第 3 讲 大数据分析模型

海量数据计算研究中心

杨东华

大数据分析的流程

1 大数据的采集和存储

2 大数据预处理

※ 3 大数据分析建模和大数据分析方法

4 大数据分析结果展示

大数据分析建模

大数据分析模型与分析方法

按照一定的方法建立大数据分析模型，并且用合适的统计分析方法对收集来的规模巨大的量数据进行分析，提取有用信息和形成结论。

大数据分析模型

1 大数据分析模型建立方法

2 基本统计量

3 推断统计

目录

大数据分析模型 —概述

- 大数据分析模型讨论的问题是“从大数据中发现什么？”。
- 尽管对大数据的分析方法林林总总，但面对一项具体应用，大数据分析非常依赖想象力。

例如：对患者进行智能导诊，为患者选择合适的医院、合适的科室和合适的医生。

- 通过患者对病症的描述建立模型选择合适的科室；
- 基于对患者位置、医院擅长病症的信息以及患者病症的紧急程度建立模型确定位置合适的医院；
- 根据医院当前的队列信息建立模型进行推荐：如果队列较长，则显示已挂号人数较少、等待时间较短的医生资料；如果队列较短，则显示那些挂号费和治疗费用较高但医术相对高明经验相对丰富的医生资料。

大数据分析模型 —概述

- 这些分析离不开一系列基本的模型与方法，大数据分析模型用于描述输入和输出之间的关系，我们经常听说的贝叶斯分类器、聚类、决策树都是大数据分析模型。
- 面向具体应用的大数据分析模型往往是这些分析方法的扩展或者叠加，
  - ✓ 例如，可以结合支持向量机（SVM）和随机森林对患有心脏病病人的重新入院率做一个预测，对那些重新入院概率高的病人提供更加周到的护理（住院期间）和出院后的跟踪护理。

大数据分析模型 —分类方法

大数据的分析模型有多种不同分类方法：

(1) 依据分析的数据类型：

- 面向结构化多维数据的多元分析；
- 面向半结构化图数据的图分析；
- 面向非结构化文本数据的文本分析。

大数据分析模型 —分类方法

(2) 根据分析过程中输出和输入的关系

- 回归分析
- 聚类分析
- 分类分析
- 关联规则分析

(3) 根据输入的特征

- 监督学习
- 无监督学习
- 半监督学习

监督学习和无监督学习区分：是否有监督(supervised)，就看输入数据是否有标签(label)，输入数据有标签，则为有监督学习，没标签则为无监督学习。

大数据分析模型 —重要性

大数据分析是一个比较广的范畴，和统计分析、机器学习、数据挖掘、数据仓库等学科都存在关系，因而Michael I. Jordan建议用”数据科学”来覆盖整个领域。其中，大数据分析模型的建立是其中最基础也是最重要的步骤。

人工智能领域泰斗、深度学习、机器学习奠基者、AlphaGo祖师爷、美国科学院、美国工程院、美国艺术与科学院三院院士。

大数据分析模型

大数据建模方法与传统建模方法的区别

两种方法同异之辩:

- 传统分析是“因果分析”，而大数据分析是“关联分析”；
- 传统分析是“假设 → 验证形式的分析”，
  大数据分析是“探索 → 关联形式的分析”；
- 也有观点认为，大数据分析并无新颖之处，只不过是将传统分析扩展到了更大规模的数据上，需要的只是一些大规模数据处理技术而不是更新的建模方法。

大数据分析模型

我们的观点：

- “大数据分析”和传统的“数据分析”并不是一个割裂的或者对立的概念，其分析模型建立方法可谓“运用之妙，存乎一心”（指运用得巧妙、灵活，全在于善于动脑筋思考。）
- 建立数据分析模型的目的是解决问题，可以根据分析的目的和所拥有的数据资源选择建模的方法论，而并不一定要区分是使用传统的数据分析建模方法还是大数据分析建模方法。

分析模型的建立步骤

例子：研究提高学生学习成绩的方法，经过老师的分析，希望研究“基于学生的行为数据预测学习成绩”这一数据分析任务。

大数据分析模型 —建立步骤

问题分析

- 传统的数据建模方法可能由专家去分析一系列可能的因素，比如上课的出勤率、作业完成率等，然后到相关的数据库中获取相应的数据，并从数据库中得到学生的成绩。
- 大数据建模方法中，试图去获取所有可能的数据，包括学生的起床时间、学生体检记录、学生的学籍等。继而通过可视化（比如做折线图）等方法分析这些因素是否可能和学生的学习成绩有关。

分析模型的建立步骤

1. 业务调研

首先向业务部门进行调研，了解业务需要解决的问题，将业务问题映射成数据分析工作和任务。

结合例子说明步骤1：

经过老师的分析，希望“通过学生的行为数据来预测学习成绩”这一个数据分析任务。

大数据分析模型 —建立步骤

2. 准备数据

- 根据业务需求准备相应的数据。需要注意的是，传统建模方法通常根据建模的目的准备数据，而大数据建模方法通常尽可能搜集全部数据，以便于从中发现此前没有发现的规律。
- 通过调研企业内外部数据，找到分析需要的数据，将数据汇聚到一个特定的区域(数据集市或数据仓库)，探索性分析。

结合例子说明步骤2：

获取所有可能的数据，包括学生的起床时间、学生体检记录、学生的籍贯等。

大数据分析模型 —建立步骤

3. 浏览数据

- 这一步是大数据建模方法所特有的。在这一步骤中，数据科学家或者用户通过浏览数据发现数据中一些可能的新关联。
- 这个步骤可以通过大数据可视化来实现。

结合例子说明步骤3：

通过可视化（比如做折线图）等方法分析这些因素（学生的起床时间、学生体检记录、学生的籍贯等）是否可能和学生的学习成绩相关。

大数据分析模型 —建立步骤

4. 变量选择

- 基于分析的目标选择模型中的自变量，并定义模型中的因变量。
- 因变量根据数据来定义。
- 自变量根据数据的模式以及和因变量之间的关系进行选择。

结合例子说明步骤4：

使用一些特征工程的方法，选择和成绩相关性比较高的变量，并排除不相关的自变量。

大数据分析模型 —建立步骤

5. 定义或发现模式

定义或发现自变量和因变量之间模型的模式。所谓模型的模式指的是模型的“样子”。例如，自变量$x_1, x_2,\ldots x_n$和因变量y之间的关系可以表示成$y=f(x_1, x_2,\ldots x_n)$，或者自变量构成决策树，因变量y在叶子上。

结合例子说明步骤5：

选择学生上课的出勤率、作业完成率和血压作为自变量。由于它们都是数值型变量，所以考虑使用多元线性回归。

大数据分析模型 —建立步骤

6. 计算模型参数

- 模型中的参数决定了模型的最终形式。
- 有些参数需要根据需求或者数据形式来定义；有些参数需要通过算法学习得到。
- 有时候，模型中的参数需要根据分析模型的实际应用结果进行调整，即“调参”。

结合例子说明步骤6：

通过算法确定多元线性回归中的参数。比如学生上课的出勤率、作业完成率和血压等。

大数据分析模型 —建立步骤

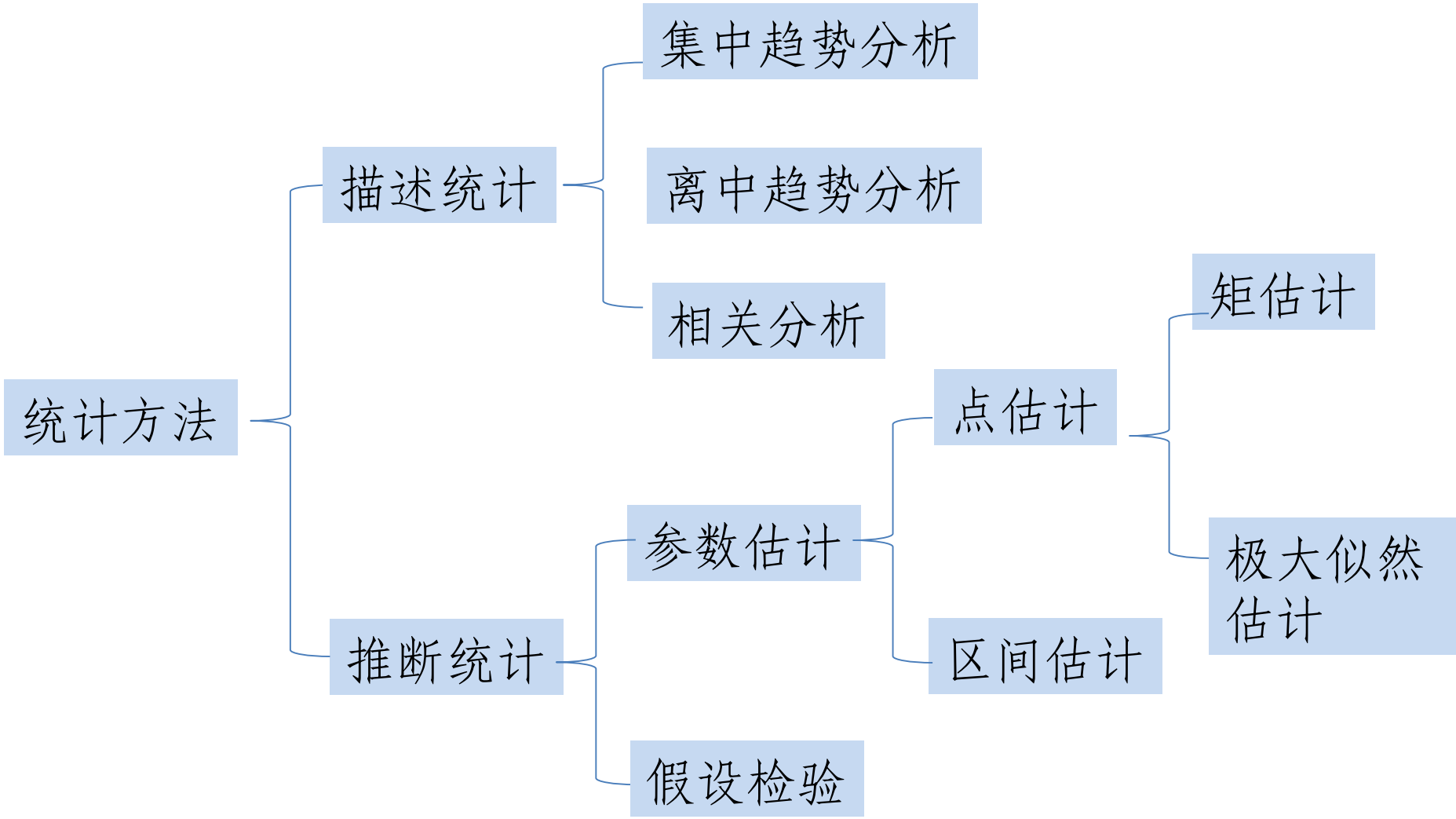
7. 模型的解释与评估

当分析模型建立之后，需要由业务专家进行业务解释和结果评价。

- 可以将分析模型应用于业务中的数据，由业务专家根据经验解释从分析模型得到的结果，看此结果是否符合业务要求；
- 也可以基于已知有效分析结果的数据对模型进行评估，自动验证模型得到的分析结果能否和有效的分析结果相符合。

结合例子说明步骤7:

一种方法是由专家来分析，比如“为什么血压的平方会对成绩有影响？”另一种方法就是用更多的数据来验证是否这个模型可以得到推广。

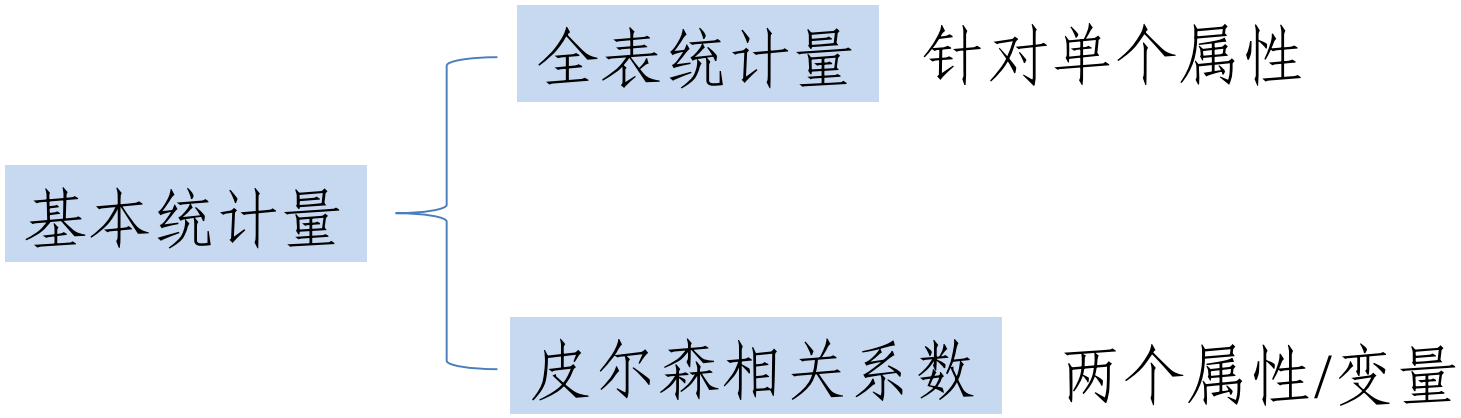

统计方法
描述统计
集中趋势分析
离中趋势分析
相关分析
推断统计
参数估计
点估计
矩估计
极大似然估计
区间估计
假设检验

目录

基本统计量

为什么需要基本统计量

基本统计量，计算简单，而且可以在一定程度上反映出数据的特征和变化趋势。



基本统计量 —全表统计量

根据反映出的数据特征类型，全表统计量可以分为两类:

- 反映数据集中趋势，包括均值、中位数和众数。
- 反应数据波动大小，包括极差和方差（标准差）。

1. 反映数据集中趋势

刻画一组“平均水平”的数据代表。

(1) 均值

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

数据集{9、8、10、7、6}，其均值为8

1. 反映数据集中趋势

(2) 加权平均

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n}} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

这里，权值反映它们所依附的对应值的意义、重要性或出现的概率。不足之处在于：受极端值的影响较大。

数据集{9、8、10、7、6}，权值分别为0.1, 0.2, 0.2, 0.3, 0.2，加权平均值为7.8

1. 反映数据集中趋势

(2) 加权平均

例：某人射击十次，其中二次射中10环，三次射中8环，四次射中7环，一次射中9环，那么他平均射中的环数为：

（10 × 2+8 × 3+7 × 4+9 × 1）/10 = 8.1

这里，7，8，9，10这四个数是射击者射中的几个不同环数，但它们出现的频数不同，分别为4，3，1，2，数据的频数越大，表明它对整组数据的平均数影响越大，实际上，频数起着权衡数据的作用，称之为权数或权重。

1. 反映数据集中趋势

(3) 中位数

- 有序数据序列的中间值，即把数据较高的一半与较低的一般分开的值。
- 如果N为奇数，中位数为该有序集的中间值。
- 如果N是偶数，中位数取作最中间两个值的平均值。
- 不足：不能充分利用所有数据信息。

有序数据序列{6、7、8、9、10}，其中位数为8;

有序数据序列{5、6、7、8、9、10}，其中位数为(7+8)/2=7.5

1. 反映数据集中趋势

(4) 众数

- 数据集中出现最频繁的值。
- 不足：当各个数据的重复次数大致相等时，众数往往没有特别的意义。

例：某人射击十次，其中二次射中10环，三次射中8环，四次射中7环，一次射中9环。

其众数为7。

基本统计量 —全表统计量

常用中心趋势度量统计量的比较

数据特征类型	统计量	意义	不足
集中趋势	平均数	平均数是反映数据集中趋势最常用的统计量，它能充分利用数据所提供的信息	受极端值的影响较大
	中位数	中位数是一个位置代表值，表明一组数据中，有一半的数据大于（或小于）中位数，计算简便，不受极端值的影响	不能充分利用所有数据信息
	众数	当一组数据有较多的重复数据时，人们往往关心众数，它提供了哪个（些）数据出现的次数最多，不受极端值的影响	当各个数据的重复次数大致相等时，众数往往没有特别的意义

2. 反映数据波动大小

刻画数据离散程度的统计量。

设$x_1, x_2, \cdots, x_n$是某数值属性X上的观测集合。

(1) 极差

- 最大值与最小值之差。
- 不足：不能充分利用所有数据信息。

数据集{9、8、10、7、6}，其极差为4

2. 反映数据波动大小

(2) 方差

令$\overline{x}$为均值，那么方差为

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$$

则称标准差为 σ_x。

- 不足：计算繁琐，且单位与原数据单位不一致。
- 低方差意味着数据观测趋向于非常靠近均值；而高方差表示数据散布在一个大的值域中。

数据集{9、8、10、7、6}，其方差为2

基本统计量 —全表统计量

常用数据散布度量统计量的比较

数据特征类型	统计量	意义	不足
波动大小	极差	反映一组数据的波动范围，计算简单	不能充分利用所有数据信息
	方差（标准差）	反映一组数据的波动大小，方差越大，数据的波动就越大，方差越小，数据的波动越小	计算繁琐，单位与原数据单位不一致

基本统计量 —例子

例子1: 下星期三就要数学竞赛了，甲、乙两名同学只能从中挑选一个参加。若你是老师，你认为挑选哪一位比较适宜?甲、乙两个同学本学期五次测验的数学成绩分别如下(单位: 分)

甲	85	90	90	90	95
乙	95	85	95	85	90

解: 计算两名同学的平均成绩， $\overline{X}_{甲}$=90(分), $\overline{X}_{乙}$=90(分)

基本统计量 —例子

谁的成绩稳定性好？应该以什么数据来衡量？

甲同学成绩与平均成绩的偏差的和:

$(85-90)+(90-90)+(90-90)+(90-90)+(95-90)=0$

乙同学成绩与平均成绩的偏差的和:

$(95-90)+(85-90)+(95-90)+(85-90)+(90-90)=0$

甲同学成绩与平均成绩的偏差的平方和:

$(85-90)^2+(90-90)^2+(90-90)^2+(90-90)^2+(95-90)^2=50$

乙同学成绩与平均成绩的偏差的平方和:

$(95-90)^2+(85-90)^2+(95-90)^2+(85-90)^2+(90-90)^2=100$

方差越大，说明数据的波动性越大，越不稳定。

基本统计量 —例子

例子2：在一次芭蕾舞比赛中，甲乙两个芭蕾舞团表演了舞剧《天鹅湖》，参加表演的女演员的身高(单位：cm)分别是：

甲团：163、164、164、165、165、165、166、167

乙团：163、164、164、165、166、167、167、168

哪个芭蕾舞团女演员的身高更整齐？

解：甲乙两团演员的平均身高分别为：

$$\bar{X}_{甲}=(163+164+164+165+165+165+166+167)/8 \approx 165$$
$$\bar{X}_{乙}=(163+164+164+165+166+167+167+168)/8 \approx 166$$

$$S^2_{甲}= [(163-165)^2+(164-165)^2+\cdots+(167-165)^2]/8 \approx 1.36$$
$$S^2_{乙}= [(163-166)^2+(164-166)^2+\cdots+(168-166)^2]/8 \approx 2.75$$

因为$S^2_{甲} < S^2_{乙}$，所以甲芭蕾舞团女演员的身高更整齐。

基本统计量 —皮尔森相关系数

引言

全表统计量都是针对单个属性的（用来描述一维数据），如何衡量两个属性(在统计学中称为变量)之间关联关系？这个关联关系可以用相关系数来衡量。

协方差

- 在概率论和统计学中，协方差用于衡量两个变量的总体误差。而方差是协方差的一种特殊情况，即当两个变量是相同的情况。
- 期望值分别为$E[X]$与$E[Y]$的两个实随机变量X与Y之间的协方差$Cov(X,Y)$定义为：

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY]-2E[Y]E[X]+E[X]E[Y] \\ &= E[XY]-E[X]E[Y] \end{aligned}$$

- 从直观上来看，协方差表示的是两个变量总体误差的期望。

协方差

- 如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值时另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值；如果两个变量的变化趋势相反，即其中一个变量大于自身的期望值时另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。
- 协方差作为描述X和Y相关程度的量，在同一物理量纲之下有一定的作用，但同样的两个量采用不同的量纲使得它们的协方差在数值上表现出很大的差异，于是引出皮尔森相关系数。

皮尔森相关系数

- 相关系数可以用许多统计值来测量，皮尔森相关系数是最常用的一种。
- 它是英国统计学家皮尔森于20世纪提出的一种计算直线相关的方法。
- 皮尔森相关又称为积差相关，积矩相关。

皮尔森相关系数

- 两个变量$X(x_1, x_1, \cdots, x_n)$和$Y(y_1, y_2, \cdots, y_n)$的皮尔森相关系数可以定义为两个变量之间的协方差和标准差的商，即

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

总体相关系数，常用希腊小写字母 ρ (rho) 作为代表符号。

样本相关系数

估算样本的协方差和标准差，可得到样本相关系数(样本皮尔森系数)，常用英文小写字母 r 代表:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

这里，协方差 $\sum[(X - u_X)(Y - u_Y)] = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$

标准差 $\sigma_X = \frac{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\sqrt{n}}$，$\sigma_Y = \frac{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}{\sqrt{n}}$

基本统计量 —皮尔森相关系数

r 亦可由(X_i, Y_i)样本点的标准分数均值估计，得到与上式等价的表达式：

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_X}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right)$$

其中 $\frac{X_i - \bar{X}}{\sigma_X}$，$\bar{X}$，及$\sigma_X$分别是 X_i样本的标准分数，样本均值和样本标准差。

这里，样本标准差：

$$\sigma_X = \frac{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\sqrt{n-1}}$$

n-1为自由度，意思为样本能自由选择的程度。当选到只剩一个时，它不可能再有自由了，所以自由度为n-1。

两个属性间的关联关系

两个属性（变量）之间的关联关系：

- 属性X增大的同时，属性Y增大，则它们为正相关，通常令其相关系数在0与1之间；
- 属性X增大的同时，属性Y减小，则它们为负相关，通常令其相关系数在-1与0之间；
- 如果X和Y没有任何关联关系，它们的相关系数为0；

相关系数的绝对值越大，相关性越强，相关系数越接近于1或-1；相关系数越接近于0，相关度越弱。

皮尔森相关系数

- 皮尔森相关系数的变化范围为[−1,1]，绝对值越大相关越强。
  - ✓如果系数的值为1，就意味着X和Y可以理想地由直线方程来描述，所有的数据点都很好地落在一条直线上，且Y随着X的增加而增加。
  - ✓相反，系数的值为−1意味着所有的数据点都落在直线上，但Y随着X的增加而减少。
  - ✓此外，系数的值为0意味着两个变量之间没有线性关系。
- 注意：高度相关并不蕴含因果关系。（比如：看见闪电和听见雷声是高度相关的，但二者之间并没有因果关系。）

皮尔森相关系数

适用条件（约束条件）：

1. 两变量均应由测量得到的连续变量；
2. 两变量所来自的总体都应是正态分布，或接近正态的单峰对称分布；取大样本进行正态分布非参数检验；
3. 变量必须是成对的数据；
4. 两变量间为线性关系；
5. 两变量独立。

皮尔森相关系数

为什么通常假设为正态分布？

- 正态分布是许多统计方法的理论基础。检验、方差分析、相关和回归分析等多种统计方法均要求分析的指标服从正态分布。

- 许多统计方法虽然不要求分析指标服从正态分布，但相应的统计量在大样本时近似正态分布，因而大样本时这些统计推断方法也是以正态分布为理论基础的。

- 皮尔森相关也不例外。在求皮尔森相关性系数以后，通常还会用*t*检验之类的方法来进行皮尔森相关性系数检验，而*t*检验是基于数据呈正态分布的假设的。

基本统计量 —皮尔森相关系数

举例

例1：计算压力x和压缩量y之间的相关系数r。

表 2-3　绝缘材料的压缩量和压力表

压力 x(10 lb/in^2)	压缩量 y(0.1 in)
1	1
2	1
3	2
4	2
5	4

1. 首先计算 $\bar{X}$和 $\bar{Y}$，

$$\bar{X}=\frac{1+2+3+4+5}{5}=3,\quad \bar{Y}=\frac{1+1+2+2+4}{5}=2$$

2. 得到：

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{5}(X_i-3)(Y_i-2)=7;$$

$$\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\sqrt{10},\quad \sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}=\sqrt{6};$$

举例

3. 从而得到:

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{7}{\sqrt{10}\sqrt{6}}=0.904$$

可以看出，压力和压缩量是高度相关的，而且是很强的正相关关系，不过需要注意的是，高度相关并不一定蕴含因果关系。

举例

例 2: 某地29名13岁男童的身高(cm)、体重(kg)如表所示，试运用相关分析法来分析其身高与体重是否相关。

分析:

- 任何事物的存在都不是孤立的，而是相互联系、相互制约的。
- 相关分析可对变量进行相关关系的分析，计算29名13岁男童的身高(cm)、体重(kg)，以判断两个变量之间相互关系的密切程度。

基本统计量 —皮尔森相关系数

举例

原始数据：

Number	height	weight
1	135.1	32
2	139.9	30
3	163.6	46
4	146.5	34
5	156.2	37
6	156.4	36
7	167.8	42
8	149.7	31
9	145.0	33
10	148.5	37
11	165.5	50
12	135.0	28
13	153.3	41
14	152.0	32

