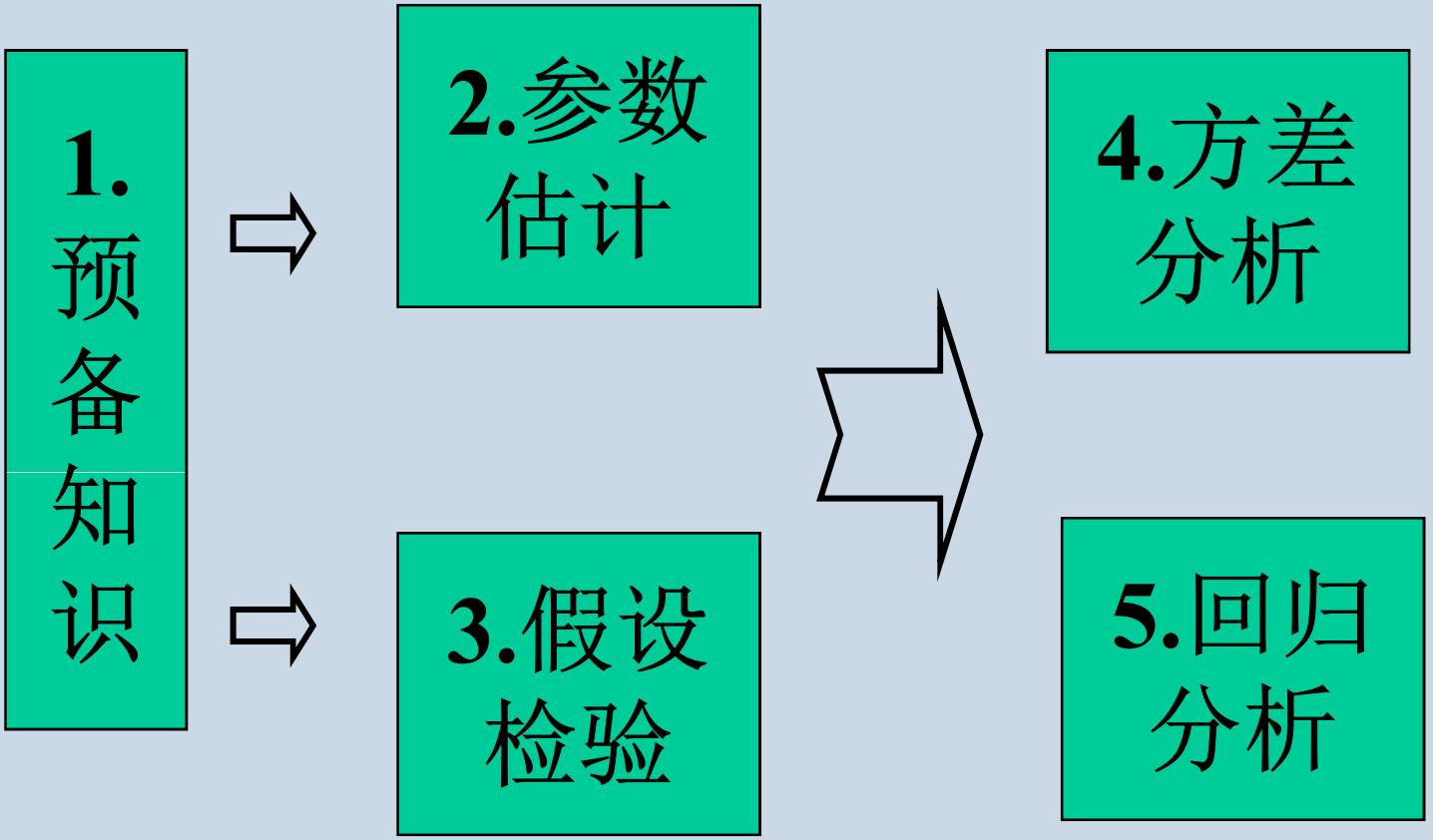
《应用数理统计》

孙　平
东北大学数学系

plsun@mail.neu.edu.cn


1.预备知识
2.参数估计
3.假设检验
4.方差分析
5.回归分析

第1章　预备知识

第1.1节　基本概念与主要内容

第1.2节　概率论基础

第1.3节　统计量与抽样分布

统计学 (***Statistics***) 是一门收集与分析数据，并且根据数据进行推断的艺术与科学。

———————— 《大英百科全书》

统计学理论主要包含三个部分：

1.数据收集，**2.**数据分析，**3.**由数据做出决策。

(数理) 统计学中的数据都是随机数据。

统计学的任务就是在随机性中去寻找规律。

一. 统计学的基本概念

1. 总体与个体 (*population*)

统计学中把所研究的对象全体称为总体，总体中的每一个元素称为一个个体。

总体与个体都用数量指标来表示

即使面临的是一个定性的实际问题，也必须把有关的资料定量化。

例如总体分成：抽烟与不抽烟两类。

1 表示 抽烟者； **0** 表示 不抽烟者。

2. 样本 (*sample*)

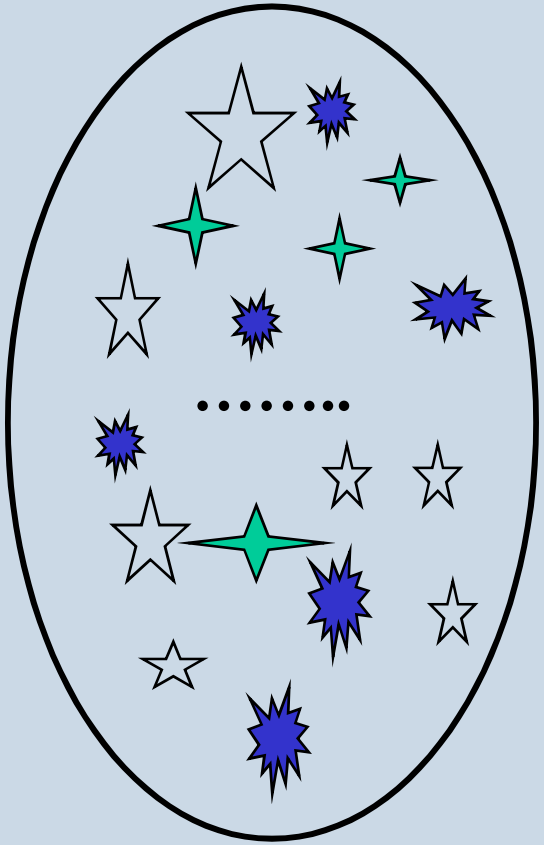
从总体中取出一个个体，称为从总体中得到一个样本。

由于各种原因与实际条件的限制，不可能得到一个总体中所有个体的数据。即样本总是总体的一小部分。

但同时在直观上又认为、或者希望做到：抽取出的每个个体 (样本) 都充分蕴涵总体信息。

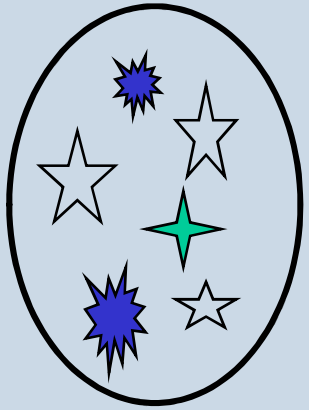
统计学的目的就是从样本去得出总体的信息。

总体



被研究的对象全体

样本



具有代表性的
部分个体

定义1.1.1　X 是具有分布函数 F 的一个随机变量，如果 X_1，X_2，…，X_n 是有同一分布函数 F 的相互独立的随机变量，则称：

X_1，X_2，…，X_n 是从总体 F（总体 X）中得到的容量为 n 的简单随机样本，简称为 样本。

这些样本随机变量各自具体的取值：

x_1，x_2，…，x_n

称为是总体随机变量 X 的样本观察值。

样本的函数称为是统计量。

总体被认为是一个服从某种概率分布 F 的随机变量。

总体分布 F 可以是未知的， 非参数统计学

总体分布 F 的类型已知，但是含有一些未知的参数。 参数统计学

样本是和总体随机变量有相同分布 F 的随机变量，样本的个数称为样本容量， n 。

独立同分布的样本称为简单随机样本。

二. 数理统计学的主要内容

1. 抽样理论：介绍如何收集数据。主要抽样方法，样本容量的确定，抽样误差，敏感问题等

2. 参数估计：如何根据数据得到总体参数信息。点估计、区间估计，**Bayes** 估计等

3. 假设检验： 如何对关于总体的一些假设做出决策。正态总体参数的检验，分布拟合检验，秩检验，列联表，统计决策等理论

4. 方差分析与回归分析：变量间效应关系。

方差分析 — 分类变量与数值变量的效应关系

回归分析 — 研究数值变量之间的效应关系

5. 多元分析： 研究若干个变量之间的关系

聚类分析、判别分析、主成分分析、

因子分析、典型相关分析等等

基本内容介绍

问题一： 希望了解某城市家庭月支出情况。

解决方法：从这个城市里随机地调查有代表性的一些家庭，根据收集到的数据去得出这个城市家庭每个月支出的有关信息。

1. 如何得到样本？ 抽样调查

不同阶层背景的家庭比例应该各占多少？样本容量应该取多少才合适？被调查者拒绝调查怎么办？

2. 如何确定总体的分布？

根据经验或者是所讨论的问题的实际背景，总体的分布类型一般可以事先确定下来。

这里的总体是这个城市的家庭月支出费用，我们有充分理由认为家庭月支出费用是一个服从正态分布的随机变量。

即，总体随机变量 $X \sim N(\mu, \sigma^2)$，而这个城市相应的两个参数 μ 与 σ^2 是未知的。

（不同城市对应的这两个参数也就不相同）

Remark

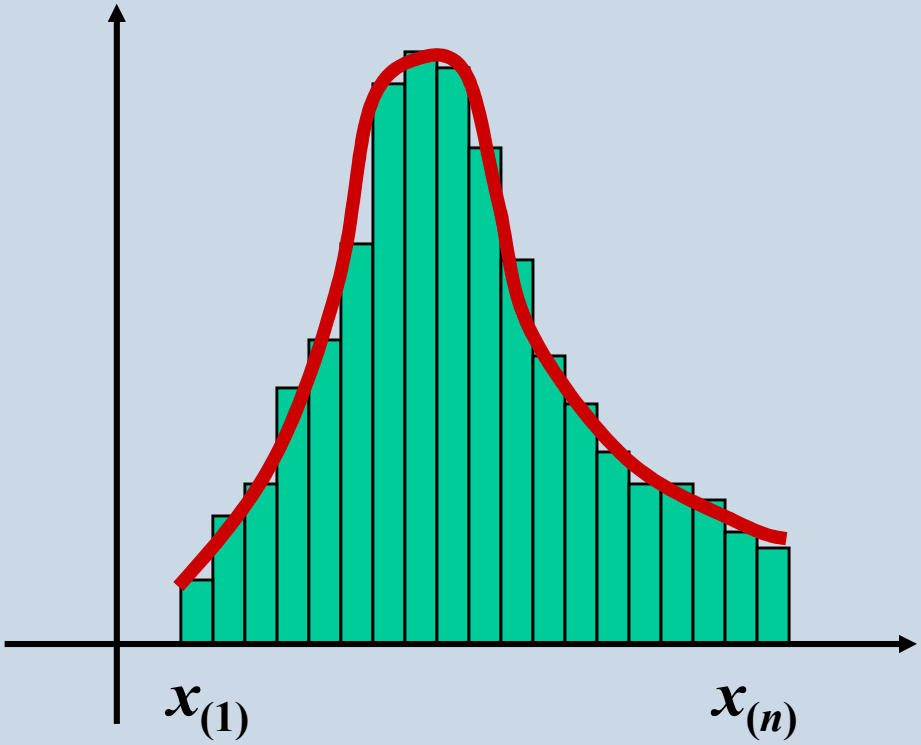
当不知道或者难以确定总体的分布类型时，在统计学中常常采用下面两种办法来近似得到总体分布的有关信息。

(1). 直方图的方法

只适用连续总体，得到的是总体密度函数近似。

把收集到的 n 个数据 x_1，x_2，…，x_n 从小到大排列：$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$；其次取区间 (a, b)，包含全部数据 $a < x_{(1)}$，$x_{(n)} < b$；

把 (a, b) 等分成若干小区间，计算每个小区间中包含的数据的频率。



根据这些频率做出相应的小区间上的矩形，则当 n 充分大时，这些小区间上矩形的面积将近似于总体的概率密度函数下曲边梯形的面积。

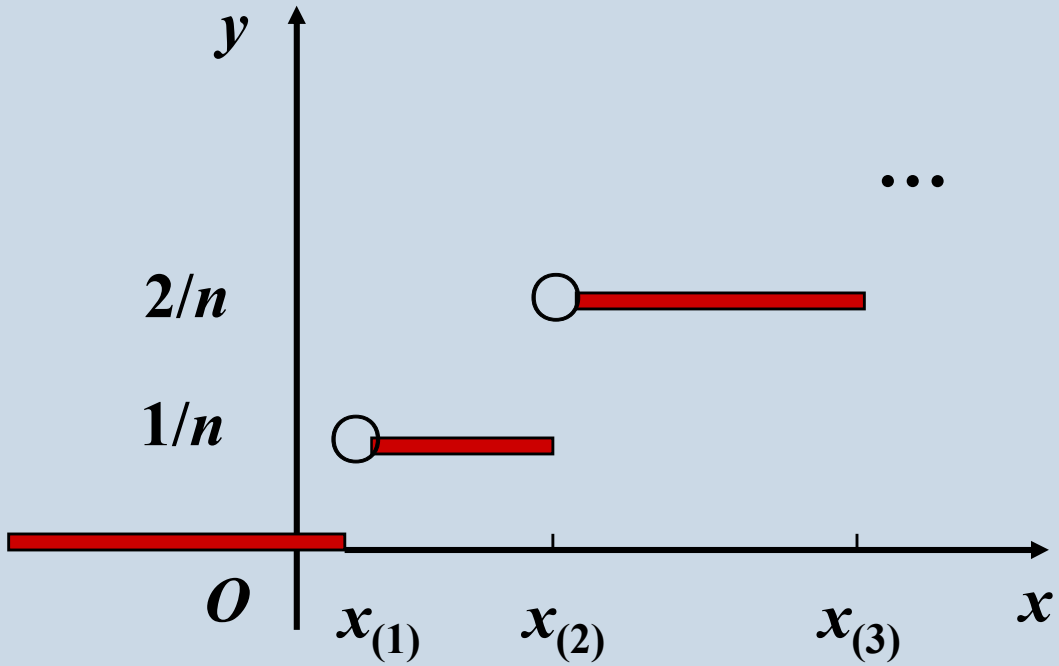
(2). 经验分布函数的方法

构造一个分布函数，得到的是总体分布函数 $F(x)$ 的近似。

$$F_n(x)=\begin{cases}0, & x\le x_{(1)}\\ \dfrac{k}{n}, & x_{(k)}<x\le x_{(k+1)}\\ 1, & x>x_{(n)}\end{cases}$$

这个函数实际上是观察值 x_1，…，x_n 中小于 x 的频率，即

$$F_n(x)=\frac{1}{n}\sum_{k=1}^{n}1\{X_{(k)}<x\}$$



可以证明，经验分布函数 $F_n(x)$ 将依概率、甚至是几乎处处收敛到 $F(x) = \mathrm{P}(\mathrm{X} < x)$。

3. 如何从样本得出总体的信息？

样本是一组与总体独立、同分布的随机变量，我们得到的数据是样本观察值，而不是样本。

调查一个家庭得到了一个数据，相当于对总体分布做了一次随机试验而观察到了这个随机变量的具体取值。

一共有 n 个数据，相当于对总体分布做了 n 次独立重复试验，而得到了这个总体随机变量在这些试验中的具体取值。

参数估计

数理统计学最重要的内容之一

利用样本观察值去估计出总体的未知参数

直观上可以利用调查到的 n 个家庭的月支出 $x_1, x_2, \ldots, x_n$ 的算术平均：

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

去估计这个城市家庭的平均月支出费用 μ。

它的合理性在哪？ 还有没有其它的办法？

这些不同的方法各有什么样的优缺点？

假设检验　　数理统计学最重要的内容之一

事先提出一个假设，利用样本观察值去检验这个假设是否可以被接受

政府和企业共同关心的一个问题：

$$\mu > \mu_0 \ ?$$

这里 μ_0 是一个已知的常数。

应该如何去做这个检验？

一种想法是：既然已经通过参数估计得到了这个城市家庭月平均支出 (即总体的参数 μ) 的估计值，自然就可以用它代替假设里的 μ 去做检验：

当估计值比 μ_0 大就接受这个假设，否则就拒绝

但是这样风险很大：样本总是随机得到的，因此估计值与真实值之间不可避免地存在着随机误差。

传统的方法是：给出一个区域 (拒绝域)，如果估计值落在这个区域内，就拒绝原来的假设，否则就接受。

除了对总体参数的检验外，还有一些重要的假设检验问题，例如：

关于总体分布的检验　　分布拟合检验

检验得到的样本数据是不是来自于某个事先给出的总体

独立性的检验

检验一些分类变量之间是否是独立的，例如：

抽烟与肺癌，睡觉打鼾与心脏病…

关于数据差异的检验

主要希望了解两组或多组数据间的差异究竟是来自于随机性，还是总体间的确存在差异？

例如：

病毒疫苗是否有效，

抽签的结果是否公平、合理？

…

以及我们在科学研究、工程实践、社会调查等等得到的数据

回归与相关分析　数理统计学重要应用之一

讨论数值变量之间的效应关系问题

一元线性回归

比如说，想了解儿子身高与父亲身高之间的关系。
在每个被调查的家庭中同时获得这两个变量的观察值，分析它们是否有某种(函数)关系，…

多元线性回归

例如，人口自然增长率与国民收入、消费价格指数、房价等等是否有关？关系如何？…

方差分析　　数理统计学重要应用之一

讨论分类变量与数值变量之间的关系

单因素方差分析

小麦产量与不同品种之间的关系。
是否某些品种的产量偏高？如果偏高，这种差异是否纯属偶然原因，…

双因素方差分析

品种与施肥等级这两个因素联合对小麦产量的影响。各自单独是否有影响？交互效应如何？…

简单的说，从概率论的角度出发，

可以把上述数理统计学的过程理解成：

有一个含有未知信息的概率分布 F

针对 F 做了 n 次独立重复的试验与观察，
得到 n 个独立同分布于 F 的随机变量的取值

根据样本的具体观察值，去推断出总体 F
所包含的未知信息，或作出进一步的决策等

问题二：如何分析与处理变量之间的关系？

简单 → 复杂

分类变量：如性别、信仰、职业等等，

顺序变量：如名次(第一、第二，…)，

数值变量：如收入、比例、产量等等

Remark

可以把复杂的变量简化为简单变量，反之不行

数值变量 → 顺序变量 → 分类变量

变量组合与相应的统计分析方法

		自变量 x		
		分类变量	顺序变量	数值变量
因变量 y	分类变量	卡方分析	←	回归与相关
	顺序变量	↑	秩方法	←
	数值变量	方差分析	↑	回归与相关

