概率统计B

Probability and Statistics

张思容
zhangsirong@buaa.edu.cn

数学与系统科学学院，北京航空航天大学
School of Mathematics and System Sciences, BUAA

December 12, 2014

Chapter 5: 经典统计推断

1. 参数估计
   - 矩估计
   - 最大似然估计MLE
   - 点估计的评估***

2. 区间估计
   - 置信区间与置信水平
   - 一个总体的区间估计

3. 假设检验
   - 实际推断原理
   - 检验错误:显著性水平
   - 一个正态总体的参数检验

点估计

假定总体分布X的模型已知但参数未知。

Definition (参数统计量 statistics)

给定简单随机样本$X_1, X_2, \ldots, X_n$, 设θ是总体分布的未知参数, 构造函数$g(x_1, x_2, \ldots, x_n)$, 称$\hat{\theta} = g$为θ的估计量或统计量;

常见估计:

- 样本均值 $\bar{X}_n = \sum X_i \to EX$ 期望
- 样本方差$S^2 = \sum(X_i - \bar{X})^2/(n-1) \to DX = EX^2 - (EX)^2$ 方差, 样本标准差$s \to \sqrt{DX}$ 标准差
- ***样本相关系数 $\hat{\rho_{xy}} = \frac{s_{xy}}{s_x s_y} \to \rho$.

Definition (矩估计 moment statistics)

给定简单随机样本$X_1, X_2, \ldots, X_n$, 如果$v_k = EX^k$存在, 则定义$\hat{V}_k = \frac{1}{n}\sum_i X_i^k$是$v_k$的矩估计统计量。更广义的: 构造函数$g(\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_n)$,去估计 未知参数$\theta$, 称为$\theta$的矩估计量。

矩估计的例子

EXAMPLE (正态分布)

$X \sim N(\mu, \sigma^2)$,参数$\mu = EX, \sigma^2 = DX = EX^2 - (EX)^2$,
则矩估计$\hat{\mu} = \hat{v}_1 = \bar{X}_n$, $\hat{\sigma^2} = \hat{v}_2 - \hat{v}_1{}^2 = \frac{1}{n}\sum(X_j - \hat{\mu})^2$.
任意分布的均值与方差的矩估计的表达式同上。
(注: 与教材说法不一样，关键在矩的定义中通常系数为$1/n$)).

EXAMPLE (均匀分布)

$X \sim U(0, b)$, 则$b = 2EX$, 矩估计$\hat{b} = 2\bar{X}_n$.
一般的$U(a, b)$有两个参数，利用$EX = (a + b)/2$, $DX = (b - a)^2/12$,解方程可得$a, b = v_1 \pm \sqrt{3(v_2 - v_1^2)}$.

- 指数分布$\epsilon(\lambda)$，$EX = 1/\lambda \rightarrow \hat{\lambda} = 1/\bar{X}_n$;
- 泊松分布,$EX = \lambda \rightarrow \hat{\lambda} = \bar{X}_n$, 注: $DX = \lambda, \hat{\lambda}_2 = \frac{1}{n}\sum(X_j - \hat{\mu})^2$???
- 二项分布$B(N, p)$，$EX = Np \rightarrow \hat{p} = \bar{X}_n/N$; 特别社会调查问题: 比例的估计$B(1, p)$, $\hat{p} = \bar{X}_n$

最大似然估计Maximum Likelihood Estimation

EXAMPLE (选择判断)

设某一天下雨的概率p为0.1或0.9，今天没有下雨，问$p=?$
则应该判断$p=0.1$；由结果判断概率的思想称为最大似然思想。如果连续几天没下雨，判断更准确!

一般存在一个似然函数(likelihood function)，已知样本值，求使得似然函数取极大值所对应解来估计参数。

Definition (MLE)

给定关于未知参数θ和样本值的似然函数$L(x_1, x_2, \ldots, x_n; \theta)$，定义$\hat{\theta}$为使得$L$取最大值的解，称$\hat{\theta}$是$\theta$的最大似然估计.

