统计学导论：R语言实验11

假设检验

主讲人：郑盼盼

Outline

1. 假设检验基本原理
2. R语言实现

11.1 假设检验原理

11.1 假设检验原理

案例 如何检验（判断）一枚硬币是否均匀？设

$$X=\begin{cases}1, & \text{掷出正面}\\ 0, & \text{掷出反面}\end{cases}$$

则 $X \sim B(1,\theta)$

$$H_0:\ \theta = 0.5 \leftrightarrow H_1:\ \theta \neq 0.5$$

11.1 假设检验原理

案例 如何检验（判断）一枚硬币是否均匀？设

$$X = \begin{cases} 1, & \text{掷出正面} \\ 0, & \text{掷出反面} \end{cases}$$

则 $X \sim B(1, \theta)$

$$H_0 : \theta = 0.5 \leftrightarrow H_1 : \theta \neq 0.5$$

问题： 已获得 X 的重复观测样本 $X_1, X_2, \ldots, X_n$，如何根据观测值进行判断？

11.1 假设检验原理

根据观测样本的判断会出现以下情形：

跟据观测数据得出结论 \ 真实情况	H0 为真	H0 为假
拒绝 H0	第一类错误 α 拒真错误	结论正确
接受 H0	结论正确	第二类错误 β 受假错误

11.1 假设检验原理

真实情况

跟据观测数据得出结论	H0 为真	H0 为假
拒绝 H0	第一类错误 α 拒真错误	结论正确
接受 H0	结论正确	第二类错误 β 受假错误

若10次观测均为正面，推断硬币不均匀，则犯I类错误的概率为：

$$\begin{aligned}\mathbb{P}(X_1 = 1, \ldots, X_{10} = 1) &= 0.5^{10} \\ &\approx 0.0009765625\end{aligned}$$

11.1 假设检验原理

真实情况

跟据观测数据得出结论

	H0 为真	H0 为假
拒绝H0	**第一类错误 α 拒真错误**	结论正确
接受H0	结论正确	**第二类错误 β 受假错误**

若10次观测均为正面，推断硬币不均匀，则犯I类错误的概率为：

$$\mathbb{P}(X_1 = 1, \ldots, X_{10} = 1) = 0.5^{10} \approx 0.0009765625$$

当原假设成立时，$|\bar{X} - 0.5|$ 应该不大，若 $|\bar{X} - 0.5|$ 太大，不利于原假设，可以考虑拒绝原假设：

11.1 假设检验原理

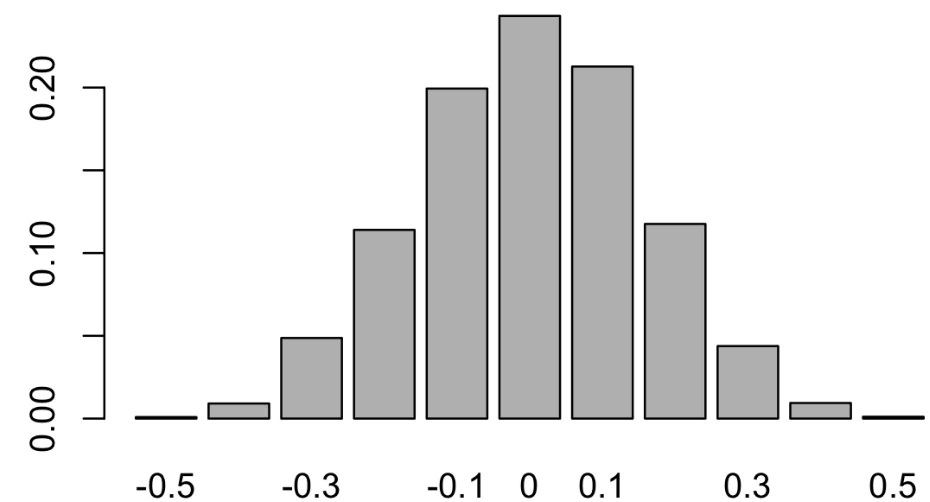
探究原假设成立的条件下，即 $X \sim B(1, 0.5)$，$\bar{X} - 0.5$ 的分布形态。

```r
n <- 10 # 样本量为10（进行10次观测）
m <- 10000 # 模拟试验次数
x <- matrix(rbinom(m*n, 1, 0.5), m, n) # 模拟H0成立时m次模拟试验的数据
xBar <- apply(x, 1, mean) # 对每次模拟试验计算样本均值
y <- as.factor(xBar - 0.5) # 计算 xBar - 0.5 并转换为因子向量
u <- table(y) # 获得y的频数统计表
v <- prop.table(u) # 计算频率统计表
barplot(v) # 绘制条形图
```

