数理统计

第一章、线性回归与相关检验

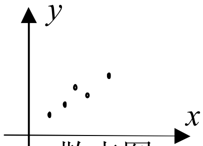
第一节　线性回归

一、两个变量间的相关关系

引例、某种产品的广告费用支出x（千元）与销售额y（万元）之间有如下的对应数据：

x	2	4	5	6	8
y	3	4	6	5	7

请画出上表数据的散点图；



散点图

1、相关关系：变量 x 与 y 有一定的关系，但没有确定函数关系

2、做散点图：把变量 x 与 y 的一组相应测量值在坐标平面上描出的一些点组成的图形

3、正负相关：在散点图中从左到右是向上的叫正相关，从左到右是向下的叫负相关

4、线性相关：当散点图中的点分布从整体上看大致分布在一条直线的附近，就称这两个变量有线性相关关系，这条直线叫回归直线

二、最小二乘法

对一个零件长度 x 的 n 次测量假设测量值为 $x_1,x_2,\cdots x_n$，如何取 $\hat{x}$ 做为 x 的估计的值更合理呢？直观的想法是 $\hat{x}$ 的值应该最接近这些测量数据，数学描述就是：$\hat{x}$ 的值应该使所有的误差平方和 $\sum_{i=1}^{n}(\hat{x}-x_i)^2$ 达到最小.

$$f(\hat{x})=\sum_{i=1}^{n}(\hat{x}-x_i)^2=n\hat{x}^2-2\hat{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}x_i^2$$

当 $\hat{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 时，$f(\hat{x})$ 达到最小. 即用测量数据的平均值作为零件估计值最好. 这种估计方法就叫最小二乘法. 最小二乘法的优点是：有效利用了全部测量数据，使误差平方和达到最小.

使误差平方和最小的估计方法叫做最小二乘法。(二乘就是平方的意思)

三、用最小二乘法推导回归直线的方程

设有线性相关系系的变量 x,y 的一组测量值是

x	x_1	x_2	…	x_n
y	y_1	y_2	…	y_n

设回归直线的方程 $\hat{y}=bx+a$，则可由最小二乘法确定 a, b 的值

误差平方和 $M=(bx_1+a-y_1)^2+(bx_2+a-y_2)^2+\cdots+(bx_n+a-y_n)^2$

按a整理得

$$M＝na^2+2a[(bx_1-y_1)+(bx_2-y_2)+\cdots+(bx_n-y_n)]+?$$

$$=na^2+2a[b(x_1+x_2\cdots+x_n)-(y_1+y_2+\cdots+y_n)]+?$$

按b整理得

$$M=nb^2+2b[(x_1(a-y_1)+x_2(a-y_2)+\cdots+x_n(a-y_n)]+?$$

$$=(x_1^2+x_2^2+\cdots+x_n^2)b^2+2b[a(x_1+x_2+\cdots+x_n)-(x_1y_1+x_2y_2+\cdots+x_ny_n)]+?$$

要使 M 取最小值，只要

$$a=\frac{(y_1+y_2+\cdots+y_n)-b(x_1+x_2\cdots+x_n)}{n}=\bar{y}-\bar{x}b(1)$$

$$b=\frac{(x_1y_1+x_2y_2+\cdots+x_ny_n)-a(x_1+x_2+\cdots+x_n)}{x_1^2+x_2^2+\cdots+x_n^2}$$

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

(1) 代入(2)得，$b\sum_{i=1}^{n} x_i^2=\sum_{i=1}^{n} x_i y_i-n\bar{x}\bar{y}+n\bar{x}^2 b$,于是$b=\dfrac{\sum_{i=1}^{n} x_i y_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2-n\bar{x}^2}$，

代入（1）得$a=\bar{y}-\bar{x}b$

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

注意由于$\bar{y}=b\bar{x}+a$，因此回归直线过点$(\bar{x},\bar{y})$，此点叫**样本中心**

例 1、相关变量 x 与 y 的一组测量值如表

x	3	4	5	6
y	2.5	3	4	4.5

(1)请画出引例数据的散点图;

(2)请根据引例提供的数据，求出 y 关于 x 的线性回归方程;

解：(2) **设线性回归方程为** $\hat{y}=bx+a,$

x	3	4	5	6	$\bar{x}=4.5$
y	2.5	3	4	4.5	$\bar{y}=3.5$
xy	7.5	12	20	27	$\sum_1=66.5$
X^2	9	16	25	36	$\sum_2=86$

$b=\dfrac{\sum_{i=1}^{4} x_i y_i-4\bar{x}\bar{y}}{\sum_{i=1}^{4} x_i^2-4\bar{x}^2}=\dfrac{66.5-4\times4.5\times3.5}{86-4\times4.5^2}$

$=\dfrac{66.5-63}{5}=0.7$

$a=3.5-0.7\times\dfrac{9}{2}=0.35$，所以线性回归方程为$y=0.7x+0.35$；

练习

1、口答、

(1)下面哪些变量是相关关系　(　　　)

A.出租车费与行驶的里程　　　B..圆的周长和它的半径之间的关系

C.身高与体重　　　　　D.铁的大小与质量

(2)有关线性回归的说法中,不正确的是　(　　)

A.相关关系的两个变量不是因果关系　　B.散点图能直观地反映数据的相关程度

C.回归直线最能代表线性相关的两个变量之间的关系　D.任一组数据都有回归方程

(3)在回归直线方程中,b 表示　　　　　　　　　　　　　　　　　　　(　　)

A.当x增加一个单位时,y增加a的数量　　B.当y增加一个单位时，x增加b的数量

C.当x增加一个单位时，y的平均增加量　　D.当y增加一个单位时，x的平均增加量

2、为了解某社区住户的年收入和年饮食支出的关系，抽取了其中 5 户家庭的调查数据如下表

年收入 x(万元)	3	**4**	5	6	7
年饮食支出 y(万元)	1	**1.3**	1.5	2	2.2

根据表中数据用最小二乘法求得回归直线方程$\hat{y}=bx+a$中的 b=0.31,请预测年收入为 9 万元家庭的年饮食支出

解:（Ⅰ）$\overline{x}=5,\overline{y}=1.6$, b=0.31,代入 $\hat{y}=bx+a$,得 $a=0.05$

故 $\hat{y}=0.31x+0.05$,当 $x=9$ 时,解得 $\hat{y}=2.84$ 万元

所以年收入为 9 万元家庭的年饮食支出约为 2.84 万元

例 2、某种产品的广告费用支出 x（千元）与销售额 y（万元）之间有如下的对应数据：

x	2	4	5	6	8
y	3	4	6	5	7

(1)请画出上表数据的散点图;（2）请根据上表提供的数据，用最小二乘法求出销售额 y 关于支出 x 的线性回归方程;

解：（1） $\overline{x}=5,\overline{y}=5,\sum_{i=1}^{5}x_iy_i=138,\ \sum_{i=1}^{5}x_i^2=145$,

$x-\overline{x}$	-3	-1	0	1	3
$y-\overline{y}$	-2	-1	1	0	2

$$b=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{6+1+0+0+6}{9+1+0+1+9}=\frac{13}{20}=0.65,\quad \hat{a}=\frac{7}{4}=1.75$$

于是 $y=0.65x+1.75$

四、相关系数 $$r=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{x}\overline{y}}{\sqrt{(\sum_{i=1}^{n}x_i^2-n\overline{x}^2)\sum_{i=1}^{n}(y_i^2-n\overline{y}^2)}}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

1°当 $r>0$ 时，y 与 x 正相关，当 $r<0$ 时，y 与 x 负相关

2°当 $|r|\geq 0.75$ 时，y 与 x 相关性较强

当 $0.25\leq|r|<0.75$ 时，y 与 x 相关性一般

当 $0\leq|r|<0.25$ 时，y 与 x 相关性较弱

例 2 中，$r=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{j=1}^{n}(y_j-\overline{y})^2}}=\dfrac{13}{\sqrt{20\times10}}=\dfrac{13\sqrt{2}}{20}=0.9191$，$y$ 与 x 相关性较强

五、相关指数

1、在回归方程 $y=0.65x+1.75$ 中，x 叫解释变量，y 叫的预报变量。

因预报值与真 y 是有判别的，故预报变量记为 $\hat{y}$，回归方程写成 $\hat{y}=0.65x+1.75$ 。

2、线性回归模型 $y=\hat{y}+e=bx+a+e$ ，这里的 e 叫做随机误差，$E(e)=0,D(e)=\sigma^2$

例 3.回归方程 $y=0.8x+2+e$ ，$|e|\leq 0.5$ ，如果 x=10，则 y 的值不会超过（ ）.

A 10　　B 9　　C 10.5　　D 9.5

3、残差，回归偏差，总偏差

解释变量 x	2	4	5	6	8
观测值 y	3	4	6	5	7

预报值 $\hat{y}$	3.05	4.35	5	5.65	6.95
残差 $\hat{e}_i = y_i - \hat{y}_i$	−0.05	−0.35	1	−0.65	0.05
回归偏差 $\hat{y}_i - \bar{y}$	−1.95	−0.65	0	0.65	1.95
总偏差 $y_i - \bar{y}$	−2	−1	1	0	2

(1)总偏差＝残差+回归偏差

(2) 残差平方和=1.55,回归偏差平方和=8.45,总偏差平方和=10

残差平方和+回归偏差平方和=总偏差平方和

相关指数 $R^2=1-\dfrac{1.55}{10}=0.845$

当相关指数 R^2 越接近 1 时，说明 $\hat{y}$ 与 y 的拟合程度越好

(3)当 R^2＝0.845 时，我们说 y 的差异有 84.5% 是由 x 引起的

（销售额的差异有 84.5% 是由广告费用支出引起的）

注：对于两个模似 R^2 大的拟合程度更好

例 3.若有一组数据的总偏差平方和为 200，相关指数为 0.8，则残差平方和为_________

例 4、关于 x 与 y 有如下数据：

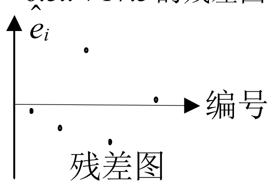
x	2	4	5	6	8
y	30	40	60	50	70

为了对 x、y 两个变量进行统计分析，现有以下两种线性模型：

$\hat{y}=6.5x+17.5$，$\hat{y}=7x+17$，

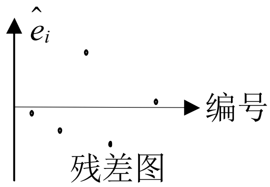
(1)试比较哪一个模型拟合的效果更好．(2)作出 $\hat{y}=6.5x+17.5$ 的残差图

解：(1) $\hat{y}=6.5x+17.5$



y	30	40	60	50	70
$\hat{y}$	30.5	43.5	50	56.5	69.5
$\hat{e}_i = y_i - \hat{y}_i$	−0.5	−3.5	10	−6.5	0.5

$\hat{y}=7x+17$



y	30	40	60	50	70
$\hat{y}$	31	45	52	59	73
$\hat{e}_i = y_i - \hat{y}_i$	−1	−5	8	−9	−3

(2) $\bar{y}=50$, 总偏差平方和 $=20^2+10^2+10^2+0^2+20^2=1000$

$\hat{y}=6.5x+17.5$ 的相关指数　$R_1^2=1-\dfrac{\text{残差平方和}}{\text{总偏差平方和}}=1-\dfrac{155}{1000}=0.845=84.5\%$，

$\hat{y}=7x+17$ 的相关指数，$R_1^2=1-\dfrac{\text{残差平方和}}{\text{总偏差平方和}}=1-\dfrac{180}{1000}=0.82=82\%$，

因为 84．5%＞82%，所以甲选用的模型拟合效果较好．

六、非线线回归转为线性回归

(1)若 $y=cx^2+d$，则令 $t=x^2$，则 $y=ct+d$，　$c=b, d=a$

(2)若 $y=ce^{dx}$，则 $\ln y=dx+\ln c$，　$d=b, \ln c=a$

注；1、回归直线的方程

设有线性相关系系的变量 x,y 的一组测量值是

设回归直线的方程 $\hat{y}=bx+a$，则可由最小二乘法确定 a, b 的值

误差平方和

$$Q(a,b)=\sum(bx_i+a-y_i)^2=\sum[b^2x_i^2+2bx_i(a-y_i)+(a-y_i)^2]$$
$$=b^2\sum x_i^2+2b\sum x_i(a-y_i)+\sum(a-y_i)^2$$
$$=b^2\sum x_i^2+2b(n\bar{x}a-\sum x_iy_i)+\sum(a-y_i)^2$$
$$Q(a,b)=\sum(a+bx_i-y_i)^2=\sum[a^2+2a(bx_i-y_i)+(bx_i-y_i)^2]$$
$$=na^2+2a(nb\bar{x}-n\bar{y})+\sum(bx_i-y_i)^2$$

要使 M 取最小值，只要

$$a=\bar{y}-\bar{x}b(1) \qquad b=\frac{\sum x_iy_i-na\bar{x}}{\sum x_i^2}\ (2)$$

(1)代入(2)得，$b\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}+n\bar{x}^2b$，于是 $b=\dfrac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}$，

求出 b 后 a 的值由(1)确定 ，由(1)得 $\bar{y}=b\bar{x}+a$，因此回归直线过**样本中心** $(\bar{x},\bar{y})$

2、如何证明用最小二乘法求出的回归方程 $\hat{y}=bx+a$ 有

残差平方和+回归偏差平方和=总偏差平方和

求证 $\sum(bx_i+a-y_i)^2+\sum(bx_i+a-\bar{y})^2=\sum(y_i-\bar{y})^2$

证明 $\because a=\bar{y}-b\bar{x}$

$$\therefore \sum(bx_i+a-y_i)^2+\sum(bx_i+a-\bar{y})^2$$
$$=\sum(bx_i-b\bar{x}+\bar{y}-y_i)^2+\sum(bx_i-b\bar{x})^2$$
$$=\sum[b(x_i-\bar{x})+(\bar{y}-y_i)]^2+\sum(bx_i-b\bar{x})^2$$
$$=\sum(\bar{y}-y_i)^2+2b^2\sum(x_i-\bar{x})^2+2b\sum(x_i-\bar{x})(\bar{y}-y_i)$$
$$=\sum(\bar{y}-y_i)^2+2b^2\sum(x_i-\bar{x})^2-2b^2\sum(x_i-\bar{x})^2=\sum(\bar{y}-y_i)^2$$

第二节、相关检验

为了解某班学生课外活动爱好运动是否与性别有关，对本班 50 人调查得到了下表：

	好运动	不好运动	合计
男生	20	5	25
女生	10	15	25
合计	30	20	50

1、分类变量：性别的值：男生和女生。爱好的值：好运动，不好运动

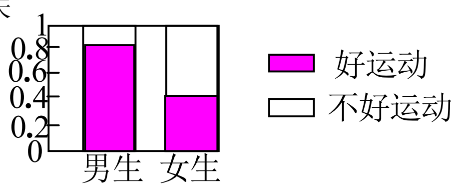
于是性别与喜好就可看成是变量，这种变量叫做分类变量，上面的表格叫做 2×2 的列联表

2、等高图

男生爱好运动的概率为 $\dfrac{20}{25}=\dfrac{4}{5}$，女生爱好运动的概率为 $\dfrac{10}{25}=\dfrac{2}{5}$

于是有理由认为男生更喜爱运动，说明喜爱运动与性别有关

可作出等高图



3、喜爱运动与性别有关的概率有多大呢？

	好运动	不好运动	合计
男生	a	b	a +b
女生	c	d	c+d
合计	a+c	b+d	n=a+b+c+d

假设性别与爱好运动没有关系，A={男生},B={好动动}，则 A 与 B 独立

$P(AB)=P(A)P(B)$，$\frac{a}{n}=\frac{a+b}{n}\times\frac{a+c}{n},a=\frac{(a+b)(a+c)}{n}$右边叫 a 的理论人数

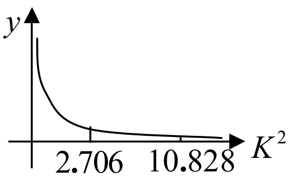
$$\frac{[a-\frac{(a+b)(a+c)}{n}]^2}{\frac{(a+b)(a+c)}{n}}=\frac{(ad-bc)^2}{n(a+b)(a+c)}$$

引入随机变量

$$K^2=\frac{(ad-bc)^2}{n(a+b)(a+c)}+\frac{(ad-bc)^2}{n(a+b)(b+d)}+\frac{(ad-bc)^2}{n(c+d)(a+c)}+\frac{(ad-bc)^2}{n(c+d)(b+d)}$$

$$=\frac{(ad-bc)^2}{n}[\frac{1}{n(a+b)(a+c)}+\frac{1}{n(a+b)(b+d)}+\frac{1}{n(c+d)(a+c)}+\frac{1}{n(c+d)(b+d)}]$$

$$=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$



当 K^2 的值越大就说明性别与爱好运动越有关系

随机变量 K^2 服从卡方分布密度曲线如图

4、当分类变量无关时

$P(K^2\ge k_0)$	**0.10**	**0. 05**	0. 025	**0.010**	0. 005	0.001
k_0	**2.706**	**3.841**	5.024	**6.635**	7.879	10.828

本例 $K^2=\frac{50(20\times15-10\times5)^2}{30\times20\times25\times25}=\frac{25}{3}\approx8.33$

假设性别与爱好运动没有关系，则 $P(K^2\geqslant7.879)=0.005$

所以做出性别与爱好运动没有关系的的判断为正确的概率不超过 0. 5%，

于是应作出性别与爱好运动有关系的判断更为合理，此判断正确的概率超过 99. 5%，

犯错误的概率不超过 0. 5%，也可说：有 99. 5%的把握认为性别与爱好运动与有关系。

注:几种回答方法

有 99. 5%的把握认为性别与爱好运动与有关系。

没有 99. 9%的把握认为性别与爱好运动与有关系。

犯错误的概率不超过 0. 5%的前提下，认为性别与爱好运动与有关系。

犯错误的概率不超过 0. 1%的前提下，没有足够的理由性别与爱好运动有关系。

例 1、某地区一种传染病与饮用水的调查，饮用干净水得病 35 人，不得病 65 人

饮用不干净水得病 70 人，不得病 30 人

(1) 作出 2×2 列联表 (2) 能否有的把握认为饮用水与得病有关？

	得病	不得病	合计
干净水	35	65	100
方法 B	70	30	100
合计	105	95	$n=200$

解 $K^2=\frac{200\times(70\times65-35\times30)^2}{100\times100\times105\times95}\approx24.56$

由于 $K^2>10.828$，所以有 99.9%的把握认为饮用水与得病有关

例 2、某厂为了检查有甲乙两条包装流水线的质量，分别从每条生产线上取 40 件产品作为样本，得下面的 2×2 列联表如下

问多大的把握认为

“产品的包装质量与两条包装流水线的选择有关”

	合格	不合格	总数
甲	30	10	40
乙	36	4	40
总数	66	14	80

解：$\because K^2=\dfrac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$

$=\dfrac{80\times(120-360)^2}{66\times 14\times 40\times 40}\approx 3.117>2.706$

∴有 90％的把握认为产品的包装质量与两条包装流水线的选择有关

例 3、对某班 50 名学生兴趣数学与兴趣地理调查，得 2×2 的列联表

	兴趣地理	不兴趣地理	总数
兴趣数学	18	9	27
不兴趣数学	8	15	23
总数	26	24	50

问是否有 97.5%的把握认为兴趣地理与兴趣数学有关系？

解：$K^2=\dfrac{50\times(18\times 15-8\times 9)^2}{27\times 23\times 26\times 24}=5.058$

假设趣地理与兴趣数学没有关系，则 $P(K^2\geqslant 5.024)=0.025$

于是应作出兴趣地理与兴趣数学有关系的判定，此判定正确的概率超过 97.5%;

故有 97.5%的把握认为兴趣地理与兴趣数学有关系。

注：当 $K^2\geq 2.706$ 应做出分类变量有关的判定。否则高考不做要求

例 3、对某班 50 名学生兴趣数学与兴趣地理调查，得 2×2 的列联表

	兴趣地理	不兴趣地理	总数
兴趣数学	18	9	27
不兴趣数学	8	15	23
总数	26	24	50

问是否有 97.5%的把握认为兴趣地理与兴趣数学有关系？

解：$K^2=\dfrac{50\times(18\times 15-8\times 9)^2}{27\times 23\times 26\times 24}=5.058$

假设趣地理与兴趣数学没有关系，则 $P(K^2\geqslant 5.024)=0.025$

于是应作出兴趣地理与兴趣数学有关系的判定，此判定正确的概率超过 97.5%;

故有 97.5%的把握认为兴趣地理与兴趣数学有关系。

6、卡方变量：

$$K^2=\frac{[a-\frac{(a+b)(a+c)}{n}]^2}{\frac{(a+b)(a+c)}{n}}+\frac{[b-\frac{(a+b)(b+d)}{n}]^2}{\frac{(a+b)(b+d)}{n}}+\frac{[c-\frac{(c+d)(a+c)}{n}]^2}{\frac{(c+d)(a+c)}{n}}+\frac{[d-\frac{(c+d)(b+d)}{n}]^2}{\frac{(c+d)(b+d)}{n}}$$

$$=\frac{(ad-bc)^2}{n(a+b)(a+c)}+\frac{(ad-bc)^2}{n(a+b)(b+d)}+\frac{(ad-bc)^2}{n(c+d)(a+c)}+\frac{(ad-bc)^2}{n(c+d)(b+d)}$$

$$=\frac{(ad-bc)^2}{n}[\frac{1}{n(a+b)(a+c)}+\frac{1}{n(a+b)(b+d)}+\frac{1}{n(c+d)(a+c)}+\frac{1}{n(c+d)(b+d)}]$$

$$=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

对于 m×n 列联表 $K^2=\dfrac{[a-\frac{(a+b)(a+c)}{n}]^2}{\frac{(a+b)(a+c)}{n}}+\dfrac{[b-\frac{(a+b)(b+d)}{n}]^2}{\frac{(a+b)(b+d)}{n}}+\cdots$ 但没有 2×2 列联表的化简表达式。应查自由度为(m-1)×(n-1)的卡方表

第二章 参数点估计

第一节 矩估计法

一、统计量

从总体 X 中取出一个样本 X_1、X_2、…、X_n.则 X_i 是与 X 同分布的随机变量，X_1、X_2、…、X_n 的一组观测值是 x_1、x_2、…、x_n 也可称为 X 的 n 个独立观测值。

常用的统计量有

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

（3）样本标准差 $S=\sqrt{S^2}$ (4)样本 k 阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^{\ k},k=1,2,3,\cdots$

（5）样本 k 阶中心矩 $B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k,k=1,2,3,\cdots$

它们的观测值分别是

（1）样本平均值 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ （2）样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n-1}(\sum_{i=1}^{n}x_i^{\ 2}-n\overline{x})$

（3）样本标准差 $S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$ (4)样本 k 阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}x_i^{\ k},k=1,2,3,\cdots$

（5）样本 k 阶中心矩 $B_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k,k=1,2,3,\cdots$

若总体 X 的 k 阶矩存在即 $E(X^k)=\mu_k$，则 $E(X_1^{\ k})=E(X_2^{\ k})=\cdots=E(X_n^{\ k})=\mu_k$

由辛钦定理知 $A_k=\frac{1}{n}\sum_{i=1}^{n}x_i^{\ k}\xrightarrow{P}\mu_K$

二、矩法估计

替换原理常指如下两句话：

1.用样本矩去替换总体矩，这里的矩可以是原点矩也可以是中心矩；

2.用样本矩的函数去替换相应的总体矩的函数。

根据这个替换原理，在总体分布形式未知场合也可以对各种参数做出估计，譬如：

1.用样本均值 $\overline{X}$ 估计总体均值 $E(X)$，即 $\hat{E}(X)=\overline{X}$；

2.用样本方差 S_n^2 估计总体方差 $D(X)$，即 $\hat{D}(X)=S_n^2$；

3.用事件 A 出现的频率估计事件 A 发生的概率；

4.用样本的 p 分位数估计总体的 p 分位数，特别，用样本中位数估计总体中位数。

矩法估计的统计思想（替换原理）十分简单明确，众人都能接受，使用场合甚广，它的

实质是用经验分布函数去替换总体分布，其理论基础是格里纹科定理。

例 1、正态总体的分布是 $N(\mu,\sigma^2)$。求 μ,σ^2 的矩估计。

解：由 $E(X)=\mu, D(X)=\sigma^2$ 可得 $\hat{\mu}=\overline{X}, \hat{\sigma}^2=S_n^{*2}$

