# 大气科学中的概率统计和决策

[美]Allan H. Murphy
Richard W. Katz 编

气象出版社

# 大气科学中的概率、统计和决策

〔美〕 Allan H. Murphy 墨菲
Richard W. Katz 编

史国宁、周诗健等译



气象出版社

## 内 容 简 介

这是一本由多位学者联合撰写的统计气象学专著，本书全面而深入地评述了对大气科学家具有潜在意义的概率和统计学方面的最新进展，这是其它书籍中所没有的，同时本书也论述了对大气科学中概率和统计学一直起着并将继续起重要作用的一些领域。全书共分13章，分别为：大气数据资料的试探分析；用多重回归发展经验模式；单批量数据资料的试探多元分析；多批量数据资料间的多元比较；时间序列分析——频率域；时间序列分析——时间域；概率模式；统计天气预报；概率天气预报；预报的评价；人工影响天气试验的设计与评价；贝叶斯推断；决策分析。本书内容新颖、全面，对于关心统计方法在大气科学中，以及地球物理学、环境科学等有关领域中的应用的研究人员、工程技术人员以及大专院校师生来说，都是一本重要的参考书。

# 大气科学中的概率、统计和决策

Allan H. Murphy
Richard W. Katz 编
史国宁、周诗健 等译
责任编辑 成秀虎

# Probability, Statistics, and Decision Making in the Atmospheric Sciences

Edited by Allan H.Murphy and Richard W.Katz

1985, Westview Press, Boulder and London

* * *

气象出版社出版
（北京西郊白石桥路46号）

北京顺义燕华印刷厂印刷
新华书店总店科技发行所发行 全国各地新华书店经销
1991年4月第一版 1991年4月第一次印刷
开本：787×1092 1/16 字数：408千字
印张：16.5 印数：1—1500
ISBN 7-5029-0600-2/P·0332
定价：11.95元

# 译 者 的 话

这是一本由多位学者联合撰写的统计气象学专著，编者是美国俄勒冈州立大学大气科学系统计学与气候影响实验室主任 A.H.Murphy 教授以及美国科罗拉多州博尔德市国家大气研究中心环境与社会影响研究组的科学家R.W.Katz。本书内容深入、全面、新颖，它包括了普通统计学专著中没有深入涉及的一些领域内的最新进展，既有大气科学家感兴趣的统计学方法中的普遍问题，也有如统计天气预报、人工影响天气试验的设计与评价之类的专门问题。如所周知，概率、统计和决策领域中的方法在大气科学的研究和业务工作中已取得越来越多的应用，占有相当重要的地位，但迄今还没有一本既全面又新颖的专著，这本《大气科学中的概率、统计和决策》，从某种意义上说弥补了这一缺憾。

本书由史国宁、周诗健等译出并相互校阅，具体分工是史国宁译前言、第一、八、九、十、十三章；周诗健译第二至七章；谷真真译第十一章；霍总会译第十二章。由于本书篇幅较大，参考文献又很多，故后者只有忍痛割爱了。我们希望本书能对从事这方面工作的气象工作者以及其他有关科技人员有所裨益，并敬请广大读者批评指正。

译者

1988.6.

# 作者序言

概率、统计和决策领域中的方法论在大气科学的基础和应用研究以及试验和业务研究中均起着越来越重要的作用。这种方法论几乎可应用于大气科学的每一个方面，从最具理论性和综合性的问题（如大气可预报性、全球气候模拟）到最实际的和局部的问题（如作物-天气模式、预报评价）。在大气科学界当前所出版的大量杂志中，几乎每一期都包含有一篇以上有概率和统计概念和（或）方法论的应用的文章。尽管这方面的应用越来越广泛，但是自从1953年Brooks和Carruthers出版了《气象学中的统计学方法手册》以及1958年Panofsky和Brier出版了《统计学在气象中的应用》这两本开创性的著作之后，至今很少见到再有什么关于大气科学的概率和统计问题的专门著作问世。因此，大气科学家对于概率和统计学的许多新进展不很了解，而在气象研究中有关概率和统计方法应用的最近的工作亦尚未为整个气象界所熟悉。

