抽样分布及预备知识

苏勤亮

中山大学计算机学院

超算中心503M

suqliang@mail.sysu.edu.cn

引言

- 统计量是样本的函数, 故是随机变量. 统计量的概率分布称为抽样分布. 当总体X的分布类型已知时, 对任一自然数n, 理论上，统计量$T = T(X_1, \cdots, X_n)$的分布都存在
- 本章主要研究内容
  - 首先研究与正态总体有关的一些统计量的精确分布，如：正态总体的样本均值和方差分布、χ^2分布、t分布和F分布
  - 研究次序统计量的分布
  - 为了后面几章的需要，本章还将介绍指数分布族、充分统计量和完全统计量的定义及性质

提纲

- 正态总体样本均值和方差的分布
- χ^2分布、t分布和F分布
- 次序统计量的分布
- 统计量的极限分布
- 指数分布族
- 充分统计量
- 完全统计量

正态变量的线性组合的分布

定理：设随机变量 $X_k \sim N\left(a_k, \sigma_k^2\right)$， $k = 1,2,\cdots,n$ 且相互独立，令$c_1, c_2, \cdots, c_n$为常数，记$T = \sum_{k=1}^{n} c_k X_k$，则

$$T \sim N(\mu, \tau^2)$$

其中$\mu = \sum_{k=1}^{n} c_k a_k$，$\tau^2 = \sum_{k=1}^{n} c_k^2 \sigma_k^2$

（可通过随机变量的特征函数$\phi_k(t) = E\left[e^{itX_k}\right]$的方法来证明）

定理：正态随机变量的线性组合仍然是正态分布

- 可以很容易计算出$T = \sum_{k=1}^{n} c_k X_k$的期望和方差是$\mu$和$\tau^2$. 利用上述结论，可以很容易得到$T \sim N(\mu, \tau^2)$

- 容易得到如下两个推论

推论：若$a_1 = a_2 \cdots = a_n = a$，$\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$，则对于统计量$T = \sum_{k=1}^{n} c_k X_k$，有

$$T \sim N\left(a \sum_{k=1}^{n} c_k , \sigma^2 \sum_{k=1}^{n} c_k^2\right)$$

推论：若$c_1 = c_2 \cdots = c_n = 1/n$，$X_1, \cdots, X_n \; i.i.d. \sim N(a, \sigma^2)$，则对于统计量$T = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$，有

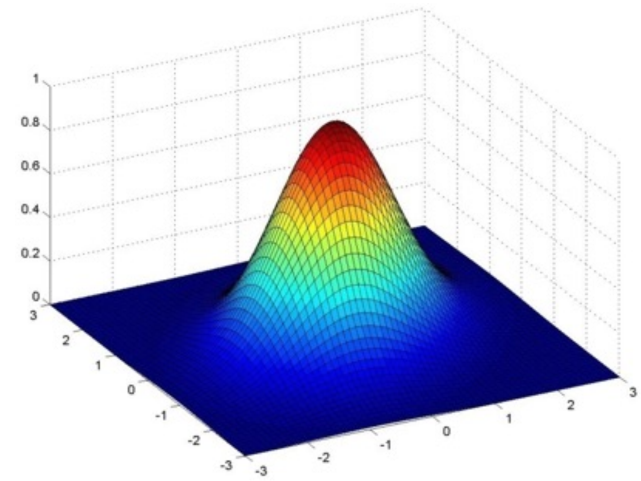
$$\bar{X} \sim N(a, \sigma^2/n)$$

补充内容——多维正态分布

- Multivariate Gaussian distribution

$$\mathcal{N}(\boldsymbol{z};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right\} \triangleq$$

  - D is the dimension
  - $\boldsymbol{\mu} \in \mathbb{R}^D$ is the mean vector
  - $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ is the covariance matrix, and $|\boldsymbol{\Sigma}|$ is its determinant



$$\boldsymbol{\Sigma} = E[(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^T]$$

- $\boldsymbol{\mu}$ controls the peak or the central point
- $\boldsymbol{\Sigma}$ controls the shapes of the distribution

- Shapes of the distributions under different kinds of $\mathbf{\Sigma}$

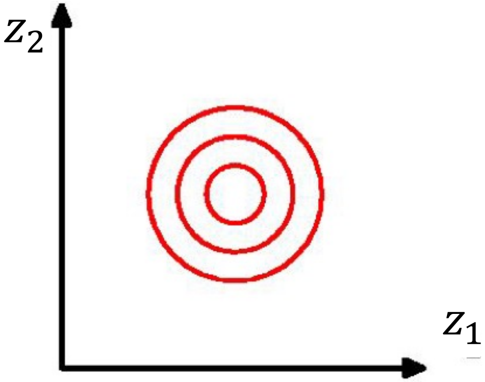
$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$\sigma_1^2 = \sigma_2^2$$

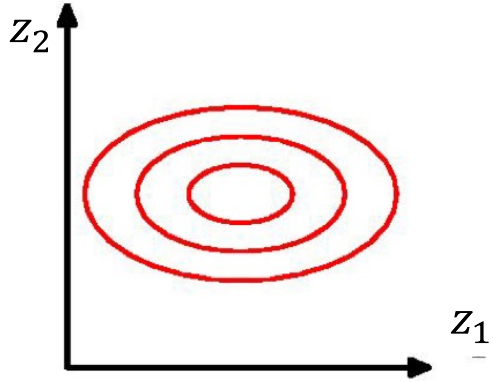
$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$\sigma_1^2 > \sigma_2^2$$

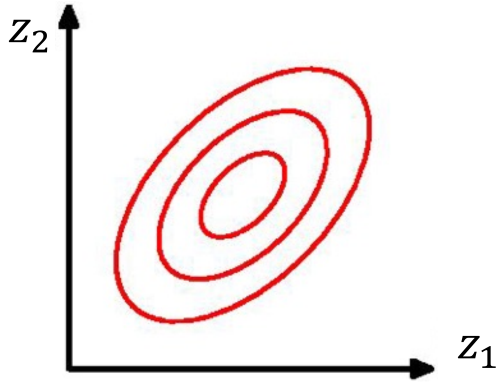
$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}$$



(a)



(b)



(c)

- No matter how $\mathbf{\Sigma}$ varies, the peak is always located at $\boldsymbol{\mu}$ (unimodal)

- For every covariance matrix $\mathbf{\Sigma}$, it can be decomposed as

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$$

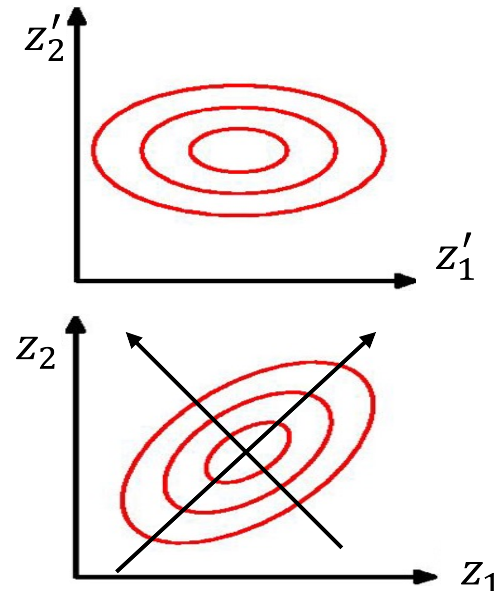
  - $\mathbf{U}$ is an orthogonal matrix, with $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$
  - $\mathbf{\Lambda}$ is a *diagonal matrix*

- By letting $\boldsymbol{z}' = \boldsymbol{U}^T\boldsymbol{z}$ and $\boldsymbol{\mu}' = \boldsymbol{U}^T\boldsymbol{\mu}$, the distribution can be expressed as

$$p(\mathbf{z}') = \frac{1}{(2\pi)^{D/2}|\mathbf{\Lambda}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}' - \boldsymbol{\mu}')^T\mathbf{\Lambda}^{-1}(\mathbf{z}' - \boldsymbol{\mu}')\right\}$$

- Thus, the shape of $p(\mathbf{z}')$ looks like

- But the shape of $p(\mathbf{z})$ is rotated by $\boldsymbol{U}$



更一般的正态随机变量线性组合分布

定理：设$\boldsymbol{X} = (X_1,\cdots,X_n)$为从正态分布$N(a,\sigma^2)$中抽取的简单随机样本, $\boldsymbol{A} = (a_{ij})$为$n \times n$的常数方阵且$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$，即：

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

则Y_i都服正态分布，且$\boldsymbol{Y}$服从多维正态分布

$$\boldsymbol{Y} \sim N(a \cdot \boldsymbol{A}\boldsymbol{1},\ \sigma^2 \cdot \boldsymbol{A}\boldsymbol{A}^T)$$

Why?

