集中趋势

均值

一般的均值计算公式如下：

$\mu = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

一般均值中蕴含了一个潜在条件，每个变量的权重相同，如果权重不同，修改为如下形式。

$\mu = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} w_i x_i$

其中

$\sum_{i=1}^{N} w_i = 1$

即常说的平均数，也叫数学期望，均值容易受极值的影响，当数据集中出现极值时，所得到的的均值结果将会出现较大的偏差。

中位数

其计算方法是将所有数据按从小到大的顺序排列，如果有基数个数据值，则位于中央的数据值就是中位数，如果有偶数个数据值，则中位数是中间两个数值的平均是。

除中位数外，还有四分位数，百分位数据等，不过和此处的中位数不同，四分位数和百分位数主要用于度量分布形状。

众数

数据中出现次数最多的数字，即频数最大的数值。众数可能不止一个，众数除能用于数值型数据，还可用于非数值型数据，不受极值影响。

离散程度

极差

极差是极大值与极小值之间的差

$\text{极差} = \text{极大值} - \text{极小值}$

极差是描述数据分散程度的量，极差描述了数据的范围，但无法描述其分布状态。一般数据统计而言为数据的极大值即为最大值，极小值即为最小值，但是从理论上而言，两者有比较直接的区别，极值是局部值，最值是全局值。

方差和标准差

方差(σ^2)是每个数据与全体数据平均数的差的平方的平均数，标准差(σ)是方差开方。方差和标准差描述数据波动离散程度和波动性。

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

$\sigma = \sqrt{\sigma^2}$

不同量纲下的数据方差和标准差有显著差异，若对比不同数据标准差，需要使用Z标准化后的数据，z标准化为

$a = \frac{x_i - \mu}{\sigma}$

四分位数极差

四分位数本身用来度量数据形态，不过其极差可以用来反应偏态分布的数据离散程度。四分位数计算方法如下：

数据从小到大排列并分成四等份，处于三个分割点位置的数值，即为四分位数，四分位数分为上四分位数（数据从小到大排列排在第75%的数字，即最大的四分位数）、下四分位数（数据从小到大排列排在第25%位置的数字，即最小的四分位数）、中间的四分位数即为中位数。四分位数可以很容易地识别异常值。箱线图就是根据四分位数做的图。

