Lec13: 非参数统计方法(二)

张伟平

May 4, 2011

§1 拟合优度检验

参数假设检验都是在假定总体是某种具体分布的条件下进行的, 但是这个假设本身不一定成立, 我们可以通过样本$(X_1, \cdots, X_n)$ 来检验它. 一般地, 检验

$$H_0 : X\text{服从某种分布}$$

可以采用Karl Pearson 提出的χ^2 拟合优度检验.

§1.1 离散总体情形

(1) 理论分布不含未知参数的情形

设某总体X 服从一个离散分布, 且根据经验得知总体落在类别$a_1, \cdots, a_k$ 的理论频率分别为$p_1, \cdots, p_k$, 现从该总体抽得一个样本量为n 的样本, 其落在类别$a_1, \cdots, a_k$ 的观测数分别为$n_1, \cdots, n_k$. 感兴趣的问题是检验理论频率是否正确, 即下面假设是否正确:

$$H_0 : P(X \in a_1) = p_1, \cdots, P(X \in a_k) = p_k.$$

这类问题只提零假设而不提对立假设, 相应的检验方法称为拟合优度检验. 显然, 在零假设下, 各类别的理论频数分别为$np_1, \cdots, np_k$, 将理论频数和观测频数列于下表:

类别	a_1	a_2	$\cdots$	a_k
理论频数	np_1	np_2	$\cdots$	np_k
观测频数	n_1	n_2	$\cdots$	n_k

由大数定律知, 在零假设成立时, n_i/n 依概率收敛于p_i, 故理论频数np_i 与观测频数n_i 接近. 而检验统计量取为

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

简单地, 就是

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

其中O 为观测频数, E 为期望频数.

这个统计量中每项的分母的选取有点讲究, 我们可以这样粗略地解释: 假设n_i 服从Poisson分布, 则n_i 的均值和方差均为np_i, 从而$(n_i - np_i)/\sqrt{np_i}$ 的极限分布为标准正态分布, 因此χ^2

近似为k 个服从自由度为1 的χ^2 分布的随机变量之和, 由于$\sum_{i=1}^{k}(n_i - np_i) = 0$, 故这$k$ 个随机变量满足一个约束, 从而χ^2 的自由度为$k-1$. 事实上, 可以严格地证明, 在一定的条件下, χ^2 的极限分布就是自由度为$k-1$ 的χ^2 分布, 但其证明超出本课程的要求范围.

下面给出一个例子来说明拟合优度检验的应用.

例 1. *有人制造一个含6 个面的骰子, 并声称是均匀的. 现设计一个实验来检验此命题: 连续投掷600 次, 发现出现六面的频数分别为97, 104, 82, 110, 93, 114. 问能否在显著性水平0.2 下认为骰子是均匀的?*

解: 该问题设计的总体是一个有6 个类别的离散总体, 记出现六个面的概率分别为$p_1, \cdots, p_6$, 则零假设可以表示为

$$H_0 : p_i = 1/6, i = 1, \cdots, 6.$$

在零假设下, 理论频数都是100, 故检验统计量χ^2 的取值为

$$\frac{(97-100)^2}{100} + \frac{(104-100)^2}{100} + \frac{(82-100)^2}{100} + \frac{(110-100)^2}{100} + \frac{(93-100)^2}{100} + \frac{(114-100)^2}{100} = 6.94,$$

跟自由度为$6-1=5$ 的χ^2 分布的上0.05 分位数$\chi^2_5(0.2) \approx 7.29$ 比较, 不能拒绝零假设, 即可在显著性水平0.2 下认为骰子是均匀的.

例 2. *孟德尔(Mendel)豌豆杂交试验。纯黄和纯绿品种杂交，因为黄色对绿色是显性的，在Mendel第一定律(自由分离定律)的假设下，二代豌豆中应该有75%是黄色的，25%是绿色的。在产生的$n = 8023$个二代豌豆中，有$n_1 = 6022$个黄色，$n_2 = 2001$个绿色。我们的问题是检验这些这批数据是否支持Mendel第一定律，要检验的假设是*

$$H_0 : \quad \pi_1 = 0.75, \quad \pi_2 = 0.25$$

解: 在Mendel第一定律(H_0)下，黄色和绿色的个数期望值为

$$\mu_1 = n\pi_1 = 8023 * 0.75 = 6017.25, \ \mu_2 = n\pi_2 = 8023 * 0.25 = 2005.75$$