三. 常用统计软件简介

利用统计方法去处理数据时，有两个必须要解决的问题：

(1) 数据量太大，因此计算复杂、繁琐；
(2) 能够应用的方法很多，因此需要反复比较不同的统计方法，找出综合的解决方案。

统计软件包(***Statistical Package***) 涵盖了应用广泛、使用频率很高的各种统计方法，是针对统计数据的特点而专门设计的软件包。

1. *SPSS*

Statistical Package for the Social Science

(社会科学统计软件包)

⇩

Statistical Product and Service Solutions

(统计产品与服务解决方案)

用户遍布于通讯、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个领域和行业，是世界上应用最广泛的专业统计软件。

2. *SAS*

Statistical Analysis System
(统计分析系统软件包)

广泛应用于经济管理、社会科学、生物医学、质量控制、以及政府和教育科研等领域，

在数据处理和统计分析领域，***SAS*** 被誉为国际上的标准软件系统。

3. *EXCEL* 统计函数

计算统计量：

***AVERAGE*，*MEDIAN*，*VAR*，*CORREL* , …**

计算区间点：***TINV*，*CHIINV*，…**

计算概率(*p*-值)：***NORMSDIST*，*CHIDIST*，*TDIST*，*FDIST*，…**

回归分析：***LINEST*，……**


苏格兰羊

习题 1.1

1. 收集生活与学习、工作中的一些统计数据。

2. 证明经验分布函数$F_n(x)$的收敛性 。

第1.2节 概率论基础

一. 随机事件 A

1. 可能发生、也可能不发生的事件

2. 事件的关系

包含、不相容、独立

3. 事件的运算

和事件、交事件、差事件、对立事件

二. 概率及有关公式

1. 概率 P(*A*)

随机事件在一次试验中发生的可能性

频率定义、主观概率

概率的数学定义：

样本空间中的一些子集到实数轴的一个集合函数，满足：非负性、规范性、可列可加性

2. 条件概率 $\mathbf{P(}\boldsymbol{B}\mathbf{|}\boldsymbol{A}\mathbf{)}$

3. 概率计算的一些公式

加法公式

减法公式

乘法公式

全概率公式

Bayes 公式

三. 随机变量及分布

1. 随机变量 X：离散型、连续型

样本空间到实数轴的函数

分布律与概率密度函数 $p(x,\theta)$

2. 随机变量与随机事件的关系

$(a \le X < b)$ 是一个随机事件；

A 是否发生可以通过两点分布表示

3. 分布函数 $F(x)$

也就是概率：$\mathbf{P}(X<x)$

离散随机变量的分布函数是阶梯型跳跃函数，

对满足$(x_k<x)$的所有 p_k 求和得到。

连续随机变量的分布函数是(0,1) 之间的非降单调函数：

$$F(x)=\int_{-\infty}^{x} p(y)dy$$

4. 重要的离散分布

两点分布：背景、分布律、期望、方差；
二项分布：背景、分布律、期望、方差；
泊松分布：背景、分布律、期望、方差。

5. 重要的连续分布

均匀分布：背景、密度函数、期望、方差；
指数分布：背景、密度函数、期望、方差；
正态分布：背景、密度函数、期望、方差。

Gamma 分布 $\Gamma(\alpha, \lambda)$

$$p(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0$$

这里参数 $\alpha > 0$， $\lambda > 0$；

$\Gamma(\alpha)$ 是Gamma 积分，$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$， $\alpha > 0$

1. 参数λ 的指数分布就是 $\Gamma(1, \lambda)$ 。

2. 自由度 n 的卡方分布 $\chi^2(n)$ 就是 $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ 。

3. Gamma 分布 $\Gamma(\alpha, \lambda)$ 对于α 具有可加性；
   而且如果 $X \sim \Gamma(\alpha, \lambda)$ ， 则 $cX \sim \Gamma(\alpha, \lambda / c)$

四. 随机向量

1. 联合分布函数、联合分布律、联合密度

2. 从联合分布到边缘分布

3. 随机变量的独立性

   两个离散随机变量的独立性

4. 二维正态与多元正态分布

5. 条件分布：条件概率的推广

从 $p(x|\theta)$ 到 $h(\theta|x)$

θ是一个具有分布$h(\theta)$的随机变量，如果X关于θ具有条件分布$p(x|\theta)$，则X与θ的联合分布是$h(\theta)\times p(x|\theta)$。

把联合分布对θ积分或求和得到X的边缘分布，再用联合分布除以X的边缘分布从而能够得到θ关于X的条件分布$h(\theta|x)$。

6. 独立同分布随机变量的和

正态分布的可加性

二项分布的可加性

卡方分布的可加性

五. 数字特征

1. 数学期望 $E(X)$

随机变量取值的加权平均

期望计算的公式：

线性变换的期望、和的期望、
乘积的期望、随机变量函数的期望。

2. 方差 $D(X)$

随机变量在期望附近取值的分散程度

方差计算的公式：

线性变换的方差、独立与一般和的方差。

3. *Chebyshev* 不等式

4. 协方差 $Cov(X, Y)$

刻化两个随机变量之间的相依关系

5. 相关系数

刻化两个随机变量之间线性关系的程度

6. 随机向量的数字特征

期望向量

协方差矩阵

7. 条件数学期望

① 离散随机变量的条件期望

Y 关于随机事件 $(X=x_i)$ 的条件期望：

$$E(Y \mid X = x_i) = \sum_{j=1} y_j \times p(Y = y_j \mid X = x_i)$$

Remark

Y 关于 X 的条件期望 $E(Y|X)$是一个随机变量，它取值为$E(Y|X=x_i)$的概率是 $P(X=x_i)$。

② 连续随机变量的条件期望

Y 关于随机事件 $(X=x)$ 的条件期望：

$$E(Y \mid X = x) = \int_{-\infty}^{+\infty} y \times p(y \mid x)\, dy$$

Remark

Y 关于 X 的条件期望 $E(Y|X)$ 是一个随机变量，它取值为 $E(Y|X=x)$，密度函数是 $p(x)$。

8. 特征函数 $f(t)=Ee^{itX},\ t\in \mathbf{R}^1$

(1) 二项分布 $f(t)=(q+pe^{it})^n$

(2) Poisson分布 $f(t)=e^{\lambda(e^{it}-1)}$

(3) 均匀分布 $f(t)=\dfrac{e^{itb}-e^{ita}}{it(b-a)}$

(4) *Gamma*分布 $f(t)=(1-\dfrac{it}{\lambda})^{-\alpha}$

(5) 正态分布 $f(t)=e^{i\mu t-\frac{1}{2}\sigma^2t^2}$

六. 大数律与中心极限定理

随机变量部分和的极限性质

1. 收敛性

依概率收敛： 频率是极限是概率

依分布收敛：分布函数的收敛

几乎处处收敛：**Borel** 强大数律

2. 中心极限定理

De moivre-Laplace 定理

独立同分布的**Lindeberg 定理**

习题 1.2

1-5. 教材 **23** 页

第 **1**、**2**、**6**、**7**、**8** 题。

第1.3节 统计量与抽样分布

一. 统计量(*statistic*)

1. 定义1.3.1 假定 X_1，…，X_n 是来自总体X的一组样本，$\varphi(\cdot)$ 是一个完全已知的函数，则称 $\varphi(X_1, \ldots, X_n)$ 是一个统计量。

当样本 X_1，…，X_n 有了观察值 x_1，…，x_n 以后，统计量 $\varphi(X_1, \ldots, X_n)$ 的相应的观察值就是 $\varphi(x_1, \ldots, x_n)$。

统计量自身带有总体中未知参数的信息，
但统计量的表达式中不能出现任何未知的参数。

例如 X_1，X_2，…，X_n 是来自总体 $X \sim N(\mu, \sigma_0^2)$ 的一组样本。其中 μ 是未知的参数，σ_0^2 已知。

思考1.

下面哪些是统计量？相应又服从什么分布？

$$\frac{1}{n}\sum_{k=1}^{n} X_k \qquad \frac{1}{n}\sum_{k=1}^{n} X_k - \mu \qquad X_k \qquad \frac{X_k}{\sigma_0}$$

Remark

把样本“加工”成统计量含有“数据压缩”的意思

对于要解决的不同的统计问题，必须构造出不同的统计量去处理。

“充分统计量”的概念：

没有损失样本所包含的总体未知参数的任何信息。

假定有统计量 $T = T(X_1, \ldots, X_n)$，如果给定 $T = t$ 时样本 $(X_1, \ldots, X_n)$ 的条件分布与总体参数 θ 无关，则称 T 是一个充分统计量。

2. 充分统计量(*Sufficient Statistic*)

1920 年左右，**Fisher** 与 **Eddington** 争论：

假定 X_1，…，X_n 来自总体 $X \sim N(\mu, \sigma^2)$，要估计反映测量精度的 σ。

Eddington 建议利用绝对平均偏差：

$$\varphi_1 = \sqrt{\frac{\pi}{2}} \frac{1}{\sqrt{n(n-1)}} \sum_{k=1}^{n} |X_k - \overline{X}|$$

而**Fisher** 建议应该用样本标准差：

$$\varphi_2 = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{2}\Gamma(\frac{n}{2})} \sqrt{\sum_{k=1}^{n} (X_k - \overline{X})^2}$$

例**1.3.1**　假定总体有 N 个个体，其中 M 个具有某种属性，从总体中采用有放回、无放回两种方式抽取 n 个样本 X_1，…，X_n。可以证明统计量 $T = X_1 + \ldots + X_n$ 是总体比例 $p = M/N$ 的充分统计量。

实际上，给定统计量 $T = t$ 时样本的条件分布是 $1/C_n^t$，所以 T 是充分统计量；而且样本比例 t/n 也是未知的总体比例 p 的充分统计量。

这个例子里充分统计量的意义在于：

如果我们希望抽取部分样本得到一批产品的次品率，(或者调查一部分人了解全体群众的观点等)，无论采用有放回还是不放回的抽样方法，我们只需要知道抽取出的产品里究竟有几个次品！没有必要了解抽取的过程中，第一个是否是次品，第二个是否是次品，…

3. 概率函数 $f(x, \theta)$ $= \prod_{k=1}^{n} p(x_k, \theta)$

离散总体时，样本 $(X_1, \ldots, X_n)$ 的联合分布律

连续总体时，样本 $(X_1, \ldots, X_n)$ 的联合密度函数

因子分解定理：当且仅当概率函数能被分解成：

$$f(x, \theta) = K(T(x), \theta)\, h(x)\ ,$$

则 $T(X)$ 是一个充分统计量。

概率函数在这里被看成是 x、θ 的函数。

例1.3.2　　总体 X ~ 两点分布 $B(1, p)$

分布律为：$P\{X=k\}=p^k(1-p)^{1-k}$ ，　$k=0,1$

概率函数为：

$$f(x,\theta)=p^{\sum x_k}(1-p)^{n-\sum x_k}$$

参数 p 的充分统计量就是样本算术平均值 $\overline{X}$

例1.3.3　　总体 X ~ 均匀分布 $U(0,\theta)$ ，

参数 θ 的充分统计量就是样本中的最大 $X_{(n)}$ 。

例1.3.4　X_1，…，X_n 是来自总体 $X \sim N(\mu, \sigma^2)$ 的一组简单随机样本，参数 μ, σ^2 都未知。

则概率函数为：

$$f(x,\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{k=1}^{n}(x_k - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right\}$$

因此，总体参数 (μ,σ^2) 的充分统计量是：

$$\left(\bar{X}, \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})^2\right)$$

或者是：$\left(\sum_{k=1}^{n} X_k, \sum_{k=1}^{n} X_k^2\right)$ 等等

4. 完备统计量 (*Complete Statistic*)

假定 T 是一个统计量，如果对于任意函数 $\varphi(\cdot)$，只要 $E_\theta \varphi(T) = 0$ 就可以推出 $P_\theta\{\varphi(T) = 0\} = 1$，对所有的参数 θ 都成立；

则统计量 T 就称为是一个完备统计量。

例1.3.5　X_1，…，X_n 是来自总体两点分布 $B(1,p)$ 的一组简单随机样本，已经知道样本均值是参数 p 的充分统计量，可以证明它也是完备的统计量。

证明. 全体样本之和 T 服从二项分布 $B(n,p)$，

如果对所有的 $0<p<1$ 下式都成立：

$$E_p[\varphi(T)]=\sum_{t=0}^{n}\varphi(t)\times C_n^t p^t(1-p)^{n-t}=0$$

即 $(p/1\text{-}p)$ 的多项式：$\sum_{t=0}^{n}\varphi(t)\times C_n^t\left(\frac{p}{1-p}\right)^t=0$

对所有的 $0<p<1$ 成立，所以每项系数为 0。

□

指数型分布族

如果总体 X 密度(或分布律) $p(x, \theta)$ 可表示成：

$$p(x,\theta) = C(\theta)\, h(x) \exp\{\sum_{i=1}^{k} b_i(\theta) T_i(x)\}$$

则称 X 的分布是一个指数型分布族 。

(1) 常见的二项分布、泊松分布、指数分布、正态分布等都属于指数型分布族。

(2) 如果 X 的总体是指数型分布族，则 $(\sum T_1(X_i), \cdots, \sum T_k(X_i))$ 是充分完备统计量。

例1.3.6　总体 $X\sim$ 泊松分布 $P(\lambda)$，因此参数 λ 的完备统计量是 $\sum_{k=1}^{n}X_k$ 或者 $\overline{X}$。

例1.3.7　总体 $X\sim N(\mu,\sigma^2)$，参数 (μ,σ^2) 的完备统计量是 $(\overline{X},\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X})^2)$。

例1.3.8　总体 $X\sim$ 均匀分布 $U(0,\theta)$，它并不是一个指数分布族，但是也可以证明参数 θ 的完备统计量仍然就是它的充分统计量 $X_{(n)}$。

二 . 常用的一些统计量

1. 表示“平均”的统计量：

样本均值、中位数、众数

2. 表示“变差”的统计量：

样本方差(或标准差)、极差

3. 特殊的统计量：顺序统计量

1.1 样本均值 (*Sample mean*)

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

反映了样本这组数据的(算术)平均值

如两组样本数据：

{2，4，6，8，10} 与 {4，5，6，7，8}

样本均值都是 6 ，即平均程度都相同。

1.2 样本中位数 (*Median*)

样本按照取值大小排列后居中的那个样本。

例如：*n* 奇数： { 2，1，6，4，3 } 3

n 偶数： { 2，1，6，4，3，7 } (3+4)/2 = 3.5

1.3 众数 (*Mode*)

样本数据中出现次数最多的样本，例如：

{1，1，3，3，4，2，3，8} 3

Remark

(1). 总体中位数与众数的定义

中位数 $M(X)$：

$$M(X)=\inf_{x}\{x:\mathrm{P}(X<x)\geq\frac{1}{2}\}\ ;$$

众数 $Mode(X)$：

分布律或者概率密度函数在此达到最大。

(2). 中位数比样本均值更为稳健，当二者相差不大时常采用样本均值表示数据平均，否则应该考虑使用中位数。

(3). 样本的众数更适用于离散的总体。

2. 样本方差(*Sample variance*)

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

S 称为是样本标准差 (***Standard deviation***),
与样本均值量纲相同。反映了样本离散程度。

如两组样本数据：

{ 2，4，6，8，10 } 与 { 4，5，6，7，8 }

样本均值都是 6 ，但 $S_1^2 = 10$， $S_2^2 = 2.5$ ；

第二组数据相对于均值 6 更为集中。

3. 顺序统计量(*Order Statistic*)

对于样本 X_1，…，X_n ，对应观察值记为 x_1，…，x_n；按照样本观察值的大小关系排序：

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

相应的样本：$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$

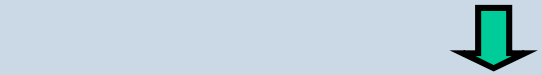
称为顺序统计量

Remark

1. 顺序统计量是充分统计量；
2. 如果观察值相同，则序号小的排在前面；
3. 样本在顺序统计量中的位置称为“秩”(*Rank*)

例如，有 5 个样本：

	X_1,	X_2 ,	X_3 ,	X_4 ,	X_5
观察值：	1,	3,	0,	3,	2



排序成：	0,	1,	2,	3,	3
原始样本：	X_3,	X_1 ,	X_5 ,	X_2 ,	X_4
顺序统计量：	$X_{(1)}$,	$X_{(2)}$,	$X_{(3)}$,	$X_{(4)}$,	$X_{(5)}$

秩：$R_1 = 2$，$R_2 = 4$，$R_3 = 1$，$R_4 = 5$，$R_5 = 3$

$X_{(1)}$ 称为 极小统计量；$X_{(n)}$称为 极大统计量 ；

极差 (*Range*) 定义成： $X_{(n)} - X_{(1)}$ 。

顺序统计量的联合分布

假定总体具有概率密度函数 $p(x)$，X_1，…，X_n 是一组样本，相应的顺序统计量记为：$Y_k = X_{(k)}$。

(1). 全体顺序统计量的联合概率密度函数：

$$p(y_1, \ldots, y_n) = n!\, p(y_1) \cdots p(y_n),\ y_1 \le y_2 \le \cdots \le y_n$$

(2). 第 k 个顺序统计量 $Y_k = X_{(k)}$ 的概率密度函数：

$$p_k(y) = \frac{n!}{(k-1)!(n-k)!}\, p(y) F(y)^{k-1} [1 - F(y)]^{n-k}$$

例1.3.9　极小统计量$X_{(1)}$的概率密度函数是：

$$p_1(y) = np(y)[1 - F(y)]^{n-1}$$

极大统计量$X_{(n)}$的概率密度函数是：

$$p_n(y) = np(y)[F(y)]^{n-1}$$

练习2

分析串连、并联系统的寿命(或者可靠性)。

(3). 任意两个顺序统计量$(k < r)$的
联合概率密度函数：

$$p_{k,r}(y_k, y_r) = \frac{n!}{(k-1)!(r-k-1)!(n-r)!} p(y_k) p(y_r)$$
$$\times F(y_k)^{k-1}[F(y_r) - F(y_k)]^{r-k-1}[1 - F(y_r)]^{n-r}, \quad y_k < y_r$$

例**1.3.10** 极差的概率密度函数是：

$$p_{range}(y) = n(n-1)\times$$
$$\int_{-\infty}^{+\infty} p(x)p(x+y)[F(x+y) - F(x)]^{n-2}dx, \quad y > 0$$

Remark

(1). 极差计算简单，但不如样本标准差稳健。

(2). 对于大多数单峰对称分布，标准差大约等于极差的四分之一。

(3). 大多数情况下，数据基本上落在“均值±2个标准差”的区间内，否则这个数据就被认为是异常的大或异常的小。
在绝大多数情况下，一组正常数据基本上落在“均值±3个标准差”的区间内。

