

第 1 章 股票市场收益率正态分布 实证方法研究

正态分布是自然界中最重要的分布，它能描述许多随机现象。在一个实际问题中，总体并不一定服从正态分布，如果不顾这个前提成立与否，盲目套用公式，必将影响统计方法的效果，影响结论的可信度。因此，正态性检验是统计方法应用中的重要问题。这个问题在应用上是如此重要，以至在国家标准体系中，也专为其制定一项标准即《GB 4882—85 数据的统计处理和解释——正态性检验》，国家标准局 1985 年批准。

1.1 引言

“正态”意味着“正常状态”，就是说在观察或实验中不出现重大的失误，则结果应服从正态分布，这个看法有大量的经验事实作为支持，也有其理论上的依据（中心极限定理），这大概就是正态分布这个名称的由来。正态分布之所以重要，除了它确实可以常用来描述实际中的观察数据外，另一个重要原因归功于 Fisher 及其同时代的若干学者的研究。他们对正态总体下一系列重要的统计量建立了形式完美、简洁，且在计算上可行的小样本理论，为统计推断提供了极大的方便，而在非正态的情况下则没有可比拟的结果。

基于这后一点原因，人们在实际使用统计分析时，总是乐于采用正态假定。在现代投资理论研究领域，收益率分布函数的确定可以说是整个金融计量研究的基础，许多金融计量模型如资产组合模型、CAPM、APT 以及 Black-Scholes 定价公式等都是以收益率服从正态分布这一假设为前提的。

马柯维兹提出的现代证券组合投资理论（均值-方差理论）将投资收益的不确定性（而非投资者的损失）定义为投资风险，这种不确定性则是用方差或标准差来度量的。但是用方差度量风险的一个条件是要求证券投资收益率及其联合分布是正态的。即该理论的证券收益率都服从正态分布，收益率用期望收益率表示，风险用收益的方差表示。通常用矩法估计证券的期望收益率和方差：

$$\bar{r}_i = \frac{1}{T} \sum_{t=1}^T r_{it}$$

$$\hat{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^T (r_{it} - \bar{r}_i)^2$$

其中 r_{it} 是证券 i 在 t 时刻的收益率, 当 r_{it} 服从正态分布时, 上述估计和最大似然估计是完全一致的, 否则, 方差估计的计算将非常复杂^[1]。由此可见, 证券收益率服从正态分布这一前提假设对均值-方差理论是非常重要的。

CAPM 是 Sharpe 在 Markowitz 均值-方差模型的基础上提出的, 是对 Markowitz 早期工作的一种简化, 也是目前金融经济学中一个重要的理论。CAPM 定义了一个系数 β 用以度量股票与市场一起变动情况下证券收益变动的程度:

$$\beta = \frac{\text{cov}(r_i, r_M)}{\sigma_M^2}$$

其中 r_i 为证券 i 的收益, r_M 为市场组合的收益; σ_M^2 为 r_M 的方差; β 为度量证券对市场资产组合方差的贡献程度。由于引入了 β 系数, 因而大大简化了协方差的计算。 β 值作为市场风险的度量, 是个别资产相对于市场证券组合风险程度的比较结果, 其本质仍是以收益的不确定性作为风险的度量。而这里的风险程度仍是用方差度量的。因此, 和均值-方差模型一样, 证券收益率服从正态分布这一假设前提对于 CAPM 同样也是不可或缺的。

目前, 引入 VAR 是金融市场风险测量的主流方法。所谓一项资产或多项资产的组合的 VAR 是这样一个值: 即在给定概率或置信水平 $1-\alpha$ 前提下, 在持有该资产 t 期间内预计损失额超过该值的概率为 α 。即

$$P(\text{损失绝对值} > \text{VAR}) = \alpha$$

可见 VAR 的计算要求期望收益率的概率密度已知, 或可用已知的分布推导出来。最常用的分布就是正态分布。收益率服从正态分布这一假设大大减轻了 VAR 的繁重的计算负担, 为风险管理者提供了一套功能强大的统计工具。而且执行巴塞尔委员会 96 协议的前提之一是估计 VAR, 并假定标准化的证券收益率服从正态分布。

在证券市场, 股票价格的波动是由众多投资者的股票买卖行为造成的, 价格的最终波动幅度 (或收益) 则是这些众多买卖行为造成价格变动的总和。投资者人数众多、分布范围广, 可以将他们的买卖行为视作同分布、相互独立的随机变量, 而且每一投资者的买卖行为对价格波动的贡献又是很微小的, 由中心极限定理, 股票收益应服从正态分布。另一层含义是, 股票价格的波动是一个维纳 (Wiener) 过程 (Brown 运动), 也就是不可预测的, 这又与有效市场假说 (EMH) 中的弱

式有效假设论点相一致。

可以看出, 价格波动或资产收益率服从正态分布这一假设对现代金融学的各种理论与方法是至关重要的, 因此在实际中, 有必要对所谓的正态性进行检验。

1.2 正态检验方法

长期以来, 在一般统计教科书中, 由于篇幅所限及体例上的习惯因素, 在谈到正态性检验时, 都只是将这个问题放在一般的分布拟合检验项下作简单处理, 没能作仔细讨论, 而且沿袭前苏联的模式, 只介绍通用的检验方法 χ^2 拟合优度检验和柯尔莫哥洛夫检验。本节详细介绍几种在金融经济学中常用的几种正态性检验方法, 并评价各种方法的检验功效。

1.2.1 分位图法

分位图(Quantile-Quantile 图)^[2]是一种直观的、用肉眼评价给定分布偏离正态性的方法。由于第 q 个百分点为这样一个数, 小于它的观察值有百分之 q , 一个 Q-Q 图描绘出了经标准化的观察收益的百分点(观察的百分点)相对于标准正态分布的百分点(正态百分点)的情况。考虑经标准化的观察回报样本, $\bar{r}_t, t=1, 2, \dots, T$ 。记第 j 个观察百分点为 q_j , 这样对所有 T 个观察百分点 $\text{Prob}(\bar{r} < q_j \approx p_j)$ 这里 $p_j = (j-0.5)/T$ 。记第 j 个标准正态的百分点为 $z_j, j=1, 2, \dots, T$ 。例如, 如果 $T=100$ 则 $z_5 = -1.645$ 。计算 Q-Q 描点得到 Q-Q 正态图的步骤如下^[2]。

