

R-8

郑泽靖 zzjstat2023@163.com

北京师范大学统计学院



假设检验问题

关于总体分布的不相容的猜测构成一个假设检验问题:

$$H_0 \longleftrightarrow H_1 \quad (1)$$

- 称 H_0 为原假设或零假设。
- 称 H_1 为备择假设或对立假设。

对于假设检验问题 $H_0 \longleftrightarrow H_1$:

- 当 p -值小于显著性水平时, 拒绝 H_0 。
- 否则, 接受 H_0 。



假设检验问题

- 设总体 $X \sim N(\mu, \sigma^2)$, 感兴趣的假设检验问题为:

$$H_0 : \mu = \mu_0, \tag{2}$$

$$H_0 : \mu \leq \mu_0, \tag{3}$$

$$H_0 : \mu \geq \mu_0, \tag{4}$$

其中 μ_0 为已知实数。

- 双边假设检验问题（简称双边检验问题）：
 - 类似于 $H_0 : \mu = \mu_0$ 的假设检验问题。
- 单边假设检验问题（简称单边检验问题）：
 - 类似于 $H_0 : \mu \leq \mu_0$ 或 $H_0 : \mu \geq \mu_0$ 的假设检验问题。



已知总体方差情况下的均值检验

- 当总体标准差 $\sigma = \sigma_0$ 已知时，检验统计量为：

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0}$$

- 由定理 5.1.1 知：

$$Z \sim N\left(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma_0}, 1\right)$$

- 特别地，当 $\mu = \mu_0$ 时， $Z \sim N(0, 1)$ 。



检验统计量的观测值

- 对于样本观测数据 x_1, x_2, \dots, x_n , 检验统计量 Z 的观测值为:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \quad z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_0}$$

```
1 n <- length(x) # 计算样本容量  
2 z <- (n^0.5) * (mean(x) - mu0) / sigma0 # 计算观测值
```



p-值计算

- 对于双边检验问题 (2), 其 p -值为:

$$P_{H_0}(|Z| > |z_0|) = 2(1 - \Phi(|z_0|))$$

对应的 R 代码:

```
1 pValue <- 2 * pnorm(abs(z), lower.tail = FALSE)
```

- 对于单边检验问题 (3), 其 p -值为:

$$P_{H_0}(Z > z_0) = 1 - \Phi(z_0)$$

对应的 R 代码:

```
1 pValue <- pnorm(z, lower.tail = FALSE)
```

- 对于单边检验问题 (4), 其 p -值为:

$$P_{H_0}(Z < z_0) = \Phi(z_0)$$

对应的 R 代码:

```
1 pValue <- pnorm(z, lower.tail = TRUE)
```



未知总体方差情况下的均值检验

- 当总体方差未知时，使用样本标准差 S 代替总体标准差，定义新的检验统计量：

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$$

- S 为样本标准差。



p-值计算

- 对于双边假设检验问题 (2)，可以使用 R 语言中的 ‘t.test’ 函数：

```
1 h <- t.test(x, alternative = "two.sided", mu = mu0)
2 pValue <- h$p.value
```

- 对于单边假设检验问题 (3)，使用以下 R 代码：

```
1 h <- t.test(x, alternative = "greater", mu = mu0)
2 pValue <- h$p.value
```

- 对于单边假设检验问题 (4)，使用以下 R 代码：

```
1 h <- t.test(x, alternative = "less", mu = mu0)
2 pValue <- h$p.value
```



双正态总体均值的检验

- 实际应用中，经常需要比较两个总体均值，例如比较两个班级学生的数学水平、两种安眠药的治疗效果等。
- 考虑总体 $X \sim N(\mu_1, \sigma^2)$ 和总体 $Y \sim N(\mu_2, \sigma^2)$ 。
- 双边假设检验问题：

$$H_0 : \mu_1 = \mu_2$$

- 单边假设检验问题：

$$H_0 : \mu_1 \geq \mu_2$$

或

$$H_0 : \mu_1 \leq \mu_2$$



检验统计量 Z

- 用样本均值之差 $\bar{Y} - \bar{X}$ 估计 $\mu_2 - \mu_1$ 。
- 构建的检验统计量：

$$Z = \frac{\bar{Y} - \bar{X}}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

