§2.4 常用分布与分布族

在统计学中常用的分布有很多, 本节只能介绍一些最重要的分布. 在概率论中已经学过的一些重要分布族有:

二项分布族(Binomial distribution family) $\{B(n,p) : 0 < p < 1\}$:

$$\mathsf{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \cdots, n.$$

泊松分布族(Poisson distribution family) $\{P(\lambda) : \lambda > 0\}$:

$$\mathsf{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \cdots.$$

均匀分布族(Uniform distribution family) $\{U(a,b) : -\infty < a < b < \infty\}$:

$$p(x;a,b) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a,b]; \\ 0, & \text{otherwise.} \end{cases}$$

指数分布族(Exponential distribution family) $\{E(\lambda) : \lambda > 0\}$:

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

正态分布族(Normal distribution family) $\{N(\mu,\sigma^2):-\infty<\mu<\infty,\sigma>0\}$:

$$p(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

下面介绍数理统计中的其它一些重要分布:

Γ分布族 $\{\Gamma(\alpha,\lambda):\alpha>0,\lambda>0\}$

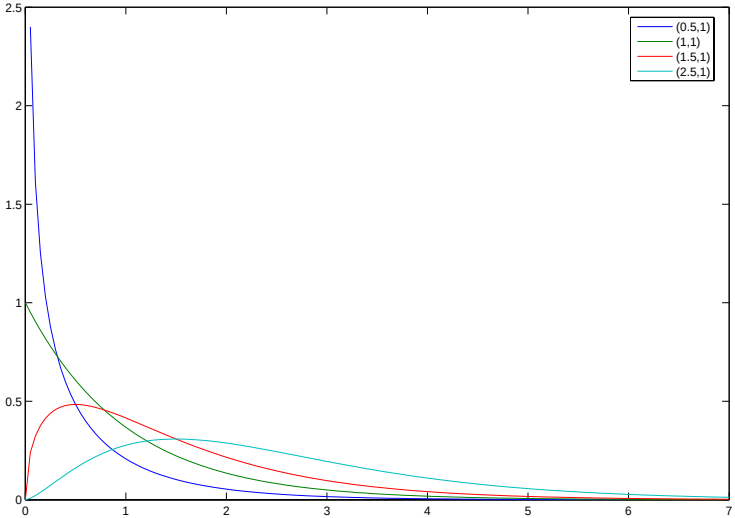
Definition

定义 具有下列密度函数的分布称为Γ分布:

$$p(x;\alpha,\lambda)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x},\qquad x>0,$$

记为$gamma(\alpha,\lambda)$、$Ga(\alpha,\lambda)$或$\Gamma(\alpha,\lambda)$, 其中$\alpha>0$称为“形状参数”, $\lambda>0$称为“尺度参数”. $(\Gamma(\alpha)=\int_0^{+\infty}t^{\alpha-1}e^{-t}dt)$

Γ分布在正值随机变量的分布中占有重要的地位. $\Gamma(\alpha,1)$称为标准Γ分布.


(0.5,1)
(1,1)
(1.5,1)
(2.5,1)
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
6
7

Γ分布的性质与“不完全Γ积分”$G(x;\alpha)=\int_0^x t^{\alpha-1}e^{-t}dt$的性质有关. 标准Γ分布的分布函数与$G(x;\alpha)$只相差一个常数因子$1/\Gamma(\alpha)$. 对一般的Γ分布$\Gamma(\alpha,\lambda)$, 其分布函数

$$\begin{aligned}F(x;\alpha,\lambda)=&\int_0^x\frac{\lambda^\alpha y^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda y}dy=\int_0^{\lambda x}\frac{t^{\alpha-1}}{\Gamma(\alpha)}e^{-t}dt\\=&\frac{G(\lambda x;\alpha)}{\Gamma(\alpha)}=F(\lambda x;\alpha,1).\end{aligned}$$

因此，只要有$\Gamma(\alpha,1)$的分布函数值$F(x;\alpha,1)$的表, 给了λ的值就可以算出一般Γ分布的分布函数值.

Γ分布的k阶矩为

$$\begin{aligned}\mathsf{E}X^k =& \int_0^\infty \frac{\lambda^\alpha x^{k+\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} dx \\ =& \frac{\Gamma(k+\alpha)}{\lambda^k \Gamma(\alpha)} \int_0^\infty \frac{\lambda^{k+\alpha} x^{k+\alpha-1}}{\Gamma(k+\alpha)} e^{-\lambda x} dx \\ =& \frac{\Gamma(k+\alpha)}{\lambda^k \Gamma(\alpha)}.\end{aligned}$$

利用$\Gamma(\alpha)$ 的性质：

$$\Gamma(\alpha+k)=(\alpha+k-1)(\alpha+k-2)...\alpha\Gamma(\alpha),$$

不难验证Γ分布的均值和方差分别为

$$E(\Gamma(\alpha,\lambda))=\frac{\alpha}{\lambda},\qquad \mathsf{Var}(\Gamma(\alpha,\lambda))=\frac{\alpha}{\lambda^2}.$$

Γ分布的负指数阶矩:

$$\begin{aligned}\mathsf{E}X^{-\beta} =& \int_0^\infty \frac{\lambda^\alpha x^{-\beta+\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} dx \\ =& \lambda^\beta \frac{\Gamma(\alpha-\beta)}{\Gamma(\alpha)} \int_0^\infty \frac{\lambda^{\alpha-\beta} x^{\alpha-\beta-1}}{\Gamma(\alpha-\beta)} e^{-\lambda x} dx \\ =& \frac{\lambda^\beta \Gamma(\alpha-\beta)}{\Gamma(\alpha)}, \quad \beta < \alpha.\end{aligned}$$

Γ分布的矩母函数为

$$\begin{aligned}\mathsf{E}e^{tX} &= \int_0^\infty \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{tx-\lambda x} dx \\ &= \left(\frac{\lambda}{\lambda-t}\right)^\alpha \int_0^\infty \frac{(\lambda-t)^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-(\lambda-t)x} dx \\ &= (1-\frac{t}{\lambda})^{-\alpha}, \quad t<\lambda.\end{aligned}$$

Γ分布的特征函数为

$$\varphi(t) = \mathsf{E}e^{itX} = (1-\frac{it}{\lambda})^{-\alpha}.$$

Theorem

定理 设随机变量X_1, X_2 相互独立，$X_i \sim \Gamma(\alpha_i, \lambda)$, $i = 1, 2$,
则$X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.