变异系数

变异系数是一种不受单位影响的表示数据离散程度的指标，比较适合在以下两种情况下比较数据差异：

- 各组数据单位不完全相同
- 各组数据的均值相差悬殊

变异系数的表示形式为

$cv = \frac{\sigma}{\mu} * 100\%$

变异系数在数据呈正态分布是效果较好，当数据呈偏态分布时，则极差和四分位数极差代表性更好。

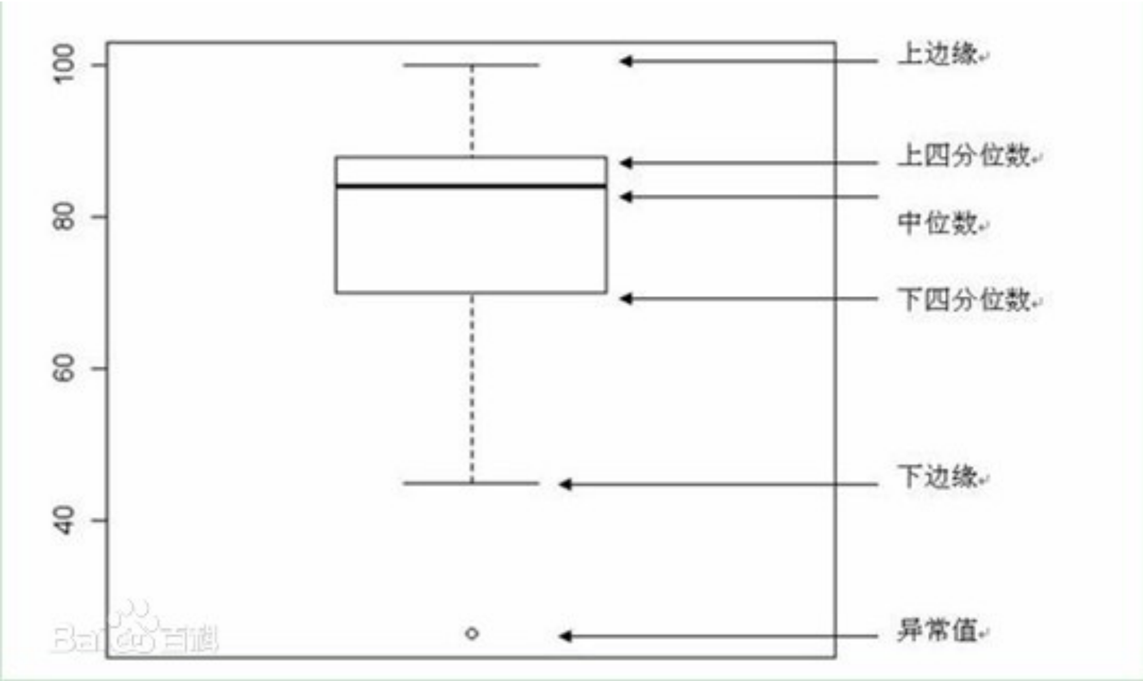
分布形态

百分位数

前面已经涉及到中位数、四分位数，而百分位数作为一种位置指标，同样可以来度量分布形态，计算方法与前面的四分位数计算方法类似。

箱线图

箱线图来源与四分位数，或者可以理解为来源与百分位数，箱子的底部为下四分位数，顶部为上四分位数，盒子高度（上四分位数和下四分位数之间的距离）记为IQR，箱上下的线不超过1.5个IQR，超过部分为异常值。箱线图示例如下。



直方图

直方图作为一种集合图形，可以处理看似无序的数据，反应数据分布情况。直方图是以组距为底边，频数或百分比为高度的一系列链接起来的直方矩形图。每个矩形图代表一组数据，矩形的高度代表落在这一组中的数据的频数或者百分比。

峰度

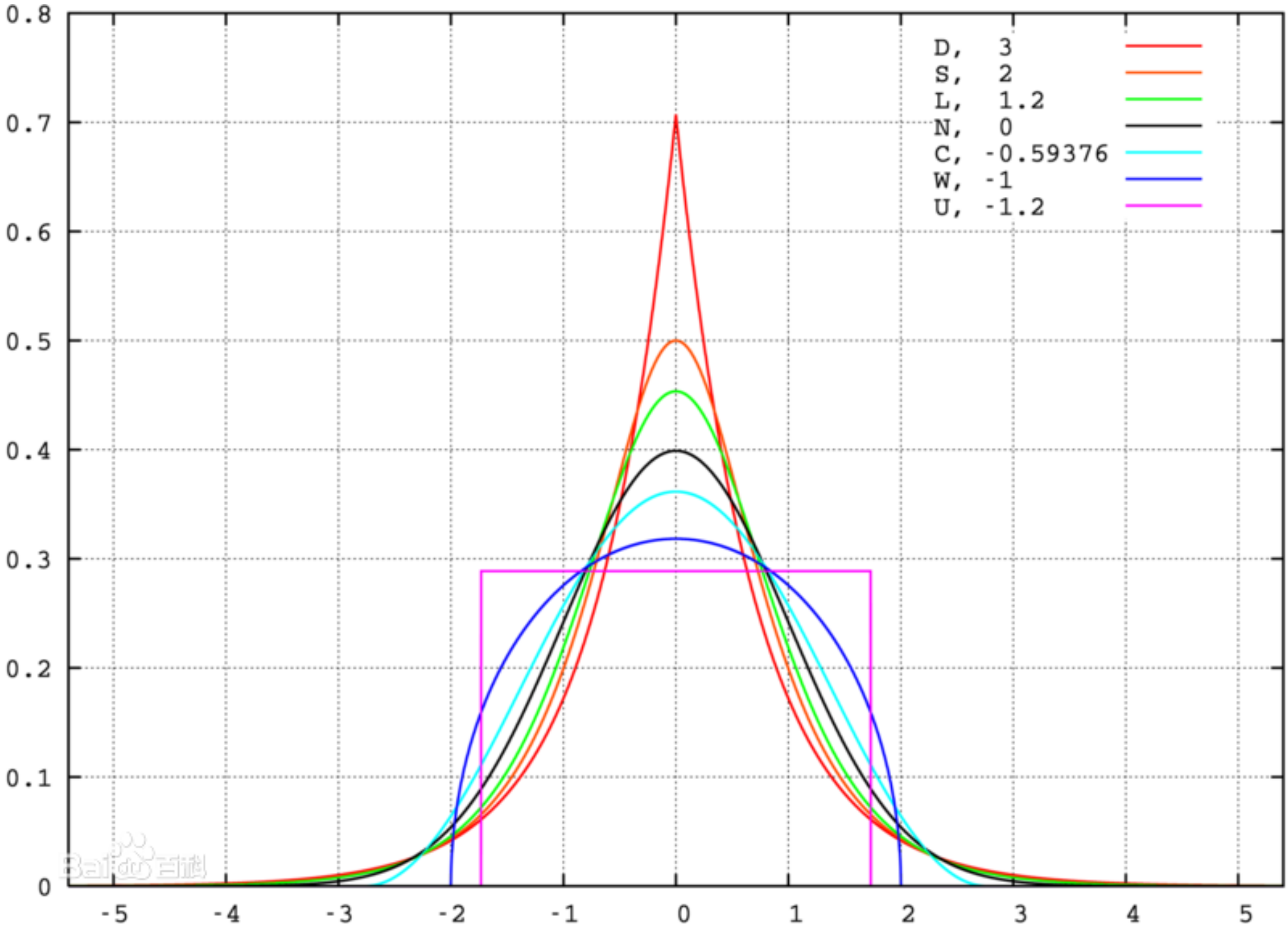
描述正态分布中曲线峰顶尖哨程度的指标。峰度系数>0，则两侧极端数据较少，比正太分布更高更瘦，呈尖哨峰分布；峰度系数<0，则两侧极端数据较多，比正太分布更矮更胖，呈平阔峰分布。

