统计计算

2016 年 3 月 1 日

目录

什么是统计

Statistics is the science of collecting, analyzing, presenting, and interpreting data. Governmental needs for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics. Currently the need to turn the large amounts of data available in many applied fields into useful information has stimulated both theoretical and practical developments in statistics.

— Encyclopedia Britannica

《统计科学百科全书》是迄今最完整的关于统计的具有权威性的百科全书:

- Samuel Kotz, Campbell B. Read, N. Balakrishnan, Brani Vidakovic, Eds., Encyclopedia of Statistical Sciences, 16 Volume Set, Second Edition, John Wiley & Sons, 9686 pages, December 2005.

(Website: http://as.wiley.com/WileyCDA/WileyTitle/productCd-0471150444.html)

其中, "统计学" 这个术语表示为 "涉及收集、表示和分析数据的普遍方法和原理的领域", 该书还列举了四十多个应用统计的领域.

统计学具体应用领域

统计广泛地应用于各种具体领域：社会、政治、经济、金融、管理、工程、农业、医学、生物、军事、航空航天等等.

由于每个领域具有各自的特点，统计方法也有很大区别.

近年来，医学统计、生物统计中不断出现新的统计理论和方法.

什么是数理统计

数理统计独立于某些具体应用领域的表面现象，使用复杂的数学工具，研究具有普遍性的统计理论和方法.

- 数学以公理系统为基础，以演绎为基本思想和方法.
  (纯) 数学可以完全脱离实际，因而国外过去一般认为 (纯) 数学是艺术 (art) 而不是科学 (science)，因为科学是研究实际事物的；
- 统计以归纳为其基本思想，归纳和演绎并用.
  统计学与数学相反，统计是以实际事物为研究和处理对象的.

统计学与数学

- 数学以公理系统为基础, 以演绎为基本思想和方法.
  (纯) 数学可以完全脱离实际, 因而国外过去一般认为 (纯) 数学是艺术 (art) 而不是科学 (science), 因为科学是研究实际事物的;
- 统计以归纳为其基本思想, 归纳和演绎并用.
  统计学与数学相反, 统计是以实际事物为研究和处理对象的.

统计与概率

- 概率: 公理体系, 演绎, 模型 → 性质, 应用;
- 统计: 应用问题 → 数据 → 原理, 方法和模型 → 应用.

- 人类社会进入大数据时代. 随着应用的发展和深入, 统计问题涉及大量 (甚至海量) 的数据及复杂的模型, 必须使用计算机进行处理;
  海量数据是指数据量极大, 往往是 Terabyte (10^{12} bytes), Petabyte (10^{15} bytes) 甚至 Exabyte (10^{18} bytes), Zettabyte (10^{21} bytes), Yottabyte (10^{24} bytes) 级的数据集合.
- 随着计算机的大量运用, 人们提出了一些新的统计方法, 推动了统计学自身的发展.

- 人类社会进入大数据时代. 随着应用的发展和深入, 统计问题涉及大量 (甚至海量) 的数据及复杂的模型, 必须使用计算机进行处理;
  海量数据是指数据量极大, 往往是 Terabyte (10^{12} bytes), Petabyte (10^{15} bytes) 甚至 Exabyte (10^{18} bytes), Zettabyte (10^{21} bytes), Yottabyte (10^{24} bytes) 级的数据集合.
- 随着计算机的大量运用, 人们提出了一些新的统计方法, 推动了统计学自身的发展.

统计推断是数理统计学理论及应用中最重要部分. 统计推断就是根据从总体中抽出的样本 (局部信息) 去推断总体的性质.

主要内容包括

- 估计: 参数和非参数;
- 检验: 参数和非参数.

一类重要应用问题 — 统计回归, 涉及统计模型、参数估计及统计检验. 可区分为

- 线性模型及线性回归;
- 非线性模型及非线性回归.

