

法律声明

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



数理统计

目录

第一章	统计量与抽样分布
第二章	点估计
第三章	区间估计
第四章	假设检验
第五章	分布的检验
第六章	Bayes统计与统计判决理论
第七章	非参数统计
第八章	抽样调查

分布的检验

到目前为止，我们所讨论的估计和检验问题，均假设总体分布属于某个特定的类型(如正态分布)，而对其未知参数或某些数字特征(如期望与方差)进行的。一个自然的问题是，我们怎么知道总体的分布属于某种类型呢？在许多情况下，我们均是利用样本来判断总体的分布，这就是所谓的分布拟合问题。

K. Pearson 于1900年发表的关于 χ^2 拟合优度(goodness-of-fit)检验的论文，常被学者们认为是近代数理统计学的发端。这篇论文对统计学应用中一个常见的重要问题，即一组随机观察数据可否合理地看成是来自一个分布完全已知的总体，提出了一个判定准则。自此之后，有关拟合优度检验得到了极大的发展，其中一个很有代表性的结果就是Kolmogorov于上世纪三十年代提出的基于经验分布的检验准则。有关拟合优度检验的详细内容，请参见杨振海(1994)。

本章主要介绍某些常见的拟合优度检验，如 χ^2 拟合优度检验，Kolmogorov检验及几个正态性检验方法。

目 录

5.1 正态概率纸检验法

5.2 拟合优度检验

5.3 列联表的独立性检验

5.4 Kolmogorov检验

5.5 正态性检验

正态概率纸检验法

设有 n 个来自分布 $F(x)$ 的观测值 $X_1, \dots, X_n$, 根据以往经验或其它方法, 我们对总体分布有一定认识, 此时我们感兴趣的为如下的假设

$$H_0 : F(x) \equiv F_0(x), \quad (6.1.1)$$

其中 $F_0(x)$ 为一形式完全已知的分布, 但可能包含有未知参数. 现常用的概率纸检验方法有: 正态概率纸、Weibull概率纸和对数正态概率纸等, 即(6.1.1)中的 F_0 为这些分布的CDF. 本节将只讲正态概率纸的构造及其应用, 其余的构造方法完全类似.

正态概率纸的构造

我们知道，正态分布的累积分布函数同其它随机变量的CDF一样，是一条介于0与1之间的单增曲线(见图6.2)，虽然不同分布之间的趋势有所不同，但肉眼很难分辨它们之间的区别。如果能将这些曲线“拉直”，则我们可以很容易地区分它们。

一般的正态分布函数

$$y = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad (6.1.2)$$

在我们熟知的直角坐标系的图形如图6.2所示。

正态概率纸的构造

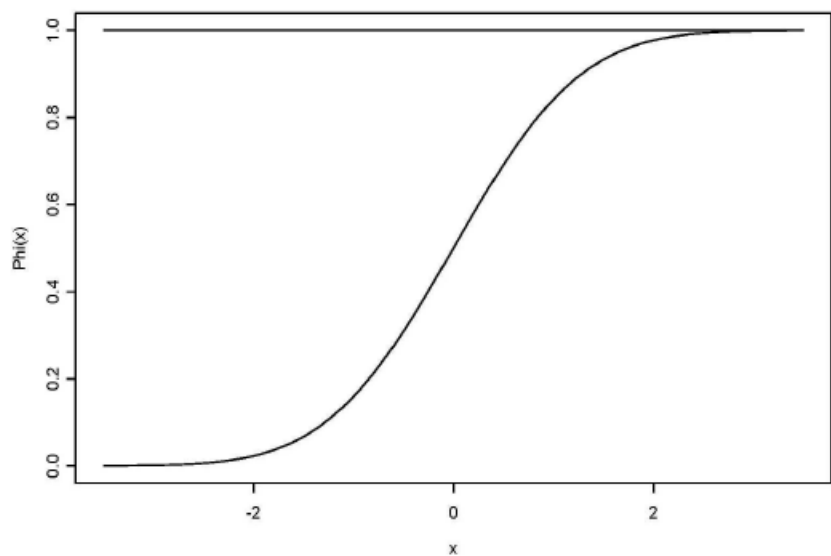


图6.2 标准正态分布函数曲线

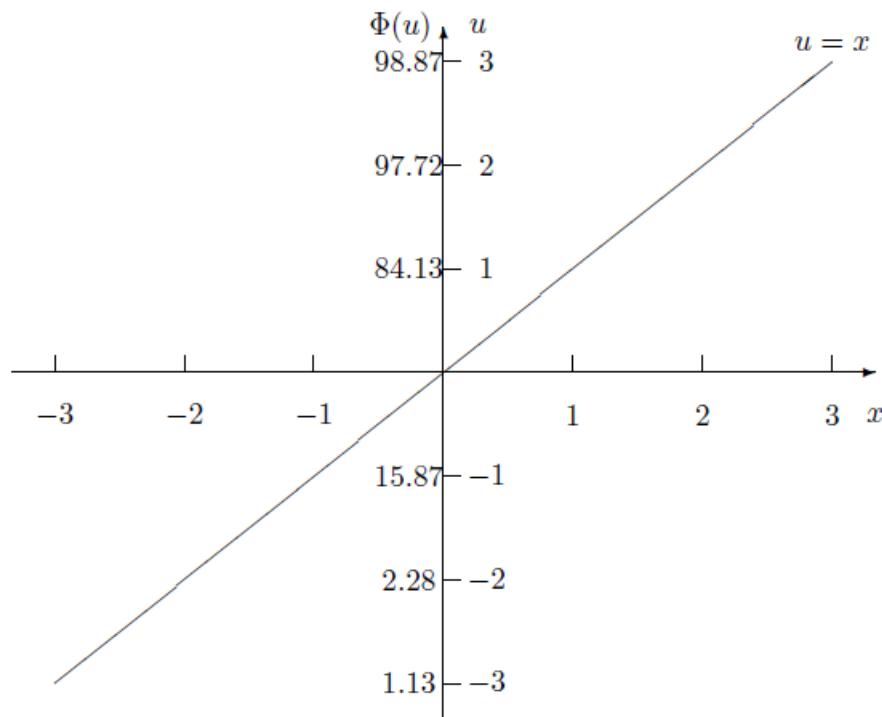


图6.3 变换刻度后的 $(x, \Phi(x))$ 图

正态概率纸的构造

显然，在图6.2中 y 关于 x 是一曲线，但如果我们做如下变换：

$$\begin{cases} y = \Phi(u), \\ u = \frac{x-\mu}{\sigma}, \end{cases}$$

则知 u 关于 x 是线性的，而 y 与 u 是一一对应的。于是，我们可以在 (x, u) 坐标系上，不改变横坐标 x 的刻度，而在 u 轴(即纵坐标)上标上相应 $\Phi(u)$ 的刻度，则得到了 x 与 $\Phi(u)$ 的一条直线了。当然，由于 $\Phi(u)$ 是非线性的，故此时纵坐标的刻度是非等距的。如果 $\mu = 0, \sigma = 1$ ，则此时的 $(x, \Phi(x))$ 的图形就如图6.3所示的一条直线，只是纵坐标的刻度不均匀而已。这就是一张正态概率纸。

