# 9.2 用数据辨识分布

在这一节中，我们将讨论当可以获得输入时，选择输入分布簇的方法。9.3节将讨论如何通过参数估计，从分布簇中确定一个特定的分布。9.6节将介绍如何处理不能获得数据的情况。

## 9.2.1 直方图

频度分布和直方图在辨识一个分布的形状时是很有用的，直方图可以按如下步骤安装：

将数据范围划分为区间(区间经常是等宽的，然而，如果调整了频度的高度，也可以使用不等的宽度)。

在水平轴上做标记，以确定所选的区间。

以每个区间里发生的维度。

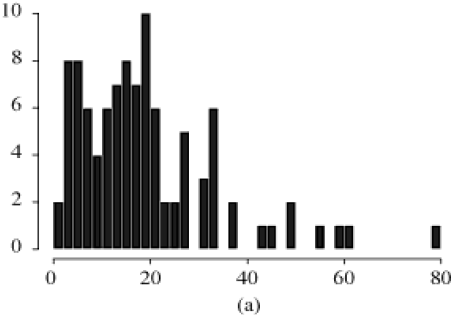
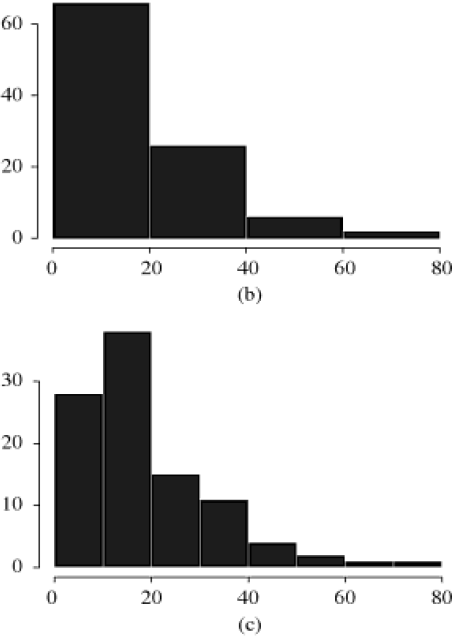
在垂直轴上做标记，以便能画出每个区间的总发生数。

在垂直轴上画出频度。

分组区间的数量依赖于所观测数据的数量和数据的分散量或离散度。Hines，Montgomery Goldsman，and Borrow[2002]指出，在实际中，当所选择的分组区间的数量近似等于样本大小的平方根时，经常会取得较好的效果。如果区间过大，直方图会比较粗糙或呈块状，并且也不能很好地表示它的形状和其他细节。如果区间过小，直方图会有毛刺，并且数据也不平滑。在图9-1中，给出了相同数据的有毛刺的，粗糙的，适当的直方图的例子。现代数据分析软件能够非常容易地和交互地改变距离的大小，直到得出很好的选择。

连续数据的直方图与理论分布的概率密度函数相对应。如果数据是连续的，那么经过每个分组区间频度的中心点画出来的线，其形状与pdf曲线的形状是相似的。

对于有大量数据点的离散数据的直方图，数据范围内的每一个值都应该有一个单元。然而，如果这里的数据点并不多，则有必要合并相邻的单元来消除直方图的毛刺。如果直方图是关于离散数据的，则它看起来应该像概率质量函数。

图9-1 有毛刺的、粗糙的、适当的直方图

例9.2 **离散数据**

在每周的5个工作日中，对每天上午7:00到下午7:05五分钟内到达十字路口西南角的车辆进行观察，一共观察20周。表9-1给出了观察结果数据，第一个条目表示有12次五分钟内没有汽车到达，第二个条目表示10次一辆车到达，等等。

汽车的数量是一个离散变量，并且能得到充足的数据，因此对于数据范围内的每一个可能的值，在直方图上都会有一个单元。相应的直方图如图9-2所示。

表9-1 在5分钟内的到达数

|  |  |  |  |
| --- | --- | --- | --- |
| 每周期的到达数 | 频度 | 每周期的到达数 | 频度 |
| 0 | 12 | 6 | 7 |
| 1 | 10 | 7 | 5 |
| 2 | 19 | 8 | 5 |
| 3 | 17 | 9 | 3 |
| 4 | 10 | 10 | 3 |
| 5 | 8 | 11 | 1 |