11.1 假设检验原理

探究原假设成立的条件下，即 $X \sim B(1, 0.5)$，$\bar{X} - 0.5$ 的分布形态。

```r
n <- 10 # 样本量为10（进行10次观测）
m <- 10000 # 模拟试验次数
x <- matrix(rbinom(m*n, 1, 0.5), m, n) # 模拟H0成立时m次模拟试验的数据
xBar <- apply(x, 1, mean) # 对每次模拟试验计算样本均值
y <- as.factor(xBar - 0.5) # 计算 xBar - 0.5 并转换为因子向量
u <- table(y) # 获得y的频数统计表
v <- prop.table(u) # 计算频率统计表
barplot(v) # 绘制条形图
```

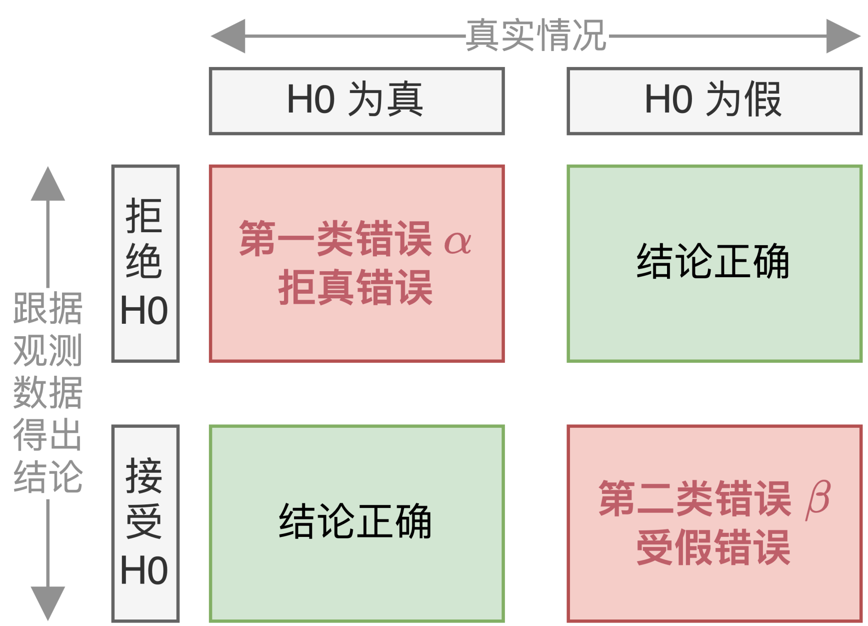


我们可以看到，当10次投掷结果均为正面(1)时，$\bar{x} - 0.5 = 0.5$，则概率 $\mathbb{P}_{\theta=0.5}(|\bar{X} - 0.5| > |\bar{x} - 0.5|)$ 十分小，因此拒绝原假设。

11.1 假设检验原理

假设检验的一般步骤：

1. 根据实际问题写出问题的原假设和备择假设
2. 给定显著性水平 α，通常为 $0.01, 0.05$ 等
3. 由样本构造统计量，计算原假设成立的条件下，**不利于原假设**的 p-值。
4. 若 p-值小于显著水平，则拒绝原假设，否则无理由拒绝原假设。



11.1 假设检验原理

例 已知某班的数学成绩服从 $N(\mu, 9)$，如何**用样本判断**“这个班级的平均成绩是否为 90 分”?

