# 数理统计

主讲教师: 冯伟

办公地点:北航国实二期E604-6

wfeng_323@buaa.edu.cn

# 数理统计

- 先修课程：高等数学 线性代数 概率统计
- 教学课时：48学时
- 教学目的：使得学生在掌握概率统计的基础知识的前提下，进一步拓展概率统计的知识面，并结合动手能力使得能够用所学知识来解决更多的实际问题。

要求和学习方法:

- 要求按时上课，遵守课堂纪律，保持安静，不影响大家听讲;
- 平时成绩占总成绩的40%.
- 学习中遇到问题解决方法：群讨论、发邮件、课前课后答疑

# 参考书目

上课教材：孙海燕，周梦，李卫国，冯伟，数理统计，北航出版社，2016.


陈希孺，*数理统计引论*，科学出版社，1997（较难）

范金城，吴可法，*统计推断导引*，科学出版社，2001.

张忠占，谢田法，杨振海，*应用数理统计*，高等教育出版社，2011.

赵选民等，*数理统计*，科学出版社，2002

汪荣鑫，*数理统计*，西安交通大学出版社，1987

吴翊等，*应用数理统计*，国防科技大学出版社，1995

李俊德，*应用数理统计方法*，河南科技大学出版社，1985

P.J.比克尔等, 李泽慧等译,陈希孺校, *数理统计—基本概念及专题*, 兰州大学出版社, 2004.

张尧庭,方开泰, *多元统计分析引论*,科学出版社, 2003

苔淑彩等, *应用数理统计*,武汉大学出版社, 2005.

茆诗松,王静龙, *高等数理统计学*, 高等教育出版社, 2006

C.R.Rao , *线性统计推断及应用*,北京：科学出版社, 1987

(Linear statistical inference and its applications )

# 成长为什么样的人

年轻人都渴望成长。但是，究竟成长为什么人呢？应该说，许多人希望成长为一个有内涵的人。身处逆境，接受锻炼，克服困难，乐在其中。身处顺境，心忧天下，兼济苍生，乐在其中。这样的人有着与众不同的人格特征，不但拥有常人所有的知识，而且还拥有常人所少有的智慧，有强烈的责任感，坚强的意志，崇高的道义追求。提炼一下就是五个关键词汇：知识，智慧，责任，意志，道义。

# 成长为什么样的人

知识--实在是太多了，且日常生活中最重要、最基本的东西都是知识无法解决的

为什么知识处理不了这些人生与社会最深刻的问题？因为知识的定义就限制了其功能。知识是分门别类的、高度精确的、标准化的、可复制的、不以人的意志为转移的，因而知识也是静态的、孤立的、无层次的。

# 成长为什么样的人

智慧--人生和社会都是时刻变化的、相互联系的、层次丰富的、以人的意志为转移的，因此，认识和处理人生、社会，知识就只能起基础性、辅助性作用，智慧将起主导性、实质性作用

即使学习知识，也要融会贯通，举一反三，即运用联想、类比、想象等“不精确、不科学、不可靠”的智慧型的思维方式。

# 成长为什么样的人

责任 -- 智慧怎么样才能开启？智慧一定是为问题而开启。你心中有问题你就会想办法解决它。比如说发现马路上噪音过大，你发明一个隔音屏，这个过程也需要运用智慧，没有标准答案。这个过程需要智慧，也自然会产生智慧。不要怕自己没有文凭，或者没有专业知识，其实，只要给时间，谦虚学习，认真钻研，任何人间奇迹都可以创造出来。

# 成长为什么样的人

•意志 --单有责任感可能还不够。因为，这个世界上无数重大的难题都不是你有责任感就可以解决的。你试一次，失败一次；再试一次，再失败一次。有些人就会说，这些责任应该政府去负担，应该科学家去解决。就你一个普通人，怎么能够担起如此巨大的责任呢？那么，怎么去让自己承担起重大、困难而持久的责任呢？这就一定要靠意志。

