第十讲
统计学
Lecture 10
Statistics

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

- 本讲义在准备过程中由于时间所限，所用材料来源并未规范标示引用来源。所引材料仅用于教学所用，作者无意侵犯原著者之知识版权，所引材料之知识版权均归原著者所有；若原著者介意之，请联系作者更正及删除。

  The time limit during the preparation of these slides incurs the situation that not all the sources of the used materials (texts or images) are properly referenced or clearly manifested. However, all materials in these slides are solely for teaching and the author is with no intention to infringe the copyright bestowed on the original authors or manufacturers. All credits go to corresponding IP holders. Please address the author for any concern for remedy including deletion.

定义
Definitions

- 数据（复数形式）通常为数字的测量值或观测值。一个数据（单数）是单个测量值或观测值，通常称为得分或原始值。

  Data (plural) are measurements or observations that are typically numeric. A datum (singular) is a single measurement or observation, usually referred to as a score or raw score.

- 样本被定义为从感兴趣的总体中选取的一组个体、物品或数据。描述样本的特征（通常为数字）被称为样本统计量。

  A sample is defined as a set of selected individuals, items, or data taken from a population of interest. A characteristic (usually numeric) that describes a sample is referred to as a sample statistic.

- 总体或群体被定义为所有个体、物品或感兴趣的数据的集合，描述总体的特征（通常为数字）被称为总体参数。总体是科学家将要进行推广的群体。

  A population is defined as the set of all individuals, items, or data of interest. A characteristic (usually numeric) that describes a population is referred to as a population parameter. This is the group about which data scientists will generalize.

定义
Definitions

- 统计学是数学的一个分支，用于总结、分析和解释一组数字或观察结果。它通常分为两个类别：描述统计学，主要用于总结观察结果；推断统计学，用于解释描述统计学的意义。

  Statistics is a branch of mathematics used to summarize, analyze, and interpret a group of numbers or observations. it is usually divided into two categories: descriptive statistics, which is mainly to summarize observations, and inferential statistics, which is used to interpret the meaning of descriptive statistics.

- 描述统计学是用于汇总、组织和理解一组得分或观测值的过程。描述统计学通常以图形方式、表格形式（在表格中）或作为汇总统计（单一值）呈现出来。

  Descriptive statistics are procedures used to summarize, organize, and make sense of a set of scores or observations. Descriptive statistics are typically presented graphically, in tabular form (in tables), or as summary statistics (single values).

- 推断统计学是允许研究人员从部分样本的观察中得出的结论推广对更大总体的一种方法。

  Inferential statistics are procedures used that allow researchers to infer or generalize observations made with samples to the larger population from which they were selected.

定义
Definitions

- 元素（或样本或总体的成员）是指收集信息的特定主题或对象。

  An element (or member of a sample or population) is a specific subject or object about which the information is collected.

- 变量是研究中的一个特征，对不同元素可以取不同的值。变量在一个元素上的值被称为观察值或测量值。在统计学中，有两种类型的变量；第一种类型被称为定量变量，第二种被称为定性变量。定量变量为我们提供代表计数或测量的数字。定性变量（或分类数据）为我们提供不是代表观察的数字的名称或标签。

  A variable is a characteristic under study that takes different values for different elements. Inferential statistics are procedures used that allow researchers to infer or generalize observations made with samples to the larger population from which they were selected. In statistics, there are two types of variables according to their elements; first type is called quantitative variable and the second one is called qualitative variable. Quantitative variable gives us numbers representing counts or measurements. Qualitative variable (or categorical data) gives us names or labels that are not numbers representing the observations.

定义
Definitions

- 变量根据它们的类别或测量方式进行分类。我们可以通过所使用的测量尺度来区分它们。有四种测量尺度：名义、有序、区间和比例。

  Variables classified according to how they are categorized or measured. We can distinguish between them through the measurement scale used. There are four levels of measurement scales: nominal, ordinal, interval, and ratio.

- 名义测量水平将数据分类为互斥的类别，数据之间不能施加任何顺序或排名。

  The nominal level of measurement classifies data into mutually exclusive (disjoint) categories in which no order or ranking can be imposed on the data.

