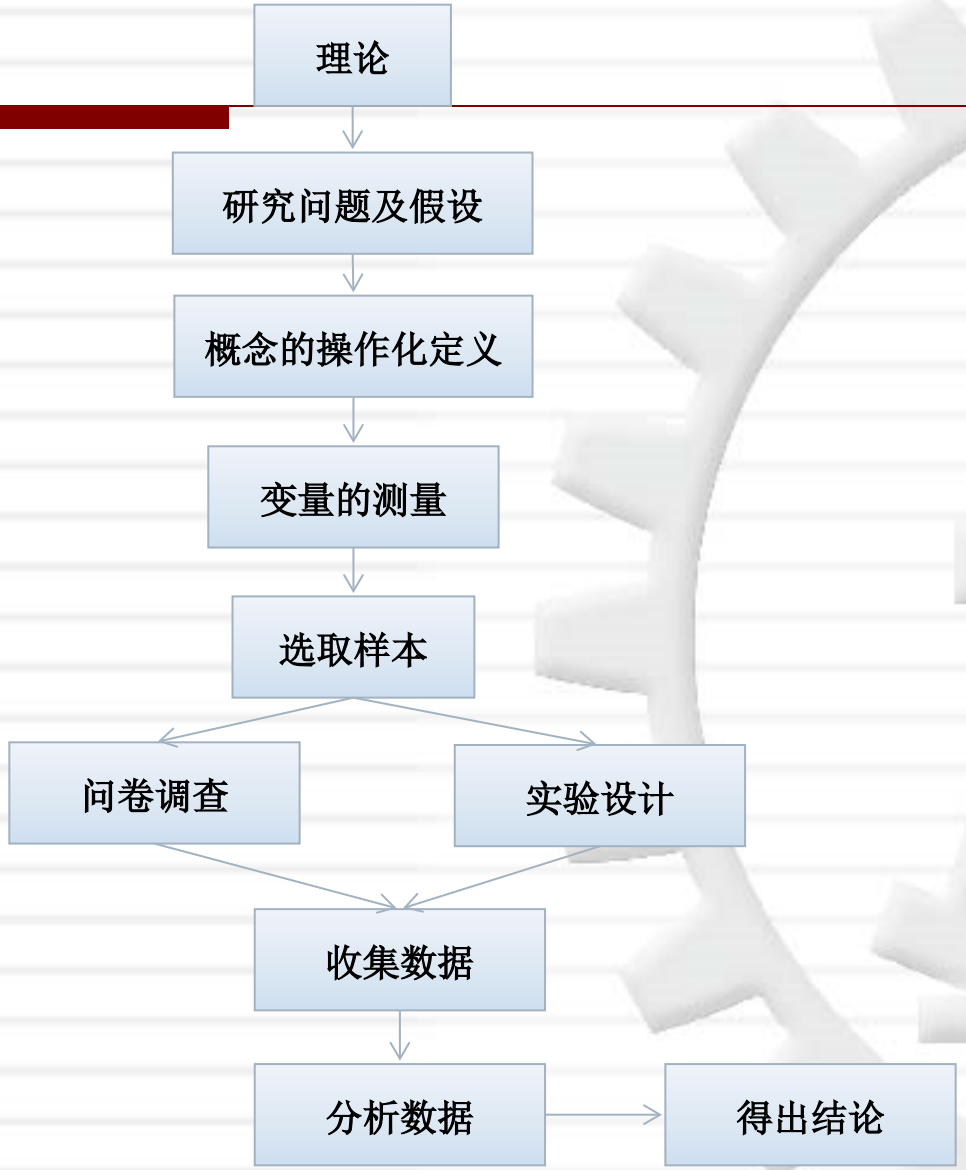
第一讲：记录基本概念及描述性记录

基本研究环节



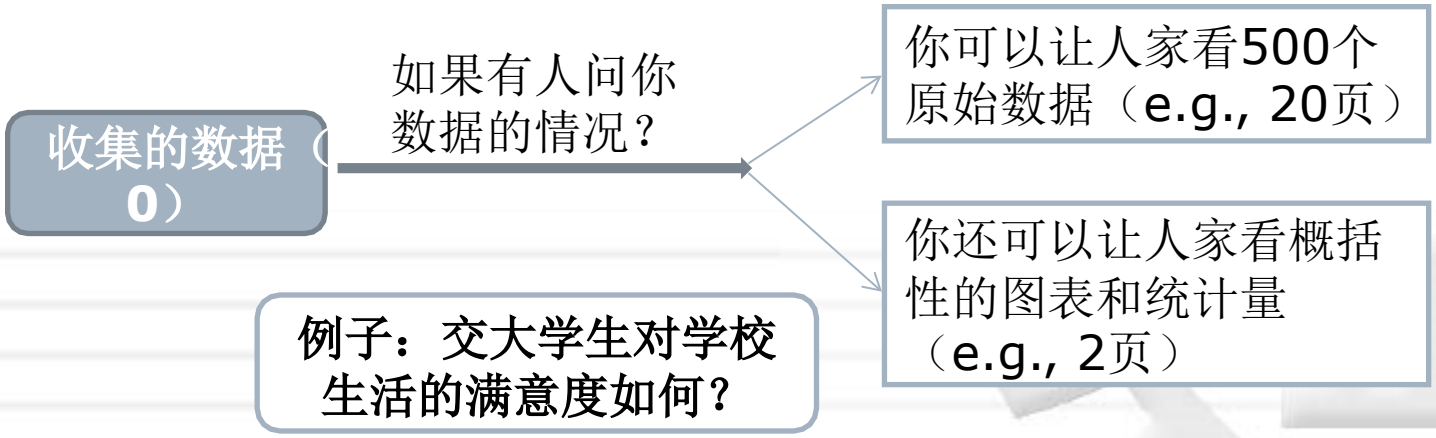
记录分析的分类 & 测量尺度

描述性 vs. 推断性

- 记录学所包括的记录分析可以分为两大类：
- 描述性记录分析 **(Descriptive Statistics)**
- 推断性记录分析 **(Inferential Statistics)**

