

# § 4 直方图

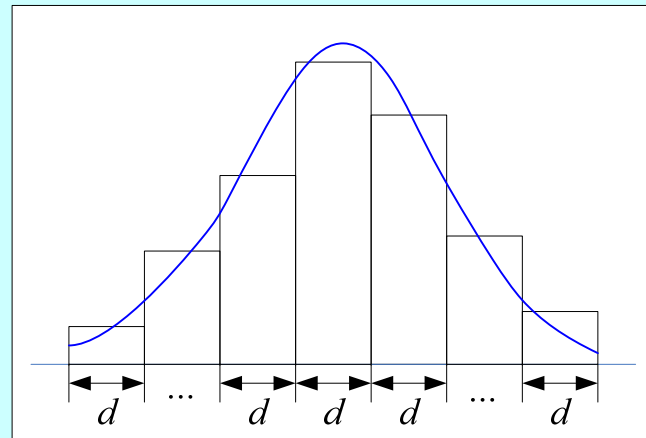
## ——总体分布的估计和检验

- (一) 等距频率直方图

- 密度函数的图解法

- (1) 说明1

- 直方图：在平面坐标上，以 $x$ 轴表示所考察的数据变量，以 $y$ 轴表示某统计量。以每组数据的区间为底边，以统计量为高画长方形，可得出数据的直方图。
- 等距：每个区间距离相等
- 频率直方图：长方形的面积表示落入此区间的频率
- 主要作用：对总体分布的密度函数进行估计。



## — (2) 原理

- 设  $X_1, X_2, \dots, X_n$  是来自分布密度函数的  $f(x)$  某总体  $X$  的样本（假设  $X$  是连续型随机变量），
- 把  $X$  的取值范围等分为  $m$  个小区间，每个区间长度为  $d$ ,
- 落入第  $i$  个小区间  $[t_{i-1}, t_i)$  ( $i=1, 2, \dots, m$ ) 的观测个数为  $\mu_i$ ,
- 观测值落入第  $i$  个小区间的概率为

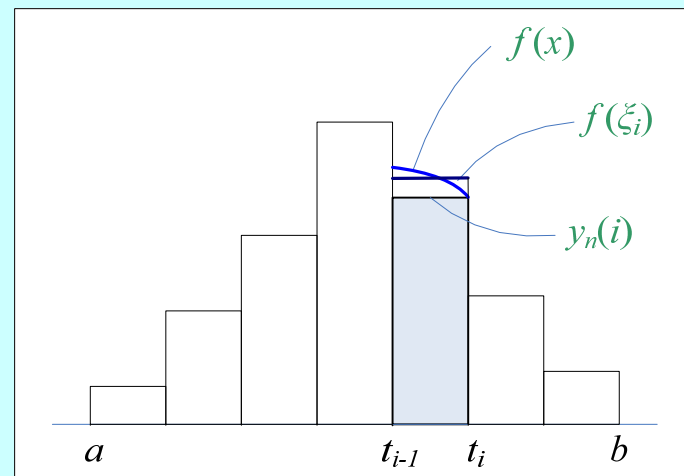
$$\begin{aligned} p_i &= P\{t_{i-1} \leq X < t_i\} = \int_{t_{i-1}}^{t_i} f(x) dx \\ &= (t_i - t_{i-1})f(\xi_i), \quad \xi_i \in [t_{i-1}, t_i) \end{aligned}$$

- 用频率作为概率的估计:

$$p_i = (t_i - t_{i-1})f(\xi_i) \approx \frac{\mu_i}{n}$$

- 故可得  $f(\xi_i)$  的估计值  $y_n(i)$  :

$$f(\xi_i) \approx \frac{\mu_i}{n(t_i - t_{i-1})} = \frac{\mu_i}{nd} @ y_n(i)$$



- 当  $f(x)$  在  $[t_{i-1}, t_i)$  上连续,  $d$  很小且样本量  $n$  充分大时, 则可用  $y_n(i)$  作为  $f(x)$  在小区间  $[t_{i-1}, t_i)$  上的近似值。
- 结论: 可通过等距频率直方图估计分布密度函数的情况。