- 以下示例包括不同情况下的名义测量水平：
  - 性别：男性、女性。
  - 眼睛颜色：黑色、棕色、蓝色、绿色，...
  - 科学类专业：统计学、数学、计算机、地理学，...

- The following examples include nominal level of measurements in different cases:
  - - Gender: Male, Female.
  - - Eye color: Black, Brown, Blue, Green, ...
  - - Scientific major field: statistics, mathematics, computers, Geography, ...

定义
Definitions

- 有序测量水平将数据分类为可以排序的类别，但排名之间没有精确的差异。

  The ordinal level of measurement classifies data into categories that can be ordered, however precise differences between the ranks do not exist.

- 当分类考虑排名时，倾向使用有序测量水平。例如，评分系统（A、B、C、D、F）提供了有序测量水平的示例。

  When the classification takes ranks into consideration, the ordinal level of measurement is preferred to be used. For instance, the grade system (A, B, C, D, F) provides an examples of ordinal level of measurements.

- 区间测量水平根据度量单位之间的精确差异对数据进行排序。另一方面，得出的测量值属于实数的一个区间。注意，此种情况下，可能没有有意义的零点。

  The interval level of measurement orders data with precise differences between units of measure. On the other hand, the resulting measurement values belong to an interval of the real numbers. Notably, in this case there is no meaningful zero.

定义
Definitions

- 比率测量水平是区间测量水平的一种，具有附加的属性，即存在一个自然的零点。在这种类型的测量中，零表示无。另一个区别在于，我们可以将一些数量归因于其他数量。

  The ratio level of measurement is the interval level with additional property that there is also a natural zero starting point. In this type of measurement zero means nothingness. Another difference lies in that we can attribute some of the quantities to others.

- 例如，当我们谈及两个城市的距离时，如城市X和Y之间的距离，我们发现该测量属于区间水平，但由于我们可以说城市X和Y之间的距离等于城市X和Z之间距离的两倍，它们变成了标准比例尺度。需要注意的是，在这里零具有意义，因为如果距离等于零，意味着城市（位置）本身。在这里我们注意到长度和高度的概念是距离概念的一个特例。

  For example, when we consider the distance between two cities, like two cities X and Y, where we find that the measurement is an interval level, but because of that we can say that the distance between the two cities X and Y is equal twice the distance between the cities X and Z, they become standard ratio scale. Note that here zero has a meaning, because if the distance is equal to zero, it means that the city (position) itself. Here we note that the concepts of length and height are a special case of the distance concept.

采样
Sampling

- 众所周知，在某些情况下，研究大规模群体以便对特定现象得出结论是困难的，因此抽样方法是进行研究并在较短时间内获得代表性结果的最佳解决方案，同时还可以节省时间、精力和金钱。

  Its known that in some cases, it’s hard to study a large population in order to make conclusions about certain phenomena, so sampling methodology is the best solution in order to perform the study and get representative results during shorter period and also it saves efforts and money.

- 抽样不仅仅是随意选择元素。根据所使用的分析类型，建立了几种抽样技术。其中一些技术包括简单随机抽样法、系统抽样法、分层抽样法和聚类抽样法。这些方法之间的差异涉及许多情况，如群体规模、研究者确定的精度程度、群体元素类型以及研究中人口类别的数量。

  Sampling is not just the selection of elements arbitrary. According to the type of analysis used, there are several techniques of sampling were established. Some of these techniques are the simple random sampling method, the systematic method, the stratified method and the clustered sampling method. Differences between such methods refer to many circumstances such as population size, degree of accuracy determined by the researcher, type of elements of population and the number of categories in the population under study.

