第五章 样本及统计量

§5.1 总体(population)与样本(sample)

1 总体、个体与总体容量

（1）把被研究的对象的全体叫做**总体**。

（2）总体中各个研究对象称为**个体**，

（3）总体中所包含的个体数称为**总体容量**。

（4）容量有限的总体称为**有限总体**，

容量无限的总体称为**无穷总体**。

实例

在研究2000名学生的年龄时, 这些学生的年龄的全体就构成一个总体, 每个学生的年龄就是个体.

总体X 即研究对象的某项数量指标 X ， 其取值在客观上有一定的分布， X是一个随机变量.

2 样本、样本容量与简单随机样本

（1）从总体中抽取一部分个体（即对r.v.X进行若干次试验)所构成的叫**样本**。

（2）样本中所包含的个体数称为**样本容量**。

（3）由总体中取出样本的过程称为**抽样**。

为使样本具有充分的代表性，
①抽样必须是随机的，
②抽样必须是独立的。

（4）这种抽样方法叫做简单随机抽样，得到的样本叫做简单随机样本。

Notes

简单随机样本满足：

（1）随机变量$X_1, X_2, \cdots, X_n$是独立的，

（2）且与总体X服从相同的分布。

$$X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} F(x)$$

总体分布（分布函数、分布律或分布密度）、数字特征或参数

从总体中取出简单随机样本（样本来自总体）

根据样本观测之作出估计推断（样本代表总体）

样本观测值的分布函数、频率分布直方图、数字特征或参数

3 样本的联合分布

若总体X是离散型的随机变量，分布函数为F(*x*)，

分布律为 $P\{X=x_i\}=p(x_i),$

则样本 $X_1, X_2, \cdots, X_n$ 的联合分布函数为：

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

联合分布率为：

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}=\prod_{i=1}^{n}p(x_i).$$

若总体X是连续型的随机变量
分布函数为F(x)，分布密度为$p(x)$,

则样本$X_1, X_2, \cdots, X_n$的联合分布函数为：

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

联合分布密度为：

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

例 设总体$X \sim B(1, p)$,求样本$X_1, \cdots, X_n$的联合分布律。

解 $\because X \sim B(1, p)$,分布律

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

$\therefore$ **联合分布律：**

$$P\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\} = \prod_{i=1}^{n} p(x_i).$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}.$$

4. 样本观测值的分布函数（经验分布）

从总体中抽取容量为n的样本，得到n个样本观测值，列表

样本观测值	频数	频率
$x_{(1)}$	n_1	f_1
$x_{(2)}$	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
$x_{(k)}$	n_k	f_k

其中$x_{(1)} < x_{(2)} < \cdots < x_{(k)} (k \le n)$

$$f_i = \frac{n_i}{n}, \quad \sum_{i=1}^{k} n_i = n, \quad \sum_{i=1}^{k} f_i = 1.$$

$$P\{X^* = x_{(i)}\} = \frac{1}{n}$$

$$F_n^*(x) = P\{X^* \le x\} = \sum_{x_{(i)} \le x} P\{X^* = x_{(i)}\} = \sum_{x_{(i)} \le x} \frac{1}{n}$$

样本观测值的分布函数

$$F_n^*(x)=P\{X^*\le x\}=\sum_{x_{(i)}\le x}P\{X^*=x_{(i)}\}=\sum_{x_{(i)}\le x}\frac{1}{n}$$

$$F_n^*(x)=\begin{cases}0,\text{当}x<x_{(1)};\\ \sum\limits_{x_{(i)}<x}f_i,\text{当}x_{(i)}\le x<x_{(i+1)};\\ 1,\text{当}x\ge x_k.\end{cases}$$

在样本容量较大时，可用样本观测值的

的分布函数$F_n^*(x)$来估计总体X的分布函数F(x).

由贝努利大数定理：$\lim_{n\to\infty} P\{|F_n^*(x)-F(x)|<\varepsilon\}=1$

经验分布函数依概率收敛于总体分布函数.

即经验分布函数是总体分布函数的近似.

例 从总体X中随机抽取8个观测值为45，46，48，51，51，64，57，62，写出样本观测值的分布函数。

解 大小重新排列 45<46<48<51=51<57<62<64

$$\text{当}x<45\text{时，}\ F_n^*(x)=0,$$

$$\text{当}45\le x<46\text{时，}\ F_n^*(x)=\frac{1}{8},(\text{仅有}45)$$

$$\text{当}46\le x<48\text{时，}\ F_n^*(x)=\frac{2}{8},(\text{有}45,46)$$

$$\text{当}48\le x<51\text{时，}\ F_n^*(x)=\frac{3}{8},(\text{有}45,46,48)$$

当$51 \le x < 57$时，$F_n^*(x) = \frac{5}{8}$,(有45,46,48,51,51)

当$57 \le x < 62$时，$F_n^*(x) = \frac{6}{8}$,(45,46,48,51,51,57)

当$62 \le x < 64$时，$F_n^*(x) = \frac{7}{8}$,

(45,46,48,51,51,57,62)

当$64 \le x$时，$F_n^*(x) = \frac{8}{8} = 1.$

5. 样本观测值的频率分布直方图

从总体中抽取容量为n的样本，得到n个样本观测值，列表

样本观测值	频数	频率
$x_{(1)}$	n_1	f_1
$x_{(2)}$	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
$x_{(k)}$	n_k	f_k

其中$x_{(1)} < x_{(2)} < \cdots < x_{(k)} (k \le n)$

$$f_i = \frac{n_i}{n}, \sum_{i=1}^{k} n_i = n, \sum_{i=1}^{k} f_i = 1.$$

样本观测值 $x_1,x_2,\ldots,x_n$

(1) 排序$x_{(1)}\le x_{(2)} \le \ldots \le x_{(n)}$

(2) 分组 $a<x_{(1)}$, $b>x_{(n)}$

组数m　$5\le m\le 15$　　组间距$\Delta_i=(b-a)/m$

(3) 统计

	$[a,a_1)[a_1,a_2)\ldots[a_{m-1},b)$
频数	n_1　n_2　　n_m
频率	n_1/n　n_2/n　n_m/n

(4)作图 以组间距为宽度,以f_i/Δ_i为高作长方形

频率直方图是总体分布密度f(x)的近似

§5.2 样本的数字特征

若总体X的一个样本为$X_1, X_2, \cdots, X_n$,

一. 样本均值

样本和: $\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n,$

样本均值: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

二. 样本方差

样本离差平方和 $$SS=\sum_{i=1}^{n}(X_i-\overline{X})^2$$

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

样本标准差 $$S=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

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

修正样本标准差 $$S^{*}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

例1. 证明: $S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2$

例2. 设总体X满足EX=μ, DX=σ².

证明: (1) $E\overline{X} = \mu,\ D\overline{X} = \frac{\sigma^2}{n}$

(2) $ES^2 = \frac{n-1}{n}\sigma^2,\ ES^{*2} = \sigma^2$

三. 样本矩

样本k阶原点矩：$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,k}$

样本k阶中心矩：$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k$

四. 其它样本数字特征

样本变异系数

$$CV = \frac{S^*}{\overline{X}} \times 100$$

众数(mode)的观测值为样本观测值中重复出现的频数最大的观测值（或组中值）；

极差(range)的观测值=最大观测值与最小观测值之差；

p分位数(0<p<1)的观测值Q为样本观测值中的某一个观测值(或组中值)，不大于Q的观测值的频率不小于p;

中位数(median)的观测值为0.5分位数的观测值，或样本观测值按大小排序后位于中间的一个观测值或两个观测值的算术平均值。

例: 设样本观测值为

1,2,2,3,3,3,4,5,6,7,8,

试计算它的数字特征:

(1)样本总和; (2)样本均值;

(3)样本标准差; (4)修正样本标准差;

(5)样本变异系数; (6)极差; (7)众数;

(8)中位数; (9)0.75分位数.

作业：P133，1　　P139，2，3

§5.3 χ^2分布、t分布及F分布

1. χ^2分布

定理　设随机变量$X_1, X_2, \cdots, X_n$相互独立且都服从$N(0,1)$,则随机变量$Z = \sum_{i-1}^{n} X_i^2$服从自由度为n的χ^2分布，记作$Z \sim \chi^2(n)$,它的分布密度为

$$p(z)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\dfrac{n}{2})}z^{\frac{n}{2}-1}e^{-\frac{z}{2}}, & z>0;\\ 0, & \text{其它.}\end{cases}$$

$$\text{其中}\Gamma(x)=\int_0^{+\infty}u^{x-1}e^{-u}du,$$

称为*Gamma*函数, 且$\Gamma(\dfrac{1}{2})=\sqrt{\pi},\Gamma(1)=1.$

$$\Gamma(n+1)=n\Gamma(n)$$

若$Y\sim\chi^2(n)$与$Z\sim\chi^2(m)$相互独立,则

$$Y+Z\sim\chi^2(n+m).$$

2. t分布

定理 设随机变量X与Y相互独立，且$X \sim N(0,1)$，$Y \sim \chi^2(n)$，则$Z = X \Big/ \sqrt{\frac{Y}{n}}$的分布称为自由度等于$n$的$t$分布，记作$Z \sim t(n)$，分布密度为

$$p(z) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{z^2}{n})^{-\frac{n+1}{2}}.$$

t分布的分布密度是偶函数，可以证明，当自由度n无限增大时，t分布将趋近于标准正态分布N(0,1).事实上，当n>30时，它们的分布曲线就差不多是相同的了。这时，t分布的分布函数值可查N(0,1)的分布函数值表得到。

3. F分布

定理 设随机变量X与Y相互独立，且$X \sim \chi^2(n)$，$Y \sim \chi^2(m)$，则$Z = \dfrac{X/n}{Y/m}$的分布称为第一自由度等于n，第二自由度等于m的F分布，

记作$Z \sim F(n,m)$,分布密度为

$$p(z)=\begin{cases}\dfrac{n^{\frac{n}{2}}m^{\frac{m}{2}}\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\cdot\dfrac{z^{\frac{n}{2}-1}}{(m+nz)^{\frac{n+m}{2}}}, & z>0;\\ 0, & \text{其它.}\end{cases}$$

F分布的分布密度与自由度的次序有关，

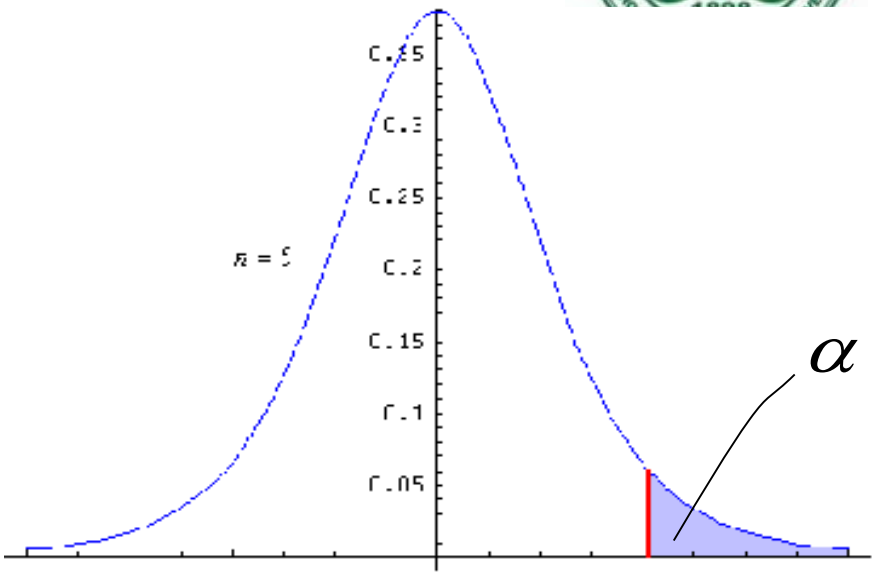
当$Z \sim F(n,m)$时，$\dfrac{1}{Z} \sim F(m,n)$.

4. t分布与F分布的关系

若$X \sim t(n)$，则$Y = X^2 \sim F(1,n)$.

4.常用分布的分位数

若随机变量X的密度为$p(x)$,$P\{X>x_\alpha\}=\alpha$,则称x_α为该分布的**上α分位数**.

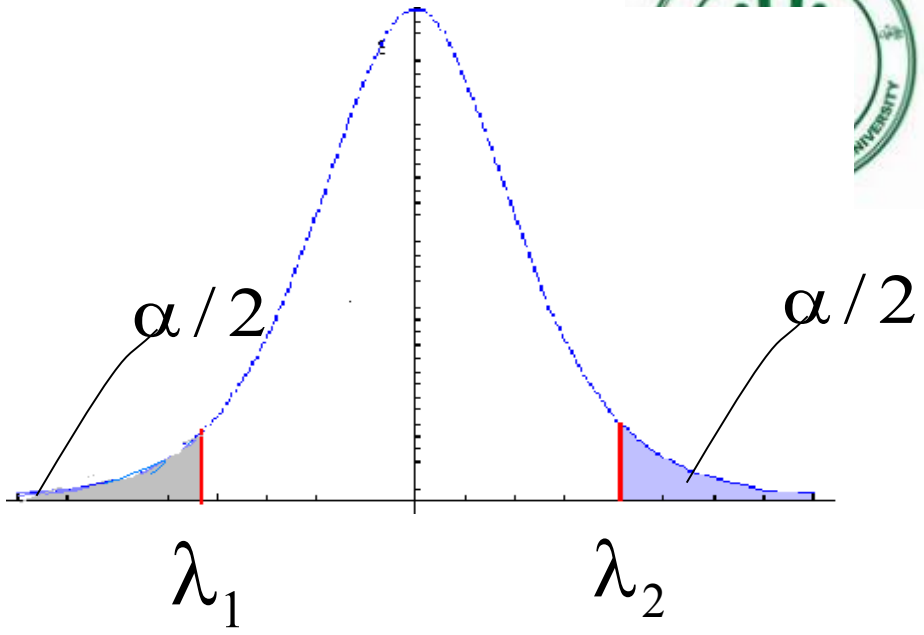


若随机变量X的密度为$p(x)$,$P\{X\le x_\alpha\}=\alpha$,则称$x_\alpha$为该分布的**下α分位数**.

注: $x_\alpha(\text{上}) = x_{1-\alpha}(\text{下})$

若随机变量X的密度为$p(x)$,$P\{X\le\lambda_1\}=\alpha/2$,
$P\{X\le\lambda_2\}=1-\alpha/2$则称

λ_1、λ_2为该分布的**双侧α分位数**.



注：三种分位数之间关系

$$1.\ x_\alpha(\text{上}) = x_{1-\alpha}(\text{下})$$

$$2.\ \lambda_1 = x_{\alpha/2}(\text{下})$$

$$\lambda_2 = x_{1-\alpha/2}(\text{下})$$

即三种分位数都能用下侧分位数表示

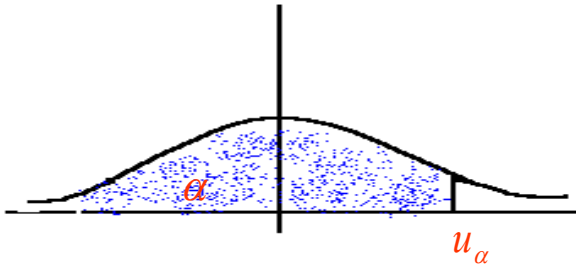
(2) 标准正态分布的α分位数

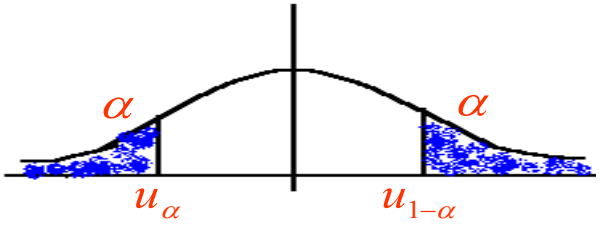
标准正态分布的α分位数记作u_α，

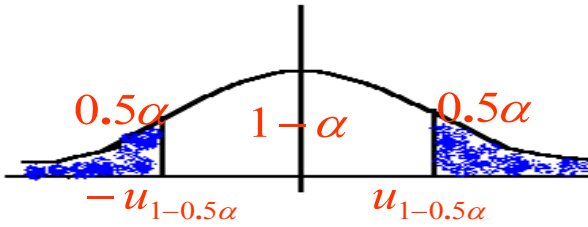
上侧α分位数$\lambda = u_{1-\alpha}$. 且由对称性有：$u_{1-\alpha} = -u_\alpha$

双侧α分位数$\lambda_1 = u_{0.5\alpha} = -u_{1-0.5\alpha}$，

双侧α分位数$\lambda_2 = u_{1-0.5\alpha}$.