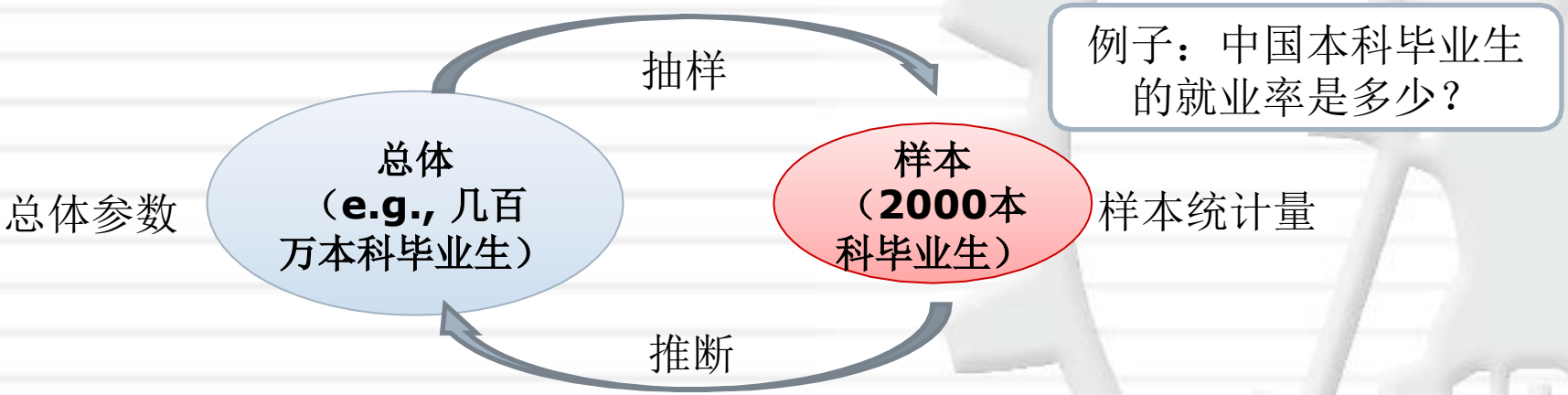
描述性记录

- **描述性记录分析： 通过制表画图及计算记录量等方式，对搜集的数据进行概括、描述、和探索。其目的是用简洁有效的方式去描述复杂繁琐的数据！**



推断性记录

- **推断性记录分析：通过样本特性来推断总体特性。或者说通过已知的样本记录量来推断未知的总体参数。**

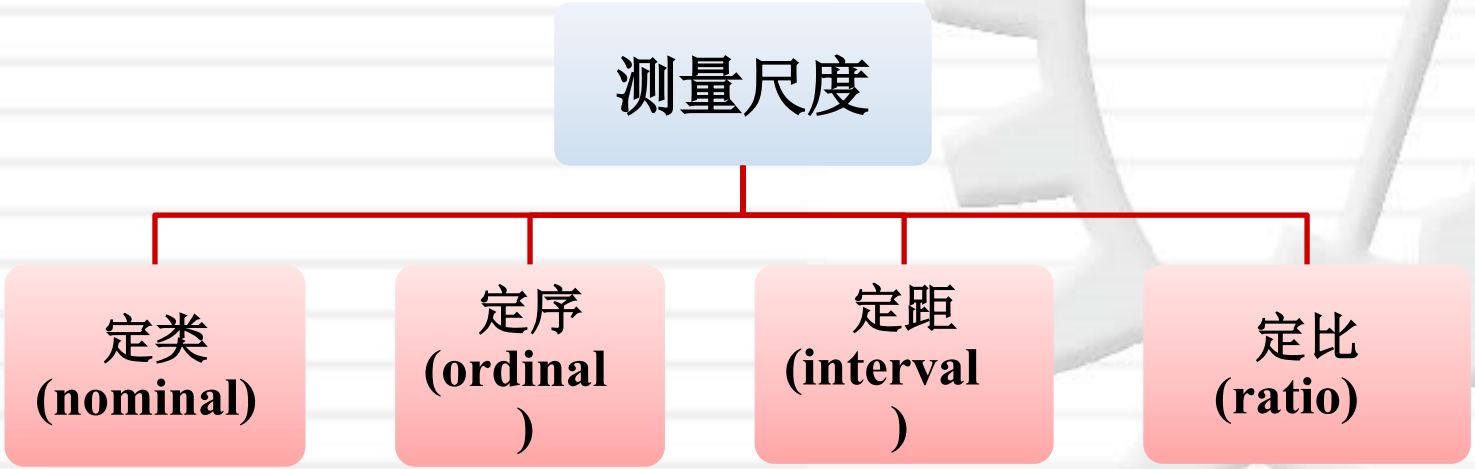


推断性记录

- 计算出的样本记录量（样本的就业比例）是描述性的，然后通过某种措施推断它和真实的总体参数的相似或靠近程度是推断性记录。
- 只需搜集部分数据就可以推断出我们感爱好的总体的特性。这就是记录学的魅力所在！推断性记录的精确性在于样本与否很好地代表总体，以及推断措施的对的性。

测量尺度 (scales of measurement)

- 测量 (measurement)：根据规则，对人或事物的特性用数值来表达。
- 数据有不一样等级的测量尺度，根据测量尺度，才能对的解释变量的赋值。



测量尺度

- 定类 (nominal : giving a name)：
- 等级最低，只是给不一样类别起个名称；
- 类别可以用名字来表达，也可以用数值来表达；