我们注意到

$$\lim_{u \rightarrow -\infty} \Phi(u) = 0, \quad \lim_{u \rightarrow \infty} \Phi(u) = 1,$$

故纵轴刻度为0和1的点在概率纸上无法标出。所以，大家所用的正态概率纸的纵轴刻度均是从0.01到99.99。反之，如在正态概率纸上有一条上升的直线，则它就是某正态分布的分布函数的图形。

正态概率纸的应用

当样本 $X_1, \dots, X_n$ 来自分布函数 $F(x)$ 时, 如果 n 个点 $(X_i, F(X_i))$ 或 $(X_{(i)}, F(X_{(i)}))$ 在正态概率纸上呈现一条直线时, 我们就有充分理由认为此组样本来自正态分布. 但现在的问题是: 由于总体分布 $F(x)$ 是未知的, 故我们无法将这些点画出来. 但是, 我们注意到, 如果 $\xi \sim F(x)$, 且 F 连续, 则 $F(\xi) \sim U(0, 1)$. 于是, 我们有

$$E(F(X_i)) = \frac{i}{n+1},$$

这样, 我们可以用 $\frac{i}{n+1}$ 来估计 $F(X_{(i)})$ 的值, 且在正态概率纸上画如下点:

$$\left(X_{(i)}, \frac{i}{n+1} \right), \quad i = 1, 2, \dots, n$$

当这 n 个点在正态概率纸上呈现一条直线时, 我们就有理由认为这些数据服从正态分布.

正态概率纸的应用

根据这些想法，我们应用正态概率纸检验正态假设的具体步骤如下：

- 把这 n 个数据 $X_1, \dots, X_n$ 由小到大排列如下： $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$.
- 在正态概率纸上画上 n 个点 $(X_{(i)}, \frac{i}{n+1})$, $i = 1, 2, \dots, n$.
- 如果这 n 个点呈现一条直线，则接受正态性假设.

如果检验之后，有理由认为这组数据服从正态分布时，一个自然的问题就是其两个参数的估计值如何？从正态概率纸的构造来看，当 $y = \Phi(u) = 0.5$ 时， $u = (x - \mu)/\sigma = 0$. 故对应着纵坐标上取值为50%的点的横坐标即为总体均值 μ 的估计. 类似地，当 $u = 1$ 时， $y = 84.13\%$ ，故直线上对应着 $y = 84.13$ 的点的横坐标就是 $\mu + \sigma$ 的估计值. 于是，就可以得到 μ 与 σ 的估计了. 当然，由于直线的不精确性，故这种估计是初步的，且无法判断它的优良性.

QQ图

现假定总体为正态分布 $N(\mu, \sigma^2)$, 对于样本 $x_1, x_2, \dots, x_n$, 其顺序统计量是 $x_{(1)}, x_{(2)}, \dots, x_{(n)}$. 设 $\Phi(x)$ 是标准正态分布 $N(0, 1)$ 的分布函数, $\Phi^{-1}(x)$ 是反函数, 对应正态分布的 QQ 图是由以下的点

$$\left(\Phi^{-1} \left(\frac{i - 0.375}{n + 0.25} \right), x_{(i)} \right), \quad i = 1, 2, \dots, n \quad (3.14)$$