当$X \sim N(0,1)$时，

$0 < \alpha < 0.5$时,先查出$u_{1-\alpha}$，然后得到$u_{\alpha} = -u_{1-\alpha}$.

例如 $u_{0.10} = -u_{0.90} \approx -1.28,$

$u_{0.05} = -u_{0.95} \approx -1.65,$

常用的上侧α分位数有

$$\alpha = 0.10,\quad u_{0.90} = 1.28;$$

$$\alpha = 0.05,\quad u_{0.95} = 1.65;$$

$$\alpha = 0.01,\quad u_{0.99} = 2.33;$$

$$\alpha = 0.025,\quad u_{0.975} = 1.96;$$

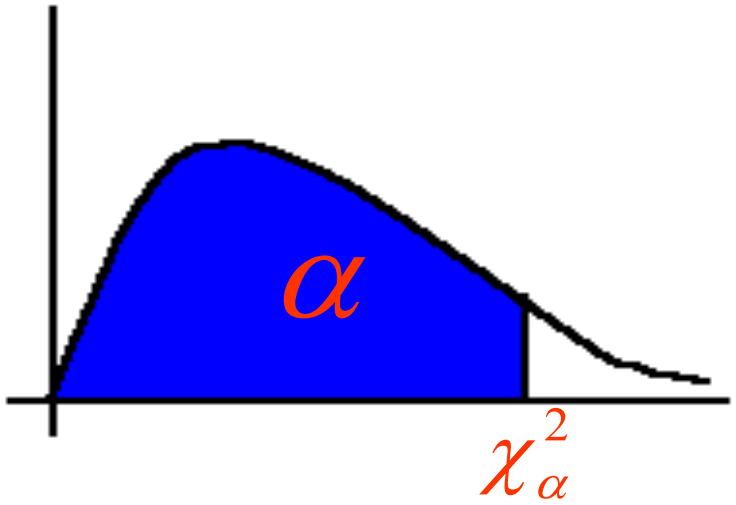
$$\alpha = 0.005,\quad u_{0.995} = 2.58.$$

(3) χ^2分布的α分位数

χ^2分布的α分位数记作$\chi^2_\alpha(\mathrm{n})$.如图，$\chi^2_\alpha(n)>0$,

当$X\sim\chi^2(n)$时，

$$P\{X<\chi^2_\alpha(\mathrm{n})\}=\alpha.$$



χ^2分布的分位数表

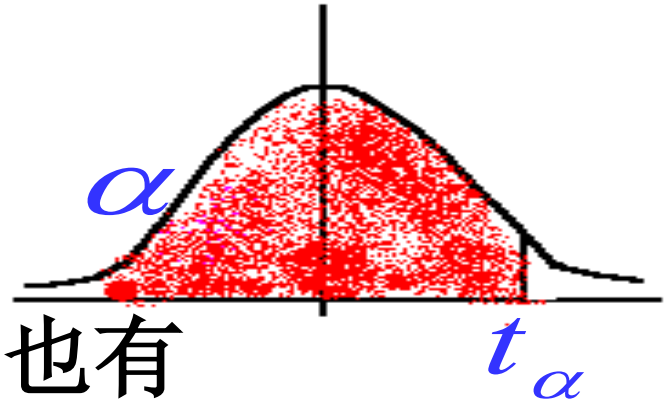
中查出$\chi^2_\alpha(\mathrm{n})$.

(4) t分布的α分位数

t分布的α分位数记作$t_\alpha(n)$.如图，

当$X \sim t(n)$时，

$$P\{X < t_\alpha(n)\} = \alpha.$$



且与标准正态分布相类似，也有

$$t_\alpha(n) = -t_{1-\alpha}(n),$$

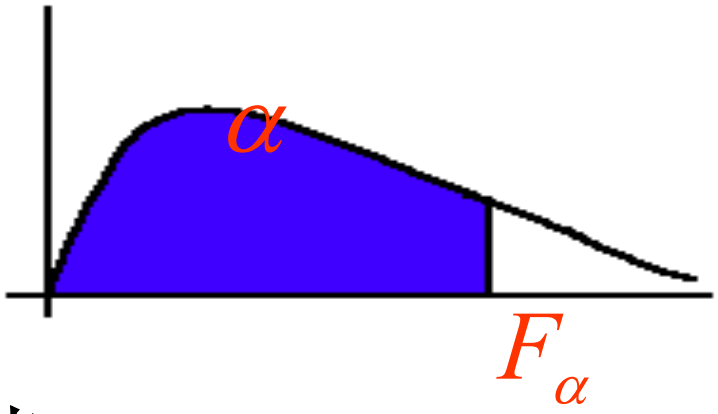
$$t_{0.95}(4) = 2.132, t_{0.005}(4) = -t_{0.995}(4) = -4.604.$$

(5) F分布的α分位数

F分布的α分位数记作$F_\alpha(n,m)$.如图,

当$X \sim F(n,m)$时,

$$P\{X < F_\alpha(n,m)\} = \alpha.$$



给出α和自由度n,m,可查表查出$F_\alpha(n,m)$.

当α较小时，表中查不出$F_\alpha(n,m)$,可先查$F_{1-\alpha}(m,n)$,

再求$F_\alpha(n,m)$.

注:颠倒自由度,查表取倒数.

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

例1. 已知随机变量$X\sim\chi^2(n)$,

(1) 求$\chi^2_{0.05}(9)$，$\chi^2_{0.975}(10)$;

(2) 当$n=10$, $\alpha=0.1$时,求c_1和c_2分别使

$P\{X>c_1\}=\alpha$, $P\{X>c_2\}=\alpha/2$成立.

例2. 已知随机变量$T\sim t(n)$,

(1) 求$t_{0.025}(8)$, $t_{0.99}(12)$;

(2) 当$n=10$, $\alpha=0.05$时,求c_1和c_2分别使

$P\{T>c_1\}=\alpha/2$, $P\{T>c_2\}=1-\alpha$成立.

例3. 已知随机变量$F\sim F(n_1,n_2)$,

(1) 求$F_{0.01}(10,12)$, $F_{0.99}(10,12)$;

(2) 当$\alpha=0.05$, $n_1=n_2+2=10$时,求c_1和c_2分别使

$P\{F>c_1\}=\alpha/2$, $P\{F>c_2\}=1-\alpha$成立.

查分位数的一般步骤：

1.将分位数用下侧分位数表示出来；

2.查表若表中没有（t分布、F分布），利用性质转换后，再查表。

作业：P145，2，7

§5.4 常用的统计量及其分布

1.统计量的定义

定义 设$X_1, X_2, \cdots, X_n$为总体X的样本，$g(X_1, X_2, \cdots, X_n)$为样本的一个函数，如果g中不包含任何未知的参数，则称g为一个统计量。如果样本的观测值为$x_1, x_2, \cdots, x_n$，则称$g(x_1, x_2, \cdots, x_n)$为统计量$g(X_1, X_2, \cdots, X_n)$的一个观测值。

注意： (1)统计量也是随机变量，也有分布；

(2)已知总体X的样本观测值，即可算出统计量的观测值。

例 设(X_1, X_2, X_3)是来自总体X的样本，且$X \sim N(\mu, \sigma^2)$,其中σ是未知参数，则

$X_1 + X_2 + X_3$, $\frac{1}{3}(X_1 + X_2 + X_3) - \mu$ 均为统计量，

而 $\frac{1}{3}(X_1 - \mu) + \frac{X_2 X_3}{\sigma}$ 不是统计量。

统计中常用分布的生成原理(典型模式)

定理1

(1) 独立的正态变量的线性函数仍为正态变量;

(2) 独立的标准正态变量的平方和 $\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$

(3) 设U, V独立且U~N(0, 1), V~χ^2 (n), 则

$$T = \frac{U}{\sqrt{V/n}} \sim t(n)$$

(4) 设U, V独立且U~$\chi^2(m)$, V~$\chi^2(n)$, 则

$$F = \frac{U/m}{V/n} \sim F(m,n)$$

2.一个正态总体的常用统计量及其分布

定理2. 设总体$X\sim N(\mu,\sigma^2)$, $X_1, X_2, \cdots, X_n$是抽自总体X的简单随机样本,则

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

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

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

(4) $\overline{X}$与S^{*2} 相互独立

(5) $\frac{\overline{X}-\mu}{s^*/\sqrt{n}}(\frac{\overline{X}-\mu}{s/\sqrt{n-1}}) \sim t(n-1)$

例1. 设总体X~N(1,4),样本容量为16,求

(1) $P\{0<X<2\}$; (2) $P\{0<\overline{X}<2\}$

解 $\because X\sim N(1,4)$,样本容量为16,

$$\frac{X-1}{2}\sim N(0,1),$$

$$\overline{X}\sim N(1,\frac{1}{4}),\qquad \frac{\overline{X}-1}{0.5}\sim N(0,1),$$

$$\therefore (1)\ \mathrm{P}\{0 < \mathrm{X} < 2\} = \Phi\left(\frac{2-1}{2}\right) - \Phi\left(\frac{0-1}{2}\right)$$

$$= \Phi(0.5) - \Phi(-0.5) = 0.383,$$

$$(2)\ \mathrm{P}\{0 < \overline{\mathrm{X}} < 2\} = \Phi\left(\frac{2-1}{0.5}\right) - \Phi\left(\frac{0-1}{0.5}\right)$$

$$= \Phi(2) - \Phi(-2) = 0.9546.$$

例1. 设总体X~N(1,4),样本容量为16,求

(3) $P\{2 < \frac{1}{16}\sum_{i=1}^{16}(X_i - 1)^2 < 8\}.$

(4) $P\{2 < \frac{1}{16}\sum_{i=1}^{16}(X_i - \bar{X})^2 < 8\}.$

3.两个正态总体的抽样分布

定理3.设总体$X\sim N(\mu_1,\sigma_1^2)$,$Y\sim N(\mu_2,\sigma_2^2)$,且X和Y相互独立.$X_1,X_2,\ldots,Xn_1$是来自总体X的样本,$Y_1,Y_2,\ldots,Yn_2$是来自总体Y的样本.则

$$(1)\ \overline{X}-\overline{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$$

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

$$(3)\ \text{当}\sigma_1^2=\sigma_2^2=\sigma^2\text{时},\ T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

$$S_W^2=\frac{(n_1-1)S_1^{*2}+(n_2-1)S_2^{*2}}{n_1+n_2-2}$$

例 当总体X~N(20,3)时，分别从X中取出容量为10及15的两个独立样本，若它们的均值为$\overline{X}_1$与$\overline{X}_2$，试求$P\{|\overline{X}_1-\overline{X}_2|>3\}$.

解 $\because X\sim N(20,3),\ \overline{X}_1\sim N(20,0.3),$

$\overline{X}_2\sim N(20,0.2),\ \overline{X}_1-\overline{X}_2\sim N(0,0.5),$

$$\frac{\overline{X}_1-\overline{X}_2}{\sqrt{0.5}}\sim N(0,1),$$

$$\therefore P\{|\bar{X}_1 - \bar{X}_2| > 0.3\}$$

$$= P\{\bar{X}_1 - \bar{X}_2 > 0.3\} + P\{\bar{X}_1 - \bar{X}_2 < -0.3\}$$

$$= P\{\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{0.5}} > \frac{0.3}{\sqrt{0.5}}\} + P\{\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{0.5}} < -\frac{0.3}{\sqrt{0.5}}\}$$

$$= 1 - \Phi(0.42) + \Phi(-0.42) = 0.6744.$$

4.非正态总体的样本均值分布

(1)当总体X的分布具有可加性时，将样本均值$\overline{X}$看成是样本总和$\sum_{i=1}^{n} X_i$的函数，由$\sum_{i=1}^{n} X_i$的的分布导出$\overline{X}$的分布。

(2) 由独立同分布的中心极限定理,当n充分大时,非正态总体的样本均值

$$\overline{X} \overset{\text{近似}}{\sim} N(\mu, \frac{\sigma^2}{n})$$

二项分布总体B(n,p)

B(n.p) 可以看成是n个相互独立且服从B(1,p)的随机变量之和。

因此，当X~B(1,p),容量为n时，

样本总和 $\sum_{i=1}^{n} X_i \sim B(n,p)$

$\overline{X}$的分布律为

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

$$=C_n^k p^k (1-p)^{n-k} \quad (k=0,1,\cdots,n).$$

结论7 当总体$X \sim B(1,p), X_1, X_2, \cdots, X_n$是$X$的一个样本，$p$和$1-p$都不是太小，$n \to \infty$时，

$$\bar{X} \sim N\left(p, \frac{p(1-p)}{n}\right),$$

$$\frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1),$$

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \sim N(0,1).$$

结论8

总体$X \sim B(1, p_1), X_1, X_2, \cdots, X_n$是$X$的一个样本，

总体$Y \sim B(1, p_2), Y_1, Y_2, \cdots, Y_m$是$Y$的一个样本，

两样本相互独立，概率都不太小且$n, m \to \infty$时

$$\bar{X} \sim N(p_1, \frac{p_1(1-p_1)}{n}), \qquad \bar{Y} \sim N(p_2, \frac{p_2(1-p_2)}{m}),$$

$$\bar{X} - \bar{Y} \sim N(p_1 - p_2, \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}),$$

$$\frac{(\bar{X} - \bar{Y}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \sim N(0,1).$$

例 将一枚均匀的硬币上抛120次，试求正面向上的频率在0.4~0.6之间的概率。

解 设总体$X \sim B(1, p_1), X_1, X_2, \cdots, X_n$是$X$的一个样本，$X_i (i = 1,2,\cdots,n)$为1时表示正面向上，为0时表示正面不向上，则

$$\overline{X} \sim N(p, \frac{p(1-p)}{n}),$$

$$n = 120, p = 0.5, \sqrt{\frac{p(1-p)}{n}} = 0.0456,$$

{频率在0.4~0.6之间}$= \{0.4 \le \overline{X} \le 0.6\}$

将0.4校正为$\dfrac{0.4\times120-0.5}{120}=0.396,$

将0.6校正为$\dfrac{0.6\times120+0.5}{120}=0.604,$

P{频率在0.4~0.6之间}$= P\{0.4 \le \overline{X} \le 0.6\}$

$\approx P\{0.396 \le \overline{X} \le 0.604\}$

$\approx P\{\dfrac{0.396-0.5}{0.0456} \le \dfrac{\overline{X}-0.5}{0.0456} \le \dfrac{0.604-0.5}{0.0456}\}$

$=\Phi(2.28)-\Phi(-2.28)=0.9774.$

5.顺序统计量及其分布

设$X_1, X_2, \cdots, X_n$为总体X的一个样本，$x_1, x_2, \cdots, x_n$为样本的观测值，由小到大排序为$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$以后，如果$X_{(i)}$总是以$x_{(i)}$为它的观测值，则称$X_{(i)}$为$X$的第$i(=1,2,\cdots,n)$个顺序统计量。

对容量为n的样本，可得n个顺序统计量$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$，

称$X_{(1)}=\min\limits_{1\le i\le n}X_i$为最小顺序统计量;

称$X_{(n)}=\max\limits_{1\le i\le n}X_i$为最大顺序统计量。

总体X的分布函数为$F(x)$.

