§5.2 统计量与抽样分布

1. 统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本, $g(X_1, X_2, \cdots, X_n)$是 $X_1, X_2, \cdots, X_n$ 的函数, 若 g 中不含未知参数, 则称 $g(X_1, X_2, \cdots, X_n)$ 是一个统计量. 统计量的分布称为**抽样分布**.

设$x_1, x_2, \cdots, x_n$是相应于样本$X_1, X_2, \cdots, X_n$的观察值, 则称$g(x_1, x_2, \cdots, x_n)$是$g(X_1, X_2, \cdots, X_n)$的观察值.

例1 设X_1, X_2, X_3是来自总体$X \sim N(\mu, \sigma^2)$的一个样本，其中μ为已知，σ^2为未知，判断下列各式些是统计量，哪些不是？

$$T_1 = X_1, \qquad T_2 = X_1 + X_2 e^{X_3},$$

$$T_3 = \frac{1}{3}(X_1 + X_2 + X_3),$$

$$T_4 = \max(X_1, X_2, X_3), \qquad T_5 = X_1 + X_2 - 2\mu,$$

是

$$T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2). \qquad T_7 = X + 2\mu.$$

不是

2. 几个常用统计量的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总

$x_1, x_2, \cdots, x_n$ 是这一样本的观

它反映了总体均值的信息

(1)**样本平均值** $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

其观察值 $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

它反映了总体方差

的信息

(2)**样本方差**

$$S_n^{\ 2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

其观察值

$$s_n^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right).$$

(3)样本标准差

$$S_n=\sqrt{S_n^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2};$$

其观察值 $$s_n=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

(4)修正样本方差

$$S_n^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right).$$

其观察值 $s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right).$

(5) 样本 k 阶(原点)矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1, 2, \cdots;$

其观察值 $a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, k = 1, 2, \cdots.$

(6)样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, k = 2, 3, \cdots;$$

其观察值 $b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k, k = 2, 3, \cdots.$

例2 从某高校一年级男生中任意抽取12名，测得他们的身高如下（单位：cm）：171，165，174，175，168，164，173，178，168，170，172，173

试估计该年级男生的平均身高，并估计其方差和标准差

解：

$$\bar{x} = \frac{1}{12}(171+165+174+\cdots+173) \approx 170.92$$

$$s_n^{*2} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

$$= \frac{1}{11}\left(171^2 + 165^2 + \cdots + 173^2 - 12\times 170.92^2\right) \approx 16.99$$

$$s_n^* = \sqrt{s_n^{*2}} \approx 4.12$$

定理5.1 设总体X（不管服从什么分布，只要均值和方差存在）的均值为 μ，方差为 σ^2，$X_1, X_2, \cdots, X_n$ 是来自总体X的样本，则有

(1) $E(\bar{X}) = \mu$;　　(2) $D(\bar{X}) = \dfrac{\sigma^2}{n}$;

(3) $E(S_n^2) = \dfrac{n-1}{n}\sigma^2$;　　$E(S_n^{*2}) = \sigma^2$;

若总体的四阶矩存在，记 $E(X-\mu)^k = \mu_k, k = 1,2,3,4,$ 则有

(4) $D(S_n^2) = \dfrac{\mu_4 - \mu_2^2}{n} - \dfrac{2(\mu_4 - 2\mu_2^2)}{n^2} + \dfrac{\mu_4 - 3\mu_2^2}{n^3}$;

(5) $\operatorname{cov}(\bar{X}, S_n^2) = \dfrac{n-1}{n^2}\mu_3$.

