第七章 样本分布

数理统计学：

运用概率论的基础知识，对要研究的随机现象进行多次观察或试验，研究如何合理地获得数据资料，建立有效的数学方法，根据所获得的数据资料，对所关心的问题作出估计与检验。

§1总体、个体与样本

对某一问题的研究对象全体称为总体。

组成总体的某个基本单元，称为个体。

总体可以是具体事物的集合，如一批产品。

也可以是关于事物的度量数据集合，如长度测量。

总体可以包含有限个个体，也可以包含无限个个体。

有限总体在个体相当多的情况下，可以作为无限总体进行研究。

总体中的个体，应当有共同的可观察的特征。该特征与研究目的有关。

例如：

总体	个体	特征
一批产品	每件产品	等级
一批灯泡	每个灯泡	寿命
一年的日平均气温	每天日平均气温	度数
数轴上某一线段	线段中每一点	坐标
一批彩票	每张彩票	号码

人们感兴趣的是总体的某一个或几个数量指标的分布情况。每个个体所取的值不同，但它按一定规律分布。

以随机变量ξ代表总体的特征。

当总体数量很大时，只能从中抽取部分个体进行研究。

从总体中取出的若干个体，称为样本。

样本中所含个体的个数，称为样本容量。

选取样本是为了从样本的特征对总体特征做出估计和推断。

抽样必须尽可能多地反映总体的特征。

要求随机抽取：

(1)独立性：抽样时互不影响。

(2)代表性：样本的分布与总体相同。

通常有两种抽样方式：

(1)不重复抽样(不放回)　　(2)重复抽样(放回)

重复抽样所得的样本，称为简单随机样本。

对总体进行n次独立试验或n次独立观察，

即是从总体中抽取容量为n的样本，

以随机变量$X_1, \ldots, X_n$代表

每个X_i应与总体ξ有相同的分布。

每次具体抽样所得的数据，是这个样本的

一组观察值，记为$(x_1, \ldots, x_n)$

一般也称为样本。

样本$(X_1,\ldots,X_n)$的函数$f(X_1,\ldots,X_n)$称为统计量，其中$f(X_1,\ldots,X_n)$不含有未知参数。如若

μ,δ未知时，$Y_i=\dfrac{X_i-\mu}{\delta}$就不是一个统计量

统计量一般是样本的连续函数，也是随机变量。

常用的统计量如：

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

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

§2 样本分布函数

- 在实际统计工作中,首先接触到的是一系列数据.数据的变异性,系统地表现为数据的分布.分布的具体表示形式为表和图.统计表有简单表和分组表之分.统计图有频率图,频率直方图和累积频率直方图等.

(一)分组数据的统计表

简单表：依数据出现先后或大小列成表。

- 例1 20名新生婴儿的体重的观察值为

2880	2440	2700	3500	3600
3080	3860	3200	3500	3100
3180	3200	3300	3040	3020
3420	2900	3440	3000	2620