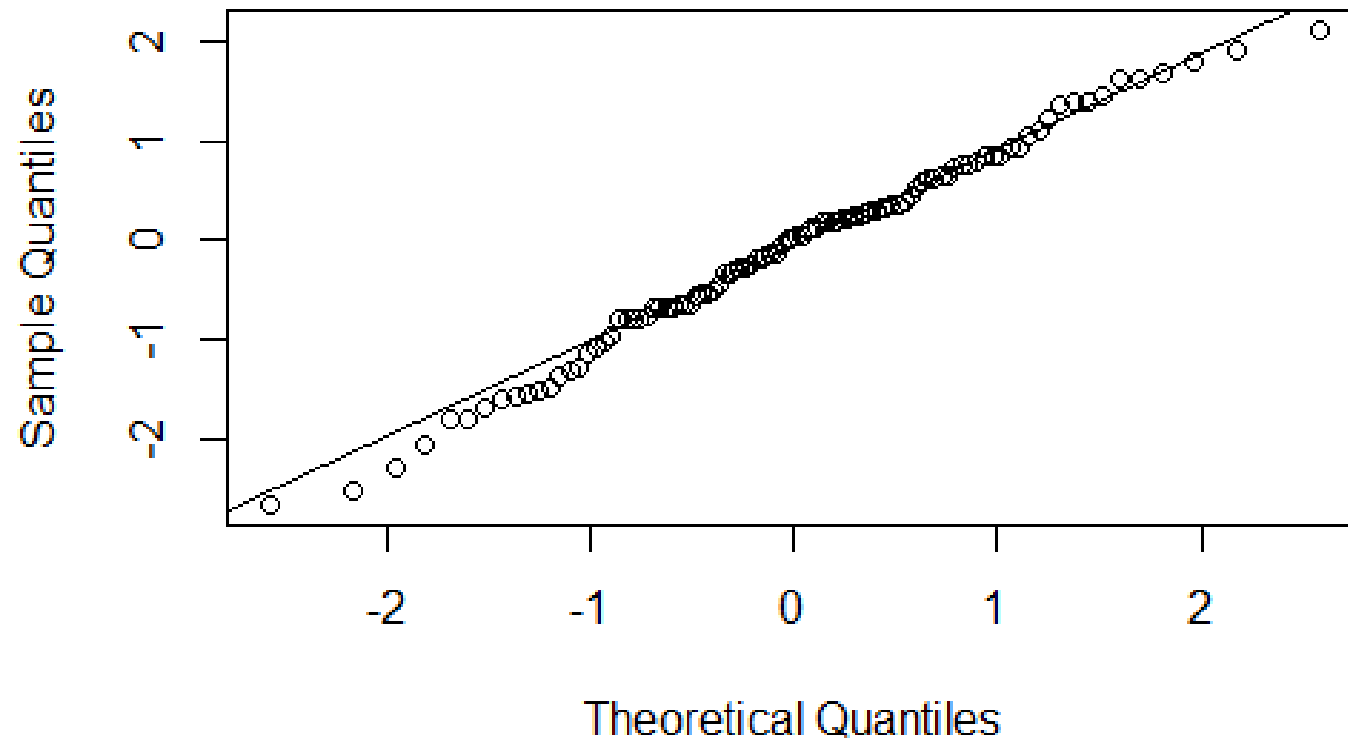
构成的散点图. 若样本数据近似于正态分布, 在 QQ 图上这些点近似地在直线

$$y = \sigma x + \mu$$

附近. 此直线的斜率是标准差 σ , 截距是均值 μ . 所以利用正态 QQ 图可以作直观的正态性检验. 若正态 QQ 图上的点近似地在一条直线附近, 可以认为样本数据来自正态分布总体.

QQ图

Normal Q-Q Plot



目 录

5.1 正态概率纸检验法

5.2 拟合优度检验

5.3 列联表的独立性检验

5.4 Kolmogorov检验

5.5 正态性检验

分类数据的拟合优度检验

例 6.2.1 在19世纪, Mendel按颜色与形状把豌豆分为四类: 黄圆、青圆、黄皱和青皱. Mendel在他观测的556颗豌豆中, 这四类豌豆的颗数分别为315, 108, 101, 32. 于是, Mendel判断这四类的比例为9:3:3:1. 请问这个比例符合遗传规律吗?

上一例子是关于分类数据的检验问题, 它的一般情形为: 根据某项指标, 总体被分成 r 类: $A_1, \dots, A_r$. 此时我们最关心的是关于各类所占的比例的假设

$$H_0: \text{第} i \text{ 类} A_i \text{ 所占的比例为} p_i, i = 1, \dots, r, \quad (6.2.1)$$

其中 $\sum_{i=1}^r p_i = 1$. 记 $X_1, \dots, X_n$ 为从此总体抽出的 n 个iid样本, 且以 n_i 记这 n 个样本中属于 A_i 的样本个数. 由于当 H_0 成立时, 在 n 个样本中属于 A_i 类的“理论个数”或“期望个数”为 np_i , 而我们实际观测到的值为 n_i , 故, 当 H_0 成立时, n_i 与 np_i 应相差不大. 于是, K. Pearson提出用统计量

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - np_i)^2}{np_i} \quad (6.2.2)$$

来衡量“理论个数”与实际观测值间的差别, 并且其拒绝域为 $\{\chi^2 \geq c\}$.

分类数据的拟合优度检验

为了控制上述检验的第一类错误，我们必须知道此检验统计量的零分布，为此，K. Pearson证明了如下定理.

定理 6.2.1 在 H_0 成立且 p_i 均已知时，我们有

$$\chi^2 \xrightarrow{\mathcal{L}} \chi^2(r-1). \quad (6.2.3)$$

分类数据的拟合优度检验

于是，对于假设(6.2.1)，我们可以采取如下的水平近似为 α 的显著性检验：

- 检验统计量如(6.2.2)所示，即

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

- 拒绝域为

$$W = \{\chi^2 \geq \chi_{\alpha}^2(r-1)\}. \quad (6.2.4)$$

这就是K. Pearson提出的最早的一个检验方法，我们称之为Pearson χ^2 拟合优度检验.

分类数据的拟合优度检验

对于例6.2.1中的数据，我们可以做如下的 χ^2 拟合优度检验. 注意到，此时

$$n = 556, n_1 = 315, n_2 = 108, n_3 = 101, n_4 = 32,$$

待检验的假设为

$$H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}.$$

由于

$$\chi^2 = 0.47 < 7.81 = \chi_{0.05}^2(4 - 1),$$

故在水平0.05下，我们没有理由拒绝 H_0 ，即可以认为Mendel的结论是可接受的.

分类数据的拟合优度检验

从以上内容可以看出,对于下面一般的分布假设

$$H_0 : F(x) \equiv F_0(x), \quad (6.2.5)$$

我们仍可利用上述的 χ^2 拟合优度检验,其中 $F_0(x)$ 为一个完全已知的分布函数(形式及参数均已知).

此时,我们可以把 $(-\infty, \infty)$ (或样本空间)分成 r 个互不相交的区域:

$$(-\infty, \infty) = \cup_{i=1}^r I_i = (-\infty, a_1) \cup [a_1, a_2) \cup \cdots \cup [a_{r-1}, \infty), \quad (6.2.6)$$

且以 n_i 记落在第 i 个区间 I_i 内的样本个数,再记

$$\begin{aligned} p_1 &= F(a_1), & p_{10} &= F_0(a_1), \\ p_2 &= F(a_2) - F(a_1), & p_{20} &= F_0(a_2) - F_0(a_1), \\ \dots & & \dots & \\ p_r &= 1 - F(a_{r-1}), & p_{r0} &= 1 - F_0(a_{r-1}), \end{aligned}$$

则我们可以利用统计量

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - np_{i0})^2}{np_{i0}} \quad (6.2.7)$$

及定理6.2.1的结论来检验假设(6.2.5).

分类数据的拟合优度检验

注 6.2.1 虽然上述的 χ^2 拟合优度检验可以用来检验一般的分布假设(6.2.5), 但通过上面的分析不难看出, 此时我们检验的假设仅为

$$H_0 : p_i = p_{i0}, i = 1, 2, \dots, r \quad (6.2.8)$$

而不是真正想检验的假设(6.2.5). 显然(6.2.8)与(6.2.5)有着一定的区别. 如果分点选的不是很好, χ^2 拟合优度检验可能会把两个有一定差别的分布检验为没有区别(参见图6.4.给出的两个累积分布函数)

注 6.2.2 在一般情形下, 分点的选取应保证落在每个区间内的样本点个数不小于5, 且总的样本量不应小于30.