- 数值自身没有实质性意义，仅是一种符号，为了辨别不一样的类别；
- 只具有等于（=）或不等于（≠）的数学特性。

- 经典例子：性别、户口、民族、婚姻状况等
- 男=0， 女=1；（也可以是其他任意数值）
- 男=M，女=F；

测量尺度

- 定序 (ordinal：ordering individuals or objects)：
- 数据体现为“类别”但有序；
- 不一样类别之间有一定的次序；
- 类别的取值反应了排列次序；
- 相邻取值之间不一定是等距的；
- 数学特性：=， ≠，>，<
- 经典例子：
- 教师的职称（讲师=1、副专家=2、专家=3）
- 满意度（非常不满意=1，不满意=2，中立=3，满意=4，非常满意=5）

测量尺度

- 满意度的取值1~5，反应了人们满意度由弱到强的排序，不过相邻数值之间的距离并不是满意度在真实程度上的差异的体现。假如张三选择5，李四选择4，王五选择3，
- 我们懂得张三比李四的满意程度高，不过高多少我们并不懂得。我们也懂得李四比王五的满意程度高，不过高多少我们也不懂得。
- 虽然5和4相差1，4和3也是相差1，但5比4高的程度与4比3高的程度并不一定是相等的。
- 成绩的排名，第一名和第二名也许仅差2分，但第二名和第三名的成绩也许差5分。

