《概率论与数理统计》习题

第十五讲　统计量

1. 设 x_1, x_2, x_n 和 $y_1, y_2, \cdots, y_n$ 是两组样本观测值，且关系如下：

$$y_i = ax_i + b, i = 1, 2, \cdots, n$$

其中，a 和 b 为非零常数。试求：

(a) 样本均值 $\bar{x}$ 和 $\bar{y}$ 间的关系；

(b) 样本方差 s_x^2 和 s_y^2 间的关系。

解：

(1)

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = a\bar{x} + b$$

.

(2)

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(ax_i + b - a\bar{x} - b)^2 \\ &= a^2 s_x^2 \end{aligned}$$

2. 设总体 X 的 3 阶矩存在，若 $x_1, x_2, \cdots, x_n$ 是取自该总体的简单随机样本，$\bar{x}$ 为样本均值，s^2 为样本方差，试证：

$$\mathrm{Cov}(\bar{x}, s^2) = \frac{\nu_3}{n},$$

其中 $\nu_3 = E(x - E(x))^3$。

证明:

$$
\begin{aligned}
Cov(\bar{x}, s^2) &= E\big[(\bar{x}-E(\bar{x}))(s^2-E(s^2))\big] \\
&= E\big[(\bar{x}-\mu)(s^2-\sigma^2)\big] \\
&= E\big[(\bar{x}-\mu)s^2\big] - E\big[(\bar{x}-\mu)\sigma^2\big] \\
&= E\big[(\bar{x}-\mu)s^2\big] \\
(\bar{x}-\mu)s^2 &= \frac{1}{n-1}(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{n-1}(\bar{x}-\mu)\sum_{i=1}^{n}((x_i-\mu)-(\bar{x}-\mu))^2 \\
&= \frac{1}{n-1}\big[(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^3\big] \\
E\big[(\bar{x}-\mu)(x_i-\mu)^2\big] &= \frac{1}{n}E\big[\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^2\big] \\
&= \frac{1}{n}E\big[(x_i-\mu)^3 + \sum_{i\neq j}(x_j-\mu)(x_i-\mu)^2\big] \\
&= \frac{v_3}{n} \\
E\big[(\bar{x}-\mu)^3\big] &= \frac{1}{n^3}E\big[\sum_{i=1}^{n}(x_i-\mu)\big]^3 \\
&= \frac{1}{n^3}\big[E\sum_{i=1}^{n}(x_i-\mu)^3 + 0\big] \\
&= \frac{v_3}{n^2}
\end{aligned}
$$

$$
Cov(\bar{x}, s^2) = E\big[(\bar{x}-\mu)s^2\big] = \frac{1}{n-1}(v_3 - \frac{v_3}{n}) = \frac{v_3}{n}.
$$

3. 设 $\bar{x}_1$ 和 $\bar{x}_2$ 是从同一正态总体 $N(\mu, \sigma^2)$ 独立抽取的容量相同的两个样本均值。试确定样本容量 n, 使得两样本均值的差超过 σ 的概率不超过 0.01。

解:

由于 $\bar{x_i} \sim N(\mu, \frac{\sigma^2}{n})$，$i = 1, 2$ 且相互独立，所以，$\bar{x_2} - \bar{x_1} \sim N(0, \frac{2\sigma^2}{n})$.

$$
P(|\bar{x_2} - \bar{x_1}| > \sigma) = P(\frac{|\bar{x_2} - \bar{x_1}|}{\sqrt{2\sigma^2/n}} > \sigma) = 2\big[1 - \Phi(\sqrt{\frac{n}{2}})\big] \leq 0.01
$$

因此,

$$
\Phi(\sqrt{\frac{n}{2}}) \geq 0.995
$$

$$n \geq 2.575^2 * 2 = 13.26$$

只要样本容量 $n \geq 14$，则可使同一正态总体的两样本均值的差超过标准差的可能性不大于 0.01.

4. 从指数总体 $Exp(1/\theta)$ 抽取了 40 个样品，试求 $\bar{x}$ 的渐近分布。

解:

由于指数总体 $Exp(\frac{1}{\theta})$ 的均值为 θ, 方差为 θ^2, 则 $\bar{x}$ 的渐近分布为 $N(\theta, \frac{\theta^2}{40})$.

5. 设总体 X 的分布函数 $F(x)$ 是连续的，$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 为取自此总体的次序统计量，设 $\eta_i = F(x_{(i)})$，试证:

(a) $\eta_1 \leq \eta_2 \leq \cdots \leq \eta_n$，$\eta_i$ 是来自均匀分布 $U(0,1)$ 总体的次序统计量;

(b) $E(\eta_i) = \frac{i}{n+1}$，$\mathrm{Var}(\eta_i) = \frac{i(n+1-i)}{(n+1)^2(n+2)}, 1 \leq i \leq n$;

(c) η_i 和 η_j 的协方差矩阵为

$$\begin{pmatrix} \frac{a_1(1-a_1)}{n+2} & \frac{a_1(1-a_2)}{n+2} \\ \frac{a_1(1-a_2)}{n+2} & \frac{a_2(1-a_2)}{n+2} \end{pmatrix}$$

其中，$a_1 = \frac{i}{n+1}, a_2 = \frac{j}{n+1}$.

