上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

上海交通大学生物医学工程学院
SJTU School of Biomedical Engineering

生物医学统计概论(BI148-2)

Fundamentals of Biomedical Statistics

2022 春季

授课老师：林关宁

课程内容安排

上课日期	章节	教学内容	教学要点	作业	随堂测	学时
2.16	1	数据可视化，描述性统计	1. 课程介绍 & 数据类型	作业1（8%）	测试1（8%）	2
2.23			2. 描述性统计Descriptive Statistics & 数据常用可视化			2
3.2			3. 大数定理 & 中心极限定理			2
3.9			4. 常用概率分布			2
3.16	2	推断性统计，均值差异检验	5. 统计推断基础-1：置信区间 Confidence Interval *	作业2（12%）	测试2（12%）	2
3.23			6. 统计推断基础-2:假设检验 Hypothesis Test			2
3.30			7. 数值数据的均值比较-1： 单样本t-检验			2
4.6			8. 数值数据的均值比较-2： 独立双样本t-检验，配对样本t-检验			2
4.13			9. 数值数据的均值比较-3：One-Way ANOVA			2
4.20			10. 数值数据的均值比较-4：Two-way ANOVA			2
4.27	3	比例差异检验	11. 类别数据的比例比较-1： 单样本比例推断 *	作业3（4%）	测试3（4%）	2
5.7（调）			12. 类别数据的比例比较-2： 联立表的卡方检验			2
5.11	4	协方差，相关分析，回归分析	13. 相关分析（Pearson r, Spearman rho, Kendal’ s tau) *	作业4 （6%）	测试4 （6%）	2
5.18			14. 简单回归分析			2
5.25			15. 多元回归Multiple Regression			2
6.1	5	Course Summary	16. 课程总结 *			2
			Total	30%	30%	32