则Pearson χ^2统计量为

$$Z = \sum \frac{(O-E)^2}{E} = (6022 - 6017.25)^2/6017.25 + (2001 - 2005.75)^2/2005.75 = 0.015$$

自由度$df = 1$，$p - value$为0.99996. 因此可以认为这些数据服从Mendel第一定律。Fisher基于Mendel的这些数据，发现其数据与理论值符合的太好，$p - value = 0.99996$，但这么好的拟合在几千次试验中才发生一次，因而Fisher断定数据可能有伪造的嫌疑。

(2) 理论分布含若干未知参数的情形

当理论总体总含有未知的参数时, 理论频数np_i 一般也与这些参数有关, 此时应该用适当的估计如极大似然估计代替这些参数以得到p_i 的估计$\hat{p}_i$, 得到的统计量记为

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

拟合优度检验的提出者Karl Pearson 最初认为在零假设下, 检验统计量的χ^2 的极限分布仍等于自由度为$k-1$ 的χ^2 分布, R. A. Fisher 发现自由度应该等于$k-1$ 减去估计的独立参数的个数r, 即$k-1-r$.

例 3. 从某人群中随机抽取*100* 个人的血液, 并测定他们在某基因位点处的基因型. 假设该位点只有两个等位基因A 和a, 这*100* 个基因型中AA, Aa 和aa 的个数分别为*30, 40, 30,* 则能否在*0.05* 的水平下认为该群体在此位点处达到*Hardy-Weinberg* 平衡态?

解: 取零假设为

$$H_0 : \text{Hardy-Weinberg 平衡态成立.}$$

设人群中等位基因A 的频率为p, 则该人群在此位点处达到Hardy-Weinberg 平衡态指的是在人群中3 个基因型的频率分别为$P(AA) = p^2$, $P(Aa) = 2p(1-p)$ 和$P(aa) = (1-p)^2$, 即零假设可等价地写成

$$H_0 : P(AA) = p^2, P(Aa) = 2p(1-p), P(aa) = (1-p)^2.$$

在H_0 下, 3 个基因型的理论频数为$100 \times \hat{p}^2$, $100 \times 2 \times \hat{p}^2(1-\hat{p})$ 和$100 \times (1-\hat{p})^2$, 其中$\hat{p}$ 等于估计的等位基因频率0.5, 代入χ^2 统计量表达式, 得统计量的值等于4. 该统计量的值大于自由度为$3-1-1=1$ (恰好一个自由参数被估计) 的χ^2 分布上0.05 分位数3.84, 故可在0.05 的水平下认为未达到Hardy-Weinberg 平衡态.

§1.2 列联表的独立性和齐一性检验

(1) 独立性检验

下面考虑很常用的列联表. 列联表是一种按两个属性作双向分类的表. 例如肝癌病人可以按所在医院(属性A) 和是否最终死亡(属性B) 分类. 目的是看不同医院的疗效是否不同. 又如婴儿可按喂养方式(属性A, 分两个水平: 母乳喂养与人工喂养) 和小儿牙齿发育状况(属性B, 分两个水平: 正常与异常) 来分类. 这两个例子中两个属性都只有两个水平, 相应的列联表称为"四格表", 一般地, 如果第一个属性有a 个水平, 第二个属性有b 个水平, 称为$a \times b$ 表(见教材p268) . 实际应用中, 常见的一个问题是考察两个属性是否独立. 即零假设是

$$H_0 : \text{属性A 与属性B 独立.}$$

这是列联表的独立性检验问题.