目录

统计计算 (Statistical Computation)

- 目的: 统计计算作为数理统计、计算数学和计算机科学三者的结合, 研究如何应用计算机更有效 (从精度、速度、存储量等方面进行评价) 地完成各种统计数据的分析和处理.
- 重点: 统计问题的计算机处理算法.
- 过程: 实际问题 → 根据要求设计实验和建立模型 → 抽样 → 求解模型参数 → 统计推断 → 回到实际应用.

计算统计学 (Computational Statistics)

- 目的: 利用计算机的高性能特点设计、开发新的统计方法 (与传统统计方法不同), 以便有效处理各种不同应用问题中的大量数据, 从这些数据中获得所需信息.
- 重点: 基于计算机的统计原理和方法.
- 过程: 大量实际应用中的海量数据 → 根据要求设计统计模型 → 开发基于计算机的算法 → 获取所关心的信息或进行各种推断.

目录

理论部分

1. 与分布函数、分位数等相关的数值计算方法
2. 随机数的产生，随机模拟方法及其应用
3. 矩阵计算，扰动分析，线性回归
4. 最优化方法 (非线性优化理论和算法)，非线性回归
5. 计算统计方法 (几种重要的现代统计计算方法，包括 EM 算法，MCMC 方法，Bootstrap 方法等)

若干问题举例

一些常用分布函数、分位数的计算.

计算机仿真，随机服务系统，用随机方法求解确定性问题.

线性回归及其最小二乘解，解的扰动分析，估计的各种改进，设计矩阵非满列秩，近似共线性，回归变量的选择等.

随机动态系统状态估计的递推算法.

若干问题举例

一些常用分布函数、分位数的计算.

计算机仿真, 随机服务系统, 用随机方法求解确定性问题.

线性回归及其最小二乘解, 解的扰动分析, 估计的各种改进, 设计矩阵非满列秩, 近似共线性, 回归变量的选择等.

随机动态系统状态估计的递推算法.

若干问题举例

一些常用分布函数、分位数的计算.

计算机仿真，随机服务系统，用随机方法求解确定性问题.

线性回归及其最小二乘解，解的扰动分析，估计的各种改进，设计矩阵非满列秩，近似共线性，回归变量的选择等.

随机动态系统状态估计的递推算法.

若干问题举例

一些常用分布函数、分位数的计算.

计算机仿真, 随机服务系统, 用随机方法求解确定性问题.

线性回归及其最小二乘解, 解的扰动分析, 估计的各种改进, 设计矩阵非满列秩, 近似共线性, 回归变量的选择等.

随机动态系统状态估计的递推算法.

若干问题举例 (续)

非线性回归分析, 非线性最小二乘解, 一些最优化问题的算法设计.

观测数据部分丢失情形下的统计推断.

Bayes 分析中有关后验分布的极值及积分计算.

在缺乏严格理论指导情形下, 对复杂系统、大规模数据的分析与处理.

若干问题举例 (续)

非线性回归分析，非线性最小二乘解，一些最优化问题的算法设计.

观测数据部分丢失情形下的统计推断.

Bayes 分析中有关后验分布的极值及积分计算.

在缺乏严格理论指导情形下，对复杂系统、大规模数据的分析与处理.

若干问题举例 (续)

非线性回归分析, 非线性最小二乘解, 一些最优化问题的算法设计.

观测数据部分丢失情形下的统计推断.

Bayes 分析中有关后验分布的极值及积分计算.

在缺乏严格理论指导情形下, 对复杂系统、大规模数据的分析与处理.

若干问题举例 (续)

非线性回归分析, 非线性最小二乘解, 一些最优化问题的算法设计.

观测数据部分丢失情形下的统计推断.

Bayes 分析中有关后验分布的极值及积分计算.

在缺乏严格理论指导情形下, 对复杂系统、大规模数据的分析与处理.

