北京大学
PEKING UNIVERSITY

信息学中的概率统计

王若松

前沿计算研究中心
北京大学

北京大学 PEKING UNIVERSITY

1. 假设检验问题

- **参数估计**：估计分布中所含有的未知参数
- 例：网络延迟服从分布$N(\theta, 16)$。网络设备工作正常的情况下，有$\theta \leq 110$。给定某天的网络延迟测试数据$x_1, x_2, \ldots, x_n$，判断网络设备是否工作正常。

- $\Theta_0 = \{\theta | \theta \leq 110\}, \Theta_1 = \{\theta | \theta > 110\}$
- 给定总体和样本，判断$\theta \in \Theta_0$或$\theta \in \Theta_1$
- 上述问题称为**假设检验**问题
- $\theta \in \Theta_0$和$\theta \in \Theta_1$被称为**假设**

- 利用总体$N(\theta, 16)$和样本$x_1, x_2, \ldots, x_n$，判断假设$\theta \in \Theta_0$是否成立
  - 结果1：假设不成立，称为**拒绝该假设**
  - 结果2：假设成立，称为**接受该假设**

1. 假设检验问题

- 假设检验的一般步骤

- 第一步：建立假设
  - **原假设** H_0：不应轻易拒绝的假设
    - 例：新的策略没有效果；网络设备没有异常
  - **备择假设**H_1：与原假设 H_0对立的假设
  - $H_0: \theta \leq 110, H_1: \theta > 110$

- 常见的假设形式
  - $H_0: \theta \leq \theta_0, H_1: \theta > \theta_0$
  - $H_0: \theta \geq \theta_0, H_1: \theta < \theta_0$
  - $H_0: \theta = \theta_0, H_1: \theta \neq \theta_0$

1. 假设检验问题

- 假设检验的一般步骤

- 第二步：选择统计量，给出拒绝域的形式
- 将样本取值$x_1, x_2, \dots, x_n$分为两个区域W和$\overline{W}$
  - 若$(x_1, x_2, \dots, x_n) \in W$，拒绝原假设 H_0
  - 若$(x_1, x_2, \dots, x_n) \in \overline{W}$，接受原假设，也即拒绝备择假设$H_1$
- W：**拒绝域**，$\overline{W}$：**接受域**
- 通常根据统计量的取值的设计拒绝域

- 刚才的例子中，可取拒绝域为$W = \{(x_1, x_2, \dots, x_n) \mid \bar{x} \geq c\}$，$c$为待定值

1. 假设检验问题

- 若拒绝域为$W = \{(x_1, x_2, \dots, x_n) \mid \overline{x} \geq c\}$
- 两种发生错误判断的情况
  - $\overline{x} \geq c$，但$\theta \leq 110$
  - $\overline{x} < c$，但$\theta > 110$

- **第一类错误**：$(x_1, x_2, \dots, x_n) \in W$，拒绝原假设$H_0$，但原假设$H_0$为真
- **第二类错误**：$(x_1, x_2, \dots, x_n) \in \overline{W}$，接受原假设$H_0$，但原假设$H_0$不为真

1. 假设检验问题

- **第一类错误**：$(x_1, x_2, \dots, x_n) \in W$，拒绝原假设 H_0，但原假设 H_0为真
  - 令α为拒绝真实的原假设的概率，也即犯第一类错误的概率
- **第二类错误**：$(x_1, x_2, \dots, x_n) \in \overline{W}$，接受原假设 H_0，但原假设 H_0不为真
  - 令β为接受错误的原假设的概率，也即犯第二类错误的概率

	H_0为真	H_0不为真
$(x_1, x_2, \dots, x_n) \in W$ 拒绝 H_0	第一类错误	决策正确
$(x_1, x_2, \dots, x_n) \in \overline{W}$ 接受 H_0	决策正确	第二类错误

- 增加样本大小可同时减少第一类错误和第二类错误
- 对于固定的样本，能否同时减少第一类错误和第二类错误?