## — (3) 作频率直方图的步骤

假设一组实验数据为： $x_1, x_2, \dots, x_n$ ,

- ①确定区间端点，分组数，组距

1> 区间端点  $[a, b]$

$$\text{设 } x_{(1)} = \min_i \{x_i\}, \quad x_{(n)} = \max_i \{x_i\}$$

$$\text{则取 } a = x_{(1)} - \varepsilon, \quad b = x_{(n)} + \varepsilon$$

$\varepsilon$  可根据实验数据的有效数字来决定（如 $x_i$ 取小数点后2位数字，则可取 $\varepsilon=0.005$ ）

2> 分组数 $m$

由样本容量 $n$ 决定，通常取值为  $m = 1.87 \times (n-1)^{\frac{2}{5}}$

3> 组距 $d$

$$d = (b - a) / m$$

- ②计算分组频数和频率

- 1> 确定每个小区间的端点

求出把区间 $[a,b]$  等分成  $m$  个小区间的  $m-1$ 个分点:

$$t_1 < t_2 < \dots < t_{m-1},$$

记 $t_0=a$ ,  $t_m=b$ , 第 $i$ 个小区间为 $[t_{i-1}, t_i)$ ,  $t_i=a+id$ 。

- 2> 求 $[t_{i-1}, t_i)$ 上的经验频数 $\mu_i$

计算满足不等式  $t_{i-1} \leq x_j < t_i$  ( $j=1,2,\dots,n; i=1,2,\dots,m$ )

的数据 $\{x_j\}$ 的个数  $\mu_i$ ,

依次扫描 $x_1, x_2, \dots, x_n$ , 对于每个 $x_j$ ,  $i = \left[ \frac{x_j - a}{d} \right] + 1$

$\mu_i = \mu_i + 1$  ( $\mu_i$ 初值=0) 。

- 3> 求 $[t_{i-1}, t_i)$ 上的经验频率 $f_i$

$$f_i = \frac{\mu_i}{n} \quad (i=1,2,\dots,m)$$

- ③画频率直方图

对每个小区间 $[t_{i-1}, t_i)$  ( $i=1,2,\dots, m$ )分别作长方形 (面积为 $f_i$ ) :

