# 探索性数据分析原理简介

## 变量类型

- **范畴变量** (Categorical)
  - **名义变量** (Nominal)： 没有逻辑顺序的标签（例如头发颜色）
  - **序数变量** (Ordinal)： 为标签，但数据可以排序（例如幸福水平）
- **数值变量** (Numeric)
  - **比率变量** (Ratio)： 具有有意义的零点（例如身高、收入、距离），可以做除法。
  - **区间变量** (Interval)： 具有任意定义的零点（例如温度），除法无意义。

### 范畴变量 (Categorical)

- 对于名义变量，每个结果都代表着质的不同。结果具有无法比较的名称或标签，并且没有相对顺序。
- 序数变量是定量数据，因为它可以有序，并允许在观测值（例如小、中、大）之间进行逻辑比较。

注意：**Likert 表** (Likert scales) 为 序数变量 而非 数值变量。

| 幸福指数 | 不幸福 | 一般 | 很幸福 |
| --- | --- | --- | --- |
| $\alpha$ | 0 | 1 | 2 |

上述 **Likert 表** 仅仅只是为范畴赋值，不改变数据类型，仍为序数变量。

我们不能对分类变量使用算术！“小 + 中 = ...” 没有意义。

范畴变量通常有许多主体，这些主体对该变量的观测值相同。每个观测值发生的 **频率** 是我们感兴趣的。

ex: 泰坦尼克号各层人数数据

| 1st | 2nd | 3rd | Crew | Sum |
|---|---|---|---|---|
| 325 | 285 | 706 | 885 | 2201 |
| 14.8 % | 12.9 % | 32.1 % | 40.2 % | 100.0 % |

我们通过计算每个唯一值的出现次数来制表表示 **范畴数据**。

以百分比形式呈现数据通常提供更多信息，但它确实隐藏了样本量。

## 数值变量 (Categorical)

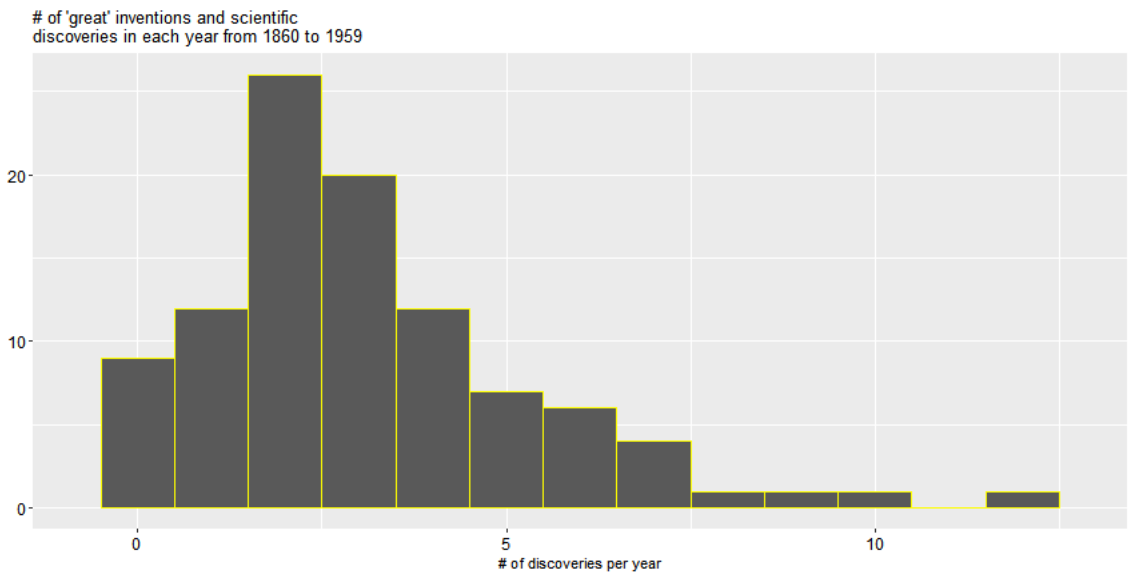
数值变量给出可以直接解释的实数。15 公里比 10 公里远 5 公里。15 摄氏度比 10 摄氏度高 5 度。