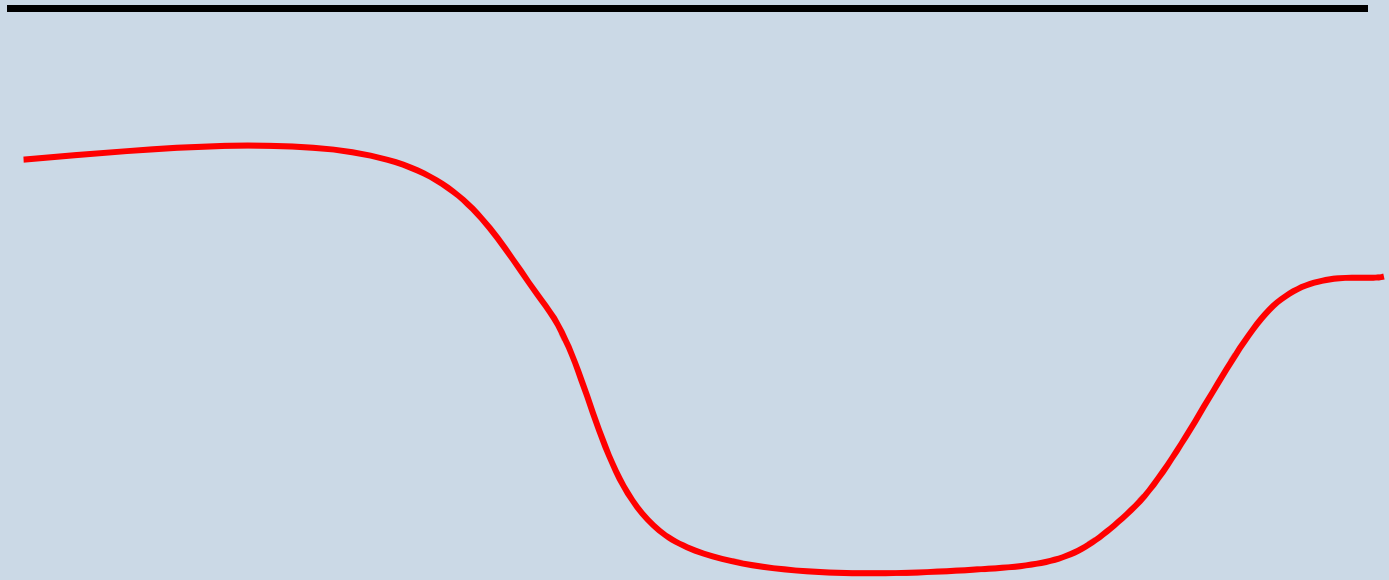
关于“平均值”的理解

样本均值是人们采用最多的一种描述数据的方法，它反映了一组数据整体上的一些信息，然而容易掩盖一些极端的情况，

所以有时候样本均值不一定合理。

思考3. 甲同学听说，有个身高 1.75 米的成年人在平均水深为 1 米的小河中淹死了，他觉得不可思议。

这件事情是否是一个玩笑？



例1.3.11　平均值与中位数

乙同学毕业后求职于一家公司。总经理说公司平均月薪是 **30,000** 元。一个月后乙同学得到工资**10,000**元，据了解公司共有**21**人，和他自己职位相同的业务员共有 **10** 人，每人的月薪都是**10,000** 元。应该如何理解乙同学的遭遇？

总经理150,000；两个副总经理每人 80, 000；
3 个部门经理每人 40,000； 5 个财务等行政人员每人 20,000； 10 个业务员每人 10,000 。
公司一共 21 人，每月支出工资 630,000。

平均值 30,000，中位数 20,000，
众数 10,000，极差 140, 000

例**1.3.12**　正确解释统计数据

下面是某高速公路上发生的交通事故有关数据：

速度(*km*/*h*)	小于 120	120～150	大于 150
车祸次数	12	32	5

丙同学由此得出结论说：统计数据显示，在高速公路上汽车速度越高，也就越安全。

三. 统计学中的三个分布

(一) 卡方分布

独立同分布于 $N(0,1)$ 的变量平方和的分布

1. 卡方分布的构造

记 $K^2 = X_1^2 + X_2^2 + \ldots + X_n^2$ ，

这里 X_1，…，X_n 独立同分布于 $N(0,1)$ ，

则称 K^2 服从参数 n 的卡方分布，记为：

$$K^2 \sim \chi^2(n)$$

2. 卡方分布的概率密度函数

$$k_n(x)=\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}\quad,\quad x>0$$

数学期望是 n，方差是 $2n$

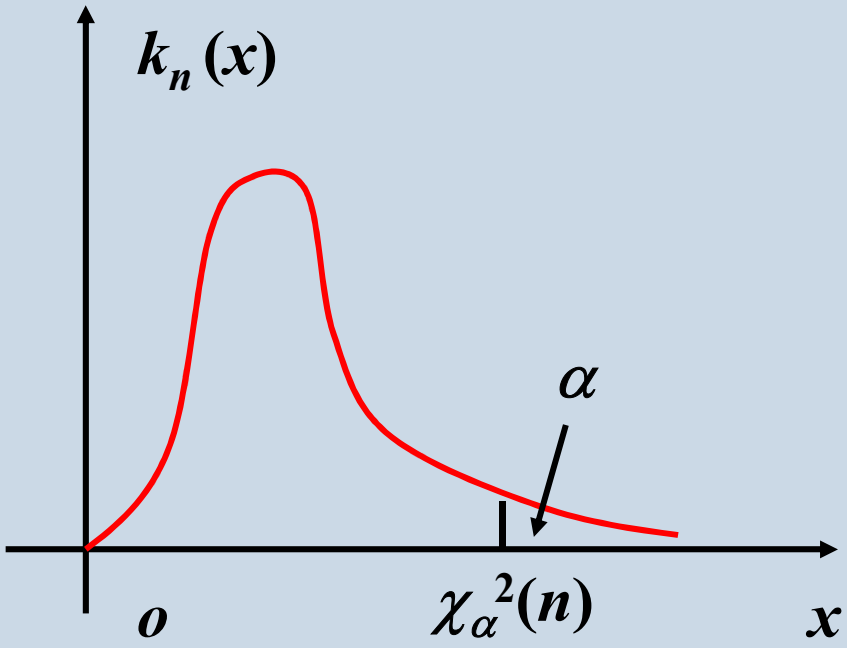
3. 卡方分布具有“可加性”

如果 X、Y 独立，$X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$

则 $X+Y\sim\chi^2(n_1+n_2)$

4. 卡方分布的上侧分位点

假定 $X \sim \chi^2(n)$，
给定：$0 < \alpha < 1$，
如果一个数 c 满足：
$P\{X > c\} = \alpha$，



则称这个数 c 是自由度 n 的卡方分布的上侧 α 分位点(数)，记成 $\chi_\alpha^2(n)$。

(二) t 分布

独立标准正态变量与卡方变量商的分布

1. t 分布的构造

如果 X、 Y 独立，并且

$$X \sim N(0,1)\,,\quad Y \sim \chi^2(n)\,;$$

则称：

$$T = \frac{X}{\sqrt{Y/n}}$$

服从自由度为 n 的 t 分布，记为 $T \sim t(n)$。

2. t 分布的概率密度函数

$$t_n(x)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$$

数学期望是 0 ($n\geq 2$)， t (1) 是Cauchy 分布

方差是 $n/(n-2)$ ($n\geq 3$)

3. $n\to\infty$时， t (n) 的极限分布是标准正态。

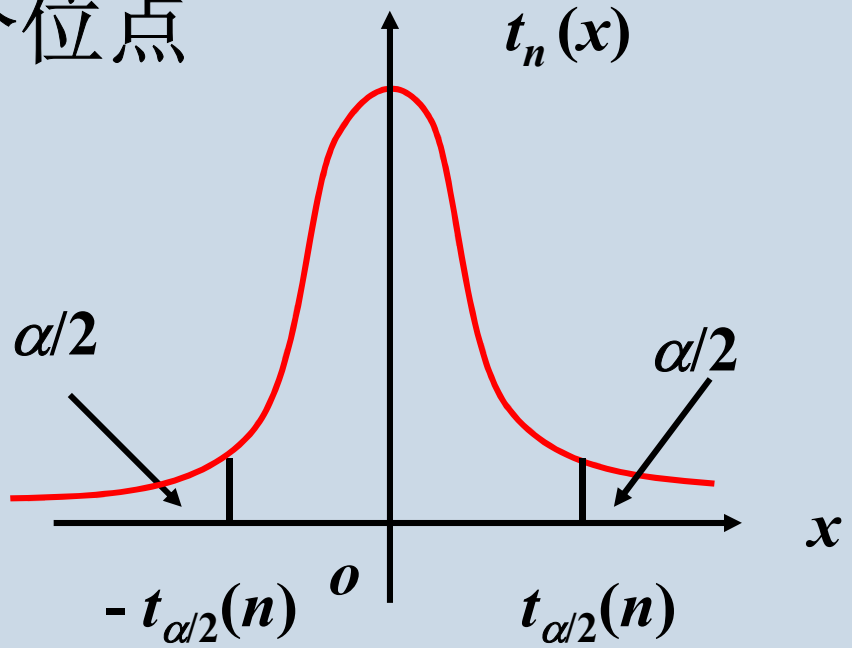
4. t 分布的双侧分位点

假定 $X \sim t(n)$，

给定：$0 < \alpha < 1$，

如果一个数 c 满足：

$P\{|X| > c\} = \alpha$，

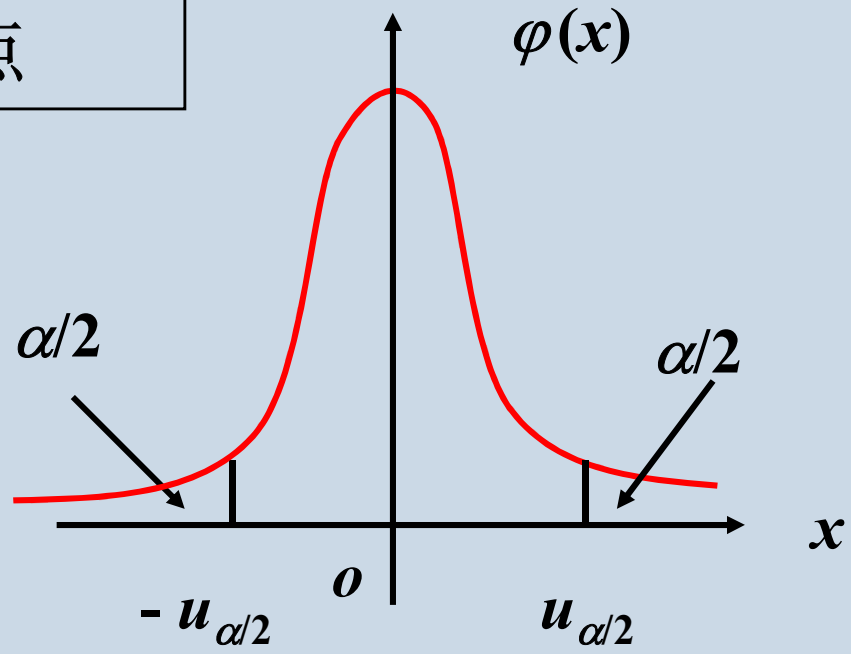


则称这个数 c 是自由度 n 的 t 分布的双侧 α 分位点(数)，记成 $t_{\alpha/2}(n)$。

对称分布的双侧 α 分位点是上侧 $\alpha/2$ 分位点

标准正态分布 $N(0,1)$
的双侧 α 分位点

记为： $u_{\alpha/2}$



如：双侧 0.05 分位点 $u_{0.025} = 1.96$

(三) F 分布

两个独立的卡方随机变量商的分布

1. F 分布的构造

如果 X、 Y 独立，并且

$$X \sim \chi^2(m)\ ,\quad Y \sim \chi^2(n)\ ;$$

则称：

$$F = \frac{X/m}{Y/n}$$

服从自由度 (m, n) 的 F 分布，记为 $F \sim F(m, n)$

2. F 分布的概率密度函数

$$f_{m,n}(x)=\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\, m^{\frac{m}{2}}\, n^{\frac{n}{2}}\, \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{m+n}{2}}}\quad ,\quad x>0$$

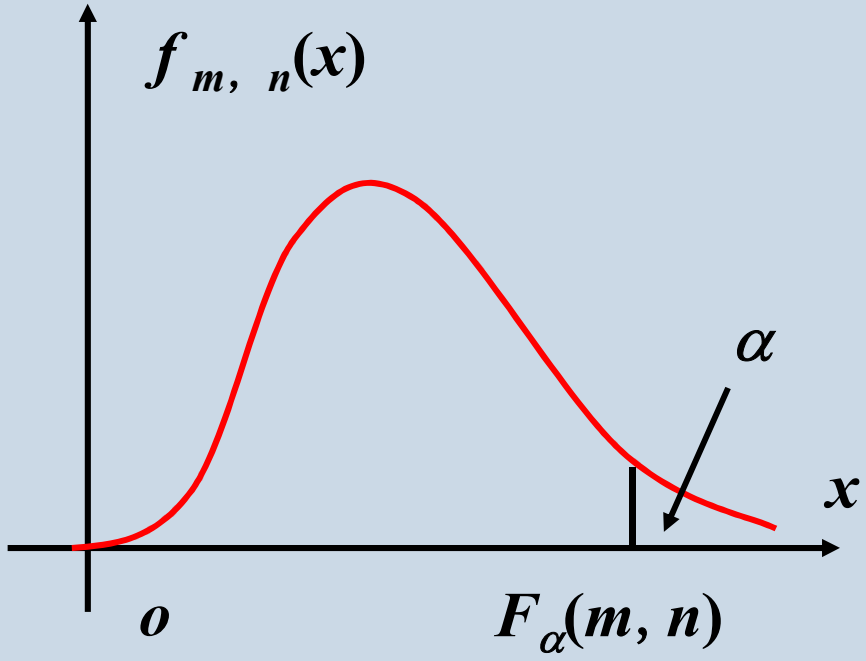
数学期望是 $n/(n-2)$ ($n \geq 3$)

3. 如果$T \sim t(n)$，则有 $T^2 \sim F(1, n)$

4. F 分布的上侧分位点

注意关系式

$$F_{1-\alpha}(m, n) = 1 / F_{\alpha}(n, m)$$



练习1.3.13　利用F分布的性质：

当 $F \sim F(m, n)$ 时，有 $1/F \sim F(n, m)$；

推导如上 F 分布的分位点的关系。

四. (正态总体) 的抽样分布

定理1.3.1　一个正态总体的抽样分布

假定 X_1，X_2，…，X_n 是来自总体 $X \sim N(\mu, \sigma^2)$ 的一组简单随机样本；$\overline{X}$ 与 S^2 分别是样本均值与样本方差。

(1). $\dfrac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1)$　　(2). $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

(3). $\overline{X}$ 与 S^2 独立　　(4). $\dfrac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t(n-1)$

多元正态分布的基本性质

1. 随机向量 **X** 服从 n 维正态分布 $N(\mu, \sum)$，
如果联合密度是：

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\}$$

2. **X** 服从 n 维正态 $N(\mu, \sum)$ 的充分必要条件是：
对任意 n 维列向量 l，有 $l^{\mathrm{T}}\mathbf{X} \sim N(l^{\mathrm{T}}\mu, l^{\mathrm{T}}\sum l)$；

3. 如果 $\mathbf{X} \sim N(\mu, \sum)$，**A** 是任意 $m \times n$ 矩阵 $(m \le n)$，
则有 $\mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{A}\sum\mathbf{A}^{\mathrm{T}})$；

定理1.3.1 的证明思路

构造一个特殊的 n 阶正交矩阵：$C = \begin{pmatrix} \frac{1}{\sqrt{n}} & .. & \frac{1}{\sqrt{n}} \\ * & .. & * \\ * & .. & * \\ & & \end{pmatrix}$

由于样本 $X \sim N(\mu 1_n, \sigma^2 I_n)$，
因此正交变换 $Y = CX \sim N(n^{1/2}\mu 1_{(1)}, \sigma^2 I_n)$，
而且保持独立性不变，这里 $1_{(1)}=(1, 0, .., 0)^T$；

(1) $Y_1 \sim N(n^{1/2}\mu, \sigma^2)$，而它实际上就是 $\sqrt{n} \times \overline{X}$；

(2) $Y_2, .., Y_n$ i.i.d 于 $N(0, \sigma^2)$，根据正交变换有
$X_1^2 + \ldots + X_n^2 = Y_1^2 + \ldots + Y_n^2$，因此得到 $(n-1)S^2$
$= Y_2^2 + \ldots + Y_n^2$。定理1.2.1 的(2)、(3) 同时得证。

定理 1.3.2　两个正态总体的抽样分布

假定两组简单 随机样本 $X_1, \ldots, X_{n_1}$ 与 $Y_1, \ldots, Y_{n_2}$ 分别来自两个独立的正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$，

X 总体的样本期望与样本方差分别是：

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \quad , \quad S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$$

Y 总体的样本期望与样本方差分别是：

$$\overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j \quad , \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_j - \overline{Y})^2$$

(1) $\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$

(2) 如果假定 $\sigma_1^2 = \sigma_2^2$，定义：

$$S_W^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

则有

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_W\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

定理 1.3.3　柯克伦(Cochren) 定理

假定 X_1，…，X_n 是来自总体 $X \sim N(0,1)$ 的一组简单随机样本，记 $\mathbf{X}=(X_1,\ldots,X_n)^{\mathrm{T}}$；

$A_i(1\le i\le r)$分别是秩为 n_i 的非负定矩阵，

满足　　$A_1+\ldots+A_r=\mathbf{I}_n$

则$(X_1,\ldots,X_n)^{\mathrm{T}}$的$r$ 个二次型 $\mathbf{X}^{\mathrm{T}}A_i\mathbf{X}$ 相互独立并且 $\mathbf{X}^{\mathrm{T}}A_i\mathbf{X}\sim\chi^2(n_i)$ 的充分必要条件是：

$$n_1+n_2+\ldots+n_r=n$$

习题 1.3

1-5. 教材 **24** 页

第 **15**、**17**、**27**、**31**、**33** 题。

第2章　参数估计

第2.1节　点估计

第2.2节　估计量的优良标准

第2.3节　区间估计

东北大学数学系

第2.1节　点估计

总体 X 的分布函数 F 含有未知的参数 θ，
θ 所有可能的取值范围称为“参数空间”，记为 Θ。
从这个总体中抽取了一组样本 X_1，…，X_n，
相应的样本观察值是 x_1，…，x_n。
应该如何估计出 θ 的具体数值？

点估计就是利用样本构造一个合理的统计量：
$\varphi(X_1, \ldots, X_n)$；用它的观察值 $\varphi(x_1, \ldots, x_n)$
去作为作为 θ 的估计值。

例**2.1.1**　政府或者企业希望了解人们的作息习惯。**Gallup** 公司做过一项调查，**56 %** 的美国人说他们习惯早起，**44 %** 的认为自己是“夜猫子”。