例 2、在泊松分布 $P(\lambda)$ 的总体中，求 λ 的矩估计

解：由 $E(X)=\lambda$ 可得 $\hat{\lambda}=\overline{X}$

例 3、在二项分布 $b(n,p)$ 的总体中，n 是已知的，求 p 的估计量。

由 $E(X)=np$，有 $\overline{X}=n\hat{p}$，所以 $\hat{p}=\dfrac{\overline{X}}{n}$

例 4、设总体 X 具有 Γ 分布，其密度为

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

其中 $\alpha>0,\beta>0$，试求 α,β 的矩估计。

解：这里 $k=2,$ 计算数学期望和方差可得 $E(X)=\dfrac{\alpha}{\beta}, D(X)=\dfrac{\alpha}{\beta^2}$

因而 $\overline{X}=\dfrac{\hat{\alpha}}{\hat{\beta}}, S_n^{*2}=\dfrac{\hat{\alpha}}{\hat{\beta}^2}$ 解方程得 $\hat{\alpha}=\dfrac{\overline{X}^2}{S_n^{*2}}, \hat{\beta}=\dfrac{\overline{X}}{S_n^{*2}}$

例 5、设总体为指数分布，其密度函数为

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

$X_1,\dots,X_n$ 是样本，此处 $k=1,$ 由于 $E(X)=\dfrac{1}{\lambda}$，即 $\lambda=\dfrac{1}{E(X)}$，故 λ 的矩法估计为 $\hat{\lambda}=\dfrac{1}{\overline{X}}$

另外，由于 $D(X)=\dfrac{1}{\lambda^2}$，其反函数为 $\lambda=\dfrac{1}{\sqrt{D(X)}}$，因此，从替换原理来看，$\lambda$ 的矩法估计也可取为 $\hat{\lambda}_1=\dfrac{1}{S_n}$

这说明矩估计可能是不唯一的，这是矩法估计的一个缺点，此时通常应该尽量采用低阶矩给出未知参数的估计。

例 6、$X_1,\dots,X_n$ 是来自 (a,b) 上的均匀分布 $U(a,b)$ 的样本，a,b 均是未知参数，这里 $k=2,$ 由于

$$E(X)=\frac{a+b}{2},D(X)=\frac{(b-a)^2}{12}$$

不难推出

$$a=E(X)-\sqrt{3D(X)},b=E(X)+\sqrt{3D(X)}$$

由此可得a,b的矩估计

$$\hat{a}=\overline{X}-\sqrt{3}S_n,b=\overline{X}+\sqrt{3}S_n$$

若从均匀总体$U(a,b)$获得如下一个容量为5的样本：4.5　5.0　4.7　4.0　4.2，经计算，有$\overline{X}=4.48,S_n^*=0.3962$，于是可得$a,b$的矩估计为

$$\hat{a}=4.48-0.3962\sqrt{3}=3.7938,$$
$$\hat{b}=4.48+0.3962\sqrt{3}=5.1662.$$

使用矩法估计的一个前提是总体存在适当阶的矩，阶数应不小于待估参数的个数（或者是参数空间的维数），但这不总是可以做到的。

例 9.1.7 柯西分布（Cauchy）设总体具有密度函数

$$f(x;\theta)=\frac{1}{\pi(1+(x-\theta)^2)},-\infty<x<+\infty$$

显然它的各阶矩皆不存在，因此不能用矩法估计来估计参数θ，另外尽管矩法估计简便易行，且只要n充分大，估计的精度也很高，但它只用到总体的数字特征的形式，而未用到总体的具体分布形式，损失了一部分很用的信息，因此在很多场合下显的粗糙和过于一般。

第二节　最大似然估计

一、极大似然估计法的意思

最大似然估计是求估计用的最多的方法，它最早是由高斯在 1821 年提出，但一般将之功归功于费希尔（R.A.Fisher），因为费希尔在 1922 年再次提出了这种想法并证明了它的一些性质而使最大似然法得到了广泛的应用。

先通过一个实例介绍最大似然估计。

例 1、设有一大批产品，其废品率为$p(0<p<1)$。今从中随意地取出100个，其中有10个废品，试估计p的数值。

若正品用“0”表示，废品用“1”表示。此总体X的分布为

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

取得的子样记为$X_1,\dots,X_n$，其中10个是“1”，90个是“0”。出现此子样的概率为

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}=P\{X_1=x_1\}P\{X_2=x_2\}\cdots P\{X_n=x_n\}$$
$$=p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n}=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}=p^{10}(1-p)^{90}$$

这个概率随 p 的数值不同而不同。自然选择使此概率达到最大的 p 值作为真正废品率的估计值。记 $L(p)=p^{10}(1-p)^{90}$。用高等数学中求极值的方法，由

$$L'(p)=10p^{9}(1-p)^{90}-90p^{10}(1-p)^{89}=p^{9}(1-p)^{89}[10(1-p)-90p]=0$$

得 $\hat{p}=\frac{10}{100}$

此例求解的思想方法是：选择参数的值使抽得的子样值出现的可能性最大，用这个值作为未知参数的估计值。这种求估计量的方法称为**最大似然估计法**，也称为**最大或然估计法**或者**极大似然估计法**。显然，如果在此例中取一个容量为 n 的子样，其中有 m 个废品，用最大似然估计法可得 $\hat{p}=\frac{m}{n}$。

下面就离散总体分布和连续总体分布两种情形分别介绍最大似然估计法。

二、离散总体分布情形

设总体 X 的分布列为 $P\{X=x_i\},i=1,2,\ldots$ 或 $P\{X=x\}=P(x;\theta_1,\ldots,\theta_k),x=x_1,x_2,\ldots$

其中 $\theta_1,\cdots,\theta_k$ 是未知参数，如果取得子样值 $x_1,\ldots,x_n$，那么出现此子样值的概率为

$$L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k)=P\{X_1=x_i,X_2=x_2,\ldots,X_n=x_n\}$$
$$=P\{X_1=x_1\}P\{X_2=x_2\}\cdots P\{X_n=x_n\}=\prod_{i=1}^{n}P(x_i)$$

选择 $\theta_1,\ldots,\theta_k$ 使 $L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k)$ 达到最大，即 $L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k)=\max$

这样获得的 $\theta_1,\cdots,\theta_k$ 值作为相应未知参数的估计值。这种求估计值的方法称为**最大似然估计法**。简记为 **MLE**（Maximum Likelihood Estimate）。求得的未知参数的估计量 $\hat{\theta}_1,\cdots,\hat{\theta}_k$ 称为**最大似然估计量**。L 称为**似然函数**。

如果 L 对 $\theta_1,\ldots,\theta_k$ 的偏导数存在，那么可以采用高等数学中求极值的方法计算估计值，只要从似然方程组 $\frac{\partial L}{\partial\theta_i}=0,i=1,2,\ldots,k$

解出 $\theta_i=\theta_i(x_1,\ldots,x_n)$，并 θ_i 将换成 $\hat{\theta}_i$ 即可。

需要指出，有时利用对数函数是单调增函数，选择 $\theta_1,\ldots,\theta_k$，使 $\ln L=\max$ 较为方便。通常 $\ln L$ 亦称为**对数似然函数**。易知 L 与 $\ln L$ 在同一处 $\hat{\theta}$ 达到极大，因此这样做不会改变极大点。

例 2、设总体 X 具有泊松分布 $P(\lambda)$，其分布列为 $\mathbb{P}\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda},k=0,1,2,\ldots$

其中$\lambda > 0$。试用最大似然估计法求未知参数λ。

解：作似然函数$L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\cdots\frac{\lambda^{x_2}}{x_n!}e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n}x_i}}{x_1!x_2!\cdots x_n!}e^{-n\lambda}$

取对数得$\ln L = \sum_{i=1}^{n}x_i\ln\lambda - n\lambda - \ln(x_1!x_2!\cdots x_n!)$

由$\frac{dLnL}{d\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n}x_i - n = 0$解出$\lambda = \frac{1}{n}\sum_{i=1}^{n}X_i = \overline{X}$

改写为$\hat{\lambda} = \overline{X}$，这里求得的$\lambda$的估计量与用矩法估计求得的结果是相同的。

三、连续总体分布情形 设总体X的分布密度是$f(x;\theta_1,\ldots,\theta_k)$，其中$\theta_1,\ldots,\theta_k$是未知参数，取得子样值为$x_1,\ldots,x_n$。我们知道当总体是连续型随机变量时，谈所谓样本值$x_1,\ldots,x_n$出现的概率是没有什么意义的，因为任何一个具体样本出现的概率都是零概率事件。这时我们考虑样本在它任意小的邻域中出现的概率，这个概率越大，就等价于此样本处的概率密度越大。因此，考虑概率

$$\begin{aligned}
&P\{x_1 - dx_1 < X_1 \le x_1, x_2 - dx_2 < X_2 \le x_2,\ldots,x_n - dx_n < X_n \le x_n\} \\
&= P\{x_1 - dx_1 < X_1 \le x_1\}\cdots P\{x_n - dx_n < X_n \le x_n\} \\
&\approx \prod_{i=1}^{n}[f(x_i;\theta_1,\ldots,\theta_k)dx_i] = \prod_{i=1}^{n}[f(x_i;\theta_1,\ldots,\theta_k)]dx_1\cdots dx_n
\end{aligned}$$

这里取的小区间$dx_1,\ldots,dx_n$长度都是固定的量。选择$\theta_1,\ldots,\theta_k$的值使此概率达到最大，亦即使$\prod_{i=1}^{n}f(x_i;\theta_1,\ldots,\theta_k)$达到最大。

令$L(x_1,\ldots,x_n;\theta_1,\ldots,\theta_k) = \prod_{i=1}^{n}f(x_i;\theta_1,\ldots,\theta_k)$。选择$\theta_1,\ldots,\theta_k$的值使$L$达到最大，即

$$L = \max$$

这样得到的值作为相应未知参数的估计值。这种方法称为**最大似然估计法**。求得的估计量亦称为**最大似然估计量**。L称为**似然函数**。

类似于离散情形如果L对$\theta_1,\cdots,\theta_k$的偏导数存在，那么只要解似然方程组

$$\frac{\partial L}{\partial\theta_i} = 0, i = 1,2,\ldots,k$$

可得最大似然估计量。

需要指出，有时选择$\theta_1,\ldots,\theta_k$的值使$\ln L = \max$较为方便，此时$\ln L$亦称为**对数似然**

函数。

例 3、设总体 X 服从负指数分布，其密度为

$$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x}, x\ge 0\\ 0, \qquad x<0\end{cases}$$

其中未知参数 $\lambda>0$ 。试求 λ 的最大似然估计。

解：由总体分布可见总体数量指标 X 非负，因而子样值 $x_1,\ldots,x_n$ 中的每一样品 x_i 非负。

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

取对数 $\ln L=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i$ 　由 $\dfrac{d\ln L}{d\lambda}=\dfrac{\lambda}{n}-\sum_{i=1}^{n}x_i=0$ 解出 $\lambda=\dfrac{1}{\overline{X}}$ 改写为 $\hat{\lambda}=\dfrac{1}{\overline{X}}$ 。

例 4、对正态总体 $N(\mu,\sigma^2),\theta=(\mu,\sigma^2)$ 是二维参数，设有子样值 $x_1,\ldots,x_n$ ，则似然函数及其对数分别为

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

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

将 $\ln L(\mu,\sigma^2)$ 分别关于两个分量求偏导并令其为零即得到似然方程组

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

解此方程组，可得 μ 的最大似然估计为

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X},$$

因此也可得到 σ^2 的最大似然估计为

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

利用二阶导函数矩阵的非正定性可以说明上述估计使得似然函数取极大值。

虽然求导函数是求最大似然估计最常用的方法，但并不是在所有场合求导都是有效的，下面的例子说明了这个问题。

例 4、设总体 X 具有均匀分布，其密度为

$$f(x;\theta)=\begin{cases}\dfrac{1}{\theta},0\le x\le\theta,\\0,\qquad 其他\end{cases}$$

其中未知参数$\theta>0$，试求θ的最大似然估计量。

解：子样值为$x_1,\ldots,x_n$，而

$$f(x_i;\theta)=\begin{cases}\dfrac{1}{\theta},0\le x_i\le\theta,\\0,\qquad 其他\end{cases}$$

似然函数

$$L=\begin{cases}\dfrac{1}{\theta^n},0\le\min\limits_{1\le i\le n}x_i\le\max\limits_{1\le i\le n}x_i\le\theta,\\0,\qquad\quad 其他\end{cases}$$

选取θ的值使L达到最大，只要取

$$\theta=\max_{1\le i\le n}X_i$$

改写成$\hat{\theta}=\max\limits_{1\le i\le n}X_i$。

和矩估计的情形一样，有时虽然能给出似然方程，也可以证明它有解，但得不到解的解析表达式。

例 5、求柯西分布中θ的最大似然估计，我们可以得到似然方程为

$$\frac{\partial\ln L(\theta)}{\partial\theta}=-\sum_{i=1}^{n}\frac{2(X_i-\theta)}{1+(X_i-\theta)}=0$$

这个方程只能求数值解。

最大似然估计有一个简单而有用的性质:如果$\hat{\theta}$是θ的最大似然估计，则对任一函数$g(\theta)$，其最大似然估计为$g(\hat{\theta})$，该性质称为最大似然估计的不变性，从而使一些复杂结构的参数的最大似然估计的获得变得容易了。

例 6、设$X_1,\ldots,X_n$是来自正态总体$N(\mu,\sigma^2)$的样本，在例 9.1.11 中已经求得μ和σ^2的最大似然估计为

$$\hat{\mu}=\bar{X},\hat{\sigma}^2=S_n^2$$

于是由最大似然估计的不变性可得如下参数的最大似然估计，它们是

标准差σ的最大似然估计是$\hat{\sigma}=S_n$;

概率$\mathbb{P}(X<3)=\Phi(\dfrac{3-\mu}{\sigma})$的最大似然估计是$\Phi(\dfrac{3-\bar{X}}{S_n})$;

第三节　估计量的评价标准

我们已经看到点估计有各种不同的求法，为了在不同的点估计之间进行比较选择，就必须对各种不同的点估计的好坏给出评价标准。

数理统计中给出了众多的估计量的评价标准，对同一估计量使用不同的评价标准可能会得到完全不同的结论，因此，在评价某一个估计好坏时首先要说明是在那一个标准下，否则所论好坏则毫无意义。

但不管怎么说，有一个基本标准是所有的估计都应该满足的，它是衡量一个估计是否可行的必要条件，这就是估计的相合性，我们就从相合性开始。

一、相合性

我们知道，点估计是一个统计量，因此它是一个随机变量，在样本量一定的条件下，我们不可能要求它完全等同于参数的真实取值。但如果我们有足够的观测值，根据格里纹科定理，随着样本量的不断增大，经验分布函数逼近真实分布函数，因此完全可以要求估计量随着样本量的不断增大而逼近参数真值，这就是相合性，严格定义如下。

定义　设$\theta\in\Theta$为未知参数，$\hat{\theta}_n=\hat{\theta}_n(X_1,\ldots,X_n)$是$\theta$的一个估计量，$n$是样本容量，若对任何一个$\varepsilon>0$，有

$$\lim_{n\to+\infty}\mathbb{P}(\left|\hat{\theta}_n-\theta\right|>\varepsilon)=0 \text{ 或 } \lim_{n\to+\infty}\mathbb{P}(\left|\hat{\theta}_n-\theta\right|<\varepsilon)=1$$

则称$\hat{\theta}_n$为参数θ的**相合估计**（**量**）或**一致估计**（**量**）。

相合性被认为是对估计的一个最基本的要求，如果一个估计量，在样本量不断增大时，它都不能把被估参数估计到任意指定的精度，那么这个估计是很值得怀疑的。通常，不满足相合性要求的估计一般不予考虑，证明估计的相合性一般可应用大数定律或直接由定义来证。

若把依赖于样本量n的估计量$\hat{\theta}_n$看作一个随机变量序列，相合性就是$\hat{\theta}_n$依概率收敛于θ，所以证明估计的相合性可应用依概率收敛的性质及各种大数定律。

例 1、设$X_1,X_2,\ldots$是来自正态总体$N(\mu,\sigma^2)$的样本，则由辛钦大数定律及依概率收敛的性质知：$\bar{X}$是μ的相合估计；S^{*2}是σ^2的相合估计；S^2也是σ^2的相合估计。由此可见参数的相合估计不止一个。

在判断估计的相合性时，下述两个定理是很有用的。

定理 1 设$\hat{\theta}_n$是θ的一个估计量，若

$$\lim_{n\to+\infty}E(\hat{\theta}_n)=\theta,\ \lim_{n\to+\infty}D(\hat{\theta}_n)=0,$$

则$\hat{\theta}_n$是θ的相合估计。

证明略。

定理 2 若$\hat{\theta}_{n1},\ldots,\hat{\theta}_{nk}$分别是$\theta_1,\ldots,\theta_k$的相合估计，$\eta=g(\theta_1,\ldots,\theta_k)$是$\theta_1,\ldots,\theta_k$的连续函数，则$\hat{\eta}_n=g(\hat{\theta}_{n1},\ldots,\hat{\theta}_{nk})$是$\eta$的相合估计。

证明略。

由大数定律及定理，我们可以看到，矩估计一般都具有相合性，比如：样本均值是总体均值的相合估计；样本标准差是总体标准差的相合估计。

例 2、设一个试验有三种可能结果，其发生概率分别为

$$p_1=\theta^2, p_2=2\theta(1-\theta), p_3=(1-\theta)^2$$

现做了n次试验，观测到三种结果发生的次数分别为n_1,n_2,n_3，可以采用频率替换方法估计θ，由于可以有三个不同的θ的表达式：

$$\theta=\sqrt{p_1},\theta=1-\sqrt{p_3},\theta=p_1+\frac{p_2}{2}$$

从而可以给出θ三种不同的频率替换估计，它们分别是：

$$\hat{\theta}_1=\sqrt{\frac{n_1}{n}},\hat{\theta}_2=1-\sqrt{\frac{n_3}{n}},\hat{\theta}_3=(n_1+\frac{n_2}{2})\Big/n$$

由大数定律，$\frac{n_1}{n},\frac{n_2}{n},\frac{n_3}{n}$分别是$p_1,p_2,p_3$的相合估计，由定理 2 知，上述三个估计都是$\theta$的相合估计。

二、无偏性

相合性是大样本下估计量的评价标准，对小样本而言，需要一些其他的评价标准，无偏性便是一个常见的评价标准。

定义 1：设$\hat{\theta}=\hat{\theta}(X_1,\ldots,X_n)$是$\theta$的一个估计，$\theta$的参数空间为$\Theta$，若对任意的$\theta\in\Theta$，有

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

则称$\hat{\theta}$是θ的**无偏估计**，否则称为**有偏估计**。

无偏性的要求可以改写为$E(\hat{\theta}-\theta)=0$，这表示无偏估计没有系统偏差。当我们使用$\hat{\theta}$估计θ时，由于样本的随机性，$\hat{\theta}$与θ总是有偏差的，这种偏差时而（对某些样本观测值）为正，时而（对另一些样本观测值）为负，时而大，时而小。无偏性表示，把这些偏差平均起来其值为0，这就是无偏估计的含义。而若估计不具有无偏性，则无论使用多少次，其平均也会与参数真值有一定的距离，这个距离就是系统误差。

定义 2、如果$E(\hat{\theta})\neq\theta$，那么$E(\hat{\theta})-\theta$称为**估计量$\hat{\theta}$的偏差**。若$\lim\limits_{n\to+\infty}E(\hat{\theta})=\theta$则称$\hat{\theta}$是$\theta$的**渐近无偏估计**。

无偏性不具有不变性，即若$\hat{\theta}$是θ的无偏估计，一般而言，$g(\hat{\theta})$不是$g(\theta)$的无偏估计，除非$g(\theta)$是θ的线性函数。

例 1、设总体X的一阶和二阶矩存在，分布是任意的。记$E(X)=\mu, D(X)=\sigma^2$。用矩法可得子样平均$\bar{X}$和子样方差S^2分别是μ和σ^2的估计量。那么$\bar{X}$和S^2是否分别是μ和σ^2的无偏估计呢？因为

$$E(\bar{X})=E(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}n\mu=\mu$$

故$\bar{X}$是μ的无偏估计。又

$$\begin{aligned}
E(S^2)&=E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2]\\
&=\frac{1}{n}E\{\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{X}-\mu)^2\}\\
&=\frac{1}{n}E\{\sum_{i=1}^{n}(X_i-\mu)^2-2\sum_{i=1}^{n}(X_i-\mu)(\bar{X}-\mu)+n(\bar{X}-\mu)^2\}\\
&=\frac{1}{n}E\{\sum_{i=1}^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2\}=\frac{1}{n}\{\sum_{i=1}^{n}D(X_i)-nD(\bar{X})\}\\
&=\frac{1}{n}\{n\sigma^2-\sigma^2\}=\frac{n-1}{n}\sigma^2
\end{aligned}$$

所以S^2不是σ^2的无偏估计量，而是σ^2的渐近无偏估计量。但由于

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

所以S^{*2}是σ^2的无偏估计量。当n很大时，$S^{*2}\approx S^2$。

三、有效性

参数的无偏估计可以很多，如何在无偏估计中进行选择？直观的想法是希望该估计围绕参数真值的波动越小越好，波动大小可以用方差来衡量，因此人们常用无偏估计的方差的大小作为度量无偏估计优劣的标准，这就是有效性。

定义 1、设$\hat{\theta}_1,\hat{\theta}_2$是$\theta$的两个无偏估计，如果对任意$\theta\in\Theta$有

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

且至少有一个$\theta\in\Theta$使得上述不等号严格成立，则称$\hat{\theta}_1$比$\hat{\theta}_2$有效。

例 1、设$X_1,\ldots,X_n$是取自某总体的样本，记总体均值为μ，总体方差为σ^2，则$\hat{\mu}_1=X_1, \hat{\mu}_2=\bar{X}$都是$\mu$的无偏估计，但

$$D(\hat{\mu}_1)=\sigma^2, D(\hat{\mu}_2)=\frac{\sigma^2}{n}.$$

显然，只要 $n>1, \hat{\mu}_2$ 比 $\hat{\mu}_1$ 有效。这表明，用全部数据的平均来估计总体均值要比使用部分数据更为有效。

例 2、我们已经计算出均匀总体 $U[0,\theta]$ 中 θ 的最大似然估计是 $\max\limits_{1\le i\le n} X_i$，由于 $E(\max\limits_{1\le i\le n} X_i)=\frac{n}{n+1}\theta$，所以 $\max\limits_{1\le i\le n} X_i$ 不是 θ 的无偏估计，但是 θ 的渐近无偏估计，经过修编后可以得到 θ 的一个无偏估计：$\hat{\theta}_1=\frac{n+1}{n}\max\limits_{1\le i\le n} X_i$。且

$$\begin{aligned} D(\hat{\theta}_1) &= (\frac{n+1}{n})^2 D(\max_{1\le i\le n} X_i) \\ &= (\frac{n+1}{n})^2 \frac{n}{(n+1)^2(n+2)}\theta^2=\frac{\theta^2}{n(n+2)} \end{aligned}$$

另一方面，由矩法，我们可以得到 θ 的另一个无偏估计 $\hat{\theta}_2=2\bar{X}$，且

$$D(\hat{\theta}_2)=4D(\bar{X})=\frac{4}{n}D(X)=\frac{\theta^2}{3n}$$

由此，当 $n>1$ 时，$\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效。

第三章、正态总体的区间估计与区间检验

第一节 可用正态分布解决的问题

一、求置信区间

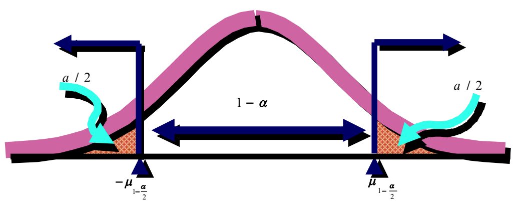
1、单态总体方差已知的 μ 区间估计

例 1. 某厂生产的化纤强度服从正态分布,长期以来其标准差稳定在 σ =0.85,现抽取了一个容量为 n=25 的样本,测定其强度,算得样本均值为 $\bar{x}$ =2.25,试求这批化纤平均强度的置信水平为 0.95 的置信区间.