- 离散情形：$L(x_1, x_2, \ldots, x_n; \theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$
- 连续情形：$L(x_1, x_2, \ldots, x_n; \theta) = f(x_1, x_2, \ldots, x_n; \theta)$;
- 求解最大值：$\frac{dL}{d\theta} = 0$，或者$\frac{d\ln L}{d\theta} = 0$. (对数似然函数).
- 多维情形$(\theta_1, \theta_2, \ldots)$：$\frac{\partial L}{\partial \theta_i} = 0, \forall i$. 现实中可能要用数值解法求解。

MLE的例子

回顾:(估计某地区的某动物总数N)
捕捉一批动物M,做标记放回,过一段时间,再捕捉一批动物n, 其中有标记的为k,利用超几何分布估计N。 概率为$P(X=k)=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$是关于$N$的一个序列， 求最大值对应的$N$,有$N=nM/k$. 称为最大似然估计。

EXAMPLE (正态分布)

$X\sim N(\mu,\sigma^2)$, 计算有$\mu=\bar{X}_n$, $\hat{\sigma^2}=\frac{1}{n}\sum(X_j-\hat{\mu})^2$,与矩估计一致。

EXAMPLE (均匀分布)

$X\sim U(0,b)$, 则MLE估计 $\hat{b}=x_{(n)}$,即最大样本值。 对于$U(a,b)$,a对应最小样本值。$L(a,b)=\frac{1}{(b-a)^n}\le\frac{1}{(x_{(n)}-x_{(1)})^n}$.

- 指数分布$\epsilon(\lambda)$， MLE $\hat{\lambda}=1/\bar{X}_n$ 与矩估计一致
- 泊松分布,$\hat{\lambda}=\bar{X}_n$,与矩估计一致
- 两点分布$B(1,p)$, $\hat{p}=\bar{X}_n$， 与矩估计一致。

无偏估计

Definition (参数统计量 statistics)

给定简单随机样本$X_1, X_2, \ldots, X_n$, 设θ是总体分布的未知参数，构造函数$g(X_1, X_2, \ldots, X_n)$, 称$\hat{\theta} = g$为θ的估计量或统计量;

Definition (无偏统计量 Unbiased)

如果以上统计量满足$E(\hat{\theta}) = \theta$, 称$\hat{\theta}$是无偏估计。

Proposition

样本均值$\bar{X}_n$和样本方差$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$是总体分布的均值和方差的无偏估计。

直接验证: $E\bar{X}_n = \mu, E(S^2) = \sigma^2$。
注：方差的矩估计(或MLE估计)$\frac{1}{n}\sum(X_i - \bar{X}_n)^2$不是无偏估计。

无偏估计的例子

Theorem (总体矩的估计)

样本的k次矩是总体的k次矩的无偏估计。
即$\hat{V}_k = \frac{1}{n}\sum_i X_i^k$是$v_k = EX^k$的无偏估计。

注: $E(X_i^k) = v_k$.

- 方差的矩估计(或MLE估计)是有偏估计(n充分大时，相差无几.)
  *** 渐进无偏估计 $\lim_n E(\hat{\theta}_n) = \theta$.
  *** 样本标准差s是有偏估计。
- 均匀分布的矩估计: $\hat{b} = 2\bar{X}_n$是无偏估计, MLE估计$Ex_{(n)} = n/(n+1)b$不是。
- 无偏估计也有很多. 比如任意$\hat{\theta} = \sum a_i X_i, \sum a_i = 1$都是期望的无偏估计。

最小方差无偏估计MVU***

Definition (有效性)

如果同一个参数θ的两个统计量满足$D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$, 称$\hat{\theta}_1$比$\hat{\theta}_2$更有效。

Definition (最小方差无偏估计MVU)

给定无偏估计$\hat{\theta}_0$, 如果任一无偏估计$\hat{\theta}$满足，$D(\hat{\theta}_0) \leq D(\hat{\theta})$, 称$\hat{\theta}_0$是最小方差无偏估计。

- 样本均值是最小方差无偏估计MVU。(验证：线性估计中的方差最小, 柯西许瓦茨不等式).
- 均匀分布的MLE估计$D\hat{b} = \frac{n}{(n+2)(n+1)^2}b^2$比矩估计$D\hat{b} = b^2/(3n)$更有效。
- *最小方差无偏估计MVU是统计量的最理想估计，有相关的存在性定理。
- ***信号处理: 维纳滤波器是线性最小方差估计器。

一致性或相合性

估计的最基本性质：收敛性。

Definition (相合性)

给定 统计量$\hat{\theta}=g(X_1,X_2,\ldots,X_n)$与样本容量$n$有关，可记为$\hat{\theta}_n$, 如果$\hat{\theta}_n$依概率收敛到$\theta$(即$\hat{\theta}_n \xrightarrow{p} \theta$), 称$\hat{\theta}_n$是$\theta$的相合估计量(或一致估计)。

Theorem (矩估计相合性)