采样
Sampling

- 简单随机抽样法：这是最简单的抽样方法，适用于群体规模较小的情况。为了得到这种类型的样本，群体的元素应满足以下条件：
  - 人口的所有元素具有相同的选择机会，
  - 人口的所有元素是独立的。
- Simple Random Sampling Method: It’s the simplest method for sampling and it is applicable when the population is slightly small. In order to get a sample of this type the elements of population should be to achieve the following conditions:
  - 1. All elements of population have the same chance of choice,
  - 2. All elements of population are independent.
- 在验证满足这些条件之后，我们可以将元素连续编号，然后使用任一随机化序列来抽取所需的样本元素。

  After verification of the fulfillment of these conditions, we assign serial numbers to the elements of population, then use one of the methods that used in randomization order to pull the required elements of the sample.

采样
Sampling

- 尽管这种简单的抽样方法易于执行，但它具有一些缺点。例如，虽然可以使用这种方法选择许多样本，但它们可能会得出相同的结果。特别是在涉及大规模人口时，这个缺点使它不是最佳选择。

  Even that simple sample method is easy to perform, but it has some disadvantages. For example, many samples can be selected using this method, but they might give same results. This is more obvious in the case of large populations, which makes it not the best choice to use.

- 系统抽样法: 假设我们想使用这种方法抽取大小为n的样本，我们需要执行以下步骤：
  - 给群体元素分配连续的编号，从1到N；
  - 确定一个间隔（称为抽取间隔）。这个间隔的宽度通过将我们感兴趣的群体大小除以所需样本大小来计算：$k = N/n$；
  - 随机选择一个位于1和k之间的数字（例如用s表示）；
  - 从群体中选取那些带有数字$s + t \cdot k$的元素，其中 $0 \leq t \leq n - 1$。

这样我们就得到了所需的样本。

采样
Sampling

- Systematic Sampling Method: Suppose we want to take a sample with size n using this method, we are including the following:
  - We giving the elements of the population serial numbers from 1 up to N;
  - Determining an interval (called the withdrawal period). This interval can be computed their width by dividing the size of the population that we are interested by the required sample size: $k = N/n$;
  - Then we randomly select number located between 1 and k (Let s, for example);
  - Take elements from population that bear numbers $s + t \cdot k$ with $0 \leq t \leq n - 1$.

  So we get the required sample.

- 前面讨论的两种方法涉及了在没有子群或子集的情况下的抽样技术。当群体由多个子群组成时，我们需要使用被称为分层抽样的技术。

  The preceding two methods discussed the sampling techniques under population without subgroups. The situation is different when the population is composed of several subgroups, a technique called stratified sampling technique is needed.

采样
Sampling

- 在统计学中，共享某些特征的群体的一个子集被称为“分层”。在这种情况下，会使用分层抽样方法，并且这些子集会被随机选择。

  In statistics, a subset of a population share some characteristics is called a ‘stratum’ (the plural is strata). In such condition, the stratified sampling method is used and these subsets are selected randomly.

- 对于分层抽样，要求子集的贡献的样本比例与子集的大小成正比。例如，如果你有三个分层，子群规模分别为100、200和300。研究者选择了抽样比例为1/2。那么，研究者必须分别从每个分层中随机抽取50、100和150个样本。

  For stratified sampling, it requires that samples withdrawn from a specific stratum is proportional to the size of it. For example, In statistics, a subset of a population share some characteristics is called a ‘stratum’ (the plural is strata). In such condition, the stratified sampling method is used and these subsets are selected randomly. If you have 3 strata with 100, 200 and 300 population sizes respectively. And the researcher chose a sampling fraction of 1/2 . Then, the researcher must randomly sample 50, 100 and 150 subjects from each stratum respectively.

采样
Sampling

- 分层抽样方法：与集群抽样方法的区别在于，在分层的情况下，研究者从人口分层中随机选择一个样本，并直接对这些元素进行分析；而在集群抽样中，分析是在从人口中随机选择的群集上进行的。通常，每个群集根据地理基础包含异质元素。这种方法的优点是比其他方法更经济。

  The difference between the stratified sampling method and the clustered is that, in case of stratified the researcher select a random sample of elements from population strata and the analyses are performed on the elements directly, while in other case; the analyses are performed on the clusters chosen randomly from the population. Usually, each cluster consists of heterogeneous elements based on geographical bases. The advantage of this method is that it’s cheaper than other methods.

