Lec2: 抽样分布

张伟平

2011 年 9 月 8 日

1 正态总体下样本均值和样本方差的分布

独立的正态随机变量的线性组合仍然服从正态分布:

定理 1. 设随机变量$X_1, \cdots, X_n$ 相互独立, 且$X_k \sim N(a_k, \sigma_k^2)$, $k = 1, \cdots, n$. $c_1, c_2, \cdots, c_n$ 为常数, 记$T = \sum_{k=1}^n c_k X_k$, 则$T \sim N(\mu, \tau^2)$,其中$\mu = \sum_{k=1}^n c_k a_k$, $\tau^2 = \sum_{k=1}^n c_k^2 \sigma_k^2$.

利用特征函数证明.

利用多元正态的定义,可以有

定理 2. 设随机向量$X \sim N_p(\mu, \Sigma)$, A为$p \times p$可逆矩阵, 令$Y = AX$, 则

$$Y \sim N(A\mu, A\Sigma A')$$

由正态分布随机变量的变换,可以得到新的分布随机变量:

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

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

是自由度为n的χ^2变量, 其分布称为自由度为n的χ^2分布, 记为$\xi \sim \chi_n^2$.

ξ的密度函数$g_n(x)$为

$$g_n(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

求密度方法: 特征函数法,归纳法,n元变换方法.

性质:

(1) 设$r.v.\ \xi \sim \chi_n^2$,则ξ的$c.f.$为$\varphi(t) = (1 - 2it)^{-\frac{n}{2}}$;

(2) $r.v.\ \xi$的数学期望和方差分别为$E(\xi) = n$, $D(\xi) = 2n$.

(3) 设$Z_1 \sim \chi_{n_1}^2, Z_2 \sim \chi_{n_2}^2$,且$Z_1$和$Z_2$独立, 则$Z_1 + Z_2 \sim \chi_{n_1+n_2}^2$.

推广: 非中心χ^2分布

定义 2. 设随机变量$X_1, X_2, \cdots, X_n$相互独立,$X_i \sim N(a_i, 1)$, $a_i\ (i=1,\cdots,n)$不全为0. 记$Y = \sum\limits_{i=1}^{n} X_i^2$, 则称$Y$的分布为自由度$n$ 非中心参数为$\delta = \sqrt{\sum\limits_{i=1}^{n} a_i^2}$的非中心 χ^2分布 ,记为$Y \sim \chi^2_{n,\delta}$. 特别当$\delta = 0$时称为中心的χ^2分布, 即前面所述χ^2_n分布.

若$Y \sim \chi^2_{n,\delta}$,则其密度函数为

$$g(x) = \begin{cases} e^{-\delta^2/2} \sum\limits_{i=0}^{\infty} \frac{1}{i!}(\frac{\delta^2}{2})^i \frac{x^{i+n/2-1}}{2^{i+n/2}\Gamma(n/2+i)} e^{-x/2}, & x > 0, \\ 0, & x \le 0 \end{cases}$$

$$= \begin{cases} e^{-\delta^2/2} \sum\limits_{i=0}^{\infty} \frac{(\delta^2/2)^i}{i!} \chi^2(x, 2i+n), & x > 0, \\ 0, & x \le 0. \end{cases} \tag{1.1}$$

此处$\chi^2(x, 2i+n)$表示自由度为$2i+n$的χ^2密度函数.

非中心χ^2密度的计算方法:

令矩阵A为正交矩阵,其第一行元素为$(a_1/\delta, \cdots, a_n/\delta)$. 从而若$Y = AX$,则$Y_1 = \frac{1}{\delta}\sum_{i=1}^{n} a_i X_i \sim N(\delta, 1)$.

非中心的χ^2变量具有下列性质：

(1) 若$Y \sim \chi^2_{n,\delta}$,则Y的$c.f.$为$\varphi(t) = (1-2it)^{-\frac{n}{2}} e^{\frac{i\delta^2 t}{1-2it}}$,

(2) 若$Y \sim \chi^2_{n,\delta}$,则$E(Y) = n + \delta^2$, $D(Y) = 2n + \delta^2$,

(3) 若$Y_1, \cdots, Y_k$相互独立, $Y_i \sim \chi^2_{n_i,\delta_i}, i = 1, 2, \cdots, k$,则$\sum\limits_{i=1}^{k} Y_i \sim \chi^2_{n,\delta}$,此处$n = \sum\limits_{i=1}^{k} n_i$, $\delta = \sqrt{\delta_1^2 + \delta_2^2 + \cdots + \delta_k^2}$.

