第一章 数理统计的基本知识

第一节 引论

第一节引论

1.数理统计学是数学的一个分支学科

2.数理统计学研究怎样有效地收集、整理和分析带有随机性的数据，以对所考察的问题作出推断和预测，直至为采取一定的决策和行动提供依据和建议。

3.数理统计的基本内容

(1)数据收集 { 全面观测 / 抽样技术 / 试验设计

(2)统计推断 { 参数估计 / 假设检验

第一节引论

1.数理统计学是数学的一个分支学科

2.数理统计学研究怎样有效地收集、整理和分析带有随机性的数据，以对所考察的问题作出推断和预测，直至为采取一定的决策和行动提供依据和建议。

3.数理统计的基本内容

(1)数据收集 { 全面观测 / 抽样技术 / 试验设计

(2)统计推断 { 参数估计 / 假设检验

第一节引论

1.数理统计学是数学的一个分支学科

2.数理统计学研究怎样有效地收集、整理和分析带有随机性的数据，以对所考察的问题作出推断和预测，直至为采取一定的决策和行动提供依据和建议。

3.数理统计的基本内容

(1)数据收集 { 全面观测 / 抽样技术 / 试验设计

(2)统计推断 { 参数估计 / 假设检验

第一节引论

例：某钢厂日产某型号钢筋10000根，质量检查员每天抽查50根的强度，于是有

（1）如何从仅有的50根钢筋强度数据估计整批钢筋强度的平均值？又如何估计整批钢筋强度偏离平均值的离散程度？——参数估计

（2）若规定了这种型号的钢筋强度，从抽查的50个强度数据如何判断整批钢筋的平均强度与规定标准有无差异？——假设检验

（3）若采用不同工艺，抽样的50个强度数据有大有小，那么强度呈现的差异是由工艺不同造成的，还是仅仅由随机因素造成的？——方差分析

（4）若钢筋强度与某种原料成分的含量有关，那么从抽查50根钢筋得到的强度与该成分含量的对应数据，如何表达整批钢筋强度与该成分含量之间的关系？——回归分析

第一节引论

4.应用：工业，农业，医药卫生，生物，环境，管理，金融，保险等等

- 《草木缘情：中国古典文学中的植物世界》–潘富俊
  植物会解决文学史上的公案？它们说《红楼梦》不是曹雪芹一个人写的。《红楼梦》前80回平均每回出现植物11种，后40回每回3.8种。
- 《魔鬼经济学(Freakonomics)》–史蒂芬·列维特，史蒂芬·都伯纳

第一节引论

5.统计计算是一门包含数理统计、计算数学及计算机科学的交叉学科。

6.软件：**Matlab, Splus, SAS, SPSS, R**等等

第二节 数理统计的基本概念

第二节数理统计的基本概念

一.总体和样本

1.总体：所研究的全部元素组成的集合（母体）

2.个体：总体中的每个元素称为个体

例：研究某批灯泡的寿命，则该批灯泡寿命数据的集合就构成一个总体，其中每个灯泡的寿命就是一个个体。

注:
(1)常用大写字母X,Y,Z表示总体。

(2) X是一随机变量，若X的分布函数为$F(x)$,则总体X为具有分布函数$F(x)$的总体。

第二节数理统计的基本概念

3.样本： 从总体抽取的部分个体，称为样本，
用$(X_1, X_2, ..., X_n)$表示，n为样本中所含的个体数，称为样本大小或样本容量。

注:(1)$(X_1, \cdots, X_n)$可以看成一个n维随机向量。

(2)每次抽样观测得到$(X_1, \cdots, X_n)$的一组确定值$(x_1, \cdots, x_n)$称作样本观测值。

(3)样本$(X_1, \cdots, X_n)$可能取值的全体称为样本空间(Sample Space),记为$\mathcal{X}$.即$(x_1, \cdots, x_n) \in \mathcal{X}$

4.简单随机样本： 满足以下两条性质

(1)代表性：$X_1, \cdots, X_n$与总体X有相同的分布.

(2)独立性：$X_1, \cdots, X_n$是相互独立的随机变量.

第二节数理统计的基本概念

设总体$X \sim F(x)$，则样本$(X_1, \cdots, X_n)$的联合分布函数为$\prod_{i=1}^{n} F(x_i)$,

- 若连续型总体X的密度函数为$f(x)$, 则样本$(X_1, \cdots, X_n)$的联合密度函数为$f_n(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i)$.
- 若总体X为离散型, 则样本$(X_1, \cdots, X_n)$的联合分布率为$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$.

例1：设某批产品共有N个，其中次品数为M，其次品率为$p = M/N$, 从这批产品中任取一件，用X 来描述其质量，$X = \begin{cases} 1, & \text{所取产品为次品} \\ 0, & \text{所取产品为正品} \end{cases}$ 则X的分布为

$P\{X = x\} = p^x(1-p)^{1-x}, x = 0, 1.$

随机抽取样本$(X_1, \cdots, X_n)$，其联合分布率为

$P\{X_1 = x_1, ..., X_n = x_n\} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, x_i = 0, 1.$

例2：某城市居民收入服从正态分布$N(\mu, \sigma^2)$,概率密度函数为$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, -\infty < x < \infty$

随机抽取样本$(X_1, \cdots, X_n)$，其联合密度为$L(x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\}$

例1：设某批产品共有N个，其中次品数为M，其次品率为$p = M/N$, 从这批产品中任取一件，用X 来描述其质量，$X = \begin{cases} 1, & \text{所取产品为次品} \\ 0, & \text{所取产品为正品} \end{cases}$ 则X的分布为

$$P\{X = x\} = p^x(1-p)^{1-x}, x = 0, 1.$$

随机抽取样本$(X_1, \cdots, X_n)$，其联合分布率为

$$P\{X_1 = x_1, ..., X_n = x_n\} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, x_i = 0, 1.$$

例2：某城市居民收入服从正态分布$N(\mu, \sigma^2)$,概率密度函数为$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, -\infty < x < \infty$

随机抽取样本$(X_1, \cdots, X_n)$，其联合密度为$L(x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\}$

例1：设某批产品共有N个，其中次品数为M，其次品率为$p = M/N$, 从这批产品中任取一件，用X 来描述其质量，$X = \begin{cases} 1, & \text{所取产品为次品} \\ 0, & \text{所取产品为正品} \end{cases}$ 则X的分布为

$P\{X = x\} = p^x(1-p)^{1-x}, x = 0, 1.$

随机抽取样本$(X_1, \cdots, X_n)$，其联合分布率为

$P\{X_1 = x_1, ..., X_n = x_n\} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, x_i = 0, 1.$

例2：某城市居民收入服从正态分布$N(\mu, \sigma^2)$,概率密度函数为$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, -\infty < x < \infty$

随机抽取样本$(X_1, \cdots, X_n)$，其联合密度为$L(x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\}$

例1：设某批产品共有N个，其中次品数为M，其次品率为$p = M/N$, 从这批产品中任取一件，用X 来描述其质量，$X = \begin{cases} 1, & \text{所取产品为次品} \\ 0, & \text{所取产品为正品} \end{cases}$ 则X的分布为

$P\{X = x\} = p^x(1-p)^{1-x}, x = 0, 1.$

随机抽取样本$(X_1, \cdots, X_n)$，其联合分布率为

$P\{X_1 = x_1, ..., X_n = x_n\} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, x_i = 0, 1.$

例2：某城市居民收入服从正态分布$N(\mu, \sigma^2)$,概率密度函数为$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, -\infty < x < \infty$

随机抽取样本$(X_1, \cdots, X_n)$，其联合密度为$L(x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\}$

例1：设某批产品共有N个，其中次品数为M，其次品率为$p = M/N$, 从这批产品中任取一件，用X 来描述其质量，$X = \begin{cases} 1, & \text{所取产品为次品} \\ 0, & \text{所取产品为正品} \end{cases}$ 则X的分布为

$P\{X = x\} = p^x(1-p)^{1-x}, x = 0, 1.$

随机抽取样本$(X_1, \cdots, X_n)$，其联合分布率为

$P\{X_1 = x_1, ..., X_n = x_n\} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, x_i = 0, 1.$

例2：某城市居民收入服从正态分布$N(\mu, \sigma^2)$,概率密度函数为$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, -\infty < x < \infty$

随机抽取样本$(X_1, \cdots, X_n)$，其联合密度为$L(x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\}$

例1：设某批产品共有N个，其中次品数为M，其次品率为$p = M/N$, 从这批产品中任取一件，用X 来描述其质量，$X = \begin{cases} 1, & \text{所取产品为次品} \\ 0, & \text{所取产品为正品} \end{cases}$ 则X的分布为

$P\{X = x\} = p^x(1-p)^{1-x}, x = 0, 1.$

随机抽取样本$(X_1, \cdots, X_n)$，其联合分布率为

$P\{X_1 = x_1, ..., X_n = x_n\} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, x_i = 0, 1.$

例2：某城市居民收入服从正态分布$N(\mu, \sigma^2)$,概率密度函数为$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}, -\infty < x < \infty$

随机抽取样本$(X_1, \cdots, X_n)$，其联合密度为$L(x_1, \cdots, x_n) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\}$

第二节数理统计的基本概念

二.直方图–分布密度函数曲线的近似

1.找出样本观测值的最小值和最大值,并把包含它们的区间$[a,\ b]$分成m等分.

注：一般分为7–18组，以$m = 1 + 3.32 \log n$作为组数,h为组距

$$a = c_0 < c_1 < \cdots < c_m = b, h = c_i - c_{i-1} = \frac{b-a}{m}, \qquad i = 1, \cdots, m,$$

2.数出样本观测值落在各区间$(c_{i-1},\ c_i]$中的个数n_i,称为第i组的组频数,$f_i = \frac{n_i}{n}$称为第i组的组频率。

3.同一组的数据都看成是相同的,它们都等于组中值$\frac{c_{i-1}+c_i}{2}$,分组整理。

4.在x轴上标出点$c_i,\ i = 0, 1, \cdots, m$,以各区间$(c_{i-1},\ c_i]$为底,组频率与组距之比$y_i = \frac{f_i}{h} = \frac{n_i}{nh}$为高作矩形,这种图称为频率直方图。

例3：表中的125 个数据表示某锅炉所炼生铁中锰的含量,每天测得5 个数据，作频率直方图。

1.40	1.28	1.36	1.38	1.44	1.40	1.34	1.54	1.44	1.46
1.80*	1.44	1.46	1.50	1.38	1.54	1.50	1.48	1.52	1.58
1.52	1.46	1.42	1.58	1.70	1.62	1.58	1.62	1.76	1.68
1.68	1.66	1.62	1.72	1.60	1.62	1.46	1.38	1.42	1.38
1.60	1.44	1.46	1.38	1.34	1.38	1.34	1.36	1.58	1.38
1.34	1.28	1.08	1.08	1.36	1.50	1.46	1.28	1.18	1.28
1.26	1.50	1.52	1.38	1.50	1.52	1.50	1.46	1.34	1.40
1.50	1.42	1.38	1.36	1.38	1.42	1.34	1.48	1.36	1.36
1.32	1.40	1.40	1.26	1.26	1.16	1.34	1.40	1.16	1.54
1.24	1.22	1.20	1.30	1.36	1.30	1.48	1.28	1.18	1.28
1.30	1.52	1.76	1.16	1.28	1.48	1.46	1.48	1.42	1.36
1.32	1.22	1.72	1.18	1.36	1.44	1.28	1.10	1.06*	1.10
1.16	1.22	1.24	1.22	1.34					

第二节数理统计的基本概念

表**1.2** 生铁含锰量频数表

各组分点	组中值	频数	各组分点	组中值	频数
0.99–1.09	1.04	3	1.39–1.49	1.44	29
1.09–1.19	1.14	9	1.49–1.59	1.54	19
1.19–1.29	1.24	18	1.59–1.69	1.64	9
1.29–1.39	1.34	32	1.69–1.79	1.74	5
			1.79–1.89	1.84	1

注:(1)频率直方图面积为1,

(2)在R中命令为hist;

第二节数理统计的基本概念

众数：数据最集中的取值，也就是最大频率所对应的组中值；

算术平均：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$；

中位数：将数据按大小顺序排列，居于中间的那个数值；

极差：最大观测值与最小观测值之差。

三 统计量

定义1：设$(X_1,\cdots,X_n)$是来自总体X的一个样本，
$T=T(x_1,\cdots,x_n)$ 是样本空间$\mathcal{X}$上的实值函数，
若$T(X_1,\cdots,X_n)$也是随机变量,且不依赖于任何未知参数,则
称$T(X_1,\cdots,X_n)$为**统计量**(Statistics).

例4.设X_1,X_2是从总体$N(\mu,\sigma^2)$中抽取的一个样本，其中参数μ已知，σ^2未知，判定以下各式哪个是统计量？

$$X_1\cdot X_2-3\sigma^2,\quad X_1^2+X_2^2+5\sigma,\quad X_1,\quad X_1+5X_2^2,\quad X_1/X_2+3\mu X_1^2$$

注：
(1)借助统计量，可以把子样所含信息进行数学加工，使其浓缩，从而使问题解决。
(2)统计量是一个随机变量。

三 统计量

定义2： 设$(X_1, \cdots, X_n)$是取自总体X的大小为n的样本,则

(1)$\overline{X} = \frac{1}{n}\Sigma_{i=1}^{n} X_i$称为样本均值;

(2)$S^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \overline{X})^2$称为样本方差,

而$S = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \overline{X})^2}$为样本均方差或样本标准差;

(3)$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$称为样本的$k$阶原点矩;

$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$为样本的$k$阶中心距.

注:二阶中心矩B_2有时记为$\tilde{S}^2$,即$\tilde{S}^2 = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \overline{X})^2$.

特别地,$A_1 = \overline{X},\ B_2 = \tilde{S}^2 = \frac{n-1}{n}S^2$.

(4)$b_s = \frac{B_3}{B_2^{3/2}}$,称为样本的偏度;

(5)$b_k = \frac{B_4}{B_2^2} - 3$称为样本峰度;

(6)$V = \frac{S}{\overline{X}}$称为样本的变异系数。

三 统计量

记$E(X) \triangleq \mu$, $D(X) \triangleq \sigma^2$, $E(X^k) \triangleq \alpha_k$, $E(X-\mu)^k \triangleq \mu_k$

定理1.1:设总体X的分布函数$F(x)$具有二阶矩,$(X_1, \cdots, X_n)$是取自这个总体的一个样本, 则对样本均值$\overline{X}$,有

$$E(\overline{X}) = \mu, \qquad D(\overline{X}) = \frac{\sigma^2}{n}.$$

定理1.2:设总体X的分布函数$F(x)$具有二阶矩,$(X_1, \cdots, X_n)$是取自这个总体的一个样本,则对样本方差S^2,有

$$E(S^2) = \sigma^2,$$

三 统计量

定理1.3: 设总体X的分布函数$F(x)$具有$2k$阶矩,$(X_1, \cdots, X_n)$是取自这个总体的一个样本,则对k 阶样本矩A_k,有

$$E(A_k) = \alpha_k, \quad D(A_k) = \frac{\alpha_{2k} - \alpha_k^2}{n}$$

第三节 次序统计量及其分布

第三节 次序统计量及其分布

一.次序统计量及经验分布函数

1.**定义**: 设$(X_1, \cdots, X_n)$是取自总体X的一个样本,将它们按大小排列为$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, 则称$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$为**次序统计量**,称$X_{(i)}$为第$i$ 个次序统计量.

特别:$X_{(1)} = \min\{X_1, \cdots, X_n\}, X_{(n)} = \max\{X_1, \cdots, X_n\}$分别称为最小次序统计量和最大次序统计量。他们的取值分别称为极小值和极大值.

2.**定义**: 设$(X_1, \cdots, X_n)$是取自分布函数为$F(x)$的总体X的一个样本,以$\nu_n(x)$ 表示$(X_1, \cdots, X_n)$ 中不超过x的观测值的个数,则称$F_n(x) = \frac{\nu_n(x)}{n}$为**经验分布函数**,简记为EDF.

第三节 次序统计量及其分布

注:$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$,则经验分布函数为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}; \\ \frac{k}{n}, & x_{(k)} \le x < x_{(k+1)},\ k = 1, \cdots, n-1; \\ 1, & x_{(n)} \le x, \end{cases}$$

第三节 次序统计量及其分布

3.性质:(1)$0 \leq F_n(x) \leq 1$;

(2)单调不减的右连续函数;

(3)在$x = x_{(k)}$有间断点,在每个间断点上有跃度$\frac{1}{n}$.

例：从总体X中抽取容量为7的样本，其观测值为1, 3, 2.5, 2, 2.5, 3, 2.5，求X的经验分布函数。

解：将样本由小到大排序，

有$1 < 2 < 2.5 = 2.5 = 2.5 < 3 = 3$,由定义得经验分布函数为

$$F_7(x) = \begin{cases} 0, & x < 1; \\ \frac{1}{7}, & 1 \leq x < 2 \\ \frac{2}{7}, & 2 \leq x < 2.5 \\ \frac{5}{7}, & 2.5 \leq x < 3 \\ 1, & x \geq 3, \end{cases}$$

第三节 次序统计量及其分布

3.性质:(1)$0 \leq F_n(x) \leq 1$;

(2)单调不减的右连续函数;

(3)在$x = x_{(k)}$有间断点,在每个间断点上有跃度$\frac{1}{n}$.

例：从总体X中抽取容量为7的样本，其观测值为1, 3, 2.5, 2, 2.5, 3, 2.5，求X的经验分布函数。

解：将样本由小到大排序，

有$1 < 2 < 2.5 = 2.5 = 2.5 < 3 = 3$,由定义得经验分布函数为

$$F_7(x) = \begin{cases} 0, & x < 1; \\ \frac{1}{7}, & 1 \leq x < 2 \\ \frac{2}{7}, & 2 \leq x < 2.5 \\ \frac{5}{7}, & 2.5 \leq x < 3 \\ 1, & x \geq 3, \end{cases}$$

第三节 次序统计量及其分布

3.性质:(1)$0 \leq F_n(x) \leq 1$;

(2)单调不减的右连续函数;

(3)在$x = x_{(k)}$有间断点,在每个间断点上有跃度$\frac{1}{n}$.

例：从总体X中抽取容量为7的样本，其观测值为1, 3, 2.5, 2, 2.5, 3, 2.5，求X的经验分布函数。

解：将样本由小到大排序，

有$1 < 2 < 2.5 = 2.5 = 2.5 < 3 = 3$,由定义得经验分布函数为

$$F_7(x) = \begin{cases} 0, & x < 1; \\ \frac{1}{7}, & 1 \leq x < 2 \\ \frac{2}{7}, & 2 \leq x < 2.5 \\ \frac{5}{7}, & 2.5 \leq x < 3 \\ 1, & x \geq 3, \end{cases}$$

第三节 次序统计量及其分布

4. 记$F(x) = P\{X \leq x\}$, $F_n(x) = \frac{\nu_n(x)}{n}$表示$\{X \leq x\}$的频率，固定$x$, $\nu_n(x), F_n(x)$是$(X_1, ..., X_n)$的函数，是随机变量.

$$P\{\nu_n(x) = k\} = P\{n\text{次独立试验中恰有 } k\text{次} X \leq x\}$$

则$\nu_n(x) \sim b(n, F(x))$.

$$= \binom{n}{k}[F(x)]^k[1 - F(x)]^{n-k}.$$

由二项分布知$E[\nu_n(x)] = nF(x)$,则$E[F_n(x)] = F(x)$.

第三节 次序统计量及其分布

4. 记$F(x) = P\{X \leq x\}$, $F_n(x) = \frac{\nu_n(x)}{n}$表示$\{X \leq x\}$的频率，固定$x$, $\nu_n(x), F_n(x)$是$(X_1, ..., X_n)$的函数，是随机变量.

$$P\{\nu_n(x) = k\} \quad = \quad P\{n\text{次独立试验中恰有 } k\text{次} X \leq x\}$$

则$\nu_n(x) \sim b(n, F(x))$.

$$= \quad \binom{n}{k}[F(x)]^k[1 - F(x)]^{n-k}.$$

由二项分布知$E[\nu_n(x)] = nF(x)$,则$E[F_n(x)] = F(x)$.

第三节 次序统计量及其分布

4. 记$F(x) = P\{X \leq x\}$, $F_n(x) = \frac{\nu_n(x)}{n}$表示$\{X \leq x\}$的频率，固定$x$, $\nu_n(x), F_n(x)$是$(X_1, ..., X_n)$的函数，是随机变量.

$$P\{\nu_n(x) = k\} \quad = \quad P\{n\text{次独立试验中恰有 } k\text{次} X \leq x\}$$

则$\nu_n(x) \sim b(n, F(x))$.

$$= \quad \binom{n}{k}[F(x)]^k[1 - F(x)]^{n-k}.$$

由二项分布知$E[\nu_n(x)] = nF(x)$,则$E[F_n(x)] = F(x)$.

第三节 次序统计量及其分布

4. 记$F(x) = P\{X \leq x\}$, $F_n(x) = \frac{\nu_n(x)}{n}$表示$\{X \leq x\}$的频率，固定$x$, $\nu_n(x), F_n(x)$是$(X_1, ..., X_n)$的函数，是随机变量.

$$P\{\nu_n(x) = k\} = P\{n\text{次独立试验中恰有 } k\text{次} X \leq x\}$$

则$\nu_n(x) \sim b(n, F(x))$.

$$= \binom{n}{k}[F(x)]^k[1 - F(x)]^{n-k}.$$

由二项分布知$E[\nu_n(x)] = nF(x)$,则$E[F_n(x)] = F(x)$.

第三节 次序统计量及其分布

4. 记$F(x) = P\{X \leq x\}$, $F_n(x) = \frac{\nu_n(x)}{n}$表示$\{X \leq x\}$的频率，固定$x$, $\nu_n(x), F_n(x)$是$(X_1, ..., X_n)$的函数，是随机变量.

$$P\{\nu_n(x) = k\} = P\{n\text{次独立试验中恰有 } k\text{次} X \leq x\}$$

则$\nu_n(x) \sim b(n, F(x))$.

$$= \binom{n}{k}[F(x)]^k[1 - F(x)]^{n-k}.$$

由二项分布知$E[\nu_n(x)] = nF(x)$,则$E[F_n(x)] = F(x)$.

第三节 次序统计量及其分布

5.**Glivenko定理**: 设总体X的分布函数为$F(x)$,经验分布函数为$F_n(x)$,记

$$D_n = \sup_{-\infty<x<+\infty} | F_n(x) - F(x) |,$$

则有$P\{\lim_{n\to\infty} D_n = 0\} = 1.$

注:(1)当n很大时，由每一组样本观测值得到的经验分布函数$F_n(x)$都是总体分布$F(x)$的一个良好近似。(样本推断总体)

(2)两种近似方法:频率直方图(密度)；经验分布函数(分布)

第三节 次序统计量及其分布

二.次序统计量的分布

(一)单个次序统计量由于$\nu_n(x) \sim b(n, F(x))$, 易知

$$\begin{aligned}
\{\nu_n(x) = 0\} &= \{X_{(1)} > x\}, \quad \{\nu_n(x) = n\} = \{X_{(n)} \le x\}, && (1)\\
\{\nu_n(x) = k\} &= \{X_{(k)} \le x < X_{(k+1)}\},\ 1 \le k \le n-1, && (2)\\
\{\nu_n(x) \ge k\} &= \{X_{(k)} \le x\},\ 1 \le k \le n-1. && (3)
\end{aligned}$$

若总体X有分布函数$F(x)$, 密度函数$f(x)$, 则$X_{(i)}$的分布函数为

$$\begin{aligned}
F_i(x) &= P\{X_{(i)} \le x\} = P\{\nu_n(x) \ge i\}\\
&= \sum_{k=i}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} && (4)
\end{aligned}$$

第三节 次序统计量及其分布

二.次序统计量的分布

(一)单个次序统计量由于$\nu_n(x) \sim b(n, F(x))$, 易知

$$\begin{aligned}
\{\nu_n(x) = 0\} &= \{X_{(1)} > x\}, \quad \{\nu_n(x) = n\} = \{X_{(n)} \le x\}, && (1)\\
\{\nu_n(x) = k\} &= \{X_{(k)} \le x < X_{(k+1)}\},\ 1 \le k \le n-1, && (2)\\
\{\nu_n(x) \ge k\} &= \{X_{(k)} \le x\},\ 1 \le k \le n-1. && (3)
\end{aligned}$$

若总体X有分布函数$F(x)$, 密度函数$f(x)$, 则$X_{(i)}$的分布函数为

$$\begin{aligned}
F_i(x) &= P\{X_{(i)} \le x\} = P\{\nu_n(x) \ge i\}\\
&= \sum_{k=i}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} && (4)
\end{aligned}$$

第三节 次序统计量及其分布

注: (1)最大次序统计量$X_{(n)}$的分布函数为$F_n(x) = [F(x)]^n$,
分布密度函数为$f_n(x) = n[F(x)]^{n-1}f(x)$.

(2)最小次序统计量$X_{(1)}$的分布函数为$F_1(x) = 1 - [1 - F(x)]^n$,
分布密度函数为$f_1(x) = n[1 - F(x)]^{n-1}f(x)$.

例1: 设总体X的密度函数
为$f(x) = \begin{cases} 2x, & 0 \le x < 1 \\ 0, & \text{其他} \end{cases}$，$(X_1, \cdots, X_4)$为取自总体的一个样本，求$X_{(3)}$的分布函数，并计算$P(X_{(3)} > \frac{1}{2})$.

第三节 次序统计量及其分布

注: (1)最大次序统计量$X_{(n)}$的分布函数为$F_n(x) = [F(x)]^n$,
分布密度函数为$f_n(x) = n[F(x)]^{n-1}f(x)$.

(2)最小次序统计量$X_{(1)}$的分布函数为$F_1(x) = 1 - [1 - F(x)]^n$,
分布密度函数为$f_1(x) = n[1 - F(x)]^{n-1}f(x)$.

例1: 设总体X的密度函数
为$f(x) = \begin{cases} 2x, & 0 \le x < 1 \\ 0, & \text{其他} \end{cases}$，$(X_1, \cdots, X_4)$为取自总体的一个样本，求$X_{(3)}$的分布函数，并计算$P(X_{(3)} > \frac{1}{2})$.

第三节 次序统计量及其分布

注: (1)最大次序统计量$X_{(n)}$的分布函数为$F_n(x) = [F(x)]^n$,
分布密度函数为$f_n(x) = n[F(x)]^{n-1}f(x)$.

(2)最小次序统计量$X_{(1)}$的分布函数为$F_1(x) = 1 - [1 - F(x)]^n$,
分布密度函数为$f_1(x) = n[1 - F(x)]^{n-1}f(x)$.

例1: 设总体X的密度函数
为$f(x) = \begin{cases} 2x, & 0 \le x < 1 \\ 0, & \text{其他} \end{cases}$，$(X_1, \cdots, X_4)$为取自总体的一个样本，求$X_{(3)}$的分布函数，并计算$P(X_{(3)} > \frac{1}{2})$.

第三节 次序统计量及其分布

注: (1)最大次序统计量$X_{(n)}$的分布函数为$F_n(x) = [F(x)]^n$,
分布密度函数为$f_n(x) = n[F(x)]^{n-1}f(x)$.

(2)最小次序统计量$X_{(1)}$的分布函数为$F_1(x) = 1 - [1 - F(x)]^n$,
分布密度函数为$f_1(x) = n[1 - F(x)]^{n-1}f(x)$.

例1: 设总体X的密度函数
为$f(x) = \begin{cases} 2x, & 0 \le x < 1 \\ 0, & \text{其他} \end{cases}$，$(X_1, \cdots, X_4)$为取自总体的一个样本，求$X_{(3)}$的分布函数，并计算$P(X_{(3)} > \frac{1}{2})$.

第三节 次序统计量及其分布

注: (1)最大次序统计量$X_{(n)}$的分布函数为$F_n(x) = [F(x)]^n$,
分布密度函数为$f_n(x) = n[F(x)]^{n-1}f(x)$.

(2)最小次序统计量$X_{(1)}$的分布函数为$F_1(x) = 1 - [1 - F(x)]^n$,
分布密度函数为$f_1(x) = n[1 - F(x)]^{n-1}f(x)$.

例1: 设总体X的密度函数
为$f(x) = \begin{cases} 2x, & 0 \le x < 1 \\ 0, & \text{其他} \end{cases}$，$(X_1, \cdots, X_4)$为取自总体的一个样本，求$X_{(3)}$的分布函数，并计算$P(X_{(3)} > \frac{1}{2})$.

