非参数统计
（在职研究生）

中国人民大学统计学院 2019年9月
主讲人: 王 星
办公电话:010-82500167
办公室: 明德主楼 1019
沟通邮箱:
wangxingwisdom@126.com

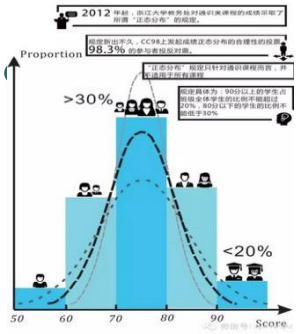
1

研讨的问题

- 为什么学习非参数统计？
- 非参数统计主要包括的内容有哪些？
- 课程特点
- 基本概念回顾

2

案例导引：ZJ大学正态成绩事件



2012年一则教育新闻报导，ZJ大学正在试行一种“正态型成绩单”，这份正态成绩单上很多成绩都将做正态化处理，这样，学生拿着成绩单无论是面试还是求学，都更容易让对方面试者知道他某门课程在学习群体中的实际水平。比如，90分表示在这项科目上低于他的学生不会少于80%（Top20），60分表示在这项科目上位于20%较低的水平。

然而在接到外界媒体询问时，大学教务处却表示近期并不准备将这项政策推广到所有课程，仅限于本科通识课，但我们会研发类似的标准成绩，但这项研发的焦点不在于扩大在校生不及格比例。

3

ZJ大学的正态成绩单分析(参考)

思考题1：正态成绩单制是一个好的构想还是一个差的构想？

正方

增强成绩可比性，管理规范

“如果不正态分布的话，压力和权力将全部转嫁到教师身上。”

“正态分布为学生提供了竞争的意识，能够有效促进学生的学习积极性。

反方

强制规定分数分布不科学的质疑，他表示，“但是如果说21% 不行，一定要降到20% 也不太好。也就是说不喜欢‘被迫’的正态分布。

对分数进行强行规定是一种粗暴的行为。“就是毫无人性的教条主义。

GPA 有了实质性的降低

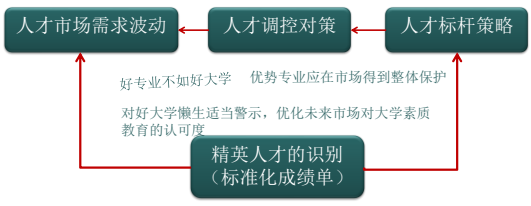
为正态分布而正态分布的做法不可取

思考题2：.成绩正态制要解决的是一个什么样的问题？这个待解决的问题一开始是怎样发生的？

4

大数据时代

唯985论，唯热门专业论，唯？？论!成绩乱象比较普遍，人才识别缺乏标准



5

这个故事接下去发展成为：

ZJ大学的校内论坛“CC98”上，一个发表于2012 年2 月21 日，投票结果是一边倒的：在177名投票者中，认为这一规定

- **“纯属扯淡”**（170 人）**“一点都不科学”（伪科学）**（4 人），达到了174 人，占总投票者的**98.3%**，
- 而选择“比较科学”与“非常科学”的分别只有1 人与2 人。

思考题4：为什么会这样？
1.是统计学的方法不够严谨吗？
2.是行业应用不了解统计学吗？
3.是成绩单导致了不及格人数的增长吗？人们到底否定的是什么？

人们否定的是纯真的师生文化被空洞化的二分精英理论所取代，到那时，绩效精英与普通民众之间的关系会真正断裂，这样的精英能否代表民众的利益?能否反映民众的要求?是否受民众的影响?一所大学的悲剧正是从人才培养战略选择二分精英论开始的?
如何识别人才：成绩单的一个作用是告诉你应该如何预备自己的未来，不是用来选拔人才的唯一指标。比起过去了的成绩而言，人才至少应该是在任何困难面前有足够思考力的人，表现出积极活跃的精英特质。

70分的学生把30分留给面试官，90分的学生要告诫自己，你不会的没考到，太不幸了，请从零开始

6

思考题3

- 分布适合作为一种标准吗？什么时候合适？什么时候不合适？合适的情况下，分布的作用是什么？
  - 分布是用来刻画不确定性的，不确定的由来：一方面是由测量误差所引起的，惟极贫无依，则械系不稍宽，为标准以警其余。——清·方苞《狱中杂记》，而另一反方面是由可见数据集的有限性所引起的，标准不能用来刻画后者的不确定性。
  - 合适的例子：标的物固定，技术标准是否达标；
  - 不合适的例子：标的物不固定。
- 在评标物分布不固定的时候，研究分布的意义是什么？
  - 分布的存在性，结构性，密度性，差异性
  - 建立分布边界，有效利用分布特点进行差异化分析

7

从统计应用来看这次失败的案例

统计失败原因之一：应用场景选择错误

这个项目的失败是由于问题的复杂性，在成绩单功能的认识上，它主要的功能还是用来反映学生学习状态，知识掌握程度的工具，若硬要将成绩单开发成一种在人才市场上精英人才快速甄别的专业占领市场策略，则还有大量的灰色地带有待开发，后来者居上的成功人士都经历过先有差成绩单而后奋起搏发的励志经历，教育不应陷入“成绩绑架”论。在危机面前策略简单化的驱动下，成绩对个体的正面激励作用被忽略了。

8

数据分析如此复杂，有哪些基本的要领

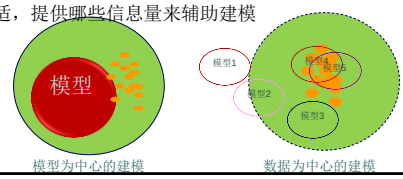
---数据分析有过程

---一步一步向上建

9

统计的能与不能之争

- 对统计需求的变化从套用、移花接木式的统计应用向对统计设计的需求
- Good of fit 检验：用于检查目前的数据是否和给定的一种正常的情况是吻合的， 如果差距较大，表示那种正常的假设是与数据目前提供的信息是不相匹配的。
- Lack of fit 检验：以数据为中心，检验当前的模型是否合适，如果不合适，提供哪些信息量来辅助建模



10

三种参数的认识

- （为了算法正常运行）环境技术参数：脚手架是为了保证分析过程顺利进行而搭设的工作平台参数，例如为算法停止而设置的参数，这些参数是程序依赖的，不是数据依赖的；
- （信号）统计参数：
  - 代表数据中稳定的信息部分，这些信息可以告诉我们应该选择怎样的模型来提取数据的模型尝试中的lack of fit检验统计量和结果，这些检验结果帮助我们尝试不同的模型空间信息；
  - ；
- （防止算法崩溃）计算参数：
  - 在提取模型的时候，需要在模型空间上进行参数的估计，然而模型的系数在全局优化的目标下会变形，在模型空间中为防止选择错误的模型而辅助性的设置的参数，也是经常调参的参数所指。

11

非参数统计的作业要求

- 陈述问题State the problem
- 描述数据Describe the data
- 翻阅并思考怎样的统计方法适合你手中的数据Review what statistical methods are available to analyze your data
- 将这些方法的优点和缺点列出来，特别是将非参数统计的方法和参数方法做一些比较List their advantages and disadvantages, in particular compare nonparametric to parameteric methods
- 用非参数方法提出一种解决方案Propose a solution using nonparametric methods
- 列出你将要完成的分析任务（收集数据，编程，模拟数据，估计和检验）List all the tasks that you plan to do: collecting data, programming, simulating data, estimating, testing, etc.

12

老木匠和学徒的对话

- 老木匠在一堆木头中选一段上好的木头做桌面，小学徒挑出一段又大又直的木头，自认为是好木料，老木匠拿来敲一敲，“这明明是块空心木头，怎么会是上好的木料”？ 老木匠说“从声音中可以分辨出来，如果声音很低很小，就是实心的，如果声音很高很大，就是空心的”。
- 又一次，师傅要找一根有用的木头，最后找了一根密的不成样子的木头，找没？里世

散文《牛轭》

牛轭，现在的人恐怕知道的不多了，但几乎过牛的人家，肯定保留着牛轭的，即使牛不在了，但牛轭仍挂在墙上。

牛轭，它是与牛、犁耙配套使用的，其状如“人”字形，有半米见方长，两肢，耕田时使农人就把它安置在牛的脖颈上。最简随最原始的牛轭是自然树木制作成的。找个“人”字形的树叉，削去树皮或者斜出的枝桠，再削去皮，稍稍加工一下即可，但不太结实、美观。有经验的农人，喜欢找木匠制作，挖槽头，凿洞眼，然后结合起来，契合得非常牢固，有板有眼。

牛轭是牛儿犁地时最重要的工具了，有了它，牛就有使劲的立足点了。但它也是农具里最挑剔的物件之一。耕田，牛儿拿着膀颈上的肌肉卷起疙瘩，在农人的鞭策下，死命地向前挣扎行走，步履沉重。为了干活后那可怜的一堪青草、稻草或者黄豆，即过架后有时累得口吐白沫，只有喘气的份，再好的牛料也吃不下呢。最让人哀怜的是，牛儿总会用沉默的大眼睛注视着你，什么话也不说。当然，无论吃好吃坏，干活时牛总是卖力气的，这是牛的本能呢。

