数理统计 week 5

学业辅导中心

离散分布参数的置信区间

一个恒等式是

$$\int_p^1 \frac{n!}{(k-1)!(n-k)!} z^{k-1}(1-z)^{n-k} dz = \sum_{x=0}^{k-1} \binom{n}{x} p^x (1-p)^{n-x}$$

从而根据Γ函数和二项式展开, 有

$$\int_0^p \frac{n!}{(k-1)!(n-k)!} z^{k-1}(1-z)^{n-k} \, \mathrm{d}z = \sum_{w=k}^{n} \binom{n}{w} p^w (1-p)^{n-w}$$

证明方法:

1. 分部积分, 拆成k项求和
2. 一个简单的办法是: 两边对p求导.

离散分布参数的置信区间

一般结论: 得到保守的置信区间, 也就是给出的置信区间的长度更长.

$$\bar{\theta} = \sup\{\theta : F_T(T;\theta) \geqslant \alpha_1\}$$
$$\underline{\theta} = \inf\{\theta : F_T(T-;\theta) \leqslant 1-\alpha_2\}$$

因此, 有

$$\theta > \bar{\theta} \Rightarrow F_T(T;\theta) \leqslant \alpha_1$$
$$\theta < \underline{\theta} \Rightarrow F_T(T-;\theta) \geqslant 1-\alpha_2$$

利用这些关系, 得出

$$\begin{aligned} P[\underline{\theta} < \theta < \bar{\theta}) &= 1 - P[\{\theta < \underline{\theta}\} \cup \{\theta > \bar{\theta}\}] = 1 - P[\theta < \underline{\theta}) - P[\theta > \bar{\theta}] \\ &\geqslant 1 - P\left[F_T(T-;\theta) \geqslant 1-\alpha_2\right] - P\left[F_T(T;\theta) \leqslant \alpha_1\right] \geqslant 1-\alpha_1-\alpha_2 \end{aligned}$$

4.4.6 设 X_1，X_2，X_3 是来自连续型分布的随机样本，此分布的 pdf 是 $f(x)=2x$，$0<x<1$，其他为 0.

(a) 计算 X_1，X_2，X_3 中最小者大于分布中位数的概率.

(b) 如果 $Y_1<Y_2<Y_3$ 是次序统计量，求 Y_2 与 Y_3 之间的相关系数.

$$
\begin{aligned}
F(x) &= \int_0^x f(x)dx = \int_0^x 2xdx \\
&= \left(x^2\right)_0^x = x^2.
\end{aligned}
$$

Ex 4.4.6.(a)

$$
\begin{aligned}
1/2 &= P\left(X \leq \xi_{1/2}\right) \\
&= \int_0^{\xi_{1/2}} 2x dx \\
&= \int_0^{\xi_{1/2}} dx^2 \\
&= \left(\xi_{1/2}\right)^2 - 1 \\
\Rightarrow \left(\xi_{l/2}\right)^2 &= 1/2 \\
\Rightarrow \xi_{1/2} &= (1/2)^{1/2}
\end{aligned}
$$

因此

$$
P\left(\xi_{1/2} \leq X_{(1)}\right) = P\left(\xi_{1/2} \leq X_1\right) P\left(\xi_{1/2} \leq X_2\right) P\left(\xi_{1/2} \leq X_3\right) = \frac{1}{8}
$$

Ex 4.4.6. (b)

只需知道$X_{(2)}, X_{(3)}$的联合分布:

$$
\begin{aligned}
g_{23}(x_2, x_3) &= \left[\frac{3!}{(2-1)!(3-2-1)!(3-3)!} \left[(F(x_2))\right]^{2-1} \left[F(x_3) - F(x_2)\right]^{3-2-1} \right] \\
&\quad \left[1 - F(x_3)\right]^{3-3} f(x_2) f(x_3) \\
&= 3! F(x_2) f(x_2) f(x_3) \\
&= 6x_2^2 2x_2 2x_3 \\
&= 24x_2^3 x_3, \quad 0 < x_2 < x_3 < 1
\end{aligned}
$$

进一步求期望(积分)是不困难的, 可以课下动手试一试.

4.4.7 设 $f(x)=1/6$，$X=1$，2，3，4，5，6，其他为 0，表示离散型分布的 pmf. 证明：若对此分布抽取样本量为 5 的随机样本，其最小观测值的 pmf 是

$$g_1(y_1)=\left(\frac{7-y_1}{6}\right)^5-\left(\frac{6-y_1}{6}\right)^5,\quad y_1=1,2,\cdots,6$$

其他为 0. 注意，本题随机样本来自于离散型分布. 书中正文中的所有公式都是随机样本来自于连续型分布假设下推导出来的，因而不可以应用它们. 为什么？

本题需要注意的地方在于连续场合和离散场合的差别: 怎样表达出顺序统计量的pmf?