解:设该厂生产的化纤强度X服从 $N(\mu,\sigma^2)$,设样本容量为25的样本为 $X_1,X_2,\cdots,X_{25}$,

则 $E(\overline{X})=E(\frac{X_1+X_2+\cdots+X_{25}}{25})=\frac{25}{25}E(X)=\mu$,

于是 μ 的点估计值 $\hat{\mu}=\bar{x}$ =2.25,



$$D(\overline{X})=D(\frac{X_1+X_2+\cdots+X_{25}}{25})=\frac{1}{25^2}\sum_{i=1}^{25}D(X_i)=\frac{25\sigma^2}{25^2}=\frac{\sigma^2}{25}$$

因此$\overline{X} \sim N(\mu,\frac{\sigma^2}{25})$,$\frac{\overline{X}-\mu}{\frac{\sigma}{5}} \sim N(0,1)$,要使

$$P(-u_{1-\frac{\alpha}{2}} \le \frac{\overline{X}-\mu}{\frac{\sigma}{5}} \le u_{1-\frac{\alpha}{2}}) = 0.95 = 1-2\times 0.025 = 1-2\times(1-0.975)$$

这里$1-\alpha = 0.95$,$\alpha = 0.05$,查表知$u_{0.975} = 1.96$,于是

$$-u_{1-\frac{\alpha}{2}} \le \frac{\overline{X}-\mu}{\frac{\sigma}{5}} \le u_{1-\frac{\alpha}{2}} \text{即} \overline{X}-\frac{\sigma}{5}u_{1-\frac{\alpha}{2}} \le \mu \le \overline{X}+\frac{\sigma}{5}u_{1-\frac{\alpha}{2}},$$

$$\overline{X}-\frac{\sigma}{5}u_{1-\frac{\alpha}{2}} = 2.25-1.96\times 0.85/5 = 2.25-0.3332=1.9168$$
$$\overline{X}-\frac{\sigma}{5}u_{1+\frac{\alpha}{2}} = 2.25+1.96\times 0.85/5 = 2.25+0.3332=2.5832$$

即这批化纤平均强度的置信水平为 0.95 的置信区间为[1.9168,2.5832].

一般的，设总体$X \sim N(\mu,\sigma^2)$ ，σ已知，从总体中抽出一个样本为$X_1, X_2, \cdots, X_n$ ，

则样本均值$\overline{X} \sim N(\mu,\frac{\sigma^2}{n})$ ，$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ ，则μ的置信水平为$1-\alpha$的置信区间是

$[\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}, \overline{X}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}]$ ，注服从标准正态分布的随机变量$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$叫枢轴量

例 2、用天平称量某物体的质量9次，得平均值为$\overline{X} = 15.4(g)$ ，已知天平称量结果为正态分布，其标准差为$0.1g$ ，试求该物体质量的0.95置信区间。

解：此处$1-\alpha = 0.95, \alpha = 0.05$ ，查表知$u_{0.975} = 1.96$ ，于是该物体质量μ的置信区间为

$$\overline{X} \pm u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}. = 15.4 \pm 1.96\times\frac{0.1}{\sqrt{9}} = 15.4 \pm 0.0653$$

从而得到该物体质量0.95的置信区间为[15.3347,15.4653]。

例 3 设总体为正态分布$N(\mu,1)$ ，为得到μ的置信水平为0.95的置信区间长度不超过

1.2，样本容量应为多大？

解：由题设条件知μ的0.95置信区间为

$$[\bar{X}-u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{X}+u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

其区间长度为$2u_{1-\alpha/2}\big/\sqrt{n}$，它依赖于样本容量$n$而与样本具体取值无关，现要求$2u_{1-\alpha/2}\big/\sqrt{n}\le 1.2$，立即有$n\ge(\frac{2}{1.2})^2u^2_{1-\frac{\alpha}{2}}$，现$1-\alpha=0.95,$ 故$u_{1-\frac{\alpha}{2}}=1.96$，从而$n\ge(\frac{3}{5})^2\times1.96^2=10.67\approx 11$。即样本容量至少为11时才能使得$\mu$的置信水平为0.95的置信区间的长度不超过1.2.

2、大样本置信区间

在样本容量充分大时，可以用渐近分布来构造近似的置信区间。

设总体X的分布是任意的，平均数$\mu=E(X),\sigma^2=D(X)$和方差都是未知的。用子样$X_1,\ldots,X_n$对总体平均数μ作区间估计。

自然可以用$\bar{X}$对μ作点估计，由中心极限定理，当n很大时，$\bar{X}$近似服从正态分布。又$E(\bar{X})=\mu,D(\bar{X})=\frac{\sigma^2}{n}$，所以$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$近似服从标准正态分布$N(0,1)$。然而，在$n$很大时，$\sigma$可用子样标准差$S^*$近似，因此上式中$\sigma$换成$S^*$后对它的分布影响不大，故当$n$很大时，$U=\frac{\bar{X}-\mu}{S^*/\sqrt{n}}$仍近似服从标准正态分布，这个$U$可作为枢轴量，给定$1-\alpha$，可找到$u_{1-\frac{\alpha}{2}}$使得$\mathbb{P}\{|U|\le u_{1-\frac{\alpha}{2}}\}=\mathbb{P}\{\frac{|\bar{X}-\mu|}{S^*/\sqrt{n}}\le u_{1-\frac{\alpha}{2}}\}\approx 1-\alpha$ 即$\mathbb{P}\{\bar{X}-u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}\le\mu\le\bar{X}+u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}\}\approx 1-\alpha$

于是μ的置信区间是$[\bar{X}-u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}},\bar{X}+u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}]$

而置信概率近似等于$1-\alpha$。需要指出的是，用这种方法求置信区间对n很大的子样适用，这是由于导出的枢轴量U的近似分布用到了中心极限定理。根据实际经验，在这里一般认为$n\ge 50$的子样是大子样。

例 4、从一台机床加工的轴中随机地抽取200根，测量其椭圆度。由测量值计算得平均值$\bar{X}=0.081$毫米，标准差$S^*=0.025$毫米。给定置信概率为95%，求此机床加工的轴平均椭圆度的置信区间。

解：由题意，可以认为$n=200$是大子样。已知$1-\alpha=0.95$，查表得$u_{1-\frac{\alpha}{2}}=1.96$，

因此，置信下限：$\bar{X}-u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}=0.081-1.96\times\frac{0.025}{\sqrt{200}}=0.078$，

置信上限：$\bar{X}+u_{1-\frac{\alpha}{2}}\frac{S^*}{\sqrt{n}}=0.081+1.96\times\frac{0.025}{\sqrt{200}}=0.084$

故置信区间是：$[0.078, 0.084]$。

例 5、另一个典型的例子是关于比例 p 的置信区间，设 $X_1,\ldots,X_n$ 是来自二点分布 $b(1,p)$ 的样本，其分布列为

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

现要求 p 的$1-\alpha$ 置信区间。从总体中抽取一个容量为n（n充分大）的子样，其中恰有m个“1”，现在对作区间估计，此时

$$\mu=E(X)=p, \bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{m}{n}$$

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

在最后一式的推导中需注意X_i仅能取“1”或“0”，因此

$$\mathbb{P}\{\frac{m}{n}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\frac{m}{n}(1-\frac{m}{n})}\le p\le\frac{m}{n}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\frac{m}{n}(1-\frac{m}{n})}\}\approx 1-\alpha$$

故 p 的置信区间是：

$$[\frac{m}{n}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\frac{m}{n}(1-\frac{m}{n})}, \frac{m}{n}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\frac{m}{n}(1-\frac{m}{n})}]$$

例 6、对某事件A作120次观察，A发生36次，试给出事件A的发生概率p的0.95置信区间

解：此处$n=120, \frac{m}{n}=\frac{36}{120}=0.3$，而$u_{0.975}=1.96$，于是$p$的$0.95$置信下限和上限分别为

$$\frac{m}{n}-u_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\frac{m}{n}(1-\frac{m}{n})}=0.3-1.96\times\sqrt{\frac{0.3\times0.7}{120}}=0.218,$$

$$\frac{m}{n}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\frac{m}{n}(1-\frac{m}{n})}=0.3+1.96\times\sqrt{\frac{0.3\times0.7}{120}}=0.382,$$

故所求 p 的置信区间为 $[0.218,0.382]$。

例 7、某传媒公司欲调查电视台某综艺节目收视率 p，为使得 p 的 $1-\alpha$ 置信区间长度不超过 d_0，问应调查多少用户？

解：这是关于二点分布比例 p 的置信区间问题，通过计算得 p 的 $1-\alpha$ 置信区间长度为 $2u_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\frac{m}{n}(1-\frac{m}{n})}=2u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$ 这是一个随机变量，但由于 $\bar{X}\in(0,1)$，所以对任意的观测值有 $\bar{X}(1-\bar{X})\le 0.5^2=0.25$。这也就是说 p 的 $1-\alpha$ 置信区间长度不会超过 $u_{1-\alpha/2}\big/\sqrt{n}$，现要求 p 的 $1-\alpha$ 的置信区间长度不超过 d_0，只需要 $u_{1-\alpha/2}\big/\sqrt{n}\le d_0$ 即可，从而 $n\ge(\frac{u_{1-\alpha/2}}{d_0})^2$，这是一类常见的寻求样本量的问题，比如，若取 $d_0=0.04,\alpha=0.05$，则

$$n\ge(\frac{u_{0.975}}{0.04})^2=2401$$

这表明，要使得综艺节目收视率 p 的置信区间的长度不超过 0.04，则需要对 2401 个用户做调查。

3、双正态正态总体 σ_1^2 和 σ_2^2 已知时的 $\mu_1-\mu_2$ 的置信区间

设 $X_1,\ldots,X_m$ 是来自 $N(\mu_1,\sigma_1^2)$ 的样本，$Y_1,\ldots,Y_n$ 是来自 $N(\mu_2,\sigma_2^2)$ 的样本，且两个样本相互独立，$\bar{X}$ 和 $\bar{Y}$ 分别是它们的样本均值，则 $\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n})$，取枢轴量为 $U=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1)$，于是 $\mu_1-\mu_2$ 的 $1-\alpha$ 置信区间为

$$[\bar{X}-\bar{Y}-u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}},\bar{X}-\bar{Y}+u_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}].$$

例 8.设从总体 $X\sim N(\mu_1,\sigma_1^2)$ 和总体 $Y\sim N(\mu_2,\sigma_2^2)$ 中分别抽取容量为 $n_1=10,\ n_2=15$ 的独立样本,算得 $\bar{x}=82,s_x^2=56.5,\bar{y}=76,s_y^2=52.4.\ \sigma_1^2=64,\sigma_2^2=49,$

求 $\mu_1-\mu_2$ 的置信水平为 95%的置信区间;

解：在 σ_1^2,σ_2^2 都已知时,$\mu_1-\mu_2$ 的 1-α 的置信区间为

$$\left[\overline{x}-\overline{y}-u_{1-a/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}},\overline{x}-\overline{y}-u_{1-a/2}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}\right]$$

经计算$\overline{x}-\overline{y}=6$，查表得 $u_{0.975}$=1.96，因而$\mu_1-\mu_2$的置信水平为 95%的置信区间为

$$\left[6-1.96\sqrt{\frac{64}{10}+\frac{49}{15}},6+1.96\sqrt{\frac{64}{10}+\frac{49}{15}}\right]=[-0.0939,12.0939]$$

二、区间检验

（一）单正态分布方差σ^2已知的μ区间检验

例 1. 某电器的平均电阻一直保持在2.64Ω，改变加工工艺后，测得 100 个零件的平均电阻为2.62Ω，如果改变工艺前后电阻的标准差保持在0.06Ω，问新工艺对此零件的电阻有无显著影响（$\alpha=0.01$）？

解 假设检验 $H_0:\mu=\mu_0=2.64$，构造统计量： $U=\dfrac{\overline{X}-\mu_0}{0.06/10}\sim N(0,1)$

$\alpha=0.01\Rightarrow 1-\dfrac{\alpha}{2}=0.995$，查正态分布表得$u_{0.995}=2.575$，拒绝域为$\{|u|>2.575\}$，

由样本值计算得$u=\dfrac{2.62-2.64}{0.06/10}=-3.33$，$|u|>2.575$，所以拒绝$H_0$，即认为新工艺对此零件的电阻有显著影响。$u_{1-\frac{\alpha}{2}}=2.575$,也可记为$z_{\frac{\alpha}{2}}=2.575$

在正态总体中已知σ^2，则可通过抽样, 假设$\mu=\mu_0$对$Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$(或$U=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$)进行区间检验

例2、 **一种罐装饮料采用自动生产线生产，每罐的容量是 255ml，标准差为 5ml。为检验每罐容量是否符合要求，质检人员在某天生产的饮料中随机抽取了 40 罐进行检验，测得每罐平均容量为 255.8ml。取显著性水平 α=0.05，检验该天生产的饮料容量是否符合标准要求？**

解：提出假设

$H0$: $\mu=255$　　　　$H1$: $\mu\neq 255$

根据 α = 0.05 查找临界值 $z_{\alpha/2}$=1.96,

计算检验统计量 $z=\dfrac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}=\dfrac{255.8-255}{5/\sqrt{40}}=1.01$

作出决策: ∵ ∣z∣=1.01≤1.96

∴不拒绝 *H*0

结论: 我们没有充足理由证明该天生产的饮料不符合标准要求

例 3、**根据过去大量资料，某厂生产的灯泡的使用寿命服从正态分布 *N*~(1020，1002)。现从最近生产的一批产品中随机抽取 16 只，测得样本平均寿命为 1080 小时。试在 0.05 的显著性水平下判断这批产品的使用寿命是否有显著提高？(α＝0.05)**

解：提出假设　H0: μ　≤ 1020　　　H1: μ　> 1020

根据α = 0.05 查找临界值: $z_\alpha = u_{1-\alpha} = u_{0.95} = 1.96$

计算检验统计量: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{1080 - 1020}{100/\sqrt{14}} = 2.4$

决策:　∵z=2.4＞ z α =1.645 ∴拒绝 *H*0

结论:　有证据表明这批灯泡的使用寿命有显著提高

例 4．有一批枪弹，出厂时，其初速 *v*~N(950,100)(单位：*m*/s).经过较长时间储存，取 9 发进行测试，得样本值(单位：*m*/s)　如下

914　　920　　910　　934　　953　　945　　912　　924　　940

据经验，枪弹经储存后其初速度仍服从正态分布，且标准差保持不变，问是否可认为这批枪弹的初速度有显著降低(α =0.05)？

解：**提出假设** $H_0: \mu \ge 950, H_1: \mu < 950$.

若取 *a*=0.05，**查找临界值:** $-u_{0.95} = -1.645$。

计算检验统计量:，$\overline{x} = 928, u = \dfrac{928 - 950}{10/3} = -6.6,$

决策:　$\because u = -6.6 < -1.645$　**∴拒绝 *H*0**

此处 μ 值落入拒绝域内，故拒绝原假设，可以判断这批枪弹的初速有显著降低。

例 5 已知某炼铁厂铁水含炭量服从正态分布 $N(4.55, 0.108^2)$.现在测定了 9 炉铁水 ,其平均含炭量为 4.484,如果铁水含炭量的方差没有变化,可否认为现在生产的铁水平均含炭量仍为 4.55(*a*=0.05)？

解：这是关于正态总体均值的双侧假设检验问题，原假设 H_0 和备择假设 H_1 分别为

$$H_0: \mu = 4.55 \quad \text{vs} \quad H_1: \mu \ne 4.55.$$

由于总体方差已知，故采用 *u* 检验，检验的拒绝域为 $\{|\mu| \ge \mu_{1-a/2}\}$ 当 *a*=0.05 时，查表知 $u_{0.975} = 1.96$，由已知条件，$\overline{x} = 4.484$，故

$$u = \frac{4.484 - 4.55}{0.108/3} = -1.83,$$

这里 *u* 值没有落入拒绝域，故不能拒绝原假设，因而可以认为生产的铁水平均含炭量仍为 4.55。

例 6、一个小学校长在报纸上看到这样的报道：“这一城市的小学学生平均每周看 8h 电视。”

她认为她所在学校的学生看电视的时间明显小于该数字。为此她，在该校随机调查了 100 个学生，得知平均每周看电视的时间$\bar{x}=6.5h$，样本标准差为 s=2h。问是否可以认为这位校长的看法是对的(取$\alpha=0.05$)？

解: 由于本题中样本量最大，可以认为样本均值服从正态分布，依题意，需要建立的原假设和备择假设为

$$H_0: \mu=8 \quad \text{vs} \quad H_1: \mu<8.$$

若取$\alpha=0.05$，则 $u_{0.05}$= -1.65，拒绝域为{$u \le -1.65$}，由样本观测值得:

$$u=\frac{10(6.5-8)}{2}=-7.5<-1.65,$$

因而拒绝原假设，认为这位校长的看法是对的。

9. 设在木材中抽取 100 根，测其小头直径，得到样本平均数为$\bar{x}=11.2cm$，样本标准差$s=2.6cm$，问该批木材小头的平均直径能否认为不低于 12*cm*(取$\alpha=0.05$)？

解: 本题与第 8 题类似，只是这里的原假设和备择假设分别为

$$H_0: \mu \ge 12 \qquad \text{vs} \qquad H_1: \mu<12,$$

拒绝域为{$u<u_a$}，当取$\alpha=0.05$时，$u_{0.05}$= -1.65，检验统计量

$$u=\frac{10(11.2-12)}{2.6}=-3.0769<-1.65,$$

u 值落入拒绝域内，因此拒绝原假设，不能认为该批木材小头的平均直径不低于 12*cm*。

（二）大样本的正态分布

设总体 X 分布任意,$E(X)=\mu, D(X)=\sigma^2$(未知),假设$H_0: \mu=\mu_0$，由中心极限定理$\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$近似服从$N(0,1)$.又因为 S 是$\sigma$的一致无偏估计，于是$S \approx \sigma$.

即$U=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$近似服从$N(0,1)$

例 1、某电器元件的平均电阻一直保持在 2.64Ω.改变加工工艺后，测量 100 个元件的电阻，计算得平均电阻为 2.62Ω，标准差 S 为 0.06Ω.问新工艺对此元件的平均电阻有无显著影响（给定显著水平 0.01）

解：**提出假设 H0: μ = 2.64　　H1: μ ≠ 2.64**

根据α = 0.01 查找临界值: $z_{\alpha/2}=u_{1-\frac{\alpha}{2}}=u_{0.995}=2.57$

计算检验统计量: n=100, $\bar{x}=2.62, S=0.06, U=\dfrac{\bar{x}-\mu_0}{S/\sqrt{n}}=\dfrac{2.62-2.64}{0.06/10}=-3.333$

决策: ∵U=-3.333 ∴| U |> $z_{\alpha/2}$ ∴拒绝 *H0*
结论: 认为新工艺对此元件的平均电阻有显著影响

三、总体比例的假设检验

例 1、某厂有一批产品，共一万件，须经过检验后方可出厂。按规定标准，次品不得超过 5%。今在其中任意选取 50 件产品进行检查，发现有次品 4 件，问这批产品能否出厂？（$\alpha = 0.01$）

解：总体服从 0-1 分布，即 $P(X=1)=p, P(X=0)=1-p$，p 为次品率

提出假设 H0:p＝p_0=0.05

设 $X_1, X_2, \ldots, X_n$ 是一个样本，则 $\overline{X}=\frac{m}{n}$，其中 m 是 n 个产品的次品数

在 H0 成立的条件下 $\overline{X}$ 近似服从正态分布，

且 $E(\overline{X})=p_0$， $D(\overline{X})=\frac{\sigma^2}{n}=\frac{p_0(1-p_0)}{n}$

$$U=\frac{\overline{X}-E(\overline{X})}{\sqrt{D(\overline{X})}}=\frac{\frac{m}{n}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}=\frac{\frac{4}{50}-0.05}{\sqrt{\frac{0.05\times 0.95}{100}}}=1.376$$

根据α = 0.01 查找临界值: $z_{\alpha/2}=u_{1-\frac{\alpha}{2}}=u_{0.995}=2.57$

于是$|U|<$ $z_{\alpha/2}$，故接受 **H0，认为这批产品的次品率是 5%，可以出产。**

注：比例检验的 *U* 统计量$U=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\sim N(0,1)$

例 2、一种以休闲和娱乐为主题的杂志，声称其读者群中有 80%为女性。为验证这一说法是否属实，某研究部门抽取了由 200 人组成的一个随机样本，发现有 146 个女性经常阅读该杂志。分别取显著性水平α =0.05 和α =0.01 ，检验该杂志读者群中女性的比例是否为 80%？

解:

提出假设 H0 ： $\pi=80\%$ H1 ： $\pi\neq 80\%$

根据α = 0.05 查找正态分布临界值$z_{\frac{\alpha}{2}}=1.96$

计算 z 检验统计量: $p=\frac{146}{200}=0.73, u=\frac{0.73-0.80}{\sqrt{\frac{0.80\times(1-0.80)}{200}}}=-2.475$

作出决策∵｜z｜=2.475＞ $z_{\alpha/2}=1.96$

∴拒绝 *H0* 结论：　该杂志的说法并不属实

第二节 可用卡方分布解决的问题

一、伽马分布

1、定义：密度函数为 $f(x)=\begin{cases}\dfrac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}, & x>0\\ 0 & ,x\le 0.\end{cases}$ 的随机变量X的分布叫服从伽马分布，

记作 $X\sim Ga(\alpha,\lambda)$.

当 α ＝1 时 $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0\\ 0 & ,x<0.\end{cases}$ 是指数分布，即 $Ga(1,\lambda)=EXP(\lambda)$

当 $\alpha=\dfrac{n}{2},\lambda=\dfrac{1}{2}$ 时 $f(x)=\begin{cases}\dfrac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x>0\\ 0 & ,x\le 0.\end{cases}$ 是自由度为卡方分布，即 $Ga(\dfrac{n}{2},\dfrac{1}{2})=\chi^2(n)$

2、可加性：若 $X\sim Ga(\alpha_1,\lambda),Y\sim Ga(\alpha_2,\lambda)$. 则 $X+Y\sim Ga(\alpha_1+\alpha_2,\lambda)$

证明：当 $z\le 0$ 显然 $f(z)=0$

当 $z>0$ 时，用卷积公式得

$$f_Z(z)=\int_0^Z f_X(z-y)f_Y(y)dy=\frac{\lambda^{\alpha_1}\lambda^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^Z(z-y)^{\alpha_1-1}e^{-\lambda(z-y)}y^{\alpha_2-1}e^{-\lambda y}dy$$
$$=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^Z(z-y)^{\alpha_1-1}y^{\alpha_2-1}dy$$

令 $t=\dfrac{y}{z}$,则 $0\le t\le 1,y=zt,dy=zt$

$$f_Z(z)=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1(z-zt)^{\alpha_1-1}(zt)^{\alpha_2-1}zdt$$
$$=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1(1-t)^{\alpha_1-1}t^{\alpha_2-1}dt=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\beta(\alpha_1,\alpha_2)$$
$$=\frac{\lambda^{\alpha_1+\alpha_2}e^{-\lambda z}z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}=\frac{\lambda^{\alpha_1+\alpha_2}z^{\alpha_1+\alpha_2-1}e^{-\lambda z}}{\Gamma(\alpha_1+\alpha_2)}$$

于是 $Z\sim Ga(\alpha_1+\alpha_2,\lambda)$

二、卡方分布

1、χ^2 分布：设 $X_1,X_2,\cdots,X_n$ 是来自总体 $N(0,1)$ 的样本，
则称统计量 $\chi^2=X_1^2+X_2^2+\cdots+X_n^2$ 服从自由度为
n 的 χ^2 分布，记为 $\chi^2\sim\chi^2(n)$.

自由度：指 $\chi^2=X_1^2+X_2^2+\cdots+X_n^2$ 中右端包含独立
变量的个数.