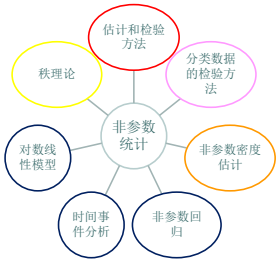
材在等放

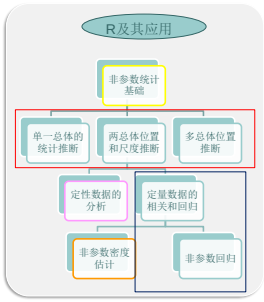
大外，边要往长度方面放大一定的尺寸，待到拼装完毕、哭 刀锯定后，再把多余的部分锯掉。

LEO Breiman也说过类似的话：统计学家就该定位于一个好木匠识人用人如此，识数用数，做“有心的正直木匠”，大体也如此。

13

知识模块和课程体系





14

课本和参考书:

1. 王星,褚挺进, 非参数统计[M], 清华大学出版社, 2015,09.
   勘误表之后会列在网站上
2. John Kloke,Joseph W.McKean,Nonparametric Statistical Methods Using R,CRC Press
3. Jeffrey D.Hart,Nonparametric Smoothing and lack-of-Fit Tests,Springer.
4. Larry Wasserman, All of non-parametric Statistics[M], Chap 2,Chap 3,Chap 5*, Chap 6*.
5. 吴喜之,2006, 非参数统计[M], 中国统计出版社;
6. John A. Rice, Mathematical Statistics and Data Analysis[M], chap 9, 10,11,13.

15

第一章 绪 论

16

主要内容:

什么是非参数统计？
非参数统计的主要内容是什么？

1. 什么是统计推断，统计推断中的基本概念？
2. 非参数统计方法简介
3. 参数统计过程与非参数统计的比较
4. 非参数统计的历史
5. 必要的准备知识

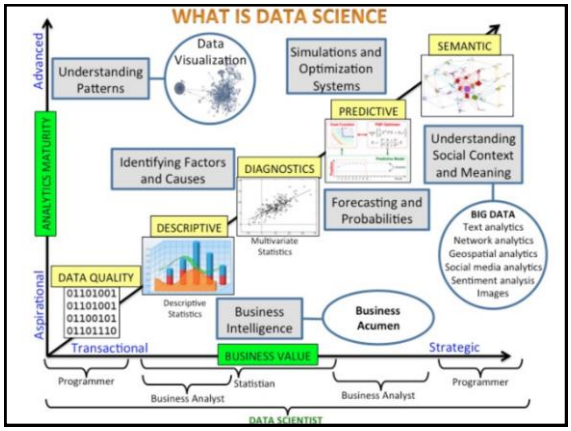
17

Role of Statistics

Role of Statistics and statisticians have always played a major role, but this has changed. It used to be almost entirely in descriptive as opposed to theoretical statistics, and observational rather than inferential. Now the processes can best be described as descriptive statistics plus modeling. However, ”It is descriptive statistics and scientific method which have to become fully one”

Ehrenberg.A.S.C.(1968)J.R.Statist.Sco.A.131,201

18



19

2. 非参数统计方法简介

20

参数方法

- 定义：样本被视为从分布族的某个参数族抽取出来的总体的代表，未知的是总体分布中具体的参数，推断问题就转化为对分布族的若干个未知参数的估计问题，用样本对这些参数做出估计或者进行某种形式的假设检验，这类推断方法称为参数方法。
- 比如：

（1）研究保险公司的索赔请求数时，可能假定索赔请求数来自泊松分布$P(a)$;

（2）研究化肥对农作物产量的影响效果时，平均意义之下，每测量单元（可能是）产量服从正态分布$N(a,b)$.

21

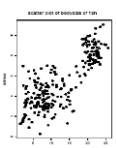
一个典型的参数检验过程

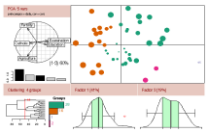
1. 总体参数
   Example: Population Mean
2. 假定数据的形态为
   Whole Numbers or Fractions
   Example: Height in Inches (72, 60.5, 54.7)
3. 有很强的假定
   Example: 正态分布
4. 例子: Z Test, t Test, χ^2 Test

22

(1) 假设检验回顾

- 问题：
  (*a*) 新引进的生产过程是否优于旧过程？
  (*b*) 几种不同的肥料哪一种更有效？
  (*c*) 大学生的就业率与城市失业率之间是否存在关系？





23

内容

（a）假设的真正涵义和作用
（b）如何选择零假设和备择假设
（c）检验的p-值和显著性水平的作用
（d）两类错误

24

统计检验的例子

- 公司在收到一批货物的时候，质检人员需要判断该批货物的属性是否与合同中规定的一致。
- 某新药的研究开发过程中，研究人员需要判断新药的药效是否比原有的药物更加有效。
- 城市中拥有汽车的人口比例是否超过30%？
- 使用理财产品A的每月新增用户数与使用理财产品B的每月新增用户数有差异吗？

25

内容

（a）假设的真正涵义和作用
（b）如何选择零假设和备择假设
（c）检验的*p*-值和显著性水平的作用
（d）两类错误
（e）置信区间和假设检验之间的关系

26

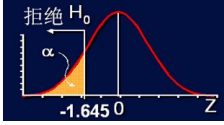
均值的单尾Z检验（实例）

【例】某批发商欲从厂家购进一批灯泡，根据合同规定，灯泡的使用寿命平均不能低于1000小时。已知灯泡使用寿命服从正态分布，标准差为20小时。在总体中随机抽取了100个灯泡，得知样本均值为996小时。批发商是否应该购买这批灯泡？（α=0.05）

27

均值的单尾Z检验（计算结果）

H0: μ ≥ 1000
H1: μ < 1000
α = 0.05
n = 100
临界值(s):



检验统计量:

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{996-1000}{20/\sqrt{100}}=-2$$

决策:
在 α = 0.05的水平上能拒绝

结论:
有证据表明这批灯泡的使用寿命低于1000小时

28

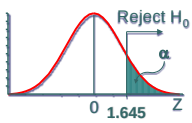
假设检验的过程和逻辑

- **寻找数据内部差异中共同的特征，甄别数据之间的本质差异是统计推断的核心内容，假设检验就是帮助我们确定显著性差异界限的最好工具。**
- **计算机软件仅仅给出*p*-值，它表示我们要对比的两个假设之间差异存在的显著性。拒绝零假设时犯错误的概率.在这个意义上，*p*-值又称为观测的显著性水平（observed significant level）。在统计软件输出*p*-值的位置，有的用“*p*-value”，有的用significant的缩写“Sig”。**

29

该问题如果两个假设对换??

- H0: μ < 1000
- Ha: μ > 1000
- α = .05
- *n* = 100
- 临界值(s):



检验统计量:

$$Z=\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{996-1000}{20/\sqrt{100}}=-2$$

决策:
在α = .05的水平上不能拒绝

结论:没有确凿证据表示零假错，也就是说可能μ< 1000。

30

单边检验和双边检验的p值??

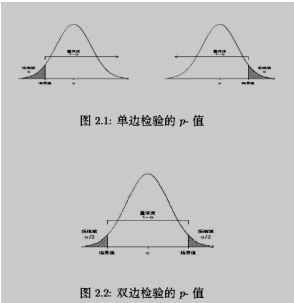


图 2.1: 单边检验的 p- 值

图 2.2: 双边检验的 p- 值

31

莫德尔和他的学术之梦

2010年，弗吉尼亚大学心理学博士生，马特·莫德尔（Matt Motyl），他的一项研究主题是政治派别对灰色色彩的敏感性，中的数据似乎表明，在涉及近2000人的研究中观察发现，与左翼或右翼人士相比，政治中立派能更准确地辨别不同色度的灰色，政治极端主义者看到的世界是确实是非黑即白的。

“实验的假设很有趣，而且数据也能够有力支持实验假设。”用来衡量统计显著性的常用指标是P值。该实验中的P值为0.01，莫德尔十分有把握能把自己的论文发表在高影响因子的刊物上。

由于担心实验结果陷入再现性争论，莫德尔和他的导师布莱恩•诺塞克（Brian Nosek）决定重复实验。添加了新的数据之后，P值变成了0.59，这个数字远未达到学界一般能接受的显著性水平0.05。莫德尔观察到的心理学效应没有了，他年少成名的梦也被打碎了。

http://www.guokr.com/article/438043/

32

p-值的作用----来自ASA

- P-值可以指示数据与一个给定模型的不相容程度
- P-值不能衡量假设为真的概率，也不能衡量数据仅由随机因素造成的概率
- 科学结论、商业决策或政策制定，不应只取决于p-值是否达到一个给定的标准
- 研究者需对研究进行完整的报告、保证透明度，才能作出合理的结论
- P-值或显著性并不能衡量效应的大小和结果的重要性
- P-值本身不能衡量模型或假设的可信度。