*** 随堂测试**

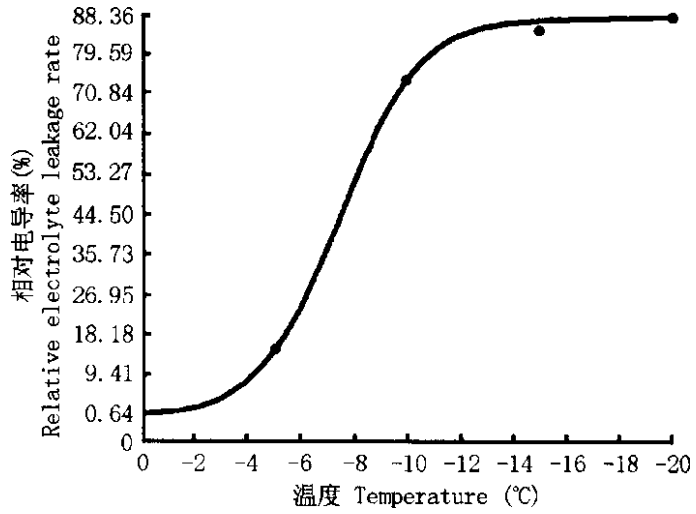
数据整理的目的

1 区别类型，剖析结构和特点，揭示内在联系。

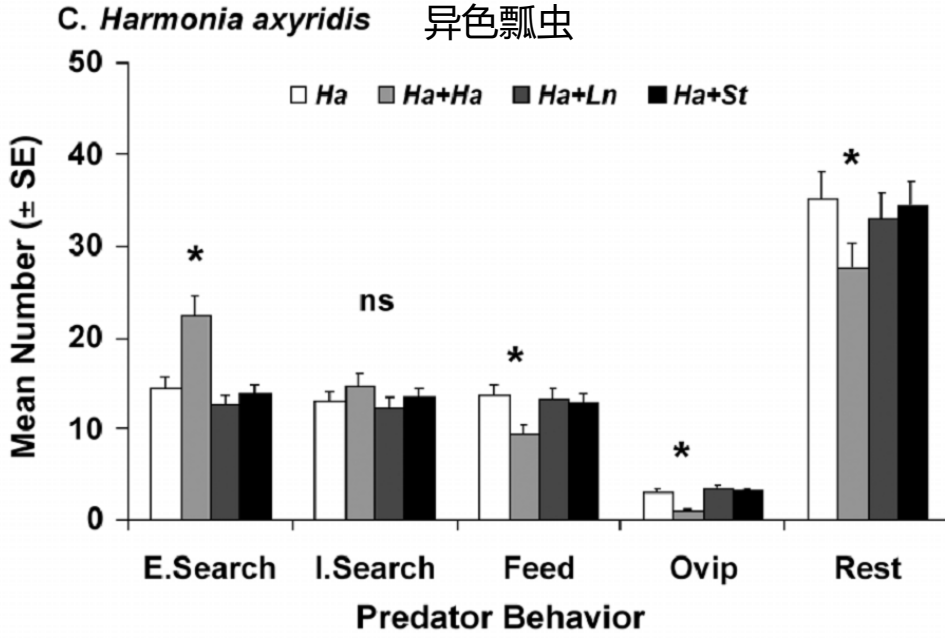
2 找出数据的重要特征： 集中情况，变异情况

3 曲线情况:变化趋势。

4 不规则情况：极端值



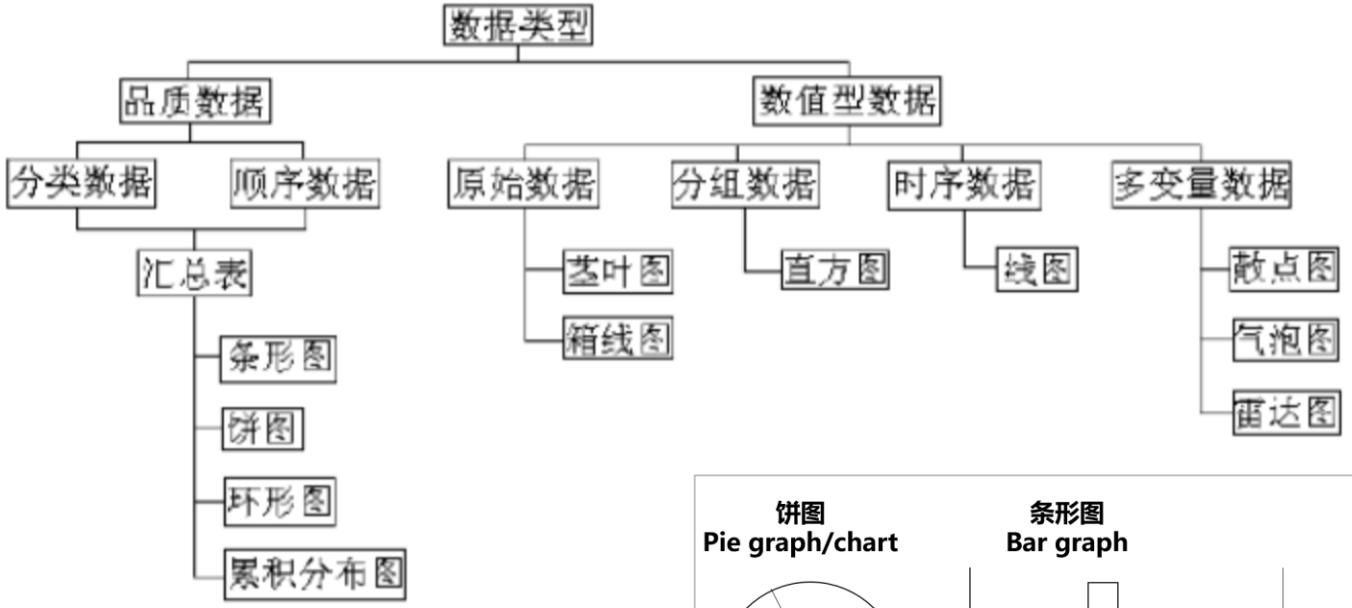
‘TifgreenⅡ’ 叶片相对电导率随温度的变化趋势

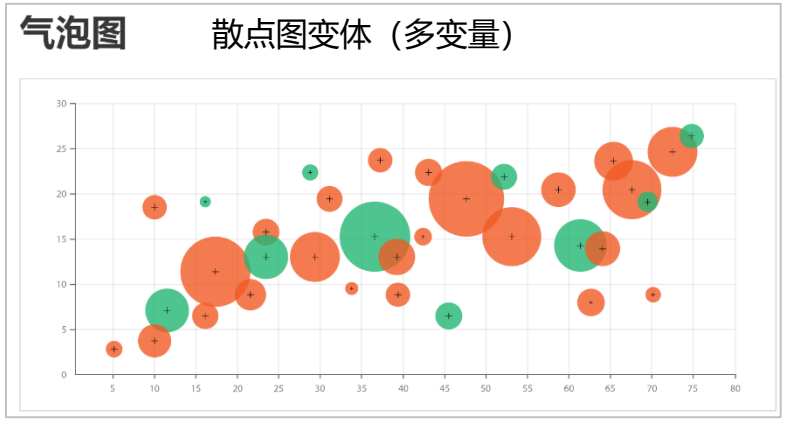


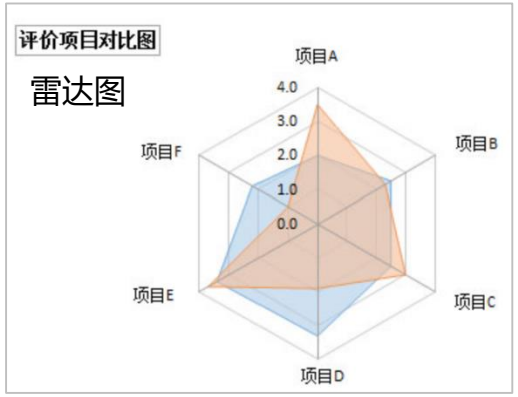
捕食行为

