习题课解答（统计部分二）

1. 设 $X_1,\cdots,X_n$ 为总体 $X\sim N(\mu,\sigma^2)$ 的一个样本，若 σ 已知，则 μ 的置信度为 $1-\alpha$ 的置信区间中，$\left(\bar{X}-u_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}},\bar{X}+u_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$ 是最短的。

证明：显然，枢轴量为 $\dfrac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}\sim N(0,1)$

对于给定的置信度为 $1-\alpha$，要定出 a,b，使得其满足

$$1-\alpha=P(a<\frac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}<b)=P(\frac{\bar{X}-b\sigma}{\sqrt{n}}<\mu<\frac{\bar{X}-a\sigma}{\sqrt{n}})$$

置信区间的长度为 $r=\dfrac{(b-a)\sigma}{\sqrt{n}}$ 最短。即求解

$$\begin{aligned}&\min r\\ &s.t.\quad \int_a^b\varphi(x)dx=1-\alpha\end{aligned}$$

这里，$\varphi(x)$ 为标准正态分布的密度函数

令 $L=\dfrac{(b-a)\sigma}{\sqrt{n}}+\lambda[\int_a^b\varphi(x)dx-(1-\alpha)]$

由

$$\frac{\partial L}{\partial a}=-\frac{\sigma}{\sqrt{n}}-\lambda\varphi(a)=0$$
$$\frac{\partial L}{\partial b}=\frac{\sigma}{\sqrt{n}}+\lambda\varphi(b)=0$$

可知 $\varphi(a)=\varphi(b)$，由于 $\varphi(x)$ 的对称性，$a=-b$，又因为 $\int_a^b\varphi(x)dx=1-\alpha$，

故$b=u_{1-\frac{\alpha}{2}}$

2. 设总体X服从均匀分布$U[0,\theta]$，$X_1,X_2,\cdots,X_n$是X的一组样本，要检验假设$H_0:\theta=c,\quad H_1:\theta>c$，其中$c>0$为常数。设统计量$M=\max\limits_{1\le i\le n}X_i$，原假设的拒绝域为$\{M>m_\alpha\}$，如果$\alpha$(0<$\alpha$<1)是犯第一类错误的概率，试证：拒绝域的临界值为$m_\alpha=c(1-\alpha)^{1/n}$。

【解】$M=\max\limits_{1\le i\le n}X_i$的分布函数为

$F_{\max}(x)=(F(x;\theta))^n=(x/\theta)^n$，当$x\in[0,\theta]$；

而犯第一类错误的概率$\alpha=P(M>m_\alpha)$.

故H_0成立时，$\alpha=1-(\frac{m_\alpha}{\theta})^n$.

3. 若总体$X\sim N(\mu_1,\sigma_1^2)$，总体$Y\sim N(\mu_2,\sigma_2^2)$，它们相互独立，而$X_1,X_2,\cdots,X_n$及$Y_1,Y_2,\cdots,Y_m$分别是它们的简单随机样本，$\bar{X},\bar{Y}$分别为它们的样本均值。如果知道$\sigma_1^2=\frac{1}{4}\sigma_2^2$，但$\sigma_1^2$和$\sigma_2^2$得具体数据未知，

（1） 证明$S_w^2=\frac{1}{n+m-2}[\sum\limits_{i=1}^{n}(X_i-\bar{X})^2+\frac{1}{4}\sum\limits_{i=1}^{m}(Y_i-\bar{Y})^2]$是$\sigma_1^2$的无偏估计。

（2） 给出假设检验$H_0:\mu_1=\mu_2,\quad H_1:\mu_1\ne\mu_2$的检验法。

【解】（1）两样本方差分别记为$S_1^2=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2, S_2^2=\frac{1}{m-1}\sum\limits_{i=1}^{m}(Y_i-\bar{Y})^2$，