假设样本量为n, 第(i,j) 格的频数为n_{ij}. 记$p_{ij} = P(\text{属性A, B 分别处于水平}i,j)$, $u_i = P(\text{属性A 有水平}i)$, $v_i = P(\text{属性B 有水平}j)$. 则零假设就是$p_{ij} = u_i v_j$. 将u_i 和v_j 看成参数, 则总的独立参数有$a-1+b-1 = a+b-2$ 个. 它们的极大似然估计为

$$\hat{u}_i = \frac{n_{i\cdot}}{n}, \hat{v}_j = \frac{n_{\cdot j}}{n}.$$

正好是它们的频率(证明参看教材) . 其中$n_{i\cdot} = \sum_{j=1}^{b} n_{ij}$, $n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$. 在$H_0$ 下, 第(i,j) 格的理论频数为$n\hat{p}_{ij} = n_{i\cdot}n_{\cdot j}/n$, 因此在$H_0$ 下, $\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij} - n\hat{p}_{ij})$ 应该较小. 故取检验统计量为

$$\chi^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{(n_{i\cdot}n_{\cdot j}/n)}.$$

在零假设下χ^2 的极限分布是有自由度为$k-1-r = ab-1-(a+b-2) = (a-1)(b-1)$ 的χ^2 分布. 对于四格表, 自由度为1.

(2) 齐一性检验

跟列联表有关的另一类重要的检验是齐一性检验, 即检验某一个属性A 的各个水平对应的另一个属性B 的分布全部相同, 这种检验跟独立性检验有着本质的区别. 独立性问题中两属性都是随机的; 而齐一性问题中属性A 是非随机的, 这样涉及到的分布实际上是条件分布. 虽然如此, 所采用的检验方法跟独立性检验完全一样.

例 4. 下面表是甲乙两医院肝癌病人生存情况. 需要根据这些数据判断两医院的治疗效果是否一样.

甲、乙两院肝癌的近期疗效

	生存	死亡	合计
甲院	$150(n_{11})$	$88(n_{12})$	$238(n_{1\cdot})$
乙院	$36(n_{21})$	$18(n_{22})$	$54(n_{2\cdot})$
合计	$186(n_{\cdot 1})$	$106(n_{\cdot 2})$	$292(n)$

解: 这是一个齐一性检验问题. 检验统计量χ^2 的观测值为0.2524, 远远小于自由度为1 的χ^2 分布的上0.05 分位数, 故可以接受零假设, 即在水平0.05 下可以认为两个医院的疗效无差别的.

当有某个格子的频数较小时, 如果允许的话可以合并格子是每个格子的频数足够大, 实际问题中不允许合并格子(合并后失去了实际意义), 此时可以用**Fisher 的精确检验法**.

§1.3 连续总体情形

设$(X_1, \cdots, X_n)$ 是取自总体X 的一个样本, 记X 的分布函数为$F(x)$, 需要检验的那种分布中含有r 个总体参数$\theta_1, \cdots, \theta_r$. 我们要在显著性水平$\alpha$ 下检验

$$H_0 : F(x) = F_0(x; \theta_1, \cdots, \theta_r),$$

其中$F_0(x; \theta_1, \cdots, \theta_r)$ 表示需要检验的那种分布的分布函数. 例如, 当我们要检验

$$H_0 : X \sim N(\mu, \sigma^2)$$

时, $r = 2$, $\theta_1 = \mu, \theta_2 = \sigma^2$.

$$F_0(x; \mu, \sigma^2) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t-\mu)^2\right\} dt.$$

上述假设可以通过适当的离散化总体分布, 采用拟合优度法来做检验. 首先把实数轴分成k 个子区间$(a_{j-1}, a_j]$, $j = 1, \cdots, k$, 其中a_0 可以取$-\infty$, a_k 可以取∞. 这样构造了一个离散总体, 其取值就是这k 个区间. 记

$$p_j \quad = \quad P_{H_0}(a_{j-1} < X \le a_j) = F_0(a_j; \theta_1, \cdots, \theta_r) - F_0(a_{j-1}; \theta_1, \cdots, \theta_r), j = 1, \cdots, k.$$

如果H_0 成立, 则概率p_j 应该与数据落在区间$(a_{j-1}, a_j]$ 的频率$f_j = n_j/n$ 接近, 其中n_j 表示相应的频数. 当p_i 的取值不含未知参数时, 取检验统计量

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j},$$

否则取

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j},$$

其中$\hat{p}_i$ 是将p_i 中的未知参数换成适当的估计后得到的p_i 的估计. 拒绝域取为

$$\left\{\chi^2 > \chi^2_{k-r-1}(\alpha)\right\}.$$

如果p_i 中不含未知参数, 则$r = 0$.

使用χ^2 进行拟合优度检验时一般要求$n \geq 50$, $n\hat{p}_j \geq 5, j = 1, \cdots, k$, 如果不满足这个条件, 最好把某些组作适当合并.

