信息学中的概率统计

王若松

前沿计算研究中心
北京大学

参数估计

1. 统计学的基本概念
2. 点估计
3. 区间估计

1. 统计学的基本概念

- 概率论：假设概率分布已知
- 统计学：利用观察到的数据，推断出数据背后的分布的性质

- 例：给定广告的点击数据，如何估计广告的平均点击率？

- **总体**：研究对象的全体
- **个体**：总体的每个成员

- 通常只关心个体的数量指标值
  - 用概率分布描述总体，数量指标值为服从该分布的随机变量X

1. 统计学的基本概念

- 从总体中随机地抽取n个个体$x_1, x_2, \dots, x_n$
- **样本**：$x_1, x_2, \dots, x_n$
- **样本量**：n

- 样本是随机变量，也写作$X_1, X_2, \dots, X_n$，X_i与总体X分布相同

- **简单随机样本**：$X_1, X_2, \dots, X_n$相互独立，且均服从与总体X相同的分布
- 若X的分布函数为$F(x)$，样本的联合分布函数为$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} F(x_i)$

- 统计学：利用样本$x_1, x_2, \dots, x_n$，研究总体X的性质

1. 统计学的基本概念

- **统计量**：不依赖于任何未知参数的样本的函数

- 样本均值：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$
- k阶矩：$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
- k阶中心矩：$B_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$
- 其他统计量：$\min\{X_1, X_2, \ldots, X_n\}$，$\max\{X_1, X_2, \ldots, X_n\}$，中位数

2. 点估计

- **点估计**：估计分布中所含有的未知参数θ，如$\theta = E(X)$和$\theta = \mathrm{Var}(X)$
- 例：给定广告的点击数据，如何估计广告的平均点击率?
  - 给定样本$x_1, x_2, \ldots, x_n$，如何估计$\theta = E(X)$?
- **估计量**：用来估计未知参数θ的统计量$\hat{\theta}(X_1, X_2, \ldots, X_n)$
- **估计值**：给定样本取值$x_1, x_2, \ldots, x_n$，$\hat{\theta}(x_1, x_2, \ldots, x_n)$为估计值
- 如何评判估计量的好坏?

2. 点估计

- 已知总体X数学期望$E(X)$和方差$\mathrm{Var}(X)$均存在

- 考虑$\theta = E(X)$的下述估计量
- $\hat{\theta}_A(X_1, X_2, \dots, X_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $\hat{\theta}_B(X_1, X_2, \dots, X_n) = X_1$
- $\hat{\theta}_C(X_1, X_2, \dots, X_n) = \frac{1}{n+1}\sum_{i=1}^{n} X_i$

- 哪个估计量最好?

2. 点估计

- 给定参数θ的估计量$\hat{\theta}\,(X_1, X_2, \ldots, X_n)$，定义**偏差**$\mathrm{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$
- 如果$\mathrm{Bias}(\hat{\theta}) = 0$，则称估计量$\hat{\theta}$是**无偏**的

- 当$n = 1$时
- $\hat{\theta}_A = \hat{\theta}_B = X_1$，$\hat{\theta}_C = \frac{1}{2} X_1$
- $E(X_1) = E(X)$

- 如果$\lim\limits_{n\to\infty} E(\hat{\theta}) = \theta$，则称估计量$\hat{\theta}$是**渐进无偏**的

2. 点估计

- 已知总体X数学期望$E(X)$和方差$\mathrm{Var}(X)$均存在，考虑$\theta=E(X)$的下述估计量
- $\hat{\theta}_A(X_1,X_2,\dots,X_n)=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$
- $\hat{\theta}_B(X_1,X_2,\dots,X_n)=X_1$

- 给定参数θ的估计量$\hat{\theta}$，定义**均方误差**$\mathrm{MSE}(\hat{\theta})=E\left(\left(\hat{\theta}-\theta\right)^2\right)$
- 若估计量$\hat{\theta}$无偏，则$\mathrm{MSE}(\hat{\theta})=\mathrm{Var}(\hat{\theta})$
  - $\mathrm{MSE}(\hat{\theta})=E\left(\left(\hat{\theta}-\theta\right)^2\right)=E\left(\left(\hat{\theta}-E(\hat{\theta})\right)^2\right)=\mathrm{Var}(\hat{\theta})$

2. 点估计

- 已知总体X数学期望$E(X)$和方差$\mathrm{Var}(X)$均存在，考虑$\theta = E(X)$的下述估计量
- $\hat{\theta}_A(X_1, X_2, \ldots, X_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $\hat{\theta}_B(X_1, X_2, \ldots, X_n) = X_1$

- 给定参数θ的估计量$\hat{\theta}$，定义**均方误差**$\mathrm{MSE}(\hat{\theta}) = E\left(\left(\hat{\theta} - \theta\right)^2\right)$
- 若估计量$\hat{\theta}$无偏，则$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$
- 计算$\hat{\theta}_A$和$\hat{\theta}_B$的均方误差
  - $\mathrm{MSE}(\hat{\theta}_A) = \mathrm{Var}(\hat{\theta}_A) = \mathrm{Var}(X)/n$
  - $\mathrm{MSE}(\hat{\theta}_B) = \mathrm{Var}(\hat{\theta}_B) = \mathrm{Var}(X)$

2. 点估计

- 给定参数θ的估计量$\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \dots, X_n)$，若有$\hat{\theta}_n \xrightarrow{P} \theta$，也即对于任意$\epsilon > 0$，$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right) = 0$，则称估计量 $\hat{\theta}_n$为**一致估计量**

- $\hat{\theta}_A(X_1, X_2, \dots, X_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $\hat{\theta}_B(X_1, X_2, \dots, X_n) = X_1$
- $\hat{\theta}_C(X_1, X_2, \dots, X_n) = \frac{1}{n+1}\sum_{i=1}^{n} X_i$

- 哪个估计量是一致的?