证明

$$(1)\ E(\overline{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(X_i\right)=\frac{1}{n}\cdot n\mu=\mu$$

$$(2)\ D(\overline{X})=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D\left(X_i\right)=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}$$

(3) $$E(S_n^2)=E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right]$$

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

$$=\frac{1}{n}\left\{\sum_{i=1}^{n}\left[D(X_i)+\left(EX_i\right)^2\right]-n\left[D(\bar{X})+\left(E\bar{X}\right)^2\right]\right\}$$

$$=\frac{1}{n}\left[n(\sigma^2+\mu^2)-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right]=\frac{n-1}{n}\sigma^2$$

$$E(S_n^{*2})=E\left(\frac{n}{n-1}S_n^2\right)=\frac{n}{n-1}E\left(S_n^2\right)=\sigma^2$$

（4） $D\left(S_n^2\right)=E\left(S_n^2\right)^2-\left[E\left(S_n^2\right)\right]^2$

$$(S_n^2)^2=\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]^2=\frac{1}{n^2}\left[\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2\right]^2$$

记$Y_i=X_i-\mu$，则$EY_i=0, D(Y_i)=\sigma^2, EY_i^4=\mu_4$，则有

$$\left(nS_n^2\right)^2=\left(\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2\right)^2=(\sum_{i=1}^{n}Y_i^2)^2-\frac{2}{n}(\sum_{i=1}^{n}Y_i^2)(\sum_{j=1}^{n}Y_j)^2+\frac{1}{n^2}(\sum_{j=1}^{n}Y_j)^4$$

$$=(\sum_{i=1}^{n}Y_i^4+\sum\sum_{i\neq j}Y_i^2Y_j^2)-\frac{2}{n}(\sum\sum_{i\neq j}Y_i^2Y_j^2+\sum_{i=1}^{n}Y_i^4+\sum_{k}\sum\sum_{i\neq j}Y_k^2Y_i\,Y_j\,)$$

$$+\frac{1}{n^2}(\sum_{j=1}^{n}Y_j^4+3\sum\sum_{i\neq j}Y_i^2Y_j^2+4\sum\sum_{i\neq j}Y_i^3Y_j$$

$$+6\sum_{k}\sum\sum_{i\neq j\neq k}Y_k^2Y_i\,Y_j\,+\sum\sum\sum\sum_{i\neq j\neq k\neq l}Y_k\,Y_i\,Y_j\,Y_l)$$

对上式两边取期望得：

$$n^2 E(S_n^2) = n\mu_4 + n(n-1)\mu_2^2 - \frac{2}{n}[n(n-1)\mu_2^2$$

$$+ n\mu_4] + \frac{1}{n^2}[n\mu_4 + 3n(n-1)\mu_2^2]$$

$$= (n-2+\frac{1}{n})\mu_4 + (n-2+\frac{3}{n})(n-1)\mu_2^2$$

由于

$$\boldsymbol{D(S_n^2) = E(S_n^2)^2 - (ES_n^2)^2}$$

得证。

(5) $\operatorname{cov}\left(\bar{X}, S_n^2\right)$

$$= E\left[\left(\bar{X}-\mu\right)\left(S_n^2-\frac{n-1}{n}\mu_2\right)\right]$$

$$= E\left[\bar{Y}\left(S_n^2-\frac{n-1}{n}\mu_2\right)\right] = E\left[\bar{Y}S_n^2\right]$$

$$= E\left[\bar{Y}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\bar{Y}^2\right)\right]$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}Y_i^2\right)\right]-E\left(\bar{Y}^3\right)$$

$$= \frac{\mu_3}{n}-\frac{\mu_3}{n^2} = \frac{n-1}{n^2}\mu_3$$

定理5.2

设$\xi^T=(X_1,X_2,\cdots,X_n)$，$\eta^T=(Y_1,Y_2,\cdots,Y_n)$是两个随机向量，

且$\eta=A\xi$，其中$A=(a_{ij})_{n\times n}$为常数方阵，则有

$$E\eta=A(E\xi),\quad D\eta=A(D\xi)A^T.$$

证明：见书P242

定理5.3（1）设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}$ 是样本均值，则有 $\bar{X} \sim N(\mu, \sigma^2/n)$.

