数理统计

基本概念

1. 总体：研究对象的全体
    - 一般为所研究对象的某个（或某些）数量指标所有可能取值的全体，它对应着一个随机变量（一维或者多维）。记为 X 。

    - 随机变量 X 的分布函数和数字特征又称为总体的分布函数和数字特征。

2. 个体：组成总体的每一个基本单元
   即总体数量指标的某次取值，亦即随机变量 X 的某次取值，用 x_i 表示。

3. 样本：从总体中抽取的部分个体（对总体进行多次观测的结果）
    - 对样本中每个个体，在获知其观测结果之前，它有可能取到总体的所有可能取值，故可用随机变量 X_i 表示（X_i 通常假定与总体 X 同分布）。

    - 样本表示为 $(X_1, X_2, \cdots, X_n)$ ，n 为样本容量，在一次试验中样本的观测值 $(x_1, x_2, \cdots, x_n)$ 称为样本的一个实现，或称为总体 X 的一个容量为 n 的样本值。

4. 简单随机样本：指总体的一个样本 $(X_1, X_2, \cdots, X_n)$ ，该样本满足：
    - $X_1, X_2, \cdots, X_n$ 与 X 有相同的分布
    - $X_1, X_2, \cdots, X_n$ 相互独立

5. 统计量
    - 设 $(X_1, X_2, \cdots, X_n)$ 是取自总体 X 的一个样本，现在有一个实值连续函数 $g(r_1, r_2, \cdots, r_n)$ 其不含有未知参数，则称随机变量 $g(X_1, X_2, \cdots, X_n)$ 为统计量。

    - 设 $(x_1, x_2, \cdots, x_n)$ 是一个样本值，称 $g(x_1, x_2, \cdots, x_n)$ 为统计量 $g(X_1, X_2, \cdots, X_n)$ 的一个观察值。

NOTE: 统计量是随机变量，比如 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$

样本的统计特征

设总体 X 的分布函数为 $F_X(x)$ ，则样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数为：

$$F(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F_X(x_i)$$

总体分布的直接近似

- 频率分布表(离散型)：用一张表来统计样本中出现每个值的频率。

- 频率直方图(连续型)：统计样本中每个点落在区间段内的频率。

经验分布函数

- 定义：设总特的样本值 $x_1, x_2, \cdots, x_n$ ，又设 Y_n 等可能地取到这 n 哥值中的每一个，称 Y_n 的分布函数 $F_n(x)$ 为总体 X 的经验分布函数。若设 $x_1 \le x_2 \le \cdots x_n$ ，则

$$F_n(x) = \begin{cases} 0, & x < x_1 \\ k/n, \quad x_k \le x < x_{k+1}, & k = 1, 2, \cdots, n-1 \\ 1, & x_n \le x \end{cases}$$

- 意义：$\forall x$ ，事件 $A = \{X \le x\}$ 在 n 次 Bernoulli 试验中发生次数，等于 $x_1, x_2, \ldots, x_n$ 中小于 x 的个数，其频率 $f_n(A)$ 等于 $F_n(x)$ ，而 $P(A) = F(x)$ ，由 Bernoulli 大数定律，$\forall \epsilon > 0$ ，有

$$\lim_{n\to+\infty} P\{|F_n(x)-F(x)|<\epsilon\}=1$$

即 $F_n(x)$ 依概率收敛于 $F(x)$ 。更强的，我们有：

$$P\{\lim_{n\to+\infty} sup_{-\infty<x<+\infty}|F_n(x)-F(x)|<\epsilon\}=1$$

常用统计量

设 $(X_1,X_2,\ldots,X_n)$ 是来自总体 X 的一个容量为 n 的样本，设 $E(X)=\mu,\ \ D(X)=\sigma^2$ ，则有统计量：

1. 样本均差：

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

   - $E(\bar{X})=\mu,\ \ D(\bar{X})=\frac{\sigma^2}{n}$

2. 样本方差：

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

   - $E(S^2)=\sigma^2$
     NOTE: 这里除以 $n-1$ 是因为 $\bar{X}$ 也是随机变量，而不是常数。只有除以 $n-1$ 才能保证 $E(S^2)$ 即 S^2 的期望值是 $D(X)$ 。