软件部分

- 通用软件:
  - Maple
  - Mathematica
  - MATLAB
- 统计软件:
  - SAS
  - SPSS
  - S-PLUS
  - R (http://www.r-project.org)

软件部分

- 通用软件:
  - Maple
  - Mathematica
  - MATLAB
- 统计软件:
  - SAS
  - SPSS
  - S-PLUS
  - R (http://www.r-project.org)

实践部分

- 直接调用软件的相应函数或模块
- 一些重要算法的具体实现
- 综合运用所学理论和方法解决较大规模的实际应用问题

理论研究与计算机软件的关系

- 理论 → 算法 → 编码
- 实际系统具有不同的条件和要求，因此一般不能直接套用现成的算法、软件模块等，需要首先在理论方面进行相关的研究.

相关课程

1. 微积分、线性代数、概率论、数理统计
2. 程序设计
3. 其它专业课程

目录

评分方法

1. 平时 (40%): 作业, 考勤, 实验题, 小测验
2. 考试 (60%)

目录

教材及参考文献

- 高惠璇, 统计计算, 北京大学出版社, 1995.
- 程兴新, 曹敏, 统计计算方法, 北京大学出版社, 1989.
- 肖云茹, 概率统计计算方法, 南开大学出版社, 1994.

教材及参考文献 (续)

- James E. Gentle, Elements of Computational Statistics, Springer, 2002.
- Wendy L. Martinez, Angel R. Martinez, Computational Statistics Handbook with MATLAB, Chapman & Hall/CRC, 2002.
- James E. Gentle, Wolfgang Karl Härdle, Yuichi Mori, Eds., Handbook of Computational Statistics: Concepts and Methods, Springer, 2004.
- Geof H. Givens, Jennifer A. Hoeting, Computational Statistics, John Wiley & Sons, 2005. (有中译版)

问　题?

目录

数理统计学的内容

数理统计有许多分支，所要解决的问题十分广泛，但其基本内容主要分为两大部分：

1. 采集样本. 包括：抽样技术，试验设计.
2. 统计推断，就是由样本推断总体，它是数理统计学的核心部分. 按内容可以进一步划分为

- 统计推断
  - 估计
    - 参数估计
      - 点估计
      - 区间估计
    - 非参数估计
  - 检验
    - 参数假设检验
    - 非参数假设检验

目录

几个基本概念

1. 总体
   表征总体的随机变量 X，总体的分布函数 $F(x)$;
2. 样本 $X_1, \cdots, X_n$
   $X_1, \cdots, X_n$ 是对随机变量 X 的观测结果，其取值也具有随机性，因此应看成随机变量;
3. 简单随机样本 $X_1, \cdots, X_n$
   - 每次观测在完全相同的条件下进行，结果 $X_i (i = 1, \cdots, n)$ 都与 X 具有相同的分布;
   - n 次观测是独立进行的，所以，$X_1, \cdots, X_n$ 应是相互独立的.

统计量

假设 $X_1, \cdots, X_n$ 是来自总体的随机样本, $T(\cdot)$ 是完全已知的 n 元 Borel 可测函数. 如果 $T = T(X_1, \cdots, X_n)$ 不依赖于其它任何未知参数, 则称 T 为统计量.

需要注意的是, 统计量是一个随机变量.

统计量在数理统计中非常重要, 估计与检验问题就是围绕着统计量来进行的.

抽样分布

统计量的分布称为抽样分布，它在研究统计量的性质和评价一个统计推断的优良性等方面十分重要. R. A. Fisher 曾将抽样分布、参数估计和假设检验列为统计推断的三个中心内容.

抽样分布分为小样本问题和大样本问题. 前者对应于研究精确分布，后者对应于近似分布.

在实际应用中，总体的分布有很多；对于不同的总体以及统计推断问题，对样本进行加工的方法多种多样，因此统计量也是多种多样的. 一般情形下，统计量的抽样分布是非常复杂的.

