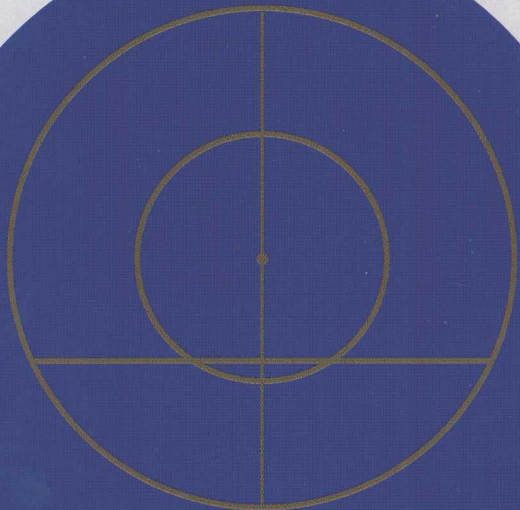
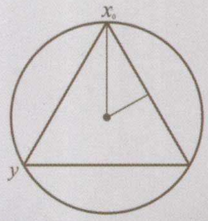
高等学校教材

数理统计

何书元





高等教育出版社
HIGHER EDUCATION PRESS

高等学校教材

数理统计

Shuli Tongji

何书元

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

内容简介

本书较系统地介绍了数理统计的基本内容，内容丰富，富有时代特色。书中有许多新的简明讲法，帮助学生更好地理解所学内容和加深对问题本质的理解.

本书以讲授数理统计的基本思想方法为主，同时介绍数理统计的诸多应用案例. 本书介绍的经验似然方法是近年来新发展的统计推断方法，在生物统计、可靠性分析、信号处理和不完全数据的统计推断方面有众多的应用，比现有教材中的似然比方法有计算简便、适用性更加广泛的优点.

为了帮助读者更快地掌握计算机的使用，本书以工程技术和科学研究中普遍使用的 Matlab 为例，在相关章节后面介绍有关的 Matlab 调用命令.

本书的内容和习题难度适中，适合作为理工科大学、师范和财经院校数学类专业和统计学专业本科生数理统计课程的教材或教学参考书. 学习本书的先修课程是数学分析、高等代数和概率论.

图书在版编目（CIP）数据

数理统计／何书元编. --北京：高等教育出版社，2012.1

ISBN 978-7-04-033793-8

Ⅰ. ①数… Ⅱ. ①何… Ⅲ. ①数理统计–高等学校–教材 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2011)第 273641 号

策划编辑 李蕊　责任编辑 李蕊　封面设计 张志奇　版式设计 余杨
插图绘制 宗小梅　责任印制 尤静

出版发行	高等教育出版社	咨询电话	400-810-0598
社　　址	北京市西城区德外大街4号	网　　址	http://www.hep.edu.cn
邮政编码	100120		http://www.hep.com.cn
印　　刷	北京宏信印刷厂	网上订购	http://www.landraco.com
开　　本	787mm×960mm 1/16		http://www.landraco.com.cn
印　　张	20.25	版　　次	2012年1月第1版
字　　数	370千字	印　　次	2012年1月第1次印刷
购书热线	010-58581118	定　　价	29.70元

物 料 号　33793-00

作者简介

何书元博士现任首都师范大学教授、北京大学数学科学学院兼职教授、教育部数学与统计学教学指导委员会副主任委员、统计学教学指导分委员会主任委员，从事概率论与数理统计的教学和科研工作，主讲的课程有概率论、数理统计、应用随机过程、应用时间序列分析等，是北京大学 2006 年国家精品课程数理统计的课程主持人和主讲教师， 1997–2009 年任北京大学教授.

前　　言

作为统计学的理论基础，数理统计已经渗透到理、工、农、医、经济、管理与人文社会科学领域，对各门科学的发展均起到了促进作用.

在北京大学和首都师范大学，编者多次为数学学院的学生讲授数理统计课程，为经济学院、生命科学学院、化学学院、物理学院、地空学院等讲授概率论与数理统计课程，在教学中参考了较多的国内外优秀教材，通过素材的积累，逐步形成了本书的框架. 本书的内容选择，包括例题和习题的选择较多地考虑了以后使用或继续学习统计学的需要. 除了数理统计的基本内容外，本书还力图通过较多的例题和习题介绍数理统计的众多应用领域.