提示

$$P(Y_1=y_1)=P(Y_1\geq y_1)-P(Y_1\geq y_1+1)$$

Ex 4.4.7.

$$
\begin{aligned}
P(Y_1 = y_1) &= P(X_1, X_2, \ldots, X_5 \geq y_1) - P(X_1, X_2, \ldots, X_5 \geq y_1 + 1) \\
&= [P(X \geq y_1)]^5 - [P(X \geq y_1 + 1)]^5 \\
&= \left(\frac{7 - y_1}{6}\right)^5 - \left(\frac{7 - y_1 - 1}{6}\right)^5 \\
&= \left(\frac{7 - y_1}{6}\right)^5 - \left(\frac{6 - y_1}{6}\right)^5, y_1 = 1, 2, 3, 4, 5, 6
\end{aligned}
$$

提示

$$P(Y_1 = y_1) = P(Y_1 \geq y_1) - P(Y_1 \geq y_1 + 1)$$

4.4.9 设 $Y_1<Y_2<\cdots<Y_n$ 表示来自下述分布样本量为 n 的随机样本次序统计量，此分布的 pdf 是 $f(x)=1$，$0<x<1$，其他为 0. 证明：第 k 个次序统计量 Y_k 具有贝塔 pdf，其参数 $\alpha=k$ 且 $\beta=n-k+1$.

注记

本题给出来一个从均匀分布到β分布的方法.

我们以另外一个称为**贝塔**(beta)分布的重要分布来结束本节，该分布将从一对独立的 Γ 随机变量中得到. 设 X_1 与 X_2 是两个独立随机变量，它们服从 Γ 分布，其联合 pdf 为

$$h(x_1,x_2)=\frac{1}{\Gamma(\alpha)\Gamma(\beta)}x_1^{\alpha-1}x_2^{\beta-1}\mathrm{e}^{-x_1-x_2},\quad 0<x_1<\infty,0<x_2<\infty$$

其他为 0，其中 $\alpha>0$，$\beta>0$. 设 $Y_1=X_1+X_2$ 并且 $Y_2=X_1/(X_1+X_2)$. 我们将证明，Y_1 与 Y_2 是独立的.

Ex 4.4.9.

$$
\begin{aligned}
g_k(y_k) &= \frac{n!}{(k-1)!(n-k)!}\left[F(y_k)\right]^{k-1}\left[1-F(y_k)\right]^{n-k} f(y_k) \\
&= \frac{n!}{(k-1)!(n-k)!}\left[\int_0^{y_k} f(x)dx\right]^{k-1}\left[1-\int_0^{y_k} f(x)dx\right]^{n-k} f(y_k) \\
&= \frac{n!}{(k-1)!(n-k)!}\left[\int_0^{y_k} 1dx\right]^{k-1}\left[1-\int_0^{y_k} 1dx\right]^{n-k} \times 1 \\
&= \frac{n!}{(k-1)!(n-k)!} y_k^{k-1}(1-y_k)^{n-k} \\
&= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} y_k^{k-1}(1-y_k)^{(n-k+1)-1},
\end{aligned}
$$

4.4.28 设 $Y_1<Y_2$ 表示来自 $N(\mu, \sigma^2)$ 分布样本量为 2 的随机样本次序统计量，其中 σ^2 是已知的.

(a) 证明 $P(Y_1<\mu<Y_2)=1/2$，并计算 Y_2-Y_1 的随机长度的期望值.

(b) 如果 $\overline{X}$ 是此样本的均值，通过解方程 $P(\overline{X}-c\sigma<\mu<\overline{X}+c\sigma)=1/2$ 求常数 c，同时把该随机区间的长度与(a)部分的期望值加以比较.

$$P(Y_2 \le y) = P(X_1 \le y)P(X_2 \le y) = F(y)^2, \quad P(Y_1 \le y) = 1 - [1 - F(y)]^2$$

Ex 4.4.28.

由于正态分布的均值就是中位数, 于是

$$P\left(Y_i < \xi_p < Y_j\right) = \sum_{w=i}^{j-1} \binom{n}{w} \left(p^w\right) (1-p)^{n-w}$$

代入即可

$$P\left(Y_1 < \mu < Y_2\right) = \binom{2}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{2-1} = \frac{1}{2}$$

进一步, 知道Y_1, Y_2的密度函数, 分别计算期望, 做差即可.