2. 点估计

- 若 $\lim_{n\to\infty} \mathrm{MSE}(\hat{\theta}_n) \to 0$，则$\hat{\theta}_n$为一致估计量
  - 对于任意$\epsilon > 0$，$P(|\hat{\theta}_n - \theta| \geq \epsilon) = P\left((\hat{\theta}_n - \theta)^2 \geq \epsilon^2\right) \leq \mathrm{MSE}(\hat{\theta}_n)/\epsilon^2$
  - 左右两侧取极限即可

- 有偏估计量是否为一致估计量?
- 渐进无偏估计量是否为一致估计量?
- 一致估计量是否为渐进无偏估计量?

2. 点估计

- 令总体 $X \sim U(0,\theta)$
- 判断估计量 $\hat{\theta}_A = 2\overline{X}$和 $\hat{\theta}_B = \max\{X_1, X_2, \dots, X_n\}$的无偏性

- $E(\hat{\theta}_A) = 2 \cdot E(X) = \theta$
- $F_{\hat{\theta}_B}(x) = \left(F_X(x)\right)^n = \left(\frac{x}{\theta}\right)^n$ 当 $0 \le x \le \theta$
- $E(\hat{\theta}_B) = \int_0^{+\infty}(1 - F(x))dx = \int_0^{\theta}\left(1 - \left(\frac{x}{\theta}\right)^n\right)dx = \frac{n}{n+1}\theta$
- $\hat{\theta}_B$渐进无偏
- $\hat{\theta}_C = \max\{X_1, X_2, \dots, X_n\} \cdot \frac{n+1}{n}$是无偏估计量

PEKING UNIVERSITY 北京大学

2. 点估计

- 令总体 $X \sim U(0, \theta)$，比较 $\hat{\theta}_A = 2\overline{X}$和 $\hat{\theta}_C = \max\{X_1, X_2, \ldots, X_n\} \cdot \frac{n+1}{n}$的均方误差
- $\mathrm{MSE}\left(\hat{\theta}_A\right) = \mathrm{Var}\left(\hat{\theta}_A\right) = \frac{\mathrm{Var}(2X)}{n} = \frac{\theta^2}{3n}$

PEKING UNIVERSITY 北京大学

2. 点估计

- 令总体 $X \sim U(0,\theta)$，比较 $\hat{\theta}_A = 2\overline{X}$和 $\hat{\theta}_C = \max\{X_1, X_2, \dots, X_n\} \cdot \frac{n+1}{n}$的均方误差
- $\hat{\theta}_B = \max\{X_1, X_2, \dots, X_n\}$
- $F_{\hat{\theta}_B}(x) = \left(F_X(x)\right)^n = \left(\frac{x}{\theta}\right)^n$ 当 $0 \le x \le \theta$
- $F_{\hat{\theta}_B^2}(x) = F_{\hat{\theta}_B}(\sqrt{x}) = \frac{x^{\frac{n}{2}}}{\theta^n}$当 $0 \le x \le \theta^2$
- $E\left(\hat{\theta}_B^2\right) = \int_0^{\theta^2} \left(1 - \frac{x^{\frac{n}{2}}}{\theta^n}\right) dx = \frac{n}{n+2}\theta^2$
- $\mathrm{Var}\left(\hat{\theta}_B^2\right) = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2$
- $\mathrm{MSE}\left(\hat{\theta}_C\right) = \mathrm{Var}\left(\hat{\theta}_C\right) = \left(\frac{n+1}{n}\right)^2 \cdot \mathrm{Var}\left(\hat{\theta}_B^2\right) = \frac{\theta^2}{n(n+2)}$

2. 点估计

- 令总体 $X \sim U(0,\theta)$，比较 $\hat{\theta}_A = 2\overline{X}$和 $\hat{\theta}_C = \max\{X_1, X_2, \dots, X_n\} \cdot \frac{n+1}{n}$的均方误差
- $\mathrm{MSE}\left(\hat{\theta}_A\right) = \frac{\theta^2}{3n}$
- $\mathrm{MSE}\left(\hat{\theta}_C\right) = \frac{\theta^2}{n(n+2)}$

- $\hat{\theta}_A$和 $\hat{\theta}_C$是否为一致估计量?

PEKING UNIVERSITY

2. 点估计

- k阶矩：$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
- k阶中心矩：$B_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$
- A_k是否为 $E\left(X^k\right)$的无偏估计?
  - $E(A_k) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^k\right) = E\left(X^k\right)$
- B_2是否为 $\mathrm{Var}(X)$的无偏估计?
  - $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}\left(X_i^2 + \overline{X}^2 - 2X_i\overline{X}\right) = \sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2$
  - $B_2 = A_2 - \overline{X}^2 \Rightarrow E(B_2) = E(X^2) - E\left(\overline{X}^2\right)$
  - $E\left(\overline{X}^2\right) = \left(E(\overline{X})\right)^2 + \mathrm{Var}(\overline{X}) = \left(E(X)\right)^2 + \frac{\mathrm{Var}(X)}{n}$
  - $E(B_2) = \frac{n-1}{n} \cdot \mathrm{Var}(X)$

2. 点估计

- $E(B_2) = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = \frac{n-1}{n} \cdot \mathrm{Var}(X)$
- 样本方差$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- $E(S^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right) = \mathrm{Var}(X)$
- 样本方差S^2是$\mathrm{Var}(X)$的无偏估计量
- 二阶中心矩B_2是$\mathrm{Var}(X)$的渐进无偏估计量

PEKING UNIVERSITY 北京大學

2. 点估计

- 总体 $X \sim N(\mu, \sigma^2)$。判断$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$和$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$的独立性。
- 考虑正交矩阵 U，第一行每个元素均为$\frac{1}{\sqrt{n}}$，其余行任取
- 令随机向量$\boldsymbol{X} = (X_1, X_2, \dots, X_n)$，令$\boldsymbol{Y} = U\boldsymbol{X}$，注意$\boldsymbol{Y}$服从$n$维高斯分布，且有
  - $E(\boldsymbol{Y}) = (\sqrt{n}\mu, 0, 0, \dots, 0)$
  - $\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2 \cdot I$
  - $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$
- $\overline{X} = Y_1/\sqrt{n}$
- $(n-1)S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \left(\sum_{i=1}^{n} X_i^2\right) - n\overline{X}^2 = \left(\sum_{i=1}^{n} Y_i^2\right) - Y_1^2 = \sum_{i=2}^{n} Y_i^2$
- $\mathrm{Cov}(\boldsymbol{Y}) = \sigma^2 \cdot I \Rightarrow Y_i$相互独立，因此$\overline{X}$与$(n-1)S^2$相互独立
- $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$。$\frac{(n-1)S^2}{\sigma^2}$服从何种分布?

