第三章 数据处理及分析

Lianghai Xiao

https://github.com/styluck/mlb

作业邮箱：alswfx@126.com

目录

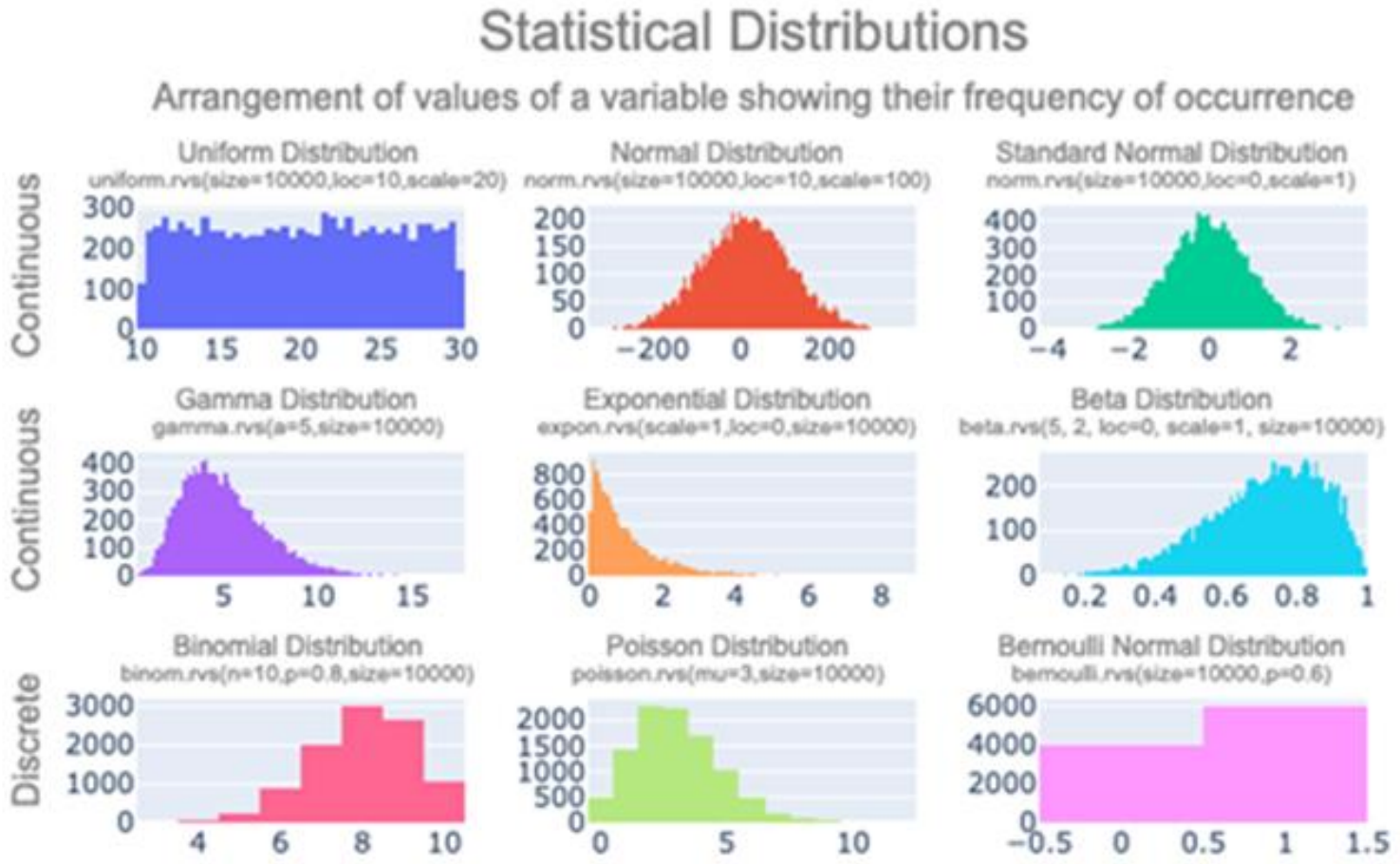
描述性统计量与参数分布估计

描述性统计量

- 描述性统计量是用于**概括和总结数据集的重要统计指标**。
- 描述性统计量的作用：
  - **帮助理解数据：**通过描述性统计量，可以快速了解数据的集中趋势、离散程度等特征，对数据有一个初步的认识。
  - **数据简化：**将大量的数据概括为几个关键的统计量，便于进行数据分析和比较。
  - **异常值发现：**通过观察描述性统计量中的极端值，可以发现数据中的异常值，以便进行进一步的检查和处理。
  - **分析准备：**描述性统计量可以为后续的推断统计分析提供基础，例如在进行假设检验、回归分析等之前，通常先进行描述性统计分析。

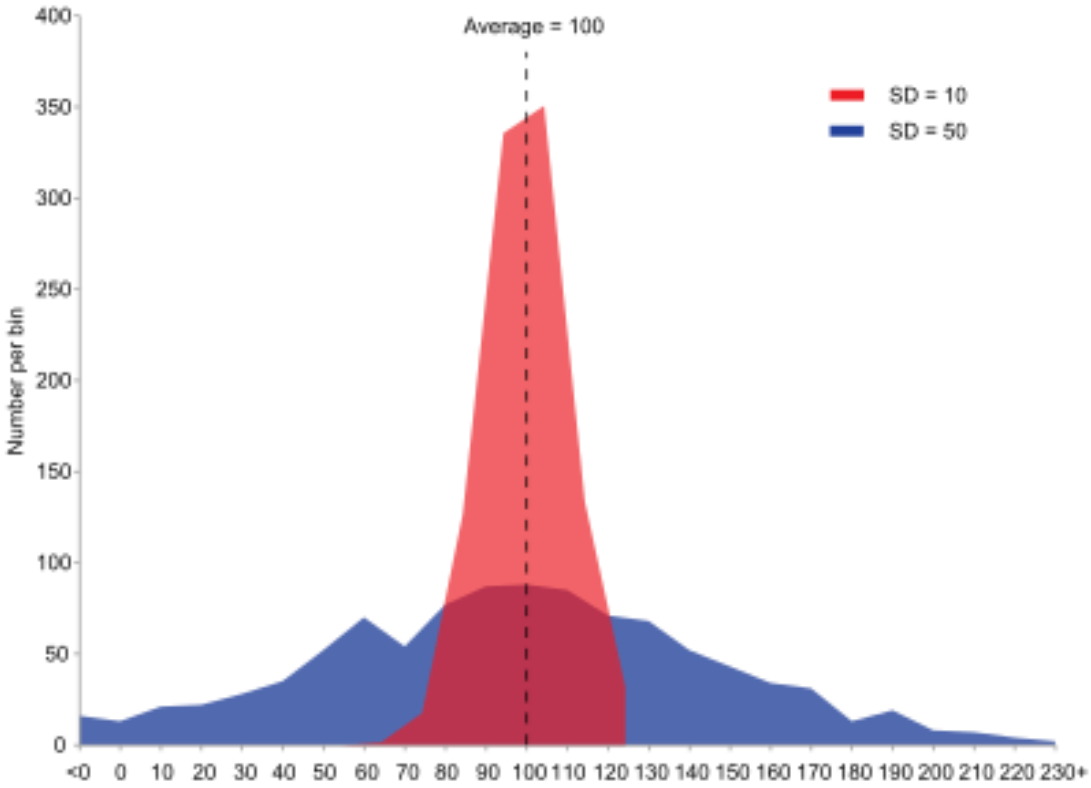
描述性统计量

- 描述性统计量是用于**概括和总结数据集的重要统计指标**。



常见的描述性统计量

- **集中趋势**:
  - **均值**: 数据的平均值。
  - **中位数**: 将数据排序后处于中间的位置。
  - **众数**: 出现频率最高的数据值。
- **离散程度**:
  - **方差**: 数据偏离均值的平方和的平均值。
  - **标准差**: 方差的平方根。
  - **极差**: 数据集中最大值和最小值之差。
  - **变异系数**：衡量变量观测值变异程度。

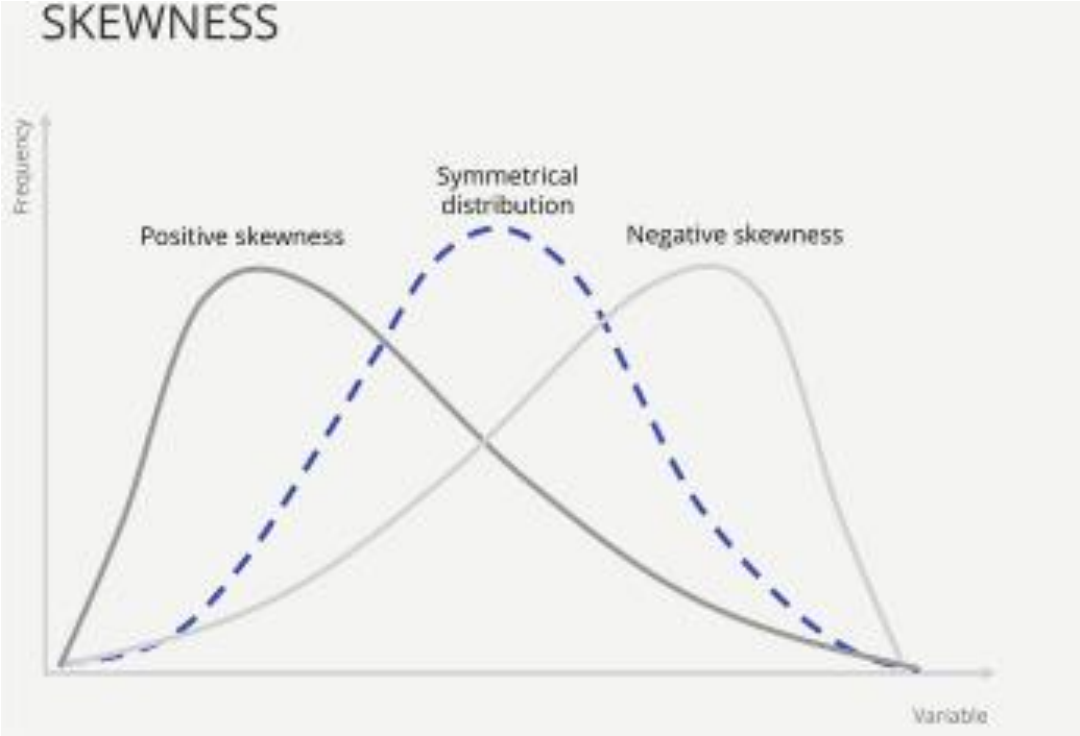


常见的描述性统计量

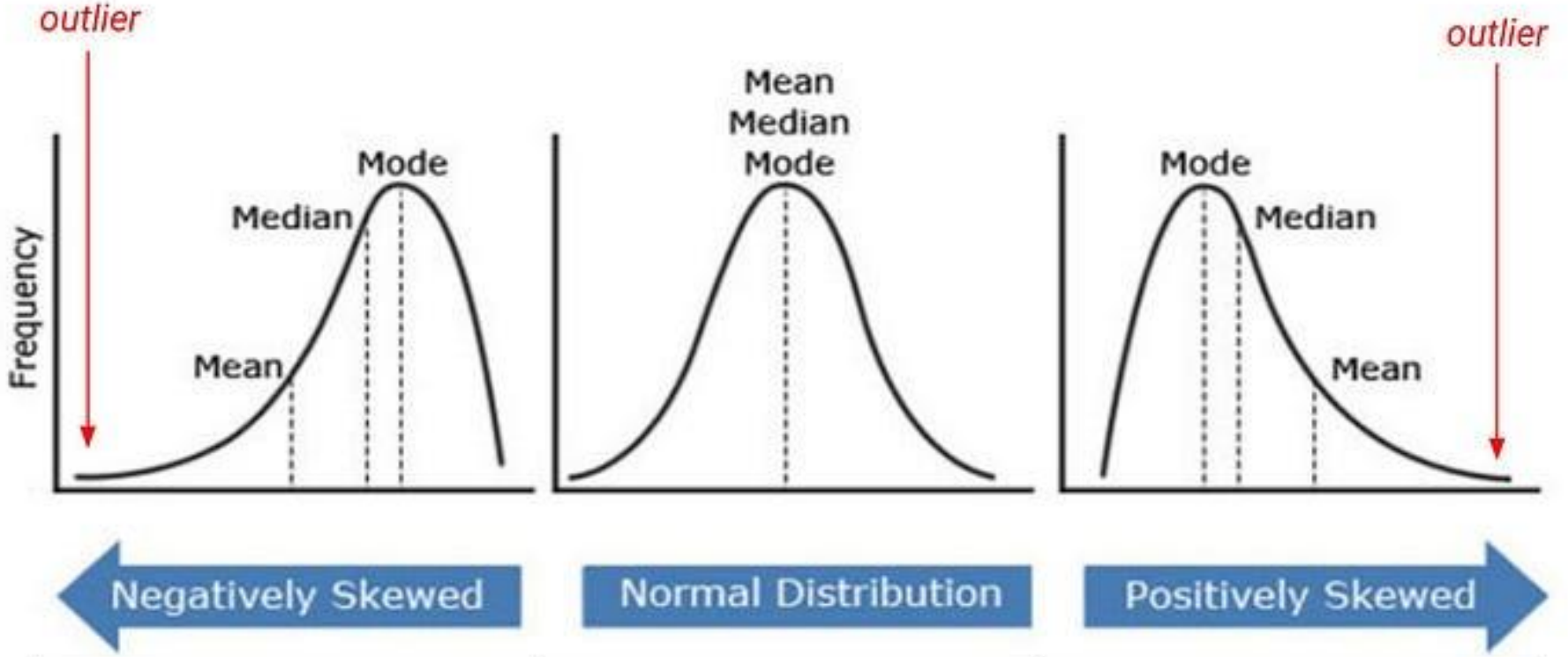
- 1. 方差 (Variance):
  - $\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- 2. 标准差 (Standard Deviation):
  - $\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$
- 3. 极差 (Range):
  - $\text{Range} = x_{\text{max}} - x_{\text{min}}$
- 4. 变异系数 (Coefficient of Variation, CV):
  - $\text{CV} = \frac{\sigma}{\bar{x}} \times 100\%$

常见的描述性统计量

- 5. 偏度 (Skewness):
- 偏度是衡量分布不对称性的统计量。它反映了数据分布相对于均值的偏斜程度。如果分布偏向左侧或右侧，则称为左偏或右偏。
- $\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3$
- - 偏度为正时，表示分布右偏，右侧尾巴较长。
- - 偏度为负时，表示分布左偏，左侧尾巴较长。
- - 偏度为0表示分布大致对称。



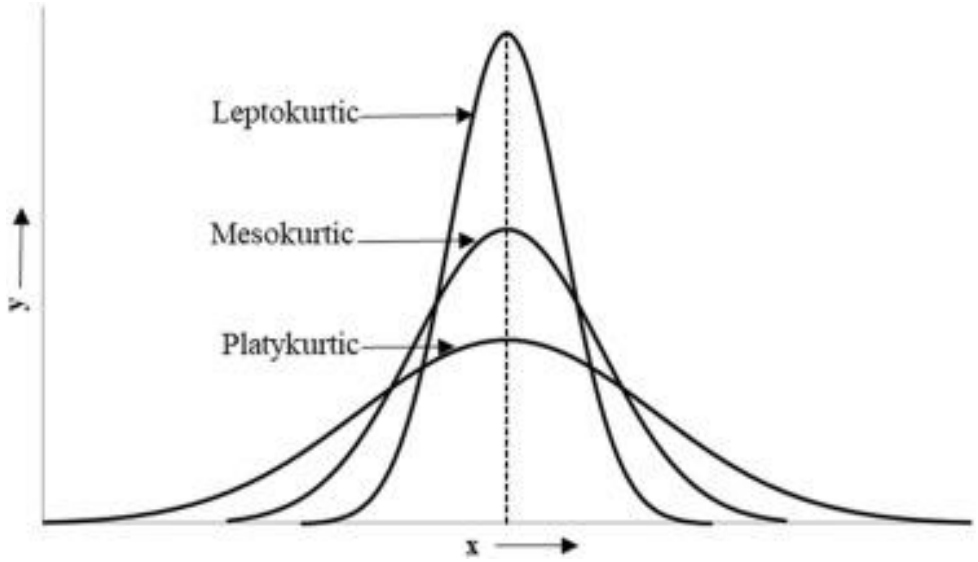
常见的描述性统计量



In negatively skewed data, the **mean is a smaller number** than the median and mode.	In a normal distribution, the **mean is the same number** as the median and mode.	In positively skewed data, the **mean is a larger number** than the median and mode.

