第六章：数理统计的基本概念

概率和统计

概率论和统计的区别：

1. **概率论**:已知随机变量X的分布(分布列或者分布密度函数)，进而根据随机变量的分布进行事件概率和随机变量矩计算
2. **统计**:总体的分布未知或者不完全已知的情形下，通过随机抽样并研究抽样得到数据的统计特征，以此来估计总体的分布。

问题提出

Question ?

对某个学校的小学生身高X进行调查，确定其分布函数？X为待研究的**总体**，构成总体的每个成员被称为**一个个体**。

研究方法

对总体进行**简单随机抽样**，并根据抽样结果**推断总体分布**

抽样 从X中随机的抽出n个个体，记为n**维样本**。由于抽取的个体是随机的，即无法提前预知其数值，样本为n维随机向量$(X_1, X_2, \cdots, X_n)$。

简单随机抽样

简单随机抽样:

1. 样本与总体**同分布**
2. $X_1, X_2, \cdots, X_n$**相互独立**

总体和样本 I

1. **总体** 是一个分布待确定的**随机变量**X。总体中包含很多个**个体**。
2. **样本** 从总体中随机抽取的n个个体，为n维**随机向量**$(X_1, X_2, \cdots, X_n)$。对样本进行一次观测，记$(x_1, x_2, \cdots, x_n)$为样本的**一个观测值**。

分布函数 I

命题6.1.1

设总体$X \sim F_X(\cdot,\theta)$，$(X_1, X_2, \cdots, X_n)$为其样本，那么

$$F_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n; \theta) = F_X(x_1,\theta)F_X(x_2,\theta)\cdots F_X(x_n,\theta)$$

经验分布函数 I

问题

由样本估计总体的分布函数。

经验分布函数，设总体的样本$(X_1, X_2, \cdots, X_n)$的一次观测值为$(x_1, x_2, \cdots, x_n)$，并将其进行升序排列$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$，并定义

$$F_n^X(x) = \begin{cases} 0, & x < x_{(1)} \\ \cdots \\ \dfrac{k}{n}, & x \in [x_{(k)}, x_{(k+1)}) \\ \cdots \\ 1, & x \geq x_{(n)} \end{cases}$$

经验分布函数 II

称F_n^X为总体X的一个**经验分布函数，或样本分布函数**。

克里汶克定理

设$X_1, X_2, \cdots, X_n$是取自总体X的容量为n的样本，总体的分布函数为$F_X(x)$，$F_n(x)$为其经验分布函数，那么

$$P\left(\lim_{n\to\infty} sup_{-\infty<x<\infty}\left|F_n^X(x) - F_X(x)\right| = 0\right) = 1$$

上面的结果说明，就是当样本容量足够大的时候，用经验分布函数来推断总体的分布函数是合理的。

练习

某食品厂生产饮料，现在从生产线(X)上任意抽取5瓶，称得重量依次为(g)

$$351 \quad 347 \quad 351 \quad 344 \quad 351$$

请写出总体X的经验分布函数

统计量

注意下面这个关系

$$\text{样本} \xrightarrow{\text{统计量}} \text{总体}$$

统计量是通过样本研究总体的手段

统计量

统计量是样本的函数，不显含总体的未知参数，即统计量由样本唯一决定。

常用的统计量 I

1. 样本均值:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2. 样本方差:

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

修正样本方差

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

常用的统计量 II

3. **样本的k阶原点矩**

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

4. **样本的k阶中心矩**

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

5. **顺序统计量** $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$，其中$X_{(1)} = \min\{X_1, X_2, \cdots, X_n\}$，$X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$，而$X_{(k)}$是将$X_1, X_2, \cdots, X_n$的取值从小到大排列的第$k$位的值。

常用的统计量 III

6. 样本中位数

$$\tilde{X} = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & 若n为奇数 \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right), & 若n为偶数 \end{cases}$$

7. 样本极差

$$R_n^X = X_{(n)} - X_{(1)}$$

抽样分布 I

设计合适的统计量可以用来研究总体的数字特征和参数。

命题6.3.1

设$(X_1, X_2, \cdots, X_n)$是来自总体X的样本，$E[X] = \mu$，$Var[X] = \sigma^2$，那么

1. $$E[\overline{X}] = \mu, \quad Var[\overline{X}] = \frac{\sigma^2}{n}$$

2. $$E[S_n^2] = \frac{n-1}{n}\sigma^2, \quad E[S_n^{*2}] = \sigma^2$$

证明：

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

再由X_i之间的独立性有

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

对于S_n

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \end{aligned}$$

所以

$$
\begin{aligned}
E[S_n^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\overline{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}(Var[X_i] + (E[X_i])^2) - (Var[\overline{X}] + (E[\overline{X}])^2) \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

$$
E[S_n^{*2}] = E\left[\frac{n}{n-1}S_n^2\right] = \frac{n}{n-1}E[S_n^2] = \sigma^2
$$