峰度计算需要涉及4阶中心矩和2阶中心矩（方差）

$$Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} (\frac{x_i - \overline{x}}{\sigma})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

也可写为

$$Kurtosis = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2)^2} - 3$$



偏度

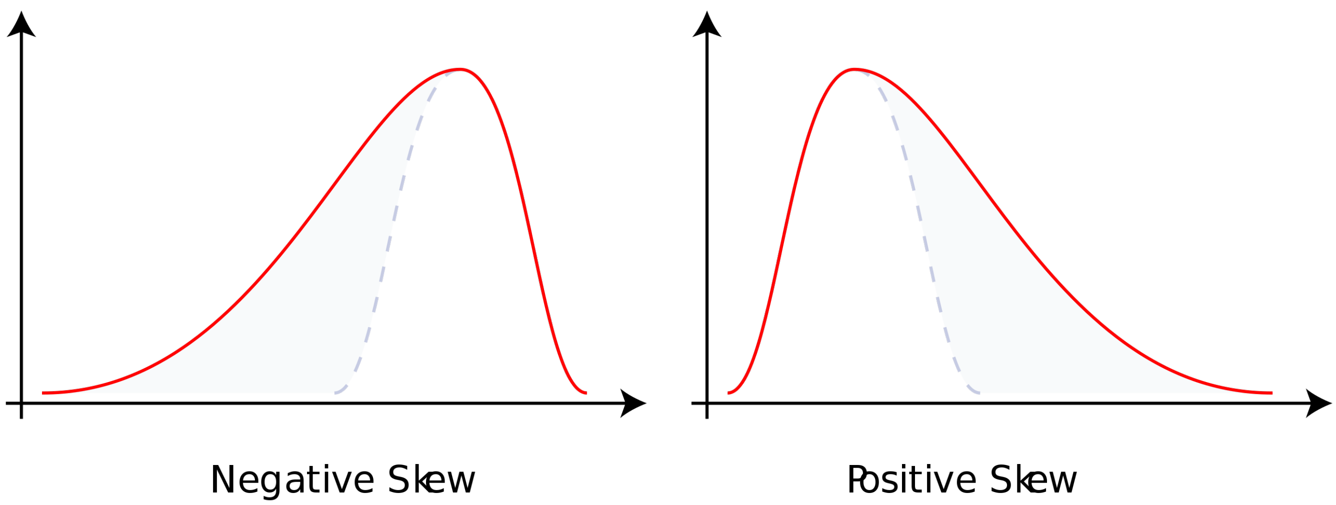
以正态分布为标准描述数据对称性的指标。偏度系数=0，则分布对称；偏度系数>0，则频数分布的高峰向左偏移，长尾向右延伸，呈正偏态分布；偏度系数<0，则频数分布的高峰向右偏移，长尾向左延伸，呈负偏态分布。

