Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

07 CHAPTER

估计(Estimation)

1

第7章 估计(Estimation)

1. 统计性估计
2. 关于总体均值的估计
3. 关于总体比例的估计
4. 关于总体方差的估计
5. 置信区间的理解

2/22

2

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

1. 统计性估计

[定义 10-1] 点估计(point estimation)

：通过估计量的观测值来估计参数的真值的过程

- 无偏性(unbiasedness) $E(\hat{\theta})=\theta$ $E(\bar{X})=\mu,\ E(S^2)=\sigma^2$

[例 10-1] 从均值为μ的总体中抽取样本，求样本均值时，可知样本均值是关于μ的无偏估计量。

$$E(\bar{X})=\frac{1}{n}E(X_1+X_2+\cdots+X_n)=\frac{1}{n}n\mu=\mu$$

[例 10-2] 从总体方差为σ^2的总体中抽取n个样本，求证样本方差S^2是总体方差σ^2的无偏估计量。

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)\ \Rightarrow E\left[\frac{(n-1)S^2}{\sigma^2}\right]=n-1\ \Rightarrow E(S^2)=\sigma^2$$

3/22

3

1.1 点估计

[例 10-3] 从均值为μ, 方差为σ^2的总体中抽取n个样本，在求样本均值X-bar和其他估计量$Y=(X_1+X_2)/2$时，探讨两估计量的无偏性，并对其方差进行比较。

$$E(\bar{X})=\mu,\ Var(\bar{X})=\sigma^2/n$$

$$E(\bar{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

$$Var(\bar{X})=Var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}Var(X_i)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{\sigma^2}{n}$$

$$E(Y)=E\left(\frac{X_1+X_2}{2}\right)=\frac{2\mu}{2}=\mu$$

$$Var(Y)=Var\left(\frac{X_1+X_2}{2}\right)=\frac{2\sigma^2}{4}=\frac{\sigma^2}{2}>\frac{\sigma^2}{n}\quad(n\geq 3)$$

4/22

4

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

1.2 区间估计

[定义 10-2] 区间估计(interval estimation)
：含参数真值的概率决定置信水平为1-α的置信区间的过程

- 置信区间(confidence interval) $\left[\hat{\theta}_L, \hat{\theta}_U\right]$
  含参数真值的概率成为置信水平为1-α的区间
- 置信水平(confidence level)
  置信区间包含参数真值的概率(1-α)
  $P\left(\hat{\theta}_L < \theta < \hat{\theta}_U\right) = 1-\alpha$
- 显著性水平(significance level)
  置信区间不包含参数真值的概率(α)
  $1-P\left(\hat{\theta}_L < \theta < \hat{\theta}_U\right) = \alpha$

5/22

5

2. 关于总体均值的估计

[定理 10-1] 总体均值的置信区间(已知总体方差的情况)
：关于正态总体的总体均值μ的置信水平(1-α)置信区间(confidence interval), 100(1-α)% 置信区间

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

[证明]

$$1-\alpha = P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = P\left(-z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right)$$

$$= P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

6/22

6

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

2.1 总体方差已知情况下的估计

[例 10-4] 一个巧克力的重量符合总体标准差为5(g)的正态分布，随机抽取50个巧克力，重量的样本均值为199.5(g)时，求95%置信区间。

$$\bar{X}=199.5,\quad z_{0.975}\doteq 1.96$$

$$\left[199.5\pm 1.96\times\frac{5}{\sqrt{50}}\right]\doteq[199.5\pm 1.386]=[198.114, 200.886]$$

- 置信区间的误差(error) $z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
- 保持误差在一定水平以下所需样本个数

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\le\delta \Rightarrow n\ge\left(z_{1-\alpha/2}\frac{\sigma}{\delta}\right)^2$$

