概率论与数理统计概览

讲师：Jeary

目录

目标

通过本章课程的学习，您将能够：

- 了解统计学的相关基本概念
- 掌握描述统计学重要方法

统计学简介

统计学简介

统计学的重要性

1. 统计学，是数据分析师**必备**的基础知识！！！

2. 统计学是在资料分析的基础上，研究测定、收集、整理、归纳和分析反映数据资料，以便给出正确消息的科学。

3. 随着大数据时代来临，统计的面貌也逐渐改变，与信息、计算等领域密切结合，是数据科学中的重要主轴之一。

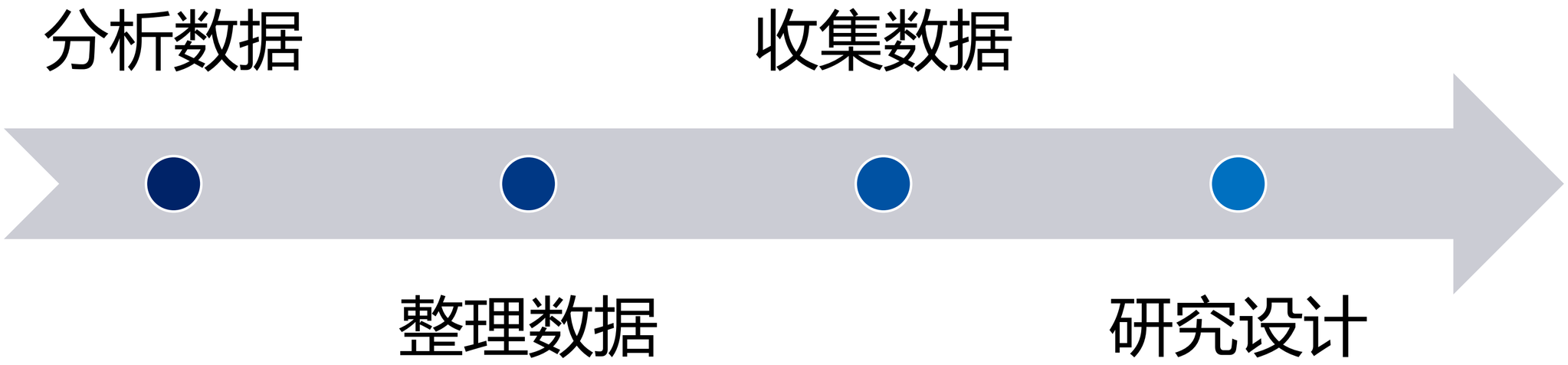
统计学简介

统计学与我们的生活有什么联系?

工作和生活中都会有大量随机现象出现:

