著名统计学家C.R.Rao(2023.8.23去世，享年102岁)：

“在终极的分析中，一切知识都是历史；在抽象的意义下，一切科学都是数学；在理性的世界里，所有的判断都是统计学。”

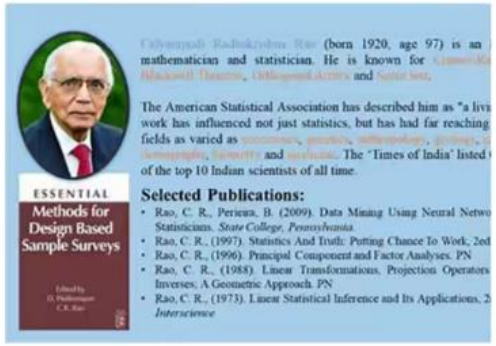
https://cosx.org/2021/08/a-century-in-statistical-science/

“统计学”一词最早来源于现代拉丁文statisticum collegium（国会）。那时，亚里士多德写了150多种纪要，这些纪要被称为“城邦纪要”，其内容包括各城邦的历史、行政、科学、艺术、人口、资源和财富等社会和经济情况的比较分析，具有社会科学的特点。

到了16世纪，意大利语用statista来称呼和政府相关的政治家；接着，德国人戈特弗里德·阿亨瓦尔开始使用statistik一词来表示对国家资料进行分析的学问；1785年，在法语中出现“统计”一词，写为statistique；1807年，丹麦语也引入statistik作为统计的名称；最终演化为现如今的“统计学”（Statistics），依然保留了城邦（state）这个词根。

C-R不等式, Rao-Blackwell定理

- 2021年8月，统计学知名期刊《International Statistical Review》发表了 Nandini Kannan 和 Debasis Kundu 撰写的《C. Radhakrishna Rao: A Century in Statistical Science》.
- 名言：在终极的分析中，一切知识都是历史；在抽象的意义下，一切科学都是数学；在理性的世界里，所有的判断都是统计学。
- 百岁老人仍然活跃在科研前线，PNAS上搜到他5篇文章
  - PNAS 117 (10) 5235-5241, https://doi.org/10.1073/pnas.1917411117, 2020.
  - PNAS 115 (23) 5914-5919, https://doi.org/10.1073/pnas.1804649115, 2018.
  - PNAS 114 (15) 3873-3878, https://doi.org/10.1073/pnas.1702654114, 2017.
  - PNAS 113 (18) 4958-4963, https://doi.org/10.1073/pnas.1604553113, 2016.
  - PNAS 111 (44) 15681-15686, https://doi.org/10.1073/pnas.141221611, 2014.
- 思政要素：学习其为事业奋斗终生的精神



https://cosx.org/2021/08/a-century-in-statistical-science/

统计学发展

19世纪初---20世纪初

不断建立和完善了统计学的理论体系并逐渐形成了以统计推断为主要内容的“数理统计学”和以描述为主的“社会计学派”。

20世纪中后期.

进化论和能量守恒定律的出现促进了描述统计学的完善，是描述统计学的顶峰。20世纪20年代以后，在细胞学的发展推动下，统计推断迈入了新阶段，直到50年代推断统计学的最迅速的发展时期，50年代在经过几代大师的努力下，数理统计学的基本框架已经建成，并逐步成为了20世纪的主流统计学。

近期

受计算机和新兴科学的影响，使得统计学越来越依靠于计算机技术和数值计算方法。而且随着大数据时代的到来，统计学针对大数据的特征，以服务和满足各个领域的需求，正在不断完善、创新、发展数据分析的新方法和理论。

推荐读物：
Computer Age Statistical Inference，
作者：**Bradley Efron and Trevor Hastie，**
出版社：**Cambridge University Press**

什么是统计学？

统计学是关于数据的科学, 统计学可分为描述统计学和推断统计学。

1. 收集数据：取得数据。
2. 处理数据：整理与图表展示。
3. 分析数据：利用统计方法和模型分析数据。
4. 检验和诊断：对方法和模型的理论假设做相应的检验和诊断。
5. 结论与决策：从数据分析中得出客观结论。

统计学(Statistics)

统计学分类

描述统计(descriptive statistics)

描述性统计是组织和总结数据(样本)的方法，即对数据进行描述，并找出数据的基本规律。

表格或图表用于组织数据，描述性的值(如平均值、标准差、中位数和分位数等)用于总结数据。

推断统计(inferential statistics)