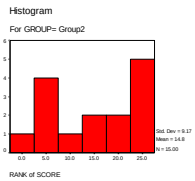
33

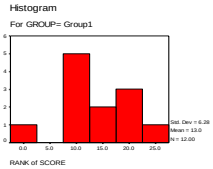
一个例子：

对两组学生进行语法测试，如何比较两组学生的成绩是否存在差异？

34

原始数据	秩
25	9.5
30	14.0
29	12.0
34	21.0
24	7.5
25	9.5
13	2.0
32	17.5
24	7.5
30	14.0
32	17.5
37	24.0
44	26.0
33	19.5
22	5.5
8	1.0
47	27.0
31	16.0
40	25.0
30	14.0
33	19.5
35	22.5
18	3.0
21	4.0
35	22.5
28	11.0
22	5.5





35

非参数检验过程

- **1.不涉及总体的分布**
  - **Example: Probability Distributions, Independence**
- **2. 数据的形态各异**
  - **定量数据**
  - **定序数据**
    - **Example: Good-Better-Best**
  - **名义数据**
    - **Example: Male-Female**
- **3.例子: Wilcoxon Rank Sum Test/Run Test**

F, F, F, F, F, F, F, M, M, M, M, M, M, M

F, M, F, M, F, M, F, M, F, M, F, M, F, M, F

36

3. 参数与非参数统计比较

37

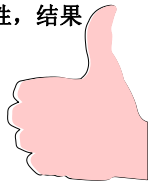
非参数统计的基本内容

- 估计
  - 分布函数估计
  - 函数估计
  - 密度估计
  - 统计关系估计:
    - 定性数据的关联分析
    - 非参数回归
- 假设检验
  - 单一总体
  - 两总体
  - 多总体

38

非参数检验的优点

- 对总体假定较少，有广泛的适用性，结果稳定性较好。
  - 1. 假定较少
  - 2. 不需要对总体参数的假定
  - 3. 与参数结果接近
- 针对几乎所有类型的数据形态。
- 强调计算
  - 在计算机盛行之前就已经发展起来;
  - 估计涉及大量数据的计算。



39

非参数检验的弱点

- 1. 可能会浪费一些信息
  - 特别当数据可以使用参数模型的时候。
  - Example: Converting Data From Ratio to Ordinal Scale
- 2. 大样本手算相当麻烦
- 3. 一些表不易得到

40

Nonparametric vs Parametric methods

- Nonparametric models
  - More flexible-no parametric model is needed
  - But require storing the entire dataset
  - And the computation is performed with all data examples
- Parametric models:
  - Once fitted, only parameters need to be stored.
  - They are much more efficient in terms of computation
  - But the model needs to be picked in advance.

41

课程大纲

- 第一讲 绪论和基本要求
- 第二讲 非参数统计基本概念,分布函数估计
- 第三讲 秩统计量及分布,连续性修正
- 第四讲 单一样本的推断问题(1) 中位数检验
- 第五讲 单一样本的推断问题(2) 趋势和随机游程检验
- 第六讲 单一样本的推断问题(3)： 置信区间计算
- 第七讲 分布的一致性检验
- 第八讲 理论部分:U统计量和渐进相对效率
- 第九讲 两样本位置检验
- 第十讲 多总体推断(一)
- 第十一讲 多总体推断(二)
- 第十二讲 多总体推断(三)
- 第十三讲 分类数据关系分析
- 第十四讲 秩相关分析
- 第十五讲 非参数密度估计
- 第十六讲 局部多项式回归**

42

4. 非参数统计的历史

43

非参数统计的历史

年代	代表性人物	代表性检验
1900	Karl Pearson	Good of fit test
1904	Spearman	Spearman等级相关系数
1937	Friedman	Friedman Q检验法
1938	Kendall	Tau相关系数
1939	Smirnov	Smirnov(K_S)检验
1939	Fisher Erwin	Fisher精确性检验
1945	Wilcoxon	Wilcoxon秩检验法
1947	Mann,Whitney	U值检验法
1949	Jackknife	Quenouille
1950	Cochran	Q检验法
1951	Brown Mood	BM中位数检验法
1951	Durbin	均衡的不完全区组设计检验法
1952	Kruskal, Wallis	KW检验
1954	Kendall	协和系数法
1958	Bross	非参数Ridit检验
1959	Mantel-Haenszel	Q_{MH}
1960	Cohen	Kappa一致性检验
1963	Hodges-Lehmann	HL估计量
1979	Efron	bootstrap
1984	Noether	渐进相对效率的Noether条件
1990		Smoothing

44

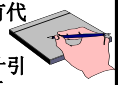
非参数统计的历史

- **非参数统计思想的形成主要归功于20世纪40年代～50年代化学家F.Wilcoxon等人的工作。Wilcoxon于1945年提出两样本秩和检验，1947年Mann和Whitney二人将结果推广到两组样本量不等的一般情况；**
- **Pitman于1948年回答了非参数统计方法相对于参数方法来说的相对效率方面的问题；**

45

非参数统计的历史（续）

- **60年代中后期，Cox和Ferguson最早将非参数方法应用于生存分析。**
- **70年代到80年代，非参数统计借助计算机技术和大量计算获得更稳健的估计和预测，以P.J.Huber以及 F.Hampel为代表的统计学家从计算技术的实现角度，为衡量估计量的稳定性提出了新准则。**
- **90年代有关非参数统计的研究和应用主要集中在非参数回归和非参数密度估计领域，其中较有代表性的人物是Silverman和J. Fan。**
- **大规模计算和自动化分析的需要将非参数统计引入机器学习领域。代表Hastie,Wasserman等。**



46

非参数统计历史（1932-1962）

在后Fisher时代1932-1962年统计思想史的历程：正是数据科学的孕育期，这个时代的特征是学科壁垒没有那么深厚，很多统计学家实际上一生都是在从事着其他学科，他们对于其他领域的眼界是很开阔的；也正是这段时间，我们看到了整个非参数话语体系的形成，它是在扩大传统统计通往机器学习的过度。他们在解决从化学、生物、心理等急速发展领域中的实际问题过程中发展出一种全新的数据分析观念，这些方法并不是来自于周密的论证，而是借着参数推断已形成的渐进理论和分布表技术，发展存在于数据本身特有的“小秩序”、“稳健性”、“小别离”和“局部特征”，这些统计方法在当时的推断文化中看似不占有核心位置，甚至也没有成为对思想来源领域认知的主流方法论，但是随着计算技术的发展，却具有动摇整个既有统计文化的强大力量，引起整个数据分析风向的深刻变革。

47

5. 基本概念

(1)分布函数和经验分布及图形表示
(2)数据的探索
(3)渐进相对效率
(4)非参数置信区间
(5)秩检验统计量
(6)U统计量*

48

经验分布的基本理论

定义： $X_1, \cdots, X_n$ F, 经验分布函数 $\hat{F}_n$ 是将每个数据点 X_i 上权重设为 1 的均匀分布分布函数，即

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i < x)$$

这里

$$I(X_i \le x) = \begin{cases} 1 & \text{如果} X_i \le x; \\ 0 & \text{如果} X_i > x. \end{cases}$$

49

[redacted]是F的一个很好估计？

- 给定x，[redacted]是一个随机变量：[redacted]服从二项分布
- 所以

50

经验分布函数的性质

定理 1: 令 $X_1, \cdots, X_n \sim F$, 经验分布函数 $\hat{F}_n$ 有以下性质：

1. 任意固定点 x,

$$\mathrm{E}(\hat{F}_n(x)) = F(x) \quad \mathrm{Var}(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

于是， $MSE(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} \to 0$, 因此 $\hat{F}_n(x) \to^{p} F(x)$.

2.(Glivenko-Cantelli) $\sup|\hat{F}_n(x) - F(x)| \to^{a.s.} 0$.