- 当人口较大或涉及分布在大地理区域的元素时，会使用集群抽样。群体抽样会进一步分为一阶段抽样与二阶段抽样。

  Cluster sampling is used when the population is large or when it involves elements residing in a large geographic area. The cluster sampling can be further categorized into single- stage-cluster sampling and two-stage cluster sampling.

研究
Study

- 任何希望研究某种现象的研究者都必须经过一个有组织的机制，以得出合理的结论。这包括从确定研究的主要目标开始，直到撰写结论和建议的阶段。一些研究涉及实验，以测试研究者所提出的主张；而另一些研究不需要涉及这些实验，因为研究者可以通过过去的观察收集信息。

  Any researcher wishes to study a certain phenomena has to go through an organized mechanism in order to get reasonable conclusions, starting from determining the main objective of the study till the stage of writing conclusions and recommendation. Some studies involve an experiment to test a proposed claim by the researcher, while others do not need to involve such experiments because the researcher can collect information using past observations.

- 根据研究者获取观察结果的方式，统计研究可分为两种类型。一种称为观察研究，另一种称为实验研究。

  Statistical studies are classified into two types of studies according to how researcher gets observations. One is called observational study, and the other is called In experimental study.

- 在观察研究中，研究者对研究过去发生的事情或在进行研究时正在发生的事情感兴趣，然后对收集到的观察结果进行统计分析，以做出正确的决策。例如，当研究者对研究两个或更多变量之间的相关性感兴趣时，会使用这种方法。例如，出勤率与成绩的关系。

研究
Study

- In an observational study, the researcher is interested in studying something happened in the past or happening at the moment of performing the study, then (he/she) makes statistical analysis on collected observations to take the right decision. For example, it’s used when the researcher interested in studying the correlation between two or more variables. For example, the relationship between absence hours of students and their GPA.

- 当研究者通过将群体元素分配到不同的组中，对每个组使用不同的处理方法进行研究时，称为实验研究。即研究者在研究中对这些组应用因子或因素，以便进行比较。这种类型的研究还关注一种变量（研究者操控的变量，称为自变量或解释变量）对其他变量（结果变量，也称为因变量）的影响。临床试验是实验研究的一个很好的例子。

  If the researchers study elements of population by distributing them into groups and each group is being studied using different treatments, it is called experimental studies. That is, the researcher applies factor(s) on the groups under study in order to compare them. This type of studies also concerns about the effect of a variable (the variable which is manipulated by researcher, called the independent variable or explanatory variable) on other variable(s) (outcome variable, called the dependent variable). Usually,, and the. Clinical trials are good example on experimental studies.

数据
Data

- 在采集之后、处理或排名之前按照采集顺序记录的数据被称为原始数据。例如，学生的原始成绩。显然，这些成绩为定量数据。当根据分数，将学生分为A、B、C、D等四档之后，得到的便是定性数据。

  the variable which is manipulated by researcher is called the independent variable or explanatory variable, and the outcome variable is called the dependent variable. For example, the original scores of the students. Obviously these scores are quantitative data. If we assign A, B, C, D as the grades for the students based on their scores, now we have the qualitative data.

- 有几种方法可以组织定性数据集。第一种叫做频数表。一个定性数据的频数表列出了所有的类别、名称或标签，以及属于每个类别、名称或标签的元素数量。此外，为了显示相应类别所占总频数的比例，我们可以使用相对频数。它可以通过将该类别的频数除以所有频数之和来计算得出。

  There are several ways to organize qualitative data set. The first is called frequency table. A frequency table for qualitative data lists all categories , names or labels and the number of elements that belong to each of the categories, names or labels. In addition, to show what proportion of the total frequency belongs to the corresponding category, we can use relative frequency. It can be calculated by dividing the frequency of that category by the sum of all frequencies.

