Chap6.样本以及抽样分布

46.样本及抽样分布

1. 总体：E 的全部可能的观察值。研究对象的总体。

个体：每个可能的观察值。

容量：总体中包含的个体的数目称为总体的容量。

有限总体：容量为有限的总体，称为有限总体。
例如：某大学某年级2000名学生的身高。

无限总体：容量为无限的总体，称为无限总体。
例如：某地点每日的最高气温。

2. 总体与随机变量的关系
   个体 $\to E$ 个观察值 $\to X$ 的值
   总体 $\leftrightarrow X$

3. 样本、样本值
   总体 $\to$ 个体 $\to$ 总体
   样本：被抽出的部分个体

在相同条件下，对总体 X 进行 n 次重复的、独立的观察。将 n 次观察结果按试验的次序记为 $X_1, X_2, ..., X_n$

由于 $X_1, X_2, ..., X_n$ 是对随机变量 X 观察的结果，且各次观察是在相同条件下独立进行的，因此可以认为 $X_1, X_2, ..., X_n$ 相互独立，并与总体 X 同分布，此时称 $X_1, X_2, ..., X_n$ 是来自总体的一个简单随机样本，也简称为样本，n 称为样本的容量。

当 n 次观察一经完成，就得到一组实数 $x_1, x_2, ..., x_n$，它们依次是随机变量 $X_1, X_2, ..., X_n$ 的观察值，称为样本值。

有限总体：取样：不放回抽样
无限总体：取样：不放回抽样

统计学的核心问题是由样本（由试验获得，已知）推断总体。

简单随机样本

定义：

设随机变量 X 的分布函数为 F，若 $X_1, X_2, ..., X_n$ 是具有同一分布函数 F 的，相互独立的随机变量，则称 $X_1, X_2, ..., X_n$ 为取自总体 X，容量为 n 的简单样本，简称为样本，它们的观察值 $x_1, x_2, ..., x_n$ 称为样本值。

注意：

- $X_1, X_2, ..., X_n$ 作为来自 X 的样本 $\Leftrightarrow X_1, X_2, ..., X_n$ 相互独立；

- $X_1, X_2, \ldots, X_n$ 与 X 相互独立。

样本的分布

定理：假设总体 X 的分布函数为 $F(x)$（概率密度为 $f(x)$）或分布律为 $P\{X = a_i\} = p_i\}$，$X_1, X_2, \ldots, X_n$ 是取自总体 X，容量为 n 的样本，则 $(X_1, X_2, \ldots, X_n)$ 的联合分布函数为

$$F(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot \cdots \cdot F_{X_n}(x_n) = \prod_{i=1}^{n} F(x_i) \tag{1}$$

相应地，对于离散型随机变量的样本 $X_1, X_2, \ldots, X_n$，联合分布为

$$P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdot \cdots \cdot P(X_n = x_n) = \prod_{i=1}^{n} P(X = x_i) \tag{2}$$

对于连续型随机变量的样本 $X_1, X_2, \ldots, X_n$，联合概率密度为

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \cdots \cdot f_{X_n}(x_n) = \prod_{i=1}^{n} f(x_i) \tag{3}$$

47.直方图

48.常用的统计量

样本 k 阶(原点)矩

随机变量 X 的 k 阶矩：$E(X^k)$

样本的 k 阶矩：

$$A_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k \tag{4}$$

定理：矩估计法理论依据

若总体的 k 阶矩 $E(X^k) = \mu^k$ 存在，则当 $n \to \infty$ 时，样本的 k 阶矩 $A_k \to \mu_k$

总体的 k 阶矩和样本的 k 阶矩的关系。

上面这个定理是依概率收敛。

即样本的 k 阶矩依概率收敛与总体的 k 阶矩

样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k, k = 2, 3.\dots \tag{5}$$

49.经验分布函数

设 $X_1, X_2, \dots, X_n$ 是总体下的一个样本，$x_1, x_2.\dots, x_n$ 是样本值，用 $S(x)$ 表示 $x_1, x_2 \dots$ 中不大于 x 的个数，可定义经验分布函数：

$$F_n(x) = \frac{1}{n}S(x) = \frac{x_1 + x_2.\dots}{n} \tag{6}$$

类似于离散型随机变量的分布函数，得到一条阶梯形曲线。

> 总体 X 具有其分布函数，是否可以提出一种和总体 X 有关的统计量，这就是经验分布函数的由来。
>
> 经验分布函数实质上是一个由样本值得到的统计量。

经验分布函数能否用来估计总计的分布呢?

