数学中国培训（之数据分析概论）

作者：韩海涛

大家好，这里是数学中国，数学建模在线培训活动。我是数学中国超级版主韩海涛。我们这次讲的内容是数据分析概论。

数学中国讲师(17515041)　19:01:33

数据分析是一个非常庞大的议题，所包含的原理和技术不可能在一次培训中得到充分涵盖。所以我们在这一次培训当中，只讲一些最基本的内容。偏向于基本概念的介绍，具体技术介绍得较少。

数学中国讲师(17515041)　19:02:24

有一些具体的计算方法同学们可以自己在相应书籍当中查到。为了达到较大的适用范围，本次培训内容尽量减少概率和统计术语的使用，偏向目的性的介绍而不追求数学上的严格性，内容可能对某些赛题有一点作用，但其设计并不针对特定赛题。

数学中国讲师(17515041)　19:03:43

统计学、数据挖掘和机器学习等学科从本质上讲都是在收集和分析数据。它们的偏向各有不同，统计有专门的统计专业，而数据挖掘和机器学习则属于计算机专业。但从学科性质上讲，完全没有必要强调它们的不同，事实上，它们的目的无非都是从数据中寻找信息，所以我们可以把它们看成统一的学科。

数学中国讲师(17515041)　19:06:24

我们通过数据分析得到的信息，从性质上和数据本身是相关的。激进一点说，我们能得到的充其量就是“数据如何生成”的信息。更多的负担不要都加在数据分析这项工作上，它可不一定能承担得起。

数学中国讲师(17515041)　19:07:38

实际的世界是先有某种机制，再通过该机制产生了许多可观测的数据。而数据分析则是这个流程的反问题，通过观察到的数据反演原来的机制是什么。

数学中国讲师(17515041)　19:08:40

所以我们的数据原则上是可靠的，但毕竟有限，所以不够全面。推得的信息要求具有普遍性，但毕竟带有推测性，不能强求其绝对吻合真实。这也是数学建模的特点，不强求“正确"，更多的是“合理”或者“有用”

数学中国讲师(17515041)　19:10:08

严格地讲，数据分析不是一次就能彻底结束的操作，而是一个不间断的流程。不仅为了回答特定的问题，而且应该能指示如何改进，以提供进一步研究的导向。

数学中国讲师(17515041)　19:11:26

正因为如此，一次完整的数据分析流程，一定要包含如下几个步骤：

1：数据收集；

2：数据清洗和交叉检验；

3：建模；

4：推断分析、决策及结论；

5：对进一步调查分析的导向；

数学中国讲师(17515041)　19:12:51

这几个步骤大致是按照时间顺序来排列的，但是其中也有交叉的部分。尤其是“建模”这里，从刚接触到数据，就已经开始有了模型的介入。直到最终，模型也拥有不可或缺的影响力。所以这个划分不是机械的。我们下面分别叙述这些步骤的主要内容。即使是概述，估计一次也讲不完。如果有必要，以后可以继续组织培训活动。

数学中国讲师(17515041)　19:14:33

（第一步）数据收集是数据分析的第一步。

在数学建模中，由于我们很少需要进行实地调查和设计实验，所以只谈谈如 何抽样就可以了。

数学中国讲师(17515041)　19:16:14

有时可以获取的数据太多，处理起来困难，而且这些数据可能并不都是必需的。所以需要进行抽样，只分析抽到的样本就可以了。这些样本需要代表总体情况，所以在设计抽样方案的时候，我们需要刻意避免加入人为的偏倚（bias）。

数学中国讲师(17515041)　19:17:50

最简单的抽样方法就是随机抽样：从总体中等概率地抽取若干样本并加以分析。如果运气不好，抽到的样本恰好是有偏的，得到的结论当然就不适用。但我们可以有效地评估风险。

数学中国讲师(17515041)　19:19:21

例如：

我们分析中国人的平均身高，随机抽取了 100 人作为样本。如果真的随机抽到了 100 个高个子，那么分析出来的结果肯定比实际偏高，不过我们可以设法评估出来出现这种“不良现象”的概率有多大。当样本的数量较大时，这个概率极小，我们也就可以忽略了。

数学中国讲师(17515041)　19:20:49

但是这里毕竟有随机产生偏差的可能，这会影响样本的“代表性”。为了尽量减小这个问题，我们可以采用一些随机抽样的变种。

数学中国讲师(17515041)　19:21:18

随机抽样的一个变种是分层抽样。

数学中国讲师(17515041)　19:21:48

这是指在抽样的时候，先把总体分成不重叠的几类，给每一类派发若干样本名额，然后在每一类中随机抽取样本。如果总体内部本身有许多不同的类别，每个类别之间有显著不同，那么为了减小随机的偏差，这个方法是很奏效的。

数学中国讲师(17515041)　19:22:44

每个类中派发的样本名额，许多时候与这个类的总体数成正比。

数学中国讲师(17515041)　19:23:32

例如：

我们试图对现代人的肌肉力量进行研究，一般来讲，男性和女性在此方面有显著的差别。为了避免样本中男女比例的随机失调，所以我们给男性和女性各分配若干样本名额。其样本数量之比应该等于总体中男性与女性的人数之比，这样才能避免引入人为的偏倚。

数学中国讲师(17515041)　19:25:24

这很类似我们做数学题时“分类讨论”的思想。不过，这个类别在分的时候，也许不同的人有不同的分法。一般来说，每类之间的区别要尽量明显，而每一类内部的区别应当较小才是。

数学中国讲师(17515041)　19:27:03

C.R.Rao 在《统计与真理》中举例：

为了推断印度西孟加拉省的城镇人口，可以在这里随机抽取一些城镇，调查其人口数，以此为数据来进行推断（一般来说，只要算出样本城镇的平均人口数，再乘以城镇数量就可以了）。