步骤 1 标准化日收益率, 即由 $r_t, t=1, 2, \dots, T$ 计算 $\bar{r}_t = (r_t - \hat{\mu})/\hat{\sigma}$;

步骤 2 将 $\bar{r}_t$ 排序并计算它们的百分点 $q_j, j=1, 2, \dots, T$;

步骤 3 计算对应于每一个 q_j 的概率 p_j ;

步骤 4 计算对应于每一个 p_j 的标准正态百分点 z_j ;

步骤 5 对每对 $(z_1, q_1), (z_2, q_2), \dots, (z_T, q_T)$ 描点。

例如在计算中选取 $T=100$ 即相当于百分点分位图 即 $\text{Prob}(q_{j-1} < \bar{r}_t < q_j) = 0.01$ 可通过在整个样本(向后 2 年)中将对应的点筛选出来。

一种度量有多少偏离正态发生的方法是计算 Q-Q 描点的相关系数, 即

$$\rho_Q = \frac{\sum_{j=1}^T (q_j - \bar{q})(z_j - \bar{z})}{\left[\sum_{j=1}^T (q_j - \bar{q})^2 \right]^{1/2} \left[\sum_{j=1}^T (z_j - \bar{z})^2 \right]^{1/2}}$$

对于大样本 ρ_Q 要求最小为 0.999 才能通过 5% 显著水平下的正态性检验。

由求做 Q-Q 正态图的基本方法可知：如果 Q-Q 正态图在左边下弯，而在右边上弯，则意味着该分布较正态分布有更长的尾部；如果 Q-Q 正态图呈现明显的“U”形，则意味该分布相对于相应的正态分布而言是有偏的；如果上述特征不明显，则需利用本章介绍的其他方法进一步检验。

封建强^[3]利用 Q-Q 图检验法对上海证券交易所综合指数的日、周、月收益率分布的正态性进行检验，得到的三张 Q-Q 正态图都呈现出明显的“S”形，表明收益率分布具有较长的尾部，但没有出现明显的“U”形。

1.2.2 χ^2 拟合检验

χ^2 拟合是由 K. Pearson(皮尔逊)提出的一种检验母体分布的方法，它能够像显著性检验一样控制犯第一类错误的概率。

设母体 ξ 的分布函数为具有明确表达式的 $F(x)$ ，例如它可以是正态分布族、指数分布族、二项式分布族、泊松分布族等。我们可以把随机变量 ξ 的值域 R 分成 k 个互不相容的区间 $A_1 = [a_0, a_1], A_2 = [a_1, a_2], \dots, A_k = [a_{k-1}, a_k]$ 。这些区间不一定有相同的长度。

设 $x_1, \dots, x_n$ 是容量为 n 的子样的一组观测值。 n_i 为子样观测值 $x_1, \dots, x_n$ 中落入 A_i 中的频数， $n = \sum_{i=1}^n n_i$ ，则 n 次试验中事件 A_i 出现的频率为 $\frac{n_i}{n}$ 。

检验原假设 $H_0: F(x) = F_0(x)$ 设在原假设 H_0 成立下，母体 ξ 落入区间 A_i 的概率为 p_i 即

$$p_i = P(A) = F_0(a_i) - F_0(a_{i-1}), i = 1, \dots, k \quad (1.1)$$

此时 n 个观测值中，恰有 n_1 个观测值落入 A_1 内， n_2 个观测值落入 A_2 内， $\dots$ ， n_k 个观测值落入 A_k 内的概率应为

$$\frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \quad (1.2)$$

这是一个多项分布。

按照大数定理在 H_0 为真时，频率与概率 p_i 的偏差不应太大，根据这个思想，皮尔逊构造了一个统计量

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i} \quad (1.3)$$

称做皮尔逊 χ^2 统计量，它的极限分布就是自由度为 $k-1$ 的 χ^2 分布。

为了能够把皮尔逊 χ^2 统计量用来做检验的统计量，我们必须指导它的抽样分布。这里先讨论 $k=2$ 的简单情形在 H_0 成立下，

$$P(a_1) = p_1, \quad P(a_2) = p_2, \quad p_1 + p_2 = 1$$

这时, 频数 $n_1 + n_2 = n$ 。我们考察

$$\chi^2 = \frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2} \quad (1.4)$$

令

$$Y_1 = n_1 - np_1, \quad Y_2 = n_2 - np_2 \quad (1.5)$$

显然

$$Y_1 + Y_2 = n_1 + n_2 - n(p_1 + p_2) = 0 \quad (1.6)$$

由此可见 Y_1 与 Y_2 不是线性独立的, 且 $Y_1 = -Y_2$ 。于是

$$\chi^2 = \frac{Y_1^2}{np_1} + \frac{Y_2^2}{np_2} = \frac{Y_1^2}{np_1 p_2} = \left[\frac{n_1 - np_1}{\sqrt{np_1(1-p_1)}} \right]^2 \quad (1.7)$$

根据德莫弗-拉普拉斯极限定理 (在 重贝努里试验中, 事件 A 在每次试验中出现的概率为 p , $0 < p < 1$, μ_n 为 n 次试验中事件 A 出现的次数, 则

$$\lim_{n \rightarrow \infty} P\left(\frac{\mu_n - np}{\sqrt{npq}} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt, \text{ 当 } n \text{ 充分大时, 随机变量 } \frac{n_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

分布是接近正态的, 从而推得皮尔逊统计量在 $k=2$ 时的分布, 当 n 充分大时, 是接近于自由度为 1 的 χ^2 分布。

对于一般情形, 有如下定理, 当 H_0 为真时, 即 $p_1, \dots, p_k$ 为母体的真实概率时, 由式 (1.3) 所定义的统计量 χ^2 的渐进分布是自由度为 $k-1$ 的 χ^2 分布, 即其密度函数为

$$f(x) = \begin{cases} \frac{1}{2^{\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right)} x^{\frac{k-3}{2}} e^{-\frac{x}{2}} & x > 0 \\ 0 & x < 0 \end{cases} \quad (1.8)$$