常见的描述性统计量

- 6. 峰度 (Kurtosis):
- 峰度衡量的是分布集中度（或尾部的厚度）。它反映了分布的尖锐程度，即相对于正态分布，数据的峰值是更尖锐还是更平坦。
- $\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$
- - 峰度为3表示分布称为中峰态，Mesokurtic。
- - 峰度大于3表示尖峰态 (Leptokurtic)，分布更尖锐。
- - 峰度小于3表示平峰态 (Platykurtic)，分布更平坦。



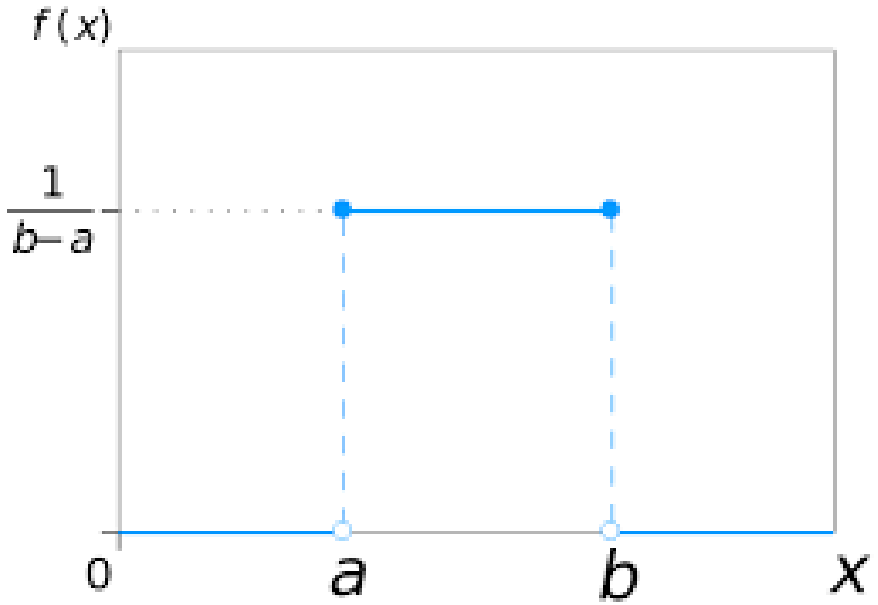
常见的概率分布

均匀分布

- 在一个特定的区间内，随机变量取任何值的可能性相等。若随机变量X在区间$[a,b]$上服从连续均匀分布，则其概率密度函数为

$$f(x)=\frac{1}{b-a}$$

- 在这个区间之外，概率密度为零。
- 对于离散均匀分布，假设随机变量Y取值为$\{x_1,x_2,\cdots,x_n\}$
- $$P(Y=x_i)=\frac{1}{n},\ i=1,2,\cdots,n。$$

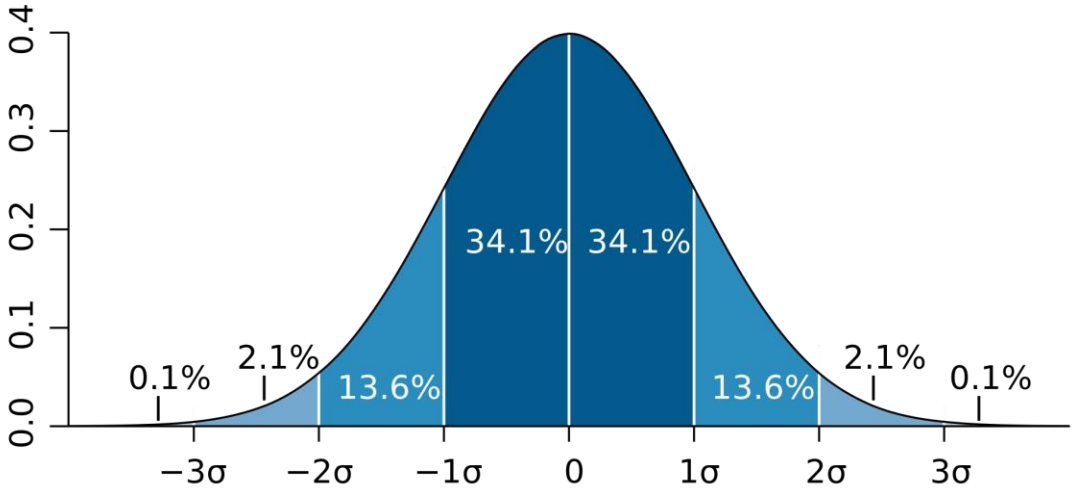


正态分布

- 正态分布，也称为高斯分布，是一种在自然界和社会科学中广泛出现的重要概率分布。其概率密度函数为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 其中，x是随机变量的值，μ是均值，决定了分布的中心位置；σ是标准差，决定了分布的离散程度。

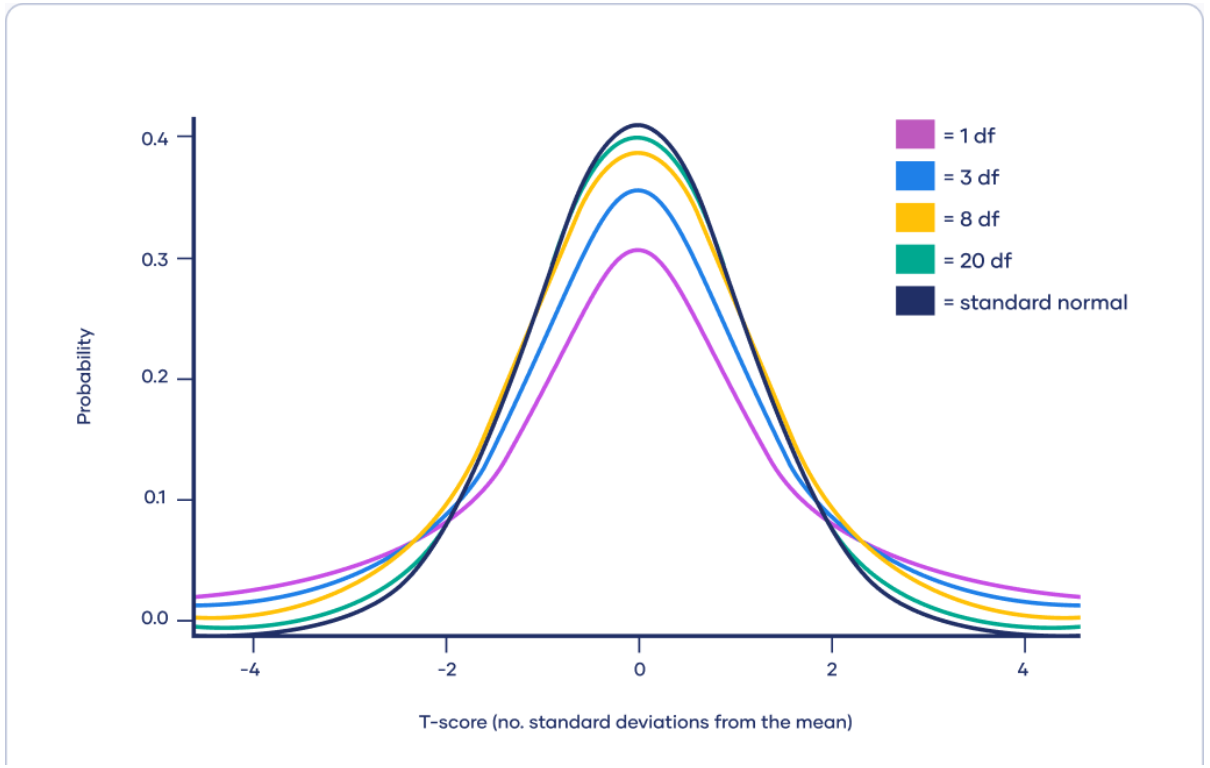


t 分布

- 如果随机变量T服从t分布，其概率密度函数为

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- 其中t为随机变量的值，ν为自由度参数。
- 与正态分布相比，分布的形状更扁平，尤其是在自由度较低时。随着自由度的增加，分布逐渐接近正态分布

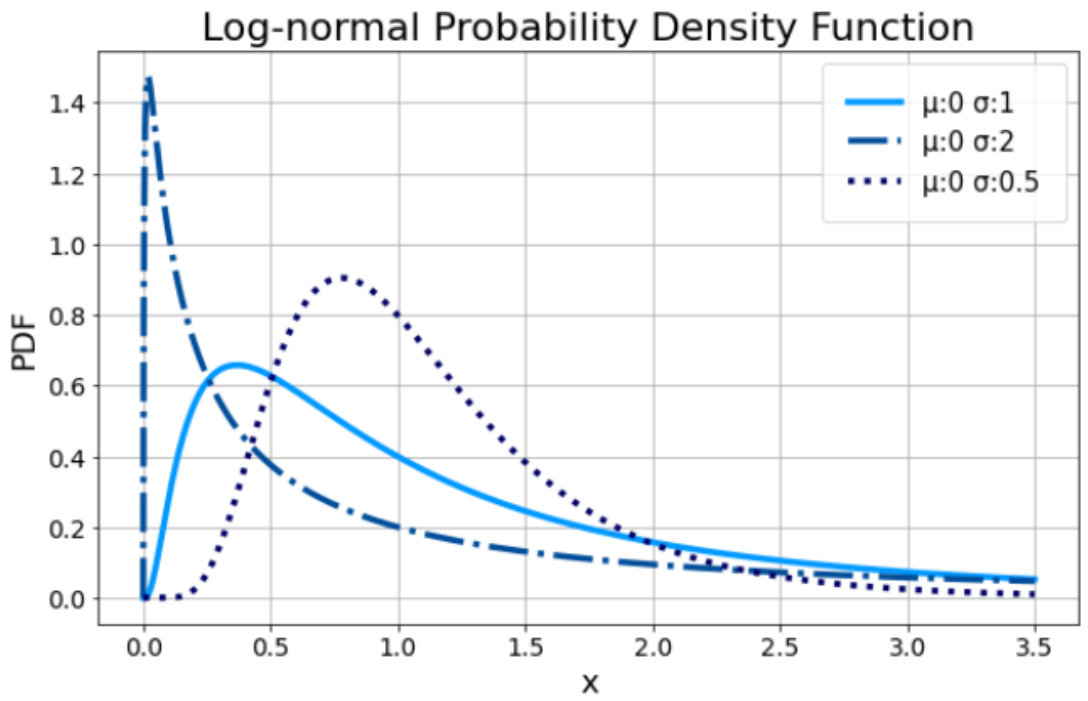


对数正态分布

- 如果随机变量X服从对数正态分布，其概率密度函数为

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

- μ和σ分别是$\ln(X)$的均值和标准差。

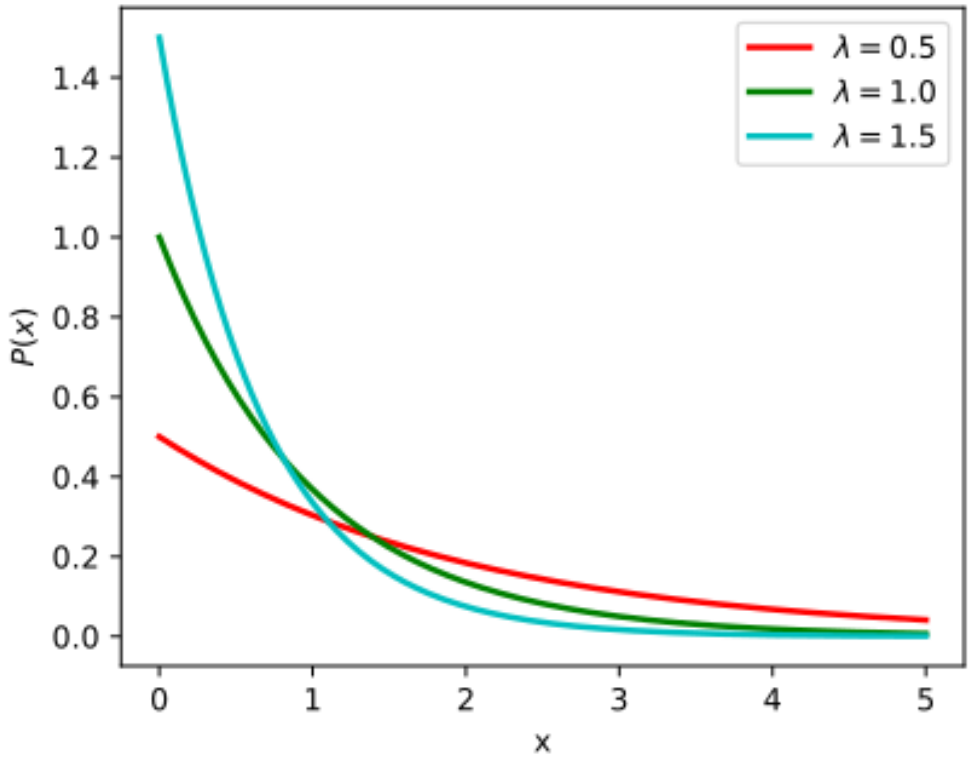


指数分布

- 如果随机变量服从指数分布，其概率密度函数为

$$f(x) = \lambda e^{-\lambda x}$$

- $\lambda > 0$是分布的参数，称为率参数（rate parameter）。



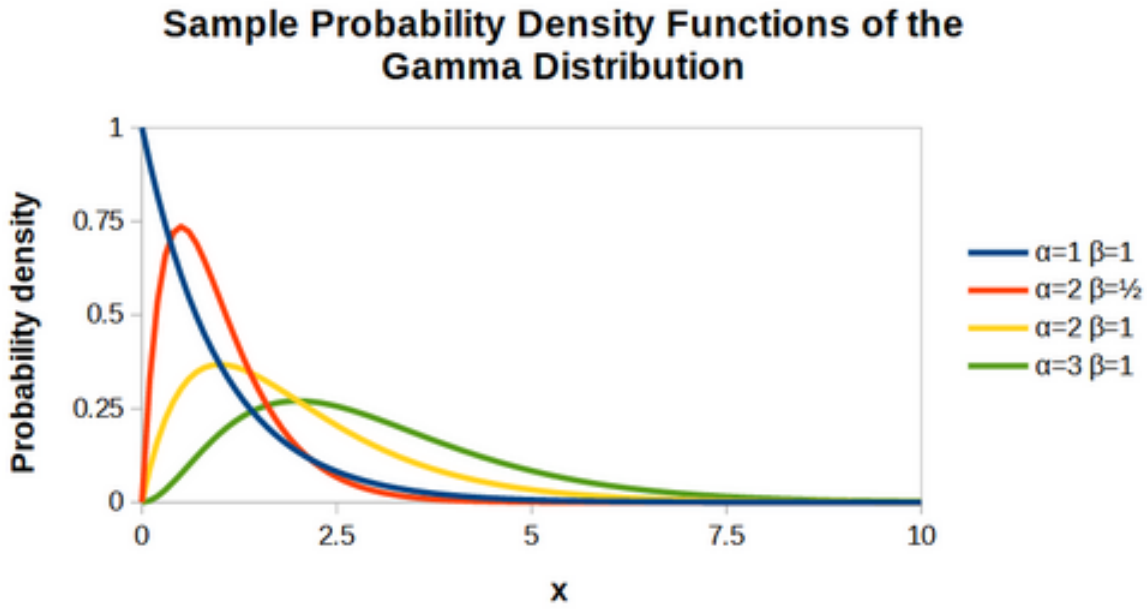
Gamma分布

- 如果随机变量X服从 Gamma 分布，其概率密度函数为

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- α和β是两个参数，$\Gamma(\cdot)$是伽马函数。
- 指数分布是 Gamma 的特例：

