直方图

有一种简单的方法能用来了解产生数据 $x_1, \ldots, x_n$ 的概率密度, 那就是直方图。

对于覆盖数据 $x_1, \ldots, x_n$ 取值范围的分区 $[a_0, a_n]$ 其中有分点 $a_0 < a_1 < \cdots < a_m$，直方图是在每个区间 $(a_{j-1}, a_j]$ (这里前开后闭前闭后开都可以, 只要不要漏掉或者重复任何一个数据点, 所以显然前开后开前闭后闭是不合适的划分)上**取值为该区间内样本点数 x_i 除以区间长度的函数**。如果区间 $(a_{j-1}, a_j]$ 的长度相同，直方图有时也可以不除以区间长度来定义。在这种情况下，直方图的柱高等于各区间内的观测次数。

为了了解给定数据(背后的概率分布的)的概率密度，使用直方图是相当有用的。可以通过将直方图每一条的高度缩放为 $1/n$（其中 n 是数据点的总数）来实现。直方图下的面积与概率密度的面积相等为 1。在 $x \in (a_{j-1}, a_j]$ 中，缩放后的直方图为

$$h_n(x) = \frac{\#\left(1 \le i \le n : x_i \in (a_{j-1}, a_j]\right)}{n\left(a_j - a_{j-1}\right)} = \frac{1}{n\left(a_j - a_{j-1}\right)} \sum_{i=1}^{n} 1_{a_{j-1} < x_i \le a_j}$$

其中指示函数 $1_{a_{j-1} < x_i \le a_j}$ 对 $a_{j-1} < x_i \le a_j$ 取 1，在其他地方取 0。另一种写法是 $1_{(a_{j-1}, a_j]}(x_i)$。

如果分区 $a_0 < a_1 < \cdots < a_m$ 选择得当且样本数量 n 足够大，缩放后的直方图可以很好地反映产生数据 $x_1, \ldots, x_n$ 的分布的密度。这一点是很好理解的，我们将 $x_1, \ldots, x_n$ 视为具有密度 f 的随机变量的实现，并在任意点 x（这个点上显然要满足 $f(x) > 0$) 算缩放后的直方图 h_n 的期望值。假设对于某个 $1 < j \le m$，我们有 $a_{j-1} < x \le a_j$，那么这个期望值等于

$$\begin{aligned} \mathrm{E}h_n(x) &= \mathrm{E}\frac{1}{n\left(a_j - a_{j-1}\right)} \sum_{i=1}^{n} 1_{a_{j-1} < X_i \le a_j} = \frac{1}{a_j - a_{j-1}} \mathrm{E}1_{a_{j-1} < X_1 \le a_j} \\ &= \frac{1}{a_j - a_{j-1}} \mathrm{P}\left(a_{j-1} < X_1 \le a_j\right) = \frac{\int_{a_{j-1}}^{a_j} f(s)ds}{a_j - a_{j-1}} \end{aligned}$$

这里第二个等号成立的原因是每一个 X_i 都是独立同分布的所以

$$E\sum_{i=1}^{n} 1_{a_{j-1} < X_i \le a_j} = nE1_{a_{j-1} < X_1 \le a_j}$$

这对于 $i.i.d.$ 的数据来说是非常重要和常见的一种转化方式.

第三个等号的成立原因可以从基础概率论中理解, 如果求某个集合的示性函数的期望, 那么就等于求数据点落到这个集合中的概率.