注 6.2.3 通过上面分析, 我们可以看到, 当 F_0 中含有未知参数时, 上述拟合优度检验无法实施.

分类数据的拟合优度检验

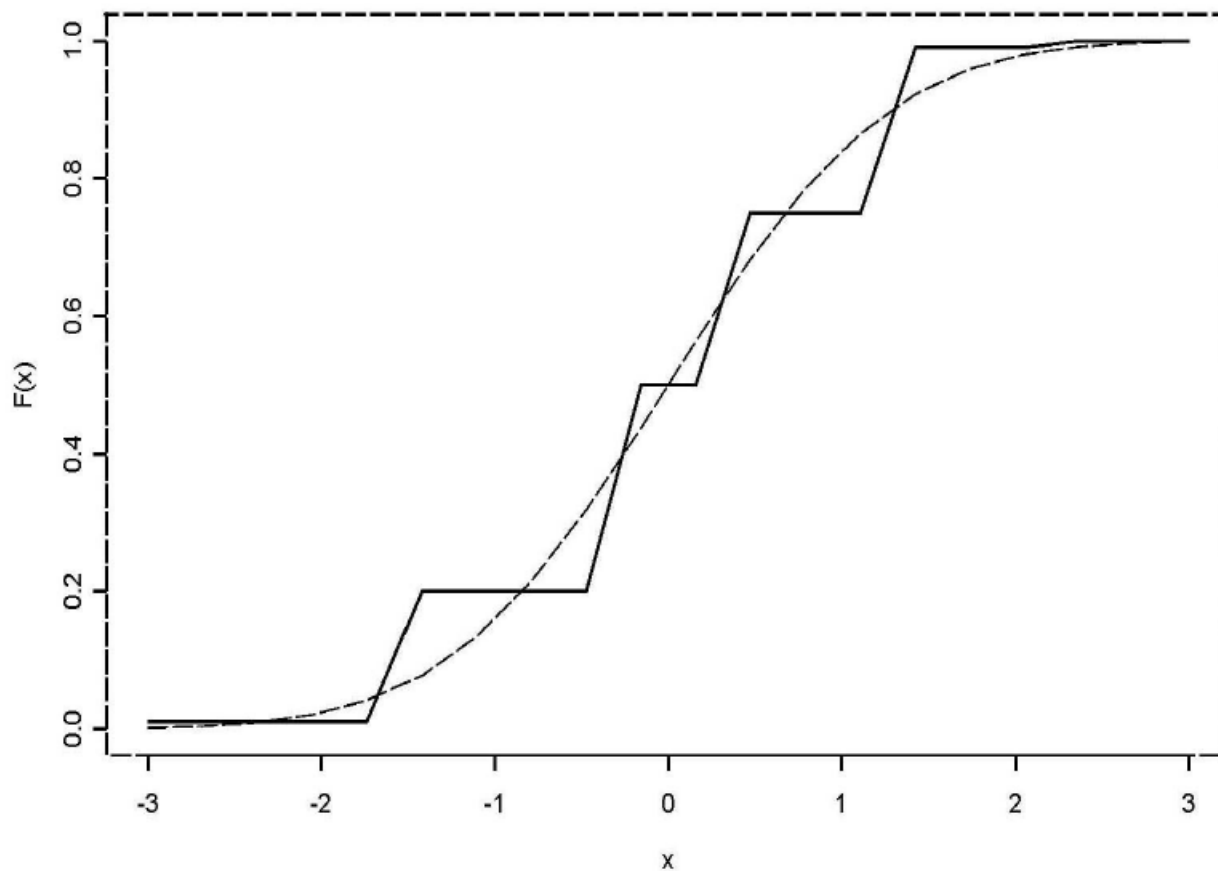


图6.4 针对连续情形的 χ^2 拟合优度检验

带有未知参数拟合优度检验

在许多实际问题中，我们感兴趣的假设可能为

$$H_0 : F(x) \equiv F_0(x; \theta_1, \dots, \theta_k), \quad (6.2.9)$$

其中 $F_0(x; \theta_1, \dots, \theta_k)$ 是依赖于 k 个未知参数的形式已知的分布，如一般的正态分布，二项分布，Poisson分布等。此时，我们无法计算(6.2.2)或(6.2.7)式的 χ^2 统计量。Pearson在其1900年的论文中注意到了这个问题，并且认为：用样本估计未知参数后，可以得到 $\hat{p}_{i0}$ 的值，之后可利用下面的统计量

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}} \quad (6.2.10)$$

作为检验统计量，且当 H_0 成立及 $n \rightarrow \infty$ 时，仍然成立 $\chi^2 \xrightarrow{\mathcal{L}} \chi^2(r-1)$ 。

带有未知参数拟合优度检验

但是, Fisher在上世纪二十年代发现了上述结论是错误的. Fisher指出, 当 H_0 成立时, (6.2.10)式的 χ^2 统计量在一定条件下有如下的极限零分布

$$\chi^2 = \sum_{i=1}^r \frac{(n_i - n\hat{p}_{i0})^2}{n\hat{p}_{i0}} \xrightarrow{\mathcal{L}} \chi^2(r-1-k), \quad (6.2.11)$$

Fisher还指出, 并不是所有的参数估计都能满足(6.2.11)式的要求, 比如一般情况下的矩估计等, 而MLE是近似可行的(请参见陈希孺(1997)). 故对于此类问题, 我们将采用MLE先估计未知参数, 之后再利用(6.2.11)的结论进行检验.

带有未知参数拟合优度检验

例 6.2.2 现在某地区随机地抽取1000人，检查其性别与色盲的数据见表6.1. 按照遗传学规律，这些数据具有如表6.2的规律. 请问这些数据符合遗传学规律吗？

表6.1 性别与色盲数据

	男	女
正常	442	514
色盲	38	6

表6.2 性别与色盲模型

	男	女
正常	$p/2$	$p^2/2 + pq$
色盲	$q/2$	$q^2/2$

带有未知参数拟合优度检验

解 对于本问题, 我们所需检验的假设为

$$H_0 : p_1 = \frac{p}{2}, p_2 = \frac{p^2}{2} + pq = \frac{p(2-p)}{2}, p_3 = \frac{1-p}{2}, p_4 = \frac{(1-p)^2}{2},$$