统计学的任务就是**找到随机现象的发生规律**，从而将不确定性事件变为可估计、可预测的确定性结果。

统计学研究步骤



统计学的意义

统计学对于数据分析工程师的意义？有什么帮助？

统计学是数据分析/机器学习工程师**必备的**先决条件！

统计学基本概念

基本概念、方法

统计学的基本概念

总体、个体、样本、变量、随机抽样、频率、概率

统计学研究方法

描述统计学、推断统计学

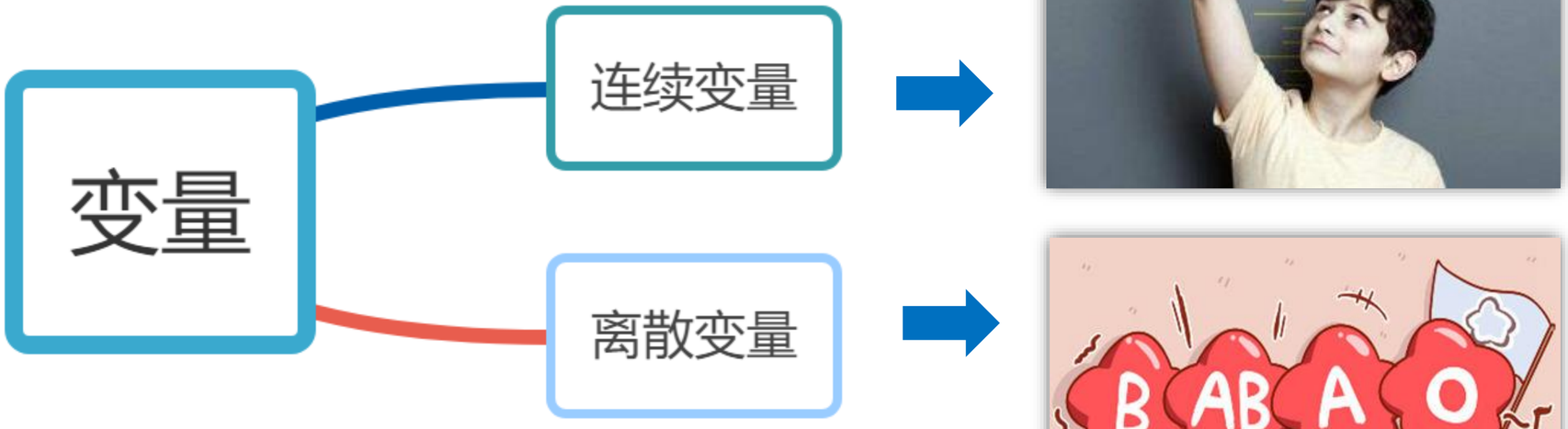
总体 & 个体

总体

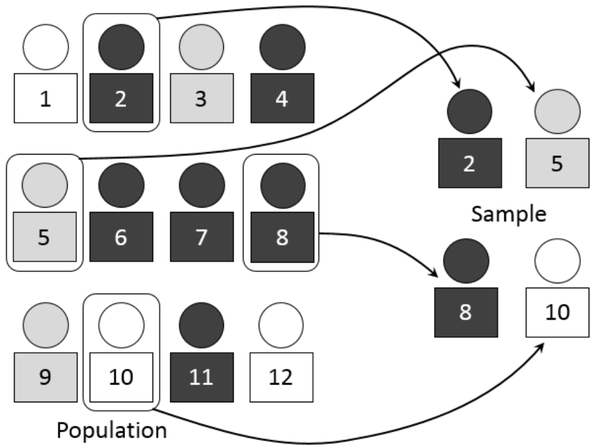
个体

样本

变量



随机抽样



频率、概率

- **频率**：某个事件出现的次数除以总的次数
- **概率**：刻画随机事件发生可能性大小的指标，概率的取值介于0~1之间

描述统计学

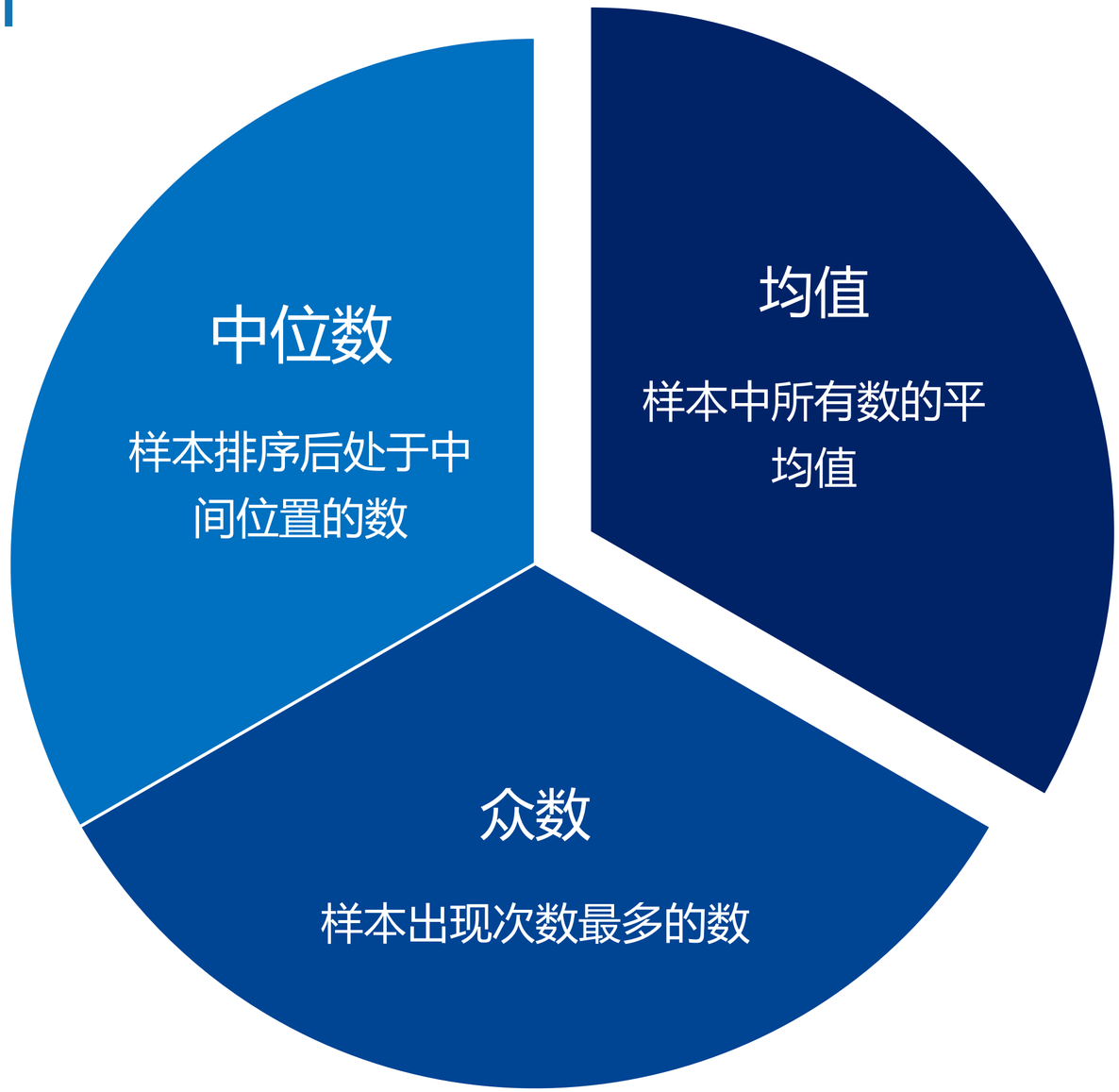
描述统计学

概念

通过图表或数学方法，对数据资料进行整理、分析，并对数据的分布状态、数字特征和随机变量之间关系进行**估计**和**描述**的方法。

分类

集中趋势分析



集中趋势分析

案例

报名**数据分析师**微专业的学员年龄：

20，23，24，25，26，25，27，30

总数：200　均值：25　众数：25　中位数：25

对比

均值

优点：充分利用所有数据，适用性强

缺点：容易受到极端值的影响

中位数

优点：不受极端值影响

缺点：缺乏敏感性

众数

优点：代表性好

缺点：缺乏唯一性

思考题

01 如果我们想用平均值描述数据，为了让结果描述的准确，我们如果操作?

02 众数的缺点是?

离中趋势分析简介

概念

指一组数据中各数据值以不同程度的距离偏离其中心（平均数）的趋势，又称**标志变动度**

离中趋势分析主要靠全距、四分差、平均差、方差、协方差、标准差等统计指标来研究数据的离中趋势

离中趋势分析方法

极差或全距（Range）： 数列 X 中最大值与最小值之间的差值，用于描述X的数字分散程度，越小则数字之间越紧密

中程数（Midrange）： 数列 X 中（最大值 + 最小值）/ 2

四分位：所有观测值从小到大排序后四等分，其中第L_1处的值记为 Q_1，第L_2处的值记为 Q_2，第L_3处的值记为Q_3，公式如下，其中N为样本量:

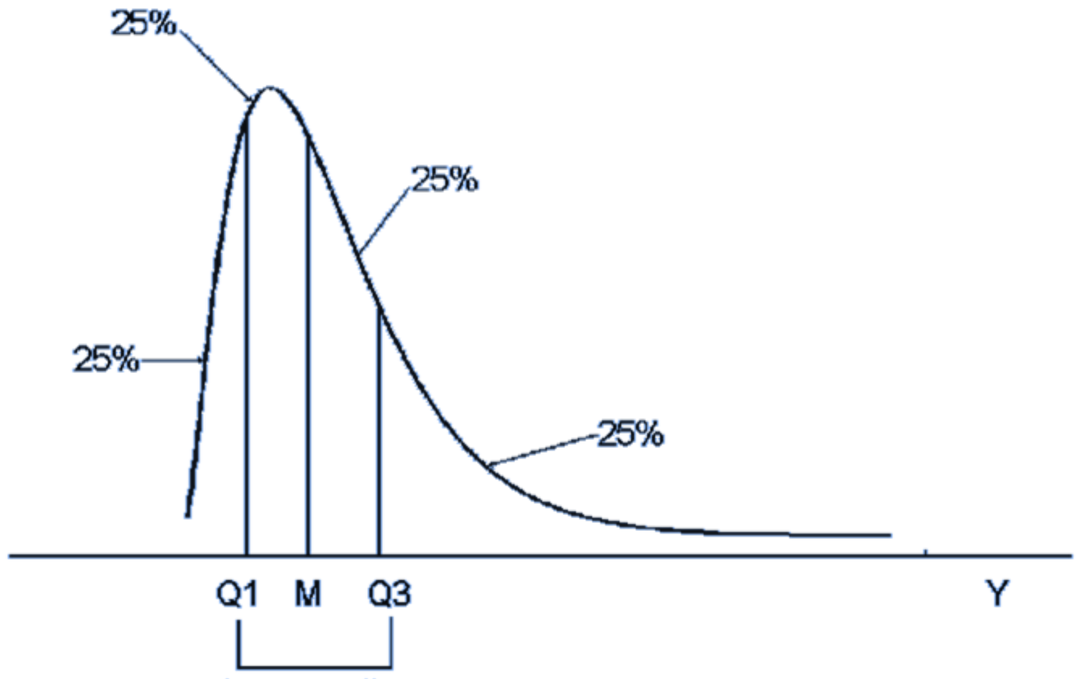
$$L_1 = (N+1) / 4$$

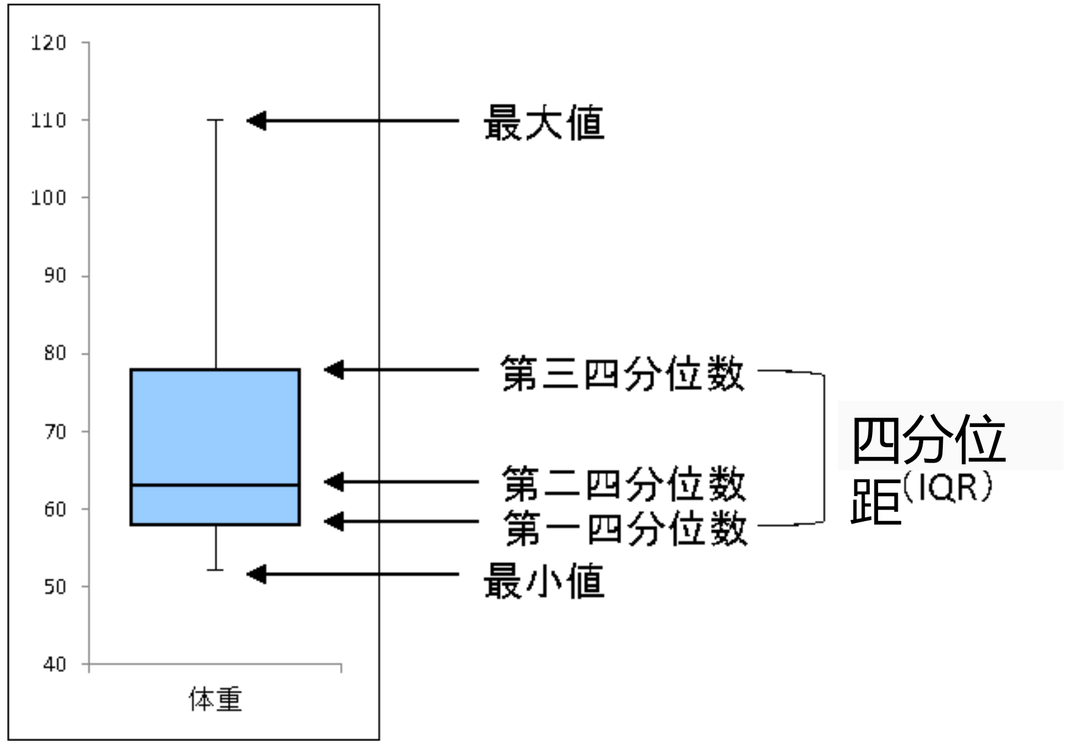
$$L_2 = 2(N+1) / 4$$

$$L_3 = 3(N+1) / 4$$

- 如果L是一个整数，则取第L_1和第L_1+1的平均值
- 如果L不是一个整数，则取下一个最近的整数（例如$L_1=1.5$， 则取2 ）

离中趋势分析方法