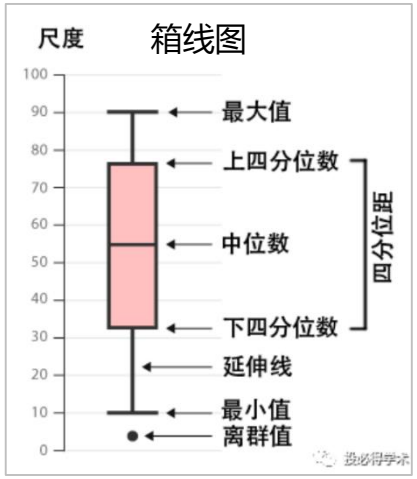
Data & presentation

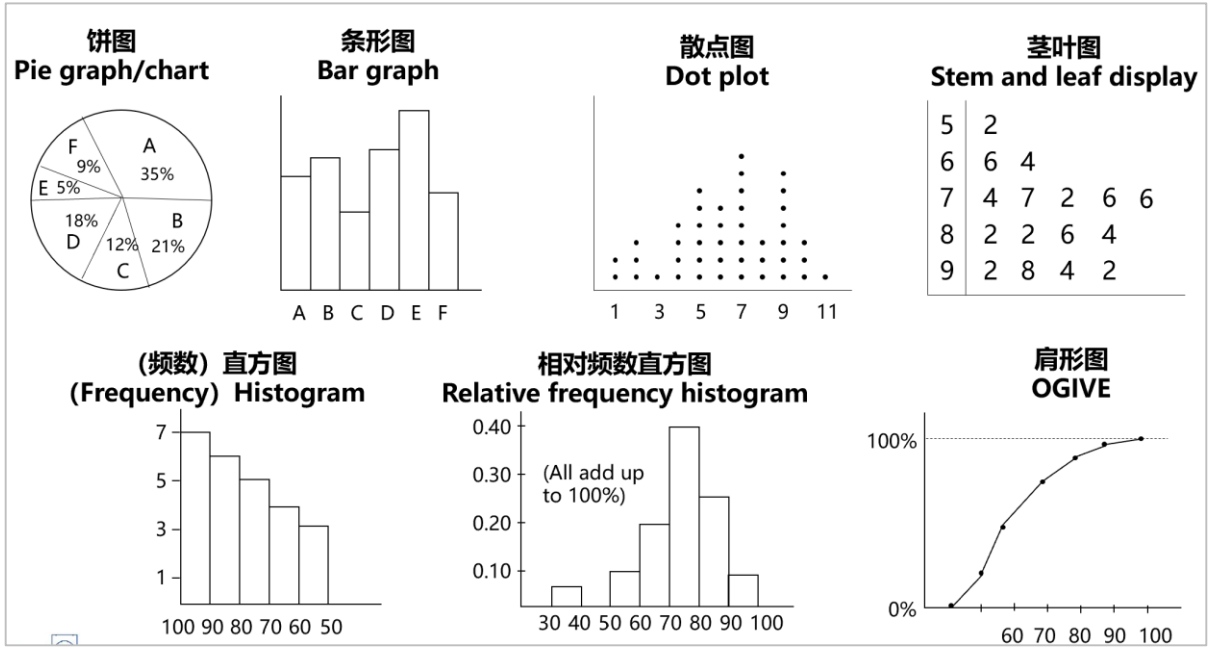
⑶图示：①分组数据——直方图、折线图、曲线图；②未分组数据——茎叶图、箱线图；③时间序列数据——线图；④多变量数据——散点图（两变量）、气泡图（三变量）、雷达图（多变量）。

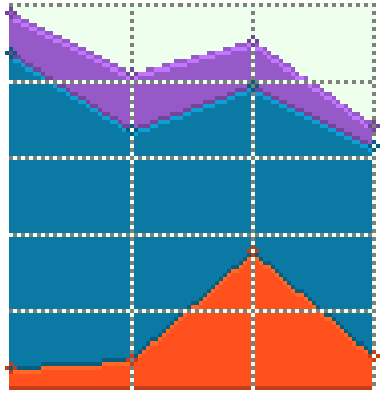
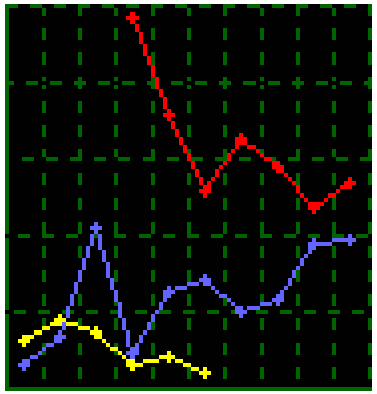
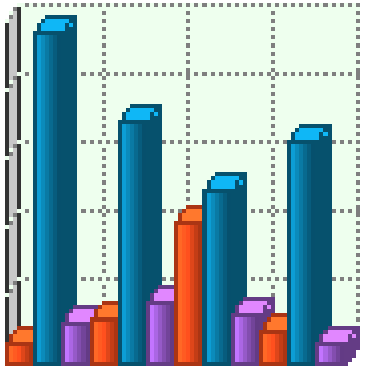
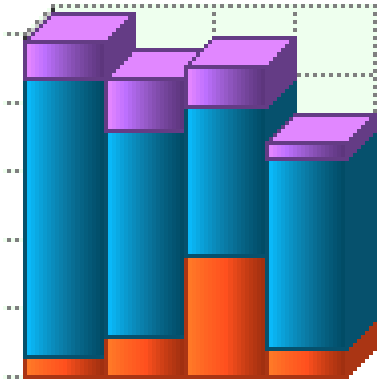