它们可以是 **离散** 的，也可以是 **连续** 的（例如，房间里的人数、考试正确答案的数量、一个人的身高、完成任务的时间）。

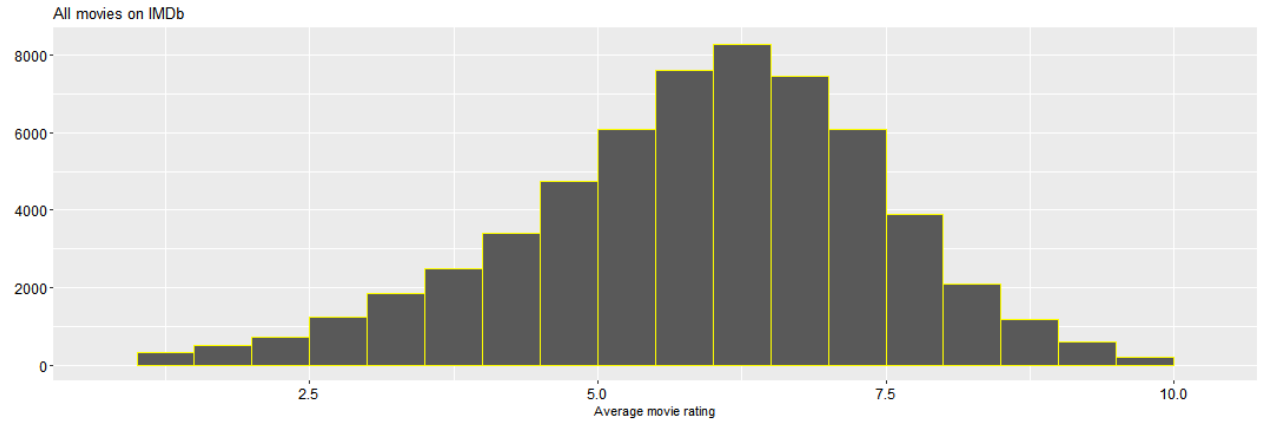
比率变量具有有意义的零，这允许解释 **相对** 值。例如，15 公里是 5 公里的 3 倍。