数学中国讲师(17515041)　19:28:27

但是一旦抽取的若干样本里包含了加尔各答（是当地最大的城市，人口总量和密度都远远超过任何其他城市），那么推断出来的结果势必有重大误差。

数学中国讲师(17515041)　19:29:15

但是要是在样本里不包括它的话，其实也会产生非常大的误差，估计的结果会严重偏低。

数学中国讲师(17515041)　19:30:05

所以合理的方法是把加尔各答独立出来，并对别的城镇进行随机抽样，去估计除加尔各答以外地区的人口数量，最后再相加。这也可以看成是一个分层抽样的案例。

数学中国讲师(17515041)　19:32:03

由于毕竟“分层”具有某种“人为性”，所以这个问题深入下去会引来许多麻烦。而且在许多复杂的问题中，说不定会引入一些人为的偏差而难以发觉 。

数学中国讲师(17515041)　19:33:09

如果在分层的时候，不同类别的“变异性”不同，可以使用这样一个原则来分配名额：各层应抽样本数与该层总体数及其标准差的乘积成正比（称为奈曼法）。也就是样本数据较为集中的类可以少抽一些样本，而本身数据之间就相差很大的类可以多抽一些样本。

数学中国讲师(17515041)　19:34:53

我们通过肌肉力量的问题来理解这个道理：如果男性的肌肉力量都大致相当，女性的肌肉力量则相差甚大（这只是假设，不是医学事实），那么分配样本名额的时候，男性的样本数量少一些，女性的则多一些，会更有利于减少（组内）随机偏差的产生。

数学中国讲师(17515041)　19:37:19

但是如果并不了解每个类别的特点（如其标准差），这个方法也就没的用了。简单地按照每组的大小来分配名额，在什么时候都是凑合能用的方法。

数学中国讲师(17515041)　19:39:01

还有一种抽样方法是有偏抽样。有的时候我们故意做成有偏的，有的时候是没办法。比如我们考察草原上某种鸟类在一年的产卵平均数量，无论怎么考察，这个数据肯定是有偏的。因为产卵 0 枚的这个现象，根本不可能观察到。

数学中国讲师(17515041)　19:40:41

但是如果没有意识到抽样的“有偏性”是很危险的，而且进行推断分析的时候会变得较为复杂。我们以后有机会再专门谈这个问题。

会员提问：

huashi3483　19:39:24

hdw　19:38:28

1、那如何能够评价一个分配方法的好坏呢

Nautilus　19:41:35

类别间的区别越大，每类内的区别越小，这个方法越好。不过一般在应用的时候，不同的类别一般都是有显著性质上区别的。这样会“名正言顺”一些。

数学中国讲师(17515041)　19:42:43

在建模问题当中，如果我们看到数据非常复杂，尤其是不易用计算机直接处理，可以首先进行抽样。对为数相对较少的样本进行分析，可能更加简单有效。如果使用手工方法进行处理，那抽样更是肯定要做的步骤了。

数学中国讲师(17515041)　19:43:54

有关数据收集的问题，可以被归类于“应用统计学”这个学科。包括设计问卷，设计实验，等等问题。它最终的目的是收集到合适的数据，公正客观，有代表性，而且还相对便于处理。

数学中国讲师(17515041)　19:45:30

（第二步）下面我们来看数据清洗和校验的问题。

数学中国讲师(17515041)　19:46:42

统计分析的目的是：从观测得到的数据中提取各种有效的信息。但是数据本身可能也有问题。可能有记录的错误，可能有异常值，甚至可能是伪造的。在数学建模问题中给的数据质量倒是一般不至于这么差，但是在任何问题当中，对原始数据的质量作一点提防是没坏处的。

数学中国讲师(17515041)　19:48:33

分辨这些现象是很困难的，并无确切的机械方法。所以这里只能讲一些笼统或使用范围极为有限的说法，希望大家能随着经验的积累，依稀辨出处理此类问题的“手感”。

数学中国讲师(17515041)　19:49:21

如果我们处理的数据只是数字而已，那么问题就会更好叙述一些。如果出现了看似异常的值，例如比其它数据明显大或小许多，那么有这么几个可能性：

数学中国讲师(17515041)　19:50:08

1：异常值是记录出错的结果；

2：异常值代表的那个个体并不属于所研究的总体，或者与样本中其它部分有本质的区别（想想加尔各答和小城镇的区别）；

3：所研究的总体本身就呈现出 Fat-tail 或者叫 Heavy-tail 的分布，例如所谓的幂律分布，这样的分布本来就容易出现有巨大差别的数字。譬如人的收入，最大者比平均值相差了若干数量级。譬如互联网上不同网站拥有的超级链接的数量，平均值<10，但同时也有 google 等远远超过通常网站能力的成员。

数学中国讲师(17515041)　19:53:24

我们可以有相应的处理方案：

1：剔除异常值，使数据更“干净”；

2：剔除异常值，但在此后的分析中要作一些相应的修正（还可以回想加尔各答的问题）；

3：认为这些看起来异常的值事实上是正常的，并选择可以产生这类值的“合适的模型”来处理问题。

数学中国讲师(17515041)　19:55:07

看：在这里其实已经涉及到模型的建立和选择了。对纯粹的数值型数据而言，所谓模型，基本上指的就是它服从的分布（当然对一个庞大的实际问题而言，模型二字有着更多的含义）。

数学中国讲师(17515041)　19:56:31

对这些数据来说，可能有某个简单的理论模型可以吻合其状态。比如某些数据“理应”是等概率分布的，或者十有八九可以猜测其是正态分布的，等等。

在某种意义上说，这个只针对数据的“小模型”也可以看成我们建模整体工作的一个部分，甚至是很重要的部分。

数学中国讲师(17515041)　19:58:41
说句脱离主线的话：整个数据分析的工作流程，有时是“数据驱动”的，有时是“模型驱动”的。

数学中国讲师(17515041)　19:59:20
或者说，有时是模型在先，或者至少有一个较为基本的模型，再通过数据将其明确化，或者作出进一步的推论，这称为“模型驱动"。