对于正态总体以及一些简单的样本函数，可以获得统计量的精确分布. 常见的包括 χ^2 分布，t 分布和 F 分布.

顺序统计量

顺序统计量以及由顺序统计量得到的其它统计量 (如样本极差, 样本中位数等) 是极为重要的一类统计量, 其理论内容十分丰富且实际应用非常广泛.

顺序统计量

设 $X_1, \cdots, X_n$ 为一个简单随机样本, 该样本的第 i 个顺序统计量 (记为 $X_{(i)}$) 是如下的样本函数: 每当该样本得到一组观测值 $x_1, \cdots, x_n$ 时, 将它们从小到大排列为

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

其中第 i 个值 $x_{(i)}$ 就是 $X_{(i)}$ 的观测值.

称 $X_{(1)}, \cdots, X_{(n)}$ 为该样本的顺序统计量.

1. 称 $X_{(1)} = \min_{1\le i\le n} X_i$ 为最小顺序统计量;
2. 称 $X_{(n)} = \max_{1\le i\le n} X_i$ 为最大顺序统计量;
3. 称顺序统计量的函数 $R_n = X_{(n)} - X_{(1)}$ 为样本极差;
4. 称

$$X_{\mathrm{med}} = \begin{cases} X_{(m+1)}, & n = 2m+1, \\ \frac{1}{2}(X_{(m)} + X_{(m+1)}), & n = 2m, \end{cases}$$

为样本中位数.

顺序统计量的分布

设 $X_1, \cdots, X_n$ 为来自总体 $X \sim F(x)$ 的简单随机样本. 对任意 x, 以 $\nu_n(x) = \#\{X_i \le x\}$ 表示不大于 x 的个数, 因此服从参数为 $(n, F(x))$ 的二项分布:

$$P\{\nu_n(x) = k\} = C_n^k [F(x)]^k [1 - F(x)]^{n-k}.$$

显然,

$$\begin{aligned}
&\nu_n(x) = 0 \Longleftrightarrow X_{(1)} > x, \\
&\nu_n(x) = n \Longleftrightarrow X_{(n)} \le x, \\
&\nu_n(x) = k \Longleftrightarrow X_{(k)} \le x < X_{(k+1)}, \qquad 1 \le k < n, \\
&\nu_n(x) \ge k \Longleftrightarrow X_{(k)} \le x, \qquad 1 \le k \le n.
\end{aligned}$$

$X_{(k)}$ 的分布

对任意 $k=1,\cdots,n$, 有

$$\begin{aligned} F_{X_{(k)}}(x) &= P\{X_{(k)} \le x\} = P\{\nu_n(x) \ge k\} \\ &= \sum_{m=k}^{n} C_n^m \big(F(x)\big)^m \big(1-F(x)\big)^{n-m} \\ &= nC_{n-1}^{k-1} \int_0^{F(x)} t^{k-1}(1-t)^{n-k}\,dt \\ &= F_{\text{Beta}}(F(x); k, n-k+1). \end{aligned}$$

因此, 顺序统计量 $X_{(k)}(k=1,\cdots,n)$ 的分布函数可以通过参数为 $(k, n-k+1)$ 的 Beta 分布的分布函数来表示.