数据
Data

- 为了对定量数据进行分组和展示，我们可以使用频数分布表。在此之前，有必要解释一些相关概念。一个类别是一个非重叠区间，包含所有在两个数字之间的值。第一个数字被称为下限，第二个数字被称为上限。这些类别始终代表一个变量，并且它们是非重叠的；也就是说，变量中的每个值属于一个且仅属于一个类别。类别通常出现在频数分布表的第一列。第二列代表数字，每个数字属于一个且仅属于一个区间，这些列出的数字被称为频数，用（f）表示，其中每个频数表示属于每个类别的值的数量。通常，以频数分布表的形式呈现的数据被称为分组数据。

  To group and display the quantitative data, we can use frequency distribution table. Before that, it is necessary to explain some related concepts. A class is an nonoverlapping interval that includes all values that fall within two numbers. The first number is called the lower limit and the second number is called the upper limit. The classes always represent a variable and they are nonoverlapping; that is, each value in the variable belongs to one and only one class. The classes usually appear in the first column in a frequency distribution table. The second column represents the number such that counting numbers belongs to one and only one interval, these listed numbers are called frequency denoted by (f) where each frequency gives the number of values that belong to each class. Usually, data presented in the form of a frequency distribution table are called grouped data.

数据
Data

- 分组数据可以以直方图、多边形或累积频数曲线的方式展示。具有相等类宽的频数分布表中的分组数据的直方图是一个图表，在水平轴上标有类别边界，垂直轴上标有频数、相对频数或百分比。频数多边形是一种图表，通过连接绘制在类别中点频数处的点的线段来显示数据。累积频数曲线是绘制在分布表中的分组数据的递增累积频数的曲线，首先通过将在各类别上界上方标记的绘制点连接起来，高度等于各类别的递增累积频数，然后通过平滑曲线连接这些点。

  Grouped data can be displayed in a histogram, a polygon or ogive. A histogram of grouped data in a frequency distribution table with equal class widths is a graph in which class boundaries are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on a vertical axis. A frequency polygon is a graph that displays the data by using line segments that connect points plotted for the frequencies at the midpoints of the classes. An ogive is a curve drawn for the ascending cumulative frequency of grouped data in a distribution table by first joining plotting dots marked above the upper boundaries of classes at heights equal to the ascending cumulative frequencies of respective classes, then joining these points by smooth curve.

中心趋势
Central Tendency

- 中心趋势的度量是一个非常重要的工具，指的是直方图或频数分布曲线的中心。通常，中心趋势的三个度量是均值、中位数和众数，分别适用于分组和未分组的数据集两种情况。

  A measure of central tendency is very important tool that refer to the centre of a histogram or a frequency distribution curve. Usually, the three measures of central tendency are the mean, the median, and the mode for the two cases (grouped and ungrouped data sets).

- 对于未分组的数据，均值是通过将所有值的总和除以数据集中的值的数量来获得的。因此，

  人口数据的均值：$\mu = \sum x/N$

  样本数据的均值：$\bar{x} = \sum x/n$

  其中$\sum x$表示所有值的总和，N是人口规模，n是样本规模，μ是人口均值，$\bar{x}$是样本均值。

- The mean for an ungrouped data is obtained by dividing the sum of all values by the number of values in that data set. Thus,

  Mean for population data: $\mu = \sum x/N$

  Mean for sample data: $\bar{x} = \sum x/n$

  Where $\sum x$ is the sum of all values, N is the population size, and n is the sample size, μ is the population mean, and $\bar{x}$ is the sample mean

均值
Mean

- 对于分组数据集，均值的计算取决于所关注数据集的所有值的总和和它们的数量。但是对于分组数据，很难找到所有值的总和或它们的数量，在这种情况下，用于计算均值的是对所有值总和的近似值。这个近似值可以通过以下定义来说明。

  Calculation of the mean for grouped data set depends on the sum of all values of the interest data set and their numbers. But for the grouped data, it is impossible to find the sum of all values or their numbers, in such case an approximate to the sum of all values is used to calculate the mean. This approximation is illustrated by the following definition.