如果原假设 H_0 只确定母体分布的类型, 而分布中还含有未知参数 $\theta_1, \dots, \theta_m$, 则我们还不能用上述定理来作为检验的理论依据。Fisher 证明了如下定理, 从而解决了含未知参数情形的分布检验问题。

下面这个定理给出构造 χ^2 统计量的理论依据, 设 $F(x; \theta_1, \dots, \theta_m)$ 为母体的真实分布 其中 $\theta_1, \dots, \theta_m$ 为 m 个未知参数。在 $F(x; \theta_1, \dots, \theta_m)$ 中用 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 的极大似然估计 $\hat{\theta}_1, \dots, \hat{\theta}_m$ 代替 $\theta_1, \dots, \theta_m$ 。并且以 $F(x; \hat{\theta}_1, \dots, \hat{\theta}_m)$ 取代式 (1.1) 中的 $F(x)$ 得到

$$p_i = F(a_i; \hat{\theta}_1, \dots, \hat{\theta}_m) - F(a_{i-1}; \hat{\theta}_1, \dots, \hat{\theta}_m) \quad (1.9)$$

则将式 (1.9) 代入式 (1.3) 所得的统计量

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \quad (1.10)$$

当 $n \rightarrow \infty$ 时有自由度 $k-m-1$ 的 χ^2 分布。

利用皮尔逊的 χ^2 检验来检验分布假设的步骤总结如下：

(1) 把母体的值域划分为 k 个互不相交的区间 $[a_i, a_{i+1}]$, $i=1, \dots, k$ 其中 a_1, a_{k+1} 可以分别取 $-\infty, +\infty$ (每个划分的区间必须包含不少于 5 个个体, 若个体少于 5 时, 则可把这种区间并入其相邻的区间, 或者把几个频数都小于 5, 但不一定相邻的区间并成一个区间)。

(2) 在 H_0 成立下, 用极大似然估计法估计分布所含的未知参数。

(3) 在 H_0 成立下, 计算理论概率

$$p_i = F_0(a_i + 1) - F_0(a_i)$$

并且算出理论频数 np_i 。

(4) 按照子样观测值 $x_1, x_2, \dots, x_n$ 落在区间 $[a_i, a_{i+1})$ 中的个数, 即实际频数 $n_i, i=1, \dots, k$ 和 (3) 中算出的理论频数 np_i 计算

$$\chi^2 = \frac{(n_i - np_i)^2}{np_i}$$

(5) 按照所给出的显著性水平 α 查自由度 $k-m-1$ 的 χ^2 分布表得 $\chi^2_{1-\alpha}(k-m-1)$ 其中 m 是未知参数的个数。

(6) 若 $\chi^2 \geq \chi^2_{1-\alpha}$ 则拒绝原假设 H_0 。若 $\chi^2 < \chi^2_{1-\alpha}$ 则认为原假设 H_0 成立。

1.2.3 柯尔莫哥洛夫与斯米尔诺夫检验—— D_n 检验

χ^2 拟合检验是比较子样与母体的概率, 尽管它对于离散型和连续型母体分布都适用, 但它依赖于区间划分。因为即使原假设 $H_0: F(x) = F_0(x)$ 不成立, 在某一划分下还是可能有

$$F(a_i) - F(a_{i-1}) = F_0(a_i) - F_0(a_{i-1}) = p_i, \quad i = 1, \dots, k$$

从而影响式 (1.3) 中 χ^2 的值, 也就是有可能把不真的原假设 H_0 接受过来。由此看到, 用 χ^2 检验实际上只是检验了 $F_0(a_i) - F_0(a_{i-1}) = p_i, i=1, \dots, k$ 是否为真, 而并未真正地检验母体分布 $F(x)$ 是否为 $F_0(x)$ 。柯尔莫哥洛夫对连续型母体的分布提出了一种检验方法, 一般称做柯尔莫哥洛夫检验或 D_n 检验。这个检验是比较子样经验分布函数 $F_n(x)$ 和母体分布函数 $F(x)$ 的。它不是在划分区间上考虑 $F_n(x)$ 与原假设的分布函数之间的偏差, 而是在每一点上考虑他们的偏差。这就克服了 χ^2 检验的依赖于区间划分的缺点, 但母体必须假定为连续的。

根据 Glivenko(格里汶科)定理(当样本容量 n 无限增大时, 经验分布函数 $F_n(x)$ 以概率 1 关于 x 一致收敛于 $F_n(x)$ 即 $P(\limsup |F_n(x) - F(x)| = 0) = 1$), 我们可以把子样经验分布函数看作实际母体分布函数的缩影, 如果原假设成立, 它与 $F(x)$ 的距离一般不应太大。由此柯尔莫哥洛夫提出一个统计量

$$D_n = \sup_x |F_n(x) - F(x)| \quad (1.11)$$

并且得到这统计量 D_n 的精确分布和极限分布 $K(\lambda)$ 。它们都不依赖于母体的分布。

柯尔莫哥洛夫定理 设母体 ξ 有连续分布函数 $F(x)$ 从中抽取容量为 n 的子样, 并设经验分布函数为 $F_n(x)$ 则

$$D_n = \sup |F_n(x) - F(x)|$$

的分布函数

$$P\left(D_n < \lambda + \frac{1}{2n}\right) = \begin{cases} 0 & \lambda < 0 \\ \int_{\frac{1}{2n}-\lambda}^{\frac{1}{2n}+\lambda} \int_{\frac{3}{2n}-\lambda}^{\frac{3}{2n}+\lambda} \cdots \int_{\frac{2n-1}{2n}-\lambda}^{\frac{2n-1}{2n}+\lambda} f(y_1, \dots, y_n) dy_1 \cdots dy_n & 0 \leq \lambda \leq \frac{2n-1}{2n} \\ 0 & \lambda \geq \frac{2n-1}{2n} \end{cases} \quad (1.12)$$

