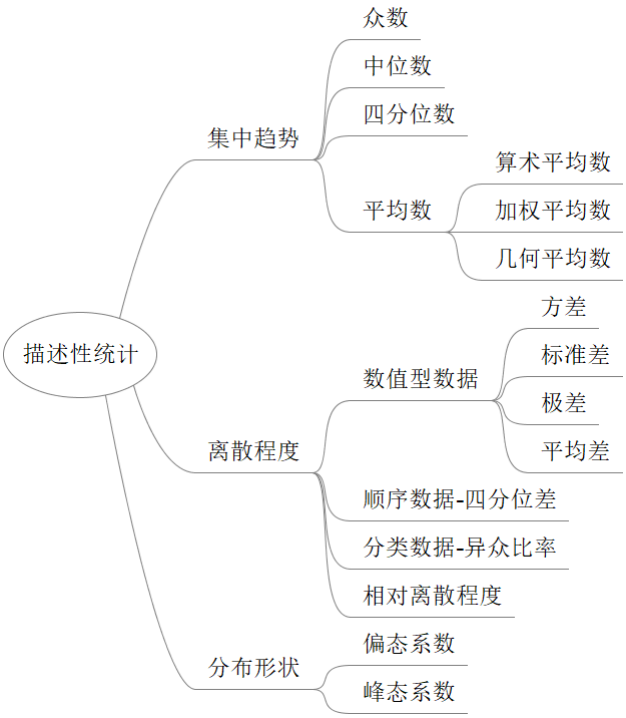
Task

描述性统计

- 集中趋势：众数、中位数、分位数、平均数(算术平均数、加权平均数、几何平均数)
- 离散程度:数值型数据(方差、标准差、极差、平均差)、顺序数据（四分位差）、 分类数据(异众比率)、相对离散程度(离散系数)
- 分布的形状:偏态系数、峰态系数

概念思维导图



集中趋势

指一组数据向中心值靠拢的程度，反应一组数据中心点的位置所在

分类数据：

众数 (Mode)

定义：一组数据中出现次数最多的变量值，使用M_0表示

主要用于测度分类数据的集中趋势， 也可作为顺序数据以及数值型数据集中趋势的测度值。

一般情况下，只有数据量比较大的情况下众数才有意义。

顺序数据：

中位数 (median)

定义： 一组数据排序后处于中间位置上的变量值，使用M_e表示

中位数主要用于测度顺序数据的集中趋势，也适用于数值型数据的集中趋势，但不适用于分类数据。

计算中位数时，先对数据进行排序，然后确定中位数的位置，最后确定中位数的具体值。

中位数计算公式：

$$M_e = \begin{cases} x_{\frac{n+1}{2}}, & n = \text{奇数} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & n = \text{偶数} \end{cases}$$

四分位数 (quartile)

定义：一组数据排序后处于 25% 位置上的数值（下四分位数）和处于 75% 位置上的数值（上四分位数）

设下四分位数为 Q_L ，上四分位数为 Q_U ，根据四分位数据的定义有：

$$Q_{L\text{位置}} = \frac{n}{4}$$
$$Q_{U\text{位置}} = \frac{3n}{4}$$

数值型数据

平均数 (mean)

定义： 平均数也称均值，是一组数据相加后除以数据的个数得到的结果。

平均数根据 ***是否分组*** 分为简单平均数和加权平均数：

- 简单平均数：根据未经分组数据计算的平均数

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- 加权平均数： 根据分组数据计算的平均数

$$\overline{x} = \frac{M_1 f_1 + M_2 f_2 + \ldots + M_k f_k}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{i=1}^{k} M_i f_i}{n}$$