# 成长为什么样的人

道义 --最后还要具备的一个东西是道义。为什么要有道义？如果一个人光有前面四个素质，但是，没有道义会怎么样？这个人他可以成功，但是生活在一个没有道义的世界里，他会惴惴不安，没有安全感。

道义这个东西，比意志还要看不见摸不着，但是它的力量比意志更大。因为意志只能成就你一个人，道义能够成就整个世界。

# 成长为什么样的人

## 知识，智慧，责任，意志，道义

把这五种品质联系起来，我们就成了知识开阔、智慧丰富、责任感强、意志坚强和道义高尚的人，就成了复杂而善良的人，就成了能干而幸福的人。这样的人，是真正的善意的精英。

通过有招学习无招，以无招胜有招

# 学习与研究漫谈

- 要不断进取,不要满足已取得的成就.

  经常地,更大的成就还在等着你呢!

- 人只有主动学习，才可能有一点点成绩，一切全等着别人逼着干，不会有什么出息。

没有什么“他学成材”，凡成材者均为自学成材。（虽有点偏激，但积极意义很明显.)

# 学习与研究

- 一定要边学习、边研究，以学习的态度研究，以研究的态度学习。
- 一定不要忽视一些小问题，要想思维清晰，必须能把大问题拆成小问题，并且精通小问题。

以有招学习无招，最终无招胜有招！

# 如何学习本课

- 学习本课离不了概念，公式，定理
- 希望我们不是以“奉天承运皇帝诏曰”的方式从天而降这些概念和结论，而是从问题出发，在尝试解决问题的过程中将所需的东西“发明”出来。

# 如何学习本课

传统的数学教学过程从一些基本的概念或定义出发，以简练的方式合乎逻辑地推演出所要求的结论，固然可以使学生在较短的时间内按部就班地学到尽可能多的内容，并体会到一种丝丝入扣、天衣无缝的美感；但是，过分强调这一点，就可能使学生误认为数学的完美无缺、无懈可击是与生俱来、天经地义的。

# 如何学习本课

其实，现在看来美不胜收的一些重要的数学理论和方法，在一开始往往是混乱粗糙、难以理解甚至不可思议的，经过许多乃至几代数学家的努力，有时甚至经过长期的激烈论争，才逐步去粗取精、去伪存真，最终才出现了现在为大家公认的系统的理论。

# 第一章 数理统计初步

经典的数理统计是以概率统计为基础的，概率论与数理统计的关系极为密切，为以后学习的方便，这里对概率论做一个简单的回顾。

- 概率论的回顾
- 数理统计简介
- 数据初步处理

# 第一节 概率论的回顾

- 确定性现象和不确定性现象、随机现象
- 概率论：研究和揭示随机现象的统计规律性的一个数学分支
- 随机事件的定义、运算及运算律
- 古典概型、几何概型、统计概型的定义及性质
- 概率的公理化定义
- 条件概率,乘法公式,全概率公式,贝叶斯公式(逆概率公式、后验公式），独立性

- 随机变量、分布函数及其性质
- 随机变量函数的分布
- 多维随机变量及其分布（以二维为例）
- 随机变量的数字特征：期望、方差、协方差、相关系数及其性质（用的很多）
- 特征函数（遇到会讲，但最好先了解一些）

# 大数定律与中心极限定理

这部分内容是整个概率论与数理统计的基础

- 依概率收敛、概率1收敛（还可了解*r*阶收敛，依分布收敛及相互之间的关系）
- 什么是大数定律（弱、强大数定律）
- 关于大数定律的几个基本定理
- 什么是中心极限定理
- 中心极限定理的应用

# 第二节 数理统计概况

- 什么是数理统计学
- 数理统计学的内容
- 数理统计方法的应用
- 数理统计学发展简史
- 数理统计学的基本概念

统计学从方法上讲有两大类:描述性统计方法和数理统计方法(即抽样统计方法).