第三节 次序统计量及其分布

解:因为$F(x)=\begin{cases}0, & x<0\\ x^2, & 0\le x<1\\ 1, & x\ge 1\end{cases}$,

而

$$\begin{aligned}F_3(x) &= \sum_{k=3}^{4}\binom{4}{k}[F(x)]^k[1-F(x)]^{4-k}\\ &= 4[F(x)]^3[1-F(x)]+[F(x)]^4\\ &= 4[F(x)]^3-3[F(x)]^4\\ &= \begin{cases}0, & x<0\\ 4x^6-3x^8, & 0\le x<1\\ 1, & x\ge 1\end{cases}\end{aligned}$$

所以

$$\begin{aligned}P(X_{(3)}>\tfrac{1}{2}) &= 1-P(X_{(3)}\le\tfrac{1}{2})=1-F_3(\tfrac{1}{2})\\ &= 1-(4(\tfrac{1}{2})^6-3(\tfrac{1}{2})^8)=1-\frac{13}{256}=\frac{243}{256}\end{aligned}$$

第三节 次序统计量及其分布

解:因为$F(x)=\begin{cases}0, & x<0\\ x^2, & 0\le x<1\\ 1, & x\ge 1\end{cases}$,

而

$$\begin{aligned}F_3(x) &= \sum_{k=3}^{4}\binom{4}{k}[F(x)]^k[1-F(x)]^{4-k}\\ &= 4[F(x)]^3[1-F(x)]+[F(x)]^4\\ &= 4[F(x)]^3-3[F(x)]^4\\ &= \begin{cases}0, & x<0\\ 4x^6-3x^8, & 0\le x<1\\ 1, & x\ge 1\end{cases}\end{aligned}$$

所以

$$\begin{aligned}P(X_{(3)}>\tfrac{1}{2}) &= 1-P(X_{(3)}\le\tfrac{1}{2})=1-F_3(\tfrac{1}{2})\\ &= 1-(4(\tfrac{1}{2})^6-3(\tfrac{1}{2})^8)=1-\frac{13}{256}=\frac{243}{256}\end{aligned}$$

第三节 次序统计量及其分布

解:因为$F(x)=\begin{cases}0, & x<0\\ x^2, & 0\le x<1\\ 1, & x\ge 1\end{cases}$,

而

$$\begin{aligned}F_3(x) &= \sum_{k=3}^{4}\binom{4}{k}[F(x)]^k[1-F(x)]^{4-k}\\ &= 4[F(x)]^3[1-F(x)]+[F(x)]^4\\ &= 4[F(x)]^3-3[F(x)]^4\\ &= \begin{cases}0, & x<0\\ 4x^6-3x^8, & 0\le x<1\\ 1, & x\ge 1\end{cases}\end{aligned}$$

所以

$$\begin{aligned}P(X_{(3)}>\tfrac{1}{2}) &= 1-P(X_{(3)}\le\tfrac{1}{2})=1-F_3(\tfrac{1}{2})\\ &= 1-(4(\tfrac{1}{2})^6-3(\tfrac{1}{2})^8)=1-\frac{13}{256}=\frac{243}{256}\end{aligned}$$

第三节 次序统计量及其分布

三. 分位数

1.定义:若随机变量X的分布函数为$F(x)$,对任$0 < p < 1$,称

$$F^{\leftarrow}(p) = inf\{x : F(x) \geq p\}$$

为随机变量X或分布$F(x)$的分位数函数.记$x_p = F^{\leftarrow}(p)$，称x_p为为随机变量X或分布$F(x)$的p分位数特别当$p = \frac{1}{2}$时,$x_{\frac{1}{2}}$称为中位数.

2.定义:设$(X_1,\cdots,X_n)$为总体X的一个样本,$(X_{(1)},\cdots,X_{(n)})$为次序统计量,对任$0<p<1$,称$x_p^*=X_{([np]+1)}$ 为样本p分位数.其中$[a]$表示不超过a的最大整数, 特别当$p=\frac{1}{2}$时,$x_{\frac{1}{2}}^*$称为样本中位数.

$$x_{\frac{1}{2}}^*=\begin{cases} X_{(\frac{n+1}{2})}, & \text{当}n\text{为奇数;} \\ \frac{1}{2}[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}], & \text{当}n\text{为偶数.} \end{cases}$$

第三节 次序统计量及其分布

四.极值分布

1.**定义**:设$(X_1, \cdots, X_n)$是取自分布为$F(x)$的总体X的一个样本,如果存在常数列a_n及$b_n > 0$, 使$(X_{(n)} - a_n)/b_n$有非退化的极限分布$G(x)$,则称$G(x)$为**极大值分布**.

注: 类似可以定义极小值分布,极大值分布和极小值分布统称极值分布.

I型极大值分布(也称贡贝尔(Gumbel)分布)

$$G_1(x) = \exp(-e^{-x}), -\infty < x < \infty,$$

II型极大值分布(也称弗莱契(Fréchet)分布)

$$G_2(x) = \begin{cases} \exp(-x^{-k}), & x > 0; \\ 0, & x \leq 0, \end{cases}$$

其中$k > 0$是参数.

第三节 次序统计量及其分布

III型极大值分布(也称威布尔(Weibull)分布)

$$G_3(x) = \begin{cases} 1, & x > 0; \\ \exp(-(-x)^k), & x \leq 0, \end{cases}$$

其中$k > 0$是参数.

第四节 统计中常用的分布族

第四节 统计中常用的分布族

$F(x;\theta)$表示X的分布,参数θ可能取值的集合称为**参数空间**, 记作Θ, 称$\{F(x;\theta):\theta\in\Theta\}$为$X$的分布函数族.

(1)正态分布族: $\{N(\mu,\sigma^2):(\mu,\sigma^2)\in\Theta\}$, 其中$\Theta=\{(\mu,\sigma^2):-\infty<\mu<\infty,\sigma^2>0\}$.

(2)二项分布族: $\{b(n,p):0<p<1\}$

(3)Possion分布族: $\{P(\lambda):\lambda>0\}$

(4)均匀分布族:$\{U(a,b):-\infty<a<b<\infty\}$

(5)指数分布族:$\{Exp(\lambda):\lambda>0\}$

第四节 统计中常用的分布族

一.Gamma分布族

1.定义：若随机变量X具有密度函数

$$f(x;\alpha,\lambda)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x},\ x>0,$$

则称X所服从的分布为Gamma分布,记作$Ga(\alpha,\lambda)$,其中$\alpha>0$是形状参数,$\lambda>0$ 是尺度参数, $\{Ga(\alpha,\lambda):\alpha>0,\lambda>0\}$称为Gamma分布族.

2.性质：

(1)特征函数:$\varphi(t)=Ee^{itX}=(1-\frac{it}{\lambda})^{-\alpha}$

(2)$E(X)=\frac{\alpha}{\lambda},\ D(X)=\frac{\alpha}{\lambda^2}$

第四节 统计中常用的分布族

(3)当$\alpha = 1$时,$Ga(1,\lambda)$就是参数为λ的指数分布.$Ga(\frac{n}{2},\frac{1}{2})$称为$n$个自由度的$\chi^2$分布,记作$\chi^2(n)$分布,$\chi^2(n)$分布的密度函数为$f(x,n) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, x > 0$

(4)可加性:设$X_1 \sim Ga(\alpha_1,\lambda), X_2 \sim Ga(\alpha_2,\lambda)$,且$X_1$与$X_2$相互独立,则$X_1 + X_2 \sim Ga(\alpha_1 + \alpha_2,\lambda)$

注: (1) $\chi^2(n)$分布是Gamma分布的特殊情况, 也具有可加性.
(2) 自由度是指独立随机变量的个数.

第四节 统计中常用的分布族

例1： 设$X \sim Ga(\alpha, \lambda)$,其中$\alpha > 0, \lambda > 0$为常数，$(X_1, \cdots, X_n)$是取自总体的样本,求样本和$\sum_{i=1}^{n} X_i$的分布密度函数。

解：因为Gamma分布具有可加性，且$X_i, i = 1, 2, ..., n$独立同分布，所以$\sum_{i=1}^{n} X_i \sim Ga(n\alpha, \lambda)$

第四节 统计中常用的分布族

例2：设$X \sim Ga(\alpha, \lambda)$，$Y = kX$,证明：$Y \sim Ga(\alpha, \frac{\lambda}{k})$，$k > 0$，并且求$\bar{X}$的分布。

解：因为$X \sim Ga(\alpha_1, \lambda)$，即密度函数为

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \ x > 0,$$

又因为$Y = kX$,所以$X = Y/k$,因而有

$$\begin{aligned} f(y; \alpha, \lambda) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} (\frac{y}{k})^{\alpha-1} e^{-\lambda \frac{y}{k}} \frac{1}{k} \\ &= \frac{(\frac{\lambda}{k})^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\frac{\lambda}{k}}, \\ &\sim Ga(\alpha, \frac{\lambda}{k}), \quad y > 0 \end{aligned}$$

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim Ga(n\alpha, n\lambda)$

第四节 统计中常用的分布族

例3： 设$X \sim N(0,\sigma^2)$,则$Y = X^2 \sim Ga(\frac{1}{2}, \frac{1}{2\sigma^2})$.

解：当$y > 0$时,Y的分布函数为

$$\begin{aligned} F_Y(y) &= P\{Y \le y\} = P\{X^2 \le y\} = P\{-\sqrt{y} \le X \le \sqrt{y}\} \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}), \end{aligned}$$

其分布密度函数为

$$f_Y(y) = [f_X(\sqrt{y}) + f_X(-\sqrt{y})]\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}\sigma} y^{-\frac{1}{2}} e^{-\frac{y}{2\sigma^2}},$$

因为$\Gamma(\frac{1}{2}) = \sqrt{\pi}$,因此$Y \sim Ga(\frac{1}{2}, \frac{1}{2\sigma^2})$.

第四节 统计中常用的分布族

定理： 设$(X_1, ..., X_n)$是取自正态总体$N(0, \sigma^2)$的一个样本，记$\chi^2 = \sum_{i=1}^{n} X_i^2$, 则$\chi^2 \sim Ga(\frac{n}{2}, \frac{1}{2\sigma^2})$.

例4： 设$X \sim U(0, 1)$, 则$Y = -\alpha \ln X \sim Exp(\frac{1}{\alpha}), \alpha > 0$.

3.χ^2分布另一种定义方式

(1)若$X_i \sim N(0, 1), i = 1, 2, ..., n, i.i.d.$,则$Y = X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$

(2)若$X_i \sim N(\mu, \sigma^2), i = 1, 2, ..., n, i.i.d.$,则$\sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma})^2 \sim \chi^2(n)$

例5： 设$(X_1, ..., X_4)$是来自正态总体$N(0, 4)$的样本，$T = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$, 求常数$a, b$, 使得$T \sim \chi^2(2)$.

第四节 统计中常用的分布族

4.χ^2分布的性质

(1)$\chi^2(n)$分布的特征函数$\phi(t) = (1-2it)^{-\frac{n}{2}}$;

(2)设$X \sim \chi^2(n)$, 则$EX = n, VarX = 2n$;

(3)设$X_i \sim \chi^2(n_i), i = 1, 2, ..., k$, 且相互独立,
则$\sum_{i=1}^{k} X_i \sim \chi^2(\sum_{i=1}^{k} n_i)$.

Cochran分解定理: 设$X_1, ..., X_n$是独立同分布的随机变量, $X_i \sim N(0,1), i = 1, 2, ..., n$, $Q_i (i = 1, ..., k)$ 是$X_1, ..., X_n$的二次型, 其秩为n_i. 如果$Q_1 + \cdots + Q_k = \sum_{i=1}^{n} X_i^2$且$\sum_{i=1}^{k} n_i = n$, 则$Q_i \sim \chi^2(n_i), i = 1, 2, ..., k$,且$Q_1, \cdots, Q_k$相互独立.

第四节 统计中常用的分布族

二.Beta分布族

1.定义:若随机变量X具有密度函数

$$f(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1},\qquad 0<x<1$$

则称X所服从的分布为Beta分布,记作$Be(a,b)$,其中$a>0,\ b>0$是两个参数.$\{Be(a,b):a>0,b>0\}$称为**Beta分布族**.

2.性质:

(1)$E(X)=\frac{a}{a+b}$,$\qquad D(X)=\frac{ab}{(a+b)^2(a+b+1)}$

(2)当$a=b=1$时,$Be(1,1)$分布就是$(0,\ 1)$上的均匀分布$U(0,1)$

第四节 统计中常用的分布族

3.定理：若$X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2)$,且$X_1, X_2$相互独立,则
(1)$F = \frac{X_1/n_1}{X_2/n_2} \sim F(n_1, n_2)$, 其中$F(n_1, n_2)$表示自由度为$n_1$, n_2的F分布.
(2)$\frac{X_1}{X_1+X_2} \sim Be(\frac{n_1}{2}, \frac{n_2}{2})$

4.结论
(1)若$X \sim F(n_1, n_2)$,则$\frac{CX}{1+CX} \sim Be(\frac{n_1}{2}, \frac{n_2}{2})$ 其中$C = n_1/n_2$,反之亦然.
(2)若$X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2)$,且$X_1, X_2$相互独立,
则$Y_1 = X_1 + X_2$与$Y_2 = X_1/X_2$相互独立.
(3)若$X \sim F(n_1, n_2)$, 则$\frac{1}{X} \sim F(n_2, n_1)$.

第四节 统计中常用的分布族

三 t分布

1.定义：设随机变量$X \sim N(0,1), Y \sim \chi^2(n)$,且$X, Y$相互独立,则称随机变量$T = \dfrac{X}{\sqrt{Y/n}}$ 所服从的分布为n个自由度的t分布,记作$T \sim t(n)$.

第四节 统计中常用的分布族

2.性质：$t(n)$分布的密度函数为

$$t(x;n)=\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}}(1+\frac{x^2}{n})^{-\frac{n+1}{2}},\qquad -\infty<x<\infty$$

(1)当$n=1$时,$t(x;1)=\frac{1}{\pi}\frac{1}{1+x^2}$即为Cauchy分布.
当$n>2$时,则$E(T)=0,\ D(T)=\frac{n}{n-2}$
(2)若$X\sim t(n)$,则$X^2\sim F(1,n)$.
(3)若$X\sim t(n)$,则$Y=\frac{n}{n+X^2}\sim Be(\frac{n}{2},\frac{1}{2})$.
(4)$\lim_{n\to\infty}t(x;n)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$
t分布的极限分布是标准正态分布。

第四节 统计中常用的分布族

四 指数型分布族

1.定义：设$\mathcal{F}=\{f(x;\theta):\theta\in\Theta\}$是分布族,若样本$(X_1,\cdots,X_n)$的密度函数(或分布列)$f(x_1,\cdots,x_n;\theta)$可以表示成

$$f(x_1,\cdots,x_n;\theta)=a(\theta)\exp\{\sum_{j=1}^{k}Q_j(\theta)T_j(x_1,\cdots,x_n)\}h(x_1,\cdots,x_n)$$

并且它的支撑$\{x:f(x_1,\cdots,x_n;\theta)>0\}$不依赖于$\theta$,则称此分布族为指数型分布族,简称**指数族**。

第四节 统计中常用的分布族

例1:证明二项分布族$\{b(m,p):0<p<1\}$是指数型分布族. 样本空间$\mathcal{X}=\{(x_1,...,x_n):x_i=0,1,...,m,i=1,2,...,n\}$.

证明: 样本$(X_1,\cdots,X_n)$的联合概率分布为

$$f(x_1,\cdots,x_n;p) = [\prod_{i=1}^{n}\binom{m}{x_i}]p^{\sum_{i=1}^{n}x_i}(1-p)^{nm-\sum_{i=1}^{n}x_i}$$

$$= (1-p)^{nm}\exp\{\sum_{i=1}^{n}x_i\ln\frac{p}{1-p}\}\prod_{i=1}^{n}\binom{m}{x_i},$$

取$a(p)=(1-p)^{nm}, Q_1(p)=\ln\frac{p}{1-p}, T_1(x_1,\cdots,x_n)=\sum_{i=1}^{n}x_i, h(x_1,\cdots,x_n)=\prod_{i=1}^{n}\binom{m}{x_i}$.

例2: 均匀分布族$\{U(-\theta,\theta):\theta>0\}$不是指数型分布族,这是因为它的支撑$\{x:f(x;\theta)>0\}=(-\theta,\theta)$依赖于未知参数$\theta$

第四节 统计中常用的分布族

例1:证明二项分布族$\{b(m,p):0<p<1\}$是指数型分布族. 样本空间$\mathcal{X}=\{(x_1,...,x_n):x_i=0,1,...,m,i=1,2,...,n\}$.
证明: 样本$(X_1,\cdots,X_n)$的联合概率分布为

$$\begin{aligned} f(x_1,\cdots,x_n;p) &= [\prod_{i=1}^{n}\binom{m}{x_i}]p^{\sum_{i=1}^{n}x_i}(1-p)^{nm-\sum_{i=1}^{n}x_i} \\ &= (1-p)^{nm}\exp\{\sum_{i=1}^{n}x_i\ln\frac{p}{1-p}\}\prod_{i=1}^{n}\binom{m}{x_i}, \end{aligned}$$

取$a(p)=(1-p)^{nm}, Q_1(p)=\ln\frac{p}{1-p}, T_1(x_1,\cdots,x_n)=\sum_{i=1}^{n}x_i, h(x_1,\cdots,x_n)=\prod_{i=1}^{n}\binom{m}{x_i}$.

例2: 均匀分布族$\{U(-\theta,\theta):\theta>0\}$不是指数型分布族,这是因为它的支撑$\{x:f(x;\theta)>0\}=(-\theta,\theta)$依赖于未知参数$\theta$

第四节 统计中常用的分布族

例3： 设X服从正态分布$N(\mu,\sigma^2)$,样本空间为n维欧氏空间R^n.记$\mathbf{x}=(x_1,\cdots,x_n)$,那么样本$(X_1,\cdots,X_n)$的分布密度为

$$\begin{aligned} f(\mathbf{x};\mu,\sigma^2) &= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\sum_{i=1}^{n}\frac{1}{2\sigma^2}(x_i-\mu)^2\} \\ &= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\frac{n\mu^2}{2\sigma^2}+\frac{n\mu}{\sigma^2}\bar{x}-\sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}\} \\ &= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\frac{n\mu^2}{2\sigma^2}\}\exp\{\frac{n\mu}{\sigma^2}\bar{x}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\}, \end{aligned}$$

第四节 统计中常用的分布族

例3： 设X服从正态分布$N(\mu,\sigma^2)$,样本空间为n维欧氏空间R^n.记$\mathbf{x}=(x_1,\cdots,x_n)$,那么样本$(X_1,\cdots,X_n)$的分布密度为

$$\begin{aligned} f(\mathbf{x};\mu,\sigma^2) &= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\sum_{i=1}^{n}\frac{1}{2\sigma^2}(x_i-\mu)^2\} \\ &= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\frac{n\mu^2}{2\sigma^2}+\frac{n\mu}{\sigma^2}\bar{x}-\sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}\} \\ &= (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\frac{n\mu^2}{2\sigma^2}\}\exp\{\frac{n\mu}{\sigma^2}\bar{x}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\}, \end{aligned}$$

第四节 统计中常用的分布族

只要使

$$a(\mu,\sigma^2)=(\frac{1}{\sqrt{2\pi}\sigma})^n\exp\{-\frac{n\mu^2}{2\sigma^2}\},\ Q_1(\mu,\sigma^2)=\frac{n\mu}{\sigma^2},$$

$$Q_2(\mu,\sigma^2)=-\frac{1}{2\sigma^2},$$

$$T_1(\mathbf{x})=\bar{x},\ T_2(\mathbf{x})=\sum_{i=1}^{n}x_i^2,\ h(\mathbf{x})=1$$

因此正态分布族$\{N(\mu,\sigma^2):-\infty<\mu<\infty,\sigma^2>0\}$是指数型分布族.

第五节 正态总体

第五节 正态总体

一 多元正态总体

1.定义：若随机向量X的联合分布密度函数为

$$f(x)=\frac{1}{(2\pi)^{\frac{n}{2}}|B|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(x-a)'B^{-1}(x-a)\}$$

其中B为正定阵,$|B|$为其行列式.B^{-1}是B的逆矩阵,则称随机向量X所服从的分布为**多元正态分布**,简记为$X\sim N_n(a,B)$.

2.特征函数：设随机向量$X\sim N_n(a,B)$,则其特征函数为

$$\varphi(t)=\exp\{ia't-\frac{1}{2}t'Bt\}$$

其中$t'=(t_1,\cdots,t_n)$

第五节 正态总体

3.期望 方差 协方差阵

设$X=(X_1,\cdots,X_n)'$, $Y=(Y_1,\cdots,Y_m)'$是两个随机向量,

$Z=(Z_{ij})_{r\times s}$是随机矩阵,记

$$\begin{aligned} E(X) &= (E(X_1),\cdots,E(X_n))', \\ E(Z) &= (E(Z_{ij}))_{r\times s}, \end{aligned}$$

$$\begin{aligned} Var(X) &= E(X-E(X))(X-E(X))' \\ &= \begin{pmatrix} D(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & D(X_2) & \cdots & Cov(X_2,X_n) \\ \cdots & \cdots & \cdots & \cdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \cdots & D(X_n) \end{pmatrix} \end{aligned}$$

其中$D(X_i)$为X_i的方差,$Cov(X_i,X_j)$为X_i和X_j的协方差.

第五节 正态总体

$$\begin{aligned} Cov(X,Y) &= (Cov(Y,X))' = E(X-E(X))(Y-E(Y))' \\ &= \begin{pmatrix} Cov(X_1,Y_1) & \cdots & Cov(X_1,Y_m) \\ \vdots & \ddots & \vdots \\ Cov(X_n,Y_1) & \cdots & Cov(X_n,Y_m) \end{pmatrix} \end{aligned}$$

4.$\rho_{ij} = \frac{Cov(X_i,X_j)}{\sqrt{D(X_i)D(X_j)}}$ 称为X_i与X_j之间的线性相关系数,简称为相关系数.

第五节 正态总体

5.性质

(1)$X \sim N_n(a, B)$, 则对X的任一子向量$\tilde{X}' = (X_{k_1}, ..., X_{k_m})(m \leq n)$, 有$\tilde{X} \sim N_m(\tilde{a}, \tilde{B})$, 其中$\tilde{a} = (a_{k_1}, ..., a_{k_m})'$, $\tilde{B}$是B中保留$k_1, ..., k_m$行列所得的m阶子矩阵. 特别地, $X_j \sim N(a_j, b_{jj}), j = 1, 2, ..., n$.

(2)$X \sim N_n(a, B)$, 则$EX = a, Var(X) = B$.

(3)若$X = (X_1, X_2)' \sim N_n(a, B)$, X_1是n_1维向量, X_2是n_2维向量, $n_1 + n_2 = n$, $X_1 \sim N_{n_1}(a_1, B_{11}), X_2 \sim N_{n_2}(a_2, B_{22})$, 其中$a_1 \in R^{n_1}, a_2 \in R^{n_2}$, $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$, 则$X_1, X_2$相互独立$\Leftrightarrow Cov(X_1, X_2) = B_{12} = 0$.

第五节 正态总体

(4)$X \sim N_n(a, B), r(A) = m, A \in R^{m\times n}, b = (b_1, ..., b_m)' \in R^m$, 则$Y = AX + b \sim N_m(Aa + b, ABA')$.

(5)若$X \sim N_n(a, B)$, 则存在一个正交变换Γ, 使得$Y = \Gamma(X - a)$的各分量是相互独立，均值都为零的正态分布随机变量。特别地，若$X \sim N_n(0, \sigma^2 I_n)$, 则$Y = \Gamma X \sim N_n(0, \sigma^2 I_n)$.

第五节 正态总体

二 正态总体统计量的分布

定理： 设$(X_1, \cdots, X_n)$是取自正态总体$N(\mu, \sigma^2)$的一个样本，$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 分别为样本均值和样本方差，则

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

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

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

(4) $\overline{X}$ 和 S^2相互独立.

推论1： $T = \frac{\overline{X}-\mu}{S}\sqrt{n} \sim t(n-1)$

第五节 正态总体

推论2: 设$(X_1,\cdots,X_{n_1})$是取自正态总体$N(\mu_1,\sigma_1^2)$的一个样本,$(Y_1,\cdots,Y_{n_2})$是取自正态总体$N(\mu_2,\sigma_2^2)$的一个样本,且$(X_1,\cdots,X_{n_1})$和$(Y_1,\cdots,Y_{n_2})$相互独立,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) 若 $\sigma_1^2=\sigma_2^2=\sigma^2$, 则

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

其中$\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i \qquad \overline{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i$,

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

$S_w^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, S_w^2称为二个样本的合并方差.

第六节 充分统计量和完备统计量

第六节 充分统计量和完备统计量

一.充分统计量

例1:了解产品的不合格率p,检验员随机抽取了10件产品进行检查,发现第3件和第9件为不合格品,记作$X_3 = 1, X_9 = 1$,其余都是合格品,记$X_i = 0,\ i = 1, 2, \cdots, 10,$ 且$i \neq 3,\ i \neq 9$.当领导问及检验结果时,检验员作了如下二种回答:
(1) 10件产品中有2件不合格品,即$\sum_{i=1}^{10} X_i = 2$;
(2)第9件产品不合格,即$X_9 = 1$.

第六节 充分统计量和完备统计量

1.定义：设$(X_1,\cdots,X_n)$是从具有分布族$\{F(x;\theta):\theta\in\Theta\}$的总体中抽取的一个样本，$T(X_1,\cdots,X_n)$是一统计量.如果在给定$T(X_1,\cdots,X_n)=t$下,$(X_1,\cdots,X_n)$的条件分布与未知参数$\theta$无关，则称统计量$T(X_1,\cdots,X_n)$是分布族$\{F(x;\theta):\theta\in\Theta\}$ 的充分统计量,或称它是θ的**充分统计量**

例2：证明$T_1=\sum_{i=1}^n X_i$是p的充分统计量,而$T_2=X_j$不是充分统计量.

第六节 充分统计量和完备统计量

对于两点分布总体,大小为n的样本的联合分布为

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

其中x_i非0即1. 统计量$T_1 = \sum_{i=1}^n X_i$服从二项分布$b(n,p)$,所以在给定$T_1 = \sum_{i=1}^n X_i = k$下,样本$(X_1, \cdots, X_n)$的条件分布

$$\begin{aligned} & \mathrm{P}\{X_1 = x_1, \cdots, X_n = x_n | T_1 = k; p\} \\ = & \frac{\mathrm{P}\{X_1 = x_1, \cdots, X_n = x_n, \sum_{i=1}^n X_i = k; p\}}{P\{T_1 = k; p\}} \\ = & \frac{p^k(1-p)^{n-k}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{1}{\binom{n}{k}}, \end{aligned}$$

与p无关,即这个条件分布已不包含有关p的任何信息了.

第六节 充分统计量和完备统计量

故$T_1 = \sum_{i=1}^{n} X_i$是p的充分统计量,或T_1是两点分布族的充分统计量.

在给定$T_2 = X_j = k$条件下,样本$(X_1, \cdots, X_n)$的条件分布

$$P\{X_1 = x_1, \cdots, X_n = x_n | T_2 = k; p\} = p^{\sum\limits_{i\neq j} x_i} (1-p)^{n-\sum\limits_{i\neq j} x_i - 1}$$

与p有关,可见$T_2 = X_j$不是充分统计量.

第六节 充分统计量和完备统计量

2.定理：(因子分解定理)设总体X为连续型随机变量,具有分布密度族$\{f(x;\theta):\theta\in\Theta\}$, $(X_1,\cdots,X_n)$取自X的一个样本,则统计量$T(X_1,\cdots,X_n)$为θ的充分统计量的充分必要条件为：样本的联合概率密度函数可分解为

$$\begin{aligned} L(x_1,\cdots,x_n;\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= h(x_1,\cdots,x_n)g(T(x_1,\cdots,x_n),\theta) \end{aligned}$$

其中$h(x_1,\cdots,x_n)$是非负函数且与θ无关,$g(T(x_1,\cdots,x_n),\theta)$仅通过$T(x_1,\cdots,x_n)=t$依赖于$(x_1,\cdots,x_n)$.

第六节 充分统计量和完备统计量

例3： 设$(X_1, \cdots, X_n)$是取自两点分布$b(1, p)$总体的一个样本,其联合概率分布为

$$\begin{aligned} P\{X = x_1, \cdots, X_n = x_n; p\} &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ &= (1-p)^n\left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} \end{aligned}$$

取

$$T(x_1, \cdots, x_n) = \sum_{i=1}^n x_i,\ h(x_1, \cdots, x_n) = 1,$$

$$g(T(x_1, \cdots, x_n), p) = (1-p)^n\left(\frac{p}{1-p}\right)^T$$

则

$$P\{X = x_1, \cdots, X_n = x_n; p\} = h(x_1, \cdots, x_n) \cdot g(T(x_1, \cdots, x_n), p),$$

由因子分解定理可知$T(X_1, \cdots, X_n) = \sum_{i=1}^n X_i$是$p$的充分统计量.

第六节 充分统计量和完备统计量

例3: 设$(X_1,\cdots,X_n)$是取自两点分布$b(1,p)$总体的一个样本,其联合概率分布为

$$\begin{aligned} P\{X=x_1,\cdots,X_n=x_n;p\} &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ &= (1-p)^n\left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} \end{aligned}$$

取

$$T(x_1,\cdots,x_n)=\sum_{i=1}^n x_i,\ h(x_1,\cdots,x_n)=1,$$

$$g(T(x_1,\cdots,x_n),p)=(1-p)^n\left(\frac{p}{1-p}\right)^T$$

则

$$P\{X=x_1,\cdots,X_n=x_n;p\}=h(x_1,\cdots,x_n)\cdot g(T(x_1,\cdots,x_n),p),$$

由因子分解定理可知$T(X_1,\cdots,X_n)=\sum_{i=1}^n X_i$是$p$的充分统计量.

第六节 充分统计量和完备统计量

3.若分布有n个参数,则θ是参数向量,若定理条件成立,则称$T(X_1,\cdots,X_n)$ 为关于θ的**联合充分统计量**

例4.设$X\sim N(\mu,\sigma^2)$,$\theta=(\mu,\sigma^2)$是未知参数向量, 证明:$(\overline{X},\sum_{i=1}^{n}X_i^2)$是$(\mu,\sigma^2)$的联合充分统计量.

第六节 充分统计量和完备统计量

证：$(X_1, \cdots, X_n)$是取自总体的一个样本,其联合密度函数为

$$\begin{aligned} L(x_1, \cdots, x_n; \theta) &= \prod_{i=1}^{n} f(x_i; \theta) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\} \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 + \frac{n\mu}{\sigma^2}\overline{x} - \frac{n\mu^2}{2\sigma^2}\}. \end{aligned}$$

取$h(x_1, \cdots, x_n) = 1$,

$$g(T(x_1, \cdots, x_n), \theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 + \frac{n\mu}{\sigma^2}\overline{x} - \frac{n\mu^2}{2\sigma^2}\}$$

由因子分解定理可知
$T(X_1, \cdots, X_n) = (\overline{X}, \sum_{i=1}^{n}X_i^2)$是$\theta = (\mu, \sigma^2)$的联合充分统计量.

