6. 数理统计的基本概念

6.1 总体与样本

总体：研究对象的全体构成的集合.（例如 整批灯泡寿命）

个体：组成总体的每个元素.（例如每个灯泡寿命）.

设 $X_1, \cdots, X_n$ i.i.d. 分布函数为 $F(x)$.（分布未知）

试验前：$X_1, X_2, \cdots X_n$ 为 r.v.

试验后：$x_1, x_2, \cdots, x_n$ 为样本观察值.（实数）

Q：由 $x_1, \cdots, x_n$ 推断分布特征.

样本：$(X_1, \cdots, X_n)$ i.i.d. 与总体 X 同分布

样本观测值（或样本实现）：$(x_1, \cdots, x_n)$

理论分布：总体 X 的分布

理论分布函数：总体 X 的分布函数.

若总体 X 分布函数为 $F(x)$，则样本 $(X_1, \cdots, X_n)$ 分布函数为

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

对于离散型总体，

$$P(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

例如，若 $X \sim B(1, p)$，则

$$P(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{(1-x_i)} = p^{\sum_{i=1}^{n} x_i}(1-p)^{\sum_{i=1}^{n}(1-x_i)}$$

对于连续型总体，

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i).$$

例如，若 $X \sim N(0,1)$，则

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = (2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}.$$

Q：如何由样本推断总体的分布？

经验分布函数 观测值 $x_1, \cdots, x_n$. 定义

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{x_i \le x\}, \quad x \in \mathbb{R}.$$

等价地，将 $(x_1, \cdots, x_n)$ 由小到大排列，$x_1^* \le x_2^* \le \cdots \le x_n^*$，则

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

由大数律，$\frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\} \xrightarrow{P} F(x)$. $\forall x \in \mathbb{R}$. 实际上，上述收敛关于 x 一致.

Glivento 定理 $P\left(\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\right) = 1$.

统计量 设 $(X_1, \cdots, X_n)$ 为总体 X 的一个样本，若

1) $T = g(x_1, \cdots, x_n)$ 为连续函数

2) $T = g(x_1, \cdots, x_n)$ 中不含有总体的未知参数

则称 $T = g(X_1, \cdots, X_n)$ 为**统计量**. 称 $t = g(x_1, \cdots, x_n)$ 为**统计量观测值**.

<u>例.</u> 设 $X_1, \cdots, X_n$ 为来自总体 $N(\mu, \sigma^2)$ 的样本. 其中，μ 已知，σ^2 未知. 则下列______为统计量. A. B. D.

A. $X_1 + X_n$.

B. $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$

C. $\sum_{i=1}^{n}\frac{|X_i|}{\sigma}$

D. $\min_{1\le i\le n} X_i$

常用统计量

	提炼EX的信息
1) 样本均值 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$	
2) 样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$	DX
样本标准差 S	
3) 样本k阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n} X_i^k$	EX^k
4) 样本k阶中心矩 $B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$	$E(X-EX)^k$

5) 顺序统计量

$$X_1^* \le X_2^* \le \cdots \le X_n^*$$

样本中位数

$$\tilde{X}=\begin{cases} X_{m+1}^*, & n=2m+1 \\ \frac{X_m^*+X_{m+1}^*}{2}, & n=2m \end{cases}$$

样本极差 $R=X_n^*-X_1^*$

6.2 抽样分布

抽样分布：统计量的分布

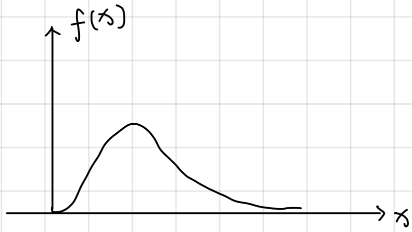
6.2.1 χ^2分布

$X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(0,1)$. 称

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

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

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



$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x}\,dx$$

数字特征. $E\chi^2 = nEX_1^2 = n$

$$D\chi^2 = nDX_1^2 = n(EX_1^4 - EX_1^2) = n(3-1) = 2n$$

$$(\ EX_1^{2n} = (2n-1)!! \ )$$

可加性. $\chi_1^2 \sim \chi^2(n)$, $\chi_2^2 \sim \chi^2(m)$, 且独立，则

$$\chi_1^2 + \chi_2^2 \sim \chi^2(m+n).$$

上侧α分位点 $\chi_\alpha^2(n)$.

$$P(\chi^2(n) > \chi_\alpha^2(n)) = \alpha.$$

n足够大时$(n>45)$,

$$\sqrt{2\chi^2(n)} \approx N(\sqrt{2n-1}, 1)$$

$$\left(\because \sqrt{2\chi^2(n)} - \sqrt{2n-1} = \frac{2\chi^2(n) - (2n-1)}{\sqrt{2\chi^2(n)} + \sqrt{2n-1}} \approx \frac{\chi^2(n) - n}{2\sqrt{2n}} \approx N(0,1) \right)$$

$\therefore \alpha = P(\chi^2(n) > \chi^2_\alpha(n)) = P(\sqrt{2\chi^2(n)} > \sqrt{2\chi^2_\alpha(n)})$