Theorem

定理 设$X \sim \Gamma(\alpha, \lambda)$, 则$Y = X/k \sim \Gamma(\alpha, k\lambda)$, 其中$k > 0$.

$\Gamma(1, \lambda)$ 为指数分布$E(\lambda)$. 当n为正整数时, $\Gamma(n, \lambda)$ 可看成n个独立、具有相同刻度参数的指数分布变量的和.

Example

某种电子产品能经受外界若干次冲击,可当第k次冲击来到的时刻产品就失效了. 这样,该产品的寿命就是第k次冲击来到的时刻. 假设在$(0,t)$时间内产品受到的冲击次数X(t)服从如下的Poisson分布:

$$\mathsf{P}(X(t)=x)=\frac{(\lambda t)^x}{x!}e^{-\lambda t},\quad x=0,1,\cdots.$$

求该产品寿命T的分布.

解: T的分布函数如下: 对$t>0$, 有

$$\begin{aligned}F(t)=&1-\mathsf{P}(T\geq t)=1-\mathsf{P}(\text{产品在}(0,t)\text{时间内没失效})\\=&1-\mathsf{P}(X(t)\leq k-1)=1-\sum_{x=0}^{k-1}\mathsf{P}(X(t)=x)\\=&\sum_{x=k}^{\infty}\frac{(\lambda t)^x}{x!}e^{-\lambda t}.\end{aligned}$$

求导,得T的密度函数为

$$f(t)=\frac{\lambda^k}{(k-1)!}t^{k-1}e^{-\lambda t}=\frac{\lambda^k}{\Gamma(k)}t^{k-1}e^{-\lambda t},\quad t>0.$$

即$T\sim\Gamma(k,\lambda)$. 特别, $k=1$时, 它是$E(\lambda)$.

$\Gamma(\alpha,\lambda)$ 的另一个重要特例是$\Gamma(n/2,1/2)$, 它就是$\chi^2(n)$分布.

χ^2分布

Definition

定义 设$X_1,X_2,...,X_n$, i.i.d.$\sim N(0,1)$，则

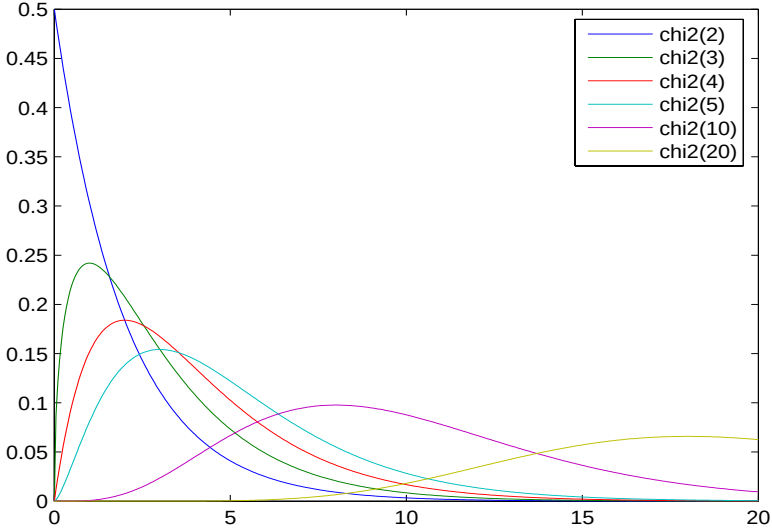
$$\xi=\sum_{i=1}^{n}X_i^2$$

的分布定义为具有自由度n的χ^2分布，记为$\xi\sim\chi^2(n)$或$\xi\sim\chi_n^2$.

χ^2分布是刻画正态变量二次型的一种重要分布，它有一系列重要而应用广泛的性质. 这里先介绍一些基本性质.

χ^2分布的密度为

$$p(x;n)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2},\quad x>0, \tag{1}$$



下面先推导$\chi^2(n)$的密度函数确实如(1)所示.

设$X_1, X_2, ..., X_n$, i.i.d.,$\sim N(0,1)$.

我们只要证$X_1^2 + X_2^2 + ... + X_n^2 \sim \Gamma(n/2, 1/2)$.

记$Y = X_1^2$的分布函数为$F_Y(y)$, 则当$y \leq 0$时, $F_Y(y) = 0$. 而当$y > 0$时,

$$\begin{aligned} F_Y(y) =& \mathsf{P}(X_1^2 < y) = \mathsf{P}(-\sqrt{y} < X_1 < \sqrt{y}) \\ =& F_{X_1}(\sqrt{y}) - F_{X_1}(-\sqrt{y}). \end{aligned}$$

而密度函数为

$$\begin{aligned} p_Y(y) =& [p_{X_1}(\sqrt{y}) + p_{X_1}(-\sqrt{y})]/(2\sqrt{y}) \\ =& \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}. \end{aligned}$$

所以$Y = X_1^2 \sim \Gamma(1/2, 1/2)$. 由样本的代表性,
知$X_i^2 \sim \Gamma(1/2, 1/2), i = 1, 2, \cdots, n$. 由于$X_1, X_2, \cdots, X_n$独立同分布, 并结合$\Gamma$ 分布的可加性得$X_1^2 + \cdots + X_n^2 \sim \Gamma(n/2, 1/2)$.