第六节 充分统计量和完备统计量

注:$T(X_1,\cdots,X_n)$是θ的联合充分统计量不能推出T_i是θ_i的充分统计量. 因此不能说明$\overline{X},\sum_{i=1}^{n}X_i^2$分别是$\mu,\sigma^2$的充分统计量.

4.定理: 设$T(X_1,X_2,\cdots,X_n)$是θ的充分统计量, $v=\psi(t)$是单值可逆函数, 则$V=\psi(T)$也是θ的充分统计量.

第六节 充分统计量和完备统计量

二.完备统计量

1.定义: 设$\{F(x;\theta):\theta\in\Theta\}$是一个分布族, 如果由$\mathrm{E}_\theta[g(X)]=0,\forall\theta\in\Theta$, 总可推出$\mathrm{P}_\theta\{g(X)=0\}=1,\forall\theta\in\Theta$, 则称分布族$\{F(x;\theta):\theta\in\Theta\}$是完备的.

例1. 二项分布族$\{b(n,p):0<p<1\}$是完备的.

例2. 正态分布族$\{N(0,\sigma^2):\sigma^2>0\}$是不完备的.

第六节 充分统计量和完备统计量

2.定义: 设$(X_1, \cdots, X_n)$是取自分布族$\{F(x;\theta) : \theta \in \Theta\}$的一个样本, 若统计量$T(X_1, \cdots, X_n)$的对应分布族$\{F^T(t;\theta) : \theta \in \Theta\}$是完备的, 则称$T$是完备的.

注: T完备时, 原分布族$\{F(x;\theta) : \theta \in \Theta\}$可能不完备.

例3. 证明: 对分布族$\{N(0,\sigma^2) : \sigma^2 > 0\}$, $T = \sum_{i=1}^{n} X_i^2$是完备统计量.

例4. 设$(X_1, \cdots, X_n)$是取自均匀分布族$\{U(0,\theta) : 0 < \theta < 1\}$的一个样本, 证明统计量$T = \max\limits_{1\le i\le n} X_i$是$\theta$的充分完备统计量.

第六节 充分统计量和完备统计量

证: $(X_1, \cdots, X_n)$的联合密度函数为

$$L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = \frac{1}{\theta^n} \mathbf{1}_{\{0 < \max\limits_{1 \le i \le n} x_i < \theta\}}.$$

由因子分解定理, $T = \max\limits_{1 \le i \le n} X_i$是$\theta$的充分统计量.

设T的分布函数为$F^T(t; \theta)$, 对$0 < t < \theta$,

$$F^T(t; \theta) = P\{\max_{1 \le i \le n} X_i \le t; \theta\} = [F(t; \theta)]^n = (\frac{t}{\theta})^n.$$

$\therefore f^T(t; \theta) = \frac{nt^{n-1}}{\theta^n}, 0 < t < \theta.$

第六节　充分统计量和完备统计量

设对$\forall 0<\theta<1$, $g(t)$满足

$$E_\theta[g(T)]=\int_0^\theta g(t)f^T(t;\theta)\mathrm{d}t=\frac{n}{\theta^n}\int_0^\theta g(t)t^{n-1}\mathrm{d}t=0,$$

即$\int_0^\theta g(t)t^{n-1}\mathrm{d}t=0,\forall 0<\theta<1$, 两边关于$\theta$求导,

$$g(\theta)\theta^{n-1}=0,a.s.,\forall 0<\theta<1,\quad \therefore g(\theta)=0,a.s.\forall 0<\theta<1.$$

所以, T的对应分布族完备, $T=\max\limits_{1\le i\le n}X_i$是$\theta$的充分完备统计量.

□

第六节 充分统计量和完备统计量

3.定理: 设$\{f(x;\theta):\theta=(\theta_1,\ldots,\theta_k)\in\Theta\}$是含$k$个参数的指数族, 样本$(X_1,\cdots,X_n)$的联合分布密度具有如下形式

$$L(x_1,\cdots,x_n;\theta)=a(\theta)\exp\{\sum_{j=1}^{k}Q_j(\theta)T_j(x_1,\cdots,x_n)\}h(x_1,\cdots,x_n),$$

如果Θ包含一k维矩形, 且$Q=(Q_1,\cdots,Q_k)$的值域包含一k维开集, 则$T(X_1,\cdots,X_n)=(T_1(X_1,\cdots,X_n),\cdots,T_k(X_1,\cdots,X_n))$是$k$维参数向量$\theta=(\theta_1,\ldots,\theta_k)$ 的充分完备统计量.

第六节 充分统计量和完备统计量

例5. 设$(X_1, \cdots, X_n)$是取自Poisson分布族$\{P(\lambda) : \lambda > 0\}$的一个样本, 证明: $T(X_1, \cdots, X_n) = \sum_{i=1}^n X_i$是$\lambda$的充分完备统计量.

证明:

$$\begin{aligned} P\{X_1 = x_1, \cdots, X_n = x_n; \lambda\} &= \mathrm{e}^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \cdots x_n!} \\ &= \mathrm{e}^{-n\lambda} \mathrm{e}^{\sum_{i=1}^n x_i \ln \lambda} \frac{1}{x_1! \cdots x_n!}. \end{aligned}$$

$\therefore \{P(\lambda) : \lambda > 0\}$是单参数指数族. $\Theta = \{\lambda : \lambda > 0\}$包含一维开区间, $Q(\lambda) = \ln \lambda$的值域包含一维开区间, 由定理,
$T(X_1, \cdots, X_n) = \sum_{i=1}^n X_i$是$\lambda$的充分完备统计量. $\square$

第六节 充分统计量和完备统计量

例6. 设$(X_1, \cdots, X_n)$是取自正态分布族$\{N(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$的一个样本, 证明: $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$是$(\mu, \sigma^2)$的充分完备统计量.

证明: $(X_1, \cdots, X_n)$的联合密度函数为

$$\begin{aligned} & L(x_1, \cdots, x_n; \mu, \sigma^2) \\ = & (\frac{1}{\sqrt{2\pi}\sigma})^n \cdot \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\} \\ = & (\frac{1}{\sqrt{2\pi}\sigma})^n \cdot \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i\} \cdot \exp\{-\frac{n\mu^2}{2\sigma^2}\}. \end{aligned}$$

$\{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$包含2维开集, $Q(\mu, \sigma^2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ 的值域包含2维开区间, 所以, $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$是$(\mu, \sigma^2)$的充分完备统计量. □

第二章　参数估计

一.分类

(一)非参数估计问题：总体的分布类型未知，需要由样本构造统计量去估计总体的分布函数或密度函数。

(二)参数估计问题：总体分布类型已知，其中含有未知参数，需由样本来估计未知参数。

例1.学生的成绩$X \sim N(\mu, \sigma^2)$, 但参数μ, σ^2未知，需通过样本来估计。

例2.已知某城市在单位时间内发生交通事故的次数$X \sim P(\lambda)$, 但λ未知，需通过样本来估计。

二.对于参数估计，按问题的性质不同分为两类

1.点估计：适当的选择一个统计量$\hat{\theta}(X_1, \cdots, X_n)$作为未知参数$\theta$的估计.

2.区间估计：求未知参数θ的范围，即求一个区间使以较大的概率包含未知参数.

第一节　矩估计和极大似然估计

第一节 矩估计和极大似然估计

一.矩估计： 用样本矩作为总体矩的一个估计量

(一)矩估计的思想：实质是由辛钦大数定律，若$E\xi^k$存在，则

$$\lim_{n\to\infty} P(|\frac{1}{n}\sum_{i=1}^{n}\xi_i^k - E\xi_i^k| \geq \epsilon) = 0.$$

即$\overline{\xi^k} \to E\xi^k$(P收敛)，所以可用样本矩代替总体矩，求得未知参数θ,即替换原则。

第一节 矩估计和极大似然估计

(二)矩估计的方法：设$X \sim F(x;\theta), \theta = (\theta_1, \cdots, \theta_k)$是未知参数向量,若$F(x;\theta)$的$k$阶矩存在,

$$\alpha_\nu(\theta) = \int_{-\infty}^{\infty} x^\nu \, dF(x;\theta), \qquad 1 \le \nu \le k$$

是$\theta = (\theta_1, \cdots, \theta_k)$的函数.设$(X_1, \cdots, X_n)$是取自总体$X$的一个样本,则由下列方程组

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n X_i = \alpha_1(\theta_1, \cdots, \theta_k); \\ \frac{1}{n}\sum_{i=1}^n X_i^2 = \alpha_2(\theta_1, \cdots, \theta_k); \\ \qquad \vdots \\ \frac{1}{n}\sum_{i=1}^n X_i^k = \alpha_k(\theta_1, \cdots, \theta_k), \end{cases}$$

得到$\theta = (\theta_1, \cdots, \theta_k)$的一组解$\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)$, 其中$\hat{\theta}_\nu = \hat{\theta}_\nu(X_1, \cdots, X_n)$, 并以$\hat{\theta}_\nu$ 作为参数θ_ν的估计量, $\nu = 1, \cdots, k$, 则称$\hat{\theta}_\nu$为未知参数θ_ν的**矩估计量**.

第一节 矩估计和极大似然估计

例1: 求总体X的均值$EX=\mu$和方差$DX=\sigma^2$的矩估计.

解: $(X_1,\cdots,X_n)$是取自总体X的一个样本,若总体的二阶矩α_2存在,则有$\alpha_2=\sigma^2+\mu^2$,

$$\begin{cases} \overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\mu; \\ \frac{1}{n}\sum_{i=1}^{n}X_i^2=\alpha_2=\mu^2+\sigma^2. \end{cases}$$

以此方程组的解作为μ,σ^2的估计

$$\begin{aligned} \hat{\mu} &= \overline{X}; \\ \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\tilde{S}^2. \end{aligned}$$

所以,总体均值μ和方差σ^2的矩估计,分别是样本均值$\overline{X}$和样本二阶中心矩$\tilde{S}^2$. 这个结论对任何总体都成立.

第一节 矩估计和极大似然估计

例2: 设X的密度函数为$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\leq 0\end{cases}$

解: $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=EX=\frac{1}{\lambda}$, 所以$\hat{\lambda}=\frac{1}{\overline{X}}$为$\lambda$的矩估计.

第一节 矩估计和极大似然估计

例3: 设$X \sim U(\theta_1, \theta_2)$,其概率密度函数为

$$f(x;\theta_1,\theta_2) = \begin{cases} \frac{1}{\theta_2-\theta_1}, \theta_1 < x < \theta_2; \\ 0, \text{其它}, \end{cases}$$

其中$\theta_1 < \theta_2$,求$\theta = (\theta_1, \theta_2)$的矩估计.

解: 由均匀分布$U(\theta_1, \theta_2)$的性质,有$EX = \frac{\theta_1+\theta_2}{2}, DX = \frac{(\theta_2-\theta_1)^2}{12}$.
因此可得方程组

$$\begin{cases} \overline{X} = \frac{\theta_1+\theta_2}{2}; \\ \tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{(\theta_2-\theta_1)^2}{12}. \end{cases}$$

以此方程组的解作为θ_1, θ_2的估计,得

$$\begin{cases} \hat{\theta_1} = \overline{X} - \sqrt{3}\tilde{S}; \\ \hat{\theta_2} = \overline{X} + \sqrt{3}\tilde{S}, \end{cases}$$

此即所求的θ_1, θ_2的矩估计.

第一节 矩估计和极大似然估计

例4:设$X \sim \Gamma(\alpha, \beta), \alpha > 0, \beta > 0$, 即密度函数为

$$f(x;\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

求α, β的矩估计.

解: 由Γ分布的性质, $EX = \frac{\alpha}{\beta}, DX = \frac{\alpha}{\beta^2}$, 因此可得方程组

$$\begin{cases} \overline{X} = \frac{\alpha}{\beta}; \\ \tilde{S}^2 = \frac{\alpha}{\beta^2}. \end{cases}$$

得

$$\begin{cases} \hat{\alpha} = \frac{\overline{X}^2}{\tilde{S}^2}; \\ \hat{\beta} = \frac{\overline{X}}{\tilde{S}^2}. \end{cases}$$

第一节 矩估计和极大似然估计

注：(1)矩估计直观简便

(2)要求总体的原点矩存在，若不存在则不能用，如柯西分布

(3)没有充分利用$F(x;\theta)$对θ所提供的信息

第一节 矩估计和极大似然估计

二.极大似然估计

(一)原理： 极大似然估计是建立在极大似然原理基础之上。

例1.设一箱子中有4个球，黑白两种，数目之比为1:3，但不知哪个多，现从中有放回的任取3个，发现有2个白球，问白球所占的比例？

解：设白球所占的比例为p，则$p=\frac{1}{4}$或$\frac{3}{4}$,若X表示任取3球所含白球个数，则$X\sim b(3,p)$,所以$P(X=2)=C_3^2p^2(1-p)$

当$p=\frac{1}{4}$时，$P(X=2)=\frac{9}{64}$,

当$p=\frac{3}{4}$时，$P(X=2)=\frac{27}{64}$. 因为$\frac{27}{64}>\frac{9}{64}$,意味$X=2$的样本来自$p=\frac{3}{4}$的总体比来自$p=\frac{1}{4}$的总体的可能性大,所以我们认为白球所占的比例为$\frac{3}{4}$。

注： 概率最大的事件最可能发生

第一节 矩估计和极大似然估计

(二)定义: 设$(X_1, \cdots, X_n)$为取自具有概率分布族$\{f(x;\theta) : \theta \in \Theta\}$的离散型总体$X$的一个样本,其中$\theta = (\theta_1, \cdots, \theta_k)$是未知的$k$维参数向量, $(X_1, \cdots, X_n)$取观测值$(x_1, \cdots, x_n)$的概率为

$$L(x_1, \cdots, x_n; \theta) \triangleq \prod_{i=1}^{n} f(x_i; \theta),$$

$L(x_1, \cdots, x_n; \theta)$称为$\theta$的**似然函数**(Likelihood Function).若$\hat{\theta}$使

$$L(x_1, \cdots, x_n; \hat{\theta}) = \sup_{\theta \in \Theta} L(x_1, \cdots, x_n; \theta)$$

并以$\hat{\theta}$作为参数θ的估计值,则称$\hat{\theta}$为θ的极大似然估计值,其相应的统计量$\hat{\theta}(X_1, \cdots, X_n)$称为$\theta$的**极大似然估计量**.简记为MLE或ML估计.

第一节 矩估计和极大似然估计

(三)求极大似然估计的方法

(1)先求似然函数$L(x_1,\cdots,x_n;\theta)\triangleq\prod_{i=1}^{n}f(x_i;\theta)\triangleq L(\theta)$

(2)因为$L(\theta)$与$\ln L(\theta)$有相同的极大值点，所以$\ln L(\hat{\theta})=\sup_{\theta\in\Theta}\ln L(\hat{\theta})$,若$L$ 可微, $\ln L(\theta)$关于$\theta_1,\cdots,\theta_k$分别求导数，并令其等于0，得$\frac{\partial \ln L(\theta)}{\partial\theta_i}=0,i=1,\cdots,k$，称为似然方程组; 求解方程组，得$\hat{\theta}$并证明$\hat{\theta}$使$L(\theta)$达到最大, $\hat{\theta}$即为θ的极大似然估计;
若L不可微, 则用其他方法求出极大似然估计值.

第一节 矩估计和极大似然估计

例2: 设X的密度函数为$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}$,求参数$\lambda$的极大似然估计。

解: 设$(x_1,\cdots,x_n)$是样本$(X_1,\cdots,X_n)$的一组观测值,似然函数为

$$L(x_1,\cdots,x_n;\lambda)=\prod_{i=1}^{n}f(x_i;\lambda)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}$$

则$\ln L(x_1,\cdots,x_n;\lambda)=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i$
令$\frac{\partial \ln L}{\partial\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i=0$
经验证, $\hat{\lambda}=\frac{1}{\overline{X}}$是$\lambda$的极大似然估计。

第一节 矩估计和极大似然估计

例3:设总体X具有均匀分布,密度函数为

$$f(x,\theta)=\begin{cases}\frac{1}{\theta}, & 0\le x\le\theta;\\ 0, & \text{其它},\end{cases}$$

其中$\theta>0$是未知参数,求θ的极大似然估计.

解: 设$(x_1,\cdots,x_n)$是样本$(X_1,\cdots,X_n)$的一组观测值,似然函数为

$$L(\theta)=\begin{cases}\frac{1}{\theta^n}, & 0\le x_i\le\theta;\\ 0, & \text{其它},\end{cases}$$

要使L最大，必须使θ最小，当$\theta=\max\limits_{1\le i\le n}x_i=x_{(n)}$时，可使$L$最大，故$\hat{\theta}=X_{(n)}$为参数$\theta$的极大似然估计量.

第一节 矩估计和极大似然估计

例4: 设总体$X \sim N(\mu, \sigma^2)$,其中$\theta = (\mu, \sigma^2)$为未知参数向量,参数空间$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$, 求$\mu, \sigma^2$的极大似然估计.

解: 设$(x_1, \cdots, x_n)$是样本$(X_1, \cdots, X_n)$的一组观测值,于是似然函数为

$$L(\mu, \sigma^2) = (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\},$$

两边取对数得

$$l(\mu, \sigma^2) = \ln L(\mu, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

分别求上式关于μ和σ^2的偏导数,并令它们为0,得似然方程组

$$\begin{cases} \frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0; \\ \frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0, \end{cases}$$

第一节 矩估计和极大似然估计

解之得

$$\begin{cases} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}; \\ \hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \tilde{s}_n^2. \end{cases}$$

容易验证$\hat{\mu}, \hat{\sigma}^2$满足关系式

$$L(\hat{\mu}, \hat{\sigma}^2) = \sup_{-\infty<\mu<\infty, \sigma^2>0} L(\mu, \sigma^2),$$

所以$\overline{X}$和$\tilde{S}_n^2$分别是μ和σ^2的极大似然估计.

第一节 矩估计和极大似然估计

注：(1)极大似然估计充分利用了总体分布所提供的信息，比矩法估计优，特别是对大样本的情况。

(2)必须知道总体的分布，且有时不易求出极大似然方程组的解。

(3)若函数$g(\theta)$具有单值反函数，且$\hat{\theta}$是θ的MLE，则$g(\hat{\theta})$是$g(\theta)$的MLE.

第二节　估计量的优良性准则

第二节　估计量的优良性准则

第二节 估计量的优良性准则

一、无偏估计

1.定义： 设总体X具有分布族$\{F(x;\theta):\theta\in\Theta\}$, $(X_1,\cdots,X_n)$是取自这个总体的一个样本, $\hat{\theta}=\hat{\theta}(X_1,\cdots,X_n)$是未知参数$\theta$的一个估计量,如果对于一切$\theta\in\Theta$,都有

$$E_\theta[\hat{\theta}(X_1,\cdots,X_n)]=\theta$$

则称$\hat{\theta}(X_1,\cdots,X_n)$为$\theta$的**无偏估计**,简记为UE.

第二节 估计量的优良性准则

例1: 设总体X的均值为μ,方差为σ^2,$(X_1,\cdots,X_n)$为取自总体的样本，则
(1)$E(\overline{X})=\mu$;
(2)总体X的k阶原点矩$m_k=E(X^k)$存在，则样本k阶原点矩A_k满足$E(A_k)=m_k$.

解: 因为$E(X_i)=E(X)=\mu, D(X_i)=D(X)=\sigma^2, i=1,\cdots,n$,且$X_1,\cdots,X_n$相互独立，则
$E(\overline{X})=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$,所以$\overline{X}$是$\mu$的无偏估计。
$E(A_k)=E(\frac{1}{n}\sum_{i=1}^{n}X_i^k)=\frac{1}{n}\sum_{i=1}^{n}E(X_i^k)=m_k$,所以$A_k$是$m_k$的无偏估计。

第二节 估计量的优良性准则

2.若$E(\hat{\theta}) \neq \theta$,则称它是有偏的，且称函数$b(\theta, \hat{\theta}) = E(\hat{\theta}) - \theta$为$\hat{\theta}$的偏.

3.若有一列θ的估计$\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$,对一切$\theta \in \Theta$满足$\lim\limits_{n\to\infty} E_\theta(\hat{\theta}_n) = \theta$，即$\lim\limits_{n\to\infty} b_n = 0$,则称$\hat{\theta}_n$为$\theta$的渐近无偏估计量。

第二节 估计量的优良性准则

例2：(续)
验证$S^2=\frac{1}{n-1}\sum_{i=1}(X_i-\overline{X})^2$，$\tilde{S}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是否为σ^2的无偏估计。

解：由Th1.2知$E(S^2)=\sigma^2$,所以S^2是σ^2的无偏估计。
$E(\tilde{S}^2)=E(\frac{n-1}{n}S^2)=\frac{n-1}{n}\sigma^2\neq\sigma^2$，所以$\tilde{S}^2$不是$\sigma^2$的无偏估计，偏为$E(\tilde{S}^2)-\sigma^2=(\frac{n-1}{n}-1)\sigma^2=-\frac{\sigma^2}{n}$。
但$\lim\limits_{n\to\infty}E(\tilde{S}^2)=\sigma^2$，所以$\tilde{S}^2$是$\sigma^2$的渐近无偏估计.

第二节 估计量的优良性准则

4.若对θ的任一实值函数$g(\theta)$,如果存在估计量$T(X_1,\cdots,X_n)$, 使得对一切$\theta\in\Theta$, 有$E_\theta(T)=g(\theta)$,则$g(\theta)$称为可估计函数.

第二节 估计量的优良性准则

例3： 设$(X_1,\cdots,X_n)$是从正态总体$N(\mu,\sigma^2)$中抽取的一个样本,由定理知$\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$, $(n-1)S^2/\sigma^2\sim\chi^2(n-1)$. 虽然$\overline{X}$是$\mu$的一个无偏估计,但$\overline{X}^2$不是$\mu^2$的无偏估计,

$$E_\mu(\overline{X}^2)=D\overline{X}+(E\overline{X})^2=\frac{\sigma^2}{n}+\mu^2\neq\mu^2.$$

同样,虽然S^2是σ^2的一个无偏估计,但S也不是σ的无偏估计.事实上

$$\begin{aligned}E_\sigma(\sqrt{n-1}S/\sigma) &= \int_0^\infty\sqrt{x}\cdot\frac{x^{\frac{n-1}{2}-1}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}e^{-\frac{x}{2}}dx\\ &= \frac{1}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}2^{\frac{n}{2}}\Gamma(\frac{n}{2})\\ &= \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})},\end{aligned}$$

即$E_\sigma(S)=\sigma[\sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}]\neq\sigma.$

第二节 估计量的优良性准则

可见,S不是σ的无偏估计,它的偏为

$$b(\sigma, S) = \sigma[\sqrt{\frac{2}{n-1}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} - 1].$$

但容易由这个有偏估计修改得到σ的无偏估计

$$\hat{\sigma} = \sqrt{\frac{n-1}{2}}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}S.$$

第二节 估计量的优良性准则

注：(1)若$\hat{\theta}$是θ的有偏估计，且$E(\hat{\theta}) = a + b\theta$,(a,$b \neq 0$为常数),则可以构造一个$\theta$的无偏估计$\hat{\theta}^* = \frac{\hat{\theta}-a}{b}$。

(2)若$\hat{\theta}$是θ的无偏估计，除了f是线性函数外，并不能推出$f(\hat{\theta})$是$f(\theta)$的无偏估计。

第二节 估计量的优良性准则

例4. 设总体X服从均匀分布,密度函数为

$$f(x;\theta)=\begin{cases}\frac{1}{\theta}, & 0<x<\theta;\\ 0, & \text{其它},\end{cases}$$

$(X_1,\cdots,X_n)$是取自这个总体的一个样本,
(1)求θ的矩估计, 验证无偏性。
(2)求θ的MLE, 验证无偏性。

解: 由于$EX=\frac{\theta}{2}$,故$\hat{\theta}=2\overline{X}$是$\theta$的矩估计,且它是无偏的.
由前知$\hat{\theta}_L=X_{(n)}$是θ的极大似然估计.
$E_\theta(\hat{\theta}_L)=\int_0^\theta x\cdot\frac{n}{\theta}(\frac{x}{\theta})^{n-1}dx=\frac{n}{n+1}\theta$, 即$\hat{\theta}_L$不是$\theta$的无偏估计,但$\hat{\theta}_L$是$\theta$的渐近无偏估计.
由$\hat{\theta}_L=X_{(n)}$可以构造θ的一个无偏估计量$\hat{\theta}^*=\frac{n+1}{n}X_{(n)}$.

第二节 估计量的优良性准则

考虑θ的这两个无偏估计量$\hat{\theta}$与$\hat{\theta}^*$的方差,

$$\begin{aligned} D_\theta(\hat{\theta}) &= D_\theta(2\overline{X}) = 4D_\theta(\overline{X}) = \frac{4}{n}D_\theta(X) = \frac{\theta^2}{3n}; \\ D_\theta(\hat{\theta}_L) &= \int_0^\theta x^2 \cdot n\frac{x^{n-1}}{\theta^n}dx - (\frac{n}{n+1}\theta)^2 \\ &= \frac{n\theta^2}{(n+1)^2(n+2)}, \end{aligned}$$

所以

$$D_\theta(\hat{\theta}^*) = D_\theta(\frac{n+1}{n}\hat{\theta}_L) = \frac{(n+1)^2}{n^2}D_\theta(\hat{\theta}_L) = \frac{\theta^2}{n(n+2)}.$$

显然,$D_\theta(\hat{\theta}^*) \le D_\theta(\hat{\theta})$,而且当$n$很大时,有$\lim_{n\to\infty}\frac{D_\theta(\hat{\theta}^*)}{D_\theta(\hat{\theta})} = \lim_{n\to\infty}\frac{3}{n+2} = 0$. 可见$\hat{\theta}^*$和$\hat{\theta}$的取值都在参数真值$\theta$的周围波动,但$\hat{\theta}^*$比$\hat{\theta}$取值更集中,作为$\theta$的估计量,$\hat{\theta}^*$比$\hat{\theta}$好.

第二节 估计量的优良性准则

二、一致最小方差无偏估计

1.定义：设$T_1(X_1,\cdots,X_n)$为可估函数$g(\theta)$的无偏估计量,若对于任意的$\theta\in\Theta$和任意的$g(\theta)$的无偏估计量$T(X_1,\cdots,X_n)$,都有

$$D_\theta[T_1(X_1,\cdots,X_n)]\le D_\theta[T(X_1,\cdots,X_n)],$$

则称$T_1(X_1,\cdots,X_n)$是$g(\theta)$的 **一致最小方差无偏估计量**,简记为UMVUE.

记$U\triangleq\{T:E_\theta(T)=g(\theta),D_\theta(T)<\infty,\ 对一切\ \theta\in\Theta\}$, 为可估函数$g(\theta)$的方差有限的无偏估计量的集合.

$U_0\triangleq\{T:E_\theta(T)=0,D_\theta(T)<\infty,\ 对一切\ \theta\in\Theta\}$, 为数学期望为零,方差有限的估计量的集合.

第二节 估计量的优良性准则

2.定理: 设$T_1 \in U$,则T_1是$g(\theta)$的一致最小方差无偏估计的充要条件为:对一切$\theta \in \Theta$和$T_0 \in U_0$,有$E_\theta(T_1 T_0) = 0$.

3.推论: 设T_1, T_2分别是参数θ的可估函数$g_1(\theta), g_2(\theta)$的一致最小方差无偏估计量, 则$b_1 T_1 + b_2 T_2$是$b_1 g_1(\theta) + b_2 g_2(\theta)$的一致最小方差无偏估计量, 其中$b_1, b_2$为固定常数.

4.定理: $U \triangleq \{T : E_\theta(T) = g(\theta), \mathrm{Var}_\theta(T) < \infty, \forall \theta \in \Theta\}$, 则至多存在一个$g(\theta)$的一致最小方差无偏估计量.

第二节 估计量的优良性准则

三、相合估计量

1.定义： 设$T_n = T_n(X_1, \cdots, X_n)$是$g(\theta)$的估计量，若对任何$\theta \in \Theta$，$T_n$依概率收敛于$g(\theta)$,则称$T_n$是$g(\theta)$的**相合估计**。

注： (1)另一种表述：$\forall \varepsilon > 0, \lim_{n\to\infty} P\{|T_n - g(\theta)| \geq \varepsilon\} = 0.$

(2)相合性是在极限意义下引入的，适用大样本情况。

(3)若T_n以概率1(几乎处处)收敛于$g(\theta)$,
即$P_\theta\{\lim_{n\to\infty} T_n(X_1, \cdots, X_n) = g(\theta)\} = 1$，则称$T_n$是$g(\theta)$的**强相合估计**。

(4)若T_n是$g(\theta)$的强相合估计,它也是$g(\theta)$的相合估计.

第二节 估计量的优良性准则

2.证明相合估计的方法

(1)定义，利用切比雪夫不等式$P\{|X - EX| \geq \varepsilon\} \leq \frac{DX}{\varepsilon^2}$.