测量尺度

- 定距 (interval: equal distance)：
- 数值的大小反应了排列次序；
- 相邻取值之间是等距的；
- 但没有真正意义上的0点 ；
- 可以对它们做加减运算，但不可以做乘除运算。
- 经典例子：温度，年份，成绩等
- 0度并不阐明没有温度；它只是人们把结冰时的温度设置为0度，不是绝对的，而是任意的；
- 25与20度之间相差5度，15度与10度之间也是差5度；（可以说：25度比20度高5度，15度比10度也是高5度）

测量等级

- 定比(ratio: equal distance)：等级最高
- 数值的大小反应了排列次序；
- 相邻取值之间是等距的；
- 有绝对的真正意义上的0点 ；
- 可以对它们做加减乘除运算。
- 经典例子：年龄，身高，体重，收入，子女个数等
- 收入为0就表达没有收入；
- 收入就比收入1500多500，收入1000比500也是多500；
- 收入是收入1000的两倍。

比较

四种计量尺度的比较				
计量尺度 / 数学特性	定类尺度	定序尺度	定距尺度	定比尺度
分类（=，≠）	√	√	√	√
排序（<，>）		√	√	√
间距（+，-）			√	√
比值（×，÷）				√

测量尺度

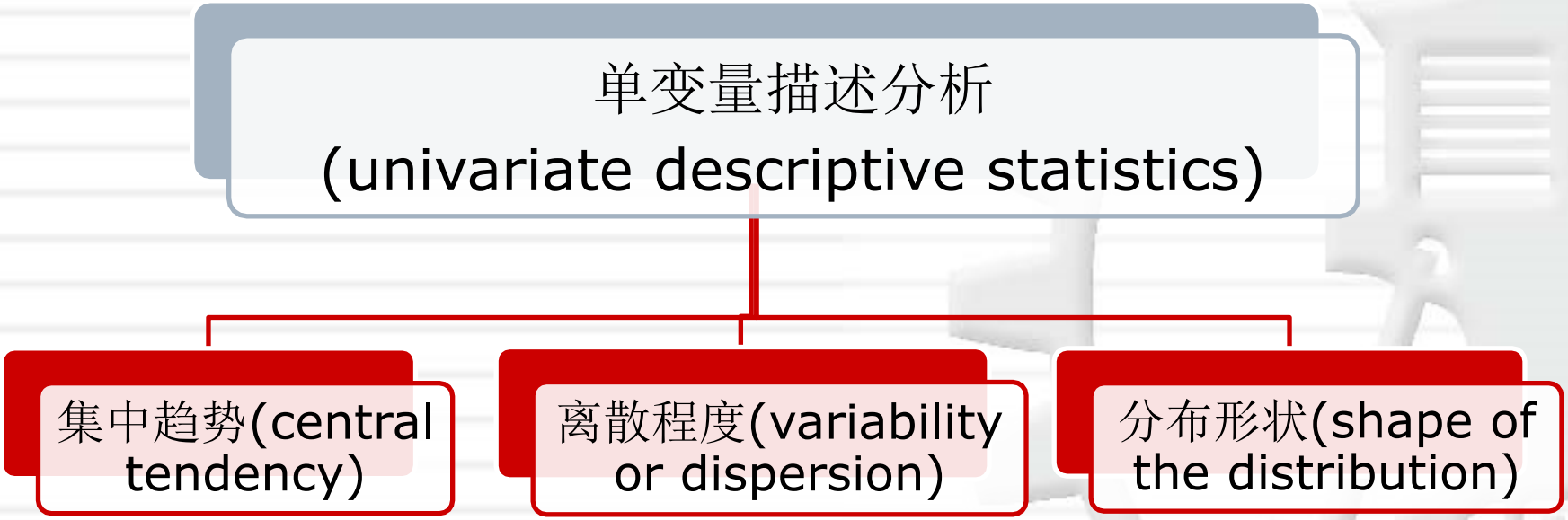
- 注意：
- 在社会学研究中，只满足"定距"而不能满足"定比"规定的变量并不多。因此，在社会学中一般不再辨别定距和定比，而是把它们当作一类，称为"定距"变量。
- 一种变量，它的层次等级并不是唯一的。假如变量是高等级的，它必然可以作为低等级来使用。但减少等级会损失信息量。
- 收入 —> 年薪多少（定距）　高中低收入（定序）
- 年龄—> 多大年龄（定距）　老中青年龄段（定序）