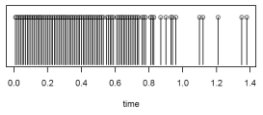
3.(Dvoretzky-Kiefer-Wolfowitz(DKW)inequality) 对任意的 $\epsilon > 0$,

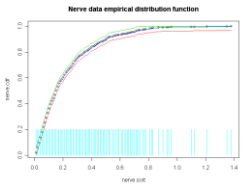
$$P\left(\sup|\hat{F}_n(x) - F(x)| > \epsilon\right) \le 2e^{-2n\epsilon^2}.$$

51

分布函数的估计

- 例:1966年Cox和Lewis的一篇研究报告给出了神经纤维细胞连续799次激活的等待时间(相邻脉冲)的分布拟合,数据的经验分布函数如图:





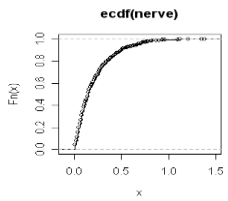

```
((1/(2*799))*log(2/0.05,exp(1)))^0.5
[1] 0.04804618
```

52

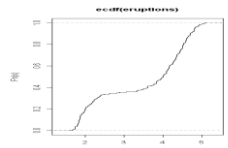
```
nerve=scan("E:\\data\\nonpar\\nerve.dat")
nerve.sort=sort(nerve)
nerve.rank=rank(nerve.sort)
nerve.cdf=nerve.rank/length(nerve)
plot(nerve.sort,nerve.cdf)
```

53

使用ecdf函数制作分布函数



plot(ecdf(nerve,verticals=T))




```
attach(faithful)
plot(ecdf(eruptions), do.points=FALSE, verticals=TRUE)
```

54

统计函数的估计

- 统计函数：F的任意函数
  - 如均值：$\mu = \int x dF(x)$
  - 如方差：$\sigma^2 = \int (x-\mu)^2 dF(x)$
  - 如中值：$m = F^{-1}(1/2)$
- 统计函数的估计：插入估计(Plug-in Estimator)
  - $\theta = T(F)$ 的插入估计为 $\hat{\theta}_n = T(\hat{F}_n)$
  - 插入 $\hat{F}_n$ 代替未知的F

55

经验分布的变形---生存函数

生存函数是生存分析里基本的概念

$$S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

$$S_n(t) = 1 - F_n(t)$$

危险函数是另一个生存分析中的重要内容，它表示一个生存超过给定时间的个体瞬时死亡率。生存图形可以非正式地表现危险函数。如果一个个体在时刻 t 仍然存活，那么个体在时间范围为 $(t, t+\delta)$ 死亡的概率为：（假设密度函数 f 在 t 上是连续的）

$$\begin{aligned}\mathbb{P}(t \le T \le t+\delta | T \ge t) &= \frac{\mathbb{P}(t \le T \le t+\delta)}{\mathbb{P}(T \ge t)} \\ &= \frac{F(t+\delta) - F(t)}{1 - F(t)} \\ &\approx \frac{\delta f(t)}{1 - F(t)}\end{aligned}$$

56

几内亚猪生存函数

例 1：（数据见光盘）数据来自受不同程度结核病毒感染的几内亚猪的死亡时间．其中实验组分为五组，每组安排 72 只猪，组内受同等程度结核病毒感染，1-5 组感染病毒的程度依次增大，标记为 1，2，3，4，5，对照组包含 107 只猪，没有受到感染．对这些试验观察两年以上，纪录猪死亡时间．这个例子中，我们用经验分布函数估计生存函数，研究受不同程度结核病毒感染几内亚猪的生存情况．如图所示，

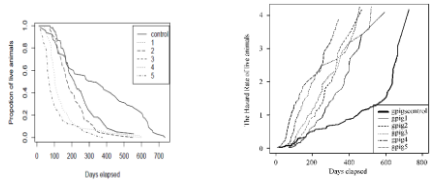


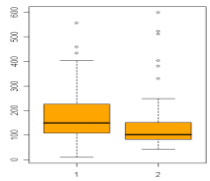
图 2.7 几内亚猪经验生存函数

57

分位数和分位数的图形表示法

$$q = \int_{-\infty}^{x_q} p(x)dx$$

boxplot(g3,g4,col="orange")



58

爱荷华大学医学院1935-1948年间26例住院精神病患者生存资料分析，该样本是对住院精神病患者进行的一项更大规模研究的一部分。Tsuang和Woolson（1977）讨论过该数据，每个患者的数据包括首次入院时的年龄、医院、性别、随访年数（从入院到死亡或检查的年数）和患者后续时间的状态。研究目标是想知道男性精神病患者和女性精神病患者比普通公众会不会更易于死亡的风险，医院的数据显示在下表中

Obs	sex	age	time	death
1	2	51	1	1
2	2	58	1	1
3	2	55	2	1
4	2	28	22	1
5	1	21	30	0
6	1	19	28	1
7	2	25	32	1
8	2	48	11	1
9	2	47	14	1
10	2	25	36	0
11	2	31	31	0
12	1	24	33	0
13	1	25	33	0
14	2	30	37	0
15	2	33	35	0
16	1	36	25	1
17	1	30	31	0
18	1	41	22	1
19	2	43	26	1
20	2	45	24	1
21	2	35	35	0
22	1	29	34	0
23	1	35	30	0
24	1	32	35	1
25	2	36	40	1
26	1	32	39	0

59

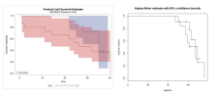
A combination of survival curves over strata (male, female) with accounting for left-truncation can be obtained in R with the following code (Diaz).

```
install.packages('survival')
install.packages('KMsurv')

library(survival)
library(KMsurv)


data(psych); attach(psych)
my.surv.object <- Surv(age, age+time, death)
my.surv.object
survfit(my.surv.object~sex)
my.fit <- survfit(my.surv.object~sex)

plot(my.fit, main="Kaplan-Meier estimate with 95% confidence bounds",
   xlab="time", ylab="survival function")
```

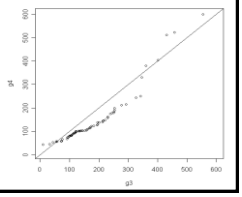


60

qqnorm, qqplot和qqline

- qqplot是比较两组数据的分位数大小的绘图工具，它使用的方法是将$(x_{(i)},y_{(i)})$做散点图，这相当于将两组数据的分位数点做二维散点图 .
- qqline和qqnorm是用来比较一组数据是否是正态分布.

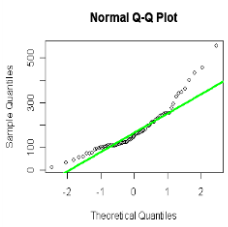
```
qqplot(g3,g4,xlim=c(8,600),ylim=c(8,600))
abline(0,1)
```

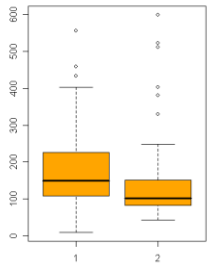


图象显示:
1.第三组大部分寿命长于第四组;
2.但在极值部分, 较为长寿的第四组比第三组略有优势.最短命的也出现在第三组.

61

```
qqnorm(g3)
qqline(g3,col="green",lwd=3)
```

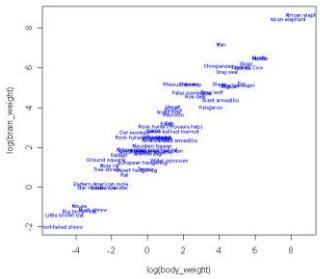




```
boxplot(g3,g4,col="orange")
```

62

散点图探察数据之间的关系



63

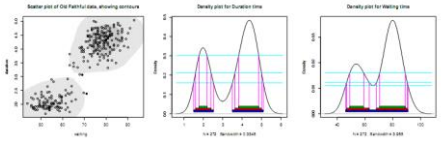


Figure 1: Scatter plot of the Old Faithful data (on the left), with bivariate probability contour. On the middle and right, univariate density plot for "waiting time" and "duration time" with 75% (green), 90% (red) and 95% (blue) confidence intervals (shaded bar).

64

练习题:

- 请对数据iris. dat进行分析,分别做出每个变量的经验分布函数,比较他们的异同.

65

势函数及其影响例2：

- Poisson分布

  H_0:λ >1　　H_1: $\lambda \le 1$

- 按照假设检验的步骤，可以选取统计量 $\sum X_i$ 为检验统计量，检验的目的是使得

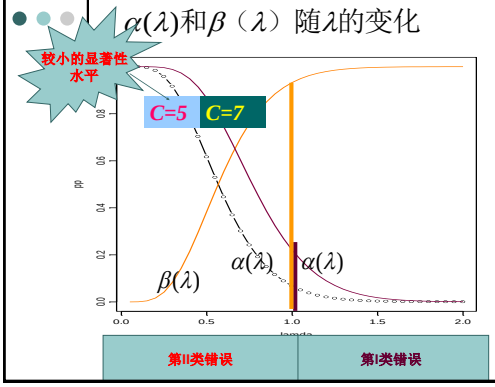
$$\alpha(\lambda) = P(\sum X_i < C)\text{足够小}$$

$$\max(\alpha(\lambda)) = \max(P(\sum X_i < C)) = \alpha_0\text{足够小}$$

$\alpha(\lambda)$和β（λ）随λ而发生变化

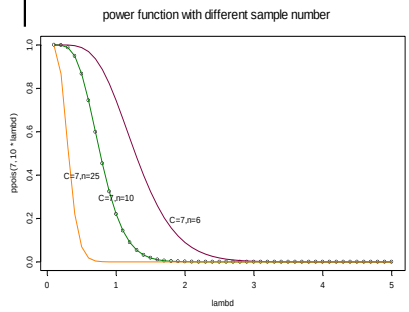
66

$\alpha(\lambda)$和β（λ）随λ的变化



67

同一个检验在不同样本量下的势函数



68

势函数

定义 2.1:（检验的势）对一般的假设检验问题： $H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$, 其中 $\Theta_0 \bigcap \Theta_1 = \phi$, 检验统计量为 T_n ， 拒绝零假设的概率，也就是样本落入拒绝域 W 的概率为检验的 **势** ，记为：

$$g_{T_n}(\theta) = \mathrm{P}(T_n \in W), \theta \in \Theta = \Theta_0 \bigcup \Theta_1.$$