这是因为: 对任意 $p \in (0,1)$, 有

$$\sum_{m=k}^{n} C_n^m p^m (1-p)^{n-m}$$

$$= \frac{n!}{(k-1)!(n-k)!} \int_0^p x^{k-1}(1-x)^{n-k} dx.$$

事实上, 记

$$S(p) = \sum_{m=k}^{n} C_n^m p^m q^{n-m}, \quad q = 1-p,$$

则有

$X_{(k)}$ 的分布

$$
\begin{aligned}
S'(p) &= \sum_{m=k}^{n-1} C_n^m \left(m p^{m-1} q^{n-m} - (n-m) p^m q^{n-m-1} \right) + n p^{n-1} \\
&= n \sum_{m=k}^{n-1} \frac{(n-1)!}{(m-1)!(n-m)!} p^{m-1} q^{n-m} \\
&\quad - n \sum_{m=k}^{n-1} \frac{(n-1)!}{m!(n-m-1)!} p^m q^{n-m-1} + n p^{n-1} \\
&= n \sum_{m=k}^{n-1} C_{n-1}^{m-1} p^{m-1} q^{n-m} - n \sum_{r=k+1}^{n} C_{n-1}^{r-1} p^{r-1} q^{n-r} + n p^{n-1} \\
&\quad (\text{令 } r = m+1) \\
&= [(m=k) + (m = k+1 \sim n-1)] \\
&\quad - [(r = k+1 \sim n-1) + (r = n)] + [(r = n)] \\
&= \frac{n!}{(k-1)!(n-k)!} p^{k-1} q^{n-k},
\end{aligned}
$$

对上式关于 p 积分即可得 $S(p) = \int_0^p S'(x) dx$.

$X_{(k)}$ 的分布

若总体有密度 $f(x)$, 则 $X_{(k)}$ 也有密度

$$f_{X_{(k)}}(x) = nC_{n-1}^{k-1}F(x)^{k-1}\left(1 - F(x)\right)^{n-k}f(x).$$

$X_{(1)}$ 和 $X_{(n)}$ 的分布

当 $k=1$ 时，$X_{(1)}$ 的分布函数和密度函数分别为

$$F_{X_{(1)}}(x) = n\int_0^{F(x)} (1-t)^{n-1}dt = -(1-t)^n\Big|_0^{F(x)}$$
$$= 1-\big(1-F(x)\big)^n,$$
$$f_{X_{(1)}}(x) = n\big(1-F(x)\big)^{n-1}f(x).$$

当 $k=n$ 时，$X_{(n)}$ 的分布函数和密度函数分别为

$$F_{X_{(n)}} = F(x)^n, \quad f_{X_{(n)}}(x) = nF(x)^{n-1}f(x).$$

目录

在数理统计中, 总体分布未知 (我们的任务就是要对总体进行推断), 这包括两个层次:

1. 分布函数的类型已知, 但其中的某些参数未知, 这对应于总体参数或其函数的估计 — 参数估计;
2. 分布函数的类型未知, 这对应于分布函数的估计 — 非参数估计.

最大似然估计 (Maximum Likelihood Estimation)

这是统计学中非常重要，应用最广泛的方法之一.

直观思想：参数的最大似然估计确定的总体分布使抽样结果最有可能来自于这个分布. 这里的“最有可能”就是要求概率最大.

假设 $X_1, \cdots, X_n$ 是来自总体 $X \sim f(x;\theta)$ 的简单随机样本，其概率密度为

$$f(x;\theta) = f(x_1, \cdots, x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

按照前述最大似然法的基本思想，如果在对总体 X 的独立观测中得到样本值 $x = (x_1, \cdots, x_n)$，那就说明 $f(x;\theta)$ 在点 x 处可能取最大值，因此有理由选择满足

$$f(x;\hat{\theta}) = \max_{\theta \in \Theta} f(x;\theta)$$

的 $\hat{\theta} = \hat{\theta}(x)$ 作为未知参数 θ 之真值的估计值，从而 $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ 作为未知参数 θ 的估计.

最大似然估计的特点

1. 最大似然估计用到了总体的分布性质，需要知道分布函数．最大似然估计可以使用分布的所有信息，因而具有许多优良性．
2. 但同时也为其实际应用带来了不便．比如实际中可能只知道总体分布的某些数字特征，最大似然原理就无能为力了．

这是统计学家 K. Pearson 于 1894 年提出来的.

直观想法: 用样本矩来估计总体理论矩, 用样本矩的函数来估计总体理论矩的函数.