证明:

(1) 由于分布函数 $F(x)$ 是递增的，所以，$\eta_1 \leq \eta_2 \leq \cdots \leq \eta_n$ 成立；又因为分布函数 $F(x)$ 具有连续性，且 $F(x)$ 服从均匀分布 $U(0,1)$，所以，η_i 是来自均匀分布 $U(0,1)$ 总体的次序统计量。

(2) 由于 η_i 是来自均匀分布 $U(0,1)$ 总体的次序统计量, 即 $\eta_i \sim Be(i, n-i+1)$. 即

$$E(\eta_i) = \frac{i}{n+1}$$

$$Var(\eta_i) = \frac{i(n+1-i)}{(n+1)^2(n+2)}$$

其中，$1 \leq i \leq n$.

(3) η_i 和 η_j 的联合分布密度为:

$$p_{ij}(y,z) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} y^{i-1}(z-y)^{j-i-1}(1-z)^{n-j}$$

其中，$y \le z$.

$$\begin{aligned}
E(\eta_i,\eta_j) &= \int_0^1\int_0^x yzp_{ij}(y,z)dydz\\
&= \int_0^1\int_0^x yz\frac{n!}{(i-1)!(j-i-1)!(n-j)!}y^{i-1}(z-y)^{j-i-1}(1-z)^{n-j}dydz\\
&= \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\int_0^1(1-z)^{n-j}z^{j-i}\int_0^x y^i(1-\frac{y}{z})^{j-i-1}dydz\\
&= \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\int_0^1(1-z)^{n-j}z^{j+1}dz\int_0^1 t^i(1-t)^{j-i-1}dt\\
&= \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\frac{\Gamma(j+2)\Gamma(n-j+1)}{\Gamma(n+3)}\frac{\Gamma(i+1)\Gamma(j-i)}{\Gamma(j+1)}\\
&= \frac{ij+i}{(n+1)(n+2)}
\end{aligned}$$

$$\begin{aligned}
Cov(\eta_i,\eta_j) &= E(\eta_i,\eta_j)-E(\eta_i)E(\eta_j)\\
&= \frac{ij+i}{(n+1)(n+2)}-\frac{ij}{(n+1)^2}\\
&= \frac{-ij+(n+1)i}{(n+1)^2(n+2)}
\end{aligned}$$

所以，η_i 和 η_j 的协方差矩阵为:

$$\begin{pmatrix} Cov(\eta_i,\eta_i) & Cov(\eta_i,\eta_j)\\ Cov(\eta_j,\eta_i) & Cov(\eta_i,\eta_i)\end{pmatrix}$$

$$= \begin{pmatrix} \frac{a_1(1-a_1)}{n+2} & \frac{a_1(1-a_2)}{n+2}\\ \frac{a_1(1-a_2)}{n+2} & \frac{a_2(1-a_2)}{n+2}\end{pmatrix}$$

其中，

$$a_1=\frac{i}{n+1}, a_2=\frac{j}{n+1}.$$

6. 假定总体的密度函数为

$$p(x)=6x(1-x), 0<x<1.$$

求样本量为 n 的样本中位数 $m_{0.5}$ 的渐近分布。解:

$$p(x)=6x(1-x)=\frac{\Gamma(2+2)}{\Gamma(2)\Gamma(2)}x^{2-1}(1-x)^{2-1}\sim Be(2,2)$$

由此可得，$p(x)$ 关于 x=0.5 对称，即 $m_{0.5}=0.5$，$Var(m_{0.5})=\frac{1}{4np^2(m_{0.5})}=\frac{1}{9n}$. 所以，$m_{0.5}$ 的渐近分布为 $N(0.5,\frac{1}{9n})$.

第十六讲　三大抽样分布

1. $x_1, x_2, \cdots, x_n$ 是来自 $N(\mu, 1)$ 的样本，试确定最小的常数 c 使得对任意的 $\mu \geq 0$ 有 $P(|\bar{x}| < c) \leq \alpha$。
   由于 $\bar{x} \sim N(\mu, \frac{1}{n})$, 所以 $P(\bar{x} < c)$ 的值依赖于 μ, 它是 μ 的函数记为 $g(\mu)$, 于是

$$g(\mu) = P_\mu(|\bar{x}| < c) = P(-c < \bar{x} < c) = \Phi(\sqrt{n}(c-\mu)) - \Phi(-\sqrt{n}(c+\mu))$$