- **描述性统计方法**: *全部资料*
- **数理统计方法**: *部分资料*

  抽样统计方法

反腐败、食品安全 .....

**定义：数理统计学**是数学的一个分支，它研究怎样用有效的方法去收集、整理、分析带随机影响的数据，并在此基础上对所讨论的问题给出统计性的估计和推断。

# 数理统计学的内容

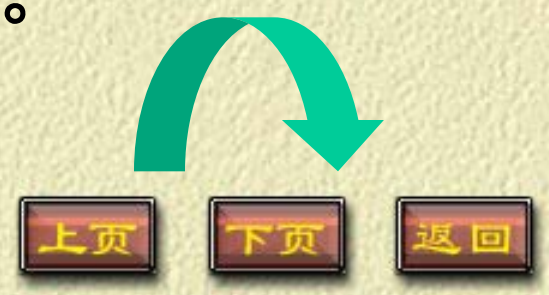
概括为两大类

- 用有效的方法去收集数据。

  抽样理论和试验设计

- 有效地使用数据。中心内容—统计推断

  它包括参数估计，假设检验，回归分析，方差分析，多元统计分析等等。



- 参数估计与假设检验
- 回归分析
- 方差分析与正交试验设计
- 多元正态分析
  - 参数估计和检验
  - 判别分析
  - 主成分分析
  - 因子分析
  - 典型相关分析

抽样方式

- 简单随机抽样
- 分层抽样
- 等距抽样
- 整群抽样
- 多阶段抽样

# 有效性的含义

上述有效性有两个含义：

- ✓可以建立一个在数学上便于处理的模型来描述所得的数据，
- ✓数据中要包含尽可能多的与所研究的问题有关的信息。

# 关于统计推断

- 由于统计推断中使用的仅仅是部分数据，且带有随机性，故所得结论只能做到尽可能而非绝对的精确可靠，而结论的正确性程度显然可以用概率来度量，因此概率论是数理统计的基础。
- 统计方法的具体使用并不需要很高深的数学知识，但不具备较多较深的数学知识，这些方法的理论依据就说不清楚。本课主要介绍数理统计方法，也给出一些必要的数学推导，但不追求其严密性和完整性。

# 数理统计方法的应用

几乎在人类活动的一切领域中都能够不同程度地发现数理统计方法的应用。

- 实验数据的处理离不开数理统计方法
- 在工农业生产中，最佳生产工艺的安排，最佳配方的确定，优良品种的对比试验，产品质量的控制管理，产品验收方案的制定，电子元器件寿命的计算等都要用到数理统计方法。

- 在医药卫生领域，流行病的研究、新药的药效试验以及某种疾病的发病率与其它因素的关系的研究都是数理统计方法的用武之地。
- 在生物遗传学、气象预报、地震研究、地质探矿等方面的研究中，数理统计方法是必备工具之一。
- 数理统计方法在社会科学方面的应用也愈来愈广泛，教育学，人口学，社会保险业，各种社会问题的抽样调查，市场预测，民意测验等都有数理统计方法涉足。

总之，只要安排试验和处理数据，就可以用数理统计方法。

# 数理统计学发展简史

- 统计学的起源：统计学起源于古代，早在公元前3050年的古埃及就为建造金字塔进行过全国国力统计。到了16世纪，西欧各国政府对收集公民有关资料发生兴趣。Statistics（统计学）源于State.
- 数理统计的正式诞生。在数学家建立了概率论后，才奠定了数理统计发展的理论基础。一般认为，它诞生于19世纪后期。

# 数理统计学发展简史（续）

- 19世纪后期到20世纪四十年代。在这时期，英国人高尔顿、皮尔逊、费歇等作了大量开创性工作。尤其，费歇于1922年的一篇论文是数理统计学建立过程的一个里程碑，该文主要观点至今仍基本有效。到了四十年代，数理统计学已发展成为一个*成熟的数学分支*，它的重要标志是瑞典统计学家H.Cramer于1949年的著作《Mathematical Methods of Statistics》