其中参数 p 是未知的, 故需要估计. 由于当 H_0 成立时, 数据(442, 514, 38, 6)服从一个多项分布, 故其似然函数为

$$L(p) \propto \left(\frac{p}{2}\right)^{442} \left(\frac{p(2-p)}{2}\right)^{514} \left(\frac{1-p}{2}\right)^{38} \left(\frac{(1-p)^2}{2}\right)^6 = p^{956}(2-p)^{514}(1-p)^{50}/2^{1000},$$

于是, 由 $\partial L(p)/\partial p = 0$ 可求得 p 的MLE为 $\hat{p} = 0.91$. 由此可算得 χ^2 的值为

$$\chi^2 = \sum_{i=1}^4 \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 3.056 < 5.99 = \chi_{0.05}^2(4 - 1 - 1),$$

因此, 在水平0.05下, 我们没有理由拒绝 H_0 , 即认为此地区的数据符合遗传学规律. □

带有未知参数拟合优度检验

例 6.2.3 某工厂生产一种滚珠，现随机地抽取了50件产品，测得其直径为（单位：mm）：

15.0, 15.8, 15.2, 15.1, 15.9, 14.7, 14.8, 15.5, 15.6, 15.3, 15.1, 15.3, 15.0, 15.6, 15.7, 15.8, 14.5,

14.9, 14.9, 15.2, 15.0, 15.3, 15.6, 15.1, 14.9, 14.2, 14.6, 15.8, 15.2, 15.9, 15.2, 15.0, 14.9, 14.8,

15.1, 15.5, 15.5, 15.1, 15.1, 15.0, 15.3, 14.7, 14.5, 15.5, 15.0, 14.7, 14.6, 14.2, 14.2, 14.5

问：滚珠直径是否服从正态分布？

带有未知参数拟合优度检验

解 设滚珠直径为 X , 其分布函数为 $F(x)$, 现感兴趣的假设为

$$H_0 : F(x) \in \left\{ \Phi \left(\frac{x - \mu}{\sigma} \right) : \mu \in R, \sigma > 0 \right\}.$$

对于此问题, 我们首先由数据求得 μ, σ^2 的MLE为: $\hat{\mu} = 15.1, \hat{\sigma}^2 = 0.4325^2$. 另外, 由于50个数据的最小值为14.2, 最大值为15.9, 故我们取分点为

$$a_0 = 14.05, a_1 = 14.35, a_2 = 14.65, a_3 = 14.95, a_4 = 15.25, a_5 = 15.55, a_6 = 15.85, a_7 = 16.15$$

由此, 可利用公式

$$\hat{p}_i = \Phi \left(\frac{a_i - 15.1}{0.4325} \right) - \Phi \left(\frac{a_{i-1} - 15.1}{0.4325} \right), \quad i = 1, \dots, 7$$

求得

$$\begin{aligned} \hat{p}_1 &= 0.0414, \hat{p}_2 = 0.1077, \hat{p}_3 = 0.2154, \\ \hat{p}_4 &= 0.2710, \hat{p}_5 = 0.2154, \hat{p}_6 = 0.1077, \hat{p}_7 = 0.0414. \end{aligned}$$

于是, 由于

$$\chi^2 = 1.7284 < 9.49 = \chi_{0.05}^2(7 - 1 - 2),$$

故我们不能拒绝 H_0 .

□

带有未知参数拟合优度检验

综合前面的内容，我们可以把 χ^2 拟合优度检验统计量概括如下：

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}, \quad (6.2.12)$$

其中 O_i, E_i 分别表示观测值和期望值.

目 录

5.1 正态概率纸检验法

5.2 拟合优度检验

5.3 列联表的独立性检验

5.4 Kolmogorov检验

5.5 正态性检验

列联表的独立性检验

比较例6.2.2与以前其它例子,我们发现,例6.2.2的数据与众不同. 对于例6.2.2,我们有两个指标(性别、色盲)对这些数据进行分类,而其它例子则只有一个指标值,对于这样的具有几个分类指标的数据表,我们就称之为列联表(Contingency Table). 对于例6.2.2,共有两个分类指标,我们称之为二维列联表. 一般的二维列联表数据如表6.3所示.

表6.3 二维列联表数据

	B_1	B_2	$\cdots$	B_s	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	n

其中 $n_{i\cdot} = \sum_{j=1}^s n_{ij}$, $n_{\cdot j} = \sum_{i=1}^r n_{ij}$. 此时, 指标 A, B 分别有 r, s 个水平, 且以 n_{ij} 表示在 n 个样本中属于 $A_i \cap B_j$ 的样本个数.

列联表的独立性检验

此时我们感兴趣的并不是如在例6.2.2中的分布假设, 而是考虑这两个指标之间是否独立, 即如下的假设:

$$H_0: \text{指标} A \text{ 与 } B \text{ 独立.} \quad (6.3.1)$$

如记

$$p_{ij} = P\{X \in A_i \cap B_j\}, \quad i = 1, \dots, r, \quad j = 1, \dots, s, \quad (6.3.2)$$

则这 n 个样本可以看成是来自多项分布 X 的样本.

如再记

$$p_{i\cdot} = P\{X \in A_i\}, \quad i = 1, 2, \dots, r, \quad (6.3.3)$$

$$p_{\cdot j} = P\{X \in B_j\}, \quad j = 1, 2, \dots, s, \quad (6.3.4)$$

则有

$$p_{i\cdot} = \sum_{j=1}^s p_{ij}, \quad p_{\cdot j} = \sum_{i=1}^r p_{ij}. \quad (6.3.5)$$

列联表的独立性检验

另外，我们还要注意到如下两个约束

$$\sum_{i=1}^r p_{i\cdot} = \sum_{j=1}^s p_{\cdot j} = 1. \quad (6.3.6)$$

当 H_0 成立时，应有 $p_{ij} = p_{i\cdot} p_{\cdot j}$. 于是，(6.3.1)式的零假设等价于如下假设：

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, s. \quad (6.3.7)$$

由于我们可以把上述列联表数据看作是多项分布的样本，故我们可以用 χ^2 拟合优度检验对其独立性假设(6.3.7)进行显著性检验. 但我们注意到，由于 $p_{i\cdot}$ 和 $p_{\cdot j}$ 是未知的，且有两个约束(6.3.6)，故当 H_0 成立时共有 $r + s - 2$ 个未知参数. 此时，其未知参数的极大似然估计为

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}, \quad i = 1, \dots, r, \quad j = 1, \dots, s.$$

