第五章　统计量及其分布

上学期所学概率论是对随机现象进行理论研究，随机变量分布是已知的，进行分析讨论。

本学期所学数理统计是概率论的实际应用，又是统计学的基础，随机变量分布是未知的，往往需要进行统计推断。如很多随机现象都服从正态分布 $N(\mu,\sigma^2)$，但参数 μ 与 σ^2 往往未知。又如某网站某时间段内的点击次数服从泊松分布 $P(\lambda)$，但参数 λ 未知。

§5.1　总体与样本

5.1.1　总体与个体

实际问题中常常需要对研究对象的某一指标进行统计研究。总体：所研究对象的全体。个体：总体中的每一个研究对象。

如果只是考察某一项指标，通常总体就体现为一系列数据。抽样检验时，取得某些数据的可能性大些，另一些的可能性又小些，表现为一个分布，称为总体分布，记为 X，这是一个随机变量。

实际问题中需要对总体分布 X 的情况进行推断，如总体期望 $E(X)$，总体方差 $\mathrm{Var}(X)$。

注意：统计学经典流派认为总体期望 $E(X)$，总体方差 $\mathrm{Var}(X)$ 都是确定的数，但通常是未知的。

5.1.2　样本

实际问题中往往难以对研究对象作全面普查，通常是抽取其中一部分进行检验，即抽样检验。每一次随机抽样的结果都是一个随机变量 X_i；所得的具体值 x_i 称为观测值。

样本：n 次随机抽样的结果是一个 n 维随机变量 $(X_1, X_2, \cdots, X_n)$，称为一个容量为 n 的样本，n 称为样本容量，样本中的每一个个体 X_i 称为一个样品；而具体值 $(x_1, x_2, \cdots, x_n)$，称为样本观测值。

注意：样本中的 X_i 是随机变量，而样本观测值中的 x_i 是数。

若样本 $(X_1, X_2, \cdots, X_n)$ 满足

（1）随机性，每一个个体在每次抽样时被抽到的机会均等，即样品 X_i 与总体 X 同分布。

（2）独立性，每次抽样互不影响，即 $X_1, X_2, \cdots, X_n$ 相互独立。

则称 $(X_1, X_2, \cdots, X_n)$ 为简单随机样本，即简单随机样本要求所有样品独立同分布。

实际工作中随机抽样通常是进行不放回抽样，在严格意义上，不满足独立性。但当总体中个体的总量 N 很大，远远大于样本容量 n 时，即 $N >> n$，此时不放回抽样可近似看作有放回抽样，独立性可以近似满足。

设总体 X 的分布函数为 $F(x)$，样本 $X_1, X_2, \cdots, X_n$，因 X_i 与总体 X 同分布，故样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数为

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

设总体 X 的质量函数或密度函数为 $p(x)$，故样本 $(X_1, X_2, \cdots, X_n)$ 的联合质量函数或联合密度函数

为

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

例 设总体 X 服从指数分布 $Exp(\lambda)$，求样本 $(X_1, X_2, \cdots, X_n)$ 的联合密度函数。

解：因总体 X 密度函数为

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0; \\ 0, & x \le 0. \end{cases}$$

即 $p(x) = \lambda \mathrm{e}^{-\lambda x} I_{x>0}$，故样本 $(X_1, X_2, \cdots, X_n)$ 的联合密度函数为

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda x_i} I_{x_i>0} = \lambda^n \mathrm{e}^{-\lambda \sum_{i=1}^{n} x_i} I_{x_1, x_2, \cdots, x_n>0}。$$

注： $I_{x\in A}$ 为示性函数，表示当 $x \in A$ 时，$I_{x\in A} = 1$；当 $x \notin A$ 时，$I_{x\in A} = 0$。

例 设总体 X 服从泊松分布 $P(\lambda)$，求样本 $(X_1, X_2, \cdots, X_n)$ 的联合质量函数。

解：因总体 X 的质量函数为

$$p(x) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \quad x = 0, 1, 2, \cdots,$$

故样本 $(X_1, X_2, \cdots, X_n)$ 的联合概率函数为

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1! x_2! \cdots x_n!} e^{-n\lambda}, \quad x_1, x_2, \cdots, x_n = 0, 1, 2, \cdots。$$

例 设总体 X 只取值 0 或 1，且

$$P\{X = 1\} = p, P\{X = 0\} = 1 - p, \quad (0 < p < 1),$$

求样本 $(X_1, X_2, \cdots, X_n)$ 的联合质量函数。

解：因总体 X 质量函数为

$$p(x) = p^x (1-p)^{1-x}, \quad x = 0, 1,$$

故样本 $(X_1, X_2, \cdots, X_n)$ 的联合质量函数为

$$p(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}, \quad x_1, x_2, \cdots, x_n = 0, 1。$$

进一步思考：多点分布的联合质量函数。

总体 X 全部可能取值 $a_1, a_2, \cdots, a_m$，且质量函数为 $p(a_i) = p_i, \quad i = 1, 2, \cdots, m$，$p_1 + p_2 + \cdots + p_m = 1$ 且 $p_i > 0$。求样本 $(X_1, X_2, \cdots, X_n)$ 的联合质量函数。

§5.2　样本数据的整理与显示

5.2.1　　经验分布函数

对于总体 X，为了反映总体分布函数 $F(x)=P\{X\le x\}$，由大数定律知频率稳定于概率，用 $\{X_i\le x\}$ 发生的频率反映 $\{X\le x\}$ 发生的概率。对于一个样本 $X_1, X_2,\cdots, X_n$，将其按由小到大顺序进行排列：$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$，称 $X_{(1)}, X_{(2)},\cdots, X_{(n)}$ 为顺序统计量。定义

$$F_n(x)=\begin{cases}0, & x<X_{(1)};\\ \cdots & \cdots\\ \dfrac{k}{n}, & X_{(k)}\le x<X_{(k+1)};\\ \cdots & \cdots\\ 1, & x\ge X_{(n)}\end{cases}$$

称为样本的经验分布函数。用于近似反映总体分布函数 $F(x)$。

可以证明，当 $n\to\infty$ 时，$F_n(x)$ 几乎处处一致收敛于总体分布函数 $F(x)$。

$P\{\sup\limits_{-\infty<x<+\infty}|F_n(x)-F(x)|\to 0\}=1$，sup 表示上确界。

5.2.2　　频数频率分布表

当样本容量 n 较大时，需要对样本数据进行分组整理。

分组：组数，组距，组限与组中值，频数。

如对期末考试成绩进行数据分组整理，分 5 组：59 分以下、60~69、70~79、80~89、90 分以上，即组数为 5，组距为 10，组限[0, 60), [60, 70), [70, 80), [80, 90), [90, 100]，组中值 ?, 65, 75, 85, 95。

5.2.3　　样本数据的图形显示

利用图形显示数据更直观。常用的有直方图、饼图、茎叶图。

§5.3 统计量及其分布

5.3.1 统计量与抽样分布

利用样本反映所研究的问题，需要进行统计分析。关于样本$(X_1, X_2, \cdots, X_n)$的函数，如果不含未知参数，则称为统计量。统计量的分布称为抽样分布。如果得到样本观测值，则可计算出统计量的观测值，称为可观测。

如总体为X，$(X_1, X_2, \cdots, X_n)$为简单随机样本。则$X_1^2 - X_2^2$，$\frac{1}{n}\sum_{i=1}^{n} X_i$是统计量，（样本容量$n$必是已知的）；$\sum_{i=1}^{n}[X_i - E(X)]^2$当$E(X)$已知时是统计量，未知时不是统计量；$\frac{X_1 - E(X)}{\sqrt{\mathrm{Var}(X)}}$当$E(X)$和$\mathrm{Var}(X)$已知时是统计量，未知时不是统计量。

常用的统计量可分两大类：

（1）平均意义统计量，如样本均值、样本方差等；

（2）排序意义统计量，如顺序统计量等。

5.3.2 样本均值及其抽样分布

为了反映总体期望$E(X)$，由大数定律知平均值稳定于数学期望。定义

$$\overline{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

称为样本均值，因此用样本均值$\overline{X}$反映总体期望$E(X)$。

定理 样本均值的性质。设$(X_1, X_2, \cdots, X_n)$为样本，则

（1）样本偏差的总和等于 0，即$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$。

（2）样本偏差的总平方和最小，即$\sum_{i=1}^{n}(X_i - \overline{X})^2$是$\sum_{i=1}^{n}(X_i - c)^2$之中最小的。

（3）$E(\overline{X}) = E(X)$，$\mathrm{Var}(\overline{X}) = \frac{1}{n}\mathrm{Var}(X)$，当$n$较大时，$\overline{X} \sim N(E(X), \frac{1}{n}\mathrm{Var}(X))$。

（4）线性性质：若$Y_i = aX_i + b$，则$\overline{Y} = a\overline{X} + b$。

证明：（1）分开求和得

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum_{i=1}^{n} X_i - n\overline{X} = \sum_{i=1}^{n} X_i - n \cdot \frac{1}{n}\sum_{i=1}^{n} X_i = 0 。$$