# 数理统计学发展简史（续）

- 二战后。这时期的一个突出特点是计算机的发明和使用。它使人们能够处理大量的数据及其运算，把数理统计的研究引入到宏观世界和微观世界，又出现了一些新的分支。

最后，特别提一下我国的许宝禄教授在极限理论、马氏过程、多元分析、正交设计、过程设计和判别函数等许多方面都有突出的贡献，他的许多研究成果都达到了世界先进水平。

# 数理统计学的基本概念

- 总体与个体
- 抽样、简单随机抽样
- 样本、简单随机样本与样本空间
- 分布族、参数空间
- 统计量与样本矩

**例** 要研究一批灯泡的寿命，这一批灯泡就是一个总体，每个灯泡就是一个个体。但我们只关心其寿命，显然是一个随机变量，记为$X$. 我们称$X$为一个总体，它的每一个取值为一个个体。如果我们对总体一点也不了解，我们可以记其分布函数为$F(x)$,然后利用已知数据来作非参数假设检验。如果已知其分布形式，只是不确定参数，我们可以记其为$\{f_\theta(x)\}$, 这里用$\theta$记总体的分布函数中的未知参数(可以是向量)，未知参数的全部可容许值组成的集合称为参数空间，记为$\Theta$。

今后我们用$f_\theta(x)$表示总体的分布，用$\{f_\theta(x)\}$表示总体所在的分布族。分布族的假定反映了我们对所研究的问题以及抽样方式的知识和规律的了解程度。

# 第三节 数据初步处理

★列表法—频率分布表

★图象法—频率分布图

★样本统计分析法—计算样本平均数,样本方差,样本标准差,中位数,众数

# 一 频率分布表

| 时间 | 含锰百分比量 | 时间 | 含锰百分比量 |
|---|---|---|---|
| 4.1 | 1.40 1.28 1.36 1.38 1.44 | 4.14 | 1.40 1.34 1.54 1.44 1.46 |
| 4.2 | 1.80 1.44 1.46 1.50 1.38 | 4.15 | 1.54 1.50 1.48 1.52 1.58 |
| 4.3 | 1.52 1.46 1.42 1.58 1.70 | 4.16 | 1.62 1.58 1.62 1.76 1.68 |
| 4.4 | 1.68 1.66 1.62 1.72 1.60 | 4.17 | 1.62 1.46 1.38 1.42 1.38 |
| 4.5 | 1.60 1.44 1.46 1.38 1.34 | 4.18 | 1.38 1.34 1.36 1.58 1.38 |
| 4.6 | 1.34 1.28 1.08 1.08 1.36 | 4.19 | 1.50 1.46 1.28 1.18 1.28 |
| 4.7 | 1.26 1.50 1.52 1.38 1.50 | 4.20 | 1.52 1.50 1.46 1.34 1.40 |
| 4.8 | 1.50 1.42 1.38 1.36 1.38 | 4.21 | 1.42 1.34 1.48 1.36 1.38 |
| 4.9 | 1.32 1.40 1.40 1.26 1.26 | 4.22 | 1.16 1.34 1.40 1.16 1.54 |
| 4.10 | 1.24 1.22 1.20 1.30 1.36 | 4.23 | 1.30 1.48 1.28 1.18 1.28 |
| 4.11 | 1.30 1.52 1.76 1.16 1.28 | 4.24 | 1.48 1.46 1.48 1.42 1.36 |
| 4.12 | 1.32 1.22 1.72 1.18 1.36 | 4.25 | 1.44 1.28 1.10 1.06 1.10 |
| 4.13 | 1.16 1.22 1.24 1.22 1.34 | | |