[例 10-5] 在上述[例 10-4]的95%置信水平中，若要保持置信区间的误差小于等于1.2，求所需的样本大小。

$$n\ge\left(1.96\times\frac{5}{1.2}\right)^2\doteq 66.7 \Rightarrow n\ge 67$$

7/22

7

2.3 总体方差未知情况下的估计

[定理 10-3] 关于总体均值的置信区间(总体方差未知的情况)
：在不知总体方差的情况下，关于正态总体均值的100(1-α)% 置信区间 (confidence interval)

$$\left[\bar{X}-t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}},\ \bar{X}+t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right]=\left[\bar{X}\pm t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right]$$

[证明]

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1) \Rightarrow P\left(-t_{1-\alpha/2;(n-1)}<\frac{\bar{X}-\mu}{S/\sqrt{n}}<t_{1-\alpha/2;(n-1)}\right)=1-\alpha$$

$$P\left(\bar{X}-t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}<\mu<\bar{X}+t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right)=1-\alpha$$

$$\left[\bar{X}-t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}},\bar{X}+t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right]=\left[\bar{X}\pm t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right]$$

8/22

8

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

2.3 总体方差未知情况下的估计

[例 10-8] 一个巧克力的重量符合正态分布，从中随机抽取16个巧克力，其重量的样本均值为199.5(g), 样本方差为25.0时，求关于总体均值的95%置信区间。

$$\bar{X}=199.5,\ S=5.0,\ t_{0.975;15}\doteq 2.131$$

$$\left[199.5\pm 2.131\frac{5}{\sqrt{16}}\right]\doteq[199.5\pm 2.664]=[196.836,202.164]$$

```
# 定义函数(样本个数、样本均值、样本标准差、显著性水平)
tci1 <- function(n, xb, s, alp) {
        err <- qt(1-alp/2, n-1)*s/sqrt(n)
        cat("[", xb-err, ",", xb+err,"]₩n")}
# 运行函数
tci1(16, 199.5, 5, 0.05)
[ 196.8357 , 202.1643 ]
```

9/22

9

2.3 总体方差未知情况下的估计

[例 10-9] 一个巧克力的重量符合正态分布，从中随机抽取50个巧克力，对其重量测量的结果如下所示，求关于总体均值的95%置信区间。

194	200	207	199	206	191	195	197	210	202
206	198	196	213	196	196	200	198	188	202
203	203	204	198	204	201	204	199	197	197
209	194	197	203	202	198	197	201	194	193
195	201	194	198	195	201	196	199	204	198

$$\sum_{i=1}^{50} x_i = 9,973 \qquad \sum_{i=1}^{50} x_i^2 = 1,990,407$$

$$\bar{X}=\frac{9,973}{50}=199.46 \qquad S=\sqrt{\frac{1,990,407-9,973^2/50}{50-1}}\doteq 4.993$$

$$t_{0.975;49}\doteq 2.010$$

$$\left[199.46\pm 2.010\times\frac{4.993}{\sqrt{50}}\right]\doteq[199.46\pm 1.402]=[198.058,200.862]$$

10/22

10

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3. 关于总体比例的估计

[定理 10-5] 总体比例的近似置信区间

：当样本大小n取值较大时，关于总体比例p的100(1-α)%近似置信区间

$$\left[\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \qquad \hat{p} = \frac{X}{n}$$

$$\frac{X-np}{\sqrt{np(1-p)}} = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \overset{a}{\sim} N(0,1)$$

$$1-\alpha = P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) \simeq P\left(-z_{1-\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{1-\alpha/2}\right)$$

$$\simeq P\left(-z_{1-\alpha/2} < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{1-\alpha/2}\right)$$

$$\Rightarrow P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \simeq 1-\alpha$$

