chisq.test(observed, p = p)
```

**例 7** 东波士顿地区 14,736 名 30-69 岁成年人的舒张压 (DBP) 的频数分布如下表所示

| 组别 (mmHg) | 观测频数 | 期望频数 |
|---|---|---|
| $< 50$ | 57 | |
| $\geqslant 50, < 60$ | 330 | |
| $\geqslant 60, < 70$ | 2132 | |
| $\geqslant 70, < 80$ | 4584 | |
| $\geqslant 80, < 90$ | 4604 | |
| $\geqslant 90, < 100$ | 2119 | |
| $\geqslant 100, < 110$ | 659 | |
| $\geqslant 110$ | 251 | |

我们想要检验这些数据是否能用正态分布进行拟合.

方法 1:

```
L<- c(-Inf,seq(50,110,10))
U<- c(seq(50,110,10),Inf)
prob<-pnorm(U,80.68,12)-pnorm(L,80.68,12)
observed=c(57,330, 2132,4584, 4604,2119,659,251)
n<-sum(observed)
(expected <- n*prob)
```

```
## [1]   77.86534  547.14930 2126.68201 4283.34879 4478.51955 2431.12761  684.08614
## [8]  107.22127
```

```
(X<- sum((observed- expected)^2/expected))
```

```
## [1] 350.198
```

```
df <- length(observed) - 1 - 2
(p_value <- 1 - pchisq(X,df))
```

```
## [1] 0
```

方法 2:

```
observed <- c(57, 330, 2132, 4584, 4604, 2119, 659, 251)
p <- pnorm(c(seq(50, 110, 10)), 80.68, 12)
p <- c(p[1], p[2]-p[1], p[3]-p[2] ,p[4]-p[3], p[5]-p[4], p[6]-p[5], p[7]-p[6], 1-p[7])
chisq.test(observed, p = p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 350.2, df = 7, p-value < 2.2e-16
```

## 4.4 Kappa 统计量

列联表的构造同上,

```
library(grid)
library(vcd)
Kappa(t22)
```

输出部分只需关注 $\kappa$ 值, $p-$ 值是对于检验 $H_0 : \kappa = 0$ 而言的.

**例 8** 两次不同调查中, 537 名美国女护士的牛肉消费情况如下表所示

| | 调查 2 | | |
|---|---|---|---|
| 调查 1 | $\leq 1$ 次/周 | $> 1$ 次/周 | 总数 |
| $\leq 1$ 次/周 | 136 | 92 | 228 |
| $> 1$ 次/周 | 69 | 240 | 309 |
| 总数 | 205 | 332 | 537 |

计算 Kappa 统计量的值并进行显著性检验

**解答:**

```
library(grid)
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.0.5
```

```
##
## Attaching package: 'vcd'
```

```
## The following object is masked from 'package:asbio':
##
##     Kappa
```

```
## The following object is masked from 'package:BSDA':
##
##     Trucks
```

```
t22 <- matrix(c(136, 92,
                69, 240), nrow = 2, byrow = T)
Kappa(t22)
```

```
##             value     ASE     z  Pr(>|z|)
## Unweighted 0.3782 0.04045 9.351 8.705e-21
## Weighted   0.3782 0.04045 9.351 8.705e-21
```

# 5 回归与相关性方法

## 5.1 线性回归

我们主要讨论的模型会有两种:

- 带常数项: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$, 其中 $\varepsilon \sim N(0, \sigma^2)$
- 不带常数项: $y = \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$, 其中 $\varepsilon \sim N(0, \sigma^2)$

我们主要使用到的命令就是 `lm`

- 带常数项

```
fit <- lm(y ~ 1 + x_1 + ... + x_m)
fit <- lm(y ~ x_1 + ... + x_m)
```

- 不带常数项

```
fit <- lm(y ~ x_1 + ... + x_m - 1)
```

通过 `summary` 命令就可以输出 `lm` 的结果

```
summary(fit)
```

### 5.1.1 参数估计 (模型拟合)

如果要得到参数的点估计, 则看 `summary()` 输出的 `Coefficients` 板块:

- 横行表示你的模型里面的参数, `(Intercept)` 为常数项.
- 纵列分别为
  - `Estimate`: 对该参数的点估计
  - `Std. Error`: 对该参数点估计的标准差
  - `t value`: 对该参数 (设为 $\beta_j$) 进行假设检验 $H_0 : \beta_j = 0$ 的检验统计量
  - `Pr(>|t|)`: 上述检验对应的 $p-$ 值

如果想要提取这个板块内的信息, 可以使用

```
fit$coefficients
```

这里得到的是一个系数向量, 如果你还想取出某一项的值, 就需要用类似于 `fit$coefficients[1]` 这种命令.

如果要得到参数的区间估计 (单独的某个参数, 如 $\beta_j$ 的, 其公式为 $\hat{\beta}_j \pm t_{1-\alpha/2, n-(k+1)} \widehat{\mathrm{SE}}\left(\hat{\beta}_j\right)$), 则用

```
confint(fit, level = 0.95)
```

其中 `level` 表示置信水平.

此外我们还可以考虑如下的一些特殊的置信区间:

- $\beta=(\beta_0,\beta_1,\cdots,\beta_n)$ 的置信域 (常称为置信椭球), 但一般比较难表示出来.
- $a'\beta=a_0\beta_0+\cdots+a_n\beta_n$ 的置信区间:

$$\left[\mathbf{a}'\hat{\beta}-\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}}\hat{\sigma}t_{n-p}(\alpha/2),\ \mathbf{a}'\hat{\beta}+\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}}\hat{\sigma}t_{n-p}(\alpha/2)\right]$$

  - 利用命令 `confint()` 求出的是这个公式的一种特殊的情形: $\mathbf{a}=(0,\cdots,0,1,0,\cdots,0)$.
- 多重比较下的置信区间:
  - Bonferroni: $\mathbf{a}_i'\hat{\beta},\ i=1,\cdots,k$ (有限个)

$$\left[\mathbf{a}_i'\hat{\beta}-\sqrt{\mathbf{a}_i'(X'X)^{-1}\mathbf{a}_i}\hat{\sigma}t_{n-p}(\frac{\alpha}{2k}),\ \mathbf{a}_i'\hat{\beta}+\sqrt{\mathbf{a}_i'(X'X)^{-1}\mathbf{a}_i}\hat{\sigma}t_{n-p}(\frac{\alpha}{2k})\right]$$

  - Scheffe: 设 $A_{d\times p}$ 行满秩, $p$ 为参数个数, $\mathbf{a}\in\mathcal{M}(A')$. (无限个)

$$\left[\mathbf{a}'\hat{\beta}-\hat{\sigma}\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}dF_{d,n-p}(1-\alpha)},\ \mathbf{a}'\hat{\beta}+\hat{\sigma}\sqrt{\mathbf{a}'(X'X)^{-1}\mathbf{a}dF_{d,n-p}(1-\alpha)}\right]$$

    * 我们讲座课课件上写的是 $\sqrt{2F_{2,n-2}(1-\alpha)}$, 这是因为这里只有两个参数, 即 $p=2$, 并且取 $d=p$.

大家可以根据实际情况进行按照上式进行编程计算 (本门课程应该用不上, 但《试验设计》, 《回归分析》等课程会涉及到.)

**例 1** 16 名婴儿的收缩压 ($y$) 与体重 ($x_1$) 以及年龄 ($x_2$) 的相关数据如下所示

```
SBP <- c(89,90,83,77,92,98,82,85,96,95,80,79,86,97,92,88)
weight <- c(135,120,100,105,130,125,125,105,120,90,120,95,120,150,160,125)
age <- c(3,4,3,2,4,5,2,3,5,4,2,3,3,4,3,3)
```

拟合回归曲线 $y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon$ 并获得 $\beta$ 的估计.

**解答:**

```
fit <- lm(SBP ~ weight + age)
summary(fit)
```

```
##
## Call:
## lm(formula = SBP ~ weight + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0438 -1.3481 -0.2395  0.9688  6.6964
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.45019    4.53189  11.794 2.57e-08 ***
## weight       0.12558    0.03434   3.657   0.0029 **
## age          5.88772    0.68021   8.656 9.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.479 on 13 degrees of freedom
## Multiple R-squared:  0.8809, Adjusted R-squared:  0.8626
## F-statistic: 48.08 on 2 and 13 DF,  p-value: 9.844e-07
```

```
fit$coefficients # 回归系数的点估计
```

```
## (Intercept)      weight         age
##  53.4501940   0.1255833   5.8877191
```

```
confint(fit) # 回归系数的区间估计
```

```
##                    2.5 %     97.5 %
## (Intercept) 43.65964398 63.2407441
## weight       0.05140441  0.1997621
## age          4.41822526  7.3572130
```

### 5.1.2 假设检验

回归分析中最 general 的假设检验框架为线性假设:

$$H_0 : A\beta = c \quad \longleftrightarrow \quad H_1 : A\beta \neq c$$

其中我们假设有 $n$ 个样本, $p$ 个自变量, $A_{q\times p}$ 为对 $\beta$ 的假设数目. 其检验统计量为

$$F = \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p)} \overset{H_0}{\sim} F_{q,n-p}$$

具体理论以及检验所用代码本课程不做过多说明. 我们关注的是它的特例:

- 检验回归方程的显著性: $H_0 : A\beta = 0$. 我们其实最关注的是 $A_{1\times p} = \mathbf{1}_p$ 的情形, 即 $H_0 : \beta_0 = \beta_1 = \cdots = \beta_{p-1}$. 检验统计量为

  $$F = \frac{\mathrm{SSR}/(p-1)}{\mathrm{RSS}/(n-p)} \overset{H_0}{\sim} F_{p-1,n-p}$$

  在 `R` 中我们看 `summary(fit)` 最后一行:

  - `F-statistic`: 检验统计量
  - `p-value`: p 值

  在一元情形下也可以用讲义中提到的 `anova(fit)` 函数, 读出来的 $F$ 值一样, 但是, 在多元情形下, 由于有多个 treatment, `anova` 只能列出每个 Treatment 各自的平方和, 得不到 SSR(所有 treatment 加在一起的), 所以这个函数无效 (但是如果是多因素双水平的试验, 两者还是等价的, 但算出来的值会不一样).

- 检验某个系数的显著性: $H_0 : \beta_j = 0$, 此时我们可以用 (从上面的统计量导出的)

  $$F = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2 d_{jj}} \overset{H_0}{\sim} F_{1,n-p}$$

  其中 $d_{jj}$ 是 $(X'X)^{-1}$ 的第 $j$ 个对角元. 这个检验统计量开根号就得到我们讲义里面用的

  $$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}} \overset{H_0}{\sim} t_{n-p}$$

这个检验没有现成的函数可用, 我们自己手写:

```
X = data.matrix(你的数据)
n <- 观测个数
p <- 参数个数
D = ginv(t(X)%*%X)
D_d = diag(D)
F = rep(0,p)
result=rep(0,p)
fit <- lm()
beta = fit$coefficients
for (i in 1:p) {
F[i] = beta[i]^2/(fit$sigma^2*D_d[i])
  if (F[i]>qf(1-alpha,1,n-p)){
  result[i] = 1
  }
}
result
```

此程序仅供参考, 最快捷的方式还是自己在 `Coefficient` 板块读出检验统计量 (上面的程序绕了一圈是因为程序不能直接从 `lm()` 命令中读取估计值的标准差).

**例 2** (接上例) 对整个回归方程以及每个回归系数进行假设检验

**解答:**

```
summary(fit)
```

```
##
## Call:
## lm(formula = SBP ~ weight + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0438 -1.3481 -0.2395  0.9688  6.6964
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.45019    4.53189  11.794 2.57e-08 ***
## weight       0.12558    0.03434   3.657   0.0029 **
## age          5.88772    0.68021   8.656 9.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.479 on 13 degrees of freedom
```

```
## Multiple R-squared:  0.8809, Adjusted R-squared:  0.8626
## F-statistic: 48.08 on 2 and 13 DF,  p-value: 9.844e-07
```

我们从最后一行可以看到 $F = 48.08$, $p$ 值为 9.844e-07$\ll 0.05$, 因此整个回归方程是显著的.

此外可以从 `Coefficient` 板块看到所有系数都是显著的 (具体描述略).

注意: `anova(fit)` 的输出其实并不能得到讲座课讲义里面的方差分析表 (我们需要的是 SSR, 但 `anova` 命令把它拆分为两个 treatment 的 SS 了).