测量尺度的重要性

不一样的记录措施是针对不一样测量尺度的数据的。只有明确了变量的测量尺度，才能对的选择适合的记录分析措施！

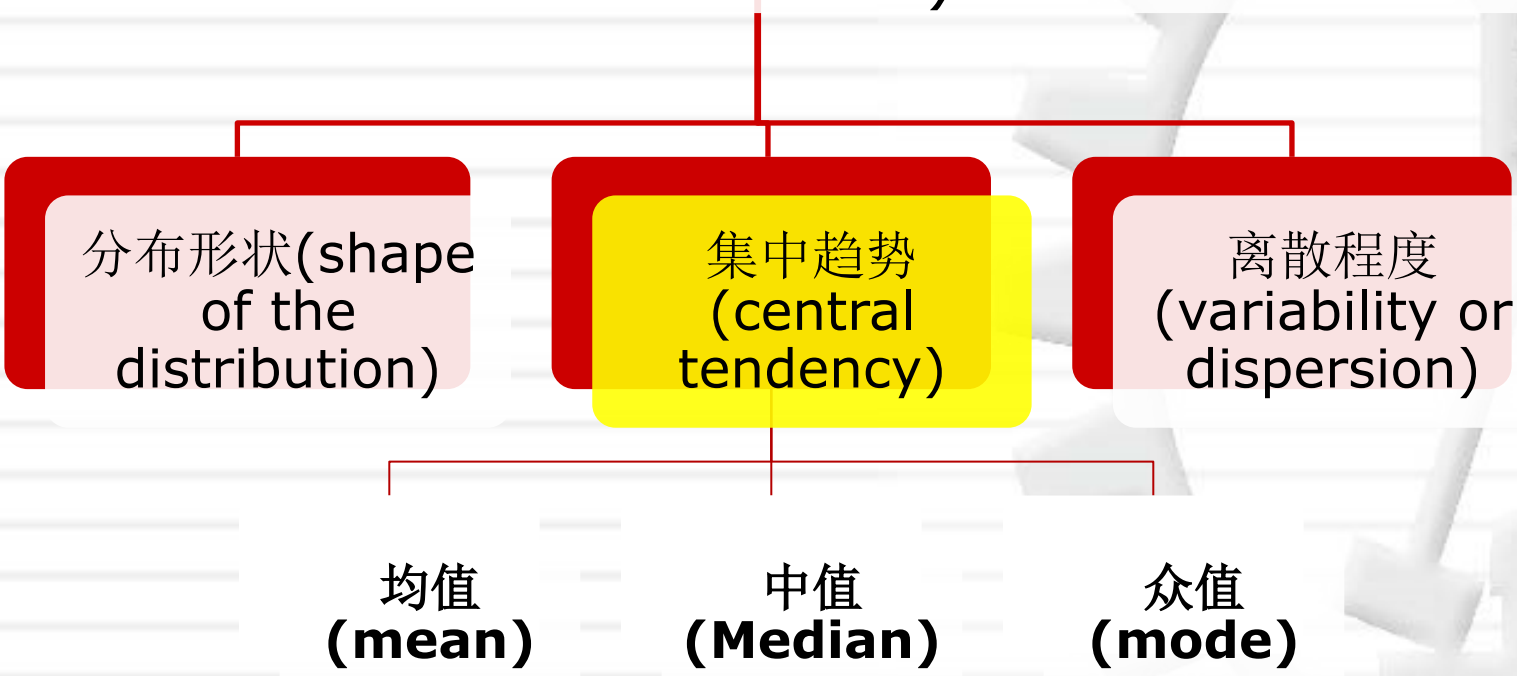
研究问题—>变量类型—>记录分析措施

描述性记录分析


单变量描述分析
(univariate descriptive statistics)
集中趋势(central tendency)
离散程度(variability or dispersion)
分布形状(shape of the distribution)

描述性记录分析——集中趋势描述

单变量描述分析 (univariate descriptive statistics)



集中趋势

- 描述集中趋势的记录量：用一种记录量去描述数据分布的中心位置，又称为"位置记录量"。常用的记录量有：
- 均值(Mean)：数据的算术平均值
- 中位数(Median)：把数据提成50%和50%的数值
- 众数(Mode)：一组数据中的出现次数最多的数值

均值 (Mean)

- 特性：
- 考虑了每个数据，因此增长或减少一种数据，均值就会发生变化；
- 很轻易受极端值(Extreme Values)的影响。比较：
- 1, 3, 5, 7, 9
- 1, 3, 5, 7, 90
- $10200, $10400, $10700, $11200, $11300, $11500

