T检验

简单的假设检验

null hypothesis	零假设、原假设
alternative hypothesis	对立假设、备选假设
significant difference	显著的差别
significance level	显著性水平
test statistic	检验统计量
***p*-value**	***P*-值**
reject	拒绝
paired design	配对设计
pooled estimation	联合估计
type I error	第一类错误
type II error	第二类错误

总体均值的假设检验

某地区280位成年男性的血红蛋白含量

$$\overline{x} = 136.0 g/L, \qquad s = 6.0 g/L$$

问题

该地区成年男性的血红蛋白含量，也就是总体平均值μ是否为140.0 g/L?

两种可能性

（1）总体平均为140，样本均值的偏离是由于采样误差造成（零假设）

$$H_0: \mu = 140$$

（2）总体均值不为140，样本均值为136.0

$$H_1: \mu \neq 140$$

问题：

（1）（2）哪个才是真实的？

—— 这就是假设检验问题

基本方法

（1）当H_0成立时，出现这种样本的可能性是?

—— 计算概率(p-value)

（2）若p-value小于给定的检验显著性水平α，拒绝 H_0；反之接受H_0

与总体均值比较

某地水质调查，抽样检查了15个地点的$CaCO_3$含量（mg/L）：

20.99，20.41，20.62，20.75，20.10，20.00，20.80，20.91，22.60，22.30，20.99，20.41，20.50，23.00，22.60

检查该地区水中碳酸钙的含量是否为20.7mg/L.

$$\bar{x} = \frac{316.98}{15} = 21.13$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{15}}{n-1}}$$

$$= \sqrt{\frac{6711.98 - \frac{(316.98)^2}{15}}{15-1}} = 0.98$$

T检验

（1）设置假设和显著性水平

$$H_0: \mu = 20.7$$
$$H_1: \mu \neq 20.7$$
$$\alpha = 0.05$$

（2）选择合适的检验方法，计算相应的统计值

（3）如果X服从高斯分布，则统计值

$$t = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

服从t_{n-1}分布。

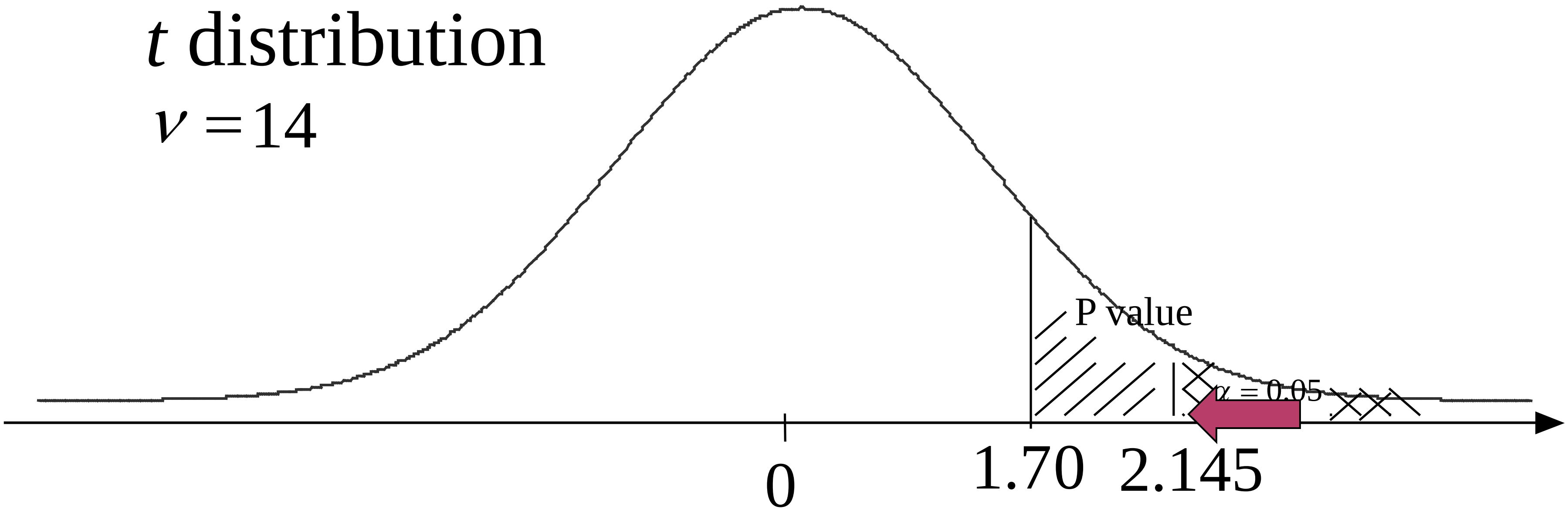
（4）决定：拒绝H_0还是接受.

