应用统计

第8讲　贝叶斯估计与区间估计

三种信息

1. 总体信息

   总体分布，例如：我国为了确认国产轴承寿命的分布服从威布尔分布，前后花了5年时间，处理了几千个数据才确定下来

2. 样本信息

3. 先验信息

贝叶斯估计 将未知参数 θ 看做随机变量

-

1. 确定参数 θ 的先验分布 $\pi(\theta)$

2. 总体分布为随机变量 θ 去某个定值时，总体的条件概率函数，记为 $p(x|\theta)$

3. 确定样本 X 参数 θ 的联合概率函数 $h(x,\theta)=p(x|\theta)\pi(\theta)$

X 的边际概率函数 $m(x)=\int_{\Theta} h(x,\theta)d\theta=\int_{\Theta} p(x|\theta)\pi(\theta)d\theta$

4. 计算 X 条件下参数 θ 的条件分布，得到参数的后验分布

$$\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta}$$

条件分布

1. 离散随机变量的条件分布

对一切使 $P(Y=y_j)=p_{\cdot j}=\sum_{i=1}^{\infty}p_{ij}>0$ 的 y_j，称

$$p_{i|j}=P(X=x_i\mid Y=y_j)=\frac{P(X=x_i,Y=y_j)}{P(Y=y_j)}=\frac{p_{ij}}{p_{\cdot j}}\quad (i=1,2,\cdots)$$ 为

给定 $Y=y_j$ 条件下 X 的分布列。在 $Y=y_j$ 条件下 X 的分布函数

$$F(x\mid y_j)=\sum_{x_i\le x}P(X=x_i\mid Y=y_j)。$$

连续随机变量的条件分布

对一切使 $p_Y(y)>0$ 的 y，给定 $Y=y$ 条件下 X 的条件分布函数与条件密度函数定义如下：

$$F(x\mid y)=P(X\le x\mid Y=y)=\lim_{h\to 0}P(X\le x\mid y\le Y\le y+h)$$

$$=\lim_{h\to 0}\frac{P(X\le x, y\le Y\le y+h)}{P(y\le Y\le y+h)}$$

$$=\lim_{h\to 0}\frac{\int_{-\infty}^{x}\int_{y}^{y+h}p(u,v)dudv}{\int_{y}^{y+h}p_Y(v)dv}=\frac{\int_{-\infty}^{x}p(u,y)du}{p_Y(y)}=\int_{-\infty}^{x}\frac{p(u,y)}{p_Y(y)}du$$

$$p(x\mid y)=\frac{p(x,y)}{p_Y(y)}$$

连续随机变量的条件分布

例. 设(X,Y)服从$G=\{(x,y);\, x^2+y^2\le 1\}$上的均匀分布，求 $p(x\,|\,y)$

解： $p(x,y)=\begin{cases}\dfrac{1}{\pi}, & x^2+y^2\le 1\\ 0, & otherwise\end{cases}$

$$p_Y(y)=\int_{-\infty}^{+\infty}p(x,y)dx=\int_{-\sqrt{1-y^2}}^{+\sqrt{1-y^2}}\frac{1}{\pi}dx=\frac{2}{\pi}\sqrt{1-y^2}\cdot I_{-1\le y\le 1}$$

$$p(x\,|\,y)=\frac{p(x,y)}{p_Y(y)}=\frac{1}{2\sqrt{1-y^2}},\quad |x|\le\sqrt{1-y^2},\ |y|<1$$

连续随机变量的条件分布计算公式

连续场合下的全概率公式

$$p(x,y)=p_X(x)p(y\mid x)\Rightarrow p_Y(y)=\int_{-\infty}^{+\infty}p_X(x)p(y\mid x)dx$$

连续场合下的贝叶斯公式

$$p(x\mid y)=\frac{p(x,y)}{p_Y(y)}\Rightarrow p(x\mid y)=\frac{p_X(x)p(y\mid x)}{\int_{-\infty}^{+\infty}p_X(x)p(y\mid x)dx}\text{。}$$

例1. 设某事件 A 的发生概率为 θ，对试验进行了n次独立观测，其中事件发生了X次，估计参数 θ

显然 $X|\theta \sim b(n,\theta)$

$$P(X=x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x=0,1,2,\dots,n$$

假设我们在试验前，对A 没有什么了解，则通常使用均匀分布

$U(0,1)$ 作为参数 θ 的先验分布。X和θ的联合分布:

$$h(x,\theta)=p(x|\theta)\pi(\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},$$

$$x=0,1,\dots n, \qquad 0<\theta<1$$

例1. 设某事件 A 的发生概率为 θ，对试验进行了n次独立观测，其中事件发生了X次，估计参数 θ

X的边际分布 $m(x)=\int_0^1 h(x,\theta)d\theta=\binom{n}{x}\int_0^1 \theta^x(1-\theta)^{n-x}\,d\theta$

$$=\binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$$

θ 的后验分布$(0<\theta<1)$

$$\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}\theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1},$$