2. 点估计

- $(n-1)S^2 = \sum_{i=2}^{n} Y_i^2 \Rightarrow \frac{(n-1)S^2}{\sigma^2} = \sum_{i=2}^{n} \left(\frac{Y_i}{\sigma}\right)^2$
- Y_i相互独立且对于$i \geq 2$，$Y_i \sim N(0, \sigma^2)$

- 若$Z_1, Z_2, \dots, Z_n$独立同分布且$Z_i \sim N(0,1)$，则$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$
  - 由作业三，$Z_i^2 \sim \chi^2(1)$，也即$Z_i^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$
  - 因此$\sum_{i=1}^{n} Z_i^2 \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$

- 因此有$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

2. 点估计

- **点估计**：估计分布中所含有的未知参数θ
  - 已知总体$X \sim U(0,\theta)$，估计θ
  - 已知总体$X \sim N(\mu,\sigma^2)$，估计μ和σ^2
- **估计量**：用来估计未知参数θ的统计量$\hat{\theta}(X_1,X_2,\ldots,X_n)$

- 设计估计量的常用方法
  - 矩法
  - 最大似然估计

PEKING UNIVERSITY 北京大学

2. 点估计

- 矩法：用样本矩去替换总体矩

- 1. 将未知参数$\theta_1, \theta_2, \dots, \theta_k$表示为总体前$k$阶矩的函数
  - $\theta_i = f_i(\mu_1, \mu_2, \dots, \mu_k)$
  - μ_i是总体X的i阶原点矩$E(X^i)$或i阶中心矩$E\left(\left(X - E(X)\right)^i\right)$

- 2. 用样本的i阶矩或i阶中心矩替换μ_i
  - $\widehat{\theta_i} = f_i(A_1, A_2, \dots, A_k)$或$\widehat{\theta_i} = f_i(B_1, B_2, \dots, B_k)$
  - k阶矩：$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
  - k阶中心矩：$B_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$
- 也可以用样本方差S^2替换$\mathrm{Var}(X)$

2. 点估计

- 矩法：用样本矩去替换总体矩

- 例1：已知总体 $X \sim U(0,\theta)$。用矩法设计θ的估计量
  - $\theta = 2E(X) \Rightarrow \hat{\theta}(X_1, X_2, \dots, X_n) = 2\overline{X}$

- 例2：已知总体 $X \sim U(a,b)$。用矩法设计a和b的估计量
  - $a + b = 2E(X)$
  - $\mathrm{Var}(X) = \frac{(b-a)^2}{12} \Rightarrow b - a = \sqrt{12 \cdot \mathrm{Var}(X)}$
  - $a = E(X) - \sqrt{3 \cdot \mathrm{Var}(X)},\ b = E(X) + \sqrt{3 \cdot \mathrm{Var}(X)}$
  - $\hat{a} = \overline{X} - \sqrt{3 \cdot S^2}, \hat{b} = \overline{X} + \sqrt{3 \cdot S^2}$

2. 点估计

- 例3：已知总体 $X \sim \mathrm{Exp}(\lambda)$。用矩法设计$\lambda$的估计量
  - $E(X) = \frac{1}{\lambda} \Rightarrow \hat{\lambda} = \frac{1}{\overline{X}}$
  - $\mathrm{Var}(X) = \frac{1}{\lambda^2} \Rightarrow \hat{\lambda} = \frac{1}{\sqrt{S^2}}$

- 例4：已知总体 $X \sim N(\mu, \sigma^2)$。用矩法设计μ和σ^2的估计量
  - $\mu = E(X), \sigma^2 = \mathrm{Var}(X) \Rightarrow \hat{\mu} = \overline{X}, \widehat{\sigma^2} = S^2$或$\widehat{\sigma^2} = B_2$
  - $\mu = E(X), E(X^2) = \mu^2 + \sigma^2 \Rightarrow \hat{\mu} = \overline{X}, \widehat{\sigma^2} = A_2 - \overline{X}^2$

2. 点估计

- 例4：已知总体 $X \sim N(\mu, \sigma^2)$。用矩法设计μ和σ^2的估计量
- σ^2的估计量： S^2， B_2 ， $A_2 - \overline{X}^2$
- 比较三者的均方误差?
- $B_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2 = A_2 - \overline{X}^2$
- $\mathrm{MSE}(S^2) = \mathrm{Var}(S^2)$，而$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$，
- $\mathrm{MSE}(S^2) = \mathrm{Var}(S^2) = \frac{2(n-1)}{(n-1)^2} \cdot \sigma^4 = \frac{2}{n-1} \cdot \sigma^4$
- $\mathrm{MSE}(B_2) = E\left((B_2 - \sigma^2)^2\right) = E(B_2^2) + \sigma^4 - 2\sigma^2 E(B_2)$
- $E(B_2^2) = \left(\frac{n-1}{n}\right)^2 \cdot E\left((S^2)^2\right) = \left(\frac{n-1}{n}\right)^2 \cdot \frac{\sigma^4}{(n-1)^2} \cdot \left((n-1)^2 + 2(n-1)\right) = \frac{n^2-1}{n^2} \cdot \sigma^4$
- $E(B_2) = \frac{\sigma^2(n-1)}{n} \Rightarrow \mathrm{MSE}(B_2) = \frac{2n-1}{n^2}\sigma^4$