其中$\boldsymbol{1} \triangleq [1,1,\cdots,1]^T$

- $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$的期望和协方差

期望： $E[\boldsymbol{Y}] = E[\boldsymbol{A}\boldsymbol{X}] = \boldsymbol{A}E[\boldsymbol{X}] = a \cdot \boldsymbol{A}\mathbf{1}$

协方差： $Cov(\boldsymbol{Y},\boldsymbol{Y}) = E[(\boldsymbol{Y} - E[\boldsymbol{Y}])(\boldsymbol{Y} - E[\boldsymbol{Y}])^T]$

$$= E[(\boldsymbol{A}\boldsymbol{X} - E[\boldsymbol{A}\boldsymbol{X}])(\boldsymbol{A}\boldsymbol{X} - E[\boldsymbol{A}\boldsymbol{X}])^T]$$

$$= \boldsymbol{A}E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T]\boldsymbol{A}^T$$

$$= \boldsymbol{A}\, diag(\sigma^2)\, \boldsymbol{A}^T$$

$$= \sigma^{\mathbf{2}} \cdot \boldsymbol{A}\boldsymbol{A}^T$$

样本均值和样本方差的分布

定理：设$X_1, X_2, \cdots, X_n$ $i.i.d. \sim N(a, \sigma^2)$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$和$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$分别为样本均值和样本方差，那么

1) $\bar{X} \sim N(a, \sigma^2/n)$

2) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$；

3) $\bar{X}$和S^2相互独立；

证 (1) 由推论 2.2.2 立得 $\bar{X} \sim N(a, \sigma^2/n)$. 下面证 (2) 和 (3). 设

$$A = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\cdot 1}} & \frac{-1}{\sqrt{2\cdot 1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\cdot 2}} & \frac{1}{\sqrt{3\cdot 2}} & \frac{-2}{\sqrt{3\cdot 2}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}$$

为一正交阵 (这一正交阵的存在性由 Schmidt 正交化方法保证), 作正交变换 $\boldsymbol{Y} = \boldsymbol{AX}$, 其中 $\boldsymbol{Y}$ 和 $\boldsymbol{X}$ 如定理 2.2.2 所示, 故有

$$Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \sqrt{n}\bar{X},$$

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

其中，$\boldsymbol{\mu} = \begin{bmatrix} \sqrt{n}a \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, $\boldsymbol{\Sigma} = \sigma^2 \cdot \boldsymbol{I}$

由正交变换保持向量长度不变可知

$$Y_1^2 + \cdots + Y_n^2 = X_1^2 + \cdots + X_n^2.$$

所以

$$
\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \\
&= \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2.
\end{aligned}
\tag{2.2.2}
$$

由定理 2.2.2 (2) 可知 $Y_1, \cdots, Y_n$ 相互独立且 $Y_i \sim N(\mu_i, \sigma^2)$, $i=2,\cdots,n$. 再由 $\boldsymbol{A}$ 的行向量正交性得

$$
\mu_i = a\sum_{k=1}^{n} a_{ik} = \sqrt{n}a \cdot \sum_{k=1}^{n}\frac{1}{\sqrt{n}} \cdot a_{ik} = 0, \tag{2.2.3}
$$

即 $Y_2, \cdots, Y_n$ i.i.d. $\sim N(0,\sigma^2)$, 故 $Y_2/\sigma, \cdots, Y_n/\sigma$ i.i.d. $\sim N(0,1)$, 由式 (2.2.2) 得

$$
\frac{(n-1)S^2}{\sigma^2} = \sum_{i=2}^{n}(Y_i/\sigma)^2 \sim \chi^2_{n-1}.
$$

故 (2) 得证.

由上述 (2) 的证明中可知 $Y_1, Y_2, \cdots, Y_n$ 相互独立, S^2 只和 $Y_2, \cdots, Y_n$ 有关, $\bar{X}$ 只和 Y_1 有关, 因此 $\bar{X}$ 和 S^2 独立, 故 (3) 得证. 定理证毕.

提纲

- 正态总体样本均值和方差的分布
- χ^2分布、t分布和F分布
- 次序统计量的分布
- 统计量的极限分布
- 指数分布族
- 充分统计量
- 完全统计量

χ^2分布

定义：设随机变量$X_1, X_2, \cdots X_n$ $i.i.d. \sim N(0,1)$则称

$$\xi = \sum_{i=1}^{n} X_i^2$$

服从自由度为n的χ^2分布，记为$\xi \sim \chi_n^2$

- χ_n^2分布的概率密度为：

$$f(y;n) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}, & y > 0 \\ 0, & y \le 0 \end{cases}$$

其中$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x} dx$是伽玛函数，用于保证该分布是归一化的

补充

- 伽玛函数的定义：

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$$

其中$\alpha > 0$

- 伽玛函数的性质

1）对于任何正整数n，有$\Gamma(n) = (n-1)!$

2）对于任何$\alpha > 1$，有$\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$

3) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

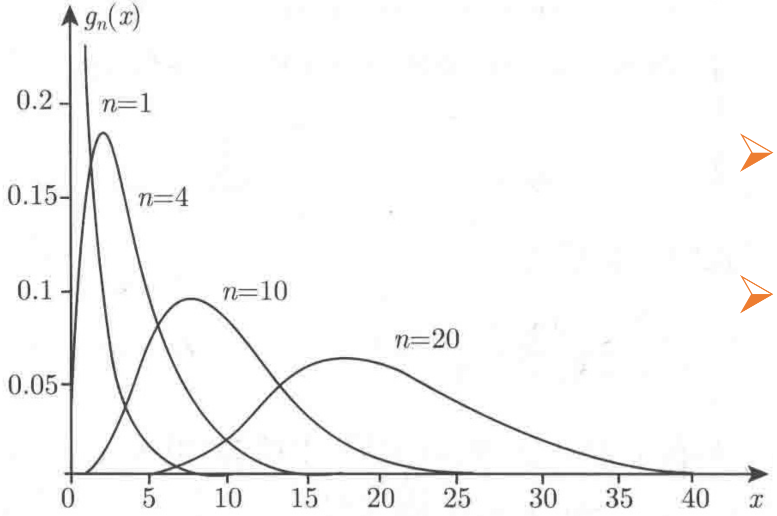


图 2.4.1　χ_n^2 的密度函数 $g_n(x)$ 形状图

- χ_n^2密度函数的支撑集为$(0, +\infty)$
- 当自由度n越大，χ_n^2的密度曲线越趋于对称

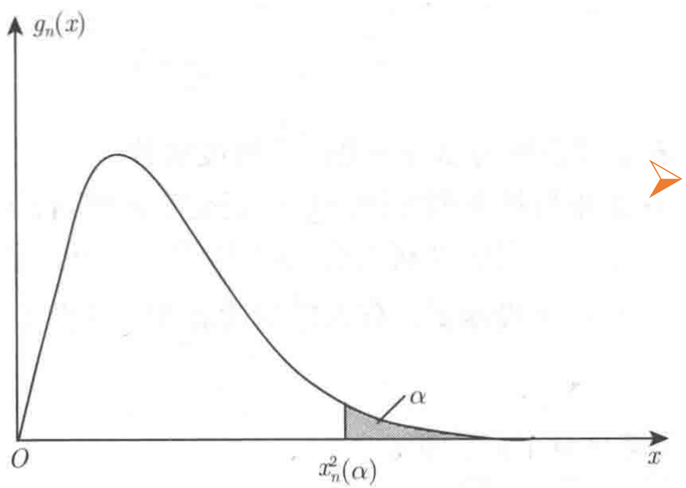


图 2.4.2　χ_n^2 的上侧 α 分位数

- 令$P(\xi > c) = \alpha$, 则称$c = \chi_n^2(\alpha)$为χ^2分布的上侧α分位数（具体数值可通过查表获得）

- 推导自由度为1的χ^2分布表达式，即：$X \sim N(0,1)$，求$Y=X^2$的分布

解：由于$Y=X^2 \geq 0$，故当$y \leq 0$时, $F_Y(y)=0$

当$y>0$时，$F_Y(y)=P(Y \leq y)=P(X^2 \leq y)$

$$= F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

$$F_X(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

因此，由$f_Y(y)=\frac{dF_Y(y)}{dy}$可得

$$f_Y(y)=\frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y})+f_X(-\sqrt{y})\right)$$

$$f_X(x)=\frac{dF(x)}{dx}=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

$$=\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}, \qquad y>0$$

- 对于有$n > 1$个自由度的卡方分布，可以类似推到得到，但复杂很多

证 由于 $X_1, X_2, \cdots, X_n$ i.i.d. $\sim N(0,1)$, 故其联合密度为

$$f(x_1, x_2, \cdots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^n x_i^2\right\}.$$