看来，没有哪一本书能够把大气科学家所关心的许许多多概率、统计和决策问题全部包括进去，这本书当然也不例外。我们在这本书内打算把重点放在两种类型的论文上：1)对于大气科学家具有潜在意义的概率和统计学新进展的全面评述，2)对于大气科学中概率和统计学一直起着并且将继续起重要作用的某些领域的评述。由于全书是由许多人的论文汇集而成，因此各章的论述范围和深度不尽相同。但是，我们相信，各章的论述均普遍适合于我们所预期的读者，即至少已学过一门概率和统计学课程的大气科学学生以及在工作中经常要用到概率和统计学基本概念和方法的气象研究人员和其他气象工作者。而且，对于许多读者所不熟悉的一些新课题（试探数据分析、决策分析）采用了浅显的叙述方法。在所有各章中，均给出了大量的有关参考文献目录，以备读者作进一步的全面了解和(或)实际应用。我们希望，查阅本书的每一个人将至少能从中找到一章是有趣的和有用的。

虽然本书每一章都可以看作一个独立完整的单位，但其中很多章是有紧密联系的，最好能放在一起学习。第一章可以看作其他章的导引，它论述了试探分析，这是数据分析两个步骤中的第一步。其他各章大部分主要涉及实证分析，这是数据分析两个步骤中的第二步。第二章论述多重回归，它无疑是大气科学中最经常使用的统计方法。本章从许许多多有关多重回归的文献出发，把重点集中在一些新发展和新方法上。第三章和第四章分别论述单批量数据和多批量数据的多元分析，可以看作是第一章的扩展，因为这两章再一次着重叙述了数据试探分析方法。

第五、六、七章全都是讨论大气测量所固有的时间相关性问题的。第五章和第六章分别论述了频率域方法和时间域方法，它们是时间序列分析的两种相关的方法。第七章的论题是概率模式，第六章中考虑了它的一个特例，即参数时间序列模式。

第八、九、十章考察了天气预报的某些问题。统计天气预报日常业务已在许多国家中实现。第八章中主要评述制作这种预报所用的各种技术方法，如回归分析和判别分析等。第九章论述以概率形式给出这种预报的方法，亦即详细给出天气预报所固有的不确定性的方法。第十章探讨天气预报的评价问题，其中包括概率形式和非概率形式的预报。

迄今关于人工影响天气试验设计和效果检验已进行了相当多的统计学研究。第十一章讨论这类试验设计和效果检验方法，着重阐述随机双盲试验的必要性以及非参数方法的应用。第十二章论述贝叶斯推断，这种统计学推断特别适用于信息量和专家判断都不够充分的情况。概率和统计学应用于大气科学的最终目的在于帮助决策，因此，第十三章，即最后一章，将介绍决策分析，它们是决策问题模式分析的数学方法。

许多个人和组织都对本计划的完成以及这本书的出版作出了大量的贡献。首先我们要感谢各章的作者，没有他们写稿本书将不可能完成。除了编辑之外，还有几位同仁阅读了本书各章并提出意见，力求使各章之间术语和记号达到统一。我们特别要感谢 Barbara G.Brown在这方面所做的工作。Sheri Kellogg打印了这些章的大部分，剩下的由Michelle Holcomb完成，对于他们的认真细心的工作表示感谢。还要感谢 Westview 出版社的几位工作人员的帮助，特别是Alice Levine，Alice Trembour和Edythe Porpa。我们有关本计划的编写工作得到了国家科学基金会（大气科学部）的部分支持，其准许证号为ATM-8004680和ATM-8209713。

A.H.Murphy
（俄勒冈州科瓦利斯）
R.W.Katz
（科罗拉多州博尔德）

# 目　录

# 第一章　大气数据资料的试探分析

Thomas E.Graedel和Beat Kleiner

## 1. 引言

大气数据资料很少适合于使用传统的统计学方法。这类数据除了具有较短时间尺度的明显变化外，还在更长的时间尺度上具有各种各样的昼夜变化、季节变化、年变化和多年变化型式。重复试验几乎是不可能的或不可行的。这些特征说明，只有使用另一种新的方法对这类资料进行处理，才有可能取得成功。这种新方法称为数据资料试探分析或 EDA 。本章中我们打算说明，这种统计学方法是在现代计算系统的可行范围内的，并且对大气数据分析使用这种方法后即可获得大量有用的信息。