$\theta|x\sim Be(x+1,n-x+1)$

后验期望估计为 $\hat{\theta}=E(\theta|x)=\frac{x+1}{n+2}$,用经典方法得到的估计量为$\bar{\theta}=\frac{x}{n}$

Beta分布

Beta分布 $X \sim Be(a,b)$

$$p(x;a,b)=\begin{cases}\dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, & 0\le x\le 1\\ 0, & \text{其他}\end{cases}$$

Beta函数 $B(a,b)=\int_0^1 y^{a-1}(1-y)^{b-1}dy$

Gamma函数 $\Gamma(a)=\int_0^{+\infty} t^{a-1}e^{-t}dt,\quad B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

- $E(X)=\frac{a}{a+b},\quad Be(1,1)$ 即为 $U(0,1)$

$\Gamma(\alpha,\beta)$分布族

Gamma函数 $\Gamma(a)=\int_0^{+\infty} t^{a-1}e^{-t}dt$,

$$g(x;\alpha,\beta)=\begin{cases}\dfrac{\beta^{\alpha}x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)}, & x\geq 0\\ 0, & x<0\end{cases}$$

$\alpha=\beta=1$时为$Exp(1)$,

$\alpha={}^{n}/_{2}$, $\beta={}^{1}/_{2}$时为 χ_n^2

Beta分布的性质

设随机变量 $X_1, X_2, \cdots, X_n$ 相互独立，且均服从$U(0,1)$，顺序统计量 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$，

则 $X_{(r)} \sim Be(r, n-r+1)$,

$$X_{(s)} - X_{(r)} \sim Be(s-r, n-s+r+1)$$

特别地，$min(X_1, X_2, \cdots, X_n) = X_{(1)} \sim Be(1, n)$;

而且，若$X_1, X_2, \cdots, X_n$独立同分布，

$min(X_1, X_2, \cdots, X_n) \sim U(0,1)$, 则 $X_1 \sim Be\left(1, \frac{1}{n}\right)$。

Beta分布的性质

设 $X \sim \chi_m^2$ 与 $Y \sim \chi_n^2$ 相互独立，则 $\mathbf{U} = \frac{X}{X+Y} \sim Be\left(\frac{m}{2}, \frac{n}{2}\right)$

证明：（X，Y）的联合密度

$$p(x,y) = \left[2^{(m+n)/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\right]^{-1} x^{\frac{m}{2}-1} y^{\frac{n}{2}-1} e^{-\frac{x}{2}} e^{-\frac{y}{2}}$$

作变换 $\begin{cases} U = \frac{X}{X+Y} \\ V = X+Y \end{cases}$ 或 $\begin{cases} X = UV \\ Y = V(1-U) \end{cases}$

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v,$$ （U，V）的联合密度

$$p(u,v) = \left[2^{(m+n)/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\right]^{-1} (uv)^{\frac{m}{2}-1} v^{\frac{n}{2}-1} (1-u)^{\frac{n}{2}-1} e^{-v} v$$

$$= \left[2^{(m+n)/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\right]^{-1} u^{\frac{m}{2}-1} (1-u)^{\frac{n}{2}-1} v^{\frac{m+n}{2}-1} e^{-v}$$

所以U，V独立，且 $\mathbf{U} \sim Be\left(\frac{m}{2}, \frac{n}{2}\right)$， $\mathbf{V} = \mathbf{X}+\mathbf{Y} \sim \chi_{m+n}^2$

Beta分布的性质

设 $X\sim\chi_m^2$ 与 $Y\sim\chi_n^2$ 相互独立，则 $\mathrm{U}=\frac{X}{X+Y}\sim Be\left(\frac{m}{2},\frac{n}{2}\right)$