- “均值”适合于描述单峰和基本对称分布的集中趋势；
- “均值”不适合用来描述严重偏态分布的集中趋势。例如，一种国家会因少数富翁的存在，使平均收入变得很高。
- 对严重偏态分布，应使用中值来描述集中趋势。

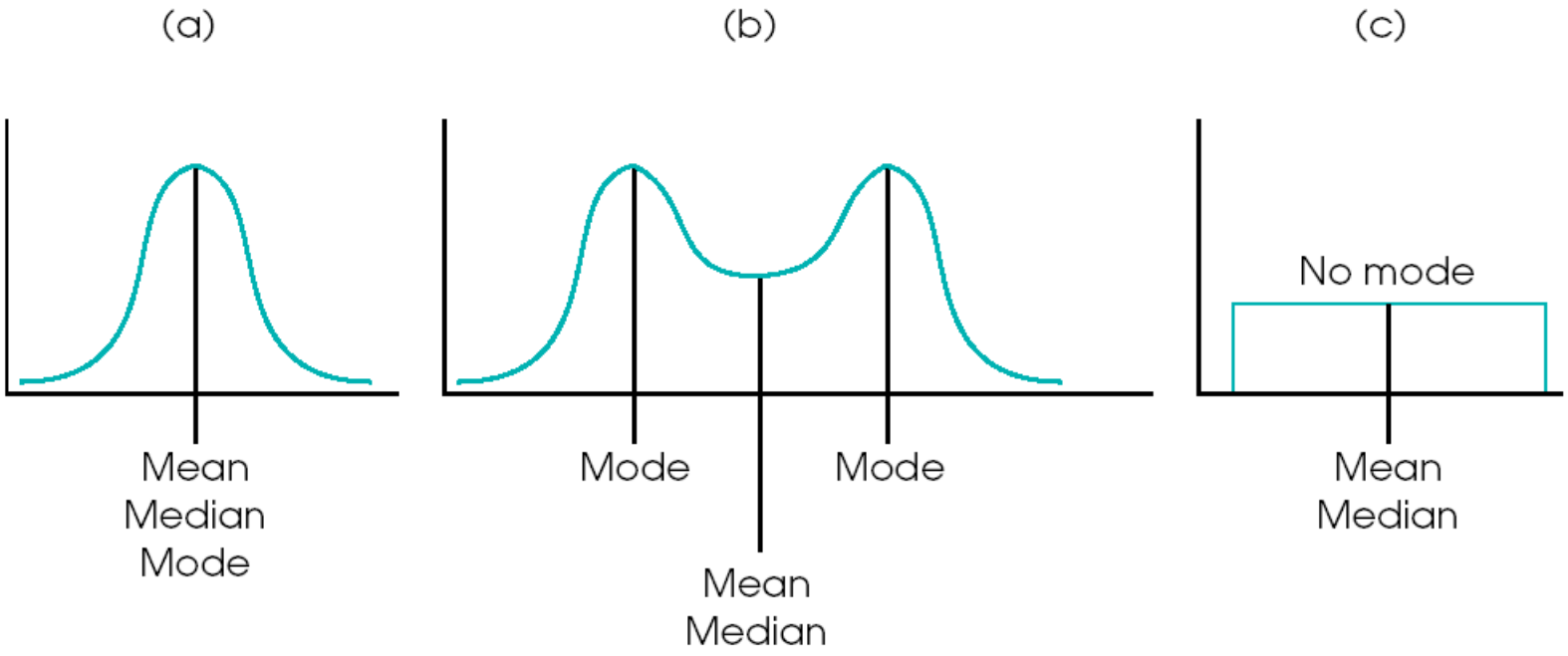
中值 (Median)

- 特性：
- 中值只是考虑了中间位置的数据值，因此仅用中位数描述数据会损失诸多信息。
- 但它受极端值的影响较小，因此对偏度较大的数据（如收入），中位数比均值更能代表数据的中心位置。比较：
- **1, 3, 5, 7, 9**
- **1, 3, 5, 7, 90**

众值 (Mode)

- 特性：
- 中值是一组数据中出现次数最多的数据值。
- 众数不一定唯一，也也许不存在；
- 1, 3, 5, 7, 9
- 1, 1, 3, 3, 7
- 众数不太稳定，数据很小的波动就可以影响到它的值；
- 1, 1, 3, 3, 7
- 1, 1, 3, 3, 7, 1
- 众数是定类数据仅能使用的集中趋势记录量。