例 5. *从某连续总体中抽取一个样本量为100 的样本, 发现样本均值和样本标准差分别为*-0.225*和*1.282*, 落在不同区间的频数如下表所示:*

区间	$(-\infty,-1)$	$[-1,-0.5)$	$[-0.5,0)$	$[0,0.5)$	$[0.5,1)$	$[1,\infty)$
观测频数	*25*	*10*	*18*	*24*	*10*	*13*
理论频数	*27*	*14*	*15*	*14*	*13*	*17*

可否在显著性水平0.05 下认为该总体服从正态分布?

解: 设理论正态分布的均值和方差分别为μ 和σ^2, 记第i 个区间为$(a_{i-1}, a_i,\ i = 1, \cdots, 6$, 则样本落在第$i$ 个格子的理论概数为$100P(a_{i-1} < X \leq a_i)$, 其中$X \sim N(\mu, \sigma^2)$. 将$\mu = -0.225$ 和$\sigma = 1.282$ 代入得到估计的理论频数, 列于上表中.

$$H_0 : \text{总体服从正态分布}$$

由此算得检验统计量χ^2 的值约为9.34, 与自由度为5 的χ^2 分布的上0.1 分位数$\chi^2_5(0.1) \approx 9.24$ 比较可以拒绝零假设, 即可以在显著性水平0.1 下认为该总体不服从正态分布.

§2 其他常用检验方法

一、 柯尔莫哥洛夫检验

尽管Pearson χ^2检验对任何类型的分布检验都可以用. 不过对于连续型的随机变量, 柯尔莫哥洛夫检验效果更好些. 这是因为Pearsonχ^2检验依赖于把$(-\infty, +\infty)$分成r个区间的具体划分方法, 包括r的选择和区间的位置. 前苏联著名数学家柯尔莫哥洛夫1933年提出了一种新的关于总体分布的拟合优度检验方法—柯尔莫哥洛夫检验(简称柯氏检验法) .

设$r.v.\ X$的分布函数$F(x)$未知, $X_1, \cdots, X_n$为从F中抽取的简单随机样本, $F_0(x)$为给定的某个分布函数. 我们来研究下列检验问题:

$$H_0 : F(x) = F_0(x). \tag{2.1}$$

首先从样本出发求出$F(x)$的经验分布函数如下:

$$F_n(X) = \begin{cases} 0, & x \leq X_{(1)}; \\ k/n, & X_{(k)} < x \leq X_{(k+1)};\ k = 1, 2, \cdots, n-1 \\ 1, & x > X_{(n)}. \end{cases} \tag{2.2}$$

这里$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$是样本$X_1, \cdots, X_n$的次序统计量. $F_n(x)$的性质见§1.3 三.

令检验统计量为

$$D_n = \sup_{-\infty<x<+\infty} |F_n(x) - F_0(x)|, \tag{2.3}$$

D_n常称为F_n与F_0之间的柯氏距离. 由Glirenko-Cantelli定理知, 如果H_0成立, 则$P(\lim\limits_{n\to\infty} D_n = 0) = 1$. 换言之, 如果$H_0$成立, n又较大, D_n的值倾向于取小值. 如果D_n值太大, 倾向于否定H_0.即检验可叙述为: 当$D_n \ge c$时否定H_0, c为临界值, 待定. 其拟合优度的计算公式如下: 在有了具体样本后, 计算出D_n的具体值D_0,则概率

$$p(D_0) = P(D_n \ge D_0 | H_0) \tag{2.4}$$

就是在柯氏距离下, 样本$X_1, \cdots, X_n$与理论分布$F_0(x)$的拟合优度. 若指定一个阀值α(亦称检验水平), 则需定出一个常数$D_{n,\alpha}$,使得

$$p\left(D_{n,\alpha}\right) = P(D_n \ge D_{n,\alpha} | H_0) = \alpha, \tag{2.5}$$

则当$D_n > D_{n,\alpha}$时否定H_0,不然就接受H_0,这就是柯氏拟合优度检验. 当n较小时, $D_{n,\alpha}$已制成表, 见附表13.

