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

06 CHAPTER

样本的分布

1

第6章
样本的分布

1. 统计量与估计量
2. 正态总体样本均值的分布
3. 正态总体样本方差的分布
4. 两正态总体样本方差比的分布
5. 中心极限定理
6. 非正态总体的样本分布

2/28

2

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

1. 统计量与估计量

[定义 9-1] 随机样本(random sample)
独立且具有相同分布的(iid: independent and identically distributed)随机变量的集合

[定义 9-2] 统计量(statistic)
不含未知(unknown)参数的随机样本的函数

[定义 9-3] 估计量(estimator)
用于估计未知参数的统计量

- 样本均值 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

3/28

3

1. 统计量与估计量

[定义 9-4] 无偏性(unbiasedness) $E(\hat{\theta}) = \theta$
：与用于计算估计量期望的参数具有相同特性，且是成为好的估计量的第一个条件

- 样本均值的无偏性

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

- 样本均值的方差

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

✓ 在正态总体中抽取一定数量的随机样本作为总体均值，对此的无偏估计量中样本均值的方差最小(UMVUE)。

4/28

4

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

2. 正态总体样本均值的分布

[定理 9-1] 正态总体样本均值的分布

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \ \Rightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$

[证明] $$Y = \sum_{i=1}^{n} X_i \sim N(n\mu,\ n\sigma^2) \ \Rightarrow \bar{X} = \frac{1}{n} Y \sim N(\mu, \frac{\sigma^2}{n})$$

[推论 9-1] 正态总体标准化样本均值的分布

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \ \Rightarrow \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

[证明] $$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \ \Rightarrow \ E(Z) = 0, Var(Z) = 1$$

5/28

5

2. 正态总体样本均值的分布

[例 9-1] 从均值为100, 方差为100的正态分布总体中每次随机抽取10个样本，为求样本均值重复取样10,000次

(1) 样本均值直方图, 检验[定理 9-1]

$$\Rightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(100, \frac{100}{10})$$

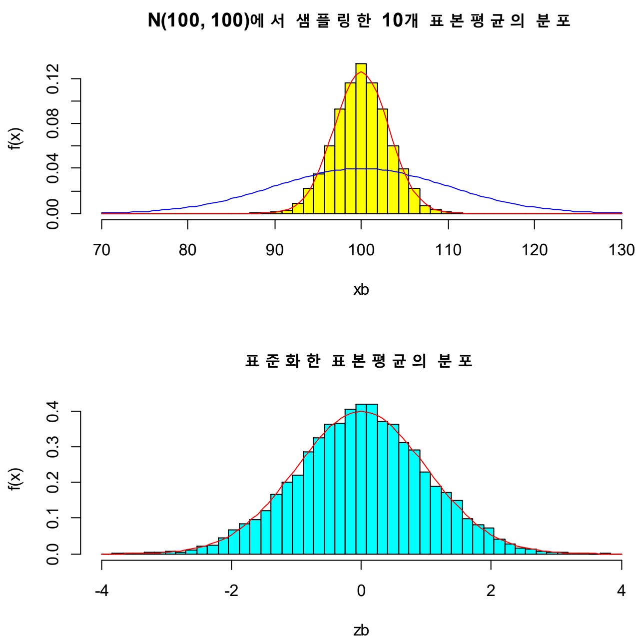
(2) 样本均值标准化直方图，检验[推论 9-1]

$$\Rightarrow \frac{\bar{X} - 100}{10 / \sqrt{10}} \sim N(0,1)$$

6/28

6

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



7/28

7

2. 正态总体样本均值的分布

[例 9-2] 一个巧克力的重量符合总体均值为μ(g), 总体标准差为5(g)的正态分布。随机抽取16个巧克力，对其重量进行测量时，样本均值与总体均值之差在2(g)以内的概率是多少?

$$X \sim N(\mu,\ 5^2) \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{5/\sqrt{16}} \sim N(0,1)$$

$$P(|\bar{X}-\mu| \le 2) = P(-2 \le \bar{X}-\mu \le 2)$$

$$= P\left(-\frac{2}{5/\sqrt{16}} \le \frac{\bar{X}-\mu}{5/\sqrt{16}} \le \frac{2}{5/\sqrt{16}}\right)$$

$$= P(-1.6 \le Z \le 1.6) \doteq 0.8904$$

```
n <- 16
# 方法1 (直接计算), mu=200 (输入任意数)
pnorm(200+2, 200, 5/sqrt(n)) - pnorm(200-2, 200, 5/sqrt(n))
[1] 0.8904014
# 方法2 (标准化)
pnorm(2/(5/sqrt(n))) - pnorm(-2/(5/sqrt(n)))
[1] 0.8904014
```