设表征总体随机变量 X 的分布函数为 $F(x)$.
定义 X 的 k 阶原点矩为

$$V_k = \mathbb{E}X^k = \int x^k dF(x), \quad k = 1, 2, \cdots$$

以及 X 的 k 阶中心矩为

$$W_k = \mathbb{E}(X - V_1)^k = \int (x - V_1)^k dF(x), \quad k = 1, 2, \cdots$$

经验分布函数

对 X 进行 n 次独立观测, 同前, 用 $\nu_n(x)$ 表示事件 $\{X \le x\}$ 出现的次数 (即不大于 x 的观测值的个数).

样本 $X_1, \cdots, X_n$ 的经验分布函数定义为

$$F_n(x) = \frac{1}{n}\nu_n(x), \quad -\infty < x < \infty.$$

经验分布函数的性质

1. $F_n(x)$ 是离散型分布函数;
2. 对任意给定的 $x \in (-\infty, \infty)$, $\nu_n(x)$ 和 $F_n(x)$ 作为随机样本的函数均是随机变量. $\nu_n(x)$ 服从参数为 $(n, F(x))$ 的二项分布, 因此

$$\mathbb{E}[\nu_n(x)] = nF(x), \quad \mathbb{E}[F_n(x)] = F(x)$$

3. (Glivenko 定理) 对任意给定的 $x \in (-\infty, \infty)$, $F_n(x)$ 是样本的函数, 因而是一个统计量, 且以概率 1 收敛到 $F(x)$.

样本矩

考虑来自总体 X 的简单随机样本, 定义样本 k 阶原点矩为

$$\widetilde{V}_k = \int x^k dF_n(x), \quad k = 1, 2, \cdots$$

样本 k 阶中心矩为

$$\widetilde{W}_k = \int (x - \widetilde{V}_1)^k dF_n(x), \quad k = 1, 2, \cdots$$

可以证明

$$\widetilde{V}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \widetilde{W}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widetilde{V}_1)^k.$$

矩方法的特点

1. 根据 Glivenko 定理，当样本容量趋于无穷时，经验分布函数以概率 1 在整个实轴上一致地收敛到总体分布．因此，用样本矩去估计总体矩是有道理的;
2. 矩估计基于经验分布函数，而经验分布函数逼近真实分布的前提条件是样本容量较大，因而理论上讲，矩方法是以大样本为应用对象的;
3. 矩方法未用总体分布全部信息，本质上是一种非参数方法．在总体分布已知情形下，矩估计不一定是一个好的估计;
4. 矩方法简单，便于应用；但估计量的优良性往往比较差．

估计的优良性标准

1. 无偏性:
$$\mathbb{E}\hat{\theta} = \theta, \quad \forall \theta \in \Theta;$$
2. 最小均方误差:
$$\hat{\theta}_{\text{opt}} = \arg\min_{\theta \in \Theta} \mathsf{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2.$$

均方误差的一个表达式

因为:

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathbb{E}(\hat{\theta}-\theta)^2 \\
&= \mathbb{E}(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta)^2 \\
&= \mathbb{E}(\hat{\theta}-\mathbb{E}\hat{\theta})^2 + 2\mathbb{E}(\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta) + (\mathbb{E}\hat{\theta}-\theta)^2 \\
&= \mathbb{E}(\hat{\theta}-\mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta}-\theta)^2 \\
&= \mathrm{Var}\hat{\theta} + (\mathbb{E}\hat{\theta}-\theta)^2.
\end{aligned}
$$

这说明均方误差是由两个量迭加而成的: (1) 估计量的方差; (2) 估计量的偏度的平方.

若 $\hat{\theta}$ 是 θ 的无偏估计, 则

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}\hat{\theta}.$$

因此, 对于无偏估计, 均方误差最小化准则等价于方差最小化准则.

无偏估计的方差下界

自然的问题是:

1. 当样本量一定时, 估计量的方差是否能任意小?
2. 如果不能任意小, 一个无偏估计的方差能小到什么程度?