研究如何利用样本数据来推断总体分布特征。

通过估计(矩估计、极大似然估计、最小二乘估计、核密度估计)、假设检验(*t*检验、*F*检验等)或方差分析、回归分析、生存分析等对总体的分布或条件分布进行推断。

统计学(Statistics)

常用的统计软件(Software)

A. **SPSS**(商业软件，下拉菜单式，社会统计，医学统计等应用领域).

B. **SAS**(商业软件，功能强大、分析全面，费用高).

C. **R语言**(开源软件，方便，功能齐全).

D. **Python**(开源软件，功能强大，文本挖掘和机器学习效率高)

E. **Matlab**(商业软件，编程容易，矩阵计算效率高，图形功能强)

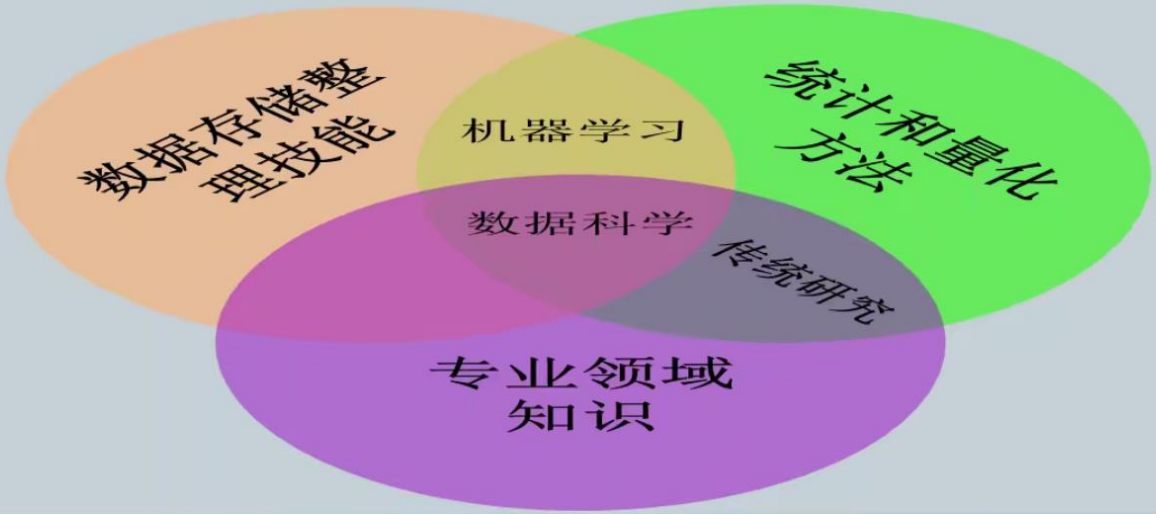
现在统计学(Statistics)的发展

统计学与数据科学的角色

现在人们似乎已经有一个共识：

一个理解数据科学的方式是认为它是在跨学科领域如商业分析结合了计算机科学、建模、统计、优化和数学的革命性的一步

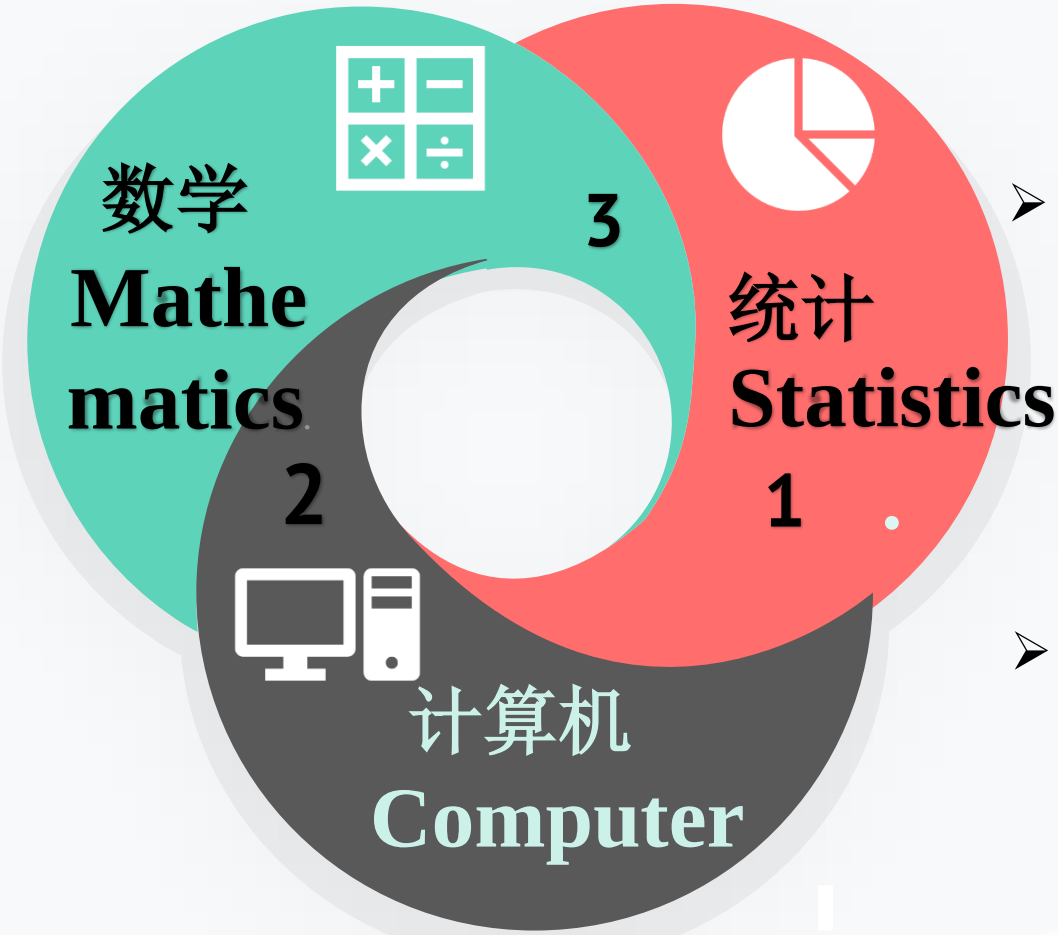
数据科学维恩图



现在统计学(Statistics)的发展

统计学(Statistics)

统计+计算机+数学=？？？



区别与联系

- 数学以公理体系为基础；统计以数据分析为基础。数学研究的是抽象的对象；统计研究的是实际的对象。