$\chi^2(n)$分布的概率密度为$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}, & y>0\\ 0 & 其他.\end{cases}$

证明： 设$Y=X^2, X\sim N(0,1)$,则

$y\le 0$时，$F_Y(y)=0$,

当$y>0$时，$F_Y(y)=P(Y\le y)=P(-\sqrt{y}\le X\le\sqrt{y})=\int_{-\sqrt{y}}^{\sqrt{y}}f_X(x)dx$

$$F_Y'(y)=\frac{1}{2\sqrt{y}}f_X(\sqrt{y})+\frac{1}{2\sqrt{y}}f_X(-\sqrt{y})=\frac{1}{2\sqrt{y}}\bullet\frac{1}{\sqrt{2\pi}}\times 2e^{-\frac{y}{2}}$$

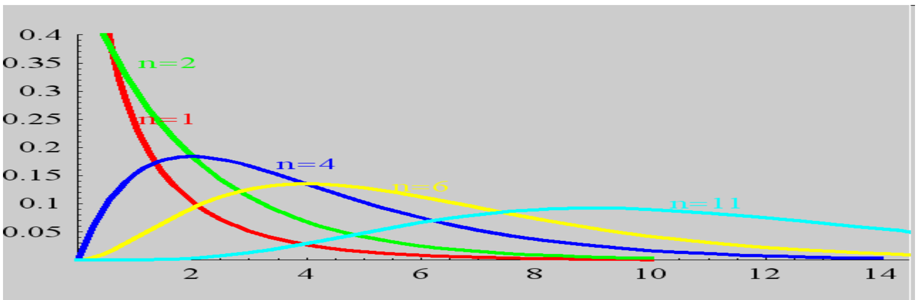
$$=\frac{1}{\sqrt{2\pi}}\times y^{1-\frac{1}{2}}e^{-\frac{y}{2}}=\frac{(\frac{1}{2})^{\frac{1}{2}}}{\Gamma(\frac{1}{2})}\times y^{1-\frac{1}{2}}e^{-\frac{y}{2}}$$

于是设$Y\sim Ga\left(\frac{1}{2},\frac{1}{2}\right)$，即 $\chi^2(1)$ 分布即为$Ga\left(\frac{1}{2},\frac{1}{2}\right)$分布,

根据Ga分布的可加性知$\chi^2(n)=\sum_{i=1}^{n}X_i^2\sim Ga\left(\frac{n}{2},\frac{1}{2}\right)$.

于是$\chi^2(n)$分布的概率密度为$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}, & y>0\\ 0 & 其他.\end{cases}$

$\chi^2(n)$分布的概率密度曲线如图.



2、χ^2 分布的性质

性质 1(χ^2 分布的可加性)　设$\chi_1^2\sim\chi^2(n_1),\chi_2^2\sim\chi^2(n_2)$,并且$\chi_1^2,\chi_2^2$独立,则$\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$.此性质可以推广到多个随机变量的情形.

设 $\chi_i^2\sim\chi^2(n_i)$,并且$\chi_i^2(i=1,2,\cdots,m)$相互独立,则$\sum_{i=1}^{m}\chi_i^2\sim\chi^2(n_1+n_2+\cdots+n_m)$.

性质 2(χ^2分布的数学期望和方差)

若 $\chi^2\sim\chi^2(n)$, 则 $E(\chi^2)=n,\quad D(\chi^2)=2n$.

证明因为 $X_i\sim N(0,1)$, 所以 $E(X_i^2)=D(X_i)=1$,

$D(X_i^2)=E(X_i^4)-[E(X_i^2)]^2=3-2=1, i=1,2,\cdots,n$.

故 $E(\chi^2)=E\left(\sum_{i=1}^{n}X_i^2\right)=\sum_{i=1}^{n}E(X_i^2)=n,\quad D(\chi^2)=D\left(\sum_{i=1}^{n}X_i^2\right)=\sum_{i=1}^{n}D(X_i^2)=2n.$

3、正态总体的样本均值与样本方差的分布

正态总体 $N(\mu,\sigma^2)$ 的样本均值和样本方差有以下两个重要定理.

定理一 设 $X_1, X_2, \cdots, X_n$ 是来自正态总体 $N(\mu,\sigma^2)$ 的样本, $\bar{X}$ 是样本均值, 则有 $\bar{X} \sim N(\mu, \sigma^2/n)$.

定理二

设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本, $\bar{X}, S^2$ 分别是样本均值和样本方差, 则有

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

(2) $\bar{X}$ 与 S^2 独立.

证　明：

已知 $X_i\left(i=1,2,\cdots,n\right) \sim N\left(\mu,\ \sigma^2\right)$，且相互独立，

$$令\ Y_i = \frac{X_i-\mu}{\sigma}\left(i=1,2,\cdots,n\right) \Rightarrow Y_1,Y_2,\cdots,Y_n \sim N\left(0,1\right)，且相互独立。$$

作下列正交变换：

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

正交变换不改变向量组的秩，由于 $Y_1,Y_2,\cdots,Y_n$ 相互独立，则 $Z_1,Z_2,\cdots,Z_n$ 相互独立，

且都服从 $N\left(0,1\right)$。

$$记\ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} = \frac{1}{\sigma}\cdot\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mu}{\sigma} = \frac{\bar{X}}{\sigma} - \frac{1}{n}\cdot n\cdot\frac{\mu}{\sigma} = \frac{\bar{X}-\mu}{\sigma}$$

由上述变换矩阵等式易得： $Z_1 = \dfrac{1}{\sqrt{n}}Y_1 + \dfrac{1}{\sqrt{n}}Y_2 + \cdots + \dfrac{1}{\sqrt{n}}Y_n = \sqrt{n}\bar{Y} \sim N\left(0,1\right)$

正交变换不改变向量的长度 $\Rightarrow \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z_i^2$，所以

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

$$=\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2+\sum_{i=1}^{n}(\frac{\overline{X}-\mu}{\sigma})^2-2\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)\left(\frac{\overline{X}-\mu}{\sigma}\right)$$

$$=\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2+n(\frac{\overline{X}-\mu}{\sigma})^2-2n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2$$

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

$$=\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2=\sum_{i=1}^{n}Z_i^2-Z_1^2=\sum_{i=2}^{n}Z_i^2\sim\chi^2(n-1)$$

评 注 $\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)$ 有重要的应用价值，如计算 $E\left(S^2\right)$；$D\left(S^2\right)$。

$$\because E\left(\chi^2(n-1)\right)=n-1,\quad D\left(\chi^2(n-1)\right)=2(n-1)$$

$$\Rightarrow E\left(S^2\right)=\frac{\sigma^2}{n-1}E\left(\frac{(n-1)S^2}{\sigma^2}\right)=\frac{\sigma^2}{n-1}E\left(\chi^2(n-1)\right)=\frac{\sigma^2}{n-1}\cdot(n-1)=\sigma^2$$

$$D\left(S^2\right)=\left[\frac{\sigma^2}{n-1}\right]^2D\left(\frac{(n-1)S^2}{\sigma^2}\right)=\left[\frac{(n-1)}{\sigma^2}\right]^2\cdot D\left(\chi^2(n-1)\right)=\left[\frac{\sigma^2}{n-1}\right]^2\cdot 2(n-1)=\frac{2\sigma^4}{n-1}$$

例 1 从正态总体 $X\sim N(\mu,\sigma^2)$ 中，抽取了 $n=20$ 的样本 $(X_1,X_2,\cdots,X_{20})$

(1) 求 $P\left(0.37\sigma^2\le\frac{1}{20}\sum_{i=1}^{20}\left(X_i-\overline{X}\right)^2\le 1.76\sigma^2\right)$

(2) 求 $P\left(0.37\sigma^2\le\frac{1}{20}\sum_{i=1}^{20}\left(X_i-\mu\right)^2\le 1.76\sigma^2\right)$

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

即 $\frac{19S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{20}\left(X_i-\overline{X}\right)^2\sim\chi^2(19)$ 故 $P\left(0.37\sigma^2\le\frac{1}{20}\sum_{i=1}^{20}\left(X_i-\overline{X}\right)^2\le 1.76\sigma^2\right)$

$$=P\left(7.4\le\frac{1}{\sigma^2}\sum_{i=1}^{20}\left(X_i-\overline{X}\right)^2\le 35.2\right)$$

$$=P\left(\frac{1}{\sigma^2}\sum_{i=1}^{20}\left(X_i-\overline{X}\right)^2\ge 7.4\right)-P\left(\frac{1}{\sigma^2}\sum_{i=1}^{20}\left(X_i-\overline{X}\right)^2\ge 35.2\right)$$

$$\overset{\text{查表}}{=}0.99-0.01=0.98$$

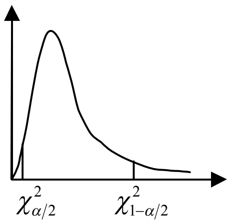
(2) $\sum_{i=1}^{20}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(20)$ 故 $P\left(0.37\sigma^2 \le \frac{1}{20}\sum_{i=1}^{20}\left(X_i-\mu\right)^2 \le 1.76\sigma^2\right)$

$$=P\left(7.4\le\sum_{i=1}^{20}\left(\frac{X_i-\mu}{\sigma}\right)^2\le 35.2\right)=P\left(\sum_{i=1}^{20}\left(\frac{X_i-\mu}{\sigma}\right)^2\ge 7.4\right)-P\left(\sum_{i=1}^{20}\left(\frac{X_i-\mu}{\sigma}\right)^2\ge 35.2\right)$$
$$=0.995-0.025=0.97$$

二、单个正态总体 σ^2 的置信区间。

（一） 当 μ 未知时， 方差 σ^2 的 置信区间

由于枢轴量 $\frac{(n-1)S^{*2}}{\sigma^2}\sim\chi^2(n-1)$，由于 χ^2 分布是偏态分布，寻找平均长度最短区间很难实现，一般都改为寻找等尾置信区间：把 α 分成为两部分，在 χ^2 分布两侧各截面积为 $\frac{\alpha}{2}$ 的部分，即采用 χ^2 的两个分位数 $\chi^2_{\frac{\alpha}{2}}(n-1)$ 和 $\chi^2_{1-\frac{\alpha}{2}}(n-1)$，它们满足



$$P(\chi^2_{\frac{\alpha}{2}}\le\frac{(n-1)S^{*2}}{\sigma^2}\le\chi^2_{1-\frac{\alpha}{2}})=1-\alpha$$

由此给出 σ^2 的 $1-\alpha$ 置信区间为 $[\frac{(n-1)S^{*2}}{\chi^2_{1-\frac{\alpha}{2}}(n-1)},\frac{(n-1)S^{*2}}{\chi^2_{\frac{\alpha}{2}}(n-1)}]$

将上式的两端开方即得到标准差 σ 的 $1-\alpha$ 置信区间。

（二）当 μ 已知时， 方差 σ^2 的 置信区间（这种情况在实际中很少）

取枢轴量 $Q=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n)$ 由下式

$$P\left(\chi^2_{\frac{\alpha}{2}}(n)\le\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}\le\chi^2_{1-\frac{\alpha}{2}}(n)\right)=1-\alpha$$

得 σ^2 的置信度为 $1-\alpha$ 置信区间为 $\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)},\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right)$

例 1、某厂生产的零件重量服从正态分布 $N(\mu,\sigma^2)$，现从该厂生产的零件中抽取 9 个，测得其质量为（单位：g）

$$45.3 \quad 45.4 \quad 45.1 \quad 45.3 \quad 45.5 \quad 45.7 \quad 45.4 \quad 45.3 \quad 45.6$$

试求总体标准差σ的0.95置信区间。

解 由数据可算得$S^{*2}=0.032, (n-1)S^{*2}=8\times 0.0325=0.26$，这里$\alpha=0.05$，查表知$\chi^2_{0.025}(8)=2.1797, \chi^2_{0.975}(8)=17.5345,$ 则σ^2的0.95置信区间为$[0.0148, 0.1193]$，从而得到σ的0.95置信区间为$[0.1218, 0.3454]$。

例 2、某工厂生产一批滚珠， 其直径 **X** 服从正态分布 **N(μ,σ2)**， 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 σ2=0.06，求 μ 的置信区间② 求方差 σ2 的置信区间． 置信度均为 0.95

解$\because \sigma^2=0.06$已知，故 μ 的置信区间为$[\bar{x}-\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}, \ \bar{x}+\frac{\sigma}{\sqrt{n}}u_{1-\frac{\alpha}{2}}],$

$$\bar{x}=\frac{1}{6}\sum_{i=1}^{6}x_i=14.95, \text{又}\alpha\text{=0. 05, 查表得}u_{1-\frac{\alpha}{2}}=u_{0.975}=1.96$$

$$\text{所求的置信区间为}[\bar{x}-\frac{\sigma}{\sqrt{n}}u_{0.975}, \ \bar{x}+\frac{\sigma}{\sqrt{n}}u_{0.975}]$$

$$=[14.95-1.96\times\frac{\sqrt{0.06}}{\sqrt{6}}, \quad 14.95+1.96\times\frac{\sqrt{0.06}}{\sqrt{6}}]=[14.754, \quad 15.146]$$

② σ^2 的置信区间为$[\ \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \quad \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\ ]$

$s^2=0.051.$

查表得 $\chi^2_{1-\alpha/2}(n-1)=\chi^2_{0.975}(5)=12.833, \quad \chi^2_{\alpha/2}(n-1)=\chi^2_{0.025}(5)=0.831$

所以 σ^2 的置信区间为$[\ \frac{5s^2}{\chi^2_{0.975}(5)}, \quad \frac{5s^2}{\chi^2_{0.025}(5)}\ ]=[0.0199, \quad 0.3069\ ]$

三、方差检验

例 1．已知维尼纶纤度在正常条件下服从正态分布，且标准差 0。048.从某天产品中 5 根纤维，测得其纤度为

1.32，1.55，1.36，1.40，1.44.

问这一天纤度的 总体标准差是否正常(取α=0.05)?

解：这是一个关于正态总体方差的双侧检验问题，待检验的原假设和备择假设分别为

$$H_0: \sigma^2=0.048^2 \quad \text{vs} \quad H_1: \sigma^2\neq 0.048^2.$$

此处$n=5$,若取显著性水平$\alpha=0.05$，查表知$\chi^2_{0.025}(4)=0.0484, \chi^2_{0.975}(4)=11.143$, 故拒绝域为$W=\left\{\chi^2 \le 0.484 或 \chi^2 \ge 11.143\right\}$，有样本数据可计算得到

$$\chi^2=\frac{(n-1)s^2}{\sigma_0{}^2}=\frac{0.03112}{0.048^2}=13.5069>11.1433,$$

因此拒绝H_0，认为这一天纤度的总体标准差不正常.

讨论：在此种 拒绝原假设的场合，人们还经常会问总体标准差σ究竟是大雨 0.048 还是小于 0.048 呢？回答这个问题要进一步考察其拒绝域，事实上，此处拒绝域$W=W_1 \cup W_2$由二部分组成：

$$W_1=\left\{\chi^2<\chi^2_{0.025}(4)\right\}=\left\{\chi^2<0.4844\right\}, 它对应\sigma<0.048;$$

$$W_2=\left\{\chi^2>\chi^2_{0.975}(4)\right\}=\left\{\chi^2>11.1433\right\}，它对应\sigma>0.048.$$

如今样本观测值落入 W_2,可以认为当天的总体标准差$\sigma>0.048$.

在实际中，产品质量的标准差通常不大可能自然减小的，人们要预防的时标准差变大而使得产品质量下降，因此，对本体的情况，一般也可建立如下的检验问题：

$$H_0:\sigma^2 \le 0.048^2 \quad \text{vs} \quad H_1:\sigma^2>0.048^2.$$

若此，则检验的统计量不变，而检验的拒绝域为$W'=\left\{\chi^2 \le \chi^2_{1-\alpha}(4)\right\}$，若仍取显著性水平为 0.05,则查表知$\chi^2_{0.95}(4)=9.4877$，由于观测值大于临界值，仍然拒绝原假设，认为当天的总体标准差$\sigma>0.048$.

例 2.某电工器材厂生产一种保险丝。测量器熔化时间，依通常情况方差为 400,今从某天的产品中抽取容量为 25 的样本，测量器熔化时间并计算得$\bar{x}=62.24, s^2=404.77$，问这天保险丝熔化时间的方差与通常有无显著差异？(取$\alpha=0.05$,假定熔化时间服从正态分布)

解：本体可归结为关于正态总体方差的双侧检验问题

$$H_0:\sigma^2=400 \quad \text{vs} \quad H_1:\sigma^2 \ne 400.$$

当 $\alpha=0.05$ 时，查表知 $\chi^2_{0.025}(24)=12.401$，$\chi^2_{0.975}(24)=39.3641$，因此拒绝域为$W=\left\{\chi^2 \le 12.401\right\} \cup \left\{\chi^2 \ge 39.3641\right\}$,由所给的条件可得出检验统计量为

$$\chi^2=\frac{24\times 404.77}{400}=24.2862,$$

可见$\chi^2 \notin W$，因此在显著性水平$\alpha = 0.05$下，接受H_0，即认为该天保险丝熔化时间的方差与通常无显著差异.

例 3. 某种导线的质量标准要求其电阻的标准差不超过 0.005Ω。今从一批导线中随机抽取 9 根，测得样本标准差为 s=0.007Ω,设总体为正态分布，问在显著性水平α=0,05 下能否认为这批导线的标准差显著地偏大？

解.本题是单侧检验问题：$H_0 : \sigma^2 \le 0.005^2$ vs $H_1 : \sigma^2 > 0.005^2$.

$\alpha = 0.05$时,查表知$\chi^2_{0.95}(8) = 15.5073$,拒绝域为$W = \{\chi^2 \ge 15.5073\}$,由所给的条件可得出检验统计量为

$$\chi^2 = \frac{8 \times 0.007^2}{0.005^2} = 15.68 > 15.5037,$$

因此拒绝H_0,在显著性水平$\alpha = 0.05$下认为这批导线的标准差显著地偏大.

四、分布拟合检验

在总体 X 的分布未知时，根据来自总体的样本，检验关于总体分布的假设的一种检验方法.

基本思想： 提出原假设：$H0$：总体 X 的分布函数为 $F(x)$

然后根据样本的经验分布和所假设的理论分布之间的吻合程度来决定是否接受原假设.这种检验通常称作拟合优度检验，它是一种非参数检验.

基本原理和步骤如下:

1. 将总体 X 的取值范围分成 k 个互不重迭的小区间（或小组）， 记作 $A_1, A_2, \ldots, A_k$.

2.把落入第 i 个小区间 Ai 的样本值的个数记作 ni ，称为实测频数. 所有实测频数之和 $n_1+ n_2+ \ldots+n_k$ 等于样本容量 n

$H0$：$P(A_i) = p_i$, i=1,2,….k.

3.根据所假设的理论分布,可以算出总体 X 的值落入每个

A_i 的概率 p_i ,于是 np_i 就是落入 A_i 的样本值的理论频数. $n_i - np_i$ 它标志着经验分布与理论分布之间的差异的大小.

4. 皮尔逊引进如下统计量表示经验分布与理论分布之间的差异： $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$

Pearson 证明了如下定理 1: 若原假设中的理论分布 $F(x)$已经完全给定，那么当 n 充分大时，统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(k-1)$$

把n_i**记为**f_0**为**实际观察次数，np_i记为f_e为理论次数，则$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \sim \chi^2(k-1)$

注:若在 $H0$ 下分布类型已知，但其参数未知，这时需要先用最大似然估计法估计参数，然后作检验.

Fisher 证明了如下定理 2:　若原假设中的理论分布 $F(x)$中有 r 个未知参数需用相应的最大似然估计来代替，那么当 n 充分大时，统计量 $\chi^2=\sum_{i=1}^{k}\frac{(n_i-n\hat{p}_i)^2}{n\hat{p}_i}\ \dot{\sim}\ \chi^2(k-r-1)$

可认为卡方检验的一般问题是要检验名义型变量的实际观测次数和理论次数分布之间是否存在显著差异。它主要应用于两种情况：

卡方检验能检验单个多项分类名义型变量各分类间的实际观测次数与理论次数之间是否一致的问题，这里的观测次数是根据样本数据得多的实计数，理论次数则是根据理论或经验得到的期望次数。这一类检验称为**拟合性检验**。

（一）检验无差假设

【例 1】 随机地将麻将色子抛掷 300 次，检验该色子的六个面是否均匀。结果 1-6 点向上的次数依次是，43，49，56，45，66，41。

解：每个类的理论次数是 300/6 = 50，代入公式：

$$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}=\frac{(43-50)^2}{50}+\frac{(49-50)^2}{50}+\frac{(56-50)^2}{50}$$
$$+\frac{(45-50)^2}{50}+\frac{(66-50)^2}{50}+\frac{(41-50)^2}{50}=8.96<\chi^2_{0.05}(5)=11.1$$

因此，在 0.05 的显著性水平下，可以说这个色子的六面是均匀的。

【例 2】 随机抽取 60 名高一学生，问他们文理要不要分科，回答赞成的 39 人，反对的 21 人，问对分科的意见是否有显著的差异。

解：如果没有显著的差异，则赞成与反对的各占一半，因此是一个无差假设的检验，于是理论次数为 60/2=30，代入公式：

$$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}=\frac{(39-30)^2}{30}+\frac{(21-30)^2}{30}=5.4>\chi^2_{0.05}(1)=3.84$$

所以对于文理分科，学生们的态度是有显著的差异的。

（二）检验假设分布的概率

这里的假设分布可以是经验性的，也可以是某理论分布。公式中所需的理论次数则按照这里假设的分布进行计算。

【例 3】 国际色觉障碍讨论会宣布，每 12 个男子中，有一个是先天性色盲。从某校抽取的 132 名男生中有 4 人是色盲，问该校男子色盲比率与上述比例是否有显著差异？

解：按国际色觉障碍讨论会的统计结果，132 人应该有 132/12=11 人是色盲，剩下的 121 人非色盲，代入公式有：

$$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}=\frac{(4-11)^2}{11}+\frac{(128-121)^2}{121}=4.86>\chi^2_{0.05}(1)=3.84$$

因此，在 0.05 和显著性水平下，该校男子色盲比率与国际色觉障碍讨论会的统计结果有显著差异，显然根据比例可知该校的色盲率小于国际色觉障碍讨论会的统计结果。

【例 4】 在英语四级考试中，某学生做对了 80 个四择一选择题中的 28 题，现在要判断该生是否是完全凭猜测做题。

解：假如该生完全凭猜测做题，那么平均而言每道题做对的可能性是 1/4，因此 80 个题中平均而能做对 80/4=20 题，代入公式有：

$$\chi^2=\sum\frac{(f_0-f_e)^2}{f_e}=\frac{(28-20)^2}{20}+\frac{(52-60)^2}{60}=4.27>\chi^2_{0.05}(1)=3.84$$

因此，该生可能会做一些题。

例 3.从 1500 年到 1931 年的 432 年间,每年爆发战争的次数可以看作一个随机变量，椐统计，这 432 年间共爆发了 299 次战争，具体数据如下:

发生战争次数 X	**0**	**1**	**2**	**3**	**4**
发生 X 次战争的年数	**223**	**142**	**48**	**15**	**4**

试检验每年爆发战争次数分布是否服从泊松分布.