11/22

11

3.1 总体比例的估计

[例 10-13] 在某一工艺流程中随机抽取200个产品进行检查，发现了15个不良品，求该关于该工艺流程不良率的95%置信区间。

$$\hat{p} = \frac{X}{n} = \frac{15}{200} = 0.075 \qquad z_{0.975} \doteq 1.96$$

$$\left[0.075 \pm 1.96 \times \sqrt{\frac{0.075(1-0.075)}{200}}\right] \doteq [0.075 \pm 0.037] = [0.038, 0.112]$$

```
# 定义正态近似置信区间函数(样本个数、成功次数、显著性水平)
pci1 <- function(n, x, alp) {
        p <- x/n
        err <- qnorm(1-alp/2)*sqrt(p*(1-p)/n)
        cat("[", p, "±", err, "] = [", p-err, ",", p+err,"]₩n")}
# 运行函数
pci1(200, 15, 0.05)
[ 0.075 ± 0.03650351 ] = [ 0.03849649 , 0.1115035 ]
# 利用二项分布估计正确的置信区间(多少会出现差异)
binom.test(15,200)$conf
[1] 0.0425828 0.1206842
attr(,"conf.level")
[1] 0.95
```

12/22

12

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

3.1 总体比例的估计

- 为使置信区间误差保持在一定水平一下，所需的样本个数

$$z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \delta \Rightarrow \quad n \ge \hat{p}(1-\hat{p})\left(\frac{z_{1-\alpha/2}}{\delta}\right)^2$$

$$p(1-p) = -(p-0.5)^2 + 0.25 \Rightarrow p^* = 0.5 \quad \Rightarrow n \ge 0.5^2 \times \left(\frac{z_{1-\alpha/2}}{\delta}\right)^2$$

[例 10-14] 根据过去的经验，某一工艺流程的不良率为7.5% 95%置信区间，若要使关于总体比例的置信区间的误差达到0.02一下，则所需的样本大小是多少。

$$p = 0.075 \quad \Rightarrow n \ge 0.075 \times (1-0.075)\left(\frac{1.96}{0.02}\right)^2 \doteq 666.3 \quad \Rightarrow n \ge 667$$

- 在完全不知总体比例p的情况下

$$\Rightarrow n \ge 0.5^2 \times \left(\frac{1.96}{0.02}\right)^2 \doteq 2400.9 \Rightarrow n \ge 2401$$

13/22

13

3.1 总体比例的估计

[例 10-14] (承上)

```
# 定义求样本个数的函数(误差限制、显著性水平、比例估计值)
nsample <- function(err, alp=0.05, ph=0.5) {
        n <- qnorm(1-alp/2)^2 * ph*(1-ph) /err^2
        cat("n ≥", n, "⇒ n ≥", ceiling(n), "₩n")}
# 运行函数
nsample(0.02, ph=15/200)
n ≥ 666.253 ⇒ n ≥ 667
nsample(0.02)
n ≥ 2400.912 ⇒ n ≥ 2401
```

14/22

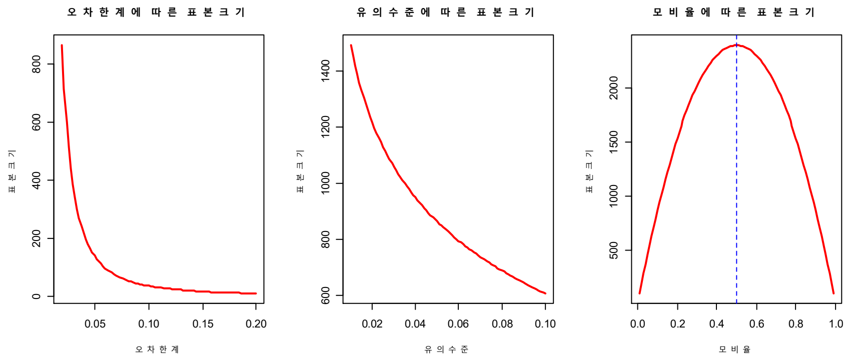
14

[例 10-15] 请绘制在下列情况下为维持关于总体比例的置信区间误差限制，所需的样本大小图表。

(1) 当p=0.1, α=0.05时，误差限制δ=0.02 ~ 0.2的变化情况

(2) 当p=0.1, δ=0.02时，显著性水平α=0.01 ~ 0.1的变化情况

(3) 当α=0.05, δ=0.02时，总体比例p=0.01 ~ 0.99的变化情况



15/22

15

4. 关于总体方差的估计

[定理 10-8] 关于总体方差的置信区间

：关于正态总体方差σ^2的100(1-α)%置信区间

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}\right]$$

- 卡方分布的分位数

$$P(\chi^2_\nu \le \chi^2_{p;\nu}) \equiv p \quad \Rightarrow P(\chi^2_{\alpha/2;\nu} \le \chi^2_\nu \le \chi^2_{1-\alpha/2;\nu}) = 1-\alpha$$