$X_{(1)}$的分布函数：

$$F^*_{\min}(x)=1-[1-F(x)]^n,$$

$X_{(n)}$的分布函数：

$$F^*_{\max}(x)=[F(x)]^n.$$

例　设总体$X \sim N(12,4)$,一个样本为$X_1, X_2, \cdots, X_5$，试求：

(1)样本的极小值小于10的概率；

(2)样本的极大值大于15的概率。

解　因为$X_1, X_2, \cdots, X_5$相互独立且都服从$N(12,4)$, 所以

$$P\{X_{(1)} < 10\} = 1 - [1 - F(10)]^5$$

$$= 1 - [1 - \Phi(\frac{10-12}{2})]^5 = 0.5785;$$

$$P\{X_{(5)} > 15\} = 1 - P\{X_{(5)} \le 15\}$$

$$= 1 - [F(15)]^5 = 1 - [\Phi(\frac{15-12}{2})]^5$$

$$= 1 - [\Phi(1.5)]^5 = 0.2923.$$

例　设总体$X \sim U(0,1)$,它的一个样本为$X_1, X_2, \cdots, X_5$，试求：

(1)样本的极小值大于0.5的概率；

(2)样本的极大值大于0.5的概率。

解　X的分布密度

$$p(x)=\begin{cases}1, & 0<x<1;\\ 0, & \text{其它.}\end{cases}$$

分布函数

$$F(x)=\begin{cases}0, & x\le 0;\\ x, & 0<x\le 1;\\ 1, & 1<x.\end{cases}$$

$$P\{X_{(1)} > 0.5\} = 1 - P\{X_{(1)} \le 0.5\}$$

$$= 1 - \{1 - [1 - F(0.5)]^5\}$$

$$= (1 - 0.5)^5 = (0.5)^5;$$

$$P\{X_{(5)} > 0.5\} = 1 - P\{X_{(5)} \le 0.5\}$$

$$= 1 - [F(0.5)]^5$$

$$= 1 - (0.5)^5.$$

第五章 小结

一、基本概念

1.总体、个体、样本、样本容量

2.简单随机样本:独立、同分布(来自总体,代表总体)

3.统计量

4.样本的数字特征

样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

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

样本k阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

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

二、三种常用抽样分布及其分位数

1.三种分布的定义与性质

2.会查分位数

三、抽样分布

定理1. (生成原理)

(1)独立的正态变量的线性函数仍为正态变量;

(2)独立的标准正态变量的平方和 $\sim\chi^2(n)$

(3)设U,V独立且$U\sim N(0,1)$,$V\sim\chi^2(n)$,

$\sim t(n)$

(4)设U,V独立且$U\sim\chi^2(n_1)$,$V\sim\chi^2(n_2)$,

$\sim F(n_1,n_2)$

定理2. (一个正态总体的抽样分布)

设总体$X\sim N(\mu,\sigma^2)$, $X_1, X_2, \cdots, X_n$是抽自总体X的简单随机样本,则

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

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

(3) $\frac{(n-1)s^{*2}}{\sigma^2}(=\frac{ns^2}{\sigma^2}) \sim \chi^2(n-1)$

(4) $\overline{X}$与S^{*2}相互独立

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

定理3.(两个正态总体的抽样分布)

设总体$X\sim N(\mu_1,\sigma_1^2)$,$Y\sim N(\mu_2,\sigma_2^2)$,且X和Y相互独立.$X_1,X_2,\ldots,Xn_1$是来自总体X的样本,$Y_1,Y_2,\ldots,Yn_2$是来自总体Y的样本.则

$$(1)\ \overline{X}-\overline{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})$$

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

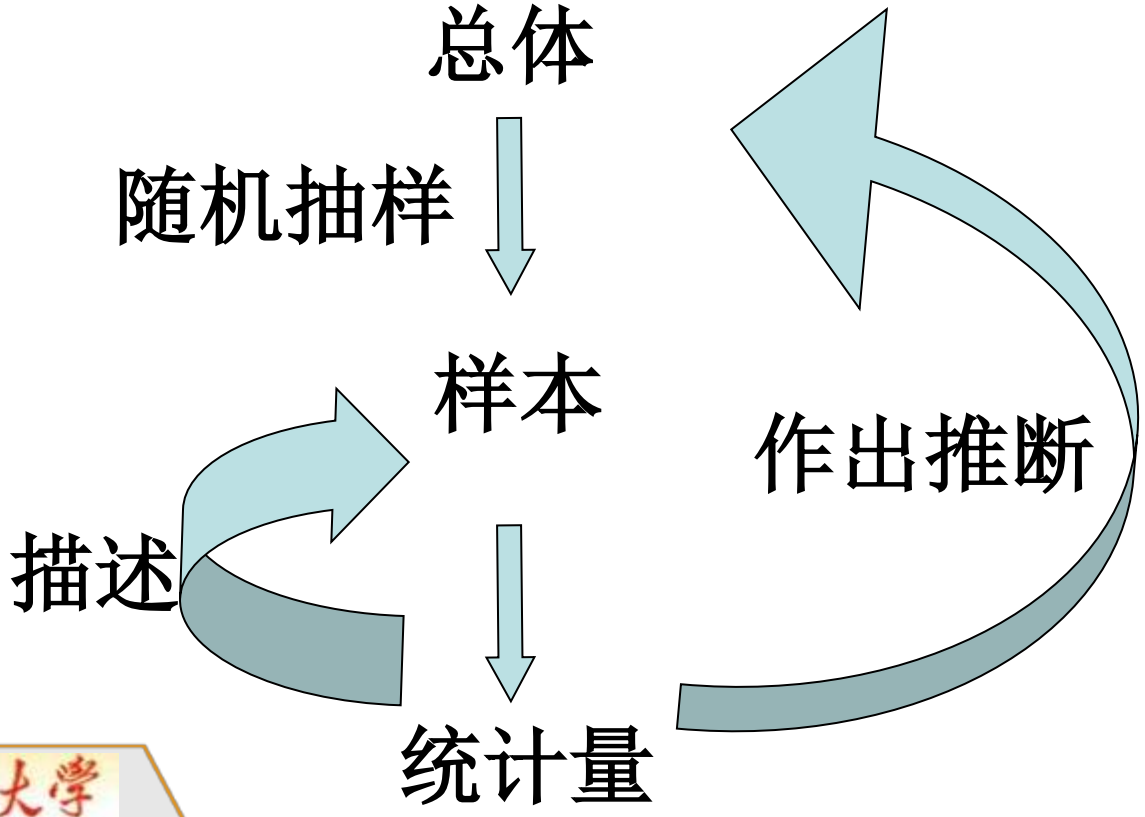
$$(3)\ \text{当}\sigma_1^2=\sigma_2^2=\sigma^2\text{时},\ T=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

$$S_W^2=\frac{(n_1-1)S_1^{*2}+(n_2-1)S_2^{*2}}{n_1+n_2-2}$$

作业：P155，1，8

第六章
总体分布中未知参数的估计

上一章，我们介绍了总体、样本、统计量和抽样分布的概念，介绍了统计中常用的**三大分布**，给出了几个重要的**抽样分布定理**．它们是进一步学习统计推断的**基础**.



假如我们要估计某班男生的平均身高.

（假定身高服从正态分布 $N(\mu, 0.1^2)$ ）

从该总体选取容量为5的样本, 样本值为

1.65　1.67　1.68　1.78　1.69

若估计μ为1.69，这是点估计.

若估计μ在区间[1.67, 1.78]内,这是区间估计.

§6.1 参数的点估计 (point estimate)

一. 点估计方法

设总体X的分布函数$F(x, \theta)$形式已知, θ为未知参数, $x_1, x_2, \ldots, x_n$是样本$X_1, X_2, \ldots, X_n$的样本观测值. 通过统计量$\hat{\theta}(X_1, X_2, \ldots, X_n)$来估计$\theta$, 称为$\theta$的估计量, $\hat{\theta}(x_1, x_2, \ldots, x_n)$为$\theta$的估计值. 估计量和估计值统称为点估计, 记作$\hat{\theta}$.

点估计：设总体的分布类型已知，但有未知参数，构造一个适当的统计量：

$\hat{\theta} = \hat{\theta}(X_1, \cdots X_n)$ 称为参数的估计量.

把样本值代入 $\hat{\theta} = \hat{\theta}(X_1, \cdots X_n)$ 得到的一个值：

$\hat{\theta} = \hat{\theta}(x_1, \cdots x_n)$ 称为参数 θ 的估计值.

注：由于 $\hat{\theta}(x_1, \cdots, x_n)$ 是实数域上的一个点，现用它来估计 θ，故称这种估计为点估计。

引例 已知某地区新生婴儿的体重$X\sim N(\mu,\sigma^2)$,

μ,σ^2未知,

随机抽查100个婴儿

得100个体重数据

10, 7, 6, 6.5, 5, 5.2, …

据此, 我们应如何估计 μ和 σ 呢?

点估计的常用方法

1. 矩估计法

2. 最大似然法

(一) 矩估计法

基于一种简单的“替换”思想，是英国统计学家 K. 皮尔逊 于1894年提出的 .

基本思想：用样本矩估计总体矩 .

理论依据： 大数定律

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} a_k = E(X^k)$$

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \xrightarrow{P} b_k = E\{[X - E(X)]^k\}$$

特别地，当总体的数学期望与方差存在时，

总体E(X)矩估计量 = 样本的均值

总体方差D(X)的矩估计量 = 样本的方差

用样本原点(中心)矩及其函数去估计总体相应的原点(中心)矩及其函数的方法称为矩估计法.

矩估计法的步骤:

设总体分布F中含有m个未知参数$\theta_1, \theta_2,\ldots,\theta_m$.

(1) 求出总体k阶原点矩$a_k=EX^k$及对应的样本k阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

(2) 令$a_k=A_k$, $k=1,2,\ldots,m$

(3) 解方程组, 其解即为$\theta_1,\theta_2,\ldots,\theta_m$的矩估计量.

例 设从某灯泡厂某天生产的灯泡中随机抽取10只灯泡，测得其寿命为(单位:小时)

1050, 1100, 1080, 1120, 1200

1250, 1040, 1130, 1300, 1200

试用矩法估计该天生产的灯泡的平均寿命及寿命分布的方差.

解 $$\hat{E(X)}_{矩} = \bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 1147(h)$$

$$\hat{D(X)}_{矩} = \frac{1}{10}\sum_{i=1}^{10} x_i^2 - \bar{x}^2 = 6821(h^2).$$

重要结论 设$X_1, X_2, \cdots, X_n$是总体X的一个样本，

(1)若$X \sim B(1, p)$, p未知。

$$\because E(X) = p, p = E(X), \therefore \hat{p} = \overline{X};$$

(2)若$X \sim P(\lambda)$, λ未知。

$$\because E(X) = \lambda, \lambda = E(X), \therefore \hat{\lambda} = \overline{X};$$

(3)若$X \sim N(\mu,\sigma^2)$, μ,σ^2未知。

$$\because E(X)=\mu, D(X)=E(X^2)-(EX)^2=\sigma^2,$$

$$\therefore \hat{\mu}=\overline{X},$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-(\frac{1}{n}\sum_{i=1}^{n}X_i)^2$$

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

例　设总体X服从参数为λ的指数分布

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

$\lambda>0$,样本为$X_1, X_2, \cdots, X_n$，求λ的矩估计量.

解　$\because E(X)=\dfrac{1}{\lambda}$,　令$\dfrac{1}{\lambda}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$,

$$\Rightarrow \hat{\lambda}=\frac{n}{\sum_{i=1}^{n}X_i}=\frac{1}{\overline{X}}.$$

例 设总体 X 在 $[a,b]$ 上服从均匀分布，其中 a，b 未知，$(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的样本，求 a, b 的估计量.

解 $E(X)=\dfrac{a+b}{2}$，$DX=\dfrac{(b-a)^2}{12}$

$$\therefore \frac{\hat{a}+\hat{b}}{2}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X},$$

$$\frac{(\hat{a}-\hat{b})^2}{12}=\frac{1}{n}\sum_{i=1}^{n}X_i^{\,2}-\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2=S^2,$$

解出 $\hat{b}=2\overline{X}-\hat{a};\quad \left(\overline{X}-\hat{a}\right)^2=3S^2;$

$$\hat{a}=\overline{X}-\sqrt{3}S;\quad \hat{b}=\overline{X}+\sqrt{3}S$$

矩估计法

基本思想：用样本矩估计总体矩

最大似然估计法

基本思想：

极大似然法的基本思想

引例：X~B(4,p)，p未知，现抽取容量为3的样本，若其观测值为1，2，1，那么对p的取值可以说什么呢？

➢最大似然法---由费歇尔引进的

Maximum Likelihood Estimation

基本思想: 选择参数使样本$X_1,X_2,\ldots,X_n$取值$x_1,x_2,\ldots,x_n$的概率(密度)最大.

设$X_1, X_2, \cdots, X_n$是总体X的一个样本，

$x_1, x_2, \cdots, x_n$为样本的观测值，则当总体

X的分布律$P\{X = x\} = p(x)$或分布密度

$p(x)$中含有未知参数θ时，

记$p(x) = p(x;\theta)$,

称$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta)$为似然函数，而称使

$L(\theta)$取极大的估计值$\hat{\theta}_L$为θ的极大似然

估计值。

最大似然估计法的步骤

(1) 写出似然函数

重要

离散型：$L(\theta)=\prod_{i=1}^{n} P\{X=x_i;\theta\}$

连续型：$L(\theta)=\prod_{i=1}^{n} f(x_i;\theta)$

(2) 取对数$\ln L(\theta)$

(3) 求$\dfrac{d[\ln L(\theta)]}{d\theta}$

(4) 解出驻点. 即为θ的极大似然估计量.

注：2.若求导方法行不通，这时可用直接求L(θ)的最大值的方法，即

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

例 设 $X \sim B(1, p)$, $X_1, X_2, \cdots, X_n$是来自X的一个样本,求 p的极大似然估计量.

解 设$x_1, x_2, \cdots, x_n$为相应于样本$X_1, X_2, \cdots, X_n$的一个样本值,

X的分布律为 $P\{X = x\} = p^x (1-p)^{1-x}, x = 0,1$

似然函数 $$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i},$$

$$\ln L(p)=\left(\sum_{i=1}^{n}x_i\right)\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p),$$

$$\text{令}\quad \frac{d}{dp}\ln L(p)=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{n-\sum_{i=1}^{n}x_i}{1-p}=0,$$

解得 p 的极大似然估计值 $\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}.$

p 的极大似然估计量为 $\hat{p}=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}.$

这一估计量与矩估计量是相同的.

例8 设 X服从参数为$\lambda\ (\lambda>0)$的泊松分布，$X_1, X_2, \cdots, X_n$是来自X的一个样本，求λ的极大似然估计量.

解 因为X的分布律为

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

所以 λ 的似然函数为

$$L(\lambda)=\prod_{i=1}^{n}\left(\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}\right)=e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}(x_i!)},$$

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\ln\lambda - \sum_{i=1}^{n}(\ln x_i!)$$

令 $$\frac{d}{d\lambda}\ln L(\lambda) = -n + \frac{\sum_{i=1}^{n}(x_i)}{\lambda} = 0,$$

解得λ的极大似然估计值

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

λ的极大似然估计量为

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

这一估计量与矩估计量是相同的.

