正态性检验

梁小筠　编著

中国统计出版社

正态性检验

梁小筠　编著

中国统计出版社

(京)新登字 041 号

图书在版编目(CIP)数据

正态性检验/梁小筠编著.
—北京:中国统计出版社,1996.12
ISBN 7-5037-2307-6

Ⅰ.正…
Ⅱ.梁…
Ⅲ.正态性检验
Ⅳ.0212.1

中国版本图书馆 CIP 数据核字(96)第 19462 号

中国统计出版社出版
(北京三里河月坛南街 75 号　100826)
新华书店经销
三河市双峰印刷厂
*
850×1168 毫米　32 开本　5.5 印张　13 万字
1997 年 5 月第 1 版　1997 年 5 月北京第 1 次印刷
印数:1—3000 册
定价:11.00 元

序

一般认为，正态分布是 Demoive 于 1730 年在研究二项式概率的逼近时发现的，有关的结果现今已载入概率论教科书。上世纪初，Gauss 在研究测量误差时，从另一个角度引进了它：用现在的术语说，若假定某个量的测量误差服从正态分布，则其最小二乘估计恰与其极大似然估计一致——自然，在 Gauss 那时还没有极大似然估计的概念。由于这个原因，文献中也常把正态分布称为 Gauss 分布。“正态”，意谓“正常的状态”，就是说若在观察或试验中不出现重大的失误，则结果应遵从正态分布。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来由。然而，在重要的应用问题中观测结果不遵从正态分布的情况，远非绝无仅有。这一点早在 K. Pearson 时代已有认识：Pearson 引进了如今就以他自己的名字命名的分布族（见本书 § 2.3），就是为了应付这个问题。虽然如此，正态分布在数理统计学中仍不失其中心的地位，这也是大家公认的。正态分布之所以如此重要，除了它确实可以常用来描述实际中的观察数据外，另一个重要原因要归功于 Fisher 及其同时代的若干杰出学者。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。但在一特定的问题中此假定是否合理，有时并无充分的把握。这时就需要使用已有的观测数据，去检验正态性假定是否可认为成立。这就是正态性检验问题，它是本书讨论的主旨。这个

问题在应用上是如此重要，以至在国家标准体系中，也专为它制定了一项标准，即《正态性标准》GB4882-85。本书作者是制定这项标准的主要参与者之一。在工作过程中曾查阅了有关这个问题的大量文献。最后写成的"标准"文本，也曾经过国内有关统计专家和应用部门同志的仔细审查。但是，作为一个国家标准，其文本不能太长，不可能把有关细节和应用上值得注意的地方都解释清楚。有鉴于此，作者撰写了这本专著，对有关正态性检验的问题作了系统的论述。

这是我国第一部论述这个重要问题的专著。在一般统计教科书中，由于篇幅所限及体例上的习惯因素，都只是把这个问题放在一般的分布拟合检验项下作简短的处理，没能作仔细的讨论。本书的问世，使得对这个问题感兴趣的实用和理论研究人员，以至数理统计课程的教师和学生，有一本较完备的参考资料，而不必从大量的文献中去寻找翻检。在这个意义上，可以说本书的出版填补了我国统计著作中的一个空白。

本书讨论的是一个极富实用价值的题目，但也有理论研究上的意义：人们可以从种种不同的角度出发探求正态性的新的检验方法；现有检验方法的统计性质研究还不充分；各种检验法的优劣比较及其在种种情况下的性能表现，目前也还知之不多，这些都是有重要实用背景的理论研究课题。但是，对本书所论课题感兴趣者，恐怕绝大多数是出于工作中的需要。为此，本书作者把书写成一本实用统计的著作，不作艰深的纯理论探讨，而对使用这些方法的步骤有仔细明白的交代，并总是有数字例子加以演示。这使本书非常适合于那些不具备高深数理统计知识的，以实用为目的的读者。另一方面，作者也没有把本书写成一个手册或Cook book的形式，对各种检验方法所依据的统计思想，在不涉及高深理论的前提下，作了清楚扼要的交代。这是本书的一个优点。另一个特点是本书于各部分的详略处置比较得宜。例如 χ^2 法和 Колмогоров 检验法，在一般统计教科书中大多有所论述，本书在这些题材上相应就