由定义 2.1 可知，当 $\theta \in \Theta_0$ 时，检验的势是犯第 I 类错误的概率，一般由显著性水平 α 控制；当 $\theta \in \Theta_1$ 时，检验的势称为检验的 **功效** 。 $1 - g(\theta)$

影响到势函数的因素有

(1)显著性水平的大小

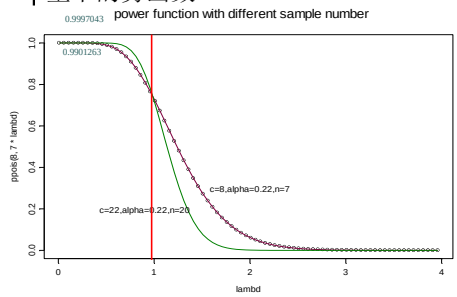
(2)参数真值

(3)样本大小

(4)检验统计量的选择

69

同样显著性水平之下的检验在不同样本量下的势函数



70

小结论

- 显著性水平大(alpha)对应较大的C, 于是对应较大的势.
- 样本量大,势大

71

例3

例 2.1: 假设 $X \sim p(x)$ ， 分布密度 $p(x)$ 有如下形式：

$$p(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & 0 < x < +\infty \\ 0, & \text{其它} \end{cases}$$

考虑假设检验问题：

$$H_0 : \theta = 2 \leftrightarrow H_1 : \theta \geq 2.$$

简单随机抽样 X_1, X_2 ，易知 $\frac{X_1+X_2}{2}$ 是 θ 的无偏估计，因此构造如下拒绝域：

$$C = \{(X_1, X_2) : 9.5 < X_1 + X_2 < +\infty\}.$$

计算该检验的 α 和 β 。

72

（2）检验的相对效率

73

渐进效率的概念

对假设检验问题

$$H_0:\theta=\theta_0 \leftrightarrow H_1:\theta\neq\theta_0,$$

取备择假设序列 $\theta_i(i=1,2,\cdots),\theta_i\neq\theta_0$，且 $\lim_{i\to\infty}\theta_i=\theta_0$。对固定的功效 $1-\beta$ 之下，我们考虑两个检验统计量 V_{n_i} 和 T_{m_i}，其中 V_{n_i} 和 T_{m_i} 分别是备择检验为 θ_i 所对应的两个检验统计量序列，n_i 和 m_i 是两个统计量分别对应的样本量，势函数满足：

$$\lim_{i\to\infty} g_{V_{n_i}}(\theta_0)=\lim_{i\to\infty} g_{T_{m_i}}(\theta_0)=\alpha$$

$$\alpha<\lim_{i\to\infty} g_{V_{n_i}}(\theta_i)=\lim_{i\to\infty} g_{T_{m_i}}(\theta_i)=1-\beta<1$$

如果极限：

$$e_{VT}=\lim_{i\to\infty}\frac{m_i}{n_i}$$

存在，且独立于 θ_i,α, 和 β，则称 e_{VT} 是 V 相对于 T 的 **渐进相对效率**，简记为 ARE(V,T)。

74

定理 2.3: 对假设检验问题 $H_0:\theta=\theta_0 \leftrightarrow H_1:\theta\neq\theta_0$

(1)V_n 和 T_m 是相容的统计量，也就是说：当 $n,m\to+\infty,\forall\theta\neq\theta_0$,

$$g(\theta_i,V_{n_i})\to 1 \quad g(\theta_i,T_{m_i})\to 1;$$

(2) 如果记 $\mathrm{E}(V_{n_i})=\mu_{V_{n_i}}$，$\mathrm{Var}(V_{n_i})=\sigma^2_{V_{n_i}}$，$\mathrm{E}(T_{m_i})=\mu_{T_{m_i}}$，$\mathrm{Var}(T_{m_i})=\sigma^2_{T_{m_i}}$，则在 $\theta=\theta_0$ 的邻域中一致有：

$$\frac{V_{n_i}-\mu_{V_{n_i}}(\theta)}{\sigma_{V_{n_i}}(\theta)}\xrightarrow{\mathcal{L}} N(0,1) \qquad \frac{T_{m_i}-\mu_{T_{m_i}}(\theta)}{\sigma_{T_{m_i}}(\theta)}\xrightarrow{\mathcal{L}} N(0,1)$$

(3) 存在导数：$\frac{d\mu_{V_{n_i}}(\theta)}{d\theta}|_{\theta=\theta_0}$ $\frac{d\mu_{T_{m_i}}(\theta)}{d\theta}|_{\theta=\theta_0}$；而且 $\mu'_{V_{n_i}}(\theta),\mu'_{T_{m_i}}(\theta)$ 在 $\theta=\theta_0$ 的某一个闭邻域中连续，导数不为 0。

(4) $\lim_{i\to\infty}\frac{\sigma_{V_{n_i}}(\theta_i)}{\sigma_{V_{n_i}}(\theta_0)}=\lim_{i\to\infty}\frac{\sigma_{T_{m_i}}(\theta_i)}{\sigma_{T_{m_i}}(\theta_0)}=1$; $\lim_{i\to\infty}\frac{\mu_{V_{n_i}}(\theta_i)}{\mu_{V_{n_i}}(\theta_0)}=\lim_{i\to\infty}\frac{\mu_{T_{m_i}}(\theta_i)}{\mu_{T_{m_i}}(\theta_0)}=1$;

(5) $\lim_{i\to\infty}\frac{\mu'_{V_{n_i}}(\theta_0)}{\sqrt{n_i\sigma^2_{V_{n_i}}(\theta_0)}}=C_V$, $\lim_{i\to\infty}\frac{\mu'_{T_{m_i}}(\theta_0)}{\sqrt{m_i\sigma^2_{T_{m_i}}(\theta_0)}}=C_T$,

则 V 相对于 T 的 Pitman 相对效率等于：

$$ARE(V,T)=\frac{mi}{ni}=\frac{C_V^2}{C_T^2}$$

75

定义 2.4: 假设检验问题：$H_0:\theta=\theta_0 \leftrightarrow H_1:\theta=\theta_1$, 上述定理中定义的极限为

$$\lim_{i\to\infty}\frac{\mu'_{V_{n_i}}(\theta_0)}{\sqrt{n}\sigma_{V_{n_i}}(\theta_0)}$$

称为 V_n 的 **效率**，记为：$\mathrm{eff}(V)$.

例 2.2: 考虑总体为正态分布

$$p(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty<x<+\infty,$$

假设检验问题：$H_0:\mu=0\leftrightarrow H_1:\mu=\mu_i(i=1,2,\cdots)$，$\lim_{i\to\infty}\mu_i=0$，考虑检验统计量：$T_n=\sqrt{n}(\bar{X}\quad)/S$ 和 $SG_n=\sum_{i=1}^n I_{(X_i>0)}$，其中：$\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$ 是样本均值，$S^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2$ 是样本方差，$I_{(X_i>0)}$ 是示性函数，计算 $ARE(T,SG)$。

76

解：根据 t 分布的性质有，

$$\mathrm{E}_\mu(T_n)=\frac{\mu}{\frac{\sigma}{\sqrt{n}}},\quad \mathrm{Var}_\mu(T_n)=1$$

$$\mathrm{E}_\mu(SG_n)=np,\quad \mathrm{Var}_\mu(SG_n)=np(1-p)$$

因而 $\mathrm{eff}(T_n)=\frac{1}{\sigma}$. 其中

$$p=\int_0^\infty \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,\mathrm{d}t.$$

容易证明他们满足 Nother 定理的条件 (1) ∼ (5)，而且，

$$[\mathrm{E}_\mu(T_n)]'=\frac{\sqrt{n}}{\sigma}$$

$$[\mathrm{E}_\mu(SG_n)]'=\frac{n}{\sqrt{2\pi}\sigma}\int_0^\infty\frac{1}{\sigma^2}(t-\mu)e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\,\mathrm{d}t=\frac{n}{\sqrt{2\pi}\sigma}\int_0^\infty \mathrm{d}\left(-e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\right)=\frac{n}{\sqrt{2\pi}\sigma}e^{-\frac{\mu^2}{2\sigma^2}}$$

$$\mathrm{eff}(SG_n)=\lim_{n\to\infty}\frac{[\mathrm{E}_0(SG_n)]'}{\sqrt{n\mathrm{Var}_0(SG_n)}}=\lim_{n\to\infty}\left[\frac{n}{\sqrt{2\pi}\sigma}\Big/\frac{\sqrt{n}}{2}\right]=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}$$

因此，T 相对于 SG 的渐进相对效率为，

$$ARE(T,SG)=\left[\frac{1}{\sigma}\Big/\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right]^2=\frac{\pi}{2}.$$