例 设总体 $X \sim N(\mu,\sigma^2)$, μ,σ^2为未知参数, $x_1,x_2,\cdots,x_n$是来自X的一个样本值, 求μ和σ^2的极大似然估计量.

解 X的概率密度为

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

X 的似然函数为

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

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

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

令
$$\begin{cases}\dfrac{\partial}{\partial\mu}\ln L(\mu,\sigma^2)=0\\[2ex]\dfrac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2)=0\end{cases}$$

$$\Rightarrow\begin{cases}\dfrac{1}{\sigma^2}\left[\sum_{i=1}^{n}x_i-n\mu\right]=0,\\[2ex]-\dfrac{n}{2\sigma^2}+\dfrac{1}{2(\sigma^2)^2}\left[\sum_{i=1}^{n}(x_i-\mu)^2\right]=0,\end{cases}$$

由 $\frac{1}{\sigma^2}\left[\sum_{i=1}^{n} x_i - n\mu\right] = 0$ 解得 $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$

由 $-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$ 解得

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

故μ和σ^2的极大似然估计量分别为

$$\hat{\mu} = \bar{X},\quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

它们与相应的矩估计量相同.

当μ已知时，$\hat{\sigma}_L^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$；

当μ未知时，$\hat{\sigma}_L^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = S^2$.

例： 设总体X的密度

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

其中θ未知，

$x_1, x_2, \ldots, x_n$为样本$X_1, X_2, \ldots, X_n$一组观测值，求θ的矩估计和最大似然估计.

4.评价估计量优劣的标准

(1)无偏性 若未知参数θ的估计量为$\hat{\theta}(X_1, X_2, \cdots, X_n)$,则当

$$E(\hat{\theta}) = \theta$$

称$\hat{\theta}$为θ的无偏估计量。

显然用未知参数θ的无偏估计量$\hat{\theta}$代替θ时所产生的误差的数学期望为零，即无系统误差(定义为:$E(\hat{\theta}) - \theta$)。

前面已证明:

若总体X数学期望E(X)存在，$E(\overline{X})=E(X)$.

若总体X方差D(X)存在，

$$E(S^2)=\frac{n-1}{n}D(X);$$

$$E(S^{*2})=D(X).$$

从而$\overline{X}, S^{*2}$是总体的期望$E(X)$和方差$D(X)$的无偏估计量,而S^2不是$D(X)$的无偏估计量.

(2)有效性

设$\hat{\theta}_1(X_1, X_2, \cdots, X_n)$和$\hat{\theta}_2(X_1, X_2, \cdots, X_n)$

都是θ的无偏估计量，如果

$$D(\hat{\theta}_1) < D(\hat{\theta}_2)$$

则称$\hat{\theta}_1$比$\hat{\theta}_2$有效。

其中: $D(\hat{\theta}_1) = E\left(\hat{\theta}_1 - \theta\right)^2; D(\hat{\theta}_2) = E\left(\hat{\theta}_2 - \theta\right)^2$

（3）. 相合性：随着样本容量的增大，估计量的值越来越接近被估计的参数

定义：设 $\hat{\theta}_n$ 是 θ 的估计量，

$$\hat{\theta}_n \xrightarrow{P} \theta \ (n \to \infty)$$

则称 $\hat{\theta}_n$ 为 θ 的相合估计量 .

注: (1) 无偏估计量不唯一, 不是每个参数都有无偏估计量.

(2) 无偏性与有效性是在n固定时讨论的, 而一致性是在n→∞时进行的.

例 测得自动车床加工的10个零件的尺寸与规定尺寸的偏差(微米)如下：

+2，+1，-2，+3，+2，+4，-2，+5，+3，+4.

求零件尺寸偏差总体的均值及方差的无偏估计值.

解 有

$$\hat{\mu} = \bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 2(\text{微米}),$$

$$\hat{\sigma}^2 = s^{*2} = \frac{1}{9}\sum_{i=1}^{10}(x_i - \bar{x})^2 = 5.78(\text{微米}^2).$$

重要结论

当$X_1, X_2, \cdots, X_n$是总体X的一个样本且$D(X) > 0$时，对任意非负常数$c_1, c_2, \cdots, c_n$，如果$\sum_{i=1}^{n} c_i = 1$,求证：$\sum_{i=1}^{n} c_i X_i$都是总体均值$E(X)$的无偏估计量，且$\overline{X}$在这些估计量中是方差最小的无偏估计量。

解 $\because E(X_i)=E(X)(i=1,2,\cdots,n)$ 且 $\sum_{i=1}^{n} c_i=1$,

$$E\left(\sum_{i=1}^{n} c_i X_i\right)=\sum_{i=1}^{n} c_i E(X_i)$$

$$=E(X)\sum_{i=1}^{n} c_i=E(X),$$

$\therefore \sum_{i=1}^{n} c_i X_i$ 都是总体均值 $E(X)$ 的无偏估计量.

又 $\because D(X_i)=D(X)(i=1,2,\cdots,n)$,

$$D(\sum_{i=1}^{n} c_i X_i) = \sum_{i=1}^{n} c_i^2 D(X_i) = D(X)\sum_{i=1}^{n} c_i^2,$$

记$f = \sum_{i=1}^{n} c_i^2$及$\sum_{i=1}^{n} c_i = 1$,得

$$f = c_1^2 + \cdots + c_{n-1}^2 + (1 - c_1 - c_2 - \cdots - c_{n-1})^2,$$

由$n-1$个方程

$$\frac{\partial f}{\partial c_i} = 2c_i - 2(1 - c_1 - c_2 - \cdots - c_{n-1}) = 0$$

解出 $c_1 = c_2 = \cdots = c_{n-1} = \dfrac{1}{n}$,

代入$\sum_{i=1}^{n} c_i = 1$中，得到$c_n = \dfrac{1}{n}$,

因此$\overline{X}$是方差最小的无偏估计量。

前面我们介绍了参数点估计，参数点估计是用一个确定的值去估计未知的参数. 但是, 点估计值仅仅是未知参数的一个近似值，它没有反映出这个近似值的误差范围，使用起来把握不大。

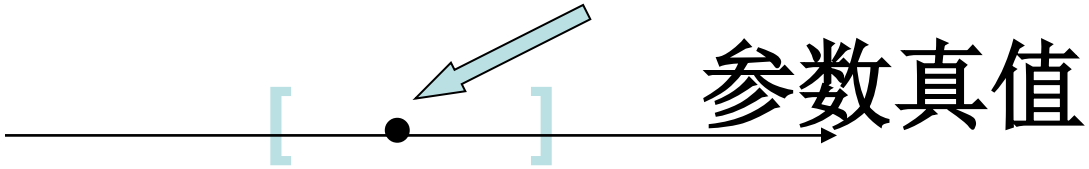
为了使估计的结论更可信，需要引入区间估计, 弥补了点估计的这个缺陷 .

§6.2 参数的区间估计
(interval estimation)

估计湖中鱼数，若根据一个实际样本，得到鱼数***N***的极大似然估计为**1000**条.

若能给出一个区间，在此区间内合理地相信 ***N*** 的真值位于其中. 这样对鱼数的估计就更有把握.

希望确定一个区间，同时给出一个可信程度，使其他人相信它包含参数真值.



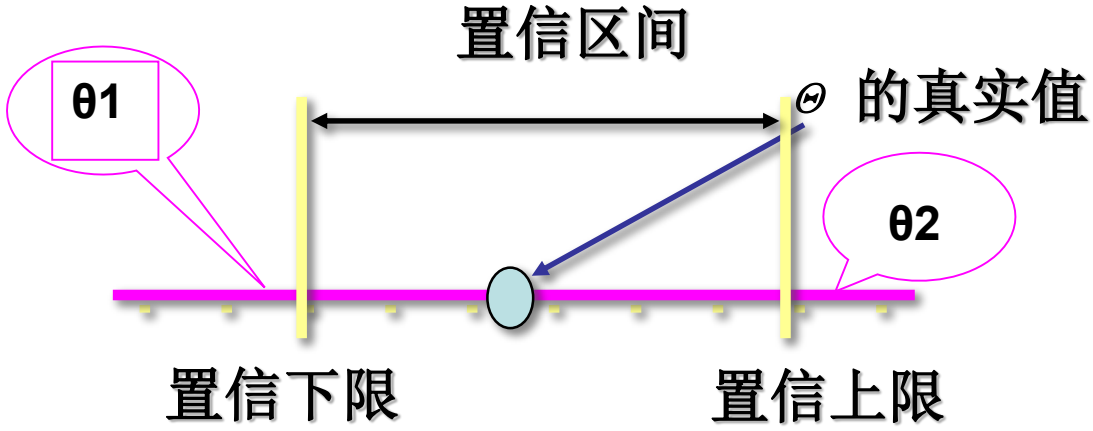
这里所说的“可信程度”是用概率来度量的，称为置信概率(置信度).

习惯上把置信水平记作$1-\alpha$，这里α是一个很小的正数.

定义：设总体X的分布中有未知参数θ，对于给定α（0<α<1），若有统计量$\theta_1=\theta_1(X_1,\cdots,X_n)$和$\theta_2=\theta_2(X_1,\ldots,X_n)$使得

$$P\{\theta_1<\theta<\theta_2\}\geq 1-\alpha$$

则称区间(θ_1,θ_2)为θ 的**置信水平**为1−α的**置信区间**



如果要找统计量$g(X_1, X_2, \cdots, X_n)$使 $P\{g < \theta\} \geq 1-\alpha$，则$(g, +\infty)$称为$\theta$的单侧$1-\alpha$置信区间，$g$为$\theta$的单侧$1-\alpha$置信下限；

如果要找统计量$h(X_1, X_2, \cdots, X_n)$使$P\{\theta < h\} \geq 1-\alpha$，则$(-\infty, h)$称为$\theta$的单侧$1-\alpha$置信区间，$h$为$\theta$的单侧$1-\alpha$置信上限。

$$P\{\theta_1 < \theta < \theta_2\} \geq 1-\alpha$$

含义: 随机区间包含参数θ的概率至少为1-α

$$1-\alpha = 0.95$$

$1-\alpha$表示区间估计的可靠性.

$1-\alpha$越大, (θ_1,θ_2)作为置信区间越可靠.

区间长度$\theta_2-\theta_1$代表区间估计的精度.

长度越小, 区间估计精度越高.

但一般情况下, 可靠性与精度不可兼得.

区间估计原则: 可靠性优于精度.

固定$1-\alpha$, 使$\theta_2-\theta_1$越小越好.

设$X_1, X_2, \ldots, X_n$是取自正态总体$N(\mu,\sigma^2)$的样本

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

$$\frac{\bar{X}-\mu}{S^*/\sqrt{n}}\sim t(n-1)$$

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

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

一个正态总体参数的区间估计 $N(\mu,\sigma^2)$

1. μ的区间估计

(1) σ^2已知 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

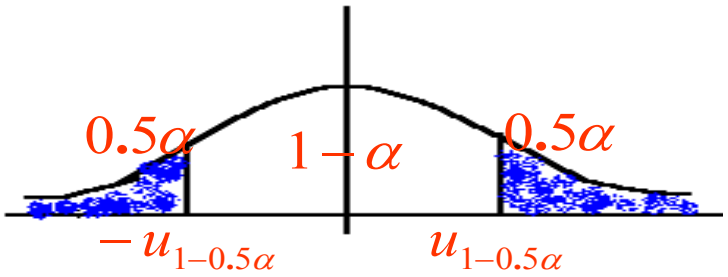
μ置信区间 $(\overline{X} \pm \dfrac{\sigma}{\sqrt{n}} u_{1-\alpha/2})$

注：单侧1-α置信下(上)限只需将双侧1- α置信下(上) 限中分位数中的0.5 α改为α即可。

$$\because \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad P\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| < u_{1-0.5\alpha}\right\} = 1-\alpha,$$

$$\text{即} P\left\{\bar{X} - \frac{\sigma}{\sqrt{n}} u_{1-0.5\alpha} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} u_{1-0.5\alpha}\right\} = 1-\alpha,$$

$$\Rightarrow \left(\bar{x} - u_{1-0.5\alpha}\frac{\sigma}{\sqrt{n}}, \bar{x} + u_{1-0.5\alpha}\frac{\sigma}{\sqrt{n}}\right).$$



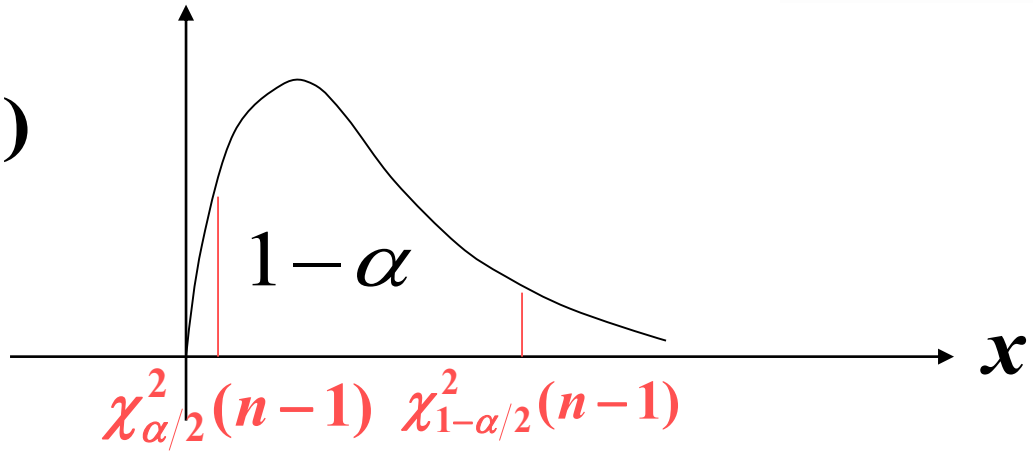
(2) σ^2未知

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

μ置信区间 $\left(\overline{X} \pm \frac{S^*}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right)$

2、σ²区间估计(μ未知)

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



σ²置信区间

$$\left(\frac{(n-1)S^{*2}}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)S^{*2}}{\chi^2_{\alpha/2}(n-1)} \right)$$

例 设总体$X \sim N(\mu,\sigma^2)$, X的一个样本的观测值为6.6; 4.6; 5.4; 5.8; 5.5.