样本矩依概率收敛到随机变量的矩。 推广的有
$g(\hat{v}_1,\hat{v}_2,\ldots,\hat{v}_k) \xrightarrow{p} g(v_1,v_2,\ldots,v_k)$.
所以所有矩估计都是相合估计。

- ***事实上矩估计以概率1收敛. 通常称为强相合估计。
- 样本方差是相合估计。 $S^2 \xrightarrow{p} \sigma^2$. 特别情形(正态分布) P146例5.
- ***一般的MLE估计在一定条件下也是相合估计。

例子

EXAMPLE

设总体分布是指数分布$\epsilon(\lambda)$. 记参数$\theta = 1/\lambda$. 则
(1) $\bar{X}_n$是θ的无偏估计； *(2)* $Z = \min(X_1, X_2, \ldots, X_n)$,则$nZ$是$\theta$的无偏估计； *(3)* $\bar{X}_n$比nZ更有效(方差更小).

解答:.

(1),(2)直接验证
(3) $D(\bar{X}_n) = \theta^2/n$, $D(nZ) = \theta^2$. □

作业

北航教材:
P155 习题八 1,2,3,4,6,7,9

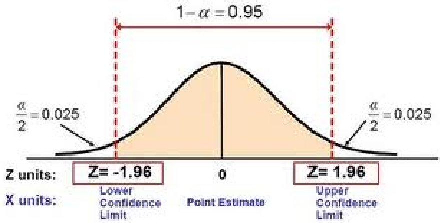
Review：回顾

Last Week:

- 参数估计: 矩估计与最大似然估计;
- 估计(统计量)的评价标准：无偏, 有效，一致。

This Week:

- 区间估计与置信度;
- 均值与方差的区间估计;



置信区间：给点估计一个误差范围。
置信水平: 0.05,0.01.

常见统计量的分布

- 样本常见特征的分布是统计推断的关键!
  一般通过样本计算(构造)出一些统计量: 样本均值，样本方差，样本相关系数，希望统计量的分布是已知的!
- 中心极限定理给出样本均值的逼近分布正态分布，其他特征呢?
- 基本假设:如果总体是正态分布，$X \sim N(\mu, \sigma^2)$,
  记$\bar{X}_n = \frac{1}{n}\sum X_i, S^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$.
- 样本均值与正态分布 $(\bar{X}_n - \mu)/\sigma/\sqrt{n} \sim N(0,1)$
- 样本方差与χ^2分布: $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$
- (未知总体方差) 样本均值与t分布: $\frac{\bar{X}_n-\mu}{S/\sqrt{n}} \sim t(n-1)$
- *(两个总体的均值差) 假设方差相同未知，P181定理五.
- *(两个总体的方差比) 两个独立总体X, Y,
  $\frac{S_X^2/\sigma_x^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1), n, m \geq 2$.

置信区间

Definition (置信区间)

假定给出两个估计$\hat{\theta}_1, \hat{\theta}_2$, 使得给定$\alpha$,满足$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) \geq 1-\alpha$,称$[\hat{\theta}_1, \hat{\theta}_2]$为$\theta$的置信水平为$1-\alpha$的置信区间。

- 注意$\hat{\theta}_1, \hat{\theta}_2$是随机变量(区间是随机区间!)，$\theta$是固定未知参数。$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = P(\hat{\theta}_1 \leq \theta \bigcap \theta \leq \hat{\theta}_2)$, 涉及两个随机变量的分布。
- 通常我们给出点估计的上下误差范围来代替两个估计,比如差值区间$[\hat{\theta}-a, \hat{\theta}+b]$ 或比例区间$[\hat{\theta}/a, \hat{\theta}*b]$, $a, b \geq 1$.特别如果$a=b$,记为$\hat{\theta} \pm b$，b又称为抽样误差。
- 选取区间满足概率等式: $P(-b \leq \hat{\theta}-\theta \leq a) = 1-\alpha$,已知$\hat{\theta}$的分布，则置信水平$1-\alpha$决定$a, b$, 从而置信区间的大小。
- 置信区间的频率解释: θ不一定在区间内! 但重复抽样得到的置信区间中有概率$(1-\alpha)$个区间包含θ.

置信水平与下分位数

Remark (由估计$\hat{\theta}$, 怎样计算抽样误差a, b?)