简略一些。第三到六章的方法是基于正态分布的特殊性，在一般教科书上很少论及，书上的处理也就比较仔细。最后一章是一个很富理论性的题材，但在实用上也很重要。这一章内容的充分处理，不涉及高深的理论不行。作者采用针对一些重要的分布族用模拟的方法来处理，颇具特色。我想，读者通过这一章的论述，也会理解到检验有一个功效问题。不同的检验，针对不同的分布类，其功效各有差别，无绝对优劣之分。这种认识能帮助我们不对种种检验方法抱一种固执的看法。

作者梁小筠同志长年活跃在教学科研第一线，并主持和参与过大量的应用项目，是我国资深的统计学家。由她这样既有深厚的理论素养又有丰富的实践经验的学者来撰写这种很富实用意义的题材，必当受到广大读者的欢迎。今值本书即将出版之际，爰志数语，表达我作为一名读者的感受，并感谢作者为我们提供了这么一部好作品。

陈希孺

1995 年 3 月 22 日

于北京

前　言

正态分布是自然界最重要的分布，它能描述许多随机现象。以总体服从正态分布为前提的统计方法已被越来越多的教学、科研工作者和工程技术人员所掌握。然而，在一个实际问题中，总体一定是正态分布吗？如果不顾这个前提成立与否，盲目套用公式，可能影响统计方法的效果。因此，正态性检验是统计方法应用中的重要问题。

长期以来，我国有关的教科书沿着前苏联的模式，在谈到正态性检验时，只介绍 χ^2 拟合优度检验和柯尔莫哥洛夫检验。这两种"万金油式"的检验方法，对正态性检验不具有特效。

我国已经制订了国家标准 GB4882-85 正态性检验，它介绍了国际上采用的先进的检验方法。在制订标准的过程中，我查阅了国外的有关文献，才知道早在 20 年代，已经有了专用的正态性检验方法，这个问题的研究一直在进行。国家标准颁布后，我经常收到来信，询问标准的原理。事实上，正态性检验方法的原理不仅对该标准的使用和研究有一定的作用，而且研究方法别具一格，是数理统计宝库的重要组成部分。现将检验方法和相应的原理编著成这本书，敬献给读者。

本书第一章介绍了关于正态分布的基本知识和利用直方图、概率纸检验分布的正态性的直观方法。第二章介绍了 χ^2 拟合优度检验和柯尔莫哥洛夫检验，这两种方法可以用来检验分布的正态性，也可以检验总体是否属于其它分布类型。第三章至第六章介绍了国家标准《正态性检验》中推荐的 W 检验、D 检验、偏度检验、峰度检验和偏度、峰度联合检验的方法及原理。第七章比较了各种检

验方法的功效。具备概率统计初步知识的读者，可以从中找到适宜的检验方法。

为了阐述这些检验方法的原理，本书介绍了用皮尔逊(Pearson)分布族、约翰逊(Johnson)分布族、柯尼西-费歇(Cornish-Fisher)展开等方法拟合统计量的分布。急于寻找检验方法的读者可以不看这些章节，想要弄清原理的读者不妨读一下，这方面国内资料很少。

魏宗舒教授审阅了本书的初稿，茆诗松教授审阅了初稿和第二稿，并自始至终关心这本书的出版。他们都提出了极其宝贵的意见和建议。陈希孺教授在百忙中为本书撰写序言。我在编著本书，尤其是关于分布拟合检验的内容时，从他的著作和讲课中受到很大的教益。在此，对以上诸位教授表示衷心的感谢。

最后，还要感谢中国统计出版社的同志为本书的出版所作的努力。

由于本人水平有限，书中一定还有许多不足之处。恳切地希望同行专家和广大读者指正。

梁小筠

1996 年 8 月

于华东师范大学统计系

目　录

第一章　正态分布

§1.1　正态分布

学过概率统计的同志都知道，正态分布是概率论中最重要的分布。这不仅由于自然界中许多现象可以用它来拟合，更重要的是它在概率论的理论研究中占有重要的地位。许多分布可以用正态分布来近似，还有一些分布由正态分布导出。所以数理统计中不少问题都是首先在总体服从正态分布的前提下研究的。