（2）设$g(c) = \sum_{i=1}^{n}(X_i - c)^2$，令

$$g'(c) = \sum_{i=1}^{n} 2(X_i - c)\cdot(-1) = \sum_{i=1}^{n} 2(c - X_i) = 2nc - 2\sum_{i=1}^{n} X_i = 0 ,$$

得 $c=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$，且 $g''(c)=2n>0$，故当 $c=\overline{X}$ 时，$g(c)=\sum_{i=1}^{n}(X_i-c)^2$ 为最小值 $\sum_{i=1}^{n}(X_i-\overline{X})^2$。

（3）因

$$E(\overline{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}E(X)=\frac{1}{n}\cdot nE(X)=E(X)，$$

$$\mathrm{Var}(\overline{X})=\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X)=\frac{1}{n^2}\cdot n\,\mathrm{Var}(X)=\frac{1}{n}\mathrm{Var}(X)，$$

由中心极限定理知，当 n 较大时，$\overline{X}\sim N(E(\overline{X}),\mathrm{Var}(\overline{X}))=N(E(X),\frac{1}{n}\mathrm{Var}(X))$。

（4）分开求和得

$$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{1}{n}\sum_{i=1}^{n}(aX_i+b)=\frac{1}{n}(a\sum_{i=1}^{n}X_i+nb)=a\cdot\frac{1}{n}\sum_{i=1}^{n}X_i+b=a\overline{X}+b。$$

5.3.3 样本方差与样本标准差

为了反映总体方差 $\mathrm{Var}(X)=E[X-E(X)]^2$，利用平均值替换数学期望。定义

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

称为未修正样本方差，又称 $S^*=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ 为未修正样本标准差。又定义

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

称为样本方差，一般用样本方差 S^2 反映总体方差 $\mathrm{Var}(X)$。又称 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$ 为样本标准差。

定理 样本方差的性质。设 $(X_1,X_2,\cdots,X_n)$ 为样本，则

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

（2）$E\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=(n-1)\mathrm{Var}(X)$，从而 $E(S^2)=\mathrm{Var}(X)$，但 $E(S^{*2})\neq\mathrm{Var}(X)$。

（3）平方性质：若 $Y_i=aX_i+b$，则 $S_Y^2=a^2S_X^2$。

证明：（1）展开计算得

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

（2）因

$$E\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=E\left[\sum_{i=1}^{n}X_i^2-n\overline{X}^2\right]=\sum_{i=1}^{n}E(X_i^2)-nE(\overline{X}^2)，$$

$$E(X_i^2)=\mathrm{Var}(X_i)+[E(X_i)]^2=\mathrm{Var}(X)+[E(X)]^2,$$

$$E(\bar{X}^2)=\mathrm{Var}(\bar{X})+[E(\bar{X})]^2=\frac{1}{n}\mathrm{Var}(X)+[E(X)]^2,$$

则

$$E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=n\{\mathrm{Var}(X)+[E(X)]^2\}-n\left\{\frac{1}{n}\mathrm{Var}(X)+[E(X)]^2\right\}=(n-1)\mathrm{Var}(X),$$

故

$$E(S^2)=\mathrm{Var}(X),$$

但

$$E(S^{*2})=\frac{n-1}{n}\mathrm{Var}(X)\neq\mathrm{Var}(X)。$$

（3）展开计算得

$$S_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\frac{1}{n-1}\sum_{i=1}^{n}(aX_i+b-a\bar{X}-b)^2=a^2\cdot\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=a^2S_X^2。$$

如一个容量为 6 的样本，观测值为 5.1, 6.7, 7.1, 5.6, 6.5, 6.2，则样本均值

$$\bar{x}=\frac{1}{6}(5.1+6.7+7.1+5.6+6.5+6.2)=6.2。$$

样本方差

$$s^2=\frac{1}{5}(5.1^2+6.7^2+7.1^2+5.6^2+6.5^2+6.2^2-6\times6.2^2)=0.544,$$

样本标准差

$$s=\sqrt{0.544}=0.7376。$$

注意：这里 $\bar{x}, s$ 都用小写，表示观测值。

5.3.4　样本矩及其函数

对应于总体原点矩 $\mu_k=E(X^k)$ 和中心矩 $\nu_k=E[X-E(X)]^k$，定义

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k，\quad B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k,$$

分别称为样本 k 阶原点矩和样本的 k 阶中心矩。特别是 $k=1$ 时，样本一阶原点矩 A_1 就是样本均值 $\bar{X}$，样本一阶中心矩 B_1 恒为 0；当 $k=2$ 时，样本二阶中心矩 B_2 是未修正样本方差 S^{*2}。

对应于总体偏度 $\beta_S=\frac{\nu_3}{\nu_2^{3/2}}$ 和峰度 $\beta_K=\frac{\nu_4}{\nu_2^2}-3$，定义

$$\gamma_S=\frac{B_3}{B_2^{3/2}}，\quad \gamma_K=\frac{B_4}{B_2^2}-3,$$

分别称为样本偏度与样本峰度。

对应于总体相关系数

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}=\frac{E\{[X-E(X)][Y-E(Y)]\}}{\sqrt{E[X-E(X)]^2}\sqrt{E[Y-E(Y)]^2}},$$

定义

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\cdot\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}},$$

称为样本相关系数。

定理 样本相关系数的性质，设$(X_1,X_2,\cdots,X_n)$与$(Y_1,Y_2,\cdots,Y_n)$为样本，则

（1）$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}$。

（2）$|r|\le 1$，当$|r|=1$时，X_i与Y_i具有完全线性关系，即存在与i无关的数a,b，使得$Y_i=aX_i+b$。

证明：（1）展开计算得

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{n}(X_iY_i-\bar{X}Y_i-X_i\bar{Y}+\bar{X}\bar{Y})=\sum_{i=1}^{n}X_iY_i-\bar{X}\sum_{i=1}^{n}Y_i-\bar{Y}\sum_{i=1}^{n}X_i+n\bar{X}\bar{Y}$$

$$=\sum_{i=1}^{n}X_iY_i-\bar{X}\cdot n\bar{Y}-\bar{Y}\cdot n\bar{X}+n\bar{X}\bar{Y}=\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}。$$

（2）因

$$\sum_{i=1}^{n}[(Y_i-\bar{Y})-\lambda(X_i-\bar{X})]^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2-2\lambda\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})+\lambda^2\sum_{i=1}^{n}(X_i-\bar{X})^2\ge 0,$$

则判别式

$$\Delta=[-2\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})]^2-4\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2\le 0,$$

故

$$[\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})]^2\le\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2,$$

$$|r|=\frac{\left|\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\right|}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\cdot\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}\le 1。$$

当$|r|=1$时，判别式$\Delta=0$，方程$\sum_{i=1}^{n}[(Y_i-\bar{Y})-\lambda(X_i-\bar{X})]^2=0$有唯一实根。设$\lambda=a$是该方程的根，

即

$$\sum_{i=1}^{n}[(Y_i-\bar{Y})-a(X_i-\bar{X})]^2=\sum_{i=1}^{n}(Y_i-aX_i-\bar{Y}+a\bar{X})^2=0,$$

故

$$Y_i - aX_i - \bar{Y} + a\bar{X} = 0, \quad i = 1, 2, \cdots, n ,$$

取 $b = \bar{Y} - a\bar{X}$ ，即

$$Y_i = aX_i + b, \quad i = 1, 2, \cdots, n 。$$

5.3.5　顺序统计量及其分布

对于一个样本 $X_1, X_2, \cdots, X_n$，将其按由小到大顺序进行排列：

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} ,$$

则称 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为顺序统计量。特别是 $X_{(1)} = \min\limits_{1 \le i \le n}\{X_i\}$ 称为最小顺序统计量， $X_{(n)} = \max\limits_{1 \le i \le n}\{X_i\}$ 称为最大顺序统计量。

需要注意的是，顺序统计量的结果与样本有关，是随机变量，可进一步考虑其分布。

定理　设总体 X 分布函数为 $F(x)$，密度函数为 $p(x)$，样本 $X_1, X_2, \cdots, X_n$，则第 k 个顺序统计量 $X_{(k)}$ 的密度函数为

$$p_k(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}\, p(x) 。$$

证明：第 k 个顺序统计量 $X_{(k)}$ 的分布函数与密度函数分别为

$$F_k(x) = P\{X_{(k)} \le x\} = P\{\text{不超过 } x \text{ 的样品个数至少有 } k \text{ 个}\}$$

$$= \sum_{m=k}^{n} P\{\text{不超过 } x \text{ 的样品个数恰有 } m \text{ 个}\} = \sum_{m=k}^{n} C_n^m [F(x)]^m [1-F(x)]^{n-m}$$

$$= C_n^k [F(x)]^k [1-F(x)]^{n-k} + C_n^{k+1} [F(x)]^{k+1} [1-F(x)]^{n-k-1} + \cdots + [F(x)]^n ,$$

$$p_k(x) = F_k'(x) = C_n^k \cdot k[F(x)]^{k-1} p(x) \cdot [1-F(x)]^{n-k} + C_n^k [F(x)]^k \cdot (n-k)[1-F(x)]^{n-k-1}[-p(x)]$$

$$+ C_n^{k+1} \cdot (k+1)[F(x)]^k p(x) \cdot [1-F(x)]^{n-k-1}$$

$$+ C_n^{k+1} [F(x)]^k \cdot (n-k-1)[1-F(x)]^{n-k-2}[-p(x)] + \cdots + n[F(x)]^{n-1}$$

$$= kC_n^k [F(x)]^{k-1} [1-F(x)]^{n-k} p(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}\, p(x) 。$$