其中:

$$f(y_1, \dots, y_n) = \begin{cases} n! & 0 < y_1 < \cdots < y_n < 1 \\ 0 & \text{其他} \end{cases}$$

在 $n \rightarrow \infty$ 时有极限分布函数

$$P(\sqrt{n}D_n < \lambda) \rightarrow K(\lambda) = \begin{cases} \sum_{j=-\infty}^n (-1)^j \exp(-2j^2\lambda^2) & \lambda > 0 \\ 0 & \lambda \leq 0 \end{cases} \quad (1.13)$$

在引用柯尔莫哥洛夫检验时, 应该注意的是, 原假设的分布参数值原则上是已知的, 而且只能用于连续型分布检验。但在参数为未知时, 如果母体分布为正态分布或指数分布时则采用下列两种方法估计: ① 可用另一个大容量子样来估计未知参数 ② 如果原子样容量很大, 也可用来估计未知参数。不过此时 D_n 检验是近似的。在检验时以取较大的显著性水平为宜, 一般取 $\alpha = 0.10 \sim 0.20$ 。

D_n 检验母体有连续分布函数 $F(x)$ 这个假设的步骤如下:

(1) 从母体抽取容量为 n (一般 $n \geq 50$) 得子样, 并把子样观察值按由小到大的次序排列。

(2) 计算经验分布函数

$$F_n(x) = \begin{cases} 0 & x < x_1 \\ \frac{n_j(x)}{n} & x_j \leq x \leq x_{j+1}, j = 1, 2, \dots, n \\ 1 & x_k \leq x \end{cases} \quad (1.14)$$

(3) 在原假设 $H_0(H_0: F(x) = F_0(x))$ 下, 计算观测值处得理论分布函数 $F(x)$

的值。

(4) 对每一个 x 计算经验分布函数与理论分布函数的差的绝对值

$$|F_n(x) - F(x)| \text{ 与 } |F_n(x_{i+q}) - F(x)|$$

(5) 由(4)算出统计量

$$D_n = \sup |F_n(x) - F(x)| = \sup\{|F_n(x_i) - F(x)|, |F_n(x_i) - F(x_i)|, i = 1, \dots, n\}$$

的值。

(6) 给出显著性水平 α ，由柯尔莫哥洛夫检验的临界值表查出

$$P(D_n \geq D_{n,\alpha}) = \alpha \quad (1.15)$$

的临界值；当 $n > 100$ 时，可通过 $D_{n,\alpha} \approx \lambda_{1-\alpha}/\sqrt{n}$ 查 D_n 的极限分布函数数值表得 $\lambda_{1-\alpha}$ 从而求出 $D_{n,\alpha}$ 的近似值。

(7) 若由(5)算出的 $D_n \geq D_{n,\alpha}$ ，则拒绝原假设 H_0 ；若 $D_n < D_{n,\alpha}$ ，则接受原假设，并认为原假设的理论分布函数与子样数据是拟合得好的。

柯尔莫哥洛夫-斯米尔诺夫检验方法在对证券价格收益率分布的正态性方面得到较多的应用。杨朝军^[4]、陶亚民等^[5]利用柯尔莫哥洛夫-斯米尔诺夫拟合优度检验法对上证综合指数收益率的正态性进行检验，样本数据从 1990 年 12 月 19 日至 1998 年 6 月 30 日，分析发现，在排除异常事件干扰的情况下，收益率服从正态分布。

在林勇等^[6]的检验中，日收益率和周收益率的柯尔莫哥洛夫统计量的显著性几乎为零，因此，他们认为上证指数和深成指日收益率和周收益率不服从正态分布；在对月收益率的考察中发现，在 1992.05.22~1996.12.13 期间，上证指数的柯尔莫哥洛夫统计量的显著性水平为 0.032，因此拒绝上证指数服从正态分布的假设；在 1996.12.16~2000.08.31 期间，上证指数的柯尔莫哥洛夫统计量的显著性水平为 0.20，因而可以接受正态性假设。但在所考察的整个时期 1992.05.22~2000.08.31，则拒绝了正态性假设。对于深圳成分指数，也存在相似的现象。不过正好相反，可以认为深成指月收益率第一阶段服从正态分布，而第二阶段却不从正态分布。

1.2.4 偏度、峰度检验

偏度和峰度是为了刻画分布的密度函数曲线的形状而引入的两个概念，偏度和峰度检验是一种更简单有效的正态性检验方法，它是通过检验经验分布曲线的形状来达到正态性检验目的的。即用偏度检验主要是检查母体分布的对称性，用峰度检验主要是检查母体分布的尖削程度。不同均值和方差取值的正态分布密度函数曲线具有固定的偏度和峰度值（分别为 0 和 3），因此可通过子样数据计算

偏度和峰度达到检验其母体是否服从正态分布的目的。

由于偏度和峰度检验是一种简单有效的正态性检验方法, 因此在实际中, 尤其是在对股票价格波动分布的正态性检验方面得到比较广泛的应用。封建强^[3]利用偏度检验法对上海证券交易所综合指数的日、周、月收益率分布的对称性进行检验, 时间间隔越长, 收益率越接近正态分布; 李亚静等^[7]对沪深两地的综合指数与成分指数 1995 年 1 月 3 日至 1999 年 12 月 30 日进行峰度、偏度检验得出两者均不服从正态分布的结论; 高桂芬等^[8]、胡晓华^[9]分别对上海综合指数和成交量(时间: 1997.1~1997.6)利用偏度、峰度检验法对其正态性进行检验, 取显著性水平 $\alpha=0.1$ 两者均服从正态分布 杨朝军^[4]、陶亚民等^[5]取从 1990 年 12 月 19 日至 1998 年 6 月 30 日的上证综合指数为样本数据, 利用偏度、峰度检验法检验其正态性, 得出在排除异常事件干扰的情况下, 收益率服从正态分布。