(1)若$\sigma^2 = 0.5$， 试求μ的双侧0.95置信区间、单侧0.95置信下限和单侧0.95置信上限。

(2)若σ^2未知， 试求μ的双侧0.95置信区间、单侧0.95置信下限和单侧0.95置信上限。

(3)若μ未知，试求σ^2的双侧0.95置信区间、单侧0.95置信下限和单侧0.95置信上限。

解(1) 有$n=5, \bar{x}=5.58, \sigma^2=0.5$

置信概率$1-\alpha=0.95$, 则$\alpha=0.05$,

查表得$u_{1-0.5\alpha}=u_{0.975}=1.96, u_{1-\alpha}=u_{0.95}=1.65$,

$$u_{0.975}\frac{\sigma}{\sqrt{n}}=1.96\times\frac{\sqrt{0.5}}{\sqrt{5}}=0.6198,$$

由公式1, 得μ的双侧0.95置信区间$(4.96,6.20)$

同理可得单侧0.95置信下限为5.06,

单侧0.95置信上限为6.10。

解（2）有$n=5,\bar{x}=5.58,$

$$s^*=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}=0.722,$$

置信概率$1-\alpha=0.95,$ 则$\alpha=0.05,$

查t分布表得

$$t_{1-0.5\alpha}(4)=t_{0.975}(4)=2.776,\quad t_{1-\alpha}(4)=t_{0.95}(4)=2.132,$$

由公式2，得μ的双侧0.95置信区间$(4.68,6.48),$

同理可得单侧0.95置信下限为4.89,

单侧0.95置信上限为6.27。

解（3）有$n=5, \bar{x}=5.58, \sum_{i=1}^{5}(x_i-\bar{x})^2=2.088$

置信概率$1-\alpha=0.95$，则$\alpha=0.05$，

查χ^2分布表得

$$\chi^2_{1-0.5\alpha}(4)=\chi^2_{0.975}(4)=11.1,$$

$$\chi^2_{0.5\alpha}(4)=\chi^2_{0.025}(4)=0.48,$$

$$\chi^2_{1-\alpha}(4)=\chi^2_{0.95}(4)=9.49,$$

$$\chi^2_{\alpha}(4)=\chi^2_{0.05}(4)=0.71,$$

由公式3，得σ^2的双侧0.95置信区间

(0.19,4.35),

同理可得单侧0.95置信下限为0.22,

单侧0.95置信上限为2.94。

统计量 $\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n}+\dfrac{\sigma_2^2}{m}}} \sim N(0,1).$

当 $\sigma_1^2=\sigma_2^2=\sigma^2$，$S_W^2=\dfrac{SSX+SSY}{n+m-2}$ 时，统计量

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}} \sim t(n+m-2).$$

统计量 $\dfrac{\dfrac{SSX}{(n-1)\sigma_1^2}}{\dfrac{SSY}{(m-1)\sigma_2^2}}=\dfrac{S_X^{*2}/\sigma_1^2}{S_Y^{*2}/\sigma_2^2} \sim F(n-1,m-1).$

两个正态总体的均值差、方差比的置信区间

$$X \sim N(\mu_1, \sigma_1^2) \quad Y \sim N(\mu_2, \sigma_2^2)$$

$X_1, \ldots, Xn_1$来自X的样本, $Y_1, \ldots, Yn_2$来自Y的样本, 两样本独立

1、 $\mu_1 - \mu_2$ 的区间估计

（1） σ_1^2、 σ_2^2 已知

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

$\mu_1 - \mu_2$ 的置信区间

$$\left(\overline{X} - \overline{Y} \pm u_{1-\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

因为当σ_1^2和σ_2^2已知时，统计量

$$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}),$$

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\sim N(0,1).$$

对给定的置信概率 $1-\alpha$，有

$$P\left\{\frac{\left|(\bar{X}-\bar{Y})-(\mu_1-\mu_2)\right|}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}<u_{1-0.5\alpha}\right\}=1-\alpha.$$

解出$\mu_1-\mu_2$所满足的不等式即可。

(2) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知

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

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

$\mu_1 - \mu_2$ 的置信区间

$$\left(\overline{X} - \overline{Y} \pm t_{1-\alpha/2}(n_1 + n_2 - 2) S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

2、σ_1^2/σ_2^2 的区间估计（μ_1、μ_2 未知）

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

σ_1^2/σ_2^2 的置信区间

$$\left(\frac{S_1^{*2}/S_2^{*2}}{F_{1-\alpha/2}(n_1-1,n_2-1)},\frac{S_1^{*2}/S_2^{*2}}{F_{\alpha/2}(n_1-1,n_2-1)}\right)$$

例 设总体$X \sim N(\mu_1, \sigma_1^2)$,它的一个样本的观测值为2.10,2.35,2.39,2.41,2.44,2.56,总体$Y \sim N(\mu_2, \sigma_2^2)$,它的一个样本的观测值为2.03,2.28,2.58,2.71.

(1)若$\sigma_1^2 = 0.02$, $\sigma_2^2 = 0.09$,试求$\mu_1 - \mu_2$的双侧0.95置信区间、单侧0.95置信下限和单侧0.95置信上限。

解 $\because n=6, m=4, \sigma_1^2=0.02, \sigma_2^2=0.09,$

由样本观测值计算出 $\bar{x}=2.375, \bar{y}=2.4,$

$$\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}=\sqrt{\frac{0.02}{6}+\frac{0.09}{4}}=0.1607,$$

查正态分布表得$u_{0.975}=1.96, u_{0.95}=1.65$,

由公式4计算出$\mu_1-\mu_2$的双侧0.95置信区间为(–0.34,0.29),

单侧0.95置信下限为– 0.29, 单侧0.95置信上限为0.24。

(2)若$\sigma_1^2=\sigma_2^2=\sigma^2$,$\sigma^2$未知，求$\mu_1-\mu_2$的双侧0.95置信区间、单侧0.95置信下限和单侧0.95置信上限。

解 $\because \mathrm{n}=6,\mathrm{m}=4,\sigma_1^2=\sigma_2^2,$

由样本观测值计算出 $\bar{x}=2.375,\bar{y}=2.4,$

$$ssx=n\cdot s_x^2=0.116,ssy=m\cdot s_y^2=0.280,$$

$$s_w=\sqrt{\frac{ssx+ssy}{n+m-2}}=0.222,\sqrt{\frac{1}{n}+\frac{1}{m}}=0.645,$$

查t分布表得$t_{0.975}(6+4-2)=2.306$

$$t_{0.95}(8)=1.860,$$

由公式5计算出$\mu_1-\mu_2$的双侧0.95置信区间为（$-0.36,0.31$），

单侧0.95置信下限为-0.29,

单侧0.95置信上限为0.24。

(3)若μ_1,μ_2未知，求$\frac{\sigma_1^2}{\sigma_2^2}$的双侧0.95置信区间、单侧0.95置信下限和单侧0.95置信上限。

解 $\because n=6, m=4$,计算出

$$s_x^{*2}=0.023, s_y^{*2}=0.0933$$

查F分布表得$F_{0.975}(5,3)=14.9$,

$$F_{0.95}(5,3)=9.01,$$

$$F_{0.025}(5,3)=\frac{1}{F_{0.975}(3,5)}=\frac{1}{7.76},$$

$$F_{0.05}(5,3)=\frac{1}{F_{0.95}(3,5)}=\frac{1}{5.41},$$

由公式6计算出$\frac{\sigma_1^2}{\sigma_2^2}$的双侧0.95置信区间为（0.02,1.93），

单侧0.95置信下限为0.03,

单侧0.95置信上限为1.35。

4. 百分比的置信区间

总体$X \sim B(1, p)$, X的一个样本: $X_1, X_2, \cdots, X_n$

n充分大时,未知参数p近似的双侧$1-\alpha$置信区间:

(公式7)

$$(\bar{x} - u_{1-0.5\alpha}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + u_{1-0.5\alpha}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}})$$

公式8

$$(\frac{-b-\sqrt{b^2-4ac}}{2a}, \frac{-b+\sqrt{b^2-4ac}}{2a})$$

其中$a = n + u_{1-0.5\alpha}^2, b = -(2n\bar{x} + u_{1-0.5\alpha}^2)$

$c = n(\bar{x})^2.$

例　设100株水稻样品中有60株长势特优，试求特优率p的双侧0.95置信区间。

解　设总体为X,则$\bar{x}=0.6, u_{1-0.5\alpha}=1.96,$

由公式7计算$\sqrt{\dfrac{\bar{x}(1-\bar{x})}{n}}=0.049,$

特优率p的双侧0.95置信区间为

（0.504,0.696）。

若用公式8计算得$a = n + u_{1-0.5\alpha}^2 = 103.842,$

$$b = -(2n\bar{x} + u_{1-0.5\alpha}^2) = -123.842,$$

$$c = n(\bar{x})^2 = 36, \sqrt{b^2 - 4ac} = 19.586,$$

特优率p的双侧0.95置信区间为

$$(0.502, 0.691).$$

作业：P161, 2, 3, 7 P172，1，4, 5

第七章 假设检验
(Hypothesis Testing)

若对参数一无所知 → 用参数估计的方法处理

参数有所了解 → 猜测需证实时 → 假设检验方法

假设检验 {
- 参数假设检验
- 总体分布假设检验
}

总体分布已知，检验关于未知参数的假设

总体分布的假设检验问题

假设检验?

假设：施加于一个或多个总体的概率分布或参数的假设.

检验：从总体中抽取样本,根据样本的观测值,按一定原则进行检验, 接受或拒绝所作的假设.

小概率原理

引例1 某产品出厂检验规定: 次品率p不超过4%才能出厂. 现从一万件产品中任意抽查12件发现3件次品, 问该批产品能否出厂? 若抽查结果发现1件次品, 问能否出厂?

解 假设 $p \le 0.04$,

$p=0.04$ 代入

$$P_{12}(3)=C_{12}^{3}p^{3}(1-p)^{9}=0.0097<0.01$$

这是 **小概率事件**, 一般在一次试验中是不会发生的, 现一次试验竟然发生, 故认为假设不成立, 即该批产品次品率 $p>0.04$, 则该批产品不能出厂.

$$P_{12}(1)=C_{12}^{1}p^{1}(1-p)^{11}=0.306>0.3$$

这不是小概率事件,没理由拒绝假设,从而接受原假设, 即该批产品可以出厂.

注1 **本检验方法是 概率意义下的反证法, 故拒绝假设是有说服力的.**

§7.1 总体分布参数的假设检验

研究对象： 总体分布函数的形式已知，

部分或全部参数未知。

例：总体$X\sim N(\mu,\sigma^2)$，其中μ,σ^2为未知参数

研究方法：概率反证法。

提出假设,利用小概率原理检验假设

定义：常把一个要检验的假设记作H_0，称为原假设（或零假设*null hypothesis*），与H_0对立的假设H_1，称为备择假设(*alternative hypothesis*）。

例 某工厂在正常情况下生产的电灯泡的寿命

X(小时)~N(1600,80^2).从该工厂生产的一批灯泡中随机抽取10个灯泡，测得它们寿命为：

1450,1480,1640,1610,1500, 1600,1420,1530,1700,1550

如果标准差不变，试检验这批灯泡的寿命均值μ也是1600，或大于1600，或小于1600.

(1) H_0为$\mu = 1600$，H_1为$\mu \neq 1600$；

(2) H_0为$\mu = 1600$，H_1为$\mu > 1600$；

(3) H_0为$\mu = 1600$，H_1为$\mu < 1600$.

假设检验的**基本原理**：

小概率原理

小概率事件在一次抽样中基本上不会出现

假设检验的**基本思想**：

根据实际问题提出原假设H_0和备择假设H_1；若H_0真，但抽样结果导致了一个小概率事件发生，拒绝H_0；否则，接受H_0。

概率反证法

假设检验会不会犯错误？

假设检验的两类错误

	H_0为真	H_0为假
拒绝H_0	**第一类** 错误	正确
接受H_0	正确	**第二类** 错误

犯两类错误的概率：

$P\{$第一类错误$\}=P\{$拒绝$H_0|H_0$为真$\}=\alpha$ ，

（弃真）

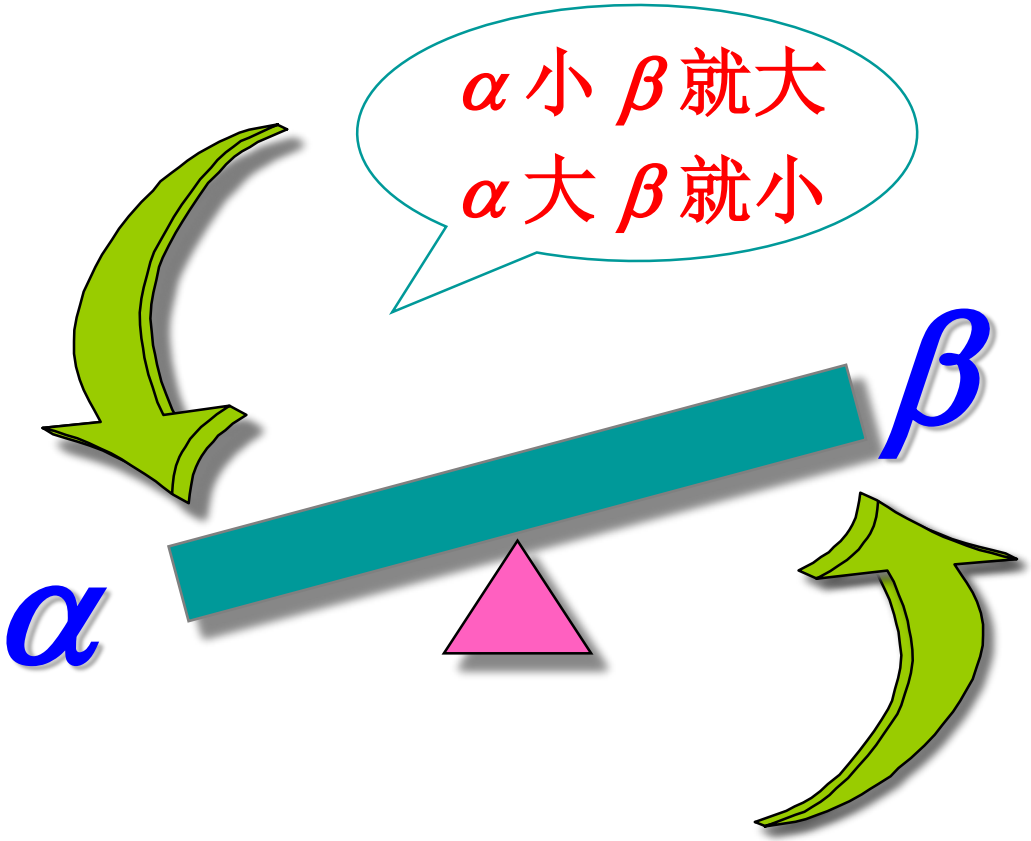
显著性水平

$P\{$第二类错误$\}=P\{$接受$H_0|H_0$为假$\}=\beta$

（存伪）

α 和 β 的关系

当样本容量固定时，一类错误概率的减少导致另一类错误概率的增加。



你不能同时减少两类错误!