设 $X\sim\chi_m^2$ 与 $Y\sim\chi_n^2$ 相互独立，则

$$\mathrm{B}=\frac{X}{X+Y}\sim Be\left(\frac{m}{2},\frac{n}{2}\right),\qquad F=\frac{n}{m}\frac{X}{Y}\sim F(m,n)$$

$$\mathrm{B}=\frac{mF}{n+mF},\qquad F=\frac{nB}{m(1-B)}$$

例2. 设某事件 A 的发生概率为 θ，对试验进行了n次独立观测，其中事件发生了X次，假设θ 的先验分布为$Be(a,b)$，估计θ

$$\mathbf{X}|\theta \sim b(n,\theta), \quad P(X=x|\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x}, \ \ x=0,1,\ldots,n$$

X和θ的联合分布:

$$h(x,\theta)=p(x|\theta)\pi(\theta)=\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

$$=\binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{x+a-1}(1-\theta)^{n-x+b-1}$$

X的边际分布

$$m(x)=\int_0^1 h(x,\theta)d\theta=\binom{n}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(n+a+b)}$$

例2. 设某事件 A 的发生概率为 θ，对试验进行了n次独立观测，其中事件发生了X次，假设θ 的先验分布为$Be(a,b)$，估计θ

- θ 的后验分布
- $\pi(\theta|x) = \frac{h(x,\theta)}{m(x)} = \frac{\Gamma(n+a+b)}{\Gamma(x+a)\Gamma(n-x+b)}\theta^{x+a-1}(1-\theta)^{n-x+b-1}$
- $\theta|x \sim Be(x+a, n-x+b)$
- 后验期望估计为 $\hat{\theta} = E(\theta|x) = \frac{x+a}{n+a+b}$
- $\hat{\theta} = \frac{x+a}{n+a+b} = \frac{n}{n+a+b}\cdot\frac{x}{n} + \frac{a+b}{n+a+b}\cdot\frac{a}{a+b}$

共轭分布族

- 设样本 $X_1, X_2, \cdots, X_n$ 对参数 θ 的条件分布为$p(x|\theta)$，如果先验分布$\pi(\theta)$决定的后验分布密度$\pi(\theta|x)$与$\pi(\theta)$是同一类型的，称先验分布$\pi(\theta)$称为 $p(x|\theta)$的共轭分布。

- Beta分布是二项分布的共轭分布

- 正态分布是自身的共轭族

例3. 设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma_0{}^2)$的一个样本，其中$\sigma_0{}^2$已知，μ未知，假设μ的先验分布亦为正态分布$N(\theta, \tau^2)$，其中先验均值θ和先验方差τ^2均为已知，试求参数μ的贝叶斯估计。

$$p(x_1, x_2, \cdots, x_n|\mu) = (2\pi\sigma_0{}^2)^{-n/2} \cdot exp\left\{-\frac{1}{2\sigma_0{}^2}\sum_{k=1}^{n}(x_k-\mu)^2\right\}$$

$$\pi(\mu) = (2\pi\tau^2)^{-1/2} \cdot exp\left\{-\frac{1}{2\tau^2}\sum_{k=1}^{n}(\mu-\theta)^2\right\}$$

$$h(x, \mu) = p(x|\mu)\pi(\mu) \qquad \mu|x \sim N\left(\frac{\frac{n\bar{x}}{\sigma_0{}^2}+\frac{\theta}{\tau^2}}{\frac{n}{\sigma_0{}^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma_0{}^2}+\frac{1}{\tau^2}}\right)$$

后验期望估计为$\hat{\mu} = E(\mu|x) = \dfrac{\frac{n}{\sigma_0{}^2}}{\frac{n}{\sigma_0{}^2}+\frac{1}{\tau^2}}\bar{x} + \dfrac{\frac{1}{\tau^2}}{\frac{n}{\sigma_0{}^2}+\frac{1}{\tau^2}}\theta$

参数区间估计

- 点估计是用一个点（即一个数）估计未知参数。顾名思义，区间估计就是用一个区间估计未知参数。
- 例如估计一个人的年龄在40至45岁之间，一个人的身高在1米75至1米80之间，估计产品的合格率在0.95至0.98之间。
- 区间估计考虑到了估计的误差，多少给人们以更大的信任感。区间估计的理论就是用明确的概率语言刻画这种“信任感”的意义，并给出得到区间估计的具体方法。

参数区间估计示例

例 样本x_1, x_2, x_3, x_4来自正态总体$N(\mu,1)$，样本均值$\bar{x}$是参数μ的一个点估计

$[\bar{x}-1, \bar{x}+1]$即为μ的一个区间估计。 $P(\bar{x}=\mu)=0$， $\bar{x}-\mu \sim N\left(0,\frac{1}{4}\right)$。

$$P(\mu \in [\bar{x}-1, \bar{x}+1]) = P(\bar{x}-1 \le \mu \le \bar{x}+1) \quad = P(\mu-1 \le \bar{x} \le \mu+1)$$
$$= P(-1 \le \bar{x}-\mu \le 1) \quad = P(-2 \le 2(\bar{x}-\mu) \le 2) \quad = \Phi(2)-\Phi(-2) \approx 0.9544$$