1. 偏度检验

股票收益率分布的对称性对于现代金融理论是有着深刻含义的。在资产组合理论和 CAPM 中, 都是将风险视作收益的不确定性, 因此用方差来度量风险, 这就意味着投资者对股价向上与向下的波动是同等看待的, 而如果收益率分布是非对称的, 这就从数据(客观)上否定了用方差度量风险大小的合理性。由此可见, 检验股票收益率分布的对称性是非常有意义的, 正是基于这一点, 这里介绍分布对称性的一种检验方法, 偏度检验。

1) 偏度

偏度反映分布的对称性。我们知道, 对称分布的奇数中心矩为零。最简单的奇数中心矩就是三阶中心矩。但是, 中心矩的大小与随机变量取值的分散程度有关, 而且中心矩是有量纲的, 三阶中心矩的量纲是随机变量的量纲的三次方。因此, 通常将三阶中心矩除以标准差 σ 的三次方, 这样得到的标准化的三阶中心矩就称为随机变量的偏度。国外文献中一般用 $\sqrt{\beta_1}$ 表示偏度, 但因为偏度可以取负值, 所以这种记法不符合人们的习惯。我们采用国家标准《正态性检验》所采用的记法 β_s 表示偏度。

左右两边不对称的频数分布, 统计上习惯称为偏态分布。若算术平均数 $\bar{x} >$ 中位平均数 $M_e >$ 众位数 M_o , 频数曲线的尾巴拖向右边, 整个曲线呈右偏态, 即该分布具有正的偏度, $\beta_s > 0$ 如果 $\bar{x} < M_e < M_o$, 频数曲线分布的尾巴拖向左方, 曲线随之呈左偏态, 即该分布具有负的偏度, $\beta_s < 0$ 如果 $\bar{x} = M_e = M_o$ 。表明频数分布是对称的, 此时没有偏态, 或者说偏度等于 0。所谓偏度就是关于频数分布偏斜程度的测量, 它不仅能反映频数分布偏斜程度的大小, 而且还能说明偏斜的方向。

在数理统计中, 以一条分布曲线的二阶中心矩 μ_2 作为衡量方差的标准, 其三

阶中心矩 μ_3 即为衡量偏度的指标。为了给出一个无量纲的纯数，以便比较不同母体的偏度，偏度定义如下：

$$\beta_s = \frac{\mu_3}{(\sqrt{\mu_2})^3} = \frac{\mu_3}{\sigma^3} \quad (1.16)$$

正态分布的 k 阶中心矩

$$\mu_k = E(X - \mu)^k = \begin{cases} 0 & k \text{ 为奇数} \\ 1 \cdot 3 \cdot 5 \cdot \dots \cdot (k-1) \sigma^k & k \text{ 为偶数} \end{cases} \quad (1.17)$$

正态分布的曲线完全对称，三阶中心矩为零，所以偏度 $\beta_s = 0$ 。偏度距零越远，分布曲线偏离对称越远。如果 $\beta_s < 0$ ，分布曲线向负的一方呈“长尾”，称为负偏；若 $\beta_s > 0$ ，分布曲线向正的一方呈“长尾”，称为正偏。

2) 偏度检验

偏度检验是在偏度这个方向上检验总体是否服从正态分布，通常称这个检验为有方向的检验。而其他无这点要求的检验（如 W 检验、）称为无方向检验。

由正态分布的偏度应该为零这一先验信息，因此原假设为

$$H_0: \beta_s = 0 \quad (\text{总体的偏度为 } 0)$$

备择假设为

$$H_1: \beta_s > 0 \text{ 或 } \beta_s < 0 \quad (\text{总体具有正或负偏度})$$

利用样本 $x_1, x_2, \dots, x_n$ 和式 (1.16) 构造偏度统计量

$$b_s = \frac{\sqrt{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}} \quad (1.18)$$

来估计总体偏度 β_s 。

为了减少计算误差，偏度统计量可由下述公式进行计算：

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \\ \sum_{i=1}^n (x_i - \bar{x})^3 &= \sum_{i=1}^n x_i^3 - \frac{3}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i^2 \right) + \frac{2}{n^2} \left(\sum_{i=1}^n x_i \right)^3 \end{aligned}$$

如果总体服从正态分布，可以证明^[10]：

$$E(b_s) = 0$$

即对于正态总体来说，偏度统计量 b_s 是总体偏度 β_s 的无偏估计量。虽然偏度统计量 b_s 的取值在一定程度上反映了总体偏度 β_s 的大小，但由于样本的随机性， b_s 的大小与 β_s 的值没有确定的大小关系。如果总体为正态分布，根据切比雪夫不等式（若随机变量 ξ 有有限方差，则对任何 $\epsilon > 0$ 有 $P\{\xi - E(\xi) \geq \epsilon\} \leq \frac{1}{\epsilon^2} D(\xi)$ ），对任

意正数 ε 有^[11]

$$P(|b_s| > \varepsilon) \leq \frac{D(b_s)}{\varepsilon^2} = \frac{6(n-2)}{\varepsilon^2(n+1)(n+3)}$$

例如：如果 $n=10$, $P(|b_s| > 2) \leq 0.0839$ ；如果 $n=50$, $P(|b_s| > 2) \leq 0.0266$, $P(|b_s| > 2) \leq 0.1065$ ；若 $n=100$, $P(|b_s| > 2) \leq 0.0565$ 。可见取 $|b_s|$ 较大值的可能性很小。所以偏度检验的拒绝域确定如下：

如果原假设不成立而 $H_1: \beta_s > 0$ 成立, b_s 的取值一般较大, 因此拒绝域为 “ $b_s > b_s(1-\alpha)$ ”, $b_s(1-\alpha)$ 为 b_s 的 $1-\alpha$ 分位数；

如果原假设不成立而 $H_1: \beta_s < 0$ 成立, b_s 的取值一般较小, 由于 b_s 的分布关于原点对称, $b_s(\alpha) = -b_s(1-\alpha)$ 因此拒绝域为 “ $b_s < -b_s(1-\alpha)$ ”。