由正态分布和χ^2分布可以构造一个t分布随机变量:

定义 3. 设$r.v.\ X \sim N(0,1), Y \sim \chi^2_n$, 且$X$和$Y$独立, 则称

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

为自由为n的t变量, 其分布称为由为n的t分布,记为$T \sim t_n$.

其密度函数为

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < x < \infty. \tag{1.2}$$

t变量具有下列的性质：

(1)若$r.v.\ T \sim t_n$,则$E(T^r)$只有当$r < n\ (n > 1)$时存在, 且

$$E(T^r) = \begin{cases} n^{\frac{r}{2}} \frac{\Gamma(\frac{r+1}{2})\Gamma(\frac{n-r}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})}, & \text{当r为偶数}, \\ 0, & \text{当r为奇数}. \end{cases}$$

特别当$n \ge 2$时, $E(T) = 0$. 当$n \ge 3$时, $D(T) = \frac{n}{n-2}$.

(2)当$n = 1$时t分布就是柯西分布, 此时(1.2)变为

$$t_1(x) = \frac{1}{\pi(1+x^2)}, \ -\infty < x < +\infty.$$

(3)当$n \to \infty$时, t变量的极限分布为$N(0,1)$.

非中心t分布:

定义 4. 设$r.v.\ X \sim N(\delta,1), Y \sim \chi^2_n$, 且$X$和$Y$独立, 则称

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

的分布服从自由度为n,非中心参数为δ的非中心 t分布, 记为$Z \sim t_{n,\delta}$. 特别当$t=0$时的分布称为中心的t分布,即前面所述的t_n分布.

其密度函数为

$$t_{n,\delta}(x) = \frac{n^{n/2}}{\sqrt{\pi}\Gamma(n/2)} \cdot \frac{e^{-\delta^2/2}}{(n+x^2)^{\frac{n+1}{2}}} \sum_{i=0}^{\infty} \Gamma\Big(\frac{n+i+1}{2}\Big)\ \frac{(\delta x)^i}{i!}\Big(\frac{2}{n+x^2}\Big)^{i/2}, \qquad -\infty < x < \infty. \tag{1.3}$$

密度函数(1.3)的推导方法也是利用求$r.v.$商的密度函数公式,经过较复杂的计算可求得.

非中心 t分布的性质如下:

(1) 若$Z_n \sim t_{n,\delta}$, 则$Z_n \xrightarrow{\mathscr{L}} N(\delta,1)$;

(2) 若$Z_n \sim t_{n,\delta}$,则有

$$\begin{aligned} E(Z_n) &= \delta \left(\frac{n}{2}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}, \quad n \ge 2; \\ D(Z_n) &= \frac{n(1+\delta^2)}{n-2} - \frac{\delta^2 n}{2}\Big(\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}\Big)^2, \quad n \ge 3. \end{aligned}$$

χ^2分布随机变量还可以构造F分布随机变量:

定义 5. 设$r.v.\ X \sim \chi^2_m,\ Y \sim \chi^2_n$,且$X$和$Y$独立,则称

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

为自由度分别是m和n的F变量，其分布称为由度分别是m和n的F分布, 记为$F \sim F_{m,n}$.

其密度函数为

$$f_{m,n}(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1}(n+mx)^{-\frac{m+n}{2}}, & x > 0, \\ 0, & \text{其它}. \end{cases} \tag{1.4}$$

F变量具有下列的性质：

(1) 若$Z \sim F_{m,n}$,则$1/Z \sim F_{n,m}$.

(2) 若$Z \sim F_{m,n}$,则对$r>0$有

$$E(X^r) = \Big(\frac{n}{m}\Big)^r \frac{\Gamma(\frac{m}{2}+r)\Gamma(\frac{n}{2}-r)}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})},\ \text{当 } 2r < n.$$

特别

$$E(X) = \frac{n}{n-2},\ n > 2.$$

$$D(X) = \frac{2n^2(m+n-2)}{m(n-2)(n-4)},\ n > 4.$$

(3) 若$T \sim t_n$,则$T^2 \sim F_{1,n}$

(4) $F_{m,n}(1-\alpha) = 1/F_{n,m(\alpha)}$

非中心F分布

定义 6. 设$r.v.\ X \sim \chi^2_{m,\delta}$, $Y \sim \chi^2_n$, 且X和Y独立, 则

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

的分布称为自由度为m, n和非中心参数为δ的非中心 F分布, 记为$Z \sim F_{m,n;\delta}$. 当$\delta = 0$时, 称Z的分布为中心的F分布,即前面定义的$F_{m,n}$.