由于$\frac{(n-1)S_1^2}{\sigma_1^2}\sim\chi^2(n-1),\frac{(m-1)S_2^2}{\sigma_2^2}\sim\chi^2(m-1)$，且由两样本的独立性知，它们是相互独立的。所以

$$\frac{(n-1)S_1^2}{\sigma_1^2}+\frac{(m-1)S_2^2}{\sigma_2^2}\sim\chi^2(n+m-2)$$

即 $$\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma_1^2}+\frac{\sum_{i=1}^{m}(Y_i-\overline{Y})^2}{4\sigma_1^2}\sim\chi^2(n+m-2),$$

故 $$E[\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma_1^2}+\frac{\sum_{i=1}^{m}(Y_i-\overline{Y})^2}{4\sigma_1^2}]=n+m-2$$

从而 $$ES_w^2=\frac{1}{n+m-2}E[\sum_{i=1}^{n}(X_i-\overline{X})^2+\frac{1}{4}\sum_{i=1}^{m}(Y_i-\overline{Y})^2]=\sigma_1^2$$

（2）由于 $\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{4\sigma_1^2}{m}}}\sim N(0,1)$，从而当 H_0 成立时，统计量

$$T=\frac{\overline{X}-\overline{Y}}{S_w\sqrt{\frac{1}{n}+\frac{4}{m}}}\sim t(n+m-2)$$

假定显著性水平为 α，则检验的拒绝域为

$$|T|\geq t_{1-\frac{\alpha}{2}}(n+m-2)$$

4. 设样本 $X_1,X_2,...,X_n$ 是来自 $U[\theta_1,\theta_2]$，求 $\theta_2-\theta_1$ 的置信度为 $1-\alpha$ 的等尾置信区间。

解：θ_1,θ_2 的极大似然估计分别为 $\hat{\theta}_1=X_{(1)},\hat{\theta}_2=X_{(n)}$，令 $\delta=\theta_2-\theta_1$，做平移变换

$Z_i=X_i+\theta_1$，$Z_i\sim U[0,\delta]$，故 $Z_{(i)}=X_{(i)}+\theta_1$，

所以 $X_{(n)}-X_{(1)}=Z_{(n)}-Z_{(1)}$

考虑枢轴量 $\zeta\triangleq\frac{X_{(n)}-X_{(1)}}{\delta}=\frac{Z_{(n)}-Z_{(1)}}{\delta}$，由于 $\frac{Z_i}{\delta}\sim U[0,1]$，

由于 $U[0,1]$ 的极差分布（为 $Beta(n-1,2)$ 分布）的密度函数为

$$\frac{Z_{(n)}-Z_{(1)}}{\delta}\sim f_R(r)=n(n-1)r^{n-2}(1-r),0<r<1$$

$B_{\frac{\alpha}{2}}$ 为该分布的 $\frac{\alpha}{2}$ 分位数

$$P(B_{\frac{\alpha}{2}} \le \frac{Z_{(n)} - Z_{(1)}}{\delta} \le B_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$[\dfrac{X_{(n)} - X_{(1)}}{B_{1-\frac{\alpha}{2}}}, \dfrac{X_{(n)} - X_{(1)}}{B_{\frac{\alpha}{2}}}]$即为所求

5. 在做某电视节目收视率调查时，甲市抽取了 **2000** 户，其中有 **541** 户收看了，乙市抽取了 **1000** 户，其中有 **285** 户收看了。若记 p_1, p_2 分别为甲乙两市对该电视节目的收视率，试在水平 $\alpha = 0.05$ 下，检验 $H_0: p_1 = p_2$， $H_1: p_1 \ne p_2$，并求其检验的 p 值。

解：设 $n_1 = 2000, \sum_{i=1}^{n_1} x_i = 541; n_2 = 1000, \sum_{i=1}^{n_2} y_i = 285$

显然，$n_1\bar{X} \sim B(n_1, p_1);\quad n_2\bar{Y} \sim B(n_2, p_2)$

由于是大样本，近似有

$$\bar{X} \dot{\sim} N(p_1, \frac{p_1(1-p_1)}{n_1});\quad \bar{Y} \dot{\sim} N(p_2, \frac{p_2(1-p_2)}{n_2}),$$

由总体独立性得

$$\bar{X} - \bar{Y} \dot{\sim} N(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})$$

