统计学笔记

统计学分类

- 描述统计
- 推论统计

基本概念

- 均值： 求平均值
- 中位数： 有序序列的中间值（个数为偶数时求中间两位平均字值）
- 众数： 次数出现最多元素为众数
- 总体均值： $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
- 样本均值： $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- 总体方差： $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
  - 简化公式 $\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$
- 样本方差： $S^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}$
  - 样本方差是来估计总体方差，由此有无偏样总体方差
  - 无偏样总体方差， $S^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$
- 总体标准差： $\sigma = \sqrt{\sigma^2}$
  - 样本标准差 $S = \sqrt{S^2}$
- 随机变量： 跟传统变量不是一个概念（连续随机变量，离散随机变量）
  - 随机过程映射到数值的函数
  - 数值是随机的
- 概率分布函数： 描述离散随机变量的概率
- 概率密度函数： 描述连续随机变量的概率
- 期望值： $E(x) = \sum_{k=1}^{\infty} x_k p_k$
  - p(x) 为该随机变量的概率值
  - 期望值就是该随机变量总体的均值
  - 当要计算总体的均值(μ)时候，总体数据量大(无穷)，又知道该随机变量概率函数，就可以计算期望值，得到总体均值

二项分布

概念

> 1.在每次试验中只有两种可能的结果，而且是互相对立的；
> 2.每次实验是独立的，与其它各次试验结果无关；
> 3.每次发生的概率不变；

概率公式

- $p(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$

二项分布期望值

- $E(X)=np$
    - n 为实验次数
    - p 为事件概率
    - 期望值可以看成最可能得到的那个结果

泊松分布

概率密度函数

- $p(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$
    - λ为期望值
    - 来源于二项分布，当二项分布的n很大而p很小时，泊松分布可作为二项分布的近似，其中λ为np
    - 概率密度函数有二项分布概率密度函数求极限推出，$\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k}$

大数定律

> - 随机变量的N次观察, 将所有观测值平均起来, 得到样本平均值，当实验次数足够多或趋于无穷，样本的平均值会趋近于随机变量的期望值
> - $\overline{x}=E(x)$

正态分布

> - 重复多次独立事件，取平均值为新的随机变量， 新的随机变量的新的概率密度函数符合正态分布
> - 二项分布实验次数足够多会趋近于正态分布

概率密度函数

- $f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- 标准正态分布概率密度函数
  - 当 $\mu = 0$, $\sigma = 1$时
  - $f(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

z分数

- z分数就是离均值有多少个标准差远
- $z = \frac{x-\mu}{\sigma}$

经验法则

- 68 - 95 - 99.7
  - 一个标准差范围的经验概率为 68%
  - 两个标准差范围的经验概率为 95%
  - 三个标准范围的经验概率为 99.7%