$$\Rightarrow P(\chi^2_{\alpha/2;n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{1-\alpha/2;n-1}) = 1-\alpha$$

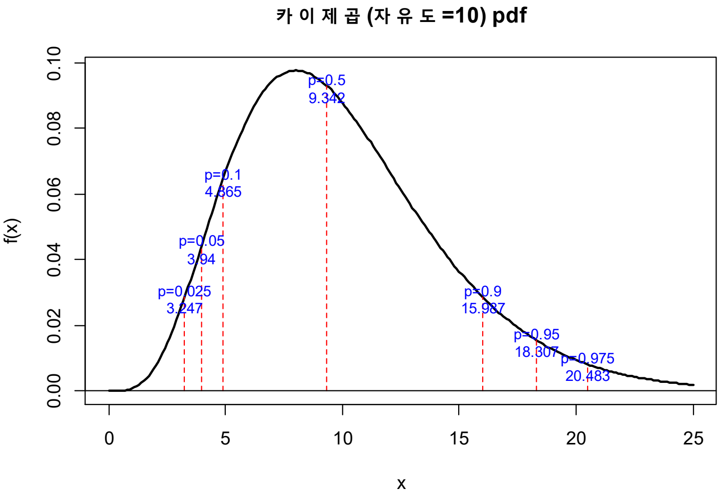
$$= P(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}})$$

16/22

16

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

[例 10-18] 请绘制自由度为10的卡方分布的概率密度函数，并在其上标示累积概率p=0.025, 0.05, 0.1, 0.5, 0.9, 0.95, 0.975所对应的分位数。



17/22

17

4.1 总体方差的估计

[例 10-19] 某一生产工艺的品质特征值符合正态分布，为估计关于品质特征的方差，随机抽取了10个产品进行检查，所得结果如下所示。求总体方差的95%置信区间。

20.0 21.5 20.9 19.8 22.5 20.3 23.6 18.0 23.3 17.8

$$S^2 = \frac{4350.13 - 207.7^2/10}{9} = \frac{36.201}{9} \doteq 4.022$$

$$\chi^2_{0.025;9} \doteq 2.700 \qquad \chi^2_{0.975;9} \doteq 19.023$$

$$\Rightarrow \left[\frac{36.201}{19.023}, \frac{36.201}{2.700}\right] \doteq [1.903, 13.408]$$

