- 样本是进行统计推断的依据
- 但是，实际应用中，往往不是直接使用样本本身，需要针对不同的问题对样本“整理”即构造样本的函数，利用这些样本构造而成的函数来进行统计推断。

——统计量

第三节　样本函数及其概率分布

一、　统计量

二、　常用统计量

一、统计量

设 $X_1, X_2, \cdots, X_n$是取自总体X的一个样本, $g(t_1, t_2, \cdots, t_n)$为一 n元实值函数, 则称随机变量$g(X_1, X_2, \cdots, X_n)$为样本函数.

若$x_1, x_2, \cdots, x_n$是样本观测值, 称$g(x_1, x_2, \cdots, x_n)$为样本函数的观测值.

若$g(X_1, X_2, \cdots, X_n)$中不含未知参数, 则称这种样本函数为统计量.

例 $X \sim N(\mu, \sigma^2), \mu, \sigma^2$是未知参数, 设$X_1, X_2, \cdots, X_n$是从总体$X$中抽取的样本, 则

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

是统计量.

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2, X_1 - \mu$$ 不是统计量, 只是样本函数.

实例2 设 X_1, X_2, X_3 是来自总体 $N(\mu, \sigma^2)$ 的一个样本，其中 μ 为已知，σ^2 为未知，判断下列各式哪些是统计量，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

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

不是

二、常用统计量

设$X_1, X_2, \cdots, X_n$是来自总体的一个样本，$x_1, x_2, \cdots, x_n$是这一样本的观察值.

1、样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$

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

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$ 是$X_1,X_2,\ldots,X_n$的算术平均值

设$E(X)=\mu, D(X)=\sigma^2, X_1,X_2,\ldots,X_n$与总体$X$同分布，所以

$E(X_i)=\mu, D(X_i)=\sigma^2$,则

$$E(\bar{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\times(n\mu)=\mu,$$

$$D(\bar{X})=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)\text{(因为}X_i\text{相互独立)}=\frac{1}{n^2}\times(n\sigma^2)=\frac{\sigma^2}{n}.$$

样本均值的期望=总体的期望；$E(\bar{X})=\mu, D(\bar{X})=\dfrac{\sigma^2}{n}$.

2. 样本方差

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

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

证明：$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2-2X_i\overline{X}+\overline{X}^2)$$

$$=\frac{1}{n-1}(\sum_{i=1}^{n}X_i^2-2\overline{X}\sum_{i=1}^{n}X_i+n\overline{X}^2)$$

$$=\frac{1}{n-1}(\sum_{i=1}^{n}X_i^2-2\overline{X}n\overline{X}+n\overline{X}^2)$$

$$=\frac{1}{n-1}[\sum_{i=1}^{n}X_i^2-n\overline{X}^2]$$

其观察值

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

3、样本标准差

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2};$$

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

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

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

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

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

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

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

样本方差的期望=总体的方差

$E(S^2)=\sigma^2=D(X)$

4、 **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 $a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$, $k = 1, 2, \cdots$

注： 样本一阶原点矩就是样本均值.

5、**样本 k 阶中心矩**

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

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

$$B_2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad VS \quad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\frac{B_2}{S^2}=\frac{n-1}{n},\ B_2=\frac{n-1}{n}S^2,$$

用S^2的期望来估计总体方差比B_2好！

$$E(B_2)=E\left(\frac{n-1}{n}S^2\right)=\frac{n-1}{n}E(S^2)=\frac{n-1}{n}\sigma^2\neq\sigma^2.$$

结论:

若总体 X 的 k 阶矩 $E(X^k) \underline{\underline{\text{记成}}} \mu_k$ 存在,

则当 $n \to \infty$ 时, $A_k \xrightarrow{P} \mu_k, k = 1, 2, \cdots.$

由依概率收敛的序列的性质知

$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k)$, 其中 g 是连续函数。

------矩估计法的理论根据.