研究上述问题的意义在于: 了解估计量的性质, 这在许多应用场合是需要的. 如果知道了估计量的方差下界, 就可以根据这个下界来衡量无偏估计的优劣: 若它达到这个下界, 则一定是最小方差的. 当然一个估计量也有可能达不到方差下界.

设 X 的密度函数为 $f(x;\theta)$, $X_1,\cdots,X_n$ 是 X 的样本, $\psi(X_1,\cdots,X_n)$ 是 $g(\theta)$ 的一个无偏估计, 且满足一定的正则性条件. 如果

$$I(\theta)=\int\left(\frac{\mathrm{d}\ln f(x;\theta)}{\mathrm{d}\theta}\right)^2 f(x;\theta)\mathrm{d}x>0,$$

那么

$$\mathrm{Var}_\theta(\psi(X_1,\cdots,X_n))\geq\frac{(g'(\theta))^2}{nI(\theta)}.$$

区间估计的提出

在参数点估计中, 我们的目标是: 对总体分布的末知参数 θ, 用样本构造一个统计量 $\hat{\theta}$ 去估计 θ. 这就好比是要根据观测来给出大海中一艘沉船确切的经纬度一样.

一个重要的问题是: 这个估计与真实位置相差多大?

在点估计中, 我们曾用均方误差来表示估计的精度, 对无偏估计而言, 其方差就是它的精度.

区间估计的提出 (续)

另一种表示精度的方法是根据样本给出一个区间 $[\hat{\theta}_1, \hat{\theta}_2]$, 其中 $\hat{\theta}_1 < \hat{\theta}_2$ 为两个统计量, 用区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 来估计 θ 的真值的范围, 这个区间也是对参数 θ 的一种估计, 称为区间估计.

对应于上面的搜索例子, 就是给出沉船在海中的一个地理范围, 并且这个范围以一个可信度包含真实位置. 显然, 这种估计在许多实际应用中更具有现实意义.

目录

一些基本概念

1. 假设: 对描述问题的集合进行划分 (在参数情形为参数空间 Θ 的一种划分);
2. 零 (原) 假设和对立 (或备择) 假设: H_0, H_1;
3. 简单假设和复杂假设;
4. 检验: 给出一种规则, 根据样本确定接受或拒绝零假设.

具体的数学提法

根据一定的优良性标准, 将样本空间 S 划分成接受域和否定 (拒绝) 域 (要求为 Borel 可测集):

1. 对于得到的样本, 如果属于否定域 W, 则检验结果为否定零假设 H_0;
2. 如果不属于否定域 W, 即属于接受域, 则检验结果为不否定零假设 H_0 (在实际应用时接受零假设 H_0).

关键是给出否定域 W.

确定否定域的统计学思想:

假设 H_0 为真, 如果样本落在某个区域 W 中的概率 p 不超过事先给定一个很小的数 α, 表明在这次试验中, 小概率事件发生了 (因为抽样得到了现在的样本), 因而根据小概率原则 (在一次观测中, 小概率事件是几乎不可能发生的), 需否定假设 H_0. 这个区域 W 就是假设 H_0 的否定域.

根据观察到样本值进行检验后有以下四种可能性:

1. H_0 真, 样本属于否定域;
2. H_0 真, 样本不属于否定域;
3. H_0 假, 样本属于否定域;
4. H_0 假, 样本不属于否定域.

其中, (2) 和 (3) 表示检验结果正确; (1) 表示虚警, 承伪; (4) 表示漏检, 拒真.

优化准则

上述四种情形下对应于四个概率:

1. P_f (第一类错误)
2. P_d
3. P_d
4. P_m (第二类错误)

自然的想法是: 选择否定域 W, 使得犯两类错误的概率均很小. 但这无法实现, 需要采取某种妥协方案.

Jerzy Neyman 和 Karl Pearson 合作建立的假设检验理论是统计判决理论中重要的理论成果. Neyman-Pearson 准则的基本思想:

首先使犯第一类错误的概率限制在某个范围内, 然后寻找使犯第二类错误的概率尽可能小的检验.