一、正态分布的密度函数

正态分布的密度函数为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.1.1}$$

其中 μ 和 σ 为参数，它们分别在范围：$-\infty<\mu<+\infty$，$\sigma>0$ 内各取一确定的值。正态分布密度函数的图形是中间高、两边低，关于直线 $x=\mu$ 对称，并在 $x=\mu$ 时达到最大值 $1/(\sigma\sqrt{2\pi})$。当 $x\to\pm\infty$ 时，$f(x)\to 0$，因此 x 轴是曲线的渐近线。通过求导不难看出：密度函数曲线在 $x=\sigma$ 和 $x=-\sigma$ 时各有一个拐点，在 $(\mu-\sigma,\mu+\sigma)$ 内曲线向上凸，在这个范围以外，曲线向下凸。

公式(1.1.1)由参数 μ 和 σ 的取值完全确定。通常用符号 $N(\mu,\sigma^2)$ 表示参数为 μ 和 σ 的正态分布。

二、正态分布的矩

参数 μ 为正态分布的数学期望，它刻划了随机变量取值的平均状态。σ^2 为方差，它反映了随机变量的取值与数学期望的平均偏离程度。σ 为随机变量的标准差，又称均方差。

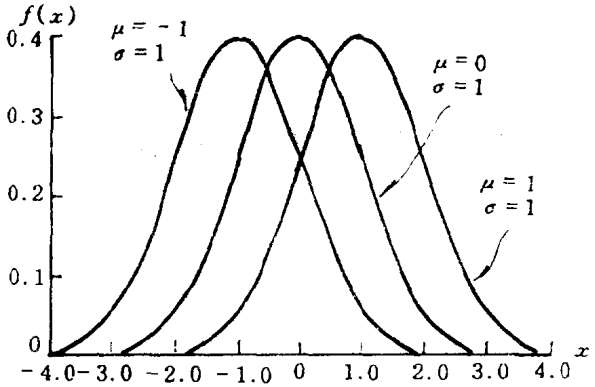

图 1-1 具有相同的 σ 值、不同的 μ 值的正态分布密度函数曲线

图 1-1 给出了 σ 值相同、μ 值不同时，正态分布的密度函数曲线。可以看出：μ 值的不同，只影响曲线的位置，故常称 μ 是正态分布的位置参数。

图 1-2 给出了 μ 值相同 σ 值不同时，正态分布的密度函数曲线。可以看出：σ 值越小，曲线越尖峭，反之，曲线就越平坦。这是因为：σ 的值小，密度函数的最大值就大，而密度函数曲线与 x 轴所夹的面积总是 1，曲线自然就“高而瘦”了；反之，σ 的值大时，密度函数的最大值较小，曲线就“矮而胖”了。但是，曲线的形状仍然是中间高、两边低、左右对称的。之所以有“胖、瘦”之分，是由于随机变量取值的分散程度不同，故又称 σ 是正态分布的尺度参数。

正态分布的 k 阶中心矩为：

$$\mu_k = E(X-\mu)^k = \begin{cases} 0 & k \text{ 为奇数} \\ 1 \cdot 3 \cdot 5 \cdots (k-1)\sigma^k & k \text{ 为偶数} \end{cases} \tag{1.1.2}$$

为了刻划各种分布的密度函数曲线的形状，还要引入偏度和峰度这两个概念。

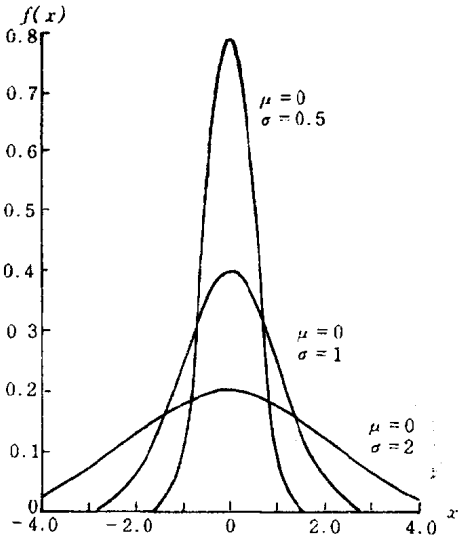

图 1-2　具有相同的 μ 值、不同的 σ 值的正态分布密度函数曲线

偏度反映分布的对称性。我们知道，对称分布的奇数阶中心矩为零。最简单的奇数阶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立方。因此，通常把三阶中心矩除以标准差 σ 的立方，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，但这个符号可能有人不习惯，因为偏度可以取负值。经研究，在国家标准《正态性检验》中用 β_s 表示偏度，即