比率对于区间变量没有意义：15 摄氏度不是 5 摄氏度的“三倍”。

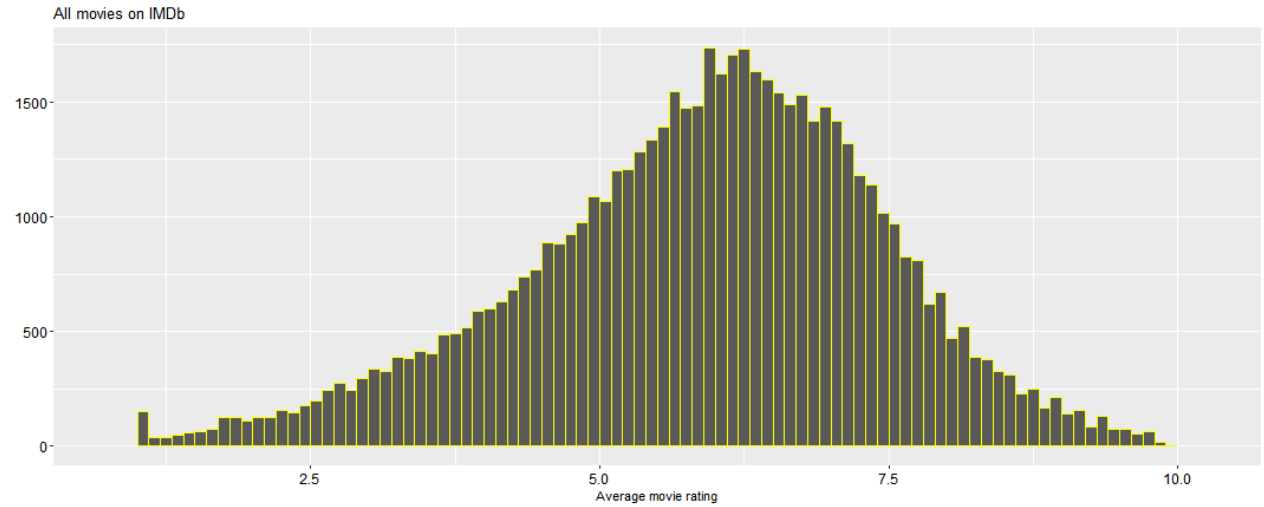
- 对于值相对较少的 **离散数值** 变量，使用 **条形图**。




- 对 **连续** 的数据，我们可以使用 **直方图** 创建小区间并计算每个区间中包含的观测值数。




注意：bin 的宽度选取很重要，即分隔连续型变量的间隔，bin越小，图像越密。

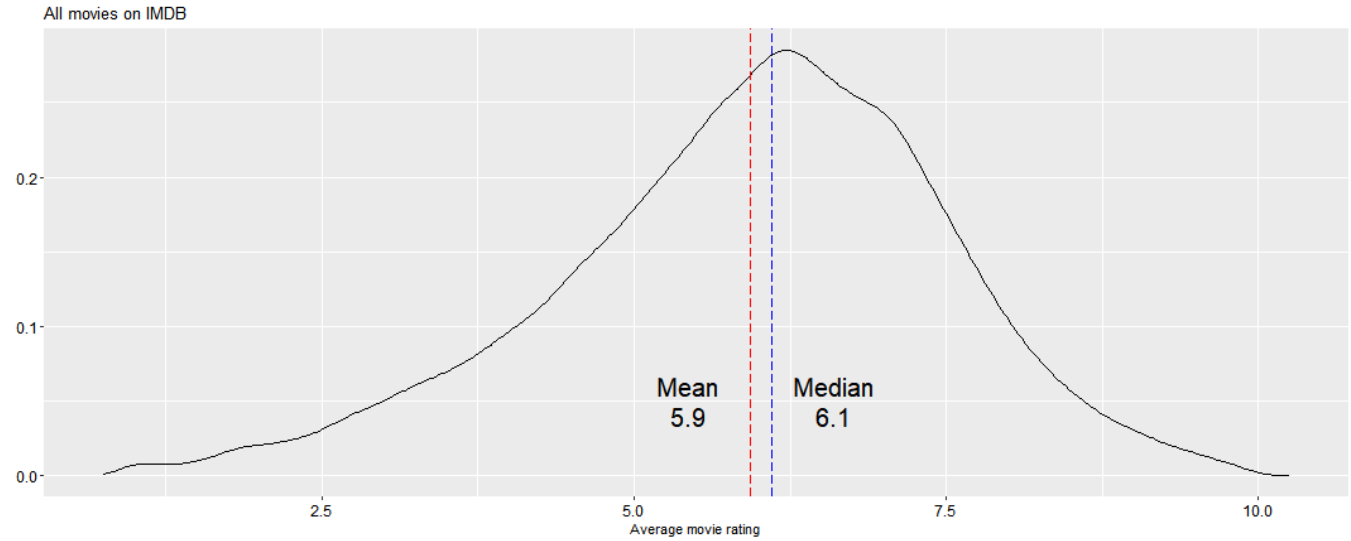



- 可以放大 bin 的宽度，得到一个大致的分布图像，而省略精细的结构。
- 也可以让 bin 趋于0，得到一条连续的 **概率曲线。**


**数值变量的特征数**

- **均值** (Mean) 常见的度量为 **样本均值** $\bar{x} = \frac{\sum_i x_i}{n}$

- **中位数** (median) 对所有数据排序后取得的中间数。




不能全局分析的情况