令 r.v. $\xi = \sum\limits_{i=1}^n X_i^2$ 的分布函数为 $G_n(x)$, 则有

$$G_n(x) = P\left(\sum_{i=1}^n X_i^2 < x\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \underset{\sum\limits_{i=1}^n x_i^2 < x}{\int \cdots \int} \exp\left\{-\frac{1}{2}\sum_{i=1}^n x_i^2\right\} dx_1 \cdots dx_n.$$

作 n 维球坐标变换

$$\begin{cases} x_1 = \rho\cos\theta_1\cos\theta_2\cdots\cos\theta_{n-2}\cos\theta_{n-1}, \\ x_2 = \rho\cos\theta_1\cos\theta_2\cdots\cos\theta_{n-2}\sin\theta_{n-1}, \\ \quad\cdots\cdots \\ x_n = \rho\sin\theta_1. \end{cases} \tag{2.4.2}$$

略……

- 卡方分布的性质

1) $\xi=\sum_{i=1}^{n}X_i^2$的均值和方差分别为

$$E[\xi]=n, \qquad Var(\xi)=2n$$

$$E[\xi]=E\left[\sum_{i=1}^{n}X_i^2\right]=\sum_{i=1}^{n}E[X_i^2]=n$$

$$Var[\xi]=Var\left[\sum_{i=1}^{n}X_i^2\right]=E\left[\left(\sum_{i=1}^{n}X_i^2-n\right)^2\right]$$

$$=\sum_{i=1}^{n}E\left[\left(X_i^2-1\right)^2\right]+\underbrace{\sum_{i\neq j}E\left[\left(X_i^2-1\right)\left(X_j^2-1\right)\right]}_{=0}$$

$$=\sum_{i=1}^{n}\left(E[X_i^4]-1\right)=2n$$

For $X\sim N(0,\sigma^2)$,
$E[X^{2n}]=(2n-1)\sigma^{2n}$

2） 设$Z_1 \sim \chi^2_{n_1}, Z_2 \sim \chi^2_{n_2}$且$Z_1$和$Z_2$独立，则

$$Z_1 + Z_2 \sim \chi^2_{n_1+n_2}$$

Why?

- 推论：设$X_1, X_2, \cdots, X_n$相互独立，$X_i \sim N\left(a_i, \sigma_i^2\right)$，则

$$\sum_{i=1}^{n} \left(\frac{X_i - a_i}{\sigma_i}\right)^2 \sim \chi^2_n$$

  - 原因：$\frac{X_i - a_i}{\sigma_i} \sim N(0,1)$且相互独立

- χ^2分布的更一般形式——Gamma分布$\Gamma(x;\alpha,\lambda)$

$$f(x;\alpha,\lambda)=\begin{cases}\dfrac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}$$

- 可以看出，$\Gamma\left(x;\frac{n}{2},\frac{1}{2}\right)$分布就是自由度为$n$的$\chi^2$分布

χ^2分布： $$f(y;n)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}y^{n/2-1}e^{-y/2}, & y>0\\ 0, & y\le 0\end{cases}$$

- χ^2分布可以看作是Gamma分布的一个特殊情况

t分布

定义：设$X \sim N(0,1), Y \sim \chi_n^2$, 并且$X, Y$相互独立，则称随机变量

$$T = \frac{X}{\sqrt{Y/n}}$$

服从自由度为n的t分布，记为$T \sim t_n$

- t_n分布的概率密度为

$$t_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

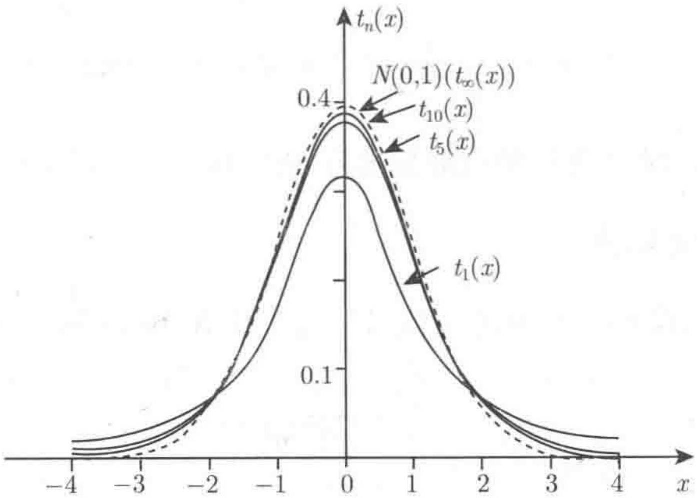


图 2.4.3 t_n 的密度函数 $t_n(x)$ 形状图

- 当$n \to \infty$时，T变量的极限分布为$N(0,1)$

- **推论**：设$X_1, X_2, \cdots, X_n \ i.i.d. \sim N(a, \sigma^2)$，则

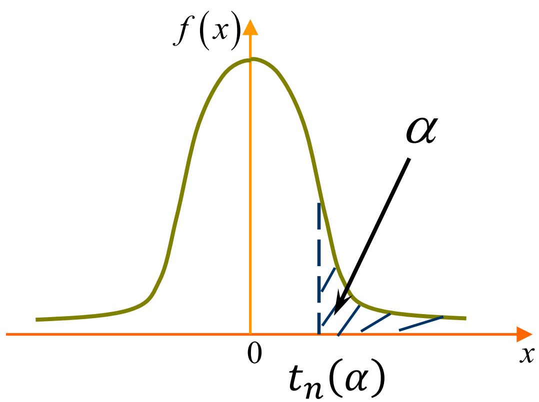
$$T = \frac{(\bar{X} - a)}{S/\sqrt{n}} \sim t_{n-1}$$

$$\frac{\bar{X} - a}{\sigma/\sqrt{n}} \sim ?$$

  - 原因：$\frac{(\bar{X}-a)}{S/\sqrt{n}} = \frac{(\bar{X}-a)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}$，而$\frac{(\bar{X}-a)}{\sigma/\sqrt{n}} \sim N(0,1)$、$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

    且$\bar{X}$与S^2相互独立

- 对给定的$\alpha\,(0 < \alpha < 1)$，称满足条件$\int_{t_n(\alpha)}^{\infty} f(t;n)dt = \alpha$的点$t_n(\alpha)$为$t_n$分布的上$\alpha$分位数。$t$分布的上$\alpha$分位数可查$t$分布表



t分布的分位数

$$t_n(1-\alpha) = -t_n(\alpha)$$

F分布

定义：设$X \sim \chi_m^2, Y \sim \chi_n^2$, 且$X,Y$相互独立，则称随机变量

$$F = \frac{X/m}{Y/n}$$

服从自由度为m和n的F分布，记为$F \sim F_{m,n}$

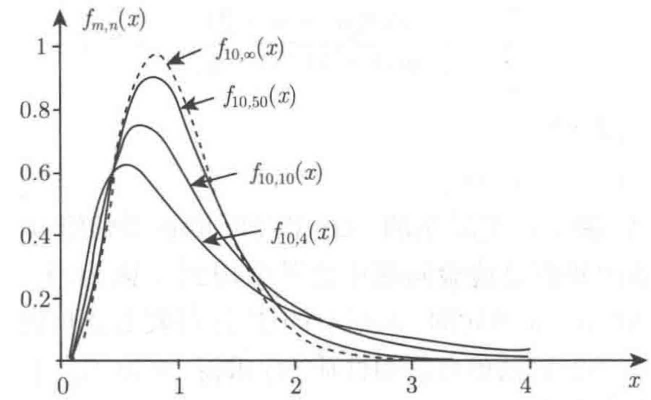


图 2.4.5 $F_{m,n}$ 的密度函数 $f_{m,n}(x)$ 形状图

- $X \sim F_{m,n}$的概率密度为

$$f_{m,n}(x) = \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1}(n+mx)^{-\frac{m+n}{2}} & x > 0 \\ 0 & x \le 0 \end{cases}$$