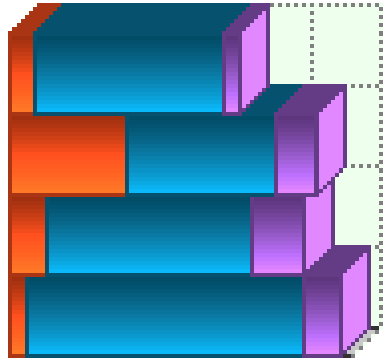










Week 2：描述性统计


描述性统计（Discrete statistics）

- **集中趋势（Measure of central tendency）**
  - 位置参数（Measures of location）
    - 算术平均数（Arithmetic mean, Average）
    - 中位数（Median）
    - 众数（Mode）

- **离散趋势（Measure of dispersion）**
  - 变异指标（Measure of variability）
    - 全距（Range, R）
    - 四分间距（Interquartile Range）
    - 方差（Variance, S^2）
    - 标准差（Standard deviation, SD）
  - 形状参数（Measures of shape）
    - 偏度（Skewness）
    - 峰度（kurtosis）

集中趋势（Measure of central tendency）

- **平均数**
  - 平均数是数量资料的代表值，表示观察值的中心位置与集中趋势，代表研究对象的一般水平，并可作为资料的代表与另一组资料相比较，借以明确两者间的差异。

$\bar{x}$ ：

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

(Mean, arithmetic mean, average)

$\bar{x}$ 基本特性：

（1）$\bar{x}$的大小受样本内每个值的影响

例1： 8 6 4 2 $\bar{x}=5$

6 6 4 2 $\bar{x}=4.5$

（2）若每个x_i都乘以相同的数K，则 $\bar{x}$ 也应乘以K

例2： 8 6 4 2 $\bar{x}=5$

×2 16 12 8 4 $\bar{x}=10=5\times2$

（3）若每个x_i都加上相同的数A，则 $\bar{x}$ 也应加上A

例3： 8 6 4 2 $\bar{x}=5$

+2 10 8 6 4 $\bar{x}=7=5+2$

（4）样本内各观察值与其平均数的差数，离均差之和等于零。

$$\sum(x_i-\bar{x})=0$$

$$\sum(x_i-\bar{x})=\sum(x_1-\bar{x})+(x_2-\bar{x})+.....+(x_n-\bar{x})$$

$$=(x_1+x_2+.....+x_n)-n\bar{x}$$

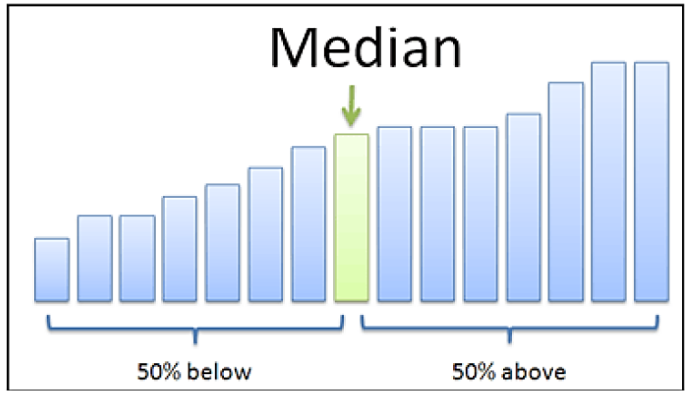
$$=\sum x_i-n\bar{x}$$

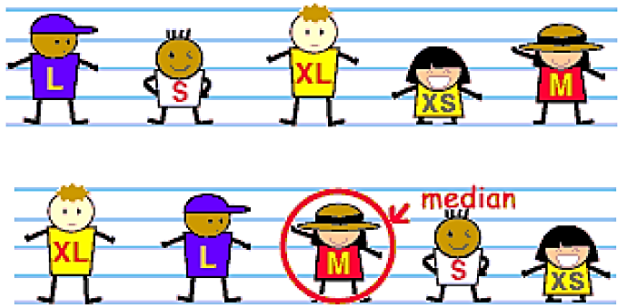
$$=\sum x_i-n\times\frac{\sum x_i}{n}$$

$$=0$$

集中趋势（Measure of central tendency）

- **中位数（Median）**
  - 一组观察值按大小顺序排列，位次居中的那个数值。在频数分配中处于中点（其上下各有相同的频数分布），**不受极端值影响。**





注意：

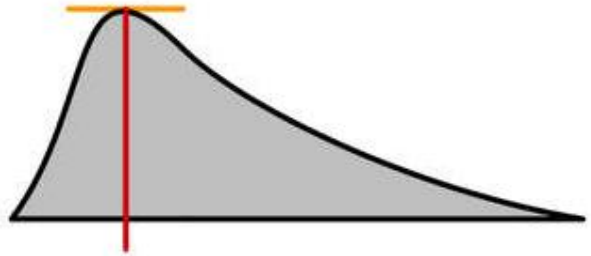
当一组观察值大部分较集中，少数甚至个别的分散在一 侧时，中位数比算术平均数可以更确切地反映频数的集 中情况。中位数是一种位置上的平均数。

集中趋势（Measure of central tendency）

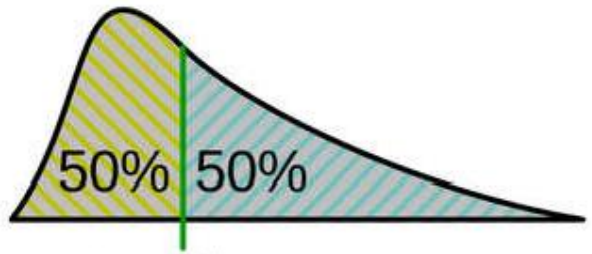
- **众数（Mode）**
  - 资料中频数最多的一个数，或次数最多一组的组值