```
library(knitr)  # 这里导入 knitr 包以及下面的 knitr::kable 只是为了能够显示在 pdf 里, 非必须
```

```
## Warning: package 'knitr' was built under R version 4.0.5
```

```
knitr::kable(anova(fit))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| weight | 1 | 130.53750 | 130.537500 | 21.23840 | 0.0004901 |
| age | 1 | 460.49814 | 460.498140 | 74.92286 | 0.0000009 |
| Residuals | 13 | 79.90186 | 6.146297 | NA | NA |

### 5.1.3 模型诊断

我们主要是对残差进行诊断分析. 取出残差 $y_i - \hat{y}_i$ 的命令为

```
res <- fit$residuals
```

或者是

```
res <- residuals(fit)
```

如果想得到标准化的残差, 则用

```
std.res <- rstandard(fit)
```

#### 5.1.3.1 正态性假设

方法 1: QQ 图

```
qqnorm(res)
qqline(res)
```

看数据点是否都在一条直线附近.

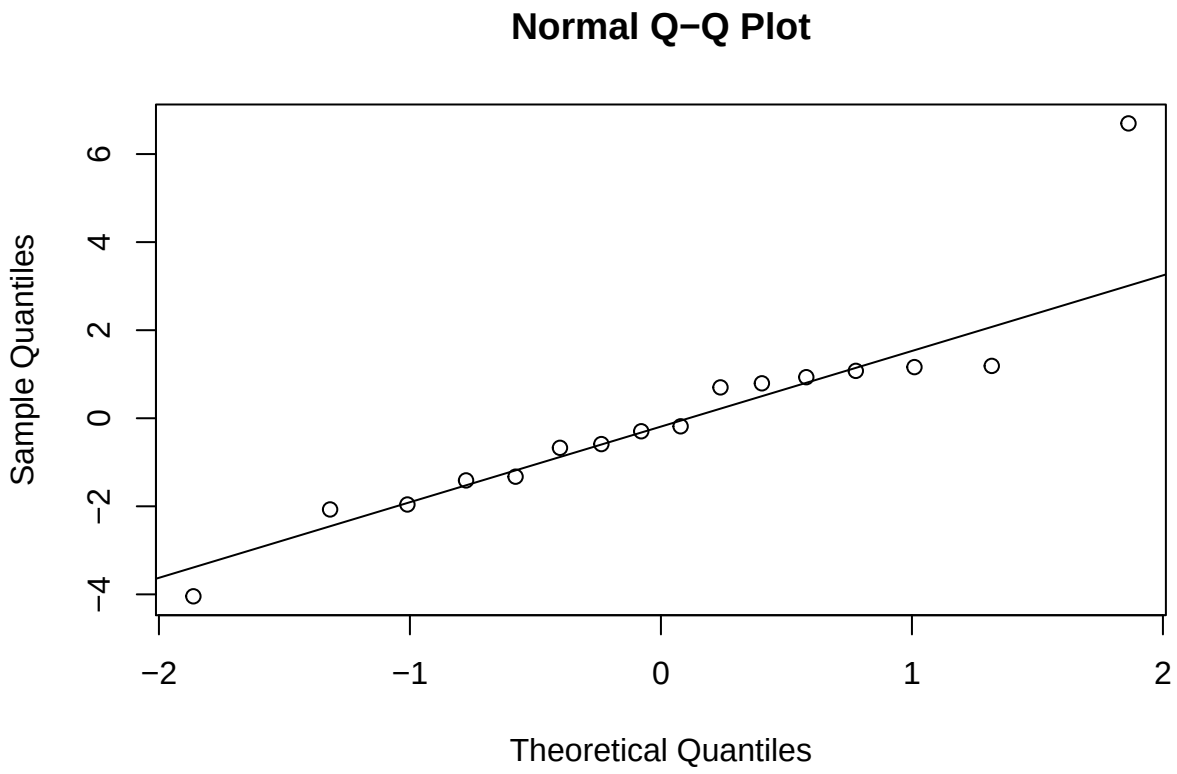
方法 2: Shapiro-Wilks 检验, 其原假设为服从正态分布 (所以我们希望 $p > 0.05$).

```
shapiro.test(res)
```

**例 3** (接例 1) 检查残差的正态性

**解答:**

```
res <- fit$residuals
qqnorm(res)
qqline(res)
```




```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.86728, p-value = 0.02472
```

The points are scattered roughly alongside the straight line, except for very few points, which means there is no severe indication of non-normality. However, the shapiro test suggest that the normality assumption is violated since $p < 0.05$.

#### 5.1.3.2 拟合效果以及同方差性 (or 方差为常数)

方法: 画拟合值 vs 残差图

```
res <- fit$residuals # 取残差
fitted <- fit$fitted.values # 取拟合值
```

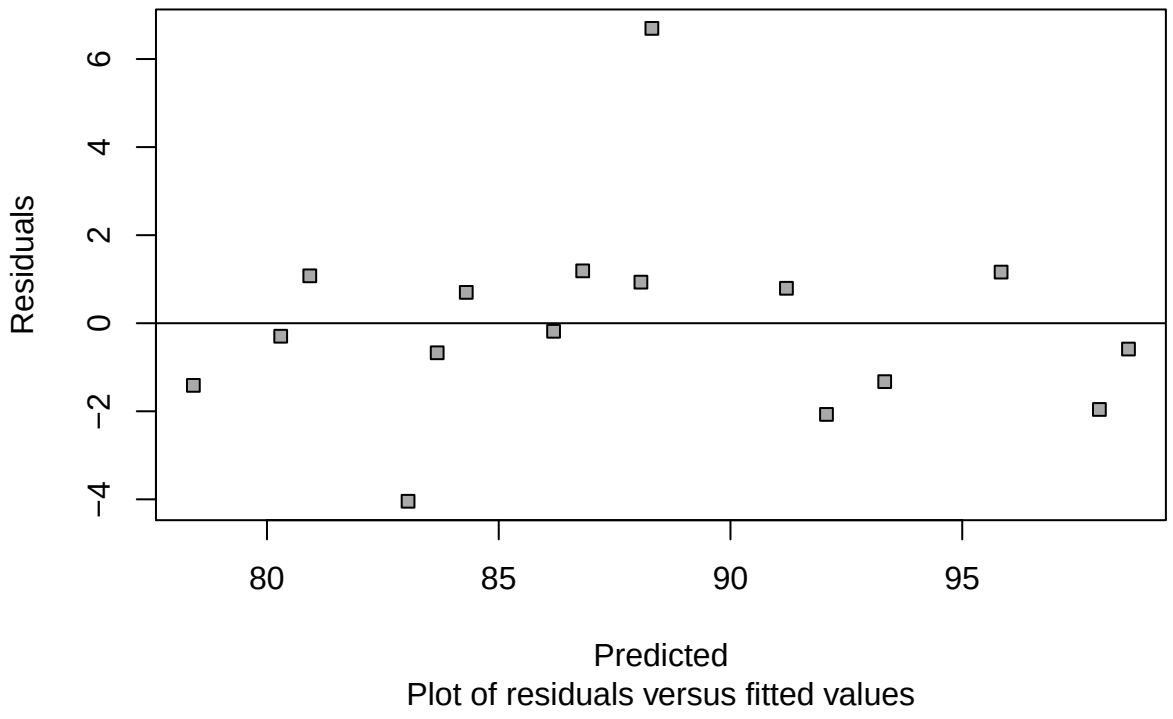
```
plot(fitted, res,
     pch = 22, bg = "darkgrey",
     xlab = "Predicted", ylab = "Residuals",
     sub="Plot of residuals versus fitted values")
abline(h = 0)
```

如果没有发现残差的异常结构则说明模型基本正确, 且同方差性没有被违背.

**例 4** (接例 1) 检查拟合效果以及同方差性

**解答:**

```
res <- fit$residuals
fitted <- fit$fitted.values # 取拟合值
plot(fitted, res,
     pch = 22, bg = "darkgrey",
     xlab = "Predicted", ylab = "Residuals",
     sub="Plot of residuals versus fitted values")
abline(h = 0)
```




There is nothing unusual about the residuals. The plot of residuals versus fitted values does not reveal any obvious pattern. Different levels of the variances of residuals are most the same. There are not any violations of independence or constant variance assumptions.

### 5.1.4 预测

两种预测:

- 预测因变量的均值 $\hat{\mu}_{y_{n+1}|x_{n+1}}$
- 预测单个因变量 $\hat{y}_{n+1}$

这两种预测得出来的点估计是一样的, 但区间估计会有所差别

- $\mu_{y_{n+1}|x_{n+1}}$ 的一个 $100(1-\alpha)\%$ 置信区间为 $\hat{\mu}_{y_{n+1}|x_{n+1}} \pm t_{1-\alpha/2,n-2}\widehat{\mathrm{SE}}(\hat{\mu}_{y_{n+1}|x_{n+1}})$.

```
pre <- data.frame(变量名 =c(你想要的值))
predict(fit,newdata=pre,interval="confidence",level=0.95)
```

- $y_{n+1}$ 的一个 $100(1-\alpha)\%$ 置信区间 (也称为**预测区间**) 为 $\hat{y}_{n+1} \pm t_{1-\alpha/2,n-2}\widehat{\mathrm{SE}}(\hat{y}_{n+1})$.

```
pre <- data.frame(变量名 =c(你想要的值))
predict(fit,newdata=pre,interval="prediction",level=0.95)
```

**例 5** (接例 1) 求出所有满足年龄为 4, 体重为 122 的儿童的平均 SBP 预测值. 如果已知的是某一个儿童年龄为 4, 体重为 122, 其 SBP 预测值又是多少? 试比较这两组结果.

**解答:**

```
pre <- data.frame(age=c(4), weight=c(122))
predict(fit,newdata=pre,interval="confidence",level=0.95)
```

```
##        fit      lwr     upr
## 1 92.32223 90.64826 93.9962
```

```
predict(fit,newdata=pre,interval="prediction",level=0.95)
```

```
##        fit     lwr      upr
## 1 92.32223 86.7108 97.93366
```

## 5.2 相关系数

### 5.2.1 单样本假设检验

假设 1:

$$H_0 : \rho_{xy} = 0 \quad \longleftrightarrow \quad H_1 : \rho_{xy} \neq 0$$

检验统计量为

$$t = r_{xy}\frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \overset{H_0}{\sim} t_{n-2}$$

如果已知数据, 可以使用

```
cor.test(x,y,alternative = c("two.sided", "less", "greater")
```

其中 x, y 就是跟我们本章所指的 $x, y$ 一致. 此外, 通过这个函数还可以获得相关系数的

- 点估计

```
cor.test(x,y)$estimate
```

- 区间估计

```
cor.test(x,y)$conf.int
```

如果是只给了相关系数和样本量, 则只能手动计算.