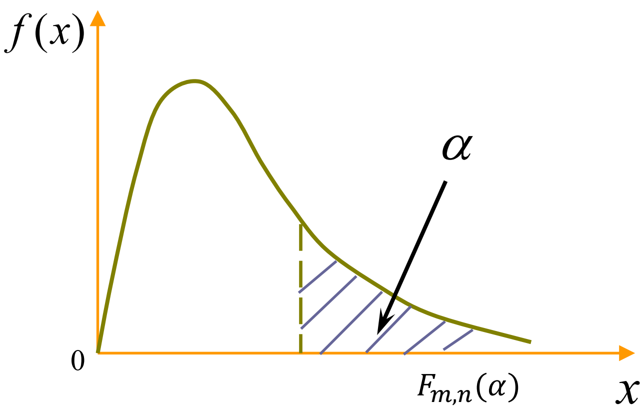
- 容易看出，如果$F \sim F_{m,n}$, 则

$$\frac{1}{F} \sim F_{n,m}$$

- 对于给定的$0 < \alpha < 1$, 称满足条件$\int_{F_{m,n}(\alpha)}^{\infty} f_{m,n}(x)dx = \alpha$的点

$F_{m,n}(\alpha)$为$F_{m,n}$分布的上α分位数

$F_{m,n}(m,n)$的值可查F分布表



F分布的分位数

$$\boxed{F_{m,n}(1-\alpha) = [F_{n,m}(\alpha)]^{-1}}$$

推论：设样本$(X_1,\cdots,X_{n_1})$和$(Y_1,\cdots,Y_{n_2})$分布来自总体$N(\mu_1,\sigma_1^2)$和$N(\mu_2,\sigma_2^2)$并且它们相互独立，其样本方差分别为S_1^2和S_2^2, 则：

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

2) $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$

3) 当${\sigma_1}^2={\sigma_2}^2=\sigma^2$时，$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$

其中$S_W=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

证明：1） $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$， $\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$

且两者独立，由F分布的定义有

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

2） $\bar{X}\sim N(\mu_1,\ \sigma_1^2/n_1)$，$\bar{Y}\sim N(\mu_2,\ \sigma_2^2/n_2)$

且$\bar{X}$与$\bar{Y}$相互独立，所以$\bar{X}-\bar{Y} \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$

即： $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,\ 1)$

3）当$\sigma_1^2=\sigma_2^2=\sigma^2$时，由2）得

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,1)$$

又由给定条件知

$$\frac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1),\qquad \frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1)$$

且它们相互独立, 故有χ^2分布的可加性知：

$$V=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$$

U与V相互独立, 于是按t分布定义知：

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}}=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{1/n_1+1/n_2}}\sim\ t(n_1+n_2-2)$$

提纲

- 正态总体样本均值和方差的分布
- χ^2分布、t分布和F分布
- 次序统计量的分布
- 统计量的极限分布
- 指数分布族
- 充分统计量
- 完全统计量

次序统计量

- 定义：从总体F 中抽取简单样本$X_1, X_2, \cdots, X_n$，按其大小排列为 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$，则称$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$或其任一子集为次序统计量

- $X_{(n)}$和 $X_{(1)}$ 的分布函数

$$
\begin{aligned}
F_n(x) &= P\left(X_{(n)} \le x\right) \\
&= P(X_1 \le x) \cdots P(X_n \le x) \\
&= [F(x)]^n
\end{aligned}
\Longrightarrow \quad f_n(x) = nf(x)[F(x)]^{n-1}
$$

$$
\begin{aligned}
F_1(x) &= P\left(X_{(1)} \le x\right) \\
&= 1 - P(X_1 > x) \cdots P(X_n > x) \\
&= 1 - [1 - F(x)]^n
\end{aligned}
\Longrightarrow \quad f_1(x) = nf(x)[1 - F(x)]^{n-1}
$$

- 定理：设总体X的密度函数为$f(x)$，分布函数为$F(x)$，$X_1, X_2, \cdots, X_n$为样本，则第k个次序统计量$X_{(k)}$的密度函数为

$$f_k(x) = \frac{n!}{(k-1)!\,(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

证：对任意的实数x，考虑次序统计量$X_{(k)}$落在区间$(x, x+\Delta x]$内这一事件。该事件等价于"样本容量为n的样本中，1个观测值落入$(x, x+\Delta x]$，$k-1$个观测值小于等于x，有$n-k$个观测值大于$x+\Delta x$"

记$X_{(k)}$的分布函数为$F_k(x)$，那么

$$F_k(x+\Delta x) - F_k(x) \approx \frac{n!}{(k-1)!\,(n-k)!}[F(x)]^{k-1}[F(x+\Delta x)-F(x)][1-F(x+\Delta x)]^{n-k}$$

$$f_k(x) = \lim_{\Delta x \to 0}\frac{F_k(x+\Delta x)-F_k(x)}{\Delta x} = \frac{n!}{(k-1)!\,(n-k)!}[F(x)]^{k-1}[1-F(x+\Delta x)]^{n-k}f(x)$$

次序统计量的联合分布

- 令$y_i = x_{(i)}$，那么

$$F(y_1,\cdots,y_n) = P\big(X_{(1)} < y_1, X_{(2)} < y_2, \cdots, X_{(n)} < y_n\big)$$

$$= \begin{cases} n!\,P(X_1 < y_1, X_2 < y_2, \cdots, X_n < y_n), & y_1 < y_2 < \cdots < y_n \\ 0, & \text{其它} \end{cases}$$

$$= \begin{cases} n!\,F(y_1)F(y_2)\cdots F(y_n), & y_1 < y_2 < \cdots < y_n \\ 0, & \text{其它} \end{cases}$$

因此，n个次序统计量$\big(X_{(1)}, X_{(2)}, \cdots, X_{(n)}\big)$的联合分布为

$$f(y_1, y_2, \cdots, y_n) = \begin{cases} n!\,f(y_1)f(y_2)\cdots f(y_n), & y_1 < y_2 < \cdots < y_n \\ 0, & \text{其它} \end{cases}$$

- n个次序统计量$\left(X_{(1)},X_{(2)},\cdots,X_{(n)}\right)$中的任意两个次序统计量 $\left(X_{(i)},X_{(j)}\right)$ $(i<j)$ 的联合分布

$$f(x,y)=\begin{cases}\dfrac{n!}{(i-1)!\,(j-i-1)!\,(n-j)!}\big(F(x)\big)^{i-1}\big(F(y)-F(x)\big)^{j-i-1} \\ \qquad\qquad\times\big(1-F(y)\big)^{n-j}f(x)f(y), & x<y \\ 0, & \text{其它}\end{cases}$$

（可通过对$f(y_1,y_2,\cdots,y_n)$积分获得）

极差的分布

- 定义：容量为n的样本的样本极差定义为

$$R = X_{(n)} - X_{(1)}$$

- 根据两个次序量的联合分布$f\left(x_{(1)}, x_{(n)}\right)$，可以得到极差$R$的概率密度分布

作如下变换

$$\begin{cases} V = X_{(j)} - X_{(i)}, \\ Z = X_{(i)} \end{cases} \Longleftrightarrow \begin{cases} X_{(i)} = Z, \\ X_{(j)} = V + Z. \end{cases}$$

得到联合分布

$$g_{ij}(v, z) = \begin{cases} \dfrac{n!}{(i-1)!(j-i-1)!(n-j)!}(F(z))^{i-1}(F(v+z)-F(z))^{j-i-1} \\ \qquad \times (1-F(v+z))^{n-j} f(z) f(v+z), & v > 0, \\ 0, & \text{其他}. \end{cases}$$

➢ 极差R的概率密度就是关于v的边缘概率密度$g_{ij}(v)$

提纲

- 正态总体样本均值和方差的分布
- χ^2分布、t分布和F分布
- 次序统计量的分布
- 统计量的极限分布
- 指数分布族
- 充分统计量
- 完全统计量

定义：当样本大小趋向无穷时，统计量的分布趋于一确定分布，则后者的分布称为统计量的极限分布，也常称为大样本分布

- 在许多情况下，统计量的精确分布很难求出、或表达式过于复杂。这时，可以研究统计量的极限分布，提供了一种近似的分布

定义：当样本容量$n \to \infty$时, 一个统计量或统计推断方法的性质称为大样本性质. 当样本大小固定时，统计量或统计推断方法的性质称为小样本性质

- 有些统计推断方法的优良性本身就是关于其极限性质，如：相合性、渐近正态性等

- 例：设$X_1,\cdots,X_n\ i.i.d.\sim\ F$，这里总体$F$有均值$a_F$和方差$\sigma_F^2$. 设$0<\sigma_F^2<\infty$，$\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$为样本均值, 求$\bar{X}_n$的大样本性质和小样本性质