2. 点估计

- 例5：从$\{1,2,\dots,N\}$中等概率且不放回地抽取n个样本。设计N的估计量
- $X_1,X_2,\dots,X_n$不独立，不符合简单随机样本的情况
- $E(X)=\frac{N+1}{2} \Rightarrow \hat{N}=2\overline{X}-1$
- 判断$\hat{N}=2\overline{X}-1$的无偏性，并计算$\hat{N}$的均方误差

2. 点估计

- $\hat{N} = \frac{2(X_1+X_2+\cdots+X_n)}{n} - 1$
- 如何计算$E(X_i)$?
  - $P(X_i = k) = \frac{1}{N} \Rightarrow E(X_i) = \frac{N+1}{2}$
- $E(\hat{N}) = N$，$\hat{N}$是无偏估计量

- $\mathrm{MSE}(\hat{N}) = \mathrm{Var}(\hat{N}) = 4\mathrm{Var}(\bar{X}) = \frac{4}{n^2} \cdot \mathrm{Var}(X_1 + X_2 + \cdots X_n)$
- $\mathrm{Var}(X_i) = \frac{N^2-1}{12}, \mathrm{Cov}(X_i, X_j) = -\frac{N+1}{12}$
- $\mathrm{MSE}(\hat{N}) = \frac{1}{3n}(N+1)(N-n)$

2. 点估计

- **点估计**：给定样本$x_1, x_2, \dots, x_n$，估计分布中所含有的未知参数θ
- **最大似然估计**：选择参数θ，使得**似然函数** $L(\theta) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$最大
- **最大似然估计量**：满足$L\left(\hat{\theta}\right) = \max_{\theta} L(\theta)$的统计量$\hat{\theta}$

- 例1：已知某网站不同用户点击广告的情况服从参数为p的伯努利分布。给定简单随机样本$x_1, x_2, \dots, x_n \in \{0,1\}$，求未知参数$p$的最大似然估计。

PEKING UNIVERSITY

2. 点估计

- **最大似然估计**：选择参数θ，使得**似然函数** $L(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$最大

- 例1：已知某网站不同用户点击广告的情况服从参数为p的伯努利分布。给定简单随机样本$x_1, x_2, \ldots, x_n \in \{0,1\}$，求未知参数$p$的最大似然估计。

- $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$
- 选择 p最大化 $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; p)$
- 对 p求导，$\sum x_i \cdot p^{\sum x_i - 1}(1-p)^{n-\sum x_i} - (n - \sum x_i) \cdot p^{\sum x_i}(1-p)^{n-\sum x_i - 1} = 0$
- $\sum x_i(1-p) - (n - \sum x_i) \cdot p = 0 \Rightarrow p = \sum x_i / n$
- $\hat{p}_{\mathrm{MLE}} = \overline{X}$

PEKING UNIVERSITY 北京大学

2. 点估计

- **最大似然估计（MLE）**：选择参数θ，使得**似然函数** $L(\theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$最大

- 例2：某超算中心每日收到的任务数量服从$\pi(\lambda)$。给定简单随机样本$x_1, x_2, \ldots, x_n$，求未知参数λ的最大似然估计。

- $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}$
- 选择 p最大化 $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \lambda)$
- 求导， $e^{-n\lambda}\lambda^{\sum x_i - 1} \cdot \sum x_i - n \cdot e^{-n\lambda}\lambda^{\sum x_i} = 0 \Rightarrow \lambda = \sum x_i / n$
- $\hat{\lambda}_{\mathrm{MLE}} = \overline{X}$

2. 点估计

- **最大似然估计（MLE）**：选择参数θ，使得**似然函数** $L(\theta) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$最大
- 对于简单随机样本， $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta) = \prod_{i=1}^{n} P(X_i = x_i; \theta)$
- 最大化$L(\theta)$等价于最大化**对数似然函数** $\ln L(\theta) = \sum_{i=1}^{n} \ln P(X_i = x_i; \theta)$

- 例2：某超算中心每日收到的任务数量服从$\pi(\lambda)$。给定简单随机样本$x_1, x_2, \dots, x_n$，求未知参数λ的最大似然估计。

- $\ln L(\lambda) = \sum_{i=1}^{n} \ln P(X_i = x_i; \lambda) = \sum_{i=1}^{n}(-\lambda + x_i \ln \lambda - \ln(x_i!))$
- 对 λ求导， $-n + \frac{\sum x_i}{\lambda} = 0 \Rightarrow \lambda = \sum x_i / n$

PEKING UNIVERSITY 北京大学

2. 点估计

- 当总体X为连续型，定义**似然函数**$L(\theta)=f(X_1=x_1,X_2=x_2,\dots,X_n=x_n;\theta)$
- **最大似然估计（MLE）**：选择参数θ，使得**似然函数**$L(\theta)=f(X_1=x_1,X_2=x_2,\dots,X_n=x_n;\theta)$最大

- 例3：令总体X服从$U(0,\theta)$。给定简单随机样本$x_1,x_2,\dots,x_n$，求未知参数θ的最大似然估计。
- $L(\theta)=\prod_{i=1}^{n}\frac{1_{x_i\le\theta}}{\theta}=\frac{1_{x_1\le\theta,x_2\le\theta,\dots x_n\le\theta}}{\theta^n}$
- 最大化 $L(\theta)\Rightarrow\theta=\max\{x_1,x_2,\dots,x_n\}$

2. 点估计

- 当总体X为连续型，定义**似然函数**$L(\theta) = f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$
- **最大似然估计（MLE）**：选择参数θ，使得**似然函数**$L(\theta) = f(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$最大

- 例4：令总体X服从$N(\mu, \sigma^2)$。给定简单随机样本$x_1, x_2, \dots, x_n$，求未知参数μ的和σ^2最大似然估计。
- $\ln L(\mu, \sigma^2) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) = \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2} - \frac{\ln\sigma^2}{2} - \frac{\ln 2\pi}{2}\right)$
- $\frac{\partial \ln L(\mu,\sigma^2)}{\partial \mu} = 0 \Rightarrow \mu = \frac{\sum x_i}{n} = \overline{X}$
- $\frac{\partial \ln L(\mu,\sigma^2)}{\partial \sigma^2} = 0 \Rightarrow \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} = 0 \Rightarrow \sigma^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\mu)^2 = B_2$