方法: 即构造概率区间满足：$P(-b \le \hat{\theta} - \theta \le a) = 1 - \alpha$.
设已知统计量$\hat{\theta} - \theta$的分布$F(x)$(其他统计量参见均值区间估计);
则记下分位数$F(x_{\alpha/2}) = \alpha/2, F(x_{1-\alpha/2}) = 1 - \alpha/2$,
取$-b = x_{\alpha/2}, a = x_{1-\alpha/2}$, 即可。

我们称$[\hat{\theta} - x_{1-\alpha/2}, \hat{\theta} - x_{\alpha/2}]$是置信水平为$1 - \alpha$关于$\theta$的置信区间的.
注意：我们取的区间使得$\hat{\theta}$在$[\theta - b, \theta + a]$两边的概率一样$(\alpha/2)$。也可以不一样。

- 应用中：如果分布是关于0点对称的,则$-x_{\alpha/2} = x_{1-\alpha/2}$，即区间为$[\hat{\theta} - x_{1-\alpha/2}, \hat{\theta} + x_{1-\alpha/2}]$。称$x_{1-\alpha/2}$是临界值。
- 一般的分布取标准正态分布或t分布。置信水平为0.95,0.99. 设$\alpha = 0.05$,则$Z_{1-0.025} = 1.96$，$\alpha = 0.01$,则$Z_{1-0.005} = 2.58$.
- 类似可定义: 单侧置信区间估计：$P(\hat{\theta} \ge \theta) = 1 - \alpha$ 或$P(\hat{\theta} \le \theta) = 1 - \alpha$. 对应区间为$[-\infty, \hat{\theta}]$或$[\hat{\theta}, +\infty]$.

案例分析一

EXAMPLE (测量误差)

我们想得到鲜牛奶的冰点μ。假设温度计的测量值服从正态分布$N(\mu,\sigma^2)$. 用一个测量误差为$e=0.0048$的温度计测量21次(数据略). 计算有样本均值$\bar{X}_n=-0.546$,样本标准差$S=0.005$.

- 可以给出估计$\hat{\mu}=\bar{X}_n=-0.546$, $\hat{\sigma}^2=S^2=0.005^2$;
- 假设已知$\sigma=e=0.0048$,则μ的置信水平为95%的置信区间为 $[\bar{X}_n-1.96\sigma/\sqrt{n},\bar{X}_n+1.96\sigma/\sqrt{n}]$;
- 假设σ未知,则μ的置信水平为95%的置信区间为 $[\bar{X}_n-2.086S/\sqrt{n},\bar{X}_n+2.086S/\sqrt{n}]$;
- 方差σ^2的置信水平为95%的置信区间为 $[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}]\approx[0.0038,0.0072]$;

对应的统计问题:

- 点估计
- (已知方差)大样本下均值的区间估计;
- (未知方差)小样本下均值的区间估计;
- 方差的区间估计;

均值估计:z, t分位数

设总体分布是正态分布$N(\mu, \sigma^2)$, 怎样给出μ的一个区间估计?
方差未知, ®假定$\sigma = 0.0048$ ®

1. 点估计$\hat{\mu} = \bar{X}_n = -0.546$,$\hat{\sigma}^2 = S^2 = 0.005^2$;
2. 统计量分布: $\bar{X}_n \sim N(\mu, \sigma^2/n)$; (一般情形由中心极限定理逼近)
3. 设定95%的置信度,选对称概率区间,查分位数表$Z_{1-0.025} = 1.96$。
   概率公式$P(|\bar{X}_n - \mu| \leq b) = P(\frac{|\bar{X}_n-\mu|}{\sigma\sqrt{n}} \leq z_{1-0.025}) = 0.95$
4. $b = 1.96\sigma/\sqrt{n}$,求解出置信区间上下限$\pm b = 0.002$。
   $\mu \in [-0.546 - 0.002, -0.546 + 0.002]$.

一般情况方差未知, 怎么办? 用样本方差。

- 统计量分布: ®假设总体是正态分布®,则 $\frac{\bar{X}_n-\mu}{s/\sqrt{n}} \sim t(n-1)$.
- 置信水平$1-\alpha = 0.95$,取对称区间$P(\frac{|\bar{X}_n-\mu|}{s\sqrt{n}} \leq t_{1-\alpha/2}(n-1)) = 0.95$
  查表$t_{1-0.025}(20) = 2.086$.
- 求解b得到置信区间为$[-0.5483, -0.5437]$.