三个值的关系

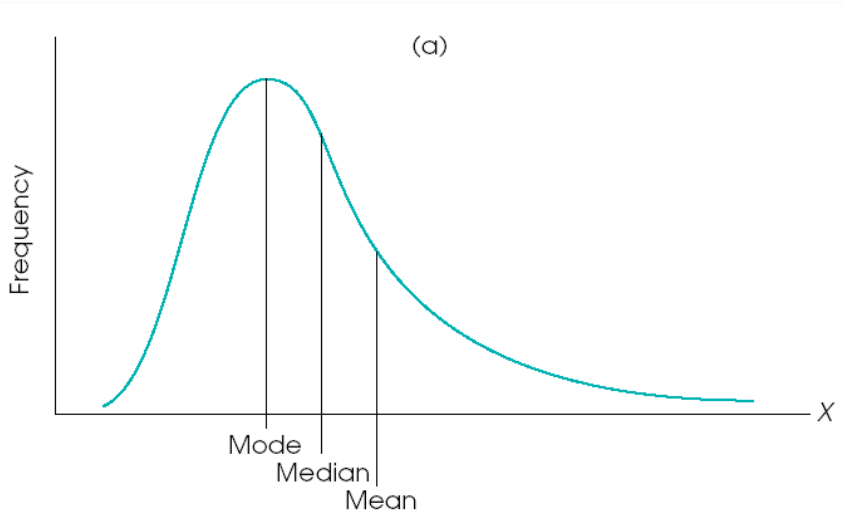


正态分布

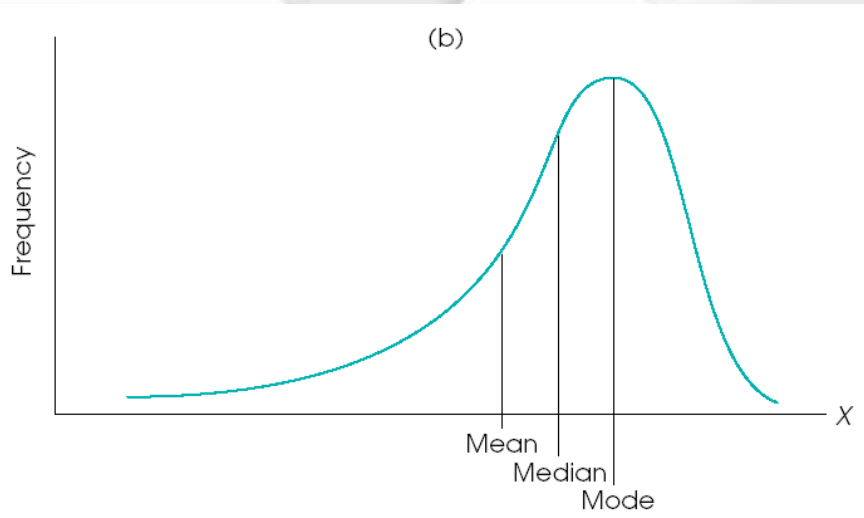
双峰对称分布

矩形分布

三个值的关系



正偏分布



负偏分布

怎样选用这三个值

- 根据变量的测量等级判断：
- 定距：均值、中值、众值
- 定序：中值、众值
- 定类：众值
- 对定距型变量，根据分布的形态判断：
- 对称或靠近对称的分布：均值、中值 （均值也许更好，由于它运用了每个数据）
- 严重偏态分布或存在一定数量的极端值：中值

描述性记录分析——离散程度描述

单变量描述分析
(univariate descriptive statistics)

- 分布形状(shape of the distribution)
- 集中趋势(central tendency)
- 离散程度(variability or dispersion)
  - 异众比率 (variation ratio)
  - 极差 (range)
  - 四分互差 (interquartile range)
  - 方差 (variance) 及标准差 (standard deviation)
  - 离散系数 (variation coefficient)

离散程度

- 离散程度：指一组数据的分散程度或者说数据之间的差异程度。
- 常用的记录量有：
- 异众比率 (Variation ratio) –定类变量
- 全距或极差或范围(Range) –定序/定距变量
- 四分位距或四分互差(Interquartile Range – IQR) – 定序/定距变量
- 方差(Variance) – 定距变量
- 原则差(Standard Deviation) – 定距变量