再次利用Γ 分布的可加性,还可得χ^2分布的可加性, 即

Corollary

推论 设$\xi_1 \sim \chi^2(n_1)$, $\xi_2 \sim \chi^2(n_2)$, 且相互独立，则$\xi_1 + \xi_2 \sim \chi^2(n_1 + n_2)$.

此性质可以推广到任意有限个相加.

另外,
设$\xi \sim \chi^2(n)$, 易得$E\xi = n$, $Var\xi = 2n$.
χ^2 分布的特征函数为

$$\varphi(t; n) = (1 - 2it)^{-n/2}. \tag{2}$$

非中心χ^2分布

Definition

定义 设$X_1, ..., X_n$相互独立, $X_i \sim N(a_i, 1)$, a_i $(i = 1, \ldots, n)$ 不全为0. 则

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

的分布定义为具有自由度n、非中心参数为$\lambda = a_1^2 + \cdots + a_n^2$(与书上不同)的非中心$\chi^2$分布，记为$\xi \sim \chi^2(n, \lambda)$或$\xi \sim \chi^2_{n,\lambda}$.

非中心χ^2分布的性质:

1. 若$\xi \sim \chi^2_{n,\lambda}$, 则$\xi$的特征函数为

$$\varphi(t;n) = (1-2it)^{-n/2} \exp\{\frac{it\lambda}{1-2it}\}.$$

2. 若$\xi_j \sim \chi^2_{n_j,\lambda_j}$, $j = 1, \ldots, k$, 且相互独立, 则$\sum_{j=1}^k \xi_j \sim \chi^2_{n,\lambda}$, 其中$n = \sum_{j=1}^k n_j$, $\lambda = \sum_{j=1}^k \lambda_j$;
3. 若$\xi \sim \chi^2_{n,\lambda}$, 则$\mathsf{E}\xi = n + \lambda$, $\mathsf{Var}(\xi) = 2(n + 2\lambda)$.

证明: (2) 由(1)即得, 或直接利用非中心χ^2分布的定义也可得到. 为证明(1)和(3), 由定义和独立性, 只需考虑$n=1$的情形, 这时可记$\xi=(\eta+\delta)^2$, 其中$\eta\sim N(0,1)$, $\delta=\sqrt{\lambda}$. 那么

$$\begin{aligned}\mathsf{E}\xi =&\mathsf{E}[\eta^2+2\delta\eta+\delta^2]=1+\delta^2=1+\lambda,\\ \mathsf{E}\xi^2 =&\mathsf{E}\left[\eta^4+4\eta^3\delta+6\eta^2\delta^2+4\eta\delta^3+\delta^4\right]\\ =&3+6\delta^2+\delta^4.\end{aligned}$$

因此

$$\mathsf{Var}(\xi)=\mathsf{E}\xi^2-(\mathsf{E}\xi)^2=2+4\delta^2=2(1+2\lambda).$$

(3) 得证.

$\xi=(\delta+\eta)^2$ 的矩母函数为

$$\begin{aligned}M(t)=&\mathsf{E}e^{t\xi}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{t(y+\delta)^2}e^{-\frac{y^2}{2}}dy\\=&\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}exp\{-(\frac{1}{2}-t)(y-\frac{\delta t}{\frac{1}{2}-t})^2+\frac{\delta^2t^2}{\frac{1}{2}-t}+t\delta^2\}dy\\=&\cdots=(1-2t)^{-1/2}\exp\left\{\frac{t\delta^2}{1-2t}\right\}\\=&(1-2t)^{-1/2}\exp\left\{\frac{t\lambda}{1-2t}\right\},t<\frac{1}{2}.\end{aligned}$$

所以ξ的特征函数为

$$M(it)=(1-2it)^{-1/2}\exp\left\{\frac{it\lambda}{1-2it}\right\}.$$

t分布(Student's t distribution)

Definition

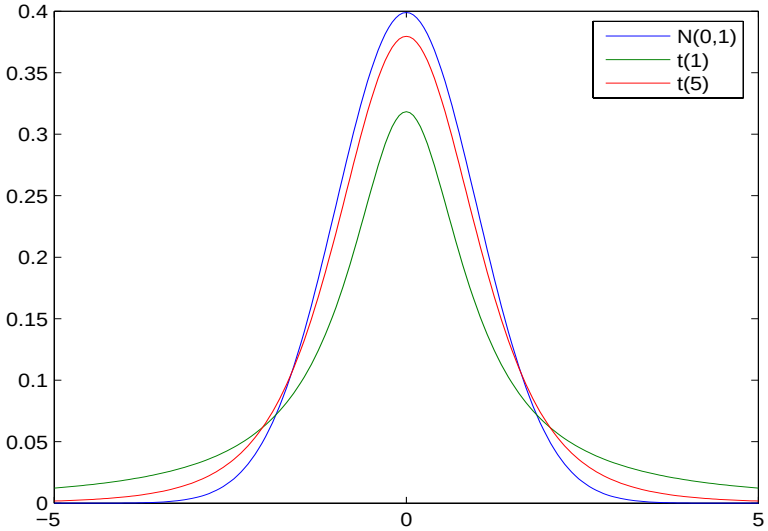
定义 设$X \sim N(0,1)$, $K \sim \chi^2(n)$, 且X 与K 相互独立, 则

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

的分布定义为自由度为n的t分布，记为$T \sim t(n)$或$T \sim t_n$.