$[\bar{x}-1, \bar{x}+1]$是μ的一个置信水平0.9544的区间估计，也可说是置信水平0.9或0.8的区间估计。给出的估计区间称为置信区间，置信水平可以取到的最大值称为置信系数。

置信水平的解释

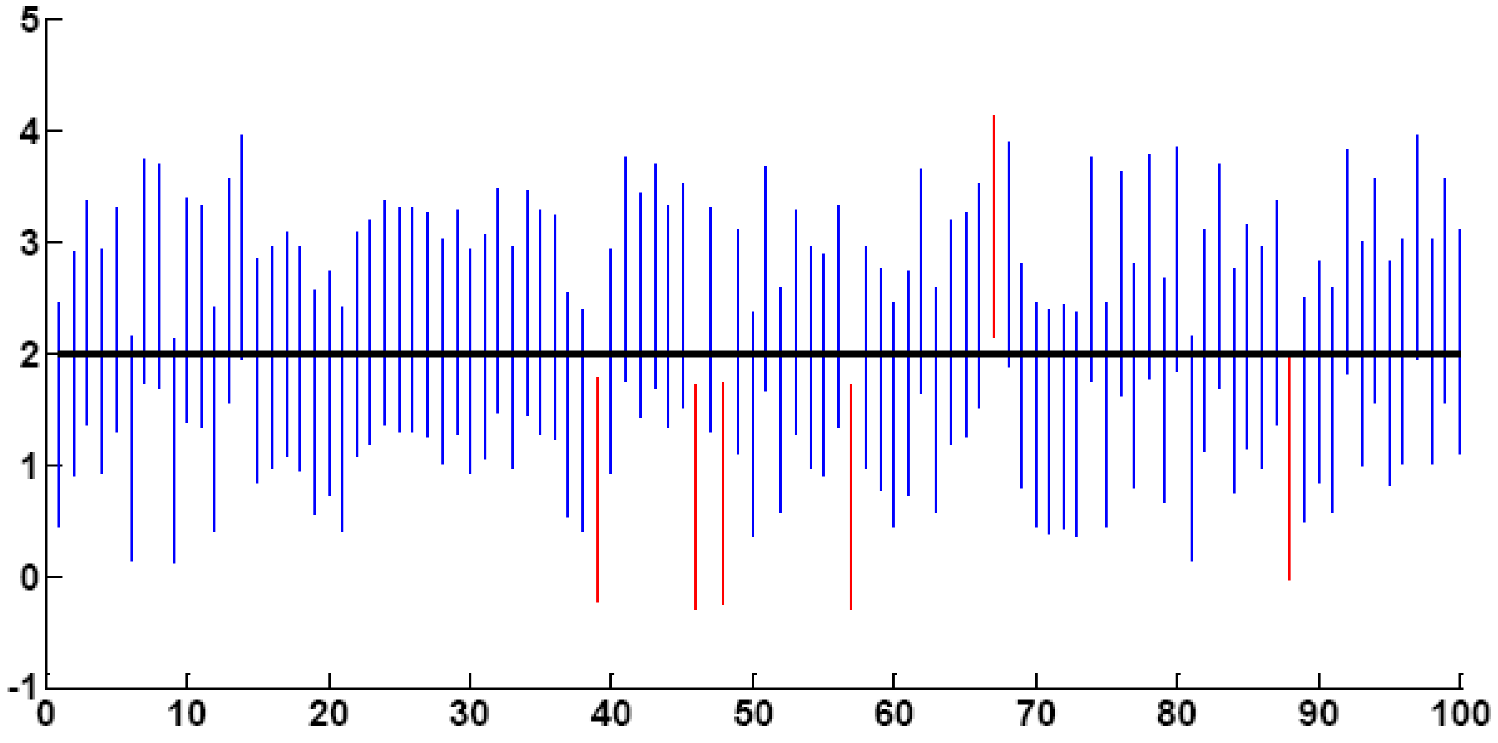
未知参数本身是确定的值，不带有随机性。随机性是由区间引入的。

一个置信水平 $1-\alpha$ 的区间估计，其含义是：得到的随机区间至少以概率 $1-\alpha$ 覆盖被估的参数。

考虑前面的问题。置信水平的含义是：每抽取 4 个样本得到一个区间估计，将这样的估计重复足够多次，至少 $1-\alpha$ 比例的估计区间包含真实的 μ 值。这里置信系数是 0.9544，即大约 95.44%估计区间包含真实的 μ 值。