特别是最小顺序统计量 $X_{(1)}$ 的分布函数与密度函数分别为

$$F_1(x) = P\{X_{(1)} \le x\} = 1 - P\{X_{(1)} > x\} = 1 - P\{X_1 > x, X_2 > x, \cdots, X_n > x\}$$

$$= 1 - P\{X_1 > x\}P\{X_2 > x\} \cdots P\{X_n > x\} = 1 - [1-F(x)]^n ,$$

$$p_1(x) = F_1'(x) = n[1-F(x)]^{n-1} p(x) ;$$

最大顺序统计量 $X_{(n)}$ 的分布函数与密度函数分别为

$$F_n(x)=P\{X_{(n)}\le x\}=P\{X_1\le x, X_2\le x,\cdots, X_n\le x\}=P\{X_1\le x\}\cdots P\{X_n\le x\}=[F(x)]^n,$$

$$p_n(x)=F_n'(x)=n[F(x)]^{n-1}p(x)。$$

进一步考虑多个顺序统计量的联合分布。两个顺序统计量 $(X_{(i)}, X_{(j)})\quad (i<j)$ 的联合分布为

$$p_{ij}(y,z)=\frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(y)]^{i-1}[F(z)-F(y)]^{j-i-1}[1-F(z)]^{n-j}p(y)p(z)I_{y\le z},$$

可以理解为“ $X_{(i)}=y, X_{(j)}=z$ ”表示有 $i-1$ 个样品观测值小于 y 而对应于 $[F(y)]^{i-1}$，有 $j-i-1$ 个样品观测值介于 y 和 z 之间而对应于 $[F(z)-F(y)]^{j-i-1}$，有 $n-j$ 个样品观测值大于 z 而对应于 $[1-F(z)]^{n-j}$，并且 $X_{(i)}=y, X_{(j)}=z$ 分别对应于 $p(y), p(z)$。

更一般地 $(X_{(i_1)}, X_{(i_2)},\cdots, X_{(i_n)})\quad (i_1<i_2<\cdots<i_n)$ 的联合分布为

$$p_{i_1i_2\cdots i_k}(x_{(i_1)}, x_{(i_2)},\cdots, x_{(i_k)})=\frac{n!}{(i_1-1)!(i_2-i_1-1)!\cdots(n-i_k)!}[F(x_{(i_1)})]^{i_1-1}[F(x_{(i_2)})-F(x_{(i_1)})]^{i_2-i_1-1}$$

$$\cdots[1-F(x_{(i_k)})]^{n-i_k}p(x_{(i_1)})p(x_{(i_2)})\cdots p(x_{(i_k)})I_{x_{(i_1)}<x_{(i_2)}<\cdots<x_{(i_k)}}。$$

特别是 $(X_{(1)}, X_{(2)},\cdots, X_{(n)})$ 的联合分布为

$$p_{12\cdots n}(x_{(1)}, x_{(2)},\cdots, x_{(n)})=n!\,p(x_{(1)})p(x_{(2)})\cdots p(x_{(n)})I_{x_{(1)}<x_{(2)}<\cdots<x_{(n)}}。$$

此外还有样本极差，最大与最小顺序统计量之差 $R_n=X_{(n)}-X_{(1)}$ 称为样本极差。

5.3.6　样本中位数与样本 p 分位数

在样本 $X_1, X_2,\cdots, X_n$ 中按大小位于中间位置（即 $\frac{n+1}{2}$ 位置）的数值，定义

$$m_{0.5}=\begin{cases}X_{\left(\frac{n+1}{2}\right)}, & n\text{ 为奇数;}\\ \dfrac{X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}}{2}, & n\text{ 为偶数.}\end{cases}$$

称为样本中位数。

在样本 $X_1, X_2,\cdots, X_n$ 中按大小位于比例为 p 的位置（即 $(n+1)p$ 位置）的数值，定义

$$m_p=\begin{cases}X_{([np]+1)}, & \text{若 } np\text{ 不是整数;}\\ \dfrac{X_{(np)}+X_{(np+1)}}{2}, & \text{若 } np\text{ 是整数.}\end{cases}$$

或

$$m_p = X_{([np])} + \{(n+1)p - [np]\} \cdot \{X_{([np]+1)} - X_{([np])}\},$$

称为样本 p 分位数。

样本中位数 $m_{0.5}$ 和样本 p 分位数 m_p 分别反映总体中位数 $x_{0.5}$ 和 p 分位数 x_p，当总体分布的密度函数为 $p(x)$ 时，$m_{0.5}$ 及 m_p 的渐近分布分别为 $m_{0.5} \sim N\left(x_{0.5}, \dfrac{1}{4n[p(x_{0.5})]^2}\right)$，$m_p \sim N\left(x_p, \dfrac{p(1-p)}{n[p(x_p)]^2}\right)$。

样本 p 分位数是按频率划分的，随机变量的 p 分位数是按概率划分的。

定义 设 X 为随机变量，$0 < p < 1$，若 x_p 满足

$$P\{X < x_p\} \le p \le P\{X \le x_p\},$$

则称 x_p 为 X 的（下侧） p 分位数。

注：若 X 为连续型随机变量，x_p 为 X 的 p 分位数，则 $F(x_p) = P\{X \le x_p\} = p$。

如 X 服从两点分布，$P\{X=0\} = 0.3$，$P\{X=1\} = 0.7$，对于区间 $(0,1)$ 内的任意实数 a，都有

$$P\{X < a\} = 0.3 = P\{X \le a\},$$

故区间 $(0,1)$ 内的任意实数 a 都是 X 的 0.3 分位数 $x_{0.3}$。而对于满足 $0.3 \le p \le 1$ 的概率 p，都有

$$P\{X < 1\} = 0.3 \le p \le P\{X \le 1\} = 1,$$

故对于满足 $0.3 \le p \le 1$ 的概率 p，X 的 p 分位数 x_p 都是 1。

标准正态分布 $N(0,1)$ 的 p 分位数记为 u_p，有 $\Phi(u_p) = p$，常见标准正态分布 p 分位数有

$$u_{0.9} = 1.28,\ u_{0.95} = 1.645,\ u_{0.975} = 1.96,\ u_{0.99} = 2.33,$$

可通过附表 2（标准正态分布函数表）查到。

定义 设 X 为随机变量，$0 < p < 1$，若 x_p^* 满足

$$P\{X > x_p^*\} \le p \le P\{X \ge x_p^*\},$$

则称 x_p^* 为 X 的（上侧） p 分位数。

显然上侧 p 分位数与下侧 p 分位数的关系为 $x_p^* = x_{1-p}$。

5.3.7 五数概括与箱线图

对于样本观测值，根据最小顺序统计量 $X_{(1)}$，样本 0.25 分位数 $m_{0.25}$，样本中位数 $m_{0.5}$，样本 0.75 分位数 $m_{0.75}$，最大顺序统计量 $X_{(n)}$ 大致描述数据的轮廓。

§5.4　三大抽样分布

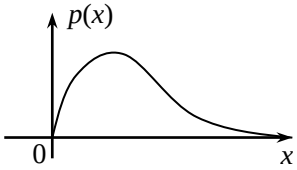
当样本容量很大时（通常 $n \geq 30$ ），则称为大样本问题。由中心极限定理知，大样本问题可用正态分布近似。在一般的统计学中，都是研究大样本问题。

当样本容量 n 较小时，则称为小样本问题。小样本问题要涉及一些特殊的抽样分布。在数理统计中既要研究大样本问题又要研究小样本问题。

5.4.1　χ^2 分布（chi square distribution）

定义　设 $X_1, X_2, \cdots, X_n$ 相互独立且都服从标准正态分布 $N(0,1)$，则 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 的分布称为自由度为 n 的 χ^2 分布，记为 $\chi^2 \sim \chi^2(n)$。其中 n 称为自由度，指所含独立随机变量平方项个数。

χ^2 分布由独立标准正态变量平方和构成。



定理　χ^2 分布的性质：

（1）设 $X \sim \chi^2(n_1)$，$Y \sim \chi^2(n_2)$，且 X 与 Y 相互独立，则 $X+Y \sim \chi^2(n_1+n_2)$；