北京大学 PEKING UNIVERSITY

2. 点估计

- 例4：令总体X服从$N(\mu,\sigma^2)$。给定简单随机样本$x_1,x_2,\ldots,x_n$，求未知参数μ的和σ^2最大似然估计。
- $\hat{\mu}_{\mathrm{MLE}}=\frac{\sum x_i}{n}=\overline{X}$
- $\widehat{\sigma^2}_{\mathrm{MLE}}=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\frac{\sum x_i}{n}\right)^2=B_2$
- 如何求σ的最大似然估计?
  - $\hat{\sigma}_{\mathrm{MLE}}=\sqrt{B_2}$
- 最大似然估计的不变性：给定θ的最大似然估计$\hat{\theta}$。若$g(\theta)$有单值反函数，则$g(\theta)$的最大似然估计为$g(\hat{\theta})$。

PEKING UNIVERSITY

2. 点估计

- 当总体X为连续型，定义**似然函数**$L(\theta)=f(X_1=x_1,X_2=x_2,\ldots,X_n=x_n;\theta)$
- **最大似然估计（MLE）**：选择参数θ，使得**似然函数**$L(\theta)=f(X_1=x_1,X_2=x_2,\ldots,X_n=x_n;\theta)$最大

- 例5：总体X的概率密度函数为
  - $f(x)=\frac{1}{\theta}\cdot e^{-(x-\mu)/\theta}$若 $x\geq\mu$
  - 否则 $f(x)=0$
- 给定简单随机样本$x_1,x_2,\ldots,x_n$，求未知参数μ和θ的最大似然估计和矩法估计。
- 矩法：$X-\mu\sim \mathrm{Exp}\left(\frac{1}{\theta}\right)\Rightarrow E(X)=\mu+\theta,\mathrm{Var}(X)=\theta^2$
- $\hat{\theta}=\sqrt{B_2},\mu=\overline{X}-\sqrt{B_2}$

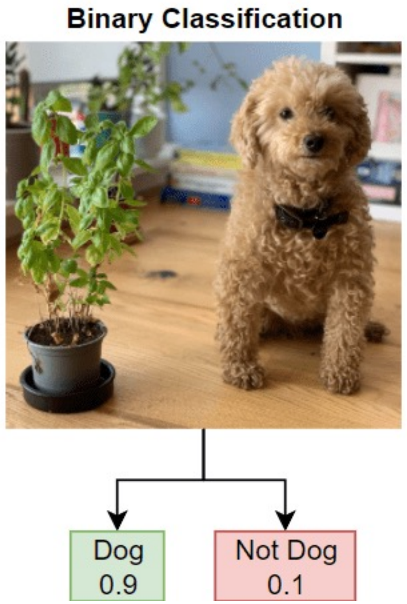
2. 点估计

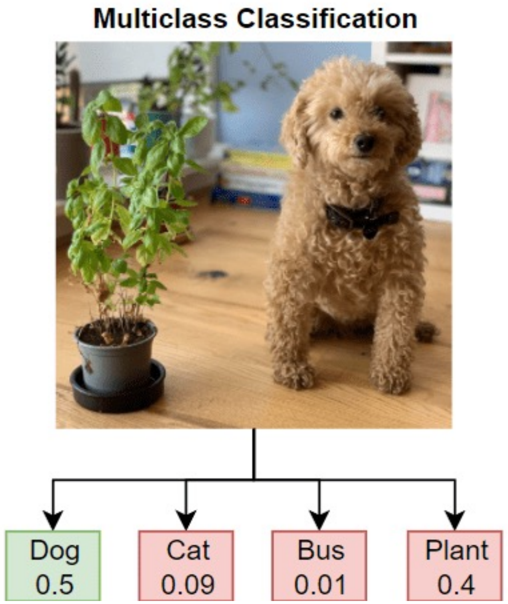
- 例5：总体X的概率密度函数为
  - $f(x) = \frac{1}{\theta} \cdot e^{-(x-\mu)/\theta}$若 $x \geq \mu$
  - 否则 $f(x) = 0$
- 给定简单随机样本$x_1, x_2, \dots, x_n$，求未知参数μ和θ的最大似然估计和矩法估计。
- $L(\mu, \theta) = \frac{1}{\theta^n} \cdot e^{-\frac{1}{\theta} \cdot \sum(x_i-\mu)} \cdot 1_{x_1 \geq \mu, x_2 \geq \mu, \dots, x_n \geq \mu}$
- 当 $x_1 \geq \mu, x_2 \geq \mu, \dots, x_n \geq \mu$， $L(\mu, \theta)$关于 μ为增函数
- $\hat{\mu}_{\mathrm{MLE}} = \min\{x_1, x_2, \dots, x_n\}$
- $\ln L(\mu, \theta) = -n \ln \theta - \frac{1}{\theta} \sum(x_i - \mu)$
- 求导， $-\frac{n}{\theta} + \frac{1}{\theta^2} \sum(x_i - \mu) = 0 \Rightarrow \hat{\theta}_{\mathrm{MLE}} = \overline{X} - \mu = \overline{X} - \min\{x_1, x_2, \dots, x_n\}$

2. 点估计

- **最大似然估计**：选择参数θ，使得**似然函数** $L(\theta) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$最大

- **分类问题**：给定数据$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$。其中$x_i \in \mathcal{X}$为数据，$y_i \in \{1, 2, \dots, C\}$为分类标签。利用数据学习分类函数$f: \mathcal{X} \to \{1, 2, \dots, C\}$。





2. 点估计

- **分类问题**：给定数据$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$。利用数据学习分类函数$f: \mathcal{X} \to \{1, 2, \dots, C\}$。

- **机器学习模型**：给定参数θ，对于给定的数据$x \in \mathcal{X}$，模型f_θ输出$\{1, 2, \dots, C\}$上的一个分布
- $f_\theta(y|x)$表示给定输入数据$x \in \mathcal{X}$，分类为y的概率

- 分类函数：$\operatorname{argmax}_y f_\theta(y|x)$

- 给定数据$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$，如何选取参数$\theta$?