如果 f 在区间 $(a_{j-1}, a_j]$ 上变化不大，那么最后一个表达式大致等于 f 在该区间的值, 如果变化太大自然会有一定偏差, 当然这也说明了把区间取细一点的用处所在。这个计算表明 $h_n(x)$ 的期望值大致等于 $f(x)$。此外，根据大数定律，我们知道 $h_n(x)$ 的值会以概率收敛到其期望值。

因此，直方图能够提供样本所对应分布的大概的直观的印象。然而，我们只有在样本足够大（例如 $n = 100$ 或更好 $n = 500$）且区间选择得当时，才能获得良好的直观图像。

这里区间的选择是一个主观问题。如果区间太短，则直方图往往过于尖锐，难以观察到真实密度的特征；如果区间太长，则细节会丢失，难以根据直方图推断出真实密度。因此，我们不能指望

从直方图中获得更多的信息。其他更复杂的技术可以提供更好的结果。

例 身高

下图显示了 44 名男性（左）和 67 名女性（右）的身高（以 cm 为单位）的直方图。这里直方图已经经过缩放，使得其下面积等于 1。两幅图也对比显示了正态分布的密度。这些正态分布的期望和方差是根据相应数据的样本均值和样本方差算出的。

不过根据直方图的形状就判断数据是否来自正态分布是存疑的。左侧的直方图明显不太对称, 这可能是由于观察数量较少所致. 从中我们也能知道进一步的研究个数据搜集是必要的。

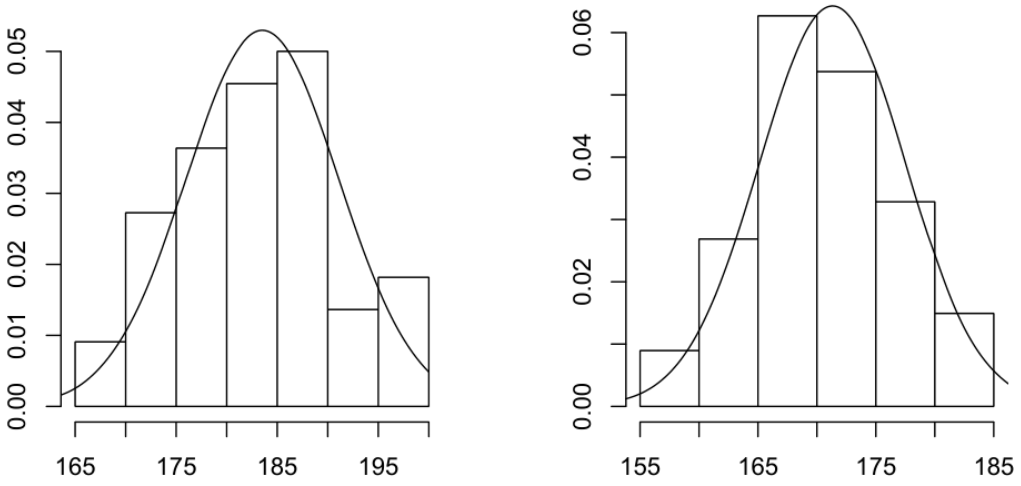

图: 名男性（左）和 67 名女性（右）的身高（单位：cm）直方图，以及期望等于样本均值和方差等于样本方差的正态分布密度。

例 正态分布

下图显示了标准正态分布的密度，以及基于 $30, 30, 100$ 和 100 次观测的直方图的四次实现，分区由统计软件包 R 选择。左上角和右下角的图很明显是偏离对称的。但是我们又清楚的知道由于数