8/28

8

2. 正态总体样本均值的分布

[例 9-3] 若要想[例 9-2]中样本均值与总体均值之差在2(g)以内的概率大于等于0.95的话，最少需要几个样本?

$$P(|\bar{X}-\mu|\le 2)=P(-2\le \bar{X}-\mu \le 2)\ge 0.95$$

$$P\left(-\frac{2}{5/\sqrt{n}}\le \frac{\bar{X}-\mu}{5/\sqrt{n}}\le \frac{2}{5/\sqrt{n}}\right)=P(-0.4\sqrt{n}\le Z\le 0.4\sqrt{n})$$

$$=\Phi(0.4\sqrt{n})-\Phi(-0.4\sqrt{n})=2\Phi(0.4\sqrt{n})-1\ge 0.95$$

$$\Rightarrow \Phi(0.4\sqrt{n})\ge 0.975 \ \Rightarrow 0.4\sqrt{n}\ge \Phi^{-1}(0.975)\equiv z_{0.975}\doteq 1.96$$

$$\Rightarrow n\ge (1.96/0.4)^2\ge 24.01 \ \Rightarrow n\ge 25$$

```
# 直接计算
ceiling((qnorm(0.975)/0.4)^2)
[1] 25
# 简单定义函数(p=置信区间, sigma=误差限制(标准差的倍数))
findsn <- function(p, sigma) ceiling((qnorm(1-(1-p)/2)/sigma)^2)
findsn(0.95, 0.4); findsn(0.99, 0.4); findsn(0.95, 0.2); findsn(0.99, 0.2)
[1] 25    [1] 42    [1] 97    [1] 166
```

9/28

9

2. 正态总体样本均值的分布

[例 9-4] 从正态总体中抽取的样本均值与总体均值之差在kσ以内的概率大于等于1-α，求此时最小样本个数，并绘制图表对其变化进行分析。

$$P(|\bar{X}-\mu|\le k\sigma)=P(-k\sigma\le \bar{X}-\mu \le k\sigma)\ge 1-\alpha$$

$$P\left(-\frac{k\sigma}{\sigma/\sqrt{n}}\le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\le \frac{k\sigma}{\sigma/\sqrt{n}}\right)=P(-k\sqrt{n}\le Z\le k\sqrt{n})$$

$$=\Phi(k\sqrt{n})-\Phi(-k\sqrt{n})=2\Phi(k\sqrt{n})-1\ge 1-\alpha$$

$$\Rightarrow \Phi(k\sqrt{n})\ge 1-\alpha/2$$

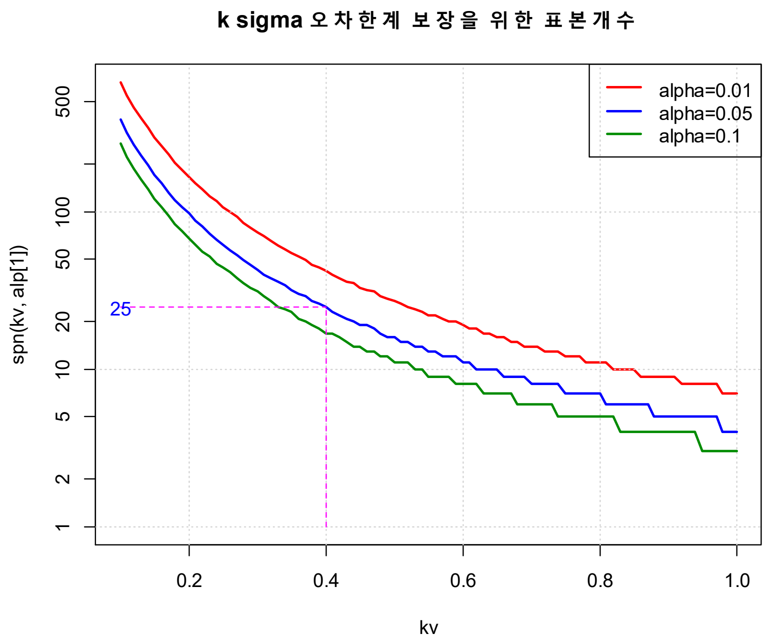
$$\Rightarrow k\sqrt{n}\ge \Phi^{-1}(1-\alpha/2)\equiv z_{1-\alpha/2}$$

$$\Rightarrow n\ge \left[\frac{\Phi^{-1}(1-\alpha/2)}{k}\right]^2$$

10/28

10

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



11/28

11

2. 正态总体样本均值的分布

[定理 9-2] 正态总体样本均值的分布(在不知σ的情况下)