设计数据资料试探分析是为了要解答“数据资料告诉我们什么”。其基本意图是要寻找一群数据内的有意义的关系和结构，并在分析结果中展示出这种关系和结构。这个过程是一个概括过程，概括的结果可以用几个简单的统计量（如一组数据的均值和方差）来表示，也可以用简单的点绘图（如散布图）来表示。这个过程也是一个表露过程，也就是说，数据的表达方式应该让人既能看出数据的预期特征，也能看出数据的意外特征。通常，寻找意外特征会比仅仅确认猜想中的特征有价值得多。

尽管EDA并不是一个新的题目，但仅仅是在最近它才成为统计分析的一个重要部分。促成这种发展的重要因素是计算机硬件和软件数量的迅速增加，并且在许多领域内都存在着处理较大数据群的困难问题，有越来越多的学科日益重视定量化。尽管EDA很有用，然而它尚未被适当地统一到形式化的统计理论中去。看来，目前所可能做到的对这一学科的最简洁的介绍是列出EDA技术上的一些重要特征，如：

（1）EDA是一种迭代过程。它依靠试误法逐步进行，每一步所获得的认识被用作下一步的指导。

（2）在作EDA时，最好是在脑子中先有一个模式，但又不可太拘泥于这个模式。在思想上应尽量作好去接受可能被发现的关系的准备。

（3）应该用各种不同的方法去检查数据，这样做是有好处的。其中大部分方法最终将会变得不合逻辑，没有意义或无使用价值。但是，应该预料到这种结果，不要过快地终结EDA。

（4）EDA要广泛地用到稳健估计/抗扰估计（robust/resistant）统计学，并且要大量借助于图示方法。

通常，人们如果事先没有有关数据组的信息，就无法在一开始时假定该组数据是正态分布的。与此相反，某一给定数据组不是正态分布的可能性却要大得多。这或者是由于在每个中等长度或较长的数据组中不可避免地混有远离中心的数据，或者是由于被监测过程本身的非正态性，或者是由于某些其他原因。因此，简单地计算出参数通常并不能得到满

意的结果，因为计数参数仅当数据组是正态时才能完全描述该组数据。所以需要使用其他的方法来概括和表露这些数据。其中最有用的方法是数据变换法（变换到近似正态或至少是近似对称）、稳健估计统计学方法以及图示法。

在本章中仅能选择数据资料试探分析技术中的某些问题进行讨论。我们是结合大气学科来选择这些问题的，并且使用大气方面的例子进行说明。进一步钻研的读者可以先查阅Tukey（1977）、Mosteller和Tukey（1977）以及Erickson和Tukey(1979)所写的书。McNeil(1977)（FORTRAN和APL语言）、Velleman和Hoaglin(1981)（FORTRAN和BASIC语言）所写的两本书中还列出了供某些数据资料试探分析方法使用的计算机程序。

在下一节中，我们将叙述概括和表露一维数据组的几种方法。在第3节中我们将讨论两个数据组的比较方法，在第4节中将给出几个数据组的比较方法。第5节是总结和结论。

## 2. 一维数据资料序列的概括和表露

对可以被看作一维资料序列（或数据组）的一组数字进行分析，可能是数据分析中最常遇到的问题。这类数字组的例子有：

（1）某一天气站几年中一月份的降雪量；

（2）闪电资料中闪击间歇的频率分布；

（3）美国超过100000人口的所有城市的平均增温数（单位为度-日）；

（4）夏季某些地点环境空气中臭氧浓度的15分钟平均读数。

一维数据组并不一定必需是一元分布的独立样本。它们可以是一个多维数据组的一部分，如例（3）和（4）所示。在一维数据组内也可以存在非零相关，如例（4）所示。在这些个例中，经常是先把这些数据看成一维样本为宜，暂时不必考虑其内部的有关结构。

人们在分析数字组时，首先需要对现有的数据组有一个“直观”了解，并提出以下一些问题：

“其最小值和最大值是什么？”

“其中哪个数能较好地代表本组数据？”

“数据的变化或扩展范围有多大？”

“这些数据是凝聚在一个或多个数值的周围，还是均匀散布在某个区间上？”

“这些数据可以被认为是对称的吗？”

在考察一维数据组时，除了简单地计算其均值作为该数据组的中心（位置）的估计值以及计算其标准差作为其变化或扩展（范围）的估计值之外，还有许多其他的方法，这些方法往往还更好一些。本章我们将叙述其中较有用的一些方法。