据来自正态分布，所以这仅仅是由于随机变化引起的。

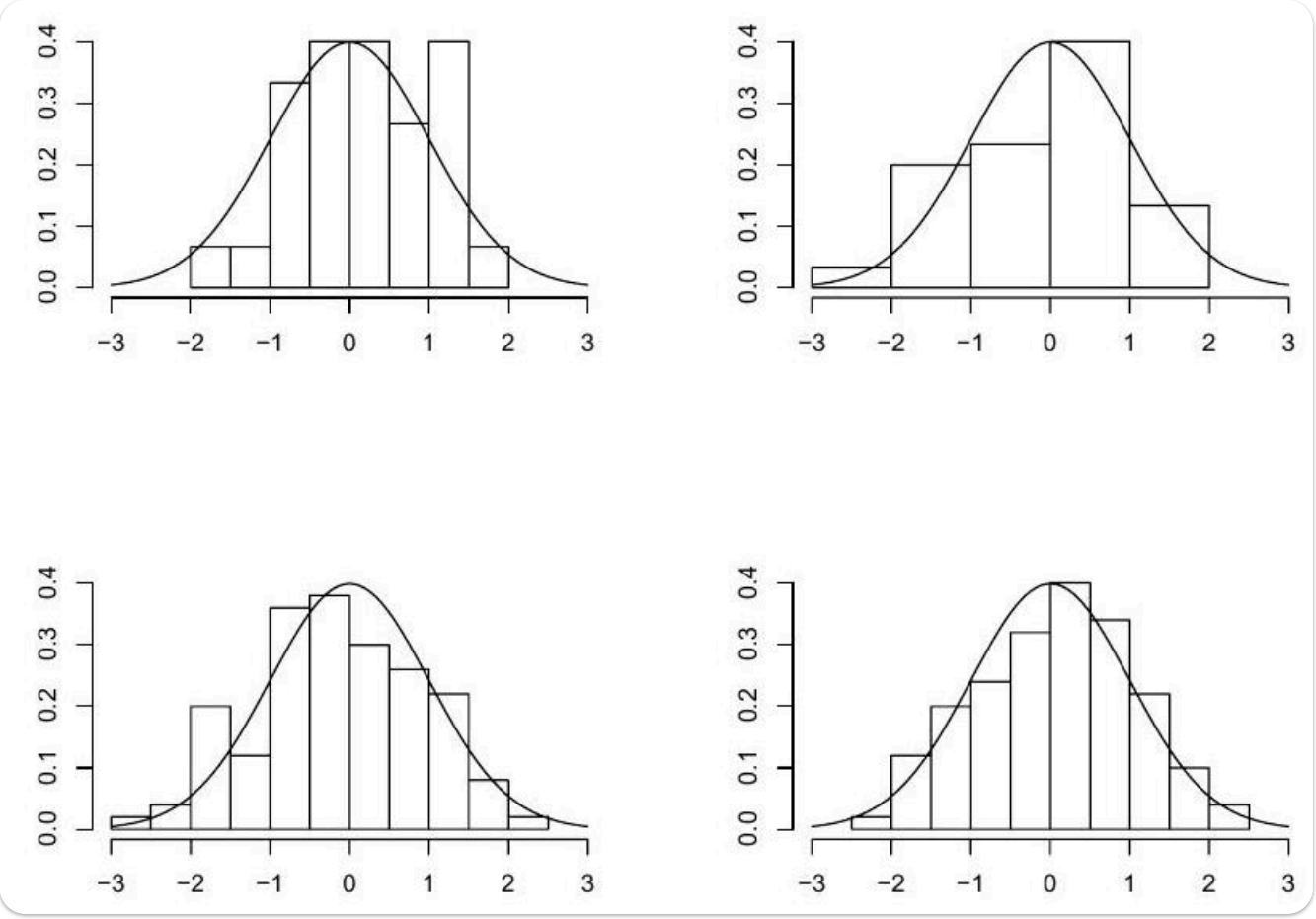

图 . 基于标准正态密度的 30（上排）和 100（下排）观测样本的直方图，展示了真实的密度。

2.2.2 箱线图

箱线图是一种数据的图形表示方法，它可以帮助我们了解数据的位置、离散程度、观测值中的潜在离群点以及生成观测值的分布的对称性。在箱线图中，观测值沿着纵轴排列。"箱子"的底部绘制在数据的下四分位数水平线上，顶部绘制在数据的上四分位数水平线上。数据的下（或上）四分位数是指其中四分之一的观测值小于（或大于）该值。箱子的宽度是任意的，箱子中间有一条水平线，表示数据的中位数。中位数是排序后的数据中的中间值。在箱子的顶部和底部绘制胡须。顶部的胡须连接到位于上四分位数的 1.5 倍四分位距内的最大观测值。四分位距是下四分位数和上四分位数之间的距离，即箱子的高度。底部的胡须则类似地绘制。超出胡须范围的观测值会被单独标出，例如用星号、小圆圈或短横线表示。