数学中国讲师(17515041)　20:00:13
而有的时候，模型根本不存在，除了数据，一概未知。模型完全是从数据得到的。

数学中国讲师(17515041)　20:01:28
一般来说，数据驱动的问题会更难一些。尝试性更强，往往都需要使用“预期以外”的方法发现“预期以外”的结果。其实，所谓数据挖掘这门学科，更多关注的是这类工作。

数学中国讲师(17515041)　20:02:43
大多数时候这两个思路需要交融在一起，而此时也正是所有问题最难的地方。

数学中国讲师(17515041)　20:04:34
在这时，我们一般都会依照经验、背景知识或一些合理的猜测，给这些数据赋予一个较明确的模型，当然是否真的和这些数据相符，还需要验证。边尝试，边修改，交叉使用这两种思路，才能得到一个比较有效的结果。

数学中国讲师(17515041)　20:05:14
幸亏大多数数学建模的问题，在这个地方没有我说得这么难。要不然，拿到数据，光开工就开不了，可就麻烦死了。

数学中国讲师(17515041)　20:07:34
我们以“检验数据是否符合正态分布”来阐述其中可能涉及到的一些技术。这些技术在此后的许多问题当中都有用，包括统计学里的几个重要部分：图形描述，参数估计和假设检验。这一段比较长，而且会谈一些特定技术，看起来会有点离题的感觉。

会员提问：
数学中国 ceo(20694876)　20:10:14
huashi3483　20:07:32
****　20:07:05*

经验、背景知识或一些合理的猜测

其中背景知识有哪些？举例说明一下？？？

Nautilus 20:09:40

所谓背景知识，一般是指所谓先验信息。譬如一些基本的科学原理，或者已经被先前资料验证过的结果，等等。比如硬币扔到地上，是否正面反面的概率是相等的?有“对称性”的道理，还有实际的实验结果。当然这个例子太简单了点，大家觉得没悬念。但复杂一些的问题，说穿了也是这么回事。

数学中国讲师(17515041) 20:12:13

我们来看看正态分布是怎么回事。一个问题中的数据，如果同时受到各种复杂因素的影响，并且各种因素对数据起作用时往往是互不相关的，经常会出现正态分布。正因为其常见，所以才有了这个称呼。

数学中国讲师(17515041) 20:13:49

正态分布的密度函数：

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

数学中国讲师(17515041) 20:15:22

密度函数是概率的基本问题，如果有同学不了解，我们可以在以后讨论，也可以自己看看概率的初级教材。

数学中国讲师(17515041) 20:16:21

总体的数据可能是服从正态分布的。我们抽出来的样本可以看成是总体数据的一部分，由于随机偏差的影响，如果光看这部分样本，不可能真正精确地服从正态分布。

数学中国讲师(17515041) 20:18:03

但当我们进行随机抽样时，可以认为这些样本大致代表了总体的情况。所以从最严格的意义上来说，我们是通过这些样本，来“推断”总体数据是否服从正态分布。

数学中国讲师(17515041) 20:18:23

如果真说的话，这已经属于“统计推断”的问题了。

数学中国讲师(17515041) 20:19:33

事实上，在实际问题中，我们也可以这样认为：不谈所谓“总体”的问题，只去谈这些样本和正态分布是否吻合。当然，些许偏差是难免的，但是如果它们的偏差在允许范围之内，我们即可认为它们相符。

数学中国讲师(17515041) 20:20:12

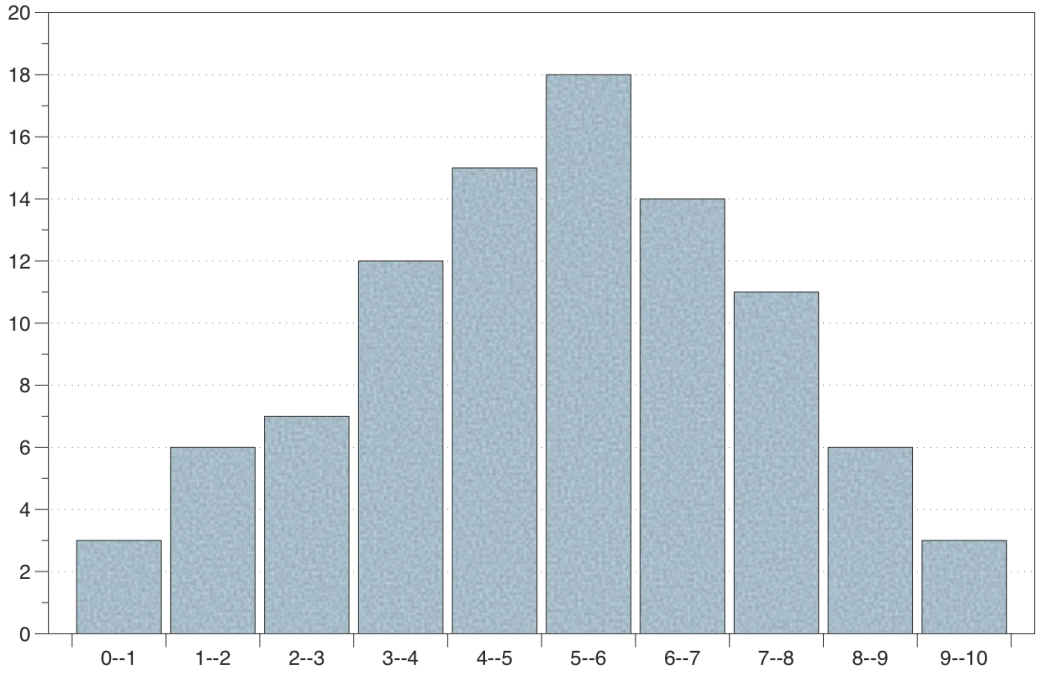
这两种观点没有实质的区别。所以不管怎么看，最重要的问题就是这些样本的情况和正态分布的理论情况是否一致，如何衡量它们之间的偏差，而多大的偏差是被允许的。