(2)定理，设T_n是$g(\theta)$的一个估计量，若$\lim_{n\to\infty} ET_n = g(\theta)$, $\lim_{n\to\infty} DT_n = 0$,则$T_n$是$g(\theta)$的相合估计。

第二节 估计量的优良性准则

例5： 设总体X服从均匀分布,密度函数为

$$f(x;\theta)=\begin{cases}\frac{1}{\theta}, & 0<x<\theta;\\ 0, & \text{其它},\end{cases}$$

$(X_1,\cdots,X_n)$是取自这个总体的一个样本，证明：$\hat{\theta}=X_{(n)}$是θ的相合估计。

解：$\hat{\theta}=X_{(n)}$的密度函数为

$$f_n(x;\theta)=\begin{cases}n(\frac{x}{\theta})^{n-1}\frac{1}{\theta}=\frac{n}{\theta^n}x^{n-1}, & 0<x<\theta;\\ 0, & \text{其它},\end{cases}$$

$$E(\hat{\theta})=\int_0^\theta x\cdot\frac{n}{\theta^n}x^{n-1}dx=\frac{n}{n+1}\theta,$$
$$E(\hat{\theta}^2)=\int_0^\theta x^2\cdot\frac{n}{\theta^n}x^{n-1}dx=\frac{n}{n+2}\theta^2,$$
$$D(\hat{\theta})=E(\hat{\theta}^2)-[E(\hat{\theta})]^2=\frac{n}{n+2}\theta^2-[\frac{n}{n+1}\theta]^2=\frac{n}{(n+1)^2(n+2)}\theta^2$$

第二节 估计量的优良性准则

(方法1)

因为$\lim_{n\to\infty} E(\hat{\theta}) = \theta$, $\lim\limits_{n\to\infty} D(\hat{\theta}) = 0$ 所以由定理$\hat{\theta} = X_{(n)}$是θ的相合估计。

(方法2)

$$
\begin{aligned}
P\{|\hat{\theta}-\theta| \geq \varepsilon\} &= P\{|\hat{\theta} - \tfrac{n}{n+1}\theta - \tfrac{1}{n+1}\theta| \geq \varepsilon\} \\
&\leq P\{|\hat{\theta} - \tfrac{n}{n+1}\theta| + \tfrac{1}{n+1}\theta \geq \varepsilon\} \\
&= P\{|\hat{\theta} - \tfrac{n}{n+1}\theta| \geq \varepsilon - \tfrac{1}{n+1}\theta\} \\
&= P\{|\hat{\theta} - E(\hat{\theta})| \geq \varepsilon - \tfrac{1}{n+1}\theta\} \\
&\leq \frac{D(\hat{\theta})}{(\varepsilon - \frac{1}{n+1}\theta)^2} \\
&= \frac{n\theta^2}{(n+1)^2(n+2)(\varepsilon - \frac{1}{n+1}\theta)^2} \\
&\longrightarrow 0, n \to \infty
\end{aligned}
$$

第二节 估计量的优良性准则

3.定理: 设$(X_1, \cdots, X_n)$是取自具有分布族$\{F(x;\theta) : \theta \in \Theta\}$的总体$X$的一个样本, 若$E|X|^p < \infty$, 其中$p$是某一正整数,则样本的$k$阶原点矩$(1 \le k \le p)$ $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$是总体$k$阶矩$\alpha_k = E_\theta(X^k)$的相合估计.

4.定理: 如果T_n是θ的相合估计量,$g(x)$在$x = \theta$连续,则$g(T_n)$也是$g(\theta)$的相合估计量.

第三节　Rao–Cramer 不等式

第三节　Rao–Cramer 不等式

第三节 Rao–Cramer 不等式

一. Rao–Cramer 不等式

1.定义: 若单参数分布密度族(或单参数概率分布族)$\{f(x;\theta):\theta\in\Theta\}$满足如下条件:
(1)参数空间Θ是数轴上的一个开区间;
(2)导数$\frac{\partial f(x;\theta)}{\partial\theta}$对一切$\theta\in\Theta$都存在;
(3)集合$S_\theta\overset{\Delta}{=}\{x:f(x;\theta)>0\}$与$\theta$无关($S_\theta$称为支撑);
(4) $\frac{\partial}{\partial\theta}\int f(x;\theta)dx=\int\frac{\partial f(x;\theta)}{\partial\theta}dx=0$,对一切$\theta\in\Theta$;
(5)对任$\theta\in\Theta$,有

$$\begin{aligned}0<I(\theta)&\overset{\Delta}{=}E_\theta[\frac{\partial}{\partial\theta}\ln f(X;\theta)]^2\\&=\int(\frac{\partial\ln f(x;\theta)}{\partial\theta})^2f(x;\theta)dx,\end{aligned}$$

则称这个分布密度族(概率分布族)为Rao–Cramer **正则分布族**,其中条件(1)–(5) 称为**正则条件**, $I(\theta)$称为该分布密度族的**Fisher**信息函数.

第三节 Rao–Cramer 不等式

例1： 验证Poisson分布族 $\{f(x;\lambda)=\frac{\lambda^x}{x!}e^{-\lambda}, x=0,1,2,\cdots:\ \lambda>0\}$ 是Rao–Cramer正则分布族.

解：(1)参数空间$\{\lambda>0\}$是一开区间;

(2) $\frac{\partial f(x;\theta)}{\partial\lambda}=\frac{\lambda^{x-1}}{x!}(x-\lambda)e^{-\lambda}$,对一切$\lambda>0$存在;

(3) $f(x;\lambda)>0$,对一切非负整数x及λ成立;

(4)

$$\sum_{x=0}^{\infty}\frac{\lambda^{x-1}}{x!}(x-\lambda)e^{-\lambda}=\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}e^{-\lambda}-\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}e^{-\lambda}=1-1=0;$$

(5) $\ln f(x;\lambda)=x\log\lambda-\ln x!-\lambda$, $\frac{\partial\log f(x;\lambda)}{\partial\lambda}=\frac{x}{\lambda}-1$, 因此

$$E_\lambda[\frac{\partial\ln f(X;\lambda)}{\partial\lambda}]=0, I(\lambda)=E_\lambda[\frac{\partial\ln f(X;\lambda)}{\partial\lambda}]^2=D_\lambda(\frac{X}{\lambda}-1)=\frac{1}{\lambda},$$

故Poisson分布族是Rao–Cramer正则分布族.

第三节 Rao–Cramer 不等式

例2： 正态分布族$\{N(\mu,1):-\infty<\mu<\infty\}$是Rao–Cramer正则分布族.

解:对所考虑的正态分布密度

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

条件(1)–(4)是显然满足的,只需验证条件(5)也满足.事实上,由于

$$\frac{\partial \ln f(x;\mu)}{\partial \mu}=x-\mu,$$

故

$$E_\mu[\frac{\partial \ln f(X;\mu)}{\partial \mu}]^2=E_\mu(X-\mu)^2=D_\mu(X)=1.$$

所以正态分布族$\{N(\mu,1):-\infty<\mu<\infty\}$是Rao–Cramer正则分布族.

第三节 Rao–Cramer 不等式

注：

(1)常用的单参数分布族都是Rao–Cramer正则分布族,但均匀分布族$\{U(0,\theta):\theta>0\}$ 不是Rao–Cramer正则分布族, 因为它的支撑$S_\theta=\{x:f(x;\theta)>0\}$是一个依赖于$\theta$的开区间.

(2)计算$I(\theta)$的一个简便方法：$I(\theta)=-E_\theta\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right)$

第三节 Rao–Cramer 不等式

2.定理：(Rao–Cramer不等式) 设总体X的概率分布密度族$\{f(x;\theta):\theta\in\Theta\}$是Rao–Cramer正则分布族,可估函数$g(\theta)$是$\Theta$上的可微函数,$(X_1,\cdots,X_n)$是取自总体$X$的一个样本.若$T(X_1,\cdots,X_n)$是$g(\theta)$的一个无偏估计量,且对一切$\theta\in\Theta$,满足

$$\begin{aligned}&\frac{\partial}{\partial\theta}\int T(x_1,\cdots,x_n)L(x_1,\cdots,x_n;\theta)dx_1\cdots dx_n\\&=\int T(x_1,\cdots,x_n)\frac{\partial}{\partial\theta}L(x_1,\cdots,x_n;\theta)dx_1\cdots dx_n,\end{aligned}\quad(*)$$

其中$L(x_1,\cdots,x_n;\theta)=\prod_{i=1}^n f(x_i;\theta)$是样本$(X_1,\cdots,X_n)$的联合分布密度函数. 则对一切$\theta\in\Theta$,有

$$D_\theta(T)\geq\frac{[g'(\theta)]^2}{nI(\theta)}\quad(1)$$

其中$I(\theta)$是该分布族的Fisher信息函数.

第三节 Rao–Cramer 不等式

特别当$g(\theta)=\theta$时, $D_\theta(T)\geq\frac{1}{nI(\theta)}$.
如果(1)中等号成立, 则$\exists C(\theta)\neq 0$, s.t.

$$\frac{\partial \ln L(x_1,\cdots,x_n;\theta)}{\partial\theta}=C(\theta)[T(x_1,\ldots,x_n)-g(\theta)]a.s.$$

注： (1)称满足上述定理中正则条件的估计量T为正规估计量。

(2)称Rao—Cramer不等式所规定的下界为**C–R**下界.

(3)若一个无偏估计T的方差达到这个下界,且$g(\theta)$的一切无偏估计都满足条件(*),则T就是$g(\theta)$的UMVUE.

(4)离散性总体：密度函数$\to$概率分布, 积分$\to$求和

第三节 Rao–Cramer 不等式

(5)$nI(\theta)$越大, $\mathrm{Var}_\theta(T)$可达下界越低, 即$g(\theta)$可能被估计得越准确. 所以$I(\theta)$在一定意义上表示对X进行观测所提供的关于未知参数θ的平均信息.

(6)C-R不等式只对C-R正则族成立. 如定理中条件之一不满足, C-R不等式可能不成立.

第三节 Rao–Cramer 不等式

例: 设$(X_1, ..., X_n)$是总体X的一个样本, X的密度函数为

$$f(x;\theta) = \mathrm{e}^{-(x-\theta)}, x > \theta.$$

因为$\{x : f(x;\theta) > 0\} = \{x > \theta\}$, 所以不是C-R正则族.
形式上计算C-R下界:

$$I(\theta) = E\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2 = E(1) = 1, \quad \text{下界}\frac{1}{n}$$

$$\hat{\theta} = X_{(1)} - \frac{1}{n}, \quad f(t;\theta) = n\mathrm{e}^{-(nt-n\theta+1)}, \quad t > \theta - \frac{1}{n}.$$

$$E(\hat{\theta}) = \theta, \quad E(\hat{\theta}^2) = \theta^2 + \frac{1}{n^2}, \quad D(\hat{\theta}) = \frac{1}{n^2} < \frac{1}{n}$$

第三节 Rao–Cramer 不等式

例3:设总体X具有Bernoulli分布族$\{b(1,p):\ p\in(0,1)\}$, $(X_1,\cdots,X_n)$是取自这一总体的样本,求p的正规无偏估计的方差下界.

解: 容易验证Bernoulli分布族是Rao–Cramer正则分布族,有

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

且

$$I(p)=\sum_{x=0}^{1}[\frac{\partial \ln f(x;p)}{\partial p}]^2 f(x;p)=\frac{1}{p(1-p)},$$

故p的正规无偏估计的方差下界为

$$\frac{1}{nI(p)}=\frac{p(1-p)}{n}.$$

$\hat{p}=\overline{X}$是p的无偏估计,而$D(\overline{X})=\frac{p(1-p)}{n}$达到Rao–Cramer不等式的下界.

第三节 Rao–Cramer 不等式

例4： 设X的密度函数为$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x}, & x>0;\\ 0, & x\le 0,\end{cases}$，$(X_1,\cdots,X_n)$ 是取自总体X的样本,未知参数$\lambda>0$, 求待估函数$g(\lambda)=\frac{1}{\lambda}$的正规无偏估计量的C–R下界.

解：指数分布族是Rao–Cramer正则分布族，且当$x>0$时，

$$\ln f(x;\lambda)=-\lambda x+\ln\lambda$$

$$\frac{\partial^2 \ln f(x;\lambda)}{\partial\lambda^2}=\frac{\partial^2(-\lambda x+\ln\lambda)}{\partial\lambda^2}=\frac{\partial}{\partial\lambda}(-x+\frac{1}{\lambda})=-\frac{1}{\lambda^2}$$

所以$I(\lambda)=-E(\frac{\partial^2 \ln f(x;\lambda)}{\partial\lambda^2})=-E(-\frac{1}{\lambda^2})=\frac{1}{\lambda^2}$

C-R下界: $\frac{[g'(\lambda)]^2}{nI(\lambda)}=\frac{1}{n\lambda^2}$.

$E(\overline{X})=E(X)=\frac{1}{\lambda}=u(\lambda)$, $D(\overline{X})=\frac{1}{n}D(X)=\frac{1}{n}\frac{1}{\lambda^2}=\frac{1}{n\lambda^2}$,

$\overline{X}$的方差达到C-R下界.

第三节 Rao–Cramer 不等式

二.有效估计量

1.定义： 设总体X的概率分布密度族$\{f(x;\theta):\theta\in\Theta\}$是Rao–Cramer正则分布族,$g(\theta)$是可估函数,$T=T(X_1,\cdots,X_n)$是$g(\theta)$的估计量.

(1)若$E_\theta(T)=g(\theta)$,则称

$$e_\theta(T)=\frac{[g'(\theta)]^2}{nI(\theta)}/D_\theta(T)$$

为$g(\theta)$的无偏估计量T的**效率**.

(2)若$E_\theta(T)=g(\theta)$,且$e_\theta(T)=1$,则称T为$g(\theta)$的**有效估计量**.

(3)设$T_n=T_n(X_1,\cdots,X_n)$是$g(\theta)$的一列无偏估计,若

$$\lim_{n\to\infty}e_\theta(T_n)=e_0,$$

则称e_0为T_n的**渐近效率**;当$e_0=1$时,称T_n是$g(\theta)$的**渐近有效估计**.

第三节 Rao–Cramer 不等式

例5： 设X服从$\Gamma(r, \frac{1}{\lambda})$,即

$$f(x;\theta) = \begin{cases} \frac{1}{\lambda^r \Gamma(r)} x^{r-1} e^{-\frac{x}{\lambda}}, x > 0; \\ 0, x \leq 0, \end{cases}$$

其中$r > 0$已知,$\lambda > 0$未知，$(X_1, \cdots, X_n)$为X的样本，证明：$\frac{1}{nr}\sum_{i=1}^{n} X_i$ 是λ的有效估计。

证： 因为$E(\frac{1}{nr}\sum_{i=1}^{n} X_i) = \frac{1}{nr} nr\lambda = \lambda$,所以$\frac{1}{nr}\sum_{i=1}^{n} X_i$是$\lambda$的无偏估计。

验证条件，它是Rao-Cramer正则分布族，$\ln f(x,\lambda) = -r\ln\lambda - \ln\Gamma(r) + (r-1)\ln x - \frac{x}{\lambda}$, 则

$$\frac{\partial \ln f(x;\lambda)}{\partial \lambda} = -\frac{r}{\lambda} + \frac{x}{\lambda^2}, \frac{\partial^2 \ln f(x;\lambda)}{\partial \lambda^2} = \frac{r}{\lambda^2} - \frac{2x}{\lambda^3}$$

第三节 Rao–Cramer 不等式

所以

$$I(\lambda) = -E(\frac{r}{\lambda^2} - \frac{2X}{\lambda^3}) = -\frac{r}{\lambda^2} + \frac{2\lambda r}{\lambda^3} = \frac{r}{\lambda^2}$$

所以C-R下界为$\frac{1}{nI(\lambda)} = \frac{1}{n\frac{r}{\lambda^2}} = \frac{\lambda^2}{nr}$

而$D(\frac{1}{nr}\sum_{i=1}^{n} X_i) = \frac{1}{n^2r^2}\sum_{i=1}^{n} D(X_i) = \frac{nr\lambda^2}{n^2r^2} = \frac{\lambda^2}{nr}$达到C-R下界，
所以$\frac{1}{nr}\sum_{i=1}^{n} X_i$ 是λ的有效估计。

第三节 Rao–Cramer 不等式

2.定义： 对可估函数$g(\theta)$的任意两个无偏估计量T_1和T_2, 称

$$e(T_1|T_2) = \frac{D_\theta(T_1)}{D_\theta(T_2)}$$

为估计量T_1关于T_2的**相对效率**;如果$e(T_1|T_2) < 1$,则称T_1比T_2有效.

注： 也就是若$D(T_1) < D(T_2)$,则T_1比T_2有效。

第四节 Rao–Blackwell定理

1.定理(Rao-Blackwell定理) 设$T(X_1,...,X_n)$是分布族$\{F(x;\theta):\theta\in\Theta\}$的一个充分统计量，$V$是$g(\theta)$的一个无偏估计量，则

$$V_0 = V_0(T) = E(V|T)$$

是$g(\theta)$的一个无偏估计量且$\mathrm{Var}_\theta(V_0)\le \mathrm{Var}_\theta(V), \forall\theta\in\Theta$. 其中等号成立的充要条件是$V$在a.s.意义下是$T$的函数.

第四节 Rao–Blackwell定理

注: (1) 定理提供了一种改善估计得方法，若已知一无偏估计量，则可构造一新的无偏估计量且其方差比原估计量的方差小;
(2) 一致最小方差无偏估计量(UMVUE)一定是充分统计量的函数，否则按定理的方法可构造比原估计量有更小方差的估计;
(3) 若$g(\theta)$的由充分统计量构造的无偏估计量是a.s.唯一的, 则此无偏估计量一定是一致最小方差无偏估计量.
(4) 由充分统计量构造的无偏估计量为充分无偏估计量.

第四节 Rao–Blackwell定理

2.定理 在Rao-Cramer正则条件下，$g(\theta)$的无偏估计量$T(X_1,...,X_n)$是有效估计量，则$T(X_1,...,X_n)$是充分统计量.

3.定理(Lehmann-Scheffe定理) 设$(X_1,...,X_n)$是取自分布族$\{F(x;\theta):\theta\in\Theta\}$的一个样本，$T(X_1,...,X_n)$是$\theta$的充分完备统计量. 如果一个只依赖于$T(X_1,...,X_n)$的估计量$h[T(X_1,...,X_n)]$为$g(\theta)$的无偏估计，则它是$g(\theta)$的唯一的一致最小方差无偏估计量.

第四节 Rao–Blackwell定理

例1： 设$(X_1, ..., X_n)$是取自两点分布族$\{b(1, p) : 0 < p < 1\}$的一个样本, 求p的一致最小方差无偏估计量.

例2： 设$(X_1, ..., X_n)$是取自均匀分布族$\{U(0, \theta) : 0 < \theta < 1\}$的一个样本, 求$\theta$的一致最小方差无偏估计量.

例3： 设$(X_1, ..., X_n)$是取自参数为λ的Poisson分布的一个样本, 其中$\lambda > 0$为未知参数, 求$P_\lambda(k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, k = 0, 1, 2...$的一致最小方差无偏估计量.

第三章　假设检验

第一节　基本概念

第一节 基本概念

一.关于总体的假设可以分为两大类

(一)设总体X的分布$F(x;\theta)$类型已知，其中参数θ未知，仅对未知参数θ提出假设，称这类假设为参数假设检验。

(二)设总体X的分布未知，只对未知分布函数类型或它的某些特征提出假设，称这类假设为非参数假设检验。

第一节 基本概念

例1：某油品公司的桶装润滑油标定重量为10Kg，检验部门从市场上随机抽取10桶，称得重量分别为10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.2, 9.8, 假设每桶油实际重量$X \sim N(\mu, \sigma^2)$, 其中$\sigma = 0.1$,问该公司的桶装油的重量是否符合标准？

例2：设从某书中抽查100页，记录各页中印刷错误的个数，能否认为一页印刷错误的个数X服从泊松分布？

假设检验：依据从总体中取得的样本观测值来判断假设是否成立的一种程序.

原假设(零假设)：根据检验结果准备予以拒绝或不予拒绝(予以接收)的假设, 以H_0表示.

备择假设(对立假设)：与原假设不相容的假设, 以H_1表示.

第一节 基本概念

二.假设检验思想

1. 从具体问题出发，构造适当的统计量$T(X_1,\ldots,X_n)$，将样本$(X_1,\ldots,X_n)$中包含的有关信息集中在一起.

2. 检验：制定一判断规则，使根据样本观测值$(x_1,\ldots,x_n)$可做出是否拒绝原假设H_0的决定，每个这样的规则就是一种检验.

W：拒绝域；$\overline{W}$：接受域；检验函数：拒绝域W的示性函数.

$$\phi(x_1,\ldots,x_n)=\begin{cases}1, & (x_1,\ldots,x_n)\in W\\ 0, & (x_1,\ldots,x_n)\in \overline{W}\end{cases}$$

$$W=\{(x_1,\ldots,x_n):\phi(x_1,\ldots,x_n)=1,(x_1,\ldots,x_n)\in \chi\}$$

第一节 基本概念

3. 给定统计量T后，确定临界值c.

(1) 原则：一次观测中小概率事件发生的可能性很小.

例: 设总体$X \sim N(\mu, 1)$, 其中μ为未知参数, 考虑假设检验问题:

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0.$$

抽取样本$(X_1, \ldots, X_{10})$, 令$T = \overline{X}$作为检验统计量, 设由样本观测值可得$\bar{x} = 0.8$, 判断是否拒绝H_0.

第一节 基本概念

(2) 显著性水平：对每次检验，确定一个值α, 当一事件发生的概率小于此α时，就认为是一个小概率事件，称此α为检验的显著性水平.

对上例，给定显著性水平α, 在原假设成立的条件下，$\overline{X} \sim N(0, \frac{1}{n})$,

$$P(|\overline{X}| \geq c) = \alpha, \quad P(\sqrt{n}\overline{X} \geq \sqrt{n}c) = \frac{\alpha}{2},$$

$$\therefore P(\sqrt{n}\overline{X} < \sqrt{n}c) = 1 - \frac{\alpha}{2}, \quad \therefore c = \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

$$n = 10, \text{取}\alpha = 0.05, \Rightarrow c = 0.62,$$

当$T(x_1, ..., x_n) = \overline{X}(x_1, ..., x_n) \geq 0.62$时，拒绝原假设$H_0$.

注：接受原假设不是在逻辑上证明了原假设的正确，且由于样本的随机性，不意味着它一定是正确的假设.

第一节 基本概念

三.两类错误

1.原假设H_0正确时,由于样本的随机性而落入了拒绝域,因而作出拒绝H_0的错误判断,称它犯了**第一类错误**,记犯第一类错误的概率为α, (拒真错误)

$$P\{\text{拒绝}H_0|H_0\text{正确}\}=\alpha$$

2.对立假设H_1正确,由于样本的随机性而落入接受域,因此作出接受原假设H_0的错误判断,称它犯了**第二类错误**,记犯第二类错误的概率为β,(纳伪错误)

$$P\{\text{接受}H_0|H_1\text{正确}\}=\beta$$

注：犯第一类错误概率的最大值就是检验的显著水平.

第一节 基本概念

总体 / 样本	H_0正确	H_1正确
接受H_0	正确	第二类错误
拒绝H_0	第一类错误	正确

3.只考虑控制犯第一类错误的概率α,而不考虑犯第二类错误的概率,这样的检验称为显著性水平为α的显著性检验.

第一节 基本概念

4.显著性检验的一般步骤

(1) 提出假设：根据问题的要求建立原假设H_0和对立假设H_1;

(2) 选统计量：根据H_0的内容选取一个合适的检验统计量T,确定它的抽样分布,算出抽样分布的分位数;

(3) 给定显著水平：α的值一般取得较小,如0.05, 0.01, 0.10等;

(4) 确定拒绝域：在原假设H_0正确的条件下,求出能使

$$P\{(X_1, \cdots, X_n) \in W | H_0\} \leq \alpha$$

成立的拒绝域W,拒绝域W与所选的统计量T和假设有关,常用检验统计量T和相应的临界值c表示.

(5) 对H_0作出推断：比较检验统计量T的观测值t和相应的临界值c,看样本观测值是否落入拒绝域. 如果样本观测值$(x_1, \cdots, x_n) \in W$,则拒绝原假设H_0,否则接受H_0.

第二节　参数假设检验

第二节 参数假设检验

一.数学期望的检验

1. U检验： 设$X \sim N(\mu, \sigma_0^2)$,$(X_1, \cdots, X_n)$为取自总体的样本,其中$\sigma_0^2$已知，要检验

(1) $H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0$

当H_0成立时,$\overline{X} \sim N(\mu_0, \frac{\sigma_0^2}{n})$,则$\frac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n} \sim N(0,1)$.

取$U = \frac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n}$为检验统计量，对给定的显著性水平$\alpha$, 为使

$$P\{(X_1, \cdots, X_n) \in W | H_0\} = \alpha,$$

应有

$$P\{u_1 < U < u_2 | H_0\} = 1 - \alpha.$$

由于$U \sim N(0,1)$具有对称性,故可取

$$P\{|U| > u_{1-\frac{\alpha}{2}} | H_0\} = \alpha.$$

其中u_α是标准正态分布的α分位数.

第二节 参数假设检验

因此得到检验的拒绝域为

$$\begin{aligned} W &= \{(x_1, \cdots, x_n) : |u| = \frac{|\overline{x} - \mu_0|}{\sigma_0}\sqrt{n} > u_{1-\frac{\alpha}{2}}\} \\ &= \{u < u_{\frac{\alpha}{2}} \text{或 } u > u_{1-\frac{\alpha}{2}}\} \end{aligned}$$

计算U的观测值u,查表得到$u_{1-\frac{\alpha}{2}}$。
若$|u| > u_{1-\frac{\alpha}{2}}$,则拒绝原假设$H_0$,即认为总体的数学期望$\mu$与$\mu_0$之间有显著差异;
否则, 若$|u| \leq u_{1-\frac{\alpha}{2}}$, 则接受$H_0$, 即认为观测结果与原假设$H_0$无显著差异.

第二节 参数假设检验

(2) $H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu > \mu_0$

取$U = \frac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n}$为检验统计量，对给定的显著性水平$\alpha$, 由

$$P\{U > u_{1-\alpha}|H_0\} = \alpha.$$

查表得临界点$u_{1-\alpha}$, 故拒绝域为$W = \{u > u_{1-\alpha}\}$,然后计算u,
若$u > u_{1-\alpha}$,则拒绝H_0,否则接受H_0.

(3) $H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu < \mu_0$

取$U = \frac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n}$为检验统计量，对给定的显著性水平$\alpha$, 由

$$P\{U < u_{\alpha}|H_0\} = \alpha.$$

查表得临界点u_{α}, 故拒绝域为$W = \{u < u_{\alpha}\}$,然后计算u,
若$u < u_{\alpha}$,则拒绝H_0,否则接受H_0.

第二节　参数假设检验

注：

(a)这种检验为U检验；

(b) (1)为双侧检验，(2),(3)为单侧检验

第二节 参数假设检验

例1：全市高三学生毕业会考，数学成绩的平均分为70分，随机抽查10名男生的会考成绩如下，

$$65, 72, 89, 56, 79, 63, 92, 48, 75, 81,$$

若已知男生的会考成绩服从正态分布，且标准差为10分，问男生的会考平均成绩是否为70分？($\alpha = 0.05$)

解：设$H_0 : \mu = 70$, $H_1 : \mu \neq 70$

采用U检验，选取统计量$U = \frac{\overline{x}-\mu_0}{\sigma_0}\sqrt{n}$, 当$H_0$成立时, $U \sim N(0,1)$

对给定的显著性水平α,由$P\{|U| > u_{0.975}|H_0\} = 0.05$，查表得$u_{1-\frac{\alpha}{2}} = u_{0.975} = 1.96$

所以拒绝域为$W = \{|u| > 1.96\}$,

而$|u| = \frac{|\overline{x}-\mu_0|}{\sigma_0}\sqrt{n} = \frac{|72-70|}{10}\sqrt{10} = 0.63 < u_{0.975}$

所以接受H_0,即在显著性水平0.05下可以认为男生的会考平均成绩为70分。

第二节 参数假设检验

例1：全市高三学生毕业会考，数学成绩的平均分为70分，随机抽查10名男生的会考成绩如下，

$$65, 72, 89, 56, 79, 63, 92, 48, 75, 81,$$

若已知男生的会考成绩服从正态分布，且标准差为10分，问男生的会考平均成绩是否为70分？$(\alpha = 0.05)$

解：设$H_0 : \mu = 70,\ H_1 : \mu \neq 70$
采用U检验，选取统计量$U = \frac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n}$，当$H_0$成立时，$U \sim N(0,1)$
对给定的显著性水平α,由$P\{|U| > u_{0.975}|H_0\} = 0.05$，查表得$u_{1-\frac{\alpha}{2}} = u_{0.975} = 1.96$
所以拒绝域为$W = \{|u| > 1.96\}$,
而$|u| = \frac{|\overline{x}-\mu_0|}{\sigma_0}\sqrt{n} = \frac{|72-70|}{10}\sqrt{10} = 0.63 < u_{0.975}$
所以接受H_0,即在显著性水平0.05下可以认为男生的会考平均成绩为70分。

第二节 参数假设检验

例2： 某厂生产一种灯泡，其寿命$X \sim N(\mu, 200^2)$,从过去来看，灯泡的平均寿命为1500h,现采用新工艺，在所生产的灯泡中抽取25只，测得平均寿命为1675h,问采用新工艺后，灯泡寿命是否有显著提高？$(\alpha = 0.05)$

解：设$H_0 : \mu = 1500,\ H_1 : \mu > 1500$

采用U检验，选取统计量$U = \frac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n}$ 当H_0成立时，$U \sim N(0,1)$

对给定的显著性水平α,拒绝域为$W = \{u > u_{1-\alpha}\}$

查表得$u_{1-\alpha} = u_{0.95} = 1.645$

而$u = \frac{\overline{x}-\mu_0}{\sigma_0}\sqrt{n} = \frac{1675-1500}{200}\sqrt{25} = 4.375 > 1.645$

所以拒绝H_0,认为灯泡寿命有显著提高。

第二节 参数假设检验

例2： 某厂生产一种灯泡，其寿命$X \sim N(\mu, 200^2)$,从过去来看，灯泡的平均寿命为1500h,现采用新工艺，在所生产的灯泡中抽取25只，测得平均寿命为1675h,问采用新工艺后，灯泡寿命是否有显著提高？$(\alpha = 0.05)$

解： 设$H_0 : \mu = 1500,\ H_1 : \mu > 1500$
采用U检验，选取统计量$U = \frac{\overline{X}-\mu_0}{\sigma_0}\sqrt{n}$ 当H_0成立时，$U \sim N(0,1)$
对给定的显著性水平α,拒绝域为$W = \{u > u_{1-\alpha}\}$
查表得$u_{1-\alpha} = u_{0.95} = 1.645$
而$u = \frac{\overline{x}-\mu_0}{\sigma_0}\sqrt{n} = \frac{1675-1500}{200}\sqrt{25} = 4.375 > 1.645$
所以拒绝H_0,认为灯泡寿命有显著提高。

第二节 参数假设检验

2. t检验： 设$X \sim N(\mu, \sigma^2)$,$(X_1, \cdots, X_n)$为取自总体的样本,其中σ^2未知，要检验

(1)$H_0 : \mu = \mu_0 \longleftrightarrow H_1 : \mu \neq \mu_0$

选取统计量$T = \frac{\overline{X}-\mu_0}{S}\sqrt{n} \overset{H_0}{\sim} t(n-1)$（用样本方差代替总体方差）

对给定的显著性水平α, 由$P\{|T| > t_{1-\frac{\alpha}{2}}(n-1)|H_0\} = \alpha$查表求出T的双侧分位点, 得拒绝域$W = \{|t| > t_{1-\frac{\alpha}{2}}(n-1)\}$, 若由样本观测值计算$|t| > t_{1-\frac{\alpha}{2}}(n-1)$, 则拒绝$H_0$, 否则接受$H_0$.