Ex 4.4.28.

$$P\left(-c < \frac{\bar{X}-\mu}{\sqrt{\sigma^2/2}} < c\right) = \frac{1}{2}$$

$$\Rightarrow 1 - 2 \times P\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/2}} < -c\right) = \frac{1}{2}$$

$$\Rightarrow P\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/2}} < -c\right) = \frac{1}{4}$$

$$P\left(\bar{X} - 0.6745\sqrt{\frac{\sigma^2}{2}} < \mu < \bar{X} + 0.6745\sqrt{\frac{\sigma^2}{2}}\right) = \frac{1}{2}$$

$$\Rightarrow P\left(\bar{X} - 0.6745\frac{\sigma}{\sqrt{2}} < \mu < \bar{X} + 0.6745\frac{\sigma}{\sqrt{2}}\right) = \frac{1}{2}$$

$$\Rightarrow P(\bar{X} - 0.4769\sigma < \mu < \bar{X} + 0.4769\sigma) = \frac{1}{2}$$

4.4.30 参照习题 4.1.1，利用式(4.4.8)求马达寿命中位数的(具有接近 90%置信水平)置信区间. 区间均值会怎样呢?

一个典型的基于二项分布的中位数置信区间问题, 找k恰好满足

$$1 - 2P(binom(n, 0.5) \leq k) \geq 0.90$$

对应区间

$$(X_{(k+1)}, X_{(n-k)})$$

4. 4. 31 设 $Y_1<Y_2<\cdots\cdots<Y_n$ 表示分布容量为 n 的随机样本顺序统计量，此分布具有 pdf $f(x)=3x^2/\theta^3$，$0<x<\theta$，其他情况为 0.

(a) 证明 $P(c<Y_n/\theta<1)=1-c^{3n}$，其中 $0<c<1$.

(b) 当 n 是 4，并且 Y_4 的观测值是 2.3 时，求 θ 的 95%置信区间？

Ex 4.4.31.

$$P(c\theta < Y_n < \theta) = P(Y_n < \theta) - P(Y_n < c\theta) = 1 - c^{3n}$$

$$P(1 < \theta/Y_n < 1/c) = P(Y_n < \theta < Y_n/c) \overset{set}{=} 0.95$$

反解出c再带入即可.

4.5.3 设 X 的 pdf 为 $f(x;\theta)=\theta x^{\theta-1}$，$0<x<1$，其他为 0，其中 $\theta\in\{\theta:\theta=1,2\}$. 为了对简单假设 $H_0:\theta=1$ vs 备择简单假设 $H_1:\theta=2$ 进行检验，使用样本量 $n=2$ 的随机样本，同时将临界区域定义成 $C=\{(x_1,x_2):3/4\leqslant x_1x_2\}$. 求检验功效函数.

Ex 4.5.3.

$$
\begin{aligned}
\gamma_C(\theta) &= P\left(x_1 x_2 \geq \frac{3}{4}\right) \\
&= \int_{3/4}^{1} \int_{3/4x_2}^{1} \theta^2 \left(x_1 x_2\right)^{\theta-1} dx_1 dx_2 \\
&= \theta^2 \int_{3/4}^{1} \frac{1}{\theta}\left(1 - \left(\frac{3}{4x_2}\right)^{\theta}\right) x_2^{\theta-1} dx_2 \\
&= 1 - \left(\frac{3}{4}\right)^{\theta} - \theta\left(\frac{3}{4}\right)^{\theta}\left(\log 1 - \log\frac{3}{4}\right) \\
&= 1 - \left(\frac{3}{4}\right)^{\theta} + \theta\left(\frac{3}{4}\right)^{\theta} \log\frac{3}{4}
\end{aligned}
$$

4.5.5 设 X_1，X_2 表示来自下面分布的样本量为 $n=2$ 的随机样本，此分布的 pdf 为 $f(x;\theta)=(1/\theta)\mathrm{e}^{-x/\theta}$，$0<x<\infty$，其他为 0. 当 X_1，X_2 的观测值(如 x_1，x_2)使得

$$\frac{f(x_1;2)f(x_2;2)}{f(x_1;1)f(x_2;1)}\leqslant\frac{1}{2}$$

那么拒绝 $H_0:\theta_0=2$，并接受 $H_1:\theta_0=1$. 这里 $\Omega=\{\theta:\theta=1,2\}$. 求当 H_0 为假时，检验显著性水平以及检验功效.

Ex 4.5.5.