数学中国讲师(17515041)　20:21:05

第一个可能的方法是图示。直方图是让肉眼看出分布的最直接方法。假设数据分布在[0,10]这个区间里。假设我们把[1,10]分成 10 个等距的区间：[0,1),…,[9,10]

数学中国讲师(17515041)　20:21:49

统计每个区间内分别落入多少个数据，并以其数量为纵坐标作图。这样的图称之为直方图。为了更清楚地看到落入每个小区间的数据占总数的比例（也称频率），我们有时将纵坐标除以数据总量。实际问题中画的直方图多是这种情况。



数学中国讲师(17515041)　20:24:15

直方图可以大致看出哪些位置的数据密度比较大。但是作直方图的时候，需要选择合适的区间大小。

数学中国讲师(17515041)　20:25:03

如果每个小区间的长度太小，可能造成相邻两个小区间的统计值区别甚大，这是扰乱视线的噪声。但长度太大，精度又太差。如果我们把整个数据范围只分成两三个区间，那就什么结果都看不出来了。

数学中国讲师(17515041)　20:27:08

直方图也有画成横坐标区间大小不相等的，但在这种情况下作图的时候一定要注意：数据落入小区间中的频率不是纵坐标，而是该区间上的长方形的面积。按理说频率在图中本来就应该表示成面积。不过在小区间都分成等距的情况下，把它看成高度也没有什么问题，而且画起来方便一些。

数学中国讲师(17515041)　20:28:54

从直方图我们可以直观地作这样一个检验：如果数据服从正态分布，理应也有其性质，绝大多数的数据应当分布在这个区间之内：

$$[\mu - 3\sigma, \mu + 3\sigma]$$

数学中国讲师(17515041)　20:30:53

我们可以看到正态分布的密度函数，大致是一个 e^{-x^2}类型的函数。（^表示乘方）

数学中国讲师(17515041)　20:31:38

这种函数在 x 增大的时候，下降的速度是极快的。也就是说，偏离 mu 值较远的数据应该极难出现。所以如果数据不吻合这个性质，几乎可以断定其不是正态分布的。（我用 mu 来表示均值）

数学中国讲师(17515041)　20:33:22

所以在并非海量数据的时候，例如数据只有 1000 个，如果在偏离均值 5 到6 倍 sigma 的地方都出现了数据点，那么基本可以断定，这批数据不应服从正态分布。更可能是被称为“Fat-tail”或者“Heavy-tail”的另外一些分布形式。如果想断定数据的分布是不是指数分布或幂律分布，只需要把直方图画到单对数或双对数坐标下，看其是否呈现出直线，就很明确了。

数学中国讲师(17515041)　20:35:24

但是直方图只能给人的肉眼提供一个初级的看法，并不见得准确。还有一个更有效的方法是 QQ 图。

数学中国讲师(17515041)　20:36:25

QQ 图（Quantile-quantile plots）是样本点与标准正态分布的分位数的散点图。

数学中国讲师(17515041)　20:37:40

分位数的概念是：把数据从小到大排序，位于序列正中的数字叫做中位数，而位于 70%位置的称为 70%分位数或 70 百分位数。其实分位数和百分位数是两个不同的名词，但其本质并无区别，所以我们在这里同时提到并不做明显的区分。

数学中国讲师(17515041)　20:39:21

大体上讲，第 p 百分位数是 k，意味着 k 是这样一个数据项，它使 p%的数据项<k，且有 1-p%的数据项>k。

数学中国讲师(17515041)　20:40:41

对分布也可以考虑分位数的概念：如果一个正态分布的随机变量，它的取值小于某个数字 k 的概率恰好为 70%，那么这个数字 k 就可以称为正态分布的 70%分位数，诸如此类。

数学中国讲师(17515041)　20:42:29

分位数的这两个说法事实上反映的是同一个道理：如果一个数字是 p%分位数，那就意味着数据比它小的可能性占 p%，而比它大的可能性占 1-p%。但由于正态分布的密度函数比较复杂，所以它的每个百分位数没有直接的计算办法，需要使用计算机或者查表。

数学中国讲师(17515041)　20:43:28

我认识一个刚毕业的学生，在给中国移动做数据分析和咨询工作，居然用分位数这个概念就唬住了客户。

数学中国讲师(17515041)　20:44:07

插一句有关坐标的问题：

数学中国讲师(17515041)　20:47:50

如果我们在作图的时候，横坐标就是 X，纵坐标就是 Y，这叫做线性坐标。如果点的横坐标取的是 logX，而纵坐标是 logY，这样的坐标系叫做双对数坐标系。一个本来是幂函数的图像，例如 Y=aX^k，我们可以化成 logY=klogX+loga，也就是说在这个坐标系下变成直线了。

数学中国讲师(17515041)　20:48:27

我们的眼睛识别曲线是否符合某个曲线是很难的，但是识别它直不直是很清楚的。

数学中国讲师(17515041)　20:49:24

回头来讲 QQ 图：QQ 图的作法是：把数据从小到大排列起来，这样每个数据就都可以看成整个数据序列中的某个分位数。然后以每个数据的值为纵坐标，以正态分布的相应的分位数为横坐标作散点图（Scatter plot）即可。这样，如果数据点都分布在直线 y=x 周围，说明大体是符合这个正态分布的。如果有一些数据点明显离开该直线，说明这些数据点和正态分布不吻合，也就是离群点，甚至完全可能是异常值。