经验似然方法是近年来新发展的统计推断方法，在生物统计、可靠性分析、信号处理和不完全数据的统计推断方面有众多的应用，比现有教材中的似然比方法有计算简便、适用性更加广泛的优点. 为了将这一方法尽快引入教学，本书的第五章对经验似然方法做了较为详细的介绍. 这部分内容供读者选择阅读.

为了帮助学生在学习中掌握计算机的使用，本书以工程技术和科学研究中广泛使用的 Matlab 为例，介绍和本书内容相关的 Matlab 调用命令. 参考这些简单的调用命令，同学们就可以处理和本书内容相关的实际数据了. 但是这些内容属于同学们的课外兴趣部分，不属于课程的要求范围。

本书的主要内容 (不带 * 的部分) 是针对每周四学时的课程设计的. 每周三学时的课程可根据需要略去第八或第九章. 带 * 的内容是为近一步提高统计学理论水平设计的，初学者可以略去.

除了编者写作的内容外，本书的部分例题和习题参考了书后所列的参考书，对这些书的作者表示感谢.

茆诗松教授对本书提出了许多重要的修改意见，使得本书的质量大有提高. 编者对茆老师表示衷心感谢.

本书的编写和出版得到教育部 2010 年度国家级教学团队项目、北京市本科数学基础课程教学团队项目和首都师范大学研究生专项经费的资助，特致谢意.

由于编者水平有限，书中不妥之处难免，敬请读者不吝指教.

何　书　元

2011 年 10 月于北京海淀蓝旗营

目　　录

第一章　描述性统计

现代生活是建立在数据之上的，没有数据，一切很难想象. 统计学是利用数据解释自然规律的科学，内容包括如何收集和分析数据.

基于统计学的数据处理方法称为统计方法. 在科学研究、工农业生产、新产品开发、产品质量的提高乃至政治、教育、社会科学等各个领域，使用统计方法和不使用统计方法获得的结果是大不相同的. 只要统计方法使用得当，就能够收到事半功倍的效果. 这也是统计学能随着科学技术和国民经济的发展而快速发展的重要原因.

统计学没有自己的基于试验的专门研究对象，但是可以为物理学家、化学家、医生、社会学家、心理学家等提供一套研究问题的有效方法. 这套方法可以帮助各个领域的研究工作者更快地获得成功.

§1.1　总体和参数

日常生活中我们总是自觉或不自觉地和总体与样本打交道. 买桔子时，先要尝尝这批桔子甜不甜. 这时称这批桔子是一个总体，单个的桔子是个体.

在仅关心桔子的甜度时，我们可以称单个桔子的甜度是个体，称所有的桔子的甜度为总体. 这样就可以把桔子的甜不甜数量化.

要了解一批桔子的甜度情况，你只需品尝一两个，然后通过这一两个桔子的甜度判断这批桔子的甜度. 这就是用个体推断总体.

为把上面的实际情况总结出来，需要引入一些术语.

A. 总体、个体和总体均值

在统计学中，我们把所要调查对象的全体叫做 **总体**(population), 把总体中的每个成员叫做 **个体** (individual).

总体中的个体总可以用数量表示. 为了叙述的简单和明确，我们把个体看成数量，把总体看成数量的集体. 我们要调查的是总体的性质.

总体中个体的数目有时是确定的，有时较难确定，但是往往并不影响总体

的确定，也不影响问题的解决．在判断一批桔子甜不甜时，你没有必要知道一共有多少个桔子.

总体平均 是总体的平均值，也称为 **总体均值**(mean). 在统计学中，常用 μ 表示总体均值．当总体含有 N 个个体，第 i 个个体是 y_i 时，总体均值

$$\mu = \frac{y_1 + y_2 + \cdots + y_N}{N}.$$

当 $y_1, y_2, \cdots, y_N$ 是总体的全部个体，μ 是总体均值时，称

$$\sigma^2 = \frac{(y_1-\mu)^2 + (y_2-\mu)^2 + \cdots + (y_N-\mu)^2}{N}$$

为 **总体方差** 或 **方差**(variance).

总体方差描述了总体中的个体向总体均值 μ 的集中程度．方差越小，个体向 μ 集中得越好．总体方差 σ^2 也描述了总体中个体的分散程度或波动幅度，方差越小，个体就越整齐.

总体标准差 是总体方差的算术平方根 $\sigma = \sqrt{\sigma^2}$, 简称为 **标准差**.

总体参数 是描述总体特性的指标，简称为 **参数**(parameter).

参数表示总体的特征，是要调查的指标．总体均值、总体方差、总体标准差都是参数．在讲到参数的时候，要明确它是哪个总体的参数.

