Q-Q图

绘制 QQ 图是一种常用的图形方法，以帮助根据样本 $x_1, \ldots, x_n$ 找到合适的分布类（即所谓的"位置-尺度族"）。

QQ-图

对于数据集 $x_1, \ldots, x_n$，针对分布函数 F 的 QQ-图是以下点的图：

$$\left\{\left(F^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right) : i = 1, \ldots, n\right\}$$

QQ-图提供了一种验证样本是否来自与 F 相关的某个位置-尺度族的图形方法。Q 代表 "quantile"（分位数）。

例 正态分布

下图显示了使用随机数生成器从 $N\left(2, 4^2\right)$ 分布模拟的六个样本的 QQ-图，并将它们与 $N(0,1)$ 分布进行比较。因为两个正态分布属于相同的位置-尺度族，我们可以预期这些点大致排列在一条直线上。顶部和底部的图分别表示 10 和 50 个观测值的数据集。QQ-图中的点并不完全在直线上，而是稍微围绕直线波动。在小样本中，这种波动比大样本更大。

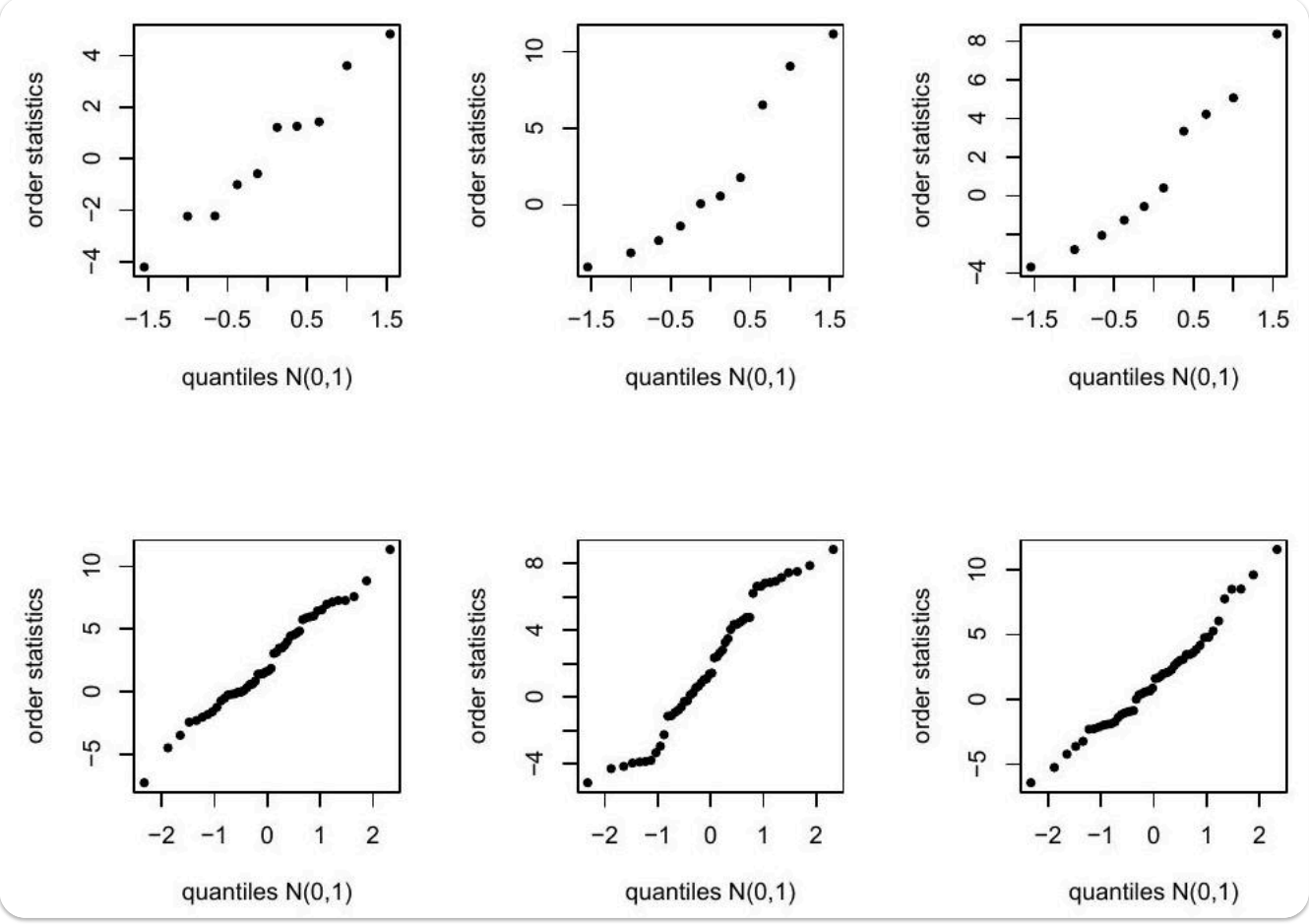

图: 六个从 $N\left(2,4^2\right)$ 分布中抽取的 10（上排）或 50（下排）个数据点的 QQ-图，与 $N(0,1)$ 分布进行比较。

如果样本 $x_1,\ldots,x_n$ 的 QQ-图相对于 F 的分位数大致显示出直线 $y=x$，那么这表明数据可能来自分布 F。与直线 $y=x$ 的偏差可以指示数据的真实分布与 F 的偏差。最简单的"意外情况"是，图中确实显示出一条直线，但不是 $y=x$ 直线。这意味着数据来自与 F 相关的位置-尺度族中的另一个成员，正如下面这个例子所说的。此时，可以通过拟合直线 $y=a+bx$ 到 QQ-图来粗略估计 a 和 b 的值。在之后章节我们将看到其他估计参数的方法。

曲线的偏差更难评估，主要反映了数据分布相对于 F 的尾部权重。为了说明线性偏差的不同类型，下图显示了“真实”分位数函数的一些 QQ-图。这些图是各种分布函数 F 和 G 的点 $\left\{\left(F^{-1}(\alpha),G^{-1}(\alpha)\right):\alpha\in(0,1)\right\}$ 的绘图。