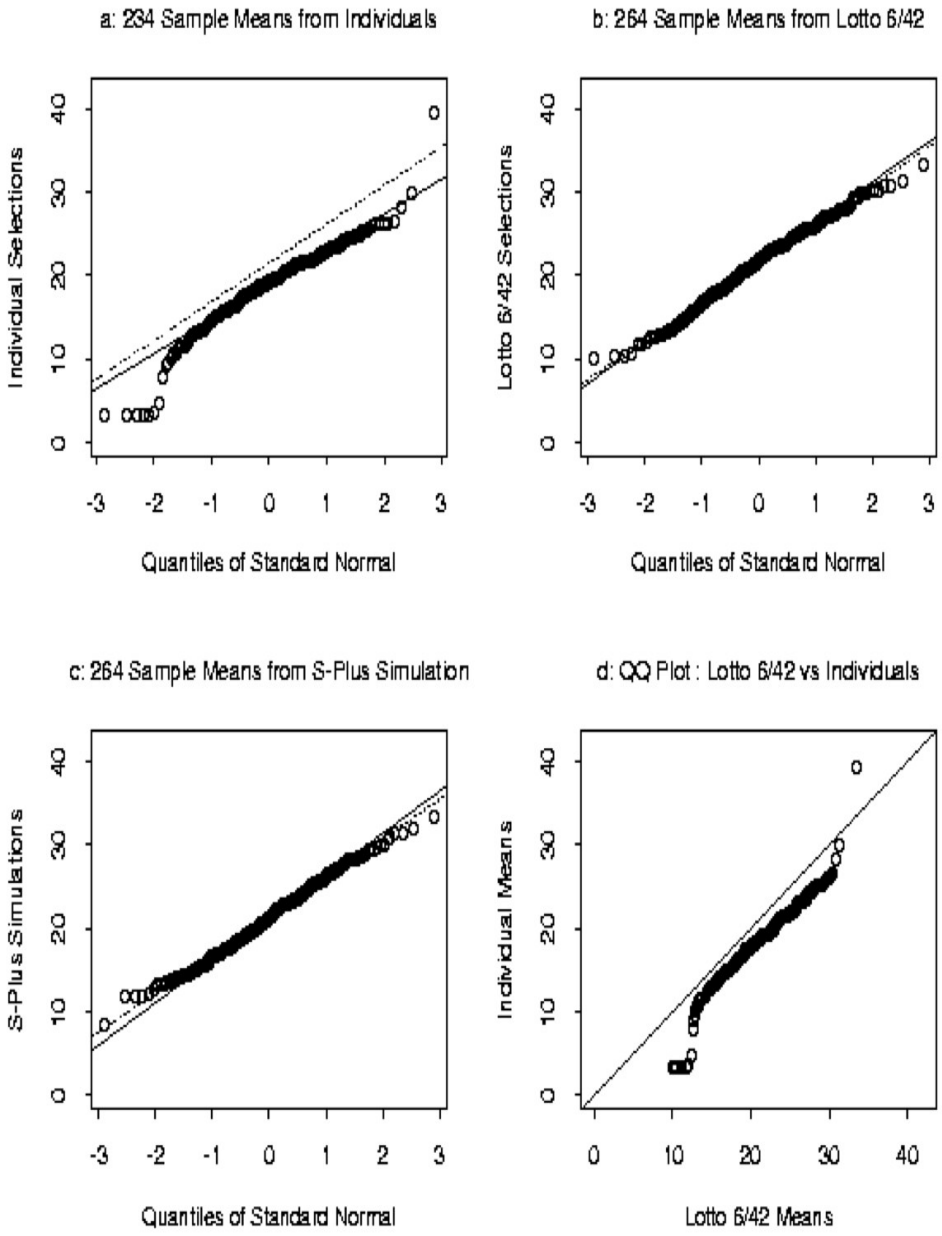
数学中国讲师(17515041)　20:51:29

或者说，如果数据是完全严格地符合正态分布的，那么数据的多少分位数和理论的多少分位数都完全一致，这样，代表数据的点在这个图上，应该完全位于直线 y=x 上。如果有偏差，那就是数据和正态分布之间的偏差了。