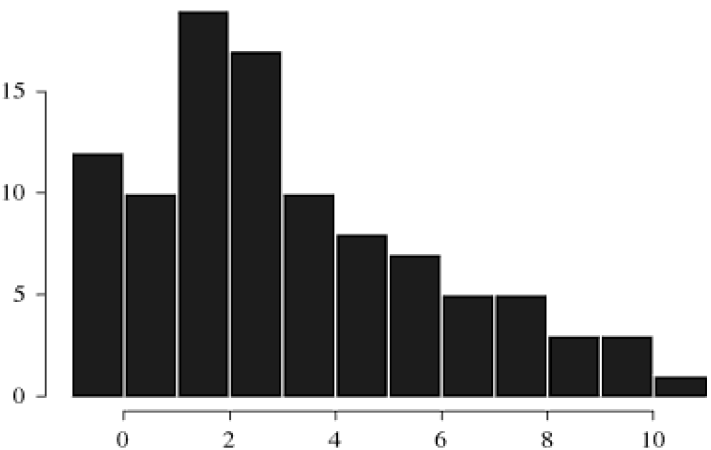
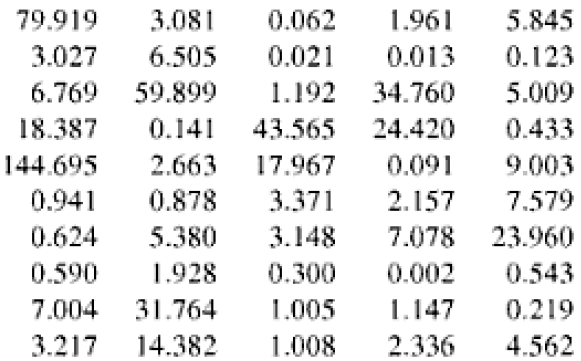
  
每个周期的到达数

图9-2 每个周期到达数直方图

例9.3 **连续数据**

在1.5倍额定电压下，对电子元件的随机样本进行寿命检验，几天中的寿命记录如下：

寿命通常被认为是连续变量。记录精确到十进制小数后三位。将数据放入分组区间中。准备做直方图。数据的范围相当大，从0.002天到144.695天。大部分的数值(50个中有30个)在0到5天的范围内。如果使用区间为三，可以得到表9-2。表9-2中的数据随后用来准备做如下9-3所示的直方图。

表9-2 电子元件数据

|  |  |
| --- | --- |
| 元件寿命(天) | 频度 |
|  | 23 |
|  | 10 |
|  | 5 |
|  | 1 |
|  | 1 |
|  | 2 |
| 18 | 0 |
|  | 1 |
|  | 1 |
|  | 0 |
| … | … |
|  | 1 |
| … | … |
|  | 1 |
| … | … |
|  | 1 |
| … | … |
|  | 1 |

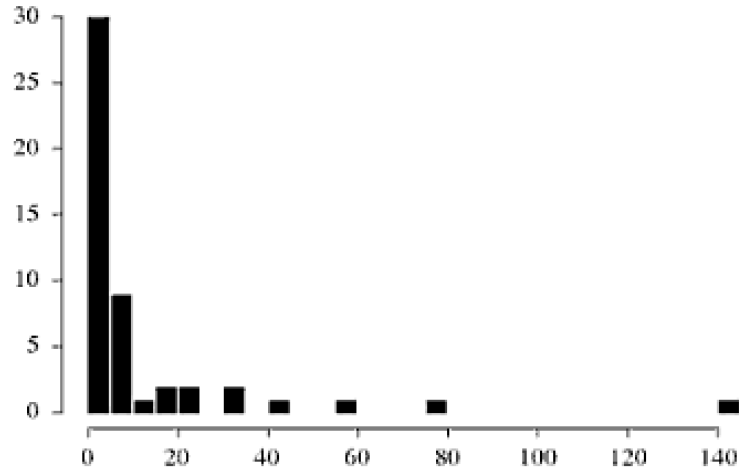


图9-3 元件寿命的直方图

## 9.2.2 选择分布簇

第五章介绍了一些在仿真中经常出现的分布，另外还展示了这些分布的形状。准备直方图的目的是推断一个已知的pdf或pmf。分布簇的选择基于对内容的研究以及直方图的形状。因此如果所收集的到达间隔时间数据的直方图的形状与图5-9中的pdf的形状相似。则有理由假设它服从指数分布。同样在测量货物托盘的重量时，如果直方图的形状像图5-11那样，关于均值对称，则有理由假设他服从正态分布。