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

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

$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{Z}{\sqrt{\chi^2_{(n-1)}/(n-1)}} \sim t(n-1)$$

12/28

12

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

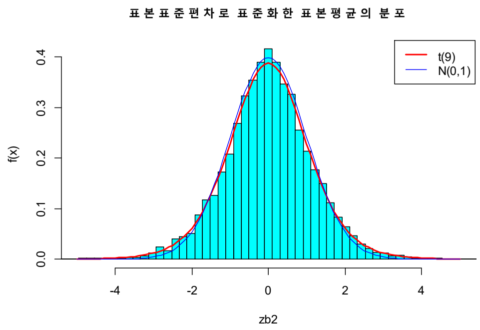
2. 正态分布样本均值的分布

[例 9-5] 从均值为100, 方差为100的正态分布总体中每次随机抽取10个样本，为求样本均值与样本方差重复取样10,000次。

(1) 在不知道总体方差的情况下，对样本均值进行标准化后，绘制相应直方图并检验[定理 9-2]。

(2) 比较其与标准正态分布的差异。

$$\frac{\bar{X}-100}{S/\sqrt{10}} \sim t(9)$$



13/28

13

3. 正态总体样本方差的分布

[定义 8-3] 卡方分布(chi-square distributiion)

$$Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} N(0,1) \Rightarrow \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

- 期望 $E(X) = \nu$
- 方差 $Var(X) = 2\nu$

[定理 9-3] 标准正态随机样本平方和的分布

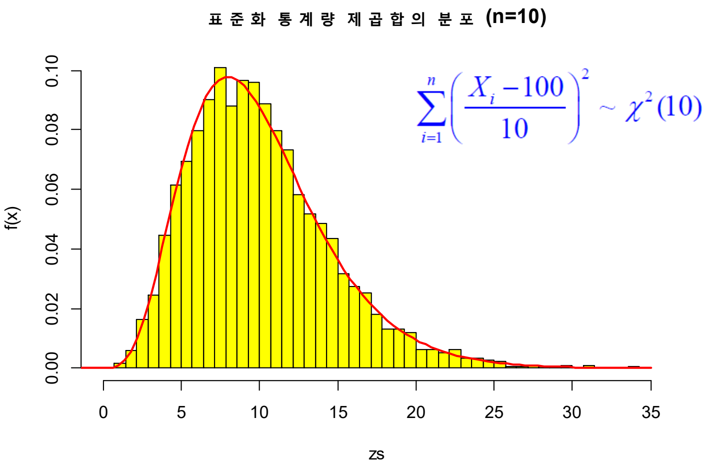
$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$$

14/28

14

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

[例 9-6] 从N(100,10^2)中随机抽取10个样本，计算标准化平方和统计量，重复取样10, 000次 → 绘制直方图并检验[定理 9-3]。



15/28

15

3. 正态总体样本方差的分布

[定理 9-4] 正态总体样本方差的分布

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \sim \chi^2(n-1)$$

[证明]

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left[(X_i-\bar{X})+(\bar{X}-\mu)\right]^2 \qquad S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2 \qquad \sum_{i=1}^{n}(X_i-\bar{X})=0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

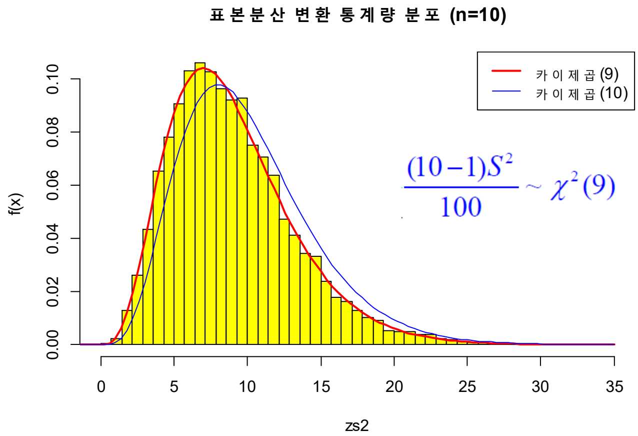
$$\Rightarrow (1-2t)^{-n} = m(t)\times(1-2t)$$

$$\Rightarrow m(t) = (1-2t)^{-(n-1)}$$

16/28

16

[例 9-7] 从均值为100, 方差为100的正态分布总体中随机抽取10个样本，计算(10-1)S^2/100，重复取样10,000次 → 绘制直方图并检验[定理 9-4]



