

## 2、数据科学(Data Science)

### 2.1 数据统计(Statistics)

#### 2.1.1 概率论(Probability Theory)

- (1) 随机性、随机变量和随机样本(Randomness, Random Variable and Random Sample)
- (2) 概率分布(Probability Distribution)
- (3) 条件概率和贝叶斯定理(Conditional Probability and Bayes's Theorem)
- (4) 统计独立性(Statistical Independence)
- (5) 独立同分布(IID, Independent and identically distributed)
- (6) cdf, pdf, pmf

1. 累积分布函数(cdf, Cumulative Distribution Function)
2. 概率密度函数(pdf, Probability Density Function)
3. 概率质量函数(pmf, Probability Mass Function)

#### 2.1.2 连续分布(Continuous Distributions)

- (1) 正态分布及高斯分布(Normal/Gaussian)
- (2) 一致连续性(Uniform Continuous)
- (3) Beta 分布(Beta Distributions)
- (4) 狄利克雷分布(Dirichlet Distribution)
- (5) 指数型分布(Exponential Distribution)
- (6) 卡方(chi-square)

#### 2.1.3 离散分布(Discrete Distributions)

- (1) 均匀离散分布(uniform discrete)
- (2) 二项式分布(Binomial Distribution)
- (3) 多项式分布(Multinomial Distribution)
- (4) 超几何分布(Hypergeometric Distribution)
- (5) 泊松分布(Poisson Distribution)
- (6) 几何分布(Geometric Distribution)

#### 2.1.4 汇总统计(Summary Statistics)

- (1) 期望和均值(expectation and mean)
- (2) 方差和标准差(Variance and Standard)
- (3) 协方差与相关性(Covariance and Correlation)
- (4) 中位数与四分位数(Median, Quartile)
- (5) 四分位数范围(Interquartile Range)
- (6) 百分位数/分位数(Percentile/Quantile)
- (7) 众数(Mode (statistics))

#### 2.1.5 重要规则(Important Laws)

- (1) 大数法则(Law of Large Numbers)
- (2) 中心极限定理(Central Limit Theorem)

#### 2.1.6 估计(Estimation)

- (1) 极大似然估计(Maximum Likelihood Estimation)
- (2) 核密度估计(Kernel Density Estimation)

#### 2.1.7 假设检验(Hypothesis Testing)

- (1) P值 (p-value)
- (2) 卡方检验(chi-square-test)
- (3) F检验(F-test)
- (4) t 检验(t-test)

#### 2.1.8 置信区间(Confidence Interval)

#### 2.1.9 蒙特卡罗法(Monte Carlo Method)

### 2.2 可视化(Visualization)

#### 2.2.1 图表建议(Chart Suggestions)

#### 2.2.2 Python库

- (1) Matplotlib
- (2) Plotnine(like ggplot in R)
- (3) Bokeh
- (4) Seaborn
- (5) ipyvolume(3D data)

#### 2.2.3 Web技术

- (1) Vega-Lite
- (2) D3.js

#### 2.2.4 仪表盘(Dashboards)

Dash

#### 2.2.5 BI

- (1) Tableau
- (2) PowerBI