置信水平的解释（图示）

下图是重复 100 次估计的模拟结果。继续模拟，将估计重复 10000 次，结果 442 次估计区间没有包含μ的真实值 2。当然，这些都是概率意义上的结果。具体的每一次估计，我们不会知道区间是否包含未知参数。



Script (Matlab):

```
n=4; m=100; mu=2; sigma=1;
% 设总体为N(mu,sigma^2), 样本容量n，重复估计m次

for k=1:m
    x(k,:)=sigma*(mu+randn(1,n));
    % randn是Matlab中生成N(0,1)随机数的命令
end

y=mean(x,2);  a=1;
hold on

for k=1:m
    if y(k)-a>mu | y(k)+a<mu
        plot([k,k],[y(k)-a,y(k)+a],'r','linewidth',2);
        % 没有覆盖参数的区间涂红色
    else
        plot([k,k],[y(k)-a,y(k)+a],'linewidth',2);      % 覆盖参数的区间涂蓝色
    end;
end

plot([1 m],[mu mu],'k','linewidth',3);                % 以参数值画一条横线
```

参数区间估计定义

设 $x_1, x_2, \cdots, x_n$ 是来自总体 $X \sim F(x;\theta)$ 的样本， $\theta \in \Theta \subset R^1$ 为未知参数。

$I(x_1, x_2, \cdots, x_n)$ 是一个随机区间，由样本值完全确定。称该区间是参数 θ 的一个置信水平为 $1-\alpha$ $(0<\alpha<1)$ 的区间估计，是指

$$P_\theta\left(\theta \in I(x_1, x_2, \cdots, x_n)\right) \geq 1-\alpha, \quad \forall\, \theta \in \Theta$$

。其中

$I = \left[\hat{\theta}_1(x_1, x_2, \cdots, x_n), \hat{\theta}_2(x_1, x_2, \cdots, x_n)\right]$ （双侧）区间估计

$I = \left[\hat{\theta}_L(x_1, x_2, \cdots, x_n), +\infty,\right]$ $\hat{\theta}_L$：置信下界 （单侧区间估计）

$I = \left[-\infty, \hat{\theta}_U(x_1, x_2, \cdots, x_n)\right]$ $\hat{\theta}_U$：置信上界 （单侧区间估计）

区间估计的两个基本要求

区间估计的两个基本要求:

1. 未知参数θ要以尽可能大的概率落在区间$I(x_1,x_2,\cdots,x_n)$中;

2. 估计的精度要尽可能高。比如，在达到一定的置信水平的前提下，要求区间的长度尽可能小，或某种能体现这个要求的其他准则。

区间估计的构造方法

枢轴量法（双侧置信区间为例）

步骤 1　构造“枢轴量”（pivot），$G(x_1,\cdots,x_n,\theta)$，G 的值完全由样本值和未知参数确定，G 的分布不依赖于未知参数

步骤 2　适当选取两个常数 c、d，对给定的 α（$0<\alpha<1$），有

$$P(c \le G \le d) \ge 1-\alpha$$

步骤 3　求解 $c \le G(x_1,\cdots,x_n,\theta) \le d$，得到

$$\hat{\theta}_1(x_1,x_2,\cdots,x_n) \le \theta \le \hat{\theta}_2(x_1,x_2,\cdots,x_n)。$$

区间估计公式

例　总体 $N(\mu,\sigma^2)$，σ^2 已知，μ 未知，简单随机样本 $x_1,x_2,\cdots,x_n$，求 μ 的 $1-\alpha$ 置信区间。

解：　μ 的点估计 $\bar{x}\sim N\left(\mu,\dfrac{\sigma^2}{n}\right)$，$\mu$ 为位置参数，$\bar{x}-\mu$ 的分布与 μ 无关，

$\dfrac{\bar{x}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}}$ 可作为枢轴量，$Z=\dfrac{\bar{x}-\mu}{\sqrt{\dfrac{\sigma^2}{n}}}=\dfrac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\sim N(0,1)$

$$\Phi(d)-\Phi(c)=P\left(c\le\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}\le d\right)\ge 1-\alpha \quad\Rightarrow\quad \bar{x}-d\frac{\sqrt{n}}{\sigma}\le\mu\le\bar{x}-c\frac{\sqrt{n}}{\sigma}。$$