77

假设$X=\frac{(n-1)S^2}{\sigma^2}\sim\Gamma\left(\frac{n-1}{2},\frac{1}{2}\right)=\chi^2(n-1)$

$$E\left(\frac{1}{S}\right)=\int_0^{+\infty}\frac{1}{S}\cdot\frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}\cdot x^{\frac{n-1}{2}-1}\cdot e^{-\frac{x}{2}}dx=\int_0^{+\infty}\frac{1}{\sqrt{\frac{x\sigma^2}{n-1}}}\cdot\frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)}\cdot x^{\frac{n-1}{2}-1}\cdot e^{-\frac{x}{2}}dx$$

$$=\int_0^{+\infty}\frac{\sqrt{n-1}}{\sigma}\cdot\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\cdot\frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{n-2}{2}\right)}\cdot\frac{\left(\frac{1}{2}\right)^{\frac{n-2}{2}}}{\Gamma\left(\frac{n-2}{2}\right)}\cdot x^{\frac{n-2}{2}-1}\cdot e^{-\frac{x}{2}}dx=\frac{1}{\sigma}\left(\frac{n-1}{2}\right)^{\frac{1}{2}}\cdot\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

由Stirling公式

$$\Gamma(x)\approx\sqrt{2\pi}\cdot e^{-x}\cdot x^{x-\frac{1}{2}}$$

$$\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}=\frac{\sqrt{2\pi}\cdot e^{-\frac{n-2}{2}}\cdot\left(\frac{n-2}{2}\right)^{\frac{n-2}{2}-\frac{1}{2}}}{\sqrt{2\pi}\cdot e^{-\frac{n-1}{2}}\cdot\left(\frac{n-1}{2}\right)^{\frac{n-1}{2}-\frac{1}{2}}}=e^{\frac{1}{2}}\cdot\left(\frac{n-2}{n-1}\right)^{\frac{n-3}{2}}\cdot\frac{\sqrt{2}}{\sqrt{n-1}}$$

所以$E\left(\frac{1}{S}\right)=\frac{1}{\sigma}\left(\frac{n-1}{2}\right)^{\frac{1}{2}}\cdot e^{\frac{1}{2}}\cdot\left(\frac{n-2}{n-1}\right)^{\frac{n-3}{2}}\cdot\frac{\sqrt{2}}{\sqrt{n-1}}\to\frac{1}{\sigma}(n\to\infty)$

78

秩检验统计量

79

无结点秩的定义

定义 2.7: 设样本 $X_1,\cdots,X_n$ 是取自总体 X 的简单随机抽样， $X_1,\cdots,X_n$ 中不超过 X_i 的数据个数，即 $R_i=\sum_{j=1}^{n} I(X_j\le X_i)$ ，称 R_i 为 X_i 的**秩**， X_i 是第 R_i 个顺序统计量， $X_{(R_i)}=X_i$. 令 $R=(R_1,\cdots,R_n)$,R 是由样本产生的统计量，称为秩统计量。

例题：某学院本科三年级由9个专业组成，统计每个专业学生每月消费数据如下，用R求消费数据的秩和顺序统计量的现值：

300 230 208 580 690 200 263 215 520

80

```
> a.ex
 [1]  10.0  12.5  15.0  17.5  20.0  22.5  25.0  27.5  30.0  32.5  35.0  37.5
[13]  40.0  42.5  45.0  47.5  50.0  52.5  55.0  57.5  60.0  62.5  65.0  67.5
[25]  70.0  72.5  75.0  77.5  80.0  82.5  85.0  87.5  90.0  92.5  95.0  97.5
[37] 100.0 102.5 105.0 107.5 110.0 112.5 115.0 117.5 120.0
> sample(a.ex,40)
 [1]  47.5  57.5  82.5  12.5  87.5  42.5  60.0 102.5  67.5 105.0  17.5  10.0
[13]  92.5  70.0 115.0 120.0  35.0  15.0 112.5  22.5  52.5 100.0  97.5  20.0
[25]  30.0  27.5  50.0  45.0  77.5  90.0  55.0  75.0  72.5  95.0  62.5  25.0
[37]  37.5  32.5  80.0  40.0
> sample(a.ex,40,rep=T)
 [1]  22.5 115.0 105.0 100.0  70.0 100.0  77.5  47.5  80.0 105.0  20.0  62.5
[13]  42.5  42.5 105.0  72.5  87.5  50.0  85.0  80.0 102.5  70.0 110.0 110.0
[25]  22.5  35.0 117.5  77.5  27.5 117.5  55.0 102.5  50.0  42.5  50.0  50.0
[37]  42.5  90.0  25.0  45.0
> b.ex=sample(a.ex,40,rep=T)
> sort(b.ex)
 [1]  10.0  15.0  15.0  15.0  17.5  25.0  25.0  25.0  27.5  30.0  32.5  35.0
[13]  37.5  40.0  62.5  62.5  65.0  70.0  72.5  75.0  75.0  75.0  75.0  77.5
[25]  77.5  77.5  77.5  80.0  87.5  90.0  90.0  92.5  92.5  97.5 100.0 107.5
[37] 115.0 115.0 117.5 120.0
> rank(b.ex)
 [1] 15.5 32.5  7.0 13.0  1.0  7.0 32.5  3.0 18.0  7.0 12.0 19.0 21.5 25.5
[15] 15.5 25.5 30.5 21.5 37.5 34.0  5.0 25.5 40.0 39.0 37.5 29.0  9.0 10.0
[29]  3.0 35.0 25.5 36.0 14.0 17.0 11.0 28.0 21.5  3.0 21.5 30.5
> order(b.ex)
 [1]  5  8 29 38 21  3  6 10 27 28 35 11  4 33  1 15 34  9 12 13 18 37 39 14
[25] 16 22 31 36 26 17 40  2  7 20 30 32 19 25 24 23
> rev(order(b.ex))
 [1] 23 24 25 19 32 30 20  7  2 40 17 26 36 31 22 16 14 39 37 18 13 12  9 34
[25] 15  1 33  4 11 35 28 27 10  6  3 21 38 29  8  5
```

81

定理 2.8: 对于简单随机样本， $R=(R_1,\cdots,R_n)$ 等可能取 $(1,\cdots,n)$ 的任意 $n!$ 个排列之一， R 在由 $(1,\cdots,n)$ 的所有排列组成的空间上是均匀分布，即：对 $(1,\cdots,n)$ 的任一排列 $(i_1,\cdots,i_n)$ 有

$$\mathrm{P}(R=(i_1,\cdots,i_n))=\frac{1}{n!}$$

推论 2.9: 对于简单随机样本，对任意 $r,s=1,\cdots,n, r\ne s$ 及 $i\ne j$ ，

$$\begin{aligned}\mathrm{P}(R_i &= r)=\frac{1}{n}\\ \mathrm{P}(R_i &= r,R_j=s)=\frac{1}{n(n-1)}.\end{aligned}$$

推论 2.10: 对于简单随机样本，

$$\begin{aligned}\mathrm{E}(R_i) &= \frac{n+1}{2}\\ \mathrm{Var}(R_i) &= \frac{(n+1)(n-1)}{12}\\ \mathrm{Cor}(R_i,R_j) &= -\frac{(n+1)}{12}.\end{aligned}$$

82

2.有结数据的秩

- **设样本 $X_1,...,X_n$ 取自总体X的简单随机抽样，将数据排序后,相同的数据点组成一个“结”，称重复数据的个数为结长。**

  例1：3.8 3.2 1.2 1.2 3.4 3.2 3.2

- **解：有4个结3.2结长为3。**

定义 2.10: 将样本 $X_1,\cdots,X_n$ 从小到大排序后，如果 $X_{(1)}=X_{(2)}=\cdots=X_{(\tau_1)}<X_{(\tau_1+1)},\cdots,X_{(\tau_1+\tau_2)}<\cdots<X_{(\tau_1+\cdots+\tau_{g-1})}=,\cdots,X_{(\tau_1+\cdots+\tau_g)}$ 其中 g 是样本中结的个数， τ_i 是第 i 个结的长度， $(\tau_1,\cdots,\tau_g)$ 是 g 个正整数，$\sum\tau_i=n$, 称 $(\tau_1,\cdots,\tau_g)$ 为结统计量. 第 i 组样本的秩都相同，如下所示：

$$r_i=\frac{1}{\tau_i}\sum_{k=1}^{\tau_i}(\tau_1+\cdots+\tau_{i-1}+k)=\tau_1+\cdots+\tau_{i-1}+\frac{1+\tau_i}{2}.$$

83

例 2.4: 样本数据有 12 个数，其值，秩和结统计量 (用 τ_i 表示，为第 i 个结中的观察值数量) 为：

表 2.1. 结的计算

观察值	2	2	4	7	7	7	8	9	9	9	9	10
秩	1.5	1.5	3	5	5	5	7	9.5	9.5	9.5	9.5	12

其中有 6 个结，每个结长分别为 2,1,3,1,4,1 。

84

假设有N个样本，记R_i为$x_i, i=1,2,\cdots,N$的不考虑平均秩下的秩。$\alpha(i),, i=1,2,\cdots,N$是一个计分函数，当结的长度为1时，$a(R_i)=R_i$,当结的长度大于1时,$a(R_i)$取平均秩。