B. 样本与估计

考虑某大学一年级 2000 个同学的平均身高 μ. 要得到这 2000 个同学的平均身高不是一件很困难的事情，只要了解了每个同学的身高就可以利用公式

$$\mu = \frac{\text{这 2000 个同学身高之和}}{2000}$$

计算得到.

但是在同一时刻要了解每个同学的准确身高也不是很容易的事情．如果让各班长在班上依次点名登记全班同学的身高，然后汇总，可能有些同学一时不能给出准确的回答，也可能有些同学受到其他同学的影响后，偏向于把自己的身高报高或报低．用这样的数据进行计算后得到的结果可能会产生偏差.

同一天对每个同学进行一次身高测量可以得到均值 μ 的准确值，但是要花费同学们较多的时间和精力. 统计上解决这类问题的最好方法是进行抽样调查，例如在 2000 个同学中只具体测量 50 个同学的身高，用这 50 个同学的平均身高

作为总体平均身高 μ 的近似. 这时我们称这 50 个同学的身高为总体的样本, 称 50 为样本量.

从总体中抽取一部分个体, 称这些个体为 **样本**(sample), 样本也叫做 **观测数据**(observation data).

称构成样本的个体数目为 **样本容量**, 简称为 **样本量**(sample size).

称从总体抽取样本的工作为 **抽样**(sampling).

在考虑身高问题时, 对于前述被选中的 50 个同学, 用 $x_1, x_2, \cdots, x_{50}$ 分别表示第 $1, 2, \cdots, 50$ 个同学在调查日的身高, 则这 50 个同学的身高

$$x_1, x_2, \cdots, x_{50}$$

是样本. 用 n 表示样本量, 则 $n = 50$.

样本均值 是样本的平均值, 用 $\overline{x}$ 表示.

给定 n 个观测数据 $x_1, x_2, \cdots, x_n$, 称

$$s^2 = \frac{1}{n-1}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2]$$

为这 n 个数据的 **样本方差**.

样本方差 s^2 是描述观测数据关于样本均值 $\overline{x}$ 分散程度的指标, 也是描述数据的分散程度或波动幅度的指标.

样本标准差 是样本方差的算术平方根 $s = \sqrt{s^2}$.

和总体均值 μ 相比较后知道, 只要抽样合理, 对于较大的样本量 n, 样本均值 $\overline{x}$ 会接近 μ. 于是, $\overline{x}$ 是总体均值 μ 的近似, 所以称为 μ 的 **估计**(estimator).

估计 是利用样本计算出的对参数的估计值. 估计能从观测数据直接计算出来.

对相同的观测数据, 不同的方法可以给出不同的估计结果, 所以估计不是唯一的. 这种不唯一性恰恰为统计学家们寻找更好的估计留下了余地.

实际问题中, 总体的容量往往是非常大的, 这时从数据本身无法看清总体的情况. 样本均值和样本方差可以提供必要的信息.

例 1.1 比赛中甲、乙两位射击运动员分别进行了 10 次射击, 成绩分别如下:

甲	9.5	9.9	9.9	9.9	9.8	9.7	9.5	9.3	9.6	9.6
乙	9.4	9.3	9.5	9.0	9.1	9.8	9.7	9.5	9.3	9.4

问哪个运动员平均水平高, 哪个运动员水平更稳定.

解 用 $\overline{x}$, s_x 和 $\overline{y}$, s_y 分别表示甲和乙成绩的样本均值和样本标准差，经过计算得到

$$\overline{x}=9.67,\ s_x=0.2058,\quad \overline{y}=9.4,\ s_y=0.2449.$$

甲的平均水平和稳定性都比乙好.

此题表明，知道样本标准差后，可以作出更好的比较结果.

§1.2 抽样调查

在日常生活中人们总是自觉或不自觉地应用抽样方法，例如在市场上买花生或瓜子时总要先尝几个看看是否饱满和新鲜；在烧菜的过程中经常要取一点尝尝味道.

在考察锅里汤的味道时，没有必要把汤喝完，只要把汤“搅拌均匀”，从中品尝一勺就可以了. 注意无论这锅汤有多多，只要一勺就够了.

记住上面的例子是大有好处的，因为它提供了抽样调查方法的最重要信息.

第一， 把汤“搅拌均匀”是说明抽样的随机性，没有抽样的随机性，样本就不能很好地反映总体的情况. 把刚加盐的地方舀出的汤作样本，你会作出汤太咸了的错误结论.