11.1 假设检验原理

例 已知某班的数学成绩服从 $N(\mu, 9)$，如何**用样本判断**“这个班级的平均成绩是否为 90 分”?

$$H_0: \mu = 90 \leftrightarrow H_1: \mu \neq 90$$

Tip

1. 根据题目写出原假设和备择假设

11.1 假设检验原理

例 已知某班的数学成绩服从 $N(\mu, 9)$，如何**用样本判断**“这个班级的平均成绩是否为 90 分”?

$$H_0: \mu = 90 \leftrightarrow H_1: \mu \neq 90$$

设定显著性水平 $\alpha = 0.05$。

> **Tip**
>
> 2. 设定显著性水平

11.1 假设检验原理

例 已知某班的数学成绩服从 $N(\mu, 9)$，如何**用样本判断**“这个班级的平均成绩是否为 90 分”?

$$H_0: \mu = 90 \leftrightarrow H_1: \mu \neq 90$$

设定显著性水平 $\alpha = 0.05$。设 $X_1, X_2, \dots, X_n$ 为重复观测样本，构造统计量 $|\bar{X} - 90|$。计算原假设成立的条件下，$|\bar{X} - 90| > |\bar{x} - 90|$ 的概率，即 p-值

$$\mathbb{P}_{H_0}(|\bar{X} - 90| > |\bar{x} - 90|)$$

若计算得到样本均值为 85，则变为计算

$$\mathbb{P}_{H_0}(|\bar{X} - 90| > 5)$$

Tip

3. 根据题目构造统计量，并计算 p-值

11.1 假设检验原理

例 已知某班的数学成绩服从 $N(\mu, 9)$，如何**用样本判断**“这个班级的平均成绩是否为 90 分”?

根据中心极限定理，当 H_0 成立的条件下，即 $\mu = 90$，样本均值 $\bar{X}$ 服从正态分布：

$$\bar{X} \sim N\left(90, \frac{9}{n}\right)$$

于是p-值为

$$\mathbb{P}_{H_0}(|\bar{X} - 90| > 5) = 2\mathbb{P}_{H_0}(\bar{X} - 90 < -5) = 2\Phi\left(-\frac{5}{3/\sqrt{n}}\right)$$

Tip

3. 根据题目构造统计量，并计算 p-值

11.1 假设检验原理

例 已知某班的数学成绩服从 $N(\mu, 9)$，如何**用样本判断**“这个班级的平均成绩是否为 90 分”?

根据中心极限定理，当 H_0 成立的条件下，即 $\mu = 90$，样本均值 $\bar{X}$ 服从正态分布：

$$\bar{X} \sim N\left(90, \frac{9}{n}\right)$$

于是p-值为

$$\mathbb{P}_{H_0}(|\bar{X} - 90| > 5) = 2\mathbb{P}_{H_0}(\bar{X} - 90 < -5) = 2\Phi\left(-\frac{5}{3/\sqrt{n}}\right)$$

比较上式所得p-值和α

Tip

4. 比较p-值和α

11.2 假设检验的R语言实现

11.2.1 总体方差已知-Z检验

例 某种子公司在销售胡萝卜种子的说明书中声称：用此种胡萝卜的平均长度为 $11.5\,\mathrm{cm}$。某人种植这种胡萝卜后得到的胡萝卜长度数据见下表，若胡萝卜长度的标准差为 $1.15\,\mathrm{cm}$，问：在显著性水平 0.05 下，可接受该种子公司关于胡萝卜平均长度的说明吗?