- 当 σ 已知时：

$$Z \sim N\left(\frac{\sqrt{mn}(\mu_2 - \mu_1)}{\sigma \sqrt{n+m}}, 1\right)$$



检验统计量的观测值

- 样本数据 x_1, x_2, \dots, x_n 和 y_1, y_2, \dots, y_m 。
- 检验统计量 Z 的观测值：

$$z_0 = \frac{\sqrt{mn}(\bar{y} - \bar{x})}{\sigma\sqrt{m+n}}$$

```
1 n <- length(x) # 计算 X 的样本容量  
2 m <- length(y) # 计算 Y 的样本容量  
3 z <- (m * n)^0.5 * (mean(y) - mean(x))  
4 / (sigma / (m + n)^0.5)
```



双正态总体均值的 t 检验

- 当 σ 未知时，不能直接使用 Z 作为检验统计量。
- 需要使用总体标准差的估计量。
- 使用样本标准差的组合估计：

$$S_p = \sqrt{\frac{1}{n+m-2} ((n-1)S_X^2 + (m-1)S_Y^2)}$$

- 构建检验统计量：

$$T = \frac{\sqrt{mn}(\bar{Y} - \bar{X})}{S_p \sqrt{m+n}}$$

- T 服从 t 分布。



p-值计算

- 对于双边假设检验问题：

```
1 h <- t.test(x, y, alternative = "two.sided")
2 pValue <- h$p.value
```

- 对于单边假设检验问题 $H_0 : \mu_1 \geq \mu_2$ ：

```
1 h <- t.test(x, y, alternative = "greater")
2 pValue <- h$p.value
```

- 对于单边假设检验问题 $H_0 : \mu_1 \leq \mu_2$ ：

```
1 h <- t.test(x, y, alternative = "less")
2 pValue <- h$p.value
```



回归模型

- 回归模型的一般形式为：

$$\begin{cases} Y = f(x | \theta) + e, \\ E(e) = 0, \quad D(e) = \sigma^2 \end{cases}$$

- 各个元素的定义：

- Y : 响应变量或预报变量
- x : 协变量或解释变量，维数可以大于 1
- θ : 模型参数，维数可以大于 1
- $f(\cdot | \theta)$: 回归函数
- e : 残差或模型误差
- $\sigma > 0$: 残差的标准差



最小二乘法估计

- 估计回归模型参数的目标是最小化样本点与回归曲线之间的距离平方和。
- 具体而言，模型参数 θ 的估计为：

$$\hat{\theta} = \arg \min_{\theta \in \Theta} Q(\theta)$$

- 其中：

$$Q(\theta) = \sum_{i=1}^n (y_i - f(x_i | \theta))^2$$

- $Q(\theta)$ 表示各个样本点与曲线 $y = f(x | \theta)$ 的整体距离。
- 回归曲线应该使得 $Q(\theta)$ 最小。
- $\hat{\theta}$ 被称为模型参数 θ 的最小二乘估计量。
- 这种估计模型参数的方法称为最小二乘法。



线性回归模型

- 当回归函数是参数的线性函数时，回归模型为：

$$\begin{cases} Y = \beta_0 + x\beta + e, \\ E(e) = 0, \quad D(e) = \sigma^2 \end{cases}$$

- 称此模型为线性回归模型。
- 线性回归模型的组成：
 - β_0 : 截距项或常数项
 - β : 列向量，表示各协变量对响应变量的影响程度



线性回归模型求解函数 lm

- 在 R 语言中，使用函数 ‘lm’ 计算线性回归模型中参数 β 的最小二乘估计。
- 函数 ‘lm’ 的结果是一个特殊结构的列表，称为 ‘lm’ 型数据。
- 该列表包含参数 β 的最小二乘估计的相关结果。



lm 函数的调用

- 可以简单地调用该函数：

```
1 S <- lm(y ~ x)
```

- 其中 ‘y’ 和 ‘x’ 分别是响应变量和协变量的观测数据。
- ‘y ~ x’ 指明回归函数的结构为 $y = a + bx$ 。
- ‘S’ 是 ‘lm’ 型变量，存放了模型参数估计的相关结果。
- ‘lm’ 型变量有多个分量，其中最常用的包括：
 - ‘coefficients’：模型参数的估计结果
 - ‘residuals’：残差估计结果