例**2.1.2**　丁同学在一个体重仪上称她的体重，假定这个体重仪没有系统误差，每次称量的结果是真实重量 μ 加上一个随机误差 ε_k 。一般认为 $\varepsilon_k \sim N(0, \sigma^2)$，因此 n 次称量的结果

$$X_k = \mu + \varepsilon_k \sim N(\mu, \sigma^2)$$

你可以用这组数据中的任何一个，或者样本均值，或者是样本中位数等，作为 μ 的估计值。

常用的点估计方法

矩估计： 用样本的有关矩去作为总体有关矩的估计。即样本均值作为总体期望的估计；样本二阶中心矩作为总体方差的估计；样本中位数 (或众数) 作为总体中位数(或众数) 的估计等 。

极大似然估计：

所有情况中 “看起来最象” 的那个估计

一. 矩估计

K.Pearson 的矩估计理论

假定总体 X 有 m 个未知参数 $\theta_1, \ldots, \theta_m$，而有关的原点矩 $V_k = EX^k$ 存在，则应该有：

$$\begin{cases} V_1 = g_1(\theta_1, \ldots, \theta_m) \\ V_2 = g_2(\theta_1, \ldots, \theta_m) \\ \cdots\cdots \\ V_m = g_m(\theta_1, \ldots, \theta_m) \end{cases}$$

理论上求解方程组可以得到

$$\begin{cases} \theta_1 = h_1(V_1, \ \ldots, \ V_m) \\ \theta_2 = h_2(V_1, \ \ldots, \ V_m) \\ \quad\quad \ldots\ldots \\ \theta_m = h_m(V_1, \ \ldots, \ V_m) \end{cases}$$

假如用样本的 k 阶原点矩作为 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 总体 k 阶矩 V_k 的估计，则可以得到总体未知参数 $\theta_1, \ \ldots, \ \theta_m$ 的估计。

理论依据：大数律。矩估计基本上都是依概率或者几乎处处收敛到未知参数。

矩估计需要注意的几个问题

(1) 总体的参数不能表示成矩的函数时 (一般是总体矩不存在) ，就不能使用矩估计；

(2) 如果能够用低阶的矩估计，就不要用高阶矩；

(3) 按照矩估计的理论应该用样本的二阶中心矩来估计总体的方差，但是在实际应用中人们总是采用样本方差作为总体方差的估计。

矩估计的最大优点是简单实用，与总体分布形式没有关系。只要知道总体随机变量一些矩存在，就可以做相应的矩估计。

例2.1.3　设总体 $X \sim U(0, \theta)$，θ 是未知参数，X_1，…，X_n 是一组样本，求 θ 的矩估计。

解．总体的未知参数 θ 可以通过期望与方差表示：

总体期望：$\frac{\theta}{2}$；　总体方差：$\frac{\theta^2}{12}$。

因此根据矩估计的思想，可以得到两个矩估计：

$$\hat{\theta} = 2\overline{X} \text{ 或者是 } \hat{\theta} = 2\sqrt{3(n-1)/n}S$$

习惯上我们采用第一个估计量 □

例**2.1.4** 简单随机样本X_1，…，X_n 来自总体 X，总体期望μ, 方差σ^2 都未知，求 μ, σ^2 的矩估计。

解. 显然有：$V_1 = \mu$，$V_2 = \sigma^2 + \mu^2$；即

$$\mu = V_1, \quad \sigma^2 = V_2 - V_1^2 \text{ 。}$$

因此得到：

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - (\frac{1}{n}\sum_{k=1}^{n} X_k)^2$$

$$= \frac{1}{n}\sum_{k=1}^{n} (X_k - \overline{X})^2$$

□

几个常见分布的矩估计

二项分布 $B(N, p)$，N 已知	$\hat{p} = \dfrac{\overline{X}}{N}$
均匀分布 $U(a, b)$	$\overline{X} \pm \sqrt{3(n-1)/n}S$
泊松分布 $P(\lambda)$	$\hat{\lambda} = \overline{X}$
参数为 λ 的指数总体	$\hat{\lambda} = 1/\overline{X}$
正态总体 $N(\mu, \sigma^2)$	$\hat{\mu} = \overline{X} \quad \hat{\sigma}^2 = \dfrac{n-1}{n}S^2$ $\hat{\sigma} = \sqrt{(n-1)/n}S$

例2.1.5　随机取 8 个零件，测得直径是 (mm)

74.001，74.005，74.003，74.001，

74.000，73.993，74.006，74.002。

设总体期望与方差存在，求它们的矩估计。

解　首先计算这组样本的样本均值和样本方差，

$$\overline{x} = 74.001375, \quad s^2 = 1.5696 \times 10^{-5}$$

根据矩估计的思想，虽然不知道这组样本来自什么样的总体，仍然可以给出总体期望和方差的矩估计：

$$\hat{\mu} = 74.001375, \quad \hat{\sigma}^2 = 1.3734 \times 10^{-5}$$

□

民意调查

采用样本中的百分比作为估计值。

总体分布被认为是一个两点分布，参数 p 或者说总体期望 p 就是总体的百分比。

因此如果希望了解总体中具有某种属性的个体的比例，只需要从总体中抽取部分样本，以样本中具有这种属性的比例 p_s 作为估计。

二. 极大似然估计 (*MLE*)

1. 极大似然估计的想法

例**2.1.6**　　假定盒子里黑、白球共 **5** 个，但是不知道黑球具体数目。现在随机有放回抽取 **3** 个小球，发现是两个黑球和一个白球。问盒子里最可能有几个黑球？

解：盒子里黑白球所有的可能有六种：

5白，**4**白**1**黑、**3**白**2**黑，**2**白**3**黑，**1**白**4**黑，**5**黑

以 p 记盒子里黑球所占的比例，

则 p 全部可能的值是：

$$\{\ 0,\ \frac{1}{5},\ \frac{2}{5},\ \frac{3}{5},\ \frac{4}{5},\ 1\ \}$$

定义三个统计量 X_1，X_2，X_3 表示抽样结果：取到黑球记为 1 ，否则记为 0 。因此 X_1，X_2，X_3 独立同分布于参数 p 的两点分布。

例题中的三个样本观察值 x_1，x_2，x_3 有两个取值是 1，一个取值为 0。

而样本的联合分布律显然是

$$L(p) = p^{x1+x2+x3}(1-p)^{3-x1-x2-x3} = p^2(1-p)$$

注意这里样本的联合分布律

$$L(p) = p^{x1+x2+x3}(1-p)^{3-x1-x2-x3}$$

其实就是概率函数 $f(x, \theta)$ 。

它的含义是：当盒中黑球比例为 p 时，随机事件“有放回取出的三个小球中有两个黑球、一个白球”的概率。

对应于参数空间中不同的 p ，样本分布 $L(p) = p^2(1-p)$ 所对应的这些概率是：

p	$0,\ \frac{1}{5},\ \frac{2}{5},\ \frac{3}{5},\ \frac{4}{5},\ 1$
$L(p)$	$0,\ \frac{4}{125},\ \frac{12}{125},\ \frac{18}{125},\ \frac{16}{125},\ 0$

既然“三个小球中包含两个黑球”是已经发生了的随机事件，因此使得这个事件发生概率取最大的那个值就是未知参数 p 最有可能的取值。

即 p 的极大似然估计就是 3/5。

□

R.A.Fisher 的极大似然估计理论

把概率函数 $f(x,\theta)$ 记为$\mathrm{L}(\theta)$，认为它是

x 固定，关于θ 的函数。

$\mathrm{L}(\theta)$ 称为“似然函数”

1. 对离散总体，它是样本联合分布律；
2. 对连续总体，它是样本联合密度函数。

如果有 $\mathrm{L}(\theta_1) < \mathrm{L}(\theta_2)$，很自然我们会认为总体参数 θ 更有可能是 θ_2，而不太可能是 θ_1。

总体参数 θ 的极大似然估计就是使得似然函数在参数空间 Θ 中达到极大者

即对于任意 $\theta \in \Theta$ 都有：

$$L(\hat{\theta}) = \max L(\theta)$$

一般采用对数似然方程 (组) 求解***MLE***

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

无法建立似然方程时，必须根据定义求***MLE***

例2.1.7　设总体 $X \sim B(N, p)$，N 已知，p 是未知参数，X_1，…，X_n 是一组简单随机样本，求总体参数 p 的极大似然估计。

解.　不妨假定样本 X_1，…，X_n 相应的观察值是 x_1，…，x_n，而二项总体的似然函数为：

$$L(\theta) = [\prod \binom{N}{x_k}] p^{\sum x_k} (1-p)^{nN - \sum x_k}$$

这里每一个 $x_k = 0$、1、…、N 中的某个值

取对数再对参数 p 求导，得到对数似然方程：

$$\frac{\partial}{\partial\theta}\ln[L(\theta)]=\frac{\bar{x}}{p}-\frac{N-\bar{x}}{1-p}=0$$

因此，当 N 已知时，二项分布 $B(N,p)$ 中参数 p 的极大似然估计就是

$$\hat{p}=\frac{\overline{X}}{N}$$

两点分布的参数 p 的 *MLE* 就是样本均值

□

例2.1.8　X_1，…，X_n 是来自总体 $X \sim N(\mu, \sigma^2)$ 的简单随机样本，求 μ、σ^2 的极大似然估计。

解. 正态总体的似然函数为

$$L(\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k - \mu)^2\}$$

注意这里总体参数 θ 是一个向量 (μ, σ^2)，因此对于似然函数取对数后分别对 μ, σ^2 求导，建立对数似然方程组：

$$\begin{cases} \dfrac{1}{\sigma^2}(\bar{x}-\mu)=0 \\ -\dfrac{n}{2\sigma^2}+\dfrac{1}{2(\sigma^2)^2}\displaystyle\sum_{k=1}^{n}(x_k-\mu)^2=0 \end{cases}$$

解方程组得到正态总体两个参数的***MLE***

$$\hat{\mu}=\bar{X} \qquad \hat{\sigma}^2=\frac{1}{n}\sum_{k=1}^{n}(X_k-\bar{X})^2=\frac{n-1}{n}S^2$$ □

练习**2.1.9**

总体标准差σ的极大似然估计是什么？如果μ已知，方差σ^2的极大似然估计又是什么？

例**2.1.10**　为了估计湖中鱼的总数 N，一次从湖中捕捉 r 条鱼做完记号后再放回，过一段时间后再次捕捉 s 条鱼，发现有 x 条带有记号，如何估计 N?

解. 总体为超几何分布，分布律为：

$$\mathrm{P}(X=k)=\frac{C_r^k\times C_{N-r}^{s-k}}{C_N^s},\quad 0\le k\le \min(r,s)$$

现在样本的观察值为 $X=x$，因此似然函数为

$$L(N)=\frac{C_r^x\times C_{N-r}^{s-x}}{C_N^s},$$

要使得 $L(N)$ 达到极大，考察如下比值：

$$\frac{L(N)}{L(N-1)}=\frac{N^2-(r+s)N+rs}{N^2-(r+s)N+xN},$$

显然 $rs > xN$ ， $L(N) > \mathrm{L}(N\text{-}1)$;

$rs < xN$ ， $L(N) < \mathrm{L}(N\text{-}1)$ 。

即 N 的极大似然估计为：

$$\hat{N}=\left[\frac{rs}{x}\right].$$

□

例2.1.11　总体 $X \sim U(\theta, \theta+1)$，$\theta$ 是未知参数，X_1，…，X_n 是一组样本，求 θ 的极大似然估计。

解. 似然函数为：

$$L(\theta) = 1, \quad \theta < x_1, \ \ldots, \ x_n < \theta+1$$

显然不能对参数 θ 求导，无法建立似然方程

注意到这个似然函数不是 0 就是 1，利用顺序统计量，把似然函数改写成如下形式：

$$L(\theta) = 1, \quad \theta < x_{(1)} < \ldots < x_{(n)} < \theta + 1$$

因此只要 $\theta < x_{(1)}$ 并且 $x_{(n)} < \theta + 1$ 同时满足，似然函数就可以达到极大值 1 。

所以 $U(\theta, \theta+1)$ 中参数 θ 的极大似然估计可以是区间 $(x_{(n)} - 1, x_{(1)})$ 里的任意一个点。 □

说明 *MLE* 可以不唯一，甚至有无穷多个

同理，总体 $U(a, b)$ 左右端点 a 、b 的 *MLE* 分别就是两个极值统计量 $x_{(1)}$、$x_{(n)}$ 。

几个常见分布的极大似然估计

二项分布 $B(N, p)$，N 已知	$\hat{p} = \dfrac{\overline{X}}{N}$
均匀分布 $U(a, b)$	$X_{(1)}$，$X_{(n)}$
泊松分布 $P(\lambda)$	$\hat{\lambda} = \overline{X}$
参数为 λ 的指数总体	$\hat{\lambda} = 1/\overline{X}$
正态总体 $N(\mu, \sigma^2)$	$\hat{\mu} = \overline{X}$　$\hat{\sigma}^2 = \dfrac{n-1}{n} S^2$ $\hat{\sigma} = \sqrt{(n-1)/n}\,S$

三. 极大似然估计与矩估计的简单比较

矩估计由 **K.Pearson** 在**1894**年提出，只要求总体的矩存在即可，不需要知道总体分布。

极大似然估计由 **R . A . Fisher** 在 **1912** 年提出的，必须要知道总体来自哪一种分布类型。

因此极大似然估计具有更多数学上的良好性质。例如

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

习题 2.1

1-5. 教材 **97** 页

第 **2**、**3**、**4**、**8**、**9** 题。

第2.2节 估计的优良标准

一般的，一个良好的点估计应该满足三个标准：

无偏性：估计量的数学期望要等于参数；

有效性：估计量的方差要比较小 (主要是限制在无偏估计的范围内) ；

相合性：当样本容量趋于无限多时，估计量应该收敛到参数。

一．无偏估计(*Unbiased estimation*)

定义**2.2.1** 参数 $g(\theta)$ 的估计量 $\varphi(X_1,..,X_n)$ 如果满足：$E\,\varphi(X_1,..,X_n)=g(\theta)$ 对 Θ 中所有的 θ 都成立，则称 $\varphi(X_1,..,X_n)$ 是 $g(\theta)$ 的一个无偏估计量(*UE*)。

Remark

无偏性是估计好坏的一个基本要求，它表明即使每一次估计都可能有误差，但是从长远来看，这种估计总的误差能够相互抵消。

例**2.2.1** 假定总体 X 的期望 μ，方差σ^2存在，则样本均值、样本方差分别是 μ、σ^2的无偏估计。

证明. 样本均值是 μ 的无偏估计很显然；

只需证明 $ES^2 = \sigma^2$ 。

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

$$E\,S^2 = \frac{1}{n-1}\{n[\mu^2+\sigma^2] - n[(E\overline{X})^2 + D\overline{X}]\}$$

$$= \frac{1}{n-1}\{n[\mu^2+\sigma^2] - n[\mu^2 + \frac{\sigma^2}{n}]\} = \sigma^2$$

□

练习2.2.2

对于总体 $X \sim N(\mu, \sigma^2)$，验证 $n \geq 2$ 时样本绝对偏差及样本标准差的统计量都是 σ 的无偏估计。

$$\varphi_1 = \sqrt{\frac{\pi}{2}} \frac{1}{\sqrt{n(n-1)}} \sum_{k=1}^{n} |X_k - \overline{X}|$$

$$\varphi_2 = \frac{\Gamma(\frac{n-1}{2})}{\sqrt{2}\Gamma(\frac{n}{2})} \sqrt{\sum_{k=1}^{n} (X_k - \overline{X})^2}$$

提示：考虑 $N(0, 1)$ 的绝对矩以及卡方分布的期望

例2.2.3　总体 $X \sim U(0, \theta)$，θ 是未知参数，讨论 θ 的无偏估计。

解. θ 的矩估计是 $2\overline{X}$，显然是无偏估计，不过样本均值不是充分统计量，因此可能不如极大似然估计 $X_{(n)}$ 好，但是 $X_{(n)}$ 是否无偏的？

$X_{(n)}$ 的分布函数显然是

$\mathrm{P}\{X_{(n)} \le x\} = (x/\theta)^n$，　$0 < x < \theta$ 。因此 $E X_{(n)} = \dfrac{n}{n+1}\theta$；修正后得到根据充分统计量构成的 θ 的 UE 是 $\dfrac{n+1}{n} X_{(n)}$。

□

利用充分统计量构造无偏估计

假定样本是 X_1，…，X_n，充分统计量为 T，参数 $g(\theta)$ 的无偏估计量是φ，则$E(\varphi|T)$ 是 $g(\theta)$ 的由充分统计量构成的无偏估计。

练习**2.2.4**

样本 X_1，…，X_n 来自参数 p 的两点分布，利用充分统计量 $T = X_1+\ldots+X_n$ 构造总体方差 $p(1-p)$ 的一个无偏估计量。

糟糕的无偏估计

反例1

设总体 X 来自泊松分布 $P(\lambda)$，现在只有一个样本 X_1，求 $g(\lambda)=e^{-2\lambda}$ 的无偏估计。

这里无偏估计只有一个：

当 X_1 的观察值为偶数时，用 1 估计 $e^{-2\lambda}$；
当 X_1 的观察值为奇数时，用 -1 估计 $e^{-2\lambda}$

反例2

总体 X 来自两点分布 $B(1, p)$，仍然只有一个样本 X_1，则 p^2 的无偏估计不存在。

如果 $\varphi(X_1)$ 是 p^2 的一个无偏估计，则多项式

$$p\,\varphi(1) + (1-p)\,\varphi(0) = p^2$$

对所有的 $0 < p < 1$ 都成立，矛盾。

所以这里无偏估计不存在。

二. 有效性

1. 如何衡量估计的偏差

$$\varphi(X_1, \ldots, X_n) + Y$$

定义**2.2.2** 假定 $\varphi(X_1,..,X_n)$ 是 $g(\theta)$ 的一个估计，则

$$MSE(\varphi) = E[\varphi(X_1,..,X_n) - g(\theta)]^2$$

称为是估计量 $\varphi(X_1,..,X_n)$ 的均方误差(*MSE*)。

MSE 越小估计就越好，

UE 的 *MSE* 就是它的方差

例2.2.5　总体 $X \sim U(0, \theta)$，比较 θ 的两个无偏估计：$\varphi_1 = 2\overline{X}$ 与 $\varphi_2 = \frac{n+1}{n} X_{(n)}$ 的*MSE*。

解. 由于都是无偏估计，因此只需计算方差

显然　$Var\ \varphi_1 = 4\ Var\ (X_1) / n = \frac{\theta^2}{3n}$，

容易计算出　$Var\ \varphi_2 = \frac{\theta^2}{n(n+2)}$；

当 $n = 1$时即只有一个样本，它们的*MSE* 相同，但事实上这时这两个估计重合；

当样本容量大于1总有 $n(n+2) > 3n$，所以从均方误差的角度看，φ_2也要比 φ_1好。

□

例2.2.6　对于总体 $N(\mu, \sigma^2)$ 中，分别就
(1) μ 未知，(2) μ 已知，讨论参数 σ^2 的
两个估计量：

$$\varphi_1(a) = a\sum_{k=1}^{n}(X_k - \overline{X})^2 \,, \varphi_2(b) = b\sum_{k=1}^{n}(X_k - \mu)^2$$

解. 根据定理1.3.1，注意到 $\varphi_1 = a\sigma^2 \times [\frac{n-1}{\sigma^2}S^2]$,

因此，$E\varphi_1 = (n-1)a\sigma^2$, $Var(\varphi_1) = 2(n-1)a^2\sigma^4$.