若$Z \sim F_{m,n;\delta}$, 则Z的密度函数为

$$f_{m,n;\delta}(x) = \begin{cases} \frac{m^{\frac{n}{2}} n^{\frac{m}{2}}}{\Gamma(\frac{n}{2})} e^{-\frac{\delta^2}{2}} x^{\frac{m}{2}-1} \sum\limits_{k=0}^{\infty} \frac{(\frac{\delta^2 mx}{2})^k \Gamma(\frac{m+n}{2}+k)}{k!\ \Gamma(\frac{m}{2}+k)(mx+n)^{\frac{m+n}{2}+k}}, & x > 0, \\ 0, & \text{其它}. \end{cases} \tag{1.5}$$

非中心 F分布具有下列性质:

(1)若$X \sim t_{n,\delta}$, 则$X^2 \sim F_{1,n;\delta}$.

(2)若$Z_n \sim F_{m,n;\delta}$, $n = 1, 2, \cdots$, δ固定, 则当$n \to \infty$时$Z_n \xrightarrow{\mathscr{L}} X^2_{m,\delta}$.

(3)若$Z \sim F_{m,n;\delta}$,则

$E(Z) = \frac{n(m+\delta)}{m(n-2)}$, 对$n > 2$.

$D(Z) = \frac{2n^2}{m^2(n-2)^2(n-4)}[(m+\delta^2)^2 + (n-2)(m+2\delta^2)]$, $n > 4$.

1.1 正态总体下样本均值和样本方差的分布

对正态总体, 我们有

定理 3. 设$X_1, X_2, \cdots, X_n$ $i.i.d. \sim N(a, \sigma^2)$, 则

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

此处我们给出一个利用特征函数的证明方法.

Proof. 做变换: $Y_1 = \bar{X}, Y_2 = X_2, \cdots, Y_n = X_n$, 因此由

$$x_1 = ny_1 - y_2 - \cdots - y_n$$
$$x_2 = y_2$$
$$\vdots$$
$$x_n = y_n$$

以及

$$\sum_{i=1}^{n}(x_i-a)^2=\sum_{1}^{n}(x_i-\bar{x})^2+n(\bar{x}-a)^2$$

可以得到$Y_1,\cdots,Y_n$的联合密度为

$$f(y_1,\cdots,y_n)=n(2\pi\sigma^2)^{-n/2}exp\left\{-\frac{(ny_1-y_2-\cdots-y_n-y_1)^2}{2\sigma^2}-\frac{\sum_2^n(y_i-y_1)^2}{2\sigma^2}-\frac{n(y_1-a)^2}{2\sigma^2}\right\}$$

则由前已知, $Y_1\sim N(a,\sigma^2/n)$, 因此$y_2,\cdots,y_n$在给定y_1的条件密度为

$$\sqrt{n}(2\pi\sigma^2)^{-(n-1)/2}exp(-q/2\sigma^2)$$

其中$q=(ny_1-y_2-\cdots-y_n-y_1)^2+\sum_2^n(y_i-y_1)^2$.

注意到

$$(n-1)S^2=\sum_{1}^{n}(X_i-\bar{X})^2=(nY_1-Y_2-\cdots-Y_n-Y_1)^2+\sum_{2}^{n}(Y_i-Y_1)^2=Q$$

因此$(n-1)S^2$在给定$Y_1=y_1$的条件特征函数为

$$\begin{aligned}E(e^{itQ/\sigma^2}|y_1) &= \int\cdots\int\sqrt{n}(2\pi\sigma^2)^{-(n-1)/2}exp(-(1-2it)q/2\sigma^2)dy_2\cdots dy_n\\ &= (1-2it)^{-(n-1)/2}\int\cdots\int\sqrt{n}\left[\frac{1-2it}{2\pi\sigma^2}\right]^{(n-1)/2}exp\{-(1-2it)q/2\sigma^2\}dy_2\cdots dy_n\\ &= (1-2it)^{-(n-1)/2}\end{aligned}$$

此即$(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$. 而且此条件分布和$Y_1$无关, 因此$S^2$ 和Y_1相互独立. □