### 2.1 一维散布图

一维散布图用标有所点绘数据的数值的一条轴线（通常用水平轴）来描述一组数字。只要看一眼散布图，就能得出上面所提某些问题的答案。它的主要优点是非常容易绘制，所占空间小，使用原始数据，而且只要这些数据不太稠密的话，就可以直接从散布图上重新读得。它的主要缺点是不易从图上得出有关该数据样本分布的推论。

图1表示1974年12月份每天上午1时在美国Newark机场所测得风速值的一维散布图。从该图可以看出，在31次测量中仅有11个不同的值，并且还能很快找到该数据组的极值。我们

可以推测，这些数据是非均匀分布的。但是，通常较难进一步推出有关数据分布的其他结论，特别是当这些数据点中有很多具有同样的数值时，或者当数据点的数目很大，使得点绘记号发生重叠时。

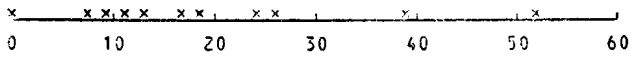


图1　1974年12月份每天上午1时在美国纽瓦克机场的风速测量值（横坐标，公里/小时）的一维散布图

（资料引自国家气候中心，1974—1975）

当数据组较短（以在100个数据点以内为宜）以及当需要对许多个数据组进行相互比较时，一维散布图是最有用的工具。由于它的简洁性，它也适合于用来描述一般（二维）散布图中每个变量的单独分布。

## 2.2　直方图

直方图是数据分析者用来考察样本总体分布的最古老和最经常使用的工具。对于一给定的数据组，构造直方图的方法是将水平轴分成一些小段（也称为分组），这些小段加在一起要能覆盖住全部数据。在每一小段上画一长方形，其面积与该小段所含数据的数目成正比。例如，图2（它所描述的数据与图1同）明显地表示出，数据不是均匀分布的，峰值在10与15公里/小时之间，并偏向正值的一侧（即偏向数据高值一侧）。约有1/3的数据位于15公里/小时的右边。绘制这个直方图的另一种方法是使用每一间隔中的数据点数目的比值而不是使用它们的数目作为面积。

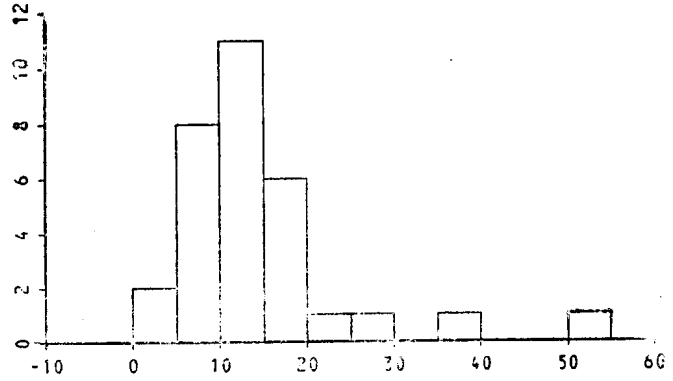


图2　图1中所示风速值的直方图

横坐标为1974年12月份0100时纽瓦克机场的风速（公里/小时）

由直方图所推得的信息在很大程度上取决绘制直方图之前所作的三个选择：分组数目、分组长度以及分组之间的分割点。这些选择的适当与否决定了一个直方图可否用来直观地表示一个数据组。例如，如果所取分组过少，则结果将表现不出特征，图形说明不了什么问题；而如果所取分组过多（特别是在样本组很短时），则落入每一分组内的观测数据很少，直方图可能会呈现不规则的形状。

通常是先确定分组长度（经常是用主观方法确定），然后用分组长度除数据总长度以确定分组数目。另一种方法是先确定分组数目，然后根据数据总长度计算出分组长度，这种方法的缺点是，只要有一个远离中心的数值，就会产生不良影响。

通常采用等长度分组。如果采用不同长度的分组，则必需保证每一矩形的面积（而不是高度）与该分组内的数据点数目成正比，这是因为直方图表示的是密度函数（这时分布中任何两点间的频率由该两点之间的曲线下的面积表示）。在选择直方图分组之间的分割点时也需要小心，这是因为仅仅为了得出一个“漂亮的”数字而任意选择分割点将会与另外两个选择发生相互作用，导至歪曲的或错误的图形。