确定 a、b 使得置信区间尽可能短。

置信区间的选取

优化问题：将选取最短置信区间的问题表达为下面优化问题。假设枢轴量的分布函数和密度函数分别为$F(x)$、$p(x)$，在约束$F(d)-F(c)=1-\alpha$约束下，求$d-c$的最小值。如果$F(x)$连续可导，由 Lagrange 乘数法可得$p(d)=p(c)$。

注： 对称分布的c和d，分别取$\frac{\alpha}{2}$和$1-\frac{\alpha}{2}$分位数。非对称分布，最优的c和d可能不易求得，通常也简单地取做枢轴量的$\frac{\alpha}{2}$和$1-\frac{\alpha}{2}$分位数。

上例中c和d分别取$c=\Phi^{-1}\left(\frac{\alpha}{2}\right)=u_{\frac{\alpha}{2}}$， $d=-c$。

统计抽样定理

设 $x_1, x_2, \cdots, x_n$ 是来自正态总体$N\left(\mu, \sigma^2\right)$的样本，其样本均值和样本方差分别为：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \text{ 和 } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2, \text{ 则有}$$

⑴ $\bar{x}$ 与 s^2 相互独立

⑵ $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

⑶ $\frac{(n-1)\cdot s^2}{\sigma^2} \sim \chi^2(n-1)$

区间估计例题

例　总体 $N(\mu,\sigma^2)$，μ 未知，σ^2 未知，求 μ 的 $1-\alpha$ 置信区间。

枢轴量 $\dfrac{\sqrt{n}(\bar{x}-\mu)}{s} \sim t(n-1) \Rightarrow P\left(t_{\alpha/2}(n-1) \le \dfrac{\sqrt{n}(\bar{x}-\mu)}{s} \le t_{1-\alpha/2}(n-1)\right) = 1-\alpha$

例　为估计某物体的质量，用一台天平测量 5 次，结果分别为（单位克）

5.52，　5.48，　5.64，　5.51，　5.43。　总体分布 $N(\mu,\sigma^2)$，μ 未知，σ^2 未知，估计 μ。　$\bar{x}=5.516$，　$s^2=\dfrac{1}{4}\sum_{k=1}^{5}(x_k-5.516)^2 \Rightarrow s=0.078$。　$t_4(0.975)=2.776$，　μ 的置信系数 0.95 的区间估计为 $[5.419, 5.613]$。

区间估计例题

例 $x_1,\cdots,x_m$ 来自正态总体 $N\left(\mu_1,\sigma_1^2\right)$， $y_1,\cdots,y_n$ 来自正态总体 $N\left(\mu_2,\sigma_2^2\right)$， μ_1、 μ_2 未知， σ_1^2、 σ_2^2 未知，求 $\mu_2-\mu_1$ 的 $1-\alpha$ 置信区间。

Behrens-Fisher 问题，统计学中至今没有完全求解的问题。

特殊情况 $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知时，

$$\bar{x}-\bar{y}\sim N\left(\mu_1-\mu_2,\left(\frac{1}{m}+\frac{1}{n}\right)\sigma^2\right),\quad \frac{(m-1)s_x^2+(n-1)s_y^2}{\sigma^2}\sim\chi^2(m+n-2)$$

$$t=\sqrt{\frac{mn(m+n-2)}{m+n}}\frac{\bar{x}-\bar{y}-(\mu_1-\mu_2)}{\sqrt{(m-1)s_x^2+(n-1)s_y^2}}\sim t(m+n-2)。$$

区间估计例题

例　总体 $N(\mu,\sigma^2)$，μ 未知，σ^2 未知，求 σ^2 的 $1-\alpha$ 置信区间。

枢轴量 $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \quad \Rightarrow \quad P\left(\chi^2_{\alpha/2} \le \dfrac{(n-1)s^2}{\sigma^2} \le \chi^2_{1-\alpha/2}\right) = 1-\alpha$.