鞋码：	38	39	40	40	40	41	Mo=40
	38	39	39	40	41	41	Mo=39、41
	38	39	40	41	42	43	无Mo

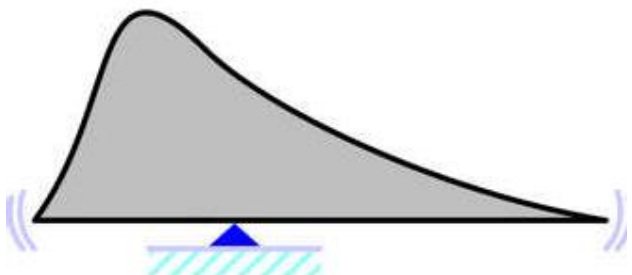
  - 众数不受极端值影响，不必计算，在定性分类中应用
  - 众数的频数多少可反映总体的集中情况。频数越大，表明总体的集中度越大，众数对总体的代表性越大。
  - 正态分布的众数接近算术平均数
  - 应用众数要求观察值很多，n 较大



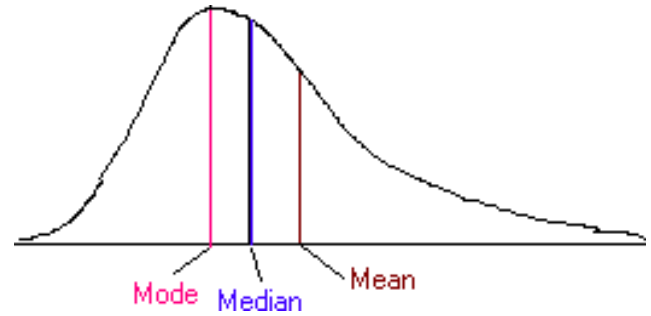
mode

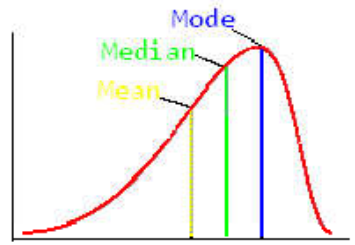


median

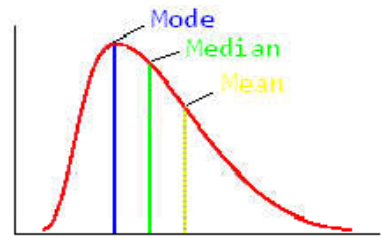


mean

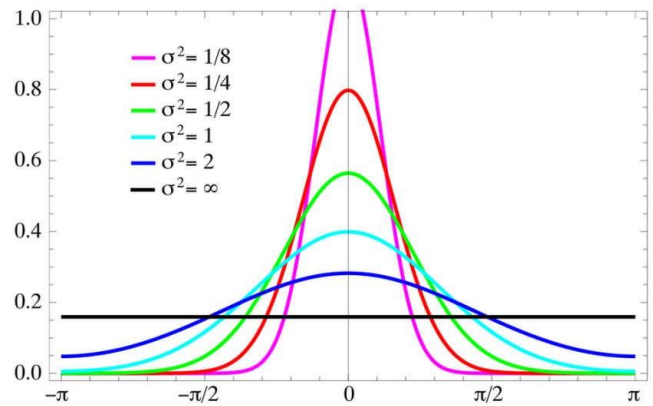




Negatively Skewed
(Skewed to the Left)



Positively Skewed
(Skewed to the Right)



定性数据、类别数据的集中趋势只能用**众数**，如眼睛颜色、收入档次、细胞类型等等。

中位数和均值最适合于定量数据，如身高、收入、年龄、 反应时间、含量等等。

如果数据中有极端值，为排除极端值，可用**中位数**。

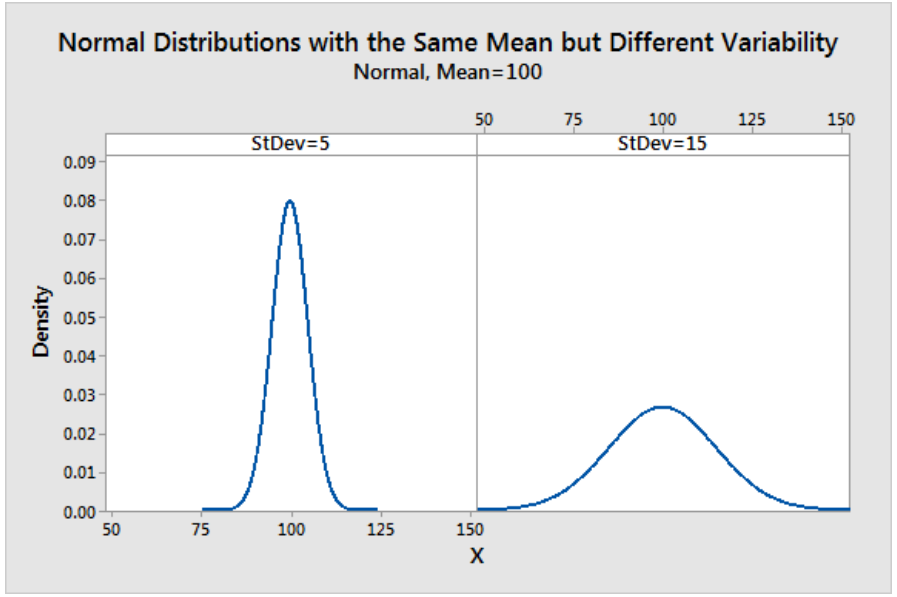
描述性统计（Discrete statistics）

- 集中趋势（Measure of central tendency）
  - 位置参数（Measures of location）
    - 算术平均数（Arithmetic mean, Average）
    - 中位数（Median）
    - 众数（Mode）