列联表的独立性检验

于是，对于二维列联表的 χ^2 拟合优度检验统计量为

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{n_{i.}n_{.j}}, \quad (6.3.8)$$

且当 H_0 成立及 $n \rightarrow \infty$ 时，我们有

$$\chi^2 \xrightarrow{\mathcal{L}} \chi^2(rs - r - s + 2 - 1) = \chi^2((r - 1)(s - 1)), \quad (6.3.9)$$

于是，此检验的近似水平为 α 的检验拒绝域为

$$\{\chi^2 \geq \chi_{\alpha}^2((r - 1)(s - 1))\}.$$

对于例6.2.2中的数据，由于 $r = s = 2, n = 1000, \chi^2 = 27.4 > 6.64 = \chi_{0.01}^2(1)$ ，故我们有理由拒绝 H_0 ，即认为色盲与性别之间存在着显著性的相关关系。

列联表的独立性检验

注 6.3.1 当分类指标为连续变量时，我们仍然可以先用前面的离散化方法对变量离散后，再用 χ^2 拟合优度检验对二者的独立性进行检验。

注 6.3.2 当带有未知参数时，Pearson认为与参数已知的极限零分布相同，而Fisher则从 $r = s = 2$ 的特殊列联表情况指出了此错误。

注6.3.3 我们还可以把上述针对二维列联表的 χ^2 拟合优度检验推广到高维列联情况，并且在列联表数据进行分析时，也要注意其维数的重要性。

列联表的独立性检验

我们可以将上述结果推广到多维的情况. 下面考虑三维列联表的独立性检验问题. 假设一个观测包含三个指标: A, B, C , 且分别有 r, s, t 个水平, 且以 n_{ijk} 表示在 n 个样本中属于 $A_i \cap B_j \cap C_k$ 的样本个数. 此时感兴趣的假设可能有三类: 三个指标相互独立; 一个指标和另两个指标独立; 一个指标给定的条件下, 另两个指标独立. 前两个类似于二维列联表的独立性检验问题, 故从略. 下面考虑第三类检验问题. 假设此时我们感兴趣的假设为

$$H_0: \text{给定指标 } C \text{ 下, 指标 } A \text{ 与 } B \text{ 独立.} \quad (6.3.10)$$

基于二维统计量的导出方法, 我们可以得到此时的似然比统计量为:

$$\Lambda = \frac{\prod_{i=1}^r \prod_{j=1}^s \prod_{k=1}^t \left(\frac{n_{i \cdot k} n_{\cdot j k}}{n n_{\cdot \cdot k}} \right)^{n_{ijk}}}{\prod_{i=1}^r \prod_{j=1}^s \prod_{k=1}^t \left(\frac{n_{ijk}}{n} \right)^{n_{ijk}}}, \quad (6.3.11)$$

且当 H_0 成立及 $n \rightarrow \infty$ 时, 我们有

$$-2 \ln \Lambda \xrightarrow{\mathcal{L}} \chi^2(t(r-1)(s-1)). \quad (6.3.12)$$

列联表的独立性检验

例 6.3.1 表6.4是1976年-1977年美国佛罗里达州我凶杀案件中，326个被告的肤色和死刑判决情况:

表6.4 美国某州的凶杀案二维数据表

被告	死刑		合计
	是	否	
白人	19	141	160
黑人	17	149	166
合计	36	290	326

表6.5 美国某州的凶杀案三维数据表

被告	被害人	死刑		死刑比例
		是	否	
白人	白人	19	132	0.126
	黑人	0	9	0.000
黑人	白人	11	52	0.175
	黑人	6	97	0.058

请问死刑与肤色有关吗？

列联表的独立性检验

解 从表6.4可以看出, 白人被判死刑的比例为 $19/160 = 11.9\%$, 而黑人被判死刑的比例为 $17/166 = 10.2\%$, 由此可见白人被判死刑的比例高于黑人, 说有肤色歧视是站不住脚的, 并且其似然比统计量 $2 \ln \lambda = 0.2214$ 比 $\chi^2(1)$ 的中位数0.455还要小, 故我们更无法认为这里有肤色歧视.

但是, 当我们把被害人的肤色也考虑进去(其数据如表6.5所示), 其情况就大不一样了. 从表6.5可以看出,

- 如果被害人是白人, 则白人被告与黑人被告被判死刑的比例分别为12.6%和17.5%;
- 如果被害人是黑人, 则白人被告与黑人被告被判死刑的比例分别为0%和5.8%.

从此可以看出, 无论被害人是白人还是黑人, 黑人被判死刑的比例都比白人被判死刑的比例高. 这就说明死刑判决与肤色有关. 事实上, 当我们把上述似然检验应用于这个三维的列联表时, 其结论与此吻合. (之所以三维列联表的结论与二维时的不同, 其原因是当我们把三维数据看成二维时, 会损失掉许多信息) □

目 录

5.1 正态概率纸检验法

5.2 拟合优度检验

5.3 列联表的独立性检验

5.4 **Kolmogorov**检验

5.5 正态性检验

Kolmogorov检验

由于样本经验分布函数 $F_n(x)$ 是 $F(x)$ 的一个很好的估计, 故我们知道, 当(6.2.5)式的 H_0 成立时, $F_n(x)$ 与 $F_0(x)$ 应相差不大, 于是, 我们可以用统计量

$$D_n = \sup_x |F_n(x) - F_0(x)|, \quad (6.4.1)$$

来衡量 $F(x)$ 与 $F_0(x)$ 间的差别, 且拒绝域为 $\{D_n \geq c\}$. 此检验是由Kolmogorov于1933年提出的, 并给出了 D_n 的极限零分布(见定理1.2.3), 于是, 我们称之为Kolmogorov检验. 对于 D_n 的精确分布, 请见下面的定理.

Kolmogorov检验

定理 6.4.1 如果 $F_0(x)$ 连续, 则当 H_0 成立时, 有

$$P\left\{D_n < \lambda + \frac{1}{2n}\right\} = \begin{cases} 0, & \lambda < 0, \\ \int_{\frac{1}{2n}-\lambda}^{\frac{1}{2n}+\lambda} \int_{\frac{3}{2n}-\lambda}^{\frac{3}{2n}+\lambda} \cdots \int_{\frac{2n-1}{2n}-\lambda}^{\frac{2n-1}{2n}+\lambda} f(\mathbf{x}) d\mathbf{x}, & 0 \leq \lambda < \frac{2n-1}{2n}, \\ 1, & \lambda \geq \frac{2n-1}{2n}, \end{cases} \quad (6.4.2)$$

$$\text{其中 } f(\mathbf{x}) = \begin{cases} n!, & 0 < x_1 < \cdots < x_n < 1, \\ 0, & \text{否则.} \end{cases}$$

Kolmogorov检验

虽然定理6.4.1及定理1.2.3的结论很复杂,但是现已有现成的分位数表可供大家查找.