例　$x_1,\cdots,x_m$ 来自正态总体 $N(\mu_1,\sigma_1^2)$，$y_1,\cdots,y_n$ 来自正态总体 $N(\mu_2,\sigma_2^2)$，

μ_1、μ_2 未知，σ_1^2、σ_2^2 未知，求 σ_1^2/σ_2^2 的区间估计。

$$\frac{(m-1)s_1^2}{\sigma_1^2} \sim \chi^2(m-1), \qquad \frac{(n-1)s_2^2}{\sigma_2^2} \sim \chi^2(n-1)$$

$$\Rightarrow \quad F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(m-1,n-1)$$

区间估计例题

例　$x_1,\cdots,x_n$ 来自指数总体 $Exp(\lambda)$，求 λ 的区间估计

可以证明 $X=2\lambda(x_1+\cdots+x_n)\sim\chi^2(2n)\ \Rightarrow\ 2n\lambda\,\overline{x}\sim\chi^2(2n)$

例　$x_1,\cdots,x_n$ 来自均匀总体 $U(0,\theta)$，求 θ 的 $1-\alpha$ 置信区间。

θ 的极大似然估计为 $\hat{\theta}=\max\{x_1,x_2,\cdots,x_n\}=x_{(n)}$，则其分布函数为

$F\left(x_{(n)},\theta\right)=\left(\dfrac{x_{(n)}}{\theta}\right)^n$，则 $\dfrac{x_{(n)}}{\theta}$ 可作为枢轴量。

$$P\left(c\le\frac{x_{(n)}}{\theta}\le d\right)=1-\alpha\ \Rightarrow\ d^n-c^n=1-\alpha,\quad \frac{x_{(n)}}{d}\le\theta\le\frac{x_{(n)}}{c}$$

可取 $d=1,c=\sqrt[n]{\alpha}$。

大样本区间估计

样本容量足够大时，可以利用渐近分布构造置信区间。（中心极限定理）

例 Behrens-Fisher 问题，$x_1,\cdots,x_m$ 来自正态总体 $N(\mu_1,\sigma_1^2)$，$y_1,\cdots,y_n$ 来自正态总体 $N(\mu_2,\sigma_2^2)$，μ_1、μ_2 未知，σ_1^2、σ_2^2 未知，求 $\mu_2-\mu_1$ 的区间估计。

$$\frac{\bar{x}-\bar{y}-(\mu_1-\mu_2)}{\sqrt{\sigma_x^2/m+\sigma_y^2/n}} \sim N(0,1), \qquad \frac{\bar{x}-\bar{y}-(\mu_1-\mu_2)}{\sqrt{s_x^2/m+s_y^2/n}} \dot\sim N(0,1)$$

大样本区间估计

例 $x_1,\cdots,x_m$ 来自两点分布总体 $b(1,p)$，求 p 的区间估计

$$E(\bar{x})=p\,,\qquad Var(\bar{x})=\frac{p(1-p)}{n}$$

$$\bar{x}\sim N\left(p,\frac{p(1-p)}{n}\right),\qquad u=\frac{\bar{x}-p}{\sqrt{p(1-p)/n}}\sim N(0,1)$$

$$P\left(\left|\frac{\bar{x}-p}{\sqrt{p(1-p)/n}}\right|\le u_{1-\frac{\alpha}{2}}\right)=1-\alpha\quad\Rightarrow\quad I=\left[\bar{x}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}},\bar{x}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right]$$

区间长度$=2u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{\bar{x}(1-\bar{x})}{n}}\le u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{1}{n}}$。

例　样本$X_1,\cdots,X_n$来自两点分布总体$b(1,p)$，求p的区间估计。

解：样本均值的期望、方差分别为 $E(\bar{X})=p$，$Var(\bar{X})=\dfrac{p(1-p)}{n}$

根据中心极限定理当n较大时，有近似分布

$\bar{X}\dot{\sim}N\left(p,\dfrac{p(1-p)}{n}\right)$，　标准化后得到枢轴量 $\dfrac{\bar{X}-p}{\sqrt{\dfrac{p(1-p)}{n}}}\dot{\sim}N(0,1)$，

$P\left(\left|\dfrac{\bar{X}-p}{\sqrt{p(1-p)/n}}\right|\le u_{1-\frac{\alpha}{2}}\right)\approx 1-\alpha$，　即 $(\bar{X}-p)^2\le u_{1-\frac{\alpha}{2}}^2\dfrac{p(1-p)}{n}$