- 统计学的基础是数学，如何基于有限的样本数据推断总体的特征正是靠数学牢固建立起来的。然而，在海量数据的信息时代，脱离计算机的统计分析几乎是不可能的。
- 理解统计思想，掌握不同统计方法在计算机中的实现过程，基于数理统计方法和模型正确解读计算机输出的结果，才更有利于推动统计学在各领域的广泛应用。

统计学与大数据(Big Data)

- 大数据(Big Data)，又称为巨量资料，也可以定义为来自各种来源的大量非结构化或结构化数据(大数据就是一切可记录信号的集合)。因此，大数据通常包含的数据大小超出传统软件在可接受的时间内处理的能力。

- 由于计算机的计算以及存储技术的进步，发布新数据的便捷性以及全球大多数政府对高透明度的要求，大数据分析在现代研究中越来越突出。

- 在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》 中大数据指不用随机分析法(抽样调查)这样捷径，而采用所有数据进行分析处理。大数据的5V特点(IBM提出)：Volume(大量)、Velocity(增速高)、Variety(多样)、Value(低价值密度)、Veracity(真实性)。

现在统计学(Statistics)的发展

统计学与大数据(Big Data)

- 由于大数据的出现使统计学更加引人注目。理由很简单，就是利用统计学方法分析大数据，在计划经营战略，市场战略，开发新产品，新业务发展等都取得了有效成果。经营不只是靠感觉，靠经验，靠勇气的东西了，而是根据以数据为基础的分析方法来进行科学决策。