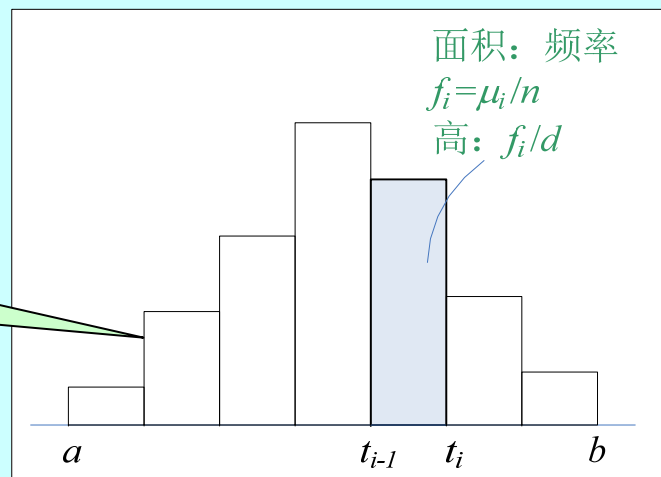
以小区间  $[t_{i-1}, t_i)$ 长度为底,

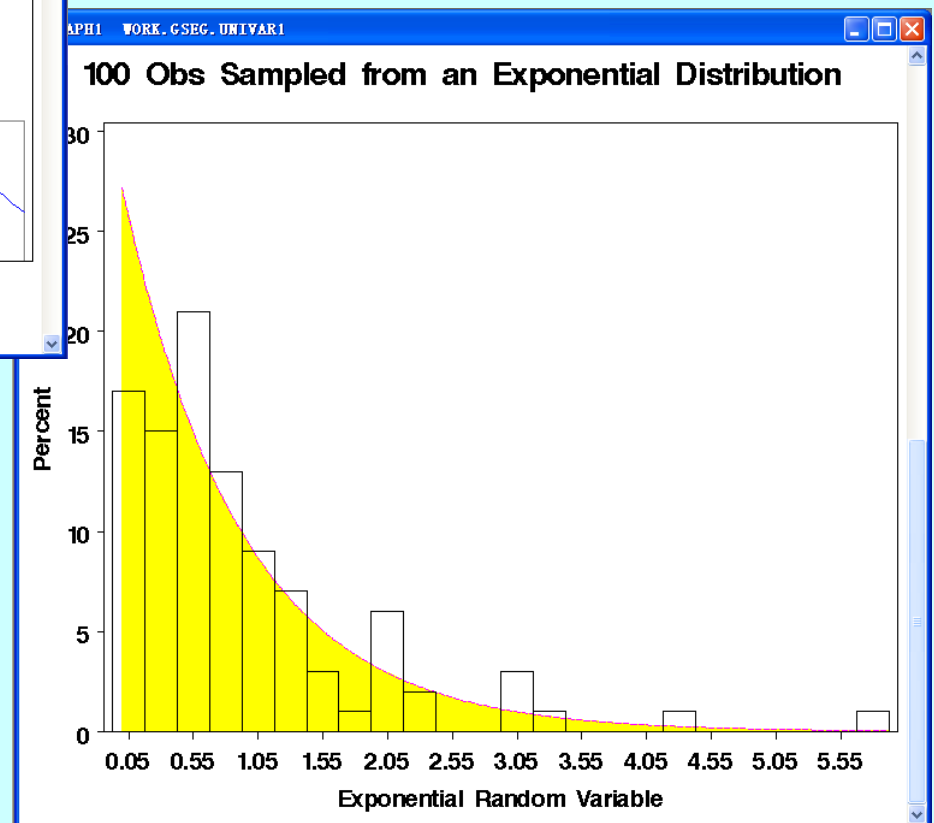
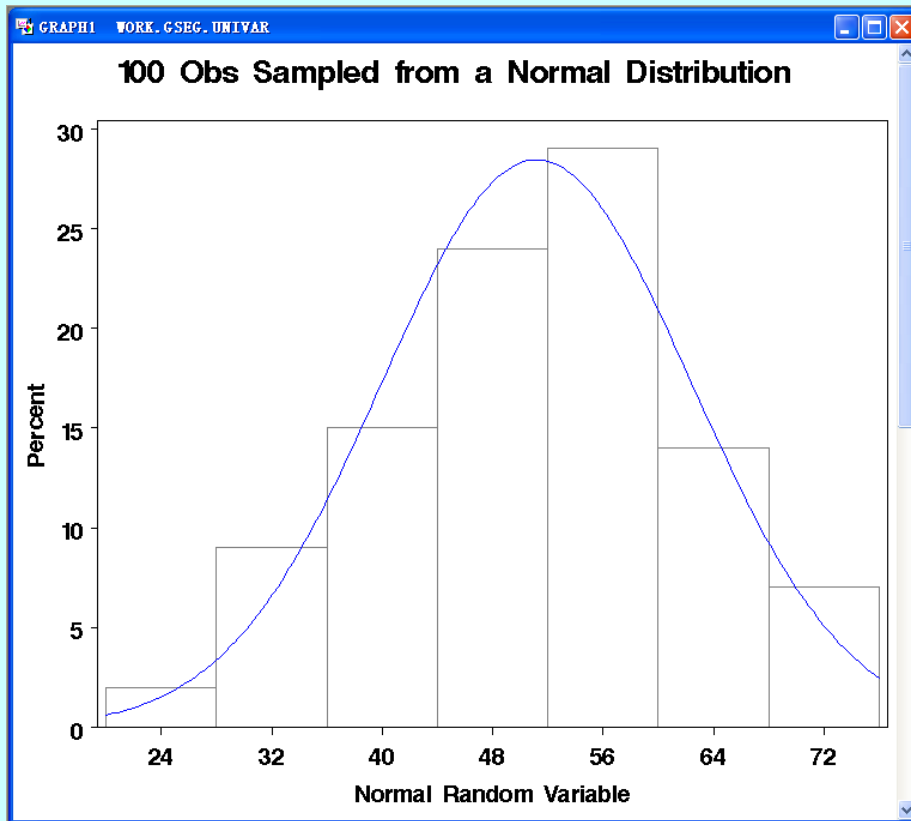
以  $y_i = f_i / d$  为高,

画出一排竖着的长方形即为频率直方图。

长方形的面积表示落入此区间的频率

长方形的高与落入此区间的频率成正比





## — (4) 方法说明

(用直方图估计总体时)