第二， 品尝一勺指出了选取的样本量不能太少，也不必太大. 太少了不足以品出味道，品尝一大碗也没有必要.

第三， “无论这锅汤有多多，只要一勺就够了”. 这里体现出抽样调查的如下基本性质：总体个数增大时，样本量不必跟着增大.

有人认为，总体数目很大时，样本量也必须跟着增大. 这种认识带有片面性.

实际的情况是这样的：在随机抽样下，一开始增加样本量会很快地增加估计的准确度，但是当样本量到达一定的时候，继续增加样本量效果就不明显了，再增加样本量就只是造成浪费了.

A. 抽样调查的必要性

抽样调查是相对于普查而言的，其含义是从总体中按一定的方式抽出样本进行考察，然后用样本的情况来推断总体的情况.

在评价 1000 个同型号的微波炉的平均工作寿命 μ 时，预备从中抽取 n 个进行工作寿命的测量试验，用这 n 个微波炉的平均工作寿命估计总体的平均工

作寿命 μ.

这里，总体是 1000 个微波炉的工作寿命，样本量是 n, 被选中的微波炉的工作寿命构成样本. 样本平均 $\bar{x}$ 是总体均值 μ 的估计.

在正确抽样的前提下，样本量越大，$\bar{x}$ 越接近总体均值 μ. 但是，较大的样本量造成的花费也很大，因为这 n 个微波炉做完寿命试验后就报废了. 在本问题中要想得到真正的总体均值 μ 是不可能的，除非把这 1000 个微波炉都拿来做工作寿命试验，报废掉这 1000 个微波炉.

在很多实际问题中，采用抽样的方法来确定总体性质不仅是必要的，也是必须的.

总体很大时，抽样调查往往可以提高调查的质量. 有人认为抽样调查不如全面调查得到的结论准确，这是不客观的. 看到抽样调查是用局部推断全体，带有抽样的误差，只是看到了问题的一个方面. 实际上调查数据的质量更重要，总体很大时进行全面调查，往往因为工作量过大、时间过长等而影响数据的质量. 一项经过科学设计并严格实施的抽样调查可能得到比全面调查更可靠的结果.

B. 随机抽样

如果总体中的每个个体都有相同的机会被抽中，就称这样的抽样方法为**随机抽样**方法. 人们经常用"**任取**"、"**随机抽取**"或"**等可能抽取**"等来表示随机抽样.

从概率论的知识知道，如果从总体中任选一个个体，这个个体是随机变量，这个随机变量的数学期望是总体均值，方差是总体方差.

随机抽样又分为无放回的随机抽样和有放回的随机抽样. 无放回的随机抽样指在总体中随机抽出一个个体后，下次在余下的个体中再进行随机抽样. 有放回的随机抽样指抽出一个个体，记录下抽到的结果后放回，摇匀后再进行下一次随机抽样.

例 2.1 设 N 件产品中有 M 件次品，N, M 都是未知的. 估计这批产品的次品率 $p = M/N$.

解 无放回地从中依次取 n 件，用 Y 表示取得的次品数，则 $Y \sim H(N, M, n)$, 按照附录 B, 有

$$\mathrm{E}Y = np, \quad \mathrm{Var}\,(Y) = np(1-p)\frac{N-n}{N-1}.$$

用样本次品率 $\hat{p} = Y/n$ 估计 p 时，有

$$\mathrm{E}\hat{p} = p, \quad \mathrm{Var}\,(\hat{p}) = \frac{1}{n}p(1-p)\frac{N-n}{N-1}. \tag{2.1}$$

如果采用有放回的随机抽样，用 X 表示取得的次品数，则 $X \sim \mathcal{B}(n,p)$, 这时有

$$\mathrm{E}X = np, \quad \mathrm{Var}\,(X) = np(1-p).$$

用这时的样本次品率 $\tilde{p} = X/n$ 估计 p 时，有

$$\mathrm{E}\tilde{p} = p, \quad \mathrm{Var}\,(\tilde{p}) = \frac{1}{n}p(1-p). \tag{2.2}$$

$\mathrm{E}\hat{p} = \mathrm{E}\tilde{p} = p$, 说明这两种方法都是较好的估计方法，没有系统偏差.

由于方差 $\mathrm{E}(\hat{p}-p)^2$ 描述的是 $\hat{p}$ 向真实参数 p 的集中程度，因而是描述估计精度的量. 方差越小，说明估计的精度越高. $\mathrm{Var}\,(\hat{p}) < \mathrm{Var}\,(\tilde{p})$ 说明无放回随机抽样的估计精度好于有放回随机抽样的估计精度. 但是当 N 比 n 大很多时，$(N-n)/(N-1)$ 接近于 1, 说明两种抽样方法差别不大.