$$\text{Exp}(\lambda) \equiv \mathit{Gamma}(\alpha = 1, \beta = \lambda).$$



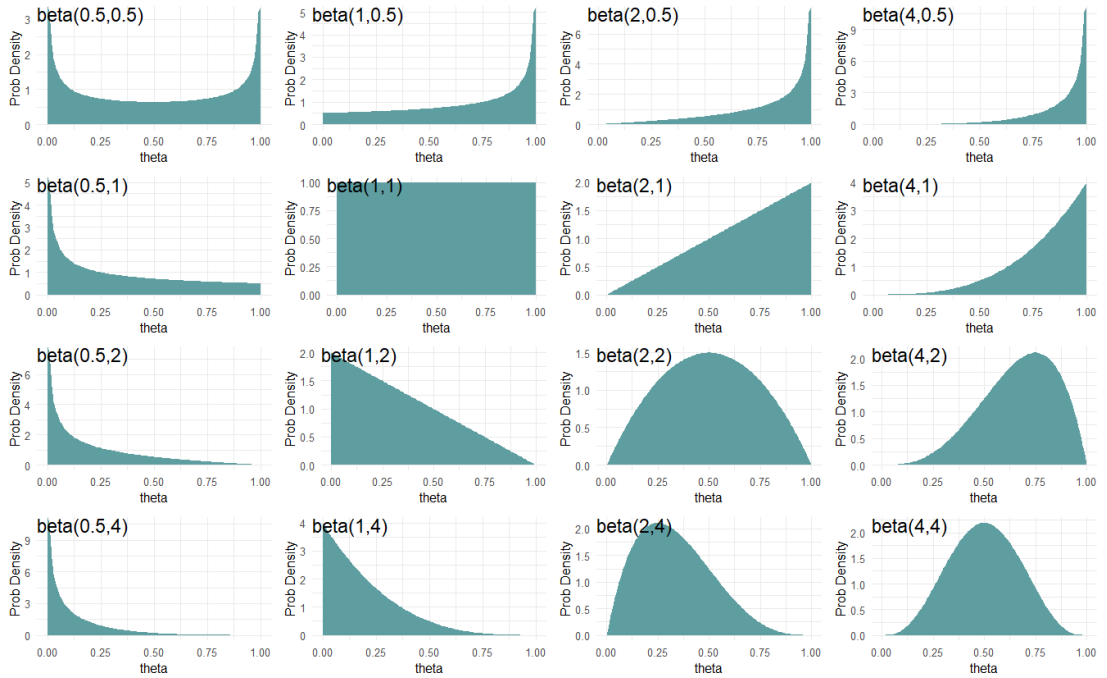
Beta分布

- 如果随机变量 X 服从 Beta 分布，其概率密度函数为

$$f(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

- 其中$x\in[0,1]$，$\alpha>0,\beta>0$，Beta 函数表达式为：

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$



数据标准化及异常值处理

数据标准化

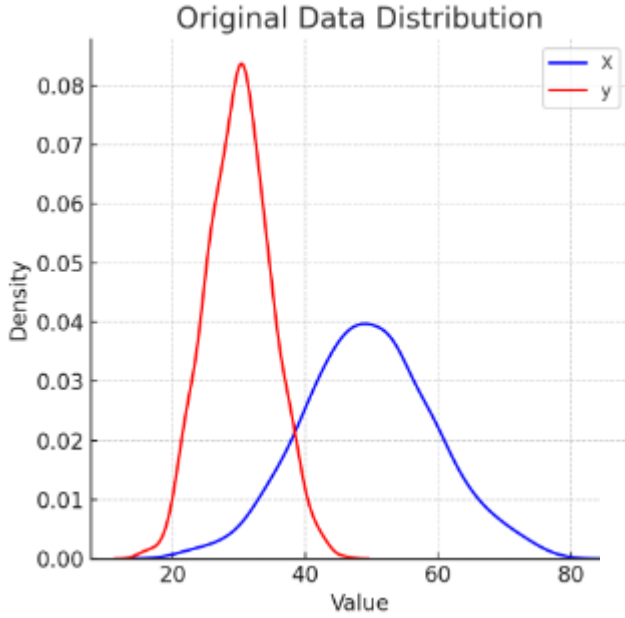
- **定义：**过特定的方法将数据的特征值转换到一个特定的范围或具有特定的分布，使得不同特征在进行分析时具有可比性和稳定性。
- **数据标准化的作用：**
  - **直观比较特征：**在金融数据分析中，经常需要比较不同特征之间的关系。经过标准化后，所有特征都被映射到相同的尺度上，可以更直观地比较各个特征对目标变量的贡献程度。
  - **增强特征解释性：**对于一些基于距离或相似度的算法，如 K 近邻算法和聚类算法，数据标准化可以使特征之间的距离或相似度计算更加合理。标准化后，距离或相似度的计算更加公平地考虑了各个特征，增强了算法结果的解释性。

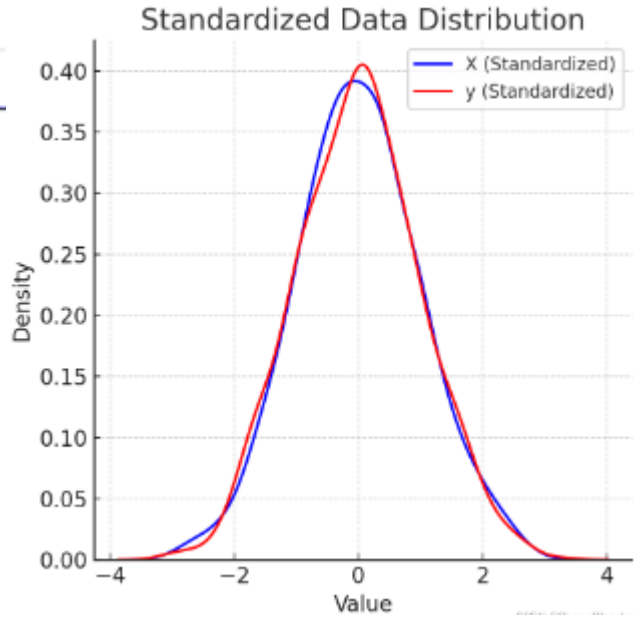
数据标准化

- **在机器学习中：**
  - **加快收敛速度：**如果数据未标准化，不同特征可能具有不同的尺度，导致算法在优化过程中在某些特征方向上进展缓慢，而在其他特征方向上进展迅速，从而延长了收敛时间。
  - **提高模型精度：**许多机器学习模型对数据的尺度非常敏感。如果数据未标准化，其模型的非线性部分可能会失效。
  - **防止数值溢出：**未标准化的数据可能会使算法在计算过程中出现数值不稳定的情况，例如梯度爆炸或消失。

常用的数据标准化方法

- **Z-score 标准化**
- 计算公式为：
  - $Z = \frac{X-\mu}{\sigma}$
- 对数据的分布形状没有改变，只是将数据进行了平移和缩放，使得数据符合标准正态分布。



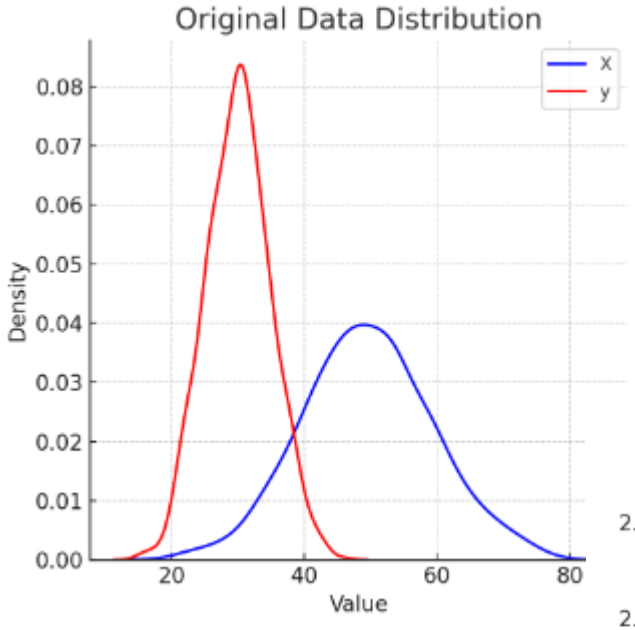


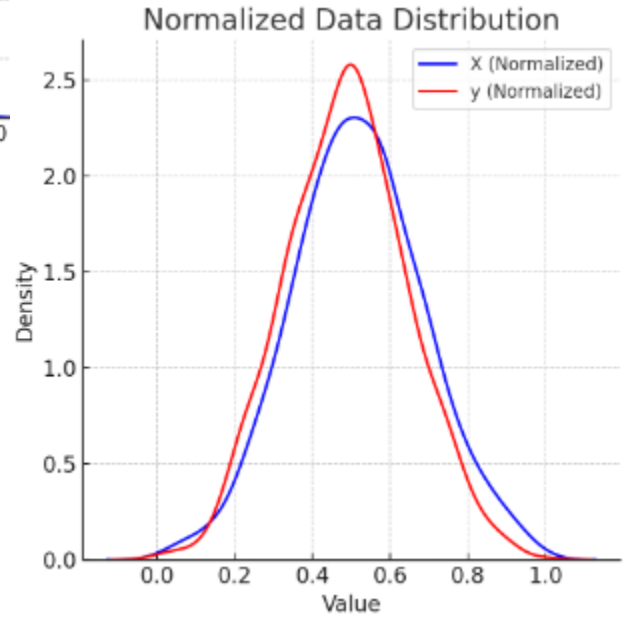
常用的数据标准化方法

- **离差标准化**（归一化）
- 计算公式为：

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

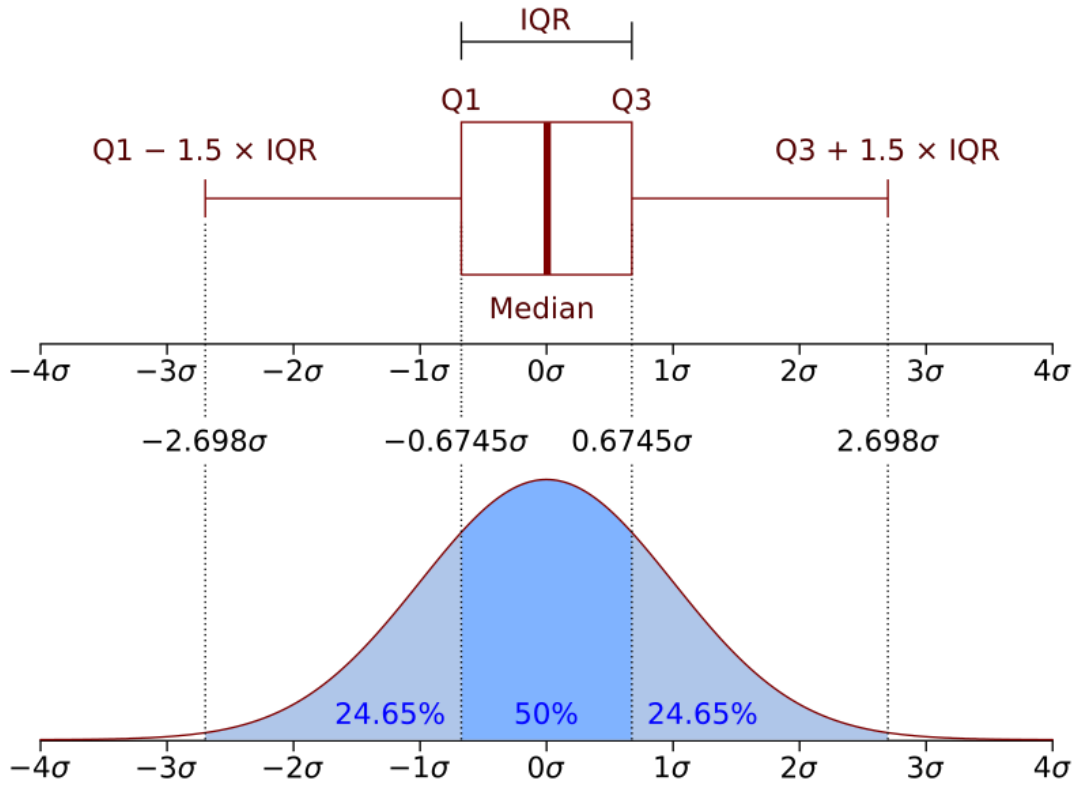
- 将数据的范围压缩到一个固定的区间内，通常是 [0,1] 区间。





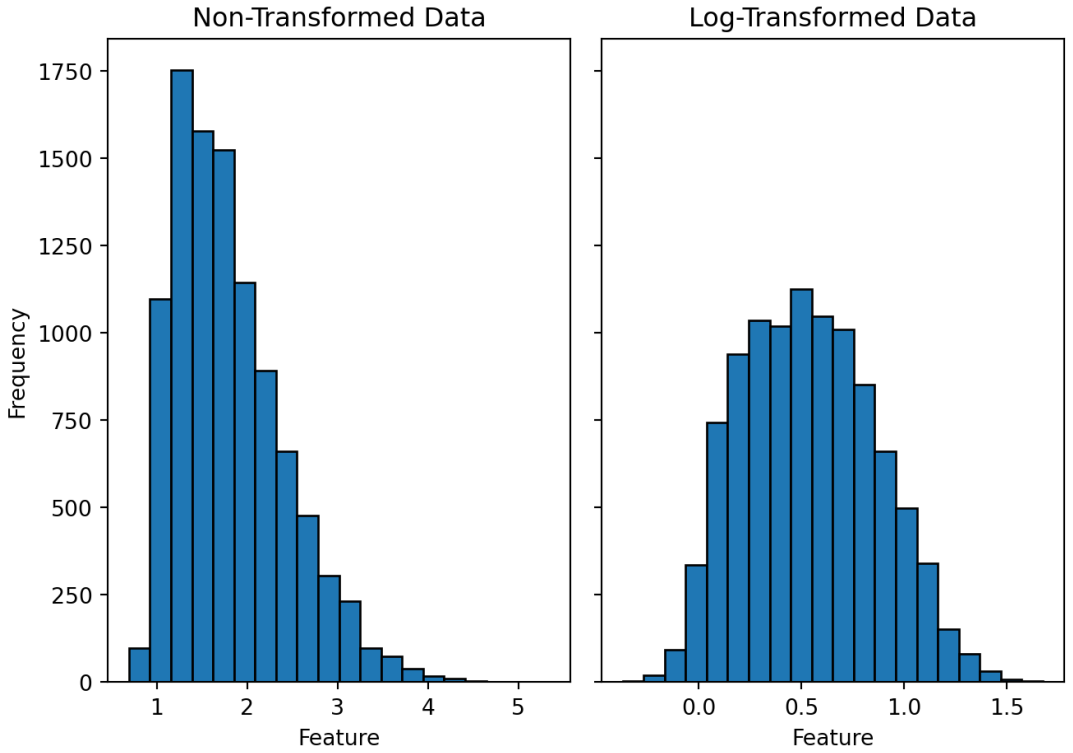
常用的数据标准化方法

- **IQR方法**（四分位距方法）：
- 1. 将数据从小到大进行排序。
- 2. 确定第一四分位数（Q1）、二四分位数（Q2）和第三四分位数（Q3）。
  - $IQR = Q3 - Q1$
  - $Z = \frac{X - Q2}{IQR}$



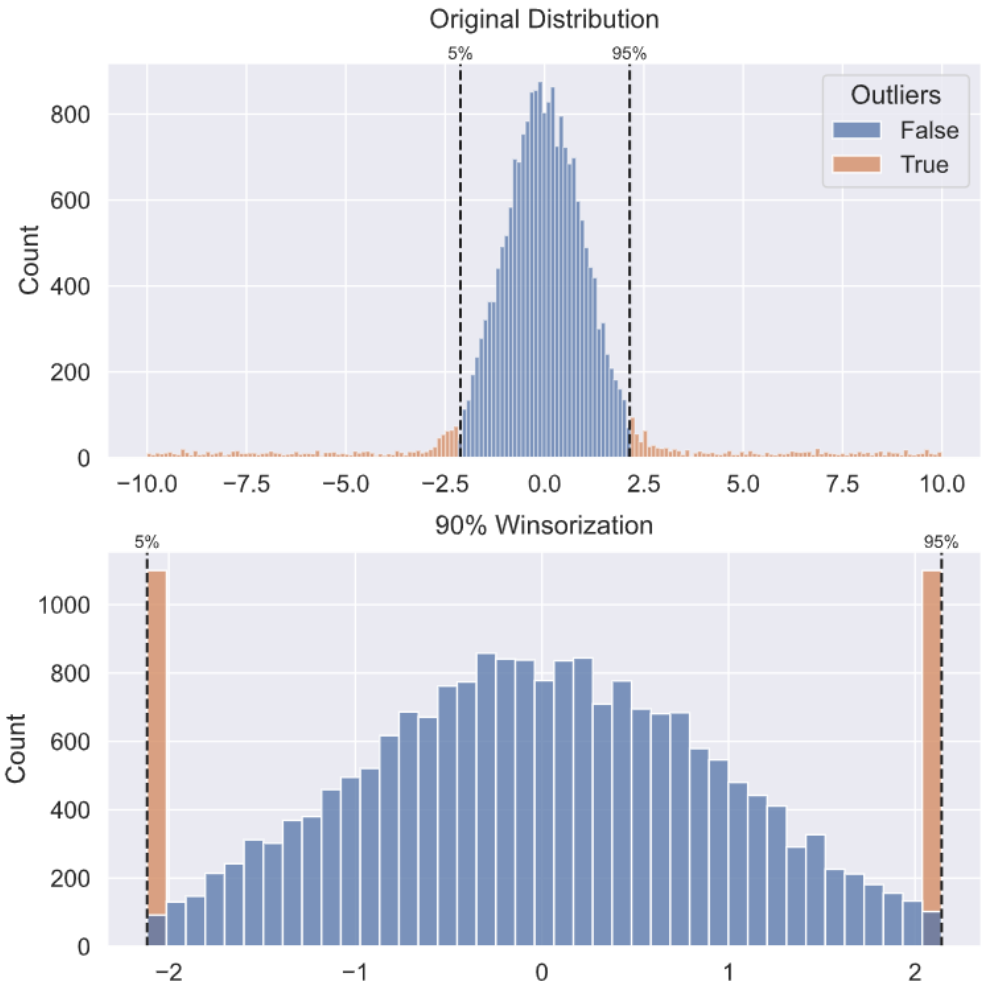
常用的数据标准化方法

- **对数标准化：**对数据取对数来实现标准化。通常使用自然对数e为底。
  - $Y = \log(X)$
- 可以将数据的分布进行压缩，对于一些具有指数增长特性的数据非常有效。