Pearson χ^2检验与柯尔莫哥洛夫检验的比较: 大体上可以这样说: 在总体X为一维且理论分布为完全已知的连续分布时, 柯尔莫哥洛夫检验优于χ^2检验. 这是因为: (i) Pearson χ^2统计量之值依赖于把$(-\infty, +\infty)$分为r个区间的具体分法, 包括r的选取和区间的位置, 柯氏距离$\sup|F_n - F_0|$则没有这个依赖性. (ii) 一般说来柯氏方法鉴别力强. 也就是说, 在F_0不是总体X的分布时, 用柯氏检验法较容易发现.

另一方面, Pearson χ^2检验也有它的优点: (i) 当总体X是多维时, 处理方法与一维一样, 极限分布的形式也与维数无关. (ii) 尤其重要的是: 对于理论分布包含未知参数时, χ^2检验容易处理, 但柯氏方法处理起来很难.

二、 斯米尔洛夫检验

设$X_{i1}, \cdots, X_{in_i}$为抽自具有一维连续分布总体F_i的简单随机样本,$i = 1, 2$,且合样本独立. 设$F_1(x)$, $F_2(x)$是未知的两个连续函数. 考虑检验问题

$$H_0 : F_1(x) = F_2(x), \quad -\infty < x < +\infty. \tag{2.6}$$

设F_{1n_1}和F_{2n_2}分别记这两组样本的经验分布函数, 令

$$D^+_{n_1,\,n_2} = \sup_{-\infty<x<+\infty} \left(F_{1n_1}(x) - F_{2n_2}(x)\right),$$

$$D_{n_1,\,n_2} = \sup_{-\infty<x<+\infty} \left|F_{1n_1}(x) - F_{2n_2}(x)\right|.$$

前苏联数学家斯米尔洛夫(Smirnov)于1936年证明了下列结果:

定理6.6.2 设原假设(2.6)成立, 则有

$$\lim_{\substack{n_1\to\infty \\ n_1\to\infty}} P\Big(\sqrt{\frac{n_1 n_2}{n_1+n_2}} D^+_{n_1,\,n_2} \le x\Big) = \begin{cases} 1 - e^{-2\lambda^2}, & \text{当} x > 0; \\ 0, & \text{当} x \le 0, \end{cases}$$

$$\lim_{\substack{n_1\to\infty \\ n_1\to\infty}} P\Big(\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,\,n_2} \le x\Big) = K(x),$$

其中$K(x)$与(??)同. $D^{+}_{n_1,\,n_2}$和$D_{n_1,\,n_2}$分别称为单边和双边的斯米尔洛夫检验统计量.

如果要检验的原假设是(2.6), 取$D_{n_1,\,n_2}$作为检验统计量, 则当$D_{n_1,\,n_2} > D_{n_1,\,n_2;\alpha}$时否定$H_0$. 临界值

$$D_{n_1,\,n_2;\alpha} = \lambda \Big/ \sqrt{\frac{n_1 n_2}{n_1+n_2}},$$

其中λ的值可由附表14查出. 这就是斯米尔洛夫检验.

若假设检验问题为

$$H_0 : F_1(x) \le F_2(x) \longleftrightarrow K : F_1(x) > F_2(x), \quad x \in (-\infty, \infty),$$

则用$D^{+}_{n_1,n_2}$作为检验统计量.

三、 正态性检验*

在实际工作中常常要检验一个随机变量是否服从正态分布, 这叫做正态性检验. 前面介绍的Pearson χ^2检验、柯氏检验法等当然可以使用. 但是由于上述方法是通用的, 适用面广, 故有针对性不强的缺点. 这些方法都没有充分利用原假设成立时的信息, 检验功效不高. 对正态分布往往可以找到针对这类特定分布功效较高的检验. 下面介绍的两种基于次序统计量的正态性检验: 小样本(样本大小在3–50之间)的W 检验和大样本(样本大小在50–1000之间)的D检验可以克服上述缺点, 提高检验的功效. 这两个方法已被列入我国统计方法的国家标准GB4882-85之中, 见参考文献[9].