1. 假设检验问题

- 总体服从$N(\theta, 16)$。
- $H_0: \theta \leq 110, H_1: \theta > 110$
- 拒绝域为$W = \{(x_1, x_2, \ldots, x_n) \mid \bar{x} \geq c\}$，$c$为待定值
- 回顾：$\frac{\bar{x}-\theta}{4/\sqrt{n}} \sim N(0,1)$
- 第一类错误：$\bar{x} \geq c$，但$\theta \leq 110$
  - $\alpha(\theta) = P(\bar{x} \geq c) = P\left(\frac{\bar{x}-\theta}{4/\sqrt{n}} \geq \frac{c-\theta}{4/\sqrt{n}}\right) = 1 - \Phi\left(\frac{c-\theta}{4/\sqrt{n}}\right)$
- 第二类错误：$\bar{x} < c$，但$\theta > 110$
  - $\beta(\theta) = P(\bar{x} < c) = P\left(\frac{\bar{x}-\theta}{4/\sqrt{n}} < \frac{c-\theta}{4/\sqrt{n}}\right) = \Phi\left(\frac{c-\theta}{4/\sqrt{n}}\right)$
- 随着c增大，α减小，β增大
- 随着c减小，α增大，β减小

1. 假设检验问题

- 假设检验的一般步骤

- 第三步：选择显著性水平α，给出拒绝域

- $H_0: \theta \in \Theta_0, H_1: \theta \in \Theta_1$
- 若检验对于任意$\theta \in \Theta_0$，都有$\alpha(\theta) \leq \alpha$，称该检验**显著性水平**为$\alpha$
  - 也即，控制第一类错误的概率不超过α，再尽量减少第二类错误的概率
  - **Neyman-Pearson原则**
  - 一般取 $\alpha = 0.05$

1. 假设检验问题

- 总体服从$N(\theta, 16)$。
- $H_0: \theta \leq 110, H_1: \theta > 110$
- 拒绝域为$W = \{(x_1, x_2, \dots, x_n) \mid \bar{x} \geq c\}$，$c$为待定值
- 回顾：$\frac{\bar{x}-\theta}{4/\sqrt{n}} \sim N(0,1)$
- 第一类错误：$\bar{x} \geq c$，但$\theta \leq 110$
  - $\alpha(\theta) = P(\bar{x} \geq c) = P\left(\frac{\bar{x}-\theta}{4/\sqrt{n}} \geq \frac{c-\theta}{4/\sqrt{n}}\right) = 1 - \Phi\left(\frac{c-\theta}{4/\sqrt{n}}\right)$
- Neyman-Pearson原则： 对于全部$\theta \leq 110$，$1 - \Phi\left(\frac{c-\theta}{4/\sqrt{n}}\right) \leq \alpha$
  - 只需要$1 - \Phi\left(\frac{c-110}{4/\sqrt{n}}\right) = \alpha \Rightarrow c = 110 + \Phi^{-1}(1-\alpha) \cdot 4/\sqrt{n}$
- 也即拒绝域为$W = \{(x_1, x_2, \dots, x_n) \mid \bar{x} \geq 110 + \Phi^{-1}(1-\alpha) \cdot 4/\sqrt{n}\}$

1. 假设检验问题

- 假设检验的一般步骤
- 第一步：建立假设$H_0: \theta \in \Theta_0, H_1: \theta \in \Theta_1$
- 第二步：选择统计量，给出拒绝域的形式
  - 计算H_0为真时统计量的分布
- 第三步：选择显著性水平α，给出拒绝域
  - 根据第二步中计算的分布，确定拒绝域的临界值
- 如何避免事先确定显著水平？

北京大学 PEKING UNIVERSITY

1. 假设检验问题

- 总体服从$N(\theta, 16)$。
- $H_0: \theta \leq 110, H_1: \theta > 110$
- 给定显著性水平α，则拒绝域为：
- $W = \{(x_1, x_2, \ldots, x_n) \mid \overline{x} \geq 110 + \Phi^{-1}(1-\alpha) \cdot 4/\sqrt{n}\}$

- 给定$\overline{x}$，称满足$\overline{x} \geq 110 + \Phi^{-1}(1-\alpha) \cdot 4/\sqrt{n}$最小的$\alpha$为$p$值
- 也即$p = 1 - \Phi\left(\frac{\overline{x}-110}{4/\sqrt{n}}\right)$
  - 若显著性水平$\alpha \geq p$，则拒绝H_0
  - 若显著性水平$\alpha < p$，则接受H_0