15	160.5	47
16	153.0	47
17	147.6	41
18	157.5	43
19	155.1	45
20	160.5	38
21	143.0	32
22	149.4	34
23	160.8	40
24	159.0	39
25	158.2	38
26	150.0	36
27	144.5	35
28	154.6	40
29	156.5	32

举例

输出结果：

Descriptive Statistics

	Mean	Std. Deviation	N
身高（cm）	152.576	8.3622	29
体重(kg)	37.65	5.746	29

图为基本的描述性统计量的输出表格，其中身高的均值(mean) 为152. 576cm、标准差(Standard deviation)为8.3622 、样本容量(number of cases)为29；体重的均值为37.65kg、标准差为5.746 、样本容量为29。

基本统计量 —皮尔森相关系数

Correlations

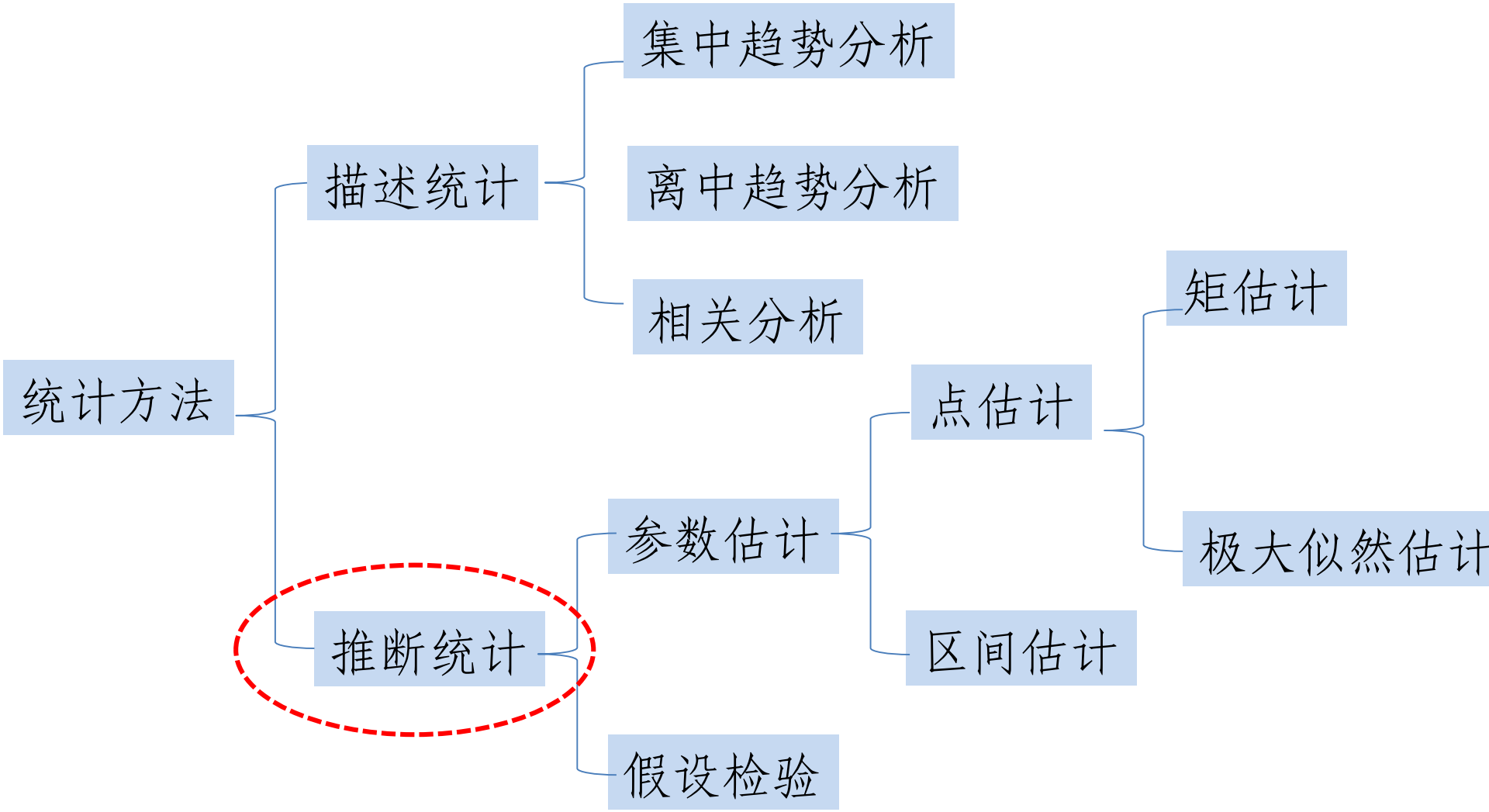
Control Variables			体重(kg)	肺活量(ml)	身高（cm）
-none-[a]	体重(kg)	Correlation	1.000	.613	.719
		Significance (2-tailed)	.	.000	.000
		df	0	27	27
	肺活量(ml)	Correlation	.613	1.000	.588
		Significance (2-tailed)	.000	.	.001
		df	27	0	27
	身高（cm）	Correlation	.719	.588	1.000
		Significance (2-tailed)	.000	.001	.
		df	27	27	0
身高（cm）	体重(kg)	Correlation	1.000	.337	
		Significance (2-tailed)	.	.079	
		df	0	26	
	肺活量(ml)	Correlation	.337	1.000	
		Significance (2-tailed)	.079	.	
		df	26	0	

a. Cells contain zero-order (Pearson) correlations.

- 图为相关分析结果表，从表中可以看出体重和身高之间的皮尔逊相关系数为0.719，表示体重与身高呈正相关关系，且两变量是显著相关的。
- 所以可以得出结论：学生的体重与身高存在显著的正相关性，当体重越高时，身高也越高。

目录

统计方法 —概览


统计方法
描述统计
集中趋势分析
离中趋势分析
相关分析
推断统计
参数估计
点估计
矩估计
极大似然估计
区间估计
假设检验

统计方法 —描述统计

- 描述统计是通过图表或数学方法，对数据资料进行整理、分析，并对数据的分布状态、数字特征和随机变量之间关系进行估计和描述的方法。
- 描述统计分为集中趋势分析、离中趋势分析和相关分析三大部分。
  - ✓ 集中趋势分析主要靠平均数、中位数、众数等统计指标来表示数据的集中趋势。例如考试的平均成绩多少？是正偏分布还是负偏分布？
  - ✓ 离中趋势分析主要靠全距、四分差、平均差、方差、标准差等统计指标来研究数据的离中趋势。例如，我们想知道两个教学班的语文成绩中，哪个班级内的成绩分布更分散，就可以用两个班级的四分差或百分点来比较。
  - ✓ 相关分析探讨数据之间是否具有统计学上的关联性。

统计方法 —推断统计

推断统计是研究如何利用样本数据来推断总体特征的统计方法。这就需要抽取部分个体即样本进行测量，然后根据获得的样本数据对所研究的总体特征进行推断，这就是推断统计要解决的问题。

- 基本特征是其依据的条件中包含带随机性的观测数据。
- 以随机现象为研究对象的概率论，是推断统计的理论基础。

统计方法 —推断统计

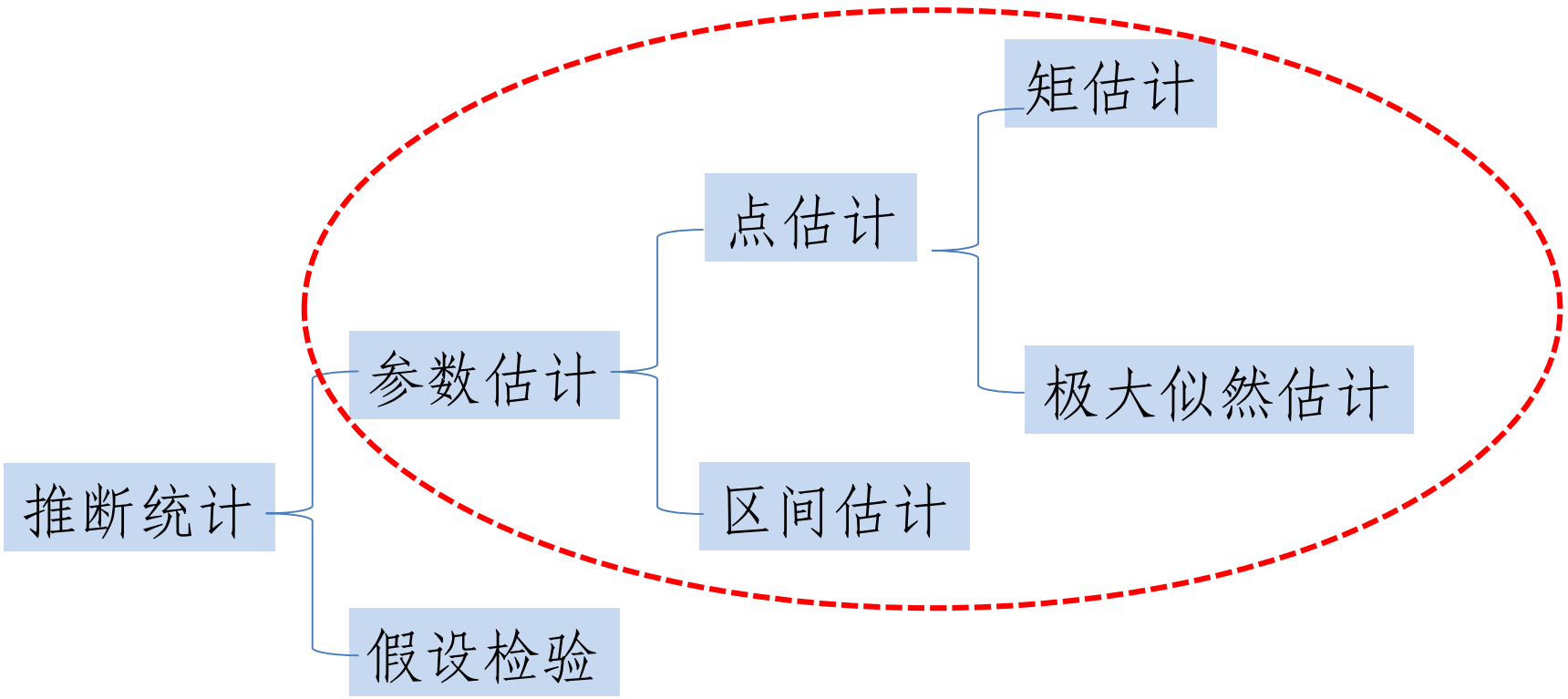
推断统计包含两个内容:

- 参数估计，即利用样本信息推断总体特征，例如某一群人的视力构成一个总体，通常认为视力是服从正态分布的，但不知道这个总体的均值，随机抽取部分人，测得视力的值，用这些数据来估计这群人的平均视力；

- 假设检验，即利用样本信息判断对总体的假设是否成立。例如，若感兴趣的问题是“平均视力是否超过4.8”，就需要通过样本检验此命题是否成立。(显著水平α默认为0.05)

推断统计的目的：利用问题的基本假定及包含在观测数据中的信息，作出尽量精确和可靠的结论。

推断统计



1. 参数估计

实际问题中，所研究的总体分布类型往往是已知的，但是要依赖于一个或者几个未知的参数。这时，求总体分布的问题就归结成了求一个或者几个未知参数的问题，这就是所谓的**参数估计**。