其中，M 为各组的组中值，n 为样本量。 $n = \sum_{i=1}^{k} f_i$

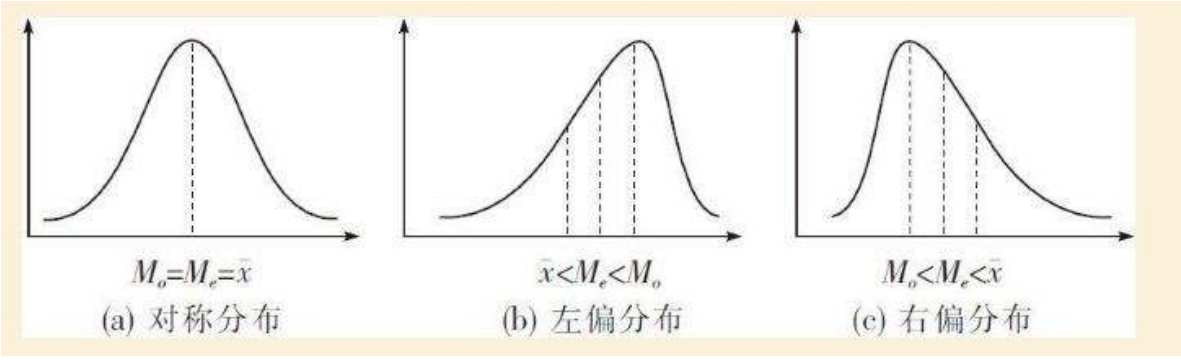
平均数根据 ***计算方式*** 又分为算数平均数和几何平均数：

- 算数平均数： 与上面简单平均数的计算方式一样。

- 几何平均数： n 个变量值连乘积的 n 次方根 ，主要用于计算平均比率，例如现象的平均增长率当掌握的变量值本身是比率形式时，采用几何平均法计算平均比率更为合理。

$$G = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

众数、中位数、平均数比较



其中 M_e 为中位数，M_o为众数， $\overline{x}$为均值

在这个图中，纵坐标表示的是「**频率**」，众数就是图中频率最大的变量值，中位数仅与样本的总数有关，在样本的中间位置，平均数在这里计算的是「**加权平均数**」。

- 如果数据的分布是对称的， 即为正态分布（也叫对称分布），即**均值=中位数=众数**
- 如果数据是偏左分布，即**均值<中位数<众数**

- 如果数据是偏右分布，即**众数<中位数<均值**

	中位数	众数	算数平均数	几何平均数
英文名	Median	Mode	Arithmetic mean	Geometric mean
别称	中值		均值	
定义	一组数据排序后处于中间位置上的变量值	一组数据中出现次数最多的变量	n个变量的和除以n	n个变量值连续乘积的n次方
优点	一组数据中间位置上代表值，不受极端值的影响	一组数据分布的峰值，不受极端值的影响	利用了全部数据，实际应用最广泛	不受极端值的影响
缺点	需要先排序	不唯一，可能有零到多个众数	易受极端值的影响	变量值不能为0或者负数，仅适用于计算平均比率
适用场景	顺序数据的集中趋势测度值	分类数据的集中趋势测度值	数值型数据的集中趋势测度值	计算现象的平均增长率

