随机变量数字特征

期望

如果 $\boldsymbol{X}$ 是在概率空间$(\boldsymbol{\Omega}, \boldsymbol{F}, \boldsymbol{P})$中的随机变量，那么它的期望值 $\mathbf{E}(\boldsymbol{X})$ 的定义是：

$$\mathbf{E}(\boldsymbol{X}) = \int_{\boldsymbol{\Omega}} \boldsymbol{X} \, \mathrm{d}\boldsymbol{P}$$

并不是每一个随机变量都有期望值的，因为有的时候上述积分不存在。

如果两个随机变量的分布相同，则它们的期望值也相同。

如果 X 是**离散**的随机变量，输出值为 $x_1, x_2, \ldots$，和输出值相应的概率为 $p_1, p_2, \ldots$（概率和为1）。若**级数** $\sum_i p_i x_i$ **绝对收敛**，那么期望值 $\mathrm{E}(X)$ 是一个无限数列的和。

$$\mathrm{E}(X) = \sum_i p_i x_i$$

如果 X 是**连续**的随机变量，存在一个相应的概率密度函数 $f(x)$，若**积分** $\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$ **绝对收敛**，那么 X 的期望值可以计算为：

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

是针对于连续的随机变量的，与离散随机变量的期望值的算法同出一辙，由于输出值是连续的，所以把求和改成了积分。

方差

定义

方差（Variance），应用数学里的专有名词。在概率论和统计学中，一个随机变量的方差描述的是它的离散程度，也就是该变量离其期望值的距离。这里把复杂说白了，就是各个误差将之平方（而非取绝对值，使之肯定为正数），相加之后再除以总数，透过这样的方式来算出各个数据分布、零散（相对中心点）的程度。继续延伸的话，方差的正平方根称为该随机变量的标准差（此为相对各个数据点间）。

设 X 为服从分布 F 的随机变量， 如果 $E[X]$ 是随机变数 X 的期望值。 随机变量 X 或者分布 F 的**方差**为（均值$\mu = E[X]$）：

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right]$$

这个定义涵盖了连续、离散、或两者都有的随机变数。方差亦可当作是随机变数与自己本身的共变异数(或协方差)：

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X)$$

离散随机变量的方差

如果随机变数X是具有机率质量函数的离散机率分布 $x_1 \mapsto p_1, \ldots, x_n \mapsto p_n$ ，则：

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2 = \sum_{i=1}^{n} (p_i \cdot x_i^2) - \mu^2$$

μ 是其期望值：

$$\mu = \sum_{i=1}^{n} p_i \cdot x_i$$

连续随机变量的方差

如果随机变量 X 是连续分布，并对应至概率密度函数 $f(x)$，则其方差为：

$$\operatorname{Var}(X) = \sigma^2 = \int (x-\mu)^2 \, f(x)\, dx = \int x^2 \, f(x)\, dx - \mu^2$$

μ 是其期望值

$$\mu = \int x \, f(x)\, dx$$

> [1] https://zh.wikipedia.org/wiki/方差

标准差

定义

标准差（又称标准偏差、均方差，英语：Standard Deviation，缩写SD），数学符号 σ（sigma），在概率统计中最常使用作为测量一组数值的离散程度之用。标准差定义：为方差开算术平方根，反映组内个体间的离散程度；标准差与期望值之比为标准离差率。测量到分布程度的结果，原则上具有两种性质：

- 为非负数值（因为开平方后再做平方根）；
- 与测量资料具有相同单位（这样才能比对）。

一个总体的标准差或一个随机变量的标准差，及一个子集合样本数的标准差之间，有所差别。

总体标准差

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

上述公式可以如下代换而简化：

$$
\begin{aligned}
\sum_{i=1}^{N}(X_i-\mu)^2 &= \sum_{i=1}^{N}(X_i^2-2X_i\mu+\mu^2) \\
&= \left(\sum_{i=1}^{N}X_i^2\right)-\left(2\mu\sum_{i=1}^{N}X_i\right)+N\mu^2 \\
&= \left(\sum_{i=1}^{N}X_i^2\right)-2\mu(N\mu)+N\mu^2 \\
&= \left(\sum_{i=1}^{N}X_i^2\right)-2N\mu^2+N\mu^2 \\
&= \left(\sum_{i=1}^{N}X_i^2\right)-N\mu^2
\end{aligned}
$$

所以：

$$
\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2}=\sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}X_i^2\right)-\frac{1}{N}N\mu^2}=\sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}X_i^2\right)-\frac{1}{N}N\mu^2}
$$