  人口数据的均值：$\mu = \sum x_m f_m / N$

  Mean for population data: $\mu = \sum x_m f_m / N$

  样本数据的均值：$\bar{x} = \sum x_m f_m / n$

  Mean for sample data: $\bar{x} = \sum x_m f_m / n$

  其中x_m是类别m的中点，f_m是类别m的频数。

  Where x_m is the midpoint and f_m is the frequency of the class m.

- 类别中点是通过将一个类别的两个边界（或两个限制）的和除以2来获得的。

  The class midpoint is obtained by dividing the sum of the two boundaries (or the two limits) of a class by 2.

中程
Midrange

- 一个变量的加权平均可以通过将每个值乘以其相应的权重，然后将乘积的总和除以权重的总和来找到，其中w_1、w_2 、…、w_n是权重，x_1、x_2、…、 x_n是值：

  The weighted mean of a variable can be found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights Where $w_1, w_2, \ldots, w_n$ are the weights and $x_1, x_2, \ldots, x_n$ are the values.

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \sum w_iw_i \Big/ \sum w_i$$

- 数据集的近似中点被称为中程。中程（MR）的定义是数据集中最低值和最高值的和除以2：

  An approximate of the middle point for a data set is called midrange. The midrange (MR) is defined as the sum of the lowest and highest values in the data set divided by 2:

$$MR = \frac{\min(Dataset) + \max(Dataset)}{2}$$

- 注意，中程作为中心趋势的衡量是较弱的，因为它仅取决于数据集中的两个值。

  Note that, MR is weak as a measure of central tendency since it is depends only on two values among of all values in the data set.

中位数
Median

- 表示排列数据集中位数的中心趋势度量被称为中位数。它是在已按升序或降序排列的数据集中的中间项的值。

  A measure of central tendency that represents the middle term of a ranked data set is called the median. It is the value of the middle term in a data set that has been ranked in increasing or decreasing order.

- 为了找到给定数据的中位数，我们需要以下三个步骤：
  - 将给定的数据集排名（按升序或降序排列）
  - 找到在步骤1中获得的排名数据集的中间项。
  - 这个项的值代表中位数。

- To find the median of a given data we need the following three steps
  - Rank the given data sets (in increasing or decreasing order)
  - Find the middle term for the ranked data set that obtained in step 1.
  - The value of this term represents the median.

众数
Mode

- 排名数据x_1、x_2、…、 x_n的中位数由以下给出：

  The median of the ranked data $w_1, w_2, \ldots, w_n$ is given by:

$$Median = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n\ is\ odd \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2} & n\ is\ even \end{cases}$$

- 众数是另一种中心趋势度量，它被称为数据集中最常见的值。按照定义，众数是数据集中出现次数最多的值。

  The mode is another measure of central tendency and it is known as the most common value in a data set. The definition of the mode is that it is the value that occurs most often in a data set.

- 对于给定的未分组数据集，众数有几种情况，一个数据集可能没有众数，也可能有一个众数，还可能有多个众数。

  For a given ungrouped data set there are several cases of mode, a data set may have none or have one mode and may have more than one mode.

众数
Mode

- 无众数的数据集：在这种数据集中，每个值只出现一次。
- 一个众数的数据集：在这种数据集中，只有一个值出现的频率最高。这种情况下的数据集被称为单峰数据集。
- 两个众数的数据集：在这种数据集中，两个值的出现频率相同（最高）。这种情况下的分布被称为双峰分布。
- 多于两个众数的数据集：在这种数据集中，多于两个值的出现频率相同（最高），则数据集包含多于两个众数，被称为多峰数据集。

- Data set with none mode: In such data set each value occurring only once.
- Data set with one mode: In such data set only one value occurring with the highest frequency. The data set in this case is called unimodal.
- Data set with two modes: In such data set two values that occur with the same (highest) frequency. The distribution, in this case, is said to be bimodal.
- Data set with more than two modes: In such data set more than two values occurs with the same (highest) frequency, then the data set contains more than two modes and it is said to be multimodal.

平均偏差
Mean Deviation

- 在某些情况下，中心趋势的度量可能无法清楚地展现数据集的分布情况。为了解决这个问题，我们需要使用变异度的度量。

  The measures of central tendency in some circumstance may not be able to give a clear picture of the distribution of a data set. to address this, we need the measures of variation.