异常值（Outlier） $> Q_3 + 1.5$ (IQR) 或

异常值（Outlier） $< Q_1 - 1.5$ (IQR)

$$IQR = Q3 - Q1$$

离中趋势分析方法

■ 平均差（Mean Deviation）

各个变量值同平均数的离差绝对值的算术平均数

■ 方差（Variance）

指每个样本值与全体样本值的平均数之差的平方值的平均数

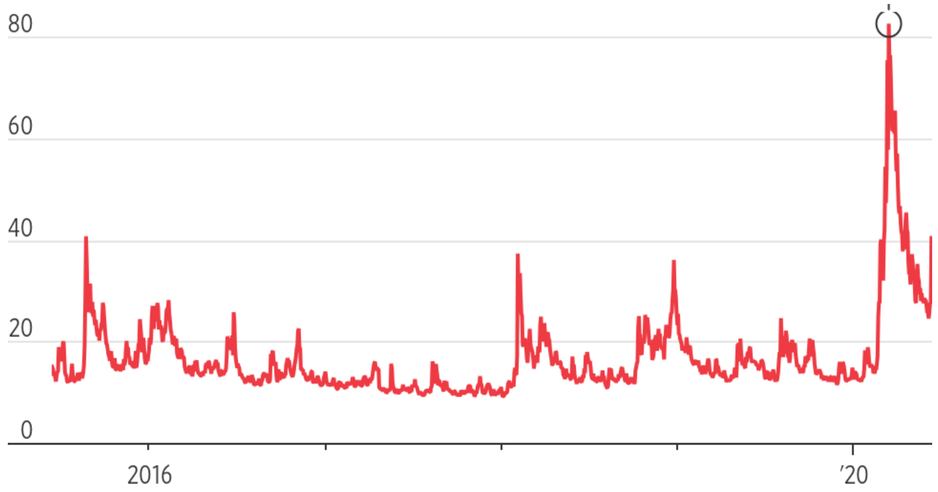
■ 标准差(Standard variance)