- ①分布的拟合检验

1> 问题：检验总体是否服从于某种确定分布函数  $F_0(x)$ 。

2> 方法：分布的  $\chi^2$  拟合优度检验法

3> 理论回顾：

假设检验问题：

$$H_0 : F(x) = F_0(x), \quad H_1 : F(x) \neq F_0(x)$$

对于连续型随机变量：

$$H_0 : f(x) = f_0(x), \quad H_1 : f(x) \neq f_0(x)$$



思想:

把样本空间  $S$  分成  $m$  个互不相交的集合

(  $A_1 \cup A_2 \cup \cdots \cup A_m = S$ ,  $A_i A_j = \emptyset$ ,  $i \neq j$ ,  $i, j = 1, 2, \dots, m$  )

在假设  $H_0$  下, 可以算出  $p_i = P(A_i)$ ,  $i = 1, 2, \dots, m$

又可统计出  $n$  次试验中,  $A_i$  出现的频数  $\mu_i$ ,

一般地, 若  $H_0$  为真, 且  $n$  充分大时,  $\mu_i$  与  $np_i$  的差异不会显著。

皮尔逊定理: 若  $n$  充分大, 如下统计量  $\chi^2$  总是近似的服从自由度为  $(m - r - 1)$  的  $\chi^2$  分布,

$$\chi^2 = \sum_{i=1}^m \frac{(\mu_i - np_i)^2}{np_i} = \sum_{i=1}^m \left( \frac{\mu_i}{n} - p_i \right)^2 \frac{n}{p_i}$$

其中  $r$  是  $F_0(x)$  中被估计参数的个数。

- 4> 在频率直方图中的应用

对于连续型随机变量,

零假设 $H_0$ : 总体的分布密度函数为  $f_0(x)$ 。

把取值范围划分为有限个互不重叠的子区间 $[t_{i-1}, t_i)$   $i=1,2,\dots, m$ ,

统计出样本中随机变量落在每个子区间的频数  $\mu_i$  ( $i=1,2,\dots, m$ ),

在假设  $H_0$  下, 求观测值落入第  $i$  个小区间  $[t_{i-1}, t_i)$  ( $i=1,2,\dots, m$ ) 的概率:

$$p_i = \int_{t_{i-1}}^{t_i} f_0(x) dx$$

并求出统计量 $\chi^2$ 的估计值:

$$\hat{\chi}^2 = \sum_{i=1}^m \left( \frac{\mu_i}{n} - p_i \right)^2 \frac{n}{p_i}$$

当  $p_i$  很小时，若出现个别小区间经验频数  $\mu_i$  与理论频数  $np_i$  相差较大的情况，将会使  $\chi^2$  的估计值增大很多，从而拒绝零假设。

- ②直方图分组数 $m$ 的选取

- 1>选取原则:

- 分组数  $m$  较大时能更好的反应样本的情况

- 使得每个小区间的频数  $\mu_i \neq 0$  , 最好  $\mu_i \geq 5$  ( $i=1,2,\dots,m$ ) ,  
当某  $\mu_i < 5$  时, 可把小频数区间与相邻区间合并, 即调整  
分组数  $m$  。

- 2>相关因素: 样本容量、实验数据的取值范围、有效数字  
的位数

- 当  $n$  大时,  $m$  也相应取较大的数;