只对犯第一类错误的概率加以控制，而不考虑犯第二类错误的假设检验，称为显著性检验。

α的选择要根据实际情况而定：

通常取 $\alpha = 0.1, \alpha = 0.05, \alpha = 0.01$

如果 α =0. 05，则称 μ 与 μ_0有显著的差异或差异显著；如果水平 α =0. 01，则称 μ 与 μ_0有极显著的差异或差异极显著。

假设检验的基本步骤：

(1)根据实际问题提出原假设H_0和备择假设H_1

(2)选取适当的统计量T,并在H_0成立条件下确定出T的分布

(3)确定拒绝域W,使$P\{T \in W | H_0\text{真}\}=\alpha$

(4)由样本值计算T的值，若$T \in W$，则拒绝H_0

否则,接受H_0

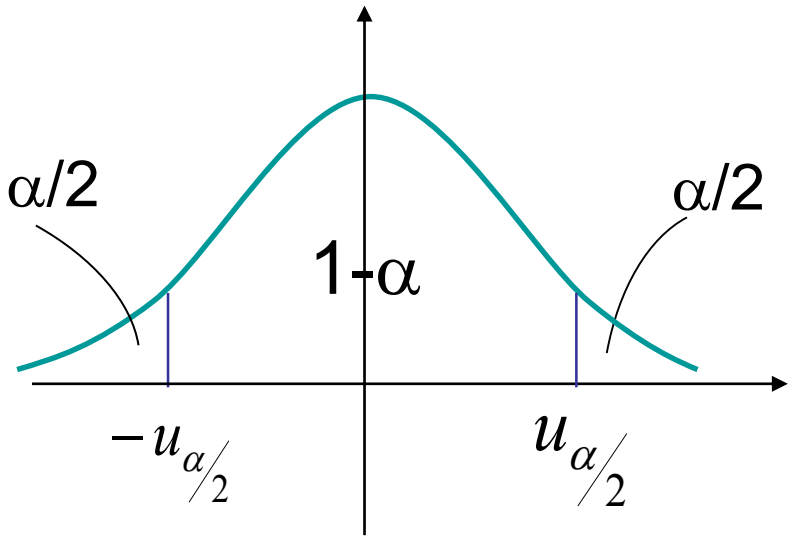
一个正态总体参数的假设检验

总体$X\sim N(\mu,\sigma^2)$，$X_1,\cdots,X_n$是来自X的样本

1. σ^2已知，μ的假设检验

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

$$U=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\overset{H_0}{\sim}N(0,1)$$

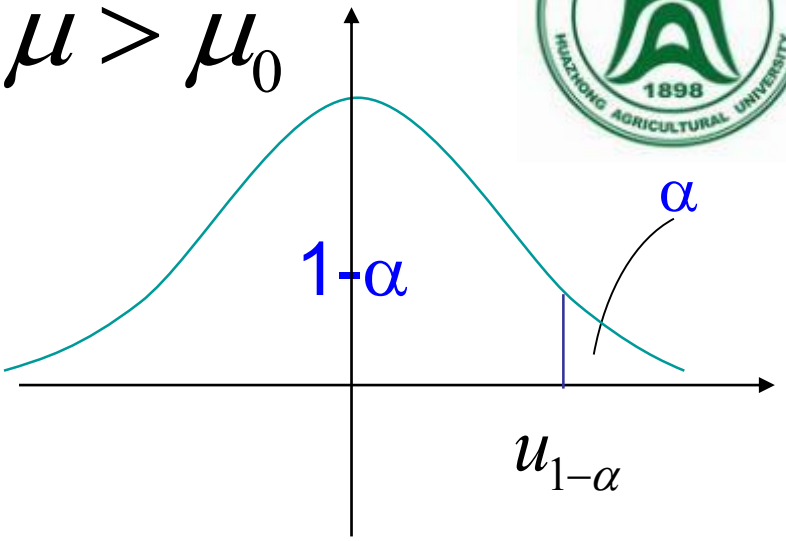


W: $|U|>u_{1-\alpha/2}$

(2) $H_0: \mu \le \mu_0, H_1: \mu > \mu_0$

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

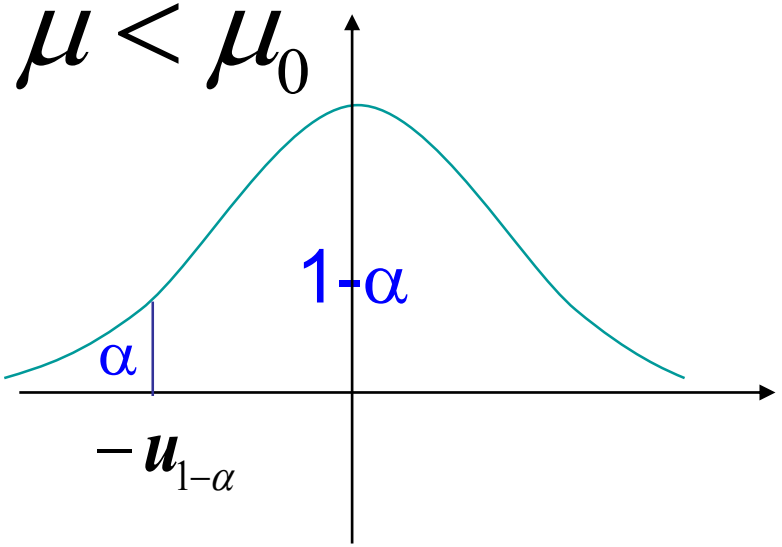
W: $U > u_{1-\alpha}$



(3) $H_0: \mu \ge \mu_0, H_1: \mu < \mu_0$

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

W: $U < -u_{1-\alpha}$



2. σ^2未知，μ的假设检验

$$T=\frac{\bar{X}-\mu_0}{S^*/\sqrt{n}} \underset{\sim}{H_0} t(n-1)$$

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

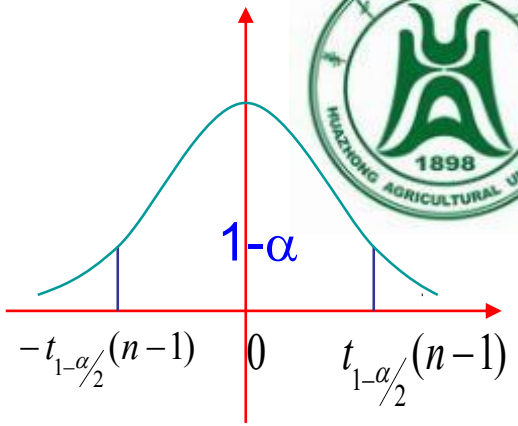
W：$|T|>t_{1-\alpha/2}(n-1)$

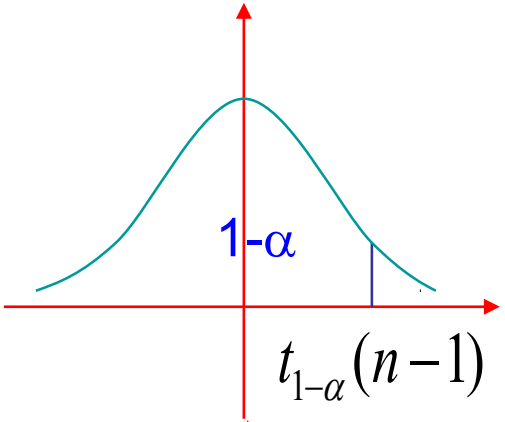
(2) $H_0:\mu\leq\mu_0, H_1:\mu>\mu_0$

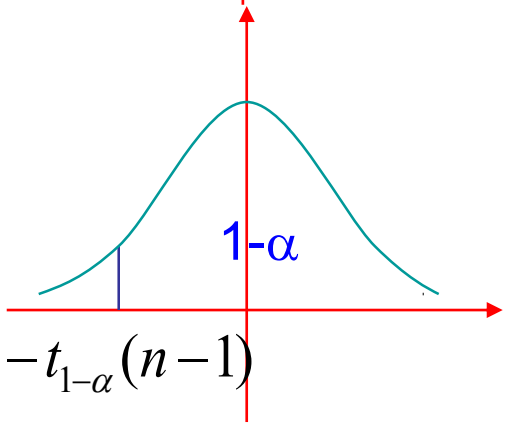
W：$T>t_{1-\alpha}(n-1)$

(3) $H_0:\mu\geq\mu_0, H_1:\mu<\mu_0$

W：$T<-t_{1-\alpha}(n-1)$







3. μ 未知，σ^2的假设检验

$$\chi^2 = \frac{(n-1)s^{*2}}{\sigma_0^2} \overset{H_0}{\sim} \chi^2(n-1)$$

(1) $H_0 : \sigma^2 = \sigma_0^2, H_1 : \sigma^2 \neq \sigma_0^2$

W： $\chi^2 > \chi^2_{1-\alpha/2}(n-1)$ 或 $\chi^2 < \chi^2_{\alpha/2}(n-1)$

(2) $H_0 : \sigma^2 \leq \sigma_0^2, H_1 : \sigma^2 > \sigma_0^2$

W： $\chi^2 > \chi^2_{1-\alpha}(n-1)$

(3) $H_0 : \sigma^2 \geq \sigma_0^2, H_1 : \sigma^2 < \sigma_0^2$

W： $\chi^2 < \chi^2_{\alpha}(n-1)$

例1 工厂用自动包装机包装葡萄糖，规定标准重量为每袋净重500克。现随机地抽取10袋，测得各袋净重为：

495，510，505，498，503，492，502，505，497，506.

设每袋净重服从~$N(\mu, \sigma^2)$，问包装机工作是否正常（$\alpha=0.05$）？

(1) 已知$\sigma=5$克；(2) 未知σ。

解 我们有 $\bar{x} = 501.3, s^* = 5.62,$

(1)已知$\sigma = 5$，试作假设检验

H_0为$\mu = 500$，$H_1 \neq 500$.

计算统计量u的观测值

$$u = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{501.3 - 500}{5/\sqrt{10}} = 0.822,$$

由α查正态分布表

$$u_{1-0.5\alpha} = u_{0.975} = 1.96,$$

$|u| < 1.96$,接受H_0，

即认为包装机工作正常

(2)未知σ，试作假设检验

H_0为$\mu = 500$，$H_1 \neq 500$.

计算统计量t的观测值

$$t = \frac{501.3 - 500}{5.62/\sqrt{10}} = 0.731,$$

由α查t分布表,自由度为ϑ,

$$t_{1-0.5\alpha} = t_{0.975} = 2.262,$$

$|t| < 2.262$,接受H_0,

即认为包装机工作正常

例1《作物栽培》已知豌豆百粒重X(单位：g)服从正态分布N(37.72,0.1089)，在改善栽培条件后随机抽出9粒，平均重量＝37.92，问改善栽培条件是否显著地提高了豌豆的百粒重，α＝0.05。

解：因为改善栽培条件不会降低豌豆籽的百粒重，所以设

H_0为$\mu=37.72$，H_1为$\mu>37.72$

$$u = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$$n = 9, \sigma^2 = 0.1089,$$
$$\mu_0 = 37.72, \overline{x} = 37.92$$

计算出u=1.818, 由α查正态分布表

$$u_{0.95} = 1.65, u > 1.65,$$

拒绝H_0,认为$\mu > 37.32.$

例2 《品种提纯》一个混杂的小麦品种，其株高的标准差为14cm，经提纯后随机地抽出10株，它们的株高(单位：cm)为90, 105, 101, 95, 100, 100, 101, 105, 93, 97，试检验提纯后的群体是否比原来的群体较为整齐，α＝0.05。

解：提纯后的群体应该比原来的群体较为整齐，故设

H_0为$\sigma^2=196$，H_1为$\sigma^2<196$，

$$\chi^2 = \frac{ss}{\sigma_0^2},$$

由观测值及$\sigma_0^2 = 196,\ n = 10,$

计算得$ss = 218.1, \chi^2 = 1.113,$

由α查χ^2分布表,自由度为9,

$\chi_\alpha^2(9) = \chi_{0.05}^2(9) = 3.33,$而$\chi^2 < 3.33,$

因此拒绝H_0,认为$\sigma^2 < 196.$

两个正态总体参数的假设检验

设总体 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$

总体X和Y相互独立，

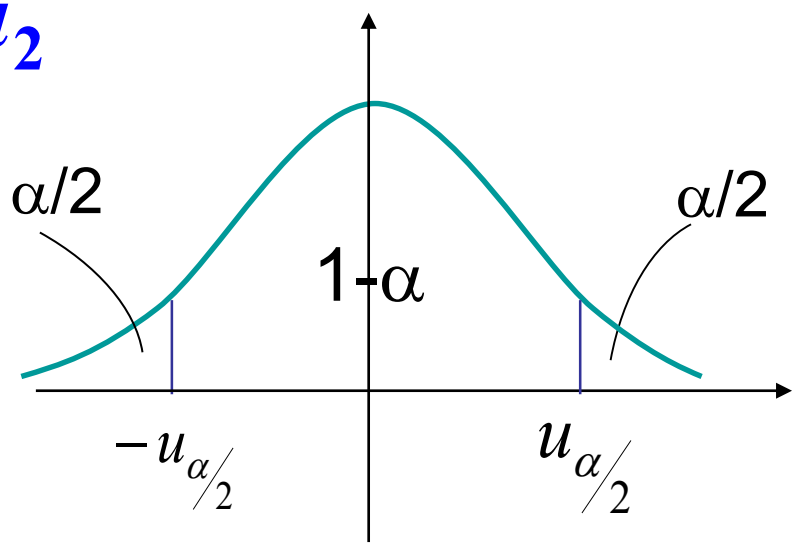
$X_1, X_2, \cdots, X_{n1}$是来自总体X的样本，

$Y_1, Y_2, \cdots, Y_{n2}$是来自总体Y的样本，

1. σ_1^2, σ_2^2已知, μ_1和 μ_2的假设检验

(1) $H_0: \mu_1=\mu_2$ $H_1: \mu_1\neq\mu_2$

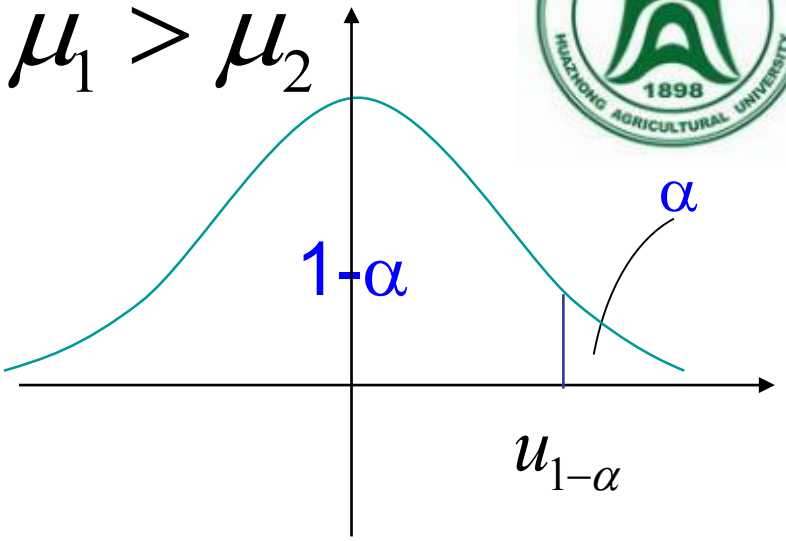
$$U=\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \overset{H_0}{\sim} N(0,1)$$



W: $|U|>u_{1-\alpha/2}$

(2) $H_0: \mu_1 \le \mu_2, H_1: \mu_1 > \mu_2$

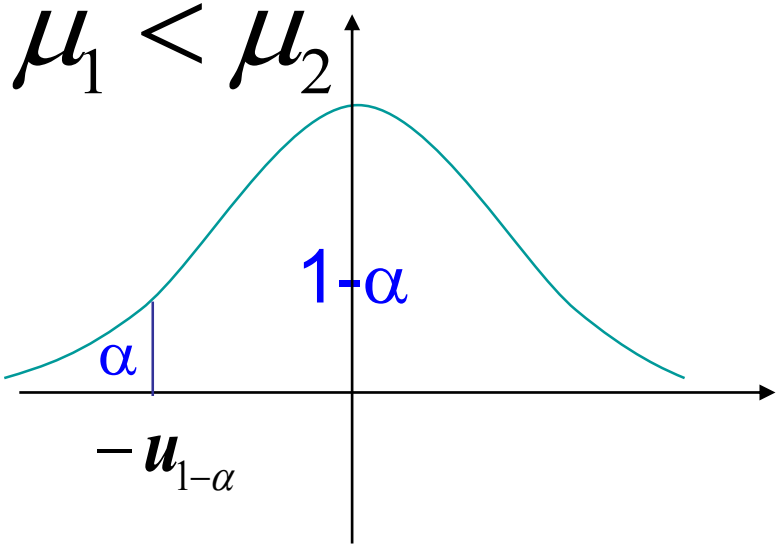
$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{H_0}{\sim} N(0,1)$$



W: $U > u_{1-\alpha}$

(3) $H_0: \mu_1 \ge \mu_2, H_1: \mu_1 < \mu_2$

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{H_0}{\sim} N(0,1)$$



W: $U < -u_{1-\alpha}$

2. σ_1^2 ,σ_2^2未知, 但$\sigma_1^2=\sigma_2^2$, μ_1 和μ_2的假设检验

$$T=\frac{\bar{X}-\bar{Y}}{S_w\sqrt{1/n_1+1/n_2}} \underset{\sim}{H_0} t(n_1+n_2-2)$$