例6.3.1 已知总体X的样本值如下表：

x_i	102	104	106
n_i	2	3	5

求样本均值，样本方差和样本标准差的观测值。

解 样本均值，样本方差和样本标准差的观测值分别为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{3}n_i x_i=\frac{1}{10}(2\times 102+3\times 104+5\times 106)$$

$$=104.6,$$

$$s^2=\frac{1}{n-1}\left(\sum_{i=1}^{3}n_i x_i^2-n\bar{x}^2\right)=\frac{1}{10-1}\left(2\times 102^2+3\times 104^2+5\times 106^2-10\times 104.6^2\right)$$

$$=2.71,$$

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\sqrt{s^2}=\sqrt{2.71}=1.646.$$

6、 样本最大值和样本最小值

$$X_{(n)} = \max(X_1, X_2, \cdots, X_n)$$

$$X_{(1)} = \min(X_1, X_2, \cdots, X_n)$$

其观察值

$$x_{(n)} = \max(x_1, x_2, \cdots, x_n)$$

$$x_{(1)} = \min(x_1, x_2, \cdots, x_n)$$

设总体X的分布函数为$F(x)$, 记$X_{(n)}$和$X_{(1)}$的分布函数分别为$F_{\max}(x)$和$F_{\min}(x)$,

$$\begin{aligned}F_{\max}(x) &= P\{\max(X_1,X_2,\cdots,X_n)\le x\}\\ &= P\{X_1\le x, X_2\le x,\cdots,X_n\le x\}\\ &= P\{X_1\le x\}P\{X_2\le x\}\cdots P\{X_n\le x\}\\ &= [F(x)]^n\end{aligned}$$

$$\begin{aligned}F_{\min}(x) &= P\{\min(X_1,X_2,\cdots,X_n)\le x\}\\ &= 1-P\{\min(X_1,X_2,\cdots,X_n)> x\}\\ &= 1-P\{X_1> x\}P\{X_2> x\}\cdots P\{X_n> x\}\\ &= 1-[1-F(x)]^n\end{aligned}$$

例 6.3.2

设总体 X的概率密度为

$$f(x)=\begin{cases}2\cos 2x, & 0\le x\le \frac{\pi}{4}\\ 0, & 其它\end{cases}$$，从总体X中抽取样本$X_1,X_2,\cdots,X_n$，

求样本容量 n 使得$P\{\max(X_1,X_2,\cdots,X_n)>\frac{\pi}{12}\}\ge\frac{15}{16}$.

解 X的分布函数为 $F(x)=\begin{cases}0, & x<0,\\ \sin 2x, & 0\le x\le\frac{\pi}{4},\\ 1, & x>\frac{\pi}{4}.\end{cases}$

由于

$$\frac{15}{16}\le P\{\max(X_1,X_2,\cdots,X_n)>\frac{\pi}{12}\}=1-P\{\max(X_1,X_2,\cdots,X_n)\le\frac{\pi}{12}\}$$

$$=1-[F(\tfrac{\pi}{12})]^n=1-\left(\sin\tfrac{\pi}{6}\right)^n=1-\left(\tfrac{1}{2}\right)^n,$$

于是有 $\left(\frac{1}{2}\right)^n\le\frac{1}{16}$，所以有$n\ge 4$.

正态总体的样本均值与样本方差的分布

定理6.3.1

设$X_1, X_2, \ldots, X_n$是取自正态总体$X \sim N(\mu, \sigma^2)$的一个样本，

则$E(\overline{X}) = \mu, D(\overline{X}) = \dfrac{\sigma^2}{n}$，且

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}).$$

进一步地，

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

证明:

由于$X_1, X_2, \ldots, X_n$相互独立，$X_i \sim N(\mu, \sigma^2)\ (i = 1, 2, \cdots, n)$,

因此$X_1, X_2, \ldots, X_n$的线性函数$\overline{X}$服从正态分布。

$$\text{而}\ E(\overline{X}) = \mu,\ D(\overline{X}) = \frac{\sigma^2}{n},\quad \text{故}\ \overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

又因$\overline{X}$的标准化随机变量$u = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$是$\overline{X}$的线性函数,

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

例 6.3.3

设总体 $X \sim N(30,\ 16)$, 从总体X中抽取容量为n的样本,

(1) 当n=25时，求$P\{|\bar{X}-30|<1\}$.