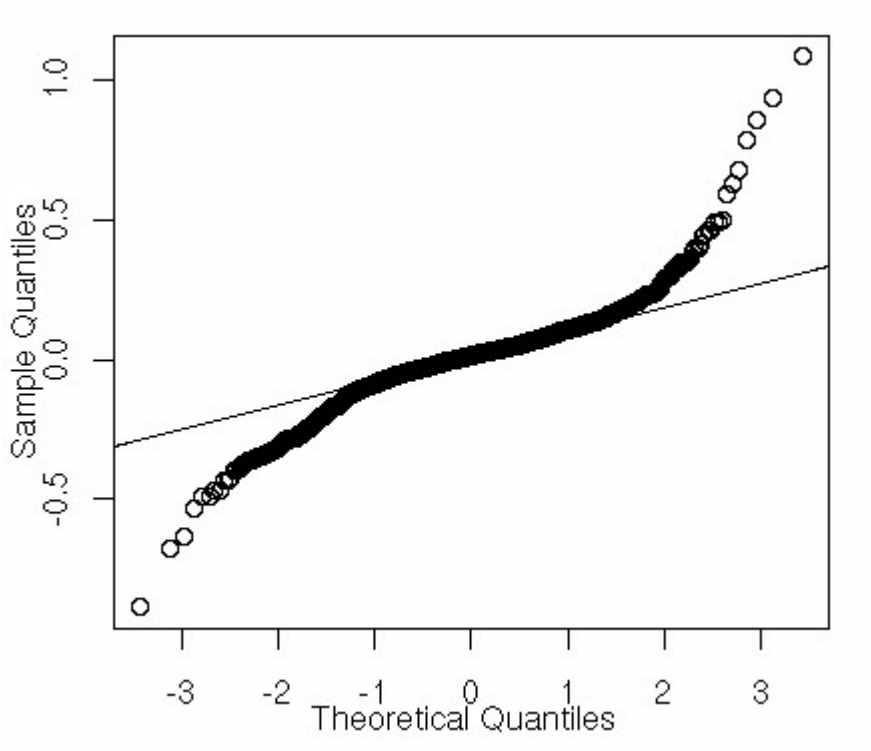
数学中国讲师(17515041)　20:53:15

这是典型的 QQ 图。

数学中国讲师(17515041)　20:53:10



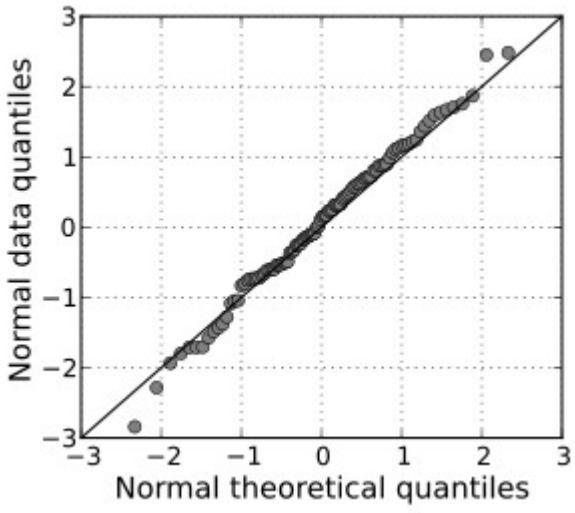
数学中国讲师(17515041)　20:53:21



数学中国讲师(17515041)　20:53:29

这个也是。我们再来看一个：

数学中国讲师(17515041)　20:53:34

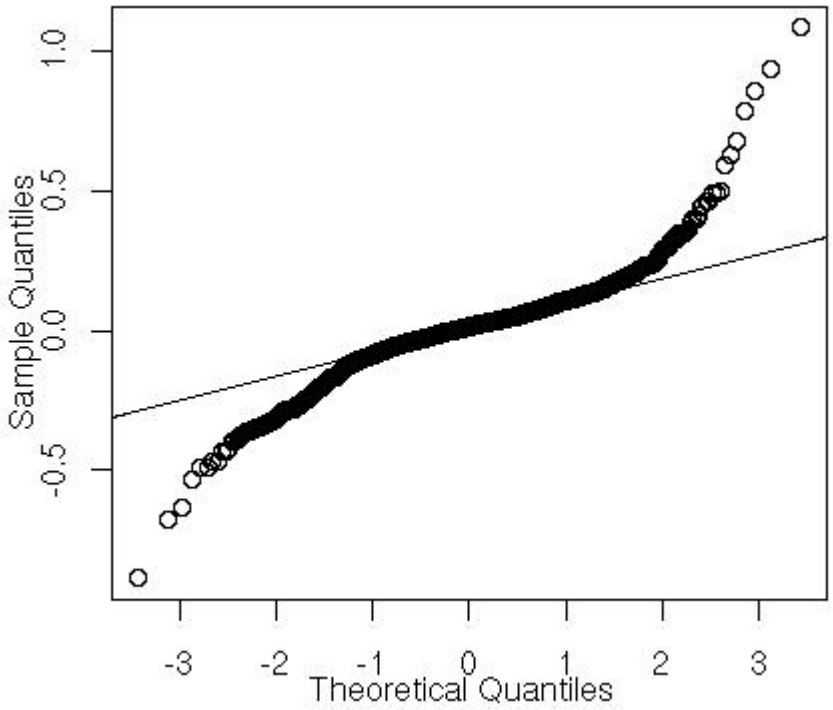


数学中国讲师(17515041)　20:57:24

OK，我们先来看看最后发的这个图。QQ 图有一个常见现象：由于“极端”值出现的次数本来就稀少，更容易和理论情况有所偏差，所以在数据点列的两端往往都会和理论直线有一些微小的偏离。偏差不大时可以忽略。这个图可以看成数据非常符合正态分布的情况。

数学中国讲师(17515041)　20:58:22

我们再来看这个图。

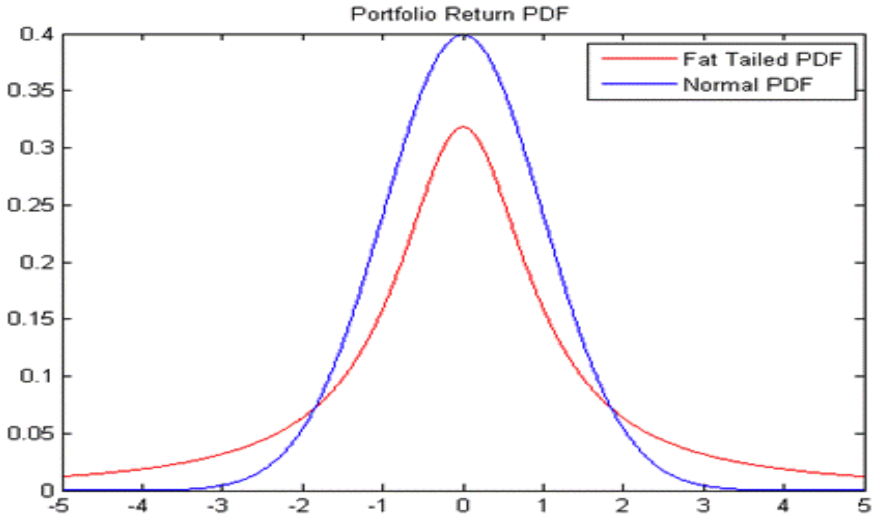


数学中国讲师(17515041)　20:59:20

代表数据的点列虽然在中段和代表正态分布的直线较为吻合，但左下角的数据点明显在 y=x 的直线下方，右上角的数据点明显在该直线上方。这意味着，在我们的数据序列中，最小的几个（它们位于左下角）比正态分布的典型情况要小，而最大的几个（它们位于右上角）比正态分布典型的情况要大。这么多点都偏离，已经不能说“离群点”或者“异常值”了。

