intro

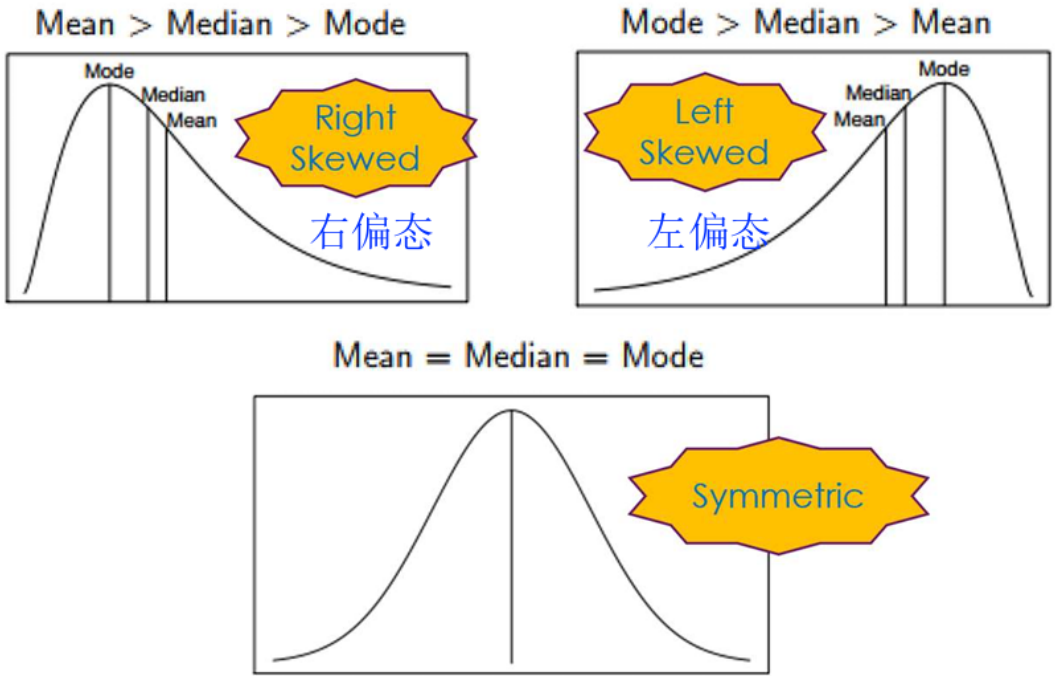
术语

- rv=random variable 随机变量
  - =X
- variation=difference
- data
  - numerical data=quantitative
    - continuous data
    - discrete data
  - categorical data=qualitative
    - nominal无序
    - ordinal
    - interval
  - count计数资料
  - survival data生存资料
- data collection methods
  - observation
  - interviews& questionnaire
  - documentary sources
  - public data service

EDA 探索性数据分析exploratory data analysis

- （文图表）
- descriptive summary

- central tendency
  - mean *sensitive*
  - median *not sensitive*
  - mode

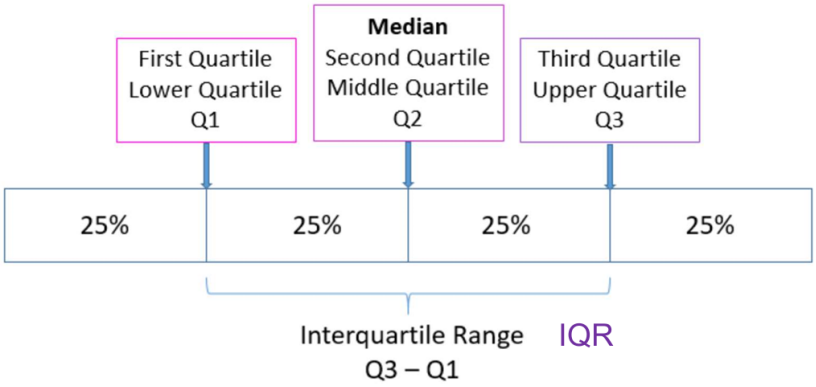


- dispersion离差
  - variance
    - standard deviation
    - (SS) 离均差平方和

$$\text{Variance } \sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

$$\text{Standard deviation } \sigma = \sqrt{\sigma^2}$$

  - range
  - interquartile range四分位数间距 IQR



data in **ascending** order

- relative position
  - Percentile • 百分位数
  - Percentages of data values are less than ***or equal to*** the r th percentile
  - Standard Z-score标准化
- **for categorical rvs** rate and proportion

- frequency table同下
- graphs

probility and distributions一组实验

Linear Functions of Random Variables and Joint distribution组合-多组不同实验

- X_1 calcium (mg)
- X_2 iron (mg)
- X_3 protein(g)
- X_4 vitamin A(μg)
- X_5 vitamin C(mg)

$$Y = 0.001X_4 + X_5$$

mean& sd

cov

- Given random variables $X_1, X_2, \ldots, X_p$ and constants $c_1, c_2, \ldots, c_p$,

$$Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

is a linear combination of $X_1, X_2, \ldots, X_p$.