$$\beta_s = \frac{\mu_3}{\sigma^3} \tag{1.1.3}$$

对正态分布来说，$\beta_s=0$。

在概率统计中，我们经常遇到连续的、单峰的、不对称的密度函数曲线，它在众数（密度函数在这一点达到最大值）的一边形成

长尾，另一边形成短尾。如果长尾是在正的一边(如图 1-3 的对数正态分布的密度函数曲线)，那么 $\beta_s>0$，称该分布具有正的偏度。反之，如果长尾是在负的一边，则 $\beta_s<0$，称该分布具有负的偏度。

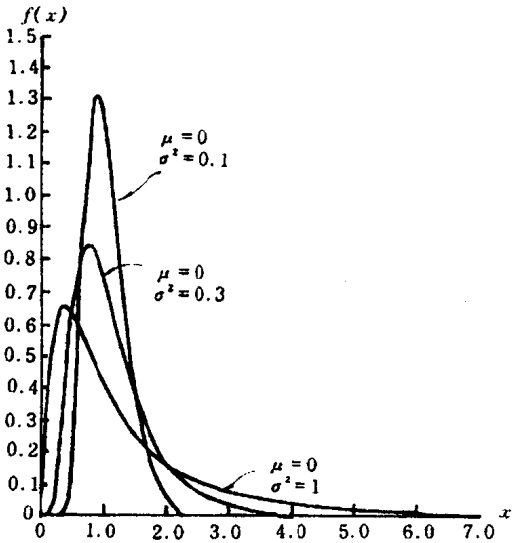

图 1-3　对数正态分布的密度函数曲线

峰度反映密度函数曲线在众数附近的“峰”的尖峭程度。如前所述，这还与方差有关。为了消除方差的影响，并把它表为一个无量纲的量，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

$$\beta_k = \frac{\mu_4}{\sigma^4} \tag{1.1.4}$$

对正态分布来说，$\beta_k=3$。

图 1-4 是利用计算机画的正态分布 $N(0,5/3)$ 和自由度为 5 的 t 分布 $t(5)$的密度函数曲线。这两个分布的标准差都是 $\sqrt{5/3}$。

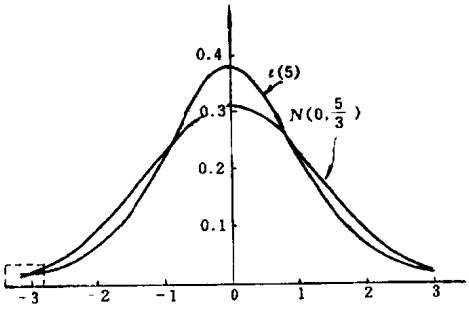

图 1-4(a)　$t(5)$分布和$N\left(0,\frac{5}{3}\right)$的分布的密度函数曲线

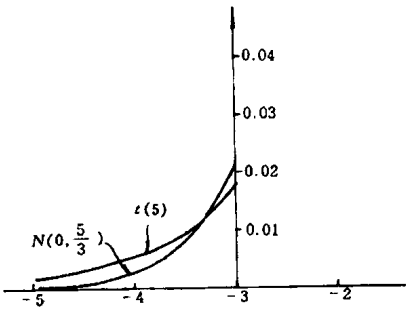

图 1-4(b)　局部放大图

$t(5)$分布的峰度$\beta_k=9$。由图 1-4(a)可见，$t(5)$分布的密度函数曲线的“峰”比正态分布的高，这两条曲线在一侧两次相交，结果$t(5)$分布密度函数曲线的“尾巴”比正态分布的“粗”(或者说前者的尾巴较长)。这一点在把图 1-4(a)左侧尾部放大后得到的图 1-4(b)上看得更清楚。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，则称该分布具有过度的峰度，如果一个分布的峰度小于 3，则称该分布具有不足的峰度。