推断统计 —参数估计

- 参数是刻画总体某方面的概率特性的数量。
- 当这个数量是未知的时候，从总体抽出一个样本，用某种方法对这个未知参数进行估计就是参数估计。

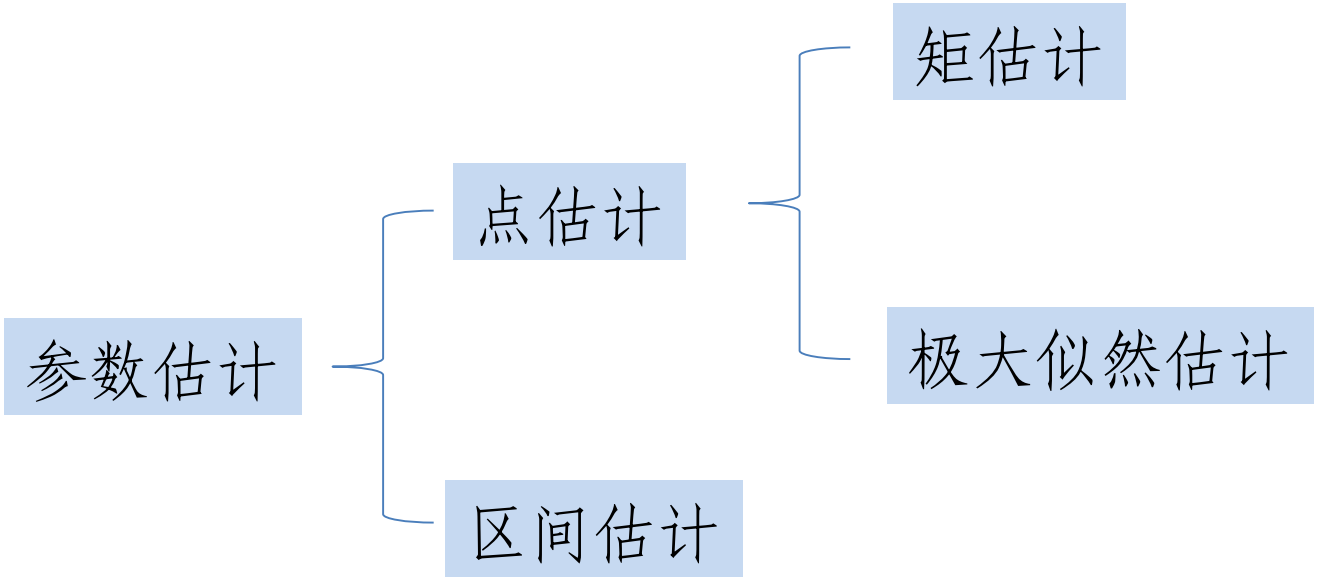
- 在总体分布类型已知的情况下，进行参数估计是推断统计的重要内容。有些实际问题中人们不关心总体分布的形式，而只是想知道均值、方差等数字特征，对这些数字特征的估计问题，也是参数估计的一部分。

例子

- 目的：为了求一款电灯的使用寿命X的分布。
- 由经验得，X服从正态分布$N(\mu,\sigma^2)$。由于μ,σ^2未知，因此我们通过构造样本的函数，给出它们的估计值或取值范围，这就是参数估计的主要内容。

点估计　　区间估计

参数估计的分类



点估计

- 简单来说，设总体X的分布函数的形式已知，且为$F(x;\ \theta)$，其中θ为未知参数，通过采集的样本$X_1, X_2, \ldots, X_n$对参数的θ值进行估计称为点估计。

- 估计量 $\hat{\theta}(X_1, X_2 \cdots, X_n)$的值称为$\theta$ $(X_1, X_2 \cdots, X_n)$的估计值。 数值 随机变量

- 点估计的目的就是寻求未知参数的估计量和估计值。

- 主要有钜估计法、极大似然估计两种方法。

区间估计的引入

- 使用点估计的方法对参数θ进行估计，得到的估计值$\hat{\theta}$只是被估参数θ的良好近似。但近似程度如何？误差范围多大？可信程度又如何？这些问题点估计都无法回答的。
- 因此，我们引入区间估计。

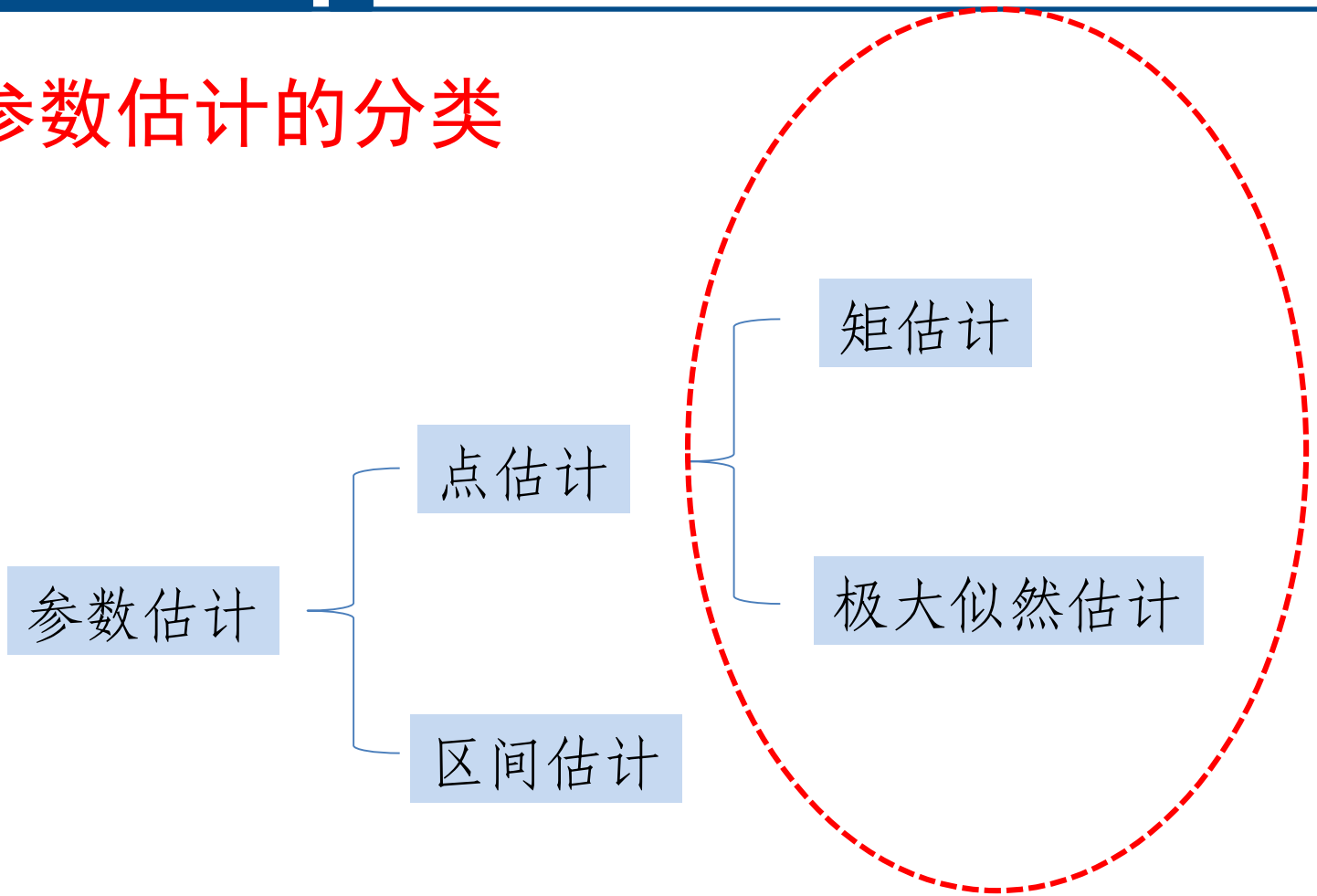
区间估计的主要思想

- 区间估计是从点估计和抽样标准误差（Standard error）出发，按照给定的概率值建立包含待估计参数的区间。
- 这个给定的概率值称为置信度或置信水平，是指总体参数值落在样本统计值某一区间内概率；
- 这个建立起来包含待估计参数的区间称为置信区间，是指在某一置信水平下，样本统计值与总体参数值之间的误差范围。
- 划定置信区间的两个值分别称为置信下限和置信上限。

点估计和区间估计的区别

参数估计
- 点估计：估计未知参数的值
- 区间估计：估计未知参数的取值范围，使得这个范围包含未知参数真值的概率为给定的值。

参数估计的分类



1. 矩估计

动机:

- 随机变量的矩是非常简单的描述随机变量统计规律的方法。
- 而且随机变量的一些参数往往本身就是随机变量的矩或者某些矩的函数。
- 因此，我们可以把未知参数θ用总体矩$\mu_k=E(X^k)$ (k=1, 2, …, m)的函数表示为$\theta = h(\mu_1, \mu_2, \cdots, \mu_n)$

1. 矩估计

- 这种用样本矩的函数作为参数θ 估计的方法，就是矩估计法。

方法实现：

用样本的k阶矩作为总体的k阶矩的估计量，建立含待估计参数的方程，从而可解出待估计参数。

1. 矩估计

一些相关的基本概念:

在数理统计学中有一类数字特征称为矩(moment)，随机变量的k阶矩就是它的k次方的数学期望。

令k为正整数（或为0），a为任何实数，X为随机变量，则期望值

$$E((X-a)^k)$$

叫做随机变量X对a的k阶矩，或叫动差。

1. 矩估计

一些相关的基本概念:

原点矩: 如果$a=0$，则有$E(X^k)$，叫做k阶原点矩，记作$\mu_k(X)=E(X^k)$，也叫k阶矩。

若X为离散随机变量，则$\mu_k(X)=\sum_i x_i^k\, p(x_i)$；

若X为连续随机变量，则$\mu_k(X)=\int_{-\infty}^{+\infty} x^k f(x)dx$

显然，一阶原点矩就是数学期望，即$\mu_1(X)=E(X)$，k=1, 2, …

1. 矩估计

中心矩：设随机变量X的函数$[X-E(X)]^k$，(k=1, 2, …)的数学期望存在，则称$\mu_k(X)=E\{[X-E(X)]^k\}$为X的k阶中心矩，记作$\mu_k(X)$，即$\mu_k(X)=E\{[X-E(X)]^k\}$, k=1, 2, …

若X为离散随机变量，则$\mu_k(X)=\sum_i[x_i-E(X)]^k\,p(x_i)$；

若X为连续随机变量，则$\mu_k(X)=\int_{-\infty}^{+\infty}[x-E(X)]^k f(x)dx$

1. 矩估计

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

易知，一阶中心矩恒等于零，即 $\mu_1(X)= E[X-E(X)]= 0;$

二阶中心矩就是方差，即 $\mu_2(X)= E\{[X-E(X)]^2\}$

$$= Var(X)$$

$$=D(X)$$

1. 矩估计

记总体k阶原点矩为：$\mu_k=E(X^k)$

样本k阶原点矩为：$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$

记总体k阶中心矩为：$\mu_k=E[X-E(X)]^k$

样本k阶中心矩为：$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$

1. 矩估计

矩估计的思想

- 以样本矩作为总体矩，求解被估参数。
- 最简单的矩估计法是用一阶样本原点矩来估计总体的期望，

$$\hat{E(X)} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

用二阶样本中心矩来估计总体的方差。

$$\hat{D(X)} = M_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

1. 矩估计

举例1：设随机变量γ在(a，b)上服从均匀分布。试求随机变量γ的k阶原点矩和三阶中心矩。

k阶原点矩:

$$E\gamma^k = \int_a^b x^k \frac{1}{b-a} dx = \frac{1}{k+1}\left(b^k + b^{k-1}a + \ldots + ba^{k-1} + a^k\right)$$

因为 $E\gamma = \frac{a+b}{2}$ 故 $E(\gamma - E\gamma)^3 = \int_a^b \left(x - \frac{a+b}{2}\right)^3 \bullet \frac{1}{b-a} dx$

令$t=x-\frac{a+b}{2}$

$$\frac{1}{b-a}\int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} t^3 dt = 0$$

1. 矩估计

例2：总体服从泊松分布$P(\lambda)$，X_1，X_2，…，X_n是来自总体的一个样本，试求参数λ的矩估计量。

解法1：泊松分布的数学期望$E(X)=\lambda$，用样本均值 $\overline{X}$ 近似期望，$\overline{X} \approx E(x) = \lambda$，得$\lambda$的矩估计量$\hat{\lambda} = \overline{X}$。

解法2：泊松分布的方差$Var(X)=\lambda$，用样本方差S^2替换总体方差，得λ的矩估计量$\hat{\lambda} = S^2$。

1. 矩估计

例3: 总体服从指数分布$E(\lambda)$，X_1，X_2，…，X_n是来自总体的一个样本，试求参数λ的矩估计。

解：指数总体的数学期望$E(X)=1/\lambda$，用样本均值 $\bar{X}$近似期望，$\bar{X}\approx E(x)=1/\lambda$，得到方程 $\bar{X}=1/\hat{\lambda}$, 求解得λ的矩估计量$\hat{\lambda}=1/\bar{X}$。

1. 矩估计

例4: 设X_1，X_2，…，X_n是来自正态分布总体$N(\mu, \sigma^2)$的样本，求参数μ和σ^2的矩估计量。

解：正态总体的数学期望$E(X)=\mu$，方差$Var(X)=\sigma^2$

用样本均值$\bar{X}$和样本方差S^2近似总体期望和方差，得到：μ的矩估计量为$\hat{\mu}=\bar{X}$

σ^2的矩估计量为$\hat{\sigma}^2=S^2$

1. 矩估计

$$\mu_2 = E\left(X^2\right) = D(X) + [E(X)]^2$$

例5: 设总体X在$[a,b]$上服从均匀分布，a,b未知，$X_1, X_2, \cdots, X_n$是总体X的一个样本，试求a,b的矩估计量。

解：

– 可知X的概率密度为 $f(x;a,b) = \begin{cases} \frac{1}{b-a} & \text{a} \le x \le \text{b} \\ 0 & \text{其他} \end{cases}$

– 又因为 $E(X) = \frac{a+b}{2}, E(X^2) = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4}$