- **离散趋势（Measure of dispersion）**
  - 变异指标（Measure of variability）
    - 全距（Range, R）
    - 四分间距（Interquartile Range, IQR）
    - 方差（Variance, S^2）
    - 标准差（Standard deviation, SD）
  - 形状参数（Measures of shape）
    - 偏度（Skewness）
    - 峰度（kurtosis）

离散趋势（Measure of dispersion）

- **变异指标（Measure of variability）**
  - 变异指标又称为标志变动度，它是反映总体中各单位标志值差异程度的综合指标
  - 变异性的度量是一种汇总统计量，代表数据集中的离散量
  - **数据中的值如何分布？虽然集中趋势的度量描述了典型值，但变异性的度量则定义了数据点偏离中心的距离。**
  - 变异指标越大，表明数据越分散、不集中；变异指标越小，表明数据越集中，变动范围越小；变异指标反映现象总体总单位变量分布的离中趋势



变异指标（Measure of variability）

- **全距（Range, R）**
  - 又称极差，它是总体各单位变量值中最大值与最小值之差
  - 公式为 $R = X_{max} - X_{min}$
  - R表示全距，X_{max}表示总体中最大的变量值；X_{min}表示总体中最小的变量值；
  - 全距说明了总体中变量值的变动范围。全距越大，说明变量值的变动范围越大，从而说明变量值的差异大；反之则小
  - 它的缺点？
  - 它适用的地方？

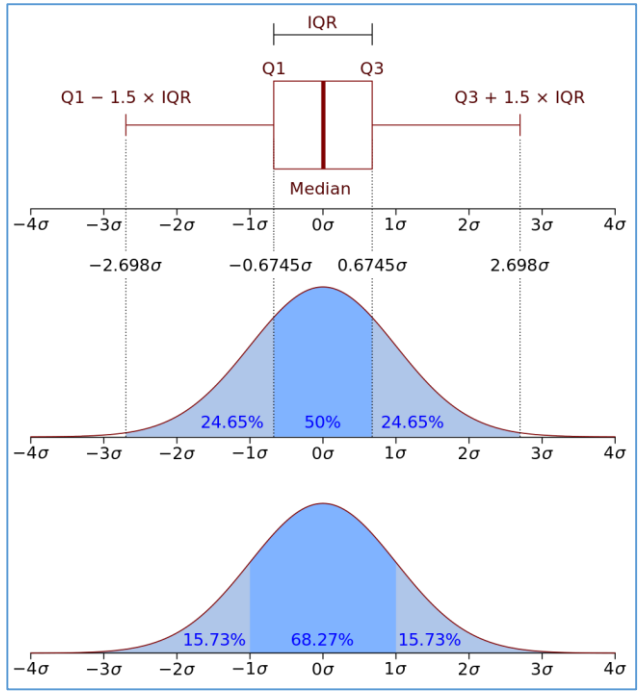
它仅基于数据集中的两个最极端值，这使其非常容易受到异常值的影响。如果这些数字之一异常高或低，会影响整个范围。

变异指标（Measure of variability）

- **四分间距（Interquartile range, IQR）**
  - **第三个**四分位数和**第一个**四分位数的差值（即 Q_1，Q_3 的差距），$IQR = Q_3 - Q_1$
  - 与方差、标准差一样，表示统计资料中各变量分散情形，但四分差更多为一种稳健统计（robust）
  - 四分位距通常是用来构建箱形图（box plot），以及对概率分布的简要图表概述。对一个对称性分布数据（其中位数必然等于第三个四分位数与第一四分位数的算术平均数）

四分差的距离为
115 – 105 = 10

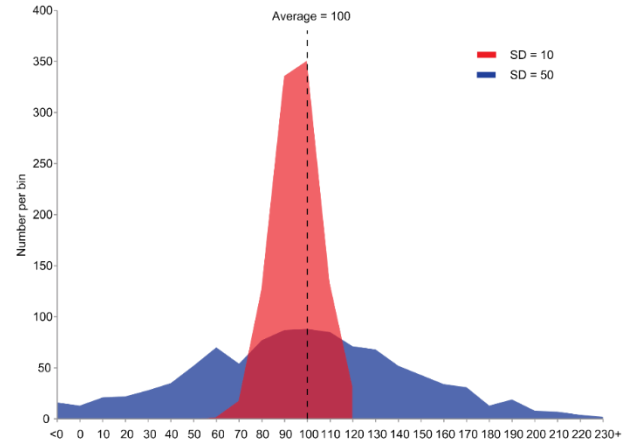
数列	参数	四分差
1	102	
2	104	
3	105	Q_1
4	107	
5	108	
6	109	Q_2（中位数）
7	110	
8	112	
9	115	Q_3
10	118	
11	118	