数学中国讲师(17515041)　21:01:16

这说明我们的真实数据具有比正态分布更“厚”的尾部，也就是远离平均值的情况（也就是比较少见而“极端”的数据），出现的可能性比正态分布的典型情况要大得多。例如下图



数学中国讲师(17515041)　21:02:37

蓝线是正态分布，而红线则是一种 fat-tail。5 或-5 这类“极端”的值比正态分布的理论情况更容易出现。

数学中国讲师(17515041)　21:03:24

几乎一切统计软件都可以画出漂亮的 QQ 图。SPSS，matlab 的统计包，R 等。推荐 R，免费的。

数学中国讲师(17515041)　21:06:48

这里的红线并不是正态分布。标准差再大的正态分布，密度函数毕竟也是 e^{-x^2}这个量级的，充其量是系数不同而已。但是这个红线，其实是一个叫做 levy 分布的东西，并不是刚才那个函数。

数学中国讲师(17515041)　21:08:24

有关 fat-tail 的问题我们以后再谈。回头来说 QQ 图，从 QQ 图的原理上我们可以看到，我们可以用它来判断这些数据是否吻合任何一个分布，不仅仅是正态分布。只要会算这个分布的分位数就可以了。统计软件一般都是有这个功能的。

数学中国讲师(17515041)　21:09:41

最后我们来讲一个检验数据是否服从某分布的最有效的方法：卡方检验。

数学中国讲师(17515041)　21:10:18

从检验数据是否服从某分布，到识别数据造假，卡方检验都有其可能的用武之地。我们来看它到底是什么原理。

数学中国讲师(17515041)　21:11:21

卡方检验不是靠图，而是靠计算。这属于统计学中“假设检验”的分支。从本质上讲，一切使用计算方法进行的检验，都是在试图反映“如果总体真的服从我们的假设（在这里，我们的假设就是“数据服从正态分布"），那么出现这些真实数据的可能性会有多大？”

数学中国讲师(17515041)　21:12:47

如果我们的假设真的成立，而我们取得的数据出现此种情况的概率有多大？如果此概率较大，那么情况尚属正常；如果此概率极小，那与其认为我们碰上了极其罕见的好运，还不如说原来的假设是错误的会更合理一些。究竟当概率小到什么地步我们就认为数据并不符合原假设，这个界限叫做“显著性水平”或叫“置信性水平"。

数学中国讲师(17515041)　21:14:09

显著性水平太大，意味着数据与假设只要稍有偏差，我们就不敢确认；反之，显著性水平太低则太冒险。对不同的问题可以有不同的显著性水平，一般日常用途习惯取在 5%左右。

数学中国讲师(17515041)　21:14:51

由于实际问题的复杂性，所以光靠这种基本想法可能不易计算，所以人们总结出了许多成型的算法，可以直接用在不同用途上。卡方检验是最常用的方法。

数学中国讲师(17515041)　21:15:40

卡方检验（Chi-Square test）是一种功能强大的检验算法，它可以用于检验数据是否符合给定的分布（不限于正态分布）。它可以想象成是衡量直方图与理论假设之间差别的一种方法。

数学中国讲师(17515041)　21:16:38

在作直方图的时候，我们记录了数据落入每个小区间的频率，我们把数据落入第 i 个小区间的频率记为 Oi。如果数据服从我们给定的分布，我们也可以算出数据落入这些小区间的理论概率，记为 Ei。（其中 i 是下标）

数学中国讲师(17515041)　21:17:06

当然它们之间一般不会完全相符，所以我们使用这个数字来衡量其距离：

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

数学中国讲师(17515041)　21:18:14

其中分子上的 O-E 衡量了在小区间上，理论概率和实际频率之间的差，平方是为了避免正负抵消。

数学中国讲师(17515041)　21:19:35

如果数据真的服从给定的分布，那么当数据总量足够大的时候，T 应当服从参数（或叫自由度）为 c-1（c 是小区间的数量）的卡方分布。这个证明比较困难，而且也只是近似的，所以卡方检验也有一定的适用范围，并涉及到复杂的研究。卡方分布的密度函数非常复杂：

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} \mathrm{e}^{-y/2} & , y > 0 \\ 0 & , y \le 0. \end{cases}$$