χ^2分布

命题6.3.2

设总体$X \sim N(0,1)$，$(X_1, X_2, \cdots, X_n)$为其简单随机样本，则

$$\boldsymbol{X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)}$$

其中

$$\chi^2(n) = \Gamma(\frac{n}{2}, \frac{1}{2}).$$

三种重要的抽样分布 II

若$X\sim\Gamma(\frac{n}{2},\frac{1}{2})$，则称$X$服从自由度为$n$的$\chi^2$分布，记为$X\sim\chi^2(n)$，其分布密度函数为

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

$$\Gamma(\alpha)=\int_0^{\infty}t^{\alpha-1}e^{-t}\,\mathrm{d}t$$

三种重要的抽样分布 III

特别的，根据例题2.4.2，当$X \sim N(0,1)$，

$$X^2 \sim \chi^2(1)$$

即X^2服从自由度为1的χ^2分布。

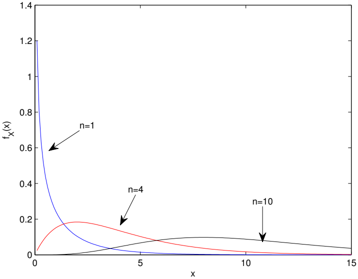


Figure : $\chi^2(n)$的分布密度

χ^2分布的性质 I

1. 若$X \sim \chi^2(n)$，那么$E[X] = n$，$Var[X] = 2n$。
2. 若$X_1 \sim \chi^2(n_1)$，$X_2 \sim \chi^2(n_2)$，且X_1与X_2独立，那么$X_1 + X_2 \sim \chi^2(n_1 + n_2)$。
3. 若$X \sim \chi^2(n)$，那么$\frac{X-n}{\sqrt{2n}}$近似服从$N(0,1)$，当$n \to \infty$

练习

1. 设$(X_1, X_2, \cdots, X_5)$是来自总体$N(0,4)$的一个样本，且

$$Y = aX_1^2 + b(2X_2 + 3X_3)^2 + c(4X_4 - X_5)^2$$

请问$a, b, c > 0$取什么值的时候，随机变量Y服从$\chi^2(n)$分布，n为多少？

2. 设$X_1, X_2, \cdots, X_{20}$是来自总体$X \sim N(0,1)$的 一个样本，记

$$Y = \frac{1}{10}\left(\sum_{i=1}^{10} X_i\right)^2 + \frac{1}{10}\left(\sum_{i=11}^{20} X_i\right)^2$$

请问Y服从什么分布？

t分布 I

命题6.3.3

设$X \sim N(0,1)$，$Y \sim \chi^2(n)$，且X与Y相互独立，令

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

则T的分布密度函数f_T为

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

t分布 II

$X \sim t(n)$表示X服从自由度为n的t分布

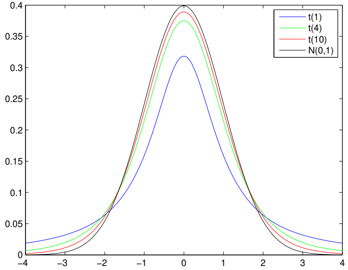


Figure : $t(n)$分布密度函数

t分布的性质

1. t分布关于原点对称
2. 当$n \to \infty$，$t(n)$趋于标准正态分布
3. $$E[X] = 0(n > 1), Var[X] = \frac{n}{n-2}(n > 2)$$

练习

设$(X_1, X_2, \cdots, X_9)$是来自总体$N(0,1)$的简单随机样本，请确定正数C，使得

$$\frac{C(X_1+X_2+X_3)}{\sqrt{(X_4+X_5)^2+(X_6+X_7)^2+(X_8+X_9)^2}}$$

服从t分布，并指出其自由度。

F分布 I

命题6.3.4

若$X \sim \chi^2(m)$，$Y \sim \chi^2(n)$，且X与Y相互独立，令

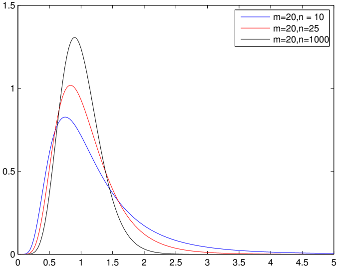
$$\boldsymbol{Z = \frac{X/m}{Y/n} \sim F(m,n)},$$