- 当 $H_0: \mu = 20.7$ 成立时，

$$t = \frac{\overline{X} - 20.7}{\frac{S}{\sqrt{n}}} \sim t \quad distribution$$

- 基于当前的样本：

$$t = \frac{\overline{X} - 20.7}{\frac{S}{\sqrt{n}}} = \frac{21.13 - 20.7}{\frac{0.98}{\sqrt{15}}} = 1.70$$



t distribution
$\nu = 14$
P value
$\alpha = 0.05$
0
1.70
2.145

与总体均值比较的T检验

```
myttest1 <- function(x, mu=20.7) {
	n <- length(x)
	xbar <- mean(x)
	s <- var(x)
	t <- (xbar-mu)/sqrt(s/n)
	pval <- 2*pt(t, df=n-1)
	return(pval)
}
```

配对数据的比较

- 8个高血压病人用药前后血压DBP的变化如下表所示

DBP variation before and after treatment

No.	Before	After	Difference
1	96	88	8
2	112	108	4
3	108	102	6
4	102	98	4
5	98	100	-2
6	100	96	4
7	106	102	4
8	100	92	8
Total			36

- $H_0 : \mu_d = 0$ $\qquad H_1 : \mu_d \neq 0$ $\qquad$ ***a*=0.05**
- $t = \dfrac{\bar{d} - 0}{s_d / \sqrt{n}} = \dfrac{4.50}{3.16 / \sqrt{8}} = 4.02$ υ**=8-1=7**
- $t > t_{0.05,7} = 2.365$, $P < 0.05$, H_0 **is rejected at significance level *a*=0.05.**

两组样本均值的比较

- 两组大鼠，分别用高蛋白饲料和低蛋白饲料喂养，增重情况如下表

用两种不同蛋白质含量饲料喂养大鼠后体重增加的克数

高蛋白组	134	146	104	119	124	161	107	83	113	129	97	123
低蛋白组	70	118	101	85	107	132	94					

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$$

设定显著性水平α=0.05

- 计算联合样本方差
- 计算统计值t
- 自由度为n_1+n_2-2

$$s_c^2 = \frac{\sum(X_1 - \overline{X}_1)^2 + \sum(X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2}$$

$$= \frac{\sum X_1^2 - (\sum X_1)^2 / n_1 + \sum X_2^2 - (\sum X_2)^2 / n_2}{n_1 + n_2 - 2}$$

$$s_c^2 = \frac{177832 - 1440^2 / 12 + 73959 - 707^2 / 7}{12 + 7 - 2} = 446.12$$

$$s_{\overline{X}_1 - \overline{X}_x} = \sqrt{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{446.12\left(\frac{1}{12} + \frac{1}{7}\right)} = 10.05$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{120 - 101}{10.05} = 1.891$$

$\upsilon = n_1 + n_2 - 2 = 12 + 7 - 2 = 17$

练习题

- 针对上述两种情况，分别写出配对和非配对t检验的函数

- 函数原型myttest2 <- function(x, y, paired=FALSE)

- 用模拟数据进行检验

假设检验需要注意的问题

a. p值的意义

***P*-value is the area of the tail(s) in the distribution of the test statistic beyond the value(s) of the test statistic calculated based on the sample.**

- **If the null hypothesis is rejected, the probability of mistake = *P***

  -- A smaller *P*-value implies the better quality of your rejection.