17/28

17

3. 正态总体样本方差的分布

[例 9-8] 从正态分布总体中抽取n=5, 10, 30, 100个的随机样本，计算样本方差 → $P(S^2 \le c\sigma^2) = 0.95$

$$P(S^2 \le c\sigma^2) = P\left(\frac{(n-1)S^2}{\sigma^2} \le (n-1)c\right) = 0.95$$

$$\Rightarrow (n-1)c = \chi^2_{0.95;n-1}$$

```
# 输入自由度、累积概率
nu <- c(5, 10, 30, 100)
p <- 0.95

# 分位数/自由度(常数 c)
qchisq(p, nu-1) / (nu-1)
[1] 2.371932 1.879886 1.467482 1.244699
```

18/28

18

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

4. 两正态总体样本方差比的分布

[定义 8-5] F-分布

$$U \sim \chi^2(\nu_1) \perp V \sim \chi^2(\nu_2) \Rightarrow F \equiv \frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2)$$

[定理 9-5] 两正态总体样本方差比的分布

$$S_1^2 \perp S_2^2 \Rightarrow \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

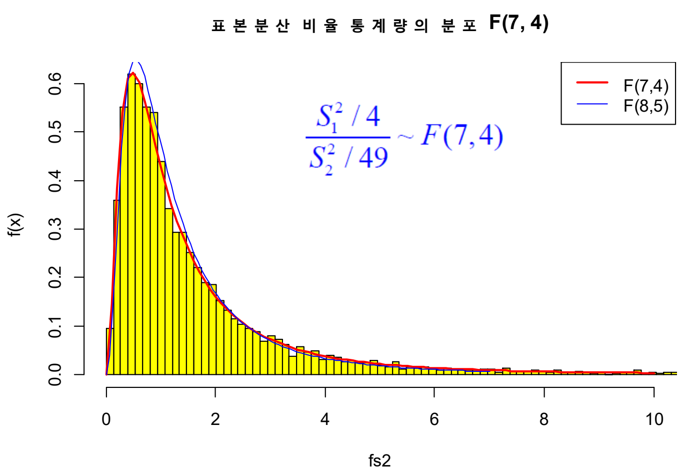
[证明] $U = \dfrac{(n_1-1)S_1^2}{\sigma_1^2}, V = \dfrac{(n_2-1)S_2^2}{\sigma_2^2} \Rightarrow U \sim \chi^2(n_1-1) \perp V \sim \chi^2(n_2-1)$

$$\frac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}} = \frac{\dfrac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2}}{\dfrac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2}} = \frac{\dfrac{U}{n_1-1}}{\dfrac{V}{n_2-1}} \sim F(n_1-1, n_2-1)$$

19/28

19

[例 9-9] 方差分别为4和49的相互独立的正态分布总体，从中各随机抽取8个、5个样本，计算样本方差比，重复取样10,000次，绘制直方图并检验[定理 9-5]。



20/28

20

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

4. 两正态总体样本方差比的分布

[例 9-10] 从总体方差相同的两个互为独立的正态总体中，各随机抽取 n1=12, n2=16的随机样本

(1) $P(S_1^2 / S_2^2 > a) = 0.05$

$$\sigma_1^2 = \sigma_2^2 \Rightarrow \quad F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F(11,15)$$

```
qf(0.05, 11, 15, lower.tail=FALSE)
[1] 2.506806
```

(2) 样本方差的比例为出现两倍以上的概率

$$P(F \le 0.5) + P(F \ge 2)$$

```
pf(0.5, 11, 15) + 1-pf(2, 11, 15)
[1] 0.2307009
```

21/28

21

5. 中心极限定理

[定理 9-6] 中心极限定理(central limit theorem)

：对于期望与方差是一定的随机样本

$$S_n = X_1 + X_2 + \cdots + X_n \overset{a}{\sim} N(n\mu, n\sigma^2)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n})$$

- 标准化 $\frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0,1)$

[例 9-11]* 对下列分布使用中心极限定理，绘制标准化统计量的直方图和正态概率图，并比较收敛情况。

$$t(3),\ Exp(5),\ \Gamma(0.5,10),\ Poi(1) \qquad (n = 10, 30, 50)$$

22/28

22

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

5. 中心极限定理

[例 9-12] 游戏技能者资格考试分数符和B(100, 0.7)分布时，从中随机抽取40名，样本均值小于等于69分以下的概率是多少。

$$E(X)=np=70,\ Var(X)=np(1-p)=21 \Rightarrow \bar{X} \overset{a}{\sim} N(70,\ 21/40)$$

$$P(\bar{X} \le 69) \simeq P(Z \le \frac{69-70}{\sqrt{21/40}}) \doteq \Phi(-1.380) \doteq 0.084$$

```
pnorm(69, 70, sqrt(21/40))
[1] 0.08377314
```