则Z的分布密度函数为

$$f_Z(x) = \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{n}{2}} n^{\frac{m}{2}} \dfrac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

F分布 II

称随机变量Z服从**第一自由度为**m，**第二自由度为**n的F分布，记作 $Z \sim F(m,n)$。



Figure： $F(m,n)$分布密度函数

F分布 III

$Z \sim F(m,n)$，那么$\boldsymbol{\frac{1}{Z} \sim F(n,m)}$。

证明：

$$
\begin{aligned}
F_{\frac{1}{Z}}(t) &= P(\frac{1}{Z} \leq t) \\
&= P(Z \geq \frac{1}{t}) \\
&= 1 - P(Z < \frac{1}{t}) \\
&= 1 - \int_0^{\frac{1}{t}} f_Z(x)\,\mathrm{d}x
\end{aligned}
$$

F分布 IV

故

$$
\begin{aligned}
f_{\frac{1}{Z}}(t) = \frac{\mathrm{d}F_{\frac{1}{Z}}(t)}{\mathrm{d}t} &= t^{-2} f_Z\left(\frac{1}{t}\right) \\
&= t^{-2} C_{m,n} \frac{t^{1-\frac{m}{2}}}{(mt^{-1}+n)^{\frac{m+n}{2}}} \\
&= t^{-2} C_{m,n} \frac{t^{\frac{m+n}{2}-\frac{m}{2}+1}}{(nt+m)^{\frac{m+n}{2}}} \\
&= C_{m,n} \frac{t^{\frac{n}{2}-1}}{(nt+m)^{\frac{m+n}{2}}}
\end{aligned}
$$

其中

$$
C_{m,n} = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{n}{2}} n^{\frac{m}{2}}
$$

练习

设$X_1, X_2, \cdots, X_8$是来自总体$N(0,1)$的简单随机样本，求常数c，使得

$$\frac{c(X_1^2 + X_2^2)}{(X_3 + X_4 + X_5)^2 + (X_6 + X_7 + X_8)^2}$$

服从F分布，并指出其自由度

分位数 I

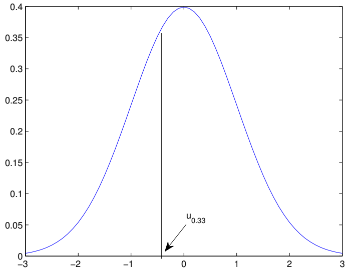
分位数

设$X \sim \phi(n)$，$0 < \alpha < 1$，称满足

$$P(X \leq \phi_\alpha(n)) = \alpha$$

的实数$\phi_\alpha(n)$为分布$\phi(n)$的α**分位数**，或叫做α**分位点**。

分位数 II



上图是标准正态分布的密度区线，其中$u_{0.33}$是其0.33分位点，表示

$$P(X \leq u_{0.33}) = 0.33$$

计算分位数 I

计算分位数时，有几个问题需要注意

1. 若$X \sim t(n)$，由于$\lim_{n\to\infty} X \sim N(0,1)$，

$$\lim_{n\to\infty} t_\alpha(n) = u_\alpha$$

2. 若$X \sim \chi^2(n)$ 分布，当$n \to \infty$，$(X-n)/\sqrt{2n} \sim N(0,1)$，所以

$$\alpha = P(X \le \chi^2_\alpha(n)) = P((X-n)/\sqrt{2n} \le (\chi^2_\alpha(n)-n)/\sqrt{2n})$$

即

$$u_\alpha = (\chi^2_\alpha(n)-n)/\sqrt{2n}$$

解出

$$\chi^2_\alpha(n) = n + \sqrt{2n}u_\alpha, \quad n \to \infty$$

计算分位数 II

3 若X服从$F(m,n)$，则有$\frac{1}{X} \sim F(n,m)$，若已知$F_\alpha(m,n)$，那么

$$\begin{aligned}\alpha = P(X < F_\alpha(m,n)) &= P(\frac{1}{X} > \frac{1}{F_\alpha(m,n)}) \\ &= 1 - P(\frac{1}{X} \leq \frac{1}{F_\alpha(m,n)})\end{aligned}$$

