统计量及其分布*

X. Ling

2022 年 6 月 4 日

第一部分　知识点

1　需要知道的概念

总体和样本以及它们的相关概念，样本的性质（**独立性**、代表性），**样本联合分布函数**，**经验分布函数**，基本的统计量[1]，**次序统计量及分布**，**三大抽样分布**.

定义 1 (Fisher 因子分解定理) *设 $x_1, x_2, \ldots, x_n$ 是来自某个总体的样本，总体的分布函数为 $F(x;\theta)$，统计量 $T = T(x_1, x_2, \ldots, x_n)$ 称为 θ 的充分统计量，如果在给定 T 的取值之后，$x_1, x_2, \ldots, x_n$ 的条件分布与 θ 无关，即*

$$\frac{\prod_{i=1}^{n} f(x_i;\theta)}{g(y,\theta)} = h(x_1, x_2, \ldots, x_n)$$

其中 $g(y;\theta)$ 是 T 的概率函数.

2　重要的定理

定理 1 *设 $x_1, x_2, \ldots, x_n$ 是来自某个总体的样本，$\bar{x}$ 是样本均值。*

(1) 若总体分布为 $N(\mu, \sigma^2)$，则 $\bar{x}$ 的精确分布为 $N(\mu, \sigma^2/n)$;

(2) 若总体分布未知或者不是正态分布，$E(x) = \mu, Var(x) = \sigma^2$ 存在，则 n 较大时可采用的渐近分布为 $N(\mu, \sigma^2/n)$，常记为 $\bar{x} \overset{A}{\sim} N(\mu, \sigma^2/n)$.

定理 2 *设总体 X 具有二阶矩，即 $E(X) = \mu, Var(X) = \sigma^2 < \infty, x_1, x_2, \ldots, x_n$ 为从该总体中得到的样本，$\bar{x}$ 和 $s*^2$ 分别是样本均值和样本方差，则 $E(\bar{x}) = \mu, Var(\bar{x}) = \sigma^2/n, E(s^2) = \sigma^2$*

*Reference:

《概率论与数理统计教程（第三版）》，茆诗松，程依明，濮晓龙，高等教育出版社;

《概率论与数理统计教程（第三版）习题与解答》，茆诗松，程依明，濮晓龙，高等教育出版社;

《统计推断（翻译版 · 原书第二版）》，（美）George Casella, Roger L. Berger，张忠占、傅莺莺译，机械工业出版社.

[1]下文的所有样本方差（标准差），如无特别指出，均为修正样本方差（标准差）.

定理 3 (单个次序统计量的分布) 设总体 X 的密度函数为 $p(x)$, 分布函数为 $F(x), x_1, x_2, \ldots, x_n$ 为样本, 则第 k 个次序统计量 $x_{(k)}$ 的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k!)} (F(x))^{k-1} (1-F(x))^{n-k} p(x)$$

Rmk: 多个次序统计量的联合分布就是上述定理的直接推论.

定理 4 设 $x_1, x_2, \ldots, x_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 其样本均值和样本方差分别为

$$\bar{x} = \frac{1}{n}\sum_{i-1}^{n} x_i, s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

则有

(1)$\bar{x}$ 和 s^2 相互独立*;*

(2)$\bar{x} \sim N(\mu, \sigma^2/n)$;

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

推论 1 设 $x_1, x_2, \ldots, x_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本, 则有

$$t = \frac{\sqrt{n}(\bar{x}-\mu)}{s} \sim t(n-1);$$

其中 $\bar{x}$ 为样本均值,s 为样本标准差.

推论 2 设 $x_1, x_2, \ldots, x_m$ 是来自 $N(\mu_1, \sigma_1^2)$ 的一个样本, 设 $y_1, y_2, \ldots, y_n$ 是来自 $N(\mu_2, \sigma_2^2)$ 的一个样本, 且这两个样本互相独立, 则有

$$F = \frac{s_x^2/\sigma_1^2}{s_y^2/\sigma_2^2} \sim F(m-1, n-1),$$

其中 s_x^2, s_y^2 分别是两个样本方差.

定理 5 (Neyman 因子分解定理) 设总体的概率函数为 $f(x;\theta), x_1, x_2, \ldots, x_n$ 为样本, 则 $T = T(x_1, x_2, \ldots, x_n)$ 为充分统计量的充要条件是*:* 存在两个函数 $g(t, \theta)$*(*注意这里的 g 的自变量是 t 和 θ*)* 和 $h(x_1, x_2, \ldots, x_n)$ 使得对任意的 θ 和任一组观测值 $x_1, x_2, \ldots, x_n$, 有

$$f(x_1, x_2, \ldots, x_n; \theta) = g(T(x_1, x_2, \ldots, x_n), \theta) h(x_1, x_2, \ldots, x_n)$$

其中,$g(t, \theta)$ 是通过统计量 T 的取值而依赖于样本的.