(1) $H_0:\mu_1=\mu_2, H_1:\mu_1\neq\mu_2$

W: $|T|>t_{1-\alpha/2}(n_1+n_2-2)$

(2) $H_0:\mu_1\leq\mu_2, H_1:\mu_1>\mu_2$

W: $T>t_{1-\alpha}(n_1+n_2-2)$

(3) $H_0:\mu_1\geq\mu_2, H_1:\mu_1<\mu_2$

W: $T<-t_{1-\alpha}(n_1+n_2-2)$

3. μ_1, μ_2未知，σ_1^2/σ_2^2的假设检验

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

(1) $H_0 : \sigma_1^2 = \sigma_2^2, H_1 : \sigma_1^2 \neq \sigma_2^2$

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

(2) $H_0 : \sigma_1^2 \leq \sigma_2^2, H_1 : \sigma_1^2 > \sigma_2^2$

W: $F > F_{1-\alpha}(n_1 - 1, n_2 - 1)$

(3) $H_0 : \sigma_1^2 \geq \sigma_2^2, H_1 : \sigma_1^2 < \sigma_2^2$

W: $F < F_{\alpha}(n_1 - 1, n_2 - 1)$

例3《作物栽培》根据资料测算，某品种小麦产量(单位：Kg/ m^2)的$\sigma^2=0.4$。收获前在麦田的四周取12个样点，得到产量的均值$\bar{x}=1.2$，在麦田的中心取8个样点，得到产量的均值$\bar{y}=1.4$，试检验麦田四周及中心处每平方米产量是否有显著的差异($\alpha=0.05$)？

解：因为要检验麦田四周及中心处每平方米产量是否有显著的差异，所以设

H_0为$\mu_1=\mu_2$，H_1为$\mu_1\neq\mu_2$，

$$u=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}},$$

容量为$n=12, m=8, \sigma_1^2=\sigma_2^2=0.4,$

计算出$\bar{x}=1.2, \bar{y}=1.4, u=-0.693,$

由α查标准正态分布的分布函数值表得到

$u_{0.975}=1.96$，$|u|<1.96$，因此应该接受H_0，认为$\mu_1=\mu_2$，即麦田四周及中心处每平方米产量没有显著的差异。

例4《产量调查》调查某地每亩30万苗和50万苗的稻田各5块，分别得到亩产量800, 840, 870, 920, 850和900, 880, 890, 890, 840，试检验两种密度的亩产量是否有显著的差异？

解：本例要检验$\mu_1 \neq \mu_2$，

由于σ_1^2和σ_2^2未知，应先检验$\sigma_1^2 = \sigma_2^2$，

例中未给出显著性水平，可认为α=0.05。设

$$f = \frac{s_x^{*2}}{s_y^{*2}},$$

根据容量为n＝m＝5的两个样本观测值算出

$$s_x^{*2} = 1930, s_y^{*2} = 550, f = 3.509,$$

则由α查F分布的分位数表得到

$F_{0.975}(4,4)＝9.60,$

$$F_{0.025}(4,4) = \frac{1}{9.60}, \text{而} \frac{1}{9.60} < f < 9.60,$$

应该接受$\sigma_1^2 = \sigma_2^2$.

下面检验$\mu_1 \neq \mu_2$，设

H_0为$\mu_1 = \mu_2$，H_1为$\mu_1 \neq \mu_2$，

$$t = \frac{\bar{x} - \bar{y}}{s_w \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

式中$s_w^2 = \dfrac{ssx + ssy}{n + m - 2}$，

根据容量为n＝m＝5的两个样本观测值算出

$$s_x^{*2} = 1930, s_y^{*2} = 550,$$

$$ssx = 7720, ssy = 2200,$$

$$\bar{x} = 856, \bar{y} = 880,$$

$$s_w^2 = 1240, t = -1.078,$$

由α查t分布表得 $t_{0.975}(8) = 2.306,$

$|t| < 2.306,$ 因此认为$\mu_1 = \mu_2,$

即两种密度的亩产量没有显著的差异。

作业：P186, 2, 7，11

第八章 方差分析

Analysis of Variance

方差分析概述

- 如农作物的产量受品种、肥料、气候、雨水、光照、土壤、播种量等众多因素的影响；
- 产品销售量受品牌、质量、价格、促销手段、竞争产品、顾客偏好、季节、居民收入水平等众多因素的影响；

因此需要了解：

- 1）哪些因素会对所研究的指标产生显著影响；
- 2）这些影响因素在什么状况下可以产生最好的结果。
- 方差分析就是解决这类问题的一种统计分析方法。

【引例】哪种促销方式效果最好？

- 某大型连锁超市为研究各种促销方式的效果，选择下属 4 个门店，分别采用不同促销方式，对包装食品各进行了4 个月的试验。

促销方式	与上年同期相比(%)			
A_1 (广告宣传)	104.8	95.5	104.2	103.0
A_2 (有奖销售)	112.3	107.1	109.2	99.2
A_3 (特价销售)	143.2	150.3	184.7	154.5
A_4 (买一送一)	145.6	111.0	139.8	122.7

超市管理部门希望了解：

(1)不同促销方式对销售量是否有显著影响？

(2)哪种促销方式的效果最好？

§8.1 单因素试验的方差分析

1. 单因素试验及有关的基本概念

在试验中，有可能影响试验指标并且有可能加以控制的试验条件称为因素。通过试验的设计，在试验中只安排一个因素有所变化、取不同的状态或水平，而其余的因素都在设计的状态或水平下保持不变的试验称为单因素试验。

单因素试验的因素为A

共有 A_1、A_2、…、A_r等r个水平、

安排了n_1、n_2、…、n_r次重复试验，

所得到的样本为 X_{i1}、X_{i2}、…、X_{ini}，

相应的观测值为 x_{i1}、x_{i2}、…、x_{ini}

$(i=1…r,\quad n_1+n_2+…+n_r= n)$

水平	观测值			
A_1	x_{11}	x_{12}	…	x_{1n1}
A_2	x_{21}	x_{22}	…	x_{2n2}
…		…		
A_r	x_{r1}	x_{r2}	…	x_{rnr}

设 $\bar{x}_{i\cdot}=\frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij}$,

$$\bar{x}_{\cdot\cdot}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}=\frac{1}{n}\sum_{i=1}^{r}n_i\bar{x}_{i\cdot}$$

在单因素试验中，假设有r个编号为i＝1至r的正态总体，它们分别服从$N(\mu_i,\sigma^2)$分布，

当μ_i及σ^2未知时，要根据取自这r个正态总体的r个相互独立且方差相同的样本检验原假设：

$$H_0:\mu_1=\mu_2=\cdots=\mu_r$$

所作的检验以及对未知参数的估计称为方差分析。

注：r个正态总体;各样本相互独立;方差相同.

$$\mu = \frac{1}{n}\sum_{i=1}^{r} n_i \mu_{i\cdot}$$

μ称为总平均值，

$\alpha_i = \mu_i - \mu$，　α_i 称为因素A的水平A_i的效应

则 (1) $x_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,

(2) $x_{ij} - \mu = \alpha_i + \varepsilon_{ij}$,

各个 ε_{ij} 称为随机误差，它们相互独立且都服从$N(0, \sigma^2)$分布，原假设H_0则等价于各 $\alpha_i=0$（i=1至r）。

2. 总离均差平方和的分解

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{..})^2,$$ 总离均差平方和

（反映全部数据之间的差异）

$$SSE=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2,$$ 误差平方和

（反映各个样本的数据与样本均值之间的差异，由各种随机因素所引起的）

$$SSA=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_{i.}-\bar{x}_{..})^2=\sum_{i=1}^{r}n_i(\bar{x}_{i.}-\bar{x}_{..})^2.$$

效应平方和或组间平方和

（反映各个样本均值之间的差异，由因素A的不同水平所引起的系统误差）

结论1）SST=SSE+SSA；

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{..})^2$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{n_i}[(x_{ij}-\bar{x}_{i.})+(\bar{x}_{i.}-\bar{x}_{..})]^2$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2+\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_{i.}-\bar{x}_{..})^2$$

$$+2\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})(\bar{x}_{i.}-\bar{x}_{..})$$

而$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})(\bar{x}_{i.}-\bar{x}_{..})$

$$=\sum_{i=1}^{r}[(\bar{x}_{i.}-\bar{x}_{..})\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})]=0.$$

结论2） $\dfrac{SSE}{\sigma^2} \sim \chi^2(n-r);$

结论3）当H_0为真时，$\dfrac{SSA}{\sigma^2} \sim \chi^2(r-1);$

结论4）当H_0为真时，SSE、SSA相互独立；

结论5）当H_0为真时，$MSA = \dfrac{SSA}{r-1}, MSE = \dfrac{SSE}{n-r}$时，

$$F = \frac{MSA}{MSE} \sim F(r-1, n-r),$$

当$F \geq F_{1-\alpha}(r-1, n-r)$时拒绝$H_0$.

3. 总体中未知参数的估计

$$(1)\hat{\mu}=\bar{x}_{..},\hat{\mu}_i=\bar{x}_{i.},\hat{\alpha}_i=\bar{x}_{i.}-\bar{x}_{..},$$

且$$E(\bar{x}_{..})=\mu,E(\bar{x}_{i.})=\mu_i,E(\bar{x}_{i.}-\bar{x}_{..})=\alpha_i.$$

$$(2)x_{ij}=\bar{x}_{i.}+(x_{ij}-\bar{x}_{i.})$$

$$=\bar{x}_{..}+(\bar{x}_{i.}-\bar{x}_{..})+(x_{ij}-\bar{x}_{i.})$$

为单因素试验的方差分析的数学模型的估计式，

而$$x_{ij}-\bar{x}_{..}=(\bar{x}_{i.}-\bar{x}_{..})+(x_{ij}-\bar{x}_{i.})$$

为效应分解的估计式。

$$(3)\hat{\sigma}^2 = MSE = \frac{SSE}{n-r},\ \text{且} E(MSE) = \sigma^2.$$

(4) 当放弃原假设H_0且u≠v时，均值差 $\mu_u - \mu_v$ 的双侧1−α 置信区间可表示为

$$(\bar{x}_{u\cdot} - \bar{x}_{v\cdot} \pm \Delta_{uv}),$$

$$\Delta_{uv} = t_{1-0.5\alpha}(n-r)\sqrt{MSE\left(\frac{1}{n_u} + \frac{1}{n_v}\right)}.$$

4. 单因素试验的方差分析的步骤

(1)计算 $T_{i\cdot}=\sum_{j=1}^{n_i}x_{ij},\bar{x}_{i\cdot},T=\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}=\sum_{i=1}^{r}T_{i\cdot}$ 及 $\bar{x}_{\cdot\cdot}$;

(2)计算 $C=\frac{T^2}{n},SST=\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}^2-C,$

$$SSA=\sum_{i=1}^{r}\frac{T_{i\cdot}^2}{n_i}-C,\ SSE=SST-SSA;$$

(3)计算 MSA,MSE 及 $F=\frac{MSA}{MSE}$;

(4)给出α,确定分位数$F_{1-\alpha}(r-1,n-r)$;

(5) 列出方差分析表：

方差来源	平方和	自由度	均方和	F值	显著性
因素A	SSA	r-1	MSA	F	
误差	SSE	n-r	MSE		
总和	SST	n-1			

其中显著性一栏应写出$F_{1-\alpha}(r-1,n-r)$比较的结果

通常取$\alpha = 0.05$或$\alpha = 0.01$,

$F > F_{0.99}(r-1, n-r)$时写**,

$F_{0.95}(r-1, n-r) < F < F_{0.99}(r-1, n-r)$

时写*，$F < F_{0.95}(r-1, n-r)$时写N;

(6) 写出假设检验的结论。

例《切胚乳试验》用小麦种子进行切胚乳试验，设计分3种处理，同期播种在条件较为一致的花盆内，出苗后每盆选留2株，成熟后测量每株粒重(单位：g)，得到数据如下：

处理	每株粒重
未切去胚乳	21,29,24,22,25,30,27,26
切去一半胚乳	20,25,25,23,29,31,24,26,20,21
切去全部胚乳	24,22,28,25,21,26

试作方差分析，估计各个总体的未知参数 μ_i 和 μ，如有必要，试求出两两总体均值差的双侧0. 95置信区间。

解： 设H_0为各个未知参数μ_i相等，也就是各个处理之间没有显著的差异。

(1)计算$T_{i\cdot}=\sum_{j=1}^{n_i}x_{ij}, \bar{x}_{i\cdot}, T=\sum_{i=1}^{r}T_{i\cdot}$及$\bar{x}_{\cdot\cdot}$并列表

处理	n_i	$T_{i\cdot}$	$\bar{x}_{i\cdot}$	$\sum_{j=1}^{n_i}x_{ij}^2$
未切去胚乳	8	204	25.50	5272
切去一半胚乳	10	244	24.40	6074
切去全部胚乳	6	146	24.33	3586
总和	24	594		14932
$\bar{x}_{\cdot\cdot}$	24.75			

(2) 计算校正数 $C=\frac{T^2}{n}=14701.5,$

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}^2=14932,$$

$$SST=\sum_{i=1}^{r}\sum_{j=1}^{n_i}x_{ij}^2-C=230.5,$$

$$\sum_{i=1}^{r}\frac{T_{i\cdot}^2}{n_i}=14708,\qquad SSA=\sum_{i=1}^{r}\frac{T_{i\cdot}^2}{n_i}-C=6.77,$$

$$SSE=SST-SSA=223.73;$$

(3) 计算r-1=2，n-r=21，MSA=3.39，MSE=10.65，F=0.32；

(4) 给出α=0.05，查表得到分位数 $F_{0.95}(2,21)=3.49$；

(5) 列出方差分析表：

方差来源	平方和	自由度	均方和	F值	显著性
A	6.77	2	3.39	0.32	N
误差	223.73	21	10.65		
总和	230.50	23			

因此接受H_0，认为各个未知参数μ_i相等，认为各个处理之间没有显著的差异。

所求的未知参数μ_i和μ的估计为：

$$\hat{\mu}_1 = \bar{x}_{1\cdot} = 25.50, \hat{\mu}_2 = \bar{x}_{2\cdot} = 24.40,$$

$$\hat{\mu}_3 = \bar{x}_{3\cdot} = 24.33, \hat{\mu} = \bar{x}_{\cdot\cdot} = 24.75.$$

又因为各个处理之间没有显著的差异，也就没有必要求两两总体均值差的双侧0.95置信区间。

【引例】哪种促销方式效果最好？

- 某大型连锁超市为研究各种促销方式的效果，选择下属 4 个门店，分别采用不同促销方式，对包装食品各进行了4 个月的试验。

促销方式	与上年同期相比(%)			
A_1 (广告宣传)	104.8	95.5	104.2	103.0
A_2 (有奖销售)	112.3	107.1	109.2	99.2
A_3 (特价销售)	143.2	150.3	184.7	154.5
A_4 (买一送一)	145.6	111.0	139.8	122.7

超市管理部门希望了解：

(1)不同促销方式对销售量是否有显著影响？

(2)哪种促销方式的效果最好？

引例 1 的方差分析表

差异源	SS	df	MS	F	P-value	F crit
组间	7925.4	3	2641.8	16.628	0.00014	3.4903
组内	1906.5	12	158.87			
总计	9831.9	15				

故不同的促销方式对商品销售额有极高度显著影响。

$$H_0 : m_1 = m_3, H_1 : m_1 > m_3$$

统计量

$$t = \frac{\bar{x}_{1\cdot} - \bar{x}_{3\cdot}}{\sqrt{\frac{SSE}{n-r}}\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}} \sim t(n-r)$$

- μ_1 μ_2 μ_4 μ_3
- (广告宣传) μ_1
- (有奖销售) μ_2
- (买一送一) μ_4 * *
- (特价销售) μ_3 * * *

- μ_1 μ_2 μ_4 μ_3
- (广告宣传) μ_1
- (有奖销售) μ_2
- (买一送一) μ_4 * *
- (特价销售) μ_3 * * *

方差分析结论：

特价销售的效果最好，
买一送一次之，
广告宣传和有奖销售的效果最差，
两者间无显著差异。

作业：P208, 1，2,

第九章
回归分析与协方差分析

§9.1 一元线性回归

1. 一元线性回归的基本概念

回归方程:

用来分析自变量x取值与因变量Y取值的内在联系

（自变量x是确定性的变量，因变量Y是随机性的变量）

进行n次独立试验，测得数据如下：

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

问题：如何根据这些观测值用“最佳的”形式来表达 变量Y与X之间的相关关系？

一般而言，在变量x取值以后，若Y所取的值服从$N(\alpha+\beta x, \sigma^2)$分布，当$\alpha$、$\beta$及$\sigma^2$未知时，根据样本$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$的观测值$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$对未知参数$\alpha$、$\beta$及$\sigma^2$所作的估计与检验称为一元线性回归分析，而$\alpha$称为截距，$\beta$称为回归系数，

$$E(Y) = \alpha + \beta x \qquad E(Y)= \hat{y}$$

称为回归方程。

由回归方程可以推出

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad (i = 1,2,\cdots,n)$$

式中的ε_i相互独立且$\varepsilon_i \sim N(0,\sigma^2)$.