偏度假设的一般步骤为：

步骤 1 根据实际问题中提供的先验信息, 给出备择假设：

$$H_1: \beta_s > 0 \quad \text{或}$$

$$H_1: \beta_s < 0$$

步骤 2 利用样本观测值 $x_1, x_2, \dots, x_n$ 由式 (1.18) 计算统计量 b_s 的值；

步骤 3 由给定的检验水平 α 和样本容量 n 查表得 b_s 的 $1-\alpha$ 分位数 $b_s(1-\alpha)$ ；

步骤 4 作出判断：

(1) 当备择假设为 $H_1: \beta_s > 0$ 时, 若 $b_s > b_s(1-\alpha)$, 则拒绝 H_0 , 否则不拒绝 H_0 ；

(2) 当备择假设为 $H_1: \beta_s < 0$ 时, 若 $b_s < -b_s(1-\alpha)$ 则拒绝 H_0 , 否则不拒绝 H_0 。

2. 峰度检验

1) 峰度

我们知道集中趋势越明显, 离散水平越小, 频数分布的形状就越高耸；而集中趋势越不明显, 离散程度越大, 反映出来的频数分布就显得越加扁平。峰度反映密度函数曲线在众数附近的“峰”的尖削程度。二阶中心矩 μ_2 即方差的大小, 与分布曲线的尖坦性有关。若方差越小, 分布曲线越尖削, 即呈“高而瘦”的形状；如果方差较大, 则分布曲线较平坦, 即呈较“矮而胖”的形状。

在数理统计中, 四阶中心矩为衡量峰度的指标。为了消除方差的影响, 并得到一个无量纲的量, 把四阶中心矩除以方差的平方, 这样得到的标准化的四阶中心矩就称为随机变量的峰度。国外文献中一般用 β_2 表示峰度, 为了与偏度的符号一致 在国家标准《正态性检验》中用 β_k 表示峰度, 即

$$\beta_k = \frac{\mu_4}{\sigma^4} \quad (1.19)$$

对正态分布来说 $\beta_k=3$ 。在说明一个分布的峰度时，通常以正态分布的峰度为标准。如果一个分布的峰度大于 3，分布曲线较尖削，则称该峰度具有过度的峰度；如果一个分布的峰度小于 3，分布曲线较平坦，则称该峰度具有不足的峰度。

2) 峰度检验

类似于偏度检验可以提出峰度检验。偏度检验利用了正态分布偏度为 0 的特性，而峰度检验则要用到正态分布峰度为 3 的特性。因此，利用峰度来检验正态性，实质上是一个关于参数 β_k 的假设检验问题。由正态分布的峰度应该为 3 这一先验信息，原假设为

$$H_0: \beta_k = 3 \quad (\text{总体的峰度为 } 3)$$

备择假设为

$$H_1: \beta_k > 3 \text{ 或 } \beta_k < 3 \quad (\text{总体具有过度或不足的峰度})$$

类似地，利用样本和式 (1.19) 构造峰度统计量

$$b_k = \frac{n \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \quad (1.20)$$

来估计母体峰度。

为了减少计算误差，峰度统计量可由下列公式进行计算：

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \\ \sum_{i=1}^n (x_i - \bar{x})^4 &= \sum_{i=1}^n x_i^4 - \frac{4}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i^3 \right) + \frac{6}{n^2} \left(\sum_{i=1}^n x_i \right)^2 \left(\sum_{i=1}^n x_i^2 \right) - \frac{3}{n^3} \left(\sum_{i=1}^n x_i \right)^4 \end{aligned}$$

如果总体服从标准正态分布，可以证明^[10]：

$$E(b_k) = \frac{3(n-1)}{n+1}$$

从上式可以看出 $E(b_k)$ 仅与样本容量 n 有关，其值小于 3。当 $n \rightarrow \infty$ 时， $E(b_k) \rightarrow 3$ 。对给定样本容量 n ， $E(b_k)$ 是一个常数。当原假设成立时， $|b_k - E(b_k)|$ 取较大值的可能性较小。所以峰度检验的拒绝域确定如下：

如果原假设不成立而 $H_1: \beta_k > 3$ 成立， b_k 的取值一般较大，因此拒绝域为“ $b_k > b_k(1-\alpha)$ ”， $b_k(1-\alpha)$ 为 b_k 的 $1-\alpha$ 分位数；

如果原假设不成立而 $H_1: \beta_k < 3$ 成立， b_k 的取值一般较小，因此拒绝域为“ $b_k < b_k(\alpha)$ ”。

峰度检验的一般步骤为：

步骤 根据实际问题中提供的先验信息 给出备择假设：

$$H_1: \beta_k > 3 \quad \text{或}$$

$$H_1: \beta_k < 3$$

步骤 2 利用样本观测值 $x_1, x_2, \dots, x_n$ 由式 (1.20) 计算统计量 b_k 的值；

步骤 3 由给定的检验水平 α 和样本容量 n 查表得 b_k 的 $1-\alpha$ 分位数 $b_k(1-\alpha)$ ；

步骤 4 作出判断：

(1) 当备择假设为 $H_1: \beta_k > 3$ 时, 若 $b_k > b_k(1-\alpha)$, 则拒绝 H_0 , 否则不拒绝 H_0 ；

(2) 当备择假设为 $H_1: \beta_k < 3$ 时, 若 $b_k < -b_k(1-\alpha)$, 则拒绝 H_0 , 否则不拒绝 H_0 。

3. 偏度统计量 b_s 和峰度统计量 b_k 的分位数

自然界中存在着大量的分布, 其密度函数或分布函数无法用解析式表示。为了掌握这些分布所蕴含的统计规律, 人们采用了许多方法。如随机模拟、Johnson 或 Pearson 分布族拟合、Cornish-Fisher 展开、Gram-Charlier 级数、Edgeworth 级数等。