简单地说就是: 控制一个, 极小化另一个.

考虑检验问题:

$$H_0 : \theta = \theta_1 \longleftrightarrow H_1 : \theta = \theta_2.$$

给定 $\alpha \in (0, 1)$, 设

$$W_0 = \{x : L(x, \theta_2) > \lambda_0 L(x, \theta_1)\}$$

满足 $P\{X \in W_0 | H_0\} = \alpha$, 则 W_0 是所有检验水平不超过 α 的否定域中犯第二类错误概率最小的一个. 这也称为似然比检验.

W_0 通常可以转换为一个检验统计量确定的区域, 求出否定域的关键是获得该统计量的分布(精确分布或渐近分布).

对于正态总体（包括单个和两个总体情形）的参数 μ 或 σ, 利用似然比检验可以推导出一些常见假设检验问题的检验统计量及其分布.

对于复杂的假设检验问题, 可以采用广义似然比方法.

　　实际中，待检验总体的分布形式未知，或者本身就是检验总体服从什么样的分布，这就需要利用非参数假设检验的思想和方法.

　　非参数检验方法很多，本课程中主要用到以下两种在实际应用常见的方法:

1. 基于中心极限定理的近似正态检验统计量
2. 拟合优度检验中的 χ^2 检验统计量

设 $X_1, \cdots, X_n$ 相互独立同分布, 且

$$\mathbb{E}X_i = \mu, \ \mathrm{Var}(X_i) = \sigma^2.$$

令 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, 则

$$U = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

以 $N(0,1)$ 为极限分布.

χ^2 检验

考虑总体随机变量 X 分布形式的假设 H_0 (例如, H_0: X 服从 Poisson 分布或 H_0: X 服从正态分布). 按以下步骤得到的检验法称为 χ^2 检验:

1. 把 X 的一切可能取值的集合 S 进行划分:

$$S = \bigcup_{i=1}^{r} A_i, \quad A_i \cap A_j = \emptyset \ (i \neq j);$$

2. 统计出 X 的观测值落入 $A_i (i = 1, \cdots, r)$ 的频数 n_i: 对于来自 X 的随机样本 $X_1, \cdots, X_n$, 以 n_i 表示落入 A_i 的观测值的个数, 即 n_i 表示事件 $\{X \in A_i\}$ 在 n 次独立观测中出现的次数.

χ^2 检验 (续)

3. 在假设 H_0 成立的前提下，分别求得观测值落入 A_i 的概率

$$p_i = P\{X \in A_i | H_0\}, \quad i = 1, \cdots, r.$$

不妨假设 $p_i > 0$ (否则可以适应改变第 1 步中的划分).

4. 考虑统计量

$$\chi_n^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}.$$

其直观意义为: χ_n^2 表示实际观测结果与理论期望结果的相对差异的总和. 当其值大于某个临界值时, 应否定所作出的假设.

χ^2 检验 (续)

5. 可以证明 (Pearson 定理): 当 $n \to \infty$ 时, 检验统计量 χ_n^2 的极限分布是 χ^2 分布, 自由度为 $r-1$. 因此, 当 n 充分大时,

$$W_0 = \{\chi_n^2 \geq C_\alpha\}$$

就是假设 H_0 的 (近似) α 水平否定域, 其中, C_α 是 $\chi^2(r-1)$ 的 α 水平上侧分位数.
如果在计算概率 p_i 时有未知参数, 可以用 Pearson 定理的推广形式 (由 Fisher 给出).

非参数假设检验问题的特点

非参数假设检验方法有如下特点:

1. 不依赖于总体分布的具体数学形式, 因而适应范围广, 但针对性差;
2. 所使用的检验统计量的精确分布难于求出, 一般只能求出其极限分布, 因而要求大样本容量.

问　题?