[例 9-13] 具体求[例 9-12]的概率。

$$Y \sim B(4000,\ 0.7)$$

$$P(\bar{X} \le 69) = P(Y \le 40 \times 69) = P(Y \le 2760)$$

```
pbinom(2760, 4000, 0.7)
[1] 0.08676363
```

23/28

23

5. 中心极限定理

[例 9-14] 某一咖啡厅的待机时间成均值为10分钟的指数分布。
随机选取40名顾客，求其平均待机时间小于等于9分的概率。

$$\bar{X} \sim N(10,\ 100/40)$$

$$P(\bar{X} \le 9) \simeq P(Z \le \frac{9-10}{\sqrt{2.5}}) \doteq \Phi(-0.6325) \doteq 0.2635$$

```
pnorm(9, 10, sqrt(100/40))
[1] 0.2635446
 pnorm(9.5, 10, sqrt(100/40))
[1] 0.3759148
```

[例 9-15]* 具体求[例 9-14]的概率。([定理 7-2])

$$Y \sim \Gamma(40,\ 10) \qquad P(\bar{X} \le 9) = P(Y \le 40 \times 9) = P(Y \le 360)$$

```
pgamma(360, 40, scale=10)
[1] 0.2736964
```

24/28

24

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

6. 非正态总体的样本分布

[定理 9-7] 样本均值的分布(样本非常大的情况)

：具有一定的期望与方差的随机样本

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n}) \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

[例 9-16] 随机投骰子40次，求出现的点数均值介于3和4之间的概率。

$$\mu = \frac{6+1}{2} = 3.5 \quad \sigma^2 = \frac{(6+1)(6-1)}{12} = \frac{35}{12} \quad \Rightarrow \bar{X} \overset{a}{\sim} N(3.5, \frac{35}{12\times 40})$$

$$P(3 < \bar{X} < 4) \simeq P\left(\frac{3-3.5}{\sqrt{35/480}} < Z < \frac{4-3.5}{\sqrt{35/480}}\right)$$

$$\doteq P(-1.8516 < Z < 1.8516) \doteq 0.9359$$

```
pnorm(4, 3.5, sqrt(35/480)) - pnorm(3, 3.5, sqrt(35/480))
[1] 0.9359225
```

25/28

25

6. 非正态总体的样本分布

[推论 9-7] 样本比例的分布

：从进行n次成功概率为p的试验中抽取随机样本

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

[例 9-17]* 绘制下列二项分布的成功次数标准化统计量的直方图和正态概率图，并比较收敛情况。

$$B(n, 0.5) \quad B(n, 0.3) \quad B(n, 0.7) \quad B(n, 0.1)$$

26/28

26

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

6. 非正态总体的样本分布

[例 9-18] 随机投骰子40次，求点数3或4出现的次数介于10和15之间的概率。

$$X \sim B(40, \frac{1}{3}) \quad P(10 \le X \le 15) = \sum_{x=10}^{15} \binom{40}{x} (\frac{1}{3})^x (\frac{2}{3})^{40-x} \doteq 0.6723$$

```
pbinom(15, 40, 1/3) - pbinom(9, 40, 1/3)
[1] 0.6722687
```

近似计算
连续性校正

$$np = 40 \times \frac{1}{3} = \frac{40}{3}, \; np(1-p) = 40 \times \frac{1}{3} \times \frac{2}{3} = \frac{80}{9}$$

$$P(10 \le X \le 15) \simeq P\left(\frac{9.5 - 40/3}{\sqrt{80/9}} < Z < \frac{15.5 - 40/3}{\sqrt{80/9}}\right)$$
$$\doteq P(-1.2857 < Z < 0.7267) \doteq 0.6670$$

```
pnorm(15.5, 40/3, sqrt(80/9)) - pnorm(9.5, 40/3, sqrt(80/9))
[1] 0.6670348
```

```
pnorm(15, 40/3, sqrt(80/9)) - pnorm(10, 40/3, sqrt(80/9))
[1] 0.5801487
```

27/28

27

6. 非正态总体的样本分布

[定理 9-8] 样本均值的分布(虽不知道总体方差，但是样本非常大的情况)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

[例 9-19]* 运用[定理 9-8]，比较下列分布。

$$t(3) \quad Exp(5) \quad \Gamma(0.5, 10) \quad Poi(5)$$

28/28

28