  - （大样本）根据中心极限定理，当$n\to+\infty$时，有

    $$\sqrt{n}(\bar{X}_n-a_F)/\sigma_F\xrightarrow{\mathcal{L}}N(0,1)$$ 渐近正态性

  - （大样本）根据大数定理，当$n\to+\infty$时，有

    $$\bar{X}_n\xrightarrow{a.s.}a_F$$ 强相合性

  - （小样本）

    $$E[\bar{X}_n]=a_F$$ 无偏性

提纲

- 正态总体样本均值和方差的分布
- χ^2分布、t分布和F分布
- 次序统计量的分布
- 统计量的极限分布
- 指数分布族
- 充分统计量
- 完全统计量

研究指数分布族的意义

- 许多常见的分布，如：正态分布、二项分布、泊松分布、指数分布、 χ^2 分布、Gamma分布等看起来形态各异、互不相关，但它们都可以统一在一更大类分布下，即：指数分布族

- 由于指数分布族抓住这些不同分布的共性，可以研究指数分布族的统计推断问题，而不用局限于某一个具体的分布

- 很多统计推断方法的优良性对这一类范围广泛的分布族都使用

定义：设$\mathcal{F} = \{f(x,\theta): \theta \in \Theta\}$ 是定义在样本空间$\mathcal{X}$上的分布族，其中Θ为参数空间. 若其概率函数$f(x,\theta)$可表示成如下形式

$$f(x,\theta) = C(\theta)\exp\left\{\sum_{i=1}^{k} Q_i(\theta)T_i(x)\right\}h(x),$$

则称此分布族为指数型分布族(简称指数族). 其中k为自然数, $C(\theta)$ 和$Q_i(\theta)$ $(i = 1,2 \cdots, k)$是定义在参数空间Θ上的函数, $h(x)$和$T_i(x)$ $(i = 1,2,\cdots,k)$是定义在$\mathcal{X}$上与θ无关的函数，其中$C(\theta) > 0$和$h(x) > 0$

$$f(x,\theta)=C(\theta)\exp\left\{\sum_{i=1}^{k}Q_i(\theta)T_i(x)\right\}h(x),$$

- 指数分布族的特点
  - 指数族分布由三个部分组成：1）只依赖参数θ的项；2）同时依赖参数θ和变量x的项；3）只依赖变量x的项
  - 将同时依赖参数θ和变量x的项写成指数形式$\exp\{\cdot\}$后，其指数幂一定可以表示成参数θ函数$[Q_1(\theta),\cdots,Q_k(\theta)]$和变量函数$[T_1(x),\cdots,T_k(x)]$内积的形式，即：

$$\sum_{i=1}^{k}Q_i(\theta)T_i(x)$$

关键所在

  - 由定义可见指数族的支撑集 $\{x: f(x,\theta)>0\}=\{x: h(x)>0\}$与$\theta$无关，且指数族中的所有分布具有共同的支撑集

例：说明正态分布族$\mathcal{F} = \{N(\mu, \sigma^2), -\infty < \mu < +\infty, \sigma^2 > 0\}$是指数分布族

解：
$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$
$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right\}$$

若取

$$c(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}$$

$$Q_1(\theta) = -\frac{1}{2\sigma^2}, \qquad Q_2(\theta) = \frac{\mu}{\sigma^2}$$

$$T_1(x) = x^2, \qquad T_2(\theta) = x$$

可看出$N(x; \mu, \sigma^2)$属于指数型分布族

例：设$\boldsymbol{X}=(X_1,\cdots,X_n)$为从正态分布$N(\mu,\sigma^2)$中抽取的简单样本，则样本$X_1,\cdots,X_n$的联合分布是指数分布族

解 样本 $\boldsymbol{X}$ 的联合密度为

$$f(\boldsymbol{x};\mu,\sigma^2)=\left(\sqrt{2\pi}\sigma\right)^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}. \tag{2.6.2}$$

记 $\theta=(\mu,\sigma^2)$, 则参数空间为 $\Theta=\{\theta=(\mu,\sigma^2):\ -\infty<\mu<+\infty,\ \sigma^2>0\}$. 将式 (2.6.2) 改写为

$$\begin{aligned} f(\boldsymbol{x},\theta)&=\left(\sqrt{2\pi}\sigma\right)^{-n}e^{-\frac{n\mu^2}{2\sigma^2}}\exp\left\{\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right\}\\ &=C(\theta)\exp\{Q_1(\theta)T_1(\boldsymbol{x})+Q_2(\theta)T_2(\boldsymbol{x})\}h(\boldsymbol{x}), \end{aligned} \tag{2.6.3}$$

其中 $C(\theta)=(\sqrt{2\pi}\sigma)^{-n}\exp\{-n\mu^2/(2\sigma^2)\}$, $Q_1(\theta)=\mu/\sigma^2$, $Q_2(\theta)=-1/(2\sigma^2)$, $T_1(\boldsymbol{x})=\sum\limits_{i=1}^{n}x_i$, $T_2(\boldsymbol{x})=\sum\limits_{i=1}^{n}x_i^2$, $h(\boldsymbol{x})\equiv 1$. 因此, 由定义可知上述样本分布族是指数族.

例：二项分布族$\{b(n,\theta): 0<\theta<1\}$是指数族

解 设 $X\sim$ 二项分布 $b(n,\theta)$, 其概率函数为

$$\begin{aligned} p(x,\theta) &= P_\theta(X=x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}\left(\frac{\theta}{1-\theta}\right)^x(1-\theta)^n, \quad x=0,1,2,\cdots,n, \end{aligned} \tag{2.6.6}$$

其中样本空间 $\mathscr{X}=\{0,1,2,\cdots,n\}$, 参数空间 $\Theta=\{\theta: 0<\theta<1\}=(0,1)$. 将上式改写为

$$\begin{aligned} p(x,\theta) &= (1-\theta)^n \exp\left\{x\log\frac{\theta}{1-\theta}\right\}\cdot\binom{n}{x} \\ &= C(\theta)\exp\{Q_1(\theta)T_1(x)\}h(x), \end{aligned} \tag{2.6.7}$$

其中 $C(\theta)=(1-\theta)^n$, $Q_1(\theta)=\log[\theta/(1-\theta)]$, $T_1(x)=x$, $h(x)=\binom{n}{x}$, 按定义二项分布族 $\{b(n,\theta):\ 0<\theta<1\}$ 也是指数族.

例：Poisson分布族$\{p_\theta(x): \theta > 0\}$是指数族

解 设 X ~Poisson 分布 $P(\theta)$, 其概率函数为

$$
\begin{aligned}
p(x,\theta) = P_\theta(X=x) &= \frac{e^{-\theta}\theta^x}{x!} = \frac{e^{-\theta}\exp\{x\log\theta\}}{x!} \\
&= C(\theta)\exp\{Q_1(\theta)T_1(x)\}h(x), \quad x = 0,1,2,\cdots,
\end{aligned} \tag{2.6.8}
$$

其中样本空间 $\mathscr{X} = \{0,1,2,\cdots\}$, 参数空间 $\Theta = \{\theta : \theta > 0\} = (0,\infty)$, $C(\theta) = e^{-\theta}$, $Q_1(\theta) = \log\theta$, $T_1(x) = x$, $h(x) = 1/x!$, 按定义 Poisson 分布族是指数族.