解: 这是一个分布的拟合优度检验，提出假设：$H0$: $X \sim P(\lambda)$

由观察值，得参数 λ 的最大似然估计为 $\hat{\lambda}=\bar{x}=0.69$

按参数为 0.69 的泊松分布，计算事件 $X=i$ 的概率 p_i ，

p_i 的估计是 i=0,1,2,3,4 $\hat{p}_i=e^{-0.69}0.69^i/i\,!$

将有关计算结果列表如下:

战争次数 *X*	实测频数 n_i	$\hat{p}_i$	$n\hat{p}_i$	$\frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$
0	**223**	**0.5016**	**216.7**	**0.183**
1	**142**	**0.346**	**149.5**	**0.376**
2	**48**	**0.1194**	**51.6**	**0.251**
⩾3	**19**	**0.03224**	**13.91**	**1.863**
合计	**299**			**2.673**

注:将频数**<5** 的组予以合并，即将发生 **3** 次及以上次数的组归并为一组**.**

检验的拒绝域为**:** $W = \{\chi^2 \geq \chi^2_{1-\alpha}(k-r-1)\}$

因 ***H*0** 所假设的理论分布中有一个未知参数，即 **r=1,**又 ***k*=4,**故自由度为 **4-1-1=2.**

又 α **=0.05，** 查 χ **2** 分布表得**:** $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(2) = 5.9915$

由样本，得检验统计量的值：$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = 2.673 < 5.9915$ 未落入拒绝域，

故认为每年发生战争次数 ***X*** 服从泊松分布**.**

例 **3.**某建筑单位宣称其麾下的建筑工地每天发生的事故不超过 **0.6** 起**.**现记录了该建筑单位工地 **200** 天的安全生产情况**,**数据如下**,**试在α **=0.05** 水平下检验其宣称是否成立 **?**

一天发生的事故数	**0**	**1**	**2**	**3**	**4**	**5**	**⩾6**	合计
天数	**102**	**59**	**30**	**8**	**0**	**1**	**0**	**200**

分析**:** 先检验每天发生的事故数服从泊松分布**,** 这是关于总体分布的拟合优度检验**,** 再进行关于总体均值的大样本检验**.**

解**:**　　这是一个分布的拟合优度检验，提出假设：***H*0: *X*～P(**λ**)**

由观察值，得参数λ的最大似然估计为 $\hat{\lambda} = \bar{x} = 0.74$

按参数为 **0.74** 的泊松分布，计算事件 ***X*=*i*** 的概率 ***pi*** ，

pi 的估计是 $\hat{p}_i = e^{-0.74} 0.74^i / i\,!$ 　　***i*=0,1,2,3,4,5,6**

将有关计算结果列表如下:

事故 次数 X	实测 频数 ni	$\hat{p}_i$	$n\hat{p}_i$	$\frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$
0	102	**0.4771**	**95.42**	**0.4537**
1	**59**	**0.3531**	**70.61**	**1.9090**
2	**30**	**0.1306**	**26.13**	**0.5732**
3	**8**	**0.0322**	**6.44**	0.172
4	**0**	**0.0060**	**1.20**	
5	**1**	**0.0009**	**0.18**	
⩾6	**0**	**0.0001**	**0.02**	
合计	200	1	200	**3.108**

X⩾3 合在一起算

检验的拒绝域为：$W = \{\chi^2 \geq \chi^2_{1-\alpha}(k-r-1)\}$,按 α =0.05，自由度为 4-1-1=2 查 χ 2 分布表得

$\chi^2_{1-\alpha}(k-r-1) = \chi^2_{0.95}(2) = 5.9915$ 由样本得：$\chi^2 = 3.108 < 5.9915$，未落入否定域.

即认为 X 服从泊松分布.

五、独立性检验

2×2 的情况前面已经讲过

例 1、 某校对学生课外活动内容进行调查，结果整理成下表，表中彩色格子里的数是原始数据的汇总数，括号内的数是理论次数（是按下面将要介绍的原理计算得来的），此外的是原始数据。

性别(因素 2)	课外活动内容(因素 1)			小计和(fx)
	体育	文娱	阅读	
男生	21(15.3)	11(10.2)	23(29.5)	55
女生	6(11.7)	7(7.8)	29(22.5)	42

小计和(fy)	27	18	52	97

由于所有学生参加三项活动的比例是 27:18:52，因此如果课外活动的选择与性别没有关系的话，男女生参加这三项活动的比例也应是这同一比例，而男女各自的人数可以计算，所以每格内的理论次数的计算方法如下：

男生中

参加体育活动的理论人数：55×27/97=15.3

参加文娱活动的理论人数：55×18/97=10.2

参加阅读活动的理论人数：55×52/97=29.5

女生中

参加体育活动的理论人数：42×27/97=11.7

参加文娱活动的理论人数：42×18/97= 7.8

参加阅读活动的理论人数：42×52/97=22.5

我们将行列的小计和分别用 f_x 和 f_y 来表示，总人数用 N 来表示时，上述计算理论次数的方法可以表示为：

$fe_{ij} = fx_i \times fy_j / N$

所以，卡方独立性检验的公式可以表示如下，其中最后一个式子比较便于计算，fxy 表示每格的原始数据。

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = \sum \frac{(f_{xy} - f_x f_y / N)^2}{f_x f_y / N} = N(\sum_{x=1}^{R} \sum_{y=1}^{C} \frac{f_{xy}^2}{f_x f_y} - 1)$$

由于在计算理论次数时，用了按每个因素分类的小计和（fx 和 fy，其个数分别记为 R 个和 C 个），和总和 N ，而总和又可由按每个因素分类的小计和计算得来，因此若从总分类个数 R×C 中减去 R+C，则将总和重复减去了，因此要补 1 个自由度回来，所以最终独立性检验的自由度表示为：

$$df = R \times C - R - C + 1 = (R-1)(N-1)$$

上述例题最终计算得：

$$\chi^2=\frac{(21-15.3)^2}{15.3}+\frac{(11-10.2)^2}{10.2}+\frac{(23-29.5)^2}{29.5}$$
$$+\frac{(6-11.7)^2}{11.7}+\frac{(7-7.8)^2}{7.8}+\frac{(29-22.5)^2}{22.5}$$
$$=8.3552$$

或者：

$$\chi^2=97\left(\frac{21^2}{55\times27}+\frac{11^2}{55\times18}+\frac{23^2}{55\times52}+\frac{6^2}{42\times27}+\frac{7^2}{48\times18}+\frac{29^2}{42\times52}-1\right)=8.3217$$

这两个公式的计算结果有一点点差异，这完全是计算误差即四舍五入引起的。

df = (3-1)(2-1) = 2，而 $\chi^2_{0.05}(2)=5.99$，所以在 0.05 的显著性水平下，拒绝零假设，即可以认为性别与课外活动内容有关联，或者说男女生在选择课外活动上存在显著的差异。

第三节 t 分布及应用

一、 t 分布

1、设 $X\sim N(0,1)$, $Y\sim\chi^2(n)$, 且 X,Y 独立, 则称随机变量 $T=\dfrac{X}{\sqrt{Y/n}}$
服从自由度为 n 的 t 分布, 记为 $t\sim t(n)$.

***t* 分布又称学生氏(Student)分布.**

$t(n)$ 分布的概率密度函数为 $h(t)=\dfrac{\Gamma\left(\dfrac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\dfrac{n}{2}\right)}\left(1+\dfrac{t^2}{n}\right)^{-\frac{n+1}{2}}$， $-\infty<t<+\infty$

证明：先求 $Z=\sqrt{\dfrac{Y}{n}}$ 的密度函数 $f(z)$，由于 Z 非负，于是 $z\le0$ 时， $f(z)=0$

当 $z>0$ 时，

$F_Z(z)=P(Z\le z)=P(Y\le nz^2)=F_Y(nz^2)$

因此 $f_Z(z)=F_Y{}'(nz^2)=f_Y(nz^2)2nz$

因为 $f_Y(y)=\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}$，

所以 $f_Z(z)=\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}(nz^2)^{\frac{n}{2}-1}e^{-\frac{nz^2}{2}}2nz=\dfrac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}n^{\frac{n}{2}}z^{n-1}e^{-\frac{nz^2}{2}}$

因 $T=\dfrac{X}{Z}$,利用独立随机变量商的密度公式可得

$$h(t)=\int_{-\infty}^{+\infty}|z|f_X(tz)f_Z(z)dz=\int_{-\infty}^{+\infty}z\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2t^2}{2}}\frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})}n^{\frac{n}{2}}z^{n-1}e^{-\frac{nz^2}{2}}dz$$

$$=\frac{n^{\frac{n}{2}}}{\sqrt{\pi}\,2^{\frac{n-1}{2}}\Gamma(\frac{n}{2})}\int_{-\infty}^{+\infty}z^ne^{-\frac{t^2+n}{2}z^2}dz,$$

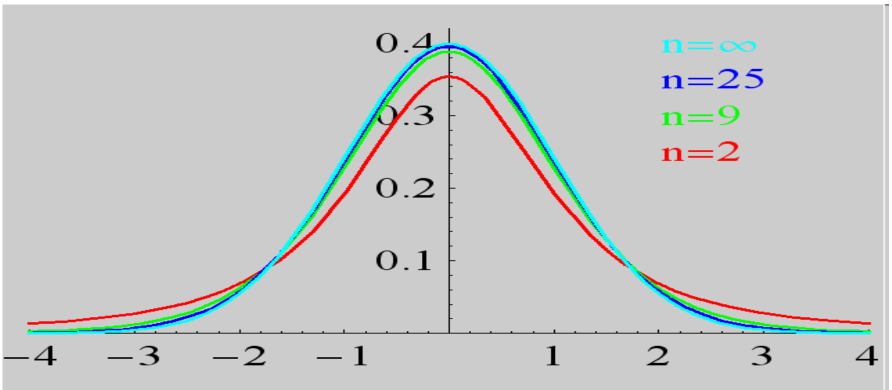
令$u=\frac{t^2+n}{2}z^2, du=z(t^2+n)dz, zdz=\frac{du}{t^2+n}, z^{n-1}=(\frac{2u}{t^2+n})^{\frac{n-1}{2}}$

$$\int_{-\infty}^{+\infty}z^ne^{-\frac{t^2+n}{2}z^2}dz=\int_{-\infty}^{+\infty}(\frac{2u}{t^2+n})^{n-1}e^{-u}\frac{du}{t^2+n}=\int_{-\infty}^{+\infty}(\frac{2u}{t^2+n})^{\frac{n-1}{2}}e^{-u}\frac{du}{t^2+n}$$

$$=\frac{2^{\frac{n-1}{2}}}{(t^2+n)^{\frac{n+1}{2}}}\int_{-\infty}^{+\infty}u^{\frac{n+1}{2}-1}e^{-u}du=\frac{2^{\frac{n-1}{2}}\Gamma(\frac{n+1}{2})}{(t^2+n)^{\frac{n+1}{2}}}$$

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

t 分布的概率密度曲线如图



显然图形是关于$t=0$对称的. 当 n 充分大时，其图形类似于标准正态变量概率密度的图形. 因为$\lim_{n\to\infty}h(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$，所以当$n$足够大时$t$分布近似于$N(0,1)$分布，但对于较小的$n$，$t$分布与$N(0,1)$分布相差很大.

对于给定的α，$0<\alpha<1$，称满足条件

t 分布的分位点，$P\{t>t_\alpha(n)\}=\int_{t_\alpha(n)}^{\infty}h(t)dt=\alpha$

的点$t_\alpha(n)$为$t(n)$分布的上α分位点.

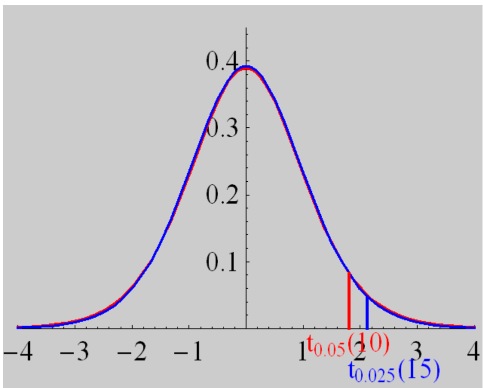
可以通过查表求得上α分位点的值.由分布的对称性知

$t_{1-\alpha}(n)=-t_\alpha(n)$. 当$n>45$时，$t_\alpha(n)\approx z_\alpha$.

例 3 设$T\sim t(n)$，$t(n)$的上α分位点满足

设$T\sim t(n)$，$t(n)$的上α分位点满足

$P\{T>t_\alpha(n)\}=\int_{t_\alpha(n)}^{+\infty}t(y;n)dy=\alpha,$



求 $t_\alpha(n)$ 的值，可通过查表完成.

$t_{0.05}(10)=1.8125,\ t_{0.025}(15)=2.1315.$

定理1：设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本，$\bar{X}, S^2$ 分别是样本均值和样本方差，则有 $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1).$

定理2：设 $X_1, X_2, \cdots, X_n$ 是总体 $N(\mu,\sigma^2)$ 的样本，$\bar{X}, S^2$ 分别是样本均值和样本方差，则有 $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1).$

证明因为 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),\ \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$

且两者独立，由 t 分布的定义知 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\Bigg/\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}} \sim t(n-1).$

二、单个正态分布总体，方差 σ^2 未知，μ 的置信区间

选取枢轴量 $t=\dfrac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$，$P\left\{-t_{1-\frac{\alpha}{2}}(n-1) \le \dfrac{\bar{x}-\mu}{s/\sqrt{n}} \le t_{1-\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$

确定 $t_{1-\frac{\alpha}{2}}(n-1)$ 故 μ 的置信区间为 $\left(\bar{x}-t_{1-\frac{\alpha}{2}}(n-1)\dfrac{S}{\sqrt{n}},\ \bar{x}+t_{1-\frac{\alpha}{2}}(n-1)\dfrac{S}{\sqrt{n}}\right)$

例 1、某工厂生产一批滚珠，其直径 X 服从正态分布 $N(\sigma,\mu 2)$，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，若 $\sigma 2$ 未知，求 μ 的置信度为 0.95 的置信区间

解：$\because \sigma$ 未知，故 μ 的置信区间为 $[\bar{x}-\dfrac{s}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1),\quad \bar{x}+\dfrac{s}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)]$

计算得 $\bar{x}=14.95$，$s=0.226,\ s^2=0.051$

$\because \alpha=0.05,\ n=6,$ 查表得 $t_{1-\alpha/2}(n-1)=t_{0.975}(5)=2.5706,$

所以 μ 的置信区间为 $[\bar{x}-\dfrac{s}{\sqrt{6}}t_{0.975}(5),\quad \bar{x}+\dfrac{s}{\sqrt{6}}t_{0.975}(5)]=[14.71,\quad 15.187]$

例 2、已知某种灯泡的寿命服从正态分布，现从一批灯泡中随机抽取 16 只，测得其使用寿命(小时)如下。建立该批灯泡平均使用寿命 95%的置信区，16 灯泡使用寿命的数据：

1510，1520，1480 ，1500，1450，1480 ，1510，1520，1480，1490，1530，1510，1460，1460，1470，1470

解：已知 $X\sim N(\mu,\ \sigma^2)$，$n=16$, $1-\alpha=95\%$，$t_{\alpha/2}=2.131$ 。根据样本数据计算得：

总体均值 μ 在 $1-\alpha$ 置信水平下的置信区间为

$$\bar{x}\pm t_{\alpha/2}\frac{s}{\sqrt{n}}=1490\pm 2.131\times\frac{24.77}{\sqrt{16}}=1490\pm 13.2=(1476.8,1503.2)$$

即：该种灯泡平均使用寿命的置信区间为 1476.8 小时～1503.2 小时

例 3、一家保险公司收集到由 36 投保个人组成的随机样本，得到每个投保人的年龄(周岁)数据如下表。试建立投保人年龄 90%的置信区间

36 个投保人年龄的数据					
23	35	39	27	36	44
36	42	46	43	31	33
42	53	45	54	47	24
34	28	39	36	44	40
39	49	38	34	48	50
34	39	45	48	45	32

解：已知 n=36, 1-α = 90%，$t_{\alpha/2}$=1.69。根据样本数据计算得：

$$\bar{x}=\frac{\sum x}{n}=39.5, s=\sqrt{\frac{\sum\left(x-\bar{x}\right)^2}{n-1}}=7.77$$

总体均值 μ 在 1-α 置信水平下的置信区间为

$$\bar{x}\pm t_{\alpha/2}\frac{s}{\sqrt{n}}=39.5\pm1.69\times\frac{7.77}{\sqrt{36}}=39.5\pm2.19=\left(37.31,41.69\right)$$

例 4、 假设轮胎的寿命服从正态分布，为估计某种轮胎的平均寿命，现随机地抽取12只轮胎试用，测得它们的寿命（单位：万公里）如下：

4.68 4.85 4.32 4.85 4.61 5.02 5.20 4.60 4.58 4.72 4.38 4.70

试求平均寿命的0.95置信区间。

解 此处正态总体标准差未知，可使用 t 分布求均值的置信区间 本例中经计算有 $\bar{X}=4.709, S^{*2}=0.615$.取 $\alpha=0.05$，查表知 $t_{0.975}(11)=2.2010$，于是平均寿命的0.95置信区间为（单位：万公里）

$$4.7902\pm2.2010\times\frac{\sqrt{0.0615}}{\sqrt{12}}=[4.5516,4.8668]$$

在实际问题中，由于轮胎的寿命越长越好，因此可以只求平均寿命的置信区间，也即构造单边的置信下限。由于

$$\mathbb{P}(\frac{\sqrt{n}(\bar{X}-\mu)}{S^*}<t_{1-\alpha}(n-1))=1-\alpha$$

由不等式变形可知 μ 的 $1-\alpha$ 置信下限为 $\bar{X}-t_{1-\alpha}(n-1)\frac{S^*}{\sqrt{n}}$。将 $t_{0.95}(11)=1.7959$ 带入计算可得平均寿命 μ 的0.95置信下限为4.5806（单位：万公里）。

例 5、从一批灯泡中随机抽取 5 只作寿命试验，测得寿命 X（单位：小时）如下：1050，1100，1120，1250，1280。设灯泡寿命服从正态分布． 求灯泡寿命均值 μ 的置信水平为 0.95 的单侧置信下限.

解： $\boldsymbol{\mu}$ 的点估计取为样本均值 $\overline{x}$ ，方差 $\boldsymbol{\sigma}^2$ 未知，$\dfrac{\overline{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$ 对给定的置信水平

$1-\boldsymbol{\alpha}$ 确定分位点 $t_{1-\alpha}(n-1)$ 使 $\boldsymbol{P}\left(\dfrac{\overline{\boldsymbol{x}}-\boldsymbol{\mu}}{\boldsymbol{s}/\sqrt{\boldsymbol{n}}} \le \boldsymbol{t}_{1-\alpha}(\boldsymbol{n}-1)\right)=1-\boldsymbol{\alpha}$ 即

$P\left(\mu \ge \overline{x}-t_{1-\alpha}(n-1)\dfrac{s}{\sqrt{n}}\right)=1-\alpha$ 即 μ 的置信水平为 **1-** α 的单侧置信下限为

$\overline{x}-t_{1-\alpha}(n-1)\dfrac{s}{\sqrt{n}}$ 由样本值得 $\overline{x}=1060,\ s=99.75$，查表得 $t_{0.95}(4)=2.1318$

$\boldsymbol{\mu}$ 的置信水平为 **0.95** 的单侧置信下限是 **1065** 小时**.**

三、双正态总体的置信区间

求两个正态总体的 $\mu_1-\mu_2$ 的置信区间

1、当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知时

此时有

$$\overline{X}-\overline{Y} \sim N(\mu_1-\mu_2,(\frac{1}{m}+\frac{1}{n})\sigma^2)$$
$$\frac{(m-1)S_X^{*2}+(n-1)S_Y^{*2}}{\sigma^2} \sim \chi^2(m+m-2)$$

由于 $\overline{X},\overline{Y},S_X^{*2},S_Y^{*2}$ 相互独立，故可构造如下服从 t 分布 $t(m+n-2)$ 的枢轴量

$$T=\sqrt{\frac{mn(m+n-2)}{m+n}}\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{(m-1)S_X^{*2}+(n-1)S_Y^{*2}}} \sim t(m+m-2)$$

记 $S_w^2=\dfrac{(m-1)S_X^{*2}+(n-1)S_Y^{*2}}{m+n-2}$，则 $\mu_1-\mu_2$ 的置信区间为

$$[\overline{X}-\overline{Y}-\sqrt{\frac{m+n}{mn}}S_w t_{1-\alpha/2}(m+n-2),\overline{X}-\overline{Y}+\sqrt{\frac{m+n}{mn}}S_w t_{1-\alpha/2}(m+n-2)].$$

2、当 $\sigma_2^2/\sigma_1^2=\theta$ 已知时

此时的处理方法与 2 中无安全类似，只须注意到

$$\overline{X}-\overline{Y} \sim N(\mu_1-\mu_2,(\frac{1}{m}+\frac{1}{n})\sigma^2)=N(\mu_1-\mu_2,\sigma_1^2(\frac{1}{m}+\frac{\theta}{n}))$$
$$\frac{(m-1)S_X^{*2}+(n-1)S_Y^{*2}/\theta}{\sigma_1^2}=\frac{(m-1)S_X^{*2}}{\sigma_1^2}+\frac{(n-1)S_Y^{*2}}{\sigma_2^2} \sim \chi^2(m+m-2)$$

由于 $\overline{X},\overline{Y},S_X^{*2},S_Y^{*2}$ 相互独立，仍可构造如下服从 t 分布 $t(m+n-2)$ 的枢轴量

$$T=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(m-1)S_X^{*2}+(n-1)S_Y^{*2}/\theta}}\sqrt{\frac{mn(m+n-2)}{m\theta+n}}\sim t(m+m-2)$$

记 $S_T^2=\frac{(m-1)S_X^{*2}+(n-1)S_Y^{*2}/\theta}{m+n-2}$，则 $\mu_1-\mu_2$ 的 $1-\alpha$ 置信区间为

$$[\bar{X}-\bar{Y}-\sqrt{\frac{mn}{m\theta+n}}S_T t_{1-\alpha/2}(m+n-2),\bar{X}-\bar{Y}+\sqrt{\frac{mn}{m\theta+n}}S_T t_{1-\alpha/2}(m+n-2)].$$

3、当 m 和 n 都很大时的近似置信区间

此时可以证明有

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_X^{*2}}{m}+\frac{S_Y^{*2}}{n}}}\dot{\sim} N(0,1)$$

由此可以给出 $\mu_1-\mu_2$ 的 $1-\alpha$ 近似置信区间为

$$[\bar{X}-\bar{Y}-u_{1-\alpha/2}\sqrt{\frac{S_X^{*2}}{m}+\frac{S_Y^{*2}}{n}},\bar{X}-\bar{Y}+u_{1-\alpha/2}\sqrt{\frac{S_X^{*2}}{m}+\frac{S_Y^{*2}}{n}}].$$

4、一般情况下的近似置信区间

当 m,n 并不都是很大时，可采用如下的近似方法：令 $S_0^2=\frac{S_X^{*2}}{m}+\frac{S_Y^{*2}}{n}$，取枢轴量

$$T=\frac{[\bar{X}-\bar{Y}-(\mu_1-\mu_2)]}{S_0}$$

此时 T 既不服从 $N(0,1)$ 也不服从 t 分布，但研究表明它与自由度为 l 的 t 分布很接近，其中 l 由公式

$$l=\frac{S_0^4}{\frac{S_X^{*4}}{m^2(m-1)}+\frac{S_Y^{*4}}{n^2(n-1)}}$$

决定，l 一般不为整数 ，可以取与 l 最接近的整数代替之，于是，近似地有 $T\sim t(l)$，从而可得 $\mu_1-\mu_2$ 的 $1-\alpha$ 近似置信区间为

$$[\bar{X}-\bar{Y}-S_0^2 t_{1-\alpha/2}(l),\bar{X}-\bar{Y}+S_0^2 t_{1-\alpha/2}(l)]$$

例 1、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 $x_1,x_2,\cdots,x_{13}$ 与 $y_1,y_2,\cdots,y_{17}$ 已知

$\bar{x}=10.6g$, $\bar{y}=9.5g$，$s_x^2=2.4g^2$，$s_y^2=4.7g^2$ **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，若它们的方差相同，**求均值差 $\mu_1-\mu_2$ 的置信度为 0.95 的置信区间；**

解： $\bar{x}=10.6g$，$\bar{y}=9.5g$，$s_x^2=2.4g^2$，$s_y^2=4.7g^2, m=13, n=17$

因为两总体未知方差相同所以 $\left((\bar{x}-\bar{y})\mp\sqrt{\frac{1}{m}+\frac{1}{n}}s_\omega t_{1-\frac{\alpha}{2}}(m+n-2)\right)$ **的置信区间为**

α $=$**0.05，查表得** $t_{0.975}(28)=2.0484$, 计算得 $s_w=1.9272$

所求置信区间为 $\left((\bar{x}-\bar{y})\mp\sqrt{\frac{1}{m}+\frac{1}{n}}s_\omega t_{1-\frac{\alpha}{2}}(m+n-2)\right)=(-0.3545，\ 2.5545)$

例 2、为比较两个小麦品种的产量，选择18块条件相似的试验田，采用相同的耕作方法做实验，结果播种甲品种的8块试验田的单位面积产量和播种乙品种的10块试验田的单位面积产量（单位：kg）分别为

甲品种：628　583　510　554　612　523　530　615

乙品种：535　433　398　470　567　480　498　560　503　426

假定每个品种的单位面积产量均服从正态分布，试求这两个品种平均单位面积产量差的置信区间（取 $\alpha=0.05$）

解：以 $X_1,\ldots,X_8$ 记为甲品种的单位面积产量，$Y_1,\ldots,Y_{10}$ 记乙品种的单位面积产量，由样本数据可计算得到

$$\bar{X}=569.38, S_X^{*2}=2140.55, m=8$$
$$\bar{Y}=487.00, S_Y^{*2}=3256.22, n=10$$

下面分两种情况讨论

⑴若已知两个品种单位面积产量的标准差相同，即 $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知时，则可采用二样本 t 区间，此处

$$S_w=\sqrt{\frac{(m-1)S_X^{*2}+(n-1)S_Y^{*2}}{m+n-2}}=52.4880,$$
$$t_{1-\alpha/2}(m+n-2)=t_{0.975}(16)=2.1199,$$
$$t_{1-\alpha/2}(m+n-2)S_w\sqrt{\frac{1}{m}+\frac{1}{n}}=52.78,$$

故 $\mu_1-\mu_2$ 的 0.95 置信区间为

$$[29.60, 135.16]$$

⑵若两个品种单位面积产量的方差不等，则可采用近似 t 区间，此处

$$S_0^2=\frac{S_X^{*2}}{m}+\frac{S_Y^{*2}}{n}=589.44, S_0=24.28,$$

$$l=\frac{589.44^2}{\frac{2110.55^2}{8^2\times 7}+\frac{3256.22^2}{10^2\times 11}}=17.74\approx 18,$$

$$S_0t_{.975}(l)=24.28\times 2.1009=51.01,$$

于是 $\mu_1-\mu_2$ 的 0.95 近似置信区间为 $[31.37,133.38]$.