定理 4. 设$X_1,X_2,\cdots,X_n$ $i.i.d.\sim N(a,\sigma^2)$, 则

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

定理 5. 设$X_1,X_2,\cdots,X_m$ $i.i.d.\sim N(a_1,\sigma_1^2)$, $Y_1,Y_2,\cdots,Y_n$ $i.i.d.\sim N(a_2,\sigma_2^2)$,且假定$\sigma_1^2=\sigma_2^2=\sigma^2$, 合样本$X_1,X_2,\cdots,X_m$ 与$Y_1,Y_2,\cdots,Y_n$相互独立, 则

$$T=\frac{(\bar{X}-\bar{Y})-(a_1-a_2)}{S_w}\cdot\sqrt{\frac{mn(n+m-2)}{n+m}}\sim t_{n+m-2},$$

此处$(n+m-2)S_w^2=(m-1)S_1^2+(n-1)S_2^2$, 此处

$$S_1^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2,\qquad S_2^2=\frac{1}{n-1}\sum_{j=1}^{n}(y_j-\bar{Y})^2.$$

定理 6. 设$X_1,X_2,\cdots,X_m$ $i.i.d.\sim N(a_1,\sigma_1^2)$, $Y_1,Y_2,\cdots,Y_n$ $i.i.d.\sim N(a_2,\sigma_2^2)$, 且合样本$X_1,X_2,\cdots,X_m$和$Y_1,Y_2,\cdots,Y_n$相互独立, 则

$$F=\frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F_{m-1,n-1},$$

此处S_1^2和S_2^2定义如前所述.

定理 7. 设$X_1,X_2,\cdots,X_n$ $i.i.d.$服从指数分布: $f(x,\lambda)=\lambda e^{-\lambda x}I_{[x>0]}$, 则有

$$2\lambda n\bar{X}=2\lambda\sum_{i=1}^{n}X_i\sim\chi^2_{2n}.$$

2 次序统计量

次序统计量: 即若$X_1, X_2, \cdots, X_n$ i.i.d. $\sim F$, 将其按大小排列为$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, 则称$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$为次序统计量, 它的任一部分, 如$X_{(i)}$,和$(X_{(i)}, X_{(j)})$ $(1 \le i < j \le n)$等也称为次序统计量.

次序统计量的分布

1. $X_{(n)}$ 的分布

$$P(X_{(n)} \le x) = P(X_1 \le x, \cdots, X_n \le x) = F^n(x)$$

2. $X_{(1)}$ 的分布

$$P(X_{(1)} \le x) = 1 - P(X_{(1)} > x) = 1 - P(X_1 > x, \cdots, X_n > x) = 1 - (1 - F(x))^n$$

3. $X_{(m)}$ 的分布$(1 < m < n)$

$$\begin{aligned} f_{X_{(m)}}(x)dx &\approx P(x < X_{(m)} \le x + dx \\ &= \frac{n!}{(m-1)!1!(n-m)!} F^{m-1}(x) f(x) dx [1 - F(x+dx)]^{n-m} \end{aligned}$$

因此两边同时除以dx, 并令$dx \to 0$,得到

$$f_{X_{(m)}}(x) = m \binom{n}{m} F^{m-1}(x) f(x) [1 - F(x)]^{n-m}$$

.

4. $X_{(i)}, X_{(j)}$ 的联合密度

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

5. $(X_{(1)}, \cdots, X_{(n)})$的联合密度为

$$g(x_{(1)}, \cdots, x_{(n)}) = \begin{cases} n! f(x_{(1)}) \cdots f(x_{(n)}), & \text{当} x_{(1)} < x_{(2)} < \cdots < x_{(n)}, \\ 0, & \text{其它}. \end{cases}$$

6. 极差$X_{(n)} - X_{(1)}$ 的分布

作下列变换

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

变换的Jacobian 行列式为:$|J| = \left| \frac{\partial(X_{(i)}, X_{(j)})}{\partial V \partial Z} \right| = 1$, $(X_{(i)}, X_{(j)})$的联合分布密度由前给出,故(V, Z)的联合密度为

$$g_{ij}(v,z) = \begin{cases} \frac{n!}{(i-1)!(j-i-1)!(n-j)!} (F(z))^{i-1} (F(v+z) - F(z))^{j-i-1} & \\ \qquad \times (1 - F(v+z))^{n-j} f(z) f(v+z), & \text{当} v > 0, \\ 0, & \text{其它}. \end{cases} \tag{2.1}$$

从而易知V的密度为

$$f_V(v)=\int_{-\infty}^{\infty}g_{i,j}(v,z)dz.$$

特别, 当取$i=1,\ j=n$得到$(R,X_{(1)})$的联合密度

$$g_{1,n}(v,z)=\begin{cases}\frac{n!}{(n-2)!}(F(v+z)-F(z))^{n-2}f(v+z)f(z), & 当v>0,\\ 0, & 当v\le 0.\end{cases} \tag{2.2}$$

而R的边缘密度为$\int_{-\infty}^{\infty}g_{1,n}(v,z)dz$.