根号里面，亦即变异数 σ^2 的简易口诀为：「平方和的平均」减去「平均的平方」。

> [1] https://zh.wikipedia.org/wiki/標準差

协方差 Covariance

定义

协方差表示的是两个变量的总体的误差，这与只表示一个变量误差的方差不同。 如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。 如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。

$$
Cov(X,Y)=E[(X-EX)(Y-EY)]
$$

可以通俗的理解为：**两个变量在变化过程中是同方向变化？还是反方向变化？同向或反向程度如何？**

- 你变大，同时我也变大，说明两个变量是同向变化的，这时协方差就是正的；
- 你变大，同时我变小，说明两个变量是反向变化的，这时协方差就是负的；
- 从数值来看，协方差的数值越大，两个变量同向程度也就越大，反之亦然。

性质

如果 X 与 Y 是实数随机变量，a 与 b 是常数，那么根据协方差的定义可以得到以下性质：

$$
\operatorname{cov}(X,X)=\operatorname{var}(X)
$$

$\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$
$\operatorname{cov}(aX, bY) = ab\ \operatorname{cov}(X, Y)$

对于随机变量序列 $X_1, \ldots, X_n$ 与 $Y_1, \ldots, Y_m$，有

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \operatorname{cov}(X_i, Y_j)$$

对于随机变量序列 $X_1, \ldots, X_n$，有

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{i,j\,:\,i<j} \operatorname{cov}(X_i, X_j)$$

相关系数

在概率论和统计学中，相关（Correlation），显示两个随机变量之间线性关系的强度和方向。在统计学中，相关的意义是用来衡量两个变量相对于其相互独立的距离。在这个广义的定义下，有许多根据数据特点而定义的用来衡量数据相关的系数。

皮尔逊积矩相关系数 Pearson's

相关性的度量有很多种，这里介绍一种最常用的皮尔逊积矩相关系数。在统计学中，**皮尔逊积矩相关系数**（英語：Pearson product-moment correlation coefficient，又称作 **PPMCC或PCCs**，文章中常用r或Pearson's r表示）用于度量两个变量X和Y之间的相关程度（线性相关），其值介于-1与1之间。在自然科学领域中，该系数广泛用于度量两个变量之间的线性相关程度。它是由卡尔·皮尔逊从弗朗西斯·高尔顿在19世纪80年代提出的一个相似却又稍有不同的想法演变而来。这个相关系数也称作“皮尔森相关系数r”。

pearson 描述的是线性相关关系，取值[-1, 1]。负数表示负相关，正数表示正相关。在显著性的前提下，绝对值越大，相关性越强。绝对值为0， 无线性关系；绝对值为1表示完全线性相关。

相关系数也可以看成协方差：一种剔除了两个变量量纲影响、标准化后的特殊协方差。既然是一种特殊的协方差，那它：

1. 也可以反映两个变量变化时是同向还是反向，如果同向变化就为正，反向变化就为负。
2. 由于它是标准化后的协方差，因此更重要的特性来了：它消除了两个变量变化幅度的影响，而只是单纯反应两个变量每单位变化时的相似程度。

总体相关系数

两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商：

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

上式定义了总体相关系数，常用希腊小写字母 ρ 作为代表符号。

样本相关系数

估算样本的协方差和标准差，可得到样本相关系数(样本皮尔逊系数)，常用英文小写字母 r 代表：

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

r 亦可由 (X_i, Y_i) 样本点的标准分数均值估算，得到与上式等价的表达式：

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}}{\sigma_Y}\right)$$

其中 $\frac{X_i - \overline{X}}{\sigma_X}$、$\overline{X}$ 及 σ_X 分别是 X_i 样本的标准分数、样本平均值和样本标准差。

> [1] https://zh.wikipedia.org/wiki/皮尔逊积矩相关系数
> [2] https://www.zhihu.com/question/20852004/answer/134902061

斯皮尔曼等级相关系数 Spearman's

衡量单调关系（无论是线性的还是非线性的）的标准，Spearman系数适用于连续和离散变量，包括序数变量（Ordinal variable）。

> [1] https://zh.wikipedia.org/wiki/斯皮尔曼等级相关系数

Kendall 等级相关系数

是用于测量两个测量量之间的序数关联的统计量。与Spearman相关性相反，Kendall相关性不受彼此等级之间的距离的影响，而仅受观察之间的等级是否相等的影响，因此仅适用于离散变量但不适用于连续变量。

矩和协方差矩阵