第二部分　习题

1. 设有 N 个产品, 其中有 M 个不合格品. 进行放回抽样, 定义 x_i 如下:

$$x_i = \begin{cases} 1, & \text{第 } i \text{ 次取得不合格品,} \\ 0, & \text{第 } i \text{ 次取得合格品.} \end{cases}$$

求样本 $x_1, x_2, \ldots, x_n$ 的联合分布.

解 总体的分布列为

$$P(X=1)=\frac{M}{N}, P(X=0)=1-\frac{M}{N}$$

也可以写成

$$P(X=x)=\left(\frac{M}{N}\right)^{x}\left(1-\frac{M}{N}\right)^{(1-x)}, x=0,1$$

因此样本 $x_1, x_2, \ldots, x_n$ 的联合分布列为

$$P(x_1, x_2, \ldots, x_n)=\prod_{i=1}^{n}\left(\frac{M}{N}\right)^{x_i}\left(1-\frac{M}{N}\right)^{(1-x_i)}=\left(\frac{M}{N}\right)^{t}\left(1-\frac{M}{N}\right)^{(n-t)}$$

其中 $t=x_1+x_2+\cdots+x_n$.

*2[2]. 设总体 X 的分布函数为 $F(x)$, 经验分布函数为 $F_n(x)$, 试证

$$E[F_n(x)]=F(x), Var[F_n(x)]=\frac{1}{n}F(x)[1-F(x)]$$

证 设 $x_1, x_2, \ldots, x_n$ 是取自总体分布函数为 $F(x)$ 的样本, 则经验分布函数为

$$F_n(x)=\begin{cases}0, & \text{当 } x<x_{(1)}, \\ \frac{k}{n}, & \text{当 } x_{(k)} \leq x<x_{(k+1)}, k=1,2, \ldots, n-1, \\ 1, & \text{当 } x \geq x_{(n)}.\end{cases}$$

若令 $y_i=I_{\{X_i \leq x\}}, i=1,2, \ldots, n$, 则 $y_1, y_2, \ldots, y_n$ 是独立同分布的随机变量 *(独立性和同分布性分别从哪来？)*, 且

$$E(y_1)=P(X_1 \leq x)=F(x), E(y_1^2)=P(x_1 \leq x)=F(x),$$

于是

$$Var(y_1)=F(x)-[F(x)]^2=F(x)[1-F(x)].$$

又 $F_n(x)$ 可写为 $F_n(x)=\frac{1}{n}\sum_{i=1}^{n} y_i$, 故有

$$E[F_n(x)]=\frac{1}{n}E(\sum_{i=1}^{n} y_i)=E(y_1)=F(x);$$

$$Var[F_n(x)]=\frac{1}{n^2}Var(\sum_{i=1}^{n} y_i)=\frac{1}{n}Var(y_1)=\frac{1}{n}F(x)[1-F(x)].$$

*3. 设 $x_1, x_2, \ldots, x_n$ 是一个样本,s^2 为样本方差, 试证:

$$\frac{1}{n(n-1)}\sum_{i<j}(x_i-x_j)^2=s^2.$$

[2]打 * 的题目会难一点

证 注意到

$$\sum_{i<j}(x_i-x_j)^2=(n-1)\sum_{i=1}^{n}x_i^2-2\sum_{i<j}x_ix_j$$

$$\left(\sum_{i=1}^{n}x_i\right)^2=\sum_{i=1}^{n}x_i^2+2\sum_{i<j}x_ix_j$$

故

$$\sum_{i<j}(x_i-x_j)^2=n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2=n\sum_{i=1}^{n}(x_i-\bar{x})^2$$

得证.

4. 设总体的密度函数为 $p(x)=6x(1-x),0<x<1,x_1,x_2,\dots,x_9$ 是来自该总体的样本, 试求该样本中位数的分布.

Hint 此即求 $x_{(5)}$ 的分布, 代入定理中验证即可.

$$p_{m_{0.5}}(x)=3780x^9(1-x)^9(3-2x)^4(2x+1)^4$$

5. 设 $X_i,i=1,2,3$ 独立且分别服从 $N(i,i^2)$ 分布, 利用 X_i 构造服从下列分布的统计量:

(1) 自由度为 3 的 χ^2 分布;

(2) 自由度为 2 的 t 分布;

(3) 自由度为 1,2 的 F 分布.

Hint 开放题, 自行构造即可. 需要注意的是服从标准正态分布的的统计量为 $\frac{X_i-i}{i}$.

6. 设随机变量 $X\sim F(n,n)$, 证明 $P(X<1)=\frac{1}{2}$.

证 若 $X\sim F(n,n)$, 则 $Y=1/X$ 也服从 $F(n,n)$, 从而

$$P(X<1)=P(Y<1)=P(1/X<1)=P(X>1).$$

又

$$P(X<1)+P(X>1)=1,$$

这就证明了 $P(X<1)=\frac{1}{2}$.

7. 设 x_1,x_2 是来自 $N(0,\sigma^2)$ 的样本, 试求 $Y=\left(\frac{x_1+x_2}{x_1-x_2}\right)^2$ 的分布.