表 5-10 40 根胡萝卜长度的数据 (单位: cm)

11.50	10.08	12.14	12.33	10.68	13.37	13.37
11.79	12.83	11.32	14.51	11.84	12.13	13.23
12.34	10.46	12.82	13.87	11.20	12.99	13.44
11.54	12.79	12.94	12.82	13.48	12.77	13.37
11.96	12.38	12.20	12.07	11.89	11.04	10.17
10.34	12.66	10.62	11.98	11.82		

11.2.1 总体方差已知-Z检验

例 某种子公司在销售胡萝卜种子的说明书中声称：用此种胡萝卜的平均长度为 $11.5\,\mathrm{cm}$。某人种植这种胡萝卜后得到的胡萝卜长度数据见下表，若胡萝卜长度的标准差为 $1.15\,\mathrm{cm}$，问：在显著性水平 0.05 下，可接受该种子公司关于胡萝卜平均长度的说明吗?

答 根据题意可知，萝卜长度 $X \sim N(\mu, 1.15^2)$。因此，原假设和备择假设为：

$$H_0: \mu = 11.5 \leftrightarrow H_1: \mu \neq 11.5$$

11.2.1 总体方差已知-Z检验

例 某种子公司在销售胡萝卜种子的说明书中声称：用此种胡萝卜的平均长度为 $11.5\,\mathrm{cm}$。某人种植这种胡萝卜后得到的胡萝卜长度数据见下表，若胡萝卜长度的标准差为 $1.15\,\mathrm{cm}$，问：在显著性水平 0.05 下，可接受该种子公司关于胡萝卜平均长度的说明吗?

答 根据题意可知，萝卜长度 $X \sim N(\mu, 1.15^2)$。因此，原假设和备择假设为：

$$H_0: \mu = 11.5 \leftrightarrow H_1: \mu \neq 11.5$$

用样本均值 $\bar{X}$ 来判断原假设是否正确，根据大数定律和题意（此处样本量为 40)，若原假设成立：

$$\bar{X} \sim N\left(11.5, \frac{1.15^2}{40}\right)$$

11.2.1 总体方差已知-Z检验

例 某种子公司在销售胡萝卜种子的说明书中声称：用此种胡萝卜的平均长度为 $11.5\,\mathrm{cm}$。某人种植这种胡萝卜后得到的胡萝卜长度数据见下表，若胡萝卜长度的标准差为 $1.15\,\mathrm{cm}$，问：在显著性水平 0.05 下，可接受该种子公司关于胡萝卜平均长度的说明吗?

因此，我们可以计算在 H_0 成立的条件下的p-值

$$\mathbb{P}(|\bar{X}-11.5| > |\bar{x}-11.5|) = 2\Phi\left(-\frac{|\bar{x}-11.5|}{1.15/\sqrt{40}}\right)$$

11.2.1 总体方差已知-Z检验

例 某种子公司在销售胡萝卜种子的说明书中声称：用此种胡萝卜的平均长度为 $11.5\,\mathrm{cm}$。某人种植这种胡萝卜后得到的胡萝卜长度数据见下表，若胡萝卜长度的标准差为 $1.15\,\mathrm{cm}$，问：在显著性水平 0.05 下，可接受该种子公司关于胡萝卜平均长度的说明吗?

因此，我们可以计算在 H_0 成立的条件下的p-值

$$\mathbb{P}(|\bar{X} - 11.5| > |\bar{x} - 11.5|) = 2\Phi\left(-\frac{|\bar{x} - 11.5|}{1.15/\sqrt{40}}\right)$$

可以利用R语言计算p-值：

```r
x <- c(11.50,10.08,12.14,12.33,10.68,13.37,13.37,
       11.79,12.83,11.32,14.51,11.84,12.13,13.23,
       12.34,10.46,12.82,13.87,11.20,12.99,13.44,
       11.54,12.79,12.94,12.82,13.48,12.77,13.37,
       11.96,12.38,12.20,12.07,11.89,11.04,10.17,
       10.34,12.66,10.62,11.98,11.82)
n <- length(x)
sigma <- 1.15
mu <- 11.5
p <- 2 * pnorm(-abs(mean(x) - 11.5)/(1.15/sqrt(n)))
p
```

11.2.2 总体方差未知-t检验

例 某种子公司在销售胡萝卜种子的说明书中声称：用此种胡萝卜的平均长度为 $11.5\,\mathrm{cm}$。某人种植这种胡萝卜后得到的胡萝卜长度数据见下表，~~若胡萝卜长度的标准差为 1.15 cm~~，问：在显著性水平 0.05 下，可接受该种子公司关于胡萝卜平均长度的说明吗?