各种正态分布，尽管μ和σ可以分别取不同的值，但偏度都等于零，峰度都等于 3，它们的密度函数曲线的形状都是一样的。因此，对正态分布不提形状参数这个名称。

三、标准正态分布

正态分布的分布函数为

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt \tag{1.1.5}$$

我们称 $\mu=0,\sigma=1$ 的正态分布为标准正态分布。它的密度函数常用 $\varphi(x)$ 表示，分布函数常用 $\Phi(x)$ 表示，即

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{1.1.6}$$

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \tag{1.1.7}$$

在(1.1.5)式中，如果令

$$u=\frac{t-\mu}{\sigma} \tag{1.1.8}$$

那么

$$F(x)=\int_{-\infty}^{\frac{x-\mu}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du=\Phi\left(\frac{x-\mu}{\sigma}\right) \tag{1.1.9}$$

因此，任何正态分布都可以通过标准正态分布表示。只要有了标准正态分布函数表，就可以获得一般正态分布的分布函数值。

从 $\varphi(-x)=\varphi(x)$ 可以得到

$$\Phi(-x)=1-\Phi(x) \tag{1.1.10}$$

上式在计算时是很有用的。

例 1.1　设一批零件的尺寸偏差(单位：微米)服从正态分布 $N(-50,100^2)$，求：

(1) 尺寸偏差的绝对值不超过 150 微米的概率；

(2) 尺寸偏差为负值的概率。

解：设随机变量 X 表示这批零件的尺寸偏差，那么：

(1) $P(|X|\leqslant 150)=P(-150\leqslant X\leqslant 150)$

$$=\Phi\left(\frac{150+50}{100}\right)-\Phi\left(\frac{-150+50}{100}\right)=\Phi(2)-\Phi(-1)$$

$$=\Phi(2)-[1-\Phi(1)]=0.9773-1+0.8413=0.8186.$$

(2) $P(X<0)=\Phi(\frac{50}{100})=\Phi(0.5)=0.6915$.

例 1.2　设随机变量 X 服从正态分布 $N(\mu,\sigma^2)$，分别求下述概率：$P(\mu-\sigma<X<\mu+\sigma)$，$P(\mu-2\sigma<X<\mu+2\sigma)$，$P(\mu-3\sigma<X<\mu+3\sigma)$.

解：$P(\mu-\sigma<X<\mu+\sigma)=\Phi(1)-\Phi(-1)$

$$=2\Phi(1)-1=68.27\%.$$

$$P(\mu-2\sigma<X<\mu+2\sigma=\Phi(2)-\Phi(-2)$$

$$=2\Phi(2)-1=95.45\%.$$

$$P(\mu-3\sigma<X<\mu+3\sigma)=\Phi(3)-\Phi(-3)$$

$$=2\Phi(3)-1=99.73\%.$$

从最后一个式子可以看出：随机变量 X 的取值落在$(\mu-3\sigma,\mu+3\sigma)$外的概率只有 0.27%。因此在一次试验中，基本上可以认为 X 的取值落在$(\mu-3\sigma,\mu+3\sigma)$中。这就是实际工作者所说的“$3\sigma$原则”。

§1.2　正态分布参数的点估计

在很多实际问题中，正态分布的参数 μ 和 σ 是未知的，我们必须从总体抽取样本 $X_1,X_2,\cdots,X_n$，通过样本估计未知参数。

一、参数 μ 的点估计

μ 是总体的数学期望，很自然地可以用样本均值去估计它。

$$\hat{\mu}=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad (1.2.1)$$

由于 $X_i(i=1,2,\cdots,n)$ 是随机变量，μ 的估计量 $\hat{\mu}$ 也是随机变量。当把 $X_1,X_2,\cdots,X_n$ 的值代入(1.2.1)式时，得到一个数，这个数就是 μ 的一个估计值。

例 1.3　已知某种灯泡的寿命服从正态分布。从一批灯泡中随机地抽取 10 只，测得其寿命(单位：小时)为：

1059,939,1193,785,1114,936,918,1156,920,945