第六章 大数据分析-统计学方法

李晚月，章舒婕，安航正

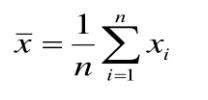
1. **数据描述性分析**（李晚月，章舒婕）

用统计学的方法，描述数据的统计特征量，分析数据的分布特性。

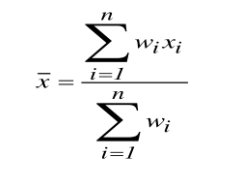
主要包括：数据的集中趋势分析、数据的离散趋势分析、数据的频度分布等。

（1）数据的集中趋势度量

数据的中心趋势度量



—>算数均值：



—>加权平均：又称加权算术平均：

* 集合中每个值与一个权值相关联

—>截断均值：可以抵消少数极端值的影响：

* 去掉最高和最低值的影响

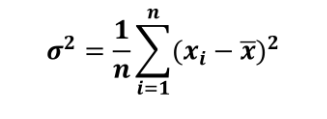
—>中位数：按要求排序后，奇数中最中间那个数，偶数中最中间两个数的平均数

相对于均值，平均数有更好的抗扰性。

—>众数：在一组数据中，出现次数最多的数据。

（2）数据的离散趋势度量

—>方差：方差值越大说明数据项波动越大。



—>极差：最大值与最小值之差。

—>四分位数：也称四分位点，把数据按大小顺序排列，并分成四等分，处于三个分割点位置的就是四分位数。

第一“四分位数”(Q1):较小四分位数

第二“四分位数”(Q2):中位数

第三“四分位数”(Q3):较大四分位数

Q1与Q3的差距又称为：四分位距

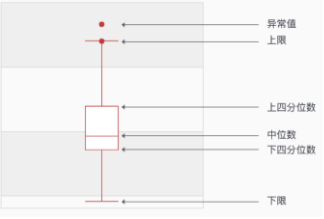
—>五数概括：数据集分布形状的完整概括：中位数，四分位数Q1和Q3，最小最大观测值

通常用箱型图（盒图）进行可视化表示。

盒图：数据分布用一个盒子来表示

盒子两端是第一和第三“四分位数”

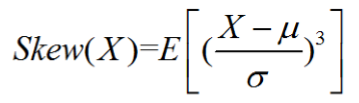
中位数在盒子用一根线表示出来

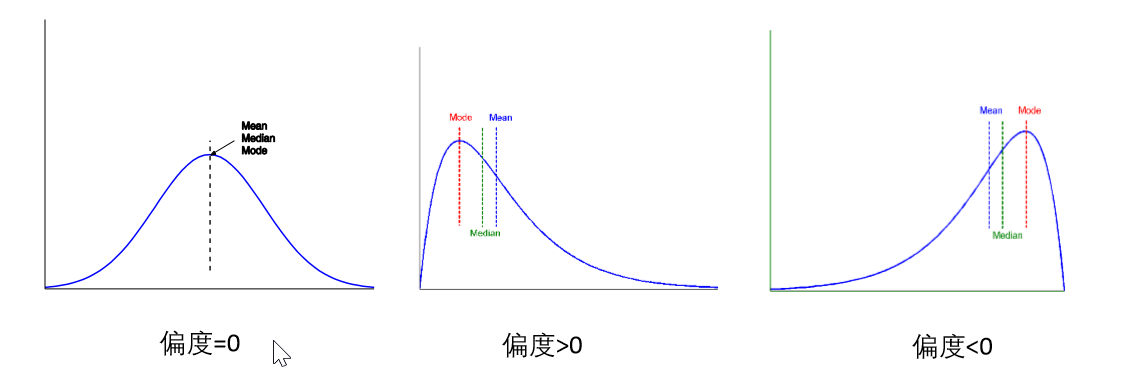
外边界：盒子外面延伸到最大值和最小值的两条线，也称“胡须”

（3）数据的偏态特性

偏度(Skewness)：也称为偏态、偏态系数，是统计数据分布偏斜方 向和程度的度量，是统计数据分布非对称程度的数字特征。

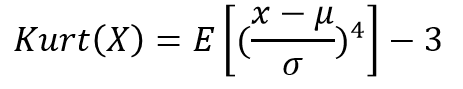
偏度定义三阶标准中心矩:

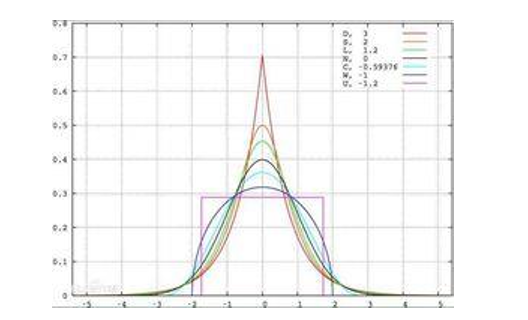




（4）数据的峰度特性

峰度(Kurtosis)：峰度系数是用来反映频数分布曲线顶端 尖峭或扁平程度的指标。通过对峰度系数的测量，我们能够 判定数据分布相对于正态分布而言是更陡峭还是平缓。峰度定义为四阶标准中心矩：