说明: 以上t分位数与z分位数的估计差不多(稍大),因为$n \geq 33$时与正态分布相近; 一般用于小样本估计。

方差估计：χ^2分位数

温度计给出的测量误差$e = 0.0048$是否准确？用样本方差去估计。

1. 点估计：$S^2 = 0.005^2$;
2. 统计量分布：总体是正态分布，样本方差服从χ^2分布，$(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$
3. 设置信水平$1-\alpha = 0.95$,选择概率区间
   $P(\chi^2_{\alpha/2}(n-1) \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}(n-1)) = 0.95$
   查表 $\chi^2_{0.025}(20) = 9.591, \chi^2_{1-0.025}(20) = 34.170,$
4. 求解区间上下限，得到σ^2的区间估计$[0.0038^2, 0.0072^2]$.

相关说明：

- 可以得到标准差σ的区间估计$[0.0038, 0.0072]$. 注意其关于$s = 0.005$的非对称性。
- 由于χ^2非对称性，区间的选择有很多种，我们的选择仅仅比较容易计算，不一定是最短的区间。

其他区间估计问题***

EXAMPLE (总体比例估计问题：早餐调查。)

设大学生吃早餐的比例p。抽样n个人，其中m个人吃早餐。
点估计$p = \bar{X}_n$，总体分布为二项分布，n充分大，用正态分布逼近，
得区间估计：$\bar{X}_n \pm z_{1-\alpha/2}\sqrt{\bar{X}_n(1-\bar{X}_n)/n}$.
区间大小与n：设抽样误差d，$d = 0.03, 0.02, 0.01$; $n = 1068, 2401, 9604$.

非正态总体的区间估计：用正态总体逼近计算:

- 均值估计：方差已知，有中心极限定理，
- 均值估计：方差未知：小样本情形t分布，可用正态分布逼近($n \geq 30$).
- 方差估计：非正态总体没法估计，有其他统计方法；χ^2检验，ANOVA方差分析.

两个(多个)正态总体的比较估计：

- 两个总体的均值差：服从正态分布或某种t分布，依赖与两个总体的方差信息。
- 两个总体的方差比：服从$F(m, n)$分布。

作业

北航教材:
P156 习题八 10.11.13.14

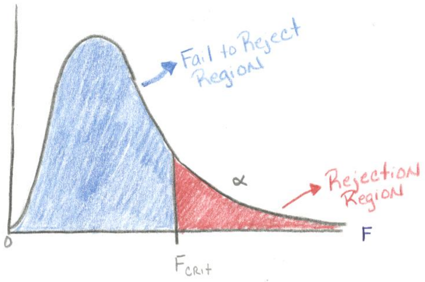
Review: 回顾

Last Week:

- 区间估计与置信度;
- 均值与方差的区间估计;

This Week:

- 假设检验的原理
- 一个正态总体的参数检验
- 其他统计方法;



区间估计 → 假设检验
计算题 → 判断题(或选择题)

简单统计决策的过程:实际推断原理

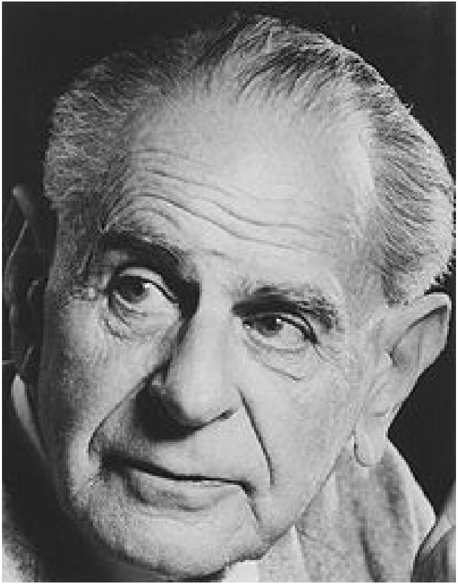
用一个样本统计量的值: "选择判断" 关于未知信息的可能.

EXAMPLE (简单假设检验的过程)

给出H_0, H_1两个假设 $\xrightarrow{sampling}$ $X_1, X_2, \ldots, X_n$ $\xrightarrow{T}$ T统计量的分布$P_{H_0}(T), P_{H_1}(T)$ $\xrightarrow{decision}$ 利用$\hat{T}$ 选择H_0或H_1.

- 实际推断原理: 如果一个很小概率事件(显著的意外)发生了， 说明假设存在不合理的。
- 假设检验与反证法:
- 例子： 抽样调查: p是吃早餐的比例可能为$0.4, 0.5, 0.6$. 调查100人有60人吃早餐。 问p更可能是什么? "肯定"不是0.4, 但不一定是$0.6, 0.5$.