偏度是三阶标准矩（m_3表示3阶中心矩），定义为

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma}\right)^3$$

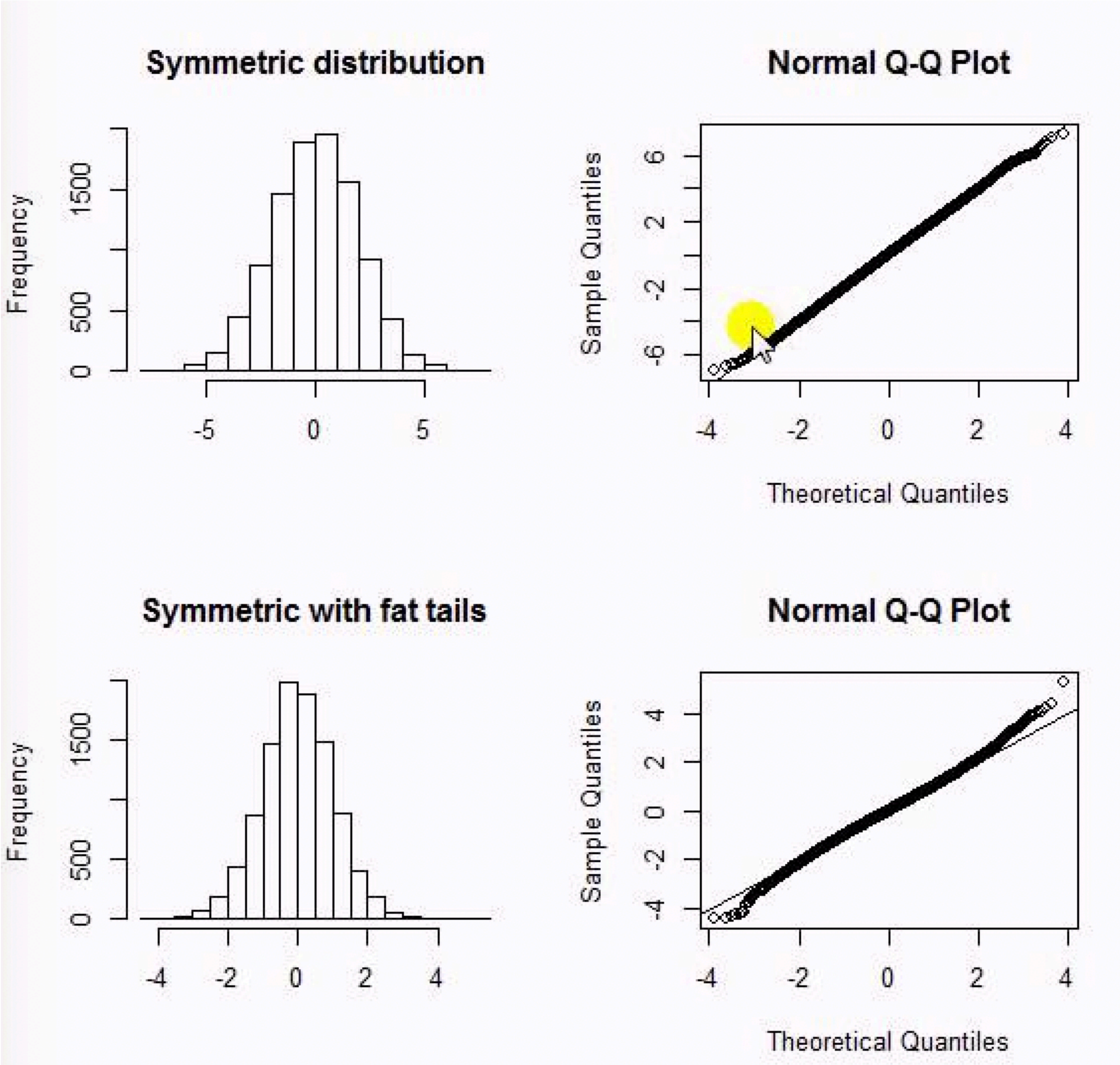
也可写为

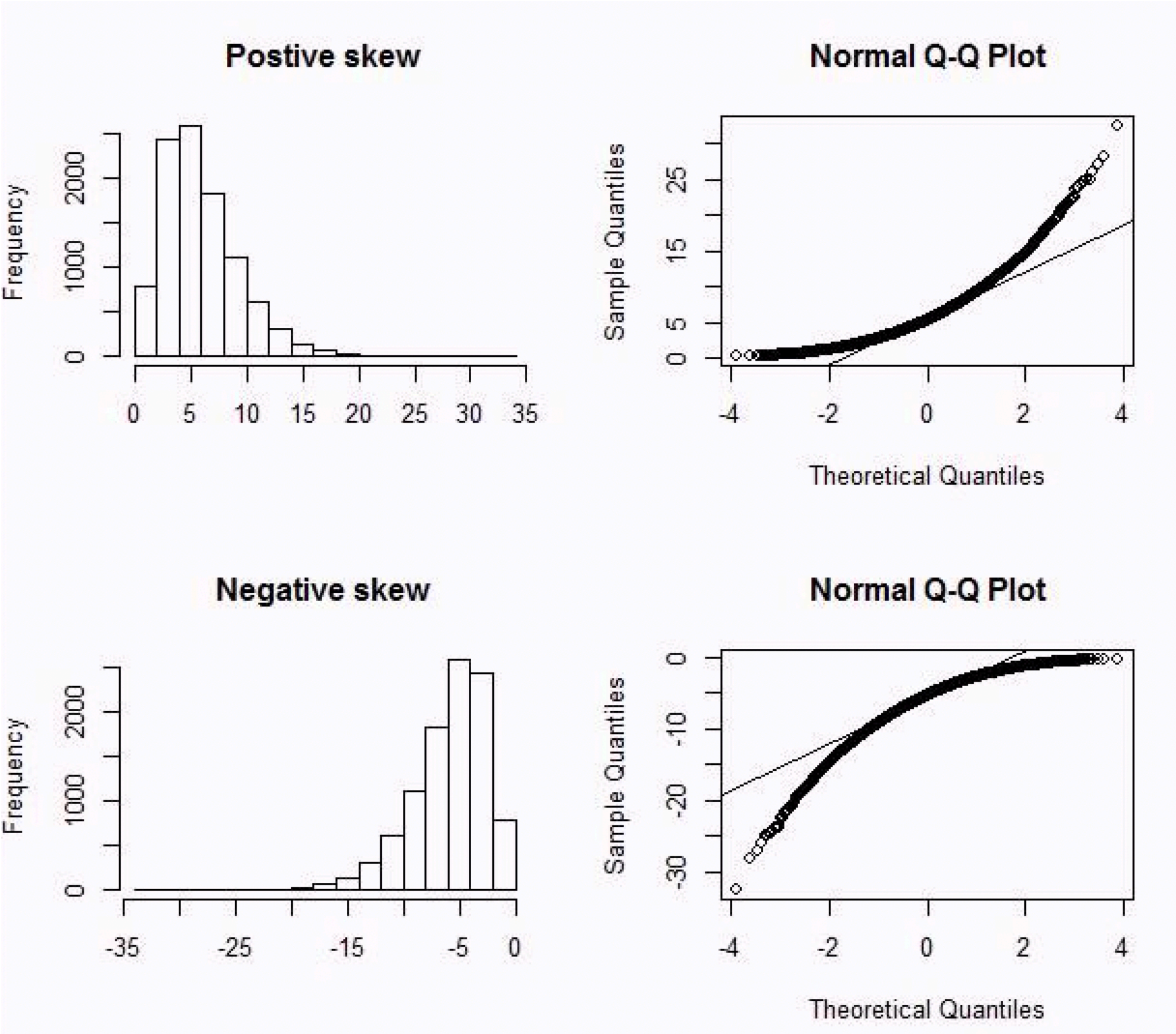
$$Skewness = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{m_3}{\sigma^3} = \frac{m_3}{m_2^{\frac{3}{2}}}$$



正态概率图

正态概率图用以检查一组数据是否服从正分布，是实际数据与正态分布分位数之间函数关系的散点图。如果一组数据服从或接近正态分布，其正态概率图中众多散点将是一条直线。如





第一个与第二个都是正态分布，第三个为正偏态分布，第四个为负偏态分布。左右对比也能看出比较明显的差异。

完～