例：均匀分布族$\{U(0,\theta), \theta > 0\}$和双参数指数分布族$\left\{p_{\mu,\sigma}(x) = \frac{1}{\sigma}\exp\left\{-\frac{x-\mu}{\sigma}\right\} I_{[x>\mu]}, -\infty < \mu < +\infty, \sigma > 0\right\}$是否是指数族？为什么？

不是，分布族中的分布没有共同的支撑集

指数族的自然形式及自然参数空间

- 在指数族的定义$C(\theta)\exp\{\sum_{i=1}^{k} Q_i(\theta)T_i(x)\}h(x)$中，如果用$\varphi_i$代替$Q_i(\theta)$，则将$C(\theta)$表示成$\varphi$的函数$C^*(\varphi)$，$\varphi=(\varphi_1,\cdots,\varphi_k)$，则指数族的表达式可以写成$C^*(\varphi)\exp\{\sum_{i=1}^{k}\varphi_i T_i(x)\}h(x)$

定义：如果指数族有如下形式

$$f(x,\theta)=C^*(\theta)\exp\left\{\sum_{i=1}^{k}\theta_i T_i(x)\right\}h(x)$$

则称它为指数族的自然形式. 此时，集合

$$\Theta^*=\left\{(\theta_1,\theta_2,\cdots,\theta_k):\int\exp\left\{\sum_{i=1}^{k}\theta_i T_i(x)\right\}h(x)dx<+\infty\right\}$$

称为自然参数空间

例：设$\boldsymbol{X}=(X_1,\cdots,X_n)$为从正态分布$N(\mu,\sigma^2)$中抽取的简单样本，试将其分布表示成指数族的自然形式，并给出其自然参数空间。

解 令 $\boldsymbol{\theta}=(\mu,\sigma^2)$. 由例 2.6.1 可知

$$
\begin{aligned}
f(\boldsymbol{x};\mu,\sigma^2) &= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{n\mu^2}{2\sigma^2}} \exp\left\{\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right\} \\
&= C(\theta)\exp\{Q_1(\theta)T_1(\boldsymbol{x})+Q_2(\theta)T_2(\boldsymbol{x})\}h(\boldsymbol{x}),
\end{aligned}
$$

其中 $C(\theta)=(2\pi\sigma^2)^{-n/2}\exp\{-n\mu^2/(2\sigma^2)\}$, 而 $Q_1(\theta)=\mu/\sigma^2$, $Q_2(\theta)=-1/(2\sigma^2)$ 以及 $T_1(\boldsymbol{x})=\sum_{i=1}^{n} x_i$, $T_2(\boldsymbol{x})=\sum_{i=1}^{n} x_i^2$, $h(\boldsymbol{x})\equiv 1$ 皆与式 (2.6.3) 中相同.

令 $\varphi_1 = Q_1(\theta)$, $\varphi_2 = Q_2(\theta)$, 解出 $\sigma^2 = -1/(2\varphi_2)$, $\mu^2/\sigma^2 = -\varphi_1^2/(2\varphi_2)$. 因此 $C(\theta)$ 变为 $C^*(\varphi) = \left(-\varphi_2/\pi\right)^{n/2} \exp\left\{n\varphi_1^2/(4\varphi_2)\right\}$, $\varphi = (\varphi_1, \varphi_2)$, 故式 (2.6.3) 变为下面的自然形式:

$$f(\boldsymbol{x}, \varphi) = C^*(\varphi) \exp\left\{\varphi_1 T_1(\boldsymbol{x}) + \varphi_2 T_2(\boldsymbol{x})\right\} h(\boldsymbol{x}), \tag{2.6.12}$$

其自然参数空间为

$$\Theta^* = \{\varphi = (\varphi_1, \varphi_2) : -\infty < \varphi_1 < +\infty,\ -\infty < \varphi_2 < 0\}. \tag{2.6.13}$$

例：将二项分布族表示成指数族的自然形式，并给出其自然参数空间

解：二项分布表示成指数族的形式为

$$f(x,\theta) = (1-\theta)^n \exp\left\{\log\frac{\theta}{1-\theta}\cdot x\right\}\binom{n}{x}, \qquad x = 0,1,\cdots,n$$

令$\log\frac{\theta}{1-\theta} = \varphi$，可解出$\theta = 1-\frac{1}{1+e^{\varphi}}$. 带入上式，可得

$$f(x,\varphi) = (1+e^{\varphi})^{-n}\exp\{\varphi\cdot x\}\binom{n}{x}$$

$$= C^*(\varphi)\exp\{\varphi\cdot T_1(x)\}\,h(x)$$

由于$0<\theta<1$，可知自然参数空间$\varphi = \log\frac{\theta}{1-\theta}$为

$$\Theta^* = \{\varphi: -\infty < \varphi < +\varphi\}$$

提纲

- 正态总体样本均值和方差的分布
- χ^2分布、t分布和F分布
- 次序统计量的分布
- 统计量的极限分布
- 指数分布族
- 充分统计量
- 完全统计量

充分统计量

- 统计量是对样本的简化，希望达到：
  1) 简化的程度高
  2) 信息的损失少

- 一个统计量能包含模型参数信息的多少，与统计量的具体形式有关，也依赖于问题的统计模型

- 理想情况是统计量包含了原样本所含有模型参数的全部信息（即：用统计量去概括原始数据时，并没有带来信息损失），称这样的统计量为充分统计量

- 一般来说，对原始数据$\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$加工后，得到的统计量$T(\boldsymbol{X})$会损失掉一部分原始数据 $\boldsymbol{X}$所蕴含的信息

- 但在有些情况下，加工得到的统计量$T(\boldsymbol{X})$可以包含原始数据$\boldsymbol{X}$所蕴含的全部关于模型参数θ的信息，所丢失的都是与模型参数θ无关的信息

- 举例说明

  设$\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$为从0-1分布抽取的简单样本且$P(X_i = 1) = \theta$，推断模型参数$\theta$的值

- 为了估计θ的值，可以使用原始观察值

$$P_\theta(\boldsymbol{X}=\boldsymbol{x}) = P_\theta(X_1 = x_1, \cdots, X_n = x_n)$$

$$= \theta^t(1-\theta)^{n-t}$$

其中$x_i \in \{0,1\}$，$t = \sum_{i=1}^n x_i$

- 也可以使用原始观测值$\boldsymbol{X}$的简化值 $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$

$$P_\theta(T(\boldsymbol{X}) = t) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$$

显然，$T(\boldsymbol{X})$会丢失一部分原始数据$\boldsymbol{X}$的信息。但是，通过比较上述两个概率分布，针对推断模型参数θ值的问题，直观上，使用$T(\boldsymbol{X})$来估计θ与使用原始数据 $\boldsymbol{X}$来估计 θ 并没有区别

- 样本$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$所包含的关于模型参数信息可以按如下方式来拆解

$$\{\text{样本 } \boldsymbol{X} \text{ 所包含的模型参数信息}\}$$

$$= \{T(\boldsymbol{X}) \text{ 中所含有的模型参数信息}\}$$

$$+ \{\text{在知道 } T(\boldsymbol{X}) \text{后，样本 } \boldsymbol{X} \text{ 仍含有} \theta \text{的信息}\}$$

- $T(\boldsymbol{X})$为充分统计量的条件是上述第二项信息为0，即：

$$\{\text{在知道 } T(\boldsymbol{X}) \text{后，样本 } \boldsymbol{X} \text{ 仍然含有关于} \theta \text{的信息}\} = 0$$

定义： 设样本$\boldsymbol{X}$的分布族$\{f(\theta,\boldsymbol{x}),\theta \in \Theta\}$, Θ是参数空间. 设$T=T(\boldsymbol{X})$为统计量，若在已知T的条件下，样本$\boldsymbol{X}$的条件分布与参数θ无关, 则称$T(\boldsymbol{X})$为充分统计量

例：设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为从0-1分布中抽取的简单样本，则$T(\boldsymbol{X})=\sum_{i=1}^{n}X_i$为充分统计量

解：联合分布

$$P(X_1=x_1,\cdots,X_n=x_n,T=t_0)$$

$$=\begin{cases} P(X_1=x_1,\cdots,X_n=x_n), & t_0=\sum_{i=1}^{n}x_i \\ 0, & otherwise \end{cases}$$

当 $\sum_{i=1}^{n} x_i = t_0$ 时

$$P(X_1 = x_1, \cdots, X_n = x_n | T = t_0) = \frac{P(X_1 = x_1, \cdots, X_n = x_n, T = t_0)}{P(T = t_0)}$$

$$= \frac{P\left(X_1 = x_1, \cdots, X_n = t_0 - \sum_{i=1}^{n-1} x_i\right)}{P(T = t_0)}$$

$$= \frac{\theta^{t_0}(1-\theta)^{n-t_0}}{\binom{n}{t_0}\theta^{t_0}(1-\theta)^{n-t_0}}$$

$$= \frac{1}{\binom{n}{t_0}}$$

因此，

$$P(X_1 = x_1, \cdots, X_n = x_n | T = t_0) = \begin{cases} \dfrac{1}{\binom{n}{t_0}}, & \displaystyle\sum_{i=1}^{n} x_i = t_0 \\ 0, & \displaystyle\sum_{i=1}^{n} x_i \neq t_0 \end{cases}$$