关于如何选择所要使用的分组数目问题，不同文献有不同的建议，意见分歧甚大。有的文献提出“我们一般建议使用5到20个分组”（Mendenhall和ott，1972，p14），有的文献则认为“分组数目应大致等于数据点数的平方根”（Davies和Goldsmith，1972，p11）。Doane（1976）认为Davies和Goldsmith的规则仅适用于对称数据，而对于非对称数据，则应该增加一些分组，数据的偏斜度越大，则增加的分组数目就应该越大。遗憾的是他的方法很容易受到远离中心的数值的影响。

Diaconis和Freedman（1981）认为，应该先确定一个近似的分组长度h如下

$$h=1.349s[(1nn)/n]^{1/3}$$

式中s是标准差的稳健估计，n表示样本长度（稳健估计将在2.7节中介绍和讨论）。对分组长度的这种选择将能近似地使所得的直方图与它所代表的密度函数之间的最大绝对偏差达到最小。然后用h去除数据总长度，从而计算出暂定分组数K，再用类似于Doane(1976)所描述的规则近似地确定K＋1个“巧妙的”分割点。如果有1/3以上的数据正好落在分割点上，则将全部分割点移动半个分组长度。如果有1/3以上的数据落在新的分割点上，则再将这些新的分割点移动1/4分组长度。

在图2中，分组长度h是按照Diaconis和Freedman（1981）的规则计算的；当n＝31且s＝5.5（取作四分位数间距/1.349；见2.4节）时，h＝3.6，由它可得出暂定的分组数K＝15。把分割点选在适当的整数上，则得出了h＝5，11个分组，分割点为0，5，10，…，50和55。确定分组的这个规则是建立在逐渐逼近方法基础上的，因此只应该看作一个合理的起始点；在每一种实际情况下，可能需要对分组数目和分割点作主观的调整，使之适应于实际情况。

直方图的基本用途在于它能表示出各分组在数据组中的相对频率，因而能提供有关数据密度函数的信息。它的一个主要缺点是由于把数据分组，使得最终结果与原来的数据多少有些偏离。已经证明，对于数据点超过300—500个的长数据组来说，以及对于已经分好组的数据来说，直方图都是最有用的分析工具。

### 2.3 干-叶图

干-叶图（Tukey，1977）是由直方图经过改造得到的。在干-叶图中，既保持了数据本身的信息，同时也给出了分组信息。我们最好还是用具体的图来进行说明。图3是1975年1月美国Newark机场31个日最高温度读数的干-叶图。干-叶图与直方图的不同之处在于，它不仅仅是画出一个矩形，使其长度与该分组内的数据点数成正比，而是用该分组内每个数据的最有效数字来代表数据点。这样就能产生一些矩形，其长度与每个分组内的数据点数成正比，而且它上面所注的数字还能让分析者看出在每个分组内数据是如何分布的。

绘制干-叶图是较简单的。必要时，每个数据点可以截断或舍入成所使用单位的整倍数。把这些整倍数分成干（通常是全部首位数）和叶（末位数）两部分。然后再从这些数据确定最大的干和最小的干，并把这些干以及所有中间的干记入竖栏。用一条竖线将干与

叶分开。然后对整个数据组继续进行下去，并把每个叶记在与它的干相对应的行上。最后，把每个干上的叶从最小到最大进行分类。图4表示绘制图3中的干-叶图的几个步骤。干-叶图所以要按竖直方向而不是按水平方向填绘，是为了方便。因为在竖直的干-叶图上的数字，填和读都要便当得多。

31个数值，单位： 2|8 = 28°F

中位数 = 44　　四分位数 = 38, 50

| | |
|---|---|
| 2 | 8 |
| 3 | 0114 |
| 3 | 688 |
| 4 | 000122244 |
| 4 | 567777 |
| 5 | 00014 |
| 5 | 7 |
| 6 | 1 |
| 6 | 6 |