- 当  $x_i$  的有效数字的位数较多且取值范围大时,  $m$  也相应  
取较大的数。

例：若只考虑  $m$  和  $n$  的关系，当总体服从于正态分布， $m$  与  $n$  的最优拟合关系：

$$m = 1.87 \times (n-1)^{\frac{2}{5}}$$

例：进一步考虑  $x_i$  的取值范围，若

$$x_{(n)} - x_{(1)} = 11.6 - 10.2 = 1.4$$

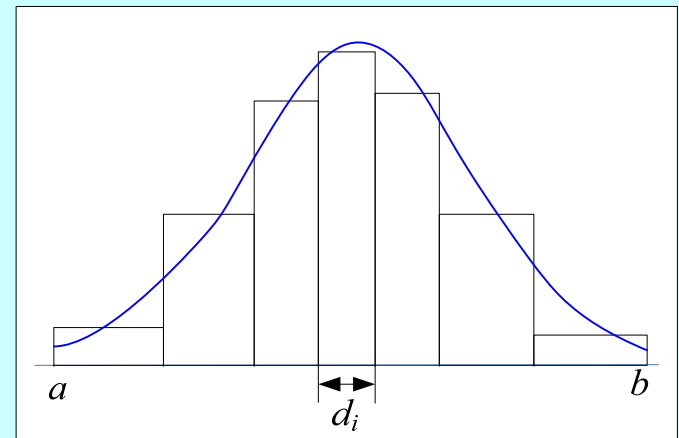
且数据的有效位数是小数点后一位，

为保证不出现频数为 0 的小区，应取  $m \leq 14$ 。

- （二）等概频率直方图——分布的拟合检验

- （1）说明1

- 等概：数据落入每个区间的概率相等
- 频率直方图：长方形的面积表示数据落入每个区间的频率
- 作用：进行分布的拟合检验



## — (2) 作图

- 设实验数据 $x_1, x_2, \dots, x_n$ 来自总体 $X$ ，其分布函数 $F(x)$ 已知，数据的取值范围是 $[a, b]$ 。
- 记 $t_0=a$ ， $t_m=b$ ，计算 $m-1$ 个分点 $t_i (i=1, 2, \dots, m-1)$ 的位置

$$P\{t_{i-1} \leq X < t_i\} = p_i = \frac{1}{m}$$

这些分点将 $[a, b]$ 分成互不相交的 $m$ 个等概率区间。

- 以下步骤同绘制等距频率直方图
  - ①确定区间端点 $[a, b]$ ，分组数 $m$ ，组距 $d_i$ （已完成）
  - ②计算分组频数 $\mu_i$ 和频率 $f_i = \mu_i / n$
  - ③画等概频率直方图

以小区间 $[t_{i-1}, t_i)$ 长度为底， $y_i = f_i / d_i$ 为高，画出一排竖着的长方形。

– (3) 作用：可对样本是否来自总体分布为已知的密度函数  $f_0(x)$  进行拟合检验。

- 零假设  $H_0$ ：总体的分布密度函数  $f(x) = f_0(x)$ 。
- 等价形式  $H_0 : p_i = \frac{1}{m} (i = 1, 2, \dots, m)$ ，其中  $p_i = P\{t_{i-1} \leq X < t_i\}$
- 区间分点  $t_i$  满足  $p_i = \int_{t_{i-1}}^{t_i} f_0(x) dx = \frac{1}{m}$
- 取统计量

$$\chi^2 = \sum_{i=1}^m \left( \frac{\mu_i}{n} - p_i \right)^2 \frac{n}{p_i} = \sum_{i=1}^m \frac{(\mu_i - np_i)^2}{np_i}$$

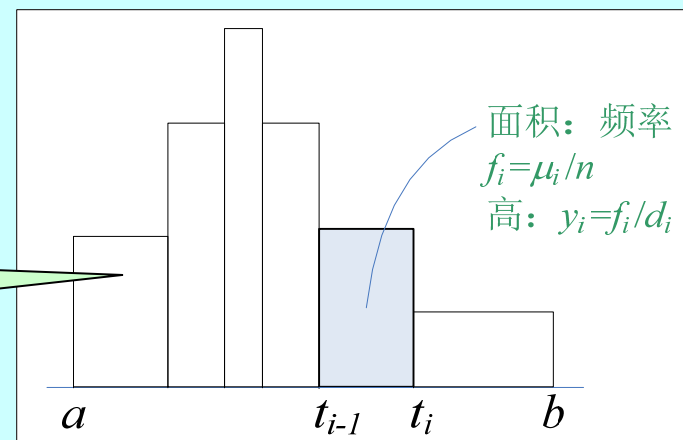


- 在上述 $H_0$ 成立时，统计量

$$V \sim \sum_{i=1}^m \frac{(\mu_i - n/m)^2}{n/m}$$

近似服从于  $m-1$  个自由度的  $\chi^2$  分布，利用统计量  $V$  可以检验总体是否来自已知分布  $f_0(x)$ 。

长方形的面积表示数据落入每个区间的频率



- (三) 累计频率直方图——分布函数的图解法

- (1) 作用：累计频率直方图可用来描述样本经验分布函数  $F_n(x)$ ， $F_n(x)$  可用来近似分布函数  $F(x)$ 。

- (2) 作图：

- 已知实验数据  $x_1, x_2, \dots, x_n$ ，区间端点  $[a, b]$ ，分组数  $m$ ，组距  $d$  的求法同频率直方图。
    - 第  $i$  个小区间  $[t_{i-1}, t_i)$  的频数为  $\mu_i (i=1, 2, \dots, m)$ ，