若要更清楚地了解数据的分布，进行分组。

每一组数据看成是相同的，它们等于组中值。

一般采用等区间分组，区间长度称为组距。

将上述数据分成五组：

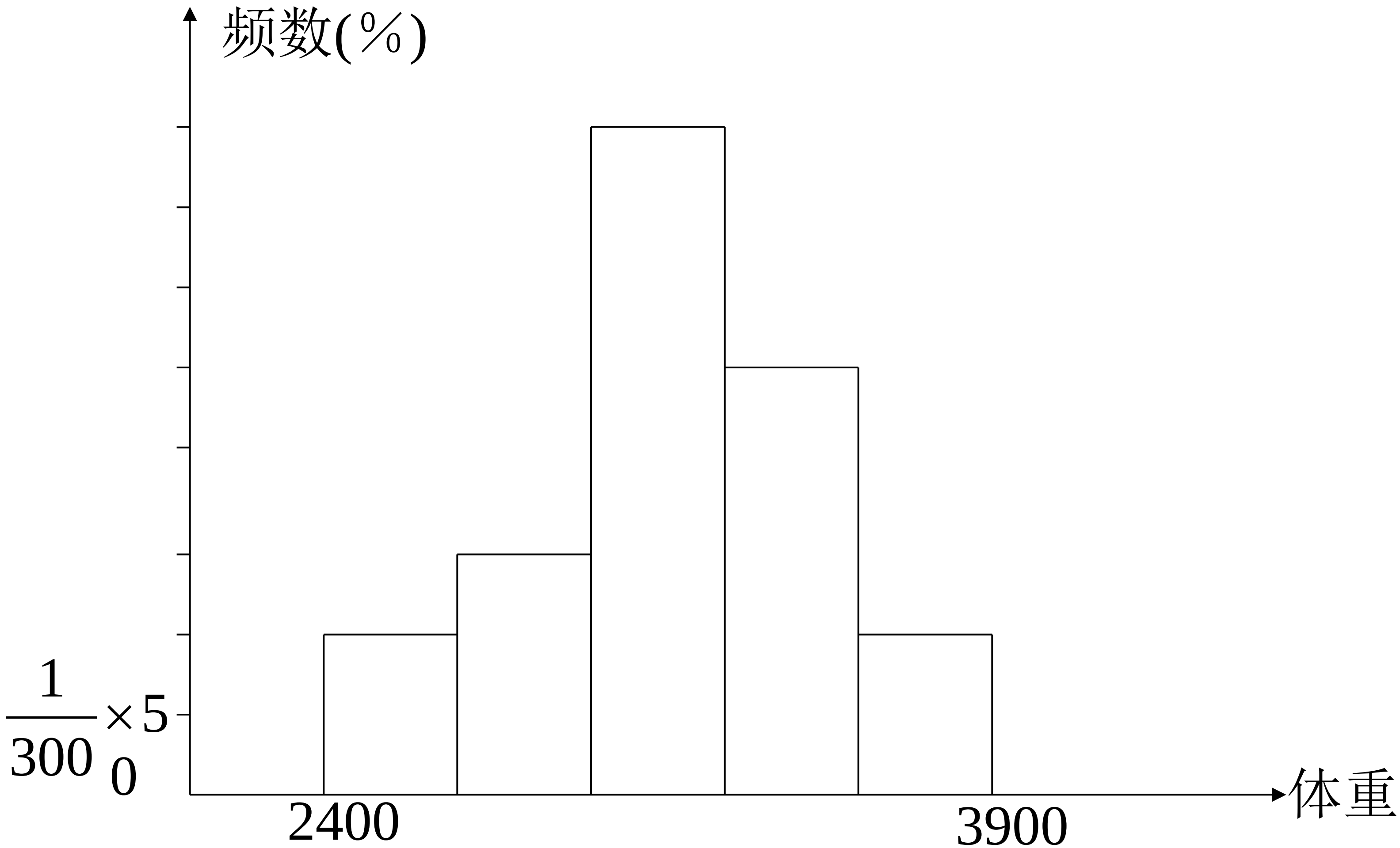
分组编号	1	2	3	4	5
组限	2400-2700	2700-3000	3000-3300	3300-3600	3600-3900
组中值	2550	2850	3150	3450	3750
组频数	2	3	8	5	2
组频率(%)	10	15	40	25	10
累积频率(%)	10	25	65	90	100