PEKING UNIVERSITY 北京大學

2. 点估计

- **最大似然估计**：选择参数θ，使得**似然函数** $L(\theta) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$最大

- **分类问题**：给定数据$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$。利用数据学习分类函数$f: \mathcal{X} \to \{1, 2, \dots, C\}$。

- **机器学习模型**： 给定参数θ，对于给定的数据$x \in \mathcal{X}$，模型f_θ输出$\{1, 2, \dots, C\}$上的一个分布

- 对于分类问题： 最大化似然函数$L(\theta) = \prod_{i=1}^{n} f_\theta(y_i|x_i)$
- 实践中，通常改为最小化$-\ln L(\theta) = \sum_{i=1}^{n} - \ln f_\theta(y_i|x_i)$

2. 点估计

- 对于两个分布 p和q，定义**交叉熵**$H(p,q)=-E_p(\ln q)$
  - 对于离散分布，$H(p,q)=-\sum_x p(x)\ln q(x)$
  - 对于连续分布，$H(p,q)=-\int P(x)\ln Q(x)\,dx$，$P(x)$和 $Q(x)$为概率密度函数

- p为数据真实标签对应的分布，也即$p(y|x_i)=1_{y=y_i}$

- $-\ln L(\theta)=\sum_{i=1}^{n}-\ln f_\theta(y_i|x_i)=\sum_{i=1}^{n}H_i(p,f_\theta)$
  - H_i为数据x_i真实标签分布和模型输出分布的交叉熵
  - $H_i(p,f_\theta)=-\sum_y p(y|x_i)\ln f_\theta(y|x_i)=-\ln f_\theta(y_i|x_i)$

- 结论：在分类问题中，最大似然估计等价于最小化交叉熵损失函数

2. 点估计

- **机器学习模型**：给定参数θ，对于给定的数据$x \in \mathcal{X}$，模型f_θ输出$\{1, 2, \dots, C\}$上的一个分布
  - 满足$f_\theta(y|x) \geq 0$和$\sum_y f_\theta(y|x) = 1$

- 如何放松上述要求?
- $\mathrm{softmax}_i(a_1, a_2, \dots, a_C) = \frac{e^{a_i}}{\sum_j e^{a_j}}$
- 给定$f_\theta: \mathcal{X} \to \mathbb{R}^C$，将softmax应用于 f_θ的输出即可得到 $\{1, 2, \dots, C\}$上的分布

- 实践中，交叉熵损失函数往往与softmax配合使用

3. 区间估计

- **点估计**：给定样本$x_1, x_2, \ldots, x_n$，估计分布中所含有的未知参数θ

- **区间估计**：设计统计量$\hat{\theta}_L(X_1, X_2, \ldots, X_n)$和$\hat{\theta}_U(X_1, X_2, \ldots, X_n)$，使得$\theta \in [\hat{\theta}_L, \hat{\theta}_U]$的概率尽量大
- $P\left(\theta \in \left[\hat{\theta}_L, \hat{\theta}_U\right]\right)$越大，则导致区间长度$\hat{\theta}_U - \hat{\theta}_L$更长

- **置信区间**：若统计量$\hat{\theta}_L(X_1, X_2, \ldots, X_n)$和$\hat{\theta}_U(X_1, X_2, \ldots, X_n)$，对于任意$\theta$，满足$P\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right) \ge 1 - \alpha$，则$[\hat{\theta}_L, \hat{\theta}_U]$为$\theta$的**置信水平**为$1 - \alpha$的**置信区间。**
  - $\hat{\theta}_L$：置信下限
  - $\hat{\theta}_U$：置信上限

PEKING UNIVERSITY

3. 区间估计

- **置信区间**：若统计量$\hat{\theta}_L(X_1, X_2, \dots, X_n)$和$\hat{\theta}_U(X_1, X_2, \dots, X_n)$，对于任意$\theta$，满足$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\right) \geq 1-\alpha$，则$[\hat{\theta}_L, \hat{\theta}_U]$为$\theta$的**置信水平**为$1-\alpha$的**置信区间。**

- **单侧置信下限**：若统计量$\hat{\theta}_L(X_1, X_2, \dots, X_n)$，对于任意$\theta$，满足$P\left(\hat{\theta}_L \leq \theta\right) \geq 1-\alpha$，则$\hat{\theta}_L$为$\theta$的**置信水平**为$1-\alpha$的**单侧置信下限。**
- **单侧置信上限**：若统计量$\hat{\theta}_U(X_1, X_2, \dots, X_n)$，对于任意$\theta$，满足$P\left(\theta \leq \hat{\theta}_U\right) \geq 1-\alpha$，则$\hat{\theta}_U$为$\theta$的**置信水平**为$1-\alpha$的**单侧置信上限。**

- 若$\hat{\theta}_L$为θ的置信水平为$1-\alpha_1$的单侧置信下限，$\hat{\theta}_U$为θ的置信水平为$1-\alpha_2$的单侧置信上限，则$[\hat{\theta}_L, \hat{\theta}_U]$为$\theta$的置信水平为$1-\alpha_1-\alpha_2$的置信区间。

3. 区间估计

- 例1：总体$X \sim N(\mu,\sigma^2)$，σ^2已知。设计μ的置信水平为$1-\alpha$的置信区间
- 利用$\overline{X}$构造置信区间?
- 注意到$\overline{X} \sim N(\mu,\sigma^2/n)$，也即$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- $P\left(c \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le d\right) = 1-\alpha \Rightarrow P\left(\overline{X}-\frac{d\sigma}{\sqrt{n}} \le \mu \le \overline{X}-\frac{c\sigma}{\sqrt{n}}\right) = 1-\alpha$
- 如何选择c与d使得区间最短?
  - 令 $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$为标准正态分布的分布函数
  - 取$c = \Phi^{-1}(\alpha/2)$，$d = \Phi^{-1}(1-\alpha/2)$，注意到$c+d=0$
- 结论：$P\left(\overline{X}-\frac{\Phi^{-1}(1-\alpha/2)\cdot\sigma}{\sqrt{n}} \le \mu \le \overline{X}+\frac{\Phi^{-1}(1-\alpha/2)\cdot\sigma}{\sqrt{n}}\right) = 1-\alpha$

