MA212 Probability and Statistics

Chapter Six: Statistics Principles and Sampling Distributions

Wenwu GONG & Qianqian WANG
Department of Statistics and Data Science
Email: 12031299@mail.sustech.edu.cn
12032005@mail.sustech.edu.cn

December 6, 2021

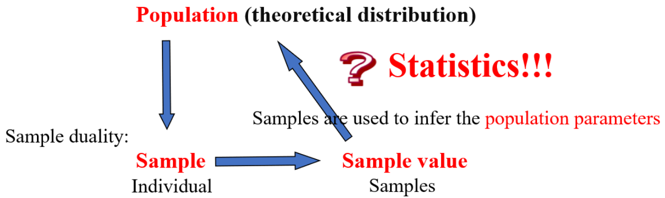
The principles of mathematical statistics

Population（总体）: The quantitative index $X \sim F(x)$ of research object.

Individual（个体）: The realization of quantitative index X, a value of R.V. X.

Sampling（抽样）: To observe the population X repeatedly and independently n times under the same condition.

Summary



Statistics is to use the obtained sample values to find out the features of the population distribution $F(x)$.

Statistic

If $X_1, X_2, \ldots, X_n$ are the samples from population $X \sim F(x)$, then $g(x_1, x_2, \ldots, x_n)$ is n-variate function. If R. V. $g(X_1, X_2, \ldots, X_n)$ does not contain any unknown parameter, then $g(X_1, X_2, \ldots, X_n)$ is statistic.

"Good" statistic can dig out the useful information from data, for examples:
If $X_1, X_2, \ldots, X_n$ are the samples from population $X \sim N(\mu, \sigma^2)$, and μ *and* σ^2 are unknown.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$$

Results!!!

Assuming the mean and variance of population X do exist.

$$E(X) \triangleq \mu, \; D(X) \triangleq \sigma^2$$

$X_1, X_2, \ldots, X_n$ are the samples of population X, define

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

then

$$\boxed{E(\overline{X}) = \mu, \;\; D(\overline{X}) = \frac{\sigma^2}{n}, \;\; E(S^2) = \sigma^2}$$

Proof!!!

Proof: $E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$

$$D(\bar{X}) = D(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}$$

$$\because \quad (n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}[(X_i - \mu) + (\mu - \bar{X})]^2$$

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

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

$$= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

$$\therefore \quad (n-1)E(S^2) = \sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2$$

$$= \sum_{i=1}^{n} E(X \ - \mu)^2 - nE(\bar{X} - E(\bar{X}))^2$$

$$= \sum_{i=1}^{n} \sigma^2 - n\frac{\sigma^2}{n} = (n-1)\sigma^2$$

Other proof? $\quad \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} {X_i}^2 - n\bar{X}.$

χ^2 distribution

If $X_1, X_2, \ldots, X_n$ are the samples of the population $X \sim N(0,1)$, let

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

χ^2 follows x^2-distribution with degree of freedom of n, noted as $x^2 \sim x^2(n)$

- **The additivity of x^2-distribution**

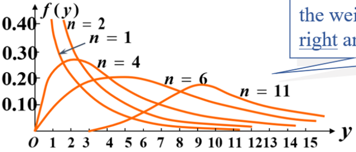
  If $x_1^2 \sim x^2(n_1)$, $x_2^2 \sim x^2(n_2)$, and x_1^2 *and* x_2^2 are independent, then

$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2)$$

- **Numerical characteristics of x^2-distribution**

$$\text{If } \chi^2 \sim \chi^2(n), \text{ then } E(\chi^2) = n,\ D(\chi^2) = 2n$$

- **The figure of x^2-distribution**



With the increase of the degree of freedom the weight point of the curve line moves to right and the curve tends to be gentle.

1.3 Sample distributions (Cont'd)

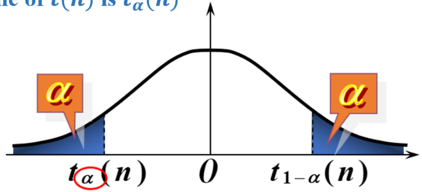
t-student distribution

If $X \sim N(0,1)$, $Y \sim x^2(n)$, X and Y are independent, let

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

t follows t-distribution with degree of freedom of n, noted as $t \sim t(n)$

The α quantile of $t(n)$ is $t_\alpha(n)$



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

F distribution

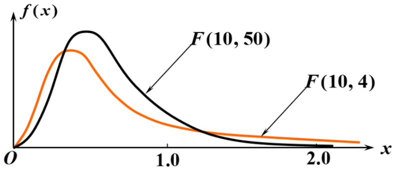
If $U \sim x^2(n_1), V \sim x^2(n_2)$, U and V are independent, let

$$F = \frac{U/n_1}{V/n_2}$$

F follows F-distribution with degree of freedom of (n_1, n_2), noted as $F \sim F(n_1, n_2)$.