第二节 参数假设检验

注：

(1)这种检验称为双侧t检验.

(2)单侧t检验

$H_0 : \mu = \mu_0,\ H_1 : \mu > \mu_0$,拒绝域为$W = \{t > t_{1-\alpha}(n-1)\}$

$H_0 : \mu = \mu_0,\ H_1 : \mu < \mu_0$,拒绝域为$W = \{t < t_{\alpha}(n-1)\}$

第二节 参数假设检验

例3：某种钢筋强度X服从正态分布，且$EX = 50(Kg/mm^2)$,今改变炼钢配方利用新法炼了9炉钢，从这9炉钢中每炉抽一根，测得强度分别为：

56.01，52.45，51.53，48.52，49.04，53.38，54.02，52.13，52.15，

问采用新法其钢筋强度均值是否有明显提高？($\alpha = 0.05$)

解：设$H_0 : \mu = 50,\ H_1 : \mu > 50$

总体方差未知，选取统计量$T = \frac{\overline{X}-\mu_0}{S}\sqrt{n}$.

对显著性水平$\alpha = 0.05$,查表得$t_{0.95}(8) = 1.86$, 因而拒绝域为$W = \{t > 1.86\}$,

由样本得$\bar{x} = 52.14, s = 2.24$,则$t = \frac{52.14-50}{2.24}\sqrt{9} = 2.87 > 1.86$

所以拒绝H_0,即认为强度有明显提高。

第二节　参数假设检验

例3：某种钢筋强度X服从正态分布，且$EX = 50(Kg/mm^2)$,今改变炼钢配方利用新法炼了9炉钢，从这9炉钢中每炉抽一根，测得强度分别为：

56.01，52.45，51.53，48.52，49.04，53.38，54.02，52.13，52.15，

问采用新法其钢筋强度均值是否有明显提高？$(\alpha = 0.05)$

解：设$H_0 : \mu = 50,\ H_1 : \mu > 50$

总体方差未知，选取统计量$T = \frac{\overline{X}-\mu_0}{S}\sqrt{n}.$

对显著性水平$\alpha = 0.05$,查表得$t_{0.95}(8) = 1.86$, 因而拒绝域为$W = \{t > 1.86\}$,

由样本得$\bar{x} = 52.14, s = 2.24$,则$t = \frac{52.14-50}{2.24}\sqrt{9} = 2.87 > 1.86$

所以拒绝H_0,即认为强度有明显提高。

第二节 参数假设检验

二.方差的检验

设$X \sim N(\mu, \sigma^2)$,$(X_1, \cdots, X_n)$为取自总体的样本,要检验$H_0 : \sigma^2 = \sigma_0^2 \longleftrightarrow H_1 : \sigma^2 \neq \sigma_0^2$

(1)若$\mu = \mu_0$已知,当H_0成立时,选取统计量

$$\chi^2 = \frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\sigma_0^2} \sim \chi^2(n)$$

故对给定的显著性水平α,由$P(k_1 < \chi^2 < k_2) = 1 - \alpha$确定$k_1, k_2$,得拒绝域

$$W = \{(x_1, \cdots, x_n) : \chi^2 < \chi^2_{\frac{\alpha}{2}}(n) 或 \chi^2 > \chi^2_{1-\frac{\alpha}{2}}(n)\}.$$

其中$\chi^2_\alpha(n)$是$\chi^2(n)$分布的α分位数.

第二节　参数假设检验

(2)若μ未知,用$\overline{X}$代替μ,选取统计量

$$\chi^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma_0^2} \sim \chi^2(n-1)$$

对给定的显著性水平α,拒绝域为

$$W = \{(x_1, \cdots, x_n) : \chi^2 < \chi^2_{\frac{\alpha}{2}}(n-1) \text{或} \chi^2 > \chi^2_{1-\frac{\alpha}{2}}(n-1)\}.$$

第二节 参数假设检验

例4: 某种导线要求其电阻的标准差不得超过0.005Ω,今在生产的一批导线中取样品9根，测得$S = 0.007\Omega$,问在$\alpha = 0.05$下，能认为这批导线的方差显著偏大吗？

解：设$H_0 : \sigma^2 = (0.005)^2$ $H_1 : \sigma^2 > (0.005)^2$

μ未知，取统计量$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(9-1)$

由$\alpha = 0.05$,查表得$\chi^2_{0.95}(8) = 15.5$,所以拒绝域$W = \{\chi^2 > 15.5\}$

由样本$\chi^2 = \frac{8*0.007^2}{0.005^2} = 15.68 > 15.5$,所以拒绝$H_0$,认为这批导线方差显著偏大。

第二节 参数假设检验

例4: 某种导线要求其电阻的标准差不得超过0.005Ω,今在生产的一批导线中取样品9根，测得$S = 0.007\Omega$,问在$\alpha = 0.05$下，能认为这批导线的方差显著偏大吗?

解：设$H_0 : \sigma^2 = (0.005)^2$ $H_1 : \sigma^2 > (0.005)^2$

μ未知，取统计量$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(9-1)$

由$\alpha = 0.05$,查表得$\chi^2_{0.95}(8) = 15.5$,所以拒绝域$W = \{\chi^2 > 15.5\}$

由样本$\chi^2 = \frac{8*0.007^2}{0.005^2} = 15.68 > 15.5$,所以拒绝$H_0$,认为这批导线方差显著偏大。

第二节　参数假设检验

三.数学期望的比较

设$(X_1,\cdots,X_{n_1})$, $(Y_1,\cdots,Y_{n_2})$是分别取自正态总体$N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$的两个简单随机样本,而且假设这两个样本之间也是相互独立的.分别用$\overline{X},\overline{Y},S_1^2,S_2^2$表示这两个样本的均值和方差

$$\overline{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i,\qquad \overline{Y}=\frac{1}{n_2}\sum_{j=1}^{n_2}Y_j,$$

$$S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\overline{X})^2,\qquad S_2^2=\frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2.$$

$$H_0:\mu_1=\mu_2\longleftrightarrow H_1:\mu_1\neq\mu_2$$

第二节 参数假设检验

(1)已知σ_1^2, σ_2^2

考虑$\overline{X} - \overline{Y}$,由于给定两个样本之间的独立性,

$$\begin{aligned} E(\overline{X} - \overline{Y}) &= \mu_1 - \mu_2, \\ D(\overline{X} - \overline{Y}) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

因此在原假设H_0 $(\mu_1 = \mu_2)$成立时,统计量

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

否则,U服从均值不为零的正态分布.故对给定的显著性水平α,应取拒绝域$W = \{|u| > u_{1-\frac{\alpha}{2}}\}$.

第二节 参数假设检验

(2) $\sigma_1^2 = \sigma_2^2 = \sigma^2$(方差未知但相等)

类似于前面t检验中的讨论,当总体方差未知时,应该用样本方差代替总体方差,而且由于给定两个总体的方差相等,我们用两个样本的合并方差$S_w^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$代替$\sigma_1^2 = \sigma_2^2$. 由定理

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_w\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

因此在原假设H_0 $(\mu_1 = \mu_2)$成立时,统计量

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

可作为H_0的检验统计量.对给定的显著水平α,取拒绝域$W = \{|t| > t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)\}$ 这就是两样本的t检验.

第二节 参数假设检验

(3) $\sigma_1^2 \neq \sigma_2^2$, $n_1 = n_2 = n$ (方差未知且不等,但样本大小相等)
定义$Z_i = X_i - Y_i$, $i = 1, \cdots, n$, 显然,Z_i相互独立.记

$$E(Z_i) = E(X_i - Y_i) = \mu_1 - \mu_2 = d,$$

$$D(Z_i) = DX_i + DY_i = \sigma_1^2 + \sigma_2^2 = \sigma^2,$$

则$Z_i \sim N(d, \sigma^2)$, $i = 1, \cdots, n$. 因此,等价于检验假设

$$H_0 : d = 0 \longleftrightarrow H_1 : d \neq 0$$

这里σ^2未知, 由t检验,记

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2,$$

则当H_0 $(d = 0)$成立时, 有

$$T = \frac{\overline{Z}}{S}\sqrt{n} \sim t(n-1)$$

这就是配对试验的t检验.

第二节 参数假设检验

四. 方差的比较

$$H_0 : \sigma_1^2 = \sigma_2^2 \longleftrightarrow H_1 : \sigma_1^2 \neq \sigma_2^2$$

1. μ_1, μ_2已知

记

$$S_1^{*2} = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - \mu_1)^2, S_2^{*2} = \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i - \mu_2)^2$$

采用F检验，选统计量

$$F = \frac{S_1^{*2}}{S_2^{*2}} \overset{H_0}{\sim} F(n_1, n_2).$$

故对给定的显著性水平α,可取拒绝域

$$W = \{F < F_{\frac{\alpha}{2}}(n_1, n_2)或\ F > F_{1-\frac{\alpha}{2}}(n_1, n_2)\}$$

第二节 参数假设检验

2. μ_1, μ_2未知

记$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$
采用F检验，选统计量

$$F = \frac{S_1^2}{S_2^2} \overset{H_0}{\sim} F(n_1 - 1, n_2 - 1).$$

故对给定的显著性水平α,可取拒绝域

$$W = \{F < F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)\text{或 } F > F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)\}$$

注：$F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)}$

第二节 参数假设检验

五. 非正态总体的参数假设检验

1. 单个非正态总体的参数检验

设总体$X \sim b(1,p)$, $(X_1,...,X_n)$取自总体X, 检验假设

$$H_0 : p = p_0 \quad H_1 : p \neq p_0$$

$$\because \mathrm{E}X = p, \quad Var(X) = p(1-p), \therefore U = \frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} \overset{H_0}{\sim} N(0,1)$$

对给定的显著性水平α, $W = \{|u| > u_{1-\frac{\alpha}{2}}\}$.

注：上述问题称为比率的假设检验.

3.3 Neyman-Pearson基本引理与随机化检验

一、功效函数

称样本观察值落在拒绝域的概率为检验的**功效函数**，记为

$$\beta(\theta)=P_\theta(X\in W),\quad \theta\in\Theta$$

注意: θ取值在全空间Θ

由定义知

在$\theta\in\Theta_0$ 时，$\beta(\theta)$ 是犯第一类错的概率.

在$\theta\in\Theta_1$ 时, $1-\beta(\theta)$ 是犯第二类错的概率.

功效函数是假设检验中最重要的概念之一. 因为同一个原假设可以有许多检验法，其中自然有优劣之分. 这区分的依据，就取决于检验的功效函数.

显然，在$\theta\in\Theta_0$ 时，即$\mathbf{H}_0$为真时, 希望功效函数$\beta(\theta)$尽量小.

在$\theta\in\Theta_1$时, 即$\mathbf{H}_1$为真时, 希望功效函数$\beta(\theta)$尽量大.

Neyman和Pearson的假设检验理论基本思想：

寻找**先控制犯第一类错误的概率**在某一个范围内，**然后寻找**使犯第二类错误的**概率尽可能小**的检验(**Neyman-Pearson**原则).

即对选定的一个较小的数$\alpha(0<\alpha<1)$在满足

$$\beta(\theta)\le\alpha,\quad \theta\in\Theta_0,$$

的检验中，寻找在$\theta\in\Theta_1$时,$\beta(\theta)$尽可能大的检验.

在$\theta\in\Theta_0$ 时,$\beta(\theta)\le\alpha$ 的检验称为水平为α的检验.

二、 检验函数和随机化检验

检验函数定义：设$\varphi(x)$为定义在样本空间χ上的可测函数，满足条件$0\leqslant\varphi(x)\leqslant 1$. 在有了样本观测值$x$后, 计算$\varphi(x)$, 然后以概率$\varphi(x)$**拒绝原假设,以概率**$1-\varphi(x)$ **接受原假设,** 则称$\varphi(x)$为检验函数.若$\varphi(x)$能取(0,1)内之值,称$\varphi(x)$为**随机化检验.** 若仅取0,1两个值，则称$\varphi(x)$为**非随机化检验**，其拒绝域由满足条件$\varphi(x)=1$的那些x构成. **功效函数**可写为

$$\beta(\theta)=P_\theta(X\in W)=E_\theta[\varphi(X)],\quad \theta\in\Theta$$

非随机化检验函数：

$$\varphi(x)=\begin{cases}1, & x\in W \longleftarrow \text{拒绝域}\\ 0, & x\notin W\end{cases}$$

三、Neyman-Pearson基本引理

最大功效检验的定义: 对如下简单假设检验，

$$H_0:\theta=\theta_0 \leftrightarrow H_1:\theta=\theta_1$$

设$\varphi(x)$是该检验水平为α的检验.如果对任意一个水平为α的检验$\varphi_1(x)$，都有

$$E_{\theta_1}[\varphi(X)]\geq E_{\theta_1}[\varphi_1(X)]\Leftrightarrow \beta_{\varphi}(\theta_1)\geq \beta_{\varphi_1}(\theta_1)$$

则称检验$\varphi(x)$是水平为α的最大功效检验，记为**MPT**.

Neyman-Pearson基本引理:

对检验问题

$$H_0:\theta=\theta_0 \quad , \quad H_1:\theta=\theta_1 \quad (\theta_0\neq\theta_1)$$

(1)对给定的α (0<α<1)存在一个检验函数$\varphi(x)$及常数$k\geqslant 0$，使

$$E_{\theta_0}[\varphi(X)]=\alpha \qquad \text{(I)}$$

$$\varphi(x)=\begin{cases}1, & L(x;\theta_1)>kL(x;\theta_0)\\ 0, & L(x;\theta_1)<kL(x;\theta_0)\end{cases} \qquad \text{(II)}$$

(2)由(I)和(II)确定的检验函数$\varphi(x)$是水平为α 的**MPT**. 反之，如果$\varphi(x)$是水平为α 的**MPT**，则一定存在常数$k\geq 0$，使$\varphi(x)$满足(II)式.

(1)对给定的$\alpha\,(0<\alpha<1)$存在一个检验函数$\varphi(x)$及常数$k\geqslant 0$，使

$$E_{\theta_0}[\varphi(X)]=\alpha \qquad \text{(I)}$$

$$\varphi(x)=\begin{cases}1, & L(x;\theta_1)>kL(x;\theta_0)\\ 0, & L(x;\theta_1)<kL(x;\theta_0)\end{cases} \qquad \text{(II)}$$

证明: 要证,存在形如(II)的检验函数，使(I)成立.

对任一实数c ,令

$$\psi(c)=P\{L(X;\theta_1)>cL(X;\theta_0)\}$$

因此

$1-\psi(c)=P\left\{\dfrac{L(X;\theta_1)}{L(X;\theta_0)}\leq c\right\}$ 是 $\dfrac{L(X;\theta_1)}{L(X;\theta_0)}$ 的分布函数,

是非降，右连续.

证$E[\varphi(X)]=\alpha$

所以$\psi(c)$非增，右连续，且

$$\psi(+\infty)=0,\quad \psi(0-0)=1,$$

$$\psi(c-0)-\psi(c)=P\left\{\frac{L(X;\theta_1)}{L(X;\theta_0)}=c\right\}\quad —(\mathrm{III})$$

给定$\alpha\in(0,1)$，有且只有下列两种情况：

(i) 存在$c_0\geq 0$,使得$\psi(c_0)=\alpha$

(ii) 存在$c_0\geq 0$,使得$\psi(c_0)<\alpha\leq\psi(c_0-0)$

在第(i)种情况下定义

$$\varphi(x)=\begin{cases}1, & L(x;\theta_1)>c_0L(x;\theta_0)\\ 0, & L(x;\theta_1)\leq c_0L(x;\theta_0)\end{cases}$$

则 $E[\varphi(X)]=P\{L(X;\theta_1)>c_0L(X;\theta_0)\}=\psi(c_0)=\alpha$

(ii) 存在$c_0 \geq 0$, 使得$\psi(c_0) < \alpha \leq \psi(c_0 - 0)$

定义

$$\varphi(x) = \begin{cases} 1, & L(x;\theta_1) > c_0 L(x;\theta_0) \\ \dfrac{\alpha - \psi(c_0)}{\psi(c_0 - 0) - \psi(c_0)}, & L(x;\theta_1) = c_0 L(x;\theta_0) \\ 0, & L(x;\theta_1) < c_0 L(x;\theta_0) \end{cases}$$

则

$$\begin{aligned} E_{\theta_0}[\varphi(X)] &= 1 \times P\{L(x;\theta_1) > c_0 L(x;\theta_0)\} \\ &\quad + \frac{\alpha - \psi(c_0)}{\psi(c_0 - 0) - \psi(c_0)} \times P\{L(x;\theta_1) = c_0 L(x;\theta_0)\} \\ &\quad + 0 \times P\{L(x;\theta_1) < c_0 L(x;\theta_0)\} \\ &= \psi(c_0) + [\alpha - \psi(c_0)] = \alpha \end{aligned}$$

于是取$k=c_0$，则以上定义的$\varphi(x)$是水平为α的检验函数.

(2)的证明:先证满足(I),(II)的检验函数$\varphi(x)$是**MPT**.

设$\varphi^*(x)$是任意一个水平为α的检验函数，　$E_{\theta_0}[\varphi^*(X)]\le\alpha$

记　$S^+=\{x:\varphi(x)>\varphi^*(x)\}$，　$S^-=\{x:\varphi(x)<\varphi^*(x)\}$

则在S^+上$\varphi^*(x)$非负$\varphi(x)>0$，因而由(II)知 $L(x;\theta_1)-kL(x;\theta_0)\ge 0$

同理在S^-上$\varphi(x)<1$，因而　$L(x;\theta_1)-kL(x;\theta_0)\le 0$

因此在 $S^+\cup S^-$上总有

$$[\varphi(x)-\varphi^*(x)][L(x;\theta_1)-kL(x;\theta_0)]\ge 0$$

于是
$$\int_{S^+\cup S^-}[\varphi(x)-\varphi^*(x)][L(x;\theta_1)-kL(x;\theta_0)]dx_1dx_2\cdots dx_n$$
$$+\int_{\chi-(S^+\cup S^-)}0\cdot[L(x;\theta_1)-kL(x;\theta_0)]dx_1dx_2\cdots dx_n\ge 0$$

从而 $E_{\theta_1}[\varphi(X)]-E_{\theta_1}[\varphi^*(X)]\ge k\{E_{\theta_0}[\varphi(X)]-E_{\theta_0}[\varphi^*(X)]\}$

$$=k\{\alpha-E_{\theta_0}[\varphi^*(X)]\}\ge 0$$

$E_{\theta_0}[\varphi(X)]=\alpha$　(I)

所以$\varphi(x)$是MPT.

$$\varphi(x)=\begin{cases}1, & L(x;\theta_1)>kL(x;\theta_0)\\ 0, & L(x;\theta_1)<kL(x;\theta_0)\end{cases}$$　(II)

最后证明:若$\varphi^*(x)$是水平为α的MPT,则**一定存在非负常数k,使$\varphi^*(x)$满足(II)式.**

设$\varphi(x)$是满足(I),(II)的检验函数,由上知$\varphi(x)$是水平为α的MPT.

若$\varphi^*(x)$是水平为α的MPT，则 $E_{\theta_1}[\varphi^*(X)] \geq E_{\theta_1}[\varphi(X)]$

又因$\varphi(x)$是水平为α的MPT,所以$E_{\theta_1}[\varphi^*(X)] \leq E_{\theta_1}[\varphi(X)]$

因此$E_{\theta_1}[\varphi^*(X)] = E_{\theta_1}[\varphi(X)]$.

由于$\varphi(x)$满足(II)，所以(由前段证明过程)

$$[\varphi(x)-\varphi^*(x)][L(x;\theta_1)-kL(x;\theta_0)] \geq 0. \quad ①$$

因$E_{\theta_0}[\varphi^*(X)] \leq \alpha = E_{\theta_0}[\varphi(X)]$,所以

$$\int [\varphi(x)-\varphi^*(x)][L(x;\theta_1)-kL(x;\theta_0)]dx_1dx_2\cdots dx_n$$

$$= E_{\theta_1}[\varphi(X)]-E_{\theta_1}[\varphi^*(X)]-k\left\{E_{\theta_0}[\varphi(X)]-E_{\theta_0}[\varphi^*(X)]\right\} \leq 0 \quad ②$$

$$\varphi(x)=\begin{cases}1, & L(x;\theta_1) > kL(x;\theta_0)\\ 0, & L(x;\theta_1) < kL(x;\theta_0)\end{cases} \quad (II)$$

但由①此式大于或等于零. 于是上式积分只能为零，因此

$$[\varphi(x)-\varphi^*(x)][L(x;\theta_1)-kL(x;\theta_0)] = 0,$$

即在集 $\{x \mid L(x;\theta_1)-kL(x;\theta_0) \neq 0\}$上$\varphi(x) = \varphi^*(x)$ (a.s.)

这证明了$\varphi^*(x)$满足(II)式(a.s.). 证完.

MPT的性质: 设$\varphi(x)$是**N-P**基本引理中水平为α的最大功效检验(**MPT**)函数，则

$$\beta_\varphi(\theta_1)=E_{\theta_1}[\varphi(X)]\geq\alpha$$

又设0<α<1, 则 $\beta_\varphi(\theta_1)=\alpha$ 的充分必要条件

$$L(x;\theta_1)=L(x;\theta_0) \quad \text{a.s.}$$

定义似然比: $\lambda(x)=\dfrac{L(x;\theta_1)}{L(x;\theta_0)}$

由N-P引理知似然比检验$\lambda(x)$是MPT,MPT是似然比检验.

$\lambda(x)$是连续时有

$$\varphi(x)=\begin{cases}1, & \lambda(x)\geq k(\text{拒绝}H_0)\\ 0, & \lambda(x)<k(\text{接受}H_0)\end{cases}$$

N-P引理的应用举例

例1 设 $X_1,\ldots,X_n$ 取自 $N(\theta,1)$. $f(x;\theta)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\theta)^2}{2}}$

要检验 $H_0:\theta=\theta_0=0$ ， $H_1:\theta=\theta_1=1.$ 取水平为 $\alpha\,(0<\alpha<1)$.

构造似然比统计量

$$\lambda(x)=\frac{\prod_{i=1}^{n}f(x_i;\theta_1)}{\prod_{i=1}^{n}f(x_i;\theta_0)}=\frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-1)^2\right\}}{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right\}}=\exp\left\{n\bar{x}-\frac{n}{2}\right\}$$

由N-P引理

拒绝域为 $W=\{\lambda(x)\geq k\}=\{\bar{x}\geq c\}$

当 H_0 为真时 $\bar{X}\sim N(0,\frac{1}{n})$

由 $E_{\theta_0}[\varphi(X)]=P_{\theta_0}\{\bar{X}\geq c\}=\alpha$ 确定 c 值

$$1-P_{\theta_0}\{\sqrt{n}\bar{X}<c\sqrt{n}\}=\alpha\Rightarrow\sqrt{n}c=u_{1-\alpha},c=\frac{u_{1-\alpha}}{\sqrt{n}}\Rightarrow W=\left\{\bar{x}\geq\frac{u_{1-\alpha}}{\sqrt{n}}\right\}.$$

3.4 一致最大功效检验

复合检验问题

$$H_0:\theta\in\Theta_0 \quad , \quad H_1:\theta\in\Theta_1$$

设$\varphi(x)$是水平为α的检验. 如果对任意一个水平为α的检验$\varphi_1(x)$，都有

$$E_\theta[\varphi(X)]\geq E_\theta[\varphi_1(X)], \quad \forall\theta\in\Theta_1$$

则称检验$\varphi(x)$是水平为α的一致最大功效检验，记为**UMPT**.

单调似然比(MLR)定义：

设$\{f(x;\theta):\theta\in\Theta\}$是含实参数$\theta$ 的概率密度族，如存在实值统计量$T(X)$，使对任意$\theta_1<\theta_2$都有

(1) $P\{f(X;\theta_1)\neq f(X;\theta_2)\}>0$

(2) 似然比 $\lambda(x)=\dfrac{f(x;\theta_2)}{f(x;\theta_1)}$ 关于$T(x)$非降(或非增).

则称概率密度族$\{f(x;\theta):\ \theta\in\Theta\}$关于$T(x)$具有非降(或非增)单调似然比(MLR).

例2 单参数指数型分布族

$$L(x;\theta)=c(\theta)\exp\{Q(\theta)\cdot T(x)\}\cdot h(x)\ \text{—}\ (\text{I})$$

其中$c(\theta)>0$, 假设$Q(\theta)$是θ 的严增(或严减)函数, 则在$\theta_1<\theta_2$时，其似然比

$$\lambda(x)=\frac{L(x;\theta_2)}{L(x;\theta_1)}=\frac{c(\theta_2)}{c(\theta_1)}\cdot\exp\{[Q(\theta_2)-Q(\theta_1)]T(x)\}$$

是$T(x)$的严增(或严减)函数. 由定义知单参数指数型分布族(I)关于充分统计量$T(x)$具有非降(或非增)MLR.

MLR分布族的性质：设$\{f(x;\theta):\theta\in\Theta\}$关于$T(x)$具有非降(增)MLR. 若$\psi(t)$是$t$的一个非降(非增)函数，则$E_\theta[\psi(T(X))]$是$\theta$的一个非降(非增)函数.

证明： 设$\theta_1<\theta_2$，需证 $E_{\theta_1}\left[\psi(T(X))\right]\le E_{\theta_2}\left[\psi(T(X))\right]$

记 $A=\{x:L(x;\theta_1)<L(x;\theta_2)\},B=\{x:L(x;\theta_1)>L(x;\theta_2)\}$

则对任意的$x_1\in A$和$x_2\in B$，有

$$\lambda(x_1)=\frac{L(x_1;\theta_2)}{L(x_1;\theta_1)}>1,\quad \lambda(x_2)=\frac{L(x_2;\theta_2)}{L(x_2;\theta_1)}<1\Rightarrow\lambda(x_1)>\lambda(x_2)$$

因为似然比$\lambda(x)$是$T(x)$的非降函数，所以可推知$T(x_1)\ge T(x_2)$.

又由于$\psi(t)$是t的非降函数,所以$\psi(T(x_1))\ge\psi(T(x_2))$

令 $$a=\inf\{\psi(T(x)):x\in A\},b=\sup\{\psi(T(x)):x\in B\}$$

则由以上过程知, $a\ge b$.于是

$$a=\inf\{\psi(T(x)):x\in A\}\,,\ b=\sup\{\psi(T(x)):x\in B\}$$

$$A=\{x:L(x;\theta_1)<L(x;\theta_2)\},B=\{x:L(x;\theta_1)>L(x;\theta_2)\}$$

$$E_{\theta_2}[\psi(T(x))]-E_{\theta_1}[\psi(T(x))]$$

$$=\int_{\Omega}\psi(T(x))[L(x;\theta_2)-L(x;\theta_1)]dx_1dx_2\cdots dx_n$$

$$=\int_{A}+\int_{B}+\int_{C}0\,dx_1dx_2\cdots dx_n,\quad C=\Omega-(A\cup B)=\{x:L(x;\theta_1)=L(x;\theta_2)\}$$

$$\geq a\int_{A}[L(x;\theta_2)-L(x;\theta_1)]dx_1dx_2\cdots dx_n+b\int_{B}[L(x;\theta_2)-L(x;\theta_1)]dx_1dx_2\cdots dx_n$$

$$=a\int_{A}[L(x;\theta_2)-L(x;\theta_1)]dx_1dx_2\cdots dx_n-b\int_{A}[L(x;\theta_2)-L(x;\theta_1)]dx_1dx_2\cdots dx_n$$

证完.

$$=(a-b)\int_{A}[L(x;\theta_2)-L(x;\theta_1)]dx_1dx_2\cdots dx_n\geq 0$$

例3. 设$\{f(x;\theta):\theta\in\Theta\}$关于$T(x)$具有非降单调似然比MLR.

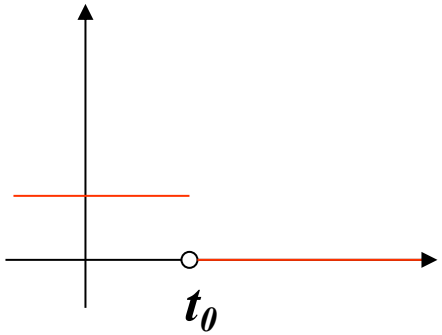
1) 取$\psi(t)=t$, 则$\psi(t)$是t的非降函数(实际严增函数). 由MLR的性质知, $T(x)$的均值$E_\theta[T(X)]$是θ的一个非降函数.