离散程度

数据的离散程度是数据分布的另一个重要特征，反应的是各变量值远离其中心值的程度。

数据的离散程度越大，集中趋势的测度值对该组数据的代表性就越差；离散程度越小，其代表性就越好。

分类数据

异众比率

定义：非众数组的频数占总频数的比率，用 V_r 表示

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i}$$

其中，$\sum f_i$ 为变量值的总频数，f_m 为众数组的频数

异众比率主要用于衡量众数对一组数据的代表程度

异众比率越大，说明非众数组的频数占总频数比重越大，众数代表性越差；

异众比率越小，说明非众数注占总频数的比重越小，众数代表性越好。

异众比率适合测度分类数据的离散程度。

对于顺序数据以及数值型数据也可以计算异众比率。

顺序数据

四分位差

也称内距或四分间距，是上四分位数与下四分位数之差， 用 Q_d 表示

$$Q_d = Q_U - Q_L$$

四分位差反映了中间50%的数据的离散程度

数值越小，说明中间数据越集中；数值越大，说明中间数据越离散。

四分位差不受极值的影响。

由于中位数处于数据的中间位置，因此，四分位差的大小一定程度上说明了中位数对一组数据的代表程度。

主要适用于测度顺序数据的离散程度。对于数值型数据也可以计算四分位差，但不适用与分类数据。

数值型数据

极差

一组数据的最大值与最小值之差，也称为全距 ， 用 R 表示

$$R = max(x_i) - min(x_i)$$

极差是最简单的描述数据离散程度的测度值，但他冗余受极端值的影响。

不能反映出中间数据的离散状况，不能准确描述出数据的分散程度。

平均值

也称平均绝对离差，是各变量值与其平均数离差绝对值的平均数 ， 用 M_d 表示

未分组数据计算平均差公式：

$$M_d = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

分组数据计算平均差公式：

$$M_d = \frac{\sum_{i=1}^{k} |M_i - \overline{x}| f_i}{n}$$

平均差以平均数为中心，反映了每个数据与平均数的平均差异程度，能全面准确的反映一组数据的离散状况。

平均差越大，说明数据的离散程度越大，反之，则说明数据的离散程度越小。

方差和标准差

方差是各变量值与其平均数离差平方的平均数。

它在数学处理上通过平方的办法消去离差的正负号，然后再进行平均。

方差的平方根称为标准差。

方差或标准差能较好的反映出数据的离散程度，是应用最广的离散程度的测度值。

未分组数据和分组数据计算样本方差公式：

$$未分组数据：s^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$$

$$分组数据：s^2=\frac{\sum_{i=1}^{k}(M_i-\overline{x})^2f_i}{n-1}$$

样本方差是用样本数据个数减1后去除离差平方和，其中样本数据个数减 1 即 $n-1$ 称为自由度。

方差开方得到标准差

$$未分组数据：s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$$

$$分组数据：s=\sqrt{\frac{\sum_{i=1}^{k}(M_i-\overline{x})^2f_i}{n-1}}$$

相对位置的度量

- 标准分数
- 经验法则
- 切比雪夫不等式

相对离散程度

为消除变量值水平高低和计量单位不同对离散程度测度值的影响，需要计算**离散系数**。

离散系数也称变异系数，它是一组数据的标准差与其相应的平均数之比。

$$计算公式：v_s=\frac{s}{\overline{x}}$$

离散系数是测度数据离散程度的统计量，主要用于比较不同样本数据的离散程度。

离散系数大，说明数据的离散程度也打；离散系数小，说明数据的离散程度也小。

当平均数接近零时，离散系数的值趋于增大，此时必须慎重解释

对于分类数据，主要用异众比率来测度其离散程度；

对于顺序数据，主要使用四分位差来测度其离散程度；

对于数值型数据，主要使用方差或标准差来测度其离散程度。

当需要不同样本数据的离散程度进行比较时，则使用离散系数。

在实际应用时，选用哪一个测度值来反映数据的离散程度，要根据所掌握的数据的类型和分析目的来确定

分布的形状

偏态和峰态是对分布形状的测度。

偏态及其测量

偏态：是对数据对称性的测度。测度偏态的统计量是偏态系数记作SK

未分组数据偏态系数计算公式：

$$SK = \frac{n\sum(x_i - \overline{x})^3}{(n-1)(n-2)s^3}$$

其中，s^3 是样本标准差的三次方

如果一组数据的分布是对称的，则偏态系数等于0； 如果偏态系数明显不等于0，表明分布是非对称的。 若偏态系数大于1或小于-1，称为高度偏态分布； 若偏态系数在0.51或-1-0.5之间，则认为是中等偏态分布； 偏态系数越接近0，偏态程度就越小。

峰态及其测量

峰态：对数据分布平峰或尖峰程度的测度。测度峰值的统计量是峰态系数，记作K 峰态通常是与标准正态分布比较而言的。 如果一组数据服从标准正态分布，则峰态系数的值等于0； 若峰态系数的值明显不等于0，则表明分布比正态分布更平或更尖，统计称为平峰分布或尖峰分布。

未分组数据计算峰态系数计算公式：

$$K = \frac{n(n+1)\sum(x_i - \overline{x})^4 - 3(\sum(x_i - \overline{x})^2)^2(n-1)}{(n-1)(n-2)(n-3)s^3}$$

分组数据计算峰态系数是用离差四次方的平均数再除以标准差的四次方，计算公式为：

$$K = \frac{\sum_{i=1}^{k}(M_i - \overline{x})^4 f_i}{ns^4} - 3$$

其中， s^4 是样本标准差的四次方