该条件分布与θ无关，因此， $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$是充分统计量

例：设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为从指数分布中抽取的简单样本，其密度函数$f(x,\theta)=\lambda e^{-\lambda x}I_{[x>0]}$, 则$T(\boldsymbol{X})=\sum_{i=1}^{N}X_i$为$\theta$的充分统计量

证：因为指数分布是Gamma分布的特例, $E(\lambda)=\Gamma(1,\lambda)$. 由Gamma分布的可加性可知

$$T\sim\Gamma(n,\lambda)\qquad\Longrightarrow\qquad f(t;\theta)=\frac{\lambda^n t^{n-1}e^{-\lambda t}}{(n-1)!}$$

另外，当$t=\sum_{i=1}^{n}x_i$, 有

$$f(x_1,\cdots,x_n,t;\theta)=f\left(x_1,\cdots,x_{n-1},t-\sum_{i=1}^{n-1}x_i\,;\theta\right)=\lambda^n e^{-\lambda t}$$

$$f(x_1,\cdots,x_n|t;\theta)=\frac{f(x_1,\cdots,x_n,t;\theta)}{f(t;\theta)}=\frac{(n-1)!}{t^{n-1}}$$

与θ无关，因此$T(\boldsymbol{X})$是参数λ的充分统计量

因子分解定理

定理：设样本$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$的概率函数$f(\boldsymbol{x},\theta)$依赖于参数$\theta$，$T=T(\boldsymbol{X})$是一个统计量，则$T$为充分统计量的充要条件是$f(\boldsymbol{x},\theta)$可以分解为如下形式

$$f(\boldsymbol{x};\,\theta)=g(t(\boldsymbol{x}),\theta)h(\boldsymbol{x})$$

- 上述充要条件的核心特点
  - $f(\boldsymbol{x};\theta)$中随机变量$\boldsymbol{x}$和模型参数θ的混合项$g(\boldsymbol{x},\theta)$一定是通过统计量$t(\boldsymbol{x})$来表示的，即可以写成$g(t(\boldsymbol{x}),\theta)$的形式
  - 只与随机变量$\boldsymbol{x}$相关的项$h(\boldsymbol{x})$不需要表示成$t(\boldsymbol{x})$的形式

证：对于任意t，令集合

$$A(t) = \{\boldsymbol{x}: T(\boldsymbol{x}) = t\}$$

充分性：由$\boldsymbol{x}, t$ 联合分布$f(\boldsymbol{x}, t; \theta)$表示的意义可知，当$t = T(\boldsymbol{x})$时，$f(\boldsymbol{x}, t; \theta) = f(\boldsymbol{x}; \theta)$；否则$f(\boldsymbol{x}, t; \theta) = 0$. 因此，$f(\boldsymbol{x}, t; \theta)$可表示成如下形式

$$f(\boldsymbol{x}, t; \theta) = f(\boldsymbol{x}; \theta) I_{[T(\boldsymbol{x})=t]}$$

若$\boldsymbol{x} \notin A(t)$，那么$f(\boldsymbol{x}|t; \theta) = 0$

若$\boldsymbol{x} \in A(t)$，那么

$$f(\boldsymbol{x}|t; \theta) = \frac{f(\boldsymbol{x}, t; \theta)}{f(t; \theta)} = \frac{g(t(\boldsymbol{x}), \theta) h(\boldsymbol{x})}{\sum_{\widetilde{\boldsymbol{x}} \in A(t)} g(t(\widetilde{\boldsymbol{x}}), \theta) h(\widetilde{\boldsymbol{x}})}$$

$$= \frac{h(\boldsymbol{x})}{\sum_{\widetilde{\boldsymbol{x}} \in A(t)} h(\widetilde{\boldsymbol{x}})}$$ 与参数θ无关

必要性：设$T(\boldsymbol{X})$是参数θ的充分统计量。由定义可知$f(\boldsymbol{x}|t;\theta)$与$\theta$无关，因此，可以将$f(\boldsymbol{x}|t;\theta)$记为$h(\boldsymbol{x})$，即：$f(\boldsymbol{x}|t;\theta)=h(\boldsymbol{x})$

对任意$\boldsymbol{x}$，当$t=T(\boldsymbol{x})$时，

$$
\begin{aligned}
f(\boldsymbol{x};\theta) &= f(\boldsymbol{x},t;\theta) \\
&= f(\boldsymbol{x}|t;\theta)f_\theta(t)
\end{aligned}
$$

其中$f_\theta(t)=\sum_x f(\boldsymbol{x},t;\theta)$

由于 $f(\boldsymbol{x}|t;\theta)$与$\theta$无关（记为$h(\boldsymbol{x})$）, 可看出

$$
\begin{aligned}
f(\boldsymbol{x};\theta) &= h(\boldsymbol{x})f_\theta(t) \\
&= h(\boldsymbol{x})f_\theta\big(T(\boldsymbol{x})\big) \\
&= h(\boldsymbol{x})g(T(\boldsymbol{x});\theta)
\end{aligned}
$$

即：$f(\boldsymbol{x};\theta)$可以分解为所给定的形式

推论：设$\boldsymbol{T}=T(\boldsymbol{X})$为$\theta$的充分统计量，$\boldsymbol{S}=\varphi(\boldsymbol{T})$是可逆函数，则$\boldsymbol{S}=\varphi(\boldsymbol{T})$也是$\theta$的充分统计量

证：由$s(\boldsymbol{x})=\varphi\big(t(\boldsymbol{x})\big)$可知

$$t(\boldsymbol{x})=\varphi^{-1}\big(s(\boldsymbol{x})\big)$$

因此，$f(\boldsymbol{x};\ \theta)=g(t(\boldsymbol{x}),\theta)h(\boldsymbol{x})$可表示成如下形式

$$f(\boldsymbol{x};\ \theta)=g\Big(\varphi^{-1}\big(s(\boldsymbol{x})\big),\theta\Big)h(\boldsymbol{x})$$

$$=\tilde{g}(s(\boldsymbol{x}),\theta)h(\boldsymbol{x})$$

由因子分解定理可知， $s(\boldsymbol{x})$是充分统计量

充分统计量的可逆变换仍然是充分统计量

例：设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为总体$b(1,\theta)$中抽取的简单样本，则$T(\boldsymbol{X})=\sum_{i=1}^n X_i$是$\theta$的充分统计量

证：
$$f(\boldsymbol{x},\theta)=P_\theta(X_1=x_1,\cdots,X_n=x_n)$$
$$=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$
$$=g(t(\boldsymbol{x}),\theta)h(\boldsymbol{x})$$

其中$h(\boldsymbol{x})=1.$

因此，根据因子分解定理可知$T(\boldsymbol{X})=\sum_{i=1}^n X_i$是充分统计量

例：设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为均匀分布$U(0,\theta)$中抽取的简单样本，则$T(\boldsymbol{X})=X_{(n)}=\max\{X_1,X_2,\cdots,X_n\}$为$\theta$的充分统计量

证：样本$\boldsymbol{X}$的联合密度为

$$f(\boldsymbol{x},\theta)=\frac{1}{\theta^n}I_{(0,\theta)}\left(x_{(n)}\right)$$

$$=g(t(\boldsymbol{x}),\theta)h(\boldsymbol{x})$$

其中$h(\boldsymbol{x})=1$；当$0<z<\theta$时，$I_{(0,\theta)}(z)=1$，否则$I_{(0,\theta)}(z)=0$

$f(\boldsymbol{x},\theta)$是否为指数分布族？

例：设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为从正态总体$N(a,\sigma^2)$中抽取的简单样本，令$\boldsymbol{\theta}=(a,\sigma^2)$，则$(\bar{X},S^2)$为$\boldsymbol{\theta}$的充分统计量。此处$\bar{X},S^2$分别表示样本均值和样本方差

证：样本$\boldsymbol{X}$的联合密度可表示为

$$f(\boldsymbol{x},\theta)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right\}$$

$$=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2-2a\sum_{i=1}^{n}x_i+na^2\right)\right\}$$

$$=g(t(\boldsymbol{x}),\boldsymbol{\theta})h(\boldsymbol{x})$$

其中$h(\boldsymbol{x})=1$，$T(\boldsymbol{X})=\left(\frac{1}{n}\sum_{i=1}^{n}X_i\,,\ \frac{1}{n-1}\sum_{i=1}^{n}X_i^2\right)=(\bar{X},S^2)$