答 根据题意可知，萝卜长度 $X \sim N(\mu, \sigma^2)$。因此，原假设和备择假设为：

$$H_0: \mu = 11.5 \leftrightarrow H_1: \mu \neq 11.5$$

同样，我们可以利用样本均值来计算p-值：

$$\mathbb{P}(|\bar{X} - 11.5| > |\bar{x} - 11.5|)$$

但此时，总体方差未知，我们通过构造统计量T来衡量p-值：

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 11.5}{S/\sqrt{40}} \sim t(n-1)$$

11.2.2 总体方差未知-t检验

例 某种子公司在销售胡萝卜种子的说明书中声称：用此种胡萝卜的平均长度为 $11.5\,\mathrm{cm}$。某人种植这种胡萝卜后得到的胡萝卜长度数据见下表，~~若胡萝卜长度的标准差为 1.15 cm~~，问：在显著性水平 0.05 下，可接受该种子公司关于胡萝卜平均长度的说明吗?

因此，我们可以利用 T 计算p-值

$$\mathbb{P}\left(\left|\frac{\sqrt{40}(\bar{X}-11.5)}{S}\right| > \left|\frac{\sqrt{40}(\bar{x}-11.5)}{S}\right|\right)$$

可以利用R语言计算p-值

```r
# H0: mu == mu0
h <- t.test(x, mu=11.5, alternative = "two.side")
h$p.value
```

11.2.2 总体方差未知-t检验

存在三种原假设：

$$
\begin{aligned}
H_0:\ \mu = \mu_0 \quad (1)\\
H_0:\ \mu \le \mu_0 \quad (2)\\
H_0:\ \mu \ge \mu_0 \quad (3)
\end{aligned}
$$

对于三种原假设，我们都可以通过构造统计量T，来计算p-值

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

其p-值分别对应于：

$$
\begin{aligned}
&(1) \quad \mathbb{P}\left(\frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} > \frac{|\bar{x} - \mu_0|}{S/\sqrt{n}}\right)\\
&(2) \quad \mathbb{P}\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > \frac{\bar{x} - \mu_0}{S/\sqrt{n}}\right)\\
&(3) \quad \mathbb{P}\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < \frac{\bar{x} - \mu_0}{S/\sqrt{n}}\right)
\end{aligned}
$$

11.2.2 总体方差未知-t检验

例 还是上面萝卜长度的例子，现在我们关心该品种的胡罗卜平均长度是否大于 $11.5\ \mathrm{cm}$ 问题。

$$H_0:\ \mu \geq 11.5$$

可以使用如下代码计算p-值

```
# H0: mu >= mu0
h <- t.test(x, mu=11.5, alternative = "less")
h$p.value
```

11.2.2 总体方差未知-t检验

例 还是上面萝卜长度的例子，现在我们关心该品种的胡罗卜平均长度是否小于 $11.5\,\mathrm{cm}$ 问题。

$$H_0:\ \mu \leq 11.5$$

可以使用如下代码计算p-值

```r
# H0: mu <= mu0
h <- t.test(x, mu=11.5, alternative = "greater")
h$p.value
```

11.2.3 双正态总体均值的检验

例 下表分别来自甲班和乙班学生的数学成绩样本数据，试对比分析两个班级学生的平均数学成绩

表 5-11　甲班学生的数学成绩数据　(单位：分)