均匀分布情形

设$X_1,X_2,\cdots,X_n$ i.i.d. $\sim$均匀分布$U(0,1)$,其分布函数和密度函数分别为

$$F(x)=\begin{cases}0, & x\le 0,\\ x, & 0<x<1,\\ 1, & x\ge 1.\end{cases}\quad 和 \quad f(x)=\begin{cases}1, & 0<x<1,\\ 0, & 其它\end{cases}$$

设$(X_{(1)},X_{(2)},\cdots,X_{(n)})$为样本$X_1,X_2,\cdots,X_n$的次序统计量, 次序统计量$X_{(m)}$ 的密度函数为

$$f_m(x)=m\binom{n}{m}x^{m-1}(1-x)^{m-n}I_{[0,1]}(x). \tag{2.3}$$

由前可知$(X_{(i)},X_{(j)})$的联合密度为

$$f_{ij}(x)=\begin{cases}\frac{n!}{(i-1)!(j-i-1)!(n-j)!}x^{i-1}(y-x)^{j-i-1}(1-x)^{n-j}, & 0<x<y<1,\\ 0, & 其它.\end{cases}$$

而$(X_{(1)},X_{(2)},\cdots,X_{(n)})$的联合密度为

$$f_{1,2,\cdots,n}(x_{(1)},x_{(2)},\cdots,x_{(n)})=\begin{cases}n!, & 当0<x_{(1)}<\cdots<x_{(n)}<1,\\ 0, & 其它.\end{cases}$$

令$F(z)=z,\ 0<z<1,\ F(v+z)=v+z,\ 0<v+z<1$,得到在均匀分布$U(0,1)$ 场合(V,Z)的联合密度

$$g_{ij}(v,z)=\begin{cases}\frac{n!}{(i-1)!(j-i-1)!(n-j)!}z^{i-1}v^{j-i-1}[1-(v+z)]^{n-j}, & \\ \qquad 当\ 0<z<1,\ 0<v+z<1, & \\ 0, & 其它.\end{cases}$$

此时$V=X_{(j)}-X_{(i)}$的边缘密度, 通过计算积分$\int_0^{1-v}g_{i,j}(v,z)dz$得

$$g_{nij}(v)=\begin{cases}\frac{n!}{(j-i-1)!(n-j+i)!}v^{j-i-1}(1-v)^{n-j+1}, & 当\ 0<v<1,\\ 0, & 其它.\end{cases}$$

特别极差 $R=X_{(n)}-X_{(1)}$的密度函数$g_{1n}(r)$为将前式中的v换成r,将j和i分别用n和1代替得到

$$g_{1n}(r)=\begin{cases}n(n-1)r^{n-2}(1-r), & 当\ 0<r<1,\\ 0, & 其它.\end{cases}$$

3 统计量的极限分布

本章引言中已指出, 在许多情形下统计量的精确分布很难求出, 因此我们要研究统计量的极限分布. 首先给出下列定义.

定义 1. *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.*

当样本大小n 充分大时, 极限分布可作为统计量的近似分布.

研究统计量的极限分布有下列意义: (1) 为了获得统计推断方法的优良性, 常常要知道统计量的分布. 但统计量的精确分布一般很难求得, 建立统计量的极限分布, 提供了一种近似方法, 总比什么方法没有要好. (2) 有时统计量的精确分布虽可求出, 但表达式过于复杂, 使用不方便. 若极限分布较简单, 宁可使用极限分布. (3) 有些统计推断方法的优良性本身就是研究其极限性质, 如相合性, 渐近正态性等.

定义 2. *t当样本大小$n \to \infty$ 时, 一个统计量或统计推断方法的性质称为*大样本性质 *(Large sample properties). 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 *(Small sample properties).*

在此要强调的是, *大样本性质* 和*小样本性质* 的差别不在于样本个数的多少, 而是在于所讨论的问题是在样本大小$n \to \infty$ 时去考虑, 还是在样本大小 n 固定时去研究, 关于*大样本性质* 的研究构成了数理统计的一个很重要的部分, 叫做统计大样本理论. 统计大样本理论, 近几十年来发展很快, 成为二次世界大战后数理统计发展的重要特点之一. 有些统计分支, 如非参数统计, 其中大样本理论占据了主导地位.