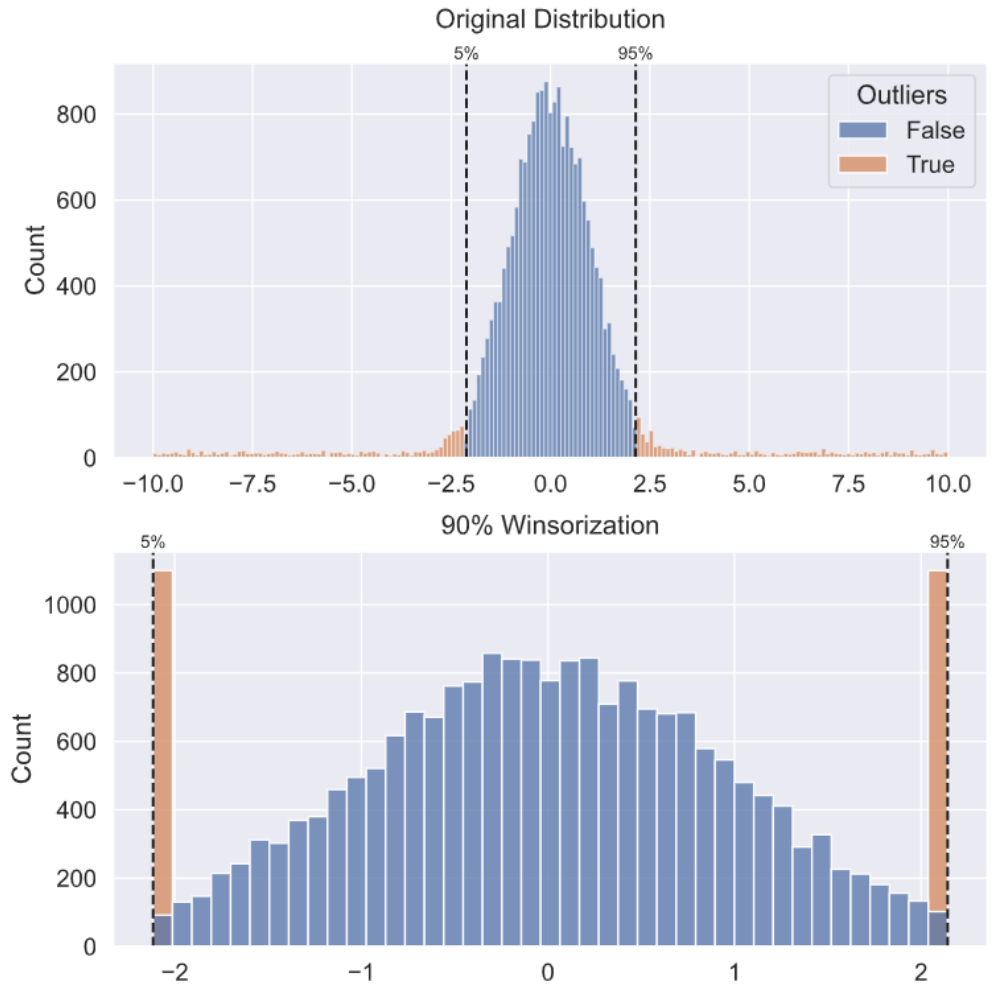
常用的异常值处理方法

- **Winsorized 法：**即缩尾法，是一种减少异常值对数据集平均数影响的统计方法。
- 若采用上下各 5% 的 Winsorized 法，那么上截断点位于第 95 个百分位数处，下截断点位于第 5 个百分位数处。
- 将数据中小于下截断点的值替换为下截断点的值，将大于上截断点的值替换为上截断点的值。而处于上下截断点之间的数据保持不变。



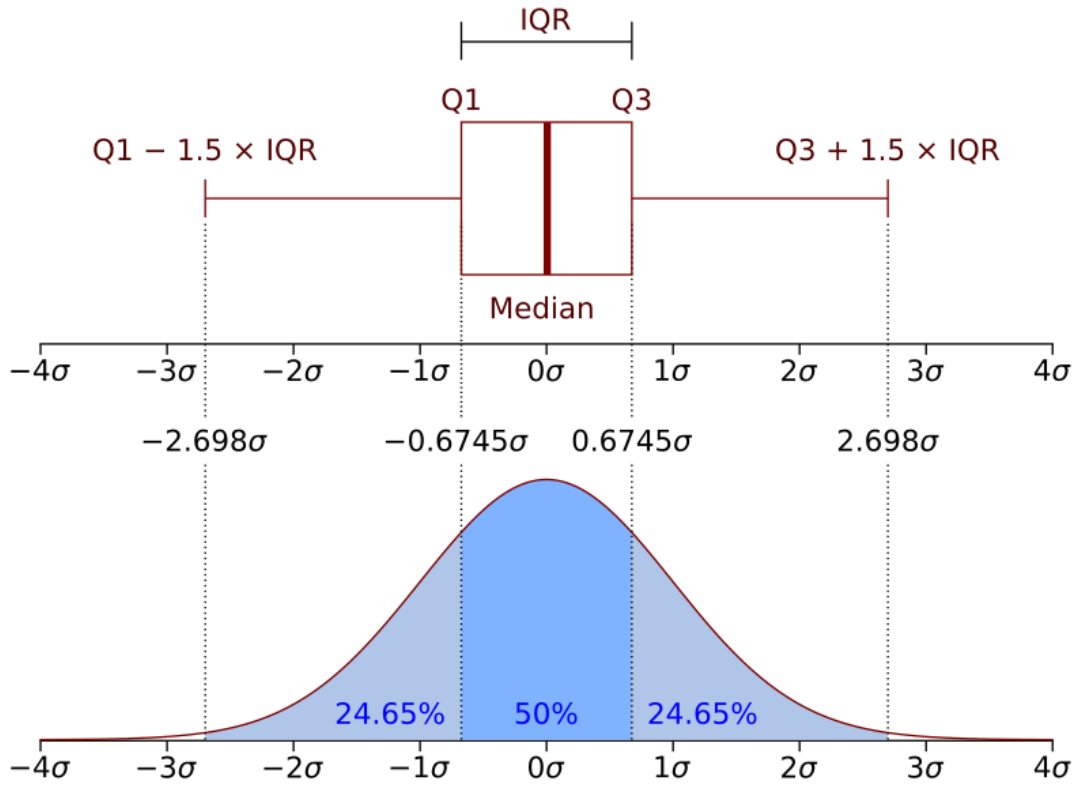
常用的异常值处理方法

- **修剪法（Trimmed）：** 这种方法涉及删除数据集中一定比例的最低和最高值，然后计算剩余数据的均值。
- 与Winsorized 不同，修剪均值直接移除极端值，而不是替换它们。

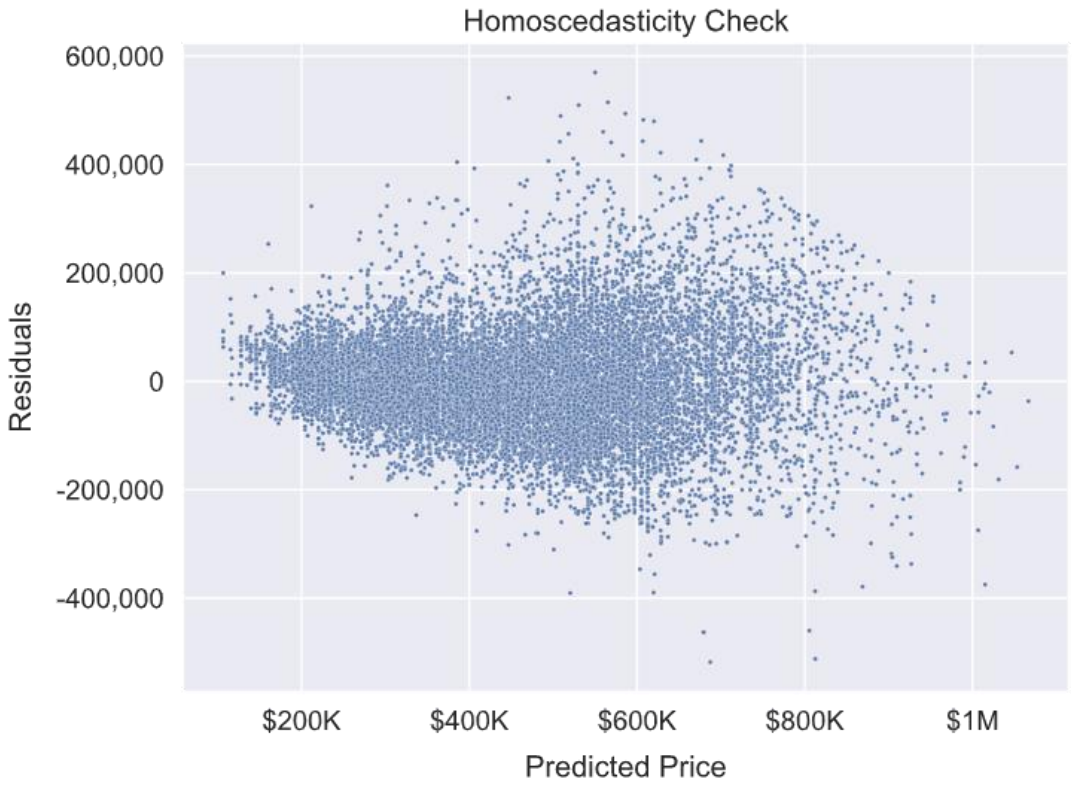


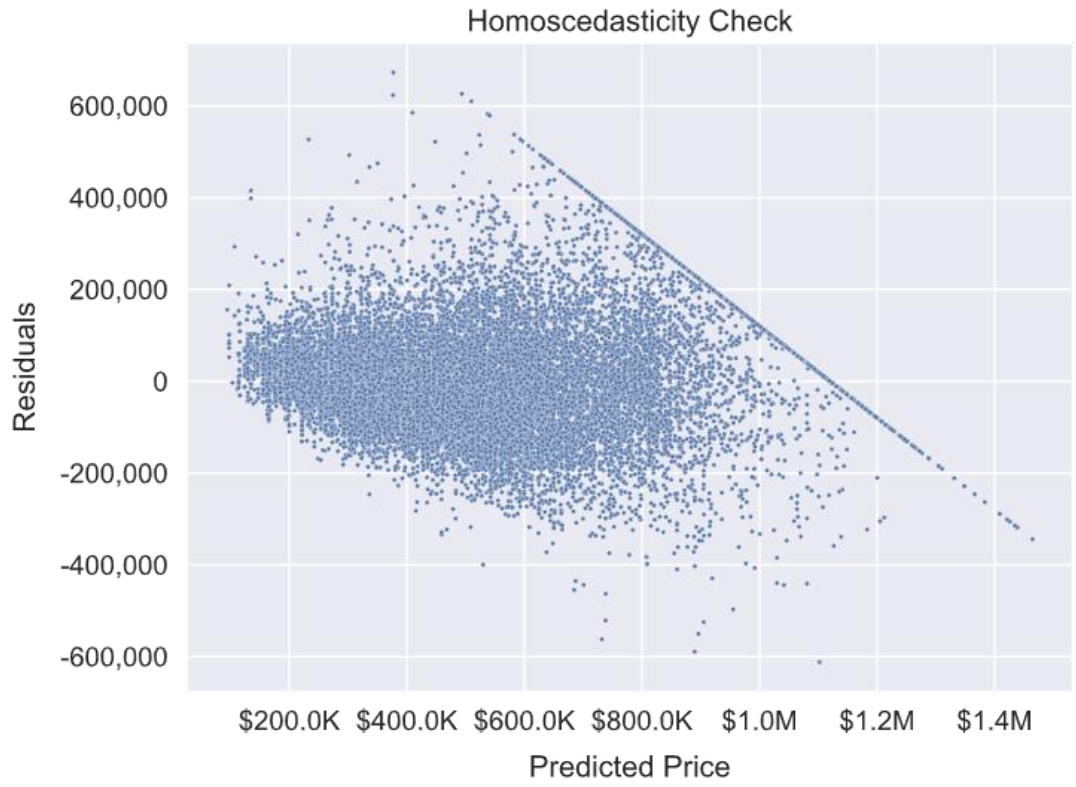
常用的异常值处理方法

- **IQR方法**（四分位距方法）：
  - $IQR = Q3 - Q1$
- 剔除小于$\mathrm{Q1} - 1.5 * IQR$的数据
- 剔除大于$Q3 + 1.5 * IQR$的数据



异常值处理的代价





ref: https://ndgigliotti.medium.com/trimming-vs-winsorizing-outliers-e5cae0bf22cb

数据中性化

数据中性化

- 数据中性化是指从数据集中删除或修改某些特定信息的过程，以确保数据能够客观、真实地反映事物的本来状态。
- 数据中性化的意义:
- 中性的数据能提供准确的信息，避免因数据偏差导致的错误判断和决策问题。
- 例如，在市场调研中，商品偏好会受到消费者群体的性别、年龄、地域显著影响。若在分析产品变量对真实需求影响时，不排除这些重要因子对分析造成的影响，企业可能会错误地判断市场需求，投入大量资源开发不受欢迎的产品。

数据中性化方法：回归法

- 确定因变量和自变量
  - 因变量：选择需要进行中性化处理的因子数据。
  - 自变量：将需要排除影响的数据作为自变量。
- 进行线性回归
  - 得到回归方程：y = a + bx，其中 y 是因变量（待中性化因子），x 是自变量（需要排除的影响），a 和 b 是回归系数。
- 计算残差
  - 残差 = 实际值 - 预测值。残差就是经过中性化处理后的数据，它消除了自变量对待中性化因子的影响。

数据中性化方法：均值方差法

- 均值方差法实际上是回归法对于0-1变量的简化版本。

- $y = bx + \epsilon$
- $\hat{b} = \frac{\sum_{i \in N} x_i y_i}{\sum_{i \in N} x_i x_i} = \frac{\sum_{i \in K} y_i}{K}$
- $\epsilon_i = y_i - \hat{b} x_i$