- 更一般的，给定样本取值，能够做出拒绝原假设的最小显著性水平称为检验的**p值**

1. 假设检验问题

- 例2：令总体服从$\mathrm{Exp}(1/\theta)$。$H_0:\theta \le \theta_0, H_1:\theta > \theta_0$
- 拒绝域为$W = \{(x_1, x_2, \dots, x_n) \mid \bar{x} \ge c\}$，$c$为待定值
- $x_i \sim \Gamma(1,1/\theta)$
- 回顾：$X \sim \Gamma(\alpha_1,\lambda), Y \sim \Gamma(\alpha_2,\lambda)$，$X,Y$相互独立，$X+Y \sim \Gamma(\alpha_1+\alpha_2,\lambda)$
- $n\bar{x} = \sum_{i=1}^{n} x_i \sim \Gamma(n,1/\theta)$
- 回顾：$X \sim \Gamma(\alpha,\lambda)$，若$k>0$，求$Y=kX$的概率密度函数
  - $f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$，$g(x)=kx$，$h(y)=\frac{y}{k}$，$|h'(y)| = 1/k$
  - $f_Y(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} (y/k)^{\alpha-1} e^{-\lambda y/k} \cdot \frac{1}{k} = \frac{(\lambda/k)^\alpha}{\Gamma(\alpha)} \cdot y^{\alpha-1} \cdot e^{-\lambda y/k}$
  - 也即$Y \sim \Gamma(\alpha,\lambda/k)$
- 因此$\frac{2n\bar{x}}{\theta} \sim \Gamma(n,1/2)$，也即$\frac{2n\bar{x}}{\theta} \sim \chi^2(2n)$

1. 假设检验问题

- $\frac{2n\bar{x}}{\theta} \sim \chi^2(2n)$
- Neyman-Pearson原则： 对于全部$\theta \le \theta_0$，$P(\bar{x} \ge c) \le \alpha$
- 也即对于全部$\theta \le \theta_0$，$P\left(\frac{2n\bar{x}}{\theta} \ge \frac{2nc}{\theta}\right) \le \alpha$
- 令Φ为$\chi^2(2n)$的分布函数，则有对于全部$\theta \le \theta_0$，$1-\Phi\left(\frac{2nc}{\theta}\right) \le \alpha$
- 也即对于全部$\theta \le \theta_0$，$\frac{2nc}{\theta} \ge \Phi^{-1}(1-\alpha) \Rightarrow c = \frac{\Phi^{-1}(1-\alpha)\theta_0}{2n}$
- 也即 $W = \left\{(x_1, x_2, \dots, x_n) \,\middle|\, \bar{x} \ge \frac{\Phi^{-1}(1-\alpha)\theta_0}{2n}\right\}$
- $p = 1-\Phi\left(\frac{2n\bar{x}}{\theta_0}\right)$

2. 正态总体参数假设检验

- 令总体服从 $N(\mu, \sigma^2)$，给定简单随机样本$x_1, x_2, \ldots, x_n$
- 三种关于μ的检验问题
  - $H_0: \mu \leq \mu_0, H_1: \mu > \mu_0$
  - $H_0: \mu \geq \mu_0, H_1: \mu < \mu_0$
  - $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$
- σ^2已知或未知
  - 若 σ^2已知，考虑统计量$\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
  - 若 σ^2未知，考虑统计量$\frac{\bar{x}-\mu_0}{\sqrt{s^2}/\sqrt{n}}$

北京大学 PEKING UNIVERSITY

2. 正态总体参数假设检验

- 令总体服从 $N(\mu,\sigma^2)$，给定简单随机样本$x_1,x_2,\dots,x_n$，σ^2已知
- $H_0:\mu\le\mu_0, H_1:\mu>\mu_0$
- 考虑统计量$u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

- 拒绝域为$W=\{(x_1,x_2,\dots,x_n)\mid u\ge c\}$，$c$为待定值
- Neyman-Pearson原则：对全部$\mu\le\mu_0$，出现第一类错误的概率不超过α
- 当$\mu=\mu_0$时，出现第一类错误的概率的概率最大，此时$u\sim N(0,1)$
- 因此$c=z_\alpha=\Phi^{-1}(1-\alpha)$，$\Phi$为标准正态分布的分布函数