图3　1975年1月份纽瓦克机场日最高温度（°F）的干-叶图（数据取自国家气候中心，1974—1975）

图中数字的含义如下：2 | 8=28°F；
3 | 0114=30, 31, 31, 34°F

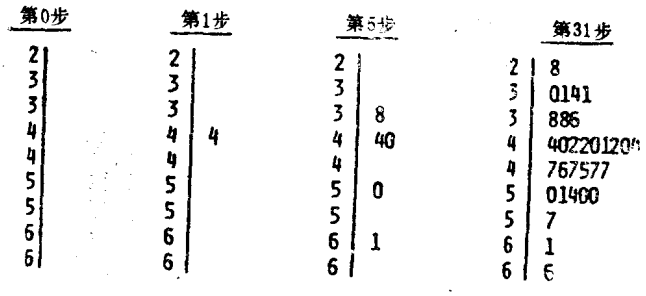


图4　制作干-叶图过程中的4个步骤。第5步后的数据值为44, 50, 61, 40和38°F。把第31步的结果进行分类，即得出图3

看来，干-叶图的最大特点在于它能保持整个数据组的分布特征，同时又能让人看出在各分组内部数据是否均匀分布的，是否有特殊的数值出现。这些特点如图5所示。图5是美国纽瓦克机场的风速干-叶图（所用的资料同图1和图2）。该图表明，某些数值（如9.2或11.1）经常出现，而另一些数值（如10.0或13.0）则完全未出现，这说明必然使用过一些特别的舍入过程。这就使我们得到一个十分肯定的设想：这些数据已经被乘以1.850，由海里/小时转换成了公里/小时。因此，图中的9.2相当于5海里/小时，11.1相当于6海里/小时，等等。Lockhart（1979）编制了一张直方图，显示出与上类似的效应；Reiss和Eversole（1978）讨论了能见度数据中的漏缺数字问题。

31个数值， 单位： 9|2 = 9.2公里/小时

中位数 = 11.1, 四分位数 = 9.25, 15.7

LO: 0.0, 0.0

| | |
|---|---|
| 7 | 444 |
| 8 | |
| 9 | 22222 |
| 10 | |
| 11 | 1111111 |
| 12 | 9999 |
| 13 | |
| 14 | |
| 15 | |
| 16 | 6666 |
| 17 | |
| 18 | 55 |
| 19 | |
| 20 | |
| 21 | |
| 22 | |
| 23 | |
| 24 | 0 |
| 25 | 9 |

HI: 38.8, 51.8

图5　1974年12月份美国纽瓦克机场0100时风速（公里/小时）干-叶图

图中数字意义如下：0 | 22=0.2, 0.2公里/小时；
7 | 444=7.4, 7.4, 7.4公里/小时

对于有些数据组，需要将基本的干-叶结构进行修改。当少数极值离开其他数据值很远时，人们可能会想要增加几个不带叶的新干以适应这少数的数据点。为了防止这种情况发生，我们可以构造一个特殊的干（通常

标以“HI”或“LO”），它能够适应所有的特大数值或特小数值；也可以改变每个干所包含的长度。不管是哪一种情况，都必须保证让看图的人知道该数据组尺度已经作了改变。

当每个干上的叶过多或过少时，也需要作调整。如果每个干上的叶过多，这时只需将每个干分作两行，0—4的叶在第一行，5—9的叶在第二行，构成一个“扩展的”干-叶图。例如

0 | 1112344556777899

应变成

0 | 1112344

0 | 556777899。

图3是一个扩展的干-叶图；图5给出了几条特殊的干以适应特大和特小的数值。如果扩展的干-叶图包含的叶仍然太多，或者标准干-叶图上的叶太少，则可以使用Tukey(1977)所谓的“压缩”干-叶图，它的每个干由5行组成。因此，上面例子就成为

0 | 111

0 | 23

0 | 4465

0 | 6777

0 | 899。

关于干-叶图的进一步变化，建议读者参看Tukey的著作（1977，第一章）。

干－叶图较之直方图能表示出更多的细节，而在图的制作上又并不比直方图多费许多事。这种图除了能使分析者对于数据组的一般特征有一个“直观”了解外，还能使分析者发现数据的某些意外的或异常的特征（诸如出现次数意外地多或意外地少的数值）。但是，干-叶图仅适用于中等长度的数据组(100—200个数据点)，并且如果数据到达分析者手上时已经分好组，则不能使用干-叶图。

## 2.4 框形图

尽管数据组的细节十分重要，但是人们经常希望用少数几个容易得到并且容易理解的数字来概括一个数据组。这些概括性数字应该包括该数据组长度、数据组的位置和尺度以及数据组偏斜度特征的信息。