统计量 b_s 和 b_k 的分布也不能用解析式表达, 需要选用上述方法求得分位数。

1) b_s 的分位数

Johnson 在文献 [12] 中给出了用 Johnson S_U 分布族进行分布拟合时, 参数的计算方法。后来 D'Agostino 圆满地解决了利用这个分布拟合 b_s 的分布的问题^[13]。由此, 只要凭借标准正态分布的分位数表, 就可以算出 $n \geq 8$ 时 b_s 的分位数。

Johnson S_U 分布族变量 X 的分布族

$$Z = \nu + \eta \sinh^{-1} \left(\frac{X - \epsilon}{\lambda} \right) \quad (1.21)$$

其中 Z 为标准正态变量；四个参数各自在范围： $-\infty < \nu < \infty, -\infty < \eta < \infty, -\infty < \epsilon < \infty, -\infty < \lambda < \infty$ 内取值。设 b_s 满足

$$Z = \eta \sinh^{-1} \left(\frac{b_s}{\lambda} \right) \quad (1.22)$$

即

$$Z = \eta \ln \left[\frac{b_s}{\lambda} + \sqrt{\left(\frac{b_s}{\lambda} \right)^2 + 1} \right] \quad (1.23)$$

为了确定参数 η 和 λ 可以先由

$$\frac{1}{2}(\omega^4 + 2\omega^2 + 3) = \beta_k \quad (1.24)$$

$$\lambda \sqrt{\frac{1}{2}(\omega^2 - 1)} = \sigma \quad (1.25)$$

求出参数 ω 和 λ 再由

$$\omega = e^{1/\eta^2} \quad (1.26)$$

求出 η 。设 b_s 的 p 分位数为 $b_s(p)$ 由式(1.22)可求出

$$b_s(p) = \lambda \sinh\left(\frac{Z_p}{\eta}\right) \quad (1.27)$$

其中 Z_p 为 $N(0,1)$ 分布的 p 分位数。

2) b_k 的分位数

b_k 的分位数也是通过随机模拟或分布拟合而获得的。D'Agostino 和 Tietjen^[14] 利用 Pearson 分布族拟合得到 $n \geq 200$ 时 b_k 的 0.01, 0.05, 0.95 和 0.99 分位数; D'Agostino 和 Pearson^[15] 给出了 $20 \leq n \leq 200$ 时关于 b_k 的经验分布的研究结果。它们对 $n=25$, (以下每一间隔为 5) 50, 100, 200 时 b_k 的分布作了随机模拟。然后, 在 $n=25$, (以下每一间隔为 5) 50 时用 Johnson S_U 族分布进行拟合, 在 $n=100, 200$ 时用 Pearson IV 型分布拟合。最后利用 Lagrangian 五点插值法和样条函数加以光滑, 给出了 $20 \leq n \leq 200$ 时 b_k 的分位数曲线族。

需要注意的是, 随机模拟实验证明: 偏度检验和峰度检验的结论具有相对性, 样本容量越大, 同一检验方法识别非正态总体的能力就越强。因此, 如果有可能, 尽量利用较大的样本。

4. 偏度和峰度联合检验

前面介绍的偏度检验和峰度检验是两种有方向的正态检验方法。当仅在偏度方向上怀疑母体偏离正态分布时用偏度检验; 若仅在峰度方向上怀疑总体偏离正态分布时用峰度检验。如果同时在偏度和峰度两个方向上都怀疑总体偏离正态分布, 可以使用偏度和峰度联合检验。

偏度和峰度联合检验的原假设为

$$H_0: \beta_s = 0, \beta_k = 3$$

备择假设为

$$H_1: \beta_s \neq 0, \beta_k \neq 3$$

与前面的方法不同, 联合检验的拒绝域是以 β_s 和 β_k 为坐标的平面图上的一个二维区域, 根据点 (b_s, b_k) 在有关图上的位置对原假设进行检验。详细内容参见文献[9]。检验的步骤为:

- (1) 分别计算统计量 b_s 和 b_k 的值。
- (2) 根据检验水平 α 查有关的图, 在图中找出与样本容量 n 相应的一条曲线, 并把点 $(|b_s|, b_k)$ 画在这张图上。
- (3) 作出判断: 若由 $(|b_s|, b_k)$ 表示的点在上述曲线外, 则拒绝原假设, 否则不拒绝原假设。

1.2.5 W 检验

W 检验由 Shapiro 和 Wilk^[16] 共同提出, 非常适合于小样本的情况, 样本容量一般在 3 至 50 之间, 即 $3 \leq n \leq 50$ 。

如果原假设成立, 即母体的分布服从 $N(\mu, \sigma^2)$ 其中 μ 和 σ^2 均未知, 那么样本 $x_1, x_2, \dots, x_n$ 经变换后, 服从标准正态分布, 将样本观测值按由小到大的次序排列:

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

于是

$$\frac{x_{(1)} - \mu}{\sigma} \leq \frac{x_{(2)} - \mu}{\sigma} \leq \dots \leq \frac{x_{(i)} - \mu}{\sigma} \leq \dots \leq \frac{x_{(n)} - \mu}{\sigma}$$

为标准正态分布的次序统计量。记

$$\xi_1 \leq \xi_2 \leq \dots \leq \xi_i \leq \dots \leq \xi_n$$

为标准正态分布的次序统计量, 这样就有

$$x_{(i)} = \mu + \sigma \xi_i \quad (i = 1, 2, \dots, n)$$

设 $E(\xi_i) = m_i (i = 1, 2, \dots, n)$ 则

$$x_{(i)} = \mu + \sigma m_i + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (1.28)$$

其中 ε_i 是均值为 0 的随机变量。

原假设为真意味着式 (1.28) 成立, 即 m 与 $x_{(i)}$ 呈线性关系。如何定量地表示点 $(m_1, x_{(1)}), (m_2, x_{(2)}), \dots, (m_i, x_{(i)}), \dots, (m_n, x_{(n)})$, 接近直线的程度? 这正是 Shapiro 和 Wilk 研究 W 分布的出发点, 求出这些点的回归直线的斜率 σ 的最佳线性无偏估计。