1. W检验(Wilk检验)

考虑检验问题:

$$H_0: \ X\text{服从正态分布} \longleftrightarrow H_1: \ X\text{不服从正态分布}. \tag{2.7}$$

设$X_1, \cdots, X_n$为来自正态总体$X \sim N(\mu, \sigma^2)$的样本, $X_{(1)} \le \cdots \le X_{(n)}$为其次序统计量. 设$Y_i = (X_i - \mu)/\sigma$, $i = 1, \cdots, n$, 则$Y_1, \cdots, Y_n$ i.i.d. $\sim N(0,1)$. 令

$$\begin{aligned} Y_{(i)} &= \frac{X_{(i)} - \mu}{\sigma}, \quad e_i = X_{(i)} - E(X_{(i)}), \\ m_i &= E(Y_{(i)}), \quad i = 1, 2, \cdots, n. \end{aligned}$$

注意$m_1, \cdots, m_n$是与μ, σ^2无关的确定的数. 显然有

$$X_{(i)} = \mu + \sigma m_i + e_i, \quad i = 1, 2, \cdots, n, \tag{2.8}$$

其中$e = (e_1, \cdots, e_n)'$是均值为0, 协方差阵为V的n维向量.

作一直角坐标系, 横轴表示$X_{(i)}$,纵轴表示m_i.由(2.8)可见,在这个坐标系中$(X_{(1)}, m_1)$, $(X_{(2)}, m_2)$, $\cdots$, $(X_{(n)}, m_n)$ 应该大致成一条直线, 微小的差别是由随机误差e_i造成的. 怎样判别n个点是否近似在一条直线上呢? 我们可以计算一下$\mathbf{X} = (X_1, \cdots, X_n)'$ 和$\mathbf{m} = (m_1, \cdots, m_n)'$ 之间的相关系数R,

$$R^2 = \frac{\left(\sum\limits_{i=1}^{n} (X_{(i)} - \bar{X})(m_i - \bar{m}) \right)^2}{\sum\limits_{i=1}^{n} (X_{(i)} - \bar{X})^2 \sum\limits_{i=1}^{n} (m_{(i)} - \bar{m})^2}.$$

显然$0 \le R^2 \le 1$,当R^2越接近1, $\mathbf{X}$与$\mathbf{m}$的线性关系越明显. 因此当H_0成立, 诸X_i服从$N(\mu, \sigma^2)$时, R^2接近1. 可见当$R^2 < c$ (c为较小的正数, 待定)时倾向于否定H_0.

由于$N(0,1)$是对称分布，所以$(Y_{(1)}, \cdots, Y_{(n)})$与$(-Y_{(n)}, \cdots, -Y_{(1)})$ 有相同的联合分布，从而$Y_{(k)}$与$-Y_{(n+1-k)}$同分布, 故$m_k = m_{n+1-k},\ k = 1, \cdots, n,\ \overline{m} = 0$,于是

$$R^2 = \frac{\Big(\sum\limits_{i=1}^{n} m_i X_{(i)}\Big)^2}{\sum\limits_{i=1}^{n}(X_{(i)} - \overline{X})^2 \sum\limits_{i=1}^{n} m_i^2} = \frac{\Big[\sum\limits_{i=1}^{[n/2]} b_i (X_{(n+1-i)} - X_{(i)})\Big]^2}{\sum\limits_{i=1}^{n}(X_{(i)} - \overline{X})^2}, \tag{2.9}$$

此处$b_i = m_{n+1-i} \big/ \sqrt{\sum_{i=1}^{n} m_i^2}$.因此$W = R^2$ 可作为检验统计量.

夏皮诺(Shapiro)和威尔克(Wilk)对(2.9)作了修正得到检验统计量(详见参考文献[4] P_{294}):

$$W = \Big\{ \sum_{i=1}^{[n/2]} a_i \big(X_{(n+1-i)} - X_{(i)}\big) \Big\}^2 \Big/ \sum_{i=1}^{n} \big(X_{(i)} - \overline{X}\big)^2 \tag{2.10}$$

在$n \le 50$时, $\{a_i :\ i \le [n/2]\}$的值已制成表, 详见附表15. 公式(2.10)可以用来简化统计量W的计算.

可以证明,检验统计量W的一个重要性质: 即在正态假设H_0成立时, W的分布仅与样本容量n有关(详见[4]中引理5.5.4). 因而在讨论有关统计量W的问题时无妨假定样本来自$N(0,1)$分布.