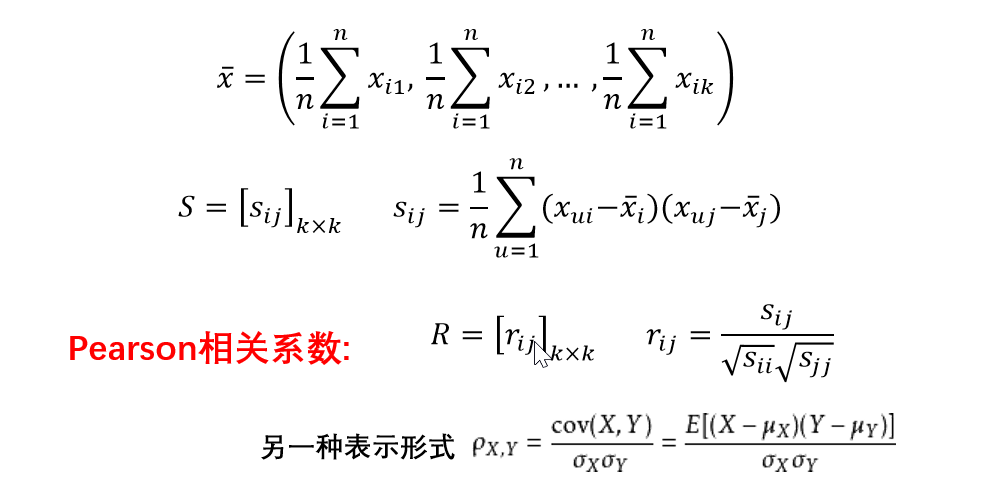


多维数据的相关性

有许多数据实际上包含了多个维度，如微博数据包含：发布时间、点赞数和转发量。想要分析两个维度数据之间的 关系，可以用协方差和Pearson相关系数。

（5）数据的相关性分析

一个k维的数据可以表示为 𝒙 = (𝒙𝟏, 𝒙𝟐,…,𝒙𝒌)，多维 数据的均值和协方差矩阵(Covariance Matrix)为：



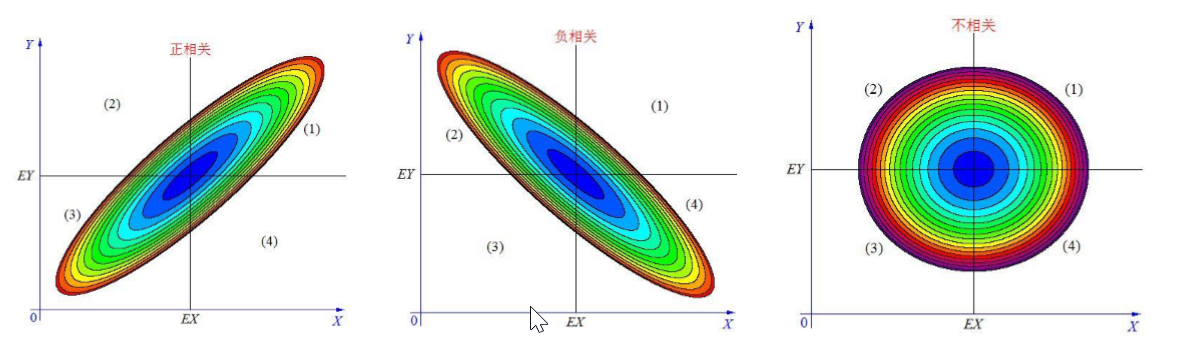
（6）多维数据的相关性

两个随机变量 X 与 Y 之间相互关系，大致有下列3种情况：

当X 与Y 正相关时，它们的分布大部分在区域（1）和（3）中，有cov(X, Y)>0。

当 X与 Y负相关时，它们的分布大部分在区域（2）和（4）中，有cov(X, Y)<0 。

当 X与 Y不相关时，它们在区域（1）和（3）中的分布，与在区域（2）和（4）中的分 布几乎一样多，有cov(X, Y)=0 。



**2. 回归分析：（安航正）**

（1）.在的数据分析中，回归分析是一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系，这种技术通常用于预测分析，时间序列模型以及发现变量之间的因果关系。

（2）.英国著名统计学家佛朗西斯.高尔顿是最先应用统计方法研究两个变量之间关系问题的人，“回归”一词就是由他引入的

（3）.一元线性回归：

线性回归使用最佳的拟合直线（也就是回归线），建立因变果（Y）和一个或多个自变量（X）之间的联系。用一个等式来表示它

即y=B0+B1x+m。其中B0表示截距，B1表示直线的斜率，m是误差项

（4）.线性回归：

回归模型是描述因变量如何以来自变量和随机误差项的方程，一元线性回归模型只涉及一个自变量

🡪最小二乘估计

🡪最小二乘法

（5）.我们在研究两个变量(x, y)之间的相互关系时，通常可以得到一系列成对的数据(x1, y1、x2,y2... xm , ym)；将这些数据描绘在x -y直角座标系中(如图1), 若发现这些点在一条直线附近，可以令这条直线方程如(式1-1)。