显然，　$\varphi_2 = b\sigma^2 \times \chi^2(n)$,

即，　$E\varphi_2 = nb\sigma^2$, $Var(\varphi_2) = 2nb^2\sigma^4$.

(1) μ 未知时， φ_2不是统计量；此时取 $a=\dfrac{1}{n-1}$，

则 $\varphi_1(\dfrac{1}{n-1})=S^2$ 为σ^2 的UE，

$$MSE(\varphi_1(\frac{1}{n-1}))=Var(S^2)=\frac{2\sigma^4}{n-1}.$$

(2) μ 已知时, 取 $b=\dfrac{1}{n}$, 则 $\varphi_2(\dfrac{1}{n})$ 也是σ^2 的UE，

$MSE(\varphi_2(\dfrac{1}{n}))=\dfrac{2\sigma^4}{n}$； φ_2 比 φ_1 有效。

一般的，如果不限于UE 的范围，则

$$MSE(\varphi_1)=\{2(n-1)a^2+[(n-1)a-1]^2\}\sigma^4,$$

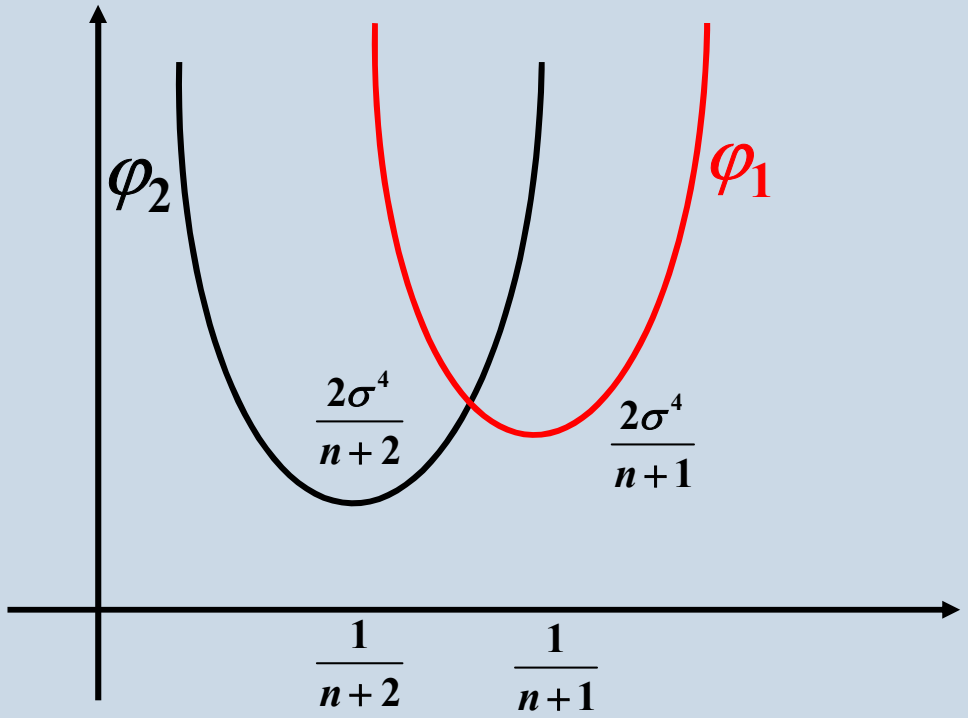
$$MSE(\varphi_2)=\{2nb^2+[nb-1]^2\}\sigma^4;$$

考察二次曲线 $f(x)$ ($c>0, x>0$):

$$f(x)=2cx^2+(cx-1)^2=(c^2+2c)x^2-2cx+1$$

曲线在 x 轴上方，

当 $x=\dfrac{1}{c+2}$ 时 $f(x)$ 达到极小 $\dfrac{2}{c+2}$.



□

2. 限制在*UE* 中的最优估计

定义**2.2.3**　一致最小方差无偏估计 (*UMVUE*)
假定 $\varphi_0(X_1,..,X_n)$ 是 $g(\theta)$ 的一个无偏估计，并且 $MSE(\varphi_0) \le MSE(\varphi)$ 对$g(\theta)$ 的任意 *UE* φ 都成立，则称$\varphi_0(X_1,..,X_n)$是$g(\theta)$ 的一致最小方差无偏估计

显然 $g(\theta)$ 的无偏估计的方差越小越好，但是这些方差不可能任意地小。 $g(\theta)$ 的所有无偏估计的方差有一个公共的下界(C-R下界)。

方差达到这个下界的 *UE* 自然就是 *UMVUE*

C-R不等式

假定 $X_1,..,X_n$ 是来自总体 $f(x,\theta)$, $\theta \in \mathbf{R}^1$的一组简单随机样本，对于统计量$\varphi(X_1,..,X_n)$，在正则性条件下有：

$$Var\,(\varphi) \geq \frac{(\frac{d}{d\theta} E\varphi)^2}{n\, I(\theta)}$$

这里 $I(\theta)$是总体的 **Fisher Information:**

$$I(\theta) = E_\theta[\frac{d}{d\theta}\log f(X,\theta)]^2$$

$$= -E_\theta[\frac{d^2}{d\theta^2}\log f(X,\theta)]$$

3. 一般情况下如何寻找 *UMVUE*

Blackwell-Lehmann-Sheffe 定理

如果T是充分、完备的统计量，$\varphi(T)$是$g(\theta)$的一个无偏估计，则$\varphi(T)$就是$g(\theta)$的*UMVUE*。

只需利用充分、完备统计量去构造无偏估计

思考1

给出计算任意$g(\theta)$的*UMVUE*的思路。

例2.2.7　求 $N(\mu, \sigma^2)$ 中参数 μ, σ^2 的*UMVUE*

解. 根据正态分布密度函数或因子分解定理，充分完备统计量是：

$$(\sum_{k=1}^{n} X_k, \sum_{k=1}^{n} X_k^2)$$

现在已知样本均值、样本方差分别是 μ, σ^2 的无偏估计，最重要的，它们还都是充分完备统计量的函数，因此 μ, σ^2 的*UMVUE* 分别就是样本均值和样本方差。

□

例2.2.8　关于一些常见分布的参数的*UMVUE*

N 已知时 $B(N, p)$，	$\hat{p} = \dfrac{\overline{X}}{N}$
泊松分布 $P(\lambda)$	$\hat{\lambda} = \overline{X}$
参数为 λ 的指数总体	$\hat{\lambda} = \dfrac{n-1}{n\overline{X}}$
$N(\mu, \sigma^2)$	$\hat{\mu} = \overline{X}$　$\hat{\sigma}^2 = S^2$

这些估计量都是无偏估计，

a . 由因子分解定理，它们都是充分统计量；

b. 总体属于指数族，它们也都是完备统计量；

因此根据**B-L-S** 定理，这些估计都是*UMVUE*。

三. 相合估计(*Consistent estimation*)

参数 $g(\theta)$ 的估计量 $\varphi(X_1,..,X_n)$ 总是与样本容量 n 有关，因此不妨记为 φ_n ；

一个好的估计直观上应该满足：

当 n 充分大时， φ_n 要充分接近 $g(\theta)$ 。

定义2.2.4 对于任意 $\varepsilon > 0$ ，如果当 $n \to \infty$ 时，

$$P\{|\varphi_n - g(\theta)| > \varepsilon\} \to 0$$

则称 φ_n 是 $g(\theta)$ 的相合估计(也称一致估计)

即， φ_n 依概率收敛到 $g(\theta)$

强相合估计：

$$P\{\varphi_n \to g(\theta)\} = 1$$

即 φ_n不收敛到 $g(\theta)$ 的那些样本点的概率为 0

一般来说矩估计都具有强相合性，而在较广泛的条件下，极大似然估计也具有强相合性。

渐近正态估计：*Asymptotically normal estimation*

如果存在一个常数 $\sigma > 0$ ，使得

$$\frac{n^{1/2}[\varphi_n - g(\theta)]}{\sigma} \to N(0,1)$$

即， φ_n的分布可以近似认为是 $N(g(\theta), \frac{\sigma^2}{n})$

习题 2.2

1-4. 教材 **99** 页

第 **10**、**12**、**15**、**19** 题。

第2.3节 区间估计

矩估计与极大似然估计，都是一种点估计。

区间估计是指用一个(随机) 区间去做未知参数 $g(\theta)$ 的估计，这个区间称为是置信区间 。

这个区间包含 $g(\theta)$的概率称为置信度或置信水平；

区间的长度称为是这个区间估计的精度，

长度越短，即精度越高，这个区间越好。

东北大学数学系

区间估计的想法是“给所做的结论留些余地”，表示我们有多大的把握肯定我们所做的结论。

显然对于总体的未知参数一个区间要比一个数值提供的信息更多，也更让人放心。

置信度越大，则区间的长度应该越长，即精度小，或者说抽样误差大。

在实际的统计应用中大多数的置信区间是由样本统计量±“抽样误差” 来构造。

一．置信区间理论(*Confidence interval*)

定义**2.3.1** 给定一个常数 $0 < \alpha < 1$，对于总体未知参数 $g(\theta)$，如果存在两个统计量 φ_1、φ_2 满足：

$$P\{\varphi_1(X) < g(\theta) < \varphi_2(X)\} \geq 1 - \alpha$$

则称(φ_1, φ_2)是$g(\theta)$的置信度$1 - \alpha$的置信区间；

φ_1、φ_2分别被称为是置信下限与置信上限。

有时也只考虑单侧区间 $(\varphi_1, +\infty)$ 或 $(-\infty, \varphi_2)$

点估计可以形式上认为是一种特殊的区间估计，这个区间的长度为 **0** 。

区间估计的置信度与精度是一对矛盾。

如果置信度越高，明显地区间应该越大，即误差大，区间的精度低。反之同理。

J . Neyman 的观点：

先考虑置信度，再去讨论估计的精度

先找出一些以**1 -** α 概率包含未知参数的区间，再从这些区间里去找长度最短者。

二. 区间估计的求解思路

置信区间主要依据统计量的抽样分布或者是大样本理论来构造。

第一步　找一个枢轴变量 $Z(X, \theta)$ 。

枢轴变量是一个随机变量，它与抽取出的样本以及待估计的 $g(\theta)$ 都有关系。但是它的分布又必须是与参数 θ 无关的已知分布 。

一般是从 $g(\theta)$ 的良好的点估计出发，去寻找枢轴变量 $Z(X, \theta)$ 。

第二步　对于给定的置信度 $1-\alpha$，求出两个常数 a、b，使得：

$$P\{a < Z(X,\theta) < b\} \geq 1-\alpha$$

第三步　变换不等式，成为等价的形式：

$$a < Z(X,\theta) < b$$

$$\longleftrightarrow \varphi_1(X) < g(\theta) < \varphi_2(X)$$

因此区间 (φ_1, φ_2) 就是 $g(\theta)$ 的一个置信度为 $1-\alpha$ 的区间估计。

三. 常见的几个区间估计

1. 总体属性比例的置信区间

假定从总体中抽取了n个观察值，以p_s记样本里具有某种属性的比例，则总体中具有这种属性的比例p的$1-\alpha$区间估计近似是：

$$p_s - u_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}} \quad \text{到} \quad p_s + u_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}}$$

$u_{\alpha/2}$ 是标准正态分布的双侧 α 分位点

依据是 De Moivre – Laplace 中心极限定理

a. 随机从总体中抽取一个样本，它具有这种属性的概率是 p；

b. 随机从大总体中抽取 n 个样本，其中具有这种属性的样本个数 X 近似有 $X \sim B(n, p)$；

c. 根据中心极限定理，又近似有：

$$\frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

把上式改写成：

$$\frac{\frac{X}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

这里(X/n) 正好是样本中的比例 p_s ，而根号符号里的 p 未知，因此用 p_s 近似替代，注意到标准正态分布是一个对称的分布，因此得到总体比例 p 的最短的近似区间估计：

$$p_s - u_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}} \quad \text{到} \quad p_s + u_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}}$$

例2.3.1　考虑容量为 1200 的一组样本构造的区间估计，假定有60% 的同学每天自习时间超过 2 小时。一个 95% 的置信区间的抽样误差是 2.8 个百分点。

这个结论的含义是：

有 95% 的把握可以肯定，学习认真同学的真实比例界于 57.2 ~ 62.8 之间。

同理，90% 的置信区间的抽样误差是2.3 个百分点，因此以 90% 的把握可以确定的这个区间是57.7 ~ 62.3 。

□

例2.3.2　随机询问500 名工作一年的大学毕业生，其中290 人表示对自己的工作还算满意，即新毕业大学生中有58%不反感自己的工作。

从区间估计的角度，95%区间的抽样误差是

$$1.96\sqrt{\frac{0.58(1-0.58)}{500}}=2.2\%$$

有 55.8～60.2 的毕业生认可他的新工作。 □

练习2.3.3

不满意而想换工作的比例大约在什么范围？

2. 指数总体参数的置信区间

假定样本 X_1，..，X_n 来自总体 $\Gamma(1, \lambda)$

根据第一章 1.2节 *Gamma* 分布 $\Gamma(\alpha, \lambda)$ 的性质：

(1) 对于 α 具有可加性；

(2) 如果 $X \sim \Gamma(\alpha, \lambda)$，则 $cX \sim \Gamma(\alpha, \lambda / c)$

$$2\lambda \sum_{i=1}^{n} X_i \sim \Gamma(\frac{2n}{2}, \frac{1}{2}) = \chi^2(2n)$$

因此有

$$P\{\chi^2_{1-\alpha/2}(2n) < 2n\lambda\overline{X} < \chi^2_{\alpha/2}(2n)\} = 1-\alpha$$

指数分布参数λ的$1-\alpha$区间估计为

$$\left(\frac{\chi^2_{1-\alpha/2}(2n)}{2n\overline{X}}, \frac{\chi^2_{\alpha/2}(2n)}{2n\overline{X}}\right)$$

注意这个区间不一定是最短区间

3. 正态总体均值的置信区间

假定样本 X_1，..，X_n 来自总体 $N(\mu, \sigma^2)$

(1) 如果总体方差已知 $\sigma^2 = \sigma_0^2$

此时样本均值的分布为 $N(\mu, \frac{\sigma_0^2}{n})$

$$P\{|\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma_0}| \le u_{\alpha/2}\} = 1-\alpha$$

总体均值 μ 的
$1-\alpha$ 的区间估计：$(\overline{X} - u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{X} + u_{\alpha/2}\frac{\sigma_0}{\sqrt{n}})$

(2) 如果总体方差σ^2未知

需要使用抽样分布中定理1.3.1

$$\frac{\sqrt{n}\,(\overline{X}-\mu)}{S} \sim t(n-1)$$

因此有：

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

总体均值μ的1 - α区间估计为：

$$(\overline{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \overline{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}})$$

这两个区间估计都是最短区间

例2.3.4　科学上的很多重大发现往往由年轻人提出，下表是 16 世纪中期到 20 世纪的 12 项重大科学突破的情况：

科学发现	科学家	时间	年龄
日心说	哥白尼	1543	40
望远镜, 天文学基本定律	伽利略	1600	43
动力学, 万有引力, 微积分	牛顿	1665	23
电的本质	富兰克林	1746	40
燃烧即氧化	拉瓦锡	1774	31

地球的演变	莱尔	**1830**	**33**
进化论	达尔文	**1858**	**49**
光的电磁特性	麦克思韦	**1864**	**33**
放射性	居里	**1896**	**34**
量子力学	普朗克	**1901**	**43**
狭义相对论	爱因斯坦	**1905**	**26**
量子力学的数学基础	薛定谔	**1926**	**39**

假定数据来自期望、方差未知时的正态总体，问什么年龄段科学家们将可能做出重要的工作？

解. 首先计算样本统计量，

$$\bar{x} \approx 36.17, \quad s \approx 7.53$$

现在有12个样本，因此抽样误差是

$$t_{0.025}(11)\frac{s}{\sqrt{12}} = 2.201 \times \frac{7.53}{3.4641} = 4.78$$

可以构造出一个区间 (31.4, 41.0) □

历史数据表明，科学家研究工作的黄金时期是31岁半到41岁间。这个年龄段他们将有可能做出杰出的工作。

这个结论的可靠程度是 95% 。

4. 正态总体方差的置信区间

只讨论μ未知的情况，由抽样分布定理1.3.1

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$P\{\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\} = 1-\alpha$$

虽然卡方分布的密度函数不是对称函数，习惯上仍然取总体方差σ^2的$1-\alpha$的区间估计为：

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

5. 两个正态总体均值差的置信区间

假定从总体 $X \sim N(\mu_1, \sigma_1^2)$ 中抽取 n_1 个样本，从另个独立的总体 $Y \sim N(\mu_2, \sigma_2^2)$ 中抽取 n_2 个样本；

相应的样本均值与样本方差分别为：

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \quad , \quad S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$$

$$\overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j \quad , \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_j - \overline{Y})^2$$

根据抽样分布中定理1.3.2，有

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

这里 $S_W^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

因此均值差 μ_1 - μ_2 的1 - α 区间估计为：

$$(\overline{X}-\overline{Y}-t_{\alpha/2}(n_1+n_2-2)S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

$$\overline{X}-\overline{Y}+t_{\alpha/2}(n_1+n_2-2)S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})$$

6. 两个正态总体方差比的置信区间

仍然根据抽样分布中定理1.3.2，有

$$\frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

因此方差比 σ_1^2/σ_2^2 的1 - α 区间估计为：

$$\left(\frac{S_1^2 / S_2^2}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}, \frac{S_1^2 / S_2^2}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)} \right)$$

置信水平的理解

由于总体参数是未知的，对于每一个计算出来的区间估计，它要么包含总体参数，要么不包含，只是我们不知道而已；

但我们可以肯定，如果采用某种方法构造出一个置信水平 **0.95** 的区间(这个区间的两个端点是统计量的函数)，当我们代入 **100** 次统计量的数据从而得到**100** 个区间时，平均有 **95** 个区间要包含总体参数。