- **Mean** of a Linear Function

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \cdots + c_pE(X_p)$$

- **Variance** of a Linear Function

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_p^2V(X_p) + 2\sum_{i<j}\sum c_ic_jcov(X_i, X_j)$$

Linear Functions of Random Variables and Joint distribution组合-多组相同实验

mean（不变）& sd（变小）

$$\text{Let } \bar{X} = \frac{(X_1 + X_2 + \cdots + X_p)}{p} = \frac{1}{p}X_1 + \frac{1}{p}X_2 + \cdots \frac{1}{p}X_p$$

??? 无分布要求

$\bar{X}$ is as mean and it is a linear combination of the p random variable we observed.

Because $E(X_i) = \mu$ for $i = 1, 2, \ldots, p$ we have $E(\bar{X}) = \mu$.

If $X_1, X_2, \ldots, X_p$ are also independent and all with the same variance of $V(X_i) = \sigma^2$ then

$$V(\bar{X}) = \frac{\sigma^2}{p}$$

Normal Dist

X_i are independent r.v. with normal distribution.

If $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$ for $i = 1, 2, \ldots, p$.

$$Y = c_1X_1 + c_2X_2 + \cdots + c_pX_p$$

is a normal random variable with

$$\mu_Y = E(Y) = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$$

and

$$\sigma_Y^2 = V(Y) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \cdots + c_p^2\sigma_p^2$$

i.e. $Y \sim N(\mu_Y, \sigma_Y^2)$ as described above.

N（μ，σ）形式相同，数值变化

statistics

Estimation

hypothesis testing

Sample size calculation

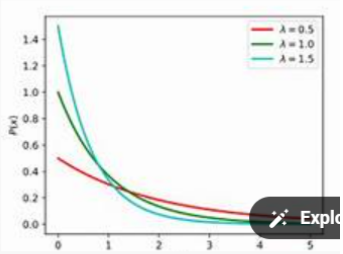
extra

在统计学中，“iid"是"Independent and Identically Distributed"的缩写，中文通常翻译为"独立同分布”。这个术语通常用于描述一组随机变量，如果满足以下两个条件，那么这组随机变量就被称为iid[1]：

1. **独立（Independent）**：在抽样的上下文中，事件是独立的，意味着观察当前的项目并不影响或提供关于你将要测量的下一个项目的值的任何信息，或者样本中的任何其他项目[2]。换句话说，它们之间没有任何联系[2]；知道一个变量的值并不能提供关于另一个变量值的任何信息，反之亦然[1]。
2. **同分布（Identically Distributed）**：同分布意味着所有的随机变量都来自同一个概率分布[1]。也就是说，样本中的所有项目都来自同一个概率分布[2]。

这两个属性在统计学中有重要的应用，例如在中心极限定理中，iid假设常常被使用，该定理指出，具有有限方差的iid变量的和（或平均值）的概率分布趋近于正态分布[1]。希望这个解释对你有所帮助！

指数分布（Exponential Distribution）是概率论和统计学中的一个重要概念，它常常用于描述某个特定事件发生的时间间隔的概率分布[1]。例如，地震发生的时间间隔、生产错误之间的时间间隔、或者织布生产过程中布卷长度等，都可以使用指数分布来描述[1]。



指数分布有两个关键特性[1]：

1. **无记忆性（Memorylessness）**：如果一个随机变量X服从指数分布，那么无论已经等待了多长时间，从现在开始再等待一段时间直到下一个事件发生的分布与原来的分布是一样的[1]。
2. **独立同分布（Independent and Identically Distributed）**：在泊松点过程中，事件连续且独立地以恒定的平均速率发生[1]。

指数分布的概率密度函数（Probability Density Function）为[2]：

$$f(x|\lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{for } x > 0 \\ 0, & \text{for } x \leq 0 \end{cases}$$

其中，λ被称为分布的速率参数[2]。

指数分布的均值（Mean）和方差（Variance）分别为[2]：

- 均值：$E[X] = \frac{1}{\lambda}$
- 方差：$Var(X) = \frac{1}{\lambda^2}$

希望这个解释对你有所帮助！

fbs of hw

“前一日睡觉时间”是离散型变量因为问卷里只有10：00 10：30 这种 离散的 选项，没法取到任意值

$$\sigma_X = E(X^2) - E^2(x)$$