四、假设检验

1.　假定条件

σ^2 未知—正态总体大样本或小样本，非正态总体大样本

2、 使用 t 检验统计量 $t=\dfrac{\overline{x}-\mu_0}{s/\sqrt{n}}\sim t(n-1)$

3、**给定显著性水平α，查表得出相应的临界值 t_α 或 $t_{\alpha/2}$**

4.将检验统计量 z 的值与α水平的临界值进行比较，然后作出决策

双侧检验：当 $|t|>t_{\alpha/2}$ 时，拒绝 H0

左侧检验：当 $t<-t_\alpha$ 时，拒绝 H0

右侧检验：当 $t>t_\alpha$ 时，拒绝 H0

例 1、**一种机床加工的零件尺寸绝对平均误差为 1.35mm。厂家采用一种新的机床进行加工以期进一步降低误差。为检验新机床加工的零件平均误差与旧机床相比是否有显著降低，从某天生产的零件中随机抽取 50 个进行检验。试检验新机床加工的零件尺寸的平均误差与旧机床相比是否有显著降低？ (α=0.01)**

50 个零件尺寸的误差数据 **(mm)**				
1.26	1.19	1.31	0.97	1.81
1.13	0.96	1.06	1.00	0.94
0.98	1.10	1.12	1.03	1.16
1.12	1.12	0.95	1.02	1.13
1.23	0.74	1.50	0.50	0.59
0.99	1.45	1.24	1.01	2.03
1.98	1.97	0.91	1.22	1.06
1.11	1.54	1.08	1.10	1.64
1.70	2.37	1.38	1.60	1.26
1.17	1.12	1.23	0.82	0.86

解:

提出假设

***H*0** ： μ ≥ **1.35**

***H*1** ： μ < **1.35**

根据 ***df*=50-1=49**、 α **=0.01** 查找临界值 $t_{\alpha}(n-1)=-2.405$

检验统计量: $t=\dfrac{\overline{x}-\mu_0}{s/\sqrt{n}}$, $t=\dfrac{1.3152-1.35}{0.365749/\sqrt{50}}=-2.6061$

决策:∵ **t=-2.6061**＜$t_{\alpha}(n-1)=-2.405$ ∴拒绝 ***H*0**

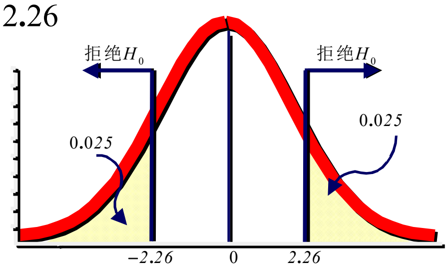
结论:新机床加工的零件尺寸的平均误差与旧机床相比有显著降低

例 2、一种汽车配件的平均长度要求为 **12cm**，高于或低于该标准均被认为是不合格的。汽车生产企业在购进配件时，通常是经过招标，然后对中标的配件提供商提供的样品进行检验，以决定是否购进。现对一个配件提供商提供的 **10** 个汽车配件进行了检验，测得这些汽车配件的平均长度为 **11.89cm,**标准差为 **0.49cm**。假定该供货商生产的配件长度服从正态分布，在 **0.05** 的显著性水平下，检验该供货商提供的配件是否符合要求?

解:

提出假设 **H0** ： μ **=12** **H1** ： μ **≠12**

根据 **df = 10-1=9**、 α **= 0.05** 查找临界值 $t_{\alpha/2}=2.26$



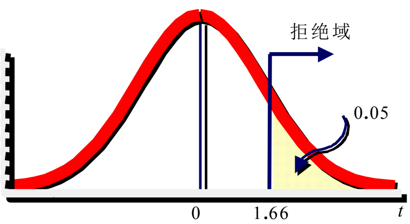
计算检验统计量: $t=\dfrac{11.89-12}{0.49/\sqrt{10}}=-0.70$

决策：因为 $|t|=0.70<t_{\alpha/2}=2.26$

∴不拒绝 **H0** 结论：该供货商提供的零件符合要求

例 3、某企业职工的月均收入服从均值为 **2400** 元，方差未知的正态分布，近来该企业的经济效益有所提高，为调查企业职工的收入是否随之提高了，从该企业随机抽取 **100** 人，得出其月平均收入为 **2450** 元，标准差为 **300** 元，试以**=0.05** 检验该企业职工的收入是否有显著性提高?

解：提出假设 **H0:** μ ≤ **2400** **H1:** μ ＞ **2400**



根据 ***df*=100-1=99**、**α = 0.05** 查找临界值 **zα=1.66**

计算检验统计量: $t=\dfrac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}=\dfrac{2450-2400}{300/\sqrt{100}}=1.67$

决策: ∵**t=1.67**＜ **zα=1.66** ∴不拒绝 ***H*0**

结论: 该企业职工的收入没有显著性提高

注：当 **n** 很大时，**t** 分布与标准正态分布很接近，此时也可采用 **z** 检验法

第四节　F分布及其应用

一、F分布

设$U \sim \chi^2(n_1)$, $V \sim \chi^2(n_2)$, 且U, V独立, 则称随机变量$F = \dfrac{U/n_1}{V/n_2}$服从自由度为(n_1, n_2)的F分布, 记为$F \sim F(n_1, n_2)$.

$F(n_1, n_2)$分布的概率密度为

$$\psi(y) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} y^{\frac{n_1}{2}-1}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)\left[1+\left(\dfrac{n_1 y}{n_2}\right)\right]^{\frac{n_1+n_2}{2}}}, & y > 0, \\ 0, & \text{其他}. \end{cases}$$

证明：先求$Z_1 = \dfrac{U}{n_1}$的密度函数$f(z_1)$，当$z_1 \le 0$时，$f_{Z_1}(z_1) = 0$

当$z_1 > 0$时，$f_U(u) = \dfrac{1}{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})} u^{\frac{n_1}{2}-1} e^{-\frac{u}{2}}$

$f_{Z_1}(z_1) = F_U{}'(n_1 z_1) = n_1 f_U(n_1 z_1)$

所以$f_{Z_1}(z_1) = \dfrac{n_1}{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})}(n_1 z_1)^{\frac{n_1}{2}-1} e^{-\frac{n_1 z_1}{2}}$

再设$Z_2 = \dfrac{V}{n_2}$的密度函数$f(z_1)$，当$z_2 \le 0$时，$f_{Z_2}(z_2) = 0$

当$z_2 > 0$时，同理得$f_{Z_2}(z_2) = \dfrac{n_2}{2^{\frac{n_2}{2}}\Gamma(\frac{n_2}{2})}(n_2 z_2)^{\frac{n_2}{2}-1} e^{-\frac{n_2 z_2}{2}}$

因$F = \dfrac{Z_1}{Z_2}$,设 F 对应的值为 y,则$y = \dfrac{z_1}{z_2}$利用独立随机变量商的密度公式可得

$$f_{Z_2}(z_2)=\frac{n_2}{2^{\frac{n_2}{2}}\Gamma(\frac{n_2}{2})}(n_2 z_2)^{\frac{n_2}{2}-1}e^{-\frac{n_2 z_2}{2}}$$

$$\psi(y)=\int_{-\infty}^{+\infty}|z_2|f_{Z_1}(yz_2)f_{Z_2}(z_2)dz_2$$

$$=\int_0^{+\infty}z_2\frac{n_1}{2^{\frac{n_1}{2}}\Gamma(\frac{n_1}{2})}(n_1 y z_2)^{\frac{n_1}{2}-1}e^{-\frac{n_1 y z_2}{2}}\frac{n_2}{2^{\frac{n_2}{2}}\Gamma(\frac{n_2}{2})}(n_2 z_2)^{\frac{n_2}{2}-1}e^{-\frac{n_2 z_2}{2}}dz_2$$

$$=\frac{n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}y^{\frac{n_1}{2}-1}}{2^{\frac{n_1}{2}}2^{\frac{n_2}{2}}\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}\int_0^{+\infty}z_2^{\frac{n_1+n_2}{2}-1}e^{-\frac{n_1 y+n_2}{2}z_2}dz,$$

令$u=\dfrac{n_1 y+n_2}{2}z_2, dz_2=\dfrac{2du}{n_1 y+n_2}, z_2=\dfrac{2u}{n_1 y+n_2}$,

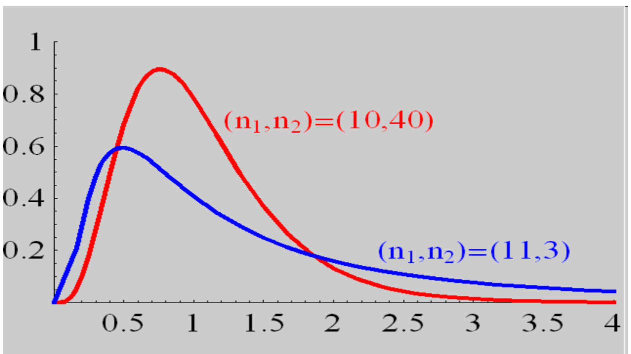
$$\int_0^{+\infty}z_2^{\frac{n_1+n_2}{2}-1}e^{-\frac{n_1 y+n_2}{2}z_2}dz=\int_{-\infty}^{+\infty}(\frac{2u}{n_1 y+n_2})^{\frac{n_1+n_2}{2}-1}e^{-u}\frac{2du}{n_1 y+n_2}$$

$$=\frac{2^{\frac{n_1+n_2}{2}}}{(n_1 y+n_2)^{\frac{n_1+n_2}{2}}}\int_{-\infty}^{+\infty}u^{\frac{n_1+n_2}{2}-1}e^{-u}du=\frac{2^{\frac{n_1+n_2}{2}}\Gamma(\frac{n_1+n_2}{2})}{(n_1 y+n_2)^{\frac{n_1+n_2}{2}}}$$

$$\psi(y)=\frac{n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}y^{\frac{n_1}{2}-1}}{2^{\frac{n_1}{2}}2^{\frac{n_2}{2}}\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})}\frac{2^{\frac{n_1+n_2}{2}}\Gamma(\frac{n_1+n_2}{2})}{(n_1 y+n_2)^{\frac{n_1+n_2}{2}}}$$

$$=\frac{\Gamma(\frac{n_1+n_2}{2})(\frac{n_1}{n_2})^{\frac{n_1}{2}}n_2^{\frac{n_1+n_2}{2}}y^{\frac{n_1}{2}-1}}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})(n_1 y+n_2)^{\frac{n_1+n_2}{2}}}==\frac{\Gamma(\frac{n_1+n_2}{2})(\frac{n_1}{n_2})(\frac{n_1}{n_2}y)^{\frac{n_1}{2}-1}(1+\frac{n_1}{n_2}y)^{\frac{n_1+n_2}{2}}}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})(n_1 y+n_2)^{\frac{n_1+n_2}{2}}}$$

F分布的概率密度曲线如图



根据定义可知，若$F\sim F(n_1,n_2)$,则$\dfrac{1}{F}\sim F(n_2,n_1)$.

F 分布的分位点

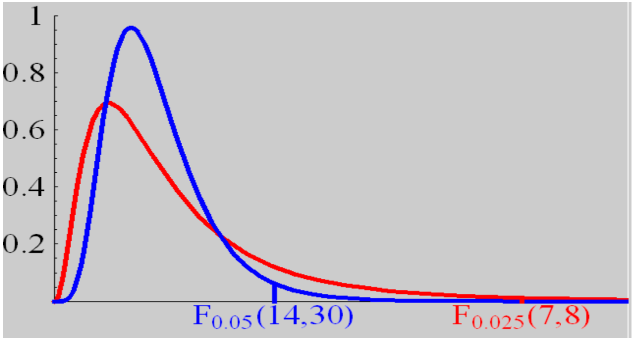
对于给定的α，$0<\alpha<1$，称满足条件

$$P\{F>F_\alpha(n_1,n_2)\}=\int_{F_\alpha(n_1,n_2)}^{+\infty}\psi(y)dy=\alpha$$

的点$F_\alpha(n_1,n_2)$为$F(n_1,n_2)$分布的上α分位点.

例 4 设 $F(n_1,n_2)$分布的上α分位点满足 $P\{F>F_\alpha(n_1,n_2)\}=\int_{F_\alpha(n_1,n_2)}^{+\infty}\psi(y)dy=\alpha,$

求 $F_\alpha(n_1,n_2)$的值,可通过查表完成. $F_{0.025}(7,8)=4.90,\ F_{0.05}(14,30)=2.31$.



F 分布的上 α 分位点具有如下性质: $F_{1-\alpha}(n_1,n_2)=\dfrac{1}{F_\alpha(n_2,n_1)}$.

证明因为$F\sim F(n_1,n_2)$, 所以 $1-\alpha=P\{F>F_{1-\alpha}(n_1,n_2)\}=P\left\{\dfrac{1}{F}<\dfrac{1}{F_{1-\alpha}(n_1,n_2)}\right\}$

$=1-P\left\{\dfrac{1}{F}\ge\dfrac{1}{F_{1-\alpha}(n_1,n_2)}\right\}$ 故 $P\left\{\dfrac{1}{F}>\dfrac{1}{F_{1-\alpha}(n_1,n_2)}\right\}=\alpha$,因为 $\dfrac{1}{F}\sim F(n_2,n_1)$,

所以 $P\left\{\dfrac{1}{F}>F_\alpha(n_2,n_1)\right\}=\alpha$, 比较后得 $\dfrac{1}{F_{1-\alpha}(n_1,n_2)}=F_\alpha(n_2,n_1)$,

即$F_{1-\alpha}(n_1,n_2)=\dfrac{1}{F_\alpha(n_2,n_1)}$. 用来求分布表中未列出的一些上 α 分位点.

例 $F_{0.95}(12,9)=\dfrac{1}{F_{0.05}(9,12)}=\dfrac{1}{0.28}=0.357$.

例证明:

$[t_{1-\frac{\alpha}{2}}(n)]^2=F_\alpha(1,n)$

证: 设 $X\sim T(n)$, $X=G/\sqrt{\dfrac{\chi^2(n)}{n}}$, $G\sim N(0,1)$

令 $Y=X^2=\dfrac{G^2}{\dfrac{\chi^2(n)}{n}}=\dfrac{\dfrac{\chi^2(1)}{1}}{\dfrac{\chi^2(n)}{n}}\sim F(1,n)$

有 $P(|X|>|t_{1-\frac{\alpha}{2}}(n)|)=P(|X|>t_{\frac{\alpha}{2}}(n))=\alpha=P(X^2>t_{\frac{\alpha}{2}}^2(n))=P(Y>t_{1-\frac{\alpha}{2}}^2(n))$

所以 $t_{1-\frac{\alpha}{2}}^2(n)=F_\alpha(1,n)$

推论 2

设 $X_1, X_2, \cdots, X_{n_1}$ 与 $Y_1, Y_2, \cdots, Y_{n_2}$ 分别是具有相同方差的两正态总体 $N(\mu_1,\sigma^2)$, $N(\mu_2,\sigma^2)$ 的样本, 且这两个样本互相独立, 设 $\bar{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i$, $\bar{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i$ 分别是这两个样本的均值,

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

分别是这两个样本的方差, 则有

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

(2) 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时,

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

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

证明(1) 由定理二

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

由假设 S_1^2, S_2^2 独立, 则由 F 分布的定义知

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

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

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

由 $\dfrac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1)$, $\dfrac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1)$,

且它们相互独立, 故由 χ^2 分布的可加性知

$$V=\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2),$$

由于 U 与 V 相互独立, 按 t 分布的定义.

$$\frac{U}{\sqrt{V/(n_1+n_2-2)}}=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2).$$

定理3.

设 $X_1,X_2,\cdots,X_{n_1}$ 与 $Y_1,Y_2,\cdots Y_{n_2}$ 分别是具有相同方差的两个正态总体 $N(\mu_1,\sigma^2),N(\mu_2,\sigma^2)$ 的样本，且它们独立。

设 $\bar{X}=\dfrac{1}{n_1}\sum_{i=1}^{n_1}X_i,\bar{Y}=\dfrac{1}{n_2}\sum_{j=1}^{n_2}Y_j$ 分别是两个样本的均值。$S_1{}^2=\dfrac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2$

$S_2{}^2=\dfrac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_j-\bar{Y})^2$ 分别是两个样本的方差；

则有：$\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$

证：$\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2})$ 所以 $\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,1)$

$\dfrac{(n_1-1)S_1^2}{\sigma^2}\sim\chi^2(n_1-1),\dfrac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_2-1),$

且它们独立。则 $\dfrac{(n_1-1)S_1^2}{\sigma^2}+\dfrac{(n_2-1)S_2^2}{\sigma^2}\sim\chi^2(n_1+n_2-2)$。

由 $t-$ 分布的定义：

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\Bigg/\sqrt{\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}/(n_1+n_2-2)}\sim t(n_1+n_2-2)$$

即：$\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$

设 $X_1,X_2,\cdots,X_n$ 与 $Y_1,Y_2,\cdots,Y_m$ 分别是来自正态总体 $X\sim N(\mu_1,\sigma^2)$ 与 $Y\sim N(\mu_2,\sigma^2)$ 的相互独立的简单随机样本。则 $\overline{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i\sim N(\mu_1,\dfrac{\sigma^2}{n})$ $\quad\overline{Y}=\dfrac{1}{m}\sum_{j=1}^{m}Y_j\sim N(\mu_2,\dfrac{\sigma^2}{m})$

故 $\overline{X}-\overline{Y}\sim N(\mu_1-\mu_2,\dfrac{\sigma^2}{n}+\dfrac{\sigma^2}{m})$，$\dfrac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma^2}{n}+\dfrac{\sigma^2}{m}}}\sim N(0,1)$

因 $\dfrac{(n-1)S_1^2}{\sigma^2}\sim\chi^2(n-1)\quad\dfrac{(m-1)S_2^2}{\sigma^2}\sim\chi^2(m-1)$

故 $\dfrac{(n-1)S_1^2}{\sigma^2}+\dfrac{(m-1)S_2^2}{\sigma^2}\sim\chi^2(n+m-2)$

$\overline{X}-\overline{Y}$ 与 $\dfrac{(n-1)S_1^2}{\sigma^2}+\dfrac{(m-1)S_2^2}{\sigma^2}$ 相互独立，于是

$$\frac{\dfrac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma^2}{n}+\dfrac{\sigma^2}{m}}}}{\sqrt{\dfrac{\dfrac{(n-1)S_1^2}{\sigma^2}+\dfrac{(m-1)S_2^2}{\sigma^2}}{n+m-2}}}=\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}\sqrt{\dfrac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}}\sim t(n+m-2)$$

例1、随机抽取机器A生产的钢管18只，测得
样本方差$s_1^2=0.34(mm^2)$;抽取机器B生产的钢管13只，测得样本方差$s_2^2=0.29(mm^2)$.设两样本相互独立,且设由机器A,B生产的钢管内径分别服从正态分布$N(\mu_1,\sigma_1^2),N(\mu_2,\sigma_2^2)$,这里$\mu_i,\sigma_i^2(i=1,2)$均未知.试求方差比$\sigma_1^2/\sigma_2^2$的置信水平为0.90的置信区间.

解： $n_1=18,\ s_1^2=0.34\quad n_2=13,\ s_2^2=0.29$

$1-\alpha=0.9\quad \alpha=0.1$，$F_{\frac{\alpha}{2}}(n_1-1,n_2-1)=F_{0.05}(17,12)=2.59$

$\dfrac{\sigma_1^2}{\sigma_2^2}$ **置信水平为0.9的置信区间：** $\left(\dfrac{S_1^2}{S_2^2}\dfrac{1}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)},\dfrac{S_1^2}{S_2^2}\dfrac{1}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}\right)$

$$=\left(\frac{0.34}{0.29}\times\frac{1}{2.59},\frac{0.34}{0.29}\times 2.38\right)=(0.45,2.79)$$

例 1、设 $r.v.\ X$ 与 Y 相互独立，$X\sim N(0,16),\ Y\sim N(0,9)$，$X_1, X_2,\dots, X_9$ 与 $Y_1, Y_2,\dots, Y_{16}$ 分别是取自 X 与 Y 的简单随机样本，求统计量 $Z=\dfrac{X_1+X_2+\cdots+X_9}{\sqrt{Y_1^2+Y_2^2+\cdots+Y_{16}^2}}$ 所服从的分布。

解 $X_1+X_2+\cdots+X_9\sim N(0,9\times16)$ $\dfrac{1}{3\times4}(X_1+X_2+\cdots+X_9)\sim N(0,1)$

$\dfrac{1}{3}Y_i\sim N(0,1)$ ，$i=1,2,\cdots,16$，$\displaystyle\sum_{i=1}^{16}\left(\frac{1}{3}Y_i\right)^2\sim\chi^2(16)$

从而 $$\frac{X_1+X_2+\cdots+X_9}{\sqrt{Y_1^2+Y_2^2+\cdots+Y_{16}^2}}=\frac{\frac{1}{3\times4}\left(X_1+X_2+\cdots+X_9\right)}{\sqrt{\dfrac{\sum_{i=1}^{16}\left(\frac{1}{3}Y_i\right)^2}{16}}}\sim t(16)$$

例 2、 设总体 $X\sim N(0,1),X_1,\cdots,X_6$ 为总体 X 的样本,
$Y=(X_1+X_2+X_3)^2+(X_4+X_5+X_6)^2$
试确定常数 c，使 cY 服从 χ^2 分布。

解 $X_1+X_2+X_3\sim N(0,3)$， $X_4+X_5+X_6\sim N(0,3)$

$\dfrac{1}{\sqrt3}\left(X_1+X_2+X_3\right),\dfrac{1}{\sqrt3}\left(X_4+X_5+X_6\right)\sim N(0,1)$

故 $\left[\dfrac{1}{\sqrt3}\left(X_1+X_2+X_3\right)\right]^2+\left[\dfrac{1}{\sqrt3}\left(X_4+X_5+X_6\right)\right]^2=\dfrac{1}{3}Y\sim\chi^2(2)$

因此$c=1/3$.