条件:

$$
\begin{aligned}
&\Rightarrow \frac{\frac{1}{4}\exp\left(-\frac{x_1}{2}-\frac{x_2}{2}\right)}{\exp\left(-x_1-x_2\right)} \leq \frac{1}{2} \\
&\Rightarrow \frac{\frac{1}{4}\exp\left(-\left(\frac{x_1+x_2}{2}\right)\right)}{\exp\left(-\left(x_1+x_2\right)\right)} \leq \frac{1}{2} \\
&\Rightarrow \frac{1}{4}\exp\left(-\frac{x_1+x_2}{2}+x_1+x_2\right) \leq \frac{1}{2} \\
&\Rightarrow \exp\left(\frac{x_1+x_2}{2}\right) \leq 2 \\
&\Rightarrow x_1+x_2 \leq 2\ln(2)
\end{aligned}
$$

功效函数

$$\gamma_C(\theta) = P_\theta\left(X_1+X_2 \leq 2\ln 2\right)$$

Ex 4.5.5.

显著性水平:

$$\alpha = P_{\theta=2}\left(X_1 + X_2 \leq 2\ln 2\right).$$

$$\begin{aligned}
\alpha &= \int_0^{2\ln(2)} \frac{1}{4} x e^{-\frac{x}{2}} dx \\
&= \frac{1}{4}\int_0^{2\ln(2)} x e^{-\frac{x}{2}} dx \\
&= \frac{1}{4}\left(-2(x+2)e^{-\frac{x}{2}}\right)_0^{2\ln(2)} \\
&= -\frac{1}{2}\left[(2\ln(2)+2)e^{-\ln(2)} + 2\right] \\
&= 1 - e^{-\ln(2)} - \ln(2)e^{-\ln(2)} \\
&= \frac{1}{2}(1 - \ln 2)
\end{aligned}$$

Ex 4.5.5.

检验功效:

$$
\begin{aligned}
\gamma(1) &= \int_0^{2\ln(2)} xe^{-x}dx \\
&= \int_0^{2\ln(2)} xe^{-x}dx \\
&= \left(-(x+1)e^{-x}\right)_0^{2\ln(2)} \\
&= \left[-(2\ln(2)+1)e^{-2\ln(2)}+1\right] \\
&= -2\ln(2)e^{-2\ln(2)} - e^{-2\ln(2)} + 1 \\
&= \frac{3}{4} - \frac{1}{2}\ln(2) \\
&= \frac{3}{4} - \frac{1}{2}(0.6931) \\
&= 0.4035
\end{aligned}
$$

4.5.8 我们认为，轮胎寿命里程设为 X，服从均值为 θ、其标准差为 5000 的正态分布. 经验表明 $\theta=30\,000$，制造者声称，由新工艺过程生产的轮胎的均值 $\theta>30\,000$. $\theta=35\,000$ 是可能的. 通过对 $H_0:\theta=30\,000$进行检验，验证此说法. 将观测 X 的 n 个独立值设为 $x_1,\cdots,x_n$，而且拒绝 H_0（因而接受 H_1）当且仅当 $\bar{x}\geqslant c$. 求 n 与 c，以使检验的功效函数 $\gamma(\theta)$满足 $\gamma(30\,000)=0.01$ 以及 $\gamma(35\,000)=0.98$.

Ex 4.5.8.

拒绝域

$$\frac{\overline{\boldsymbol{x}} - \boldsymbol{\theta}_0}{\sigma/\sqrt{n}} \geq z_\alpha$$

功效函数:

$$\gamma(\theta) = \Phi\left(-z_\alpha - \frac{\sqrt{n}\,(\theta_0 - \theta)}{\sigma}\right)$$

由于

$$\theta_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = c$$

于是

$$\begin{aligned}\gamma(30000) &= 1 - \Phi\left(\frac{c - \theta_0}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(\frac{c - 30000}{5000/\sqrt{n}}\right) \\ &= 0.01\end{aligned}$$

Ex 4.5.8.

$$1-\Phi\left(\frac{c-30000}{5000/\sqrt{n}}\right)=0.01$$

$$\begin{aligned}\gamma(35000)&=1-\Phi\left(\frac{c-30000}{5000\sqrt{n}}+\frac{\sqrt{n}(30000-35000)}{5000}\right)\\&=0.98\end{aligned}$$

两式联合解出c, n.

4.5.11 设 $Y_1<Y_2<Y_3<Y_4$ 是来自下面分布的样本量为 $n=4$ 的随机样本次序统计量，此分布的 pdf 为 $f(x;\theta)=1/\theta$，$0<x<\theta$，其他为 0，其中 $0<\theta$. 若观测值 $Y_4\geqslant c$，则拒绝 $H_0:\theta=1$ 而接受 $H_1:\theta>1$.