其中频率是频数除以总频数。

累积频率是指相应的组频率之和。

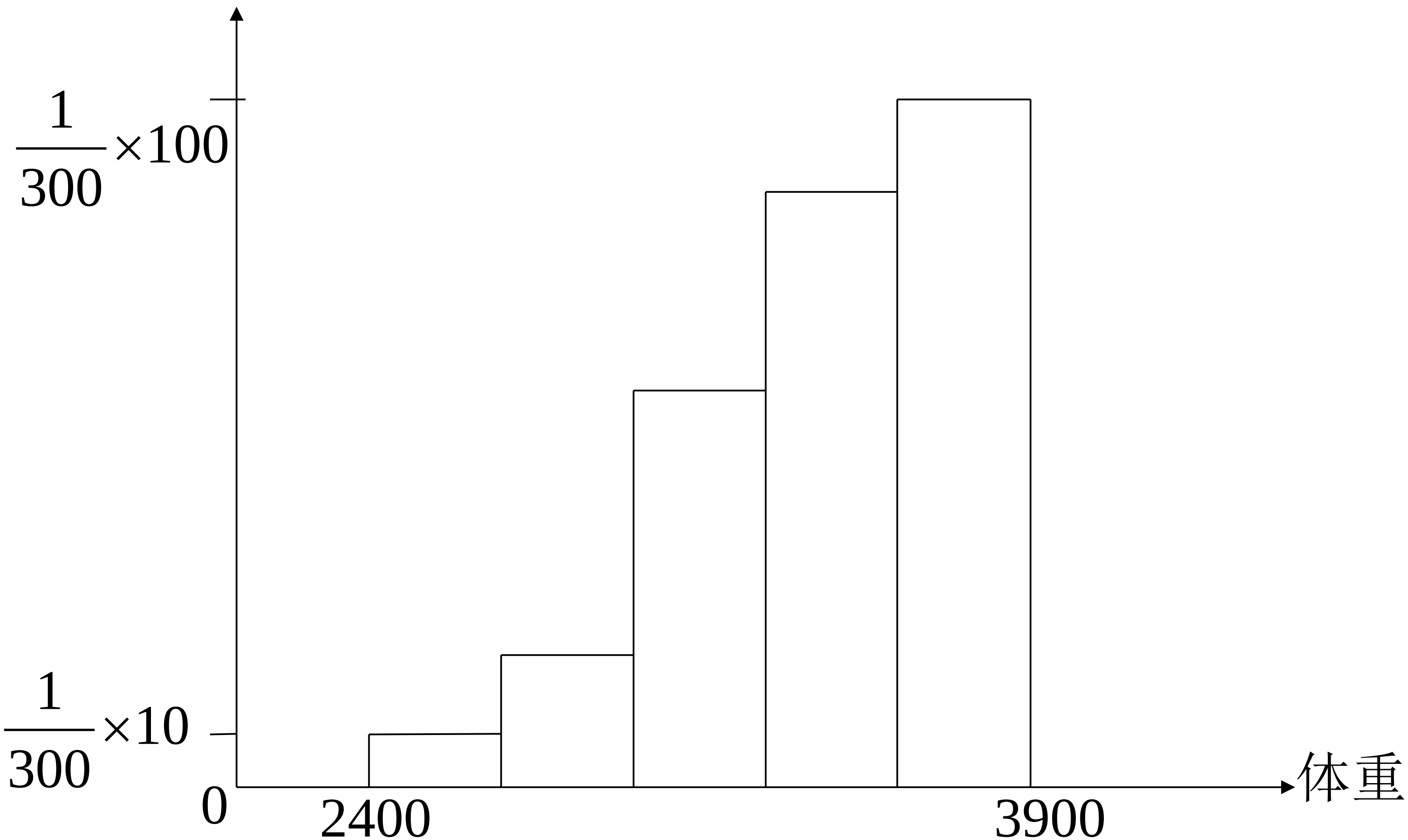
为了直观，一般用直方图表示：

频率直方图：



累积频率直方图：

第i个长方形的面积表示累积频率。



(二)样本分布函数

总体是一个随机变量ξ，ξ的分布就是总体的分布。ξ的分布函数$F(x)$是总体分布函数。

设$(x_1,\dots,x_n)$是总体ξ的一个样本观察值，按大小排列为：$x_1^* \le x_2^* \le \dots \le x_n^*$

$$令F_n(x)=\begin{cases} 0 & 当x < x_1^* \\ 1/n & 当x_1^* \le x < x_2^* \\ \dots & \\ k/n & 当x_k^* \le x < x_{k+1}^* \\ \dots & \\ 1 & 当x \ge x_n^* \end{cases}$$

$F_n(x)$的图形是累积频率曲线。

它是跳跃上升的一条阶梯曲线。

若观测值不重复，跃度为1/n

若重复，按1/n的倍数跳跃上升。

当$n \to \infty$时，

$F_n(x)$的极限为总体分布函数$F_\xi(x)$

称$F_n(x)$为样本分布函数或经验分布函数。

例2 随机观察总体ξ，得10个数据如下：

3.2，2.5，-4，2.5，0，3，2，2.5，4，2

求样本分布函数$F_{10}(x)$

解：将数据由小到大排列为

-4<0<2=2<2.5=2.5=2.5<3<3.2<4

$$
\text{其样本分布函数为：} F_{10}(x) = \begin{cases} 0 & \text{当} x < -4 \\ 1/10 & \text{当} -4 \le x < 0 \\ 2/10 & \text{当} 0 \le x < 2 \\ 4/10 & \text{当} 2 \le x < 2.5 \\ 7/10 & \text{当} 2.5 \le x < 3 \\ 8/10 & \text{当} 3 \le x < 3.2 \\ 9/10 & \text{当} 3.2 \le x < 4 \\ 1 & \text{当} x \ge 4 \end{cases}
$$