例 3、 设$(X_1,X_2,\cdots,X_n)$是来自 $N(\mu,\sigma 2)$的简单随机样本，$\bar X$ 是样本均值

$$S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar X)^2,\ S_2^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar X)^2,$$

$$S_3^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2,\ S_4^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2,$$

则服从自由度为 n - 1 的 t 分布的随机变量为

$(A)\ \dfrac{\bar{X}-\mu}{S_1}\sqrt{n-1}$ $(B)\ \dfrac{\bar{X}-\mu}{S_2}\sqrt{n-1}$ $(C)\ \dfrac{\bar{X}-\mu}{S_3}\sqrt{n}$ $(D)\ \dfrac{\bar{X}-\mu}{S_4}\sqrt{n}$

解 $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$，$\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2(n-1)$

$$\frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}} = \frac{\sqrt{n(n-1)}(\bar{X}-\mu)}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}} \sim t(n-1)$$ 故应选 (B)

例 4、他能根据股票价格的历史图表预报未来股市的涨跌，若在一场测试中，他工作了 10 次预测，报对了 8 次

(1) 在显著性水平α=0.05 下，能否相信他具有这种能力？

(2) 对什么样的显著性水平，可相信他具有这种能力？

解：我们先对问题做一个简单的分析：若该人有预报能力，则他预报正确的概率因该大于 1/2,若他没有这种能力，则他胡乱猜测也有 50%的可能，现以 X 表示在 10 次预测中预测正确的次数，则 $X\sim B(10,p)$,要检验的一对假设为：

$$H_0: p=0.5 \qquad \text{vs} \qquad H_1: p>0.5$$

若拒绝原假设，则可相信他有这个预报能力，否则则认为他没有预报能力。由于检验的拒绝域形如 $W=\{x\ge c\}$，故检验的 p 值为

$$\begin{aligned} p &= P_{0.5}(X\ge 8) = P_{0.5}(X=8) - P_{0.5}(X=9) - P_{0.5}(X=10) \\ &= 45\times 0.5^8\times 0.5^2 - 10\times 0.5^9\times 0.5 - 0.5^{10} = 0.0547 > 0.05 = \alpha \end{aligned}$$

现对 p 值做一些讨论：

(1)由于检验的 p 值大于显著性水平 0.05,故不应拒绝原假设，不能相信他有预报能力。这是可能犯第二类错误，其概率为 $\beta(\theta)=1-P_\theta(X\ge 8)$，对具体的$\theta$，可以计算出$\beta(\theta)$的值，如$\theta=0.6$，则

$$\begin{aligned} \beta(0.6) &= 1-P_{0.6}(X\ge 8) = 1-P_{0.6}(X=8)-P_{0.6}(X=9)-P_{0.6}(X=10) \\ &= 1-45\times 0.6^8\times 0.4^2 - 10\times 0.6^9\times 0.4 - 0.6^{10} = 0.8327 \end{aligned}$$

类似地可算出 $\beta(0.7)=0.6172, \beta(0.8)=0.3222, \beta(0.9)=0.0702,$

可见随着θ 的增加，犯第二类错误的概率在减小。，

(2)我们知道，当 $p<\alpha$ 时应拒绝原假设，因此当$\alpha>0.0547$ 时拒绝原假设，譬如，若取 $\alpha=0.06$，因为 $p=0.0547<0.06=\alpha$,则拒绝原假设，可相信他有这种能力。

9.设 $x_1,\cdots,x_n$ 是来自指数分布 E$xp(\lambda)$的一个样本，$y_1,\cdots,y_m$ 是来自另一指数分布 E$xp(\mu)$的一个样本，且两样本相互独立.若设 $\Delta=\lambda/\mu$，对如下检验问题：

$$H_0:\Delta=1 \quad \text{vs} \quad H_1:\Delta\neq 1$$

在显著性水平为 α 的场合给出拒绝域.

解：由于指数分布是特殊的伽玛分布，具体是 E$xp(\lambda)$=G$a(1,\lambda)$,于是

$$\sum_{i=1}^{n}x_i\sim \mathrm{G}a(n,\lambda),2n\lambda\sum_{i=1}^{n}x_i\sim \mathrm{G}a(n,\frac{1}{2})=\chi^2(2n).$$

同理可得 $2m\mu\bar{y}\sim\chi^2(2m)$,由两样本相互独立可知 $\dfrac{\lambda\bar{x}}{\mu\bar{y}}=\dfrac{2n\lambda\bar{x}/n}{2m\mu\bar{y}/m}\sim F(2n,2m)$，

在原假设 $H_0:\Delta=1$ 成立下，有 $\lambda=\mu$，从而有 $\dfrac{\bar{x}}{\bar{y}}\sim F(2n,2m)$,检验拒绝域 $w=\{\bar{x}/\bar{y}\le F_{\alpha/2}(2n,2m)\}\cup\{\bar{x}/\bar{y}\ge F_{1-\alpha/2}(2n,2m)\}$.譬如，若两样本量与样本均值分别为

$$n=5,\qquad \bar{x}=1600,\qquad m=6,\qquad \bar{y}=1200.$$

在给定显著性水平 α =0.05，可查表得

$$F_{0.975}(10，12)=3.37，F_{0.025}(10，12)=\frac{1}{F_{0.975}(12,10)}=\frac{1}{3.62}=0.276.$$

从而得拒绝域 w= $\{\bar{x}/\bar{y}\le 0.276$ 或 $\bar{x}/\bar{y}\ge 3.37\}$，如今 $\dfrac{\bar{x}}{\bar{y}}=\dfrac{1600}{1200}=1.333$, 它不在拒绝域内，故不能拒绝原假设.

第五节　小结及例题

一、　三大抽样分布的分布函数

综 述：a) 根据大数定理和中心极限定理，但样本容量 n 较大时（数学上一般要求 $n>45$），任

何分布都依概率收敛于正态分布 $N\left(\mu,\ \sigma^2\right)$，并可标准化为 $N\left(0,\ 1\right)$。

b) 现实世界和工程技术中的任何数据样本流到目前为止，不外乎 $N\left(0,\ 1\right)$ 的函数分布，

集中表现为 3 大抽样分布规律。

c) 考研数学中规定：$N(0,\ 1)$的分位数定义为下分位数（从图形上看为左边面积），3

大抽样分布的分位数定义都为上分位数（从图形上看为右边面积）

1. $\chi^2(n)$分布（分布函数不要求掌握）

量纲模型：$\boxed{\sum_{i=1}^{n}\left[N_i(0,\ 1)\right]^2 \sim \chi^2(n)}$

性 质：

(1) $\{X_i\}$独立同，$\boxed{X_i \sim N(0,1) \Rightarrow \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)}$

(2) 可加性 $\boxed{\chi_1^2+\chi_2^2+\cdots \sim \chi^2(n_1+n_2+\cdots)}$

(3) $\boxed{E\left[\chi^2(n)\right]=n; \quad D\left[\chi^2(n)\right]=2n}$

证 明(3)：由于$X_i \sim N(0,1) \Rightarrow E(X_i)=0;\ D(X_i)=1$

$$E\left(X_i^2\right)=E\left[X_i-E\left(X_i\right)\right]^2=D\left(X_i\right)=1 \quad (i=1,2,\cdots,n)$$

$$E\left(X_i^4\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^4 e^{-\frac{x^2}{2}}dx=3$$

$$D\left(X_i^2\right)=E\left(X_i^4\right)-\left[E\left(X_i^2\right)\right]^2=3-1=2$$

$$E\left(\chi^2(n)\right)=E\left(\sum_{i=1}^{n} X_i^2\right)=\sum_{i=1}^{n} E\left(X_i^2\right)=n$$

$$D\left(\chi^2(n)\right)=D\left(\sum_{i=1}^{n} X_i^2\right)=\sum_{i=1}^{n} D\left(X_i^2\right)=2n$$

评 注 样本函数中的必需记住的数字特征

$$E\left(\overline{X}\right)=\mu; \qquad D\left(\overline{X}\right)=\frac{\sigma^2}{n}$$

$$E\left(S^2\right)=\sigma^2; \qquad D\left(S^2\right)=\frac{2\sigma^4}{n-1}$$

$$E\left(B_2\right)=E\left(S^{*2}\right)=E\left(\frac{n-1}{n}S^2\right)=\frac{n-1}{n}\sigma^2;$$

$$D\left(B_2\right)=D\left(\frac{n-1}{n}S^2\right)=\left(\frac{n-1}{n}\right)^2\frac{2\sigma^4}{n-1}=\frac{2(n-1)\sigma^4}{n}$$

(4) 上分位点 α 定义为 $\chi^2(n)$ 分布的分位数

2． $t(n)$ 分布（分布函数不要求掌握）

$\{X_i\}$ 独立同分布　$X_i \sim N(0,1),\ Y \sim \chi^2(n);\ X$ 和 Y 独立

量纲模型：$\boxed{T = \dfrac{X}{\sqrt{\dfrac{Y}{n}}} \sim t(n)}$

性 质：

(1) t 分布密度函数 $n \to \infty \Rightarrow f_{t(n)}(x) \sim N(0,1)$

(2) 上分位点 α 定义为 $t(n)$ 分布的分位数 $\boxed{P\{t(n) \ge t_\alpha(n)\} = \int_{t_\alpha(n)}^{+\infty} f_t(x)dx = \alpha}$

(3) $EX = 0,\quad DX = \dfrac{n}{n-2}(n > 2)$

(4) 性质　T 分布具有对称性，　$t_{1-\alpha}(n) = -t_\alpha(n);\ n > 45$ 时，$t_\alpha(n) \approx Z_\alpha$

3． $F(m,\ n)$ 分布（分布函数不要求掌握）

X、Y 相互独立，$X \sim \chi(m);\ Y \sim \chi^2(n)$；量纲模型：$\boxed{F = \dfrac{\dfrac{X}{m}}{\dfrac{Y}{n}} \sim F(m,\ n)}$

$\boxed{\text{评 注}}$ 特别地，但 $\boxed{m = n = 1 \Rightarrow F(1,\ 1) = \dfrac{1}{\pi(1+x)\sqrt{x}}}$。

例：假定 $(X_1,\ X_2)$ 来自正态整体 $X \sim N(0,\ \sigma^2)$ 的一个样本，求 $P\left[\dfrac{(X_1+X_2)^2}{(X_1-X_2)^2} < 4\right]$。

解：$X_i \sim N(0,\ \sigma^2) \Rightarrow X_1+X_2 \sim N(0,\ 2\sigma^2);\ X_1-X_2 \sim N(0,\ 2\sigma^2)$

$$\frac{X_1+X_2}{\sqrt{2}\sigma} \sim N(0,\ 1);\ \frac{X_1-X_2}{\sqrt{2}\sigma} \sim N(0,\ 1) \Rightarrow \left(\frac{X_1+X_2}{\sqrt{2}\sigma}\right)^2 \sim \chi^2(1);\ \left(\frac{X_1-X_2}{\sqrt{2}\sigma}\right)^2 \sim \chi^2(1)$$

$$\Rightarrow \frac{(X_1+X_2)^2}{(X_1-X_2)^2} = \frac{\left(\frac{X_1+X_2}{\sqrt{2}\sigma}\right)^2 \Big/ 1}{\left(\frac{X_1-X_2}{\sqrt{2}\sigma}\right)^2 \Big/ 1} \sim F(1,\ 1) = \frac{1}{\pi(1+x)\sqrt{x}}$$

$$\Rightarrow P\left[\frac{(X_1+X_2)^2}{(X_1-X_2)^2} < 4\right] = \int_0^4 \frac{1}{\pi(1+x)\sqrt{x}}dx = \frac{2}{\pi}\arctan 2.$$

① 上 分 位 点 α 定 义 为 $t(n)$ 分 布 的 分 位 数

$$\boxed{P\{F \geq F_\alpha(n,m)\} = \int_{F_\alpha(n,m)}^{+\infty} f_F(x)dx = \alpha}$$

② 性 质 $\boxed{F_{1-\alpha}(n,\ m) = \frac{1}{F_\alpha(m,\ n)}}$ $\boxed{F(n,\ m) = \frac{1}{F(m,\ n)}}$

$\boxed{X \sim t(n) \Rightarrow X^2 \sim F(1,\ n)}$ $\boxed{X \sim F(n,\ m) \Rightarrow \frac{1}{X} \sim F(m,\ n)}$

● 证明结论 $t^2(n) \sim F(1,\ n)$

$$U \sim N(0,1), \qquad V \sim \chi^2(n)$$

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

$$T^2 = \frac{U^2}{V/n};\ \text{而}\ U^2 \sim \chi^2(1)\ \text{时} \Rightarrow T^2 \sim F(1,\ n) \Rightarrow t^2(n) \sim F(1,\ n)$$

● 证明结论 $F_{1-\alpha}(n,\ m) = \frac{1}{F_\alpha(m,\ n)}$ 如下

$$P\left\{X \ge F_{1-\alpha}(m,\ n)\right\} = 1-\alpha \Rightarrow P\left\{\frac{1}{X} \le \frac{1}{F_{1-\alpha}(m,\ n)}\right\} = 1-\alpha$$

$$\Rightarrow P\left\{\frac{1}{X} > \frac{1}{F_{1-\alpha}(m,\ n)}\right\} = \alpha$$

又根据分位数的定义，$\xrightarrow{X\sim F(m,\ n)Y=\frac{1}{X}\sim F(n,\ m)} P\left\{\frac{1}{X} \ge F_{\alpha}(n,\ m)\right\} = \alpha$

而连续分布对一点的概率取值为零，则

$$P\left\{\frac{1}{X} > \frac{1}{F_{1-\alpha}(m,\ n)}\right\} = P\left\{\frac{1}{X} \ge F_{\alpha}(n,\ m)\right\} \Rightarrow F_{\alpha}(n,\ m) = \frac{1}{F_{1-\alpha}(m,\ n)}$$

二、数理统计中 8 大样本函数的分布（枢轴量）的详细证明

1. 单个正态总体

设 $\{X_n\} \sim N(\mu,\ \sigma^2)$ 为一系列简单随机样本，则有

(1) 若 σ 已知，需要估计 μ 的范围，则使用枢轴量

$$\boxed{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})\text{；} \qquad \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}$$

证明一：

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

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot n \cdot \mu = \mu$$

$$D(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

证明二：

$$E\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = \frac{\sqrt{n}}{\sigma} E(\overline{X}-\mu) = 0$$

$$D\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)=\frac{n}{\sigma^2}D(\overline{X}-\mu)=\frac{n}{\sigma^2}[E(\overline{X}-\mu)^2-E^2(\overline{X}-\mu)]$$

$$=\frac{n}{\sigma^2}[E\left(\overline{X}^2+\mu^2-2\mu\overline{X}\right)-0]$$

$$=\frac{n}{\sigma^2}[E(\overline{X}^2)+\mu^2-2\mu^2]=\frac{n}{\sigma^2}(E(\overline{X}^2)+\mu^2-2\mu^2)$$

$$=\frac{n}{\sigma^2}[(\mu^2+\frac{\sigma^2}{n})-\mu^2]=1$$

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

评 注　公式 ① 是标准化随机变量的手段，也是确定复合随机变量分布的基础。

(2) 若 μ 未知，需要估计 σ 的范围，则使用枢轴量

$\boxed{\frac{(n-1)S^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\sim\chi^2(n-1)\text{；且 }\overline{X}\text{ 与 }S^2\text{ 独立}}$ （$\overline{X}$ 是随机变量）

证 明：

已知 $X_i\left(i=1,2,\cdots,n\right)\sim N\left(\mu,\ \sigma^2\right)$，且相互独立，

令 $Y_i=\frac{X_i-\mu}{\sigma}\left(i=1,2,\cdots,n\right)\Rightarrow Y_1,Y_2,\cdots,Y_n\sim N\left(0,1\right)$，且相互独立。

作下列正交变换：

$$\begin{pmatrix}Z_1\\Z_2\\\vdots\\Z_n\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}}\\ c_{21} & c_{22} & \cdots & c_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ c_{n1} & c_{n2} & \cdots & c_{nn}\end{pmatrix}\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix}$$

正交变换不改变向量组的秩，由于 $Y_1,Y_2,\cdots,Y_n$ 相互独立，则 $Z_1,Z_2,\cdots,Z_n$ 相互独立，

且都服从 $N\left(0,1\right)$。

记 $$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i=\frac{1}{n}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}=\frac{1}{\sigma}\cdot\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\frac{\mu}{\sigma}=\frac{\overline{X}}{\sigma}-\frac{1}{n}\cdot n\cdot\frac{\mu}{\sigma}=\frac{\overline{X}-\mu}{\sigma}$$

由上述变换矩阵等式易得：$Z_1=\frac{1}{\sqrt{n}}Y_1+\frac{1}{\sqrt{n}}Y_2+\cdots+\frac{1}{\sqrt{n}}Y_n=\sqrt{n}\overline{Y}\sim N\left(0,1\right)$

正交变换不改变向量的长度$\Rightarrow \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z_i^2$，所以

$$\begin{aligned}
\frac{(n-1)S^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma}-\frac{\overline{X}-\mu}{\sigma})^2 \\
&= \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 + \sum_{i=1}^{n}(\frac{\overline{X}-\mu}{\sigma})^2 - 2\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)\left(\frac{\overline{X}-\mu}{\sigma}\right) \\
&= \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 + n(\frac{\overline{X}-\mu}{\sigma})^2 - 2n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 - n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 \\
&= \sum_{i=1}^{n}Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^{n}Z_i^2 - Z_1^2 = \sum_{i=2}^{n}Z_i^2 \sim \chi^2(n-1)
\end{aligned}$$

$\boxed{\text{评　注}}$　$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 \sim \chi^2(n-1)$ 有重要的应用价值，如计算 $E\left(S^2\right)$；$D\left(S^2\right)$。

$$\because E\left(\chi^2(n-1)\right) = n-1,\quad D\left(\chi^2(n-1)\right) = 2(n-1)$$

$$\Rightarrow E\left(S^2\right) = \frac{\sigma^2}{n-1}E\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{\sigma^2}{n-1}E\left(\chi^2(n-1)\right) = \frac{\sigma^2}{n-1}\cdot(n-1) = \sigma^2$$

$$D\left(S^2\right) = \left[\frac{\sigma^2}{n-1}\right]^2 D\left(\frac{(n-1)S^2}{\sigma^2}\right) = \left[\frac{(n-1)}{\sigma^2}\right]^2 \cdot D\left(\chi^2(n-1)\right) = \left[\frac{\sigma^2}{n-1}\right]^2 \cdot 2(n-1) = \frac{2\sigma^4}{n-1}$$

(3) 若 σ 未知，需要估计 μ 的范围，则使用枢轴量

$$\boxed{\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)}$$

证明：

$$\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\frac{\overline{X}-\mu}{\sigma}}{\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1)$$

(4) 若 μ 已知，需要估计 σ 的范围，则使用枢轴量

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

证明：

$$Y_i=\frac{X_i-\mu}{\sigma}\sim N(0,1)$$

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

2. 两个正态总体 （X和Y独立同分布）

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

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

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

则有：

(5) 若σ_1^2，σ_2^2已知，需要估计$\mu_1-\mu_2$的范围，则使用枢轴量

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

证明：

$$\overline{X}-\overline{Y}-(\mu_1-\mu_2)\sim N\left[0,\ \frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}\right]\Rightarrow\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}\sim N(0,1)$$

(6)若σ_1^2，σ_2^2未知，但$\sigma_1=\sigma_2=\sigma$时，需要估计$\mu_1-\mu_2$的范围，则使用枢轴量

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

其中： $S_W=\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}$

证明：

$$\overline{X}-\overline{Y}-(\mu_1-\mu_2)\sim N\left[0,\ \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right]$$

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n}+\frac{1}{m}}}=\frac{N\left[0,\ \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right]}{\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}\cdot\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

$$=\frac{\frac{N\left[0,\ \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)\right]}{\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}}}=\frac{\sigma\cdot N(0,\ 1)}{\sqrt{\frac{\chi^2(n-1)\sigma^2+\chi^2(m-1)\sigma^2}{n+m-2}}}$$

$$=\frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)+\chi^2(m-1)}{n+m-2}}}=\frac{N(0,1)}{\sqrt{\frac{\chi^2(n+m-2)}{n+m-2}}}\sim t(n+m-2)$$

(7) 如 $\mu_1,\ \mu_2$ 已知，需要估计 $\frac{\sigma_1^2}{\sigma_2^2}$ 的范围，则使用枢轴量

$$\boxed{\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_1)^2}{\frac{1}{m}\sum_{i=1}^{m}(Y_i-\mu_2)^2}\frac{\sigma_2^2}{\sigma_1^2}\sim F(n,m)}$$

证明：

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_1)^2}{\frac{1}{m}\sum_{i=1}^{m}(Y_i-\mu_2)^2}\frac{\sigma_2^2}{\sigma_1^2}=\frac{\frac{\sum_{i=1}^{n}(X_i-\mu_1)^2}{n\sigma_1^2}}{\frac{\sum_{i=1}^{n}(Y_i-\mu_2)^2}{m\sigma_2^2}}=\frac{\frac{\chi^2(n)}{n}}{\frac{\chi^2(m)}{m}}\sim F(n,m)$$

根据 F 分布的意义，可以推知 $\begin{cases}\sum_{i=1}^{n}(X_i-\mu_1)^2/\sigma_1^2\sim\chi^2(n)\\ \sum_{i=1}^{m}(Y_i-\mu_2)^2/\sigma_2^2\sim\chi^2(m)\end{cases}$

(8) 如 $\mu_1,\ \mu_2$ 未知，需要估计 $\frac{\sigma_1^2}{\sigma_2^2}$ 的范围，则使用枢轴量

$$\boxed{\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}\sim F(n-1,m-1)}$$

证明：

$$\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}=\frac{\frac{(n-1)S_1^2}{(n-1)\sigma_1^2}}{\frac{(m-1)S_2^2}{(m-1)\sigma_2^2}}=\frac{\frac{\chi^2(n-1)}{n-1}}{\frac{\chi^2(m-1)}{m-1}}\sim F(n-1,m-1)$$

三、先进题型与求解秘技

$\boxed{\text{陈氏密技}}$ 量纲法求复合统计量的抽样分布。

3 种抽样源正态；量纲法则判类型。

根据定义凑模式；标准变量容量值。

【例 1】设 x_1, x_2, x_3, x_4 来自正态总体 $N\left(0,\ 2^2\right)$ 的简单随机样本，求 a, b，使得

$$X=a\left(x_1-2x_2\right)^2+b\left(3x_3-4x_4\right)^2\sim\chi^2\text{。}$$

解：

$$X=a\left(x_1-2x_2\right)^2+b\left(3x_3-4x_4\right)^2=\left[\sqrt{a}\left(x_1-2x_2\right)\right]^2+\left[\sqrt{b}\left(3x_3-4x_4\right)\right]^2$$

$$\sqrt{a}\left(x_1-2x_2\right)\sim N\left[0,\ \left(\sqrt{a}\cdot 2\right)^2+\left(2\sqrt{a}\cdot 2\right)^2\right]=N\left[0,\ 20a\right]\Rightarrow a=\frac{1}{20}$$

$$\sqrt{b}\left(3x_3-4x_4\right)\sim N\left[0,\ \left(3\sqrt{b}\cdot 2\right)^2+\left(4\sqrt{a}\cdot 2\right)^2\right]=N\left[0,\ 100a\right]\Rightarrow b=\frac{1}{100}$$

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

【例 2】$\{N_i\}\sim N(0,2^2)$，

$$Q=aX_1^2+b(X_2+X_3)^2+c(X_4+X_5+X_6)^2+d(X_7+X_8+X_9+X_{10})\text{ 服从 }\chi^2$$

分布，求 a, b, c, d 和自由度 m。

解： $X_1\sim N(0,2^2)\Rightarrow\frac{1}{\sigma}X_1=\frac{1}{2}X_1\sim N(0,1)\Rightarrow\frac{1}{4}X_1^2\sim x^2(1)$

同理 $X_2+X_3\sim N(0,8), X_1+X_5+X_6\sim N(0,12), X_7+X_8+X_9+X_{10}\sim N(0,16)$

$$\frac{1}{8}(X_2+X_3)^2\sim\chi^2(1)$$

$$\frac{1}{12}(X_4+X_5+X_6)\sim\chi^2(1)$$

$$\frac{1}{16}(X_7+X_8+X+X_{10}\ )\sim\chi^2(1)$$

由$\chi^2(n)$的可加性知　　$Q\sim\chi^2(4)$

所以　$a=\frac{1}{4}$　$b=\frac{1}{8}$　$c=\frac{1}{12}$　$d=\frac{1}{16}$　$m=4$

【例 3】设X, Y相互独立，都服从$N\left(0,\ 3^2\right)$，则统计量$U=\frac{x_1+x_2+\cdots+x_9}{\sqrt{y_{_1}^2+y_{_2}^2+\cdots+y_{_9}^2}}$服从什么分布。