Karl Popper: 批判理性主义



卡尔 · 雷蒙德 · 波普尔爵士
(1902年7月28日－1994年9月17日）
二十世纪最伟大科学哲学家;
科学研究 → 个人选择 → 社会
证伪主义 → 试错 → 开放社会。
In search of a better world.
"通过知识获得解放".

惊人的预测：统计思考

网络骗局:
一封电子邮件: 存在高级统计方法以0.95概率预测(足球或篮球)比赛结果。
第一封预测： 成功!
第二封预测： 成功!
第三封预测： 成功!
第四封预测： 成功!
第五封预测： 成功!
第六封信： 请汇款200元买我们的方法。

- 概率思考: 五次成功的概率$0.5^5 \sim 0.0313$;
  可能有内幕? 值得买!!!
- 统计思考: 骗子发8000封邮件，一半猜A赢，一半猜B赢；
  第二次： 仅发新预测给4000个得到正确结果的人；
  第三次： 发给2000人； 第四次： 发给1000人；
  第五次： 发给500人。最后要钱的电邮发给250人即可!

实际推断原理: 仅仅是一次的判断。
但大量试验则小概率事件必然发生。

决策错误与显著性水平

给定两个假设H_0, H_1, 简单随机样本和检验统计量$T = g(X_1, X_2, \ldots, X_n)$.

Definition (拒绝域)

指定区域W, 假定H_0成立，如果统计量的值$t = g_{H_0}(x_0, x_1, \ldots, x_n) \in W$, 则推断$H_0$不成立$(H_1$成立$)$。称$W$为拒绝域。

- 一般称H_0原假设或零假设，H_1为备择假设；希望证伪原假设!
- 通常区域写成$W = \{t : |g(X_1, X_2, \ldots X_n)| > t_0\}$. t_0称为临界值。
- 临界值由事先确定的事件的概率大小决定。即$P_{H_0}(t > t_0) = \alpha$, 称为(统计)显著性水平。通常取为0.05, 001, 0.1.

Definition (简单决策的两类错误)

如果真实情况下，原假设是正确的，我们抽样得到的结论否定原假设，称为 第一类错误；如果真实情况下，原假设是错误的，我们抽样得到的结论不拒绝原假设，称为 第二类错误；记第一类错误的概率为α,第一类错误的概率为β; 显著性水平即第一类错误的概率。

例子: 硬币是否是均匀的?

EXAMPLE (硬币的公平性)

给定一个硬币，问其是否是公平的？即投硬币为正面的概率为0.5.
设投硬币100次，其中31次为正面.

解答:

- 假设$H_0 : p = 0.5$ vs $H_1 : p \neq 0.5$
- 定义$T = \frac{1}{n}\sum X_i$, $X_i \sim B(1, p)$. 则$nT \sim B(n, p)$.
- 设显著水平$\alpha = 0.05$,则拒绝域: $P(|T - p| > t_0) = \alpha = 0.05$.
  利用$Z = \frac{T-p}{\sigma/\sqrt{n}} \sim N(0, 1)$, $z_{0.975} = 1.96$. 和$p = 0.5$!!!
  $|\frac{T-p}{\sigma/\sqrt{n}}| > 1.96$, 计算有$|T - p| > 1.96\sigma/\sqrt{n}$, $t_0 = 0.098$.
- 判断: $|T - p| = 0.19 > t_0$，所以原假设否定，(统计显著意义下) 硬币不是公平的。

说明：如果$T = 0.41$呢？不知道。显著性水平$0.1, 0.01$？P值为0.07.

假设检验的类型

常见假设的类型:

- 参数检验: $H_0 : \theta \in D$ vs. $H_1 : \theta \notin D$,D是一个集合。
- 非参数检验: $H_0 : F = F(x)$ vs. $H_1 : F \neq F(x)$,$F(x)$是一个分布函数。

参数检验的类型

- 双边参数检验: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- 单边检验;
  左边检验 $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$.
  右边检验 $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$.