$t(n)$分布的密度为

$$p(t;n) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1+\frac{t^2}{n})^{-(n+1)/2}, \quad -\infty < t < \infty.$$


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−5
0
5
N(0,1)
t(1)
t(5)

t分布与标准正态分布非常相似, 由t分布的密度式不难看出，t分布与标准正态分布有相似之处：它的密度也是以原点为对称中心的“钟形”曲线. 英国统计学家哥塞特(W. S. Gosset) 于1908年首先发现了这个分布, 并以学生(Student)的笔名发表了他的研究结果. 因此, t分布又称为“学生氏”分布.

William Gosset(1876-1937)

- 1908年提出t-分布

t分布的发现, 与正态分布$N(\mu,\sigma^2)$的均值μ的估计与检验问题相关. 如果$X_1,\ldots,X_n$是取自$N(\mu,\sigma^2)$ 的样本, 那么

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}\sim N(0,1)\ \text{ 即}\overline{X}-\mu\sim N(0,\sigma^2/n).$$

当σ已知时, 由上式可以大致知道μ与$\overline{X}$的距离.
当σ未知时, $\overline{X}-\mu$的分布与未知参数σ相关, 无法进行统计推断. 自然的办法是用样本方差S^2代替总体方差σ^2. 问题是, 这时对应的分布是什么? 即

$$T=:\frac{\sqrt{n}(\overline{X}-\mu)}{S}\sim?$$

统计学家E.S. Pearson一直认为仍然是正态分布.

当样本容量n比较大时, Pearson的结论差不多是正确的, 因为

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S} \xrightarrow{D} N(0,1), \quad n \to \infty.$$

相当一段时间没有人敢怀疑Pearson 的论断, 而且在实践中也是这样使用的. 但是, W. S. Gosset 恰好遇到的是小样本问题, 他发现, 在小样本场合用标准正态分布来近似T的分布效果不好，会低估误差. 他导出了T的分布是自由度为$n-1$的t 分布(见下面定理). 这个发现导致了对抽样分布的深入研究,并产生了丰富的成果. 因此, t分布的发现被认为是统计学发展史上的一件大事. 由此发现了t分布.

Theorem

定理T1 设$X_1, ..., X_n$为来自总体$X \sim N(\mu, \sigma^2)$的样本, $\overline{X}$为样本均值,S^2为样本方差, 则

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

定理T1的证明 记

$$T = \frac{n^{1/2}(\overline{X} - \mu)/\sigma}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} = \frac{U}{\sqrt{K/(n-1)}},$$

其中$U = n^{1/2}(\overline{X} - \mu)/\sigma$, $K = (n-1)S^2/\sigma^2$. 由定理2.2.3, $U \sim N(0,1)$, $K \sim \chi^2(n-1)$, 且U与K相互独立. t 分布的定义知$T \sim t(n-1)$.

Theorem

定理T2 设$X_1, X_2, ..., X_m$, $i.i.d. \sim N(\mu_X, \sigma^2)$, $Y_1, Y_2, ..., Y_n$ $i.i.d. \sim N(\mu_Y, \sigma^2)$ *(即两个总体的方差相等)*, 且$X_1, X_2, ..., X_m$与$Y_1, Y_2, ..., Y_n$ 相互独立.
记$\overline{X}, S_X^2$分别为$X_1, X_2, ..., X_m$ 的样本均值和样本方差, $\overline{Y}, S_Y^2$为$Y_1, Y_2, ..., Y_n$的样本均值和样本方差. 并记

$$S_W^2 = \frac{1}{m+n-2}\{(m-1)S_X^2 + (n-1)S_Y^2\},$$

则有

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_W\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2).$$

定理T2的证明 由定理2.2.3, 可知

$$\overline{X} \sim N(\mu_X, \sigma^2/m),$$

$$\overline{Y} \sim N(\mu_Y, \sigma^2/n),$$

由$(X_1, ..., X_m)$与$(Y_1, ..., Y_n)$相互独立知$\overline{X}$与$\overline{Y}$相互独立, 因此

$$\overline{X} - \overline{Y} \sim N(\mu_X - \mu_Y, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}).$$

即

$$U := \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0, 1).$$

再次利用定理2.2.3, 得

$$(m-1){S_X}^2/\sigma^2 \sim \chi^2(m-1),$$

$$(n-1){S_Y}^2/\sigma^2 \sim \chi^2(n-1),$$

且由$(X_1,...,X_m)$与$(Y_1,...,Y_n)$相互独立知${S_X}^2$ 与${S_Y}^2$相互独立, 结合χ^2分布的可加性, 可知

$$V := \frac{(m-1){S_X}^2 + (n-1){S_Y}^2}{\sigma^2} \sim \chi^2(m+n-2).$$

再由$(X_1,...,X_m)$与$(Y_1,...,Y_n)$的独立性以及定理2.2.3的第三个结果, 可知U与V独立.

利用T分布的定义, 可知

$$\frac{U}{\sqrt{V/(m+n-2)}} \sim t(m+n-2).$$

结合记号U和V的含义, 即得定理的结论.

Theorem

定理T3 当$n \to \infty$时, $t(n)$依分布收敛到$N(0,1)$分布.

这说明, 当n足够大时, t分布与标准正态分布没有什么太大的区别.