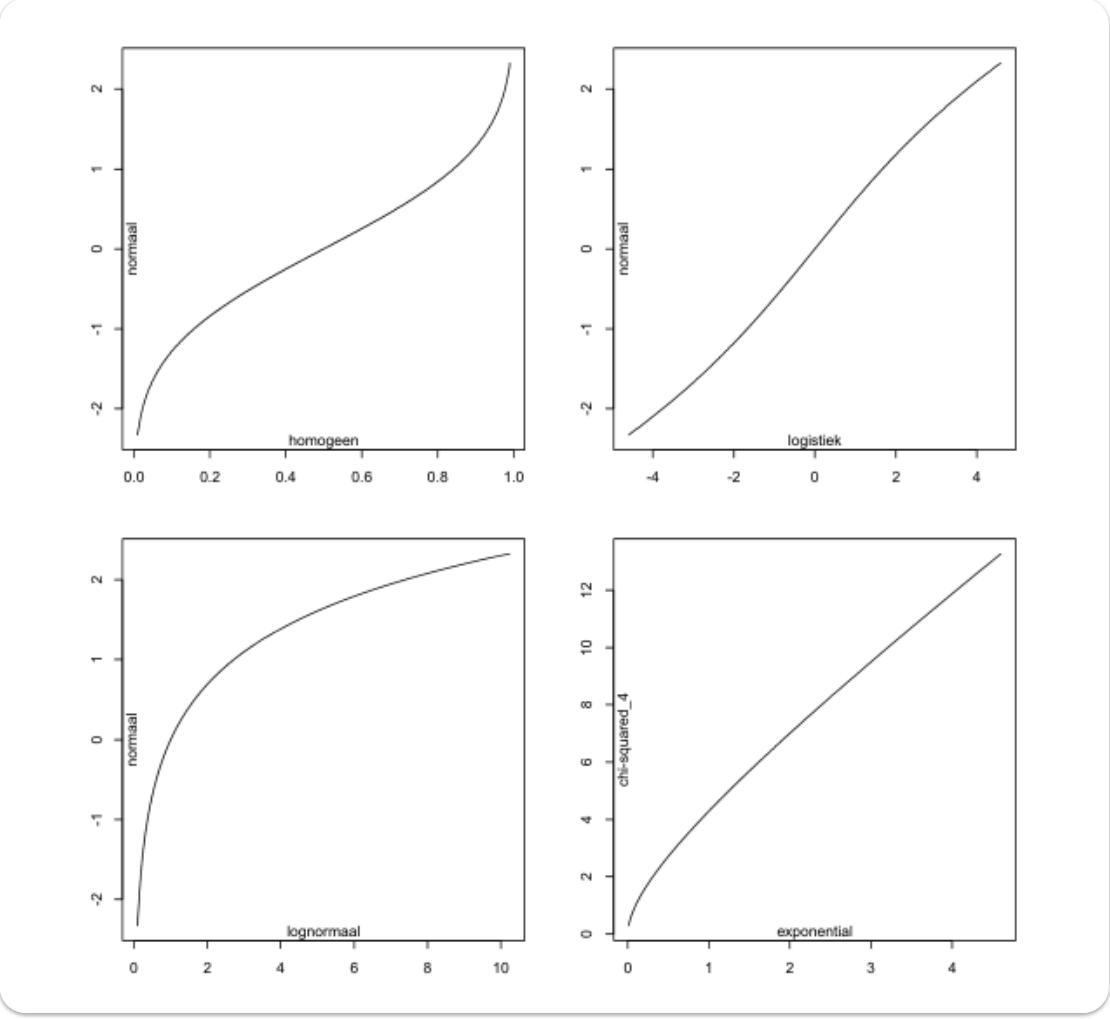

图: 分位函数对照.

例 身高

根据之前直方图部分数据的形状，对于身高是否来自正态分布存在一些怀疑。为了进一步研究，下图显示了绘制的 QQ-图，分别将男性（左侧）和女性（右侧）的身高与标准正态分布进行比较。为了研究这些点是否位于直线上，在两个图中都绘制了适当的直线 $y=a+bx$。对于男性，这条线是 $y=183.5+7.5x$，而对于女性，则是 $y=171.3+6.2x$。