解得a,b的矩估计量分别为

$$\begin{cases} \hat{a} = E(X) - \sqrt{3E(X^2) - 3E(X)^2} \\ \hat{b} = E(X) + \sqrt{3E(X^2) - 3E(X)^2} \end{cases}$$

用解得参数a, b的矩估计量为$\hat{a} = \overline{X} - \sqrt{3}S$, $\hat{b} = \overline{X} + \sqrt{3}S$

1. 钜估计

例6: 有一批零件，其长度X~N(μ, σ2)，现从中任取4件，测的长度(单位: mm) 为12.6, 13.4, 12.8, 13.2 。试估计μ 和σ2 的值。

解: 由

$$\bar{x} = \frac{1}{4}(12.6+13.4+12.8+13.2) = 13$$

$$s^2 = \frac{1}{4-1}[(12.6-13)^2 + (13.4-13)^2 + (12.8-13)^2 + (13.2-13)^2] = 0.133$$

得μ和σ^2的估计值分别为13 (mm) 和0. 133 $(mm)^2$

1. 矩估计

例7: 设总体X的概率密度为，$f(x;\theta)=\begin{cases}\theta x^{\theta-1}, & 0<x<1\\ 0, & \text{其它}\end{cases}$

X_1，X_2，…，X_n是来自总体X的样本，x_1，x_2，…，x_n为样本值求参数θ的矩估计。

解：先求总体矩

$$E(X)=\int_0^1 x\cdot\theta x^{\theta-1}dx=\theta\int_0^1 x^{\theta}dx=\frac{\theta}{\theta+1}x^{\theta+1}\Big|_0^1=\frac{\theta}{\theta+1}$$

$$\text{解之：}\ \theta=\frac{E(X)}{1-E(X)}$$

令 $$A_1=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$$

$\hat{\theta}=\frac{\overline{X}}{1-\overline{X}}$ 为θ的矩估计量,

$\hat{\theta}=\frac{\overline{x}}{1-\overline{x}}$ 为θ的矩估计值.

用样本的一阶原点矩来估计数学期望

1. 矩估计

例8: 设总体X的概率密度为, $f(x,\theta)=\frac{1}{2\theta}e^{-\frac{|x|}{\theta}}$, $-\infty<x<+\infty$, $\theta>0$

求θ的估计量$\hat{\theta}$。

解法1: 虽然$f(x,\theta)$中仅含有一个参数θ, 但是因为

$$EX=\int_{-\infty}^{+\infty}x\cdot\frac{1}{2\theta}e^{-\frac{|x|}{\theta}}dx=0$$

不含θ，不能由此解出θ，就需要继续求总体的二阶原点矩。

$$EX^2=\int_{-\infty}^{+\infty}x^2\cdot\frac{1}{2\theta}e^{-\frac{|x|}{\theta}}dx=\frac{1}{\theta}\int_{0}^{+\infty}x^2e^{-\frac{x}{\theta}}dx=\theta^2\Gamma(3)=2\theta^2$$

用 $A_2=\frac{1}{n}\sum_{i=1}^{n}X_i^2$ 替换 EX^2 即 $A_2=\frac{1}{n}\sum_{i=1}^{n}X_i^2=2\theta^2$

得θ的矩估计量为

$$\hat{\theta}=\sqrt{\frac{1}{2}\cdot\frac{1}{n}\sum_{i=1}^{n}X_i^2}=\sqrt{A_2/2},\quad \theta>0$$

2. 极大似然估计

引言：某位同学与一位猎人一起外出打猎，一只野兔从前方窜过。只听一声枪响，野兔应声到下，如果要你推测，这一发命中的子弹是谁打的？你就会想，只发一枪便打中，由于猎人命中的概率一般大于这位同学命中的概率，看来这一枪是猎人射中的。

这个例子所做的推断就体现了极大似然法的基本思想：在只有概率的情况下，忽略低概率事件直接将高概率事件认为是真实事件。

2. 极大似然估计

例1、离散的小球问题：箱子里有一定数量的小球，每次随机拿取一个小球，查看颜色以后放回，已知拿到白球的概率p为0.7或者0.3，拿了三次，都不是白球，想要求拿到白球的概率的极大似然估计。

分析：此处从数学上来讲，想要准确的求出拿到白球的概率是不可能的，所以此处求的是概率的极大似然估计。而这里的有放回的拿取，是经典的独立重复事件，可以很简单的分别求出白球概率为0.7和0.3的时候拿三次都不是白球的概率。

解：若拿到白球的概率为0.7，拿三次都不是白球的概率为：

P_0.7=0.3*0.3*0.3=0.027

若拿到白球的概率为0.3，拿三次都不是白球的概率为：

P_0.3=0.7*0.7*0.7=0.343

P_0.3>P_0.7，可知当前情况下白球概率为0.3的概率大于白球概率为0.7

综上所述：拿到白球的概率的极大似然估计为0.3。

2. 极大似然估计

例2、连续的小球问题：箱子里有一定数量的小球，每次随机拿取一个小球，查看颜色以后放回，已知拿到白球的概率p的范围是[0.3,0.7]，拿了三次，都不是白球，想要求拿到白球的概率的极大似然估计。

分析：与例1相同，想要知道小球的极大似然估计，就是要先求在已知条件下，发生已知事件的概率，然后据此求出小球的极大似然估计。

解：记拿到白球的概率为p，取白球的事件为Y，取到时Y=1，未取到时Y=0，小球颜色不是白色的事件Y重复3次的概率为：$P(Y=0;p)=(1-p)^3$

欲求p的极大似然估计，即要求P(Y=0;p)的极大值：令$Q(p)=(1-p)^3$，$Q'(p)=-3*(1-p)^2$，令$Q'(p)=0$

求得Q的极值点为p=1，且当p<1时,$Q'(p)<0$，p>1时，$Q'(p)<0$，可知Q(p)为单调减函数，可知$0.3\leq p\leq 1$的条件下，p=0.3时，Q(p)取得最大值。

综上所述：小球概率的极大似然估计为0.3。

2. 极大似然估计

似然函数

- 设总体X是离散型随机变量，具有分布率 $\mathrm{p}(x;\theta)$，其中θ为未知参数，$X_1, X_2, \cdots, X_n$是总体X的一个样本。
- 称以下函数为样本的似然函数

$$\mathrm{L}(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} \mathrm{p}(x_i; \theta)$$

若X是连续性随机变量，则似然函数定义为：

$$\mathrm{L}(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

2. 极大似然估计

极大似然估计

– 使得似然函数最大的参数$\hat{\theta}$，称之为参数θ的极大似然估计。即

$$\hat{\theta} = arg\,max\,\mathrm{L}(x_1, x_2, \cdots, x_n; \theta)$$

2. 极大似然估计

极大似然估计

– 若$p(x;\theta)$或$f(x;\theta)$关于θ可微，则我们常通过下面的对数似然方程求解$\hat{\theta}$

$$\frac{\partial L(\theta)}{\partial \theta}=0$$

又因为lnx为x的单调函数，因此，最大最大似然估计$\hat{\theta}$可通过下列方程求得

$$\frac{\partial}{\partial \theta}\ln L=0$$

2. 极大似然估计

例1：设总体X服从参数为p的0-1分布，$X_1, X_2, \cdots, X_n$是总体X的一个样本，试求 p 的极大似然估计。

可知X的概率分布为

$$p(x=k)=\begin{cases}1-p & \text{k=0}\\ p & \text{k=1}\end{cases}$$

解：似然函数为

$$L(X_1, X_2, \cdots, X_n; \theta)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

取对数得$\ln L=\sum_{i=1}^{n} x_i \ln p+\left(n-\sum_{i=1}^{n} x_i \ln(1-p)\right)$

2. 极大似然估计

即， $\ln L = \sum_{i=1}^{n}(1-x_i)\ln(1-p) + x_i \ln p$

令 $$\frac{d\ln L}{dp} = \frac{1}{p}\sum x_i - \frac{1}{1-p}\sum(1-x_i) = 0$$

$$(1-p)\sum x_i - p\sum(1-x_i) = 0$$

$$np = \sum x_i$$

$$\Rightarrow \hat{p} = \bar{x}$$

解得

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}$$

2. 极大似然估计

例2: 设样本服从正态分布 $N(\mu, \sigma^2)$，则似然函数为:

$$L(\mu,\sigma^2)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=(2\pi\sigma^2)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

它的对数:

$$\ln L(\mu,\sigma^2)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

求导，得方程组:

$$\begin{cases}\dfrac{\partial \ln L(\mu,\sigma^2)}{\partial\mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu) &=0\\ \dfrac{\partial \ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases}$$

2. 极大似然估计

联合解得:

$$\begin{cases} \mu^* = \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \sigma^{*2} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{cases}$$

似然方程有唯一解：(μ^*, σ^{*2})

而且它一定是最大值点，这是因为，当$|\mu| \to \infty$,

或$\sigma^2 \to \infty$或0，非负函数$L(\mu, \sigma^2) \to 0$。

于是μ和σ^2的极大似然估计为(μ^*, σ^{*2})。

2. 极大似然估计

例3: 设总体X服从参数为λ的指数分布，即有概率密度:

$$f(x,\lambda)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases},(\lambda>0)$$

又$X_1,X_2,\cdots,X_n$为来自于总体的样本值，试求λ的极大似然估计。

解: 第一步，似然函数为

$$L=L(x_1,x_2,\cdots,x_n;\lambda)=\lambda^n\prod_{i=1}^{n}e^{-\lambda x_i}=\lambda^n\exp(-\lambda\sum_{i=1}^{n}x_i)$$

于是，

$$\ln L=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i$$

2. 极大似然估计

第二步，$$\frac{d\ln L}{d\lambda}=\frac{d}{d\lambda}(n\ln\lambda-\lambda\sum_{i=1}^{n}x_i)=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i$$

第三步，令 $$\frac{d\ln L}{d\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i=0\Rightarrow\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}x_i}=\frac{1}{\bar{x}}$$

经验证，$\ln L(\lambda)$在$\lambda=\hat{\lambda}=1/\bar{X}$处达到最大，所以$\hat{\lambda}$是$\lambda$的极大似然估计。

例4: 设X为离散型随机变量，其分布律如下(0<θ<1/2)，

X	0	1	2	3
P	θ^2	$2(\theta-\theta^2)$	θ^2	$1-2\theta$

随机抽样得3, 1, 3, 0, 3, 1, 2, 3, 分别利用矩估计和极大似然估计法估计参数θ

解:

$EX=0\times\theta^2+1\times2\theta(1-\theta)+2\times\theta^2+3\times(1-2\theta)=3-4\theta$，

故：$\theta=\frac{1}{4}(3-EX)$，

θ的矩估计量为：$\hat{\theta}=\frac{1}{4}(3-\overline{X})$，

根据给定的样本观察值计算：$\overline{x}=\frac{1}{8}(3+1+3+0+3+1+2+3)=2$，

因此θ的矩估计值为：$\hat{\theta}=\frac{1}{4}(3-\overline{x})=\frac{1}{4}$.

对于给定的样本值，

似然函数为：$L(\theta)=4\theta^{6}(1-\theta)^{2}(1-2\theta)^{4}$，

取对数可得：

$lnL(\theta)=ln4+6ln\theta+2ln(1-\theta)+4ln(1-2\theta)$，

从而：$\frac{dlnL(\theta)}{d\theta}=\frac{6}{\theta}-\frac{2}{1-\theta}-\frac{8}{1-2\theta}=\frac{24\theta^{2}-28\theta+6}{\theta(1-\theta)(1-2\theta)}$，

令：$\frac{dlnL(\theta)}{d\theta}=0$，

得方程：$12\theta^{2}-14\theta+3=0$，

解得：$\theta=\frac{7-\sqrt{13}}{12}$（$\theta=\frac{7+\sqrt{13}}{12}>\frac{1}{2}$，不合题意），

于是θ的最大似然估计值为：$\hat{\theta}=\frac{7-\sqrt{13}}{12}$.