累积频数为 
$$v_i = \sum_{j=1}^i \mu_j \quad , \quad (i=1, 2, \dots, m)$$

累积频率为 
$$g_i = v_i / n \quad (i=1, 2, \dots, m)$$

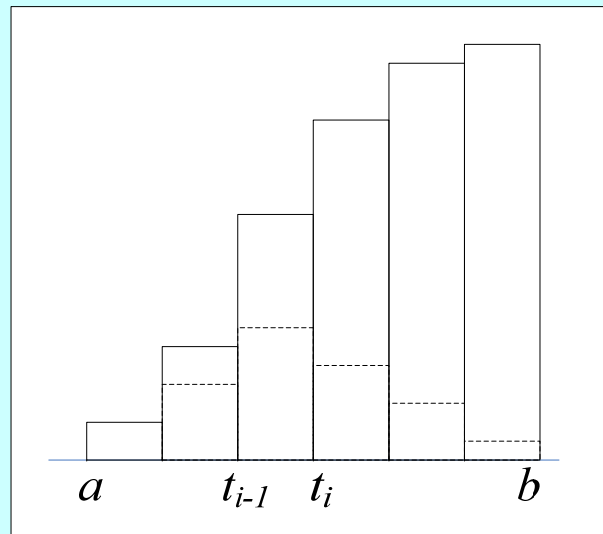
- 画直方图

对每个小区间  $[t_{i-1}, t_i)$  ( $i=1,2,\dots, m$ ) 分别作长方形:

以小区间  $[t_{i-1}, t_i)$  长度为底,

以  $y_i = g_i / d$  为高,

画出一排竖着的长方形即为累计频率直方图。



## § 5 正态性检验

- 正态性检验：检验某随机变量是否服从正态分布。  
即检验观测数据与正态总体差异是否显著。
- 问题：设  $X_1, X_2, \dots, X_n$  是来自总体  $X$  的样本，我们要检验

$H_0$ ：总体  $X$  的分布是正态分布  $N(\mu, \sigma^2)$ 。

- 1.  $\chi^2$  检验法

- 方法:

- 用  $m$  个点 ( $t_1 < t_2 < \dots < t_m$ ) 把实轴分成  $m+1$  段, 分别为:

$$(-\infty, t_1), [t_1, t_2), \dots, (t_m, \infty),$$

- 用  $\mu_i$  表示观测数据落入第  $i$  段 ( $i=1, 2, \dots, m$ ) 的频数,

- $\mu_i / n$  表示频率,

- 用  $p_i$  表示样本来自正态分布总体时落入第  $i$  段的概率，即

$$p_1 = P\{X < t_1\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t_1} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$p_i = P\{t_{i-1} \leq X < t_i\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t_{i-1}}^{t_i} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$(i = 2, \dots, m)$$

$$p_{m+1} = P\{X \geq t_m\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{t_m}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

- 当假设成立而且  $n$  充分大（ $n > 30$ ）时，计算统计量

$$V = \sum_{i=1}^{m+1} \left( \frac{\mu_i}{n} - p_i \right)^2 \frac{n}{p_i} = \sum_{i=1}^{m+1} \frac{(\mu_i - np_i)^2}{np_i}$$

统计量  $V$  是随机变量，当  $n$  充分大时近似的服从  $\chi^2(m)$  分布。

注意：在实际应用中，一般正态分布总体的参数  $\mu, \sigma^2$  未知，需要先求参数： $\bar{x}, s^2$  分别作为  $\mu, \sigma^2$  的估计值；

再进行  $\chi^2$  检验，这时统计量  $V$  近似的服从  $\chi^2(m-2)$  分布。

- 求出满足  $P\{V > \lambda \mid H_0\} = \alpha$  的  $\lambda$ ，其中  $\alpha$  是显著水平，
- 若  $V > \lambda$  就拒绝  $H_0$ ，否则接受  $H_0$ 。