**例 6** 用力呼气量 (FEV) 常用于衡量肺功能, 推测 FEV 与身高有关系, 现收集了 10-15 岁的男孩的身高以及他们的 FEV 数据如下所示:

```
height <- seq(134,178,by=4)
FEV <- c(1.7,1.9,2.0,2.1,2.2,2.5,2.7,3.0,3.1,3.4,3.8,3.9)
```

检验假设 $H_0 : \rho_{xy} = 0 \quad \longleftrightarrow \quad H_1 : \rho_{xy} \neq 0$

**解答:**

```
cor.test(height,FEV)
```

```
##
##  Pearson's product-moment correlation
##
## data:  height and FEV
## t = 20.366, df = 10, p-value = 1.797e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9569489 0.9967804
## sample estimates:
##      cor
## 0.988159
```

假设 2:

$$H_0 : \rho_{xy} = \rho_0 \quad \longleftrightarrow \quad H_1 : \rho_{xy} \neq \rho_0 \quad \text{其中} \rho_0 \neq 0$$

检验统计量为:

$$\lambda = (z_{xy} - z_0)\sqrt{n-3} \overset{H_0}{\to} N(0,1) \quad \text{其中} z_{xy} = \frac{1}{2}\log\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$$

使用的代码需要自己手写 (这里用的是双边检验, 需要单边检验的话可自行修改代码):

```
cor.test.ztrans <- function(x,y,alpha=0.05,rho_0){
  r<-cor(y,x)
```

```
  n<-length(x)
  z<-(1/2)*log((1+r)/(1-r))
  z_rho_0 <-(1/2)*log((1+rho_0)/(1-rho_0))
  lambda=(z-z_rho_0)*sqrt(n-3)
  p_value <- 2*(1-pnorm(abs(lambda)))
  conf_z <- z + c(-1,1)*qnorm(1-alpha/2)/sqrt(n-3)   # 先得到 z 的置信区间
  conf_rho <- (exp(2*conf_z)-1)/(exp(2*conf_z)+1) # 再逆变换得到相关系数的置信区间

  return(list(test_statistic=lambda,p_value=p_value,confidence_interval=conf_rho))
}
```

**例 7** (接上例) 检验假设 $H_0 : \rho_{xy} = 0.98 \quad \longleftrightarrow \quad H_1 : \rho_{xy} \neq 0.98$

**解答:**

```
cor.test.ztrans(height,FEV,rho_0=0.98)
```

```
## $test_statistic
## [1] 0.792416
##
## $p_value
## [1] 0.4281182
##
## $confidence_interval
## [1] 0.9569489 0.9967804
```

### 5.2.2 两样本假设检验

假设

$$H_0 : \rho_1 = \rho_2 \quad \text{vs} \quad H_a : \rho_1 \neq \rho_2.$$

检验统计量为

$$\lambda = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \overset{H_0}{\sim} N(0,1)$$

当 $|\lambda| > z_{1-\alpha/2}$ 时拒绝 $H_0$, $p$ 值 $= 2 \times P(Z > \lambda_{\text{computed}})$.

这部分也没有代码, 需要自己手写:

```
cor.test.ztrans.twosample <- function(x1,y1,x2,y2){
  r1 <- cor(x1,y1); r2 <- cor(x2,y2)
  n1 <-length(x1); n2 <-length(x2)
  z1<-(1/2)*log((1+r1)/(1-r1)); z2<-(1/2)*log((1+r2)/(1-r2))
  lambda=(z1-z2)/sqrt(1/(n1-3)+1/(n2-3))
  p_value <- 2*(1-pnorm(abs(lambda)))
```

```
  return(list(test_statistic=lambda,p_value=p_value))
}
```

**例 8** 假设现在有两组儿童, 其中一组与他们的生父母住在一起, 另一组与养父母住在一起. 研究的问题是母亲的血压与孩子的血压的相关性在两个组中是否相同. 倘若不一样, 那就说明是遗传因素在起作用. 假设第一组中共有 1000 对母子, 相关系数为 0.35; 第二组中有 100 组, 相关系数为 0.06. 通过这些数据能得出什么结论?

**解答:**

这里已经给好 $r_1$ 与 $r_2$ 以及样本量了, 所以就用不了我们刚写的函数 (只能拆解出来用)

```
r1 <- 0.35; r2 <- 0.06
n1 <- 1000; n2 <- 100
z1<-(1/2)*log((1+r1)/(1-r1)); z2<-(1/2)*log((1+r2)/(1-r2))
(lambda=(z1-z2)/sqrt(1/(n1-3)+1/(n2-3)))
```

```
## [1] 2.871134
```

```
(p_value <- 2*(1-pnorm(abs(lambda))))
```

```
## [1] 0.004090024
```

### 5.2.3 偏相关系数

命令

```
library(ggm)
pcor(u,s)
```

输入的参数为

- u: 向量, 前两个位置表示要计算相关系数的变量名 (或下标), 后面的位置为要控制的条件变量名 (或下标).
- s: cov(你的数据)

注意: 这里是一整个数据, 即 $(x, y)$ 是放在同一个数据里的, 而非分开.

**例 9** (接例 1) 求 $R^2_{y\cdot x_1,\ x_2}$ 以及求 $R^2_{y\cdot x_2,\ x_1}$

**解答:**

```
library(ggm)
```

```
## Warning: package 'ggm' was built under R version 4.0.5
```

```
egdata <- cbind(SBP,weight,age)
pcor(c(1,2,3),cov(egdata)) # 求 y 关于 x1
```

```
## [1] 0.7121422
```

```
pcor(c(1,3,2),cov(egdata)) # 求 y 关于 x2
```

```
## [1] 0.923116
```

# 6 假设检验: 多样本推断 (方差分析)

## 6.1 单因素方差分析

### 6.1.1 固定效应模型

模型为:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \begin{cases} i = 1, \cdots, k \\ j = 1, \cdots, n_i \end{cases}$$

其中

- $\mu$ 为所有总体 (处理水平) 的共同参数, 称为总均值 (未知常数).
- $\tau_i$ 为第 $i$ 个总体 (处理水平) 的效应 (未知常数).
  - $k$ 个处理可以由实验者具体选定, 此时所得结论仅适用于该分析中所考虑的 $k$ 个因子水平, 而不能推广到未曾明确考虑的相似的因子水平中去 (因此称为固定效应模型)
- $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ 为误差项.

#### 6.1.1.1 假设检验

我们要检验的假设为

$$H_0 : \tau_1 = \cdots = \tau_k = 0 \quad \longleftrightarrow \quad H_1 : \exists\ i,\ \tau_i \neq 0$$

检验统计量为

$$F = \frac{\mathrm{SS}_{Trt}/(k-1)}{\mathrm{SS}_E/(n-k)}$$

注: 我们讲义上也把处理的平方和 ($\mathrm{SS}_{Trt}$) 称为组间平方和 ($\mathrm{SS}_B$), 把误差平方和 ($\mathrm{SS}_E$) 称为组内平方和 ($\mathrm{SS}_W$).

方差分析的核心内容在于如下的**方差分析表**

| 差异来源 | SS | $df$ | MS | $F$ 统计量 | $p$ 值 |
|---|---|---|---|---|---|
| 组间 (处理水平) | SSB | $k-1$ | $s_B^2 = \dfrac{\mathrm{SSB}}{k-1}$ | $F = \dfrac{s_B^2}{s_W^2}$ | $\mathbb{P}(F_{k-1,n-k} > F)$ |
| 组内 (误差) | SSW | $n-k$ | $s_W^2 = \dfrac{\mathrm{SSW}}{n-k}$ | | |
| 总和 | TSS | $n-1$ | | | |

使用如下命令可以输出 ANOVA 表

```
data.aov <- aov(观测~factor(水平))
summary(data.aov)
```

其中放在括号里的内容称为 `formula`, 根据需要会有多种不同的结构, 我们会在多因素方差分析中详细讲解.

注意: `~` 后面放的是 factor, 不然默认为连续变量, 会使得自由度出错. 如果使用 `factor()` 函数来进行转化的话, 一定要注意转化前后因子次序有无改变 (看看是否会对应出错).

**例 1** 22 名年轻的哮喘病患者被选中以研究不同条件下暴露于二氧化硫 ($SO_2$) 的短期效应. 下面给出的是筛选时根据肺功能 (定义为 FEV/FVC) 进行分组的不同患者对 $SO_2$ 的支气管反应数据 (单位: cm $H_2O/s$):

- FEV/FVC $\leqslant 74\%$ (A 组): 20.8, 4.1, 30.0, 24.7, 13.8,
- FEV/FVC $75-84\%$ (B 组): 7.5, 7.5, 11.9, 4.5, 3.1, 8.0, 4.7, 28.1, 10.3, 10.0, 5.1, 2.2
- FEV/FVC $\geqslant 85$ (C 组): 9.2, 2.0, 2.5, 6.1, 7.5

上述数据整理到 R 中为:

```
BR<-c(20.8, 4.1, 30.0, 24.7, 13.8, 7.5, 7.5,
      11.9, 4.5, 3.1, 8.0, 4.7, 28.1, 10.3, 10.0,
      5.1, 2.2, 9.2, 2.0, 2.5, 6.1, 7.5)
LF<-c(rep("A",5),rep("B",12),rep("C",5))
```

检验假设: 三个肺功能组的患者的支气管反应数据均值存在差异.

**解答:**

```
data.aov <- aov(BR ~ factor(LF))
summary(data.aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## factor(LF)   2  503.5  251.77   4.989 0.0181 *
## Residuals   19  958.8   50.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#### 6.1.1.2 区间估计

我们接下来考虑的是均值的置信区间, 我们从简单开始 (为了代码表示方便, 我们下面把观测向量记为 Y, 处理各个水平组成的向量表示为 A):

- $\mu_i$ 的 $100(1-\alpha)\%$ 置信区间:

$$\overline{y}_{i\cdot} \pm t_{1-\alpha/2}(n-k)\sqrt{\frac{\mathrm{MS}_E}{n_i}}$$

代码 (注意自己调整里面的 n,k,i,ni):

```
alpha = 0.05
upper.bound = (aggregate(Y ~ factor(A), mean)$Y[i]
               + qt(1 - alpha / 2, n - k) * sqrt(MSE / ni))
lower.bound = (aggregate(Y ~ factor(A), mean)$Y[i]
               - qt(1 - alpha / 2, n - k) * sqrt(MSE / ni))
round(lower.bound, 2);round(upper.bound, 2)
```

- $\mu_i - \mu_j$ 的 $100(1-\alpha)\%$ 置信区间:

$$\overline{y}_{i\cdot} - \overline{y}_{j\cdot} \pm t_{1-\alpha/2}(n-k)\sqrt{\mathrm{MS}_E \times (\frac{1}{n_i} + \frac{1}{n_j})}$$

```
alpha = 0.05
upper.bound = (aggregate(Y ~ factor(A), mean)$Y[i]
                - aggregate(Y ~ factor(A), mean)$Y[j]
                + qt(1 - alpha / 2, n - k) * sqrt(MSE * (1/n1 + 1/n2)))
lower.bound = (aggregate(Y ~ factor(A), mean)$Y[i]
                - aggregate(Y ~ factor(A), mean)$Y[j]
                - qt(1 - alpha / 2, n - k) * sqrt(MSE * (1/n1 + 1/n2)))
round(lower.bound, 2);round(upper.bound, 2)
```

#### 6.1.1.3 对照

我们考虑一个特殊的假设检验: 对照

$$H_0 : \sum_{i=1}^{k} c_i\mu_i = 0 \quad \longleftrightarrow \quad H_1 : \sum_{i=1}^{k} c_i\mu_i \neq 0$$

检验统计量为

$$F_0 = \frac{\left(\sum_{i=1}^{a} c_i\bar{y}_i\right)^2}{\frac{MS_E}{n}\sum_{i=1}^{a} c_i^2} = \frac{\mathrm{SS}_C/1}{\mathrm{MS}_E} \overset{H_0}{\sim} F_{1,n-k}$$

补充: 为什么要定义对照? 因为我们无法满足于 $\mu_i - \mu_j = 0$ 这样简单的假设, 有时候我们会需要到 $(\mu_1 - \mu_2) - (\mu_3 - \mu_4)$ 这种复杂的假设. 什么时候会出现? 处理出现了交叉, 只能用减法分离出单独的处理效应.

代码:

```
library(multcomp)
A <- factor(A)
data.aov = aov(Y~A)
c.1 <- c(1,-1,1,-1)
mc <- glht(data.aov, linfct = mcp(A = c.1))
summary(mc)   # 检验结果
confint(mc, level = 0.95) # 置信区间
```

**例 2** 病人在服用治疗高血压的药物后的收缩压下降量是衡量其对于这种药物的反应的关键指标. 在治疗中, 药物的副作用也备受关注. 某项研究要评估两种药物 A, B 对于减轻高血压标准用药 S 的副作用的效果. A 或 B 要与 S 一起服用. 这项研究采取了完全随机设计, 总共有 5 个处理组, 如下表所示.

| 处理组 | 处方 |
|---|---|
| 1 | 标准 (S) |
| 2 | S 与低剂量 A (S+AL) |
| 3 | S 与高剂量 A (S+AH) |
| 4 | S 与低剂量 B (S+BL) |
| 5 | S 与高剂量 B (S+BH) |

每个处理组都重复进行 4 次实验. 用药四周后, 患者的收缩压下降量 (单位: mmHg) 如下表所示.

| 处理组 | 下降量 (mmHg) | 均值 ($\overline{y}_{i\cdot}$) |
|---|---|---|
| 1 | 27, 26, 21, 26 | 25.00 |
| 2 | 19, 13, 15, 16 | 15.75 |
| 3 | 15, 10, 10, 11 | 11.50 |
| 4 | 22, 15, 21, 18 | 19.00 |
| 5 | 20, 18, 17, 16 | 17.75 |

检验三种对照:

- 低剂量 A 与高剂量 A 带来的效应有差别吗?
- 低剂量 B 与高剂量 B 带来的效应有差别吗?
- 两种剂量的 A 带来效应的均值与两种剂量的 B 带来效应的均值有差别吗?

**解答:**

我们只做第一种对照, 剩下两种同理

```
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 4.0.5
```

```
BP<-c(27, 26, 21, 26,
      19, 13, 15, 16,
      15, 10, 10, 11,
      22, 15, 21, 18,
      20, 18, 17, 16)
Trt <- rep(c("S","AL","AH","BL","BH"),each=4)
Trt <- factor(Trt, levels = c("S","AL","AH","BL","BH"))
data.aov <- aov(BP ~ Trt)
contr <- rbind("AL-AH"=c(0,1,-1,0,0))
mc <- glht(data.aov,linfct=mcp(Trt=contr))
summary(mc,test=adjusted(type=c("none")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
```

```
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = BP ~ Trt)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## AL-AH == 0    4.250      1.794    2.37   0.0316 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

#### 6.1.1.4 多重比较

我们在方差分析 F 检验法下拒绝了等处理均值的原假设之后想要对一切 $i \neq j$ 检验到底是哪对不一样.

- *Tukey 法

  检验方法: $\hat{\ell}_i$ 是显著的, 若

  $$|\hat{\ell}_i| \geq q_\alpha(k, df_{MSE})\sqrt{\hat{V}\left(\hat{\ell}_i\right)}$$

  置信区间

  $$\hat{\ell}_i \pm q_\alpha(k, df_{MSE})\sqrt{\hat{V}\left(\hat{\ell}_i\right)}$$

```r
# 方法 1
TukeyHSD(data.aov)

# 方法 2
confint(glht(data.aov, linfct = mcp(A = "Tukey")), level = 0.95)
```

- Fisher LSD 法

  检验方法: $\hat{\ell}_i$ 是显著的, 若

  $$|\hat{\ell}_i| \geq t_{1-\alpha/2, n-k}\sqrt{\hat{V}\left(\hat{\ell}_i\right)}$$

  置信区间

  $$\hat{\ell}_i \pm t_{1-\alpha/2, n-k}\sqrt{\hat{V}\left(\hat{\ell}_i\right)}$$

```r
# 方法 1(只能处理简单的两均值差)
library(agricolae)
LSD.test(data.aov, A, alpha = 0.05,
         DFerror = n - k, MSerror = MSE,
         group = FALSE, console = TRUE)

# 方法 2(自己填 df_SSE)
```

```
contr <- rbind(c.1,c.2,...)
mc <- glht(data.aov, linfct = mcp(A = contr))
summary(mc, test=adjusted(type=c("none"))) # 默认就是 Fisher
confint(mc, calpha=qt(1-0.05/2,df_SSE))
```

- Bonferroni 法

  检验方法: $\hat{\ell}_i$ 是显著的, 若

$$|\hat{\ell}_i| \geq t_{1-\alpha/(2m),n-k}\sqrt{\hat{V}\left(\hat{\ell}_i\right)}$$

  置信区间

$$\hat{\ell}_i \pm t_{1-\alpha/(2m),n-k}\sqrt{\hat{V}\left(\hat{\ell}_i\right)}$$

```
contr <- rbind(c.1,c.2,...)
mc <- glht(data.aov, linfct = mcp(A = contr))
summary(mc, test = adjusted("bonferroni"))
confint(mc, calpha=qt(1-0.05/(2*m),df_SSE)) # m 为参与比较的数量
```

- Scheffe 法

  检验方法: $\hat{\ell}_i$ 是显著的, 若

$$|\hat{\ell}_i| \geq \sqrt{(k-1)F_{1-\alpha,k-1,n-k}}\sqrt{\hat{V}\left(\widehat{\ell_i}\right)}$$

  置信区间

$$\widehat{\ell_i} \pm \sqrt{(k-1)F_{1-\alpha,k-1,n-k}}\sqrt{\hat{V}\left(\widehat{\ell_i}\right)}$$

  代码不完备, 大家可以根据自己需要来选用或者是自行编写

```
# 方法 1 (只能处理简单的两均值差)
library(agricolae)
scheffe.test(data.aov, A, alpha = 0.05,
         DFerror = n - k, MSerror = MSE,
         group = FALSE, console = TRUE)

# 方法 2 (只能处理简单的两均值差)
library(DescTools)
ScheffeTest(data.aov)

# 方法 3 (能处理复杂的对照, 但求不了置信区间)
mc <- glht(data.aov, linfct = mcp(A = contr))
pf((summary(mc)$test$tstat)^2 / 2, k-1, n-k, lower.tail = FALSE)
```

**例 3** (接上例) 对三种对照同时进行检验, 求置信区间

**解答:**

```r
contr <- rbind("AL-AH"=c(0,1,-1,0,0),
               "BL-BH"=c(0,0,0,1,-1),
               "A-B"=c(0,1/2,1/2,-1/2,-1/2))
mc <- glht(data.aov,linfct=mcp(Trt=contr))

# Tukey
TukeyHSD(data.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BP ~ Trt)
##
## $Trt
##         diff         lwr        upr     p adj
## AL-S   -9.25 -14.7882136 -3.7117864 0.0009464
## AH-S  -13.50 -19.0382136 -7.9617864 0.0000156
## BL-S   -6.00 -11.5382136 -0.4617864 0.0307779
## BH-S   -7.25 -12.7882136 -1.7117864 0.0080172
## AH-AL  -4.25  -9.7882136  1.2882136 0.1773046
## BL-AL   3.25  -2.2882136  8.7882136 0.4026253
## BH-AL   2.00  -3.5382136  7.5382136 0.7962763
## BL-AH   7.50   1.9617864 13.0382136 0.0061171
## BH-AH   6.25   0.7117864 11.7882136 0.0235783
## BH-BL  -1.25  -6.7882136  4.2882136 0.9540324
```

```r
# Fisher
summary(mc,test=adjusted(type=c("none")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = BP ~ Trt)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## AL-AH == 0    4.250      1.794   2.370  0.03164 *
## BL-BH == 0    1.250      1.794   0.697  0.49649
## A-B == 0     -4.750      1.268  -3.745  0.00195 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

```
confint(mc, calpha=qt(1-0.05/2,15))
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = BP ~ Trt)
##
## Quantile = 2.1314
## 95% confidence level
##
##
## Linear Hypotheses:
##            Estimate lwr     upr
## AL-AH == 0  4.2500   0.4272  8.0728
## BL-BH == 0  1.2500  -2.5728  5.0728
## A-B == 0   -4.7500  -7.4531 -2.0469
```

```
# Bonferroni
summary(mc, test = adjusted("bonferroni"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = BP ~ Trt)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## AL-AH == 0    4.250      1.794   2.370  0.09493 .
## BL-BH == 0    1.250      1.794   0.697  1.00000
## A-B == 0     -4.750      1.268  -3.745  0.00585 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- bonferroni method)
```

```r
confint(mc, calpha=qt(1-0.05/(2*3),15))
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = BP ~ Trt)
##
## Quantile = 2.6937
## 95% confidence level
##
##
## Linear Hypotheses:
##             Estimate lwr     upr
## AL-AH == 0  4.2500  -0.5812  9.0812
## BL-BH == 0  1.2500  -3.5812  6.0812
## A-B == 0   -4.7500  -8.1662 -1.3338
```

```r
# Scheffe
pf((summary(mc)$test$tstat)^2 / 2, 4, 15, lower.tail = FALSE)
```

```
##       AL-AH       BL-BH         A-B
## 0.063682952 0.909522514 0.002164035
```

```r
## 这里只求第一个对照的, 剩下的同理
## 4.25 与 1.794 可以在其他方法输出的 Estimate 以及 Std. Error 中读出来
4.25+c(-1,1)*1.794*sqrt(4*qf(0.95,4,15))
```

```
## [1] -2.02189 10.52189
```

### 6.1.2 随机效应模型

模型为:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \begin{cases} i = 1, \cdots, k \\ j = 1, \cdots, n_i \end{cases}$$

其中

- $\mu$ 为所有总体 (处理水平) 的共同参数, 称为总均值 (未知常数).
- $\tau_i$ 为第 $i$ 个总体 (处理水平) 的效应 (随机变量).
  - $k$ 个处理可以看作是来自一个较大总体的一个随机样本. 在这种情况下, 能够把所得结论推广到总体的所有处理中去. 这里 $\tau_i$ 是随机变量, 假设其服从 $N(0, \sigma_\tau^2)$(因此称为随机效应模型).
- $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ 为误差项.

#### 6.1.2.1 假设检验

注意到随机效应模型与固定效应模型的检验统计量完全一致, 可以用上面提到的 `aov()` 函数的方法来做.

#### 6.1.2.2 参数估计

相比之前, 这里最关键是求 $\sigma_\tau^2$, 这可以使用 `lme4` 包中的函数 `lmer()` 函数可以方便地求得.

```
data.lme = lmer(Y~(1|A))
summary(data.lme)
```

**例 4** “护士健康研究” 是约有 100,000 名美国护士参加的一个大型的前瞻性研究项目. 从 1976 年起, 每两年这些护士就会收到关于她们健康状况的调查问卷. 其中一个项目是收集一小部分护士的血样, 用于研究血清中激素水平与乳腺癌的关系. 在研究的第一步中, 研究人员从 5 名绝经期的女性中获取血样, 每份血样都被等分成两份, 并用双盲的方式送到某实验室进行分析. 接下来的四个实验室也进行相同的操作. 这个研究的目的在于评估人与人之间的差异与每个人自己的差异占总差异的多少. 要在不同激素以及不同实验室之间作比较. 下面的表格为某实验室中化验的血浆雌二醇水平的重复性数据.

| | 重复 | | 重复之间的 | |
|---|---|---|---|---|
| 人 | 1 | 2 | 差异绝对值 | 均值 |
| 1 | 25.5 | 30.4 | 4.9 | 27.95 |
| 2 | 11.1 | 15.0 | 3.9 | 13.05 |
| 3 | 8.0 | 8.1 | 0.1 | 8.05 |
| 4 | 20.7 | 16.9 | 3.8 | 18.80 |
| 5 | 5.8 | 8.4 | 2.6 | 7.10 |

从表中数据估计出不同人之间的差异程度吗.

**解答:**

```
logPE <- log(c(25.5, 30.4, 11.1, 15.0, 8.0,
               8.1, 20.7, 16.9, 5.8, 8.4))
Group <- as.factor(rep(1:5, each=2))

library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.0.5
```

```
data.lme = lmer(logPE~(1|Group))
summary(data.lme)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: logPE ~ (1 | Group)
##
## REML criterion at convergence: 8.7
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.2321 -0.4461 -0.1258  0.6986  0.9061
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Group    (Intercept) 0.3172   0.5632
##  Residual             0.0300   0.1732
## Number of obs: 10, groups:  Group, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.5683     0.2578   9.964
```

我们也可以用讲义上的方法求出来对比一下

```
logPE <- log(c(25.5, 30.4, 11.1, 15.0, 8.0,
               8.1, 20.7, 16.9, 5.8, 8.4))
Group <- as.factor(rep(1:5, each=2))

model <- aov(logPE ~ Group)
a <- summary(model)
(sigma2 <- a[[1]]$"Mean Sq"[2])
```