（2）若总体分布未知或不是正态分布，根据中心极限定理，当n较大时，$\bar{X}$的渐近分布为$N\left(\mu, \frac{\sigma^2}{n}\right)$，即$\bar{X} \sim AN\left(\mu, \frac{\sigma^2}{n}\right)$。

证明: (1) $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$

根据性质：独立正态分布r.v.的线性组合

仍然服从正态分布；

又$E\left(\bar{X}\right) = \mu, \quad D\left(\bar{X}\right) = \dfrac{\sigma^2}{n},$

故$\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right).$

标准化：$U = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N\left(0, 1\right).$

定理4.3.2 (林德贝格-勒维中心极限定理)

设$\xi_1, \xi_2, \cdots, \xi_n, \cdots$是独立同分布的随机变量列，具有有限数学期望和方差：

$$E\xi_k = \mu,\ D\xi_k = \sigma^2,\ k = 1, 2, \cdots.$$

则对任意实数x，有

$$\lim_{n\to\infty} P\left(\frac{\sum\limits_{k=1}^{n} \xi_k - n\mu}{\sqrt{n}\sigma} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy = \Phi(x).$$

（2）由中心极限定理，$\sqrt{n}(\bar{x}-\mu)/\sigma \xrightarrow{L} N(0,1)$，这表明 n 较大时 $\bar{x}$ 的渐近分布为 $N(\mu,\sigma^2/n)$，证明完成.

定理5.4（Fisher引理）

设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}, S_n^2$ 分别是样本均值和样本方差, 则有

(1) $\dfrac{nS_n^2}{\sigma^2} = \dfrac{(n-1)S_n^{*2}}{\sigma^2} = \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1)$;

(2) $\bar{X}$ 与 S_n^2, $\bar{X}$ 与 S_n^{*2} 分别独立.

推论1 设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本, $\overline{X}, S_n^{*2}$ 分别是样本均值和修正样本方差, 则有

$$\frac{\overline{X}-\mu}{S_n/\sqrt{n-1}}=\frac{\overline{X}-\mu}{S_n^*/\sqrt{n}}\sim t(n-1).$$

证明 $\because \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1),\qquad \dfrac{(n-1)S_n^{*2}}{\sigma^2}\sim\chi^2(n-1),$

且两者独立, 由 t 分布的定义知

$$\frac{\overline{X}-\mu}{S_n^*/\sqrt{n}}=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\bigg/\sqrt{\frac{(n-1)S_n^{*2}}{\sigma^2(n-1)}}\sim t(n-1).$$

推论2

设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是来自两个正态总体 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$的样本, 且这两个样本互相独立, 设 $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$ 分别是这两个样本的均值,

$$S_1^{*2} = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2, \quad S_2^{*2} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$$

分别是这两个样本的修正样本方差, 则有

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

(2) 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时,

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t(n_1+n_2-2),$$

其中 $S_w^2=\dfrac{(n_1-1)S_1^{*2}+(n_2-1)S_2^{*2}}{n_1+n_2-2}$, $\quad S_w=\sqrt{S_w^2}$.

证明 (1) 由定理5.4

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

由假设 S_1^{*2}, S_2^{*2} 独立, 则由 F 分布的定义知

$$\frac{(n_1-1)S_1^{*2}}{(n_1-1)\sigma_1^2}\Bigg/\frac{(n_2-1)S^{*2}_2}{(n_2-1)\sigma_2^2}\sim F(n_1-1,n_2-1),$$

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

(2) 因为 $\bar{X}-\bar{Y} \sim N\left(\mu_1-\mu_2, \dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2}\right)$

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

由 $\dfrac{(n_1-1)S_1^{*2}}{\sigma^2} \sim \chi^2(n_1-1), \dfrac{(n_2-1)S_2^{*2}}{\sigma^2} \sim \chi^2(n_2-1),$

且它们相互独立, 故由 χ^2 分布的可加性知

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

由于 U 与 V 相互独立,按 t 分布的定义

$$\frac{U}{\sqrt{V/(n_1 + n_2 - 2)}}$$

$$= \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

作业：5.1，5.3，5.14（2），

5.22，5.6，5.9，5.10