当n较小时, t分布与标准正态分布的区别还是不能忽略的. 注意观察t分布的密度式可以看出, 当$|x| \to \infty$时,$p(x;n)$是$|x|^{-(n+1)}$数量级的; 而标准正态分布的密度函数为$e^{-x^2/2}$数量级的. 我们可以形象地说: t分布的“尾重”; 而标准正态分布的“尾轻”.

t分布的矩:

只有当$r < n$ $(n > 1)$时，r阶矩才存在. $t(n)$的密度函数是偶函数, 故其奇数阶矩为0,而偶数阶矩为

$$\begin{aligned}\mathsf{E}T^r =& \mathsf{E}N(0,1)^r \mathsf{E}\left(\chi^2(n)/n\right)^{-r/2} \\ =& \frac{n^{\frac{r}{2}}}{\sqrt{\pi}} \frac{\Gamma(\frac{r+1}{2})\Gamma(\frac{n-r}{2})}{\Gamma(\frac{n}{2})}, \quad r < n \text{ 为偶数.}\end{aligned}$$

特别地

$$\mathsf{E}(t(n)) = 0, \quad n \geq 2;$$

$$\mathsf{Var}(t(n)) = \frac{n}{n-2}, \quad n \geq 3.$$

因此, t分布的方差（当存在时）比标准正态分布的方差大.

非中心t分布

Definition

定义 设$X \sim N(\delta, 1)$, $K \sim \chi^2(n)$, 且X 与K 相互独立, 则

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

的分布定义为自由度为n、非中心参数为δ的非中心t分布，记为$T \sim t(n, \delta)$或$T \sim t_{n,\delta}$.

F分布

(Snedecor's F distribution/Fisher – Snedecor distribution)

Definition

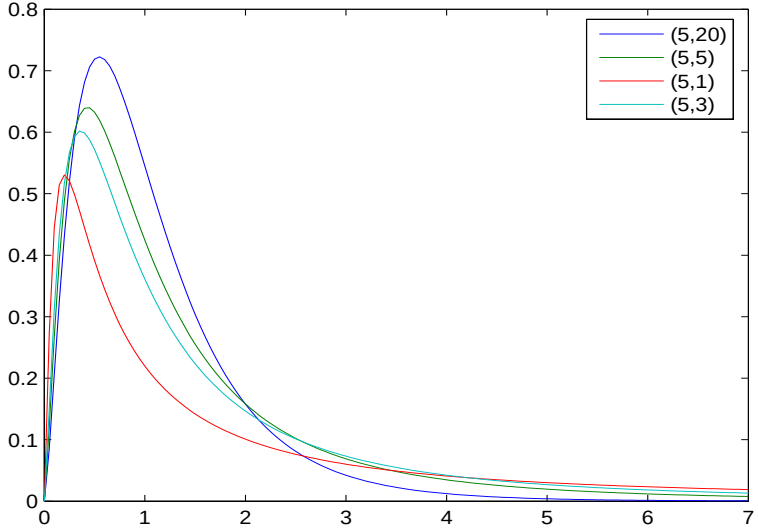
定义 设$K_1 \sim \chi^2(m)$, $K_2 \sim \chi^2(n)$, 且K_1与K_2相互独立, 则

$$F = \frac{K_1/m}{K_2/n}$$

的分布定义为具有自由度(m, n)(或称为第一自由度为m, 第二自由度为n)的F分布, 记为$F \sim F(m, n)$或$F \sim F_{m,n}$.

$F(n,m)$的密度函数为

$$p(x;n,m)=\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}m^{m/2}n^{n/2}\frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}},\qquad x>0.$$



性质: 若$F \sim F(n,m)$, 则$1/F \sim F(m,n)$.

F分布的主要用途是在方差分析中. 在两正态总体的方差比检验中要用到下面的定理.

Theorem

定理F1 设$X_1, X_2, ..., X_m$, $i.i.d.\sim N(\mu_X, \sigma_X^2)$, $Y_1, Y_2, ..., Y_n$ $i.i.d.$ $\sim N(\mu_Y, \sigma_Y^2)$, 且$X_1, X_2, ..., X_m$与$Y_1, Y_2, ..., Y_n$ 相互独立.
记S_X^2为$X_1, X_2, ..., X_m$ 的样本方差, S_Y^2 为$Y_1, Y_2, ..., Y_n$的样本方差. 则

$$F = \frac{{S_X}^2/\sigma_X^2}{{S_Y}^2/\sigma_Y^2} \sim F(m-1, n-1).$$

定理F1的证明 由定理2.2.3, 可知

$$K_1 = (m-1){S_X}^2/\sigma_X^2 \sim \chi^2(m-1),$$

$$K_2 = (n-1){S_Y}^2/\sigma_Y^2 \sim \chi^2(n-1),$$

且由$X_1, ..., X_m$与$Y_1, ..., Y_n$相互独立知${S_X}^2$ 与${S_Y}^2$相互独立，因而K_1与K_2相互独立. 由F分布定义有

$$F = \frac{K_1/(m-1)}{K_2/(n-1)} \sim F(m-1, n-1).$$

非中心F分布

Definition

定义 设$K_1 \sim \chi^2(m, \lambda)$, $K_2 \sim \chi^2(n)$,且K_1与K_2相互独立, 则

$$F = \frac{K_1/m}{K_2/n}$$

的分布定义为具有自由度(m, n)和非中心参数为λ的非中心F分布.记为$F \sim F(m, n, \lambda)$或$F \sim F_{m,n,\lambda}$.

分布的上α分位点

Definition

定义 设X的概率密度函数为$f(x)$, 对于给定的正数α, $0 < \alpha < 1$, 若存在实数x_α满足

$$\mathsf{P}(X \geq x_\alpha) = \int_{x_\alpha}^{+\infty} f(x)dx = \alpha, \qquad (*)$$

则称点x_α为X的上侧α分位点(或上侧α分位数), 简称上α分位点(或上α分位数).

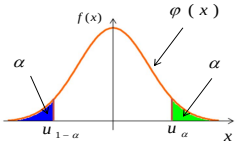
若X服从某分布, 则称x_α为该分布的上α分位点.