§3 样本分布的数字特征

(一)样本平均值

对于样本$(X_1,\ldots,X_n)$

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

对于具体样本值$(x_1,\ldots,x_n)$

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

若样本观察值已整理成分组数据，分成k组。

属于同一组的数据以组中值x_i'代表，组频数为m_i

则样本平均值的计算公式为 $\overline{x}=\frac{1}{n}\sum_{i=1}^{k}m_i x_i'$

若观察值为

5，6.5，7，4，5.4，6.3，5.8，6.9

则 $\bar{x} = \frac{1}{8}(5+6.5+7+4+5.4+6.3+5.8+6.9)$

$= \frac{1}{8} \times 46.9 = 5.8625$

再如婴儿的体重

组中值	2550	2850	3150	3450	3750
组频数	2	3	8	5	2

则 $\bar{x} = \frac{1}{20}(2\times2550+3\times2850+8\times3150+5\times3450+2\times3750)$

$= 3180$

(二)样本方差

对于样本$(X_1,\ldots,X_n)$

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

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

实际计算时，

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

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

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

若$(x_1,\ldots,x_n)$为样本观察值

$$s^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)$$

若数据已分成k组

$$s^2=\frac{1}{n-1}\left(\sum_{i=1}^{k}m_i(x_i')^2-n\overline{x}^2\right)$$

如观察值为

5，6.5，7，4，5.4，6.3，5.8，6.9

$$\sum_{i=1}^{8}x_i^2=5^2+6.5^2+\ldots+6.9^2＝282.35$$

$$\therefore s^2=\frac{1}{7}(282.35-8\times5.8625^2)\approx1.057$$

又如婴儿体重

组中值	2550	2850	3150	3450	3750
组频数	2	3	8	5	2

$$\sum_{i=1}^{5} m_i x_i^2 = 2\times 2550^2 + 3\times 2850^2 + 8\times 3150^2 + 5\times 3450^2 + 2\times 3750^2$$

$$=204390000$$

$$s^2 = \frac{1}{19}(204390000 - 20\times 3180^2) = 112736.84$$

或 $$s^2 = \frac{1}{19}\Big[2(2550-3180)^2 + 3(2850-3180)^2 + 8(3150-3180)^2$$

$$+5(3450-3180)^2 + 2(3750-3180)^2\Big]$$

$$=112736.84$$

(三)样本平均值与样本方差的简单公式

设$(x_1,\ldots,x_n)$为样本的n个观察值

对任意常数a及非零常数c

记$z_i = {(x_i - a)}/{c}$ $\qquad i = 1,\ldots,n$

即 $x_i = cz_i + a$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(cz_i + a) = c\cdot\frac{1}{n}\sum_{i=1}^{n} z_i + a = c\overline{z} + a$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(cz_i - c\overline{z})^2$$

$$= c^2\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{z})^2 = c^2 s_z^2$$

a与c选取应使z尽可能简单。

例3 在婴儿体重数据中，令$z_i = {(x_i - 3150)} \big/ {300}$

组中值	2550	2850	3150	3450	3750
z_i	-2	-1	0	1	2
组频数m_i	2	3	8	5	2
$m_i z_i$	-4	-3	0	5	4
z_i^2	4	1	0	1	4
$m_i z_i^2$	8	3	0	5	8

故 $\bar{z} = \frac{1}{20} \times 2 = 0.1$ $\quad s_z^2 = \frac{1}{19} \times (24 - 20 \times 0.1^2) = \frac{1}{19} \times 23.8$

$\bar{x} = 300 \times 0.1 + 3150 = 3180$

$s_x^2 = 300^2 \times \frac{1}{19} \times 23.8 = 112736.84$

§4 几个常用统计量的分布

数理统计中，较多使用正态总体，其样本$X_1,\ldots,X_n$的统计量$\overline{X}$与S^2及其函数的分布很重要。

- 定理1 设$X_1,\ldots,X_n$相互独立，X_i服从正态分布$N(\mu_i,\sigma_i^2)$，则它们的线性函数$\eta=\sum_{i=1}^{n}a_iX_i$($a_i$不全为零)也服从正态分布，且 $E\eta=\sum_{i=1}^{n}a_i\mu_i,\ D\eta=\sum_{i=1}^{n}a_i^2\sigma_i^2$