根据样本及其观测值可以得到α、β及σ^2的估计量及估计值 $\hat{\alpha},\hat{\beta}$和$\hat{\sigma}^2$,

得到回归方程的估计式或经验回归方程

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

确定未知参数 $\hat{\alpha}, \hat{\beta}$和$\hat{\sigma}^2$， 最常用的是最小二乘法

即求出

$$\hat{\alpha} = a, \hat{\beta} = b及\hat{y}_i = a + bx_i使$$

$$\min Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

a称为截距，b称为回归系数.

$$\hat{\sigma}^2 = \frac{Q}{n-2}是\sigma^2的无偏估计。$$

2. 总体中未知参数的估计

由 $\frac{\partial Q}{\partial a}=0,\frac{\partial Q}{\partial b}=0,Q=\sum_{i=1}^{n}(y_i-a-bx_i)$得

$$\begin{cases}-2\sum_{i=1}^{n}[y_i-(a+bx_i)]=0,\\ -2\sum_{i=1}^{n}[y_i-(a+bx_i)]x_i=0,\end{cases}$$

得到一元线性回归的正规方程组

$$\begin{cases}na+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i,\\ a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i,\end{cases}$$

并求出

$$b=\frac{l_{xy}}{l_{xx}},a=\bar{y}-b\bar{x},$$

$$l_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2,$$

$$l_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i),$$

$$l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\frac{1}{n}(\sum_{i=1}^{n}y_i)^2.$$

建立一元线性回归方程的具体步骤：

(1)计算$\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} y_i^2, \sum_{i=1}^{n} x_i y_i$；

(2)计算l_{xx}, l_{xy}, l_{yy}；

(3) 计算b和a，写出一元线性回归方程。

$$b = \frac{l_{xy}}{l_{xx}}, a = \bar{y} - b\bar{x},$$

将a、b和SSE以及$\hat{Y}$和$\hat{Y}_i$看作是统计量，它们的表达式分别为

$$b=\frac{l_{xy}}{l_{xx}},a=\bar{y}-b\bar{x},$$

$$SSE=\sum_{i=1}^{n}[Y_i-(a+bx_i)]^2,$$

$$\hat{Y}=a+bx=\overline{Y}+b(x-\bar{x}),$$

$$\hat{Y}_i=\overline{Y}+b(x_i-\bar{x}).$$

这些统计量之间以及它们与总体参数之间有以下的结论：

(1)$\bar{Y}, b$与SSE相互独立；

(2)$E(b)=\beta, E(a)=\alpha,$

$$D(b)=\frac{\sigma^2}{l_{xx}}, D(a)=(\frac{1}{n}+\frac{\bar{x}^2}{l_{xx}})\sigma^2;$$

注：①为提高a的估计精度，最理想的选择是使 $\bar{x}=0$，其绝对值越小越好；

②为提高b的估计精度，应该使l_{xx} 取较大的数值，x_1、x_2、…、x_n越分散越好；

③观测值的个数n不能太少。

$$(3)E(SSE)=(n-2)\sigma^2,$$

即$\hat{\sigma}^2=\dfrac{SSE}{n-2}$是$\sigma^2$的无偏估计

(4)b和a以及$\hat{Y}$都服从正态分布

而$\dfrac{SSE}{\sigma^2}\sim\chi^2(n-2).$

$$SSE=\sum_{i=1}^{n}[y_i-\hat{y}_i]^2$$

3. 线性回归方程的显著性检验

建立回归方程的前提：

在变量x取值以后，Y所取的值要服从$N(\alpha+\beta x, \sigma^2)$分布. 然后，根据最小二乘法估计方程中的未知参数。

在建立回归方程的时候，并不知道Y所取的值是否服从$N(\alpha+\beta x, \sigma^2)$分布。因此，必须对回归方程的拟合情况或效果作显著性检验。

显著性检验理论基础：总平方和的分解.

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

$$\because \sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}),$$

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}(y_i-a-bx_i)(a+bx_i-\bar{y})$$

$$=\sum_{i=1}^{n}[y_i-(\bar{y}-b\bar{x})-bx_i][b(x_i-\bar{x})]$$

$$=\sum_{i=1}^{n}b(x_i-\bar{x})(y_i-\bar{y})-\sum_{i=1}^{n}b^2(x_i-\bar{x})^2$$

$$=bl_{xy}-b^2l_{xx}=0.$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

$$l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$ 称为总平方和. 记作SST

$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$是$y_i$与$\hat{y}_i$之间的偏差

称为剩余平方和，记作SSE.

$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$表示$n$个$\hat{y}_i$之间的差异，

称为回归平方和，记作SSR。

$$SST = SSE + SSR$$

如果SSR的数值较大，SSE的数值便比较小，说明回归的效果好；

如果SSR的数值较小，SSE的数值便比较大，说明回归的效果差。

$$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(a+bx_i-\bar{y})^2$$

$$=\sum_{i=1}^{n}(\bar{y}-b\bar{x}+bx_i-\bar{y})^2=\sum_{i=1}^{n}b^2(x_i-\bar{x})^2$$

$$=b^2l_{xx}=bl_{xy},$$

$$\therefore SSE=SST-bl_{xy}.$$

$$SSE=SST-bl_{xy}=l_{yy}-\frac{l_{xy}^2}{l_{xx}}=l_{yy}\left(1-\frac{l_{xy}^2}{l_{xx}l_{yy}}\right)$$

现引进$r^2=\dfrac{l_{xy}^2}{l_{xx}l_{yy}}$，$r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$.

如果|r|较大，SSE的数值便比较小，说明回归的效果好或者说x与Y的线性关系密切；

如果|r|较小，SSE的数值便比较大，说明回归的效果差或者说x与Y的线性关系不密切；

称r为x与Y的观测值的相关系数

由r及回归系数的计算公式

$$b = \frac{l_{xy}}{l_{xx}},$$

r>0时b>0，*x*增加时Y的观测值呈增加的趋势；

r<0时b<0，*x*增加时Y的观测值呈减少的趋势。

因此r>0时称*x*与Y正相关，r<0时称*x*与Y负相关。

设H_0为β＝0，也就是假设*x*与Y不是线性关系. 则可以用以下三种实质相同的方法检验线性回归方程的显著性，且当检验的结果显著时*x*与Y的线性关系显著，回归方程可供应用；当检验的结果不显著时*x*与Y的线性关系不显著，回归方程不可应用。

（1）F检验法：$\because \dfrac{SSE}{\sigma^2} \sim \chi^2(n-2),$

当H_0为$\beta=0$真时，$\dfrac{SSR}{\sigma^2} \sim \chi^2(1);$

且SSR与SSE相互独立；

$$F = \frac{SSR/1}{SSE/(n-2)} \sim F(1, n-2),$$

当$F \geq F_{1-\alpha}(1, n-2)$时应该放弃原假设$H_0$。

(2) t检验法：

$$\because b \sim N(\beta, \frac{\sigma^2}{l_{xx}}), \frac{SSE}{\sigma^2} \sim \chi^2(n-2),$$

当H_0为β＝0为真时，

$$T = b\sqrt{\frac{l_{xx}}{SSE/(n-2)}} \sim t(n-2),$$

当$|t| \geq t_{1-0.5\alpha}(n-2)$时应该放弃原假设$H_0$。

(3) r检验法：根据x与Y的观测值的相关系数

$$r=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}},r^2=\frac{l_{xy}^2}{l_{xx}l_{yy}},$$

可以推出 $r^2=\frac{SSR}{SST}$.

当H_0为$\beta=0$为真时，

$$F=\frac{r^2}{(1-r^2)/(n-2)}\sim F(1,n-2),$$

当$F \geq F_{1-\alpha}(1,n-2)$或$|r| \geq r_{\alpha}(n-2)$时应该放弃原假设$H_0$，式中的

$$r_{\alpha}(n-2)=\sqrt{\frac{F_{1-\alpha}(1,n-2)}{F_{1-\alpha}(1,n-2)+(n-2)}}$$

可由r检验用表中查出。

4. 利用回归方程进行点预测和区间预测

若线性回归作显著性检验的结果是放弃H_0，也就是放弃回归系数 $\beta = 0$的假设，便可以利用回归方程进行点预测和区间预测.

(1) 当$x = x_0$时，用$\hat{y}_0 = a + bx_0$预测Y_0的观测值

y_0称为点预测

由于$E(\hat{y}_0) = \alpha + \beta x_0 = E(Y_0)$,

Y_0的观测值y_0的点预测是无偏的。

(2) 若Y与样本中的各Y相互独立，则根据$Z=Y_0-(a+bx_0)$服从正态分布，$E(Z)=0$，

$$D(Z)=\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right),$$

及$\frac{SSE}{\sigma^2}\sim\chi^2(n-2)$，Z与SSE相互独立，

$$t=\frac{Z}{\sqrt{\frac{SSE}{n-2}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)}}\sim t(n-2).$$

因此，Y_0的1-α预测区间为 $a+bx_0\pm\Delta(x_0)$，

$$\Delta(x_0)=t_{1-0.5\alpha}(n-2)\sqrt{\frac{SSE}{n-2}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)}.$$

例《吸附方程》某种物质在不同温度下可以吸附另一种物质，如果温度x(单位：℃)与吸附重量Y(单位：mg)的观测值如下表所示：

温度x	1.5	1.8	2.4	3.0	3.5	3.9	4.4	4.8	5.0
重量y	4.8	5.7	7.0	8.3	10.9	12.4	13.1	13.6	15.3

试求线性回归方程并用三种方法作显著性检验，若$x_0=2$，求Y_0的0.95预测区间。

解：根据上述观测值得到n＝9，

$$\sum_{i=1}^{9} x_i = 30.3, \sum_{i=1}^{9} y_i = 91.11,$$

$$\sum_{i=1}^{9} x_i^2 = 115.11, \sum_{i=1}^{9} x_i y_i = 345.09,$$

$$\sum_{i=1}^{9} y_i^2 = 1036.65,$$

$$l_{xx} = 13.100, l_{xy} = 38.387, l_{yy} = 114.516,$$

$$\bar{x} = 3.367, \bar{y} = 10.122,$$

$$b = \frac{l_{xy}}{l_{xx}} = 2.9303, a = \bar{y} - b\bar{x} = 0.2569(0.2558),$$

所求的线性回归方程为 $\hat{y} = 0.2569 + 2.9303x$

显著性检验方法

(1) F检验法：

SST＝l_{yy}＝114.516，　SSR＝bl_{xy}＝112.485，

SSE＝SST-bl_{xy}＝2.031，n-2＝7，

$F_{0.99}(1,7)=12.2$，

$$F=\frac{SSR}{SSE/(n-2)}=387.69, F>12.2,$$

所以回归方程极显著；

(2) t检验法：

$$|t| = |b|\sqrt{\frac{l_{xx}}{SSE/(n-2)}} = 19.69,$$

$$t_{0.995}(7) = 3.499, |t| > 3.499,$$

所以回归方程极显著；

(3) r检验法：

$$r^2 = \frac{l_{xy}^2}{l_{xx}l_{yy}} = 0.9823, r = 0.9911,$$

$$r_{0.01}(7) = 0.7977, |r| > 0.7977,$$

所以回归方程极显著.

当$x_0 = 2$时，$\hat{y}_0 = 6.12, \Delta(x_0) = 1.43,$

Y_0的0.95预测区间为(4.09,8.15)。

这说明当温度为2时，应该预测吸附另一种物质的重量在4. 09至8. 15之间，并且预测100次将有95次是正确的。

例《植物保护》 一些夏季害虫的盛发期与春季温度有关，现有1956-1964年间3月下旬至4月中旬旬平均温度的累计数x和一代三化蛾盛发期Y(以5月10日为0)的观测值如下：

温度x	35.5	34.1	31.7	40.3	36.8	40.2	31.7	39.2	44.2
盛发期y	12	16	9	2	7	3	13	9	-1

试求线性回归方程并用三种方法作显著性检，若$x_0=40$，求Y_0的0.95预测区间。

解： 根据上述观测值得到n＝9，

$$\sum_{i=1}^{9} x_i = 333.7, \sum_{i=1}^{9} y_i = 70,$$

$$\sum_{i=1}^{9} x_i^2 = 12517.49, \sum_{i=1}^{9} x_i y_i = 2436.4,$$

$$\sum_{i=1}^{9} y_i^2 = 794,$$

$$l_{xx} = 144.6356, l_{xy} = -159.0444, l_{yy} = 249.5556,$$

$$\bar{x} = 37.077, \bar{y} = 7.7778,$$

$$b = \frac{l_{xy}}{l_{xx}} = -1.0996, a = \bar{y} - b\bar{x} = 48.5493,$$

所求的线性回归方程为

$$\hat{y} = 48.5 - 1.1x;$$

显著性检验方法

(1) F检验法：

SST＝l_{yy}＝249.5556，SSR＝bl_{xy}＝174.8886

SSE＝SST-bl_{xy}＝74.6670，n-2＝7，

$F_{0.99}(1,7)=12.2$，

$$F=\frac{SSR}{SSE/(n-2)}=16.40, F>12.2,$$

所以回归方程极显著；

(2) t检验法：

$$|t| = |b|\sqrt{\frac{l_{xx}}{SSE/(n-2)}} = 4.05,$$

$$t_{0.995}(7) = 3.499, |t| > 3.499,$$

所以回归方程极显著；

(3) r检验法：

$$r^2 = \frac{l_{xy}^2}{l_{xx} l_{yy}} = 0.7008, r = -0.8371,$$

$$r_{0.01}(7) = 0.7977, |r| > 0.7977,$$

所以回归方程极显著.

$$当 x_0 = 40 时, \hat{y}_0 = 4.56, \Delta(x_0) = 8.36,$$

Y_0的0.95预测区间为(-3.80,12.92)。

这说明当3月下旬至4月中旬旬平均温度的累计数为40时，应该预测一代三化蛾盛发期为5月6日至5月23日之间，并且预测100次将有95次是正确的。

作业：P232, 2, 4