案例演示

- 下面通过一个简单的案例演示如何使用 ‘lm’ 函数。
- 示例代码：

```
1 # 示例数据
2 x <- c(1, 2, 3, 4, 5)
3 y <- c(2, 4, 5, 4, 5)
4
5 # 线性回归
6 S <- lm(y ~ x)
7
8 # 查看结果
9 summary(S)
```



线性回归模型结构的表达方式

- 在 R 语言中，线性回归模型的结构可以通过公式表达式来定义。
- 示例：在某些情况下，回归模型中没有截距项，响应变量 y 与协变量 x^2 的关系为线性。



无截距项的线性回归模型

- 回归函数结构的表达式为：

$$y \sim 0 + I(x^2)$$

- 其中：
 - '0' 表示模型中没有截距项。
 - 'I(x^2)' 表示协变量为 x^2 ，即回归函数为 $y = \beta x^2$ 。
 - 'I()' 函数用于指示 ' x^2 ' 是一个数学运算，而不是变量名。
- 也可以使用 '-1' 代替 '0' 来表示没有截距项。



R 代码示例

```
1 # 示例数据
2 x <- c(1, 2, 3, 4, 5)
3 y <- c(1, 4, 9, 16, 25)
4
5 # 无截距项的线性回归
6 model <- lm(y ~ 0 + I(x^2))
7
8 # 查看结果
9 summary(model)
```



提取模型参数和估计结果

- 在得到 ‘lm’ 型变量后，可以使用 R 函数 ‘coef’ 和 ‘predict’ 提取模型参数和响应变量的估计结果。
- 示例代码：

```
1 # 生成线性回归模型  
2 myReg <- lm(y ~ x)  
3  
4 # 提取参数估计结果  
5 coef(myReg) # 提取 lm 列表 myReg 中的参数估计结果  
6  
7 # 计算响应变量的估计结果  
8 predict(myReg) # 计算响应变量的估计结果
```

- 其中 ‘myReg’ 存储的是函数 ‘lm’ 的计算结果。



练习题 1：假设检验

题目：

假设你有一组数据表示某个城市中一小部分人的每日步数（单位：步）。你想知道这些人的平均每日步数是否显著大于 10000 步。请使用 R 语言进行单样本 t 检验来检验这一假设。

```
1 # 步数数据
2 steps <- c(9500, 10200, 9800, 11000, 12000, 9700,
3           10300, 10800, 11500, 9900)
4
5 # 使用显著性水平为 0.05
```



练习题 1：答案

答案：

```
1 # 单样本 t 检验
2 t_test_result <- t.test(steps, mu = 10000,
3                             alternative = "greater")
4
5 # 输出结果
6 t_test_result
```

解释：

- 使用 ‘t.test’ 函数进行单样本 t 检验。
- 参数 ‘mu = 10000’ 指定检验的均值假设。
- 参数 ‘alternative = "greater"' 用于检验平均步数是否大于 10000 步。
- 检验结果包括 t 值、自由度、p 值等信息。如果 p 值小于 0.05，可以拒绝原假设。



练习题 2：线性回归

题目：

你是一名数据分析师，正在研究某城市的房价与房屋面积之间的关系。你有以下数据：

```
1 # 房屋面积（平方米）
2 area <- c(50, 60, 80, 100, 120, 150)
3
4 # 房价（万元）
5 price <- c(50, 60, 80, 90, 110, 140)
```

使用 R 语言建立一个线性回归模型来预测房价，并回答以下问题：

1. 建立线性回归模型，并输出模型的系数。
2. 根据模型预测一个面积为 90 平方米的房子的价格。



练习题 2：答案

答案：

```
1 # 建立线性回归模型  
2 house_model <- lm(price ~ area)  
3  
4 # 输出模型系数  
5 coef(house_model)  
6  
7 # 预测面积为 90 平方米的房价  
8 predicted_price <- predict(house_model,  
9     newdata = data.frame(area = 90))  
10  
11 # 输出预测结果  
12 predicted_price
```



练习题 2：答案

解释：

- 使用 ‘lm’ 函数建立线性回归模型，‘price’ 作为响应变量，‘area’ 作为解释变量。
- ‘coefficients(housemodel)’ 返回模型的截距和斜率。
- 使用 ‘predict’ 函数预测面积为 90 平方米的房价。
- ‘newdata’ 参数用于传递新数据进行预测。

Thanks!