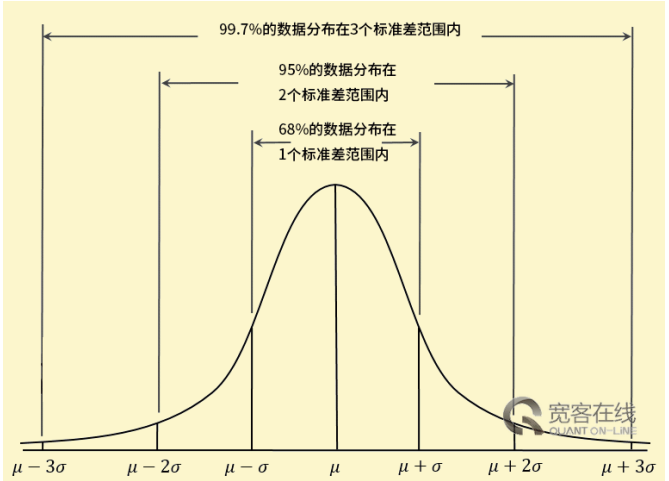
图示中箱形图（有四分位数及四分位距）和概率密度函数 为描述一个常规总量 $N(0, 1, \sigma^2)$ 的分布情况

变异指标（Measure of variability）

- **方差（Variance, s^2, σ^2 ）**
  - 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**
  - 将各个误差之平方（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
  - σ^2 为总体方差，X为变量，μ为总体均值，N为总体例数 $\sigma^2 = \frac{\sum (X-\mu)^2}{N}$
  - 实际工作中，总体均数难以得到时，用样本统计量代替总体参数
  - 经校正后，样本方差计算公式 s^2为样本方差 $S^2 = \frac{\sum (X-\bar{X})2}{n-1}$



- **标准差（均方差，Standard Deviation, SD, σ）**
  - 在概率统计中**最常使用**作为测量一组数值的离散层度之用
  - 定义为方差平方根
  - **反映个体间的离散层度**
  - 标准差是一组数值自平均值分散开来的程度的一种测量观念
    - 一个较大的标准差，代表大部分的数值和其平均值之间差异较大
    - 一个较小的标准差，代表这些数值较接近平均值
  - **通常不用作独立的指标，而与其它指标配合使用**



离散趋势（Measure of dispersion）

- 均值、方差、标准差（统计学里最基本的概念）

一个含有n个样本的集合

均值

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

均值描述的是样本集合的中间点（信息是很有限）

方差（variance）

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

用来度量随机变量和其数学期望（即均值）之间的偏离程度，描述的则是样本集合的各个样本点到均值的距离之平均

之所以除以n-1而不是除以n，是因为这样能使我们以较小的样本集更好的逼近总体的标准差，即统计上所谓的“无偏估计”

标准差（SD）

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

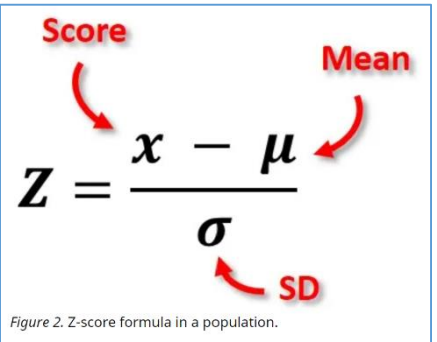


Figure 2. Z-score formula in a population.

[0, 8, 12, 20] => 均值 10，方差 8.3
[8, 9, 11, 12] => 均值 10，方差 1.8

本周

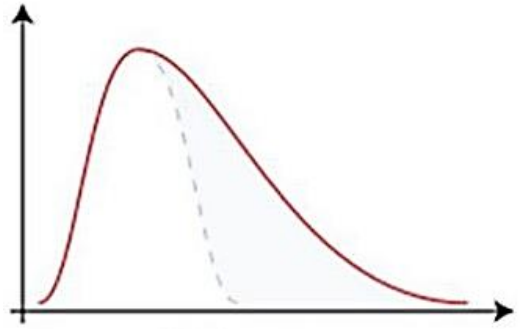
- **集中趋势（Measure of central tendency）**
  - 位置参数（Measures of location）
    - 算术平均数（Arithmetic mean, Average）
    - 中位数（Median）
    - 众数（Mode）
- **离散趋势（Measure of dispersion）**
  - 变异指标（Measure of variability）
    - 全距（极差，Range, R）
    - 四分间距（Interquartile Range）
    - 方差（Variance, S^2）
    - 标准差（Standard deviation, SD）
  - 形状参数（Measures of shape）
    - 偏度（Skewness）
    - 峰度（kurtosis）

形状参数（Measure of shape)

- **偏度（Skewness)**
  - 偏度衡量随机变量概率分布的不对称性，是相对于平均值不对称程度的度量
  - 通过对偏度系数的测量，我们能够判定数据分布的不对称程度以及方向
  - 定义上偏度是样本的三阶标准化矩

$$Skew(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{k_3}{\sigma^3} = \frac{k_3}{k_2^{3/2}}$$

    - 偏度定义中包括正态分布（偏度=0），右偏分布（也叫正偏分布，其偏度>0），左偏分布（也叫负偏分布，其偏度<0）
  - 【注意】数据分布的左偏或右偏，指的是数值拖尾的方向，而不是峰的位置