- 近些年，由于信息技术的发展迅速，通过分析大量数据有助于企业的经营，从而统计学得到了人们的注目。

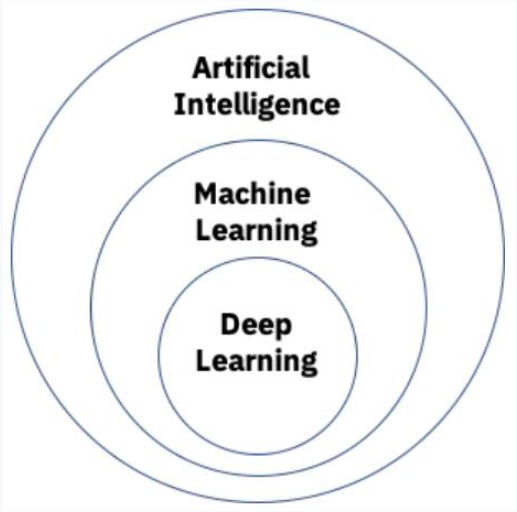
- 统计学注重的是方式方法，而大数据则更关注于整个数据价值化的过程，大数据不仅需要统计学知识，还需要具备数学知识和计算机知识。从另一个角度来说，统计学为大数据进行数据价值化奠定了一定的基础。

- 从体系结构来看，统计学可应用于大数据分析领域，统计学是大数据分析的两种主要途径之一，另一种途径则是机器学习。

第 统计学与人工智能(AI) 本与统计量

人工智能是结合了计算机科学和强大数据集的领域，能够实现问题解决。 它还包括机器学习和深度学习等子领域，这些子领域经常与人工智能一起提及。 这些学科由 AI 算法组成，这些算法旨在创建基于输入数据进行预测或分类的专家系统。

人工智能：语音识别(如Siri等)、客户服务(如在线聊天机器人等);
计算机视觉(如自动驾驶汽车等)、推荐引擎(如淘宝等)、自动股票交易等。



推荐书籍：《数学之美》
作者：吴军博士，他是当前谷歌中日韩文搜索算法的主要设计者。2010年至2012年，吴军加盟腾讯公司，出任负责搜索和搜索广告的副总裁，同时担任国家重大专项“新一代搜索引擎和浏览器”项目的总负责人。2012年回到谷歌，负责开发了谷歌自动回答系统。

书中关键词：条件概率，贝叶斯公式，隐含马尔科夫链，逻辑回归...

统计学与机器学习

统计学专家Aleks Jakulin:“机器学习是人工智能领域人员做数据分析，数据挖掘是数据库领域人员做数据分析，统计学习是统计学领域人员做数据分析”

严谨地讲，混淆机器学习和统计学习两个概念会被认为是一种过于简单的表述，不太合理。主流更倾向于接受，机器学习方法是建立在统计学习基础之上的。

统计学(Statistics)的应用

- 农作物的产量与施肥量之间是否真的存在相关关系？
- 金融市场的波动是否存在一定规律？能否用这种规律对未来波动率进行预测？
- 一个网购者，如果JD知道他的年龄，消费历史，商品浏览时间或点击率，能否找到更有效的产品推介方案？
- 某银行能否根据申请贷款人的信用卡消费还款信息，收入水平，工作性质，学历程度及其他特征评估他的违约概率？

第六章　数理统计的基本概念

§1　总体、样本与统计量

一、引言

数理统计是以概率论为理论基础(姐妹学科),研究

1) 如何以有效的方式收集和整理受随机因素影响的数据;

2) 如何合理地分析这些数据从而作出科学的推断(称为统计推断).

这两部分有密切联系, 实际应用中更应前后兼顾. 将主要介绍统计推断方面的内容.

二、总体与个体

总体： 研究对象的全体 (或母体).

个体： 组成总体的每个基本元素(或对象).

- 例如, 要考察本校男生的身体情况, 则将本校的所有男生视为一个总体, 而每一位男生就是一个个体.
- 又如, 考察某厂生产的电子元器件的质量, 将全部产品视为总体, 每一个元器件即为一个个体.

通常研究的是总体的一项或几项数量指标.

- 如仅考虑男生的身高X数量指标, 不考虑体重, 视力等;
- 如关心电子元件的寿命Y数量指标, 不考虑重量等;

总体与个体：

总体：研究对象的全体所组成的集合。

个体： 组成总体的每个基本元素。

通常研究的是总体的一项或几项数量指标.

- 如仅考虑男生的身高X数量指标, 不考虑体重, 视力等;
- 如关心电子元件的寿命Y数量指标, 不考虑重量等;

个体x_i :第i个男生的身高 或元件寿命(数量指标的值)

总体X:$\{x_1, x_2, \ldots\}$(数量指标的值的全体)

能否将（实际）总体和随机变量 X 等同起来?

个体x_i: 第i个男生的身高 或元件寿命(数量指标的值)

总体X:$\{x_1, x_2, \ldots\}$(数量指标的值的全体)

男生身高X服从正态分布, 电子元件寿命Y服从指数分布

将刻画总体的数量指标X看作一个随机变量,

即把随机变量取值的全体当作总体,

把每个个体的数值看成X的一个观测值.