（2）设 $\chi^2 \sim \chi^2(n)$，则 $E(\chi^2)=n$，$\mathrm{Var}(\chi^2)=2n$。

证明：（1）因 $X \sim \chi^2(n_1)$，存在 $X_1, X_2, \cdots, X_{n_1}$ 相互独立且都服从 $N(0,1)$，使得

$$X = X_1^2 + X_2^2 + \cdots + X_{n_1}^2,$$

又 $Y \sim \chi^2(n_2)$，存在 $Y_1, Y_2, \cdots, Y_{n_2}$ 相互独立且都服从 $N(0,1)$，使得

$$Y = Y_1^2 + Y_2^2 + \cdots + Y_{n_2}^2,$$

故 $X_1, X_2, \cdots, X_{n_1}, Y_1, Y_2, \cdots, Y_{n_2}$ 相互独立且都服从 $N(0,1)$，且

$$X+Y = X_1^2 + X_2^2 + \cdots + X_{n_1}^2 + Y_1^2 + Y_2^2 + \cdots + Y_{n_2}^2 \sim \chi^2(n_1+n_2)。$$

（2）因 $\chi^2 \sim \chi^2(n)$，存在 $X_1, X_2, \cdots, X_n$ 相互独立且都服从 $N(0,1)$，使得

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

则

$$E(\chi^2) = E(X_1^2) + E(X_2^2) + \cdots + E(X_n^2) = nE(X_1^2),$$

$$\mathrm{Var}(\chi^2) = \mathrm{Var}(X_1^2) + \mathrm{Var}(X_2^2) + \cdots + \mathrm{Var}(X_n^2) = n\,\mathrm{Var}(X_1^2),$$

因 $X_1 \sim N(0,1)$，有 $E(X_1)=0$，$\mathrm{Var}(X_1)=1$，则

$$E(X_1^2) = \mathrm{Var}(X_1) + [E(X_1)]^2 = 1,$$

$$E(X_1^4)=\int_{-\infty}^{+\infty}x^4\cdot\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}dx=\int_{-\infty}^{+\infty}x^3\cdot\frac{1}{\sqrt{2\pi}}d(-\mathrm{e}^{-\frac{x^2}{2}})=-\frac{x^3}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\cdot 3x^2dx$$

$$=0+3\int_{-\infty}^{+\infty}x^2\cdot\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}dx=3E(X_1^2)=3,$$

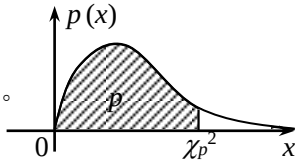
$$\mathrm{Var}(X_1^2)=E(X_1^4)-[E(X_1^2)]^2=3-1=2,$$

故 $E(\chi^2)=nE(X_1^2)=n$， $\mathrm{Var}(\chi^2)=n\,\mathrm{Var}(X_1^2)=2n$。

可见，当 n 很大时， χ^2 分布 $\chi^2(n)$ 可用正态分布 $N(n,2n)$ 近似。

χ^2 分布 $\chi^2(n)$ 的 p 分位数记为 $\chi_p^2(n)$。若 $\chi^2\sim\chi^2(n)$，则 $P\{\chi^2\le\chi_p^2(n)\}=p$。分位数 $\chi_p^2(n)$ 可由附表 3（χ^2 分布分位数表）查到，也可用 MATLAB 中命令 chi2inv(p,n)或 Excel 中命令 chiinv(1-p,n)计算。$\chi^2(n)$ 的分布函数值与密度函数值可分别用 MATLAB 中命令 chi2cdf(x,n)与 chi2pdf(x,n)计算。

如 $\chi^2(10)$ 的 0.95 分位数 $\chi_{0.95}^2(10)=18.3070$，0.05 分位数 $\chi_{0.05}^2(10)=3.9403$；又如 $\chi^2(17)$ 的 0.9 分位数 $\chi_{0.9}^2(17)=24.7690$，0.1 分位数 $\chi_{0.1}^2(17)=10.0852$。



当 n 很大时， $\chi^2(n)$ 的 p 分位数不能直接查表，可利用正态分布近似计算。

例 求 $\chi^2(50)$ 的 0.975 分位数 $\chi_{0.975}^2(50)$ 的近似值。

解：设 $X\sim\chi^2(50)$，因 $n=50$ 较大，有 $X\overset{\cdot}{\sim}N(50,100)$，即 $\dfrac{X-50}{10}\overset{\cdot}{\sim}N(0,1)$，则

$$P\{X\le\chi_{0.975}^2(50)\}=P\left\{\frac{X-50}{10}\le\frac{\chi_{0.975}^2(50)-50}{10}\right\}=0.975,$$

即 $\dfrac{\chi_{0.975}^2(50)-50}{10}\approx u_{0.975}=1.96$，故 $\chi_{0.975}^2(50)\approx 69.6$。

注：用 MATLAB 中命令 chi2inv(0.975,50)或 Excel 中命令 chiinv(0.025,50)计算可得 71.4202。一般当 n 很大时， $\chi^2(n)$ 的 p 分位数 $\chi_p^2(n)\approx n+u_p\sqrt{2n}$。

此外，还可由 p 分位数查表求概率。

例 设 $X\sim\chi^2(28)$，求 $P\{19<X<41\}$。

解：查 χ^2 分布分位数表，可得

$$P\{19<X<41\}\approx P\{\chi_{0.1}^2(28)<X<\chi_{0.95}^2(28)\}=0.95-0.1=0.85。$$

注：用 MATLAB 中命令 chi2cdf(41,28)-chi2cdf(19,28)计算可得 0.8444（即 $0.9463-0.1019$）。

例 设 $X_1,X_2,\cdots,X_{15}$ 相互独立且都服从 $N(0,0.2^2)$，求 $P\left\{\sum_{i=1}^{15}X_i^2<1\right\}$。

解：因 $X_i \sim N(0,0.2^2)$，有 $\dfrac{X_i-0}{0.2}=5X_i \sim N(0,1)$，则 $\sum_{i=1}^{15}(5X_i)^2=25\sum_{i=1}^{15}X_i^2 \sim \chi^2(15)$，故

$$P\left\{\sum_{i=1}^{15}X_i^2<1\right\}=P\left\{25\sum_{i=1}^{15}X_i^2<25\right\}\approx P\left\{25\sum_{i=1}^{15}X_i^2<\chi^2_{0.95}(15)\right\}=0.95\text{。}$$

注：用 MATLAB 中命令 chi2cdf(25,15)计算可得 0.9501。

例 设 X_1,X_2,X_3,X_4,X_5 相互独立且都服从 $N(0,1)$，且 $a(X_1+X_2)^2+b(X_3+2X_4-3X_5)^2$ 服从 χ^2 分布（a,b 均不为 0），求自由度及常数 a,b。

解：式中独立平方项个数为 2，即自由度为 2。因

$$X_1+X_2 \sim N(0,2)\text{，}\quad X_3+2X_4-3X_5 \sim N(0,14)\text{，}$$

标准化，可得

$$\frac{X_1+X_2}{\sqrt{2}} \sim N(0,1)\text{，}\quad \frac{X_3+2X_4-3X_5}{\sqrt{14}} \sim N(0,1)\text{，}$$

故 $\dfrac{(X_1+X_2)^2}{2}+\dfrac{(X_3+2X_4-3X_5)^2}{14} \sim \chi^2(2)$，即 $a=\dfrac{1}{2},b=\dfrac{1}{14}$。

例 设 X_1,X_2,X_3 相互独立且都服从 $N(0,1)$，且 $a[(X_1-X_2)^2+(X_2-X_3)^2+(X_3-X_1)^2]$ 服从 χ^2 分布（$a\neq 0$），求自由度及常数 a。

解：虽然式中有 3 个平方项，**但不独立**，不能判断自由度为 3。

为了使得平方项独立，用正交变换法化二次型为标准型（后面将证明相互独立且方差相同的 n 维正态随机变量经过正交变换后仍为相互独立且方差相同的 n 维正态随机变量）。

二次型

$$(X_1-X_2)^2+(X_2-X_3)^2+(X_3-X_1)^2=2X_1^2+2X_2^2+2X_3^2-2X_1X_2-2X_1X_3-2X_2X_3$$

的系数矩阵

$$A=\begin{pmatrix}2&-1&-1\\-1&2&-1\\-1&-1&2\end{pmatrix}\text{，}$$

特征方程为

$$|\lambda E-A|=\begin{vmatrix}\lambda-2&1&1\\1&\lambda-2&1\\1&1&\lambda-2\end{vmatrix}=\lambda(\lambda-3)^2=0\text{，}$$

特征值为 $\lambda_1=\lambda_2=3$，$\lambda_3=0$，可得原二次型在正交变换下的标准型为

$$(X_1-X_2)^2+(X_2-X_3)^2+(X_3-X_1)^2=3Y_1^2+3Y_2^2\text{。}$$

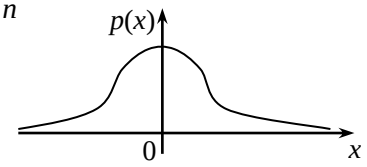
因 X_1,X_2,X_3 相互独立且都服从 $N(0,1)$，可以证明 Y_1,Y_2,Y_3 相互独立且都服从 $N(0,1)$，即

$$\frac{1}{3}[(X_1-X_2)^2+(X_2-X_3)^2+(X_3-X_1)^2]=Y_1^2+Y_2^2 \sim \chi^2(2)\text{，}$$