45	54	41	45	80	60	49	44	55	87	73	64	35	89	85	55	84	80	67	70
84	76	51	56	83	83	45	58	46	65	0	76	79	75	86	62	43	57	75	47

表 5-12　乙班学生的数学成绩数据　(单位：分)

45	46	78	45	47	83	67	36	80	55	45	0	75	68	53	85	44	51	67	64
73	62	65	72	84	75	58	56	55	38	83	73	67	72	76	67	83	43	57	60

11.2.3 双正态总体均值的检验

例 下表分别来自甲班和乙班学生的数学成绩样本数据，试对比分析两个班级学生的平均数学成绩

1. 原假设：甲乙两班平均成绩无明显差异，即

$$H_0: \mu_1 = \mu_2$$

11.2.3 双正态总体均值的检验

例 下表分别来自甲班和乙班学生的数学成绩样本数据，试对比分析两个班级学生的平均数学成绩

1. 原假设：甲乙两班平均成绩无明显差异，即

$$H_0: \mu_1 = \mu_2$$

```r
x <- c(45,54,41,45,80,60,49,44,55,87,73,64,35,89,85,55,84,80,67,70,
       84,76,51,56,83,83,45,58,46,65,0,76,79,75,86,62,43,57,75,47) # 甲班成绩
y <- c(45,46,78,45,47,83,67,36,80,55,45,0,75,68,53,85,44,51,67,64,
       73,62,65,72,84,75,58,56,55,38,83,73,67,72,76,67,83,43,57,60) # 乙班成绩
h <- t.test(x,y, alternative = "two.side")
h$p.value
```

11.2.3 双正态总体均值的检验

例 下表分别来自甲班和乙班学生的数学成绩样本数据，试对比分析两个班级学生的平均数学成绩

1. 原假设：甲乙两班平均成绩无明显差异
2. 原假设：甲班成绩优于乙班，即

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

11.2.3 双正态总体均值的检验

例 下表分别来自甲班和乙班学生的数学成绩样本数据，试对比分析两个班级学生的平均数学成绩

1. 原假设：甲乙两班平均成绩无明显差异
2. 原假设：甲班成绩优于乙班，即

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

```r
x <- c(45,54,41,45,80,60,49,44,55,87,73,64,35,89,85,55,84,80,67,70,
       84,76,51,56,83,83,45,58,46,65,0,76,79,75,86,62,43,57,75,47) # 甲班成绩
y <- c(45,46,78,45,47,83,67,36,80,55,45,0,75,68,53,85,44,51,67,64,
       73,62,65,72,84,75,58,56,55,38,83,73,67,72,76,67,83,43,57,60) # 乙班成绩
h <- t.test(x,y, alternative = "less")
h$p.value
```

11.2.3 双正态总体均值的检验

例 下表分别来自甲班和乙班学生的数学成绩样本数据，试对比分析两个班级学生的平均数学成绩

1. 原假设：甲乙两班平均成绩无明显差异
2. 原假设：甲班成绩优于乙班
3. 原假设：甲班成绩差于乙班，即

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

11.2.3 双正态总体均值的检验

例 下表分别来自甲班和乙班学生的数学成绩样本数据，试对比分析两个班级学生的平均数学成绩

1. 原假设：甲乙两班平均成绩无明显差异
2. 原假设：甲班成绩优于乙班
3. 原假设：甲班成绩差于乙班，即

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

```r
x <- c(45,54,41,45,80,60,49,44,55,87,73,64,35,89,85,55,84,80,67,70,
       84,76,51,56,83,83,45,58,46,65,0,76,79,75,86,62,43,57,75,47) # 甲班成绩
y <- c(45,46,78,45,47,83,67,36,80,55,45,0,75,68,53,85,44,51,67,64,
       73,62,65,72,84,75,58,56,55,38,83,73,67,72,76,67,83,43,57,60) # 乙班成绩
h <- t.test(x,y, alternative = "greater")
h$p.value
```