格里汶科提出了：对于任意的 x，当 $n \to \infty$，$F_n(x) \to F(x)$

实际上，可以用经验分布函数当做总体的 $F(x)$ 来使用。

50.三大抽样分布

从总体 X 抽样 $\to$ 样本 $x_1, x_2, \dots x_n \to g(x_1, x_2, \dots x_n)$ 统计量，不能含有任何的未知参数。

意义在于，用样本对总体 X 进行统计推断。

考虑标准正态 $N(0,1) \to x_1, x_2, \dots x_n$ 抽样分布

χ^2 分布

定义：设 $X_1, X_2, \dots, X_n$ 是来自标准正态体 $N(0,1)$ 的样本，则称统计量

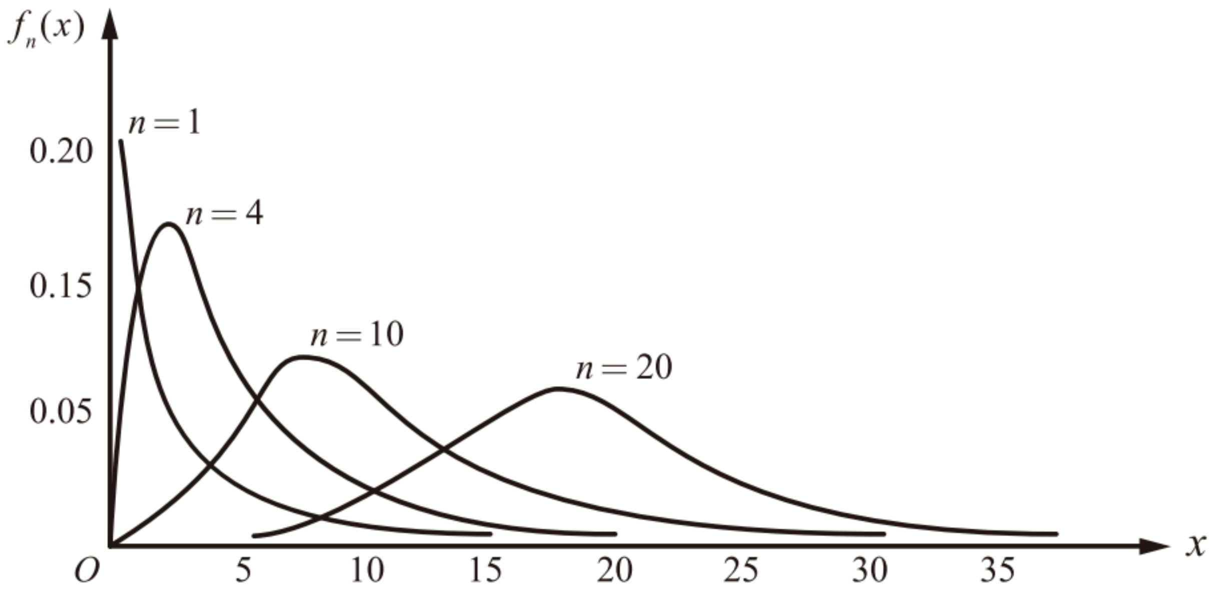
$$X^2 = X_1^2 + X_2^2 + \cdots + X_n^2 \tag{7}$$

服从自由度为n的 X^2 分布，记为 $X^2 \sim \chi^2(n)$。

$X^2(n)$ 分布的概率密度

$$f(y) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1} e^{-y/2}, & y > 0 \\ 0, & \text{其他} \end{cases} \tag{8}$$

概率密度函数的图形，如下图：



性质：

可加性：$\chi_1^2 \sim \chi^2(n_1)$，$\chi_2^2 \sim \chi^2(n_2)$，$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1+n_2)$