从 D_n 的定义可以看出,对于某些连续型分布检验,Kolmogorov检验优于 χ^2 拟合优度检验.但是,它的计算看起来比较复杂.然而,由于 $F(x)$ 及 $F_n(x)$ 均是单增的,且 $F_n(x)$ 是一阶梯函数,于是其上确界仅可能在 n 个点 $x_{(i)}$ 处达到,即(如果 F_0 连续)

$$D_n = \max_{1 \leq i \leq n} \left\{ F_0(x_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F_0(x_{(i)}) \right\}. \quad (6.4.3)$$

如果 F_0 是右连续的,则 D_n 可以写成如下便于计算的公式:

$$D_n = \max_{1 \leq i \leq n} \left\{ F_0(x_{(i)} - 0) - \frac{i-1}{n}, \frac{i}{n} - F_0(x_{(i)}) \right\}. \quad (6.4.4)$$

Kolmogorov检验

例 6.4.1 问在水平 $\alpha = 0.1$ 下是否可以认为下列10个数

0.437, 0.863, 0.034, 0.964, 0.366, 0.469, 0.637, 0.623, 0.804, 0.261

来自于均匀分布 $U(0, 1)$?

Kolmogorov检验

解 对于此组数据，为了便于计算，我们有如下表：

表6.6 例6.4.1的计算表

i	$x_{(i)}$	$F_0(x_{(i)})$	$\frac{i-1}{n}$	$\frac{i}{n}$	δ_i	i	$x_{(i)}$	$F_0(x_{(i)})$	$\frac{i-1}{n}$	$\frac{i}{n}$	δ_i
1	0.034	0.034	0.0	0.1	0.066	6	0.623	0.623	0.5	0.6	0.123
2	0.261	0.261	0.1	0.2	0.161	7	0.637	0.637	0.6	0.7	0.063
3	0.366	0.366	0.2	0.3	0.166	8	0.804	0.804	0.7	0.8	0.104
4	0.437	0.437	0.3	0.4	0.137	9	0.863	0.863	0.8	0.9	0.063
5	0.469	0.469	0.4	0.5	0.069	10	0.964	0.964	0.9	1.0	0.064

其中 $\delta_i = \max \{F_0(x_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F_0(x_{(i)})\}$. 从表6.6可以看出，本问题的Kolmogorov统计量的值为 $D_n = 0.166 < 0.3687 = D_{0.1}(10)$ ，故我们没有理由拒绝 H_0 ，即可以认为这组数是来自 $U(0,1)$ 的样本. □

Kolmogorov检验

关于我们前面讲述的两个分布检验方法： χ^2 拟合优度检验与Kolmogorov检验，我们有如下几点注解：

注 6.4.1 当 F_0 为完全已知的连续分布时，Kolmogorov检验优于 χ^2 拟合优度检验。但如果 F_0 是离散的或含有未知参数，则Kolmogorov检验无法应用。

注 6.4.2 当 F_0 为含有未知参数的连续分布时，如用样本去估计未知参数，则定理1.2.3及定理6.4.1的结论均不成立，故此时我们无法利用Kolmogorov检验，但 χ^2 拟合优度检验是可以的。（如果 F_0 是正态分布和指数分布，则有人给出了 D_n 的分布，有兴趣的请参见茆诗松、王玲玲(1984)）

注 6.4.3 对于两样本分布假设检验问题，即设 $X_1, \dots, X_m$ 和 $Y_1, \dots, Y_n$ 为分别来自总体 F 和 G 的iid样本，且全样本独立，则我们可以用统计量

$$D_{m,n} = \sup_x |F_n(x) - G_m(x)|, \quad (6.4.5)$$

进行分布检验，且拒绝域为 $\{D_{m,n} \geq c\}$ ，其中 F_n, G_m 分别表示总体 X, Y 的经验分布函数。由于Smirnov给出了此统计量的极限分布，故我们称之为Smirnov检验。由于它是Kolmogorov检验的自然推广，故有的书也统称之为Kolmogorov-Smirnov检验。

Kolmogorov检验

注 6.4.4 Kolmogorov检验很难处理高维数据的分布检验，而 χ^2 拟合优度检验则与一维的类似。

注 6.4.5 如取 F_0 为 $U(0, 1)$ 的分布函数，则Kolmogorov统计量 D_n 还可作为均匀性的一种度量，我们称这种度量为偏差(discrepancy)，而 χ^2 拟合优度检验统计则不然。

关于分布检验，还有许多类似于Kolmogorov统计量的许多检验统计量，如

- $\sup_x |w(x)(F_n(x) - F_0(x))|$, 其中 $w(x)$ 为给定的权函数.
- Cramér-Von Mises统计量: $C_n = n \int (F_n(x) - F_0(x))^2 dF(x)$.
- Anderson-Darling 统计量: $n \int w(F_0(x))(F_n(x) - F_0(x))^2 dF(x)$, 其中 $w(t) = [t(1 - t)]^{-1/2}$.

目 录

5.1 正态概率纸检验法

5.2 拟合优度检验

5.3 列联表的独立性检验

5.4 Kolmogorov检验

5.5 正态性检验

正态性检验

虽然我们在本章第一节讲述了正态分布的正态概率纸检验法,但由于那种方法比较粗糙,故本节将介绍另外两种更有效的常用检验方法.当然,关于正态性检验现有许多种,有兴趣的读者请参见梁小筠(1997).另外,经过国际标准化组织(ISO)统计标准分委员会的研究,最后认为Wilk-Shapiro的W检验和D'Agostino的D检验法最好(它们的第二类错误概率最小),于是它们向世界各国推荐使用这两个正态性检验方法.我国已把这两个方法列入了国家标准,见GB488-85.当样本量 $n \leq 50$ 时,用W检验;当样本量 $n > 50$ 时,用D检验.