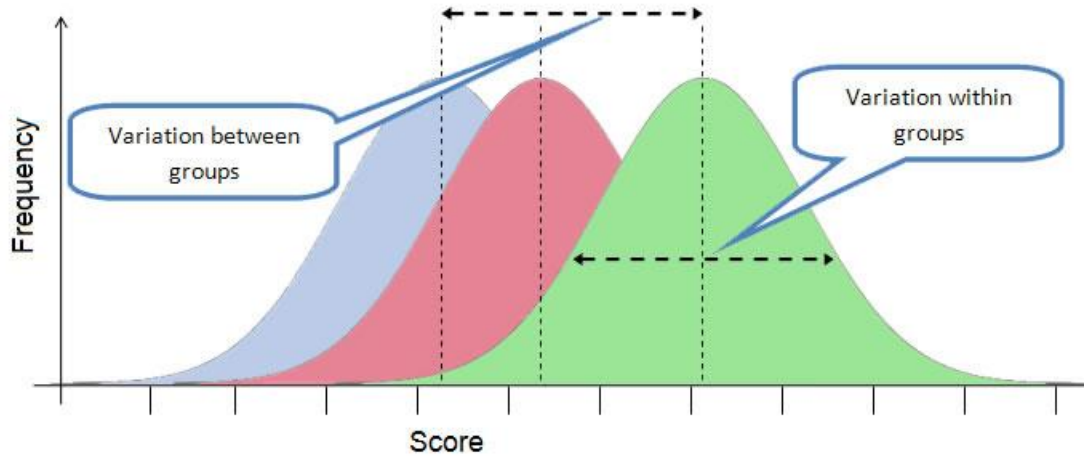
属于分类A	属于分类B
$\frac{y_i - \mu_A}{\sigma_A}$	$\frac{y_i - \mu_B}{\sigma_B}$

方差分析

方差分析

- 方差分析（ANOVA）是一种用于比较两个或多个总体均值是否存在显著差异的统计方法。

1. 原假设（H0）：所有组的总体均值相等。
2. 备择假设（H1）：并非所有的均值都相等。



方差分析的假设

- ANOVA有三大前提假设：
  - **正态性**：数据服从正态分布或者是逼近正态分布。
  - **方差齐性**：各组样本的总体方差相等。
  - **独立性**：各组样本是互相独立的随机样本。

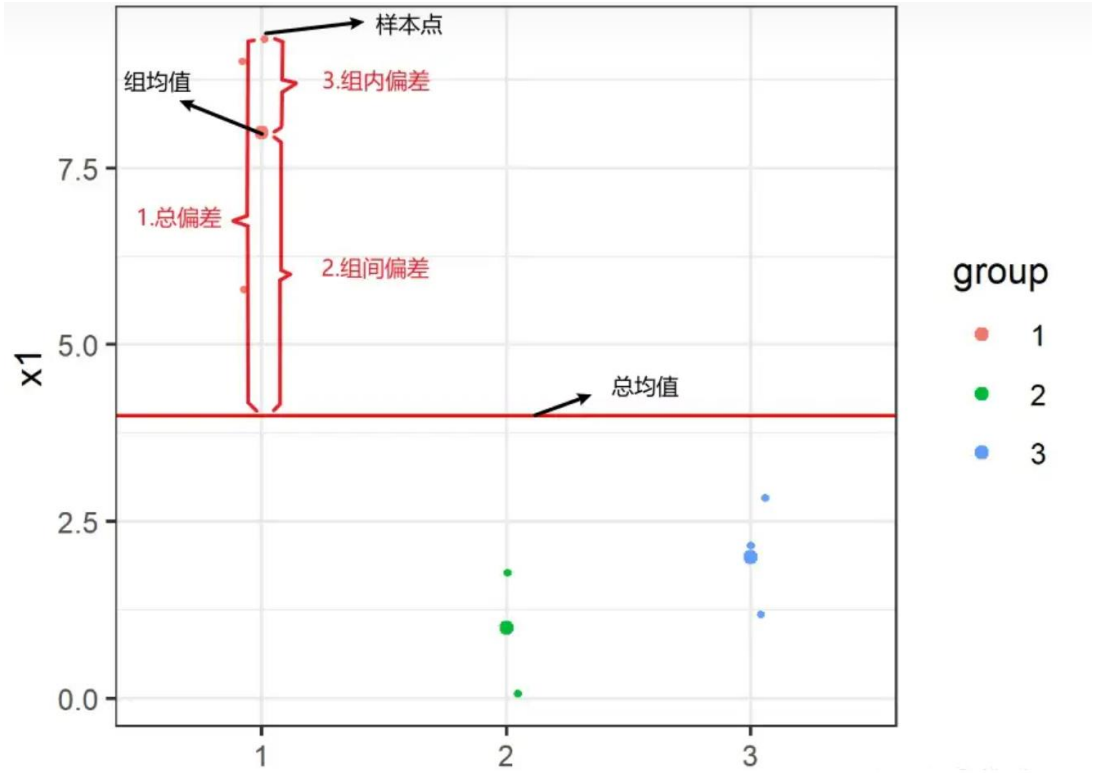
方差分析

- **基本原理：**
- 将**总变异**分解为各个因素引起的偏差和随机误差引起的偏差。通过比较不同因素水平下的**组间偏差**和**组内偏差**的大小，来判断因素对因变量是否有显著影响。

- 总偏差：$\sum_i \sum_j (Y_{ij} - \overline{Y}_{total})^2$
- 组间偏差：$\sum_i n_i (\overline{Y}_i - \overline{Y}_{total})^2$
- 组内偏差：$\sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2$

方差分析

- 总偏差(TSS)：$\sum_i \sum_j (Y_{ij} - \overline{Y}_{total})^2$
- 组间偏差(BSS)：$\sum_i n_i (\overline{Y}_i - \overline{Y}_{total})^2$
- 组内偏差(WSS)：$\sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2$
- TSS = BSS + WSS



方差分析的步骤:

- **建立假设**:
  1. 原假设（H0）： 所有组的总体均值相等。
  2. 备择假设（H1）： 并非所有的均值都相等。
- **计算方差**:
  1. 计算组间偏差、组内偏差。
  2. 计算组间均方、组内均方。

$$BMSS = \frac{BSS}{k-1} = \frac{\sum_i n_i(\overline{Y}_i - \overline{Y}_{total})^2}{k-1}$$

$$WMSS = \frac{WSS}{N-k} = \frac{\sum_i \sum_j (Y_{ij} - \overline{Y}_i)^2}{N-k}$$

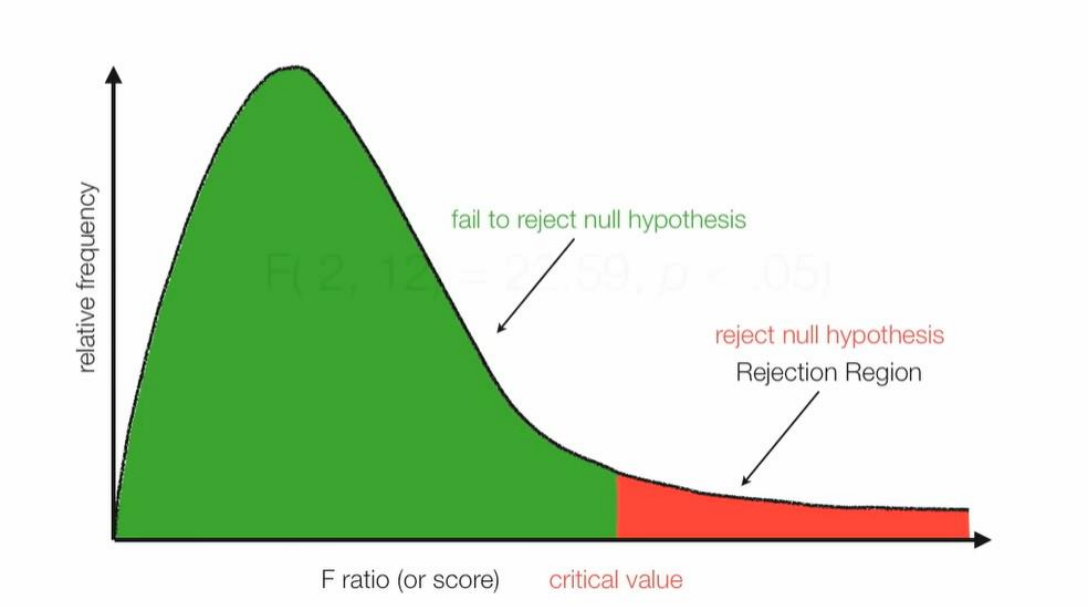
3. 计算F统计量： F = BMSS/ WMSS

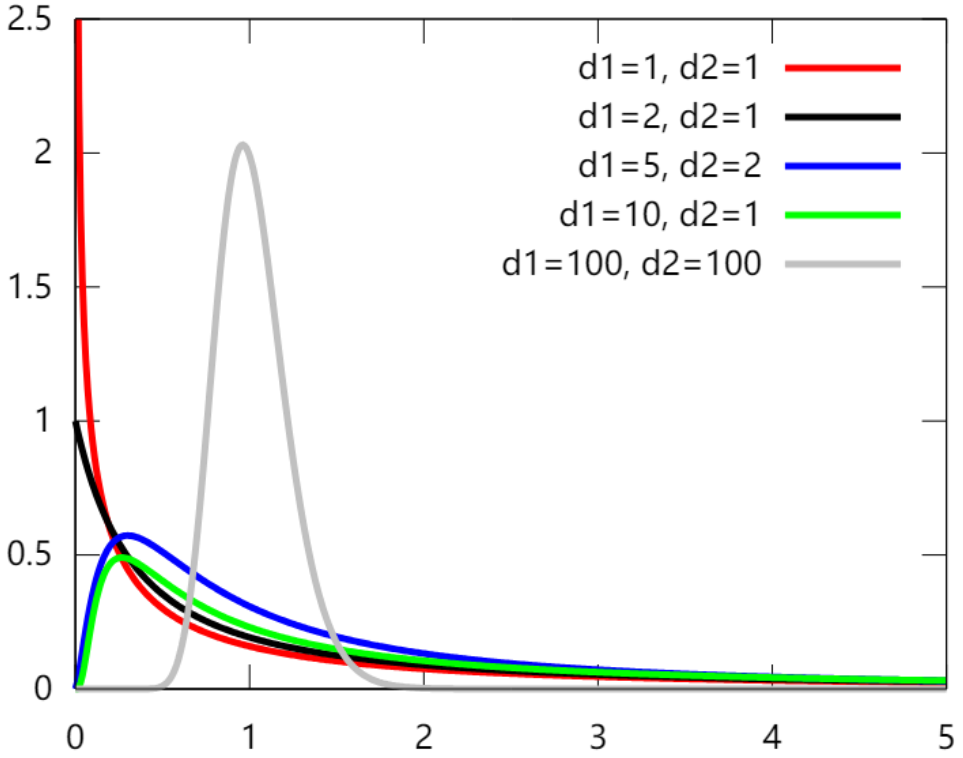
方差分析的步骤:

- **确定显著性水平并作出决策**:
  1. 使用F分布表或统计软件来确定F统计量的显著性水平。
  2. 如果F值大于临界值，或者对应的p值小于显著性水平（例如0.05），则拒绝原假设，认为至少有两组的均值存在显著差异。
  3. 如果F值小于临界值，或者p值大于显著性水平，则不能拒绝原假设。

- $F(2, 12) = 22.59, p<.05$

F 分布


relative frequency
fail to reject null hypothesis
reject null hypothesis
Rejection Region
F ratio (or score)
critical value


2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=10, d2=1
d1=100, d2=100

例：狗的体重

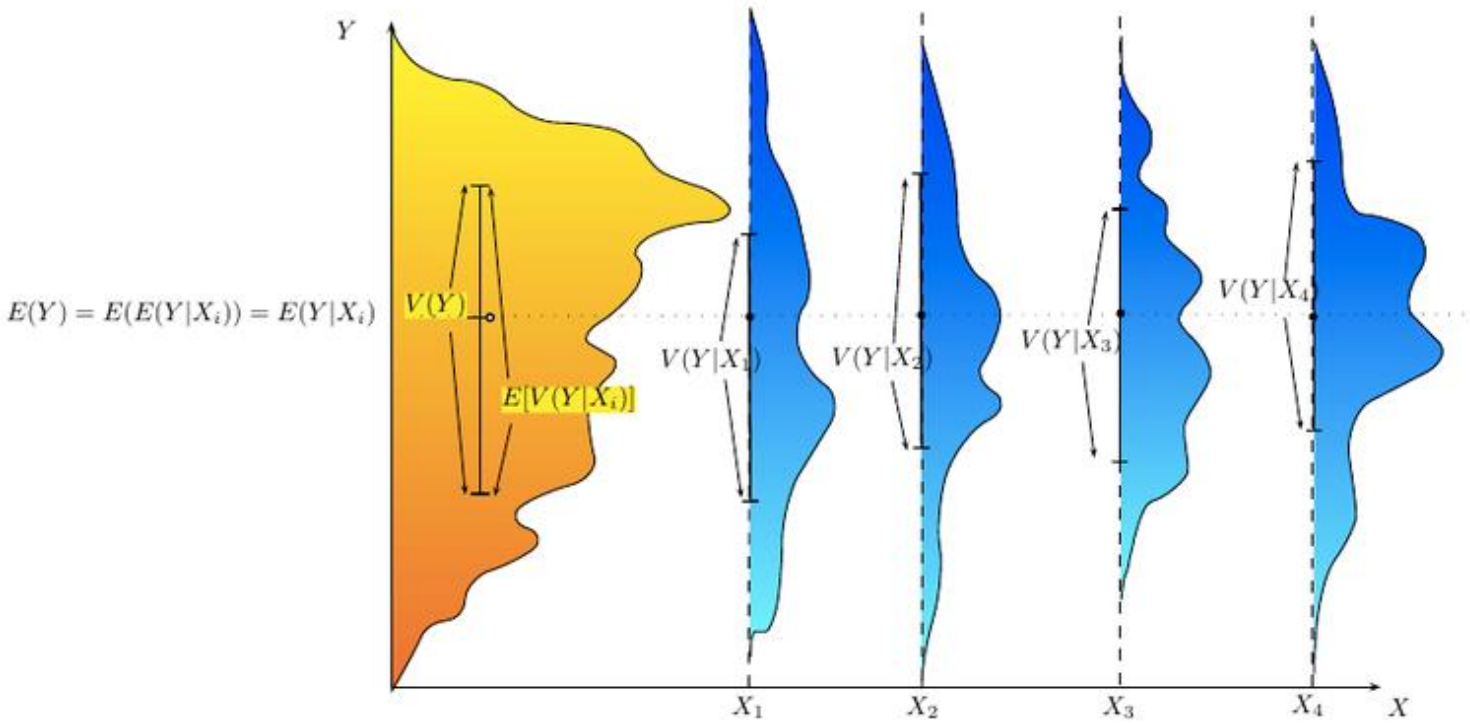
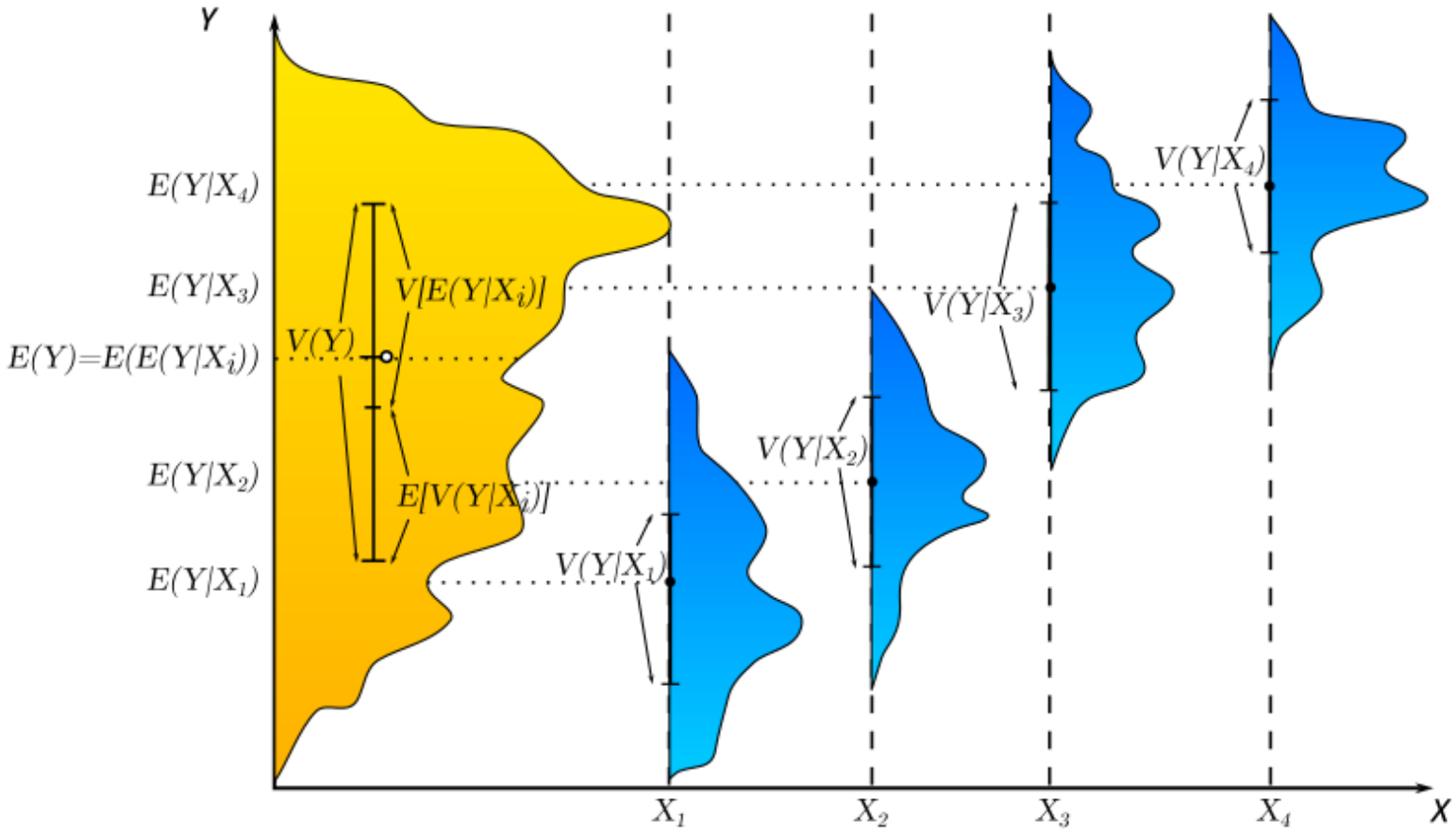