3. 区间估计

- 例1：总体$X \sim N(\mu,\sigma^2)$，σ^2已知。设计μ的置信水平为$1-\alpha$的置信区间
- $\overline{X}$为无偏估计量。基于集中不等式构造置信区间?
- 回顾：若$X \sim N(\mu,\sigma^2)$
  - $P(X-E(X) \geq k\sigma) \leq e^{-\frac{k^2}{2}}$
  - $P(X-E(X) \leq -k\sigma) \leq e^{-\frac{k^2}{2}}$
- $\overline{X}-\mu \sim N(0,\sigma^2/n) \Rightarrow P\left(\left|\overline{X}-\mu\right| \geq \frac{k\sigma}{\sqrt{n}}\right) \leq 2e^{-\frac{k^2}{2}}$
- $2e^{-\frac{k^2}{2}} = \alpha \Rightarrow k = \sqrt{2\ln(2/\alpha)}$
- 结论：$P\left(\overline{X}-\frac{\sigma\cdot\sqrt{2\ln(2/\alpha)}}{\sqrt{n}} \leq \mu \leq \overline{X}+\frac{\sigma\cdot\sqrt{2\ln(2/\alpha)}}{\sqrt{n}}\right) \leq 1-\alpha$
- 与$P\left(\overline{X}-\frac{\Phi^{-1}(1-\alpha/2)\cdot\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}+\frac{\Phi^{-1}(1-\alpha/2)\cdot\sigma}{\sqrt{n}}\right) = 1-\alpha$的联系?

PEKING UNIVERSITY

3. 区间估计

- **置信区间**：若统计量$\hat{\theta}_L(X_1, X_2, \dots, X_n)$和$\hat{\theta}_U(X_1, X_2, \dots, X_n)$，对于任意$\theta$满足$P\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right) \ge 1-\alpha$，则$[\hat{\theta}_L, \hat{\theta}_U]$为$\theta$的**置信水平**为$1-\alpha$的**置信区间。**

- **枢轴量法**
- 1. 设计枢轴量$G = G(x_1, x_2, \dots, x_n)$，使得$G$的分布与未知参数$\theta$无关。
- 2. 选择c和d，$P(c \le G \le d) = 1-\alpha$，$c \le G \le d$可变形为 $\hat{\theta}_L \le \theta \le \hat{\theta}_U$
- 3. 得到结论$P\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right) = 1-\alpha$

- 设计枢轴量通常从θ的点估计出发
- 通常选择c和d使得 $P(G < c) = P(G > d) = \alpha/2$

3. 区间估计

- 例2： 总体 $X \sim U(0,\theta)$。设计θ的置信水平为$1-\alpha$的置信区间。

- 令 $\hat{\theta} = \max\{X_1, X_2, \dots, X_n\}$为 θ的最大似然估计
- $F_{\hat{\theta}}(x) = \left(F_X(x)\right)^n = \left(\frac{x}{\theta}\right)^n$ 当 $0 \le x \le \theta$

- 枢轴量$G = \hat{\theta}/\theta$
- $F_G(x) = x^n$当 $0 \le x \le 1$
- $P(c \le G \le d) = 1-\alpha \Rightarrow P\left(\frac{\hat{\theta}}{d} \le \theta \le \frac{\hat{\theta}}{c}\right) = 1-\alpha$
- $d^n - c^n = 1-\alpha \Rightarrow d = 1, c = \alpha^{1/n}$

- 结论：$P\left(\hat{\theta} \le \theta \le \hat{\theta} \cdot \alpha^{-1/n}\right) = 1-\alpha$

3. 区间估计

- 例3：总体$X \sim N(\mu,\sigma^2)$。设计σ^2的置信水平为$1-\alpha$的置信区间。

- 回顾：S^2是σ^2的无偏估计，$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

- 令$F(x)$为$\chi^2(n-1)$的分布函数，取$c = F^{-1}(\alpha/2), d = F^{-1}(1-\alpha/2)$

- 则有$P\left(c \le \frac{(n-1)S^2}{\sigma^2} \le d\right) = 1-\alpha \Rightarrow P\left(\frac{(n-1)S^2}{d} \le \sigma^2 \le \frac{(n-1)S^2}{c}\right) = 1-\alpha$

3. 区间估计

- 例4：总体$X \sim N(\mu, \sigma^2)$，μ, σ^2均为未知参数。设计μ的置信水平为$1-\alpha$的置信区间。
- $\frac{\overline{X}-\mu}{\sqrt{S^2}}$服从何种分布?
- 回顾：$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$，且$\overline{X}$和$S^2$相互独立

- ***t*分布**：令$X_1 \sim N(0,1)$，$X_2 \sim \chi^2(n)$，X_1和X_2相互独立。称$T=\frac{X_1}{\sqrt{X_2/n}}$的分布为自由度为$n$的***t*分布**，记为$T \sim t(n)$

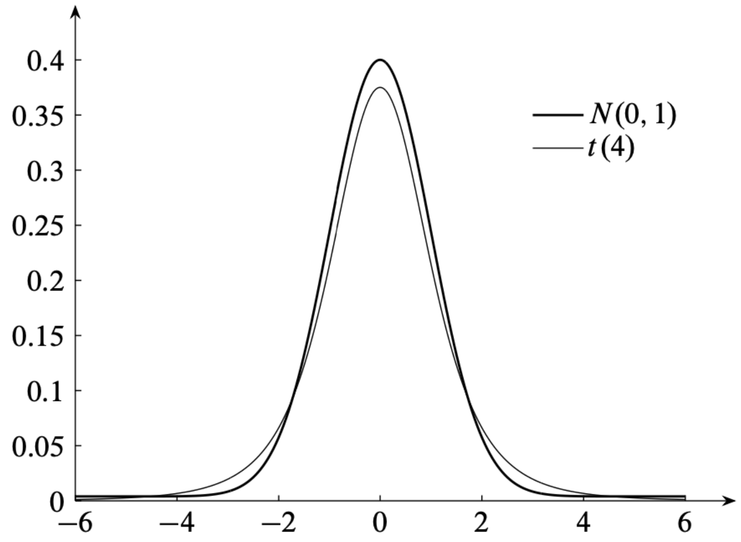
3. 区间估计

- ***t*分布**：令 $X_1 \sim N(0,1)$，$X_2 \sim \chi^2(n)$，X_1和X_2相互独立。称$T=\frac{X_1}{\sqrt{X_2/n}}$的分布为自由度为$n$的***t*分布**，记为$T \sim t(n)$