这些线通过使用最大似然估计法估计 a 和 b^2 来确定（参见之后章节）。由于数据较好地遵循这些线，我们可以得出结论，与标准正态分布相关的位置-尺度族是这两个数据集的良好拟合。由于该族仅包含正态分布，这支持了数据来自正态分布的假设。

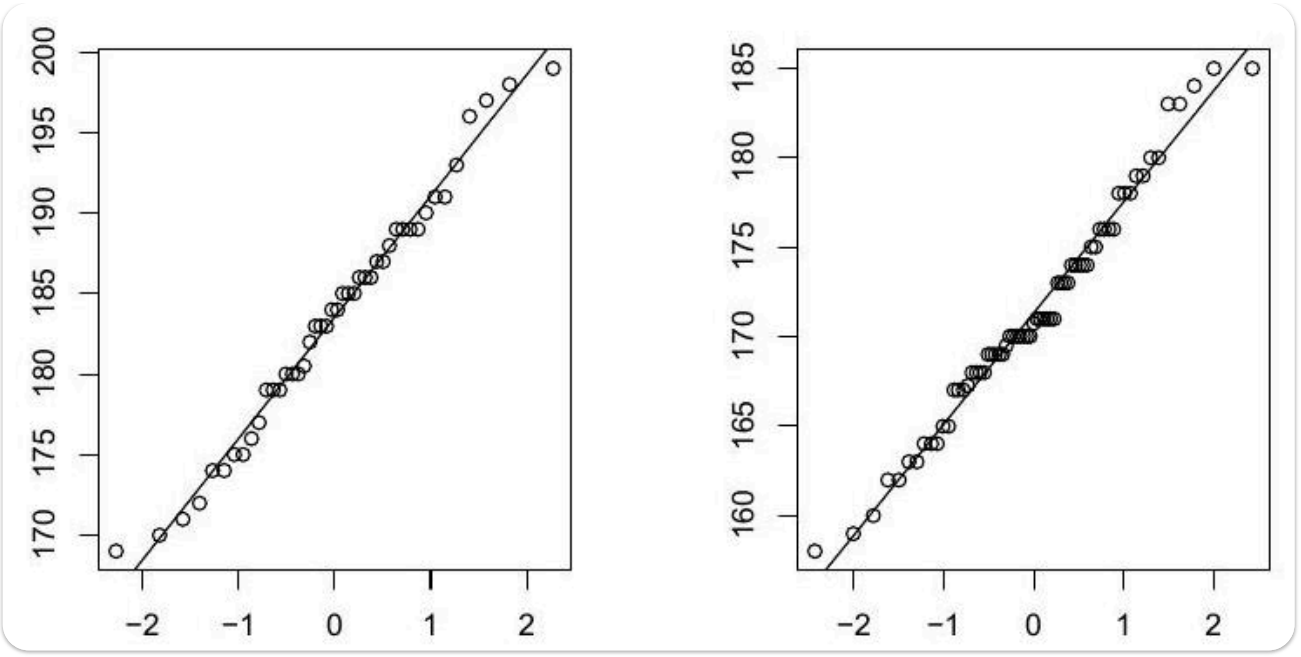

图:将 44 名男性（左图）和 67 名女性（右图）的身高绘制在标准正态分布的分位数 QQ-图中。