2)对任意给定的t_0,定义

$$\psi(t)=\begin{cases}1, & t\le t_0\\ 0, & t>t_0\end{cases}$$

则$\psi(t)$是t的非增函数.

$\{f(x;\theta):\theta\in\Theta\}$关于$T(x)$具有非降MLR.
若$\psi(t)$是t的一个非降(非增)函数,
则$E_\theta[\psi(T(X))]$是θ的一个非降(非增)函数.



由MLR的性质知$E_\theta[\psi(T(X))]$是θ的非增函数.

即对$\theta_1<\theta_2$有 $E_{\theta_1}[\psi(T(X))]\ge E_{\theta_2}[\psi(T(X))]$

即 $P_{\theta_1}\{T(X)\le t_0\}\ge P_{\theta_2}\{T(X)\le t_0\}$

此不等式说明概率密度族关于$T(x)$具有非降MLR时,
$T(x)$的分布函数$F_\theta(t)=P_\theta\{T(X)\le t\}$是$\theta$的非增函数.

3)设$X\sim b(n,p)$,其中$0<p<1$为未知参数.其概率密度为

$$f(x;p)=\binom{n}{k}p^k(1-p)^{n-k}=(1-p)^n\exp\left\{\ln\left(\frac{p}{1-p}\right)\cdot k\right\}\cdot\binom{n}{k}$$

是单参数指数型分布族.这里 $Q(p)=\ln\left(\frac{p}{1-p}\right)$ 是p的严增函数,

此分布关于$T(x)=x$具有非降MLR.

由**1)**二项分布均值$E_p[T(X)]=E_p[X]=np$是p的非降函数(实际严增);

由**2)**知其分布函数

$$P_p\{X\le x\}=\sum_{k\le x}\binom{n}{x}\cdot p^x\cdot(1-p)^{n-x}$$

是***p***的非增函数(严降).

简单假设的最佳检验法:

N-P基本引理： $f(x;\theta_0), f(x;\theta_1)$是两个不同总体的密度函数,

则关于假设 $H_0:\theta=\theta_0,\quad H_1:\theta=\theta_1(\theta_0\neq\theta_1)$

(1) 对给定的$\alpha(0<\alpha<1)$, $\exists\varphi(x)$及$k\geq 0$

使

$$E_{\theta_0}\varphi(x)=\alpha \qquad \text{(I)}$$

$$\varphi(x)=\begin{cases}1, & L(x;\theta_1)>kL(x;\theta_0)\\ 0, & L(x;\theta_1)<kL(x;\theta_0)\end{cases} \qquad \text{(II)}$$

(2) $\varphi(x)$是水平为α的MPT.(犯第二类错误的概率最小)

反之,如$\varphi(x)$是水平为α的MPT,则一定$\exists\ k\geq 0$,使$\varphi(x)$满足(II)(a.s.)

为给出复杂假设的最佳检验法：

1.先扩充最大功效检验的定义(UMPT)

为找出UMPT还要利用似然比统计量 $\lambda(x)=\dfrac{L(x;\theta_2)}{L(x;\theta_1)}$，$(\theta_2>\theta_1)$

2.定义单调似然比(MLR)：存在统计量$T(X)$,对$\forall\,\theta_1<\theta_2$,

$$\lambda(x)=\frac{L(x;\theta_2)}{L(x;\theta_1)}$$ 是$T(X)$的非降(非增)函数.

则称密度族$\{f(x;\theta):\theta\in\Theta\}$ 关于$T(X)$具有非降(非增)单调似然比.

单边假设检验(复合假设的检验)

定理. 设$\{f(x;\theta):\theta\in\Theta\}$关于实值统计量$T(x)$具有非降MLR，则关于单边检验问题

$$H_0:\theta\le\theta_0,\quad H_1:\theta>\theta_0$$

(1) 存在水平为α的**UMPT**检验函数

$$\varphi(T(x))=\begin{cases}1, & T(x)>c\\ \delta, & T(x)=c\\ 0, & T(x)<c\end{cases}\qquad (\mathrm{I})$$

其中常数 $\delta(0\le\delta\le1)$和c由下式确定

$$E_{\theta_0}\left[\varphi(T(X))\right]=\alpha\qquad (\mathrm{II})$$

(2) 此检验的功效函数$\beta(\theta)=E_\theta\left[\varphi(T(X))\right]$

是非降的,且集合$\{\theta:0<\beta(\theta)<1\}$上是严格增加的.

(3) 若$\overline{\varphi}$ 是满足 $E_{\theta_0}\left[\overline{\varphi}(x)\right]=\alpha$ 的任一检验函数,

则 $\beta_\varphi(\theta)\le\beta_{\overline{\varphi}}(\theta),\quad \theta<\theta_0,$ 即检验(I)犯第一类错误的概率最小.

前提

证明(1):先考虑简单原假设$H_0:\theta=\theta_0$,对简单备择假设$H_1:\theta=\theta_1$ $(\theta_1>\theta_0)$检验问题.从N-P基本引理，应在

$$\lambda(x)=\frac{L(x;\theta_1)}{L(x;\theta_0)}$$ 比较大时,拒绝原假设.由定理条件

$\lambda(x)$是$T(x)$的非降函数，所以应在$T(x)$比较大时，拒绝原假设.

- 当$T(x)$为连续型r.v时，取

$$\varphi(T(x))=\begin{cases}1, & T(x)\geq c\\ 0, & T(x)<c\end{cases}$$

c由 $E_{\theta_0}\left[\varphi(T(x))\right]=P_{\theta_0}\{T(x)\geq c\}=\alpha$ 确定, 拒绝域为 $W=\{T(x)\geq c\}$.

- 当$T(x)$为离散型r.v时，取

$$\varphi(T(x))=\begin{cases}1, & T(x)>c\\ \delta, & T(x)=c\\ 0, & T(x)<c\end{cases}$$

c由 $P_{\theta_0}\{T(x)>c\}\leq\alpha\leq P_{\theta_0}\{T(x)\geq c\}$ 确定, 取$\delta=\dfrac{\alpha-P_{\theta_0}\{T(x)>c\}}{P_{\theta_0}\{T(x)=c\}}$

因此无论$T(x)$连续还是离散都满足(I)和(II).

下面接着证明检验(I)是UMPT。

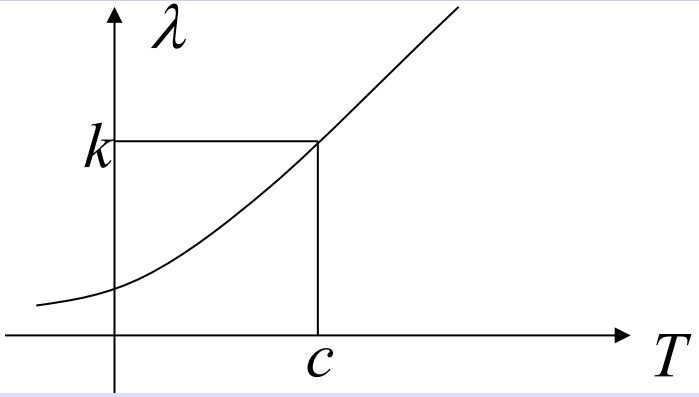
$$\varphi(T(x))=\begin{cases}1, & T(x)>c\\ \delta, & T(x)=c\\ 0, & T(x)<c\end{cases}\qquad (\mathrm{I})$$

在$T(x)=c$时,设$\lambda(x)=k$.

由于$\lambda(x)$是$T(x)$的非降函数

$$\{x:\lambda(x)>k\}\subseteq\{x:T(x)>c\}$$

$$\{x:\lambda(x)<k\}\subseteq\{x:T(x)<c\}$$



满足(I)的检验$\varphi(T(x))\Rightarrow$
必满足N-P基本引理中(II),
满足(II)的检验$\varphi(T(x))\Rightarrow$
必满足N-P基本引理中(I),

$$\varphi(x)=\begin{cases}1, & L(x;\theta_1)>kL(x;\theta_0)\\ 0, & L(x;\theta_1)<kL(x;\theta_0)\end{cases}\qquad (\mathrm{II})$$

于是由N-P基本引理的推论, $\varphi(T(x))$是简单假设检验问题

$$H_0:\theta=\theta_0 \text{ 对 } H_1:\theta=\theta_1\ (\theta_1>\theta_0)$$

水平为α的MPT.由以上过程知检验$\varphi(T(x))$与θ_1无关,只要求$\theta_1>\theta_0$,
所以$\varphi(T(x))$也是单边假设检验问题

$$\mathrm{H}_0:\theta=\theta_0 \text{ 对 } \mathrm{H}_1:\theta>\theta_0$$

的水平为α的UMPT.

$\mathrm{H}_0:\theta\le\theta_0$ 对 $\mathrm{H}_1:\theta>\theta_0$

$H_0:\theta\le\theta_0$ 对 $H_1:\theta>\theta_0$

$\varphi(T(x))$关于$T(x)$非降, 由MLR分布族的性质知$\beta(\theta)=E_\theta[\varphi(T(x))]$是$\theta$的非降函数.故$\theta\le\theta_0$时

$$E_\theta[\varphi(T(x))]\le E_{\theta_0}[\varphi(T(x))]=\alpha$$

对原假设$H_0:\theta\le\theta_0$, $\varphi(T(x))$是水平为α的检验.

综上所述, $\varphi(T(x))$是单边假设检验问题

$$H_0:\theta\le\theta_0 \text{ 对 } H_1:\theta>\theta_0$$

的水平为α的UMPT.

(2) $\beta(\theta)=E_\theta[\varphi(T(x))]$是非降的,且在$\{\theta:0<\beta(\theta)<1\}$上是严格增加的.

证明:前面已知, $\beta(\theta)=E_\theta[\varphi(T(x))]$ 是θ的非降函数.下证$\beta(\theta)$ 在$\{\theta:0<\beta(\theta)<1\}$上严格增加,即对$\theta_2>\theta_1$,要证$\beta(\theta_2)>\beta(\theta_1)$.

由N-P基本引理, $\varphi(T(x))$也是 $H_0:\theta=\theta_1$ ⇔ $H_1:\theta=\theta_2$ $(\theta_2>\theta_1)$ 的水平为$\alpha_1=E_{\theta_1}[(\varphi(x))]=\beta(\theta_1)$的MPT. 因此在$0<\alpha_1<1$时,由MPT性质有

$$E_{\theta_2}[(\varphi(x))]=\beta(\theta_2)>\alpha_1=\beta(\theta_1).$$

(3)若 $\bar{\varphi}$ 是满足$E_{\theta_0}\left[\bar{\varphi}(x)\right]=\alpha$的任一检验函数,

则 $\beta_{\varphi}(\theta)\le\beta_{\bar{\varphi}}(\theta),\quad \theta<\theta_0$

证明:任取$\theta_2<\theta_0$, 考虑 $H_0:\theta=\theta_0$ ↔ $H_1:\theta=\theta_2$ (III)
的水平为$\alpha^*=1-\alpha$的检验.

由NP基本引理, $\dfrac{L(x;\theta_2)}{L(x;\theta_0)}$ 比较大时拒绝原假设.

由条件, $\lambda(x)=\dfrac{L(x;\theta_0)}{L(x;\theta_2)}$ 关于$T(x)$非降,于是 $\dfrac{L(x;\theta_2)}{L(x;\theta_0)}$ 关于$T(x)$非增.

因此$T(x)$比较小时拒绝原假设, 即拒绝域$W=\{T(x)<c\}$.

如定义$\varphi^*=1-\varphi$, 则 $E_{\theta_0}(\varphi^*)=\alpha^*=1-\alpha$

即φ^*是(III)的水平为α^*的MPT.

因此对任一水平为α^*的检验$\bar{\varphi}^*$, $E_{\theta_0}\left[\bar{\varphi}^*(x)\right]\le\alpha^*$,都有

$$\beta_{\varphi^*}(\theta_2)\ge\beta_{\bar{\varphi}^*}(\theta_2)$$

今取检验 $\bar{\varphi}^* = 1-\bar{\varphi}$ 则

已有 $\beta_{\varphi^*}(\theta_2) \geq \beta_{\bar{\varphi}^*}(\theta_2)$

要证 $\beta_{\varphi}(\theta) \leq \beta_{\bar{\varphi}}(\theta),\ \theta < \theta_0$

即 $\bar{\varphi}^* = 1-\bar{\varphi}$ 是水平为α^*的检验

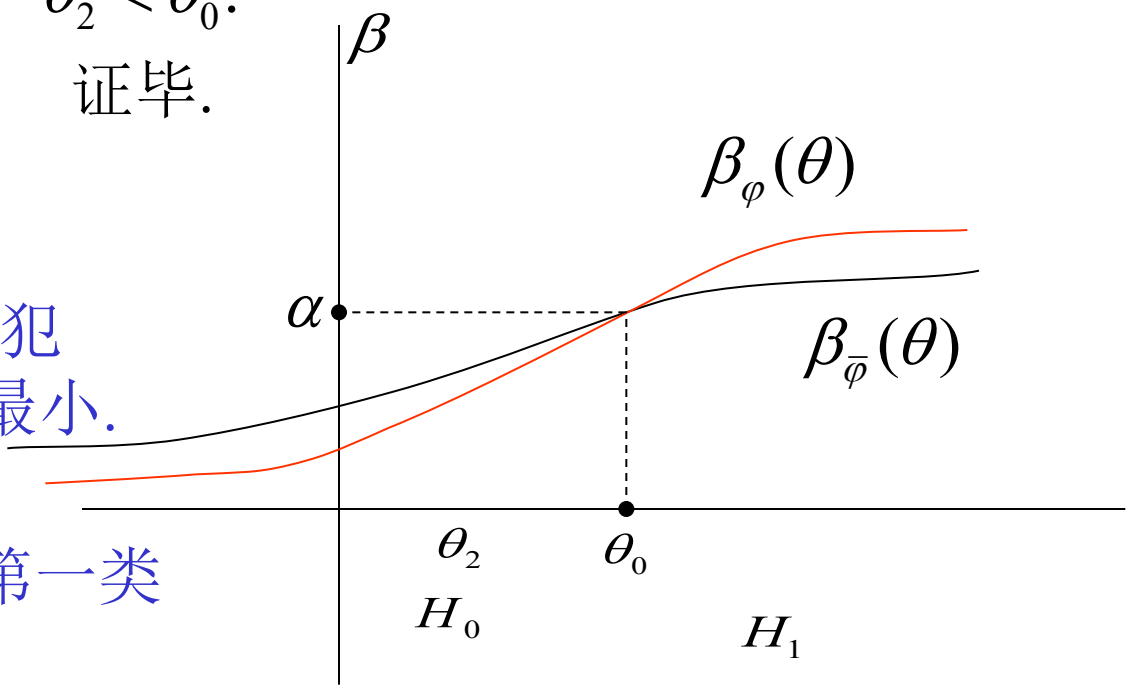
所以也有

$$1-\beta_{\varphi}(\theta_2) = \beta_{\varphi^*}(\theta_2) \geq \beta_{\bar{\varphi}^*}(\theta_2) = E_{\theta_2}[\bar{\varphi}^*(X)] = 1-\beta_{\bar{\varphi}}(\theta_2)$$

即 $\beta_{\varphi}(\theta_2) \leq \beta_{\bar{\varphi}}(\theta_2),\ \theta_2 < \theta_0.$

证毕.

满足(I),(II)的检验φ,
在水平为α的检验中犯
第 二类错误的概率最小.
在$E_{\theta_0}[\bar{\varphi}(x)] = \alpha$
的检验中$\theta<\theta_0$时犯第一类
错误的概率最小。



§3.5 区间估计

引例 已知 $X \sim N(\mu, 1)$,

μ 的无偏、有效点估计为 $\overline{X}$

μ：常数　　$\overline{X}$：随机变量

不同样本算得的 μ 的估计值不同，因此除了给出 μ 的点估计外，还希望根据所给的样本确定一个随机区间，使其包含参数真值的概率达到指定的要求.

如引例中，要找一个区间,使其包含 μ 的真值的概率为0.95. （设 $n=5$）

$$\bar{X} \sim N\left(\mu, \frac{1}{5}\right) \Rightarrow \frac{\bar{X}-\mu}{\sqrt{1/5}} \sim N(0, 1)$$

取 $\alpha = 0.05$

查表得 $z_{\alpha/2} = 1.96$

这说明 $P\left\{\left|\dfrac{\bar{X}-\mu}{\sqrt{1/5}}\right| \geq 1.96\right\}=0.05$

即 $P\left\{\bar{X}-1.96\sqrt{1/5} \leq \mu \leq \bar{X}+1.96\sqrt{1/5}\right\}=0.95$

称随机区间 $\left(\bar{X}-1.96\sqrt{1/5},\ \bar{X}+1.96\sqrt{1/5}\right)$

为未知参数 μ 的置信度为0.95的置信区间.

置信区间的意义

反复抽取容量为5的样本,都可得一个区间,此区间不一定包含未知参数 μ 的真值, 而包含真值的区间占95%.

若测得 一组样本值, 算得 $\bar{x} = 1.86$

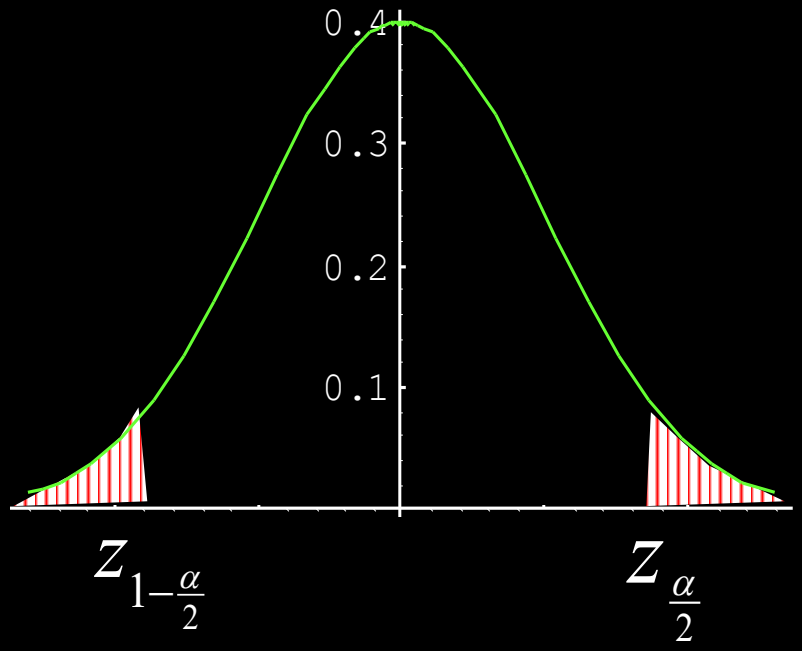
则得一区间$(1.86 - 0.877,\ 1.86 + 0.877)$

它可能包含也可能不包含μ 的真值, 反复抽样得到的区间中有95%包含 μ 的真值.

为何要取 $z_{\alpha/2}$？

当置信区间为 $(\bar{X}-z_{\frac{\alpha}{2}}\sqrt{1/5}\,,\ \bar{X}+z_{\frac{\alpha}{2}}\sqrt{1/5})$ 时

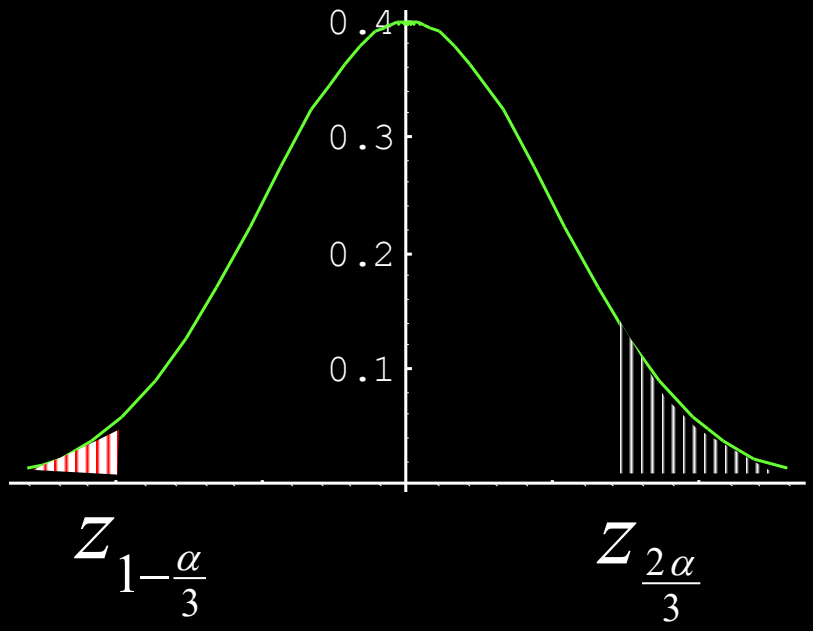
区间的长度为 $2z_{\frac{\alpha}{2}}\sqrt{1/5}$ —— 达到最短



取 $\alpha = 0.05$

$$z_{\frac{\alpha}{2}} - z_{1-\frac{\alpha}{2}} = 1.96 - (-1.96)$$

$$= 3.92$$



$$z_{\frac{2\alpha}{3}} - z_{1-\frac{\alpha}{3}} = 1.84 - (-2.13)$$

$$= 3.97$$

置信区间的定义

设 θ 为待估参数, α 是一给定的数, ($0<\alpha<1$). 若能找到统计量 $\hat{\theta}_1, \hat{\theta}_2$, 使

$$P\left\{\hat{\theta}_1<\theta<\hat{\theta}_2\right\}=1-\alpha, \qquad \theta\in\Theta$$

则称 $\left(\hat{\theta}_1, \hat{\theta}_2\right)$ 为 θ 的置信水平为 $1-\alpha$ 的置信区间或区间估计.

$\hat{\theta}_1$ —— 置信下限

$\hat{\theta}_2$ —— 置信上限

几点说明

- 置信区间的长度 $\hat{\theta}_2 - \hat{\theta}_1$ 反映了估计精度 $\hat{\theta}_2 - \hat{\theta}_1$ 越小，估计精度越高.
- α 反映了估计的可靠度, α 越小, 越可靠. α 越小, $1\text{-}\alpha$ 越大, 估计的可靠度越高,但这时, $\hat{\theta}_2 - \hat{\theta}_1$ 往往增大, 因而估计精度降低.
- α 确定后, 置信区间 的选取方法不唯一, 常选最小的一个.

处理“可靠性与精度关系”的原则

先

再

求参数置信区间

保证可靠性

提高精度

求置信区间的步骤

- 寻找一个样本的函数

$$g(X_1, X_2, \cdots, X_n, \theta)$$ —— 称为**枢轴量**

它含有待估参数，不含其它未知参数，它的分布已知，且分布不依赖于待估参数 (常由θ的点估计出发考虑).

例如 $\overline{X} \sim N(\mu, 1/5)$

取枢轴量

$$g(X_1, X_2, \cdots, X_n, \mu) = \frac{\overline{X} - \mu}{\sqrt{1/5}} \sim N(0, 1)$$

- 给定置信度 $1-\alpha$，定出常数 a，b，使得

$$P(a < g(X_1, X_2 \cdots, X_n, \theta) < b) = 1-\alpha$$

（引例中 $a = -1.96, b = 1.96$）

- 由 $a < g(X_1, X_2, \cdots X_n, \theta) < b$ 解出

得置信区间

引例中 $$P\left\{-1.96 < \frac{\overline{X}-\mu}{\sqrt{1/5}} < 1.96\right\} = 0.95$$

$$(\hat{\theta}_1, \hat{\theta}_2) = (\overline{X} - 1.96\sqrt{1/5},\quad \overline{X} + 1.96\sqrt{1/5})$$

置信区间常用公式

(一) 一个正态总体 $X \sim N(\mu, \sigma^2)$的情形

(1) 方差$\sigma^2=\sigma_0^2$已知, μ的置信区间

$$\left(\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}\right) \quad \cdots\cdots\cdots \quad (1)$$

推导 由 $\overline{X} \sim N(\mu, \frac{\sigma_0^2}{n})$ 选取枢轴量

$$g(X_1, X_2, \cdots, X_n, \mu) = \frac{\overline{X} - \mu}{\sigma_0 / \sqrt{n}} \sim N(0,1)$$

由 $$P\left\{\left|\frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}}\right| \geq z_{\frac{\alpha}{2}}\right\}=\alpha$$

解 $$\left|\frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}}\right| < z_{\frac{\alpha}{2}}$$

得 μ 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}},\quad \overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right)$$

(2) 方差σ^2未知，μ的置信区间

推导　选取枢轴量

由 $P\left\{\left|\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\right| \geq t_{n-1}\left(\frac{\alpha}{2}\right)\right\}=\alpha$ 确定

故 μ 的置信区间为

(3) 当 μ 已知时, 方差 σ^2 的 置信区间

取枢轴量 ,由概率

$$P\left\{\chi_n^2(1-\frac{\alpha}{2})<\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}<\chi_n^2(\frac{\alpha}{2})\right\}=1-\alpha$$

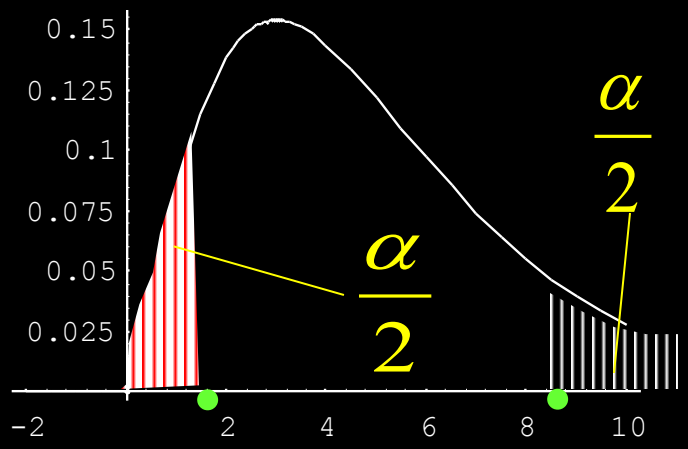
得 σ^2 的置信度为 $1-\alpha$ 置信区间为

(4) 当 μ 未知时，方差 σ^2 的置信区间

选取

则由

$$P\left\{\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)<\frac{(n-1)S^2}{\sigma^2}<\chi^2_{n-1}\left(\frac{\alpha}{2}\right)\right\}=1-\alpha$$



得 σ^2 的置信区间为

例1 某工厂生产一批滚珠，其直径 X 服从正态分布 $N(\mu,\sigma^2)$，现从某天的产品中随机抽取 6 件，测得直径为

15.1，14.8，15.2，14.9，14.6，15.1

(1) 若$\sigma^2=0.06$，求μ的置信区间
(2) 若σ^2未知，求μ的置信区间
(3) 求方差σ^2的置信区间.

} 置信度均为0.95

解 (1) $\bar{X} \sim N(\mu, 0.06/6)$ 即$N(\mu,0.01)$

$$\frac{\bar{X}-\mu}{0.1} \sim N(0,1) \quad z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$$

由给定数据算得 $\bar{x}=\frac{1}{6}\sum_{i=1}^{6}x_i=14.95$

由公式 (1) 得 μ 的置信区间为

$$(14.95-1.96\times 0.1,\quad 14.95+1.96\times 0.1)$$

$$=(14.75,\quad 15.15)$$

(2) 取　　查表

由给定数据算得　$\bar{x}=14.95$

$$s^2=\frac{1}{5}(\sum_{i=1}^{6}x_i^2-6\bar{x}^2)=0.051.\quad s=0.226$$

由公式 (2) 得 μ 的置信区间为

(3) 选取枢轴量 $s^2 = 0.051.$

查表得

由公式 (4) 得 σ^2 的置信区间为

(二) 两个正态总体的情形

$(X_1, X_2, \cdots, X_n)$为取自总体 $N(\mu_1, \sigma_1^2)$ 的样本,

$(Y_1, Y_2, \cdots, Y_m)$ 为取自总体 $N(\mu_2, \sigma_2^2)$ 的样本,

$\bar{X}, S_1^2$; $\bar{Y}, S_2^2$ 分别表示两样本的均值与方差

置信度为 $1-\alpha$

(1) σ_1^2, σ_2^2 已知, $\mu_1 - \mu_2$ 的置信区间

$$\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n}), \quad \bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{m})$$ 相互独立,

$$\longrightarrow \quad \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

$\mu_1 - \mu_2$ 的置信区间为

$$\left((\bar{X} - \bar{Y}) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \quad (\bar{X} - \bar{Y}) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right)$$

$\cdots\cdots\cdots$ (5)

(2) σ_1^2, σ_2^2未知(但$\sigma_1^2 = \sigma_2^2 = \sigma^2$) $\mu_1 - \mu_2$的置信区间

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$$

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sigma} \sim N(0,1)$$

$$\frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(m-1)$$

$$\frac{(n-1)S_1^2}{\sigma^2} + \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2(n+m-2)$$

$$\rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}} \sim t(n+m-2)$$

$$P\left\{\left|\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}}\right|<t_{\frac{\alpha}{2}}\right\}=1-\alpha$$

$\mu_1-\mu_2$ 的置信区间为

$$\left((\bar{X}-\bar{Y})\pm t_{\frac{\alpha}{2}}\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}\right)$$

(3) 方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间 (μ_1 , μ_2 未知)

取枢轴量

因此, 方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间为

. . . . . (9)

例2 某厂利用两条自动化流水线罐装番茄酱. 现分别 从两条流水线上抽取了容量分别为13与17的两个相互独立的样本

$$X_1, X_2, \cdots, X_{13} \text{ 与 } Y_1, Y_2, \cdots, Y_{17}$$

已知 $\bar{x} = 10.6g, \quad \bar{y} = 9.5g,$

$$s_1^2 = 2.4g^2, \quad s_2^2 = 4.7g^2$$

假设两条流水线上罐装的番茄酱的重量都服从正态分布, 其均值分别为 μ_1 与 μ_2

(2) 枢轴量为

查表得

由公式(9)得方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 的置信区间为

（三） 单侧置信区间

定义 对于给定的 $\alpha(0<\alpha<1)$，θ 是待估参

$(X_1, X_2, \cdots, X_n)$ 是总体 X 的样本，

若能确定一个统计量

$$\hat{\theta}_1 = \hat{\theta}_1(X_1, X_2, \cdots, X_n) \quad (\text{或} \hat{\theta}_2 = \hat{\theta}_2(X_1, X_2, \cdots, X_n))$$

使得 $P\{\hat{\theta}_1 < \theta\} = 1-\alpha$ （或 $P\{\theta < \hat{\theta}_2\} = 1-\alpha$）

则称 $(\hat{\theta}_1, +\infty)$ （或 $(-\infty, \hat{\theta}_2)$）

为置信度为 $1-\alpha$ 的单侧置信区间.