Figure 2: ANOVA : No fit

例：狗的体重


Y
$E(Y|X_4)$
$E(Y|X_3)$
$E(Y)=E(E(Y|X_i))$
$E(Y|X_2)$
$E(Y|X_1)$
$V(Y)$
$V[E(Y|X_i)]$
$E[V(Y|X_i)]$
$V(Y|X_1)$
$V(Y|X_2)$
$V(Y|X_3)$
$V(Y|X_4)$
X_1
X_2
X_3
X_4
X

例：狗的体重

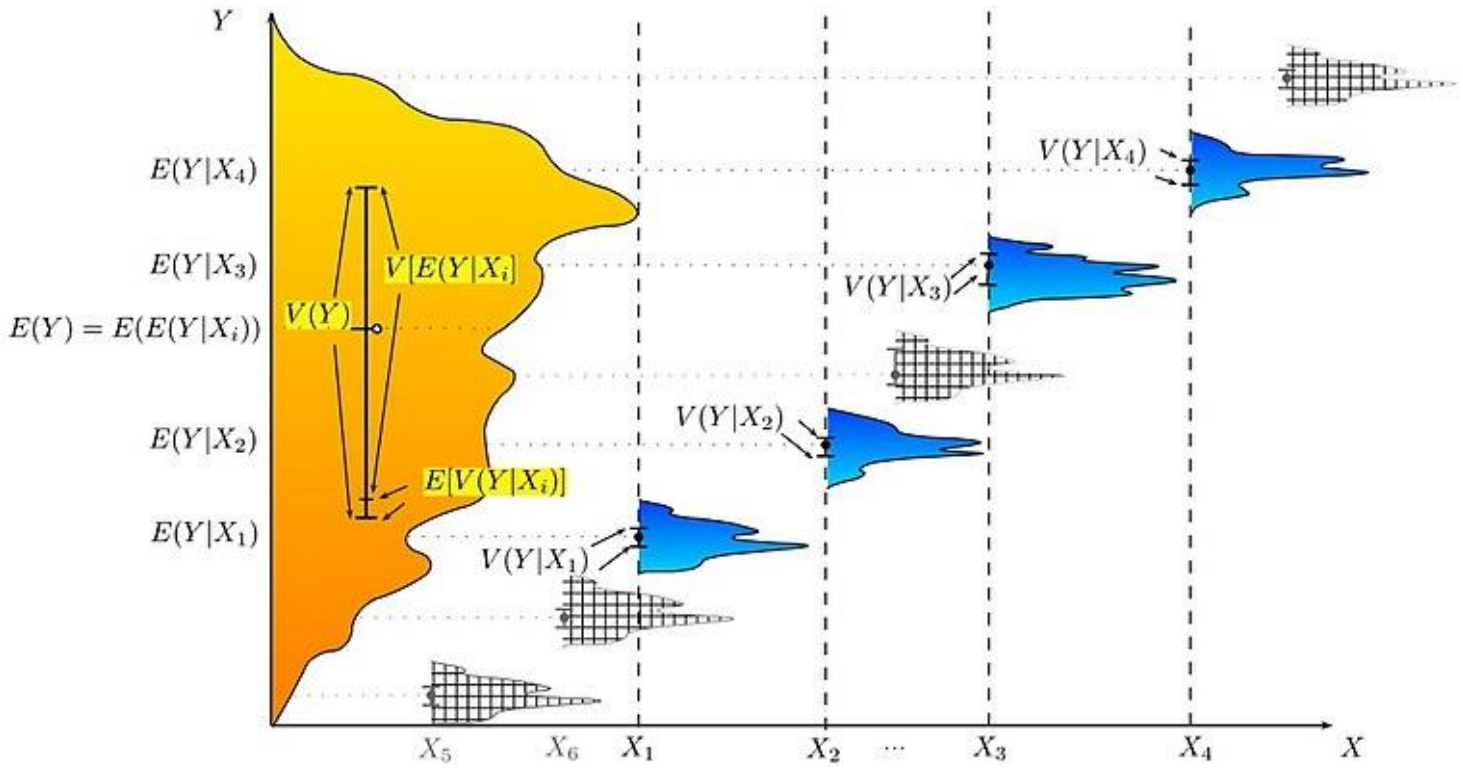


Figure 3: ANOVA : very good fit

双因素方差分析

- 某一变量可能并非仅受单一因素影响，甚至存在另一个因素的效应。
- 在双因素方差分析中，除了考虑双因素彼此的效应之外，也可能存在因素之间的联合效应，也就是因素间的交互作用。这也使得双因素方差分析变得比较复杂。
- 例如，要比较五个城市的空气污染总指标差异，除了城市本身的因素之外，还必须考虑机动车密度的因素。在这种情况下，城市与机动车密度可能存在某种效应影响着空气污染的多少。

双因素方差分析

- 总偏差(TSS):
- A因子的组间偏差(ASS):
- B因子的组间偏差(BSS):
- 相互作用(ABSS) :
- 组内偏差(WSS):
- TSS=ASS+BSS+ABSS +WSS

$$\sum_i \sum_i \sum_z (Y_{ijz} - \overline{Y}_{total})^2$$

$$nb \sum_i (\overline{Y}_i - \overline{Y}_{total})^2$$

$$na \sum_j (\overline{Y}_j - \overline{Y}_{total})^2$$

$$n \sum_i \sum_j (\overline{Y}_{ij} - \overline{Y}_i - \overline{Y}_j + \overline{Y}_{total})^2$$

$$\sum_i \sum_j \sum_z (Y_{ijz} - \overline{Y}_{ij})^2$$

双因素方差分析

- A因子的F检验：$\frac{ASS/(a-1)}{\mathrm{WSS}/(ab(n-1))}$
- B因子的F检验：$\frac{BSS/(b-1)}{\mathrm{WSS}/(ab(n-1))}$
- 相互作用的F检验：$\frac{ABSS/((a-1)(b-1))}{\mathrm{WSS}/(ab(n-1))}$

方差分析的假设检验

- 正态分布检验
  - 卡方拟合优度检验
  - Jarque-Bera检验
  - Kolmogorov-Smirnov检验
  - Lilliefors检验
- 方差齐次性检验
  - Bartlett检验

非参数方差分析

- 当数据不满足正态性和方差齐次性假定时，方差分析可能会给出错误的结果。

- 非参数方差分析检验：
  - Kruskal-Wallis检验
  - Friedman检验

Kruskal-Wallis检验

- Kruskal-Wallis 检验，也称 H 检验。与ANOVA不同的是，H检验不使用原始数据，而是将样本数据进行排序，对秩次进行分析。
- **计算步骤**：
  1. 将所有样本的数据合并在一起，按照从小到大的顺序进行排序，对每个数据赋予秩次（排序后的序号）。
  2. 计算每组样本的秩次和，以及所有样本秩次的总和。
  3. 计算Kruskal-Wallis H统计量，公式如下：

$$H = \left(\frac{12}{N(N+1)}\right) \sum_j^K \left(\frac{R_j^2}{n_j}\right) - 3(N+1)$$

其中N是所有样本数据的总数，K是分类个数，R_j是第个样本的秩和，n_j是第个样本的大小。

Kruskal-Wallis检验

- **不满足正态分布和方差齐次性的数据**：当数据不满足正态分布、方差齐次性，Kruskal-Wallis检验提供了一种有效的替代方法。
- **小样本数据**：适用于样本量较小的情况，特别是当样本量不足以进行正态性检验时。
- **不同尺度的数据**：当数据来自不同的尺度或单位时，Kruskal-Wallis检验可以消除尺度的影响。

Friedman检验

- Friedman 检验也是一种非参数检验方法，主要用于分析同一组样本在不同条件下进行测量，检验这些样本的总体分布是否相同。
- **计算步骤**：
  1. **数据排序**：对于每一组样本，将观测值从小到大排序，并为每个观测值分配一个秩次。
  2. **计算秩次和**：对于每一组样本，计算所有观测值的秩次总和。
  3. **计算检验统计量**： Friedman检验的统计量是基于n组秩次和的分布。统计量通常表示为一个卡方分布的值，计算公式如下：
     - $Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$
       - 样本数为 k，组数为 n， R_j为第j个样本的秩和。

回归分析

回归分析与方差分析

- 都是用于研究变量之间关系的统计方法。
- 对于方差分析，可以通过设置虚拟变量的方式将其纳入回归模型中，通过回归分析的方法来研究这些因素对因变量的影响。
- 方差分析中也将总偏差分解为组间偏差和组内偏差。组间偏差对应于回归分析中的回归平方和，反映了因素的影响；组内偏差对应于残差平方和，反映了随机误差。
- 在实际研究中，方差分析和回归分析常常结合使用。
  - 可以先通过方差分析判断各个因素的主效应是否显著，再进一步使用回归分析来确定该因素与因变量之间的具体关系。
  - 在回归分析中发现某个自变量的影响不显著时，可以考虑使用方差分析来进一步检验该自变量的不同水平对因变量是否有差异。

线性回归

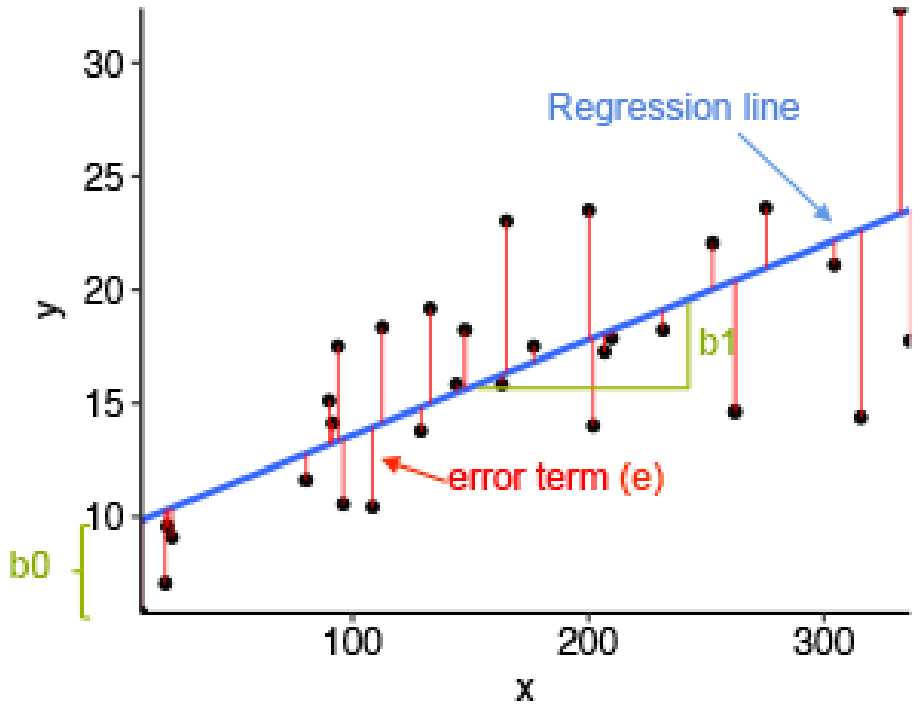
- 一元线性回归（Univariate Linear Regression）用于研究两个变量之间的线性关系。有一个因变量（通常表示为Y）和一个自变量（通常表示为X），我们假设自变量X的变化会以线性方式影响因变量Y的变化。

一元线性回归

- 一元线性回归模型的基本形式是：

$$Y = \beta_0 + \beta_1 X + \epsilon$$

其中：

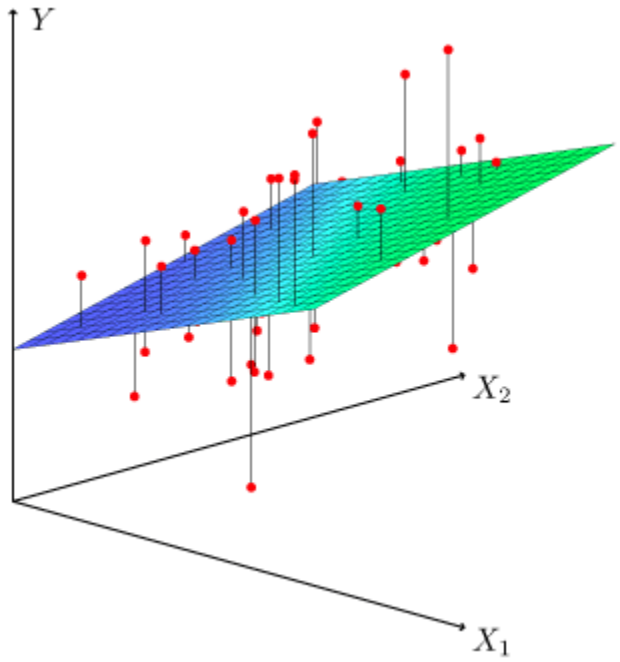


- Y 是因变量，也就是我们想要预测的变量。
- X 是自变量，也就是用来预测因变量的变量。
- β_0 是截距项，表示当自变量X为0时因变量Y的预期值。
- β_1 是斜率系数，表示自变量X每变化一个单位，因变量Y预期会变化的量。
- ϵ 是误差项，表示模型未能解释的随机变异。

多元线性回归

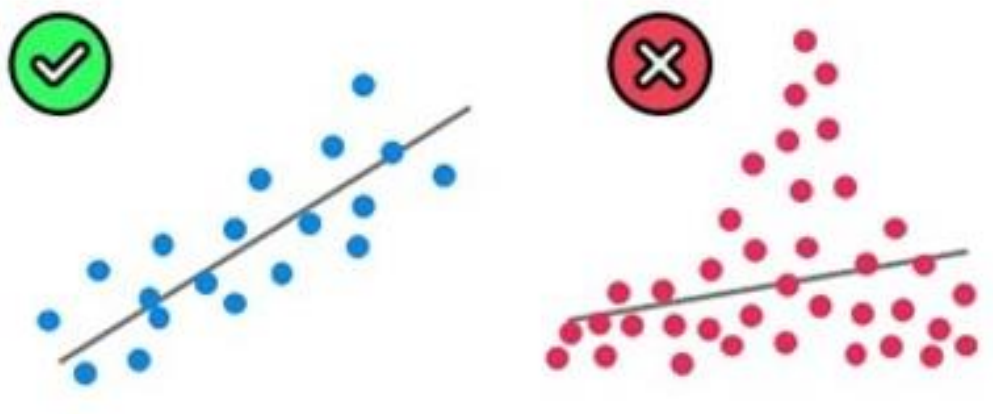
- 多元线性回归模型的一般形式是

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$



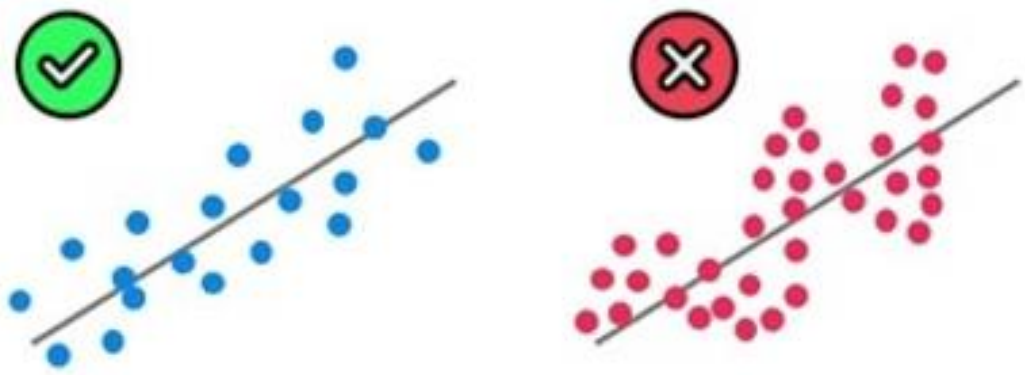
线性回归的假设

- **线性关系假设：**因变量y与自变量x之间存在线性关系。
- **独立性假设：**误差项ϵ_i之间相互独立。自变量x和误差项ϵ相互独立。
- **同方差性假设：**误差项ϵ的方差在不同的x值上是相同的，称为同方差性。
- **正态性假设：** 误差项ϵ服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