3. 样本标准差：

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

4. 样本 k 阶原点矩：

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$$

5. 顺序统计量与极差：
   对样本 $(X_1,X_2,\ldots,X_n)$ ，将其由小到大进行排序 $X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$ ，则称 $(X_{(1)},X_{(2)},\cdots,X_{(n)})$ 为顺序统计量，并称 $X_{(k)}$ 为第 k 个顺序统计量。称 $X_{(1)}$ 和 $X_{(n)}$ 为最大和最小顺序统计量，称 $D_n=X_{(n)}-X_{(1)}$ 为极差（统计量）。

来自正态总体常用统计量及其分布

$\chi^2(n)$ 分布(n 为自由度)

- 定义：设 $X_1,X_2,\cdots,X_n$ 为标准正态总体 $N(0,1)$ 的容量为 n 的样本，则称如下统计量为 χ^2 统计量：

$$\chi^2\triangleq\sum_{i=1}^{n}X_i^2\sim\chi^2(n)$$

  其分布称为自由度为 n 的 χ^2 分布。其概率密度满足：

$$f(x)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}, & x>0\\ 0, & x\le 0\end{cases}$$

- 性质：
  1. $E(\chi^2)=nE(X_i^2)=n$
  2. $D(\chi^2)=nD(X_i^2)=2n$
  3. 设 $X_1\sim\chi^2(n_1),\ \ X_2\sim\chi^2(n_2)$ ，$X_1,\ X_2$ **相互独立**，则

$$X_1+X_2\sim\chi^2(n_1+n_2)$$

（事实上 X_1+X_2 可以看作 n_1+n_2 个正态分布的和。）

4. $n\to+\infty$，$\chi^2\to$ 正态分布

t(n) 分布（n 为自由度）

- 定义：设 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，X,Y 相互独立，则称统计量：

$$T=\frac{X}{\sqrt{Y/n}}\sim t(n)$$

为 t 统计量，其分布称为自由度为 n 的 t 分布，其密度函数为：

$$f(t)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{t^2}{n})^{\frac{n+1}{2}}\quad -\infty<t<+\infty$$

- 性质：
  1. $f_n(t)$ 是偶函数
  2. $n\to+\infty$，$f_n(t)\to\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$

$F(n,m)$ 分布（n,m 分别为第一、二自由度）

- 定义：设 $X\sim\chi^2(n)$，$Y\sim\chi^2(m)$，X,Y 相互独立，则称：

$$F\triangleq\frac{X/n}{Y/m}\sim F(n,m)$$

为 F 统计量，其分布称为第一、二自由度分别为 n,m 的 F 分布。其分布密度函数为：

$$f(t,n,m)=\begin{cases}\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}(\frac{n}{m})^{\frac{n}{2}}t^{\frac{n}{2}-1}(1+\frac{n}{m}t)^{-\frac{n+m}{2}}, & t>0\\ 0, & t\le 0\end{cases}$$

- 性质：
  1. 若 $F\sim F(n,m)$，则 $\frac{1}{F}\sim F(m,n)$
  2. 设 $F(n,m)$ 的上侧分位数为 $F_\alpha(m,n)$，则 $F_{1-\alpha}(n,m)=\frac{1}{F_\alpha(m,n)}$
     - 证明：

$$\begin{aligned}&P(F>F_{1-\alpha}(n,m))\\=&1-\alpha\\=&P(\frac{1}{F}\le\frac{1}{F_{1-\alpha}(n,m)})\\=&1-P(\frac{1}{F}>\frac{1}{F_{1-\alpha}(n,m)})\\ \\ \Rightarrow&P(\frac{1}{F}>\frac{1}{F_{1-\alpha}(n,m)})=\alpha\end{aligned}$$

由于

$$\frac{1}{F}\sim F(n,m)$$

故结论成立

正态总体的样本均值与样本方差的一些结论

设总体 $X\sim N(\mu\sigma^2)$，样本为 $(X_1,X_2,\cdots,X_n)$

1. 样本均值：

$$\bar{X}\sim N(\mu,\frac{\sigma^2}{n})$$

- 证明：直接利用 $\bar{X}=\frac{1}{n}(X_1+X_2+\cdots+X_n)$ 以及 X_i 相互独立的性质即可。

2. $$\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}(\frac{X_i-\bar{X}}{\sigma^2})\sim\chi^2(n-1)$$

3. $\frac{(n-1)S^2}{\sigma^2}$ 与 $\bar{X}$ 相互独立，即样本均值与样本方差相互独立。

4. $$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\div\frac{S}{\sigma}=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$