## – 2. 偏峰检验法

- 原理

设正态随机变量  $X$  的偏度为  $g_1$ ，峰度为  $g_2$ ：

$$g_1 = \frac{E(X - E(X))^3}{\sigma^3}, \quad g_2 = \frac{E(X - E(X))^4}{\sigma^4} - 3$$

则  $g_1=0$ ， $g_2=0$ 。

根据样本数据可计算出偏度和峰度的估计量：

$$G_1 = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^3}{s^3}, \quad G_2 = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4}{s^4} - 3$$

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



可以证明，当总体服从于正态分布且样本容量相当大（ $n > 30$ ）时，统计量  $G_1$  和  $G_2$  近似正态分布，且有

$$E(G_1) \approx 0, \quad Var(G_1) \approx 6/n;$$

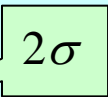
$$E(G_2) \approx 0, \quad Var(G_2) \approx 24/n$$

- 实现:

取检验水平  $\alpha=0.05$

由样本值计算统计量  $G_1$  和  $G_2$

判断: 如果以下不等式

$$\begin{aligned} -2\sqrt{\frac{6}{n}} \leq G_1 \leq 2\sqrt{\frac{6}{n}} \\ -2\sqrt{\frac{24}{n}} \leq G_2 \leq 2\sqrt{\frac{24}{n}} \end{aligned}$$


有一个不成立, 就拒绝  $H_0$ , 认为总体不服从正态分布;  
如果不等式均成立, 就不能否认总体服从正态分布。

- 3. Q-Q图检验法

分位数-分位数 (Quantile-Quantile) 图

- (1) 原理

- 将观测数据从小到大排列，得次序统计量：

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} ,$$

- 则经验分布函数为：

$$F_n(x) = \begin{cases} 0, & \text{当 } x < x_{(1)} \\ \frac{k}{n}, & \text{当 } x_{(k)} \leq x < x_{(k+1)} \\ 1, & \text{当 } x \geq x_{(n)} \end{cases} \quad k = 1, \dots, n-1$$

- 对于正态分布总体，分布函数近似等于样本经验分布函数，即

$$F(x) = P\{X < x\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$
$$= \Phi\left(\frac{x-\mu}{\sigma}\right) \approx F_n(x)$$

从而

$$\frac{x-\mu}{\sigma} = \Phi^{-1}(F_n(x)) @u$$

故有

$$x = \sigma u + \mu$$

在  $Oux$  平面上，表示斜率为  $\sigma$ ，截距为  $\mu$  的直线。

- 根据样本数据在在  $Oux$  平面上画点集  $(u_i, x_{(i)})(i=1,2,\dots, n)$ ，其中

$$u_i = \Phi^{-1}(F_n(x_{(i)}))$$

- 当  $x = x_{(i)}$  时,  $x_{(i)}$  是经验分布函数的样本分位点  
经验分布函数  $F_n(x_{(i)}) = \frac{i}{n}$

实际应用中, 常用  $F_n(x_{(i)}) = \frac{i-0.5}{n}$  (作“连续性”修正)

- 相应的  $u_i = \Phi^{-1}(\frac{i-0.5}{n})$ , 是标准正态分布函数的  $\frac{i-0.5}{n}$  分位点
- 所以称点集  $(u_i, x_{(i)})(i=1,2,\dots,n)$  为分位数-分位数 (Q-Q) 图,  $n$  个点应该近似分布在  $x = \sigma u + \mu$  的直线上。此时样本来自正态总体的假设成立; 否则不成立。

## — (2) 步骤

- 将样本数据从小到大排列，得次序统计量： $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ ，求事件 $\{X \leq x_{(i)}\}$ 的概率 $p_i (i=1,2,\dots, n)$

$$p_i = F_n(x_{(i)}) = \frac{i-0.5}{n}$$

- 对  $p_i$  计算相应的标准正态分位数  $u_i (i=1,2,\dots, n)$
- 把点  $(u_i, x_{(i)})(i=1,2,\dots, n)$  画在平面坐标系上，并考察他们是否在同一条直线上；
- 计算相关系数  $r$ ，并检验其正态性

$$r = \frac{\sum (x_{(i)} - \bar{x})(u_i - \bar{u})}{\sqrt{\sum (x_{(i)} - \bar{x})^2} \sqrt{\sum (u_i - \bar{u})^2}}$$