***假设检验的若干问题

1. 错误是不可避免的。给出样本量大小, 缩小第一类错误α, 会加大第二类错误β. 我们通常控制第一类错误的大小。
2. 假设的选择: H_0选为历史数据给定或保守选择。 仅仅当有充分证据否定原假设, 才采用备择假设,这样错误的代价较小。
3. 结论: 如果小概率事件发生,可以否定原假设. 如果小概率事件没有发生,不能肯定原假设.只能说明没有充分证据, 需要进一步考察!
4. 显著性水平: $\alpha = 0.05$,或0.06有区别吗? 必须考察实际应用的背景。

应用的谬误:

- 统计显著性与实际显著性:
  样本量充分大后, 一般会得到统计显著性, 拒绝原假设。
  统计显著性必须与实际问题相结合考虑问题。 比如差异是否重要(经济考虑)。 统计显著性仅仅是第一步, 我们需要进一步发现为什么有差异?
- 不能对一个数据做多次假设检验。比如做二十次检验, 很可能有一次的检验结果是显著的。

案例：质量检验

EXAMPLE (case study: 白糖的质量检验)

厂家: 希望生产的每包白糖的质量服从正态分布$N(500, 0.8)$.
商家: 希望得到的白糖不要短斤少两。每包白糖约为500克。
抽样9包，计算有$\bar{X}_n = 499.412, s = 0.676$。

- 厂家: 假设检验机器正常工作。
- 均值检验: $H_0 : \mu = 500 vs H_1 : \mu \neq 500$
  正态检验结果: 机器不正常。$\alpha = 0.05$
- 方差检验: $H_0 : \sigma^2 = 0.8 vs H_1 : \sigma^2 \neq 0.8$
  或 单边方差检验: $H_0 : \sigma^2 \leq 0.8 vs H_1 : \sigma^2 > 0.8$
  χ^2检验结果: 机器是否正常不确定。
- 商家：均值检验: $H_0 : \mu = 500 vs H_1 : \mu \neq 500$
  t检验结果: 质量不是500克。$\alpha = 0.05$
- 单边均值检验: $H_0 : \mu \geq 500 vs H_1 : \mu < 500$
  t检验结果: 质量不足500克。$\alpha = 0.05$

均值检验：正态检验

厂家: 假设检验机器正常工作。

- 假设 $H_0 : \mu = 500 vs H_1 : \mu \neq 500$
- ®假设总体是正态分布$N(500, 0.8)$ ®, 检验统计量
  $Z = \frac{\bar{X}_n - 500}{\sqrt{0.8/n}} \sim N(0, 1)$
- 如果H_0成立，$|Z|$应该比较小，反之应该比较大；
  特别取$\alpha = 0.05$，使得$P(|Z| > z_{\alpha/2}) = \alpha = 0.05$, 即 $|Z| > 1.96$时为小概率事件.
- 代入数据，$|Z| = |(499.412 - 500)/\sqrt{0.8/9}| = 1.97 > 1.96$,
  小概率事件发生了，应该否定原假设，即$\mu \neq 500$.

说明:

- $\{|Z| > z_{\alpha/2}\}$称为拒绝域，检验统计量位于其中则拒绝原假设。
  如果该事件发生，称检验为显著的。使用正态分布决定拒绝域，称为正态检验。
- 现代统计软件可以计算$|Z| \geq 1.97$对应的概率$P = 0.0488$。
  由于$P < 0.05$否定原假设，称为P值检验。

小样本均值检验：t检验

商家不可能知道方差，要用样本方差，对应有t检验

- 假设 $H_0 : \mu = 500 vs H_1 : \mu \neq 500$
- ®假设总体是正态分布®，则样本方差服从χ^2分布；则 检验统计量$T = \frac{\bar{X}_n - \mu}{s\sqrt{n}} \sim t(n-1)$.
- 如果H_0成立，$|T|$应该比较小，反之应该比较大；
- 类似查表有$t_{0.025}(8) = 2.306$, 计算$|T| = 2.609$, 处于拒绝域， 否定原假设。
- 用P检验，$P(|T| > 2.609) = 2P(T > 2.609) = 2 * 0.01559 == 0.03118 < 0.05$,否定原假设。

商家可以有理由拒绝该批白糖吗?
其实$\mu > 500$是可以接受的。
我们更关心$\mu < 500$,称为单边检验。

小样本均值检验：单边t检验

$\bar{X}_n = 499.412$,前面t检验显著，我们怀疑短斤少两!