样本容量对区间长度的影响

以95%的区间估计为例，

总体比例 $2\times 1.96\sqrt{\dfrac{p_s(1-p_s)}{n}}$

方差未知正态总体 $2\times t_{0.025}(n-1)\dfrac{s}{\sqrt{n}}$

方差已知正态总体 $2\times 1.96\dfrac{\sigma_0}{\sqrt{n}}$

4 倍的样本容量，抽样误差才可能缩减一半

习题 2.3

1-2. 教材 **100** 页

第 **23**、**24**题。

第3章 假设检验

第3.1节 基本理论

第3.2节 重要参数检验

第3.3节 非参数检验

第3.4节 统计决策与Bayes 理论

东北大学数学系

一. 假设检验的背景

1. 假设检验(*Test of Hypothesis*) 含义

它是如下的一种统计推断：

对于一个统计模型，我们提出一个假设，根据抽取到的样本，来作出是接受还是拒绝这个假设。

小概率事件在一次试验中不应该发生。

2. 假设检验理论的重要历史事件

(1) K. Pearson 的拟合优度检验，1900年

(2) R. A. Fisher 的显著性检验，1920年代

(3) J. Neyman 与E. S. Pearson 理论，1928年开始

(4) A. Wald 的统计决策理论，1950年

(5) Bayes 方法，基本观点从 T. Bayes (1702~1761) 开始，二战后得到巨大发展。

女士品茶

有一种饮料由 **Tea** 和 **Milk** 混合而成，
按照顺序的不同，分为 **TM**、 **MT** 两种，

有位女士声称她有能力品尝出是 **TM** 还是**MT** 。

为了检验她的说法是否可信，准备 **8** 杯饮料，**TM** 和 **MT** 各一半，并且把这一点告诉她。

现在随机的让这位女士品尝，指出哪些是 **TM** ，
最终的结果是她全部说对了。

R.A.Fisher 的推理过程如下：

引进一个假设，

H_0：这位女士没有鉴别能力

如果 H_0 是正确的，她只能随机从 **8** 杯饮料中猜测 **4** 杯说是 **TM** 。全部猜对的概率为：

$$\frac{1}{C_8^4} = \frac{1}{70} \approx 0.014$$

现在她正确的说出了全部的 **TM**，要解释这种现象，只能有下面两种可能：

(1) H_0 不成立，即：她的确有鉴别能力；

(2) H_0 成立，意味着一件概率为 0.014 的随机事件在一次试验中发生了。

一个概率不到 2% 的随机事件在一次试验中发生了，这是比较稀奇或者说不太可能的。

Fisher 认为，随机试验的结果(或样本) 构成不利于假设 H_0 的显著性证据，因此应该否定H_0 。

这种推理过程就称为：显著性检验

显著性是统计意义上的显著，意思是一个小概率事件是否发生。

思考1　假如这位女士只说对了 3 杯？

一个人纯粹靠随机的猜测，能够说对至少 3 杯的概率（即 H_0 成立的情况下，出现这种试验结果的可能性）：

$$\frac{1+C_4^3 C_4^1}{C_8^4} = \frac{17}{70} \approx 0.243$$

显然我们不会对一个概率接近 25% 的随机事件在一次试验中发生而感到惊讶。

试验结果并没有提供不利于H_0 的显著性证据，因此不能否定零假设，而应该接受H_0，即应该认为这位女士没有鉴别能力。

二. 假设检验的基本过程

通过一个简单的例题来说明

例**3.1.1** 当机器正常工作时，生产的滚珠的直径应该是一个服从均值 **15.1 mm**，方差**0.05mm** 的正态随机变量。有一天随机地抽取了 **6** 个成品滚珠，测量出它们的平均直径是 **14.95 mm**，问这台机器是否正常工作？

(假定即使工作异常标准差也不会改变)

1. 提出一个统计假设

根据题意每个滚珠直径 $X \sim N(\mu, 0.05)$，

如果机器正常工作，应该是 $\mu = 15.1$，反之应该是 $\mu \neq 15.1$。

因此首先提出统计假设：

$$H_0: \mu = \mu_0 (= 15.1) \Leftrightarrow H_1: \mu \neq \mu_0 (\neq 15.1)$$

假设检验的任务就是要根据抽取出的样本，来决定是接受零假设，还是拒绝零假设 (接受对立假设) 。

2. 选取一个合适的检验统计量

它的分布当零假设成立时应该是已知的，而且一般是从待检验的总体参数的良好的点估计中去寻找。

在例题中需要检验的是总体期望 μ，因此考虑样本均值，

$$\overline{X} \sim N(\mu, \frac{0.05}{6})$$

零假设成立时 $(\mu = 15.1)$ 则有：

$$z = \frac{\sqrt{6}(\overline{X} - 15.1)}{\sqrt{0.05}} \sim N(0,1)$$

3. 利用零假设成立时检验统计量的分布构造出一个小概率事件

这个小概率就是给定的显著性水平(也称检验水平)，而这个小概率事件就是零假设的拒绝域，并且拒绝域必须和对立假设有关：零假设的拒绝域相当于对立假设的接受域 。

在例题中由于样本均值是总体期望 μ 的一个良好的点估计，因此零假设成立($\mu = 15.1$)时，偏差$|\overline{X} - 15.1|$ 应该比较小，不能够太大。

而如果$|\overline{X}-15.1|$比较大时，自然我们会认为零假设不成立，所以应该接受对立假设。所以零假设 $(\mu = 15.1)$ 的拒绝域的形式就是：

$$\text{统计量}|z| = \frac{\sqrt{6}|\overline{X}-15.1|}{\sqrt{0.05}} > \text{某个常数 } z_0$$

根据检验统计量的分布，有：

$$P\{\frac{\sqrt{6}|\overline{X}-15.1|}{\sqrt{0.05}} > u_{\alpha/2}\} = \alpha$$

这个常数 z_0 就可以取为 $u_{\alpha/2}$

4. 代入样本观察值，如果使得这个小概率事件发生，就否定零假设而去接受对立假设。否则说明样本没有提供否定零假设的显著性证据，因此应该接受零假设。

在这个例题里，检验统计量

$$|z| = \frac{\sqrt{6}\,|14.95 - 15.1|}{\sqrt{0.05}} \approx 1.6432,$$

$\mathbf{H_0}$：$\mu = \mu_0\ (=15.1) \Leftrightarrow \mathbf{H_1}$：$\mu \neq \mu_0\ (\neq 15.1)$

的显著水平α的拒绝域就是$\{1.6432 > u_{\alpha/2}\}$。

假设检验默认的显著水平是α=0.05

(1) 如果取 α=0.05，则1.64＜常数 $z_0 = 1.96$ ，说明一个小概率事件没有发生(此时认为 $p \le 0.05$的事件为小概率事件)，样本没有 提供机器异常的显著证据，应该接受零假设；

(2) 如果取$\alpha = 0.15$，则1.64＞常数 $z_0 = 1.43$ ，说明一个小概率事件发生了(此时认为 $p \le 0.15$的事件为小概率事件)，样本提供了机器异常的显著证据，应该否定零假设；

在不同的显著水平下，可以导致最终得出的检验结论完全不同。这个现象说明了显著水平α对于 H_0的保护：α越小越不容易否定零假设。

三. Neyman – Pearson 理论

1. 零假设与对立假设 *null hypothesis* 与 *alternative hypothesis*

如果有样本 X，取值于样本空间 Λ。我们只知道 X 的分布属于一个分布族$\{F_\theta, \theta \in \Theta\}$。

Θ_0 是参数空间Θ中的一个非空子集， Θ_1 是与Θ_0不相交的参数空间Θ中另一个非空子集。

则称： H_0： $\theta \in \Theta_0$ $\Leftrightarrow$ H_1： $\theta \in \Theta_1$

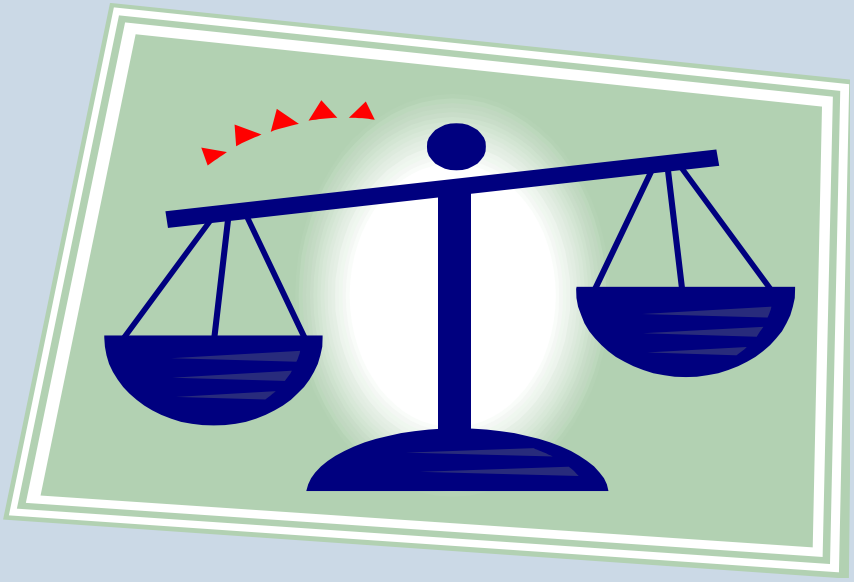
是一个假设检验问题。

H_0 称为零假设， H_1 称为对立假设

Remark

一般来说，零假设总是“受到保护的假设”，没有充分的证据是不能拒绝零假设的。

因此在实际检验问题中检验者总是把自己的倾向性观点作为零假设、再辅助以比较低的检验水平，从而使得没有充分的证据就不能否定它。



2. 否定域(或拒绝域)

有了样本以后需要制定一个法则，根据这个法则去决定究竟是接受零假设，还是拒绝它而接受对立假设，因此一个假设检验的处理，实际上是等价于把样本空间分解成两个不相交的部分：

A_0 (接受域) 与 A_1 (否定域)

3. 两类检验错误

无论我们采用什么决策，任何一个假设检验都必然可能犯两种错误之一：

第一类错误“拒真”：零假设本身是正确的却被检验拒绝，接受的是一个错误的对立假设即真实参数 $\theta \in \Theta_0$，但是我们认为它在 Θ_1 中；

第二类错误“采假”：零假设本身是错误的而没有被拒绝，最后接受的是一个错误的零假设即真实参数 $\theta \in \Theta_1$，但是我们认为它在 Θ_0 中。

在理论上可以证明，假设检验的这两类错误的关系如同区间估计中置信度与精度的关系。

如果样本容量固定，则一个检验方法不可能同时使得犯这两类错误的概率都很小。

假设检验和区间估计具有一种内在的联系

实际上，可以把估计未知参数的区间就取成零假设的接受域，反之同理，一个关于参数的假设检验的接受域同样可以作为估计参数的区间。

从发展历史来说先有假设检验的 **N-P** 理论，后有 **Neyman** 的置信区间(**1934**年开始)

4. 功效函数 $\beta_\phi(\theta)$

$\beta_\phi(\theta) = P_\theta\{\text{否定零假设 } H_0\}$，$\theta \in \Theta$

	$\theta \in \Theta_0$	$\theta \in \Theta_1$
P_I	$\beta_\phi(\theta)$	0
P_{II}	0	$1 - \beta_\phi(\theta)$

一个好的检验应该是功效函数 $\beta_\phi(\theta)$在Θ_0中越小越好，而在 Θ_1中越大越好。

Neyman 和 Pearson 提出的解决办法

首先用一个很小的数 $\alpha > 0$ 控制犯第一类错误的概率，在这个基础上再考虑使得犯第二类错误的概率尽量的小。

“犯第一类错误的概率” α 在统计学中称为显著性水平，或者又叫检验水平。

而检验的水平体现了对零假设的保护程度，

显著性水平越小，越不容易否定零假设

例**3.1.2**　教师给学生出了**10** 个判断题。为了检验学生是否在猜答案，他采用了如下的决策：

如果答对了**7** 个或以上的题，就不是凭猜测；

如果答对的题数少于**7** 个，学生就是在猜答案。

这个决策原则犯第一类错误的概率是多少？

解.　以 p 记答对一个题的概率，则零假设应是：

$$H_0：p = 0.5$$

显然答对的全部题数 $X \sim B(10，p)$。

第一类错误是指：教师认为学生没有猜题而大家的确是靠猜测做题的概率，

因此当 $p = 0.5$ 时，概率 $P\{X \geq 7\}$ 即为所求，

$$P\{X \geq 7\} = P\{X = 7\} + \ldots + P\{X = 10\}$$
$$= 0.1172 + 0.0439 + 0.0098 + 0.0010$$
$$= 0.1719。$$

当学生确实在猜答案时，而认为他们不是猜题的概率是 17% (犯第一类错误的概率)。

如果把决策原则改成“答对题数少于 8 个才认为在猜答案”，则第一类错误的概率降为 0.0547。

□

$P_{p=0.7}\{X \leq 6\}$ 是真实值0.7时犯第二类错误的概率

例3.1.3　如果一组样本X_1，…，X_n 来自总体 $X \sim N(\mu, \sigma_0^2)$，$\sigma_0^2$ 已知。对于假设：

$$\mathbf{H}_0: \ \mu \le \mu_0 \Leftrightarrow \mathbf{H}_1: \ \mu > \mu_0$$

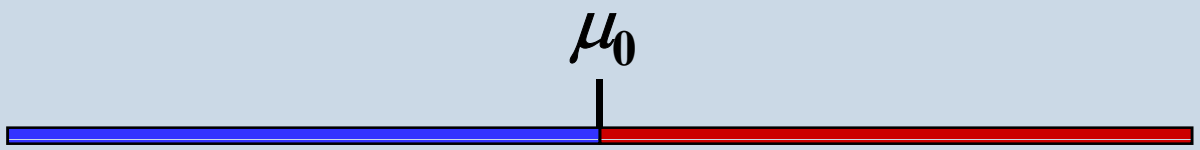
类似例 3.1.1 的讨论，要去构造一个检验：

$\overline{X}$ ＞ 某个常数 C 时拒绝$\mathbf{H}_0$，否则接受。

因为参数空间$\Theta = \mathbf{R}^1$，这种检验的功效函数为：

$$\beta_\phi(\mu) = \mathbf{P}_\mu\{\overline{X} > C\} = 1 - \Phi\left(\frac{\sqrt{n}(C-\mu)}{\sigma_0}\right)$$

$\Phi(x)$ 是$N(0,1)$的分布函数，单调非降，故功效函数 $\beta_\phi(\mu) = 1 - \Phi\left(\frac{n^{1/2}(C-\mu)}{\sigma_0}\right)$ 随着μ的增加而增大，而现在参数空间Θ被分解成两个不相交的部分 $H_0: \mu \le \mu_0$ 与 $H_1: \mu > \mu_0$



要希望第一类错误概率越小，即在Θ_0中μ应该越向左边偏离 μ_0来取值；而要希望第二类错误概率越小，在Θ_1中μ应该越向右边远离 μ_0来取值。

矛盾:两类错误的概率不可能同时都很小

1. $\beta_\phi(\mu)$是μ的单调递增函数，则$\mu \in \Theta_0 (\mu \le \mu_0)$时 $\beta_\phi(\mu)$在$\mu = \mu_0$达到最大，要使得检验犯第一类错误的概率$\le \alpha$，只需要取 $\beta_\phi(\mu_0) = \alpha$即可。

$$\text{解出 } C = \mu_0 + \frac{u_\alpha \sigma_0}{n^{1/2}}$$

因此例题中的一个水平α的检验 ϕ 就是：

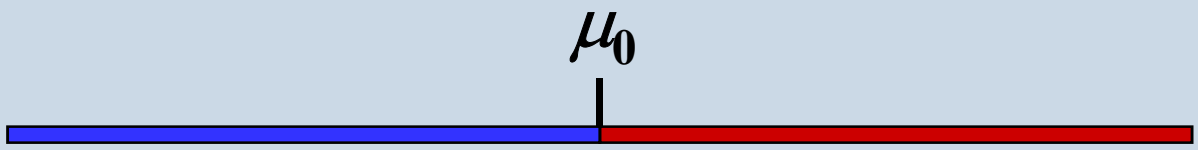
$$\overline{X} > \mu_0 + \frac{u_\alpha \sigma_0}{\sqrt{n}} \text{ 时拒绝} \mathbf{H}_0\text{，否则接受。}$$

而这个检验的功效函数是：

$$\beta_\phi(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} + u_\alpha\right) = \Phi\left(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma_0} - u_\alpha\right)$$

2. 详细分析这个检验的功效函数：

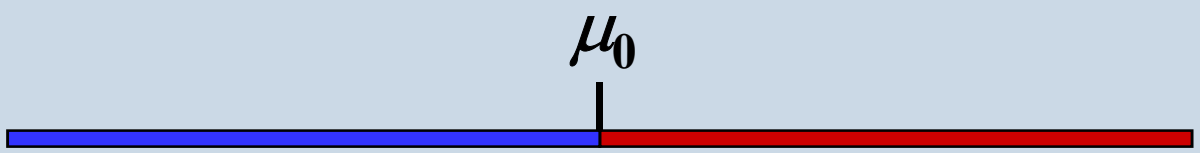
$$\beta_\phi(\mu) = \Phi\left(\frac{\sqrt{n}(\mu-\mu_0)}{\sigma_0} - u_\alpha\right)$$



(1). 真实的μ在$\Theta_1(\mu > \mu_0)$中越大，则$\beta_\phi(\mu)$越大，第二类错误的概率就越小，直观理解，这时μ离零假设μ_0越远也就越容易把它们区别开。

如果真实的μ在$\Theta_1(\mu > \mu_0)$中越靠近μ_0，则检验犯第二类错误的概率$\approx 1 - \alpha$。

$$\beta_\phi(\mu) = \Phi\left(\frac{\sqrt{n}(\mu - \mu_0)}{\sigma_0} - u_\alpha\right)$$



(2). α 越大 u_α 将越小则 $\beta_\phi(\mu)$ 也就越大；反之，α 越小 u_α 将越大，则 $\beta_\phi(\mu)$ 也就越小。直观理解 α 大即允许犯第一类错误的概率增加，作为一种补偿，相应地犯第二类错误的概率就应该要降低。

如何控制犯第二类错误的概率？

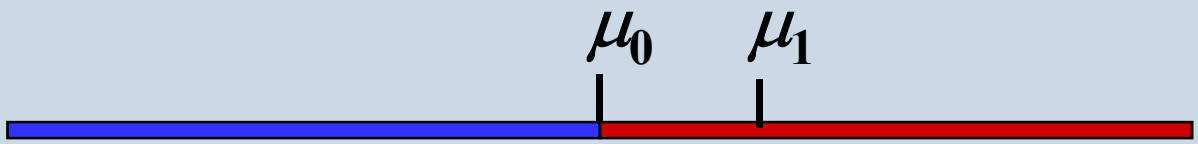
思想基本上和序贯区间估计的想法一样，
对样本容量的要求

如果在原来的检验问题里，还希望犯第二类错误的概率 $\leq$ 事先指定的一个小的正数 β，即

当 $\mu > \mu_0$ 时 $(\mu \in \Theta_1)$ 有 $\beta_\phi(\mu) \geq 1 - \beta$ ，

按照前面的分析，当 μ 非常接近 μ_0 时这个要求实际上办不到，所以可行做法是放松条件：

$\beta_\phi(\mu) \geq 1 - \beta$，当 $\mu > \mu_1$ (μ_1 比 μ_0 稍大)



同样道理，因为 $\beta_\phi(\mu)$是μ 的单调递增函数，因此只要 $\beta_\phi(\mu_1) \geq 1-\beta$ ，就可以保证：

当 $\mu > \mu_1$ 时，都有 $\beta_\phi(\mu) \geq 1-\beta$。即，

$$\beta_\phi(\mu_1) = \Phi\left(\frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma_0} - u_\alpha\right) \geq 1-\beta$$

所以只要 $n \geq \dfrac{\sigma_0^2 (u_\alpha + u_\beta)^2}{(\mu_1-\mu_0)^2}$ ，就可以保证犯第一类错误的概率 $\leq \alpha$，同时当参数真实值大于μ_1 时，犯第二类错误的概率 $\leq \beta$。 □

例3.1.4　如果一组样本X_1，…，X_n 来自总体 $X \sim N(\mu, 2.6^2)$，对于假设：

$$H_0: \ \mu \le 12 \Leftrightarrow H_1: \ \mu = 13$$

(1) 构造此假设的一个水平α的检验，并计算 n=100时该检验的两类错误 p_1，p_2；

(2) n至少为多少才能使得 $p_2 \le 0.01$？

解. 类似上题的讨论，水平α的检验 ϕ 是

$$\overline{X} > 12 + \frac{u_\alpha \times 2.6}{\sqrt{n}}$$ 时拒绝H_0，否则接受。

n=100，α=0.05时拒绝域为：$\overline{X} > 12.4264$.