标准正态分布的上α分位数

设$X \sim N(0,1)$, 对于给定的α, $0<\alpha<1$, 称满足条件

$$P\{X \geq u_\alpha\} = \int_{u_\alpha}^{\infty} \varphi(x)\,dx = 1-\Phi(x)$$

的点u_α为标准正态分布的上α分位数, u_α值可查标准正态分布表.



标准正态分布的上α分位数

$$u_{1-\alpha} = -u_\alpha$$

标准正态分布函数表http://en.wikipedia.org/wiki/Standard_normal_table

z	+0.00	+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53980	0.54380	0.54776	0.55172	0.55567	0.55966	0.56360	0.56749	0.57142	0.57535
0.2	0.57930	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900

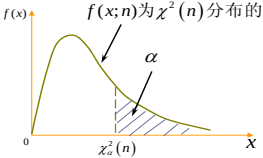
$\Phi(1.96) = 0.975 \Leftrightarrow u_{0.025} = 1.96, \quad \Phi(1.645) = 0.95 \Leftrightarrow u_{0.05} = 1.645$,等等.

χ^2分布的上α分位数

对给定的$\alpha, 0<\alpha<1$,称满足条件

$$\int_{\chi_\alpha^2(n)}^{\infty} f(x;n)dy=\alpha$$

的点$\chi_\alpha^2(n)$为$\chi^2(n)$分布的上α分位数,其中$f(x;n)$为$\chi^2(n)$分布的概率密度函数.上α分位数$\chi_\alpha^2(n)$的值可查χ^2分布表得到.



χ^2分布的上α分位数

当$n>45$时,有

$\chi_\alpha^2(n)\approx\frac{1}{2}(u_\alpha+\sqrt{2n-1})^2$,

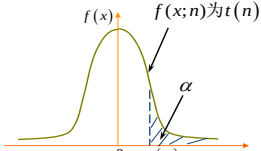
其中u_α为标准正态分布的上α分位数.

t分布的上α分位数

对给定的α，$0<\alpha<1$,称满足条件

$$\int_{t_\alpha(n)}^{\infty} f(x;n)\,dx=\alpha$$

的点$t_\alpha(n)$为$t(n)$分布的上α分位数，其中$f(x;n)$为$t(n)$分布的概率密度函数. 上α分位数$t_\alpha(n)$可查t分布表得到.



t分布的上α分位数

$$t_{1-\alpha}(n)=-t_\alpha(n)$$

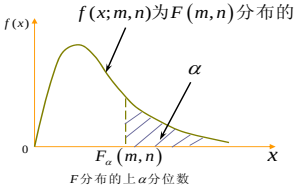
当$n>45$时，有
$t_\alpha(n)\approx u_\alpha$.

F分布的上α分位数

对于给定的α，$0<\alpha<1$,称满足条件

$$\int_{F_\alpha(m,n)}^{\infty} f(x;m,n)\,dx=\alpha$$

的点$F_\alpha(m,n)$为$F(m,n)$分布的上α分位数, 其中$f(x;m,n)$为$F(m,n)$的概率密度函数. $F_\alpha(m,n)$的值可查F分布表.



F分布的上α分位数

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

设$F \sim F(m,n)$, 由于$\mathsf{P}(F>0)=1$, 且结合上α分位数的定义知

$$1-\alpha = \mathsf{P}\big(F \geq F_{1-\alpha}(m,n)\big) = \mathsf{P}\Big(\frac{1}{F} \leq \frac{1}{F_{1-\alpha}(m,n)}\Big).$$

且注意到F为连续型的随机变量, 所以

$$\alpha = 1-\mathsf{P}\Big(\frac{1}{F} \leq \frac{1}{F_{1-\alpha}(m,n)}\Big) = \mathsf{P}\Big(\frac{1}{F} > \frac{1}{F_{1-\alpha}(m,n)}\Big) = \mathsf{P}\Big(\frac{1}{F} \geq \frac{1}{F_{1-\alpha}(m,n)}\Big),$$

由F分布的性质可知$\dfrac{1}{F} \sim F(n,m)$, 故$\frac{1}{F_{1-\alpha}(m,n)}$为$F(n,m)$的上$\alpha$分位数, 即$\frac{1}{F_{1-\alpha}(m,n)} = F_{\alpha}(n,m)$, 因此

$$F_{1-\alpha}(m,n) = \frac{1}{F_{\alpha}(n,m)}.$$

beta分布

Definition

定义 具有下列密度函数的分布称为beta分布:

$$p(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},\qquad 0<x<1.$$

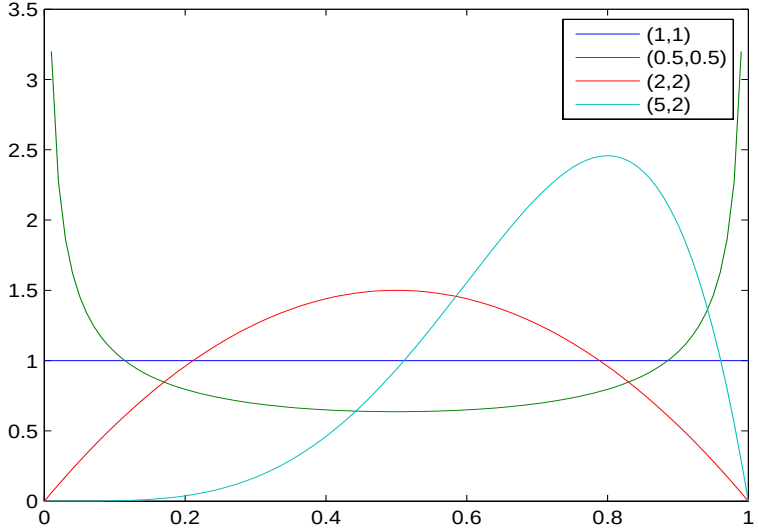
其中$a>0$, $b>0$为形状参数.