$\chi^2 \sim \chi^2(n)$，则 $E[\chi^2(n)] = n, D(\chi^2(n)) = 2n$

t 分布

定义：设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，且 X 和 Y 相互独立，则统计量

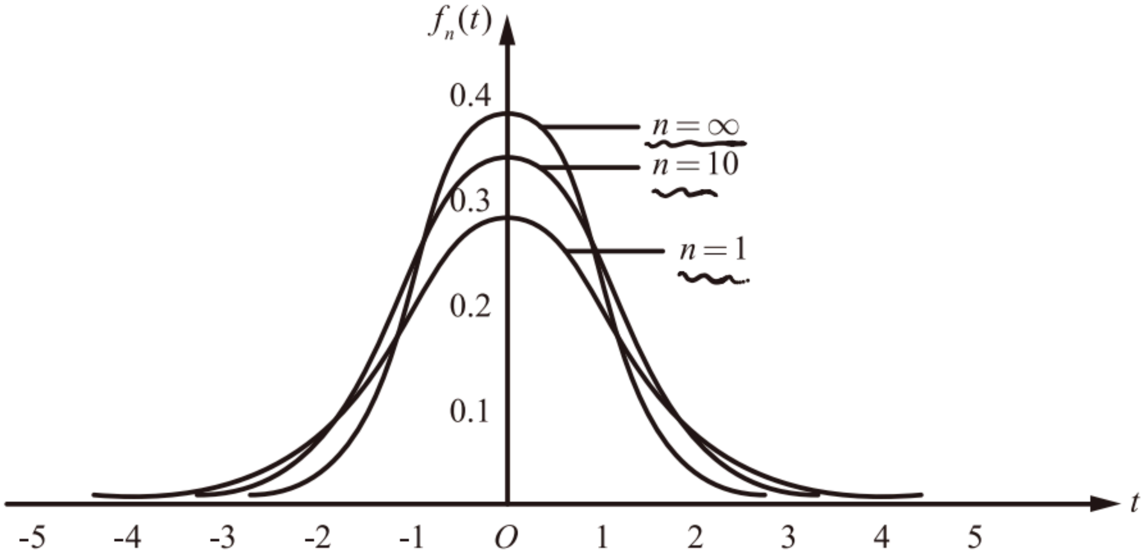
$$t = \frac{X}{\sqrt{Y/n}} \tag{9}$$

服从自由度为n的 t 分布，记为 $t \sim t(n)$。

$t(n)$ 分布的概率密度函数为

$$h(t) = \frac{\Gamma\left(\frac{(n+1)}{2}\right)}{\sqrt{\pi n}\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < +\infty \tag{10}$$