指数分布，正态分布，泊松分布是经常会遇到的，从计算的角度来讲，他们并不难分析。分析分布伽马分布和韦布尔分布，虽然难度会稍大，但是它们的形状范围很宽，在对隐含的概率过程进行建模时不应该忽略它们。或许假设了一个指数分布，但是却发现它与数据不拟合，下一步应该检查不拟合的位置。如果不拟合位于分布的尾部，那么伽马分布和韦布尔分布也许会与数据有较好的拟合。

已经建立的概率分布有数百种，很多事考虑到特定的物理过程而建立的，以分布的物理基础作为指导，有助于选择分部下面举几个例子：

**二项分布** 对n试验的成功次数进行建模，此时试验是独立的，具有共同的成功概率p，例如从数量很大的n块计算机芯片中找出的残缺芯片。

**负二项分布**(包括几何分布)对取得k次成功所需要的试验次数进行建模，例如为找到四块有残缺的芯片，我们必须检查多少块芯片。

**泊松分布** 对在固定长度时间和固定大小空间内发生的独立事件的次数进行建模。例如，在一个小时内到达商店的顾客人数。或者30㎡的金属板上所发现的残缺芯片数。

**正态分布** 对一个可以看成是许多子过程之和的过程的分布及建模。例如一个产品的组装过程所需要的时间是每个组装操作的时间的总和。注意正态分布允许负值，这对于过程时间是不可能的。

**对数正态分布** 对一个可以看成是许多子过程之积的过程的分布进行建模。例如，如果利润是复合的，则投资率是许多阶段回报的乘积。

**指数分布** 为独立事件之间的时间进行建模和对无记忆的过程时间性建模。在过程结束之前知道已经过去了多少时间，并不能为得出还需要多少事提供信息。例如数量巨大的相互独立的隐含顾客的到达之间的时间指数分布是一个变化很强的分布，因为它经常得出易于处理的模型。因此在某些时候他被过度的使用。记住：如果时间的间隔时间服从指数分布，则一个固定时期内的事件发生次数服从泊松分布。

**伽马分布** 一个用于为非负随机变量建模的及其灵活的分布。通过加一个常数，伽马分布可以位移离开0。

**分布** 这是一个用于为有界随机变量(有固定的上限和下限)建模的极其灵活的分布。通过加一个常数，分布可以位移离开零，也可以乘一个常数而得到一个比[0,1]大的范围。

**爱尔朗分布** 对一个可以看成是几个指数分布过程之和的过程进行建模。例如当一台计算机和两台备用计算机都损坏时，计算机网络会出现故障，如果每台计算机的故障时间服从指数分布，这时计算机网络故障就服从爱尔朗分布。埃尔朗分布是伽马分布的一个特例。

**韦布尔分布** 对部件的故障时间进行建模，例如一个硬盘驱动器的故障时间。指数分布是伪布尔分布的一个特例。

**离散和连续均匀分布** 对完全不确定性建模：所有的结果是等可能性的。当没有数据时，这种分布经常被不适当的使用。

**三角分布** 当只知道一个过程的分布的最小值和最大值时，用它来建模。例如检测一个产品需要的最小值(最有可能)和最大值时，这种模型通常是对均匀分布的一个明显改进。

**经验分布** 当收集到的实际数据再取样。常常在没有合适的理论分布的情况下使用。

在选择分布时不要忽略过程的物理特性，过程的值在本质上是离散的还是连续的？它是有界的还是没有自然界限的？这些依靠数据的认识，有助于缩小备选的分布簇的范围。记住：对于任何随机输入过程没有“真的”分布。一个输入模型是对现实的近似，因此，目标是得到一个近似，从而由仿真实验产生有用的结果。希望读者能够完成习题6到习题11，以便了解本节所提到的关于分布形状的更多知识。检查形状随参数改变而变化很有意义的。