18/22

18

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

5. 置信区间的理解

- 对置信区间的误解

在置信区间[a, b]，包含参数θ的概率为100(1-α)%。

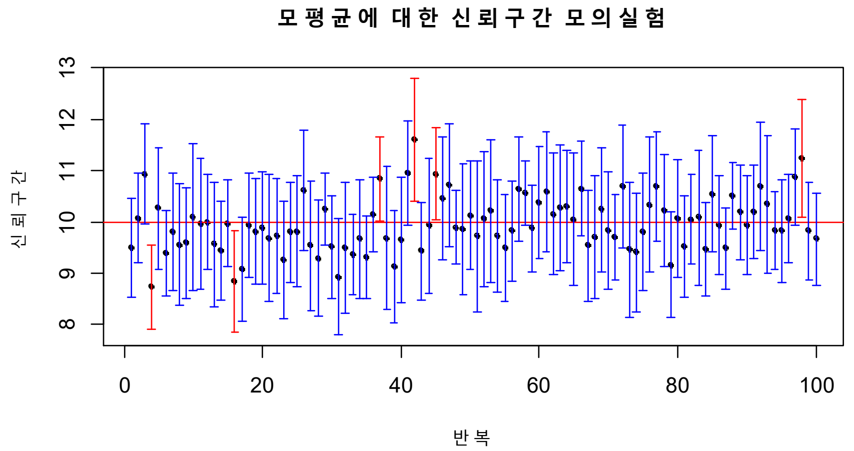
- 对置信区间的正确解析

任意置信区间[L, U]，包含参数θ的真值的概率为100(1-α)%，其置信区间中区间[a, b]为观测值。

19/22

19

[例 10-22] 从均值为10, 方差为4的正态总体中，随机抽取16个样本，求关于总体均值的95%置信区间，并绘图表示。重复取样100次，求不含总体均值10的置信区间出现的次数。

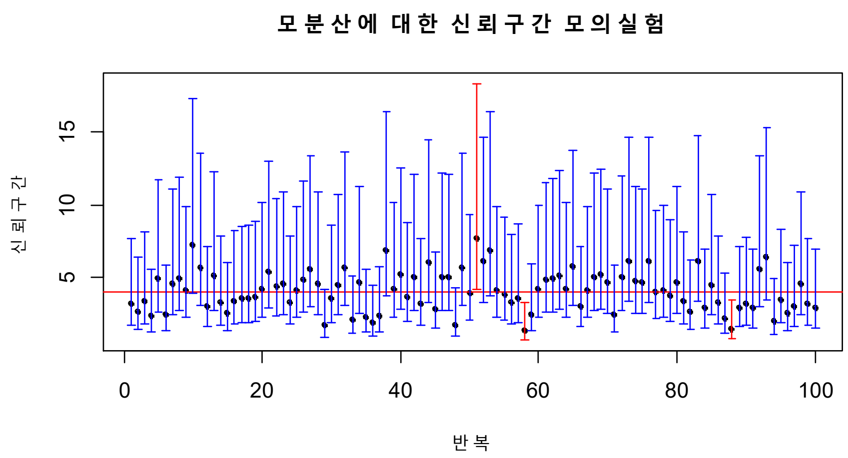


20/22

20

Kim, Dae Ho (Mokwon University)
mis@mokwon.ac.kr / mis@acm.org

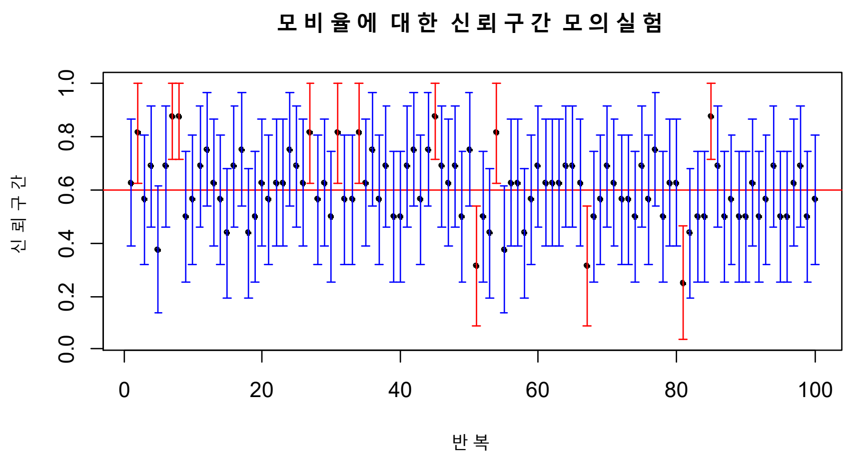
[例 10-23] 从均值为10，方差为4的正态总体中，随机抽取16个样本，求关于总体均值的95%置信区间，并绘图表示。重复取样100次，求不含总体均值4的置信区间出现的次数。



21/22

21

[例 10-24] 从总体比例为0.6的总体中，随机抽取16个样本，求关于总体均值的95%置信区间，并绘图表示。重复取样100次，求不含总体比例0.6的置信区间出现的次数。



22/22

22