解：　U的分子是N分布，分母是$\sqrt{\chi^2}$分布，则U必是T分布。

根据T分布定义，需要把分子和分母标准化$N\left(0,\ 1\right)$，这需要利用公式①

$$\overline{X}=\frac{\sum_{i=1}^{n}X_{_i}}{n}\sim N\left(\mu,\ \frac{\sigma^2}{n}\right)\Rightarrow\overline{X}=\frac{\sum_{i=1}^{9}X_{_i}}{9}\sim N\left(0,\ 1\right)$$

$$\frac{Y_i}{3}\sim N\left(0,\ 1\right)\Rightarrow\frac{\sum_{i=1}^{9}Y_{_i}^2}{9}=\sum_{i=1}^{9}\left(\frac{Y_i}{3}\right)^2\sim\chi^2\left(9\right)$$

$$U=\frac{x_1+x_2+\cdots+x_9}{\sqrt{y_{_1}^2+y_{_2}^2+\cdots+y_{_9}^2}}=\frac{\dfrac{\sum_{i=1}^{9}X_{_i}}{9}}{\dfrac{1}{3}\times\sqrt{\sum_{i=1}^{9}\left(\dfrac{Y_i}{3}\right)^2}}\sim\frac{N\left(0,\ 1\right)}{\sqrt{\dfrac{\chi^2\left(9\right)}{9}}}=t\left(9\right)$$

【例 4】设$X\sim$正态分布，又设$X_{n+1}\sim N\left(\mu,\ \sigma^2\right)$，且$X_{n+1}$与$X_1,\ X_2,\cdots,\ X_n$相互独立，求

$T=\frac{x_{n+1}-\overline{x}}{S}\sqrt{\frac{n}{n+1}}$的分布。

解：含有S，可以预计容量应该是$n-1$，分子量纲为N分布，分母$S=\sqrt{S^2}$相当于χ^2，根据量纲法，可以推知结果是t分布。

$$x_{n+1}-\overline{x}\sim N\left(0,\ \sigma^2+\frac{\sigma^2}{n}\right)=N\left(0,\ \frac{n+1}{n}\sigma^2\right)\Rightarrow\frac{x_{n+1}-\overline{x}}{\sqrt{\frac{n+1}{n}\sigma^2}}=\frac{x_{n+1}-\overline{x}}{\sigma\cdot\sqrt{\frac{n+1}{n}}}\sim N(0,\ 1)$$

$$T=\frac{x_{n+1}-\overline{x}}{S}\sqrt{\frac{n}{n+1}}=\frac{\frac{x_{n+1}-\overline{x}}{\sigma\cdot\sqrt{\frac{n+1}{n}}}\cdot\sigma}{S}=\frac{\frac{x_{n+1}-\overline{x}}{\sigma\cdot\sqrt{\frac{n+1}{n}}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\cdot\frac{1}{n-1}}}=\frac{N(0,\ 1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}}\sim t(n-1)$$

【例 5】

设 $x_1, x_2,\cdots, x_n\sim N(\mu,\ \sigma^2)$，记 $S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$，$S_2^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$，

$S_3^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$，$S_4^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$。则下列正确的是（ ）。

$(A)\ t(n-1)=\frac{\overline{x}-\mu}{\frac{S_1}{\sqrt{n-1}}}$ $\qquad$ $(B)\ t(n-1)=\frac{\overline{x}-\mu}{\frac{S_2}{\sqrt{n-1}}}$

$(C)\ t(n-1)=\frac{\overline{x}-\mu}{\frac{S_3}{\sqrt{n}}}$ $\qquad$ $(D)\ t(n-1)=\frac{\overline{x}-\mu}{\frac{S_4}{\sqrt{n}}}$

解：由于容量为$(n-1)$的分布含样本方差，而S_1^2, S_2^2是样本方差，S_3^2, S_4^2不是，故立即可以否定$(C), (D)$。又只有S_1^2才是标准的样本方差，由标准的$t(n-1)=\frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}}$推知$(A)$不对。故选$(B)$。事实上

$$\frac{\overline{x}-\mu}{\frac{S_2}{\sqrt{n-1}}}=\frac{\overline{x}-\mu}{\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}}{\sqrt{n-1}}}=\frac{\overline{x}-\mu}{\frac{\sqrt{\frac{n-1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}}{\sqrt{n-1}}}$$

$$=\frac{\overline{x}-\mu}{\frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}}{\sqrt{n}}}=\frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)$$

【例 6】$X_1, X_2,\cdots, X_n\ (n\geq 2)$为来自$N(0,\ 1)$的简单随机样本，则下列哪个正确。

$(A)\ n\overline{X} \sim N(0,\ 1)$　　　　$(B)\ nS^2 \sim \chi^2(n)$

$(C)\ \dfrac{(n-1)\overline{X}}{S} \sim t(n-1)$　　　　$(D)\ \dfrac{(n-1)X_1^2}{\sum_{i=2}^{n} X_i^2} \sim F(1,\ n-1)$

解：选(D)

$\dfrac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{\overline{X}-0}{\frac{1}{\sqrt{n}}} = \sqrt{n}\,\overline{X} \sim N(0,\ 1)$，故排除$(A)$；

$\dfrac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} = \dfrac{\overline{X}-0}{\frac{1}{\sqrt{n}}} = \dfrac{\sqrt{n}\,\overline{X}}{S} \sim t(n-1)$，故排除$(C)$；

$\dfrac{(n-1)S^2}{1^2} = (n-1)S^2 \sim \chi^2(n-1) \not\Rightarrow nS^2 \sim \chi^2(n)$，故排除$(B)$；

$\dfrac{(n-1)X_1^2}{\sum_{i=2}^{n} X_i^2} = \dfrac{X_1^2}{\frac{\sum_{i=2}^{n} X_i^2}{n-1}} \sim \dfrac{\chi^2(1)}{\frac{\chi^2(n-1)}{n-1}} \sim F(1,\ n-1)$，故$(D)$正确。

【例 7】$X \sim N(\mu,\ \sigma^2)$，S_1^2和S_2^2独立，求$D(S_1^2-2S_2^2)$。

解：　$D(S_1^2-2S_2^2) = D(S_1^2)+4D(S_2^2) = \dfrac{2\sigma^4}{n_1-1}+\dfrac{8\sigma^4}{n_2-1} = 2\left(\dfrac{1}{n_1-1}+\dfrac{4}{n_2-1}\right)\sigma^4$

【例 8】$\{X_i\}$是来自正态整体X的简单随机样本，已知$Y_1 = \dfrac{1}{6}(X_1+X_2+\cdots X_6)$,

$Y_2 = \dfrac{1}{3}(X_7+X_8+X_9)$, $S^2 = \dfrac{1}{2}\sum_{i=1}^{9}(X_i-Y_2)^2$。求$Z = \dfrac{\sqrt{2}(Y_1-Y_2)}{S}$分布。

解：

$$E\left(Y_1-Y_2\right)=E\left(Y_1\right)-E\left(Y_2\right)=\frac{1}{6}E\left(X_1+X_2+\cdots X_6\right)-\frac{1}{3}E\left(X_7+X_8+X_9\right)=0$$

$$D\left(Y_1-Y_2\right)=D\left(Y_1\right)+D\left(Y_2\right)=\frac{1}{36}D\left(X_1+X_2+\cdots X_6\right)+\frac{1}{9}E\left(X_7+X_8+X_9\right)$$

$$=\frac{1}{36}\times 6\sigma^2+\frac{1}{9}\times 3\sigma^2=\frac{\sigma^2}{2}$$

$$\Rightarrow Y_1-Y_2\sim N\left(\mu,\ \frac{\sigma^2}{2}\right)\Rightarrow\frac{\left(Y_1-Y_2\right)-0}{\sqrt{\frac{\sigma^2}{2}}}=\frac{\sqrt{2}\left(Y_1-Y_2\right)}{\sigma}\sim N\left(0,\ 1\right)$$

$$\Rightarrow Z=\frac{\sqrt{2}\left(Y_1-Y_2\right)}{S}\sim\frac{\frac{\sqrt{2}\left(Y_1-Y_2\right)}{\sigma}}{\frac{S}{\sigma}}=\frac{N\left(0,\ 1\right)}{\sqrt{\frac{S^2}{\sigma^2}}}$$

$$\sim\frac{N\left(0,\ 1\right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\cdot\frac{1}{n-1}}}\sim\frac{N\left(0,\ 1\right)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}}\sim t(n-1)\ \left(n=2,\ 3,\cdots\right)$$

【例 9】$X\sim N\left(0,\ 2^2\right)$，$x_1, x_2,\cdots, x_{15}$ 来自 X 的简单随机样本，则 $Y=\frac{x_1^2+x_2^2+\cdots x_{10}^2}{2\left(x_{11}^2+x_{12}^2+\cdots x_{15}^2\right)}$ 服从什么分布。

解：分子量纲为 $N^2\to\chi^2$ 分布，分母量纲为也为 $N^2\to\chi^2$ 分布，根据量纲法，可以推知结果是 F 分布。下面具体计算如下

$$Y=\frac{x_1^2+x_2^2+\cdots x_{10}^2}{2\left(x_{11}^2+x_{12}^2+\cdots x_{15}^2\right)}=\frac{1}{2}\cdot\frac{\left(\frac{x_1}{2}\right)^2+\left(\frac{x_2}{2}\right)^2+\cdots\left(\frac{x_{10}}{2}\right)^2}{\left(\frac{x_{11}}{2}\right)^2+\left(\frac{x_{12}}{2}\right)^2+\cdots\left(\frac{x_{15}}{2}\right)^2}$$

$$\sim\frac{1}{2}\cdot\frac{\chi^2(10)}{\chi^2(5)}\sim\frac{\frac{\chi^2(10)}{10}}{\frac{\chi^2(5)}{5}}\sim F\left(10,\ 5\right)$$

【例 10】设 $X_1, X_2,\cdots, X_n\sim N\left(\mu,\ \sigma^2\right)$，$d=\frac{1}{n}\sum_{i=1}^{n}\left|X_i-\mu\right|$，求 Ed 和 Dd。

解：记 $Y_i=X_i-\mu$，则 $Y_i\sim N\left(0,\ \sigma^2\right)$。

$$E\left|Y_i\right|=\int_{-\infty}^{+\infty}|y|\cdot\frac{1}{\sqrt{2\pi}\cdot\sigma}e^{-\frac{y^2}{2\sigma^2}}dy=2\cdot\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_0^{+\infty}ye^{-\frac{y^2}{2\sigma^2}}dy=\sqrt{\frac{2}{\pi}}\sigma$$

$$D\left|Y_i\right|=EY_i^2-\left(E\left|Y_i\right|\right)^2=DY_i+\left(EY_i\right)^2-\left(\sqrt{\frac{2}{\pi}}\sigma\right)^2=\sigma^2+0-\frac{2}{\pi}\sigma^2=\left(1-\sqrt{\frac{2}{\pi}}\right)\sigma^2$$

$$\Rightarrow Ed=E\left(\frac{1}{n}\sum_{i=1}^{n}\left|X_i-\mu\right|\right)=\frac{1}{n}E\left(\sum_{i=1}^{n}\left|Y_i\right|\right)=\frac{1}{n}\cdot n\cdot\sqrt{\frac{2}{\pi}}\sigma=\sqrt{\frac{2}{\pi}}\sigma$$

$$Dd=D\left(\frac{1}{n}\sum_{i=1}^{n}\left|X_i-\mu\right|\right)=\frac{1}{n^2}D\left(\sum_{i=1}^{n}\left|Y_i\right|\right)=\frac{1}{n}\cdot n\cdot\left(1-\sqrt{\frac{2}{\pi}}\right)\sigma^2=\left(1-\sqrt{\frac{2}{\pi}}\right)\frac{\sigma^2}{n}$$

【例 11】 $X\sim t(n)$，求$\frac{1}{X^2}$的分布。

解：

$$X\sim t(n)\Rightarrow X\sim\frac{N(0,\ 1)}{\sqrt{\frac{\chi^2(n)}{n}}}\Rightarrow\frac{1}{X^2}\sim\frac{\frac{\chi^2(n)}{n}}{N^2(0,\ 1)}\sim\frac{\frac{\chi^2(n)}{n}}{\chi^2(1)}\sim\frac{\frac{\chi^2(n)}{n}}{\frac{\chi^2(1)}{1}}\sim F(n,\ 1)$$

【例 12】 $X\sim N\left(\mu,\ \sigma^2\right),\ Y\sim N\left(\mu,\ \sigma^2\right)$，均为简单随机样本，求

$$E\left[\frac{\sum_{i=1}^{n_1}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{n_2}\left(Y_i-\overline{Y}\right)^2}{n_1+n_2-2}\right]$$。

解：

$$E\left[\frac{\sum_{i=1}^{n_1}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{n_2}\left(Y_i-\overline{Y}\right)^2}{n_1+n_2-2}\right]=E\left[\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right]=E\left[\frac{(n_1-1)+(n_2-1)}{n_1+n_2-2}\right]\sigma^2=\sigma^2$$

。

【例 13】 $X\sim N(0,\ 1),\ Y=\left(X_1+X_2+X_3\right)^2+\left(X_4+X_5+X_6\right)^2,\ CY\sim\chi^2$。求$C$。

解：

$$X_1+X_2+X_3\sim N(0,\ 3)\Rightarrow\frac{X_1+X_2+X_3}{\sqrt{3}}\sim N(0,\ 1)$$

$$X_4+X_5+X_6\sim N(0,\ 3)\Rightarrow\frac{X_4+X_5+X_6}{\sqrt{3}}\sim N(0,\ 1)$$

$$\begin{aligned}\Rightarrow CY&=C(X_1+X_2+X_3)^2+C(X_4+X_5+X_6)^2\\&=C\times3\left[\left(\frac{X_1+X_2+X_3}{\sqrt{3}}\right)^2+\left(\frac{X_4+X_5+X_6}{\sqrt{3}}\right)^2\right]\sim\chi^2(2)\Rightarrow C=\frac{1}{3}\end{aligned}$$

【例 14】设总体$X\sim N(0,\ 1)$，$X_1,X_2,\cdots,X_{2n}$是来自总体的简单随机样本，求下列分布。

$$(1)Y_1=\frac{\sqrt{2n-1}X_1}{\sqrt{\sum_{i=2}^{2n}X_i^2}};\quad(2)Y_2=\frac{(2n-3)\sum_{i=1}^{3}X_i^2}{3\sum_{i=4}^{2n}X_i^2};\quad(3)Y_3=\frac{1}{2}\sum_{i=1}^{2n}X_i^2+\sum_{i=1}^{n}X_{2i-1}X_{2i}$$

解：

$$(1)Y_1=\frac{\sqrt{2n-1}X_1}{\sqrt{\sum_{i=2}^{2n}X_i^2}}=\frac{X_1}{\sqrt{\frac{\sum_{i=2}^{2n}X_i^2}{2n-1}}}\sim\frac{N(0,1)}{\sqrt{\frac{\chi^2(2n-1)}{2n-1}}}\sim t(2n-1)$$

$$(2)Y_2=\frac{(2n-3)\sum_{i=1}^{3}X_i^2}{3\sum_{i=4}^{2n}X_i^2}=\frac{\frac{\sum_{i=1}^{3}X_i^2}{3}}{\frac{\sum_{i=4}^{2n}X_i^2}{2n-3}}\sim\frac{\frac{\chi^2(3)}{3}}{\frac{\chi^2(2n-3)}{2n-3}}\sim F(3,\ 2n-3)$$

$$\begin{aligned}(3)Y_3&=\frac{1}{2}\sum_{i=1}^{2n}X_i^2+\sum_{i=1}^{n}X_{2i-1}X_{2i}\\&=\frac{1}{2}(X_1^2+\cdots+X_{2n}^2)+X_1X_2+X_3X_4+\cdots+X_{2n-1}X_{2n}\\&=\frac{1}{2}(X_1+X_2)^2+\frac{1}{2}(X_3+X_4)^2+\cdots+\frac{1}{2}(X_{2n-1}+X_{2n})^2\\&=\sum_{i=1}^{n}\left(\frac{X_{2i-1}+X_{2i}}{\sqrt{2}}\right)^2\sim\sum_{i=1}^{n}\left[N_i(0,1)\right]^2\sim\chi^2(n).\end{aligned}$$

【例 15】已知$X\sim N(\mu_1,\ \sigma^2),\ Y\sim N(\mu_2,\ \sigma^2)$

求 $(1)\dfrac{(n-1)\left(S_1^2+S_2^2\right)}{\sigma^2}$；$(2)\dfrac{n\left[\left(\overline{X}-\overline{Y}\right)-\left(\mu_1-\mu_2\right)\right]^2}{S_1^2+S_2^2}$。

解：

$$(1)\ \frac{(n-1)\left(S_1^2+S_2^2\right)}{\sigma^2}\sim\frac{(n-1)S_1^2}{\sigma^2}+\frac{(n-1)S_2^2}{\sigma^2}\sim\chi^2(n-1)+\chi^2(n-1)\sim\chi^2(2n-2)$$

$$(2)\frac{n\left[\left(\overline{X}-\overline{Y}\right)-\left(\mu_1-\mu_2\right)\right]^2}{S_1^2+S_2^2}=\frac{\left[\dfrac{\overline{X}-\mu_1}{\frac{\sigma}{\sqrt{n}}}-\dfrac{\overline{X}-\mu_2}{\frac{\sigma}{\sqrt{n}}}\right]^2}{\dfrac{(n-1)\left(S_1^2+S_2^2\right)}{(n-1)\sigma^2}}\sim\frac{\left[N(0,2)\right]^2}{\dfrac{\chi^2(n-1)+\chi^2(n-1)}{(n-1)}}$$

$$\sim\frac{\left[\dfrac{N(0,2)}{\sqrt{2}}\right]^2}{\dfrac{\chi^2(2n-2)}{2(n-1)}}\sim\frac{\left[\dfrac{N(0,1)}{1}\right]^2}{\dfrac{\chi^2(2n-2)}{2(n-1)}}\sim F(1,\ \ 2n-2)$$

四、**总体比例问题**

（一）总体比例的区间估计

1.基于标准正态分布的估计

假定条件

- 总体服从二项分布
- 样本比例可以由正态分布来近似（n≥30)

使用正态分布统计量$Z=\dfrac{p-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}\sim N(0,1)$，$P(-z_{\alpha/2}\le\dfrac{p-\pi}{\sqrt{\dfrac{\pi(1-\pi)}{n}}}\le z_{\alpha/2})=1-\alpha$

总体比例 π 在 1-α 置信水平下的置信区间为

不重复抽样：$p\pm z_{\alpha/2}\sqrt{\dfrac{\pi(1-\pi)}{n}(\dfrac{N-n}{N-1})}$ 或 $p\pm z_{\alpha/2}\sqrt{\dfrac{p(1\text{-}p)}{n}(\dfrac{N-n}{N-1})}$

例 1、

某城市想要估计下岗职工中女性所占的比例，随机抽取了 100 个下岗职工，其中 65 人为女性职工。试以 95%的置信水平估计该城市下岗职工中女性比例的置信区间

解：已知 n=100，p＝65% , 1-α= 95%，$z\alpha/2$=1.96

$$p\pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}=65\%\pm1.96\times\sqrt{\frac{65\%(1-65\%)}{100}}=65\%\pm9.35\%=\left(55.65\%,74.35\%\right)$$

即：该城市下岗职工中女性比例的置信区间为 55.65%-74.35%

2.基于 t 分布的估计

- 假定条件
  - 总体服从二项分布
  - 样本比例可以由正态分布来近似（n≥30)
- **使用正态分布统计量 *t***

$$t=\frac{p-\pi}{\sqrt{\frac{p(1-p)}{n}}}\sim t(n-1)，\quad P\left(-t_{\alpha/2}\le\frac{p-\pi}{\sqrt{\frac{p(1-p)}{n}}}\le t_{\alpha/2}\right)=1-\alpha$$

总体比例 π 在 1-α 置信水平下的置信区间为不重复抽样：$p\pm t_{\alpha/2,n-1}\sqrt{\frac{p(1\text{-}p)}{n}(\frac{N-n}{N-1})}$

例 2、某高校有教职工 1000 人，按随机原则以不重复抽样方式抽取 201 名教职工进行调查，其中表示支持某候选人的有 110 人，试以 95%的置信标准估计该候选人的全校支持率。

解：已知　***N*=1000, *n*=201，**

　　***p*＝110/201=54.73% , 1-α= 95%查 t 分布表得 *tα/2 n-1*=1.972**

$$E_p=t_{\alpha/2,n-1}\cdot\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)}$$
$$=1.972\times\sqrt{\frac{0.5437(1-0.5437)}{201}\left(\frac{1000-201}{1000-1}\right)}\approx 6.2\%$$

$$p\pm E_p=\left(54.37\%\pm 6.2\%\right)=(48.17\%,60.57\%)$$

即：以 **95%**的置信度估计该候选人的全校支持率介于 **48.17%～60.57%**之间。

（二）总体比例的假设检验

（检验统计量）

z 检验 $z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$　　**t** 检验 $t=\frac{p-\pi_0}{\sqrt{\frac{p(1-p)}{n}}}$

二种方法的检验步骤基本一致，主要区别在于检验统计量的计算公式与查找的临界值不同。前者最为常见。在此，仅介绍总体比例的 **z** 检验法

总体比例的 **z** 检验法

1. 假定条件

   是非标志总体，计量结果只有两种情况

   大样本

   样本比例的抽样分布可用正态分布来近似

2. 比例检验的 ***Z*** 统计量 $Z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}\sim N(0,1)$

π0 为假设的总体比例

3.给定显著性水平α，查表得出相应的临界值 $z\alpha$ 或 $z\alpha/2$

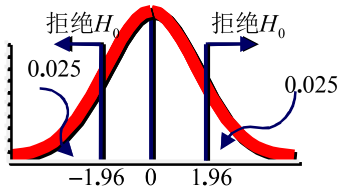
4.将检验统计量 **z** 的值与α水平的临界值进行比较，然后作出决策

双侧检验：当 $|z| > z_{\alpha/2}$ 时，拒绝 **H0**

左侧检验：当 $z < -z_{\alpha}$ 时，拒绝 **H0**

右侧检验：当 $z > z_{\alpha}$ 时，拒绝 **H0**

例 3、一种以休闲和娱乐为主题的杂志，声称其读者群中有 80%为女性。为验证这一说法是否属实，某研究部门抽取了由 200 人组成的一个随机样本，发现有 146 个女性经常阅读该杂志。分别取显著性水平 α =0.05 和 α =0.01 ，检验该杂志读者群中女性的比例是否为 **80%**？



解：（**1**） α = **0.05**

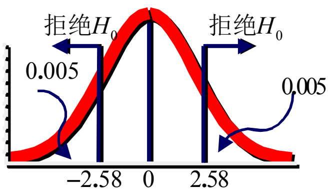
提出假设 ***H0*** ： $\pi = 80\%$ ***H1*** ： $\pi \neq 80\%$

根据 α = **0.05** 查找正态分布临界值,如图

计算 **z** 检验统计量：$p = \frac{146}{200} = 0.73, z = \frac{0.73 - 0.80}{\sqrt{\frac{0.80 \times (1 - 0.80)}{200}}} = -2.475$

作出决策∵ | **z** | =**2.475**> $z_{\alpha/2} = 1.96$

∴拒绝 ***H0*** 结论： 该杂志的说法并不属实



（**2**） α = **0.01**

提出假设 ***H0*** ： $\pi = 80\%$ ***H1*** ： $\pi \neq 80\%$

根据 α = **0.01** 查找正态分布临界值

计算 **z** 检验统计量：$p = \frac{146}{200} = 0.73, z = \frac{0.73 - 0.80}{\sqrt{\frac{0.80 \times (1 - 0.80)}{200}}} = -2.475$

作出决策∵ | **z** | =**2.475**< $z_{\alpha/2} = 2.58$

∴不拒绝 ***H0***

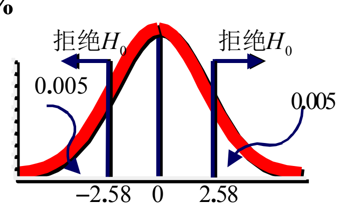
结论： 该杂志的说法属实

例 4、某产品的不合格品率是 **8%**，企业对该产品的生产设备进行了技术改造，从改造后的产品中随机抽取 **36** 件进行检验，发现有次品 **1** 件，即次品率为 **2.778%**。能否认为该次技术改造提高了产品的质量？ （α =**0.01**）

解：提出假设 ***H0*** ： $\pi \geqslant$ **8%** ***H1*** ： $\pi <$ **8%**

根据 α = **0.01** 查找正态分布临界值

计算 **z** 检验统计量：



$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} = \frac{0.02778 - 0.08}{\sqrt{\frac{0.08 \times (1 - 0.08)}{26}}} = -1.155$$

作出决策

∵**z=-1.155**＞ $-z_{\alpha/2} = -2.58$

∴不拒绝 ***H*0**

结论:　该技术改造没有显著地提高产品质量