- 自由度为n的t分布概率密度函数：$f(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$

- 当 $n=1$，$f(t)=\frac{1}{1+t^2}\cdot\frac{1}{\pi}$

- 当自由度n较大，近似为标准正态分布



3. 区间估计

- 例4：总体$X \sim N(\mu,\sigma^2)$且μ,σ^2均为未知参数。设计μ的置信水平为$1-\alpha$的置信区间。
- ***t*分布**：令$X_1 \sim N(0,1)$，$X_2 \sim \chi^2(n)$，X_1和X_2相互独立。称$T=\frac{X_1}{\sqrt{X_2/n}}$的分布为自由度为$n$的***t*分布**，记为$T \sim t(n)$
- 回顾：$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$，且$\overline{X}$和$S^2$相互独立

- $\frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{S^2}} \sim t(n-1)$
- $P\left(c \le \frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{S^2}} \le d\right)=1-\alpha \Rightarrow P\left(\overline{X}-d\cdot\frac{\sqrt{S^2}}{\sqrt{n}} \le \mu \le \overline{X}-c\cdot\frac{\sqrt{S^2}}{\sqrt{n}}\right)=1-\alpha$
- 令$F(x)$为$t(n-1)$的分布函数，取$c=F^{-1}(\alpha/2), d=F^{-1}(1-\alpha/2)$
- 注意到$c+d=0$

3. 区间估计

- 例5：总体$X \sim B(1,p)$。设计p的置信水平为$1-\alpha$的置信区间。

- 由中心极限定理，$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$近似服从 $N(0,1)$

- $P\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \le \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \le \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) \approx 1-\alpha$
  - $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$为标准正态分布的分布函数

- 解方程可得到置信水平为$1-\alpha$的近似置信区间

3. 区间估计

- 例5：总体$X \sim B(1, p)$。设计p的置信水平为$1-\alpha$的置信区间。

- $\bar{X}$为无偏估计
- Chernoff bound：$P\left(\left|\bar{X}-p\right| > \epsilon\right) \le 2e^{-2n\epsilon^2}$
- $2 \cdot e^{-2n\epsilon^2} = \alpha \Rightarrow \epsilon = \sqrt{\ln(2/\alpha)/2n}$
- $P\left(\bar{X} - \sqrt{\ln(2/\alpha)/2n} \le p \le \bar{X} + \sqrt{\ln(2/\alpha)/2n}\right) \ge 1-\alpha$

3. 区间估计

- 有n台游戏机，第i台游戏机中奖概率为未知参数p_i。
- 每一轮可从n台游戏机选择一台并进行游戏
- 用最少的轮数，找到中奖概率最高的游戏机
- 应用：推荐系统，医疗，...

- 第t轮中：
  - 选择游戏机i
  - 观测到结果 $X_t \sim B(1, p_i)$
- 输出：中奖概率最高的游戏机$o \in \{1,2,\dots,n\}$

- 均匀采样：对每台游戏机采集相同数量的样本N，返回样本均值最大的游戏机
- 要求：返回游戏机$o \in \{1,2,\dots,n\}$，使得$P(p_o \geq \max_i p_i - \epsilon) \geq 2/3$。 N应当取多大?

3. 区间估计

- 若对第i台游戏机采集N个样本
- $P\left(\overline{X}_i - \sqrt{\ln(2/\alpha)\,/2N} \le p_i \le \overline{X}_i + \sqrt{\ln(2/\alpha)\,/2N}\right) \ge 1-\alpha$
- 若取 $N = O(\ln(1/\alpha)/\epsilon^2)$，则有$P\left(\left|p_i - \overline{X}_i\right| \le \epsilon\right) \ge 1-\alpha$

- α应当如何选择?
- 取$\alpha = \frac{1}{3n}$， $N = O(\ln n\,/\epsilon^2)$，则$P\left(\left|p_i - \overline{X}_i\right| \le \epsilon\right) \ge 1-\frac{1}{3n}$
- Union bound： $P\left(\forall i,\ \left|p_i - \overline{X}_i\right| \le \epsilon\right) \ge 2/3$

- 此时，样本均值最大的游戏机o满足$p_o \ge \max_i p_i - 2\epsilon$
- $p_o \ge \overline{X}_o - \epsilon \ge \overline{X}_i - \epsilon \ge p_i - 2\epsilon$

3. 区间估计

- 均匀采样：对每台游戏机采集相同数量的样本N，返回样本均值最大的游戏机
- 如何使$P(p_o \geq \max_i p_i - \epsilon) \geq 1 - \delta$?

- $P\left(\overline{X}_i - \sqrt{\ln(2/\alpha)/2N} \leq p_i \leq \overline{X}_i + \sqrt{\ln(2/\alpha)/2N}\right) \geq 1 - \alpha$
- 取$\alpha = \delta/n$，$N = O(\ln(n/\delta)/\epsilon^2)$，则$P\left(\left|p_i - \overline{X}_i\right| \leq \epsilon\right) \geq 1 - \delta/n$
- Union bound：$P\left(\forall i,\ \left|p_i - \overline{X}_i\right| \leq \epsilon\right) \geq 1 - \delta$

- 对每台游戏机采集样本数量：$O(\ln(n/\delta)/\epsilon^2)$