即

$$P(\frac{1}{X} \leq \frac{1}{F_\alpha(m,n)}) = 1 - \alpha$$

也即

$$\boldsymbol{\frac{1}{F_\alpha(m,n)} = F_{1-\alpha}(n,m)}$$

课堂练习

求下列分位数

1. $\chi^2_{0.99}(10), \chi^2_{0.05}(20), \chi^2_{0.95}(60)$
2. $t_{0.95}(10), t_{0.1}(20), t_{0.9}(50)$
3. $F_{0.99}(5,4), F_{0.05}(3,7)$

正态总体的抽样分布 I

抽样分布基本定理

设总体$X \sim N(0,1)$，$(X_1, X_2, \cdots, X_n)$为其样本，则样本均值$\overline{X} \sim N(0, \frac{1}{n})$，$nS_n^2 \sim \chi^2(n-1)$，并且$\overline{X}$与$S_n^2$相互独立。

证明略。

1. 对于服从正态分布的总体，样本均值$\overline{X}$服从一维正态分布。
2. $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$，即$S_n^2$是$\overline{X}$的函数，但是它们却相互独立。

几个推论 I

1. **推论6.3.1** 设总体$X \sim N(\mu, \sigma^2)$，$(X_1, X_2, \cdots, X_n)$为其样本，则$\overline{X}$与S_n^2**相互独立**，且

$$\overline{X} \sim N(\mu, \sigma^2/n)$$
$$\frac{(n-1)S_n^{*2}}{\sigma^2} = \frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1)$$

2. **推论6.3.2** 设总体$X \sim N(\mu, \sigma^2)$，$(X_1, X_2, \cdots, X_n)$为其样本，那么

$$\frac{\overline{X} - \mu}{S_n^*}\sqrt{n} = \frac{\overline{X} - \mu}{S_n}\sqrt{n-1} \sim t(n-1)$$

几个推论 II

3 **推论6.3.3** 设总体$X \sim N(\mu_1, \sigma_1^2)$，$(X_1, X_2, \cdots, X_m)$为其样本，样本均值为$\overline{X}$，样本方差为$S_{1m}^2$；另有与$X$独立的总体$Y \sim N(\mu_2, \sigma_2^2)$，$(Y_1, Y_2, \cdots, Y_n)$为其样本，样本均值为$\overline{Y}$，样本方差为$S_{2n}^2$，那么

$$\frac{mS_{1m}^2}{nS_{2n}^2}\frac{\sigma_2^2}{\sigma_1^2}\frac{n-1}{m-1} = \frac{S_{1m}^{*2}}{S_{2n}^{*2}}\frac{\sigma_2^2}{\sigma_1^2} \sim F(m-1, n-1)$$

4 **推论6.3.4** 在推论6.3.3的假定中增加一个条件$\sigma_1 = \sigma_2$，那么

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{mS_{1m}^2 + nS_{2n}^2}{m+n-2}}\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2)$$

5 **命题6.3.5** 设总体X的分布函数为F_X，分布密度函数为f_X，那么

$$\boldsymbol{f_{X_{(k)}}(x) = \frac{n!}{(n-k)!(k-1)!}[F_X(x)]^{k-1}[1-F_X(x)]^{n-k}f_X(x)}$$

$k = 1, 2, \cdots, n$

练习

1. 设$X_1, X_2, \cdots, X_{10}$是来自总体$X \sim N(\mu, \sigma^2)$的一个样本，且

$$\overline{X} = \frac{1}{9}\sum_{i=1}^{9} X_i, \quad {S_9^*}^2 = \frac{1}{8}\sum_{i=1}^{9}(X_i - \overline{X})^2, \quad T = \frac{3(X_{10}-\overline{X})}{S_9^*\sqrt{10}}$$

   请问T服从何种分布

2. 设总体$X \sim N(0, \sigma^2)$，从X中抽取样本$(X_1, X_2, \cdots, X_{14})$，且

$$\begin{array}{cc} Y_1 = \frac{1}{5}\sum_{i=1}^{5} X_i, & Y_2 = \frac{1}{5}\sum_{i=10}^{14} X_i, \\ Z_1 = \sum_{i=1}^{5}(X_i - Y_1)^2, & Z_2 = \sum_{i=10}^{14}(X_i - Y_2)^2, \\ Z_3 = \sum_{i=6}^{9} X_i^2, & T = \frac{Z_1+Z_2}{2Z_3} \end{array}$$

   请问T服从什么分布