概率密度函数的图形，如下图：



性质

1. $h(t)$ 概率密度函数关于 $t=0$ 对称
2. 当 $n\to\infty$，$t(n)\sim N(0,1)$

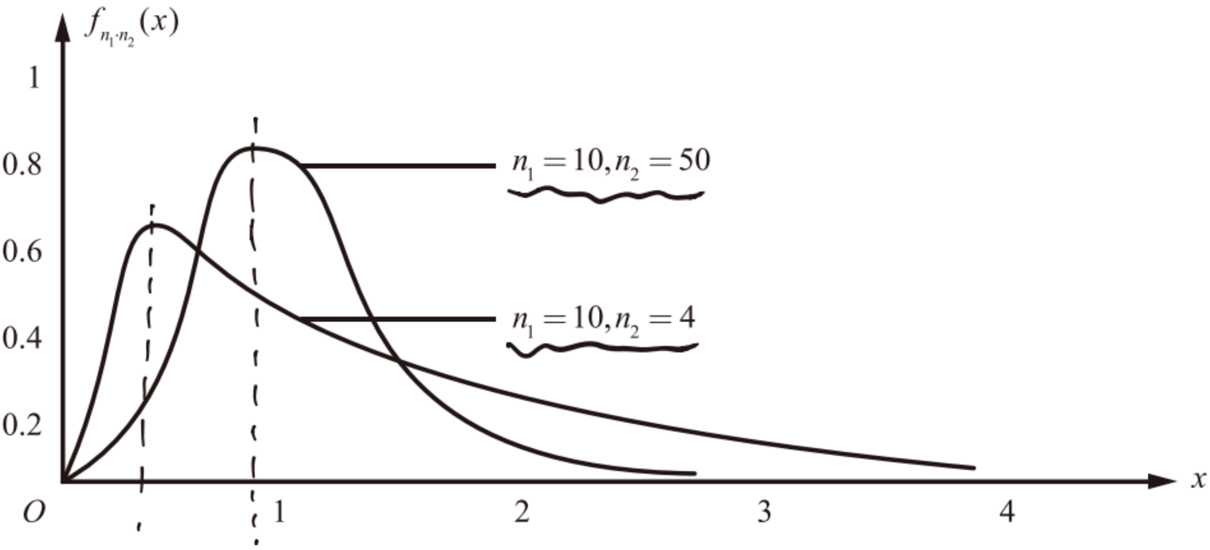
F 分布

定义：设 $U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，且 U 和 V 相互独立，则称随机变量

$$F=\frac{U/n_1}{V/n_2} \tag{11}$$

服从自由度为 (n_1,n_2) 的 F 分布，记为 $F\sim F(n_1,n_2)$。

$F(n_1,n_2)$ 分布的概率密度函数略



单峰图形。

性质

若 $F \sim F(n_1, n_2)$，则 $\frac{1}{F} \sim F(n_2, n_1)$

51.分位点

上 α 分位点

设 $X \sim f(x)$，若对给定的正数 $\alpha(0,1)$，有数 Z_α，满足

$$P\{X > Z_\alpha\} = \int_{Z_\alpha}^{+\infty} f(x)dx = \alpha \tag{12}$$

则称 Z_α 是 X 的上 α 分位点。

$N(0,1)$ 分位点

设 $X \sim N(0,1)$，若对给定的正数 $\alpha(0,1)$，有数 Z_α，满足

$$P\{X > Z_\alpha\} = \int_{Z_\alpha}^{+\infty} f(x)dx = \alpha \tag{13}$$

则称 Z_α 是 $N(0,1)$ 的上 α 分位点。

与 x 轴的交点就是 Z_α

χ^2 分位点

设 $X^2 \sim \chi^2(n)$，若对给定的正数 $\alpha, 0 < \alpha < 1$，满足条件

$$P\{X^2 > \chi^2(n,\alpha)\} = \int_{\chi^2(n,\alpha)}^{+\infty} f(y)\, dy = \alpha \tag{14}$$

的点 $\chi^2(n,\alpha)$ 为 $\chi^2(n)$ 的上分位点。

t 分位点

设 $t \sim t(n)$，若对给定的正数 $\alpha, 0 < \alpha < 1$，满足条件

$$P\{t > t_\alpha(n)\} = \int_{t_\alpha(n)}^{+\infty} h(t)\, dt = \alpha \tag{15}$$

的点 $t_\alpha(n)$ 为 $t(n)$ 的上分位点

F 分位点

设 $F \sim F(n_1, n_2)$，若对给定的正数 $\alpha, 0 < \alpha < 1$，满足条件

$$P\{F > F_\alpha(n_1, n_2)\} = \int_{F_\alpha(n_1,n_2)}^{+\infty} f(y)\, dy = \alpha \tag{16}$$

的点 $F_\alpha(n_1, n_2)$ 为 $F(n_1, n_2)$ 分布的上分位点。

52. 正态总体的样本均值与样本方差的分布

设总体 X 的 $E(X)=\mu, D(X)=\sigma^2$，$X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本，则有

$$E(\overline{X})=\mu, \quad D(\overline{X})=\frac{\sigma^2}{n}, \quad E(S^2)=\sigma^2. \tag{17}$$

1. 样本均值 $\bar{X}$ 的期望等于总体 X 的期望 $E[X]$
2. 样本均值 $\bar{X}$ 的方差等于总体 X 的方差除以样本容量
3. 样本方差 S^2 的期望等于总体的方差 $D(X)$

设 $X_1, X_2, \ldots, X_n$ 是来自正态总体 $N(\mu_1, \sigma_1^2)$ 的样本，$Y_1, Y_2, \ldots, Y_{n_2}$ 是来自正态总体 $N(\mu_2, \sigma_2^2)$ 的样本，并且这两个样本相互独立，其中

$$S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2, \quad S_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2, \tag{18}$$

则有：

1. $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$;
2. 当两总体的方差相同，即 $\sigma_1^2=\sigma_2^2$，有

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_\omega\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2) \tag{19}$$

其中

$$S_\omega=\sqrt{\frac{1}{n_1+n_2-2}\left[(n_1-1)S_1^2+(n_2-1)S_2^2\right]}. \tag{20}$$