$\hat{\theta}_1$ —— 单侧置信下限 $\hat{\theta}_2$ —— 单侧置信上限

例3 已知灯泡寿命X服从正态分布，从中随机抽取 5 只作寿命试验，测得寿命为

1050 , 1100 , 1120 , 1250 , 1280 (小时)

求灯泡寿命均值的单侧置信下限与寿命方差的单侧置信上限.

解 $X \sim N(\mu,\sigma^2)$, μ,σ^2未知 $n=5$, $\bar{x}=1160$,

$$s^2=\frac{1}{4}\left(\sum_{i=1}^{5}x_i^2-5\bar{x}\right)=9950.$$

(1) 选取枢轴量

(2) 选取枢轴量

第四章　线性模型

1.变量之间的关系一般分为两类:

(1)完全确定的关系，也就是变量之间的关系可以用函数解析式表达出来；如$y = f(x)$

(2)非确定的关系，y与x的取值有关，但不能完全确定，也称为相关关系.

2.回归分析研究的主要内容:

(1)通过观察或实验数据的处理，找出变量间相关关系的定量数学表达式–经验公式，即进行参数估计，并确定经验回归方程的具体形式;

(2)检验所建立的经验回归方程是否合理;

(3)利用合理的回归方程对随机变量Y进行预测和控制;

例1：为研究商品价格与销售量之间的关系，现在收集了某商品在一个地区25个时间段内平均价格x(元) 和销售总额y(万元)，试研究y与x之间的关系.

x	35.3	29.7	30.8	58.8	61.4	71.3	74.4	76.7	70.7
y	10.98	11.13	12.51	8.40	9.27	8.73	6.36	8.50	7.82
x	46.4	28.9	28.1	39.1	46.8	48.5	59.3	70.0	70.0
y	8.24	12.19	11.88	9.57	10.94	9.58	10.09	8.11	6.83
x	72.1	58.1	44.6	33.4	28.6	57.7	74.5		
y	7.68	8.47	8.86	10.36	11.08	9.14	8.88		

第一节　一元线性回归模型

第一节　一元线性回归模型

一.基本概念

1.定义:设回归变量x与响应变量(因变量)y之间有这样的关系，$y = \beta_0 + \beta_1 x + \varepsilon$,其中$\beta_0, \beta_1$ 是未知参数，ε是随机项，且假定$E(\varepsilon) = 0, D(\varepsilon) = \sigma^2$,则称此模型为一元线性回归模型.
若$\varepsilon \sim N(0, \sigma^2)$，则称为一元正态线性回归模型.

2.设n对观测值$(x_i, y_i), i = 1, \cdots, n$，每一对$(x_i, y_i)$都有$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$,且$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2, i = 1, \cdots, n$，$Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$. 我们希望选取一条回归直线$\tilde{y} = \beta_0 + \beta_1 x$使它接近这$n$个点.
由此可以估计β_0, β_1，记为$\hat{\beta}_0, \hat{\beta}_1$，称其为回归系数.

第一节 一元线性回归模型

二.参数β_0, β_1的估计

对于每一个x_i,可以得到回归直线$\tilde{y} = \beta_0 + \beta_1 x$上对应点的纵坐标$\tilde{y}_i = \beta_0 + \beta_1 x_i$, $|y_i - \tilde{y}_i|$刻画了y_i与$\tilde{y}_i$之间的偏离程度.

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

表示总的偏离平方和

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0. \end{cases}$$

此方程组称为正规方程组, 即

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i, \end{cases}$$

第一节 一元线性回归模型

解得

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

记$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

则$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$分别称为$\beta_0, \beta_1$的最小二乘估计，这种方法称为最小二乘法, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 为经验回归直线$(\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x})$.

第一节　一元线性回归模型

性质1： 对于一元线性回归模型有

(1) $E(\hat{\beta}_0) = \beta_0$，$E(\hat{\beta}_1) = \beta_1$,即$\hat{\beta}_0, \hat{\beta}_1$分别是$\beta_0, \beta_1$的无偏估计；

(2) $Var(\hat{\beta}_0) = (\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}})\sigma^2 = \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}\sigma^2$

$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}}{S_{xx}}\sigma^2$

性质2： $Cov(\hat{\beta}_1, \overline{y}) = 0$, 即$\hat{\beta}_1$与$\overline{y}$不相关.

性质3： $\hat{\beta}_0$，$\hat{\beta}_1$ 分别是β_0，β_1的最小方差线性无偏估计.

第一节 一元线性回归模型

三.参数σ^2的估计

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \cdots, n$为回归值，则称$\hat{\varepsilon}_i = y_i - \hat{y}_i$为第i个残差，$i = 1, \cdots, n$. **定义**

$$\begin{aligned} Q_e &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \sum_{i=1}^{n} [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2 = S_{yy} - \hat{\beta}_1 S_{xy} \end{aligned}$$

为残差平方和,它代表y_i与经验回归直线上点的纵坐标$\hat{y}_i$的离差平方和,反映了试验的随机误差.

性质4. $\hat{\sigma}^2 = \frac{Q_e}{n-2}$为$\sigma^2$的无偏估计.

第一节　一元线性回归模型

四.回归方程的显著性检验

$$H_0: \beta_1 = 0 \leftrightarrow H_1: \beta_1 \neq 0$$

当H_0成立时，说明y与x之间无线性相关关系；

当H_0不成立时，说明y与x之间线性相关关系显著.

第一节　一元线性回归模型

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2 \\ &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = Q_e + U \end{aligned}$$

其中$Q_e = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$,

$$U = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1 x_i-\hat{\beta}_0-\hat{\beta}_1\bar{x})^2 = \hat{\beta}_1{}^2 S_{xx},$$

S_{yy}为总的离差平方和，U为回归平方和，表示回归值$\hat{y}_i$的波动，Q_e为剩余(残差)平方和，反映了随机误差的存在而引起的因变量的波动.

第一节 一元线性回归模型

定理:若随机误差$\varepsilon \sim N(0, \sigma^2)$, 则

$$(1)\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}), \quad (2)\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\sigma^2),$$

$$(3)\frac{Q_e}{\sigma^2} \sim \chi^2(n-2), Q_e \text{与} \hat{\beta}_1 \text{相互独立},$$

$$(4)\text{在} H_0(\beta_1 = 0) \text{成立条件下}, \frac{U}{\sigma^2} \sim \chi^2(1),$$

$$F = \frac{U}{Q_e/(n-2)} = (n-2)\frac{U}{Q_e} \sim F(1, n-2),$$

$$(5)\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}}\sqrt{S_{xx}} \sim t(n-2), \text{其中} \hat{\sigma}^2 = \frac{Q_e}{n-2}$$

第一节　一元线性回归模型

设$\varepsilon \sim N(0, \sigma^2)$

1. F检验法

统计量$F = (n-2)\frac{U}{Q_e} \overset{H_0}{\sim} F(1, n-2)$, 对给定的显著性水平$\alpha$,拒绝域为$W = \{F > F_{1-\alpha}(1, n-2)\}$.

2. t检验法

$T = \frac{\hat{\beta}_1\sqrt{S_{xx}}}{\hat{\sigma}} \overset{H_0}{\sim} t(n-2)$ 对给定的显著性水平α,拒绝域为$W = \{|t| > t_{1-\frac{\alpha}{2}}(n-2)\}$

注:

(1)当落入拒绝域时，拒绝H_0,即认为y与x之间线性关系显著，或者说回归方程是有意义的;

(2)否则认为回归方程不合理，这种情况由多种原因引起.

第一节　一元线性回归模型

3. 相关系数检验法(判定系数法)

$$R=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}},\quad R^2=\hat{\beta}_1{}^2\frac{S_{xx}}{S_{yy}}=\frac{U}{S_{yy}}$$

R: 线性相关系数; R^2: 相关指数(判定系数), $|R|, R^2$越接近1, 说明线性相关程度越高. 对给定的显著性水平α, 拒绝域为$W=\{|R|>c\}$.

第一节 一元线性回归模型

注: $R^2 = U/S_{yy}$: 回归平方和在总离差平方和中的比例,
T, F, R^2之间的关系:

$$U = R^2 S_{yy}$$

$$Q_e = S_{yy} - U = (1 - R^2)S_{yy}, \quad \therefore F = (n-2)\frac{R^2}{1-R^2},$$

$T^2 = F = (n-2)\frac{R^2}{1-R^2}$, 故三种检验在本质上是一致的,大部分软件都采用$F$检验.

第一节　一元线性回归模型

五.预测

1.点预测： 当给定x_0时，$\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$就是$y_0$的点预测.
2.区间预测： 当给定x_0时，y_0的置信度为$1-\alpha$的置信区间，称为预测区间，即寻找y_1, y_2，使$P\{y_1 \leq y_0 \leq y_2\} = 1-\alpha$.
当$\varepsilon \sim N(0, \sigma^2)$，$y_0$的置信水平为$1-\alpha$的置信区间(即预测区间)为

$$[\hat{y}_0 - \delta(x_0),\ \hat{y}_0 + \delta(x_0)]$$

其中

$$\delta(x_0) = \hat{\sigma} t_{1-\frac{\alpha}{2}}(n-2)\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

第一节 一元线性回归模型

注：

(1)这个区间长度为$2\delta_0$，中心在$\hat{y}_0$，n, α固定，$Q_e \downarrow, S_{xx} \uparrow$，可提高预测精度.

(2)由样本观测值可以作两条曲线$y_1(x) = \hat{y}(x) - \delta(x), y_2(x) = \hat{y}(x) + \delta(x)$,这两条曲线把回归直线$\hat{y}(x) = \beta_0 + \beta_1 x$夹在中间，形成一条宽窄不等的区域，这个区域在$x = \bar{x}$处最窄.

(3)在利用回归方程进行预测时，样本容量不能太小，因为小样本也许不能真实反映变量之间的结构关系.

(4) n很大时，$t_{1-\frac{\alpha}{2}}(n-2) \approx u_{1-\frac{\alpha}{2}}$，若$\bar{x}$离$x_0$不太远，$y_0$的置信水平近似为$1-\alpha$的置信区间(即预测区间)为$[\hat{y}_0 - \hat{\sigma} u_{1-\frac{\alpha}{2}}, \hat{y}_0 + \hat{\sigma} u_{1-\frac{\alpha}{2}}]$

第一节　一元线性回归模型

3.控制：

控制问题是预测的反问题，若要y在某个范围$y_1 \le y \le y_2$，则变量x应控制在何处，即确定x_1, x_2 使

$$\begin{cases} \hat{y}(x_1) - \delta(x_1) \ge y_1 \\ \hat{y}(x_2) + \delta(x_2) \le y_2 \end{cases}$$

则当$x \in [\min\{x_1, x_2\}, \max\{x_1, x_2\}]$时，就以至少$1-\alpha$的概率保证x相对应的y落在区间$[y_1, y_2]$内.

第一节 一元线性回归模型

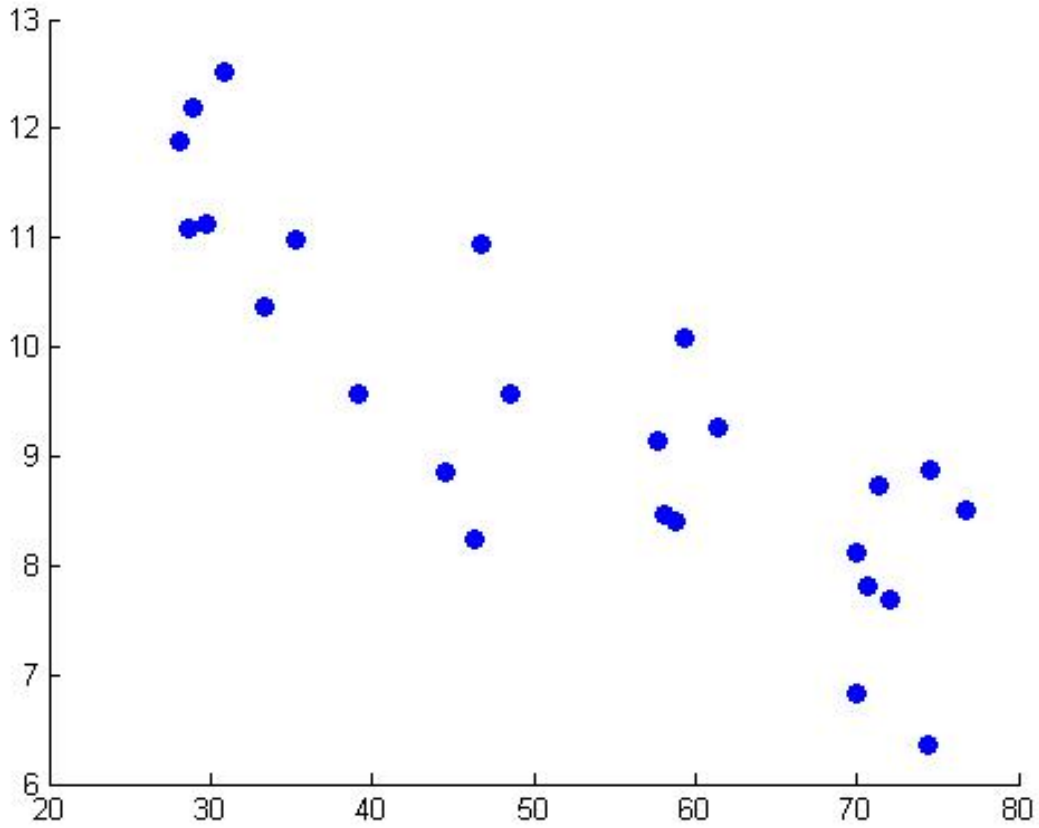
例2：为研究商品价格与销售量之间的关系，现在收集了某商品在一个地区25个时间段内平均价格x(元) 和销售总额y(万元)，试研究y与x之间的关系.

x	35.3	29.7	30.8	58.8	61.4	71.3	74.4	76.7	70.7
y	10.98	11.13	12.51	8.40	9.27	8.73	6.36	8.50	7.82
x	46.4	28.9	28.1	39.1	46.8	48.5	59.3	70.0	70.0
y	8.24	12.19	11.88	9.57	10.94	9.58	10.09	8.11	6.83
x	72.1	58.1	44.6	33.4	28.6	57.7	74.5		
y	7.68	8.47	8.86	10.36	11.08	9.14	8.88		

一元线性回归模型例1

解：

(1)画散点图



第一节　一元线性回归模型

(2)确定回归方程

由样本数据$\bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i = 52.60$, $\bar{y} = \frac{1}{25}\sum_{i=1}^{25} y_i = 9.424$

因而

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{25} x_i y_i - \frac{1}{25}(\sum_{i=1}^{25} x_i)(\sum_{i=1}^{25} y_i)}{\sum_{i=1}^{25} x_i^2 - 25\bar{x}^2} = -0.0798$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 13.623$$

所以$\hat{y} = 13.623 - 0.0798x$

第一节 一元线性回归模型

3)对回归方程进行显著性检验
利用F检验法：$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$

$$F = \frac{U}{Q_e/(n-2)} \overset{H_0}{\sim} F(1, n-2)$$

$S_{yy} = \sum_{i=1}^{25}(y_i - \bar{y})^2 = 63.82$， $U = \sum_{i=1}^{25}(\hat{y}_i - \bar{y})^2 = 45.59$,
$Q_e = \sum_{i=1}^{25}(y_i - \hat{y}_i)^2 = 18.82$
取$\alpha = 0.05$， 则$F_{0.95}(1, 23) = 4.28$,而$f = 57.56 > 4.28$,所以拒绝H_0 即认为回归方程反映了该商品销售总额y与其价格x之间的相关关系.
(4)同样可计算$R^2 = \frac{U}{S_{yy}} = 0.714$

第一节 一元线性回归模型

(5)利用回归方程对销售额进行预测

当$x = x_0 = 28.6$元时，则$\hat{y}_0 = 11.34$万元

对$\alpha = 0.05$,查表$t_{1-\frac{\alpha}{2}}(23) = 2.069$,可得$\delta(x_0) = 1.95$

所以y_0的置信水平为95%的预测区间

为$[(\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0))] = [9.39, 13.29]$，即把价格定在28.6元时，销售总额落在区间[9.39,13.29] 的概率为95%.

(6)若使销售额控制在[10,12]之间，则由

$$\begin{cases} 10 = 13.623 - 0.0798x_1 \\ 12 = 13.623 - 0.0798x_2 \end{cases}$$

得到x的范围为[20.34,45.40].

第一节　一元线性回归模型

六.可化为一元线性回归的模型

1. 双曲线方程:$\frac{1}{y} = a + \frac{b}{x}$;
令$y' = \frac{1}{y}, x' = \frac{1}{x}$,则有$y' = a + bx'$.

2. 幂函数方程:$y = ax^b$;
令$y' = \ln y, x' = \ln x, a' = \ln a$,则$y' = a' + bx'$.

3. 指数曲线方程:$y = ae^{bx}$;
令$y' = \ln y, a' = \ln a$,则$y' = a' + bx$.

第一节 一元线性回归模型

4. 指数曲线方程:$y = ae^{\frac{b}{x}}$;
   令$y' = \ln y, x' = \frac{1}{x}, a' = \ln a$,则$y' = a' + bx'$.

5. 对数曲线方程:$y = a + b\ln x$;
   令$x' = \ln x$,则$y = a + bx'$.

6. S型曲线方程:$y = \frac{1}{a+be^{-x}}$.
   令$y' = \frac{1}{y}, x' = e^{-x}$,则$y' = a + bx'$.

第二节　多元线性回归模型的参数估计

第二节　多元线性回归模型的参数估计

一.数学模型

假设随机变量y与k个变量$x_1, \cdots, x_k$之间存在下面的线性关系

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon,$$

其中ε是一个随机变量,满足$E\varepsilon = 0, D\varepsilon = \sigma^2(\sigma^2$为未知常数),称为随机误差, $\beta_0, \beta_1, \cdots, \beta_k$是未知参数.
设有n组独立的观测值$(y_i, x_{i1} \cdots x_{ik}), i = 1, \cdots, n$,则有

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1, \\ \quad \cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n, \end{cases}$$

成立.

第二节 多元线性回归模型的参数估计

或写成矩阵形式为

$$Y = X\beta + \varepsilon,$$

其中

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

这里Y表示随机变量y的n次观测值组成的列向量,称为观测向量,X的元素是k个自变量$x_1, \cdots, x_k$在n次观测中的取值, β称为未知参数向量,ε称为随机误差向量,满足

$$E(\varepsilon) = 0, \quad Cov(\varepsilon, \varepsilon) = \Sigma.$$

第二节　多元线性回归模型的参数估计

称模型

$$Y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad Cov(\varepsilon, \varepsilon) = \Sigma$$

为n元线性回归模型.

若

$$Y = X\beta + \varepsilon, \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n),$$

则称为多元正态线性回归模型.

第二节 多元线性回归模型的参数估计

二.参数β的估计

仍采用最小二乘法，求$y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik})$的平方和.
令$Q(\beta)=\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik})]^2$达到最小，若用矩阵表示，

$$Q(\beta)=(Y-X\beta)'(Y-X\beta)=Y'Y-2\beta'X'Y+\beta'X'X\beta$$

$$\frac{\partial Q}{\partial \beta}=-2X'Y+2X'X\beta=0$$

所以

$$X'X\beta=X'Y \tag{1}$$

若X是列满秩矩阵，$(X'X)^{-1}$存在，则$\hat{\beta}=(X'X)^{-1}X'Y$

第二节　多元线性回归模型的参数估计

正规方程组的解与β的最小二乘估计有以下关系：

定理：
(1)正规方程组的解必是β的最小二乘估计；
(2)β的最小二乘估计必是正规方程组的解.

因而$\hat{\beta}$即为β的最小二乘估计，$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)'$.

第二节 多元线性回归模型的参数估计

最小二乘估计的性质

性质1: $E(\hat{\beta}) = \beta$, $\hat{\beta}$是β的线性无偏估计.

性质2: $Cov(\hat{\beta}, \hat{\beta}) = \sigma^2(X'X)^{-1}$.

注: $\hat{\beta}$的各分量$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$一般并不独立.

$$S^{-1} := (X'X)^{-1} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{pmatrix}$$

第二节　多元线性回归模型的参数估计

$$Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}, \quad Var(\hat{\beta}_i) = \sigma^2 c_{ii}, i, j = 0, 1, ..., k.$$

选择使c_{ij}尽可能小的设计矩阵X. 如

$$c_{ij} = \delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \quad i, j = 0, 1, ..., k,$$

则称相应的X为正交的.

第二节 多元线性回归模型的参数估计

定义: 对任一$k+1$维向量$C=(c_0,c_1,...,c_k)'$, 若存在n维列向量L, 使$E(L'Y)=C'\beta$, 则称$C'\beta$为可估函数, 而可估函数$C'\beta$的最小方差线性无偏估计, 称为它的最好线性无偏估计(BLUE).

性质3(Guass-Markov定理): $C'\hat{\beta}$是$C'\beta$的最好线性无偏估计, 其中$\hat{\beta}$是β的最小二乘估计.

注: $C'\hat{\beta}$不一定是$C'\beta$的一切无偏估计中方差最小的.

性质4: 若$\varepsilon\sim N(0,\sigma^2 I_n)$, 则$C'\hat{\beta}$是$C'\beta$的UMVUE.

第二节　多元线性回归模型的参数估计

三.参数σ^2的估计

用残差向量$\hat{\varepsilon} = Y - X\hat{\beta}$来构造方差$\sigma^2$的估计.

性质5: (1) $E(\hat{\varepsilon}) = 0$, (2)$Cov(\hat{\varepsilon}, \hat{\varepsilon}) = \sigma^2 \left[I_n - X(X'X)^{-1}X'\right]$,

(3)$Cov(\hat{\beta}, \hat{\varepsilon}) = 0$. 残差平方和$Q_e = \|\hat{\varepsilon}\|^2 = \hat{\varepsilon}'\hat{\varepsilon}$.

第二节　多元线性回归模型的参数估计

$$
\begin{aligned}
Q_e &= \|\hat{\varepsilon}\|^2 = \hat{\varepsilon}'\hat{\varepsilon} \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
&= (AY)'(AY) = Y'AY \\
&= Y'Y - Y'X(X'X)^{-1}(X'X)(X'X)^{-1}X'Y \\
&= Y'Y - \hat{\beta}'X'X\hat{\beta} = Y'Y - \hat{Y}'\hat{Y}
\end{aligned}
$$

第二节　多元线性回归模型的参数估计

性质6: 记$\hat{\sigma}^2 = Q_e/(n-k-1)$, 称为残差方差, 则有$E(\hat{\sigma}^2) = \sigma^2$.

性质7: 若$\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I_n})$, 则
(1) $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$, $\hat{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{A})$ 且二者相互独立.

(2) $\hat{\beta}, Q_e$相互独立.　　(3) $\frac{Q_e}{\sigma^2} \sim \chi^2(n-k-1)$

(4) β的最小二乘估计$\hat{\beta}$也是β的极大似然估计, σ^2的极大似然估计为$\frac{Q_e}{n}$.

第三节　多元线性回归模型的假设检验

一.回归方程的显著性检验

提出假设:

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

(1)若接受H_0, 则表明诸变量与y之间确实无线性相关关系;

(2)若拒绝H_0,则认为回归方程是有意义的，但是这个结论只说明至少有一个$\beta_i \neq 0$,也就是说在所选自变量中，至少有一部分对y来说是重要的，但不表示所有自变量都是重要的.

第三节 多元线性回归模型的假设检验

1.平方和分解： $\hat{Y} = X\hat{\beta}$为n个试验点处Y的回归值, 总的变差平方和定义为

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = (Y - \mathbf{1}\bar{y})'(Y - \mathbf{1}\bar{y}) \\ &= \|Y - \mathbf{1}\bar{y}\|^2 = \|Y - \hat{Y} + \hat{Y} - \mathbf{1}\bar{y}\|^2 \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \mathbf{1}\bar{y}\|^2 + 2(Y - \hat{Y})'(\hat{Y} - \mathbf{1}\bar{y}) \end{aligned}$$

$$(Y - \hat{Y})'(\hat{Y} - \mathbf{1}\bar{y}) = (Y - \hat{Y})'\hat{Y} - (Y - \hat{Y})'\mathbf{1}\bar{y}$$

$$= (AY)'(PY) - (Y - \hat{Y})'\mathbf{1}\bar{y} = Y'APY - (Y - \hat{Y})'\mathbf{1}\bar{y} = 0.$$

$$S_{yy} = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \mathbf{1}\bar{y}\|^2 = Q_e + U.$$

$Q_e = \|Y - \hat{Y}\|^2$为剩余(残差)平方和, $U = \|\hat{Y} - \mathbf{1}\bar{y}\|^2$为回归平方和, 若$U >> Q_e$, 则拒绝$H_0$.

第三节 多元线性回归模型的假设检验

2.构造统计量:

(1) F统计量

当$\varepsilon \sim N(0, \sigma^2 I_n)$, $Q_e/\sigma^2 \sim \chi^2(n-k-1)$, $S_{yy}/\sigma^2 \overset{H_0}{\sim} \chi^2(n-1)$.

$$S_{yy} = Q_e + U, \quad Q_e = Y'AY, \quad r(A) = n-k-1,$$

由Cochran分解定理可证明$U/\sigma^2 \overset{H_0}{\sim} \chi^2(k)$, Q_e与U相互独立, 则选取

$$F = \frac{U/k}{Q_e/(n-k-1)} \sim F(k, n-k-1)$$

对给定的显著性水平α,拒绝域$W = \{F > F_{1-\alpha}(k, n-k-1)\}$.

第三节　多元线性模型的假设检验

(2) $R^2 = \frac{U}{S_{yy}}$, 称R^2为全相关系数.
它刻画了全体自变量$x_1, \ldots, x_k$对于因变量y的线性相关程度.R^2越大,越接近于1,说明上述线性相关程度越显著,R^2可作为衡量回归方程总效果的一个数量指标.

注：$F = \frac{n-k-1}{k} \frac{R^2}{1-R^2}$, 所以$F$检验与$R^2$检验是等价的.

第三节 多元线性模型的假设检验

二.回归系数的显著性检验

检验x_i对y的影响是否显著，等价于检验回归系数$H_{0i}:\beta_i=0$
(1)若接受H_{0i}，则表明x_i对y的影响相对于整个模型来说比较小；
(2)若拒绝H_{0i}，则表明x_i对y确有一定的影响，称x_i为显著因子.

当$\varepsilon\sim N(0,\sigma^2 I_n)$, 若记$S=(c_{ij})=(X'X)^{-1}$,则由性质知$\hat{\beta}_i\sim N(\beta_i,\sigma^2 c_{ii})$, 且$\hat{\beta}_i$与$Q_e$相互独立，其中$c_{ii}$表示矩阵$S$的第$i$个对角元.

第三节 多元线性模型的假设检验

1. T检验：选取

$$T = \frac{\hat{\beta}_i}{\sqrt{c_{ii}}} \Big/ \sqrt{\frac{Q_e}{n-k-1}} \overset{H_{0i}}{\sim} t(n-k-1),$$

对给定的显著性水平α, 拒绝域

$$W = \{|t| > t_{1-\alpha/2}(n-k-1)\}$$

2. F检验：选取

$$F = \frac{\hat{\beta}_i^2}{c_{ii}} \Big/ \frac{Q_e}{n-k-1} \overset{H_{0i}}{\sim} F(1, n-k-1)$$

对给定的显著性水平α, 拒绝域

$$W = \{F > F_{1-\alpha}(1, n-k-1)\}$$

第三节 多元线性模型的假设检验

三.偏回归平方和

自变量对y的影响,是指从回归方程剔除了这个自变量后所造成的影响，称回归平方和的减少部分为y对这个自变量的偏回归平方和.

若在$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$中剔除自变量$x_i$,不能简单地抹去这一项而得到

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{i-1} x_{i-1} + \hat{\beta}_{i+1} x_{i+1} + \cdots + \hat{\beta}_k x_k,$$

应该重新估计回归系数,建立新的回归方程

$$\hat{y}^* = \hat{\beta}_0^* + \hat{\beta}_1^* x_1 + \cdots + \hat{\beta}_{i-1}^* x_{i-1} + \hat{\beta}_{i+1}^* x_{i+1} + \cdots + \hat{\beta}_k^* x_k.$$

第三节 多元线性模型的假设检验

一般地$\hat{\beta}_j^* \neq \hat{\beta}_j$, 可以证明

$$\hat{\beta}_j^* = \hat{\beta}_j - \frac{c_{ij}}{c_{ii}}\hat{\beta}_i, i \neq j,$$
$$\hat{\beta}_0^* = -\sum_{j \neq i} \hat{\beta}_j^* \bar{x}_j,$$

且y对x_i的偏回归平方和$U_i = \frac{\hat{\beta}_i^2}{c_{ii}}$.