方差的平方根

方差

概念

衡量随机变量或一组数据时离散程度的度量

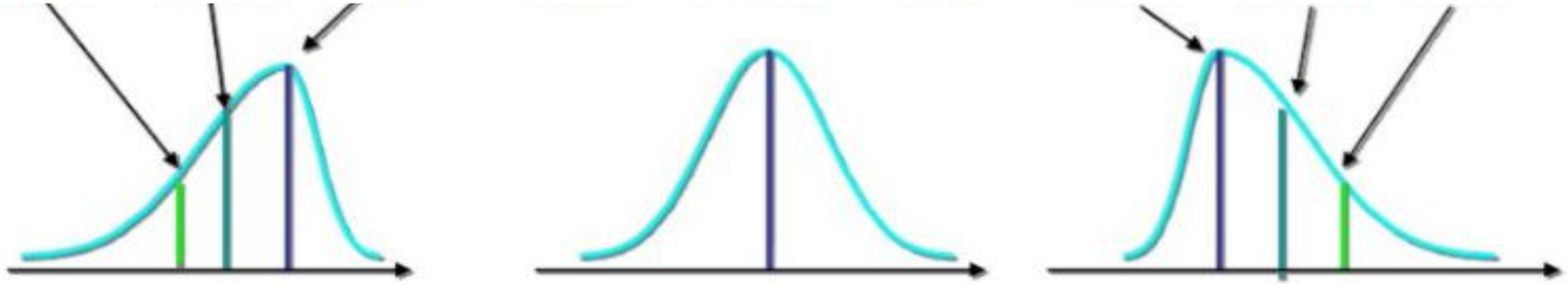


$$s^2 = \frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$$

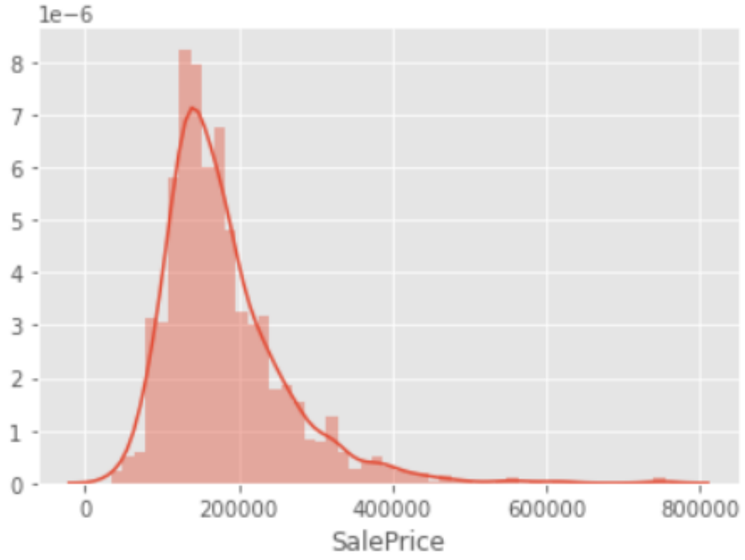
$$s = \sqrt{\frac{1}{n}[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]}$$