为了表达方便起见, 记 $y_i = x_{(i)} (i = 1, 2, \dots, n)$ 显然

$$y_1 \leq y_2 \leq \dots \leq y_i \leq \dots \leq y_n$$

令

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & m_1 \\ 1 & m_2 \\ \vdots & \vdots \\ 1 & m_n \end{bmatrix} = (Im), \beta = \begin{pmatrix} \mu \\ \sigma \end{pmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

则式 (1.28) 可以写成

$$Y = X\beta + \varepsilon \quad (1.29)$$

且

$$\begin{aligned} E(\varepsilon) &= \theta \\ \text{Var}(\varepsilon) &= \sigma^2 \hat{V} \end{aligned}$$

其中：

$$\begin{aligned} m &= E([\xi_1, \xi_2, \dots, \xi_n]^T) \\ \hat{V} &= (v_{ij}) \end{aligned}$$

表示相应的协方差矩阵。

由于 ε 不一定相互独立，因此由推广的最小二乘法，参数的最佳线性无偏估计为：

$$\hat{\beta} = (X^T \hat{V} X)^{-1} X^T \hat{V}^{-1} Y$$

即

$$\begin{aligned} \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} &= \left[\begin{pmatrix} I^T \\ m^T \end{pmatrix} \hat{V}^{-1} (I \ m) \right]^{-1} \begin{pmatrix} I^T \\ m \end{pmatrix} \hat{V}^{-1} Y \\ &= \frac{\begin{bmatrix} m^T \hat{V}^{-1} m & -m^T \hat{V}^{-1} I \\ -I \hat{V}^{-1} m & I^T \hat{V}^{-1} I \end{bmatrix} \begin{pmatrix} I^T \\ m^T \end{pmatrix} \hat{V}^{-1} Y}{I^T \hat{V}^{-1} I m^T \hat{V}^{-1} m - (I^T \hat{V}^{-1} m)^2} \end{aligned}$$

所以

$$\hat{\sigma} = \frac{(-I^T \hat{V}^{-1} m I^T + I^T \hat{V}^{-1} I m^T) \hat{V}^{-1} Y}{I^T \hat{V}^{-1} I m^T \hat{V}^{-1} m - (I^T \hat{V}^{-1} m)^2} \quad (1.30)$$

关于对称分布，有

$$I^T \hat{V}^{-1} m = 0 \quad (1.31)$$

故

$$\hat{\sigma} = \frac{m^T \hat{V}^{-1} Y}{m^T \hat{V}^{-1} m} \quad (1.32)$$

由式(1.32)可以看出： $\hat{\sigma}$ 仍是 y_i 的线性函数。为了使线性系数标准化，令

$$b = \frac{R^2}{C} \hat{\sigma}$$

其中： $R^2 = m^T \hat{V}^{-1} m$, $C^2 = m^T \hat{V}^{-1} \hat{V}^{-1} m$ 。又令

$$a^T = (a_1, a_2, \dots, a_n) = \frac{m^T \hat{V}^{-1}}{\sqrt{m^T \hat{V}^{-1} \hat{V}^{-1} m}} \quad (1.33)$$

则

$$b = a^T Y = \sum_{i=1}^n a_i y_i \quad (1.34)$$

b 是回归直线的斜率的最佳线性无偏估计与一个使线性系数标准化的常数的乘积。关于式 (1.31) 的证明可参阅文献 [10] 或 [16]。

设

$$S^2 = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (1.35)$$

它是 $(n-1)\sigma^2$ 的无偏估计。定义统计量

$$W = \frac{b^2}{S^2} = \frac{\left(\sum_{i=1}^n a_i y_i\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1.36)$$

实际应用中, 通常将 a_i 制成表格, 以方便使用。注意

$$-a_i = a_{n+1-i}$$

我们知道, 如果总体服从正态分布时, y 与 $m_i (i=1, 2, \dots, n)$ 之间应该有线性关系。而统计量 W 中并不含有 m_i 只含有 a 和 y , 那么一个自然的问题就是 W 能定量地刻画 y_i 与 m_i 之间的关系吗?

可以证明, 在原假设成立的前提下, a 与 m_i 基本上成正比关系 [15]; 另外 a 与 m 的值都是有表可查的, 通过计算 $\{m_i\}_{i=1}^n$ 与 $\{a_i\}_{i=1}^n$ 的相关系数 r 也可以证实这一点。例如 [16], $n=15$ 时, $r=0.9962$; $n=30$ 时, $r=0.9954$; $n=50$ 时, $r=0.9926$ 。

由式 1.33 得

$$\mathbf{a}^T \mathbf{a} = \sum_{i=1}^n a_i^2 = 1$$

另外, 由式 1.336 可以得出 W 统计量具有关于线性变换的不变性, 即 $W(y_i) = W(c + d y_i)$ 。这样不妨假设 $\bar{y} = 0$ 因而

$$W = \frac{\left(\sum_{i=1}^n a_i y_i\right)^2}{\sum_{i=1}^n y_i^2} = \frac{\left[\frac{\sum_{i=1}^n a_i y_i}{\sqrt{\sum_{i=1}^n a_i^2} \sqrt{\sum_{i=1}^n y_i^2}}\right]^2}{\frac{\sum_{i=1}^n (a_i - \bar{a})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}\right]^2}$$

可见统计量 W 就是 $\mathbf{Y}$ 与 $\mathbf{a}$ 的相关系数的平方。从而得出结论:

(1) 原假设 H_0 成立 $\Rightarrow y_i$ 与 m_i 基本上有线性关系 $\Rightarrow y_i$ 与 a_i 基本上有线性关系 $\Rightarrow W$ 的取值接近于 1。

(2) 给定显著水平 α 如果 $W < W_\alpha$ 拒绝 H_0 。 W_α 为统计量 W 的 α 分位数。

W 检验的最后一个问题就是关于 W 的分布, 或它的分位数的计算问题, 只有解决了这一问题才能决定拒绝域。但是, 对一般的 n , W 的密度函数没有明显的解析式。这种情况下, 一个可以采用的方法是通过计算机进行随机模拟, 用样本分布函数来近似母体分布函数。由于 W 统计量具有关于线性变换的不变性, 因此