例： 设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为从指数族中抽取的简单样本，则$T(\boldsymbol{X})=\left(T_1(\boldsymbol{X}),\cdots,T_k(\boldsymbol{X})\right)$为充分统计量

证： 指数族密度函数可以表示为

$$f(\boldsymbol{x},\theta)=C(\theta)\exp\left\{\sum_{i=1}^{k}Q_i(\theta)t_i(\boldsymbol{x})\right\}h(\boldsymbol{x})$$

$$=g(t(\boldsymbol{x}),\theta)h(\boldsymbol{x})$$

由因子分解定理可知$T(\boldsymbol{X})=\left(T_1(\boldsymbol{X}),\cdots,T_k(\boldsymbol{X})\right)$为$\theta$的充分统计量

指数族中的分布一定有充分统计量，且$T(\boldsymbol{X})=\left(T_1(\boldsymbol{X}),\cdots,T_k(\boldsymbol{X})\right)$一定是其中一个充分统计量

提纲

- 正态总体样本均值和方差的分布
- χ^2分布、t分布和F分布
- 次序统计量的分布
- 统计量的极限分布
- 指数分布族
- 充分统计量
- 完全统计量

函数的正交性

- 统计量的完全性的统计背景不像充分统计量那样好说明. 这个概念与正交函数理论中的完全性概念相似

- 函数正交：设$f(x)$和$g(x)$为定义在某有限或无限区间上的两个可积函数，若$\int f(x)g(x)dx = 0$，则称函数$f(x)$与$g(x)$正交

- 设$\{\varphi_i(x)\}_{i=1}^{\infty}$为相互正交的函数构成的集合。若此集合外的任一函数$\varphi(x)$与 $\{\varphi_i(x)\}_{i=1}^{\infty}$中的每一个函数正交，则必有$\varphi(x) \equiv 0$，那么称正交函数集 $\{\varphi_i(x)\}_{i=1}^{\infty}$ 是完备（或完全）的

统计量完全性的定义

- 定义：设$\mathcal{F}=\{f_\theta(\boldsymbol{x}),\theta\in\Theta\}$为一分布族，$\Theta$是参数空间. 设$T=T(\boldsymbol{X})$为一统计量，对任一满足

$$\mathrm{E}_{\boldsymbol{X}}\left[\varphi\left(T(\boldsymbol{X})\right)\right]=0,\ \ \forall\,\theta\in\Theta$$

的$\varphi(\cdot)$（其中$\boldsymbol{X}\sim f_\theta(\boldsymbol{x})$），如果都有如下结论

$$P_\theta\left[\varphi\left(T(\boldsymbol{X})\right)=0\right]=1,\quad \forall\,\theta\in\Theta$$

则称$T(\boldsymbol{X})$是一完全统计量

- 直白点的理解：只有当$\varphi(\cdot)=0$时，$\mathrm{E}_{\boldsymbol{X}}\left[\varphi\left(T(\boldsymbol{X})\right)\right]$对一切$\theta\in\Theta$都等于0才成立，则称$T(\boldsymbol{X})$是一完全统计量

- 统计量$T(\boldsymbol{X})$是随机变量，有概率密度分布，记为$g_\theta(t)$. 因此， “对于一切$\theta \in \Theta$有$\mathrm{E}_{\boldsymbol{X}}\left[\varphi\left(T(\boldsymbol{X})\right)\right] = 0$” 可等价表示成

$$\int \varphi(t) g_\theta(t) dt = 0，对于一切\theta \in \Theta$$

- 因此，对于一切$\theta \in \Theta$，积分$\int \varphi(t) g_\theta(t) dt = 0$表示“$\varphi(t)$与$\{g_\theta(t)\}_{\theta \in \Theta}$函数集合正交”

- 只有当$\varphi(\cdot) = 0$时，$\int \varphi(t) g_\theta(t) dt = 0 \ \forall \theta \in \Theta$才成立，说明集合$\{g_\theta(t)\}_{\theta \in \Theta}$是完备的

- 统计量的完全性实质上指的是统计量所诱导产生的分布族$\{g_\theta(t), \theta \in \Theta\}$的完全性

例：设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为从总体$b(1,\theta)$中抽取的简单样本，则$T(\boldsymbol{X})=\sum_{i=1}^n X_i$是完全统计量.

证：可以知道 $T(\boldsymbol{X})\sim b(n,\theta)$, 即：

$$P(T(\boldsymbol{X})=k)=\binom{n}{k}\theta^k(1-\theta)^{n-k}$$

设$\varphi(t)$为任一实函数. 对于任一$0<\theta<1$有$E_\theta[\varphi(T)]=0$等价于

$$\sum_{k=0}^{n}\varphi(k)\binom{n}{k}\theta^k(1-\theta)^{n-k}=0$$

$$\Leftrightarrow \sum_{k=0}^{n}\varphi(k)\binom{n}{k}\left(\frac{\theta}{1-\theta}\right)^k=0,\qquad 0<\theta<1$$

$$\Leftrightarrow \sum_{k=0}^{n}\varphi(k)\binom{n}{k}\delta^k=0,\qquad 0<\delta<+\infty$$

其中$\delta=\frac{\theta}{1-\theta}$. 上式左边是$\delta$的多项式，若系数$\varphi(k)\binom{n}{k}\neq 0$，那么最多只能有$n$个根. 因此，必定有 $\varphi(k)\binom{n}{k}=0$，即： $\varphi(k)=0$

例：设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$为从均匀分布$U(0,\theta)$中抽取的简单样本，则$T(\boldsymbol{X})=X_{(n)}=\max\{X_1,X_2,\cdots,X_n\}$为完全统计量.

证：$T(\boldsymbol{X})=X_{(n)}$的密度函数为

$$g_\theta(t)=\begin{cases} nt^{n-1}/\theta^n, & 0<t<\theta \\ 0, & \text{其他} \end{cases}$$

设$\varphi(t)$为t的任一实函数，满足$E_\theta[\varphi(T)]=0$，即：

$$\frac{n}{\theta^n}\int_0^\theta \varphi(t)t^{n-1}dt=0，对一切\theta>0都成立$$

$$\Leftrightarrow \int_0^\theta \varphi(t)t^{n-1}dt=0，对一切\theta>0都成立$$

对上式两边关于θ求导得

$$\varphi(\theta)\theta^{n-1}=0，对一切\theta>0都成立$$

$$\Leftrightarrow \varphi(\theta)=0，对一切\theta>0都成立$$

因此，$\varphi(t)=0,t>0$，所以$T(\boldsymbol{X})=X_{(n)}$为完全统计量

指数族中统计量的完全性

定理：设样本$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$的概率函数

$$f(\boldsymbol{x},\boldsymbol{\theta})=C(\boldsymbol{\theta})\exp\left\{\sum_{i=1}^{k}\theta_i T_i(\boldsymbol{x})\right\}h(\boldsymbol{x}),\qquad \boldsymbol{\theta}=(\theta_1,\cdots,\theta_k)\in\Theta^*$$

为指数族的自然形式．令$T(\boldsymbol{X})=\big(T_1(\boldsymbol{X}),\cdots,T_k(\boldsymbol{X})\big)$，若自然参数空间$\Theta^*$作为$R_k$的子集有内点，则$T(\boldsymbol{X})$是完全统计量

例：设$\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$为从总体$b(1,\theta)$中抽取的简单样本，则$T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$是完全统计量

证：样本 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$的联合分布表示成指数族的自然形式

$$f(\boldsymbol{x}, \theta) = \theta^{T(\boldsymbol{x})}(1-\theta)^{n-T(\boldsymbol{x})}$$

$$= (1+e^{\varphi})^{-n} \exp\{\varphi \cdot T(\boldsymbol{x})\}\, h(\boldsymbol{x})$$

其中$h(\boldsymbol{x}) \equiv 1$, $\varphi = \log\left(\frac{\theta}{1-\theta}\right)$. 自然参数空间为

$$\Theta^* = \{\varphi: -\infty < \varphi < +\infty\}$$

显然，自然参数空间Θ^*作为R_1的子集有内点，因此，$T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$是完全统计量

本章作业

- 1.3节： 11， 12， 13
- 1.4节： 2， 4，
- 1.5节： 1， 3， 9， 13
- 1.5节： 3， 7， 8