$\approx P(N(\sqrt{2n-1}, 1) > \sqrt{2\chi^2_\alpha(n)})$

$= P(N(0,1) > \sqrt{2\chi^2_\alpha(n)} - \sqrt{2n-1})$

$\therefore \quad u_\alpha \approx \sqrt{2\chi^2_\alpha(n)} - \sqrt{2n-1}$

$P(N(0,1) > u_\alpha) = \alpha. \qquad u_\alpha = \Phi^{-1}(1-\alpha)$

即 $\quad \chi^2_\alpha(n) \approx \frac{1}{2}(u_\alpha + \sqrt{2n-1})^2$

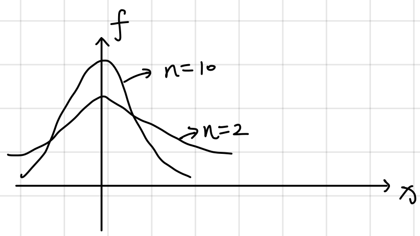
6.2.2. t 分布

$X \sim N(0,1)$, $Y \sim \chi^2(n)$ 且独立. 称

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

服从自由度为 n 的 t (student) 分布, 记为 $T \sim t(n)$.

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R}.$$



数字特征 $\quad ET = 0$

渐近正态 $\quad \lim_{n\to\infty} f(x) = e^{-\frac{x^2}{2}}$.

上侧 α 分位点 $\quad t_\alpha(n)$

$$P(T > t_\alpha(n)) = \alpha.$$

$n > 45$ 时, $\quad t_\alpha(n) \approx u_\alpha$.

6.2.3 F 分布

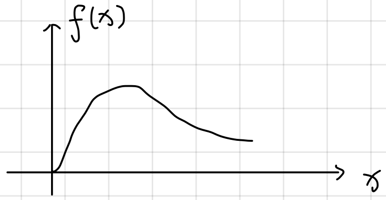
$X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且独立. 称

$$F = \frac{X/n_1}{Y/n_2}$$

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

由定义, $\quad \frac{1}{F} \sim F(n_2, n_1)$.

$$f(x)=\begin{cases}\dfrac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\dfrac{n_1}{n_2}\right)\left(\dfrac{n_1}{n_2}x\right)^{\frac{n_1}{2}-1}\left(1+\dfrac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x\geqslant 0\\ 0, & x<0.\end{cases}$$



上侧α分位点 $F_\alpha(n_1,n_2)$.

$$P(F(n_1,n_2)>F_\alpha(n_1,n_2))=\alpha.$$

$\because \alpha=P(F(n_1,n_2)>F_\alpha(n_1,n_2))=P\left(\dfrac{1}{F(n_1,n_2)}<\dfrac{1}{F_\alpha(n_1,n_2)}\right)$

$=P\left(F(n_2,n_1)<\dfrac{1}{F_\alpha(n_1,n_2)}\right)$

$\therefore \quad P\left(F(n_2,n_1)\geqslant\dfrac{1}{F_\alpha(n_1,n_2)}\right)=1-\alpha$

即 $\quad F_{1-\alpha}(n_2,n_1)=\dfrac{1}{F_\alpha(n_1,n_2)}$.

6.2.4. 基本抽样定理.

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

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

则 (1) $\bar{X} \sim N(\mu, \frac{1}{n}\sigma^2)$

(2) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

(3) $\bar{X}$ 与 S^2 独立.

解释. (1) $E\bar{X} = \mu$, $D\bar{X} = \frac{1}{n}\sigma^2$.

(2) $X_i - \bar{X} = -\frac{1}{n}\sum_{j\neq i} X_j + (1-\frac{1}{n}) X_i \sim N(0, [\frac{n-1}{n^2} + (1-\frac{1}{n})^2]\sigma^2)$

$\sim N(0, \frac{n-1}{n}\sigma^2)$

(3) $\bar{X} \xrightarrow{P} \mu$, $S^2 \xrightarrow{P} \sigma^2$. $\square$

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

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

且 独立

$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

$\therefore \quad \frac{\bar{X} - \mu}{\sigma}\cdot\sqrt{n} \Big/ \frac{S}{\sigma} = \sqrt{n}\,\frac{\bar{X} - \mu}{S} \sim t(n-1).$

推论2. 设 $X_1, \cdots, X_{n_1} \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$: $\bar{X}$, S_1^2

$Y_1, \cdots, Y_{n_2} \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$: $\bar{Y}$, S_2^2

且独立. 则

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

(2) $$\frac{1}{n_1}\sum_{i=1}^{n}\left(\frac{X_i-\mu_1}{\sigma_1}\right)^2 \Big/ \frac{1}{n_2}\sum_{i=1}^{n}\left(\frac{Y_i-\mu_2}{\sigma_2}\right)^2 \sim F(n_1, n_2).$$

证. 由 $F(n_1, n_2) = \dfrac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2}$ 易证. $\square$

推论3. 条件同推论2, 且 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 则

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

其中 $$S_w^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

证. $$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$$

$$\frac{\bar{X} - \bar{Y} - (\mu_1-\mu_2)}{\sigma} \sim N\left(0, \frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$t(n) = \frac{N(0,1)}{\sqrt{\chi^2(n)/n}}.$$ $\square$