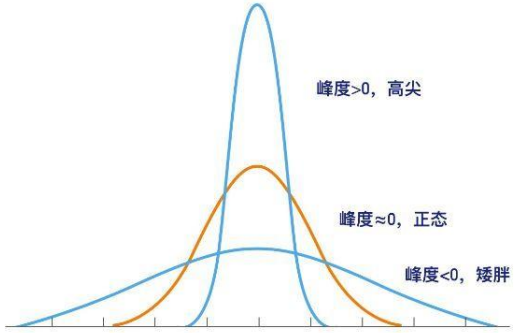


The skewness values can be interpreted in the following manner:

- Highly skewed distribution: If the skewness value is less than −1 or greater than +1.
- Moderately skewed distribution: If the skewness value is between −1 and −½ or between +½ and +1.
- Approximately symmetric distribution: If the skewness value is between −½ and +½.

形状参数（Measure of shape)

- **峰度（Kurtosis)**
  - 峰度，是研究数据分布陡峭或者平滑的统计量
  - 定义上是样本的标准四阶中心矩（standardized 4rd central moment)
  - 峰度计算：随机变量的四阶中心矩与方差平方的比值



  - 这个统计量需要与正态分布相比较，
    - **峰度为0**表示该总体数据分布与正态分布的陡缓程度相同；
    - **峰度大于0**表示该总体数据分布与正态分布相比较为陡峭，为**尖顶峰**；
    - **峰度小于0**表示该总体数据分布与正态分布相比较为平坦，为**平顶峰**

Definition of Kurtosis

For univariate data $Y_1, Y_2, ..., Y_N$ the formula for kurtosis is:

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4}$$

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the kurtosis, the standard deviation is computed using N in the denominator rather than N - 1.

Alternative Definition of Kurtosis

The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

$$\text{kurtosis} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} - 3$$

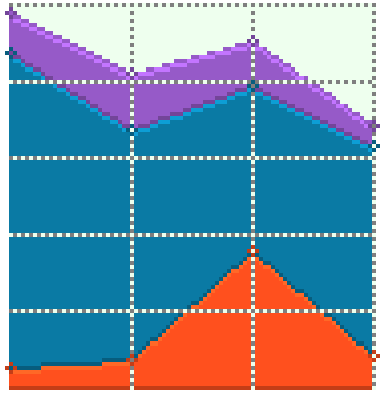
This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "heavy-tailed" distribution and negative kurtosis indicates a "light tailed" distribution.

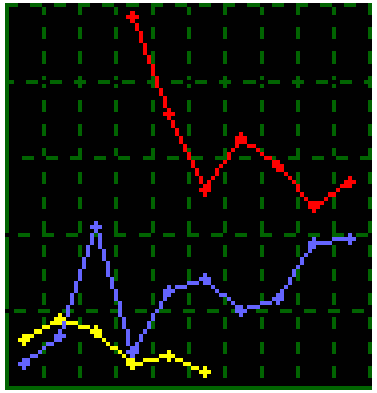
Which definition of kurtosis is used is a matter of convention (this handbook uses the original definition). When using software to compute the sample kurtosis, you need to be aware of which convention is being followed. Many sources use the term kurtosis when they are actually computing "excess kurtosis", so it may not always be clear.

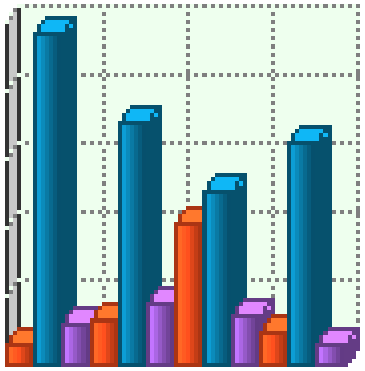
https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

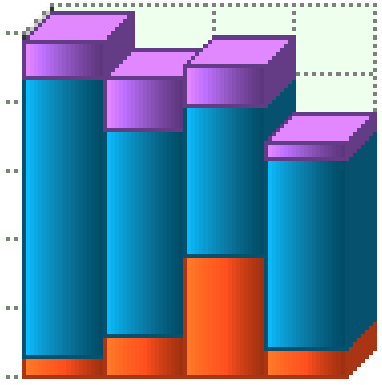
Python 练习

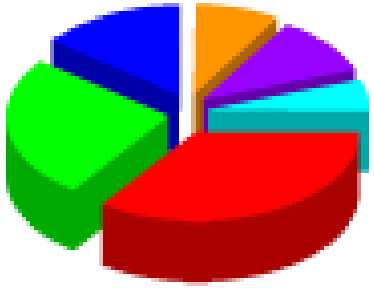
- Box plot
- Histogram
- Dot plot
- Bar chart
- Measure of central tendency

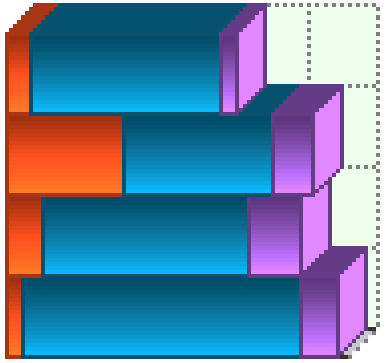












谢谢！