以后将(实际)总体和随机变量X等同起来

(可以借助随机变量来研究总体)

所谓总体分布是指随机变量 X 的分布.

思考:人寿命与电子元件寿命两总体都服从指数分布,

两总体关系?分布类型相同, 统计中可视为同一类总体.

总 体 是 随 机 变 量

三、样本(Sample)

一般, 从总体中抽取一部分(取n个)进行观测, 再依据这n个个体的试验(或观察)的结果去推断总体的性质.

样本：按照一定的规则从总体中抽取的一部分个体.

“按照一定规则”：

采取措施使每个个体抽取的机会均等.

抽样：抽取样本的过程.

样本容量：样本中个体的数目n.

将第 i 个个体的对应指标记为 X_i, $i=1,2,\ldots,n$, 构成随机向量: $(X_1, X_2, \cdots, X_n)$.

容量为n的一组样本：$(X_1, X_2, \cdots, X_n)$

•样本是一组随机变量，其具体试验(观察)数值记为：

$x_1, x_2, \cdots, x_n$，称为样本观测值，简称样本值.

为使样本具有代表性，抽样应满足：

（1）代表性：X_i与总体X相同分布；

$$E(X_i)=E(X),\ D(X_i)=D(X) \qquad E(X_i^k)=E(X^k),\ E(|X_i^k|)=E(|X^k|)$$

$$E(\overline{X}_n)=E(X) \qquad D(\overline{X}_n)=\frac{1}{n}D(X) \Leftarrow \overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

（2）独立性：$X_1, X_2, \cdots, X_n$相互独立.

称与总体同分布且相互独立的样本为**简单随机样本**，简称**样本**.

举例：研究全校学生的身高，全校学生的身高是总体X，每个学生的身高是个体X_i，随机抽取的学生的身高X_i，其分布同总体X的分布

容量为n的一组样本：$(X_1, X_2, \cdots, X_n)$

•样本是一组随机变量，其具体试验(观察)数值记为：

$x_1, x_2, \cdots, x_n$，称为样本观测值，简称样本值.

称与总体同分布且相互独立的样本为**简单随机样本**，简称**样本.**

实际中，如何得到简单随机样本？

1) 对总体有放回抽样得到的n个个体；

2) 当样本容量n相对总体很小时，连续抽取的n个个体；

3) 测量一个物体重量，重复测量n次得到的；

思考：还有哪些类型的样本？

从民意测验看抽样

1936年, *F*ranklin Delano *R*osevelt(罗斯福)与共和党的候选人—*K*ansas州州长*A*lfred *L*andon(兰登)竞选总统. 绝大多数观测家认为罗斯福会是获胜者,

但《文学摘要》却预测兰登会以 *57%*：*43%* 的优势获胜.

《文学摘要》自1916年以来的历届总统选举中都正确地预测出获胜的一方, 但这次罗斯福以 *62%*：38%的压倒优势取胜！

(不久，《文学摘要》就垮了)

从民意测验看抽样

1936年, *F*ranklin Delano *R*osevelt(罗斯福)与共和党的候选人－*K*ansas州州长*A*lfred *L*andon(兰登)竞选总统. 绝大多数观测家认为罗斯福会是获胜者,但

《文学摘要》却预测兰登会以 *57%:43%* 的优势获胜.

(不久,《文学摘要》就垮了)

《文学摘要》调查的过程是将问卷寄给一千万人, 这些人的名字和地址摘自电话簿或俱乐部会员名册, 这就筛掉了不属俱乐部或未装电话的穷人.

这在1936年前影响不大, 因为穷人富翁以类似的思考投票; 但1936年经济正在从大萧条中恢复, 故穷人选罗斯福, 而富翁们选兰登。

总体,样本,统计量都是随机变量

样本: $(X_1, X_2, \cdots, X_n)$

(1) X_i与总体X相同分布; (2) $X_1, X_2, \cdots, X_n$相互独立.

$E(X_i) = E(X), D(X_i) = D(X)$

如果矩存在的话: $E(X_i^k) = E(X^k)$,

$E[X_i - E(X_i)]^k = E[X - E(X)]^k$

$E\left(\overline{X}_n\right) = E\left(X\right)$ $D\left(\overline{X}_n\right) = \frac{1}{n} D\left(X\right)$

$\Leftarrow \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

统计量: $T = g(X_1, X_2, \cdots, X_n)$不包含未知参数的样本的函数.