由于$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$,所以beta分布的密度函数可写为:

$$p(x;a,b)=\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1},\qquad 0<x<1,$$

其中$B(a,b)=\int_0^1 x^{a-1}(1-x)^{b-1}dx$.

以后我们用$beta(a,b)$或$Be(a,b)$或$\beta(a,b)$来记beta分布.



β分布的k阶矩为

$$\mathsf{E}X^k = \frac{B(a+k,b)}{B(a,b)} = \frac{\Gamma(a+b)\Gamma(a+k)}{\Gamma(a)\Gamma(a+b+k)}.$$

β分布的均值和方差分别为

$$\mathsf{E}(beta(a,b)) = \frac{a}{a+b},$$

$$\mathsf{Var}(beta(a,b)) = \frac{ab}{(a+b)^2(a+b+1)}.$$

在β分布中，

当$a=b=1$时, 我们就得到$(0,1)$区间上的均匀分布$U(0,1)$.

当$a=b=1/2$时, 我们就得到$\beta(\frac{1}{2},\frac{1}{2})$的概率密度函数为

$$p(x)=\frac{1}{\pi\sqrt{x(1-x)}},\quad 0<x<1,$$

此分布为反正弦分布.

Example

设$X_1, X_2, \cdots, X_n$ i.i.d. $\sim U(0,1)$. 则

$$X_{(k)} \sim \beta(k, n-k+1).$$

事实上, $X_{(k)}$的密度函数为

$$p(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, \quad 0 < x < 1.$$

一般地, 设$X \sim F(x)$, $F(x)$是连续函数, 可以证明, $F(X) \sim U(0,1)$. 因此

$$F(X_{(k)}) \sim \beta(k, n-k+1).$$

Fisher Z分布(βII型分布)

Definition

定义 具有下列密度函数的分布称为Z分布:

$$p(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{x^{a-1}}{(1+x)^{a+b}},\qquad x>0.$$

其中$a>0,\ b>0$. 记为$Z(a,b)$.

Z分布的密度函数可写为:

$$p(x;a,b)=\frac{1}{B(a,b)}\frac{x^{a-1}}{(1+x)^{a+b}},\qquad x>0.$$

Z分布的k阶矩为

$$\begin{aligned}\mathsf{E}X^k =& \frac{1}{B(a,b)}\int_0^\infty \frac{x^{k+a-1}}{(1+x)^{(k+a)+(b-k)}}dx\\ =& \frac{B(a+k,b-k)}{B(a,b)} = \frac{\Gamma(a+k)}{\Gamma(a)}\cdot\frac{\Gamma(b-k)}{\Gamma(b)}\\ =& \frac{(a+k-1)(a+k-2)\cdots a}{(b-1)(b-2)\cdots(b-k)}.\end{aligned}$$

特别地

$$\mathsf{E}X = \frac{a}{b-1},\ \ b>1;\quad \mathsf{E}X^2 = \frac{(a+1)a}{(b-1)(b-2)},\ \ b>2.$$

Z分布与β分布的关系

$$Y \sim \beta(a,b) \implies X = \frac{Y}{1-Y} \sim Z(a,b),$$

$$X \sim Z(a,b) \implies Y = \frac{X}{1+X} \sim \beta(a,b).$$

Z分布与Γ分布的关系

Theorem

定理 设$X_1 \sim \Gamma(\alpha_1, \lambda)$,$X_2 \sim \Gamma(\alpha_2, \lambda)$, 且$X_1$与$X_2$独立, 则

$$\begin{aligned} Y_1 =& X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda), \\ Y_2 =& X_1/X_2 \sim Z(\alpha_1, \alpha_2), \end{aligned}$$

且Y_1与Y_2独立.

Corollary

推论1 设随机变量X_1, X_2相互独立，$X_i \sim \Gamma(\alpha_i, \lambda)$, $i = 1, 2$, 则

$$\frac{X_1}{X_1 + X_2} = \frac{X_1/X_2}{1 + X_1/X_2} \sim \beta(\alpha_1, \alpha_2).$$

Z分布与F分布的关系

Corollary

推论2

$$F \sim F(n,m) \implies \frac{n}{m}F \sim Z(\frac{n}{2},\frac{m}{2}).$$

证: 记$F = \frac{X_1/n}{X_2/m}$, 其中$X_1 \sim \chi^2(n) = \Gamma(n/2,1/2)$, $X_2 \sim \chi^2(m) = \Gamma(m/2,1/2)$, 且$X_1$与$X_2$独立. 从而可得

$$\frac{n}{m}F = \frac{X_1}{X_2} \sim Z(\frac{n}{2},\frac{m}{2}).$$

因此,F分布的密度函数可利用Z分布的密度函数导出, 其数学期望也可导出.

$$\mathsf{E}F(n,m)=\frac{m}{n}\mathsf{E}Z(n/2,m/2)=\frac{m}{n}\frac{n/2}{m/2-1}=\frac{m}{m-2}.$$

此外

Corollary

推论3 设随机变量$F\sim F(n,m)$, 则$(nF/m)/(1+nF/m)\sim\beta(n/2,m/2)$.

§2.5 统计量的极限分布

参考《概率论》第四章

§2.6 指数型分布族(Exponential family)

Definition

定义 设有参数分布族$\mathcal{F}=\{p(x;\theta):\theta\in\Theta\}$, $p(x;\theta)$ 为分布的密度函数(pdf: probability density function) 或分布列(pmf: probability mass function).
若$p(x;\theta)$可表示成如下形式

$$p(x;\theta)=c(\theta)\exp\left\{\sum_{j=1}^{k}Q_j(\theta)T_j(x)\right\}h(x),$$

则称此分布族称为指数型分布族, 或简称为指数族. 其中$c(\theta)>0$, $Q_j(\theta)$ $(j=1,...,k)$为定义在参数空间Θ上的函数,与x无关, $h(x)\geq 0$, $T_j(x)$ $(j=1,...,k)$ 为与θ无关的函数.