第三节　多元线性模型的假设检验

注:

(1) 回归系数显著性检验的F统计量的分子即为偏回归平方和. 偏回归平方和越大, 此变量对y 的影响越显著.

(2) 得到回归方程后, 计算每个变量的偏回归平方和, 对偏回归平方和最小的变量, 如果相应的回归系数检验又不显著, 可将此变量剔除.

第三节 多元线性模型的假设检验

四.“最优”回归方程的选择

最优回归方程: 如回归方程中包含所有对y有显著影响的自变量, 不包含对y影响不显著的自变量, 同时在同类方程中残差平方和Q_e达到最小, 则称此回归方程为最优的.

(1) 全部比较法: 从所有可能的自变量组合的回归方程中选择最优者.
注: 总可找到最优方程; 但计算量大.

(2) 只出不进法: 从包含全部自变量的回归方程中逐个剔除不显著的自变量, 直到回归方程中所含自变量全部都是显著的为止. 首先考虑含所有自变量的回归方程, 剔除不显著自变量中偏回归平方和最小的, 再对其中的每个自变量进行显著性检验, 继续剔除, 直到所有自变量都显著.

第三节 多元线性模型的假设检验

四.“最优”回归方程的选择

注: 每剔除一次自变量就得重新计算回归系数, 考虑自变量不多时, 不显著自变量不多时, 可采用. 不显著自变量较多时, 计算量大.

(3) 只进不出法: 从一个自变量开始, 把显著的自变量逐个引入回归方程, 直到在余下的自变量中选出一个与已引入的自变量一起组成回归方程有最大偏回归平方和的自变量, 至经检验为不显著, 因而不被引入时为止.
注: 计算量少, 但不一定能得到最优方程. 由于自变量间的相关关系, 引入新的自变量后, 使原来引入的自变量成为不显著的.

第三节 多元线性模型的假设检验

四.“最优”回归方程的选择

(4) 逐步回归法: 综合方法(2)、(3), 将自变量按其对y的影响一个个引入, 同时每引入一个新的自变量, 即对原已引入的自变量逐个检验, 将不显著的剔除, 直到回归方程再也不能引入新的自变量, 同时也不能剔除任一自变量为止.

注: 有计算技巧, 计算量相对较小, 有较好的计算程序.

第四节　非线性回归模型

第四节 非线性回归模型

多项式回归

若随机变量y与自变量x之间的相关系数为

$$\begin{cases} y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon; \\ (\varepsilon \sim N(0, \sigma^2)). \end{cases}$$

称此模型为(正态)多项式回归模型.

只需令$x_i = x^i, i = 1, \cdots, k$，则可转化为多元线性回归模型

$$\begin{cases} Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon; \\ (\varepsilon \sim N(0, \sigma^2)). \end{cases}$$

第四节 非线性回归模型

例： 某种合金钢中的两种主要成分之和x与它的膨胀系数y之间有一定的数量关系，给出实验所得的13 组数据，求y与x的回归方程.

x	37	37.5	38	38.5	39	39.5	40	40.5	41	41.5	42	42.5	43
y	3.4	3	3	3.27	2.1	1.83	1.53	1.70	1.8	1.9	2.35	2.54	2.9

解： 先画散点图

设回归方程为$y = \beta_0 + \beta_1 x + \beta_2 x^2$

令$x_1 = x, x_2 = x^2$,则可确定回归方程为$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$,

经计算，$\hat{\beta}_0 = 257.063, \hat{\beta}_1 = -12.620, \hat{\beta}_2 = 0.156$

所以y对x的多项式回归方程为

$$\hat{y} = 257.063 - 12.620x + 0.156x^2$$

第五节　单因子试验方差分析

第五节　单因子试验方差分析

一.基本概念

1.指标–试验的结果(如产品的性能，质量，产量等)，用Y表示；

2.因子–试验中变化的因素(即影响指标的原因)，用A, B, C表示；

3.水平–因子在试验中所处的不同状态，如因子A有p个水平，用$A_1, A_2, \cdots, A_p$表示；

4.若试验中只有一个因素在变化，其他条件不变，则称为单因子试验，处理单因子试验的统计推断方法称为单因子方差分析.

第五节 单因子试验方差分析

例1：5个水稻产品比较试验，在成熟期随机抽取样本测定产量，每个品种取3个点，结果如下表

A	1	2	3	$\overline{Y}_{i.}$
A_1	41	39	40	40
A_2	33	37	35	35
A_3	38	35	35	36
A_4	37	39	38	38
A_5	31	34	34	33

指标：产量；因子：品种；水平：A_1, A_2, A_3, A_4, A_5

第五节　单因子试验方差分析

由表中数据我们可以分析:
(1)在同一水平$A_i(i=1,\cdots,5)$下，生产的条件虽然相同，但产量却有所不同，产生这种差异的原因是由于试验过程中随机因素的干扰及测量误差所致，称这类差异为随机误差或试验误差.说明试验结果是一个随机变量.
(2)5个不同的品种，从平均产量来看，它们是参差不齐的，其原因主要是由于品种的不同引起的差异（除了随机波动），称这类差异为系统误差.
(3)对同一品种进行3次重复试验的结果可看成是取自同一个总体的样本，表中的5组数据可以看成是来自5个不同总体的样本，记这些总体为$Y_1, Y_2, \cdots, Y_5$,每个试验结果记为$Y_{ij}, i=1,\cdots,5, j=1,2,3$

第五节　单因子试验方差分析

通常假定：$1^*.Y_i \sim N(\mu_i, \sigma^2), i = 1, \cdots, 5,$(其它试验条件不变，因而认为所有试验的方差是相同的).

$2^*.Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$是来自$Y_i$的样本，$n_i = 3, i = 1, \cdots, 5,$且$Y_1, Y_2, \cdots, Y_5$相互独立.

(4)设因子A有a个水平，每个水平A_i重复n_i次，若重复数n_i全相等，则称这类试验为等重复的单因子试验；反之，则称为不等重复的单因子试验.

(5)本例分析判断5个不同品种的产量之间的差异主要是由随机误差还是由于不同品种造成的问题，可归结为判定5个正态总体的均值是否相等的问题.若5个正态总体的均值相等，则认为产量之间的差异是由随机误差造成的；否则，认为产量之间的差异是由不同品种造成的.

第五节 单因子试验方差分析

二.数学模型

设因子A有p个不同水平$A_1, \cdots, p_p$,它们对应的总体$Y_1, \cdots, Y_p$相互独立，且$Y_i \sim N(\mu_i, \sigma^2), i = 1, \cdots, p$. 在水平$A_i$下进行$n_i$次独立观测，获得容量为$n_i$的一个样本$Y_{i1}, Y_{i2}, \cdots, Y_{in_i}, i = 1, \cdots, p$

水平	总体	样本
A_1	$Y_1 \sim N(\mu_1, \sigma^2)$	$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$
A_2	$Y_2 \sim N(\mu_2, \sigma^2)$	$Y_{21}, Y_{22}, \cdots, Y_{2n_2}$
$\vdots$	$\vdots$	$\vdots$
A_p	$Y_p \sim N(\mu_p, \sigma^2)$	$Y_{p1}, Y_{p2}, \cdots, Y_{pn_p}$

第五节　单因子试验方差分析

令$\varepsilon_{ij} = Y_{ij} - \mu_i$,则

$$\begin{cases} Y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \cdots, p, j = 1, \cdots, n_i; \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且}\varepsilon_{ij}\text{相互独立}. \end{cases}$$

为了找因子各水平对试验指标的影响，将μ_i分解.
令$\mu = \frac{1}{n}\sum_{i=1}^{p} n_i\mu_i$，$n = \sum_{i=1}^{p} n_i$，$\alpha_i = \mu_i - \mu, i = 1, \cdots, p$
其中μ为所有Y_{ij}的总的平均值，α_i为第i个水平对试验指标的效应，简称为水平A_i的效应，它反映了因子的第i个水平A_i对试验指标作用的大小. 可以验证：$\sum_{i=1}^{p} n_i\alpha_i = \sum_{i=1}^{p} n_i(\mu_i - \mu) = 0$.
于是有

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \cdots, p, j = 1, \cdots, n_i; \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且}\varepsilon_{ij}\text{相互独立}; \\ \sum_{i=1}^{p} n_i\alpha_i = 0. \text{称为单因子方差分析模型}. \end{cases}$$

第五节　单因子试验方差分析

三.统计分析

判定因子A的p个水平下均值是否相等，归结为检验假设

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$$

或

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p$$

是否成立

1.显著性检验

记$\overline{Y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$,表示所有$Y_{ij}$的总平均值.

$\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$,表示第$i$个水平下的样本均值.

考虑统计量$S_T = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$ 称为总离差平方和，反映全部试验数据之间的差异(离散程度).

第五节 单因子试验方差分析

将S_T分解：

$$\begin{aligned} S_T &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}[(Y_{ij}-\overline{Y}_{i\cdot})+(\overline{Y}_{i\cdot}-\overline{Y})]^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\cdot})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot}-\overline{Y})^2 \\ &= S_E + S_A \end{aligned}$$

其中$S_E = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i\cdot})^2$反映了在相同条件下各次试验的差异,称为误差平方和或组内平方和；$S_A = \sum_{i=1}^{p} n_i(\overline{Y}_{i\cdot}-\overline{Y})^2$反映了来自不同总体的样本之间的差异,也就是反映了因子各水平效应α_i的影响,称为组间平方和, S_A也与试验误差有关.

第五节　单因子试验方差分析

(1)若H_0成立，则$\alpha_i = 0, i = 1, \cdots, p$,模型变为

$$\begin{cases} Y_{ij} = \mu + \varepsilon_{ij}, i = 1, \cdots, p, j = 1, \cdots, n_i; \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且}\varepsilon_{ij}\text{相互独立} \end{cases}$$

这时S_T表示仅由随机误差ε_{ij}所引起的偏差.

(2)若H_0不成立，则S_T中除包含由随机误差ε_{ij}所引起的偏差外，还应包含由α_i不全为0所引起的偏差.

若能把S_T中由ε_{ij}所引起的偏差和因子α_i不全为0所引起的偏差分开，并选取适当的统计量作为衡量它们之间差异的度量尺度，就可以检验假设H_0.

第五节 单因子试验方差分析

可以推出

$$\frac{S_A}{\sigma^2} \overset{H_0}{\sim} \chi^2(p-1), \frac{S_E}{\sigma^2} \sim \chi^2(n-p)$$

且S_A和S_E相互独立.
选取统计量

$$F = \frac{S_A/(p-1)}{S_E/(n-p)} \overset{H_0}{\sim} F(p-1, n-p)$$

拒绝域为

$$W = \{F > F_{1-\alpha}(p-1, n-p)\}$$

若拒绝H_0,则认为因子A的p个水平效应之间有显著性差异；否则认为因子A的p个水平效应之间没有显著性差异.

第五节 单因子试验方差分析

方差来源	平方和S	自由度f	均方$\overline{S}$	F值	显著性
因子A	$S_A = \sum_{i=1}^{p} n_i(\overline{Y}_{i\cdot} - \overline{Y})^2$	$p-1$	$\overline{S}_A = \frac{S_A}{p-1}$	$F = \frac{\overline{S}_P}{\overline{S}_E}$	
误差	$S_E = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$	$n-p$	$\overline{S}_E = \frac{S_E}{n-p}$		
总和	$S_T = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2$	$n-1$			

第五节 单因子试验方差分析

例2： 研究5个品种产量之间是否有显著性差异($\alpha = 0.1$)

A	1	2	3	$\overline{Y}_{i.}$
A_1	41	39	40	40
A_2	33	37	35	35
A_3	38	35	35	36
A_4	37	39	38	38
A_5	31	34	34	33

第五节 单因子试验方差分析

解： $H_0 : \mu_1 = \cdots = \mu_5, F = \frac{S_A/(5-1)}{S_E/(15-5)} \sim F(4,10)$

$S_A = \sum_{i=1}^{p} \frac{1}{n_i}(\sum_{j=1}^{n_i} Y_{ij})^2 - \frac{1}{n}(\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij})^2 = 19962 - 19874.4 = 87.6$

$S_E = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij})^2 - \sum_{i=1}^{p} \frac{1}{n_i}(\sum_{j=1}^{n_i} Y_{ij})^2 = 19986 - 19962 = 24$

$S_T = S_A + S_E = 111.6$

方差来源	平方和S	自由度f	均方$\overline{S}$	F值	显著性
因子A	$S_A = 87.6$	4	21.9	9.13	
误差	$S_E = 24$	10	2.4		
总和	$S_T = 111.6$	14			

对于$\alpha = 0.1$,查表$F_{0.9}(4,10) = 2.641 < 9.13$,所以拒绝$H_0$,说明5个品种有显著差异.

第五节 单因子试验方差分析

2.参数估计

p个水平效应之间有显著差异，也就是说$\mu_1, \cdots, \mu_p$不完全相同，还需对每一对μ_i, μ_j之间的差异程度作出估计，也就是对效应之差$\mu_i - \mu_j$进行区间估计.

(1)$E(\overline{Y}_{i\cdot}) = \mu + \alpha_i, \quad i = 1, 2, \cdots, p, \quad E(\overline{Y}) = \mu$. 所以$\hat{\mu} = \bar{Y}, \hat{\alpha}_i = \overline{Y}_{i\cdot} - \overline{Y}$分别是$\mu$和$\alpha_i$的无偏估计.

(2) $Y_i \sim N(\mu_i, \sigma^2), Y_j \sim N(\mu_j, \sigma^2)$ $(i \neq j)$,求均值差$\mu_i - \mu_j = \alpha_i - \alpha_j$的区间估计.

第五节 单因子试验方差分析

$\overline{Y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{n_i}), i \neq j$时，$\overline{Y}_{i\cdot}$ 与$\overline{Y}_{j\cdot}$相互独立.
所以

$$\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot} \sim N(\mu_i - \mu_j, (\frac{1}{n_i} + \frac{1}{n_j})\sigma^2)$$

又因为$S_E/\sigma^2 \sim \chi^2(n-p)$,且$\hat{\sigma}^2 = \frac{S_E}{n-p}$

$$\frac{(\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}) - (\alpha_i - \alpha_j)}{\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\hat{\sigma}} \sim t(n-p)$$

于是均值差$\mu_i - \mu_j = \alpha_i - \alpha_j$的置信水平为$(1-\alpha)$置信区间为

$$[\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot} - \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\hat{\sigma} t_{1-\frac{\alpha}{2}}(n-p), \overline{Y}_{i\cdot} - \overline{Y}_{j\cdot} + \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\hat{\sigma} t_{1-\frac{\alpha}{2}}(n-p)]$$

第五节　单因子试验方差分析

注：

(1)若置信区间包含0，则以$(1-\alpha)$概率认为μ_i与μ_j没有显著差异；

(2)若置信区间上限小于0，则以$(1-\alpha)$概率认为$\mu_i < \mu_j$

(3)若置信区间下限大于0，则以$(1-\alpha)$概率认为$\mu_i > \mu_j$

第五节 单因子试验方差分析

例3： (续例2)由$\overline{Y}_{1\cdot}=40$, $\overline{Y}_{2\cdot}=35$, $\overline{Y}_{3\cdot}=36$, $\overline{Y}_{4\cdot}=38$, $\overline{Y}_{5\cdot}=33$及$\overline{Y}=36.4$, 可以算出因子A各个效应的估计为：
$\hat{\alpha}_1=\overline{Y}_{1\cdot}-\overline{Y}=3.6$, $\hat{\alpha}_2=-1.4$, $\hat{\alpha}_3=-0.4$, $\hat{\alpha}_4=1.6$, $\hat{\alpha}_5=-3.4$及σ^2的无偏估计为$\hat{\sigma}^2=2.4$.
均值差$\mu_i-\mu_j=\alpha_i-\alpha_j$的置信水平为95%置信区间如下表所示, $i,j=1,2,3$ 这里$t_{0.975}(10)=2.2281$, $n_i=n_j=5$. 这些置信区间均包含0, 也说明不同品种对产量没有显著影响.

$\mu_i-\mu_j$	$\overline{Y}_{i\cdot}-\overline{Y}_{j\cdot}$	置信区间
$\mu_1-\mu_2$	5	[-1.9035, 11.9035]
$\mu_1-\mu_3$	4	[-2.9035, 10.9035]
$\mu_2-\mu_3$	-1	[-7.9035, 5.9035]

第五节　单因子试验方差分析

3.当观测值过大或过小，可以经过线性变换使计算简单，令$Y'_{ij}=\frac{Y_{ij}-c}{b}, b\neq 0$, 用$Y'_{ij}$与$Y_{ij}$计算的$F$值是相同的.

第五章　非参数假设检验

非参数假设检验

若假设$H_0: F(x) = F_0(x;\theta)$, 其中$F_0(x;\theta)$为一个指定的分布,$\theta$是参数向量.
(1)若$\theta = \theta_0$已知, 即总体分布完全确定,此时H_0称为**简单假设**.
(2)若θ部分或完全未知,即$F_0(x;\theta)$形式上确定,但含有未知参数,此时H_0称为**复合假设**.

当假定某一理论分布$F_0(x;\theta)$,实际数据$x_1, \cdots, x_n$与理论分布$F_0(x;\theta)$偏离的量用$\Delta(x_1, \cdots, x_n; F)$表示, 规定一个界限$\Delta_0$,若$\Delta$超过这个界限$\Delta_0$,则认为理论分布与数据$x_1, \cdots, x_n$ 不符,因而拒绝H_0,否则接受H_0. 这个Δ就称为**“拟合优度”**(Goodness of Fit), 这种检验称为**“拟合优度检验”**

χ^2拟和优度检验

一.χ^2拟和优度检验

1.简单假设

设总体分布为$F(x)$.$(X_1, \cdots, X_n)$为取自总体的样本,提出假设$H_0 : F(x) = F_0(x; \theta_0)$, θ_0为已知参数。

(1)选取$r-1$个实数$-\infty < y_1 < y_2 < \cdots < y_{r-1} < +\infty$,它们将随机变量$X$的一切可能取值的集合分为$r$个区间,并用$n_i$表示样本观测值落入第$i$个区间$(y_{i-1},\ y_i]$的**观测频数** (这里设$y_0 = -\infty$, $y_r = +\infty$). $\sum_{i=1}^{r} n_i = n$.

(2)在H_0为真下,则由给定的分布函数$F_0(x; \theta_0)$可以求出$p_i = F_0(y_i; \theta_0) - F_0(y_{i-1}; \theta_0),\ i = 1, 2, \cdots, r$.

$\sum_{i=1}^{r} p_i = 1$,称np_i为样本落入第i个小区间的**理论频数**.

χ^2拟和优度检验

(3)考虑统计量$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$，它表示实际观测频数$n_i$与理论频数$np_i$的相对差异的总和. 由Pearson定理:$\chi^2 \sim \chi^2(r-1)$,因此当$n$充分大时,对给定的显著性水平$\alpha$,检验的拒绝域为

$$W = \{\chi^2 \geq \chi^2_{1-\alpha}(r-1)\}.$$

χ^2拟和优度检验

2.复合假设

$H_0 : F(x) = F_0(x;\theta)$,$\theta$未知,为$s$维向量,$\theta = (\theta_1, \cdots, \theta_s)$

(1)由MLE法得$\hat{\theta}$ 代替未知参数,

(2)$\hat{p}_i = F_0(y_i;\hat{\theta}) - F_0(y_{i-1};\hat{\theta}), i = 1, 2, \cdots, r$

(3)$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2(r - s - 1)$

χ^2拟和优度检验

3.假设检验步骤

(1)将观测值(数据)分为r个互不相容的区间，算出n_i,每个区间至少有5个样本，区间长度可以不一样。

(2)在H_0为真下，用MLE估计法去估计分布中所含的未知参数。

(3)在H_0为真下，计算$p_i = F(y_i) - F(y_{i-1})$和$np_i, i = 1, \cdots, r.$

(4)计算$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$

(5)对给定的显著性水平α,检验的拒绝域为$W = \{\chi^2 \geq \chi^2_{1-\alpha}(r-s-1)\}$，查表得临界值，$s$为未知参数的个数。

(6)若$\chi^2 > \chi^2_{1-\alpha}$,则拒绝$H_0$,否则接受$H_0$。

χ^2拟和优度检验

注：(1)χ^2拟和优度检验必需在大样本下进行。
(2)要求$np_i \geq 5$
(3)在简单假设检验中，分区间时最好各区间概率相同。

例1. 有一正二十面体的20个面上分别标以数字$0, 1, \cdots, 9$,每个数字在两个对称的面上标出，为检验其均匀性，共做800次投掷试验，数字$0, 1, \cdots, 9$朝上的次数为

数字	0	1	2	3	4	5	6	7	8	9
频数	74	92	83	79	80	73	77	75	76	91

问：该正20面体是否均匀？$(\alpha = 0.05)$

χ^2拟和优度检验

解: H_0:该正20面体均匀,
即$p_i = P(X = i) = \frac{1}{10}, i = 0, 1, \cdots, 9$,则$np_i = 80$

数字	n_i	np_i	$n_i - np_i$	$(n_i - np_i)^2$
0	74	80	-6	36
1	92	80	12	144
2	83	80	3	9
3	79	80	-1	1
4	80	80	0	0
5	73	80	-7	47
6	77	80	-3	9
7	75	80	-5	25
8	76	80	-4	16
9	91	80	11	121
$\sum$	800			410

$$\chi^2 = \sum_{i=1}^{10} \frac{(n_i - np_i)^2}{np_i} = \frac{1}{80} \times 410 = 5.125,$$

对$\alpha = 0.05$,查表$\chi^2_{0.95}(10-1) = 16.9$,因为$\chi^2 < \chi^2_{0.95}(9)$,所以接受$H_0$,即认为该正20面体是均匀的。

χ^2拟和优度检验

例2. 遗传学中常常有考虑拟合检验的例子.例如某种动物身上的毛可分成三种类型:极卷,中等卷曲,正常,而毛的卷曲由二个遗传基因F, f所控制,(F, F)的后代身上的毛是极卷的,(F, f)的后代是中等卷曲,(f, f)则为正常,并且两个基因随机结合,因此极卷,中等卷曲,正常的比例应是$1:2:1$.现在进行了93次试验,所得下面结果.

极卷	中等卷曲	正常
23	50	20

设$p_1 = P\{\text{后代的毛是极卷的}\}$, $p_2 = P\{\text{后代的毛中等卷曲}\}$, $p_3 = P\{\text{后代的毛正常}\}$,

χ^2拟和优度检验

则假设检验为

$$H_0: \ p_1 = p_{10} = \frac{1}{4}, \ p_2 = p_{20} = \frac{1}{2}, \ p_3 = p_{30} = \frac{1}{4}$$
$$\leftrightarrow \quad H_1: \ \text{至少有一个} p_i \neq p_{i_0}.$$

$$\chi^2 = \frac{(23 - 93 \times \frac{1}{4})^2}{93 \times \frac{1}{4}} + \frac{(50 - 93 \times \frac{1}{2})^2}{93 \times \frac{1}{2}} + \frac{(20 - 93 \times \frac{1}{4})^2}{93 \times \frac{1}{4}} = 0.72,$$

对$\alpha = 0.05$, $\chi^2_{0.95}(2) = 5.991$ $\chi^2 = 0.72 < \chi^2_{0.95}(2)$,因此不能拒绝原假设,即毛的卷曲程度是由遗传基因$(F, F)$, (F, f)和(f, f)所控制的遗传学理论是站得住脚的.

χ^2拟和优度检验

例3.电话交换台在某一小时内接到用户的呼唤次数，按每分钟计

呼叫次数n_i	0	1	2	3	4	5	6	≥ 7
频数n_i	8	16	17	10	6	2	1	0

问：电话每分钟呼叫次数是否服从泊松分布？($\alpha = 0.05$)

解：由MLE得$\hat{\lambda} = \bar{X}$,由样本观测值得

$$\bar{x} = (0 \times 8 + 1 \times 16 + \cdots + 6 \times 1)/60 = 2$$

$H_0 : \hat{p}_i = P(X = i) = \frac{2^i e^{-2}}{i!}, i = 0, 1, \cdots$
在H_0为真下，每分钟接到呼唤次数的理论频数

$$n\hat{p}_i = 60 \times \frac{2^i e^{-2}}{i!}, i = 0, 1, \cdots$$

χ^2拟和优度检验

i	n_i	$n\hat{p}_i$	$n_i - n\hat{p}_i$	$(n_i - n\hat{p}_i)^2$	$(n_i - n\hat{p}_i)^2/n\hat{p}_i$
0	8	8.1204	-0.1204	0.0145	0.0018
1	16	16.2402	-0.2402	0.0577	0.0036
2	17	16.2402	0.7598	0.5773	0.0355
3	10	10.8264	-0.8264	0.6829	0.0631
4	6	5.4134			
5	2	2.1654	0.4932	0.2432	0.0286
6	1	0.7218			
7	0	0.2062			
$\sum$	60				0.1326

$\chi^2 = \sum_{i=1}^{5} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 0.133, r = 5, s = 1$,对$\alpha = 0.05$,查表$\chi^2_{0.95}(3) = 7.815$,因为$\chi^2 < \chi^2_{0.95}(3)$,所以接受$H_0$,即认为每分钟呼叫次数服从参数为2的泊松分布。

列联表的独立性检验

二.列联表的独立性检验

“对所考察的总体中每一个元素同时测定两个指标X, Y,要检验这两个指标是否有关.”

例:考虑对某种疾病的几种治疗方法与治疗结果之间的关系.
将n个病人按不同的治疗方法(第一个指标)分组,观察各组内病人的不同效果(第二个指标).设X可能取值为$1, 2, \cdots, p$, Y可能取值为$1, 2, \cdots, q$, 现在对(X, Y) 进行了n次独立观测而得$(x_1, y_1), \cdots, (x_n, y_n)$, 用$n_{ij}$表示样本观测值中“X取$i$, Y取j”的次数. 检验假设

$$H_0 : X\text{与}Y\text{独立}.$$

把数据排列成表的形式,这种表称为**列联表**

列联表的独立性检验

X \ Y	1	2	$\cdots$	q	
1	n_{11}	n_{12}	$\cdots$	n_{1q}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2q}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
p	n_{p1}	n_{p2}	$\cdots$	n_{pq}	$n_{p\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot q}$	n

其中$n_{i\cdot} = \sum_{j=1}^{q} n_{ij}$, $n_{\cdot j} = \sum_{i=1}^{p} n_{ij}$, $n = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$.

检验X与Y是否相互独立$\Leftrightarrow$ $H_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$,对于所有$(i,j)$ 都成立$\leftrightarrow$ $H_1 : p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j}$,对于某个$(i,j)$ 成立

列联表的独立性检验

由MLE得

$$\begin{cases} \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, & i = 1, \cdots, p; \\ \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}, & j = 1, \cdots, q. \end{cases}$$

选取统计量

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = n(\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1) \sim \chi^2((p-1)(q-1))$$

其中$\hat{n}_{ij} = n \times \hat{p}_{ij} = n \times \hat{p}_{i\cdot} \times \hat{p}_{\cdot j} = n \times \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n}$. 对给定的$\alpha$，拒绝域$W = \{\chi^2 \geq \chi^2_{1-\alpha}((p-1)(q-1))\}$.

列联表的独立性检验

例5. 某校甲乙两班进行某种技能训练，测验成绩按优，良，及格及不及格四级给分，结果如下表，问成绩与班级有无关系？($\alpha = 0.05$)

班级	优	良	及格	不及格	合计
甲	14	20	15	11	60
乙	18	10	20	12	60
合计	32	30	35	23	120

列联表的独立性检验

解:H_0:成绩与班级无关。
在H_0为真下，理论频数表如下，$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$

班级	优	良	及格	不及格	合计
甲	16	15	17.5	11.5	60
乙	16	15	17.5	11.5	60
合计	32	30	35	23	120

$$\chi^2 = n\sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(n_{ij}-\hat{n}_{ij})^2}{n_{i\cdot} n_{\cdot j}}$$

$$= \frac{(14-16)^2}{16} + \frac{(20-15)^2}{15} + \cdots + \frac{(12-11.5)^2}{11.5} = 4.592$$

对$\alpha = 0.05$，$p = 2, q = 4$,查表得$\chi^2_{0.95}(3) = 7.815$,所以$W = \{\chi^2 > 7.815\}$。由于$\chi^2 < \chi^2_{0.95}(3)$,因此接受$H_0$,即在显著性水平0.05下，认为成绩与班级无关。

列联表的独立性检验

注: (1)列联表检验独立性时，实际上是χ^2拟和优度检验中χ^2检验统计量极限定理的应用。

(2)也可用于连续型，将变量值分成若干个互不相容的区间。

(3)当$p = q = 2$时，得到2×2列联表，也叫四格表，用a, b, c, d表示观测值。

	1	2	$\sum$
1	a	b	$a+b$
2	c	d	$c+d$
$\sum$	$a+c$	$b+d$	n

有一简便的计算公式，

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(c+d)(a+b)} \sim \chi^2(1)$$

列联表的独立性检验

例6.调查339名50岁以上吸烟习惯与患慢性气管炎的关系，得下表，问吸烟与患慢性气管炎是否有关？($\alpha = 0.01$)

	患	不患	$\sum$
吸烟	43	162	205
不吸烟	13	121	134
$\sum$	56	283	339

解:设H_0 :吸烟与患慢性气管炎无关，由四格表检验

$$\chi^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(c+d)(a+b)} = 7.469$$

对$\alpha = 0.01$，查表$\chi^2_{0.99}(1) = 6.635$,因为$\chi^2 > \chi^2_{0.99}(1)$,所以拒绝$H_0$,即认为吸烟与患慢性气管炎有关。