The property of F-distribution

If $F \sim F(n_1, n_2)$, then $\frac{1}{F} \sim F(n_2, n_1)$



Based on Normal distribution

Theorem 1: If $X_1, X_2, \dots, X_n$ are the samples of population $X \sim N(\mu, \sigma^2)$, $\bar{X}$ *and* S^2 are sample mean and sample variance, then

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\bar{X}, S^2 \quad \text{are independent} \qquad \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

Based on Normal distribution

Theorem 2: Assuming $X_1, X_2, \ldots, X_{n_1}$ is the samples of population $X \sim N(\mu_1, \sigma_1^2)$. $Y_1, Y_2, \ldots, Y_{n_2}$ are the samples of population $Y \sim N(\mu_2, \sigma_2^2)$. They are independent. Their sample mean and sample variance are $\bar{X}, \bar{Y}, S_1^2, S_2^2$. Then

- **1)**

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

$$S_\omega^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}, \quad S_\omega = \sqrt{S_\omega^2}.$$

- **2)**

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

1.4 Basic concepts (Cont'd)

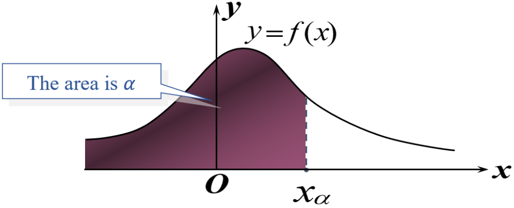
Quantile

α − *quantile* （分位点）

Assuming $X \sim f(x)$, if $\forall 0 < \alpha < 1$, constant x_α has

$$P\{X \le x_\alpha\} = \int_{-\infty}^{x_\alpha} f(x)\mathrm{d}x = \alpha$$

Then x_α is the α quantile of distribution density $f(x)$.



Sample distribution

- Suppose that $\{X_n\}$ are sampled from $N(7.6, 4)$, try to find out n such that $P(5.6 < \bar{X} < 9.6) \geq 0.95$ (Sample duality).
- Suppose that $\{X_n\}$ are sampled from a continuous and increasing population $F(x)$. Try to prove that statistic $T = -2\sum_{i=1}^{n} \ln F(x_i) \sim \chi^2(2n)$ ($\chi^2(n) = \Gamma(n/2, 1/2) = \sum_{i=1}^{n} \mathsf{exp}_i(1/2)$).

2. There are 10000 elder people participating in a type of insurance. The premium is 200 yuan per year. If the insured person passes away within the insured year, the beneficiary will receive 10000 yuan. Assume that the probability of death is 0.017, what is the probability that the insurance company will suffer a deficit in a year?

3. Let $\overline{X}$ be the average of a sample of 16 independent normal random variables with mean 0 and variance 1. Determine c such that

$$P(|\overline{X}| < c) = .5$$

6. Show that if $T \sim t_n$, then $T^2 \sim F_{1,n}$.

8. Show that if X and Y are independent exponential random variables with $\lambda = 1$, then X/Y follows an F distribution. Also, identify the degrees of freedom.

Supplementary questions:

1. Suppose that the population distribution is $N(240, 20^2)$. Draw a sample of size 36 and another of size 49 independently from the population. Compute the probability that the absolute value of the difference between the two sample means does not exceed 10.

2. Suppose that $X_1, X_2, \dots, X_{10}$ are a sample from population $X \sim N(0, 0.3^2)$. Compute the constant C such that $P\left(\sum_{i=1}^{10} X_i^2 \leq C\right) = 0.95$.

3. Suppose that $X_1, X_2, \dots, X_n$ are an independent sample from population $X \sim N(0, \sigma^2)$. Find the constant n such that $P\left(\left|\overline{X} - \mu\right| < 1\right) \geq 0.95$.

4. Suppose that X_1, X_2 are a sample from population $X \sim N(0, \sigma^2)$.
   a) Find the distribution of $\frac{(X_1-X_2)^2}{(X_1+X_2)^2}$
   b) Find the constant k such that $P\left\{\frac{(X_1+X_2)^2}{(X_1+X_2)^2+(X_1-X_2)^2} > k\right\} = 0.1$

5. Suppose that $X_1, X_2, \dots, X_n, X_{n+1}$ are a sample from population $X \sim N(\mu, \sigma^2))$, $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$. Compute the constant c such that $t_c = c\frac{X_{n+1} - \bar{X}_n}{S_n}$ follows the t-distribution and find the degree of freedom.