## 9.2.3 分位点-分位点(q-q)图

9.2.1节讨论的直方图的构造和在9.2.2节中讨论的分布形状的识别，对于选择表示数据样本的分布簇来说都是必要的。然而，直方图在估计分布的拟合时，并不是那么有用。在只有少量数据点。比如说30个或更少。直方图的毛刺情况会很严重，更进一步，我们对拟合的理解依赖于直方图区间的宽度。但是，即使区间选得好，将数据归组到单元中也会使直方图与连续概率密度之间的比较变得很困难。对于评价分布拟合来说，q-q图是一种非常有用的工具。它不会遇到这些情问题。

如果X是cdf为F的随机变量，则X的q-分位点是使的的值。当F有逆时，我们记为。

现在，设{}是一个来自于X的数据样本。将观察值按从小到大顺序排列，并且用{}表示排列好的结果，这里。令j表示顺序号。因此，对于最小值，j=1；对于最大值，j=n。q-q图建立在是X的(j-1/2)/n分位点的一个估计这一事实基础上的，换句话说：

是的近似

现在假定我们已经选择了一个cdf为F的分布作为X的分布的一个可能的表示。如果F是合适的分布簇的一个成员，则对于的图形将近似是一条直线。如果F来自于一个合适的分布簇，并且有一个合适的参数值，那么直线的斜率是1。另一方面，如果假设的分布不合适，则这些点经常会以一种系统的方式从直线偏离。是否拒绝某个假设模型的决定是主观做出的。

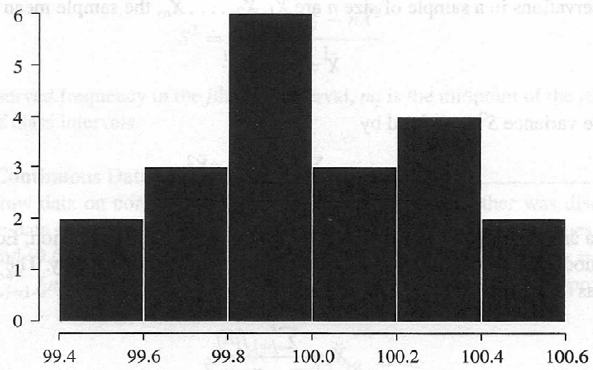
例9.4 **正态q-q图**

一台机器人用于沿装配线在汽车上安装车门。认为安装时间服。从正态分布。机器人能精确地测量安装时间。机器人自动获得如下的20个安装时间的样本，其值以秒为单位：

|  |  |  |  |
| --- | --- | --- | --- |
| 99.79 | 99.56 | 100.17 | 100.33 |
| 100.26 | 100.41 | 99.98 | 99.83 |
| 100.23 | 100.27 | 100.02 | 100.47 |
| 99.55 | 99.62 | 99.65 | 99.82 |
| 99.96 | 99.90 | 100.06 | 99.85 |

样本均值是99.99秒，样本方差是，可以作为正态分布的均值和方差的估计。观察值从小到大的排列如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| j | 值 | j | 值 | j | 值 | j | 值 |
| 1 | 99.55 | 6 | 99.82 | 11 | 99.98 | 16 | 100.26 |
| 2 | 99.56 | 7 | 99.83 | 12 | 100.02 | 17 | 100.27 |
| 3 | 99.62 | 8 | 99.85 | 13 | 100.06 | 18 | 100.33 |
| 4 | 99.65 | 9 | 99.90 | 14 | 100.17 | 19 | 100.41 |
| 5 | 99.79 | 10 | 99.96 | 15 | 100.23 | 20 | 100.47 |

然后对有序观察值做对于)的图形得到q-q图，这里的F是均值为99.99、方差为的正态分布的cdf。所画出的值与数据的直方图一起被表示在图9-9中。该数据样本具有阶梯状的正态分布密度函数。注意，本例难于通过观察直方图来分辨数据是否可以由正态分布很好地表示，但是q-q图相当清晰地给了我们一个直线的整体感觉，并支持了正态分布的假设。