Y计= a0 + a1 X

　　　　　　　　　　　　　　　　 (式1-1)

其中：a0、a1 是任意实数

（6）.为建立这直线方程就要确定a0和a1，应用《最小二乘法原理》，将实测值Yi与利用(式1-1)计算值(Y计=a0+a1X)的离差(Yi-Y计)的平方和〔∑(Yi -Y计)2〕最小做为“优化判据”。

令: φ = ∑(Yi - Y计)2

(式1-2)

把(式1-1)代入(式1-2)中得: φ = ∑(Yi - a0 - a1 Xi)2

(式1-3)

当∑(Yi-Y计)平方最小时，可用函数 φ 对a0、a1求偏导数，令这两个偏导数等于零。

亦即：

m a0 + (∑Xi ) a1 = ∑Yi

(式1-6)

(∑Xi ) a0 + (∑Xi2 ) a1 = ∑(Xi,Yi)

(式1-7)

得到的两个关于a0、 a1为未知数的两个方程组，解这两个方程组得出：

a0 = ∑Yi / m - a1 \* ∑Xi / m

(式1-8)

a1 = [m∑(XiYi) - ∑Xi \* ∑Yi] / [m∑(Xi^2) - (∑Xi)^2]

　 (式1-9)

其中m为样本容量

这时把a0、a1代入(式1-1)中, 此时的(式1-1)就是我们回归的元线性方程即：数学模型。

（7）.在回归过程中，回归的关联式是不可能全部通过每个回归数据点(x1, y1、 x2, y2...xm,ym),为了判断关联式的好坏,可借助相关系数“R”，统计量“F”，剩余标准偏差“S”进行判断；“R”越趋近于

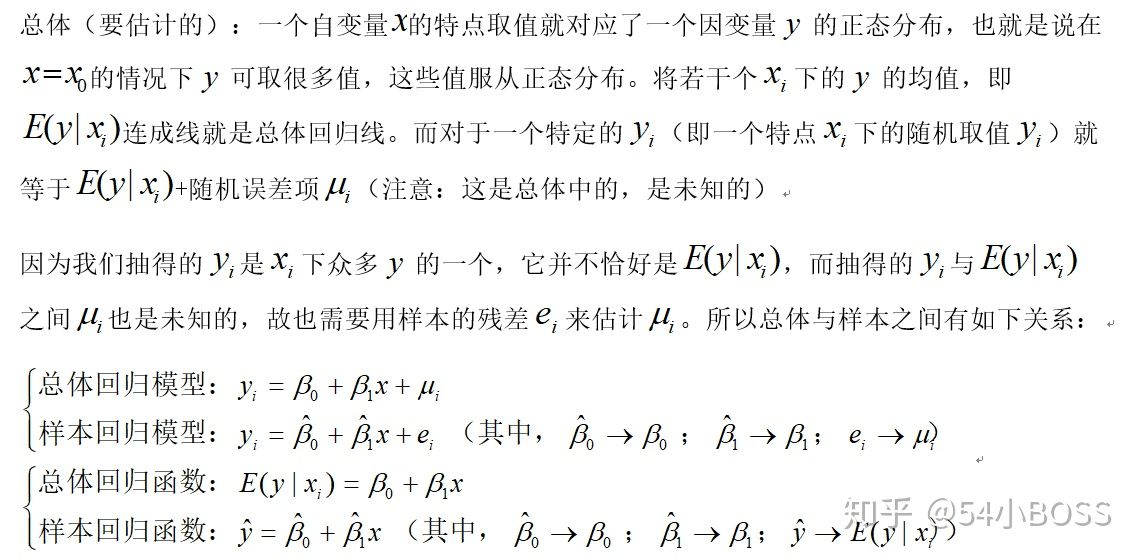
1 越好；“F”的绝对值越大越好；“S”越趋近于 0 越好。 其中：

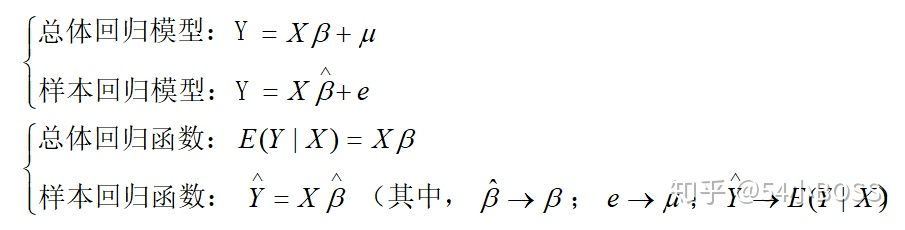
R = [∑XiYi - m (∑Xi / m)(∑Yi / m)]/ SQR{[∑Xi2 -m (∑Xi / m)2][∑Yi2 - m (∑Yi / m)2]}

　　　　　　　 (式1-10)

普通的最小二乘法，使得全部观察值的残差平方和最小

（8）.回归：





（9）.非线性回归：

1. 适配曲线问题：

（1）确定变量间的依存关系，根据实际资料做散点图

（2）确定回归模型中的未知参数

2，常见的函数：

双曲线，幂函数，二次曲线，对数函数等