故自由度为 2，$a=\dfrac{1}{3}$。

5.4.2　　t 分布（t-distribution）

定义　设 $X \sim N(0,1)$，$Y \sim \chi^2(n)$，且 X 与 Y 相互独立，则 $T = \dfrac{X}{\sqrt{Y/n}}$ 的分布称为自由度为 n 的 t 分布，记为 $T \sim t(n)$。



t 分布由独立的标准正态变量与 χ 变量之商构成。

定理　t 分布的性质：

（1）t 分布的密度函数为偶函数，关于 y 轴对称；

（2）当 $n \to +\infty$ 时，t 分布 $t(n)$ 极限分布是标准正态分布 $N(0,1)$（即 $t(n)$ 依分布收敛于 $N(0,1)$）。

此定理可通过 t 分布的密度函数

$$p_t(y) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\dfrac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < y < +\infty$$

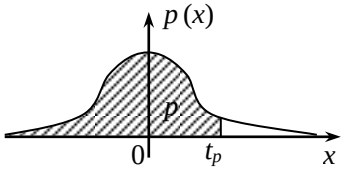
验证。

t 分布 $t(n)$ 的 p 分位数记为 $t_p(n)$。若 $T \sim t(n)$，则 $P\{T \le t_p(n)\} = p$。分位数 $t_p(n)$ 可由附表 4（t 分布分位数表）查到，也可用 MATLAB 中命令 tinv(p,n)或 Excel 中命令 tinv(2(1-p),n)计算。$t(n)$ 的分布函数值与密度函数值可分别用 MATLAB 中命令 tcdf(x,n)与 tpdf(x,n)计算。

由对称性知 $t_{1-p}(n) = -t_p(n)$。

如 $t(15)$ 的 0.95 分位数 $t_{0.95}(15) = 1.7531$，0.01 分位数 $t_{0.01}(15) = -t_{0.99}(15) = -2.6025$；又如 $t(35)$ 的 0.9 分位数 $t_{0.9}(35) = 1.3062$，0.025 分位数 $t_{0.025}(35) = -t_{0.975}(35) = -2.0301$。

当 n 很大时，$t(n)$ 的 p 分位数不能直接查表，可利用标准正态分布的 p 分位数近似。如 $t(200)$ 的 0.975 分位数 $t_{0.975}(200) \approx u_{0.975} = 1.96$。



此外，还可由 p 分位数查表求概率。

例　设 $T \sim t(9)$，求 $P\{0.7 < X < 3.25\}$。

解：查 t 分布分位数表，可得

$$P\{0.7 < X < 3.25\} \approx P\{t_{0.75}(9) < T < t_{0.995}(9)\} = 0.995 - 0.75 = 0.245。$$

例　设 X_1, X_2, X_3 相互独立且都服从 $N(0,1)$，问 $\xi = \dfrac{\sqrt{3}X_1}{\sqrt{X_1^2 + X_2^2 + X_3^2}}$ 是否服从 t 分布。若是，自由度是多少？若不是，构造一个服从 t 分布的随机变量 T，并将 ξ 表示为 T 的函数。

解：虽然 ξ 是正态变量与 χ 变量之商，**但分子分母不独立**，不能判断 ξ 服从 t 分布。

为了使得构造的 t 分布随机变量 T 分子分母独立，在分母的根号中只保留 $X_2^2 + X_3^2$。根据 t 分布的构成可知

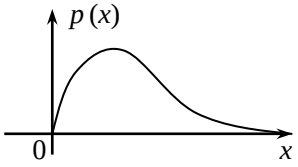
$$T = \frac{\sqrt{2}X_1}{\sqrt{X_2^2 + X_3^2}} \sim t(2)。$$

并且可得

$$\xi=\frac{\sqrt{3}X_1}{\sqrt{X_1^2+X_2^2+X_3^2}}=\frac{\frac{\sqrt{3}X_1}{\sqrt{X_2^2+X_3^2}}}{\sqrt{\frac{X_1^2}{X_2^2+X_3^2}+1}}=\frac{\frac{\sqrt{3}\cdot\sqrt{2}X_1}{\sqrt{X_2^2+X_3^2}}}{\sqrt{\frac{2X_1^2}{X_2^2+X_3^2}+2}}=\frac{\sqrt{3}T}{\sqrt{T^2+2}}。$$

5.4.3　　F 分布（F-distribution）

定义　设 $X\sim\chi^2(n)$，$Y\sim\chi^2(m)$，且 X 与 Y 相互独立，则 $F=\frac{X/n}{Y/m}$ 的分布称为自由度为 (n,m) 的 F 分布，记为 $F\sim F(n,m)$。其中 n 为第一自由度，m 为第二自由度。



F 分布由独立的 χ^2 变量之商构成。

定理　F 分布的性质：

（1）若 $T\sim t(n)$，则 $T^2\sim F(1,n)$；

（2）若 $F\sim F(n,m)$，则 $\frac{1}{F}\sim F(m,n)$。

证明：（1）因 $T\sim t(n)$，存在 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，且 X 与 Y 相互独立，使得

$$T=\frac{X}{\sqrt{Y/n}}，\quad T^2=\frac{X^2}{Y/n}。$$

因 $X\sim N(0,1)$，有 $X^2\sim\chi^2(1)$，故 $T^2=\frac{X^2/1}{Y/n}\sim F(1,n)$。

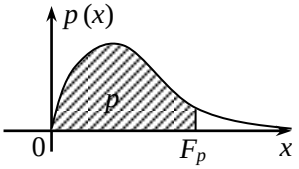
（2）因 $F\sim F(n,m)$，存在 $X\sim\chi^2(n)$，$Y\sim\chi^2(m)$，且 X 与 Y 相互独立，使得 $F=\frac{X/n}{Y/m}$，故

$$\frac{1}{F}=\frac{Y/m}{X/n}\sim F(m,n)。$$

F 分布 $F(n,m)$ 的 p 分位数记为 $f_p(n,m)$。若 $F\sim F(n,m)$，则 $P\{F\le f_p(n,m)\}=p$。分位数 $f_p(n,m)$ 可由附表 5（F 分布分位数表）查到，也可用 MATLAB 中命令 finv(p,n,m)或 Excel 中命令 finv(1-p,n,m)计算。$F(n,m)$ 的分布函数值与密度函数值可分别用 MATLAB 中命令 fcdf(x,n,m)与 fpdf(x,n,m)计算。附表 5 分成四个分表分别给出不同的概率（p 为 0.9, 0.95, 0.975, 0.99）下的分位数表。

如 $F(8,10)$ 的 0.95 分位数 $f_{0.95}(8,10)=3.07$，$F(24,16)$ 的 0.99 分位数 $f_{0.99}(24,16)=3.18$。

此外可以根据下面的定理计算概率 p 为 0.1, 0.05, 0.025, 0.01 的分位数。



定理　$f_{1-p}(m,n)=\frac{1}{f_p(n,m)}$。

证明：设 $F\sim F(n,m)$，有 $\frac{1}{F}\sim F(m,n)$，则

$$P\{F \le f_p(n,m)\} = P\left\{\frac{1}{F} \ge \frac{1}{f_p(n,m)}\right\} = p\ ,\quad P\left\{\frac{1}{F} < \frac{1}{f_p(n,m)}\right\} = 1-p\ 。$$

因

$$P\left\{\frac{1}{F} \le f_{1-p}(m,n)\right\} = 1-p\ ,$$

故

$$f_{1-p}(m,n) = \frac{1}{f_p(n,m)}\ 。$$

如 $F(6,10)$ 的 0.1 分位数

$$f_{0.1}(6,10) = \frac{1}{f_{0.9}(10,6)} = \frac{1}{2.94} \approx 0.34\ ,$$

$F(12,7)$ 的 0.025 分位数

$$f_{0.025}(12,7) = \frac{1}{f_{0.975}(7,12)} = \frac{1}{3.61} \approx 0.277\ 。$$

分布	记号	构成	图形	性质	p 分位数及其关系
χ^2	$\chi^2(n)$	正态变量平方和	单边	可加性, 期望是n, 方差是$2n$	$\chi_p^2(n)$
t	$t(n)$	正态与χ变量之商	双边	对称性, 极限分布是$N(0,1)$	$t_p(n),\quad t_{1-p}(n) = -t_p(n)$
F	$F(n,m)$	χ^2变量之商	单边	倒数性, 与t 分布的平方关系	$f_p(n,m),$ $f_{1-p}(n,m) = \frac{1}{f_p(m,n)}$

5.4.4　抽样分布定理

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

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right),\ \text{即}\ U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\ 。$$

证明：因 $X_1, X_2, \cdots, X_n$ 相互独立且都服从 $N(\mu,\sigma^2)$，它们的线性组合 $\bar{X}$ 也服从正态分布，且