为了方便,本节始终假设 $X_1, \dots, X_n$ 为来自总体为 F 的iid样本,现感兴趣的假设为

$$H_0 : F(x) \in \left\{ \Phi \left(\frac{x - \mu}{\sigma} \right) : \mu \in R, \sigma > 0 \right\} \quad (6.5.1)$$

W检验

W检验是Wilks和Shapiro于1965年提出的，其检验统计量为

$$W = \frac{\left[\sum_{k=1}^{[n/2]} a_k(n) (X_{(n+1-k)} - X_{(k)}) \right]^2}{\sum_{i=1}^n (X_{(i)} - \bar{X})^2}, \quad (6.5.2)$$

其中 $X_{(1)} \leq \cdots \leq X_{(n)}$ 为次序统计量， $\{a_k(n)\}$ 是一组给定的与 n 有关的常数，可查表求得.

W检验的拒绝域为

$$\{W < c\}, \quad (6.5.3)$$

其中临界值 c 满足 $P\{W < c|H_0\} = \alpha$.

W检验

此检验的理由大致如下:

设 $X_1, \dots, X_n$ 为来自 $N(\mu, \sigma^2)$ 的iid样本, 如记 $Y_i = \frac{X_{(i)} - \mu}{\sigma}$, $m_i = EY_i$, $\epsilon_i = X_{(i)} - EX_{(i)}$, 则知 m_i 是与 μ, σ 无关的常数, 且有

$$X_{(i)} = \mu + \sigma m_i + \epsilon_i. \quad (6.5.4)$$

由于当 n 较大时, $X_{(i)} \approx EX_{(i)}$, 因而 ϵ_i 较小. 于是 n 个点 $\{(X_{(i)}, m_i) : i = 1, 2, \dots, n\}$ 近似在一条直线上.

为区别这 n 个点是否在一条直线上, 我们可采用下面的 R^2 统计量

$$R^2 = \left[\frac{\sum_{i=1}^n (X_{(i)} - \bar{X})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^n (X_{(i)} - \bar{X})^2 \sum_{i=1}^n (m_i - \bar{m})^2}} \right]^2, \quad (6.5.5)$$

其中 $\bar{m} = \sum_{i=1}^n m_i / n$.

当 H_0 成立时, 即 X_i 服从 $N(\mu, \sigma^2)$ 时, (6.5.5)式的 R^2 应接近1. 于是, 当 R^2 远离1时, 我们有理由拒绝 H_0 .

W检验

由于 $N(0,1)$ 是对称分布, 所以当 H_0 成立时, $(Y_1, \dots, Y_n)$ 与 $(-Y_1, \dots, -Y_n)$ 有相同的联合分布, 从而 Y_k 与 $-Y_{n+1-k}$ 也有相同的分布, 故 $m_k = -m_{n+1-k}$, $\bar{m} = 0$. 于是,

$$R^2 = \frac{[\sum_{i=1}^n X_{(i)} m_i]^2}{\sum_{i=1}^n (X_{(i)} - \bar{X})^2 \sum_{i=1}^n m_i^2} = \frac{\left[\sum_{k=1}^{[n/2]} m_{n+1-k} (X_{(n+1-k)} - X_{(k)}) \right]^2}{\sum_{i=1}^n (X_{(i)} - \bar{X})^2 \sum_{i=1}^n m_i^2}, \quad (6.5.6)$$

如取

$$a_k(n) = \frac{m_{n+1-k}}{[\sum_{i=1}^n m_i^2]^{1/2}}, \quad (6.5.7)$$

则知 $W = R^2$.

为理解(6.5.6)式, 我们定义

$$Q^2 = \sum_{i=1}^n (X_{(i)} - \bar{X})^2, \quad \hat{\sigma} = \sum_{i=1}^n b_i X_{(i)}, \quad b_i = m_i / \sum_{i=1}^n m_i^2,$$

W检验

则 R^2 可以重写成

$$R^2 = \sum_{i=1}^n m_i^2 \hat{\sigma}^2 / Q^2.$$

我们注意到，无论样本是否正态， Q^2 经修正后都可以成为 σ^2 的一个UE，而分子 $\hat{\sigma}$ 是样本的线性函数，且当 H_0 成立时它是 σ 的UE，于是 $\hat{\sigma}^2$ 也可成为 σ^2 的一个估计。这样，在正态性假设下，二者的差别不应很大。这就是我们采用 R^2 来检验正态性假设的一个直观解释。为了扩大这个差异，Wilks 和Shapiro把上述的 $\hat{\sigma}$ 换为 σ 的BLUE

$$\hat{\sigma}' = \sum_{i=1}^n c_i X_{(i)}, \quad (6.5.8)$$

就得到了W检验统计量。

D检验

W检验虽有效,但可惜的是它只适用于容量3-50的样本,其主要原因是,对超过50的样本容量 n ,我们很难计算其分位数值.为此,Dagostino提出了D检验.

由W统计量的定义不难看出,如果不计常数因子,它是正态分布的标准差 σ 的BLUE与样本标准差的比值的平方,即关于 σ 的两个估计量的比值的平方.由于BLUE的系数的计算是繁杂的,且随着样本量的增加,这种繁杂在增加.于是,随着样本量的增加,我们不得不放弃用BLUE来构造检验统计量的打算.这样,Dagostino建议,在样本容量超过50时,用 $(i - \frac{n+1}{2})$ 来代替BLUE(6.5.8)式中的 c_i ,即考虑用

$$T = \sum_{i=1}^n \left(i - \frac{n+1}{2} \right) X_{(i)}$$

作为 σ 的一个线性无偏估计.再考虑到 n 的阶,我们用统计量

$$D_n = \frac{\sum_{i=1}^n \left(i - \frac{n+1}{2} \right) X_{(i)}}{n^{3/2} \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}, \quad (6.5.9)$$

作为正态性检验统计量,且当它远离1时,我们有理由拒绝正态性假设.

D检验

为求取其极限零分布，我们可以证明，当 H_0 成立时

$$E(D_n) = 0.28209479, \quad \sqrt{\text{Var}(D_n)} = 0.02998598/\sqrt{n},$$

故我们采用如下的标准化随机变量

$$Y_n = \frac{\sqrt{n}(D_n - 0.28209479)}{0.02998598} \quad (6.5.10)$$

作为检验统计量，且其拒绝域为

$$\{Y_n < c_1\} \cup \{Y_n > c_2\},$$

其中 $c_1 < c_2$.

关于D检验的分位数现已有表可查.

我们在这里

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