2. 极大似然估计

例5: 考虑一个抛硬币的例子。假设这个硬币正面跟反面轻重不同。我们把这个硬币抛80次(即, 我们获取一个采样$x_1=H$, $x_2=T$, …, $x_{80}=T$ 并把正面的次数记下来，正面记为H，反面记为T)。并把抛出一个正面的概率记为p, 抛出一个反面的概率记为$1-p$。假设我们抛出了49个正面，31 个反面，即49次H，31次T。假设这个硬币是我们从一个装了三个硬币的盒子里头取出的。这三个硬币抛出正面的概率分别为p = 1/3, p = 1/2, p = 2/3。 这些硬币没有标记，所以我们无法知道哪个是哪个。

2. 极大似然估计

解：使用最大似然估计，通过这些试验数据(即采样数据)，我们可以计算出哪个硬币的可能性最大。这个可能性函数取以下三个值中的一个:

$$\mathbb{P}(\text{H=49, T=31} \mid p = 1/3) = \binom{80}{49}(1/3)^{49}(1-1/3)^{31} \approx 0.000$$

$$\mathbb{P}(\text{H=49, T=31} \mid p = 1/2) = \binom{80}{49}(1/2)^{49}(1-1/2)^{31} \approx 0.012$$

$$\mathbb{P}(\text{H=49, T=31} \mid p = 2/3) = \binom{80}{49}(2/3)^{49}(1-2/3)^{31} \approx 0.054$$

从上面的算式可以看出，$\hat{p}$ =2/3时，似然函数取得最大值，为0.054。

2. 极大似然估计

例6: 设总体 X 的概率密度为

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

$(X_1,\cdots,X_n)$ 是取自 X 的样本，其中 $0<p<1$ 是未知参数；试分别用矩法和最大似然估计法给出 p 的估计量.

解 (1) 矩估计法：

X 服从几何分布， $E(X)=\dfrac{1}{p}$

所以 p 的矩估计量为 $\hat{p}=\dfrac{1}{\overline{X}}$

2. 极大似然估计

解 (2) 最大似然估计法:

$$L(p)=\prod_{i=1}^{n}p(1-p)^{x_i-1}=p^n(1-p)^{\sum_{i=1}^{n}x_i-n},$$

$$\ln L=n\ln p+\left(\sum_{i=1}^{n}x_i-n\right)\ln(1-p),$$

$$\frac{\mathrm{d}\ln L}{\mathrm{d}p}=\frac{n}{p}+\frac{n-\sum_{i=1}^{n}x_i}{1-p}\overset{令}{=\!=}0,$$

解得 p 的最大似然估计量为 $\hat{p}=\dfrac{n}{\sum_{i=1}^{n}X_i}=\dfrac{1}{\overline{X}}$.

3. 点估计的评价

- 对于同一个总体的同一个参数，可能得到不同的点估计。
- 有下面三个标准对其进行评价

(1) 无偏性：若估计量$\hat{\theta}$的数学期望存在，且对于任意的$\theta \in \Theta$，满足$E(\hat{\theta}) = \theta$，则称$\hat{\theta}$是θ的无偏估计量；

(2) 有效性：对于θ的两个无偏估计$\hat{\theta}_1, \hat{\theta}_2$，如果$D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$，则称$\hat{\theta}_1$比$\hat{\theta}_2$有效。$D(\hat{\theta})$是$\hat{\theta}$的方差。

3. 点估计的评价

(3) 相合性（一致性）：设n为样本容量，如果对任意$\varepsilon > 0$，都有

$$\lim_{n\to\infty} P\{\left|\hat{\theta} - \theta\right| < \varepsilon\} = 1$$

则称$\hat{\theta}$为θ的相合估计（一致估计）。

引言

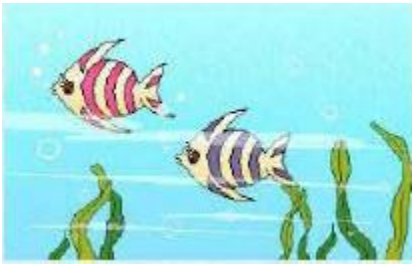
前面，我们讨论了参数的点估计。它的本质是：用样本k阶矩代替总体k阶矩，即用样本算得的一个值去估计总体中的未知参数。但是，点估计值仅仅是未知参数的一个近似值，它没有反映出这个近似值的误差范围，使用起来把握不大。区间估计正好弥补了点估计的这个缺陷。

引言

比如，在估计湖中鱼数的问题中，若我们根据一个实际样本，得到鱼数的估计为1000条。

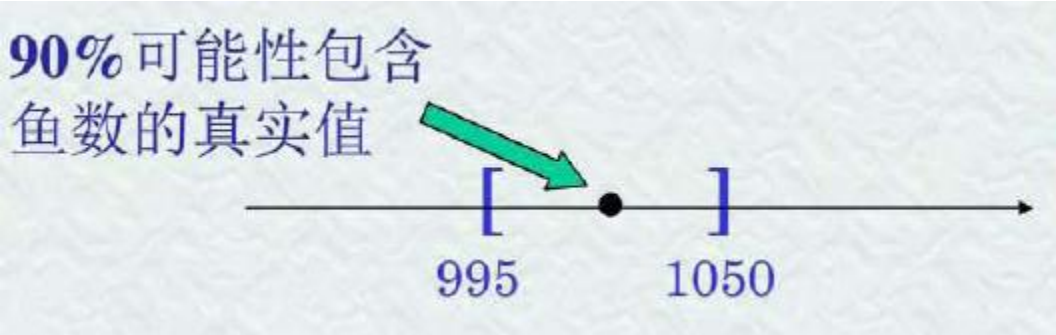
实际上，N的真值可能大于1000条，也可能小于1000条。

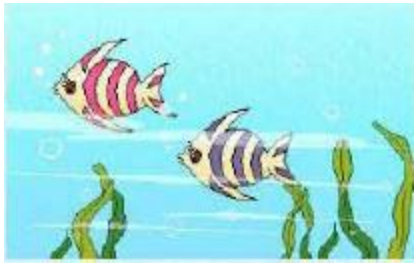
若我们能给出一个区间，在此区间内我们合理地相信N的真值位于其中。这样对鱼数的估计就有把握多了。



引言

也就是说，我们希望确定一个区间，使我们能以比较高的可靠程度相信它包含真参数值。





所说的“可靠程度”是用概率来度量的，称为置信度或置信水平。习惯上把置信水平记作1-α，这里α是一个很小的正数。

推断统计 —区间估计

- 区间估计的原理是样本分布理论。即在进行区间估计值的计算及估计正确概率的解释上，是依据该样本统计量的分布规律和样本分布的标准误差。
- 也就是说，只有知道了样本统计量的分布规律和样本统计量分布的标准误差，才能计算总体参数可能落入的区间长度，才能对区间估计的概率进行解释，可见标准误差及样本分布对于总体参数的区间估计是十分重要的。
- 样本分布可提供概率解释，而标准误差的大小决定区间估计的长度，标准误差越小置信区间的长度越短，而估计成功的概率仍可保持较高水平。一般情况下，加大样本容量可使标准误差变小。

置信水平和置信区间的定义

定义：设θ是一个待估参数，对于给定的实数$\alpha(0<\alpha<1)$，若存在由样本$X_1, X_2, \cdots, X_n$确定的两个统计量$\underline{\theta}$和$\overline{\theta}$，其中，$\underline{\theta}=\underline{\theta}(X_1, X_2, \cdots, X_n)$，$\overline{\theta}=\overline{\theta}(X_1, X_2, \cdots, X_n)$；$(\underline{\theta}<\overline{\theta})$，满足$P\{\underline{\theta}<\theta<\overline{\theta}\}=1-\alpha$,则称区间$(\underline{\theta}, \overline{\theta})$为未知参数$\theta$的置信区间(置信度)，概率$1-\alpha$为置信水平。$\underline{\theta}$和$\overline{\theta}$分别称为置信下限和置信上限。

- 置信水平，也称置信度，表示了区间估计的可靠性、可信度。
- 置信区间，则表示区间估计的精度，$\overline{\theta}-\underline{\theta}$越大表示估计的精度越低。

说明：

由于参数θ的区间估计的意义可以解释为：随机区间$[\underline{\theta}(X_1, X_2, \cdots, X_n), \overline{\theta}(X_1, X_2, \cdots, X_n)]$包含参数$\theta$的真值的概率为1-α，因此可认为“区间$[\underline{\theta}, \overline{\theta}]$包含着参数$\theta$的真值”，则犯错误的概率为α。

由于θ不是随机变量，所以不能说参数θ以1-α的概率落入随机区间$[\underline{\theta}, \overline{\theta}]$，而只能说区间$[\underline{\theta}, \overline{\theta}]$以1-α的概率包含$\theta$。

统计量的构造

可见，对参数θ做区间估计，就是要设法找出两个只依赖于样本的界限（构造统计量）。

$\underline{\theta} = \underline{\theta}(X_1, X_2, \cdots, X_n)$，$\overline{\theta}=\overline{\theta}(X_1, X_2, \cdots, X_n)$; $(\underline{\theta}< \overline{\theta})$

一旦有了样本，就把θ估计在区间$(\underline{\theta}, \overline{\theta})$内。

这里有两个要求：

1. 要求θ以很大的可能被包含在区间$(\underline{\theta}, \overline{\theta})$内，就是说，概率$P\{\underline{\theta} < \theta < \overline{\theta}\}$要尽可能大，即要求估计尽量可靠。

2. 估计的精度要尽可能的高。如要求区间长度$\overline{\theta} - \underline{\theta}$尽可能短，或能体现该要求的其它准则。

可靠度与精度是一对矛盾，一般是在保证可靠度的条件下尽可能提高精度.

单个正态总体参数的区间估计有以下几种情况：

1）σ^2 已知，求 μ 的置信区间。

2）σ^2 未知，求 μ 的置信区间。

两个正态总体参数的区间估计有以下几种情况：

1）已知 $\sigma_1^2=\sigma_2^2$，求 $\mu_1-\mu_2$ 的置信区间。

2）求 $\dfrac{\sigma_1^2}{\sigma_2^2}$ 的置信区间。

以上的区间估计问题都有公式可以直接使用。

（一）方差 σ^2 已知时总体均值的区间估计

设 $x_1, x_2, \mathrm{L}, x_n$ 为来自正态总体 $N(\mu,\sigma^2)$ 的一个样本，其中方差 σ^2 已知，$\bar{x}$ 和 S^2 分别是样本均值和样本方差。

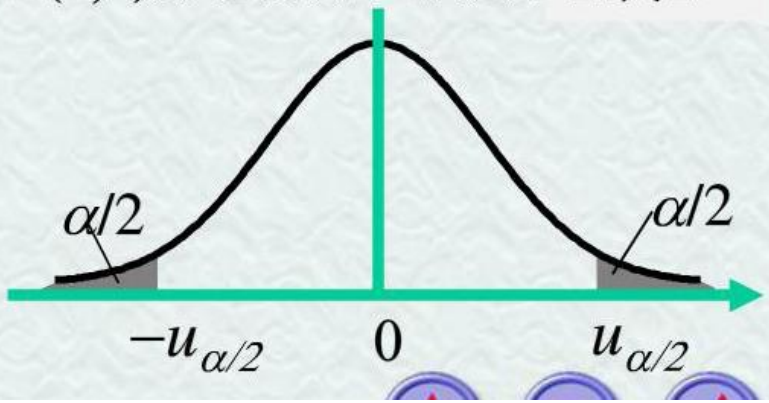
由于 $\bar{x}$ 是总体均值μ的无偏估计，选择统计量

$$u=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

对于给定的置信水平 **1**－α，查 **N(0,1)**分位数表（附表 1或2

得到临界值 $u_{\alpha/2}$，使得

$$P\{|u|<u_{\alpha/2}\}=1-\alpha$$ （图）



即 $$P\{\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right|<u_{\alpha/2}\}=1-\alpha$$

或 $$P\{\bar{x}-u_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}=1-\alpha$$

故总体均值μ的置信水平为 **1**－α的置信区间为

$$\left(\bar{x}-u_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x}+u_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

也可简记为 $\bar{x}\pm u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$。

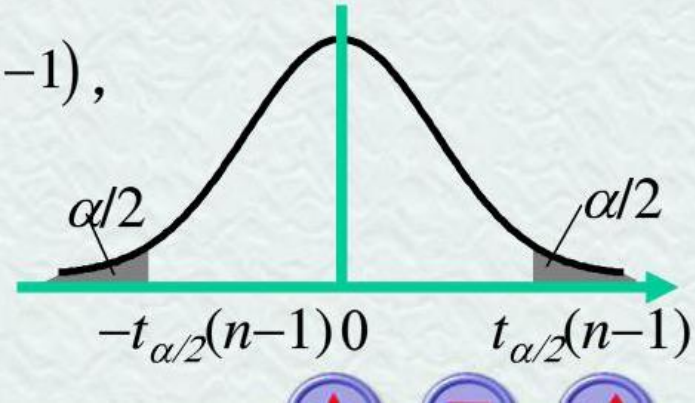
（二）方差 σ^2 未知时总体均值的区间估计

由于总体方差 σ^2 未知，用 σ^2 的无偏估计量——样本方差 S^2 代替 σ^2，可得到统计量

$$T=\frac{\bar{x}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

对于给定的置信度 $1-\alpha$ 和自由度 $n-1$，查 t 分布分位数表（附表 4），可得到临界值 $t_{\alpha/2}(n-1)$，

使得 $$P\{|T|<t_{\alpha/2}(n-1)\}=1-\alpha$$



即 $$P\{\left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right|<t_{\alpha/2}(n-1)\}=1-\alpha$$