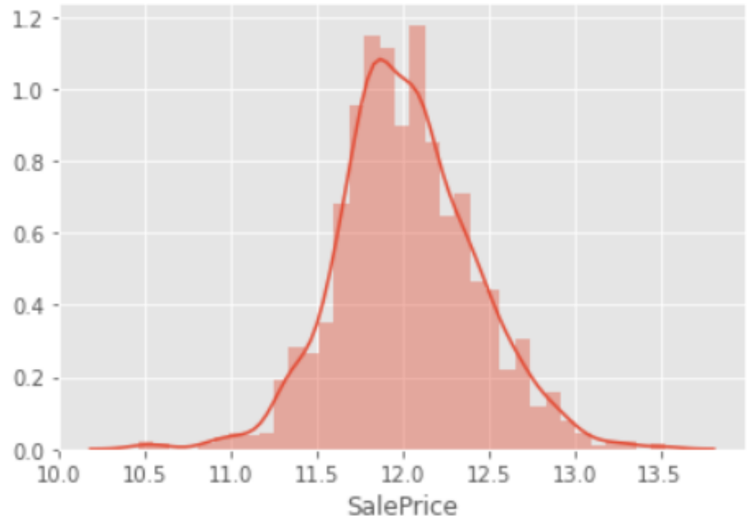
偏态

均值　中位数　众数　　均值 = 中位数 = 众数　　众数　中位数　均值



左偏分布　　对称分布　　右偏分布





偏度计算

$$\mu = EX$$

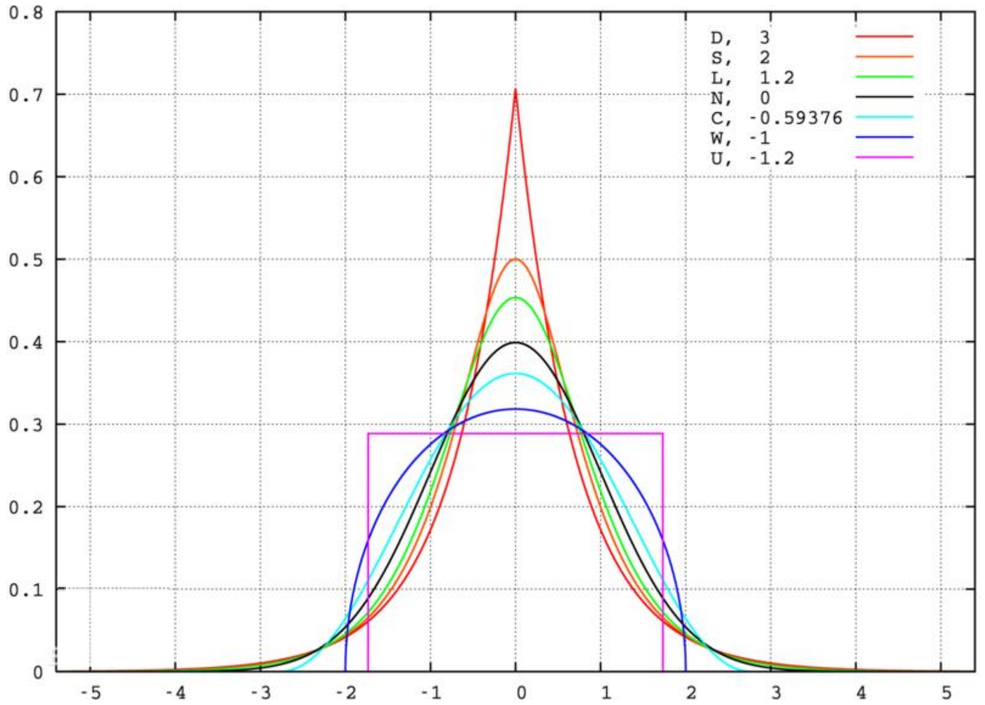
$$\sigma^2 = E(X - EX)^2 = EX^2 - E^2X = EX^2 - \mu^2$$

$$\begin{aligned} Skew(X) &= E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \\ &= \frac{E\left[(X-\mu)^3\right]}{\sigma^3} \\ &= \frac{EX^3 - 3\mu EX^2 + 3\mu^2 EX - \mu^3}{\sigma^3} \\ &= \frac{EX^3 - 3\mu EX^2 + 2\mu^3}{\sigma^3} \\ &= \frac{EX^3 - 3EXEX^2 + 2E^3X}{(EX^2 - E^2X)^{3/2}} \end{aligned}$$

```
#转换后，偏度skewness and 峰度kurtosis计算
print("Skewness: %f" % data['SalePrice'].skew())
print("Kurtosis: %f" % data['SalePrice'].kurt())
```

```
Skewness: 0.121347
Kurtosis: 0.809519
```

峰度



$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\text{StDev}}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

```
#转换后，偏度skewness and 峰度kurtosis计算
print("Skewness: %f" % data['SalePrice'].skew())
print("Kurtosis: %f" % data['SalePrice'].kurt())
```

```
Skewness: 0.121347
Kurtosis: 0.809519
```

思考题

01 偏度的作用是什么?

02 集中趋势的方法都有哪些?

03 离中趋势的方法都有哪些？检测异常值可能会用到哪个方法?

总结

总结

本章包含三小节内容：

统计学简介

- 了解了学习统计学的重要性

统计学基本概念

- 学习了统计学的研究方法、基本概念

描述统计学

- 学习了趋势分析的方法

思考题

01 统计学的基本概念：总体、样本、变量、随机事件、频率?

02 集中趋势的方法和作用?

03 离中趋势的方法和作用?

谢谢观看