- **If the null hypothesis is not rejected, the bigger *P*-value implies the better quality of your acceptation.**

b. 显著性水平α的意义

α 体现推断的质量，也就是当你拒绝零假设时，犯错误的概率被限制在 α

c. t检验适用的条件

（1）变量服从正态分布；

（2）样本量较小；

（3）样本具有相同的方差

其他常见的假设检验

- 假设检验主要包括两种大类
  - 一类是参数型的假设检验
    - Z检验
    - t检验
    - 二项式检验（binomial test）
    - 卡方检验
  - 另一类则是非参数型的假设检验
    - Wilcoxon秩和检验或Mann-Whitney检验
    - Wilcoxon符号秩检验
    - Kolmorov-Smirnov检验
    - Kaplan-Meier检验
    - Logrank检验

Z-TEST: TESTING THE NORMAL MEAN WHEN THE VARIANCE IS KNOWN

$$N\left(\mu, \sigma^2\right)$$

$$\begin{array}{lll} H_0 : \mu = \mu_0 & \text{vs} & H_1 : \mu = \mu_1 \\ H_0 : \mu = \mu_0 & \text{vs} & H_1 : \mu \neq \mu_0 \quad \triangleleft \\ H_0 : \mu = \mu_0 & \text{vs} & H_1 : \mu \leq \mu_0 \\ H_0 : \mu = \mu_0 & \text{vs} & H_1 : \mu \geq \mu_0 \end{array}$$

$$T = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

```
function(x,mu0,sigma) {
		n = length(x)
		t = abs( (mean(x)-mu0)  /  (sigma/sqrt(n)) )
		pnorm(-t) + pnorm(t,lower.tail=FALSE)
}
```

CHISQ.TEST: TESTING A NORMAL VARIANCE WHEN THE MEAN IS UNKNOWN

$$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

```
function(x,sigma0) {
        n = length(x)
        t = sum( ( (x-mean(x))/sigma0 )^2 )
        pchisq(t,n-1,lower.tail=FALSE)
}
```

T.TEST: TESTING A NORMAL MEAN WHEN THE VARIANCE IS UNKNOWN

$$T = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \sim t_{n-1}$$

```
function(x,mu0) {
        n = length(x)
        s2 = var(x)
        t = abs( (mean(x) - mu0) / sqrt(s2/n) )
        pt(-t,n-1,lower.tail=TRUE) + pt(t,n-1,lower.tail=FALSE)
}
```

CONFIDENCE INTERVAL FOR MEAN

$$\overline{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

```
function(x, a=0.95) {
        n = length(x)
        m = mean(x)
        s2 = var(x)
        c( m + qt((1-a)/2,n-1)*sqrt(s2/n) ,
                m + qt(1-(1-a)/2,n-1)*sqrt(s2/n) )
}
```

CONFIDENCE INTERVAL FOR VARIANCE

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\chi^2_{n-1,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}$$

```
function(x,a=0.95) {
        n = length(x)
        s2 = var(x)
        c( (n-1) * s2 / qchisq(1-(1-a)/2,n-1) ,
                (n-1) * s2 / qchisq((1-a)/2,n-1) )
}
```

Comparing Two Normal Variances

$$(X_1, \ldots, X_n) \sim N(\mu_x, \sigma_x^2)$$

$$(Y_1, \ldots, Y_m) \sim N(\mu_y, \sigma_y^2)$$

$$\frac{\dfrac{(n-1)s_x^2}{\sigma_x^2} / (n-1)}{\dfrac{(m-1)s_y^2}{\sigma_y^2} / (m-1)} = \frac{s_x^2}{\sigma_x^2} \times \frac{\sigma_y^2}{s_y^2} \sim F_{n-1,m-1}$$

```
function(x,y) {
        n = length(x)
        m = length(y)
        t = var(x)/var(y)
        lt = t < 1
        pf(t,n-1,m-1,lower.tail=lt) + pf(t,m-1,n-1,lower.tail=lt)
}
```

COMPARING TWO NORMAL MEANS

$$(X_1,\ldots,X_n) \sim N(\mu_x,\sigma_x^2)$$

$$(Y_1,\ldots,Y_m) \sim N(\mu_y,\sigma_y^2)$$

$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

$$s^2 = \frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}$$

```
ttest <- function(x,y) {
        n = length(x); m = length(y)
        sp = ( (n-1)*var(x) + (m-1)*var(y) ) / (n+m-2)
        t = abs(mean(x)-mean(y)) / sqrt(sp*(1/n+1/m))
        pt(-t,n+m-2) + pt(t,n+m-2,lower.tail=FALSE)
}
```