异众比率

- 当用“众值”来描述数据的集中趋势，“异众比率”表达非众数在总数N中所占的比例：

$$\gamma = \frac{N - f_{mode}}{N}$$

- 当γ=0时，阐明变量只有一种取值，那就是众值；这时，众值可以完全代表变量。
- 当γ->1时，阐明数据非常分散，众值几乎没有代表性。
- 当γ=1时？

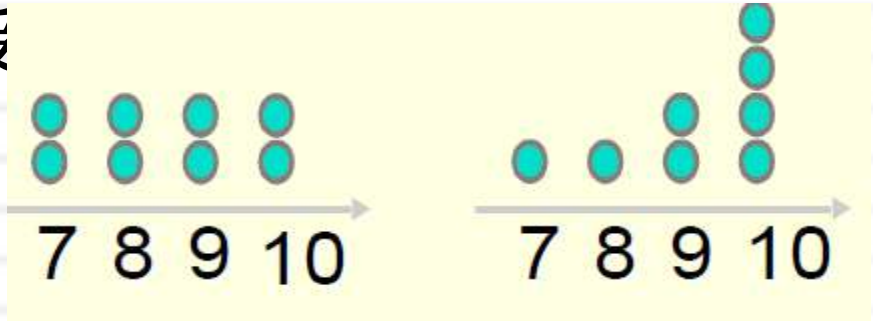
极差

- 一组数据最大值和最小值的差，又称“全距”：

$$R = Max - Min$$

  - 最简单的测量离散程度的统计值
  - 未考虑数据的分布
  - 受极端值的影响很大

- 最简朴的测量离散程度的记录值
- 未考虑数据的分布
- 受



四分位距

- 四分位数(quartiles)：将数据从小到大进行排序，然后分为四等份，处在三个分割点的数据就是四分位数：**Q1 Q2 Q3**
- 四分位距：
- **IQR = Q3 - Q1**
- 测量了中间**50%**的数据的范围，反应了中间**50%**数据的离散程度。
- 长处：**IQR**优于极差和原则差在于它不易受极端值的影响！因此当分布偏度很大或者说有少部分极端值时，适合用**IQR**描述离散程度！

方差和原则差

- 对定距变量，方差和原则差是最常用也是最重要的描述离散程度的措施。
- 反应了各变量值与均值的平均差异。
- 和均值同样，计算方差和原则差需要用到每个数据值。
- 根据总体数据计算的，称为总体方差和原则差；根据样本数据计算的，称为样本方差和原则差。

方差和原则差

- **总体的方差和原则差：**

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \qquad SD = \sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

- **样本的方差和原则差：**

$$s^2 = \frac{\sum(X-\bar{X})^2}{n-1} \qquad SD = s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$$

方差和原则差

- 方差和原则差均不小于等于0；值越大阐明数据越分散；等于0时，数据所有相等，无差异。
- 原则差的单位和原始数据的单位相似，因此，它比方差轻易解释。
- 不能根据原则差来比较不一样变量的离散程度，由于原则差和原始数据的尺度有关，比较：
- 100、200、300　（SD=100）
- 10、20、30　（SD=10）

离散系数

- 数据的标准差与其相应均值之比
- 也称为“变异系数”
- 测量了数据的相对离散程度
- 用于对不同样组别数据离散程度的比较
- 计算公式为：
  - $V_\sigma = \frac{\sigma}{\mu}$　　$V_s = \frac{S}{\bar{X}}$

离散系数

【例3.16】某管理局抽查了所属的8家企业，其产品销售数据如表3-16。试比较产品销售额与销售利润的离散程度

表3-16　某管理局所属8家企业的产品销售数据

企业编号	产品销售额（万元）X_1	销售利润（万元）X_2
1	170	8.1
2	220	12.5
3	390	18.0
4	430	22.0
5	480	26.5
6	650	40.0
7	950	64.0
8	1000	69.0

离散系数

$\overline{X}_1$=536.25（万元）　　$\overline{X}_2$=32.5215（万元）

S_1=309.19（万元）　　S_2=23.09（万元）

$$V_1=\frac{309.19}{536.25}=0.577 \qquad V_2=\frac{23.09}{32.5215}=0.710$$

结论： 计算结果表明，$V_1<V_2$，说明产品销售额的离散程度小于销售利润的离散程度

描述性记录分析——分布形态描述