```
## [1] 0.03000244
```

```
(sigmaA2 <- (a[[1]]$"Mean Sq"[1]-a[[1]]$"Mean Sq"[2])/2)
```

```
## [1] 0.3172171
```

## 6.2 多因素方差分析

多因素方差分析的具体内容会在《试验设计》课程中涉及到, 这里我们略去背景和理论, 主要阐释如何使用命令. 与单因素 ANOVA 命令相近, 最主要是 formula `Y~A` 稍有变动:

- `+` 添加新的因素, 如 `Y~A+B`
- `:` 交互项, 如 `A:B` 代表 A 与 B 的交互
- `*` 产生所有交互项, 如 `Y~A*B` 相当于 `Y~A+B+A:B`
- `^` 交互项的最高次数, 如 `Y~(A+B+C)^2` 相当于 `Y~A+B+C+A:B+B:C+A:C`

# 7 流行病研究中的设计与分析技术

## 7.1 分类数据的效应测度

### 7.1.1 危险度差

设 $p_1, p_2$ 分别为暴露及未暴露的被试者中患病的比例, 我们的任务是得到危险度差 (risk difference)$p_1 - p_2$ 的

- 点估计:

$$\hat{p}_1 - \hat{p}_2$$

- 区间估计:

$$\begin{cases} \hat{p}_1 - \hat{p}_2 - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}, & \text{if } \hat{p}_1 > \hat{p}_2 \\ \hat{p}_1 - \hat{p}_2 + \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}, & \text{if } \hat{p}_1 \leqslant \hat{p}_2 \end{cases}$$

代码:

```
library(fmsb)
riskdifference(a, b, N1, N0, CRC=FALSE, conf.level=0.95)
```

输入的参数为

- a: 暴露组中的患病人数
- b: 非暴露组中的患病人数
- N1: 暴露组中的总人数
- N0: 非暴露组中的总人数
- CRC: 是否要计算单个危险度的置信区间 (选 FALSE)
- conf.level: 置信水平

**例 1** 一项心血管疾病研究关注口服避孕药 (OC) 对 40-44 岁女性的心脏疾病的影响. 研究发现, 在 5000 名服用口服避孕药的女性中, 3 年内有 13 人患上了心肌梗塞 (MI); 而在 10,000 名未服用口服避孕药的女性中, 3 年内只有 7 人患上心肌梗塞. 求危险度差

**解答:**

```
library(fmsb)
```

```
## Warning: package 'fmsb' was built under R version 4.0.5
```

```
riskdifference(13, 7, 5000, 10000, conf.level=0.95)
```

```
##                  Cases People at risk         Risk
## Exposed    1.300000e+01   5.000000e+03 2.600000e-03
## Unexposed 7.000000e+00   1.000000e+04 7.000000e-04
## Total      2.000000e+01   1.500000e+04 1.333333e-03

##
##  Risk difference and its significance probability (H0: The difference
```

```
##  equals to zero)
##
## data:  13 7 5000 10000
## p-value = 0.01327
## 95 percent confidence interval:
##  0.0003963116 0.0034036884
## sample estimates:
## [1] 0.0019
```

### 7.1.2 危险度比

设 $p_1, p_2$ 分别为暴露及未暴露的被试者中患病的比例, 我们的任务是得到危险度差 (risk ratio)$p_1/p_2$ 的

- 点估计:

$$\hat{p}_1/\hat{p}_2$$

- 区间估计:

$$\hat{p}_1/\hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\widehat{\mathrm{RR}})} \approx \hat{p}_1/\hat{p}_2 \pm z_{1-\alpha/2}\hat{p}_1/\hat{p}_2(\frac{b}{an_1}+\frac{d}{cn_2})$$

代码:

```
library(fmsb)
riskratio(X, Y, m1, m2, conf.level=0.95)
```

输入的参数为

- `X`: 暴露组中的患病人数
- `Y`: 非暴露组中的患病人数
- `m1`: 暴露组中的总人数
- `m2`: 非暴露组中的总人数
- `conf.level`: 置信水平

**例 2** (接上例) 求危险度比

**解答:**

```
library(fmsb)
riskratio(13, 7, 5000, 10000, conf.level=0.95)
```

```
##            Disease Nondisease Total
## Exposed         13       4987  5000
## Nonexposed       7       9993 10000

##
##  Risk ratio estimate and its significance probability
##
## data:  13 7 5000 10000
```

```
## p-value = 0.002646
## 95 percent confidence interval:
##  1.482854 9.303627
## sample estimates:
## [1] 3.714286
```

### 7.1.3 优势比

设 $p_1, p_2$ 分别为暴露及未暴露的被试者中患病的比例, 我们的任务是得到优势比 (odds ratio)$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$ 的

- 点估计:

$$\frac{ad}{bc}$$

- 区间估计:

$$\exp\left\{\log(\mathrm{OR}) \pm z_{\alpha/2}\sqrt{\mathrm{Var}\{\log(\mathrm{OR})\}}\right\} \approx \exp\left\{\log(\mathrm{OR}) \pm z_{\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right\}$$

代码:

```
library(fmsb)
oddsratio(a, b, c, d, conf.level=0.95)

# 或者是
oddsratio(A, conf.level=0.95)
```

输入的参数为

- a,b,c,d:

| | | 患病 | | |
|---|---|---|---|---|
| | | 是 | 否 | 总数 |
| 暴露 | 是 | $a$ | $c$ | $a+c=n_1$ |
| | 否 | $b$ | $d$ | $b+d=n_2$ |
| | 总数 | $a+b=m_1$ | $c+d=m_2$ | |

  或者可以把上述表格做成一个矩阵, 填入 `A` 中
- `conf.level`: 置信水平

注意: `vcd` 包里也有同名函数, 因此如果使用报错, 请改为 `fmsb::oddsratio` 来强制使用 `fmsb` 包里的函数.

**例 3** (接例 1) 求优势比

**解答:**

```
library(fmsb)
oddsratio(13, 7, 4987, 9993)
```

```
##            Disease Nondisease Total
## Exposed         13        4987  5000
## Nonexposed       7        9993 10000
## Total           20       14980 15000

##
##  Odds ratio estimate and its significance probability
##
## data:  13 7 4987 9993
## p-value = 0.002646
## 95 percent confidence interval:
##  1.483814 9.333060
## sample estimates:
## [1] 3.721361
```

## 7.2 分层分类数据 (三维列联表) 的统计推断

### 7.2.1 Mantel-Haenszel 检验

我们要检验的假设为

$$H_0 : \mathrm{OR}_1 = \cdots = \mathrm{OR}_k = 1 \quad \text{vs} \quad H_1 : \neg H_0$$

我们主要要用到的命令为 `mantelhaen.test`, 但有两种输入的方式:

- 制作一个三维的列联表. 以 $2 \times 2 \times 2$ 列联表为例: (为了防止出错, 输入的时候请对照下表以及代码)

层 1

| | 列属性 1 | 列属性 2 | |
|---|---|---|---|
| 行属性 1 | $a_{11}$ | $a_{12}$ | — |
| 行属性 2 | $a_{21}$ | $a_{22}$ | — |
| | — | — | — |

层 2

| | 列属性 1 | 列属性 2 | |
|---|---|---|---|
| 行属性 1 | $b_{11}$ | $b_{12}$ | — |
| 行属性 2 | $b_{21}$ | $b_{22}$ | — |
| | — | — | — |

```
table <- array(c(a11,a21,a12,a22,b11,b21,b12,b22),
               dim = c(2,2,2),
               dimnames = list(行属性名 = c(" 行属性 1", " 行属性 2"),
                               列属性名 = c(" 列属性 1", " 列属性 2"),
                               层属性名 = c(" 层属性 1", " 层属性 2"))
```

$2 \times 2 \times k$ 的列联表也同理.

然后使用命令

```
mantelhaen.test(table)
```

- 分开两层, 每层各制作一个列联表 (矩阵). $2 \times 2 \times 2$ 表中推荐使用这种方法!

```
table1 <- matrix(c(a11,a12,
                   a21,a22), nrow = 2, ncol = 2, byrow = TRUE)
table2 <- matrix(c(b11,b12
                   b21,b22), nrow = 2, ncol = 2, byrow = TRUE)
```

然后使用命令

```
mantelhaen.test(table1, table2)
```

**例 4** 1985 年开展的一项研究向 15-59 岁间的 518 名癌症患者以及年龄与性别都相匹配的另外 518 名对照邮寄了问卷. 研究的目的在于被动吸烟对患癌风险的影响. 在这项研究中, 暴露变量是被动吸烟 (具体指被研究者的配偶每天至少吸 1 支烟且烟龄至少 6 个月). 一个可能的混杂变量就是被调查者本人是否吸烟 (即, 本人吸烟), 这是因为本人吸烟与患癌风险以及配偶是否吸烟都有关系. 因此在研究被动吸烟与患癌风险的关系时, 控制本人吸烟这一混杂变量很重要.

| | 本人吸烟 | | | 本人不吸烟 | | |
|---|---|---|---|---|---|---|
| | 被动吸烟 | | | 被动吸烟 | | |
| | 是 | 否 | | 是 | 否 | |
| 病例 | 120 | 111 | 231 | 161 | 117 | 278 |
| 对照 | 80 | 155 | 235 | 130 | 124 | 254 |
| | 200 | 266 | 466 | 291 | 241 | 532 |

控制了混杂变量后, 被动吸烟与患癌有关系吗?

**解答:**

```
table<- array(c(120,80,111,155,161,130,117,124),
              dim = c(2,2,2), dimnames = list(
              status = c("case", "control"),
              passive.smoker = c("yes", "no"),
              active.smoker = c("yes", "no")))
table
```

```
## , , active.smoker = yes
##
##          passive.smoker
## status    yes  no
##   case    120 111
##   control  80 155
##
## , , active.smoker = no
##
```

```
##          passive.smoker
## status    yes  no
##   case    161 117
##   control 130 124
```

```
mantelhaen.test(table)
```

```
##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  table
## Mantel-Haenszel X-squared = 13.942, df = 1, p-value = 0.0001885
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.263955 2.090024
## sample estimates:
## common odds ratio
##          1.625329
```

### 7.2.2 不同层间 OR 齐性的卡方检验 (Woolf 方法)

我们要检验的假设为

$$H_0 : \mathrm{OR}_1 = \cdots = \mathrm{OR}_k \quad \text{vs} \quad H_1 : \text{存在} i \neq j \text{ 使得} \mathrm{OR}_i \neq \mathrm{OR}_j$$

使用的命令为: (注意: 这里仅支持输入三维列联表, 即上述的方法 1)

```
library(vcd)
woolf_test(table)
```

**例 5** (接上例) 在不同层的被动吸烟与患癌有关系吗?

**解答:**

```
library(vcd)
woolf_test(table)
```

```
##
##  Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)
##
## data:  table
## X-squared = 3.2697, df = 1, p-value = 0.07057
```

### 7.2.3 有混杂下的趋势检验

设 $p_{ij}$ 为第 $i$ 层第 $j$ 个暴露组中的患病率, 设 $p_{ij} = \alpha_i + \beta x_j$. 要检验的假设为

$$H_0 : \beta = 0 \quad \longleftrightarrow \quad H_1 : \beta \neq 0$$

代码:

```
Mantel.Extension.test <- function(x,score){
  b <- dim(x)[1]; k <- dim(x)[2]; s <- dim(x)[3]
  Ni <- array(0,s); ni <- array(0,s); mi <- array(0,s)
  for (i in 1:s) {
    Ni[i] <- sum(x[,,i])
    ni[i] <- sum(x[2,,i])
    mi[i] <- sum(x[1,,i])
  }
  O <- sum(x[2,,1:s]*score)
  s1 <- colSums(colSums(x[,,1:s])*score)
  s2 <- colSums(colSums(x[,,1:s])*score^2)
  E <- sum(s1*ni/Ni)
  V <- sum(ni*mi*(Ni*s2-s1^2)/(Ni^2*(Ni-1)))
  XTR2 <- (abs(O-E)-0.5)^2/V
  p <- 1-pchisq(q=XTR2,df=1)
  return(list(O=O,E=E,V=V,test_statistic=XTR2,p_value=p))
}
```

**例 6** 一项关于睡眠呼吸障碍的调查旨在研究 30–60 岁人群中这种病的患病率. 受访者如果有书上所列的两类现象, 则被归类为习惯性打鼾者. 调查的结果如下表所示:

| | 女性 | | | 男性 | |
|---|---|---|---|---|---|
| | 是否为习惯性打鼾者 | | | 是否为习惯性打鼾者 | |
| 年龄 | 是 | 否 | 年龄 | 是 | 否 |
| 30-39 | 196 | 603 | 30-39 | 188 | 348 |
| 40-49 | 223 | 486 | 40-49 | 313 | 383 |
| 50-59 | 103 | 232 | 50-59 | 232 | 206 |

**解答:**

```
x <- array(c(603,196,486,223,232,103,
             348,188,383,313,206,232),
           dim = c(2,3,2),
           dimnames = list('ill'=c('no','yes'),
                           'age'=c('30-39','40-49','50-59'),
                           'sex'=c('M','F')))
Mantel.Extension.test(x, score=1:3)
```

```
## $O
## [1] 2461
##
## $E
## [1] 2335.565
##
## $V
## [1] 445.2057
##
## $test_statistic
## [1] 35.05958
##
## $p_value
## [1] 3.197706e-09
```

## 7.3 多重 Logistic 回归

### 7.3.1 参数估计 (模型拟合)

使用的代码为

```
lrf1 <- glm(Y ~ A + B + C, family=binomial(link = "logit"), data = 你的数据集名称)
summary(lrf1)
confint(lrf1, level = 0.95)
```

Logistic 回归得出系数估计很简单, 难点在于如何诠释系数的含义.

**例 7** (热身, 这是简单 Logistic 回归) 西部联合研究组 (WCGS) 在一项 10 年的研究中记录了吸烟状态 (x = 1 或 0) 以及冠心病 (CHD, chd69) 等相关数据.

**解答:**

```
wcgs<-read.csv("wcgs.csv",header=T)
lrf1<-glm(chd69 ~ smoke,family=binomial(link = "logit"), data=wcgs)
summary(lrf1)
```

```
##
## Call:
## glm(formula = chd69 ~ smoke, family = binomial(link = "logit"),
##     data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4731  -0.4731  -0.3497  -0.3497   2.3769
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7636     0.1042  -26.54  < 2e-16 ***
## smoke         0.6299     0.1337    4.71 2.47e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1781.2  on 3153  degrees of freedom
## Residual deviance: 1758.4  on 3152  degrees of freedom
## AIC: 1762.4
##
## Number of Fisher Scoring iterations: 5
```

```
confint(lrf1, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -2.9741104 -2.5653968
## smoke        0.3697937  0.8945827
```

**例 8** 使用多重 Logistic 回归拟合 WCGS 数据, 通过年龄, 胆固醇水平 (chol), 收缩压 (sbp), 身体质量指数 (bmi) 以及目前的吸烟状况 (smoke) 来预测是否患有冠心病 (chd69).

**解答:**

```
res<-glm(chd69~factor(smoke)+age+chol+sbp+bmi,family=binomial,data=wcgs)
summary(res)
```

```
##
## Call:
## glm(formula = chd69 ~ factor(smoke) + age + chol + sbp + bmi,
##     family = binomial, data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1512  -0.4408  -0.3276  -0.2398   2.8824
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -12.336305   0.975011 -12.652  < 2e-16 ***
## factor(smoke)1   0.632693   0.140079   4.517 6.28e-06 ***
## age              0.064356   0.011905   5.406 6.45e-08 ***
## chol             0.010839   0.001491   7.267 3.67e-13 ***
## sbp              0.019305   0.004091   4.719 2.37e-06 ***
```

```
## bmi              0.057671   0.026350   2.189   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1779.2  on 3141  degrees of freedom
## Residual deviance: 1614.6  on 3136  degrees of freedom
##   (12 observations deleted due to missingness)
## AIC: 1626.6
##
## Number of Fisher Scoring iterations: 6
```

```
confint(res, level = 0.95)
```

```
##                         2.5 %       97.5 %
## (Intercept)    -14.268675353 -10.44387932
## factor(smoke)1   0.360144855   0.90983108
## age              0.041020176   0.08772358
## chol             0.007934249   0.01378028
## sbp              0.011198159   0.02725461
## bmi              0.005722808   0.10904913
```

### 7.3.2 假设检验

对于系数的检验与线性回归是类似的, 都可以通过 `summary()` 中的输出得到.

### 7.3.3 预测

我们还是使用 `predict` 命令基于拟合出来的模型进行预测

```
pre <- data.frame(pre <- data.frame(变量名 1=c(你想要的值 1), 变量名 2=c(你想要的值 2),...))
lgt <- predict(拟合的模型, newdata=pre, type="link", se.fit=TRUE)
CI <- lgt$fit + 1.96 * c(-1, 1) * lgt$se.fit
plogis(c(lgt$fit,CI))
```

**例 9** (接上例) 根据下面的条件预测 10 年内患冠心病的概率, 给出 95% 置信区间: 60 岁的吸烟者; 总胆固醇水平为 253 mg/dL; 收缩压为 136 mmHg; BMI 为 25.

**解答:**

```
res <- glm(chd69~factor(smoke)+age+chol+sbp+bmi,family=binomial,data=wcgs)
pre <- data.frame(age=60,chol=253,sbp=136,bmi=25,smoke=1)
lgt <- predict(res, newdata=pre, type="link", se.fit=TRUE)
```

```
CI <- lgt$fit + 1.96 * c(-1, 1) * lgt$se.fit
plogis(c(lgt$fit,CI))
```

```
##         1
## 0.2625415 0.2041442 0.3307024
```

# 8 假设检验：人-时数据

## 8.1 发病率的单样本检验

设在 $t$ 人年的随访中一共出现 $a$ 个事件, ID = 潜在的发病密度 (率).

要检验的假设为

$$H_0 : \mathrm{ID} = \mathrm{ID}_0 \quad \longleftrightarrow \quad H_1 : \mathrm{ID} \neq \mathrm{ID}_0$$

- 大样本下 ($\mathrm{ID}_0 \times t > 10$) 检验统计量为:

$$X^2 = \frac{(a-\mu_0)^2}{\mu_0} \overset{H_0}{\sim} \chi_1^2,$$

  其中 $\mu_0 = \mathrm{ID}_0 \times t$. $p-$ 值 $= \mathbb{P}(\chi_1^2 > X^2)$

- 小样本下 ($\mathrm{ID}_0 \times t \leq 10$), 使用精确法计算 $p$ 值

$$\min\left\{2 \times \sum_{k=0}^{a} \frac{e^{-\mu_0}\mu_0^k}{k!},\ 1\right\}, \quad \text{若 } a < \mu_0,$$
$$\min\left\{2 \times \left(1 - \sum_{k=0}^{a-1} \frac{e^{-\mu_0}\mu_0^k}{k!}\right),\ 1\right\}, \text{ 若 } a \geqslant \mu_0.$$

所用代码需要自己手写:

```
incidence_onesample<-function(a,t,ID0,alpha=0.05)
{
  if(t*ID0>10)
  { # 大样本
    u0=t*ID0
    X=(a-u0)^2/u0
    p=1-pchisq(X,1)
    if (p <=alpha) con="reject" else con="accept"
    result <- list(X_square=X,p_value=p,conclusion=con)
  }
  else
  { # 小样本
    if (a<t*ID0) p=min(2*ppois(a,t*ID0),1)
    else p=min(2*(1-ppois(a-1,t*ID0)),1)
```

```
    if (p <=alpha) con="reject" else con="accept"
    result <- list(p_value=p,conclusion=con)
  }
  return(result)
}
```

**例 1** 1990-1994 年间对疑似有乳腺癌标记基因但未患病的女性进行了登记. 500 名 60 ~ 64 岁的女性被选中并随访至 2000 年 12 月 31 日. 随访总长度为 4000 人年, 在此之间乳腺癌一共新发病 28 例. 这组女性的乳腺癌发病密度是否与 60 ~ 64 岁普通人群中的平均发病密度 ($400/10^5$ 人年) 相同?

**解答:**

```
incidence_onesample(28,4000,400/10^5)
```

```
## $X_square
## [1] 9
##
## $p_value
## [1] 0.002699796
##
## $conclusion
## [1] "reject"
```

可以与讲义上的手算结果进行对比.

## 8.2 发病率的两样本检验

假设有如下所示的表格

| 暴露组 | 事件数 | 人-时 |
|---|---|---|
| 1 | $a_1$ | $t_1$ |
| 2 | $a_2$ | $t_2$ |
| 总数 | $a_1+a_2$ | $t_1+t_2$ |

假设为

$$H_0: \mathrm{ID}_1 = \mathrm{ID}_2 \quad \longleftrightarrow \quad H_1: \mathrm{ID}_1 \neq \mathrm{ID}_2$$

其中 $\mathrm{ID}_i$ = 第 $i$ 组的真实发病密度

- 大样本下 ($V_i > 5$), 检验统计量为

$$z = \begin{cases} \frac{a_1 - E_1 - .5}{\sqrt{V_1}}, & \text{若 } a_1 > E_1, \\ \frac{a_1 - E_1 + .5}{\sqrt{V_1}}, & \text{若 } a_1 \leqslant E_1, \end{cases} \quad z \overset{H_0}{\sim} N(0,1)$$

其中 $V_1=(a_1+a_2)t_1t_2/(t_1+t_2)^2$, $p$ 值为

$$p-\text{value}=\begin{cases}2\times[1-\Phi(z)], & \text{若 } z\geqslant 0,\\ 2\times\Phi(z), & \text{若 } z<0.\end{cases}$$

- 小样本下 ($V_i\leq 5$), 设 $p$ 为第一组中事件发生的真实比例. 假设为 (其中 $p_0=t_1/(t_1+t_2)$ )

$$H_0: p=p_0 \quad \text{vs.} \quad H_a: p\neq p_0$$

精确 $p$ 值为 ($q_0=1-p_0$)

$$\begin{cases}2\times\sum_{k=0}^{a_1}\binom{a_1+a_2}{k}p_0^k q_0^{a_1+a_2-k}, & \text{若 } a_1<(a_1+a_2)p_0,\\ 2\times\sum_{k=a_1}^{a_1+a_2}\binom{a_1+a_2}{k}p_0^k q_0^{a_1+a_2-k}, & \text{若 } a_1\geq(a_1+a_2)p_0.\end{cases}$$

这个检验方式对于一般情况下的两个发病密度的比较都适用, 但对于 $V_1<5$ 的情形更为有用 (因为无法使用正态近似).

所用代码需要自己手写:

```
incidence_twosample<-function(a1,a2,t1,t2,alpha=0.05)
{
  E1=(a1+a2)*t1/(t1+t2)
  V1=(a1+a2)*t1*t2/(t1+t2)^2
  if (V1>=5)
  {
    if (a1>E1) z=(a1-E1-0.5)/sqrt(V1)
    else z=(a1-E1+0.5)/sqrt(V1)
    p=2*(1-pnorm(abs(z)))
    if (p<=alpha) con="reject" else con="accept"
    result <- list(p_value=p,conclusion=con,z=z)
  }
  else
  {
    p0=t1/(t1+t2);q0=1-p0;w=(a1+a2)*p0
    if (a1<w) p=2*pbinom(a1,a1+a2,p0)
    else p=2*(1-pbinom(a1-1,a1+a2,p0))
    if (p<=alpha) con="reject" else con="accept"
    result <- list(p_value=p,conclusion=con)
  }
  return(result)
}
```

**例 2** 在 1976 年开展的护士健康研究中, 根据不同的 OC 使用情况 (目前服用/曾经服用/从未服用) 对一些未患乳腺癌的女性进行分类. 每两年通过邮寄问卷更新这些女性的 OC 使用情况, 并且关注她们此后两年内的乳腺癌发病状况. 计算每位女性目前服用 OC 或从不用 OC 的总时长 (忽略曾经服用). 下表是 45-49 岁的女性的相关数据.

| OC 使用情况 | 病例数 | 人年数 |
|---|---|---|
| 目前服用 | 9 | 2,935 |
| 从未服用 | 239 | 135,130 |

如何确定两组女性的乳腺癌发病密度是否有显著性差异?

**解答:**