数据组的长度是用它的极大与极小值来表示的。关于位散、尺度和偏斜度的信息可以从一小组样本分位数得到。粗略地说，与概率p（$0\leqslant p\leqslant 1$）相联系的样本分位数q是这样一个数字：有 $p\times 100\%$ 的数据均位于q值的下边。与 $p=0.5$ 相联系的分位数称为中位数（它位于数据的“中间”，也就是说有一半的数据位于它的下边，一半的数据位于它的上边）。与$p=0.25$和$p=0.75$相联系的分位数分别称为下四分位数和上四分位数。与$p=(0.1)k$（$k=1, 2, \cdots, 9$）相联系的分位数称为十分位数。

中位数是表明数据位置的一个很好的指标。当 n 为奇数时，即以中间的那个数据值作为中位数，当n为偶数时，则以中间两个数据的平均值作为中位数。虽然中位数并不是正态假设下的一个可能的最佳估计量，但它在各种各样的情况下都能保持稳定。也就是说，中位数是一个稳健估计（2.7节将详细地讨论稳健估计）。

关于尺度和偏斜度的信息由下四分位数和上四分位数($p=0.25$和$0.75$)提供。上四分位数和下四分位数的差值（四分位数间距）是尺度的一个稳健估计量(对于正态分布的数据，

四分位数间距等于1.349乘以标准差），而两个四分位数与中位数的差值可以给出数据偏斜度的某些信息。

因此，利用两个极值、两个四分位数和一个中位数这样五个数值就能很好地概括一维数据组，并能用于框形图方法对一组数据进行描述（Tukey，1977）。如图6所示，制作框形图第一步是从一个四分位数到另一四分位数画一框形，并在中位数处画一横线穿过框。然后从框形的每一端画一竖直的虚线，直至相应的极值处为止，用一短的横线标明每个极值。图7左边是图5的风速数据的框形图。数据明显向正的一侧偏斜，这从中位数离开两个

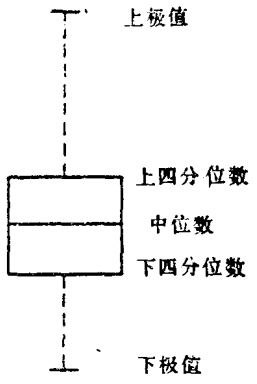


图 6　框形图的结构

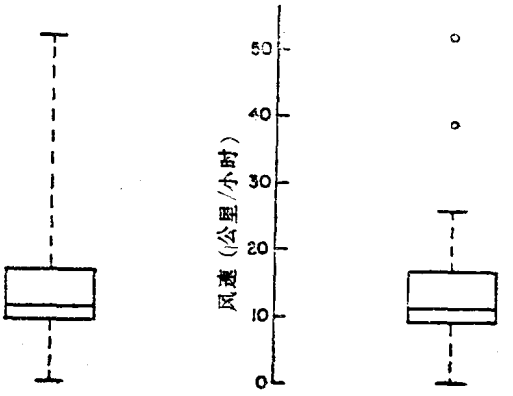


图 7　图5中风速数据的框形图（左）和框形示意图（右）

四分位数的距离可以看出。此外，上极值与上四分位数的距离要比下极值与下四分位数的距离大得多。

这里所给出的基本框形图只着重讨论数据组中间那50%的数据（即框形内的数据），但并未给出位于四分位数与极值之间的那一部分数据的信息。例如，在图7中我们无法知道极值51.8究竟是个别的远离中心的值还是在其他日子里也出现过类似的高风速值。

为了增加框形图的信息，并对框形外面一些特殊的点予以注意，Tukey（1977）提出了框形示意图。在示意图中框形仍然是从下四分位数画到上四分位数，并通过中位数画一条线横穿框形，但竖直线却并不一定要画到极值为止。在框形示意图中，竖直线终止在这样一个区间的极值上，该区间为（下四分位数－1.5×四分位数间距，上四分位数＋1.5×四分位数间距）。在此区间外的所有测值均单独点绘，并经常具体地标出。之所以选择四分位数间距的乘数1.5，是为了要使得在正态分布中有95%的数据均位于虚线的范围内。因此框形示意图表示出有多少（以及哪一些）数据点形成了长尾。图7右边是纽瓦克机场风速数据的框形示意图，在该图中可以看出，上部的长尾仅是由两个数值形成的。