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot n\mu = \mu\ ,\quad \mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}\ ,$$

故

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

再标准化得

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

引理 设 $X_1, X_2, \cdots, X_n$ 相互独立且都服从方差同为 σ^2 的正态分布，记 $\vec{X} = (X_1, X_2, \cdots, X_n)^T$。又设 C 是 n 阶正交阵且 $\vec{Y} = (Y_1, Y_2, \cdots, Y_n)^T = C\vec{X}$，则 $Y_1, Y_2, \cdots, Y_n$ 相互独立且都服从方差同为 σ^2 的正态分布。

证明：因 $Y_1, Y_2, \cdots, Y_n$ 分别是独立正态随机变量 $X_1, X_2, \cdots, X_n$ 的线性组合，则 $(Y_1, Y_2, \cdots, Y_n)$ 服从 n 维正态分布，且 $(Y_1, Y_2, \cdots, Y_n)$ 的协方差矩阵为

$$\mathrm{Cov}(\vec{Y}, \vec{Y}) = \mathrm{Cov}(C\vec{X}, C\vec{X}) = C \cdot \mathrm{Cov}(\vec{X}, \vec{X}) \cdot C^T = C \cdot \sigma^2 E \cdot C^T = \sigma^2 CC^T = \sigma^2 E ,$$

故 $Y_1, Y_2, \cdots, Y_n$ 相互独立且都服从方差同为 σ^2 的正态分布。

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

（1）$\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n)$。

（2）样本均值 $\bar{X}$ 与样本方差 S^2 相互独立，且 $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1)$。

证明：（1）因 $X_i \sim N(\mu, \sigma^2)$，有 $\dfrac{X_i - \mu}{\sigma} \sim N(0,1)$，且 $\dfrac{X_i - \mu}{\sigma}$， $i = 1, 2, \cdots, n$ 相互独立，故

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n) 。$$

（2）可以按如下方法理解：用样本均值 $\bar{X}$ 替换上式中的总体期望 μ，因 $X_i - \bar{X}$ 服从正态分布，有 $\chi^2 = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$ 也服从 χ^2 分布，但因

$$(X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_n - \bar{X}) = X_1 + X_2 + \cdots + X_n - n\bar{X} = 0 ,$$

这是 n 个独立随机变量，但受到一个约束条件，自由度减少一个，变为 $n-1$，有

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

具体证明：因二次型 $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}X_i^2 - n\bar{X}^2$，则通过正交变换 $\vec{Y} = C\vec{X}$，使得 $\sum_{i=1}^{n}X_i^2 = \sum_{i=1}^{n}Y_i^2$，且若使得 $Y_1^2 = n\bar{X}^2$，就有 $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}Y_i^2 - Y_1^2 = \sum_{i=2}^{n}Y_i^2$，再结合引理可证明此结论。

因

$$\sqrt{n}\bar{X} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i = \frac{1}{\sqrt{n}}(1, 1, \cdots, 1) \cdot \vec{X} ,$$

取向量 $\alpha_1 = \dfrac{1}{\sqrt{n}}(1, 1, \cdots, 1)^T$，显然 α_1 是单位向量。将单位向量 α_1 扩充为 n 维向量空间的一组标准正交基 $\alpha_1, \alpha_2, \cdots, \alpha_n$，并令正交阵 $C = (\alpha_1, \alpha_2, \cdots, \alpha_n)$，正交变换 $\vec{X} = C\vec{Y}$，即

$$\vec{Y} = (Y_1, Y_2, \cdots, Y_n)^T = C^T\vec{X} 。$$

因 $X_1, X_2, \cdots, X_n$ 相互独立且都服从方差同为 σ^2 的正态分布，由引理可知 $Y_1, Y_2, \cdots, Y_n$ 相互独立且都服从方差同为 σ^2 的正态分布。

因

$$\sum_{i=1}^{n} X_i^2 = \vec{X}^T \vec{X} = (C\vec{Y})^T C\vec{Y} = \vec{Y}^T C^T C\vec{Y} = \vec{Y}^T \vec{Y} = \sum_{i=1}^{n} Y_i^2 ,$$

且由

$$\vec{Y} = (Y_1, Y_2, \cdots, Y_n)^T = C^T \vec{X} = (\alpha_1^T, \alpha_2^T, \cdots, \alpha_n^T)\vec{X} = (\alpha_1^T \vec{X}, \alpha_2^T \vec{X}, \cdots, \alpha_n^T \vec{X}) ,$$

可知 $Y_i = \alpha_i^T \vec{X}$，特别是

$$Y_1 = \alpha_1^T \vec{X} = \sqrt{n}\bar{X} ,$$

则

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2 ,$$

故样本均值 $\bar{X} = \frac{1}{\sqrt{n}} Y_1$ 与样本方差 $S^2 = \frac{1}{n-1}\sum_{i=2}^{n} Y_i^2$ 相互独立。由 $Y_i = \alpha_i^T \vec{X}$，可得

$$E(Y_i) = \alpha_i^T E(\vec{X}) = \alpha_i^T E\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \alpha_i^T \begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix} = \alpha_i^T \cdot \mu\sqrt{n}\alpha_1 = \mu\sqrt{n} \cdot \alpha_i^T \alpha_1 ,$$

当 $i \geq 2$ 时，$E(Y_i) = 0$，则 $Y_2, \cdots, Y_n$ 相互独立且都服从正态分布 $N(0, \sigma^2)$，即 $\frac{Y_i}{\sigma}$ 服从标准正态分布 $N(0,1)$，故

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=2}^{n} Y_i^2 = \sum_{i=2}^{n}\left(\frac{Y_i}{\sigma}\right)^2 \sim \chi^2(n-1) 。$$

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

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

证明：可看做将统计量 U 分母中的总体标准差 σ 替换为样本标准差 S，抽样分布就由标准正态分布变成 t 分布。因

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) ,\quad \chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) ,$$

且 $\bar{X}$ 与 S^2 相互独立，即 U 与 χ^2 相互独立，则由 t 分布的定义可知

$$T = \frac{U}{\sqrt{\chi^2/(n-1)}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1) 。$$

定理 设总体 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$ 相互独立。$X_1, X_2, \cdots, X_{n_1}$ 为 X 的样本，$Y_1, Y_2, \cdots, Y_{n_2}$ 为 Y 的样本，则

（1）$U = \dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$。

（2）当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 但未知时，$T = \dfrac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_w \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$，其中 S_w^2 是 S_x^2 与 S_y^2 关于自由度的加权平均，即

$$S_w = \sqrt{\frac{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2}{n_1 + n_2 - 2}}$$。

（3）$F = \dfrac{S_x^2 / \sigma_1^2}{S_y^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$。

证明：（1）因 $X \sim N(\mu_1, \sigma_1^2)$，$Y \sim N(\mu_2, \sigma_2^2)$，且相互独立，有 $\bar{X} \sim N\left(\mu_1, \dfrac{\sigma_1^2}{n_1}\right)$，$\bar{Y} \sim N\left(\mu_2, \dfrac{\sigma_2^2}{n_2}\right)$，且相互独立，$\bar{X} - \bar{Y}$ 也服从正态分布，则

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2,$$

$$\mathrm{Var}(\bar{X} - \bar{Y}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

故

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

再标准化，得

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$$。

（2）可看做当 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 但未知时，

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1),$$

再将分母中的总体标准差 σ 替换为样本标准差 S_w，它的平方是 S_x^2 与 S_y^2 关于自由度的加权平均，抽样分布就由标准正态分布变成 t 分布。

因 $X \sim N(\mu_1, \sigma^2)$，$Y \sim N(\mu_2, \sigma^2)$，且相互独立，有

$$\frac{(n_1-1)S_x^2}{\sigma^2}\sim\chi^2(n_1-1)\,,\quad \frac{(n_2-1)S_y^2}{\sigma^2}\sim\chi^2(n_2-1)\,,$$

且相互独立，则

$$\chi^2=\frac{(n_1-1)S_x^2}{\sigma^2}+\frac{(n_2-1)S_y^2}{\sigma^2}=\frac{(n_1-1)S_x^2+(n_2-1)S_y^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)\,。$$

又因当$\sigma_1^2=\sigma_2^2=\sigma^2$时，

$$U=\frac{(\bar X-\bar Y)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}}}=\frac{(\bar X-\bar Y)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim N(0,1)\,,$$

且$\bar X$、$\bar Y$、S_x^2、S_y^2相互独立，即U与χ^2相互独立，则由t分布的定义可知

$$T=\frac{U}{\sqrt{\chi^2/(n_1+n_2-2)}}=\frac{\dfrac{(\bar X-\bar Y)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\dfrac{(n_1-1)S_x^2+(n_2-1)S_y^2}{\sigma^2}\Big/(n_1+n_2-2)}}$$

$$=\frac{(\bar X-\bar Y)-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_x^2+(n_2-1)S_y^2}{n_1+n_2-2}}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)\,。$$

