深入浅出统计学

信息图形化

统计：通过某种有意义的方式对原始事实和数字进行提炼，使得仅仅通过观察原始数据无法立即水落石出的一些理念得以昭示

统计量是样本的函数
在样本未知的情况下是随机变量

对统计的研究：

- 统计数据来源
- 计算方法
- 有效使用方法得出结论

数据是指所搜索的原始事实与数字
信息是指加上某种意义的数据

- 饼图
- 条形图
  - 堆积条形图
  - 分段条形图
- 直方图
  - 长方形的面积与频数成比例
  - 图上的长方形之间没有间隔
  - 分组数值型数据
- 累积频数图
- 折线图

集中趋势的度量

均值：$\mu = \frac{\sum x}{n}$
中位数

- 数据右倾：均值>中位数
- 数据左倾：均值<中位数

众数

分散性与变异性的度量

分散性

极差/全距 = 上界 - 下界
迷你距：忽略异常值

- 四分位数
  - 最小四分位数 下四分位数 第一四分位数
  - 最大四分位数 上四分位数 第三四分位数
  - 中间的四分位数 中位数
- 四分位距 = 上四分位数 - 下四分位数
  - 与全距相比较少受到异常值的影响
  - 得到一种对几个数据集进行比较且比较结果不会被异常值扭曲的方法
- 下四分位数的位置
  - 首先计算 $\frac{n}{4}$

    - 如果结果为整数，则下四分位数位于$\frac{n}{4}$这个位置和下个位置的中间
    - 如果不是整数，则向上取整
- 上四分位数的位置
    - 首先计算 $\frac{3n}{4}$
    - 如果结果为整数，则下四分位数位于$\frac{3n}{4}$这个位置和下个位置的中间
    - 如果不是整数，则向上取整
- 百分位数---将数据一分为百的数值
    - 第k百分位数就是位于数据$k\%$处的数值，P_k
    - 求百分位数
        - 首先将数值按升序排序
        - 为了求出n个数字的第k百分位数的位置，先计算$k\left(\frac{n}{100}\right)$
        - 如果结果为整数，则百分位数处于第$k\left(\frac{n}{100}\right)$位和下一位数之间。取这两个位置上的数字平均值，得出百分位数
        - 如果不是整数，则将其向上取整，结果即为百分位数的位置
- 箱线图
    - 全距
    - 四分位距
    - 中位数

变异性

方差：$\sigma^2=\frac{\sum(x-\mu)^2}{n}=\frac{\sum x^2}{n}-\mu^2$

标准差：$\sigma=\sqrt{方差}$

概率计算

- 概率是度量某事发生几率的一种数量指标
- 事件：有概率可言的一个结果或一件事
- $P(A)=\dfrac{n(A)}{n(S)}$
    - S样本空间
    - 事件是S的子集
- 对立事件
    - $P(A') = 1 - P(A)$
- 互斥事件
    - $P(A\cap B)=0$
- 相交事件
- $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
- 条件概率
    - $P(A|B)=\dfrac{P(A\cap B)}{P(B)}$
    - $P(A\cap B)=P(A|B)P(B)$
    - $P(B\cap A)=P(B|A)P(A)$
- 全概率公式
    - $P(B)=P(A)P(B|A)+P(A')P(B|A')$
- 贝叶斯定理
    - $P(A|B)=\dfrac{P(A)P(B|A)}{P(A)P(B|A)+P(A')P(B|A')}$
- 相关事件
- 独立事件
    - $P(A|B)=P(A)$
    - $P(A\cap B)=P(A)P(B)$

离散概率分布的运用

- 期望：$E(X)=\sum xP(X=x)$
    - $E(X)=\mu$

- 方差：$Var(X) = E(X-\mu)^2 = E(X^2) - \mu^2$
- 标准差：$\sigma = \sqrt{Var(X)}$
- $E(aX+b) = aE(X)+b$
- $Var(aX+b) = a^2Var(X)$
- 独立观察值：
  - $E(X_1+X_2+\cdots+X_n) = nE(X)$
  - $Var(X_1+X_2+\cdots+X_n) = nVar(X)$
- X和Y是相互独立的随机变量
  - $E(X+Y) = E(X)+E(Y)$
  - $Var(X+Y) = Var(X)+Var(Y)$
  - $E(X-Y) = E(X)-E(Y)$
  - $Var(X-Y) = Var(X)+Var(Y)$

排列与组合

- n个对象排列数目：$n! = n*(n-1)*\cdots*3*2*1$
- n个对象作圆形排位：$(n-1)!$
- 按类型排位：
  - 第一类对象k个，第二类对象j个，第三类对象m个...
  - $\frac{n!}{j!\,k!\,m!\ldots}$
- 排列：与顺序有关
  - $P_n^r = \frac{n!}{(n-r)!}$
- 组合：与顺序无关
  - $C_n^r = \frac{n!}{r!\,(n-r)!}$

几何分布、二项分布及泊松分布

- 几何分布
  - X表示取得第一次成功所需进行的试验次数
  - $P(X=x) = q^{x-1}p$
  - $P(X>x) = q^x$
  - $X \sim Geo(p)$
  - $E(X) = \frac{1}{p}$
  - $Var(X) = \frac{q}{p^2}$
- 二项分布
  - X表示n次试验中的成功次数
  - $P(X=x) = C_n^x p^x q^{n-x}$
  - $X \sim B(n,p)$
  - $E(X) = np$
  - $Var(X) = npq$
- 泊松分布
  - 泊松分布满足以下条件
    - 单独事件在给定区间内随机、独立的发生
    - 已知该区间内的事件平均发生次数
  - $X \sim Po(\lambda)$
  - $E(X) = \lambda$
  - $Var(X) = \lambda$
  - $P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$
- 泊松分布近似二项分布
  - $\lambda \sim np$
  - $\lambda \sim npq$
  - 当n很大且p很小时，可以用$X \sim Po(np)$代替$X \sim B(n,p)$

正态分布的运用

正态分布

- $X \sim N(\mu, \sigma^2)$

再谈正态分布的运用

二项分布的近似

如果$X \sim B(n,p)$，且$np > 5$，$nq > 5$，则可以使用$X \sim N(np, npq)$近似代替二项分布

泊松分布的近似

如果$X \sim Po(\lambda)$且$\lambda > 15$，则可用$X \sim N(\lambda, \lambda)$进行近似

统计抽样的运用

- 总体
  - 指的是对其进行测量、研究或分析的整个群体
- 样本
  - 从样本中选取的一部分对象
- 抽样单位
- 抽样空间
  - 列出总体中的所有独立单位
- 抽样方法
  - 简单随机抽样
    - 重复抽样
    - 不重复抽样
  - 分层抽样
  - 整群抽样
  - 系统抽样

总体和样本的估计

- 点估计量
  - 可用于估计总体参数数值的某个函数或算式
- 样本均值
  - $\bar{x} = \frac{\sum x}{n}$
- 样本均值估计总体均值
  - $\hat{\mu} = \bar{x}$
- 估计总体方差
  - $\widehat{\sigma^2} = \frac{\sum (x - \bar{x})^2}{n - 1}$