```
incidence_twosample(9,239,2935,135130)
```

```
## $p_value
## [1] 0.1553016
##
## $conclusion
## [1] "accept"
##
## $z
## [1] 1.421052
```

**例 3** 下表是 30-34 岁女性的 OC 使用情况与乳腺癌发病的相关数据.

| OC 使用情况 | 病例数 | 人年数 |
|---|---|---|
| 目前服用 | 3 | 8,250 |
| 从未服用 | 9 | 17,430 |

检验两组女性的乳腺癌发病密度是否有显著性差异.

**解答:**

```
incidence_twosample(3,9,8250,17430)
```

```
## $p_value
## [1] 0.8564199
##
## $conclusion
## [1] "accept"
```

### 8.2.1 率比的点估计与区间估计

率比 (RR) 定义为 $\mathrm{RR}=\dfrac{\text{暴露组的发病密度}}{\text{非暴露组的发病密度}}$. 它的一个点估计为

$$\widehat{\mathrm{IRR}}=\frac{a_1/t_1}{a_2/t_2}=\frac{a_1t_2}{a_2t_1}.$$

若 $V_1 \geq 5$, 则 RR 的双侧 \$100% ×(1- ) \$ 置信区间为

$$(c_1, c_2) = \exp\left\{\ln(\widehat{\text{IRR}}) \pm z_{1-\alpha/2}\sqrt{\frac{1}{a_1}+\frac{1}{a_2}}\right\}$$

```
rate_ratio<-function(a1,a2,t1,t2,alpha=0.05)
{
  if ((a1+a2)*t1*t2/(t1+t2)^2<5) stop("V1<5, the data doesn't match the function")
  RR=a1*t2/a2/t1
  interval=qnorm(1-alpha/2)*sqrt(1/a1+1/a2)
  d1=log(RR)-interval
  d2=log(RR)+interval
  return(c(rate_ratio=RR,CI=c(exp(d1),exp(d2))))
}
```

**例 4** 基于例 2 的乳腺癌案例的数据,

| OC 使用情况 | 病例数 | 人年数 |
|---|---|---|
| 目前服用 | 9 | 2,935 |
| 从未服用 | 239 | 135,130 |

求 RR 的点估计以及 95% 置信区间.

**解答:**

```
rate_ratio(9,239,2935,135130)
```

```
## rate_ratio        CI1        CI2
##   1.733757   0.891172   3.372990
```

## 8.3 生存分析

在生存分析中, 我们要用到的最基本的包为 `survival`

```
library(survival)
```

使用时都是要基于 individual-level 的数据集

### 8.3.1 创建生存对象

```
Surv(time, time2, event,
     type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate'))
```

输入的参数为:

- `time`: 对于单侧的删失 (通常右删失), 指随访的时间. 若为区间数据, 则为区间数据的开始时间.

- time2: 区间数据的结束时间.
- event: 结局变量
  - 默认 0 为删失, 1 为出现终点事件; 也可以 1 为删失, 2 为出现终点事件.
  - 可以自定义终点事件, 比如 df$status==2 就是将 status 为 2 定义为终点事件, 其他值代表删失.
- type: 声明这是何种删失

Surv 函数创建出来的是将时间数据和删失标记相结合的对象.

**例** 我们使用自带的 lung 数据集进行尝试.

```
data("lung")
knitr::kable(head(lung))
```

| inst | time | status | age | sex | ph.ecog | ph.karno | pat.karno | meal.cal | wt.loss |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 306 | 2 | 74 | 1 | 1 | 90 | 100 | 1175 | NA |
| 3 | 455 | 2 | 68 | 1 | 0 | 90 | 90 | 1225 | 15 |
| 3 | 1010 | 1 | 56 | 1 | 0 | 90 | 90 | NA | 15 |
| 5 | 210 | 2 | 57 | 1 | 1 | 90 | 60 | 1150 | 11 |
| 1 | 883 | 2 | 60 | 1 | 0 | 100 | 90 | NA | 0 |
| 12 | 1022 | 1 | 74 | 1 | 1 | 50 | 80 | 513 | 0 |

我们将基于 time 和 status 来创建生存对象:

```
sobject <- Surv(lung$time, lung$status)
head(sobject)
```

```
## [1]  306   455  1010+  210   883  1022+
```

可以看到, status=1 的都被打上了标签 + 来说明这个数据是右删失的.

### 8.3.2 创建生存曲线

如果只是想创建一条生存曲线 (Kaplan-Meier 估计曲线):

```
fit <- survfit(Surv(time, censor) ~ 1, conf.type="none")
```

如果想按某个属性的不同水平 (如: 男/女或者实验/对照) 创建多条生存曲线进行比较:

```
fit <- survfit(Surv(time, censor) ~ sex, conf.type="none")
```

查看整体拟合的结果, 可以用

```
summary(fit)
```

生存曲线的可视化可以直接用

```
plot(fit)
```

如果想要好看一点, 则可以采用 survminer 包的 ggsurvplot() 函数

```
library(survminer)
ggsurvplot(fit, pval = TRUE, conf.int = TRUE,
           risk.table = TRUE, # Add risk table
           risk.table.col = "strata", # Change risk table color by groups
           linetype = "strata", # Change line type by groups
           surv.median.line = "hv", # Specify median survival
           ggtheme = theme_bw(), # Change ggplot2 theme
           palette = c("#E7B800", "#2E9FDF"))
```

**例** 我们首先创造单一条的 KM 曲线

```
fit <- survfit(Surv(time, status) ~ 1, data = lung)
fit
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = lung)
##
##       n  events  median 0.95LCL 0.95UCL
##     228     165     310     285     363
```

```
summary(fit)
```

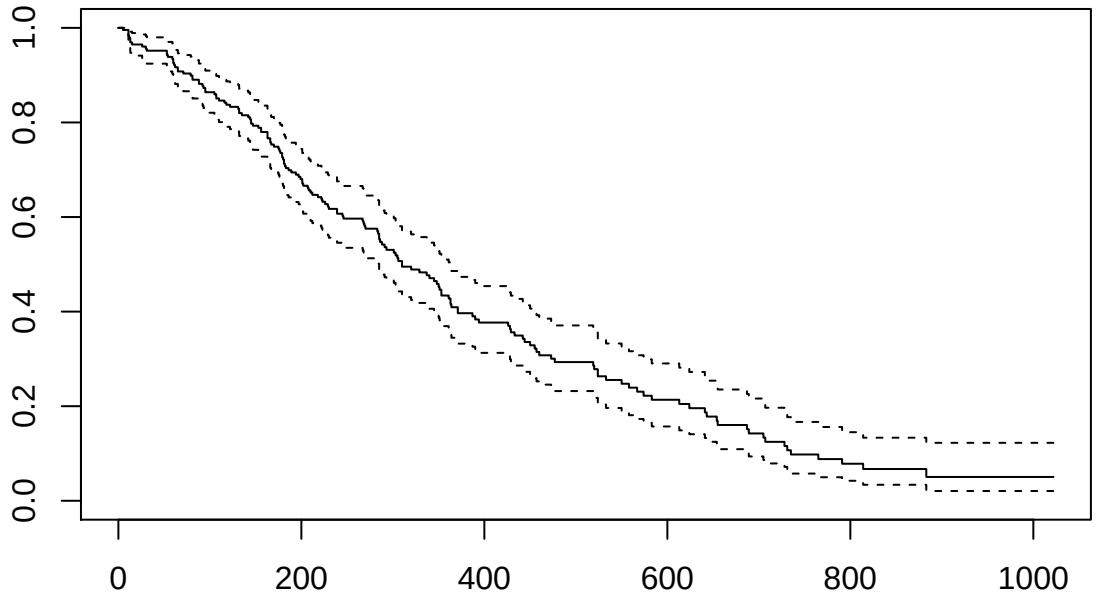
```
## Call: survfit(formula = Surv(time, status) ~ 1, data = lung)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     5    228       1   0.9956 0.00438       0.9871        1.000
##    11    227       3   0.9825 0.00869       0.9656        1.000
##    12    224       1   0.9781 0.00970       0.9592        0.997
##    13    223       2   0.9693 0.01142       0.9472        0.992
##    15    221       1   0.9649 0.01219       0.9413        0.989
##    26    220       1   0.9605 0.01290       0.9356        0.986
##    30    219       1   0.9561 0.01356       0.9299        0.983
##    31    218       1   0.9518 0.01419       0.9243        0.980
##    53    217       2   0.9430 0.01536       0.9134        0.974
##    54    215       1   0.9386 0.01590       0.9079        0.970
##    59    214       1   0.9342 0.01642       0.9026        0.967
##    60    213       2   0.9254 0.01740       0.8920        0.960
##    61    211       1   0.9211 0.01786       0.8867        0.957
##    62    210       1   0.9167 0.01830       0.8815        0.953
##    65    209       2   0.9079 0.01915       0.8711        0.946
##    71    207       1   0.9035 0.01955       0.8660        0.943
##    79    206       1   0.8991 0.01995       0.8609        0.939
```

```
##    81    205       2   0.8904 0.02069       0.8507       0.932
##    88    203       2   0.8816 0.02140       0.8406       0.925
##    92    201       1   0.8772 0.02174       0.8356       0.921
##    93    199       1   0.8728 0.02207       0.8306       0.917
##    95    198       2   0.8640 0.02271       0.8206       0.910
##   105    196       1   0.8596 0.02302       0.8156       0.906
##   107    194       2   0.8507 0.02362       0.8056       0.898
##   110    192       1   0.8463 0.02391       0.8007       0.894
##   116    191       1   0.8418 0.02419       0.7957       0.891
##   118    190       1   0.8374 0.02446       0.7908       0.887
##   122    189       1   0.8330 0.02473       0.7859       0.883
##   131    188       1   0.8285 0.02500       0.7810       0.879
##   132    187       2   0.8197 0.02550       0.7712       0.871
##   135    185       1   0.8153 0.02575       0.7663       0.867
##   142    184       1   0.8108 0.02598       0.7615       0.863
##   144    183       1   0.8064 0.02622       0.7566       0.859
##   145    182       2   0.7975 0.02667       0.7469       0.852
##   147    180       1   0.7931 0.02688       0.7421       0.848
##   153    179       1   0.7887 0.02710       0.7373       0.844
##   156    178       2   0.7798 0.02751       0.7277       0.836
##   163    176       3   0.7665 0.02809       0.7134       0.824
##   166    173       2   0.7577 0.02845       0.7039       0.816
##   167    171       1   0.7532 0.02863       0.6991       0.811
##   170    170       1   0.7488 0.02880       0.6944       0.807
##   175    167       1   0.7443 0.02898       0.6896       0.803
##   176    165       1   0.7398 0.02915       0.6848       0.799
##   177    164       1   0.7353 0.02932       0.6800       0.795
##   179    162       2   0.7262 0.02965       0.6704       0.787
##   180    160       1   0.7217 0.02981       0.6655       0.783
##   181    159       2   0.7126 0.03012       0.6559       0.774
##   182    157       1   0.7081 0.03027       0.6511       0.770
##   183    156       1   0.7035 0.03041       0.6464       0.766
##   186    154       1   0.6989 0.03056       0.6416       0.761
##   189    152       1   0.6943 0.03070       0.6367       0.757
##   194    149       1   0.6897 0.03085       0.6318       0.753
##   197    147       1   0.6850 0.03099       0.6269       0.749
##   199    145       1   0.6803 0.03113       0.6219       0.744
##   201    144       2   0.6708 0.03141       0.6120       0.735
##   202    142       1   0.6661 0.03154       0.6071       0.731
##   207    139       1   0.6613 0.03168       0.6020       0.726
##   208    138       1   0.6565 0.03181       0.5970       0.722
##   210    137       1   0.6517 0.03194       0.5920       0.717
```