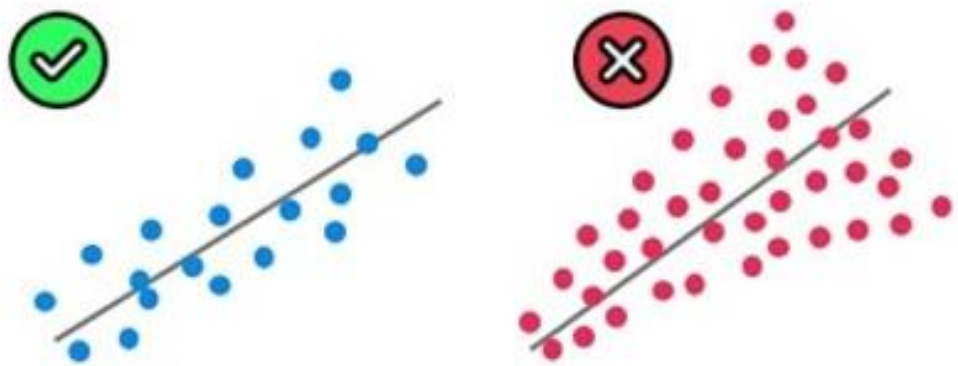
线性回归的假设

- **线性关系假设：**因变量y与自变量x之间存在线性关系。
- **独立性假设：**误差项ϵ_i之间相互独立。自变量x和误差项ϵ相互独立。
- **同方差性假设：**误差项ϵ的方差在不同的x值上是相同的，称为同方差性。
- **正态性假设：** 误差项ϵ服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



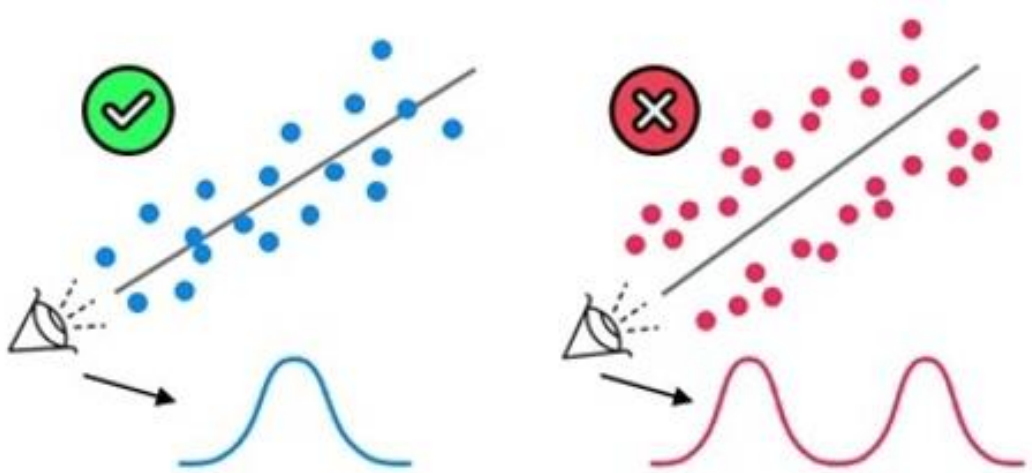
线性回归的假设

- **线性关系假设：**因变量y与自变量x之间存在线性关系。
- **独立性假设：**误差项ϵ_i之间相互独立。自变量x和误差项ϵ相互独立。
- **同方差性假设：**误差项ϵ的方差在不同的x值上是相同的，称为同方差性。
- **正态性假设：** 误差项ϵ服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



线性回归的假设

- **线性关系假设：**因变量y与自变量x之间存在线性关系。
- **独立性假设：**误差项ϵ_i之间相互独立。自变量x和误差项ϵ相互独立。
- **同方差性假设：**误差项ϵ的方差在不同的x值上是相同的，称为同方差性。
- **正态性假设：** 误差项ϵ服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



线性回归的假设

- **线性关系假设：**因变量y与自变量x之间存在线性关系。
- **独立性假设：**误差项ϵ_i之间相互独立。自变量x和误差项ϵ相互独立。
- **同方差性假设：**误差项ϵ的方差在不同的x值上是相同的，称为同方差性。
- **正态性假设：**误差项ϵ服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。

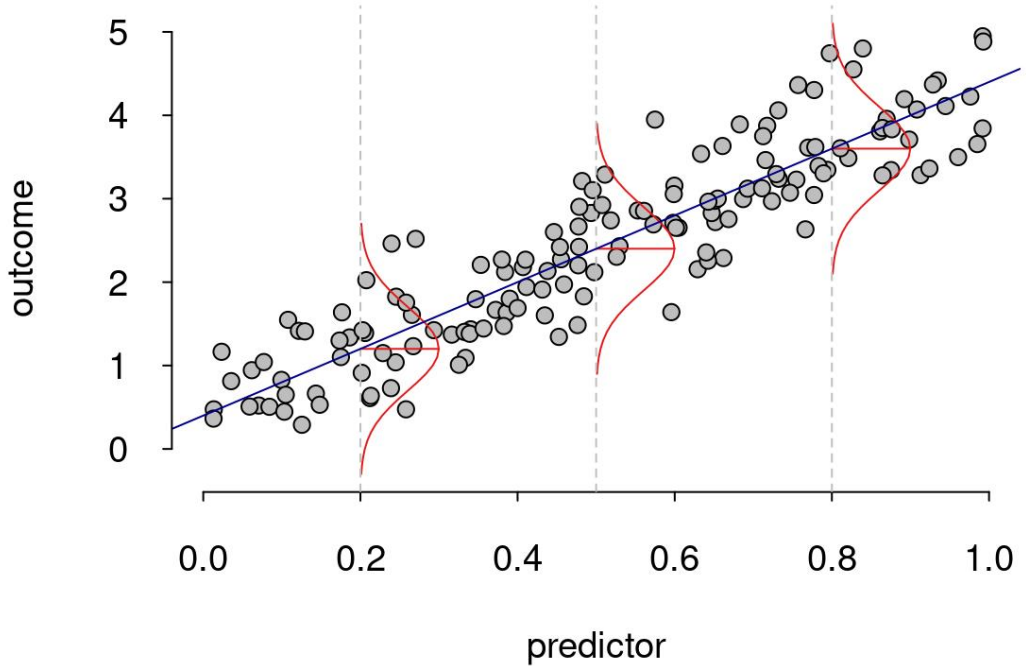
$X_1 \nsim X_2$

$X_1 \sim X_2$

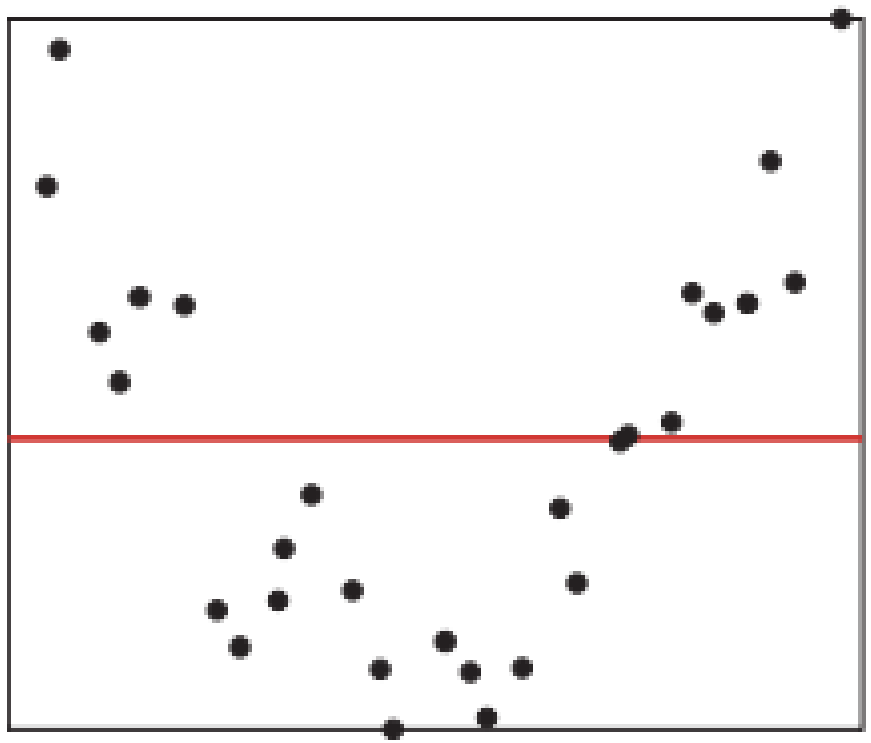
最小二乘法

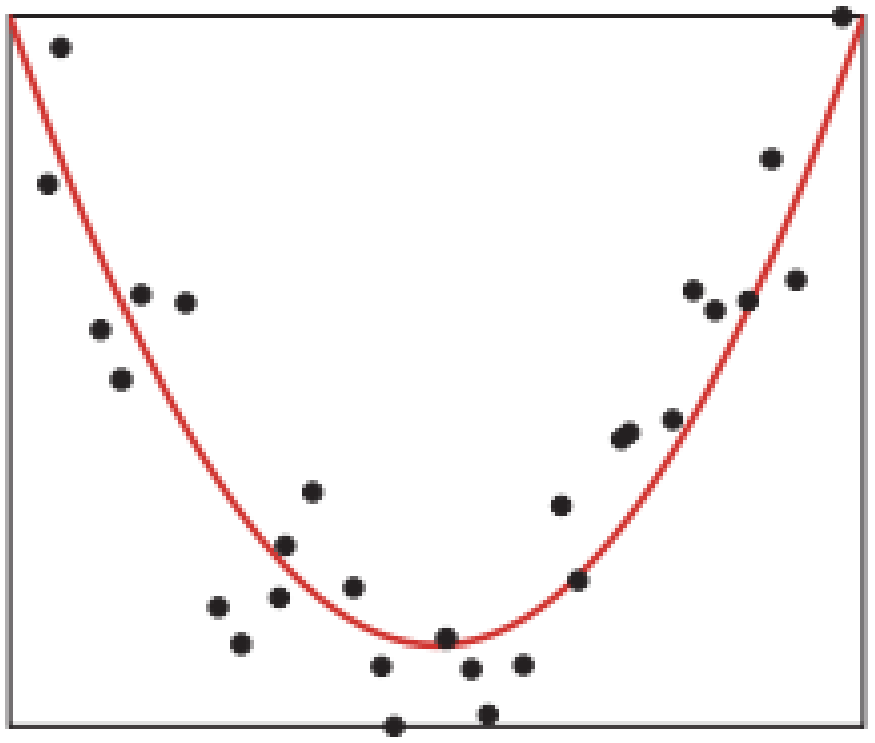
- 最小二乘法（Least Squares Method）：找到一个超平面，使得各个数据点到这个超平面的垂直距离的平方和最小。

- 最小二乘法的有效性严重依赖上述假设!



最小二乘法





线性关系假设不满足直接导致模型失效。

最小二乘法

- 总平方和

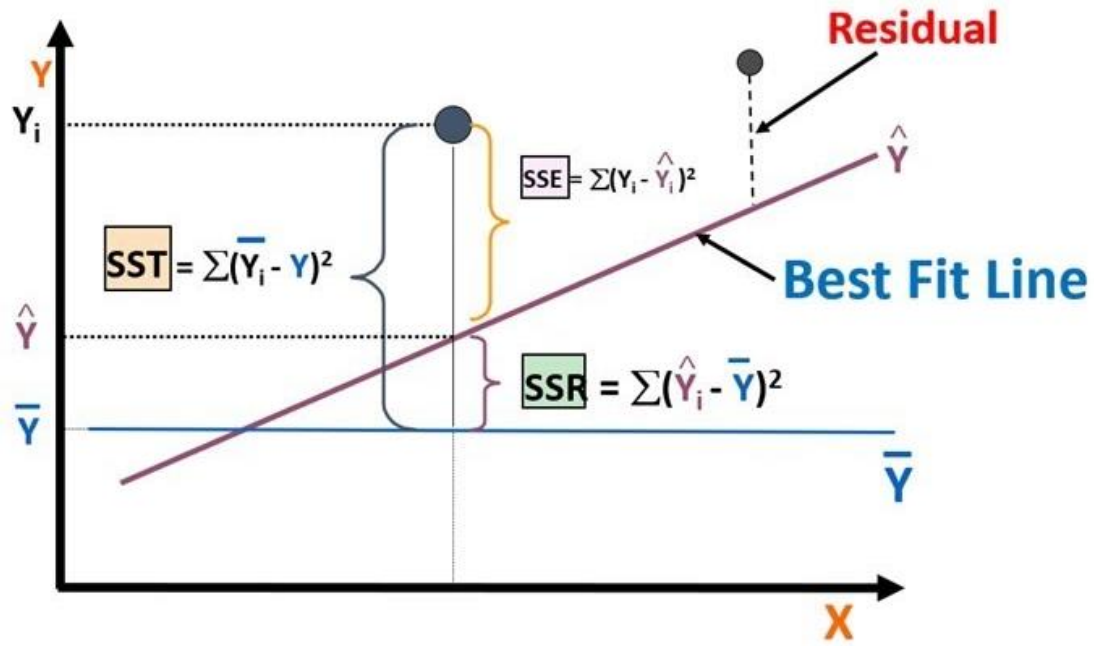
$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- 回归平方和

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

- 残差平方和

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



$$SST = SSR + SSE$$

$$SST = \sum(Y_i - \bar{Y})^2 = SSR = \sum(\hat{Y}_i - \bar{Y})^2 + SSE = \sum(Y_i - \hat{Y}_i)^2$$

最小二乘法

- 最小二乘法的目标是找到一组回归系数$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n$，使得残差平方和

$$Q = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_n x_{in})^2$$最小，其

中m是观测值的数量，$(x_{i1}, x_{i2}, \ldots, x_{in}, y_i)$是第$i$个观测值。

- 回归系数的估计值$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{y}$。

最小二乘法

- 具体求解步骤：

$$
\begin{aligned}
\mathrm{Q}(\boldsymbol{\beta}) &= \frac{1}{\mathrm{n}}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 = \frac{1}{\mathrm{n}}\mathrm{tr}[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}] \\
&= \frac{1}{\mathrm{n}}\mathrm{tr}[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y}^{\mathrm{T}}-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}})] \\
&= \frac{1}{\mathrm{n}}\mathrm{tr}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}-\boldsymbol{y}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{y}^{\mathrm{T}}+\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}]
\end{aligned}
$$

$$
\begin{aligned}
\text{再令}\frac{\partial \mathrm{Q}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\mathrm{n}}\frac{\partial \mathrm{tr}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}-\boldsymbol{y}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{y}^{\mathrm{T}}+\boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}]}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{\mathrm{n}}[0-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}+2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}] \\
&= \frac{2}{\mathrm{n}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}) = 0
\end{aligned}
$$

$$
\text{得}\boldsymbol{\beta}^{\mathrm{OLS}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}
$$

有效性检验

- 回归分析的有效性检验是评估回归模型是否准确、可靠的重要步骤。常见的检验方法：
- **R^2检验、调整后R^2的检验**
- **F 检验**
- **t 检验**
- **回归诊断**
  - **模型假设的检验：正态性、同方差性、独立性检验**
  - **异常数据的发现：**异常值、高杠杆点和强影响力点
  - **自变量关系的分析：**多重共线性问题

R平方（R-squared）

- R平方（R-squared），也称为决定系数（Coefficient of Determination），衡量回归模型对观测数据拟合程度的统计量。它表示因变量的总变异中可以由自变量解释的比例。
- 取值范围在0到1之间。如果$R^2 = 0.6$，这意味着因变量的变异中有60%可以由模型中的自变量来解释，剩下的40%则是由其他未包含在模型中的因素或随机误差所导致。计算公式为：

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST}$$

- 如果其他条件相同，一般会优先选择R^2值较高的模型，因为它能更好地解释因变量的变异。

调整R平方

- R^2显著的缺点：**过度拟合风险。**总是随着自变量个数的增加而增大，即使新增加的自变量在实际中对因变量没有真正的解释作用，R^2也会增加。
- 调整R^2（Adjusted R-squared）是R^2的一种修正，它对自变量的数量进行了惩罚，以防止过拟合。
- 调整R^2的值可能会随着自变量个数的增加而增加、减少或保持不变，这取决于新增加的自变量是否真正对模型的解释能力有贡献。
- 计算公式为：

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

F检验

- F检验（F-test）在回归分析中基于方差分析的原理。我们将因变量的y变异分解为两个部分：
  - 由自变量x的变化所引起的变异：回归平方和SSR;
  - 是不能由自变量解释的变异：残差平方和SSE。
- F检验的基本思想是通过比较数据组内和组间的方差来判断各组之间是否存在显著的差异。计算公式：

$$F = \frac{(SSR/k)}{(SSE/(n-k-1))}$$

- F值越大，自变量对因变量的解释能力越强，模型的整体显著性越高。

F检验

- **零假设：**为所有自变量的回归系数都为0:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

- **备择假设**：至少有一个自变量的回归系数不为0。

- 当拒绝零假设时，说明模型中至少有一个自变量对因变量有显著的解释作用，模型是有效的。

t检验

- 在零假设下，回归分析的总体回归系数为 0。
- t 检验实际上是在检验样本所估计的回归系数与总体回归系数之间的差异是否显著。通过判断样本回归系数是否显著偏离 0，来推断自变量对因变量是否有显著影响。

- **零假设：**单个自变量x_j的回归系数β_j为0: $\beta_j = 0$.