1 结数据秩与秩平方和的一般性质

在一个由N个已排序好的数列中，其中有一段由τ个数组成的结数据，如果这个结的第一个数的秩$R_{r+1}=r+1$,

1.当这τ个数完全不同时，这些数的秩和为

$$(r+1)+(r+2)+\cdots+(r+\tau)=\tau r+\frac{\tau(\tau+1)}{2} \tag{1}$$

这些数的秩的平方和为

$$(r+1)^2+(r+2)^2+\cdots+(r+\tau)^2=\tau r^2+\tau r(\tau+1)+\frac{\tau(\tau+1)(2\tau+1)}{6}$$

2.当这τ个数完全相同时，这些数的秩和为

$$(r+\frac{\tau+1}{2})+(r+\frac{\tau+1}{2})+\cdots+(r+\frac{\tau+1}{2})=\tau r+\frac{\tau(\tau+1)}{2} \tag{2}$$

这些数的秩的平方和为

$$(r+1)^2+(r+2)^2+\cdots+(r+\tau)^2=\tau r^2+\tau r(\tau+1)+\frac{\tau(\tau+1)^2}{4}.$$

式（1）和（2）可以发现不论这τ个数是否全相同，秩的和都是相同的，但是秩的平方和不同。完全相同的数列比完全不同的数列的秩平方和大$\frac{\tau^3-\tau}{12}$,

85

2 结数为g的数据秩的一般性质

1. 对于N个有结数据的秩和仍然为

$$\sum_{i=1}^{N}\alpha(R_i)=\sum_{i=1}^{N}\alpha(i)=\frac{(N+1)N}{2}$$

由于无结数据的秩的平方和为

$$\sum_{i=1}^{N}\alpha(R_i)^2=\frac{N(N+1)(2N+1)}{6},$$

所以结数为g的数据的秩平方和为

$$\sum_{i=1}^{N}\alpha(R_i)^2=\sum_{i=1}^{N}\alpha(i)^2=\frac{N(N+1)(2N+1)}{6}-\sum_{j=1}^{g}\frac{\tau_j^3-\tau_j}{12}$$

2.对于$x_1,x_2,\cdots,x_N$,i.i.d的情况，$\alpha(R_i)$等可能的取$\alpha(i)$, 有

$$\mathrm{E}(\alpha(R_i))=\bar{\alpha}=\frac{\sum_{j=1}^{N}\alpha(i)}{N}$$

$$\mathrm{Var}(\alpha(R_i))=\frac{\sum_{i=1}^{N}(\alpha(i)-\bar{\alpha})^2}{N}$$

对于协方差$\mathrm{cov}(\alpha(R_i),\alpha(R_j))=E(\alpha(R_i)\alpha(R_j))-E(\alpha(R_i)E(\alpha(R_j)$

$$\mathrm{E}(\alpha(R_i)\alpha(R_j))=\frac{\sum_{i\neq j}\alpha(i)\alpha(j)}{N(N-1)}=\frac{N^2\bar{\alpha}^2-\sum_{i=1}^{N}a(i)^2}{N(N-1)}$$

$$\mathrm{cov}(\alpha(R_i),\alpha(R_j))=\frac{N^2\bar{\alpha}^2-\sum_{i=1}^{N}a(i)^2}{N(N-1)}-\bar{\alpha}^2$$

$$\mathrm{cov}(\alpha(R_i),\alpha(R_j))=\frac{N\bar{\alpha}^2-\sum_{i=1}^{N}a(i)^2}{N(N-1)}=-\frac{\sum_{i=1}^{N}(\alpha(i)-\bar{\alpha})^2}{N(N-1)}$$

86

令$x_1,x_2,...,x_n$为N个i.i.d数列中的任意n个数，则

$$\mathrm{E}(\sum_{i=1}^{n}\alpha(R_i))=\sum_{i=1}^{n}\mathrm{E}(\alpha(R_i))=n\bar{\alpha} \tag{3}$$

$$\mathrm{Var}(\sum_{i=1}^{n}\alpha(R_i)=\sum_{i=1}^{n}\mathrm{Var}(\alpha(R_i))+2\sum_{i<j}\mathrm{cov}(\alpha(R_i),\alpha(R_j)) \tag{4}$$

$$=n\mathrm{Var}(\alpha(R_i))+n(n-1)\mathrm{cov}(\alpha(R_i),\alpha(R_j)) \tag{5}$$

$$=n\frac{\sum_{i=1}^{N}(\alpha(i)-\bar{\alpha})^2}{N}-n(n-1)\frac{\sum_{i=1}^{N}(\alpha(i)-\bar{\alpha})^2}{N(N-1)} \tag{6}$$

$$=\frac{n(N-n)\sum_{i=1}^{N}(\alpha(i)-\bar{\alpha})^2}{N(N-1)} \tag{7}$$

87

带入之前关于有结数据秩和和秩的平方和的结论，可以得到，$\bar{\alpha}=\frac{N+1}{2}$

$$\begin{aligned}\sum_{i=1}^{N}(\alpha(i)-\bar{\alpha})^2&=\sum_{i=1}^{N}\alpha(i)^2-N\bar{\alpha}^2\\&=\frac{N(N+1)(2N+1)}{6}-\sum_{j=1}^{g}\frac{\tau_j^3-\tau_j}{12}-\frac{N(N+1)^2}{4}\\&=\frac{N(N+1)(N-1)}{12}-\frac{\sum_{j=1}^{g}(\tau_j^3-\tau_j)}{12}\end{aligned}$$

代入(3)式和(4)式可得，

$$\mathrm{E}(\sum_{i=1}^{n}\alpha(R_i))=\frac{n(N+1)}{2}$$

$$\mathrm{Var}(\sum_{i=1}^{n}\alpha(R_i))=\frac{n(N-n)(N+1)}{12}-\frac{n(N-n)\sum_{j=1}^{g}(\tau_j^3-\tau_j)}{12N(N-1)}$$

88

U统计量

89

核的概念

定义 2.11: 设 $X_1,\cdots,X_n$ 取自分布族 $\mathcal{F}=\{F(\theta),\theta\in\Theta\}$, 如果待估参数 θ 存在样本量为 k 的无偏估计量 $h(X_1,\cdots,X_k),k<n$. 即满足:

$$\mathrm{E}h(X_1,\cdots,X_k)=\theta,\forall\theta\in\Theta$$

使上式成立的最小的样本量为 k , 则称参数 θ 是 k 可估的, 此时 $h(X_1,\cdots,X_k)$ 称为参数 θ 的核（ Kernel ）。

例：总体期望有无偏估计X_1，总体期望是1可估的， X_1是总体期望的核。

JunShao Chap3.2.1

The use of U-statistics is an effective way of obtaining unbiased estimators. In nonparametric problems, U-statistics are often UMVUE's, whereas in parametric problems, U-statistics can be used as initial estimators to derive more efficient estimators.

90

对称核和U统计量的概念

一般，还要求核有对称的形式，也就是说：对 $\forall(1,\cdots,k)$ 的任何一个排列 $(i_1,\cdots,i_k)$，有 $h(X_1,\cdots,X_k)=h(X_{i_1},\cdots,X_{i_k})$. 如果核本身不对称，可以构造对称的核函数：

$$h^*(X_1,\cdots,X_k)=\frac{1}{k!}\sum_{(i_1,\cdots,i_k)}h(X_{i_1},\cdots,X_{i_k})$$

定义 2.12: 设 $X_1,\cdots,X_n$ 取自分布 $\mathcal{F}=\{F(\theta),\theta\in\Theta\}$ 的样本，可估参数 θ 存在样本量为 k 的无偏估计量 $h(X_1,\cdots,X_k)$， θ 有对称核 $h^*(X_1,\cdots,X_k)$, 则参数 θ 的 U 统计量如下定义：

$$U(X_1,\cdots,X_n)=\frac{1}{\binom{n}{k}}\sum_{(i_1,\cdots,i_k)}h^*(X_{i_1},\cdots,X_{i_k})$$

91

例 2.6: 设 $\mathcal{F}=\{F(\theta),\theta\in\Theta\}$ 为全体二阶矩有限的分布类，则方差 $\theta=\mathrm{E}(X-\mathrm{E}X)^2$ 是 2 阶可估参数.