- 一种变异度的度量被称为平均偏差，它是每个值与均值之间距离的平均值：$Mean\ Deviation = \sum|x-\bar{x}|/n$，其中$x$是一个值，$\bar{x}$是均值，$n$是值的数量：

  measure of variation is called mean deviation; it is the mean of the distances between each value and the mean, it is given by the by the formula $Mean\ Deviation = \sum|x-\bar{x}|/n$, where x is a value, $\bar{x}$ is the mean, and n is the number of values :

- 分组数据在k个类别中的平均偏差计算公式为：$Mean\ Deviation = \sum f_m|x_m-\bar{x}|/\sum f_m$，其中$x_m$是第$m$个类别的中点，$f_m$是第$m$个类别的频数，$\bar{x}$是分组数据的均值。

  We use the formula $Mean\ Deviation = \sum f_m|x_m-\bar{x}|/\sum f_m$ to calculate the mean deviation of grouped data in k classes, where x_m is the midpoint of the m-th class, f_m is the frequency of the m-th class, and $\bar{x}$ is the mean of grouped data.

标准差
Standard Deviation

- 最常用的变异度度量为前面讲过的标准差，用符号σ表示（用于总体）和S表示（用于样本）。该度量的数值告诉我们数据集中与该度量相对应的值在均值周围是如何紧密分布的。标准差平方称为方差，用符号σ^2表示（用于总体）和S^2表示（用于样本）。

  A most used measure of variation is called standard deviation denoted by σ (for the population) and S (for the sample). The numerical value of this measure helps us how the values of the dataset corresponding to such measure are relatively closely around the mean. To squared the standard deviation leads to the variance, denoted by σ^2 (for the population) and S^2 (for the sample).

- 用于计算总体方差和总体标准差的公式如下，其中N与n分别为总体大小与样本个数：

  The formulas that are used to calculate the population variance and the population standard deviation are as below, N and n denote the population size and sample size respectively.

- 总体（Population）

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

- 样本（Samples）

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

变异系数
Coefficient of Variation

- 对于分组数据或频数分布，总体方差和总体标准差的公式如下：

  For the grouped data or frequency distribution, The formulas for the population variance and the population standard deviation are given by:

- 总体（Population）

$$\sigma^2 = \frac{\sum f_m (x_m - \mu)^2}{N}$$

- 样本（Samples）

$$\sigma^2 = \frac{\sum f_m (x - \bar{x})^2}{n - 1}$$

- 标准差的一个缺点是它是绝对变异性的度量，而不是相对变异性的度量。我们引入变异系数（CV）来比较具有不同计量单位的两个不同数据集的变异性。其计算如下：

  One disadvantage of the standard deviation that its being a measure of absolute variability and not of relative variability. Therefore, we introduce the coefficient of variation (CV) to compare the variability of two different data sets that have different units of measurement. it can be calculated:

- 总体（Population）

$$CV = \frac{\sigma}{\mu} \times 100\%, \mu \neq 0$$

- 样本（Samples）

$$CV = \frac{S}{\bar{x}} \times 100\%, \bar{x} \neq 0$$

标准分数
Standard Score

- 对于总体或样本数据集，位置度量是用来确定单个值与其他值的位置关系的工具。常用的度量有标准分数、四分位数、百分位数和百分位等级等。

  For a population or a sample data sets, the measures of position are a tool that used to determine the position of a single value in relation to other values. Commonly-used measures include the standard Scores, quartiles, percentiles, and percentile rank, etc.

- 比较常用的一个位置度量被称为标准分数。一个值的标准分数（或z分数）由以下公式计算。

  A pretty commonly-used measure of position is called standard score, A standard score (or z - score) for a value is obtained by the formula:

$$z = \frac{value\ \ - mean}{standard\ deviation}$$

- 总体（Population）

$$z = \frac{x-\mu}{\sigma}$$

- 样本（Samples）

$$z = \frac{x-\bar{x}}{S}$$