另外 $\mathrm{Var}\,(\tilde{p})$ 与 N 无关，说明要达到一定的估计精度，只需要适当地增加 n. 并不是说总体数目 N 越大，就需要多抽样. 无放回随机抽样下的情况也是类似的，因为实际问题中 N 通常都很大，而相比之下 n 较小.

在相同的总体中和相同的样本量下，无放回随机抽样得到的结果比有放回的随机抽样得到的结果要好. 但是当总体的数量很大，样本量相对总体的数量又很小时，这两种抽样方法得到的结果是相近的.

试验和理论都证明：在随机抽样下，样本均值 $\overline{x}$ 是总体均值 μ 很好的估计，样本标准差 s 是总体标准差 σ 很好的估计. 在样本量不大时，增加样本量可以比较好地提高估计的精确度.

考虑某大学一年级 2000 个同学的平均身高 μ 时，需要调查 50 同学的身高. 实现无放回的随机抽样的方法是先将 2000 个同学的学号分别写在 2000 张小纸片上，然后放入一个大纸箱进行充分地摇匀，最后从纸箱中无放回地抽取 50 个纸片，纸片上的学号就是被选中的同学的学号.

C. 随机抽样的无偏性

样本均值是对总体均值的估计. 在总体中任取一个个体 X, X 是随机变量，从数学期望的定义知道 $\mathrm{E}X = \mu$ 是总体均值. 这说明随机抽样是无偏的. 如果用 $X_1, X_2, \cdots, X_n$ 表示依次随机抽取的样本，则样本均值

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

是总体均值 μ 的估计. 下面证明 $\mathrm{E}\overline{X} = \mu$. 在有放回的随机抽样下，$X_1, X_2, \cdots$,

X_n 有相同的数学期望 μ, 于是有

$$\mathrm{E}\overline{X} = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}X_j = \frac{1}{n}\sum_{j=1}^{n}\mu = \mu.$$

在无放回的随机抽样下，根据抽签的原理 (参考 [6]), 第 j 次抽到每个个体的概率是相同的，所以 X_1 和 X_j 是同分布的，因而有相同的数学期望，于是也有 $\mathrm{E}\overline{X} = \mu$. 说明在随机抽样下，样本均值 $\overline{X}$ 总是总体均值 μ 的无偏估计.

下面是没有采取正确的抽样方案导致调查结论严重失真的著名案例.

例 2.2 1936 年是美国总统选举年. 这年罗斯福 (Roosevelt) 任美国总统期满，参加第二届的连任竞选，对手是堪萨斯州州长兰登 (Landon). 当时美国刚从经济大萧条中恢复过来，失业人数仍高达 900 多万，人们的经济收入下降了三分之一后开始逐步回升. 当时，观察家们普遍认为罗斯福会当选. 而美国的《文学摘要》杂志的调查却预测兰登会以 57% 对 43% 的压倒优势获胜.

《文学摘要》的预测是基于对 240 万选民的民意调查得出的. 自 1916 年以来，在历届美国总统的选举中《文学摘要》都做了正确的预测. 《文学摘要》的威信有力地支持着它的这次预测.

但是选举的结果是罗斯福以 62% 对 38% 的压倒优势获胜，此后不久《文学摘要》杂志就破产了.

要了解《文学摘要》预测失败的原因就必须检查他们的抽样调查方案.《文学摘要》是将问卷寄给了 1 000 万个选民，基于收回的 240 万份问卷得出的判断. 这些选民的地址是在诸如电话簿、俱乐部会员名单等上查到的.

分析：1936 年只有大约四分之一的家庭安装了电话. 由于有钱人才更有可能安装家庭电话和参加俱乐部，所以《文学摘要》的调查方案漏掉了那些不属于俱乐部的穷人和没有安装电话的穷人，这就导致了调查结果有排除穷人的偏向.

在 1936 年，由于经济开始好转，穷人普遍有赞同罗斯福当选的倾向，富人有赞同兰登当选的倾向. 《文学摘要》的调查结果更多地代表了富人的意愿，导致了预测的失败.

抽样的方案应当公平地对待每一位选民和每一个群体，以便得到选民的真实情况. 将哪一个群体排除在外的抽样方案都可能导致有偏的样本，从而导致错误的结论.