由 $\mathrm{E}(X-\mathrm{E}X)^2=\mathrm{E}X^2-(\mathrm{E}X)^2$，可知:

$$h(X_1,X_2)=X_1^2-X_1X_2$$

是参数 θ 的无偏估计，显然它不具有对称性，如下构造对称核:

$$h^*(X_1,X_2)=\frac{1}{2}[(X_1^2-X_1X_2)+(X_2^2-X_1X_2)]=\frac{1}{2}(X_1-X_2)^2,$$

相应的 U 统计量为:

$$\begin{aligned}&U(X_1,\cdots,X_n)\\&=\frac{1}{\binom{n}{2}}\sum_{i<j}\frac{1}{2}(X_i-X_j)^2\\&=\frac{1}{n(n-1)}\sum_{i<j}(X_i^2+X_j^2-2X_iX_j)\\&=\frac{1}{n(n-1)}\left[\frac{1}{2}\sum_{i\neq j}(X_i^2+X_j^2)-\sum_{i\neq j}X_iX_j\right]\end{aligned}$$

$$\begin{aligned}&=\frac{1}{n(n-1)}\left[\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(X_i^2+X_j^2)-\frac{1}{2}\sum_{i=1}^n(X_i^2+X_i^2)-\sum_{i\neq j}X_iX_j\right]\\&=\frac{1}{n(n-1)}\left[n\sum_{i=1}^nX_i^2-\left(\sum_{i=1}^nX_i\right)^2\right]\\&=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2.\end{aligned}$$

92

对称性检验的例子

3. Testing Symmetry. In some situations, it is important to test for symmetry about an unknown center. Here is one method based of the observation that for a sample of size 3, X_1, X_2, X_3 from a continuous distribution, symmetric about a point ξ, $P(X_1 > (X_2 + X_3)/2) = P((X_1 - \xi) > ((X_2 - \xi) + (X_3 - \xi))/2) = 1/2$. Because of this, $f(X_1, X_2, X_3) = \mathrm{sgn}(2X_1 - X_2 - X_3)$ is an unbiased estimate of $\theta(P) = P(2X_1 > X_2 + X_3) - P(2X_1 < X_2 + X_3)$. Here, $\mathrm{sgn}(x)$ represents the sign function, which is 1 if $x > 0$, 0 if $x = 0$ and -1 if $x < 0$. When P is symmetric, $\theta(P)$ has value zero . The corresponding symmetric kernel is

$$h(x_1,x_2,x_3)=\frac{1}{3}[\mathrm{sgn}(2x_1-x_2-x_3)+\mathrm{sgn}(2x_2-x_1-x_3)+\mathrm{sgn}(2x_3-x_1-x_2)]. \quad (12)$$

对称分布的性质：

```
Bt
[1]  0.48 0.40 0.64 0.40 0.54 0.54 0.46
0.46 0.48 0.40 0.44 0.42 0.60 0.40 0.48
0.50 0.54 0.50 0.54 0.54
```

```
Bt=NULL
for (i in 1:20)
{
x=rt(200,4)
n=50
x1=sample(x,n)
x2=sample(x,n)
x3=sample(x,n)
greater=sum(x1>(x2+x3)/2)/n
sum(x1<(x2+x3)/2)/n
Bt=c(Bt,greater)
}
Bt
```

93

从正态中产生U-stat程序：

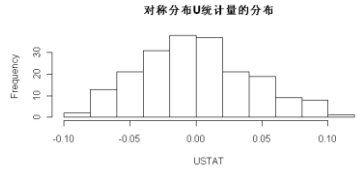
```
kk=200  #  sample number of U statistics
USTAT=NULL  #  kk refers to bootstrap number
for (mu in 1:kk)
{
a=rnorm(20,0,1) # generate normal sample from N(0,1) number is 20
x=a
#x=exp(a)
n=length(x)
H=NULL   # H refers to different kernels to generate U statistics
for (i in 1:(n-2))
for (j in (i+1):(n-1))
for (k in (j+1):(n))
 {
  a1=sign(2*x[i]-x[j]-x[k])
  a2=sign(2*x[j]-x[i]-x[k])
  a3=sign(2*x[k]-x[i]-x[j])
  h=1/3*(a1+a2+a3)  # h refers to symmetric kernel
  H=c(H,h)
  }
Ustat=(1/choose(length(x),3))*sum(H)  # Ustat refers to the Ustat of one time
USTAT=c(USTAT,Ustat)
}
USTAT2=USTAT
 hist(USTAT)
```

94

从非对称分布中产生U-stat程序：

```
kk=200  #  sample number of U statistics
USTAT=NULL  #  kk refers to bootstrap number
for (mu in 1:kk)
{
x=rnorm(20,0,1) # generate normal sample from N(0,1) number is 20
#x=exp(a)
n=length(x)
H=NULL   # H refers to different kernels to generate U statistics
for (i in 1:(n-2))
for (j in (i+1):(n-1))
for (k in (j+1):(n))
 {
  a1=sign(2*x[i]-x[j]-x[k])
  a2=sign(2*x[j]-x[i]-x[k])
  a3=sign(2*x[k]-x[i]-x[j])
  h=1/3*(a1+a2+a3)  # h refers to symmetric kernel
  H=c(H,h)
  }
Ustat=(1/choose(length(x),3))*sum(H)  # Ustat refers to the Ustat of one time
USTAT=c(USTAT,Ustat)
}
USTAT2=USTAT
 hist(USTAT,main="非对称分布U统计量的分布")
```

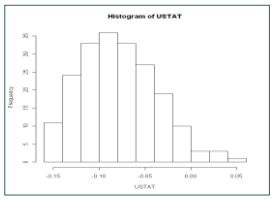
95



96

非对称总体分布：均值在0.8

Min. 1st Qu. Median Mean 3rd Qu. Max.
0.03656 0.44600 0.79680 0.87290 1.19200 2.35900



97

U统计量的特征计算

定理 2.13: 设 $X_1,\cdots,X_n$ 是取自分布 $\mathcal{F}=\{F(\theta),\theta\in\Theta\}$ 的简单随机样本，θ 是 k 可估参数，$U(X_1,\cdots,X_n)$ 是 θ 的 U 统计量，它的核是 $h(X_1,\cdots,X_k)$，有 $\mathrm{E}(U(X_1,\cdots,X_n)=\theta$,

$$\mathrm{Var}(U(X_1,\cdots,X_n))=\frac{1}{\binom{n}{k}}\sum_{i=1}^{k}\binom{k}{i}\binom{n-k}{k-i}\zeta_i$$

其中 $\zeta_i=\mathrm{cov}[h(X_1,\cdots,X_i,X_{i+1},\cdots,X_k),h(X_1,\cdots,X_k,X_{k+1},\cdots,X_{2k-i})]$.
特别，$\zeta_0=0,\zeta_r=\mathrm{Var}\{h(X_1,\cdots,X_r)\}$。

98

U统计量的大样本性质

定理 2.14: 设 $X_1,\cdots,X_n$ 取自分布 $\mathcal{F}=\{F(\theta),\theta\in\Theta\}$ 的简单随机样本，θ 是 k 可估参数，$U(X_1,\cdots,X_n)$ 是 θ 的 U 统计量，它的核 $h(X_1,\cdots,X_k)$，有

$$\mathrm{E}[h(X_1,\cdots,X_k)]^2<\infty,$$

则

$$\lim_{n\to\infty}\frac{n}{k^2}\mathrm{Var}[U(X_1,\cdots,X_n)]=\zeta_1$$

其中 $\zeta_1=\mathrm{cov}[h(X_1,\cdots,X_k),h(X_1,X_{k+1},\cdots,X_{2k-1}]>0$.

定理 2.15:(Hoeffding 定理) 设 $X_1,\cdots,X_n$ 取自分布 $\mathcal{F}=\{F(\theta),\theta\in\Theta\}$ 的简单随机样本，θ 是 k 可估参数，$U(X_1,\cdots,X_n)$ 是 θ 的 U 统计量，它的核是 $h(X_1,\cdots,X_k)$，

$$\mathrm{E}[h(X_1,\cdots,X_k)]^2<\infty,$$

当 $\zeta_1=\mathrm{cov}[h(X_1,\cdots,X_k),h(X_1,X_{k+1},\cdots,X_{2k-1}]>0$ 时，则

$$\lim_{n\to\infty}\sqrt{n}[U(X_1,\cdots,X_n)-\theta]\to N(0,k^2\zeta_1)(n\to+\infty).$$

99

Corollary 3.2. Under the condition of Theorem 3.4,
(i) $\frac{m^2}{n}\zeta_1\le\mathrm{Var}(U_n)\le\frac{m}{n}\zeta_m$;
(ii) $(n+1)\mathrm{Var}(U_{n+1})\le n\mathrm{Var}(U_n)$ for any $n>m$;
(iii) For any fixed m and $k=1,...,m$, if $\zeta_j=0$ for $j<k$ and $\zeta_k>0$, then

$$\mathrm{Var}(U_n)=\frac{k!\binom{m}{k}^2\zeta_k}{n^k}+O\left(\frac{1}{n^{k+1}}\right).\ \blacksquare$$

It follows from Corollary 3.2 that a U-statistic U_n as an estimator of its mean is consistent in mse (under the finite second moment assumption on h). In fact, for any fixed m, if $\zeta_j=0$ for $j<k$ and $\zeta_k>0$, then the mse of U_n is of the order n^{-k} and, therefore, U_n is $n^{k/2}$-consistent.

100

U统计量举例

例 2.7: 设 $X_1,\cdots,X_n$ 取自连续分布 $\mathcal{F}=\{F(\theta),\theta\in\Theta\}$ 的简单随机样本，固定 p, 假设 m_p 是样本的 p 分位数，$\forall i=1,\cdots,n$, 令 $Y_i=I_{(X_i<m_p)}$，定义计数统计量 $T=\sum_{i=1}^{n}Y_i$，证明：T/n 是 1 阶可估参数 $\mathrm{P}(X>m_p)$ 的 U 统计量，T/n 有渐进正态分布。

101