(2) 要使$P\{|\bar{X}-30|<1\}\geq 0.95$, 问样本容量$n$至少应取多大?

解 根据定理6.3.1, 样本均值$\bar{X} \sim N(30,\dfrac{16}{n})$, 则

$$u=\frac{\bar{X}-30}{4/\sqrt{n}} \sim N(0,1).$$

$$P\left\{\left|\bar{X}-30\right|<1\right\}=P\left\{\frac{\left|\bar{X}-30\right|}{4/\sqrt{n}}<0.25\sqrt{n}\right\}=P\left\{\left|u\right|<0.25\sqrt{n}\right\}=2\Phi(0.25\sqrt{n})-1,$$

当$n=25$时，$P\{|\bar{X}-30|<1\}=2\Phi(0.25\sqrt{25})-1=2\Phi(1.25)-1$

$=2\times 0.8944-1=0.7888.$

例 6.3.3

设总体 $X \sim N(30,\ 16)$, 从总体X中抽取容量为n的样本,

(1) 当n=25时，求$P\{|\bar{X}-30|<1\}$.

(2) 要使$P\{|\bar{X}-30|<1\} \geq 0.95$, 问样本容量$n$至少应取多大?

(2)

要使$P\{|\bar{X}-30|<1\} \geq 0.95$, 则有

$$2\Phi(0.25\sqrt{n})-1 \geq 0.95,\ \ \Phi(0.25\sqrt{n}) \geq 0.975.$$

查表得, $\Phi(1.96)=0.975$, 由于$\Phi(x)$单调增加, 故应有

$0.25\sqrt{n} \geq 1.96$, 即 $n \geq 61.4656$, 故样本容量至少应取为62.

定理 6.3.2

设 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$. 分别独立地从总体$X$和总体$Y$中抽取样本$X_1, X_2, \cdots, X_{n_1}$及$Y_1, Y_2, \cdots, Y_{n_2}$, 样本均值分别为$\overline{X}$ 和$\overline{Y}$,

则 $\overline{X}$ 和 $\overline{Y}$ 相互独立且 $\overline{X} \sim \mathrm{N}(\mu_1, \frac{\sigma_1^2}{n_1}), \overline{Y} \sim \mathrm{N}(\mu_2, \frac{\sigma_2^2}{n_2})$,

从而有 $\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$,

所以 $\dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$.

例 6.3.4

设总体 $X\sim N(20, 25)$, 总体 $Y\sim N(10, 4)$, 从总体X中抽取容量为 $n_1=10$ 的样本,样本均值为 $\overline{X}$; 从总体Y中抽取容量为 $n_2=8$ 的样本,样本均值为 $\overline{Y}$.假设这两个样本是各自独立抽取的,求$P\{\overline{X}-\overline{Y}>6\}$.

解 根据定理6.3.2可得

$$u=\frac{\overline{X}-\overline{Y}-(20-10)}{\sqrt{\frac{5^2}{10}+\frac{2^2}{8}}}=\frac{\overline{X}-\overline{Y}-10}{\sqrt{3}}\sim N(0,1).$$

$$P\left\{\overline{X}-\overline{Y}>6\right\}=P\left\{\frac{\overline{X}-\overline{Y}-10}{\sqrt{3}}>\frac{6-10}{\sqrt{3}}\right\}=P\{u>-2.31\}$$

$$=1-P\{u\le -2.31\}=1-\Phi(-2.31)=\Phi(2.31)=0.9896.$$