如前所述, W是n个数对之间的相关系数的平方, 因此$0 \le W \le 1$.由线性模型理论可知在正态假设H_0下, 这n个数对之间基本上存在线性关系, 故W取值应接近于1. 因此, 给定检验水平α后, 检验问题(2.7)的W检验是

$$\text{当} W \le W_\alpha \text{时, 否定} H, \text{否则接受} H. \tag{2.11}$$

其中W按公式(2.10)计算,临界值W_α可由附表16查出. 附表16是根据W的分布仅与样本容量n有关的这个性质, 利用随机模拟法编制而成的.

例6.6.2　为了检验一批煤灰砖中各块砖的抗压强度的变化是否服从正态分布, 从中随机取10块得抗压强度数(由小到大排列) 为:

$$57,\ 66,\ 74,\ 77,\ 81,\ 87,\ 91,\ 95,\ 97,\ 109$$

试检验这些数据是否与正态分布相等?($\alpha = 0.05$)

解　将数据填入下表

i	$x_{(i)}$	$x_{(11-i)}$	$x_{(11-i)} - x_{(i)}$	a_i
1	57	109	52	0.5739
2	66	97	31	0.3291
3	74	95	21	0.2141
4	77	91	14	0.1224
5	81	87	6	0.0399

其中, a_i这一列的值由附表15根据$n=10$查得. 经计算得

$$\sum_{i=1}^{10} x_{(i)} = 834, \qquad \frac{1}{10}\sum_{i=1}^{10} x_{(i)} = \overline{x} = \frac{1}{10}834 = 83.4\,,$$
$$\sum_{i=1}^{10}(x_{(i)}-\overline{x})^2 = \sum_{i=1}^{10} x_{(i)}^2 - 10\overline{x}^2 = 71736 - 10\times 6955.56 = 2180.4\,,$$
$$\sum_{i=1}^{5} a_i(x_{(11-i)} - x_{(i)}) = 46.494, \qquad \Big[\sum_{i=1}^{5} a_i(x_{(11-i)} - x_{(i)})\Big]^2 = 2161.692$$

于是有

$$W = \Big[\sum_{i=1}^{5} a_i(x_{(11-i)} - x_{(i)})\Big]^2 \Big/ \sum_{i=1}^{10}(x_{(i)}-\overline{x})^2 = \frac{2161.7}{2180.4} = 0.99$$

由$\alpha = 0.05,\ n=10$,查附表16得$W_{0.05} = 0.842 < 0.99 = W$,所以不能拒绝正态性假定.

2. D**检验**

检验问题和样本仍如W检验中所述. W检验是有效的, 可惜它只适用于样本容量$3 \le n \le 50$的样本. 当$n>50$时很难计算附表15中的相应的值. 为此人们提出了D检验. 达戈斯底纳(Dagostino)建议在$n>50$时用

$$D = \frac{\sum_{i=1}^{n}(i - \frac{n+1}{2})X_{(i)}}{(\sqrt{n})^3\sqrt{\sum_{i=1}^{n}(X_{(i)}-\bar{X})^2}} \tag{2.12}$$

作为检验统计量. 由此导出的检验方法, 称为D检验.

可以证明, 在正态假设H_0成立时, D的分布仅与样本容量n有关, 且

$$E(D) \approx 0.28209479, \quad \sqrt{Var(D)} \approx 0.02998598/\sqrt{n}$$

将D标准化得

$$Y = \frac{\sqrt{n}(D - 0.28209479)}{0.02998598}$$

可以证明: 当正态性假定H_0成立, 且$n\to\infty$时有

$$Y \xrightarrow{\mathscr{L}} N(0,1)$$

但是统计量Y趋向于标准正态分布的速度很慢, 以致于$n=100$时, Y的分布与标准正态分布仍有不可忽略的偏差. 故Dagostina 用随机模拟法获得Y的分位数值(见附表17) .

大量模拟表明, 在H_0成立时, Y的值集中在零左右, 在正态性假定不成立时, Y的值不是偏小就是偏大, 因此检验问题(2.7)水平为α的D检验是

$$\text{当}Y \le Y_{1-\alpha/2}\text{或}Y \ge Y_{\alpha/2}\text{时, 否定}H\text{;否则就接受}H. \tag{2.13}$$

其中$Y_{\alpha/2}$和$Y_{1-\alpha/2}$分别是Y的上侧$\alpha/2$和$1-\alpha/2$ 分位数, 其值可从附表17查出.