或 $$P\{\bar{x}-t_{\alpha/2}\frac{S}{\sqrt{n}}<\mu<\bar{x}+t_{\alpha/2}\frac{S}{\sqrt{n}}\}=1-\alpha$$

故总体均值μ的置信水平为 **1**－α的置信区间为

$$\left(\bar{x}-t_{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{x}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

也可简记为 $\bar{x}\pm t_{\alpha/2}\frac{S}{\sqrt{n}}$。

四、小结

二、单个正态总体$X \sim N(\mu,\sigma^2)$的区间估计：（1- a 为置信度）

待估计参数	条件	构造估计用的统计量	置信区间	查临界值
μ	σ^2已知	$U=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$	$\overline{x}\pm\dfrac{\sigma}{\sqrt{n}}\cdot u_{\frac{\alpha}{2}}$	双侧（表2）查 α
	σ^2未知	$t=\dfrac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$	$\overline{x}\pm\dfrac{S}{\sqrt{n}}\cdot t_{\frac{\alpha}{2}}(n-1)$	$\alpha/2$
σ^2	$\chi^2=\dfrac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$		$\left(\dfrac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},\ \dfrac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\right)$	$\alpha/2$ 和 $1-\alpha/2$

计算步骤：

对于给定的置信度1-α，怎样根据样本来确定未知参数θ的置信区间$(\underline{\theta}, \overline{\theta})$，就是参数$\theta$的区间估计问题。求未知参数$\theta$的置信区间的步骤如下：

(1)构造一个已知其分布的，含有未知参数θ而不含有其他未知参数的样本函数$W = W(X_1, X_2, \cdots, X_n, \theta)$

(2) 对给定的置信度1 -α，根据$W(X_1, X_2, \cdots, X_n, \theta)$的分布定出分位点a 和b，使得$P\{a < W(X_1, X_2, \cdots, X_n) < b\} = 1 - \alpha$。

计算步骤：

(3) 从不等式$a < W(X_1, X_2, \cdots, X_n) < b$中解出$\theta$，得出其等价形式 $\underline{\theta}(X_1, X_2, \cdots, X_n) < \theta < \overline{\theta}(X_1, X_2, \cdots, X_n)$,

这时必有

$P\{\underline{\theta}(X_1, X_2, \cdots, X_n) < \theta < \overline{\theta}(X_1, X_2, \cdots, X_n)\} = 1-\alpha$

于是$(\underline{\theta}, \overline{\theta})$即为$\theta$的置信度为$1-\alpha$的置信区间。

点估计与区间估计的异同

- **相同点:**
  - ✓都可以给出未知参数的估计;
  - ✓估计的准确度都依赖取样的质量。
- **不同点:**
  - ✓ 点估计需要的信息少(矩估计仅需要样本信息)，得到的估计值也比较粗略;
  - ✓ 区间估计需要的信息更多(除了样本，还需要知道总体或样本的某些数字特征的分布形式)，得到的结果是包含置信水平的一个区间.

例1: 设轴承内环的锻压零件的平均高度X服从正态分布$N(\mu, 0.4^2)$。现在从中抽取20只内环，其平均高度为32.3毫米。求内环平均高度的置信度为95% 的置信区间。

σ^2已知	$U=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$	$\bar{x}\pm\frac{\sigma}{\sqrt{n}}\cdot u_{\frac{\alpha}{2}}$	双侧（表2）查 α

解: 1-α= 0.95，查表得$Z_{\alpha/2} = Z_{0.025} = \Phi(0.975) = 1.96$

又$\overline{x} = 32.3$，σ =0.4, $n = 20$，算得

$$\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}=32.3-1.96\times\frac{0.4}{\sqrt{20}}=32.12$$

$$\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}=32.3+1.96\times\frac{0.4}{\sqrt{20}}=32.48$$

所以μ的一个置信度95%的置信区间为(32.12, 32.48)。

例2: 设有一批胡椒粉，每袋净重x (单位: 克）服从正态分布。从中任取8 袋，测得净重分别为: 13.1, 11.9, 12.4，12.3，11.9, 12.1, 12.4, 12.1. 试求μ的置信度为0.99 的置信区间。

σ^2未知 $t=\frac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$ $\overline{x}\pm\frac{S}{\sqrt{n}}\cdot t_{\frac{\alpha}{2}}(n-1)$ $\alpha/2$

解: 这里， l-α=0.99， α =0.01， α /2=0.005, n-l=7

经计算$\overline{x}$ =12.275， s=0.3882

查表可得

$$t_{\alpha/2}(n-1)=t_{0.005}(7)=3.4995$$

从而，

$$\overline{x}-\frac{s}{\sqrt{n}}t_{\alpha/2}(n-1)=12.275-\frac{0.3882}{\sqrt{8}}\times 3.4995=11.80$$

$$\overline{x}+\frac{s}{\sqrt{n}}t_{\alpha/2}(n-1)=12.275+\frac{0.3882}{\sqrt{8}}\times 3.4995=12.75$$

所以μ的置信度为0.99 的置信区间为（11.80, 12.75）

例3：在一项关于软塑料管的实用研究中，工程师们想估计软管所承受的平均压力。他们随机抽取了9个压力读数，样本均值和标准差分别为3.62kg 和0.45。假定压力读数近视服从正态分布，试求总体平均压力的置信度为0.99 时的置信区间。

解：因为 $\dfrac{\overline{X}-\mu}{\dfrac{S}{\sqrt{n}}}\sim t(n-1)$

σ^2未知 $t=\dfrac{\overline{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$ $\overline{x}\pm\dfrac{S}{\sqrt{n}}\cdot t_{\frac{\alpha}{2}}(n-1)$ $\alpha/2$

所以，

$$P\left\{-t_{\frac{\alpha}{2}}(n-1)\le\frac{\overline{X}-\mu}{\dfrac{S}{\sqrt{n}}}\le t_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

于是，总体压力平均压力μ的

1-α置信区间为：

$$\left[\overline{x}-\frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1),\overline{x}+\frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right]$$

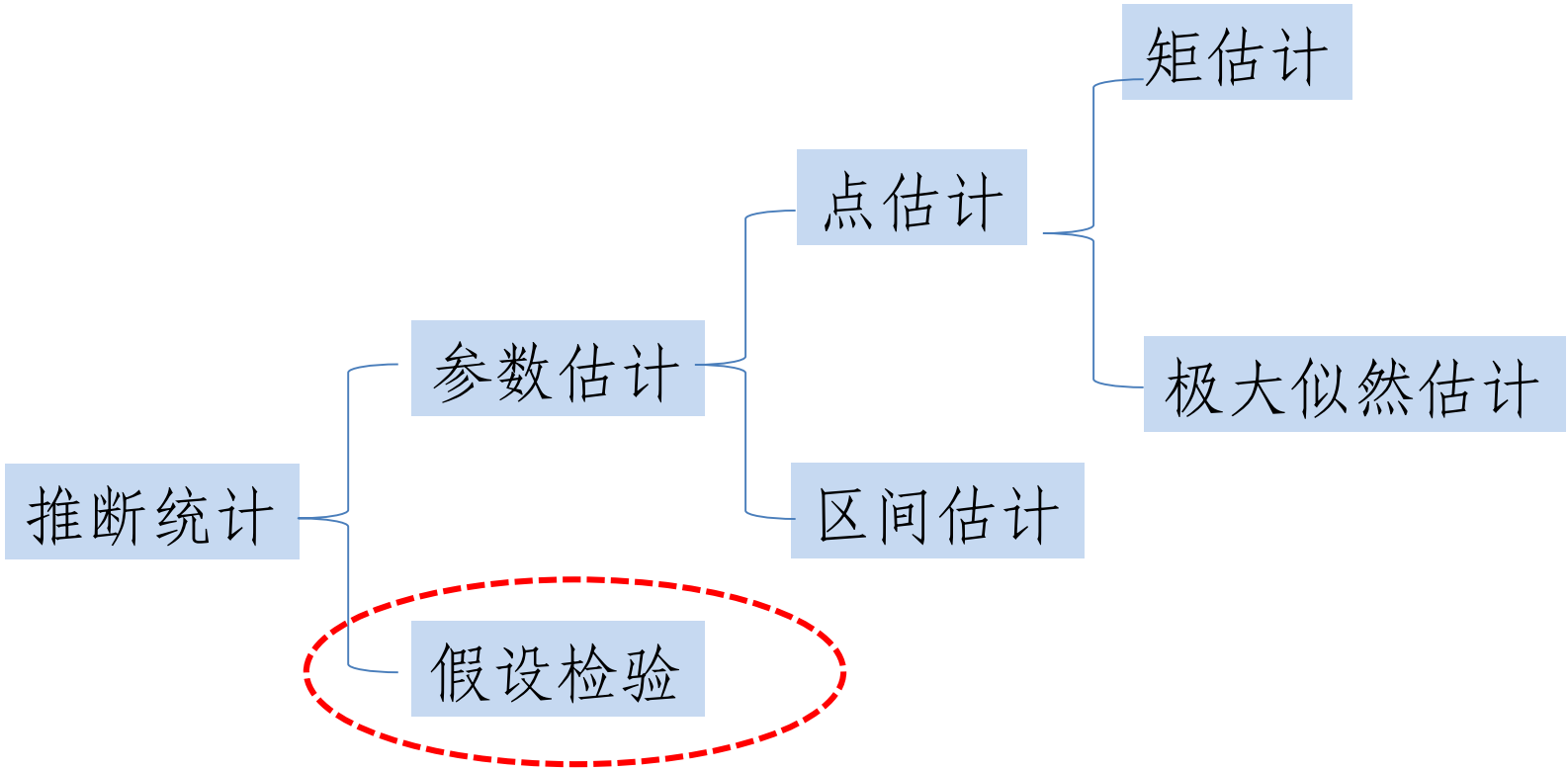
由题意知，$n=9, \bar{x}=3.62, s_{n-1}=0.45, 1-\alpha=0.99, t_{\alpha/2}(n-1)=t_{0.005}(8)=3.3554$,

代入上式，得总体平均压力μ的99%置信区间为

$$\left[3.62-\frac{0.45}{\sqrt{9}}\times 3.3554, 3.62+\frac{0.45}{\sqrt{9}}\times 3.3554\right]$$

$=[3.12, 4.12]$

推断统计 —假设检验



推断统计 —假设检验

什么是假设? (hypothesis)

对总体参数的数值所作的一种陈述。

- 总体参数包括总体均值、比例、方差等
- 分析之前必需陈述

什么是假设检验?(hypothesis testing)

- 事先对总体参数或分布形式作出某种假设，然后利用样本信息来判断原假设是否成立(根据样本的信息检验关于总体的某个命题是否正确)。

假设检验

根据一定假设条件由样本推断总体的一种方法。

引例：设一箱有红白两种颜色的球共100个，甲说这里有98个白球，乙从箱中任取一个，发现是红球，问甲说的是否正确?