- 假设$H_0: \mu \geq 500 vs H_1: \mu < 500$
- ®假设总体是正态分布®，同样 $T_0 = \frac{\bar{X}_n - \mu}{s\sqrt{n}} \sim t(n-1)$.
- 检验统计量$T = \frac{\bar{X}_n - 500}{\sqrt{0.676 * 9}}$的分布未知!
  但是如果假设H_0成立，$T \geq T_0 = \frac{\bar{X}_n - \mu}{\sqrt{0.676 * 9}}$.
- 要控制T的概率, 如果$T > T_0$, 选左边区间!!!(见后面解释)
  有 $P(T < -t_{0.05}(8)) \leq P(T_0 < -t_{0.05}(8)) = 0.05$
- 查表$-t_{0.05}(8) = -1.86$, $T = -2.609 < -1.86$,所以否定原假设，即该批产品质量不足。

相关说明：

- 一般的假设可以是 $H_0: \mu \leq \mu_0, H_0: \mu = \mu_0, H_0: \mu \geq \mu_0$
  第一，三情形都用单边(t)检验。
- 单边与双边检验查表时分位数不同!
- 第一种情形(案例一）：牛奶兑水会改变冰点-0.545 (变高)。

方差检验：χ^2检验

厂家关心质量控制的重点是方差的大小。 机器不正常工作的另一个标志是质量不稳定，实际方差大。

- 假设$H_0 : \sigma^2 \leq 0.8 vs H_1 : \sigma^2 > 0.8$
- 总体是正态分布，样本方差服从χ^2分布，$\xi_0 = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$
- 如果假设H_0成立，检验统计量$\xi = (n-1)S^2/0.8 \leq \xi_0$.
  要控制ξ的概率，如果$\xi \leq \xi_0$，选右边区间!!! (见后面解释)
  $P(\xi > \chi^2_{0.05}) < P(\xi_0 > \chi^2_{0.05}) = 0.05$
  取单边拒绝域$\xi \geq \chi^2_{0.05}$，则其发生概率小于0.05.
- 查表$\chi^2_{0.05}(8) = 15.507$,计算$\xi = 8 * (0.676)^2/0.8 = 4.57 < 15.507$ 不能拒绝原假设。
- 说明该机器的质量比较稳定，主要问题是均值!(度量问题？)

相关说明：

- 对应的也可以用P值检验；
- 也可以先做χ^2双边检验.结果同样不显著。

单边检验的拒绝域：解释

设单边检验$H_0: \mu \le \mu_0$或$H_0: \mu \ge \mu_0$

- 单边检验的统计量T是未知分布(μ有不同取值)。但一般与最大或最小值对应的分布$T_0: \mu = \mu_0$有大小关系。 具体情况见前面两页：
- 如果随机变量$T > T_0$,则有分布函数$F_T(x) \le F_{T_0}(x)$, 即$P(T < x) \le P(T_0 < x)$; (证明：事件$\{T < x\} \subset \{T_0 < x\}$.) 分布密度图像看： T的分布密度函数在T_0的右边。
- 如果随机变量$T > T_0$, 应该选左边区域为拒绝域，因为拒绝域的概率$P(T < x)$有$P(T_0 < x) = \alpha$控制. 若选右边区域，则无法控制。
- (类似)随机变量$T < T_0$,有$F_T(x) \ge F_{T_0}(x)$即$P(T > x) \le P(T_0 > x)$. 说明：事件$\{T > x\} \subset \{T_0 > x\}$. 图像上$T$的分布密度函数在$T_0$的左边。
- 如果随机变量$T < T_0$, 应该选右边区域为拒绝域，因为拒绝域的概率$P(T > x)$有$P(T_0 > x) = \alpha$控制. 若选左边区域，则无法控制。

其他检验与统计问题***

两个或多个正态总体的假设检验(非正态总体用正态总体逼近计算):

- 两个成对总体的均值差: 服从t分布;
- 两个总体的均值差：服从正态分布或某种t分布，依赖与两个总体的方差信息。
- 两个总体的方差比：服从$F(m, n)$分布。

其他检验: 分布检验，秩检验(Wilcox), 偏度，峰度检验;

- 假设检验与区间估计的关系: α是一个;
  如果假设成立$\theta = \theta_0$,则拒绝域可定义为置信区间的外面!
- 相关分析: 用s_{xy} 估计σ_{xy};
  线性回归分析:$Y = aX + b$ 假设检验: $b = 0$;
- 方差分析: 非正态总体没法估计，有其他统计方法; χ^2检验, ANOVA方差分析.
- 抽样方法改进： bootstrap 从一个样本中得到若干个自助样本(每一个自助样本通过在样本中有放回抽取n次得到);

作业

北航教材:
P177 习题九 1.2.3.4.5.6
补充: 根据以下历史数据(正面数/总次数)，判断其硬币是否公平$p = 0.5$。 $\alpha = 0.1, 0.05, 0.01$.

- De Morgan 德摩根： 1061/2048
- 蒲丰Buffon： 2048/4040
- John Kerrich: 5067/10000
- 费勒 Feller: 4979/10000;
- 罗曼洛夫斯基 40173/80664;