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

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

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

这是因为$\overline{X}$是$X_1,\ldots,X_n$的线性函数

故$\overline{X}$是正态分布

$$E\overline{X}=\frac{1}{n}\sum_{i=1}^{n}EX_i=\frac{1}{n}\sum_{i=1}^{n}\mu=\mu$$

$$D\overline{X}=\frac{1}{n^2}\sum_{i=1}^{n}DX_i=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{\sigma^2}{n}$$

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

标准化可得 $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

- 定理2 设$X_1,\dots,X_n$相互独立，都服从标准正态分布，

则$\chi^2=\sum_{i=1}^{n}X_i^2$服从具有n个自由度的$\chi^2$分布，记为$\chi^2(n)$

- 定理3 设$X_1,\dots,X_n$相互独立，都服从标准正态分布，

则$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$与$\sum_{i=1}^{n}(X_i-\overline{X})^2$相互独立，并且

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

若X_i不是标准正态分布，而是$N(\mu,\sigma^2)$

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

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

对Y_i应用定理3得到

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

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

(2)$\overline{X}$与$\sum_{i=1}^{n}(X_i-\overline{X})^2$相互独立

- 定理4　设两个随机变量ξ与η相互独立，且

$\xi\sim N(0,1),\eta\sim\chi^2(n)$，则$T=\xi\Big/\sqrt{\eta/n}$

服从具有n个自由度的t分布

记为　$T\sim t(n)$

若$X_1,\ldots,X_n$是取自正态总体的样本，

则 $\xi = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

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

且ξ与η独立，应用定理4

$$\frac{\xi}{\sqrt{\eta/(n-1)}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)}$$

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

即

推论1 设$X_1,\ldots,X_n$是取自正态总体$N(\mu,\sigma^2)$的样本

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

推论2 设$X_1,\ldots,X_m$和$Y_1,\ldots,Y_n$分别是来自两个独立正态总体$N(\mu_1,\sigma^2)$及$N(\mu_2,\sigma^2)$，则

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

由于 $\bar{X}\sim N\left(\mu_1,\frac{\sigma^2}{m}\right)$ $\quad\bar{Y}\sim N\left(\mu_2,\frac{\sigma^2}{n}\right)$

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

故 $$\xi=\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{1}{m}+\frac{1}{n}}\sigma} \sim N(0,1)$$

而 $$\frac{(m-1)S_1^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{m}(X_i-\overline{X}) \sim \chi^2(m-1)$$

$$\frac{(n-1)S_2^2}{\sigma^2}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 \sim \chi^2(n-1)$$

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

对ξ与η应用定理4得到推论2

- 定理5 设有两个随机变量ξ_1和ξ_2相互独立，

且$\xi_1 \sim \chi^2(n_1)$, $\xi_2 \sim \chi^2(n_2)$, 则

$$F = \frac{\xi_1 / n_1}{\xi_2 / n_2} \sim F(n_1, n_2)$$

$F(n_1, n_2)$是第一个自由度为n_1,第二个自由度为n_2的$F-$分布

推论 设$X_1, \dots, X_{n_1}$和$Y_1, \dots, Y_{n_2}$是分别取自两个独立正态总体$N(\mu_1, \sigma_1^2)$与$N(\mu_2, \sigma_2^2)$的样本，S_1^2, S_2^2分别为它们的样本方差，则

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

由定理4的推论可知

$$\xi_1=\frac{1}{\sigma_1^2}\sum_{i=1}^{n_1}(X_i-\overline{X})^2=\frac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1)$$

$$\xi_2=\frac{1}{\sigma_2^2}\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2=\frac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$$

总体独立，故ξ_1与ξ_2也独立。

对ξ_1与ξ_2用定理5得到推论。

例, 设 X 服从自由度为n的t分布, 求 X^2 的分布.

解 假设

$$X=\frac{Y}{\sqrt{Z/n}},Y\text{服从}N(0,1)\text{分布},Z\text{服从}\chi^2(n)\text{分布},$$

且Y和Z是相互独立的.则

$$X^2=\frac{Y^2}{Z/n},\text{故服从}F(1,n)\text{分布}$$