故 $\dfrac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \dot{\sim} N(0,1)$

在 $p_1 = p_2 = p$ 未知时，可以用混合样本 $\hat{p} \triangleq \dfrac{n_1\bar{X} + n_2\bar{Y}}{n_1 + n_2}$ 作为 p 的点估计，所以，假设 $H_0: p_1 = p_2$， $H_1: p_1 \ne p_2$ 的检验统计量为

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \dot{\sim} N(0,1)$$

拒绝域为

$$W=\{|u|>u_{1-\frac{\alpha}{2}}\}$$

由于

$$\hat{p}\triangleq\frac{n_1\bar{x}+n_2\bar{y}}{n_1+n_2}=\frac{541+285}{2000+1000}=0.2753$$

$$u=\frac{\bar{x}-\bar{y}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}=\frac{0.2705-0.285}{\sqrt{0.2753\times0.7247(\frac{1}{2000}+\frac{1}{1000})}}=-0.838$$

因为$|u|=0.838<1.96=u_{1-\frac{0.05}{2}}$**，故接受原假设。**

此检验的 p **值**

$$p_{value}=2P(U\le u)=2P(U\le -0.838)=2(1-\Phi(0.838))>0.05=\alpha$$

6. 设 $X_1,\cdots,X_n$ 是总体 X 的一个样本， X 的密度函数为

$$f(x;\sigma)=\begin{cases}\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}e^{-\frac{x^2}{2\sigma^2}}, & x>0\\ 0, & x\le 0\end{cases},\sigma>0\text{为未知参数}$$

（1）试证明：$\frac{X}{\sigma}$ 与 $|Z|$ 同分布，这里 $Z\sim N(0,1)$；

（2）试给出假设 $H_0:\sigma=1\leftrightarrow H_1:\sigma=2$ 的似然比检验的拒绝域（水平为 α ）。

解：（1）对 $z>0$

$$P(|Z|\le z)=2\Phi(z)-1$$

故 $f_{|Z|}(z)=2\varphi(z)$

而 $f_{\frac{X}{\sigma}}(z)=2\varphi(z)$，事实上，

$$P(\frac{X}{\sigma}\le z)=P(X\le\sigma z)$$

$$=\int_0^{\sigma z}\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}e^{-\frac{x^2}{2\sigma^2}}dx$$

$$\overset{t=\frac{x}{\sigma}}{=}\int_0^{z}\sqrt{\frac{2}{\pi}}e^{-\frac{t^2}{2}}dt$$

（2）似然比统计量为：

$$\lambda(X_1,X_2,\cdots,X_n)=\frac{L(X_1,X_2,\cdots,X_n,2)}{L(X_1,x_2,\cdots,X_n,1)}$$

$$=\frac{[\frac{1}{2}\sqrt{\frac{2}{\pi}}]^n e^{-\frac{\sum_{i=1}^n X_i^2}{2\cdot 2^2}}}{[\sqrt{\frac{2}{\pi}}]^n e^{-\frac{\sum_{i=1}^n X_i^2}{2}}}=\frac{1}{2^n}e^{\frac{3}{8}\sum_{i=1}^n X_i^2}$$

拒绝域形式为

$$W=\{(x_1,x_2,\cdots,x_n):\lambda(x_1,x_2,\cdots,x_n)\ge C\}$$

$$=\{(x_1,x_2,\cdots,x_n):\frac{1}{2^n}e^{\frac{3}{8}\sum_{i=1}^n x_i^2}\ge C\}$$

$$=\{(x_1,x_2,\cdots,x_n):\sum_{i=1}^n x_i^2\ge\frac{8}{3}\ln(2^nC)\}$$

由于H_0为真时，$\frac{X}{\sigma}=X$与$|Z|$同分布，所以$\sum_{i=1}^n X_i^2\sim\chi^2(n)$，所以

$$W=\{(x_1,x_2,\cdots,x_n):\sum_{i=1}^n x_i^2\ge\chi^2_{1-\alpha}(n))\},\quad \chi^2_{1-\alpha}(n)=\frac{8}{3}\ln(2^nC)$$