（3）因$X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，且相互独立，有

$$\chi_1^2=\frac{(n_1-1)S_x^2}{\sigma_1^2}\sim\chi^2(n_1-1)\,,\quad \chi_2^2=\frac{(n_2-1)S_y^2}{\sigma_2^2}\sim\chi^2(n_2-1)\,,$$

且相互独立，则由F分布的定义可知

$$F=\frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)}=\frac{\dfrac{(n_1-1)S_x^2}{\sigma_1^2}\Big/(n_1-1)}{\dfrac{(n_2-1)S_y^2}{\sigma_2^2}\Big/(n_2-1)}=\frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2}\sim F(n_1-1,n_2-1)\,。$$

单总体统计量	抽样分布	双总体统计量	抽样分布
$U=\dfrac{\bar X-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$	$U=\dfrac{(\bar X-\bar Y)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$N(0,1)$
$T=\dfrac{\bar X-\mu}{S/\sqrt{n}}$	$t(n-1)$	$T=\dfrac{(\bar X-\bar Y)-(\mu_1-\mu_2)}{\sqrt{\frac{(n_1-1)S_x^2+(n_2-1)S_y^2}{n_1+n_2-2}}\cdot\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$，$(\sigma_1^2=\sigma_2^2)$	$t(n_1+n_2-2)$
$\chi^2=\dfrac{(n-1)S^2}{\sigma^2}$	$\chi^2(n-1)$	$F=\dfrac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2}$	$F(n_1-1,n_2-1)$

例 设总体 $X \sim N(1,4)$，且 $X_1, X_2, \cdots, X_{16}$ 为样本，求 $P\{\bar{X} > 0\}$， $P\{S^2 < 6\}$。

解：因 $\mu = 1$， $\sigma^2 = 4$，样本容量 $n = 16$，有

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = 2(\bar{X} - 1) \sim N(0,1),$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{15S^2}{4} \sim \chi^2(15),$$

故

$$P\{\bar{X} > 0\} = P\{U = 2(\bar{X} - 1) > -2\} = 1 - \Phi(-2) = \Phi(2) = 0.9772,$$

$$P\{S^2 < 6\} = P\left\{\chi^2 = \frac{15S^2}{4} < 22.5\right\} = P\{\chi^2 < \chi^2_{0.9}(15)\} = 0.9。$$

例 设总体 $X \sim N(2,2)$ 与 $Y \sim N(3,5.5)$，且相互独立。$X_1, X_2, \cdots, X_9$ 为 X 的样本，$Y_1, Y_2, \cdots, Y_{26}$ 为 Y 的样本，求 $P\{\bar{X} < \bar{Y}\}$， $P\{S_x^2 < S_y^2\}$。

解：因 $\mu_1 = 2$， $\sigma_1^2 = 2$， $\mu_2 = 3$， $\sigma_2^2 = 5.5$，样本容量 $n_1 = 9$， $n_2 = 26$，有

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X} - \bar{Y} + 1}{0.6586} \sim N(0,1),$$

$$F = \frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2} = \frac{2.75S_x^2}{S_y^2} \sim F(8,25),$$

故

$$P\{\bar{X} < \bar{Y}\} = P\left\{U = \frac{\bar{X} - \bar{Y} + 1}{0.6586} < \frac{1}{0.6586} = 1.52\right\} = \Phi(1.52) = 0.9357,$$

$$P\{S_x^2 < S_y^2\} = P\left\{F = \frac{2.75S_x^2}{S_y^2} < 2.75\right\} = P\{F < f_{0.975}(8,25)\} = 0.975。$$

§5.5 充分统计量

5.5.1 充分性的概念

为了估计总体 X 的某未知参数 θ，抽取一个样本 $X_1, X_2, \cdots, X_n$，样本的联合分布函数是一个 n 元函数 $F(x_1, x_2, \cdots, x_n; \theta)$，直接处理往往非常困难。通常建立统计量 $T = T(X_1, X_2, \cdots, X_n)$，其分布函数是一元函数 $F_T(t; \theta)$，处理相对容易。并且希望一元函数 $F_T(t; \theta)$ 中包含 n 元函数 $F(x_1, x_2, \cdots, x_n; \theta)$ 中关于参数 θ 的全部信息。

当取定 $T = t$ 时，样本联合质量函数或联合密度函数为

$$p(x_1, x_2, \cdots, x_n; \theta) = p_T(t; \theta)\, p(x_1, x_2, \cdots, x_n; \theta \mid T = t),$$

要求统计量 T 的一维分布包含样本 $X_1, X_2, \cdots, X_n$ 的 n 维分布中关于参数 θ 的全部信息，即要求条件质量函数或条件密度函数 $p(x_1, x_2, \cdots, x_n; \theta \mid T = t)$ 与参数 θ 无关。

定义 设总体 X 的分布函数为 $F(x; \theta)$，$X_1, X_2, \cdots, X_n$ 为样本，$T = T(X_1, X_2, \cdots, X_n)$ 为一个统计量。如果在给定 T 的取值条件下，样本 $X_1, X_2, \cdots, X_n$ 的条件分布与未知参数 θ 无关，则称 T 为参数 θ 的一个充分统计量。

例 总体 X 服从泊松分布 $P(\lambda)$，$X_1, X_2, \cdots, X_n$ 为样本，验证 $T = \sum_{i=1}^{n} X_i$ 为参数 λ 的充分统计量。

解：因总体 X 的质量函数为

$$p(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \cdots,$$

则样本联合质量函数为

$$p(x_1, x_2, \cdots, x_n; \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{x_1! x_2! \cdots x_n!} e^{-n\lambda}, \quad x_1, x_2, \cdots, x_n = 0, 1, 2, \cdots。$$

因 $X_1, X_2, \cdots, X_n$ 相互独立且都服从泊松分布 $P(\lambda)$，根据泊松分布可加性可知 $T = \sum_{i=1}^{n} X_i$ 服从泊松分布 $P(n\lambda)$，即

$$p_T(t; \lambda) = \frac{(n\lambda)^t}{t!} e^{-n\lambda}, \quad t = 0, 1, 2, \cdots,$$

则当 $T = t$ 时，即 $t = \sum_{i=1}^{n} x_i$，有

$$p(x_1, x_2, \cdots, x_n; \lambda \mid T = t) = \frac{p(x_1, x_2, \cdots, x_n; \lambda)}{p_T(t; \lambda)} = \frac{\dfrac{\lambda^t}{x_1! x_2! \cdots x_n!} e^{-n\lambda}}{\dfrac{n^t \lambda^t}{t!} e^{-n\lambda}} = \frac{t!}{n^t x_1! x_2! \cdots x_n!},$$

这与参数 λ 无关，故 $T = \sum_{i=1}^{n} X_i$ 为参数 λ 的充分统计量。

5.5.2　因子分解定理

根据定义寻找或判断充分统计量通常比较困难，可以用更加方便的因子分解定理处理。

定理　（因子分解定理）设总体 X 的质量函数或密度函数为 $p(x;\theta)$，$X_1, X_2, \cdots, X_n$ 为样本，则统计量 $T=T(X_1, X_2, \cdots, X_n)$ 为参数 θ 一个充分统计量的充分必要条件是样本联合质量函数或联合密度函数 $p(x_1, x_2, \cdots, x_n;\theta)$ 可以分解为两个函数的乘积 $g(t;\theta)h(x_1, x_2, \cdots, x_n)$，其中 $t=T(x_1, x_2, \cdots, x_n)$ 为统计量 T 的观测值，函数 $g(t;\theta)$ 只与观测值 t 有关，与其它形式的观测值 x_i 无关，而函数 $h(x_1, x_2, \cdots, x_n)$ 与未知参数 θ 无关。

证明：必要性，设 $T=T(X_1, X_2, \cdots, X_n)$ 为参数 θ 一个充分统计量。

当取定 $T=t$ 时，样本联合质量函数或联合密度函数为

$$p(x_1, x_2, \cdots, x_n;\theta)=p_T(t;\theta)\,p(x_1, x_2, \cdots, x_n \mid T=t)\,,$$

且条件分布 $p(x_1, x_2, \cdots, x_n \mid T=t)$ 与参数 θ 无关。取

$$g(t;\theta)=p_T(t;\theta)\,,\quad h(x_1, x_2, \cdots, x_n)=p(x_1, x_2, \cdots, x_n \mid T=t)\,,$$

必要性得证。

充分性，设样本联合质量函数或联合密度函数

$$p(x_1, x_2, \cdots, x_n;\theta)=g(t;\theta)h(x_1, x_2, \cdots, x_n)\,,$$

其中 $h(x_1, x_2, \cdots, x_n)$ 与参数 θ 无关。

统计量 T 的边际分布为固定 $T=t$ 的情况下对样本联合质量函数或联合密度函数积分，形式上为

$$p_T(t;\theta)=\int_{T=t} g(t;\theta)h(x_1, x_2, \cdots, x_n)dS=g(t;\theta)\int_{T=t} h(x_1, x_2, \cdots, x_n)dS\,,$$