- p值为满足$u\ge z_\alpha=\Phi^{-1}(1-\alpha)$最小的$\alpha$，也即$p=1-\Phi(u)$

2. 正态总体参数假设检验

- 令总体服从$N(\mu,\sigma^2)$，给定简单随机样本$x_1,x_2,\dots,x_n$，σ^2已知
- $H_0:\mu\geq\mu_0, H_1:\mu<\mu_0$
- 考虑统计量$u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

- 拒绝域为$W=\{(x_1,x_2,\dots,x_n)\mid u\leq c\}$，$c$为待定值
- Neyman-Pearson原则：对全部$\mu\geq\mu_0$，出现第一类错误的概率不超过α
- 当$\mu=\mu_0$时，出现第一类错误的概率的概率最大，此时$u\sim N(0,1)$
- 因此$c=-z_\alpha=-\Phi^{-1}(1-\alpha)$，$\Phi$为标准正态分布的分布函数

- p值为满足$u\leq-z_\alpha=-\Phi^{-1}(1-\alpha)=\Phi^{-1}(\alpha)$最小的$\alpha$，也即$p=\Phi(u)$

2. 正态总体参数假设检验

- 令总体服从 $N(\mu,\sigma^2)$，给定简单随机样本$x_1,x_2,\dots,x_n$，σ^2已知
- $H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$
- 考虑统计量$u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$

- 拒绝域为$W=\{(x_1,x_2,\dots,x_n)\mid |u|\geq c\}$，$c$为待定值
- Neyman-Pearson原则：出现第一类错误的概率不超过α
- 当$\mu=\mu_0$时，$u\sim N(0,1)$
- 因此$c=z_{\alpha/2}=\Phi^{-1}(1-\alpha/2)$，$\Phi$为标准正态分布的分布函数

- p值为满足$|u|\geq z_{\alpha/2}=\Phi^{-1}(1-\alpha/2)$最小的$\alpha$，也即$p=2(1-\Phi(|u|))$

2. 正态总体参数假设检验

- 令总体服从 $N(\mu,\sigma^2)$，给定简单随机样本$x_1,x_2,\dots,x_n$，σ^2未知
- $H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$
- 考虑统计量$t=\frac{\bar{x}-\mu_0}{\sqrt{s^2}/\sqrt{n}}$

- 拒绝域为$W=\{(x_1,x_2,\dots,x_n)\mid |t|\geq c\}$，$c$为待定值
- Neyman-Pearson原则：出现第一类错误的概率不超过α
- 当$\mu=\mu_0$时，$t\sim t(n-1)$
- 因此$c=t_{\alpha/2}(n-1)=\Phi^{-1}(1-\alpha/2)$，$\Phi$为$t(n-1)$的分布函数

- p值为满足$|t|\geq t_{\alpha/2}(n-1)=\Phi^{-1}(1-\alpha/2)$最小的$\alpha$，也即$p=2(1-\Phi(|t|))$

2. 正态总体参数假设检验

- 令总体服从 $N(\mu,\sigma^2)$，给定简单随机样本$x_1,x_2,\dots,x_n$，μ和σ^2均未知
- $H_0:\sigma^2=\sigma_0^2, H_1:\sigma^2\neq\sigma_0^2$
- 考虑统计量$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$

- 拒绝域为$W=\{(x_1,x_2,\dots,x_n)\mid \chi^2\leq a$或$\chi^2\geq b\}$，$a,b$为待定值
- Neyman-Pearson原则：出现第一类错误的概率不超过α
- 当$\sigma^2=\sigma_0^2$时，$\chi^2\sim\chi^2(n-1)$
- 令$F(x)$为$\chi^2(n-1)$的分布函数，$a=F^{-1}(\alpha/2), b=F^{-1}(1-\alpha/2)$

2. 正态总体参数假设检验

- 与置信区间的联系
- 令总体服从 $N(\mu,\sigma^2)$，给定简单随机样本$x_1,x_2,\dots,x_n$，σ^2已知

- $H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$
- 令$u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$，拒绝域$W=\{(x_1,x_2,\dots,x_n)\mid |u|\geq c\}$
- 接受域$\overline{W}=\{(x_1,x_2,\dots,x_n)\mid |u|\leq c\}$
- 当 $\mu=\mu_0$，$P(|u|\leq c)=1-\alpha$，对应于置信水平为$1-\alpha$的置信区间