- **备择假设**: $\beta_j \neq 0$。

t检验

- t 统计量的计算公式为

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

$SE(\hat{\beta}_j)$是$\hat{\beta}_j$的标准误差。

- 回归系数的标准误差的计算较为复杂，一般通过以下公式计算:

$$SE(\hat{\beta}_j) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-k-1)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2(1-R_j^2)}}$$

- 与残差平方和、自变量的取值以及自变量之间的相关性有关。

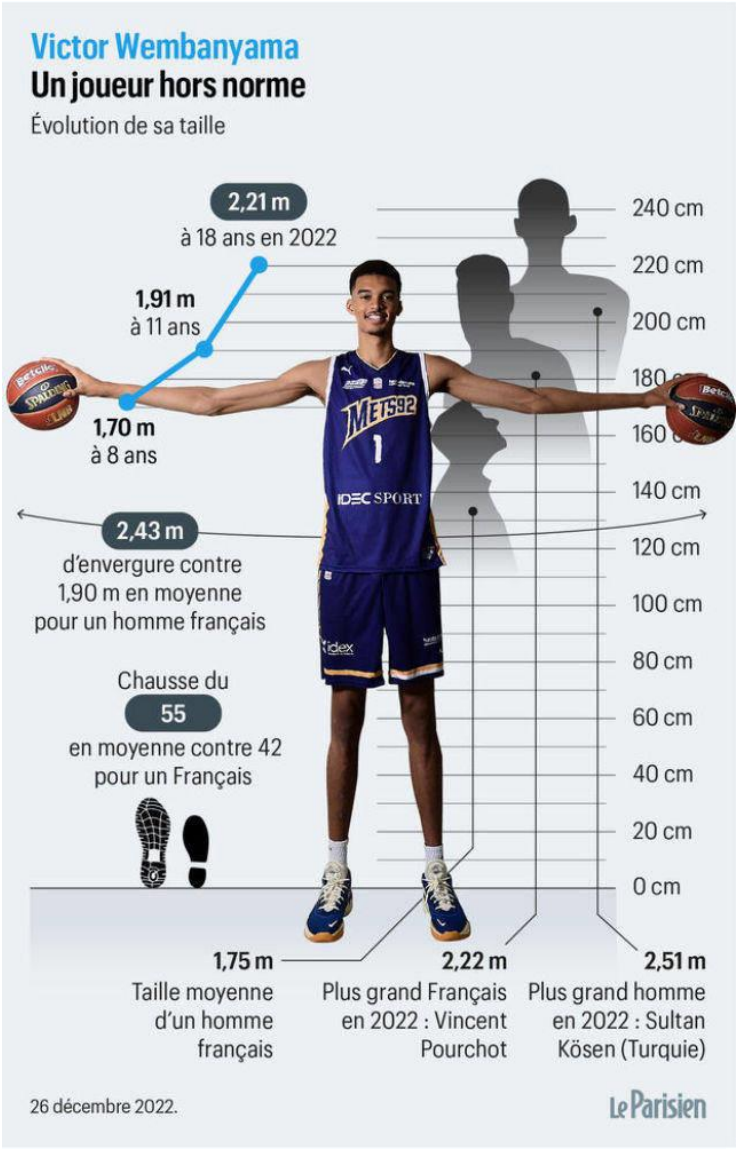
回归诊断

- 对回归模型进行全面评估和分析的一系列过程。
- 确定回归模型的合理性、可靠性和有效性。

- **主要内容包括：**
- **模型假设的检验：**模型假设不成立，可能会导致模型的参数估计不准确、假设检验失效以及预测结果不可靠。
- **异常数据的发现：**特殊的数据点可能对回归模型的参数估计和拟合效果产生较大的影响。
- **自变量关系的分析：**多重共线性可能导致回归系数的估计不稳定、标准误差增大以及对单个自变量的重要性评估出现偏差。

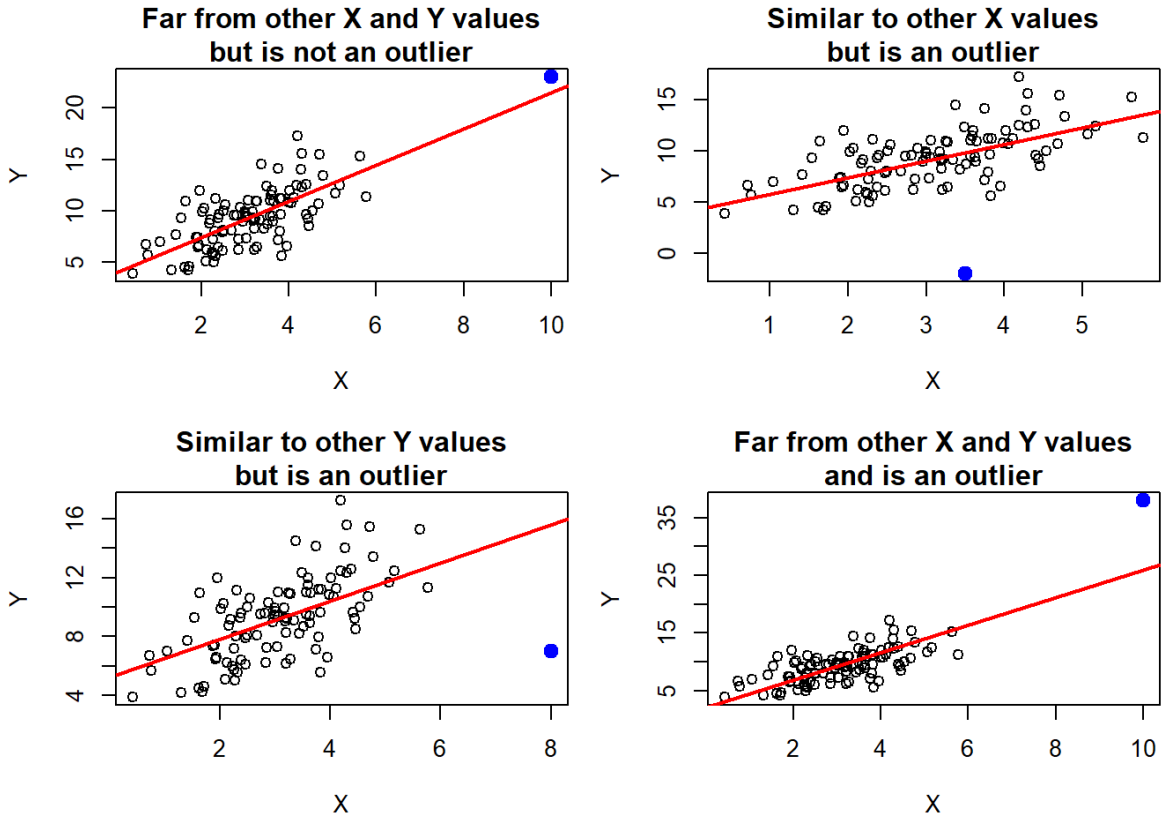
异常值和强影响点检测

- **异常值和强影响点：**与其他观测值相比显著偏离的点。它们可能由于测量错误、数据录入错误或其他原因而与数据集中的其他观测值不一致。



异常值和强影响点检测

- 识别1：**标准化残差**的绝对值大于 3 （通常情况）的观测值常被视为异常值。
- 识别2：Cook’s距离、DFITS等统计量。
- 处理方法：删除、替换、变换数据。



正态性检验

- Shapiro - Wilk 检验：小样本
- Kolmogorov - Smirnov 检验：小样本和大样本
- Q-Q图（Quantile-Quantile Plot）

- **如果正态性检验不通过：**
- 利用 Box - Cox 变换对因变量数据y进行处理。

Box-Cox变换

- 用于改善数据的正态性质，使数据更接近正态分布。
- Box - Cox 变换的公式为：

$$\text{当}\lambda \neq 0\text{时，}\ y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda}$$

$$\text{当}\lambda = 0\text{时，}\ y^{(\lambda)} = \ln(y)$$

- 可以通过极大似然估计等方法来确定λ，使得变换后数据的对数似然函数达到最大，即找到最有可能使数据服从正态分布的变换。

同方差性检验

- Breusch - Pagan 检验：残差平方和与自变量之间的关系。
- White 检验：残差平方和与自变量、自变量的平方项和交叉项之间的关系。

- **如果同方差性检验不通过：**
- 采用加权最小二乘法（WLS）进行修正，即根据误差项方差的不同对观测值赋予不同的权重。

加权最小二乘法

- 目标是最小化加权残差平方和 $\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$

- 对于多元线性回归模型，

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

- 求解正规方程组 $(X^T W X)\hat{\beta} = X^T W y$ 得到回归系数的估计值$\hat{\beta}$。W是对角矩阵，其对角元素为$w_1, w_2, \ldots, w_n$

独立性检验

- Durbin - Watson 检验：
  - 时间序列数据，
  - 残差项的一阶自回归模型。
  - 检验相邻误差项之间是否存在自相关性
  - 可以配合残差的自相关图（观察残差在不同时间点或不同观测值之间的相关性）进行辅助验证。
- Breusch - Godfrey 检验：
  - 时间序列数据或横截面数据，
  - 残差项的多元回归模型，包括了残差的滞后项、自变量

独立性检验

- **如果独立性检验不通过：**

- 采用自回归模型（AR）、移动平均模型（MA）或自回归移动平均模型（ARMA）等来处理。

多重共线性检验

- 检验模型中自变量之间是否存在高度相关性，
  - 方差膨胀因子（Variance Inflation Factor, VIF）
  - 条件指数（Condition Index）等方法。

- **如果多重共线性检验不通过：**
  - **删除变量：**某个自变量的VIF值很高，可以考虑从模型中删除。
  - **变量转换：**对自变量进行对数、平方根或差分，或能降低相关性。
  - **正交化*：**对自变量矩阵Gram-Schmidt正交化，得到线性无关的自变量矩阵。
  - **逐步回归*、岭回归***或Lasso**回归***等方法来选择重要的变量或转换变量。

实验1：数据分布分析及数据处理

- 分析收益率、市净率(pb)、市盈率(pe_ttm)、净资产回报率(roe)、总市值(total_mv)的数据分布类型，并对数据进行处理
- 步骤1：载入最后一期的收益率、市净率(pb)、市盈率(pe_ttm)、净资产回报率(roe)、总市值(total_mv)数据。
- 步骤2：计算数据的各项统计量指标，绘制上述数据的直方图，判断数据符合哪种类型的分布。
- 步骤3：基于数据的分布类型，选择正确的方法，对数据进行标准化处理。
- 步骤4：绘制标准化后的数据直方图，与原始数据的直方图进行对比。

实验2：bp因子的构建与分析

- bp因子简介：
- 在金融中，PB 即市净率（Price-to-Book Ratio），而 bp 因子就是基于市净率的一种投资分析因子。市净率指的是每股股价与每股净资产的比率。计算公式为：
  - 市净率 = 每股市价 / 每股净资产
- 市净率反映了公司股票的市场价值与账面价值之间的关系。市净率较低，意味着投资者可以用相对较低的价格买入公司的资产。这可能表明公司被低估，或者市场对公司的前景较为悲观。市净率较高，则说明投资者为获得公司的资产需支付较高的价格。可能意味着公司具有较强的竞争力和良好的发展前景，市场对其较为看好，但也可能存在高估的风险。

实验2：bp因子的构建与分析

- 注意：
- 很多行业的利润往往呈现周期性，在景气期间，行业的盈利水平高涨，使得行业市盈率大幅下降，而市净率的变化较小。采用市净率估值可以避免因盈利大幅变化导致的风格偏移问题。
- 不同行业的市净率水平差异较大，这是由于行业的特点和发展阶段不同所致。例如，传统制造业的市净率通常较低，因为这些行业的资产较重，盈利能力相对较弱；而高科技行业的市净率往往较高，因为这些行业具有较高的成长性和创新性，市场对其未来的盈利能力有较高的预期。因此，不同行业公司之间的bp因子往往不适合直接进行比较。

实验2：bp因子的构建与分析

- 因子分析目标:
- 分析bp因子与股票涨跌幅之间的相关性，从而指导投资组合构建的决策过程。

实验2：bp因子的构建与分析

- 步骤1，数据载入及因子计算：
  - 载入股票的pb值数据
  - 计算bp因子，即pb值的倒数。
    - bp因子 $= \frac{1}{pb值}$
  - 将bp因子数据调整到周度频率

实验2：bp因子的构建与分析

- 步骤2，数据的预处理：
  - **去空值：**将数据缺失的地方设为上一期的数据值
  - **去极值：**将每期bp因子的数据采用winsor方法去极值，区间为$[\mu - 3\sigma, \mu + 3\sigma]$。
  - **标准化：**将去极值处理后的因子序列进行z-score标准化处理。即，减去其现在的均值、除以其标准差，得到一个新的近似服从N(0,1)分布的序列
  - **行业中性化：**基于申万1级行业分类标准，对经过上述步骤处理后的因子值进行行业中性化处理

实验2：bp因子的构建与分析

- 步骤3，因子评价1：
  - 以预处理后的t期bp因子作为解释变量，t+1期的收益率数据作为被解释变量，采用加权最小二乘回归（WLS），使用个股市值(total_mv)的平方根作为权重，旨在消除异方差性。

  - 评价指标：
    - t 值序列绝对值平均值
    - t 值序列绝对值大于 2 的占比
    - t 值序列均值的绝对值除以 t 值序列的标准差

实验2：bp因子的构建与分析

- 步骤3，因子评价2：
  - 计算因子 IC 值：因子的 IC 值是指因子在第 T 期的暴露度与 T+1 期的股票收益的相关系数，即

$$IC_d^T = \mathrm{corr}(R^{T+1}, d^T)$$

  - 其中，IC_d^T代表因子 d 在第 T 期的 IC 值，R^{T+1}代表所有个股第 T+1 期的收益率向量，d^T代 表所有个股第 T 期在因子 d 上的暴露度向量。
  - normal IC值：采用Spearman相关性系数计算的IC值
  - rank IC值：采用Pearson相关性系数计算的IC值
  - 评价指标： IC 值序列的均值大小、 IC 值序列的标准差、 IC 值累积曲线、IC 值序列大于零的占比

实验2：bp因子的构建与分析

- 步骤4：可视化
  - 将上述评价指标绘制图表进行展示