   其导函数为

$$g(\mu)' = -\sqrt{n}[\varphi(\sqrt{n}(c-\mu)) - \varphi(-\sqrt{n}(c+\mu))]$$

   由于 $g(\mu)' \leq 0, g(\mu)$ 为减函数, 并在 $\mu = 0$ 处取最大值
   故只需 $2\Phi(\sqrt{n}c) - 1 \leq \alpha$

2. 设随机变量 $X \sim F(n, m)$，证明:

$$Z = \frac{\frac{n}{m}X}{(1+\frac{n}{m}X)}$$

   服从贝塔分布，并指出其参数。
   证明:
   若 $X \sim F(n, m)$，则 X 的密度函数为:

$$p_X(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1}\left(1+\frac{n}{m}x\right)^{-\frac{m+n}{2}}$$

   因为 $Z = \frac{\frac{n}{m}X}{1+\frac{n}{m}X}$ 在 $(0, +\infty)$ 上严格单调增函数，其反函数为 $x = \frac{mz}{n(1-z)}$.
   所以，Z 的密度函数为:

$$\begin{aligned} p_Z(z) &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\left(\frac{n}{m}\right)^{\frac{n}{2}}\left(\frac{mz}{n(1-z)}\right)^{\frac{n}{2}-1}\left(1+\frac{z}{1-z}\right)^{-\frac{m+n}{2}}\frac{m}{n(1-z)^2} \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} z^{\frac{n}{2}-1}(1-z)^{\frac{m}{2}-1} \end{aligned}$$

   其中，$0 < z < 1$. 说明 Z 服从 $Be(\frac{n}{2}, \frac{m}{2})$，其两个参数分别为 F 分布两个自由度的一半。

3. 设 $x_1, x_2, \cdots, x_n$ 是来自 $N(\mu_1, \sigma^2)$ 的样本，$y_1, y_2, \cdots, y_m$ 是来自 $N(\mu_2, \sigma^2)$ 的样本，c, d 是任意两个不为 0 的常数，证明

$$t = \frac{c(\bar{x} - \mu_1) + d(\bar{y} - \mu_2)}{s_w \sqrt{\frac{c^2}{n} + \frac{d^2}{m}}} \sim t(n + m - 2)$$

其中 $s_w^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$。证明：

可得：

$$c(\bar{x} - \mu_1) \sim N(0, \frac{c^2\sigma^2}{n}), d(\bar{y} - \mu_2) \sim N(0, \frac{d^2\sigma^2}{m}),$$
$$\frac{(n-1)s_x^2}{\sigma^2} \sim \chi^2(n-1), \frac{(m-1)s_y^2}{\sigma^2} \sim \chi^2(m-1),$$

且 $\bar{x}, \bar{y}, s_x^2, s_y^2$ 相互独立，可得：

$$c(\bar{x} - \mu_1) + d(\bar{y} - \mu_2) \sim N(0, \frac{c^2\sigma^2}{n} + \frac{d^2\sigma^2}{m}).$$

$$\frac{(n+m-2)s_w^2}{\sigma^2} = \frac{(n-1)s_x^2}{\sigma^2} + \frac{(m-1)s_y^2}{\sigma^2} \sim \chi^2(n+m-2).$$

即：

$$\begin{aligned} t &= \frac{c(\bar{x} - \mu_1) + d(\bar{y} - \mu_2)}{s_w \sqrt{\frac{c^2}{n} + \frac{d^2}{m}}} \\ &= \frac{\Big[c(\bar{x} - \mu_1) + d(\bar{y} - \mu_2)\Big] / \sqrt{\frac{c^2}{n} + \frac{d^2}{m}}}{\sqrt{\frac{(n+m-2)s_w^2}{\sigma^2} / (n+m-2)}} \\ &\sim t(n+m-2) \end{aligned}$$

(编程题) 使用随机模拟的方法得到样本峰度和样本偏度的分布（输出分位数）。

仅供参考

```python
def k_order_moment(Sample,k_order):
    result = 0
    for i in Sample:
        result += (i-np.mean(Sample))**k_order
    result = result/len(Sample)
    return result
```

```python
def skewness(Sample):#偏度
    return k_order_moment(sample,3)/((k_order_moment(sample,2))**1.5)
```

```python
def kurtosis(Sample):#峰度
    return k_order_moment(Sample,4)/(k_order_moment(Sample,2)**2) - 3
```

```python
f=[0.1,1,2.5,5,10,20,30,40,50,60,70,80,85,90,95,97.5,99,99.5,99.9]
for i in range(10,25):
    skewness_list = []
    kurtosis_list = []
    for j in range(10000):
        sample = []
        for k in range(i):
            sample.append(np.random.rand(1))

        sample = np.array(sample)
        sample_skewness = skewness(sample)
        sample_kurtosis = kurtosis(sample)
        skewness_list.append(sample_skewness)
        kurtosis_list.append(sample_kurtosis)
    for l in f:
        print(np.percentile(skewness_list,l))
    for m in f:
        print(np.percentile(kurtosis_list,m))
```