解 由条件,$x_1+x_2\sim N(0,2\sigma^2),x_1-x_2\sim N(0,2\sigma^2)$, 故

$$\left(\frac{x_1+x_2}{\sqrt{2}\sigma}\right)^2\sim\chi^2(1),\left(\frac{x_1-x_2}{\sqrt{2}\sigma}\right)^2\sim\chi^2(1),$$

又

$$Cov(x_1+x_2,x_1-x_2)=Var(x_1)-Var(x_2)=0,$$

且 x_1+x_2 与 x_1-x_2 服从二元正态分布, 因此 x_1+x_2 与 x_1-x_2 独立, 于是

$$Y=\left(\frac{x_1+x_2}{x_1-x_2}\right)^2=\frac{\left(\frac{x_1+x_2}{\sqrt{2}\sigma}\right)^2}{\left(\frac{x_1-x_2}{\sqrt{2}\sigma}\right)^2}\sim F(1,1).$$

8. 设 $x_1,x_2,\ldots,x_n,x_(n+1)$ 是来自 $N(\mu,\sigma^2)$ 的样本,$\bar{x}_n=\frac{1}{n}\sum\limits_{i=1}^{n}x_i,s_n^2=\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x}_n)^2$, 试求常数 c 使得 $t_c=c\frac{x_{(n+1)}-\bar{x}_n}{s_n}$ 服从 t 分布, 并指出分布的自由度.

解 由条件:$x_{n+1}\sim N(\mu,\sigma^2),\bar{x}_n\sim N\left(\mu,\frac{\sigma^2}{n}\right),\frac{(n-1)s_n^2}{\sigma^2}\sim\chi^2(n-1)$, 且 $x_{n+1},\bar{x}_n,s_n^2$ 相互独立, 因而 $x_{n+1}-\bar{x}_n\sim N\left(0,\sigma^2+\frac{\sigma^2}{n}\right)=N\left(0,\frac{n+1}{n}\sigma^2\right)$, 故

$$t=\frac{(x_{n+1}-\bar{x}_n)/\sqrt{\frac{n+1}{n}}}{s_n}=\frac{(x_{n+1}-\bar{x}_n)/\sqrt{\frac{n+1}{n}\sigma^2}}{\sqrt{\frac{(n-1)S_n^2}{\sigma^2}/(n-1)}}\sim t(n-1)$$

这说明当 $c=\sqrt{\frac{n}{n+1}}$ 时,$t_c=c\frac{x_n-\bar{x}_n}{s_n}\sim t(n-1)$, 自由度为 $n-1$.

9. 设 $x_1,x_2,\ldots,x_n$ 是来自泊松分布 $P(\lambda)$ 的样本, 证明:$T=\sum\limits_{i=1}^{n}x_i$ 是充分统计量.

Hint 利用 poisson 分布的可加性和概率分布, 根据 Fisher 因子分解定理除一下就得到了.

10. 设 $x_1,x_2,\ldots,x_n$ 是来自正态分布 $N(\mu,1)$ 的样本，证明:$T=\bar{x}$ 是充分统计量. 如果正态分布方差未知呢?

解 由条件,$T=\bar{x}\sim N(n\mu,n)$, 在给定 $T=t$ 下由条件密度函数

$$\begin{aligned}p_\mu(x_1,x_2,\ldots,x_n|T=t)&=\frac{p_\mu(x_1,x_2,\ldots,x_n)}{p_\mu(t)}\\&=\frac{(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i-\mu)^2\right\}}{(2\pi n)^{-1/2}\exp\left\{-\frac{1}{2n}(t-n\mu)^2\right\}}\\&=\frac{(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum\limits_{i=1}^{n}(x_i)^2-2n\mu t+n\mu^2\right\}}{(2\pi n)^{-1/2}\exp\left\{-\frac{1}{2n}t^2-2n\mu t+n^2\mu^2\right\}}\\&=\sqrt{n}(2\pi)^{-(n-1)/2}\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2-\frac{t^2}{n}\right)\right\}\end{aligned}$$

这个条件概率与 μ 无关.

11. 设 $x_1,x_2,\ldots,x_n$ 是来自韦布尔分布

$$p(x;\theta)=mx^{m-1}\theta^{-m}e^{-(x/\theta)/m},x>0,\theta>0$$

的样本 ($m>0$ 已知), 试给出一个充分统计量.

解 样本的联合密度函数为

$$p(x_1,x_2,\ldots,x_n;\theta)=m^n(x_1x_2\ldots x_n)^{m-1}\theta^{-mn}\exp\left\{\sum_{i=1}^{n}\frac{x_i^m}{\theta^m}\right\}.$$

若令 $T=\sum\limits_{i=1}^{n}x_i^m$, 取 $g(t,\theta)=\theta^{-mn}\exp\{-t/\theta^m\},h(x_1,x_2,\ldots,x_n)=m^n\left(\prod\limits_{i=1}^{n}x_i\right)^{m-1}$, 由因子分解定理,$T=\sum\limits_{i=1}^{n}x_i^m$ 是 θ 的充分统计量.