表一

上页 下页 返回

# 一 频率分布表(续)

**例** 有一鼓风炼铁炉,每天生产5炉生铁(生产数据见上表). 试研究每炉生铁中,锰的百分比含量的统计规律性.

**解**: 从表一的125个数中首先找出最小数1.06%和最大数1.80%,令$S$=1.80%-1.06%,再将代表S的线段均分成9个小区间,从而确定每组上下限.再仔细地一一观察125个数据,确定落在每组中的数据个数, 可得下表2.

# 一 频率分布表

| 组上下限(%) | 组中值 | 组频数 | 组频率 | 累计频率 |
|---|---|---|---|---|
| 0.99-1.09 | 1.04 | 3 | 0.024 | 0.024 |
| 1.09-1.19 | 1.14 | 9 | 0.072 | 0.096 |
| 1.19-1.29 | 1.24 | 18 | 0.144 | 0.24 |
| 1.29-1.39 | 1.34 | 32 | 0.256 | 0.496 |
| 1.39-1.49 | 1.44 | 29 | 0.232 | 0.728 |
| 1.49-1.59 | 1.54 | 19 | 0.152 | 0.880 |
| 1.59-1.69 | 1.64 | 9 | 0.072 | 0.952 |
| 1.69-1.79 | 1.74 | 5 | 0.04 | 0.952 |
| 1.79-1.89 | 1.84 | 1 | 0.008 | 1 |
| 合计 | | 125 | 1 | |

上页 下页 返回

注: 依序对一批数据分组后,各组所含数据个数称为组频数,各组组频数之和称为总频数(即样本容量$n$),各组的组频数去除总频数,所得之商称为该组的组频率,组中值=组下限+组上限.

# 一 频率分布表（续）

频率分布表的一般制作步骤

- 找出该批数据的最大最小值；
- 决定组距和分组数目.组数太少，估计效果较差，多于20时，会增加计算上的麻烦；
- 决定分点.第一组一定要包含最小值.最后一组要包含最大值；
- 数出每组的组频数(左闭右开)；
- 计算出每组的组中值、组频数（率）、累计频率

# 二 频率分布图

- 从某种意义上讲，频率分布图是频率分布表的几何表示。
- 常用的频率分布图有下列两种：
  - **频率分布直方图**（以各组距为底边，以对应的组频为高，做长方形即得）
  - **频率分布多边形**（以各组的组中值为横坐标，纵坐标仍是各组的组频率，就得到频率多边形的各个定点，依次连接即得）
- 频率分布直方图和频率分布多边形可用来估计总体的概率分布密度曲线。茎叶图(stem-and-leaf display)与盒子图(box-plot)也较有用.

# 三 样本统计分析法

为了掌握一批数据的主要特性，还需要计算这批数据的一些数字特征。这里介绍平均数，中位数，众数，方差，均方差。其中前三个表示数据集中趋势，后两个反应数据的离散程度。

下面假设$(x_1,x_2,\ldots,x_n)$为来自总体$X$的样本。

平均数： $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

中位数：把样本按大小重新排列成：$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$，则排在正中间的一个 $x_{(k)}$ 就称为这组数据的中位数，记作$Me$.

当$n$为正奇数时，$Me=x_{(\frac{n+1}{2})}$

当$n$为偶数时，$Me=\frac{1}{2}\left[x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right]$

**众数**：在频数分布中出现次数最多的数据，或者说是具有最大频率的数据，用符号$M_0$表示。在频率分布图上很容易看出。

对于单峰曲线，曲线最高点的横坐标就是众数。对于双峰曲线，有时认为每个高峰的横坐标都是众数，这样的频率分布就有两个众数；有时取这两个峰中最高的一个的横坐标为众数。

求众数的方法包括：作图法和插入法

**极差**：$R=\max\{x_i\}-\min\{x_i\}$

极差$R$计算简便，所以常用极差作为检验产品质量离散程度的指标。

**样本方差**：$S^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

**样本均方差**：样本方差的算术平方根$S$称为样本均方差。