(a) 求常数 c，以使显著性水平是 $\alpha=0.05$.

(b) 求检验功效函数.

$$f_{Y_4}(y_4)=4\frac{y_4^3}{\theta^4}0<y_4<\theta$$

Ex 4.5.11 (a)

$$P(\text{ Type I error }) = 0.05$$
$$P(\text{ reject } H_0 \mid H_0 \text{ is true }) = 0.05$$
$$P(Y_4 \geq c \mid \theta = 1) = 0.05$$
$$\int_c^{\theta=1} 4\frac{y_4^3}{(1)^4} dy_4 = 0.05$$

解出c

$$4\int_c^1 y_4^3 dy_4 = 0.05$$
$$c^4 = 0.95$$

Ex 4.5.11 (b)

为求检验功效函数

$$
\begin{aligned}
\Upsilon(\theta) &= P\left(Y_4 \geq 0.9873 \mid \theta > 1\right) \\
&= \int_{0.9873}^{\theta} 4\frac{y_4^3}{\theta^4} dy_4 \\
&= \frac{4}{\theta^4}\left(\frac{y_4^4}{4}\right)_{0.9873}^{\theta} \\
&= \frac{\left(\theta^4 - 0.9873^4\right)}{\theta^4} \\
&= 1 - \frac{0.95}{\theta^4} \quad \theta > 1
\end{aligned}
$$

4.6.5 假定“每盒 10 磅”燕麦片的重量服从 $N(\mu, \sigma^2)$. 为了检验 $H_0: \mu=10.1$ vs $H_1: \mu>10.1$，我们抽取样本量 $n=16$ 的随机样本，发现 $\bar{x}=10.4$ 与 $s=0.4$.

(a) 在 5% 的显著性水平上，是接受还是拒绝 H_0 呢?

(b) 此检验的近似 p 值是多少?

统计量: 3, df: 15

```
> pt(3, 15, lower.tail = F)
[1] 0.004486369
> pnorm(3, lower.tail = F)
[1] 0.001349898
```

4.6.7 由世界卫生组织空气质量监测项目所收集的数据是对悬浮颗粒以 $\mu g/m^3$ 形式加以测量. 设 X 与 Y 分别等于墨尔本与休斯敦市中心(商业中心)地区以 $\mu g/m^3$ 度量的悬浮颗粒. 利用 X 的 $n=13$ 个观测值以及 Y 的 $m=16$ 个观测值, 检验 $H_0: \mu_X=\mu_Y$ vs $H_1: \mu_X<\mu_Y$.

(a) 当假定未知方差相等时, 定义检验统计量与临界区域. 设 $\alpha=0.05$.

(b) 如果 $\bar{x}=72.9$, $s_x=25.6$, $\bar{y}=81.7$, $s_y=28.3$, 计算检验统计量的值, 并阐述你的结论.

Ex 4.6.7.

$$
\begin{aligned}
T &= \frac{(\bar{x}-\bar{y})-\left(\mu_X-\mu_Y\right)}{\sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}} \\
&= \frac{(72.9-81.7)-0}{\sqrt{\frac{(13-1)(25.6)^2+(16-1)(28.3)^2}{13+16-2}}\sqrt{\frac{1}{13}+\frac{1}{16}}} \\
&= \frac{-8.8}{10.13137} \\
&\approx -0.869
\end{aligned}
$$

下面两题直接计算即可.

4.6.9 在习题 4.2.18 中，从 $N(\mu, \sigma^2)$ 分布抽取样本量为 n 的随机样本，并利用方差 S^2 给出方差 σ^2 的置信区间，其中均值 μ 是未知的. 在检验 $H_0: \sigma^2=\sigma_0^2$ vs $H_1: \sigma^2>\sigma_0^2$ 时，利用由 $(n-1)S^2/\sigma_0^2 \geqslant c$ 定义的临界区域. 也就是，当 $S^2 \geqslant c\sigma_0^2/(n-1)$ 时，则拒绝 H_0 且接受 H_1. 当 $n=13$ 且显著性水平 $\alpha=0.025$ 时，求 c.

$c = 23.34$

4.6.10 习题 4.2.27 建立两个正态分布方差之比的置信区间时，运用统计量 S_1^2/S_2^2，当那两个正态分布方差相等时，它服从 F 分布. 如果用 F 表示该统计量，那么利用临界区域 $F \geqslant c$ 对 $H_0: \sigma_1^2=\sigma_2^2$ vs $H_1: \sigma_1^2>\sigma_2^2$ 进行检验. 若 $n=13$，$m=11$ 以及 $\alpha=0.05$，求 c.

$c = 2.913$