- $H_0:\mu\leq\mu_0, H_1:\mu>\mu_0$
- 令$u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$，接受域$\overline{W}=\{(x_1,x_2,\dots,x_n)\mid u\leq c\}$
- 当 $\mu=\mu_0$，$P(u\leq c)=1-\alpha$，对应于置信水平为$1-\alpha$的单侧置信上限

3. 大样本假设检验

- 若总体均值为θ，方差为$\sigma^2(\theta)$
- $H_0: \theta = \theta_0, H_1: \theta \neq \theta_0$

- 当样本容量$n \to \infty$，当$\theta = \theta_0$，$W = \frac{\bar{x} - \theta_0}{\sqrt{\sigma^2(\theta_0)}/\sqrt{n}}$近似服从$N(0,1)$
- 拒绝域为$W = \{(x_1, x_2, \ldots, x_n) \mid |W| \geq z_{\alpha/2}\}$

- 可替换$\sigma^2(\theta_0)$为对应的估计量

3. 大样本假设检验

- 例1：令总体服从 $B(1,p)$，给定简单随机样本$x_1, x_2, \dots, x_n$
- $H_0: p = p_0, H_1: p \neq p_0$
- $W = \frac{\bar{x}-p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}}$或 $W = \frac{\bar{x}-p_0}{\sqrt{\bar{x}(1-\bar{x})}/\sqrt{n}}$
- 拒绝域为$W = \{(x_1, x_2, \dots, x_n) \mid |W| \geq z_{\alpha/2}\}$

- 例2：令总体服从 $\pi(\lambda)$，给定简单随机样本$x_1, x_2, \dots, x_n$
- $H_0: \lambda = \lambda_0, H_1: \lambda \neq \lambda_0$
- $W = \frac{\bar{x}-\lambda_0}{\sqrt{\lambda_0}/\sqrt{n}}$
- 拒绝域为$W = \{(x_1, x_2, \dots, x_n) \mid |W| \geq z_{\alpha/2}\}$

北京大学 PEKING UNIVERSITY

后续课程

- 随机算法：球与桶模型，Johnson Lindenstrauss Lemma，…
- 机器学习：交叉熵损失函数，多臂老虎机，…
- 计算复杂性： bounded-error probabilistic polynomial time
- 统计学/高维统计学：参数估计，假设检验，…
- 概率证法：discrepancy，Ramsey number，…

北京大学 PEKING UNIVERSITY

期末考试

- 时间：2025.01.03， 8:30-10:30。至少提前10分钟到达考场
- 地点：二教109

- 考试内容：全部内容，后半学期内容（从多维连续随变量开始）为主
  - 不包含：特征函数、中心极限定理、交叉熵损失函数、多元线性回归及其扩展

- 考试形式：闭卷。 可携带一张**双面有手写或打印内容**的A4纸。不允许使用包括计算器在内的任何电子设备。

北京大学 PEKING UNIVERSITY

期末考试前后安排

- 第八次作业（今晚或明天中午发布），截止日期为期末考试当天

- 期末考试复习课，第十七周，待投票确定具体时间
- 答疑：下周二下午2-3pm，静园五院206-2

- 期末查卷：如有，为第十八周某个上午+下午。
- 成绩：全部成绩（期中+期末+作业）会上传至教学网。有问题可查卷日处理

本科期末课程评估指导语

各位同学：

课程评估是学校本科教学质量保障的重要环节，对保障和提升教学质量至关重要。课程评估结果对学校院系规范教学管理和提升教学质量有着重要作用，同时也是任课教师改进和调整教学的重要依据。只有各位同学认真负责，提供有意义的反馈意见，才能够为教学管理和课程教学提供有效信息，真正促进教学改进和提升。

衷心感谢各位同学参加本学期课程评估，同时希望同学们给予课程更多的改进和提升建议。具体参与方式如右所示。

一、电脑端登录

1、登录网上评估系统（kcpg.pku.edu.cn）。

2、输入『学号』及『密码』（与校内门户一致）完成登录。

3、填写任务列表中对应的课程评估任务，填写问卷并点击『提交』。

二、手机端登录

1、用微信扫描如下二维码，关注“本科课程评估”。

2、点击首页——输入『学号』及『密码』登录——任务评价；非本校同学请点击个人设置——校外绑定——输入『学号』、『密码』默认为学号。

3、根据我的任务中的课程，填写问卷并点击『提交』。

教务部教育教学评估办公室