特别地,当$k=1$时,此分布族称为单参数指数型分布族.

如果令$\lambda_j = Q_j(\theta)$, 若$c(\theta)$可表示成$\widetilde{\lambda} = (\lambda_1, \ldots, \lambda_k)$的函数$c^*(\widetilde{\lambda})$, 那么$p(x;\theta)$可表示成

$$p(x;\widetilde{\lambda}) = c^*(\widetilde{\lambda}) \exp\left\{\sum_{j=1}^{k} \lambda_j T_j(x)\right\} h(x).$$

这种形式称为指数型分布族的自然形式(natural form). 此时

$$\Lambda = \left\{\widetilde{\lambda} : \int \exp\left\{\sum_{j=1}^{k} \lambda_j T_j(x)\right\} h(x)dx < \infty\right\}$$

称为自然参数空间(natural parametric space).

有很多常用分布族是属于指数型分布族的.

Example

正态分布族$\mathcal{F}=\{N(\mu,\sigma^2),-\infty<\mu<\infty,\sigma>0\}$是指数型分布族.因为它的概率密度可以表示为

$$\begin{aligned} p(x;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\exp\left\{-\frac{x^2}{2\sigma^2}+\frac{\mu x}{\sigma^2}\right\} \end{aligned}$$

若取

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

$$Q_1(\mu,\sigma)=-\frac{1}{2\sigma^2},\quad Q_2(\mu,\sigma)=\frac{\mu}{\sigma^2},\qquad T_1(x)=x^2,\qquad T_2(x)=x.$$

根据定义,即可看出这是一个指数型分布族.

Example

二项分布族$\{B(n,\theta): 0<\theta<1\}$ 是指数型分布族.

解. 二项分布$B(n,\theta)$的概率分布列为

$$\begin{aligned} p(x;\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x}\cdot I\{x=0,1,2,\cdots,n\} \\ &= (1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x\binom{n}{x}\cdot I\{x=0,1,2,\cdots,n\} \\ &= (1-\theta)^n\exp\left\{x\log\frac{\theta}{1-\theta}\right\}\binom{n}{x}\cdot I\{x=0,1,2,\cdots,n\}, \end{aligned}$$

若取

$$c(\theta) = (1-\theta)^n, \quad Q_1(\theta) = \log \frac{\theta}{1-\theta},$$

$$T_1(x) = x, \quad h(x) = \binom{n}{x} \cdot I\{x = 0, 1, 2, \cdots, n\},$$

根据定义,即可看出这是一个指数型分布族.

若令$\log \frac{\theta}{1-\theta} = \lambda$, 那么二项分布的自然指数族形式为

$$p(x;\lambda) = (1+e^{\lambda})^{-n} \exp\{\lambda x\} \binom{n}{x} \cdot I\{x = 0, 1, 2, \cdots, n\}.$$

自然参数空间为$(-\infty, \infty)$.

Example

均匀分布族$\mathcal{F}=\{U(0,\theta),\theta>0\}$不是指数型分布族.因为它的概率密度函数为

$$p(x;\theta)=\frac{1}{\theta},\quad 0<x<\theta,$$

其支撑为$(0,\theta)$依赖于参数θ,所以均匀分布族不是指数型分布族.

Theorem

如果总体分布族是指数型分布族, 那么从中抽取的简单随机样本的分布族也是指数型分布族.

因为如果总体X的pdf或pmf为

$$c(\theta)\exp\left\{\sum_{j=1}^{k}Q_j(\theta)T_j(x)\right\}h(x),$$

那么样本$(X_1,X_2,\ldots,X_n)$的joint pdf或pmf为

$$c^n(\theta)\exp\left\{\sum_{j=1}^{k}Q_j(\theta)\{T_j(x_1)+\ldots+T_j(x_n)\}\right\}h(x_1)\cdots h(x_n).$$

指数族有一些重要的性质. 首先, 我们定义一个随机变量分布的支撑集为集合$S = \{x : p(x) > 0\}$, 其中$p(x)$为pdf或pmf. 分布的支撑集也就是在求概率时实质上起作用的集合. 指数族的第一个重要性质是:

指数族分布的支撑集与参数θ无关.

由定义不难看出,对于指数族, 支撑集$\{x : p(x;\theta) > 0\} = \{x : h(x) > 0\}$, 显然与$\theta$无关.

指数族的第二个重要的性质是: 它有较好的解析性质.

若在指数族的自然形式中, 自然参数空间有内点, 其内点集合记为Λ_0. $g(x)$为任一实函数, 并满足: 积分

$$G(\lambda_1,\cdots,\lambda_k)=\int_{\mathcal{X}} g(x)\exp\left\{\sum_{j=1}^{k}\lambda_j T_j(x)\right\}h(x)dx$$

在集合Λ_0内均有限, 则$G(\lambda_1,\cdots,\lambda_k)$的任意阶偏导在$\Lambda_0$内存在且有

$$\begin{aligned}&\frac{\partial^m G(\lambda_1,\cdots,\lambda_k)}{\partial\lambda_1^{m_1}\cdots\partial\lambda_k^{m_k}}\\ =\quad&\int_{\mathcal{X}}\frac{\partial^m}{\partial\lambda_1^{m_1}\cdots\partial\lambda_k^{m_k}}\left(g(x)\exp\left\{\sum_{j=1}^{k}\lambda_j T_j(x)\right\}h(x)\right)dx,\end{aligned}$$

其中$m=m_1+\cdots+m_k\geq 1$.