分析：先作假设H_0: 箱中确有98个白球。

如果假设H_0是正确，则从箱中任取一个球是红球的概率只有0.02，是小概率事件。通常认为在一次随机试验中概率小的件不易发生，因此，若乙从箱中任取一个，发现是白球，则没有理由怀疑假设H_0的正确性。今乙从箱中任取一个，发现是红球，即小概率事件竟然在一次试验中发生了，故有理由拒绝假设H_0，即认为甲的说法不正确。

假设检验的基本思想

- 假设检验的基本思想实质上是带有某种概率性质的反证法。为了检验一个假设H_0是否正确：
  - 首先假设该假设H_0正确；
  - 然后根据抽取到的样本对假设H_0作出接受或拒绝的决策；
  - 如果样本观察值导致了不合理的现象发生，就应拒绝假设H_0，否则应接受假设H_1。

假设检验的基本思想

- 假设检验中所谓的“不合理”，并非逻辑中的绝对矛盾，而是基于人们在实践中广泛采用的原则，即小概率事件在一次实验中是几乎不发生的。
- 但概率到什么程度才能算作“小概率事件”，“小概率事件”的概率越小，否定原假设H_0就越有说服力。常记这个概率值为$\alpha(0<\alpha<1)$，称为检验的显著性水平。对不同的问题，检验的显著性水平α不一定相同，但一般应取为较小的值，如0.1、0.05或0.01等

什么是小概率？

- 概率是从0 到1 之间的一个数，因此小概率就应该是接近0的一个数。
- 著名的英国统计家Ronald Fisher 把1/20作为标准，这也就是0.05，从此0.05或比0.05小的概率都被认为是小概率。
- Fisher没有任何深奥的理由解释他为什么选择0.05，只是说他忽然想起来的。

什么是小概率？

- 在一次试验中， 一个几乎不可能发生的事件发生的概率。
- 在一次试验中小概率事件一旦发生，我们就有理由拒绝原假设。
- 小概率由研究者事先确定。

检验假设的步骤

1. 提出假设（原假设H_0和备择假设H_1）；
2. 确定适当的检验统计量;
3. 规定显著性水平α和计算检验统计量的值;
4. 做出结论。

1. 提出假设（原假设H_0和备择假设H_1）

- 根据问题的需要对所研究总体做出某种假设，记作H_0。
- 原假设(H_0)：又称零假设，它是接受检验的假设，没有充分根据，是不会被轻易否定的。
- 备择假设(H_1)：是拒绝原假设后可供选择的假设。
- 原假设和备择假设是相对立的，在任何情况下只能有一个成立。如果接受H_0就必须拒绝H_1；拒绝H_0就必须接受H_1。
- 若原假设$H_0: \mu = \mu_0$，根据具体问题，备择假设有三种选择:

  (1)备择假设H_1：$\mu \neq \mu_0$，称为双侧检验。

  (2)备择假设H_1：$\mu > \mu_0$，称为右侧检验。

  (3)备择假设H_1：$\mu < \mu_0$，称为左侧检验。

2. 确定适当的检验统计量

- 样本统计量即样本特征值，也称检验统计量。
- 选取合适的统计量（这个统计量的选取要使得在假设H_0成立时，其分布为已知）
- 所设计的检验统计量应与原假设相关，即与待检验的参数相关，选择的方法与参数估计相同，需考虑：
  - ✓ 是大样本还是小样本
  - ✓ 总体方差已知还是未知
- 基本形式为：$$Z=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$$

3. 显著性水平α和相应的临界值

- 显著性水平α
  - ✓ 是一个概率值，
  - ✓ 原假设H_0成立条件下，规定的小概率的标准
  - ✓ 被称为抽样分布的拒绝域
  - ✓ 一般取值很小，常取0.1、0.05、0.01等
  - ✓ 由研究者事先确定。
- 临界值
  - ✓ 由有关的概率分布表查得，从而可确定H_0的接受域和拒绝域。
  - ✓ 临界值就是接受域和拒绝域的分界点。
- 注意:
  - ✓ 同一显著性水平， 选择不同的检验统计量，得到的临界值是不同的:
  - ✓ 同一显著性水平和同一的统计量， 双侧检验和单侧检验的临界值也是不同的。

4. 做出结论

将临界值与由样本资料计算出的检验统计量的数值比较，视统计量落入接受域还是拒绝域，做出接受或拒绝原假设H_0的结论。

- 计算检验的统计量;
- 根据给定的显著性水平α，查表得出相应的临界值Z_α或$Z_{\alpha/2}$，t_α或$t_{\alpha/2}$;
- 将检验统计量的值与α水平的临界值进行比较;
- 得出拒绝或不拒绝原假设的结论。

两种常用的假设检验方法

– *t*检验和*u*检验，其适用条件如下

检验方法	t检验	u检验
适用条件	1. 单因素小样本（n<50）数据 2. 样本来自正态分布 3. 总体标准差未知 4. 两样本均数比较时，要求两样本对应的总体方差相等	1. 大样本 2. 样本小，但总体标准差已知

– 下面主要阐述*t*检验，包括单样本*t*检验、两个独立样本均数*t*检验和配对样本均数*t*检验（非独立两样本均数*t*检验）。

1. 单样本t检验

- 使用前提
  - 只有一个总体，并且总体呈正态分布。
- 适用场合
  - 检验总体均值是否与给定的值存在显著差异（不相等）。
- 检验过程
  - H_0: $\mu=\mu_0$ 可以认为样本是从已知总体中抽取的
  - 令$t=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$，其中$S$为样本方差。
  - 计算t值，然后与$t_{\left(\frac{\alpha}{2}\right),n-1}$（查表可得）比较大小。如果$t$值较小，则拒绝原本假设。

1. 单样本t检验

例1：以往通过大规模调查已知某地新生儿出生体重3.30kg，从该地难产儿中随机抽取35名新生儿作为研究样本，平均出生体重为3.42kg，标准差为0.40kg，问该地难产儿出生体重是否与一般新生儿体重不同?

解：经过分析，已知总体均数μ_0=3.30kg，尽管知道研究样本的标准差S=0.40kg，但总体标准差σ未知，而且n=35为小样本，故选用单样本t检验。

1. 单样本t检验

检验的步骤如下:

(1) 建立检验假设，确定检验水准

$H_0: \mu = \mu_0$，该地难产儿与一般新生儿平均出生体重相同;

$H_1: \mu \neq \mu_0$，该地难产儿与一般新生儿平均出生体重不同;

(2) 计算检验统计量

在$\mu = \mu_0$成立的前提条件下，计算统计量为:

$$t = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{3.42 - 3.30}{0.40/\sqrt{35}} = 1.77$$

确定P值，作出推断结论。

1. 单样本t检验

检验的步骤如下:

（3）通过查表得知 $t_{0.05/2,\ 34}=2.032$，因为$t < t_{0.05/2,\ 34}$，故$P>0.05=\alpha$，则根据检验水准$\alpha=0.05$，不拒绝H_0

该差别无统计学意义，根据现有样本信息，尚不能认为该地难产儿与一般新生儿平均出生体重不同。

2. 两个独立样本均数t检验

- 使用前提
  - 来自两个总体的独立样本，两个样本所代表的总体均服从正态分布，并且两个总体方差相同，而两组样本数量可以不同。
- 目的
  - 考察两个总体的均值是否存在显著差异。
- 检验过程
  - H_0: $\mu_1=\mu_0$（不存在差异）
  - $t=\frac{|\overline{X}_1-\overline{X}_2|}{S_{\overline{X}_1-\overline{X}_2}}$，其中$S_{\overline{X}_1-\overline{X}_2}=\sqrt{S_c^2(\frac{1}{n_1}+\frac{1}{n_2})}$而$S_c^2=\frac{\sum X_1^2-\frac{(\sum X_1)^2}{n_1}+\sum X_2^2-\frac{(\sum X_2)^2}{n_2}}{n_1+n_2-2}$

检验过程中，具体计算t值，和$t_{(\alpha/2),(n_1+n_2-1)}$比较大小。计算的$t$值小的话拒绝原本假设。

2. 两个独立样本均数*t*检验

该方法可用于判断两个样本是否来自不同总体，即是否不同：因素作用在另一组后，判断因素是否起作用（使其不再来自原来总体）。

2. 两个独立样本均数*t*检验

例子：25例糖尿病患者随机分成两组，甲组单纯用药物治疗，乙组采用药物治疗合并饮食疗法，两个月后测空腹血糖，如下表所示，问两种疗法治疗后患者血糖值是否相同？

编号	甲组血糖值(X_1)	编号	乙组血糖值（ ）
1	8.4	1	5.4
2	10.5	2	6.4
3	12.0	3	6.4
4	12.0	4	7.5
5	13.9	5	7.6
6	15.3	6	8.1
7	16.7	7	11.6
8	18.0	8	12.0
9	18.7	9	13.4
10	20.7	10	13.5
11	21.1	11	14.8
12	15.2	12	15.6
		13	18.7

25名糖尿病患者两种疗法治疗后二个月血糖值

2. 两个独立样本均数t检验

可知：

甲组：$n_1=12, \overline{X_1}=15.21$　　乙组：$n_2=13, \overline{X_1}=10.85$

检验步骤如下：

（1）建立检验假设，确定检验水准

$H_0:\mu=\mu_0$，两种疗法治疗后患者血糖值的总体均数相同；

$H_1:\mu\neq\mu_0$，两种疗法治疗后患者血糖值的总体均数不同；

α=0.05。

2. 两个独立样本均数t检验

检验步骤如下:

（2）计算检验统计值。

由原始数据可得:

$$n_1 = 12, \sum X_1 = 182.5, \sum X_1^{\,2} = 2953.43$$

$$n_2 = 13, \sum X_1 = 141.0, \sum X_2^{\,2} = 1743.16$$

$$\overline{X_1} = \frac{\sum X_1}{n_1} = \frac{182.5}{12} = 15.21 \qquad \overline{X_2} = \frac{\sum X_2}{n_2} = \frac{141.0}{13} = 10.85$$

2. 两个独立样本均数*t*检验

代入公式，得到:

$$S_C^2 = \frac{2953.43 - \frac{(182.5)^2}{12} + 1743.16 - \frac{(141.10)^2}{13}}{12+13-2} = 17.03$$

$$S_{\overline{X}_1 - \overline{X}_2} = \sqrt{17.03(\frac{1}{12} + \frac{1}{13})} = 1.652$$

按公式计算，得:

$$t = \frac{15.21 - 10.85}{1.652} = 2.639$$

2. 两个独立样本均数t检验

检验步骤如下:

（3）确定P值，作出推断结论

查表可知：$t_{0.05,(23)}=2.069$

由于$t>t_{0.05/2,(23)}$，$P<0.05$，按$\alpha=0.05$的水准拒绝H_0，接受H_1，有统计学意义。故可认为该地两种疗法治疗糖尿病患者二个月后测得的空腹血糖值的均数不同。

3. 配对样本均数*t*检验

来自两个总体的配对样本，两两配对样本*t*检验应用的前提与单样本*t*检验类似，只是抽样不是独立的，而是两两配对相互关联的。

- 使用前提
  - 两组样本数量相同，并且两组样本的观测值是一一对应的。
- 目的
  - 考察两个总体的均值是否存在显著差异。
- 检验过程
  - H_0: $\mu_1-\mu_0=0$(总体均值不存在差异)

3. 配对样本均数t检验

- 检验过程
  - 首先计算各对数据间的差值d，将d作为变量计算均数，可以将该检验理解为差值样本均值与已知总体均数$\mu_d(\mu_d = 0)$比较的单样本t检验，公式为：

$$t = \frac{\bar{d} - \mu_d}{S_{\bar{d}}} = \frac{\bar{d} - 0}{S_{\bar{d}}} = \frac{\bar{d}}{S_d/\sqrt{n}}$$

  - 具体计算t值，和$t_{(\alpha/2),(n-1)}$比较大小。如果t值较小，拒绝原本假设。

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

3. 配对样本均数*t*检验

例子：有12名接种卡介苗的儿童，8周后用两批不同的结核菌素，一批是标准结核菌素，一批是新制结核菌素，分别注射在儿童的前臂，两种结核菌素的皮肤侵润反应平均直径如下表所示，问两种结核菌素的反应性有无差别。

编号	标准品	新制品	差值d	d^2
1	12.0	10.0	2.0	4.00
2	14.5	10.0	4.5	20.25
3	15.5	12.5	3.0	9.00
4	12.0	13.0	-1.0	1.00
5	13.0	10.0	3.0	9.00
6	12.0	5.5	6.5	42.25
7	10.5	8.5	2.0	4.00
8	7.5	6.5	1.0	1.00
9	9.0	5.5	3.5	12.25
10	15.0	8.0	7.0	49.20
11	13.0	6.5	6.5	42.25
12	10.5	9.5	1.0	1.00
合计			39($\sum d$)	195($\sum d^2$)

12名儿童分别用两种结核菌素的皮肤侵润反应结果

3. 配对样本均数t检验

检验步骤如下:

（1）建立检验假设，确定检验水准

$H_0: \mu_d = 0$，两种结核菌素的皮肤侵润反应总体平均直径差异为0;

$H_1: \mu_d \neq 0$，两种结核菌素的皮肤侵润反应总体平均直径差异不为0；

α=0.05。

3. 配对样本均数t检验

检验步骤如下:

（2）先计算差值d及d^2，如表第4、5列所示，$\Sigma d=39$, $\Sigma d^2=195$

计算差值的标准差 $$S_d=\sqrt{\frac{\sum d^2-\frac{\left(\sum d\right)^2}{n}}{n-1}}=\sqrt{\frac{195-\frac{(39)^2}{12}}{12-1}}=2.4909$$

计算差值均值的标准差 $$S_{\bar{d}}=\frac{S_d}{\sqrt{n}}=\frac{2.4909}{3.464}=0.7191$$

按公式计算，得 $$t=\frac{\bar{d}}{S_{\bar{d}}}=\frac{3.25}{0.7191}=4.5195$$

3. 配对样本均数t检验

检验步骤如下:

（3）确定P值，作出推断结论

查表可知：$t_{0.05/2,(11)}=2.201$，因为$t>t_{0.05/2,(11)}$，$P<0.05$，按$\alpha=0.05$的水准，拒绝H_0，接受H_1，差异有统计学意义。可以认为两种方法皮肤浸润反应结果不同。

本章小结

1. 大数据分析建模

2. 统计方法
   - 描述统计
     - 集中趋势分析
     - 离中趋势分析
     - ※ 相关分析
   - 推断统计
     - 参数估计
       - 点估计
         - ※ 矩估计
         - ※ 极大似然估计
       - ※ 区间估计
     - ※ 假设检验