其中 $\int_{T=t} \bullet\, dS$ 表示在 $T=t$ 时的超曲面上进行积分，则条件分布

$$p(x_1, x_2, \cdots, x_n \mid T=t)=\frac{p(x_1, x_2, \cdots, x_n;\theta)}{p_T(t;\theta)}=\frac{g(t;\theta)h(x_1, x_2, \cdots, x_n)}{g(t;\theta)\int_{T=t} h(x_1, x_2, \cdots, x_n)dS}$$

$$=\frac{h(x_1, x_2, \cdots, x_n)}{\int_{T=t} h(x_1, x_2, \cdots, x_n)dS}\,,$$

可见条件分布 $p(x_1, x_2, \cdots, x_n \mid T=t)$ 与参数 θ 无关，故 $T=T(X_1, X_2, \cdots, X_n)$ 为参数 θ 一个充分统计量。

根据因子分解定理求具体问题中参数 θ 的一个充分统计量时，首先写出样本联合质量函数或联合密度函数，再将与参数 θ 有关的部分集中在一起，并观察其中所有的观测值 x_i 能否用一个统计量的观测值 $t=T(x_1, x_2, \cdots, x_n)$ 表示出来。若能这样表示，则这部分可记为 $g(t;\theta)$，而剩下的部分与参数 θ 无关，则记为 $h(x_1, x_2, \cdots, x_n)$。从而得到参数 θ 的一个充分统计量 $T=T(X_1, X_2, \cdots, X_n)$。

例 用因子分解定理验证 $T=\sum_{i=1}^{n}X_i$ 为泊松分布 $P(\lambda)$ 中参数 λ 的一个充分统计量。

解：因泊松分布 $P(\lambda)$ 的质量函数为

$$p(x;\lambda)=\frac{\lambda^x}{x!}e^{-\lambda},\quad x=0,1,2,\cdots,$$

则样本联合质量函数为

$$p(x_1,x_2,\cdots,x_n;\lambda)=\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}=\frac{\lambda^{\sum_{i=1}^{n}x_i}}{x_1!x_2!\cdots x_n!}e^{-n\lambda},\quad x_1,x_2,\cdots,x_n=0,1,2,\cdots。$$

对于其中与参数 λ 有关的部分 $\lambda^{\sum_{i=1}^{n}x_i}e^{-n\lambda}$，令 $t=\sum_{i=1}^{n}x_i$，取

$$g(t;\lambda)=\lambda^t e^{-n\lambda},\quad h(x_1,x_2,\cdots,x_n)=\frac{1}{x_1!x_2!\cdots x_n!},$$

根据因子分解定理可知 $T=\sum_{i=1}^{n}X_i$ 为泊松分布 $P(\lambda)$ 中参数 λ 的一个充分统计量。

例 求指数分布 $Exp(\lambda)$ 中参数 λ 的一个充分统计量。

解：因指数分布 $Exp(\lambda)$ 的密度函数为

$$p(x;\lambda)=\lambda\,\mathrm{e}^{-\lambda x}I_{x>0},$$

则样本联合密度函数为

$$p(x_1,x_2,\cdots,x_n;\lambda)=\prod_{i=1}^{n}\lambda\,\mathrm{e}^{-\lambda x_i}I_{x_i>0}=\lambda^n\,\mathrm{e}^{-\lambda\sum_{i=1}^{n}x_i}I_{x_1,x_2,\cdots,x_n>0}。$$

对于其中与参数 λ 有关的部分 $\lambda^n\,\mathrm{e}^{-\lambda\sum_{i=1}^{n}x_i}$，令 $t=\sum_{i=1}^{n}x_i$，取

$$g(t;\lambda)=\lambda^n e^{-\lambda t},\quad h(x_1,x_2,\cdots,x_n)=I_{x_1,x_2,\cdots,x_n>0},$$

根据因子分解定理可知 $T=\sum_{i=1}^{n}X_i$ 为指数分布 $Exp(\lambda)$ 中参数 λ 的一个充分统计量。

例 求均匀分布 $U(0,\theta)$ 中参数 θ 的一个充分统计量。

解：因均匀分布 $U(0,\theta)$ 的密度函数为

$$p(x;\theta)=\frac{1}{\theta}I_{0<x<\theta},$$

则样本联合密度函数为

$$p(x_1,x_2,\cdots,x_n;\lambda)=\prod_{i=1}^{n}\frac{1}{\theta}I_{0<x_i<\theta}=\frac{1}{\theta^n}I_{0<x_1,x_2,\cdots,x_n<\theta}=\frac{1}{\theta^n}I_{0<x_{(1)}\le x_{(n)}<\theta}=\frac{1}{\theta^n}I_{x_{(n)}<\theta}\cdot I_{x_{(1)}>0}。$$

对于其中与参数 θ 有关的部分 $\frac{1}{\theta^n}I_{x_{(n)}<\theta}$，令 $t=x_{(n)}=\max\{x_1,x_2,\cdots,x_n\}$，取

$$g(t;\theta)=\frac{1}{\theta^n}I_{t<\theta},\quad h(x_1,x_2,\cdots,x_n)=I_{x_{(1)}>0},$$

根据因子分解定理可知 $T=X_{(n)}=\max\{X_1,X_2,\cdots,X_n\}$ 为均匀分布 $U(0,\theta)$ 中参数 θ 的一个充分统计量。

充分统计量可推广到多个参数与多个统计量的情形。

定义 设总体 X 的分布函数为 $F(x;\theta_1,\theta_2,\cdots,\theta_r)$，记 $\theta=(\theta_1,\theta_2,\cdots,\theta_r)$，又设 $X_1,X_2,\cdots,X_n$ 为样本，$T_i=T_i(X_1,X_2,\cdots,X_n),\ \ i=1,2,\cdots,s$ 为 s 个统计量，记 $T=(T_1,T_2,\cdots,T_s)$。如果给定 $T=(T_1,T_2,\cdots,T_s)$ 的取值后，样本 $X_1,X_2,\cdots,X_n$ 的条件分布与未知参数 $\theta=(\theta_1,\theta_2,\cdots,\theta_r)$ 无关，则称 $T=(T_1,T_2,\cdots,T_s)$ 是参数 $\theta=(\theta_1,\theta_2,\cdots,\theta_r)$ 的一组充分统计量。

定理 （因子分解定理）设总体 X 的质量函数或密度函数为 $p(x;\theta_1,\theta_2,\cdots,\theta_r)$，记 $\theta=(\theta_1,\theta_2,\cdots,\theta_r)$，又设 $X_1,X_2,\cdots,X_n$ 为样本，$T_i=T_i(X_1,X_2,\cdots,X_n),\ \ i=1,2,\cdots,s$ 为 s 个统计量，记 $T=(T_1,T_2,\cdots,T_s)$，则 $T=(T_1,T_2,\cdots,T_s)$ 为 $\theta=(\theta_1,\theta_2,\cdots,\theta_r)$ 充分统计量的充分必要条件是样本联合质量函数或联合密度函数 $p(x_1,x_2,\cdots,x_n;\theta_1,\theta_2,\cdots,\theta_r)$ 可以分解为两个函数的乘积 $g(t_1,t_2,\cdots,t_s;\theta_1,\theta_2,\cdots,\theta_r)h(x_1,x_2,\cdots,x_n)$，其中 $t_i=T_i(x_1,x_2,\cdots,x_n)$ 为统计量 T_i 的观测值，$i=1,2,\cdots,s$，函数 $g(t_1,t_2,\cdots,t_s;\theta_1,\theta_2,\cdots,\theta_r)$ 只与观测值 $t=(t_1,t_2,\cdots,t_s)$ 有关，与其它形式的观测值 x_i 无关，而函数 $h(x_1,x_2,\cdots,x_n)$ 与未知参数 $\theta=(\theta_1,\theta_2,\cdots,\theta_r)$ 无关。

例 求正态分布 $N(\mu,\sigma^2)$ 中参数 (μ,σ^2) 的一组充分统计量。

解：因正态分布 $N(\mu,\sigma^2)$ 的密度函数为

$$p(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

则样本联合密度函数为

$$p(x_1,x_2,\cdots,x_n;\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$=\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2-2\mu\sum_{i=1}^{n}x_i+n\mu^2\right)}。$$

对于其中与参数 (μ,σ^2) 有关的部分 $\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2-2\mu\sum_{i=1}^{n}x_i+n\mu^2\right)}$，令 $t_1=\sum_{i=1}^{n}x_i$，$t_2=\sum_{i=1}^{n}x_i^2$，取

$$g(t_1,t_2;\mu,\sigma^2)=\frac{1}{(\sqrt{2\pi}\sigma)^n}e^{-\frac{1}{2\sigma^2}(t_2-2\mu t_1+n\mu^2)},\quad h(x_1,x_2,\cdots,x_n)=1,$$

根据因子分解定理可知 $(T_1,T_2)=\left(\sum_{i=1}^{n}X_i,\sum_{i=1}^{n}X_i^2\right)$ 为正态分布 $N(\mu,\sigma^2)$ 中参数 (μ,σ^2) 的一组充分统计量。