$P\left(\left|\dfrac{\bar{X}-p}{\sqrt{p(1-p)/n}}\right|\le u_{1-\frac{\alpha}{2}}\right)\approx 1-\alpha$，　即 $(\bar{X}-p)^2\le u_{1-\frac{\alpha}{2}}^2\dfrac{p(1-p)}{n}$，解得

$$P\left(\left|\frac{\bar{X}-p}{\sqrt{p(1-p)/n}}\right|\le u_{1-\frac{\alpha}{2}}\right)\approx 1-\alpha$$，即 $\left(\bar{X}-p\right)^2\le u^2_{1-\frac{\alpha}{2}}\dfrac{p(1-p)}{n}$，解得

$$\frac{1}{1+c}\left(\bar{X}+\frac{c}{2}-\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}u^2_{1-\frac{\alpha}{2}}+\frac{c^2}{4}}\right)\le p\le\frac{1}{1+c}\left(\bar{X}+\frac{c}{2}+\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}u^2_{1-\frac{\alpha}{2}}+\frac{c^2}{4}}\right)$$

其中 $c=\dfrac{u^2_{1-\frac{\alpha}{2}}}{n}$，当 n 较大时，c 的值很小可略去，得到参数 p 的 $1-\alpha$ 置信水平的近似估计区间 $\left[\bar{X}-u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}\left(1-\bar{X}\right)}{n}},\bar{X}+u_{1-\frac{\alpha}{2}}\sqrt{\dfrac{\bar{X}\left(1-\bar{X}\right)}{n}}\right]$。

例 一个网络产品的运营商希望了解该产品在某个城市的用户占有率，进行了随机抽样调查，样本容量为 500，调查结果有 84 人是该产品的用户，求这个网络产品在该城市占有率的一个 95%置信区间。

解：总体分布 $X \sim b(1,p)$， 样本均值 $\bar{X}=\dfrac{X_1+\cdots+X_{500}}{500}$

期望、方差分别为 $E(\bar{X})=p$，$Var(\bar{X})=\dfrac{p(1-p)}{n}=\dfrac{p(1-p)}{500}$

$$\frac{(\bar{X}-p)}{\sqrt{p(1-p)/500}} \stackrel{\cdot}{\sim} N(0,1)\text{，}\left[\bar{X}-u_{0.975}\sqrt{\frac{\bar{X}(1-\bar{X})}{500}},\bar{X}+u_{0.975}\sqrt{\frac{\bar{X}(1-\bar{X})}{500}}\right]$$

$\bar{x}=\dfrac{84}{500}=0.168$， $u_{0.975}=1.96$， 近似估计区间为：$[0.135,0.201]$。

例 一个网络产品的运营商希望了解该产品在某个城市的用户占有率，进行了随机抽样调查，样本容量为 500，调查结果有 84 人是该产品的用户。根据抽样调查的信息，运营商希望得到参数 p 的 95%置信水平的区间估计，且估计区间长度不超过 0.02，问至少需要多大的样本容量？

解：参数 p 的 $1-\alpha$ 置信区间为 $\left[\bar{X}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right]$

估计区间长度不超过 0.02，即 $u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \le 0.01$。

$$n \ge \left(\frac{u_{1-\alpha/2}}{0.01}\right)^2 \bar{x}(1-\bar{x}) = \left(\frac{u_{0.975}}{0.01}\right)^2 \bar{x}(1-\bar{x}) = \left(\frac{1.96}{0.01}\right)^2 0.168(1-0.168) = 5369.6$$

例 一个网络产品的运营商希望了解该产品在某个城市的用户占有率，希望得到参数 p 的 95%置信水平，且估计区间长度不超过 0.02 的区间估计，问在没有任何先验知识的情况下，至少需要多大的样本容量才能保证达到所希望的估计精度？

解：参数 p 的 $1-\alpha$ 置信区间为 $\left[\bar{X}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}}, \bar{X}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}}\right]$

对任何 $\bar{X}\in(0,1)$，$\bar{x}(1-\bar{x})\leq\frac{1}{4}$

$$n\geq\left(\frac{u_{1-\alpha/2}}{0.01}\right)^2\frac{1}{4}\geq\left(\frac{u_{1-\alpha/2}}{0.01}\right)^2\bar{x}(1-\bar{x}),\qquad n\geq\left(\frac{u_{0.975}}{0.01}\sqrt{\frac{1}{4}}\right)^2=9604$$

问题没有变，可是现在这个情况下估计的所需样本容量数比上一题多了不少。因为本题中先验信息少于上一题， 所以得到估计不如上题精确也是很自然的。一般而言，得到信息越多，越有可能得到更好的估计。

作业

1. 例3的详细推导过程

例3. 设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, {\sigma_0}^2)$的一个样本，其中${\sigma_0}^2$已知，μ未知，假设μ的先验分布亦为正态分布$N(\theta, \tau^2)$，其中先验均值θ和先验方差τ^2均为已知，试求参数μ的贝叶斯估计。

习题二 23，24