(1) 这个水平α的检验 ϕ 的功效函数仍然是

$$\beta_\phi(\mu)=\Phi(\frac{\sqrt{n}(\mu-\mu_0)}{\sigma_0}-u_\alpha)=\Phi(\frac{10(\mu-12)}{2.6}-1.64);$$

因此检验 ϕ 的第一类错误为:

$$p_1=\Phi(\frac{10(\mu-12)}{2.6}-1.64)|_{\mu\le 12}\le\Phi(-1.64)=0.05;$$

而检验 ϕ 的第二类错误为:

$$p_2=1-\Phi(\frac{10(13-12)}{2.6}-1.64)=1-\Phi(2.206)=0.01369\ .$$

(2) 要使得第二类错误的概率不超过0.01， 即

$$p_2 = 1 - \Phi\left(\frac{\sqrt{n}(13-12)}{2.6} - 1.64\right) \le 0.01,$$

$$\Phi\left(\frac{\sqrt{n}}{2.6} - 1.64\right) \ge 0.99,$$

$$\frac{\sqrt{n}}{2.6} - 1.64 \ge 2.3263,$$

$$n \ge 106.6$$

所以样本容量 n至少需要 107 。 □

习题 3.1

1-2. 教材 **175** 页

第 **1**、**2** 题。

第3.2节 重要的参数检验

这一节主要讨论

1. 正态总体均值、方差的检验问题，包括两个独立正态总体之间均值差、方差比的检验；
2. 指数总体参数的检验问题；
3. 两点分布参数的检验，主要是一个总体百分比的检验。

只构造水平为α的检验，不讨论犯第二类错误的概率

一. 正态总体参数的检验

1. 均值的检验

样本X_1，…，X_n 来自总体 $X \sim N(\mu, \sigma_0^2)$

一个总体的均值检验问题

1.1 方差已知 σ_0^2 (u 检验)

1.1a. $H_0: \mu = \mu_0 \Leftrightarrow H_1: \mu \neq \mu_0$

1.1b. $H_0: \mu \leq \mu_0 \Leftrightarrow H_1: \mu > \mu_0$

1.1c. $H_0: \mu \geq \mu_0 \Leftrightarrow H_1: \mu < \mu_0$

利用样本均值来作为检验统计量，分布是：

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma_0} \sim N(0,1)$$

当零假设 $\mathbf{H_0}$：$\mu=\mu_0$ 成立时有：

$$\mathbf{P}\{\frac{\sqrt{n}\,|\overline{X}-\mu_0|}{\sigma_0} > u_{\alpha/2}\} = \alpha$$

因此假设 *1.1a*. $\mathbf{H_0}$：$\mu=\mu_0 \Leftrightarrow \mathbf{H_1}$：$\mu\neq\mu_0$

一个水平α的拒绝域应该是

$$\frac{\sqrt{n}\,|\overline{X}-\mu_0|}{\sigma_0} > u_{\alpha/2}$$

假设 *1.1b*. $\mathbf{H_0}$: $\mu \le \mu_0 \Leftrightarrow \mathbf{H_1}$: $\mu > \mu_0$

例题3.1.3中已经讨论，一个水平α的拒绝域是

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma_0} > u_\alpha$$

假设 *1.1c*. $\mathbf{H_0}$: $\mu \ge \mu_0 \Leftrightarrow \mathbf{H_1}$: $\mu < \mu_0$

同理讨论，一个水平α的拒绝域如下

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma_0} < -u_\alpha$$

p-值 (*p-value*)

p-值是零假设成立时得到所观测数据或者更为极端数据的概率。

p-值越小越应该否定零假设。有时候 *p*-值又被理解成拒绝零假设的最小显著性水平。

p-值的计算依赖于所使用的检验统计量以及对立假设是单边假设还是双边假设。

拟合优度检验中的 *p*-值又称为是拟合优度，即数据与总体分布拟合的程度。

例3.2.1　从正态总体 $X \sim N(\mu, 10.1^2)$ 中随机抽取 100 个样本，样本均值是 59.22。检验以下假设：

$$H_0: \ \mu = 60 \Leftrightarrow H_1: \ \mu < 60$$

解. 从对立假设的形式知道零假设的水平 α 的拒绝域应该是样本均值偏小，即

$$z = \frac{10 \times (59.22 - 60)}{10.1} < -u_\alpha$$

根据样本数据计算出的 $z = -0.77$，

$$u_{0.05} = 1.64, \quad u_{0.20} = 0.84, \quad u_{0.25} = 0.68$$

甚至在0.2的水平下也不会拒绝零假设，
但是在水平0.25下将会拒绝零假设。

如果采用 ***EXCEL***或 ***SPSS*** 等软件处理这个问题，则将输出一个检验统计量 z 对应的 p-值：

$$p = \mathrm{P}\{z < -0.77\} = 0.22065$$

只要这个p-值小于事先给出的水平α，就否定零假设

□

Remark 如果对立假设是 H_1： $\mu \neq 60$，则检验的拒绝域的形式应该为 $|z| > u_{\alpha/2}$。

因此在**0.4**的水平下也不会拒绝零假设，但是在水平**0.50**下将会拒绝零假设。

相应的 p-值将是 $p = \mathrm{P}\{|z| > 0.77\} = 0.4413$

1.2 方差σ^2未知 (t 检验)

样本X_1，…，X_n 来自总体 $X \sim N(\mu, \sigma^2)$

1.2a. H_0: $\mu = \mu_0 \Leftrightarrow H_1$: $\mu \neq \mu_0$

1.2b. H_0: $\mu \leq \mu_0 \Leftrightarrow H_1$: $\mu > \mu_0$

1.2c. H_0: $\mu \geq \mu_0 \Leftrightarrow H_1$: $\mu < \mu_0$

检验的构造与方差已知时基本相同，需要使用抽样分布中的定理**1.3.1**

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t(n-1)$$

假设 *1.2a*. $H_0: \mu = \mu_0 \Leftrightarrow H_1: \mu \neq \mu_0$

一个水平α的拒绝域是 $\dfrac{\sqrt{n}\,|\overline{X}-\mu_0|}{S} > t_{\alpha/2}(n-1)$

假设 *1.2b*. $H_0: \mu \leq \mu_0 \Leftrightarrow H_1: \mu > \mu_0$

一个水平α的拒绝域是 $\dfrac{\sqrt{n}(\overline{X}-\mu_0)}{S} > t_{\alpha}(n-1)$

假设 *1.2c*. $H_0: \mu \geq \mu_0 \Leftrightarrow H_1: \mu < \mu_0$

一个水平α的拒绝域是 $\dfrac{\sqrt{n}(\overline{X}-\mu_0)}{S} < -t_{\alpha}(n-1)$

两个总体均值差的检验

假定两组简单随机样本 $X_1, \ldots, X_{n_1}$ 与 $Y_1, \ldots, Y_{n_2}$ 分别来自两个独立的正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$，

1.3 方差 σ_1^2、σ_2^2已知 (u 检验)

1.3a. $H_0: \mu_1 - \mu_2 = \delta \Leftrightarrow H_1: \mu_1 - \mu_2 \neq \delta$

1.3b. $H_0: \mu_1 - \mu_2 \leq \delta \Leftrightarrow H_1: \mu_1 - \mu_2 > \delta$

1.3c. $H_0: \mu_1 - \mu_2 \geq \delta \Leftrightarrow H_1: \mu_1 - \mu_2 < \delta$

因为两个总体的样本均值分别服从正态分布，

$$\overline{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1}), \quad \overline{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

所以有

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

则当$\mathbf{H_0}$：$\mu_1 - \mu_2 = \delta$成立时，统计量 $z \sim N(0, 1)$

$$z = \frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

假设 *1.3a*. H_0: $\mu_1 - \mu_2 = \delta \Leftrightarrow H_1$: $\mu_1 - \mu_2 \neq \delta$
的一个水平α的拒绝域是 $$\frac{|\overline{X} - \overline{Y} - \delta|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > u_{\alpha/2}$$

假设 *1.3b*. H_0: $\mu_1 - \mu_2 \leq \delta \Leftrightarrow H_1$: $\mu_1 - \mu_2 > \delta$
的一个水平α的拒绝域是 $$\frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > u_{\alpha}$$

假设 *1.3c*. H_0: $\mu_1 - \mu_2 \geq \delta \Leftrightarrow H_1$: $\mu_1 - \mu_2 < \delta$
的一个水平α的拒绝域是 $$\frac{\overline{X} - \overline{Y} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -u_{\alpha}$$

1.4 方差 σ_1^2、σ_2^2未知但相等(t 检验)

$\sigma_1^2 \neq \sigma_2^2$ 时称为Behrens-Fisher问题

使用抽样分布中的定理1.3.2，定义：

$$S_W^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

则有 $$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

则当H_0：$\mu_1 - \mu_2 = \delta$ 成立时，统计量：$$\frac{\overline{X}-\overline{Y}-\delta}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

假设 ***1.4a***. $H_0: \mu_1 - \mu_2 = \delta \Leftrightarrow H_1: \mu_1 - \mu_2 \neq \delta$
的一个水平α的拒绝域是 $$\frac{|\overline{X} - \overline{Y} - \delta|}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{\alpha/2}(n_1+n_2-2)$$

假设 ***1.4b***. $H_0: \mu_1 - \mu_2 \leq \delta \Leftrightarrow H_1: \mu_1 - \mu_2 > \delta$
的一个水平α的拒绝域是 $$\frac{\overline{X} - \overline{Y} - \delta}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{\alpha}(n_1+n_2-2)$$

假设 ***1.4c***. $H_0: \mu_1 - \mu_2 \geq \delta \Leftrightarrow H_1: \mu_1 - \mu_2 < \delta$
的一个水平α的拒绝域是 $$\frac{\overline{X} - \overline{Y} - \delta}{S_W\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < -t_{\alpha}(n_1+n_2-2)$$

2. 方差的检验

一个总体的方差检验问题

2.1 均值 μ_0已知(χ^2 检验)

2.1a. $H_0: \sigma^2 = \sigma_0^2 \Leftrightarrow H_1: \sigma^2 \neq \sigma_0^2$

2.1b. $H_0: \sigma^2 \leq \sigma_0^2 \Leftrightarrow H_1: \sigma^2 > \sigma_0^2$

2.1c. $H_0: \sigma^2 \geq \sigma_0^2 \Leftrightarrow H_1: \sigma^2 < \sigma_0^2$

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} \sim N(0,1) \quad or \quad \sum_{k=1}^{n}\frac{(X_k-\mu_0)^2}{\sigma^2} \sim \chi^2(n)?$$

当零假设 $\mathbf{H_0}$：$\sigma^2 = \sigma_0^2$ 成立时有：

$$\sum_{k=1}^{n} \frac{(X_k - \mu_0)^2}{\sigma_0^2} \sim \chi^2(n)$$

因此假设 *2.1a*. $\mathbf{H_0}$：$\sigma^2 = \sigma_0^2 \Leftrightarrow \mathbf{H_1}$：$\sigma^2 \neq \sigma_0^2$
一个水平α的拒绝域应该是

$$\sum_{k=1}^{n} \frac{(X_k - \mu_0)^2}{\sigma_0^2} > \chi^2_{\alpha/2}(n) \quad or \quad \sum_{k=1}^{n} \frac{(X_k - \mu_0)^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}(n)$$

同理假设 *2.1b*. $\mathbf{H_0}$: $\sigma^2 \le \sigma_0^2 \Leftrightarrow \mathbf{H_1}$: $\sigma^2 > \sigma_0^2$

一个水平α的拒绝域是

$$\sum_{k=1}^{n} \frac{(X_k - \mu_0)^2}{\sigma_0^2} > \chi_\alpha^2(n)$$

假设 *2.1c*. $\mathbf{H_0}$: $\sigma^2 \ge \sigma_0^2 \Leftrightarrow \mathbf{H_1}$: $\sigma^2 < \sigma_0^2$

一个水平α的拒绝域是

$$\sum_{k=1}^{n} \frac{(X_k - \mu_0)^2}{\sigma_0^2} < \chi_{1-\alpha}^2(n)$$

2.2 均值 μ 未知(χ^2 检验)

2.2a. $H_0: \sigma^2 = \sigma_0^2 \Leftrightarrow H_1: \sigma^2 \neq \sigma_0^2$

2.2b. $H_0: \sigma^2 \leq \sigma_0^2 \Leftrightarrow H_1: \sigma^2 > \sigma_0^2$

2.2c. $H_0: \sigma^2 \geq \sigma_0^2 \Leftrightarrow H_1: \sigma^2 < \sigma_0^2$

根据抽样分布中定理1.3.1中关于样本方差的分布，有

$$\sum_{k=1}^{n} \frac{(X_k - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

当零假设 $\mathbf{H_0}$：$\sigma^2 = \sigma_0^2$ 成立时有：

$$\sum_{k=1}^{n} \frac{(X_k - \overline{X})^2}{\sigma_0^2} \sim \chi^2(n-1)$$

因此假设 *2.2a*. $\mathbf{H_0}$：$\sigma^2 = \sigma_0^2 \Leftrightarrow \mathbf{H_1}$：$\sigma^2 \neq \sigma_0^2$

一个水平α的拒绝域应该是

$$\sum_{k=1}^{n} \frac{(X_k - \overline{X})^2}{\sigma_0^2} > \chi^2_{\alpha/2}(n-1) \quad or \quad \sum_{k=1}^{n} \frac{(X_k - \overline{X})^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}(n-1)$$

同理假设 *2.2b*. $\mathbf{H_0}$: $\sigma^2 \le \sigma_0^2 \Leftrightarrow \mathbf{H_1}$: $\sigma^2 > \sigma_0^2$
一个水平α的拒绝域是

$$\sum_{k=1}^{n} \frac{(X_k - \overline{X})^2}{\sigma_0^2} > \chi_\alpha^2(n-1)$$

假设 *2.2c*. $\mathbf{H_0}$: $\sigma^2 \ge \sigma_0^2 \Leftrightarrow \mathbf{H_1}$: $\sigma^2 < \sigma_0^2$
一个水平α的拒绝域是

$$\sum_{k=1}^{n} \frac{(X_k - \overline{X})^2}{\sigma_0^2} < \chi_{1-\alpha}^2(n-1)$$

两个总体方差比的检验

假定两组简单 随机样本 X_1，…，X_{n_1} 与 Y_1，…，Y_{n_2} 分别来自两个独立的正态总体 $X \sim N(\mu_1, \sigma_1^2)$ 与 $Y \sim N(\mu_2, \sigma_2^2)$，

2.3 均值 μ_1、μ_2 都未知 (F 检验)

2.3a. H_0: $\sigma_1^2 = \sigma_2^2 \Leftrightarrow H_1$: $\sigma_1^2 \neq \sigma_2^2$

2.3b. H_0: $\sigma_1^2 \leq \sigma_2^2 \Leftrightarrow H_1$: $\sigma_1^2 > \sigma_2^2$

2.3c. H_0: $\sigma_1^2 \geq \sigma_2^2 \Leftrightarrow H_1$: $\sigma_1^2 < \sigma_2^2$

由于两个总体的样本方差 S_1^2、S_2^2 分别是它们各自方差 σ_1^2、σ_2^2的无偏估计，因此，

对于假设 *2.3a*. H_0：$\sigma_1^2 = \sigma_2^2 \Leftrightarrow H_1$：$\sigma_1^2 \neq \sigma_2^2$ 拒绝域的形式应该是 S_1^2/S_2^2 偏大或偏小；

同理假设 *2.3b*. H_0：$\sigma_1^2 \leq \sigma_2^2 \Leftrightarrow H_1$：$\sigma_1^2 > \sigma_2^2$ 拒绝域的形式应该是 S_1^2/S_2^2 偏大；

而假设 *2.3c*. H_0：$\sigma_1^2 \geq \sigma_2^2 \Leftrightarrow H_1$：$\sigma_1^2 < \sigma_2^2$ 拒绝域的形式应该是 S_1^2/S_2^2 偏小。

利用抽样分布中定理1.3.2，

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

假设 *2.3a*. H_0: $\sigma_1^2 = \sigma_2^2 \Leftrightarrow H_1$: $\sigma_1^2 \neq \sigma_2^2$
的一个水平α的拒绝域是:

$$S_1^2/S_2^2 > F_{\alpha/2}(n_1 - 1, n_2 - 1) \text{ 或}$$
$$S_1^2/S_2^2 < F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$$

同理假设 *2.3b*. H_0: $\sigma_1^2 \leq \sigma_2^2 \Leftrightarrow H_1$: $\sigma_1^2 > \sigma_2^2$
一个水平α的拒绝域是

$$S_1^2/S_2^2 > F_{\alpha}(n_1 - 1, n_2 - 1)$$

假设 *2.3c*. H_0: $\sigma_1^2 \geq \sigma_2^2 \Leftrightarrow H_1$: $\sigma_1^2 < \sigma_2^2$
一个水平α的拒绝域是

$$S_1^2/S_2^2 < F_{1-\alpha}(n_1 - 1, n_2 - 1)$$

例3.2.2　从一批矿砂中随机抽取了5 个样品来测量镍的百分比含量，数据如下：

3.25，3.27，3.24，3.26，3.24

设测量值总体是正态分布，在显著性水平0.01下能不能认为这批矿砂的镍含量均值为 3.25 ？

解.　这是一个总体方差未知，关于均值的检验：

$$H_0: \mu = 3.25 \Leftrightarrow H_1: \mu \neq 3.25$$

采用t 检验，拒绝域为：

$$\frac{\sqrt{5}\,|\overline{X} - 3.25|}{S} > t_{0.01/2}(4)$$

代入数据，计算出

样本均值为 **3.252**，样本标准差为**0.01304**；

查表得到 t 分布的上侧 **0.005** 分位点

$$t_{0.005}(4) = 4.6041$$

因此有统计量 **0.343** ＜ **4.6041**

即样本没有提供否定零假设的显著证据。 □

$t_{0.25}(4) = 0.7407$ 说明水平 **0.5** 下也不会否定零假设

相应的 p-值是 $p = 0.74887$

例3.2.3　从马克·吐温 (Mark Twain) 和斯诺特格拉斯(Snodgrass) 作品中各选了 8 篇和 10 篇小品文。得到 3 个字母组成单词的比例如下：