个体值为总体X的一个观测值.

样本值$(x_1, x_2, \cdots, x_n)$为样本的观察值.

统计值$t = g(x_1, x_2, \cdots, x_n)$为统计量$T$的观察值.

个体值、样本值和统计值为确定的数.

四、统计量

统计量: $T = g(X_1, X_2, \ldots, X_n)$不包含未知参数的样本的函数.

例 6.1.1:

设总体 $X \sim B(1, p)$, 其中 p 是未知参数, $(X_1, X_2, \ldots, X_5)$ 是来自 X 的简单随机样本.

(1) 指出以下变量哪些是统计量, 为什么?

$$X_1 + X_2, \quad \max_{1 \le i \le 5} X_i, \quad X_5 + 2p, \quad (X_5 - X_1)^2$$

(2) $(X_1, X_2, \ldots, X_5)$ 的联合概率分布?

解 (1) 只有 $X_5 + 2p$ 不是统计量, 因 p 是未知参数.

例6.1.1: 设总体 $X \sim B(1, p)$, 其中 p 是未知参数, $(X_1, X_2, \ldots, X_5)$ 是来自 X 的简单随机样本.

(2) $(X_1, X_2, \ldots, X_5)$ 的联合概率分布?

(2) 因 $P\{X = x\} = p^x (1-p)^{1-x} \qquad x = 0,1$

故 $(X_1, X_2, \ldots, X_5)$ 的联合分布律为:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_5 = x_5)$$

$$= \prod_{i=1}^{5} P\{X_i = x_i\} = \prod_{i=1}^{5} p^{x_i} (1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^{5} x_i} (1-p)^{5-\sum_{i=1}^{5} x_i} \qquad x_i = 0,1 \quad i = 1,2,\cdots,5.$$

(3) $(X_1, X_2, \ldots, X_5)$ 的联合分布函数?

$$F(x_1, \cdots, x_5) = \prod_{i=1}^{5} F(x_i)$$

四、统 计 量

统计量: $T= g(X_1, X_2, \cdots, X_n)$不包含未知参数的样本的函数.

对于相应的样本值$(x_1, x_2, \cdots, x_n)$,

$t= g(x_1, x_2, \cdots, x_n)$称为统计量的**统计值**.

总 体 是 随 机 变 量.

样 本 是 随 机 向 量.

统计量 是 随机变量 (或向量)

常见统计量：

样本均值：

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

$\gamma_k = E(X^k)$ $k=1,2,3\ldots..$

为X的 k 阶原点矩.

样本方差：

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

样本 k 阶原点矩：

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i{}^k$$

$\mu_k = E\{[X-E(X)]^k\}$

为X的k阶中心矩.

样本 k 阶中心矩：

$$M_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k$$

$$A_1=\overline{X}$$

$\gamma_1 = E(X)$

$D(X)=E(X^2)-[E(X)]^2$

统计学中最常用的公式

$$(1)\sum_{i=1}^{n}(X_i-\overline{X})=0;$$

$$(2)\sum_{i=1}^{n}(X_i-A)^2=\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-A)^2;$$

$$(3)\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}X_i^2-n\overline{X}^2.$$

思考：样本矩与总体矩 (即第四章中定义的随机变量矩)的概念有什么区别？

$\gamma_k = E(X^k)$, $k=1,2,3\ldots$ 为X的 k 阶原点矩.

(第六章)为总体的 k 阶原点矩.

$\mu_k = E\{[X-E(X)]^k\}$, $k=1,2,3\ldots$ 为X的k阶中心矩.

(第六章)为总体的 k 阶中心矩.

样本矩：

是随机变量, 包含了部分个体的信息,

取值是变化的, 由具体的实验结果决定.

总体矩 (第四章中定义的矩)：

包含了所有个体信息, 取值是确定的数.

思考1： 设 $E(X)=\mu$ ， $D(X)=\sigma^2$

则 $E(A_1)=\mu$

又若 $E(S^2)=\sigma^2$ ，则 $E(M_2)=$

思考2： $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 由独立同分布中心极限定理，

当n 充分大的时，

样本矩 是 随机变量, 总体矩 是 数值.