经常是将框形图与一维散布图结合起来使用为宜。散布图表示原始的数据，而框形图则概括出它们的主要特征。这两种图除了能很好地描述出数据组的主要方面外，还具有十简洁的特点，因此它们在对几个数据组的分布进行比较时很有用（见4.2节）。

## 2.5　概率图

上面所介绍的图示方法都能让人较好地看出所考察数据的分布型式，但是却不能确定一个数据组与一给定的理论分布的一致程度。用拟合优度试验可以确定数据组与一给定的理论分布之间是否有显著差异。然而，这种试验并不能告诉我们该数据组与该理论分布之

间在何处发生差异，怎样发生差异。而概率图则不仅能表示出一个经验分布与一给定的理论分布在总体上的拟合程度，而且从图上可以一眼看出这两个分布是怎样发生差异的。

概率图有两种基本类型：P-P图和Q-Q图。这两种图都可以用来把两个分布作相互比较。在比较两个理论分布，一个经验（或样本）分布与一个理论分布，或两个经验分布时，这两种类型的概率图所用的基本原理是相同的。图8可以用来对 P-P和Q-Q 图作很好的解释。在图8中，画了两条光滑的累积分布函数曲线 $F_x$ 和 $F_y$。在制作P-P图时，从共同的分位数q开始对x和y的样本分布计算概率 $P_x(q)$ 和 $P_y(q)$（表示在图8的y轴上）。因此，P-P图是由以概率〔$P_x(q)$，$P_y(q)$〕为坐标的那些点（对于不同的分位数值q）所组成的。

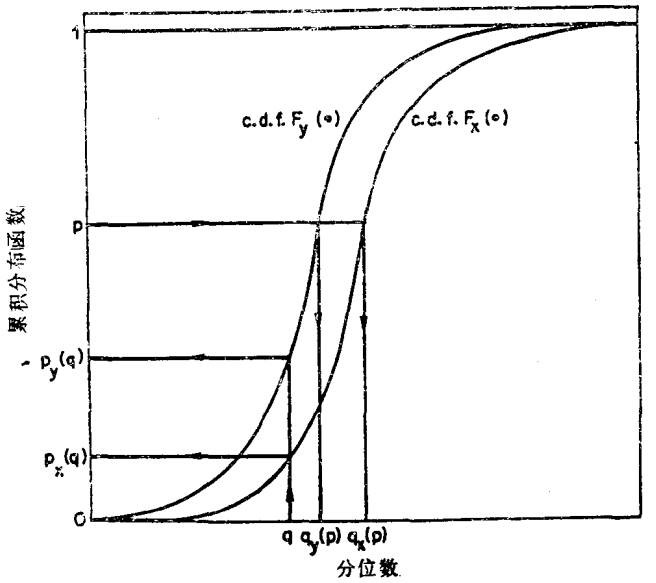


图 8　概率图示例

图中c、d、f表示累积分布函数

（摘自Gnanadesikan，1977，承蒙威利父子出版公司的允许）

为了制作Q-Q图，可以用同样的方法从共同概率p开始计算分位数 $q_x(p)$ 和 $q_y(p)$（表示在图8的x轴上）。因此，Q-Q 图是由以四分位量〔$q_x(p)$，$q_y(p)$〕为坐标的点（对于不同的概率值p）所组成的。

如果两个密度函数曲线大致呈钟形（它导致图8中那样形状的分布），则 P-P 图对于分布中部的差异比较敏感，而Q-Q图则对于尾部的差异比较敏感。如果 $F_x=F_y$，则P-P图和Q-Q图都变成为一条截距为0、斜率为1的直线。如果随机变量 X 和Y具有关系 $Y=ax+b$，则除了一个线性变换外，这两个分布相同。在这种情况下，$q_y$ 与 $q_x$ 的 Q-Q图仍然是一条直线（这时截距为b，斜率为a），这是因为对于每个固定的p和正数a有

$$p=Pr\{X\leqslant q_x(p)\}=Pr\{aX+b\leqslant aq_x(p)+b\}=Pr\{Y\leqslant q_y(p)\}$$

因此有

$$q_y(p)=aq_x(p)+b$$

这个线性不变性对于P-P图却不能成立，这可能正是P-P图远不及 Q-Q图样那被广泛使用的原因。

Q-Q图的最广泛用途是将一个长度为 n 的数据组的经验累积分布与一个假设的标准化