```
##   212   135       1 0.6469 0.03206       0.5870       0.713
##   218   134       1 0.6421 0.03218       0.5820       0.708
##   222   132       1 0.6372 0.03231       0.5769       0.704
##   223   130       1 0.6323 0.03243       0.5718       0.699
##   226   126       1 0.6273 0.03256       0.5666       0.694
##   229   125       1 0.6223 0.03268       0.5614       0.690
##   230   124       1 0.6172 0.03280       0.5562       0.685
##   239   121       2 0.6070 0.03304       0.5456       0.675
##   245   117       1 0.6019 0.03316       0.5402       0.670
##   246   116       1 0.5967 0.03328       0.5349       0.666
##   267   112       1 0.5913 0.03341       0.5294       0.661
##   268   111       1 0.5860 0.03353       0.5239       0.656
##   269   110       1 0.5807 0.03364       0.5184       0.651
##   270   108       1 0.5753 0.03376       0.5128       0.645
##   283   104       1 0.5698 0.03388       0.5071       0.640
##   284   103       1 0.5642 0.03400       0.5014       0.635
##   285   101       2 0.5531 0.03424       0.4899       0.624
##   286    99       1 0.5475 0.03434       0.4841       0.619
##   288    98       1 0.5419 0.03444       0.4784       0.614
##   291    97       1 0.5363 0.03454       0.4727       0.608
##   293    94       1 0.5306 0.03464       0.4669       0.603
##   301    91       1 0.5248 0.03475       0.4609       0.597
##   303    89       1 0.5189 0.03485       0.4549       0.592
##   305    87       1 0.5129 0.03496       0.4488       0.586
##   306    86       1 0.5070 0.03506       0.4427       0.581
##   310    85       2 0.4950 0.03523       0.4306       0.569
##   320    82       1 0.4890 0.03532       0.4244       0.563
##   329    81       1 0.4830 0.03539       0.4183       0.558
##   337    79       1 0.4768 0.03547       0.4121       0.552
##   340    78       1 0.4707 0.03554       0.4060       0.546
##   345    77       1 0.4646 0.03560       0.3998       0.540
##   348    76       1 0.4585 0.03565       0.3937       0.534
##   350    75       1 0.4524 0.03569       0.3876       0.528
##   351    74       1 0.4463 0.03573       0.3815       0.522
##   353    73       2 0.4340 0.03578       0.3693       0.510
##   361    70       1 0.4278 0.03581       0.3631       0.504
##   363    69       2 0.4154 0.03583       0.3508       0.492
##   364    67       1 0.4092 0.03582       0.3447       0.486
##   371    65       2 0.3966 0.03581       0.3323       0.473
##   387    60       1 0.3900 0.03582       0.3258       0.467
##   390    59       1 0.3834 0.03582       0.3193       0.460
##   394    58       1 0.3768 0.03580       0.3128       0.454
```

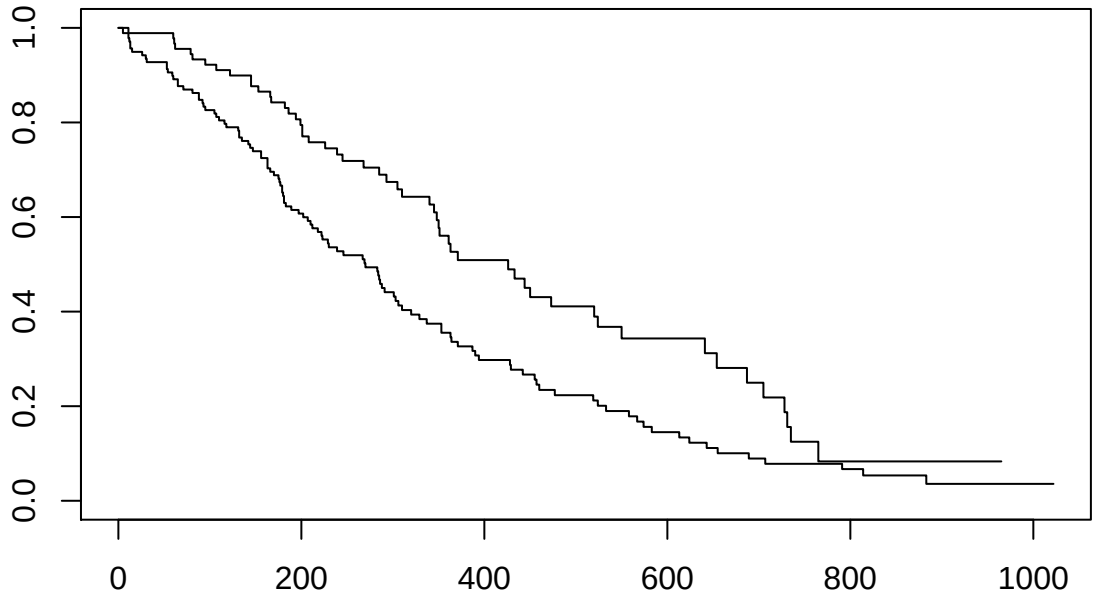
```
##    426      55        1    0.3700 0.03580       0.3060        0.447
##    428      54        1    0.3631 0.03579       0.2993        0.440
##    429      53        1    0.3563 0.03576       0.2926        0.434
##    433      52        1    0.3494 0.03573       0.2860        0.427
##    442      51        1    0.3426 0.03568       0.2793        0.420
##    444      50        1    0.3357 0.03561       0.2727        0.413
##    450      48        1    0.3287 0.03555       0.2659        0.406
##    455      47        1    0.3217 0.03548       0.2592        0.399
##    457      46        1    0.3147 0.03539       0.2525        0.392
##    460      44        1    0.3076 0.03530       0.2456        0.385
##    473      43        1    0.3004 0.03520       0.2388        0.378
##    477      42        1    0.2933 0.03508       0.2320        0.371
##    519      39        1    0.2857 0.03498       0.2248        0.363
##    520      38        1    0.2782 0.03485       0.2177        0.356
##    524      37        2    0.2632 0.03455       0.2035        0.340
##    533      34        1    0.2554 0.03439       0.1962        0.333
##    550      32        1    0.2475 0.03423       0.1887        0.325
##    558      30        1    0.2392 0.03407       0.1810        0.316
##    567      28        1    0.2307 0.03391       0.1729        0.308
##    574      27        1    0.2221 0.03371       0.1650        0.299
##    583      26        1    0.2136 0.03348       0.1571        0.290
##    613      24        1    0.2047 0.03325       0.1489        0.281
##    624      23        1    0.1958 0.03297       0.1407        0.272
##    641      22        1    0.1869 0.03265       0.1327        0.263
##    643      21        1    0.1780 0.03229       0.1247        0.254
##    654      20        1    0.1691 0.03188       0.1169        0.245
##    655      19        1    0.1602 0.03142       0.1091        0.235
##    687      18        1    0.1513 0.03090       0.1014        0.226
##    689      17        1    0.1424 0.03034       0.0938        0.216
##    705      16        1    0.1335 0.02972       0.0863        0.207
##    707      15        1    0.1246 0.02904       0.0789        0.197
##    728      14        1    0.1157 0.02830       0.0716        0.187
##    731      13        1    0.1068 0.02749       0.0645        0.177
##    735      12        1    0.0979 0.02660       0.0575        0.167
##    765      10        1    0.0881 0.02568       0.0498        0.156
##    791       9        1    0.0783 0.02462       0.0423        0.145
##    814       7        1    0.0671 0.02351       0.0338        0.133
##    883       4        1    0.0503 0.02285       0.0207        0.123
```

```
plot(fit)
```




我们接下来按性别来分别画 KM 曲线

```
fit <- survfit(Surv(time, status) ~ sex, data = lung)
plot(fit)
```




想要画的更好看一些, 可以用

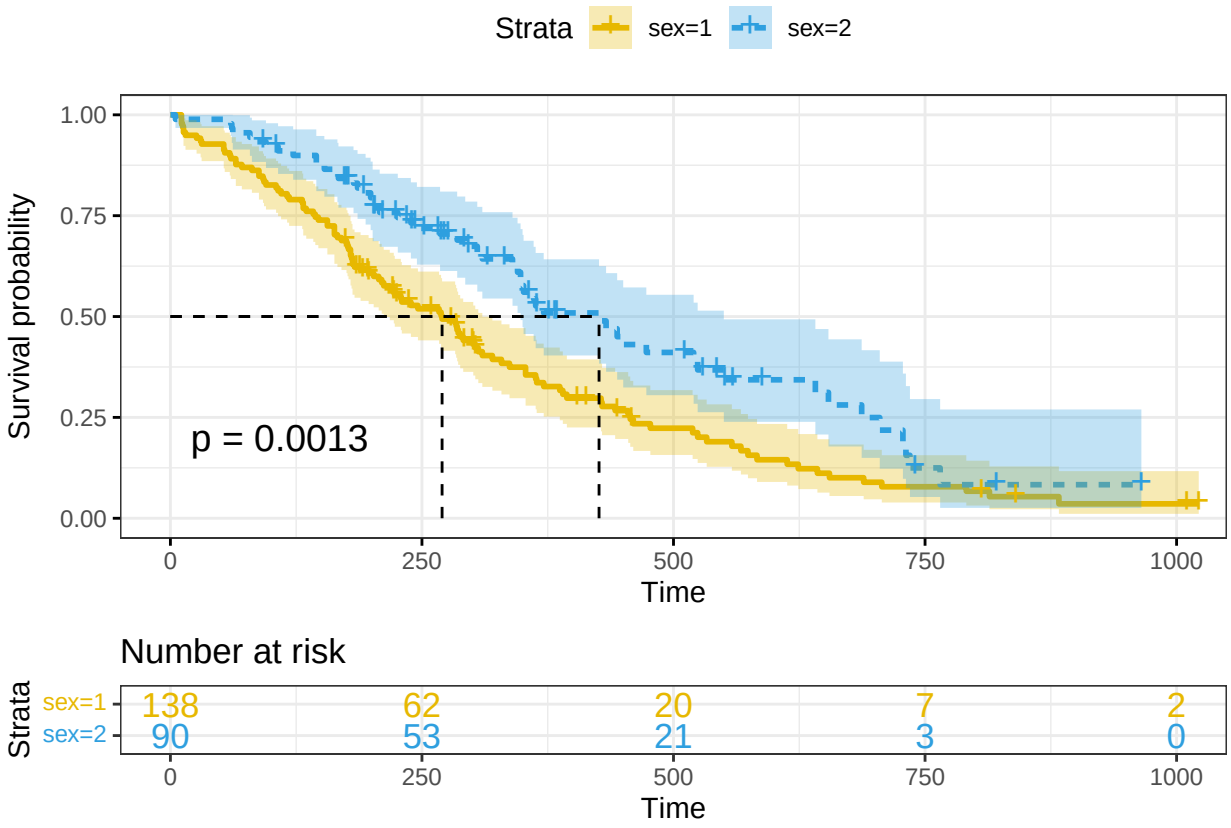
```
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 4.0.5
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'ggpubr' was built under R version 4.0.5
```

```
ggsurvplot(fit, pval = TRUE, conf.int = TRUE,
           risk.table = TRUE, # Add risk table
           risk.table.col = "strata", # Change risk table color by groups
           linetype = "strata", # Change line type by groups
           surv.median.line = "hv", # Specify median survival
           ggtheme = theme_bw(), # Change ggplot2 theme
           palette = c("#E7B800", "#2E9FDF"))
```




### 8.3.3 假设检验

这里我们指的是 Log-Rank 检验. 用的是 `survdiff` 函数

```
survdiff(Surv(time, censor)~drug, rho=0)
```

特殊的参数设置:

- `rho`: `rho = 0` 表示 log-rank (CMH) 检验, `rho = 1` 表示 Gehan-Wilcoxon 检验的检验的 Peto & Peto 修正

**例** 在 `lung` 数据中, 不同性别人群的患病率是否有差别? 我们进行 log-rank 检验:

```
surv_diff <- survdiff(Surv(time, status) ~ sex, data = lung)
surv_diff
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = lung)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 138      112     91.6      4.55      10.3
## sex=2  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```