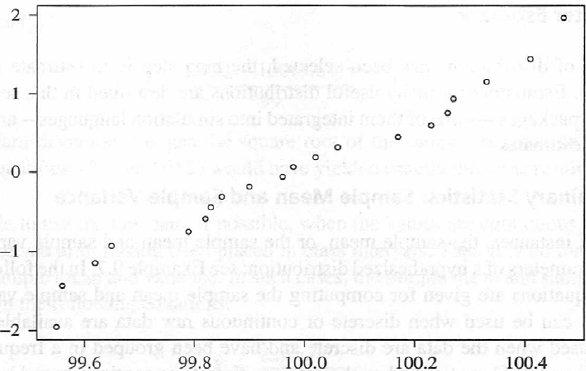


图9-4 安装时间的直方图和q-q图

在评价一个q-q图的直线性时，应该考虑下面的观点：

1. 观察值永远不会正好落在一点。

2. 观察值不是独立的，它们已经被排序。如果一个点在直线的上方，那么下一个点很可能也位于直线的上方。这些点是不大可能关于直线分散。

3. 极大值和极小值的方差比。图中部点的方差要大的多。在极点处可以接受较大的差异。图中部点的直线性比极值点处的直线性要重要多。

现代数据分析软件经常包含产生q-q曲线的工具。尤其是正态分布。q-q曲线也可以用来比较两个数据样本，看看他们是否可以用相同的分布来表示。设是随机变量X的一个样本，是随机变量Z的一个样本，如果两个样本可以被相同的分布很好地表示的话，那么做有序的X值对于有序的Z值的图形，会看出它近似是一条直线(Chamners，Cleveland，and Tukey[1983])。

# 9.3 参数估计

在选定分布簇后，下一步就是估计这个分布的参数。本节将介绍许多有用分布的估计。另外现在许多软件包(其中的一些被集成在仿真语言中)可以用来计算这些估计。

## 9.3.1 统计学原理：样本均值和样本方差

在许多例子中，样本均值被用来估计假设分布的参数。请看例9.4。在下面的段落中给出了三组等式，用来计算样本均值和样本方差。能够得到离散的或者连续的数据时，采用式(9-1)和式(9-2)。当数据是离散的。并且这些数据被归组到一个频度分布中时，采用式(9-3)和式(9-4)。当数据是离散的或连续的，并且这些数据被放入分组区间中时， 采用式(9-5)和式(9-6)。式(9-5)和式(9-6)是近似并且应该只用在得不到原始数据的情况下。

如果一个容量为n的样本的观察值是，则样本均值由下式定义：



而样本方差(S2)由下式定义：

(9-2)

如果数据是离散的，并且被归属到一个频度分布中。那么可以修改式(9-1)和式(9-2)以提供更好的计算效率。样本均值可以由下式计算：

(9-3)

而样本方差可以由下式计算：

(9-4)

这里，k是不同的X值的个数，是X的值的观察频度。

例9.5 **分组数据**

分析表9-1中的数据可以得到n=100，f1=12，X1=0，f2=10，X2=1，…，=364和。根据式(9-3)有

而根据式(9-4)有

样本标准差S正好是样本方差的平方根。在这种情况下，S=。式(9-1)和式(9-2)能精确地提供相同的结果。

如果值是连续的，在可能的情况下，使用原始数据是更好的选择。然而有时数据是在被放入分组区间之后才获得的，那么就不再有可能获得精确的样本均值和样本方差了。在这种情况下，样本均值和样本方差可由下面的等式近似获得：

(9-5)

和

(9-6)

这里，fj是第j个分组区间的观察频度，mj是第j个分组区间的中点，c是分组区间的个数。

例9.6 **分组区间中的连续数据**

设立9.3中的元件寿命的原始取被扔掉了或是被弄丢了，然而表9-2所示的数据依然可以使用。为了求和S2的近似值，可以使用式(9-5)和式(9-6)。由表9-2得到下列数值：f1=23，m1=1.5，f2=10，m2=4.5，…，=614和。当n=50时，由式(9-5)求得的近似值是：

由式(9-6)求得S2的近似值是：

且

24.614

对例9.3中的原始数据使用式(9-1)和式(9-2)得到和S=24.953。因此，若原始数据被扔掉或者丢失时，会导致不精确的结果。