数学中国讲师(17515041)　21:20:05

这个密度函数不知道也罢，反正该算的时候一定要用计算机。

数学中国讲师(17515041)　21:20:44

其中那个希腊字母(我写作 nu)是参数，又称自由度。它的图像看起来倒是很简单：

数学中国讲师(17515041)　21:21:53

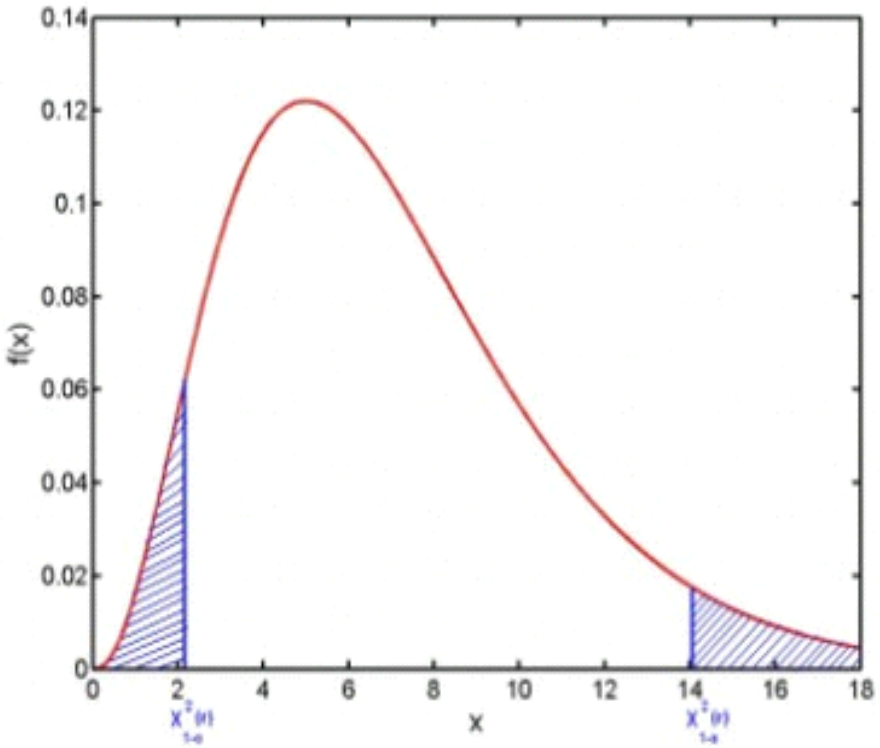
这就是自由度为 1 到 5 时的卡方分布密度函数。

数学中国讲师(17515041)　21:22:49

可以直观地看到，自由度>3 的卡方分布图像有一个峰值，当 x 增大的时候，

函数下降的速度也非常快。结论是：T 出现较大值的概率很低，所以我们算得的 T 如果较大，可以判定为数据和我们给定的分布不符。我们也可以直观地看到，如果 T 太大，意味着实际频率与理论概率总的来说偏差太大，可判定为不相符。

数学中国 ceo(20694876) 21:23:58



数学中国讲师(17515041) 21:24:38

具体来说，如果我们事先选择的显著性水平是 a，那么如果算出来的 T 大于相应卡方分布的 1-a 分位数，则可认为数据与给定的分布不相符。见上图，如果我们算的 T 已经进入右侧的阴影区域，可以判定为“发生了罕见的事情”，进而判定为数据和假设的分布不相符。

数学中国讲师(17515041) 21:26:27

卡方检验还有另外一个用途：如果算得的 T 值非常小会意味着什么？首先，T 如果非常小，意味着数据在每个小区间中出现的频率和理论概率相差无几，也就是说数据非常吻合我们给定的分布。但是我们看上图，左侧的阴影区域高度也甚低，导致了阴影区域的面积也很小，这意味着 T 的取值落在左侧阴影区的概率事实上相当小。这提醒我们，此时数据和给定的分布吻合得过分完美，或许是我们好运，但也可能是数据被伪造或做过针对性修正的结果。

数学中国讲师(17515041) 21:28:05

试试看，譬如我要求你掷硬币 50 次，并记录下正反面。但你别真扔硬币，自己编一个结果，编得尽量像一点。然后卡方检验一下，看看 T 会是多大？

数学中国讲师(17515041)　21:29:10

以上讨论的是我们完全清楚假设的分布，来检验数据是否与之相符。如果我们只想知道这些数据是不是满足正态分布的，并不知道这个正态分布的均值和方差，那就需要先用数据来估计分布的这几个参数。所用的估计方法称为“极大似然估计法”。

数学中国讲师(17515041)　21:30:30

这种做法的思路是：暂且认同数据真的服从正态分布。我们设 mu 和 sigma 取了某个特定的值，然后计算出现这些真实数据的概率。例如我们取得的真实数据是 11,13,12...，而如果 mu=0,sigma=1，产生这几个数据的概率则微乎其微，完全难以置信。

数学中国讲师(17515041)　21:31:13

既然这几个数据已经既成事实了，所以想必未知的 mu 和 sigma 的值，应该较有利于这些数据的出现才合理。所以我们取合适的 mu 和 sigma，使这些数据出现的概率最大。这就是所谓极大似然估计法，在估计参数的时候效果很好。统计软件也基本都有这个方法的内置函数。

数学中国讲师(17515041)　21:33:06

估计完了以后，我们再来看数据是不是符合我们估计出来的这个正态分布。在这里要看明白：这个正态分布的参数，已经是我们为了迎合数据而尽量去取了，如果此时这些数据仍然是“比较罕见的现象”，那可就别怪我不客气了。

数学中国讲师(17515041)　21:34:31

此时检验的方法和前述一样，但区别在于：此时算出来的 T 服从的是自由度为 c-1-r 的卡方分布。其中 r 是估计的参数数量。或者说，当我们多估计一个参数的时候，自由度将减少 1。

数学中国讲师(17515041)　21:35:40

这个问题可以大致地理解为：估计一个参数时，虽然我们用的是什么“极大似然估计法”，但真正用掉的信息量大概相当于“一个小区间的情况”，所以在检验时，“有效的”小区间就少了一个，体现在自由度上就少了 1。（别忘了，自由度就是所用的小区间的数量）

数学中国讲师(17515041)　21:38:56

所以卡方检验的流程就是：1：估计理论分布的参数（如果需要的话）。2：算 T。3：看 T 到底在相应的卡方分布的哪个位置上，是否进入了上图中右侧的阴影区域（那里我们称为拒绝域），如果进入了，则判定为数据与理论分布的偏差过大。当然多提防一点：如果过小，则可以怀疑怀疑数据是否被伪造过？

数学中国讲师(17515041)　21:40:49

卡方检验的原理是这样的，而在软件中实现起来一般只有一个函数，很简单。

最后我们多说一句话，这个事情知道就行了：卡方检验在使用的时候有以下注意事项。首先是数据总量 N 要够大，较保守的估计需要 N>50。

数学中国讲师(17515041)　21:41:24

然后是如果一些 Ei 太小，T 可能较严重地偏离卡方分布，这样就没法使用卡方检验了。但究竟小到什么程度还不清楚，较易掌握的标准是 Cochran（1952）提出来的，他建议所有的 Ei 都不要小于 1，而且小于 5 的 Ei 不要超过 20%。否则我们可以把一些小区间合并起来以满足此要求。

数学中国讲师(17515041)　21:42:58

但一些新的研究表明这个条件可以放宽，Koehler 和 Larntz（1980）的研究称：如果数据总量 N>10,c>3,(N^2)/c>10，并且所有的 Ei>0.25，用卡方检验就都是合适的。

数学中国讲师(17515041) 21:43:28

我们今天的培训就到这里吧。