图 2.3 显示了来自指数分布（参数为 1）、标准正态分布和标准柯西分布的样本的箱线图。指数分布和柯西分布的样本中都有离群点（用胡须外的小圆圈表示）。中间的标准正态分布的箱线图

显示数据相对于中位数相当对称，且不包含离群点。

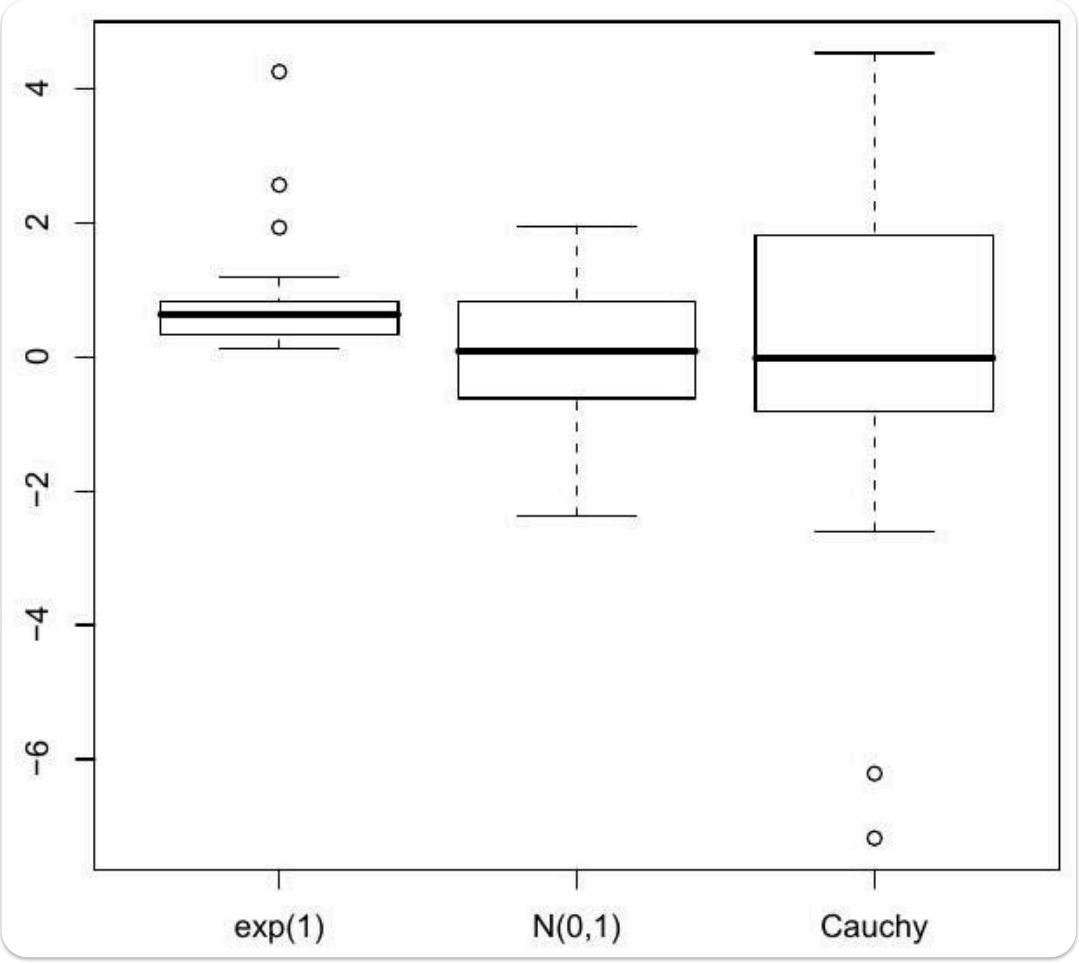

图 2.3. 来自标准指数分布（左）、标准正态分布（中）和标准柯西分布（右）的样本（样本量为 20）的箱线图。