Mark Twain	Snodgrass
0.225, 0.262, 0.217, 0.240, 0.230, 0.229, 0.235, 0.217。	0.209, 0.205, 0.196, 0.210, 0.202, 0.207, 0.224, 0.223, 0.220, 0.201。

设这两组数据都来自正态总体而且独立的，问在检验水平 0.05 下，它们的方差是否相同？

解. 这是关于两个独立总体方差是否相同的检验，两个总体的期望都未知，采用*2.3a.*的 F 检验：

$$\mathrm{H}_0:\ \sigma_1^2=\sigma_2^2 \Leftrightarrow \mathrm{H}_1:\ \sigma_1^2\neq\sigma_2^2$$

取检验统计量

$$F=S_1^2/S_2^2$$

$$=\frac{2.12125\times10^{-4}}{9.3344\times10^{-5}}=2.2725$$

拒绝域的形式为：

$$F>F_{0.025}(7,9)\ \text{或者}\ F<F_{0.975}(7,9)$$

查表得到　$F_{0.025}(7, 9) = 4.20$，

但是一般查不到 $F_{0.975}(7, 9)$的数值

必须利用 F 分布的性质，

$$F_{1-\alpha}(m, n) = 1/F_{\alpha}(n, m)$$

所以有

$$F_{0.975}(7, 9) = 1/F_{0.025}(9, 7) = 1/4.82 = 0.2075$$

只有当检验统计量 F (= 2.2725) 大于 4.20 或小于0.2075 时才否定零假设，因此可以认为两个总体方差相同。

相应的 p-值是 $p = 0.125$ □

例3.2.4　在刚才例题里已经通过检验认为方差相同。即马克·吐温与斯诺特格拉斯 的作品中 3 个字母的单词占的比例来自相同方差的正态总体。 那么在水平 0.05 下，能不能认为这两个总体均值有显著差异？

解. 采用*1.4a.* 的 t 检验($\delta = 0$)，要检验的是：

$$H_0: \mu_1 = \mu_2 \Leftrightarrow H_1: \mu_1 \neq \mu_2$$

取检验统计量为

$$T = \frac{\overline{X} - \overline{Y}}{S_W \sqrt{\frac{1}{8} + \frac{1}{10}}}$$

拒绝域是 $|T| > t_{0.025}(16) = 2.1199$

根据例题4.2.3 提供的数据，算出：

$$\bar{x} = 0.2319\,,\quad \bar{y} = 0.2097$$

$$S_w = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{1.4531\times 10^{-4}} = 0.0121$$

最后得到检验统计量 T 的观察值

$$|T| = 3.8679 \ > \ t_{0.025}(16) = 2.1199$$

否定零假设，认为这两位作家写作习惯显著不同。

注：$t_{0.005}(16) = 2.9208$，即水平0.01下也否定。

相应的 p-值是 $p = 0.0014$ □

Remark

假如现在新发现了一部佚名的文集，怀疑可能是马克·吐温或者是斯诺特格拉斯的作品。那么从统计学的角度又应该如何来处理？

这里需要检验的是：一组数据是否来自于一个已知的总体。统计理论中称为非参数检验。

下一节的拟合优度检验能够解决这类问题，“拟合优度”的含义是：一组样本数据和某个总体拟合的程度。

3. 成对数据的检验问题

这也是一种 t 检验，常常用来检验性能、方法或者属性是否有显著差异。

例如要比较不同的药品的疗效，有两个办法：
A. 把患者分成两组，一组服用甲另一组服用乙；
B. 同一组患者各服用甲、乙。

A 方法要检验的就是两个独立正态总体均值是否相同的问题 (应该事先检验方差是否相同)

B 方法中的数据来自正态总体，但不应该认为是独立的，因此要检验的是这些成对数据的差来自的正态总体的均值是否为 **0** 。

例3.2.5　10 个失眠患者分别服用了甲、乙两种安眠药，延长睡眠的时间如下(单位：小时)：

甲	1.9	0.8	1.1	0.1	0.1	4.4	5.5	1.6	4.6	3.4
乙	0.7	–1.6	–0.2	–1.2	–0.1	3.4	3.7	0.8	0.0	2.0

解.　这十组成对数据的差(甲-乙) 是：

1.2　2.4　1.3　1.3　0.2　1　1.8　0.8　4.6　1.4

可以认为它们来自正态总体 $X \sim N(\mu, \sigma^2)$

需要检验的是*1.2a*.　H_0：$\mu = \mu_0 \Leftrightarrow H_1$：$\mu \neq \mu_0$

由于总体方差未知，采用 t 检验，

否定域为：$$|T| = \frac{\sqrt{10}\,|\overline{X}|}{S} > t_{0.025}(9)$$

计算出数据的

样本均值为 1.6 ，样本标准差 1.2028

查表得到 $t_{0.025}(9) = 2.2622$

因为 $|T| = 4.2066 > 2.2622$，在水平 0.05 下否定零假设，认为两种药的疗效有显著差异。

注：$t_{0.005}(9) = 3.2498$，即水平0.01下也否定。

相应的 p-值是 $p = 0.00228$

Remark

如果题目所给不是来自10 个失眠患者的成对数据，而是甲、乙两种药物分别由10人试验所得到的数据，则必须类似例题3.2.4，使用检验*1.4a*

$$H_0: \mu_1 = \mu_2 \Leftrightarrow H_1: \mu_1 \neq \mu_2$$

得到检验统计量 T 的观察值

$$|T| = 1.8981 < t_{0.025}(18) = 2.1009$$

即在水平 0.05 下接受零假设，认为两种药物的疗效没有显著差异。

相应的 p-值是 $p = 0.07384$ □

二. 指数总体参数的检验

如果一组样本X_1，…，X_n 来自参数为 λ 的指数总体，密度函数是：

$$f(x) = \lambda e^{-\lambda x}\ ,\ x > 0$$

可以证明，对于样本均值，有 $2n\lambda\overline{X} \sim \chi^2(2n)$

需要对参数λ (平均寿命的倒数) 作检验：

a. H_0: $\lambda = \lambda_0 \Leftrightarrow H_1$: $\lambda \neq \lambda_0$

b. H_0: $\lambda \leq \lambda_0 \Leftrightarrow H_1$: $\lambda > \lambda_0$

c. H_0: $\lambda \geq \lambda_0 \Leftrightarrow H_1$: $\lambda < \lambda_0$

假设 ***a***. $\mathbf{H_0}$: $\lambda = \lambda_0 \Leftrightarrow \mathbf{H_1}$: $\lambda \neq \lambda_0$
的一个水平α检验的拒绝域为:

$$2n\lambda_0 \overline{X} < \chi^2_{1-\alpha/2}(2n) \quad or \quad 2n\lambda_0 \overline{X} > \chi^2_{\alpha/2}(2n)$$

假设 ***b***. $\mathbf{H_0}$: $\lambda \leq \lambda_0 \Leftrightarrow \mathbf{H_1}$: $\lambda > \lambda_0$
的一个水平α的检验的拒绝域为:

$$2n\lambda_0 \overline{X} < \chi^2_{1-\alpha}(2n)$$

假设 ***c***. $\mathbf{H_0}$: $\lambda \geq \lambda_0 \Leftrightarrow \mathbf{H_1}$: $\lambda < \lambda_0$
的一个水平α的检验的拒绝域为:

$$2n\lambda_0 \overline{X} > \chi^2_{\alpha}(2n)$$

例3.2.6　一种电子元件要求寿命不得低于1000小时，现在随机抽取15个，测量出样本平均寿命为900小时，不妨假定这种元件寿命是指数分布，在0.05的水平下能否认为这批元件合乎要求？

解. 需要检验 H_0：$\lambda \le 0.001 \Leftrightarrow H_1$：$\lambda > 0.001$

拒绝域为 $K^2 < \chi_{0.95}^2(30)$ ($= 18.493$)，

这里统计量 $K^2 = \dfrac{30 \times 900}{1000} = 27$。

样本没有提供元件不合格的显著证据，即应该接收这批元件。

相应的 p-值是 $p = 0.3767$　□

三. 两点分布参数的检验

总体属性比例的检验

假定总体中具有某种属性的比例是 p，从总体抽取了 n 个观察值，p_s 记样本中这种属性的百分比，

$$H_0: p = p_0 \Leftrightarrow H_1: p \neq p_0$$

的水平 α 检验的拒绝域近似为：

$$|z| = \frac{|p_s - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} > u_{\alpha/2}$$

两个总体比例差的检验

假定两个独立总体某种属性比例分别是 p_1 和 p_2，分别抽取n_1、n_2个观察值，样本比例是 p_{s1}、p_{s2}。则

$$H_0: p_1 - p_2 = 0 \Leftrightarrow H_1: p_1 - p_2 \neq 0$$

的水平α 检验的拒绝域近似为：

$$|z| = \frac{|p_{s1} - p_{s2}|}{\sqrt{\frac{p_{s1}(1-p_{s1})}{n_1} + \frac{p_{s2}(1-p_{s2})}{n_2}}} > u_{\alpha/2}$$

四. 似然比检验(*Likelihood ratio test*)

这是基于 **Fisher** 的“似然原理”建立的构造参数检验的一种比较统一的方法

由于似然函数 $L(\theta)$ 的大小体现了在样本给定以后参数θ的“似然性”，考虑如下假设：

$$H_0: \ \theta \in \Theta_0 \ \Leftrightarrow \ H_1: \ \theta \in \Theta_1$$

这里Θ_0与Θ_1构成Θ的一个划分。

“似然比”(统计量)定义成

$$LR = \frac{\sup\{L(\theta), \ \theta \in \Theta_0\}}{\sup\{L(\theta), \ \theta \in \Theta\}}$$

显然有 $0 < LR \leq 1$，根据极大似然估计的性质，当样本容量n 足够大时，极大似然估计应该以很大的概率非常接近于参数 θ 的真实值，因此如果零假设为真，则极大似然估计应该很接近Θ_0，甚至就落在Θ_0里，因此LR应该以相当大的概率接近甚至等于 1。反之如果零假设不真，则LR应该远离 1 而接近于 0 。所以**LR偏小就否定零假设**。

只要事先给出一个检验水平α，计算常数C：

$$\sup\{P_\theta(LR < C),\ \theta \in \Theta_0\} = \alpha$$

从而零假设 H_0：$\theta \in \Theta_0$的拒绝域就是 $\{LR < C\}$

习题 3.2

1-5. 教材 **175** 页

第 **4**、**6**、**8**、**10**、**11** 题。

第3.3节　非参数检验

参数检验是在假定总体分布类型已知，对于总体的未知参数或者数字特征的检验。而非参数检验则不同，它是对于总体分布类型的检验。

其中最重要的是拟合优度检验 (***Goodness-of-fit test***)，即一组样本数据是否来自某种分布的问题。

此外，还有一些重要的非参数检验，包括判断独立性的列联表检验(卡方分析)，基于顺序统计量的秩检验、符号检验等等。

一. 卡方 检验

1. 卡方检验的想法

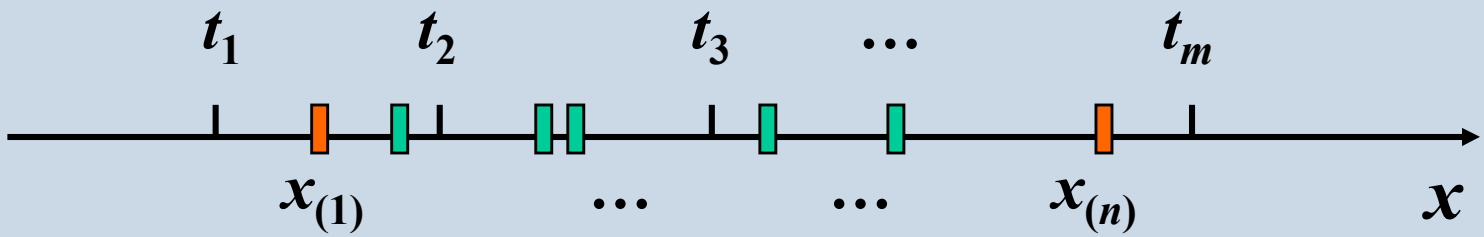
如果一组样本 X_1，…，X_n 来自分布 F，需要检验是如下问题：

$$H_0: F = F_0 \Leftrightarrow H_1: F \neq F_0$$

从理论上来说无论 F 是离散还是连续分布，卡方检验都可以处理；不过它更适用于离散的总体，对于连续的总体 F，采用 **Kolmogrov** 检验更好。

K.Pearson 的拟合优度检验思想

在实数轴上取 m 个点把 $\mathbf{R}^1$ 分成 $m+1$ 个部分，以 v_i 表示落在第 i 个区间里的样本个数，p_i 是总体随机变量 X 在这个区间中的概率：



Remark

F_0 是离散分布时，m 及分点 t_i 的选取很自然；而 F_0 是连续分布时，一般建议取 $m \le n/5$，再取一个稍大的区间 $[a, b] \supset (x_{(1)}, x_{(n)})$，把 $[a, b]$ m 等分得到分点 t_i。

当零假设 $\mathbf{H_0}$：$F = F_0$ 成立时p_i 可以计算出：

$$p_i = F_0(t_i) - F_0(t_{i-1})，\ 1 \le i \le m+1;$$

这里 $F_0(t_0) = 0$，$F_0(t_{m+1}) = 1$

n 充分大时，频率 v_i/n 与概率 p_i 应该相当接近，

因此如果零假设成立则统计量：

$$K^2 = \sum_{i=1}^{m+1} \frac{n}{p_i}\left(\frac{v_i}{n} - p_i\right)^2$$

应该偏小，反之则可以否定零假设 $\mathbf{H_0}$：$F = F_0$。

1900年K.Pearson 证明了极限分布 $K^2 \to \chi^2(m)$，

因此 $\mathbf{H_0}$ 的一个水平α拒绝域近似为 $K^2 > \chi_\alpha^2(m)$ 。

2. 完全已知离散分布的卡方检验

总体 X 只可能取有限个值a_i，$1 \le i \le k$。相应地，样本 X_1，…，X_n中取值为a_i 的个数为v_i，$1 \le i \le k$。需要检验：

$$H_0: P\{X = a_i\} = p_i, \quad 1 \le i \le k$$

取检验统计量：

$$K^2 = \sum_{i=1}^{k} \frac{n}{p_i}\left(\frac{v_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{(v_i - np_i)^2}{np_i} = \frac{1}{n}\sum_{i=1}^{k} \frac{v_i^2}{p_i} - n$$

则H_0 的一个水平α 检验的拒绝域为 $K^2 > \chi_\alpha^2(k-1)$

3. 含未知参数离散分布的卡方检验

如果总体 X 的分布含有 r 个未知参数， Fisher 证明了在 H_0 成立的条件下用极大似然估计估计出这 r 个参数，计算出相应的概率 $\hat{p}_i$，$1 \le i \le k$，则修改自由度后Pearson 的结论仍然成立，即：

取检验统计量：

$$K^2 = \sum_{i=1}^{k} \frac{n}{\hat{p}_i}\left(\frac{v_i}{n} - \hat{p}_i\right)^2$$

则H_0 的一个水平α 的检验的拒绝域为 $K^2 > \chi_\alpha^2(k - r - 1)$

例3.3.1 Mendel 的遗传学例子

Mendel 研究豌豆时发现豌豆有两种特性：圆与皱、黄与绿，他观察了 **556** 颗豌豆：

圆黄	皱黄	圆绿	皱绿	(总数)
315	**101**	**108**	**32**	**(556)**

而根据他的遗传学理论，**Mendel** 认为这些组合关系应该有理论上的概率：

圆黄	皱黄	圆绿	皱绿	(概率)
9/16	**3/16**	**3/16**	**1/16**	**(1)**

解. 总体分布的 $k=4$，对应 K^2 统计量为：

$$K^2=\frac{1}{n}\sum_{i=1}^{k}\frac{v_i^2}{p_i}-n$$

$$=\frac{1}{556}\left\{\frac{315^2}{9/16}+\frac{101^2}{3/16}+\frac{108^2}{3/16}+\frac{32^2}{1/16}\right\}-556=0.47$$

$\chi_{0.05}^2(3)=7.815$，$\chi_{0.90}^2(3)=0.584$，$\chi_{0.95}^2(3)=0.352$

甚至在水平0.90下都可以接受零假设，即认为Mendel 的遗传学理论是正确的。

从p-值的角度拟合优度 $p=\mathrm{P}\{\chi^2(3)>0.47\}$这个值是0.9254，理论分布与实际数据拟合得太好。

□

例3.3.2　**Rutherford** 与**Geiger** 的放射性衰变

观测一个放射源，每**7.5**秒记录一次 α 粒子个数，一共进行了 **2608** 次观测，数据如下：

放射出的粒子数	观测到的次数	放射出的粒子数	观测到的次数
0	57	6	273
1	203	7	139
2	383	8	45
3	525	9	27
4	532	≥10	16
5	408		

是否能认为衰变产生粒子个数服从泊松分布？

解. 以 X 记每 7.5 秒衰变产生的 α 粒子个数，

则如果 X 服从泊松分布，应该有：

$$\mathbf{H_0}:\ \mathrm{P}\{X=i\}=\frac{\lambda^i}{i!}e^{-\lambda},\ i\geq 0$$

为了方便，实际检验的是截尾泊松分布：

$$\mathbf{H_0}:\ p_i=\mathrm{P}\{X=i\}=\frac{\lambda^i}{i!}e^{-\lambda},\ 0\leq i\leq 9$$

$$p_{10}=\mathrm{P}\{X=10\}=1-(p_0+..+p_9)$$

λ 的极大似然估计为

$$=\frac{0\times 57+1\times 203+\ldots+10\times 16}{2608}=3.87$$

从而算出 p_0，…，p_{10}，得到检验统计量 K^2

$$K^2 = \sum_{i=0}^{10} \frac{n}{\hat{p}_i}\left(\frac{v_i}{n} - \hat{p}_i\right)^2$$

$n = 2608$，v_i 是 2608 次观测中粒子个数恰好是 i 个的观测次数：$v_0 = 57$，$v_1 = 203$，…，$v_{10} = 16$。

$$\hat{p}_i = \frac{3.87^i}{i!} e^{-3.87}, \quad 0 \le i \le 9 \qquad \hat{p}_{10} = 1 - \sum_{i=0}^{9} \hat{p}_i$$

检验统计量 $K^2 = 13.064$，自由度是 11 - 1 - 1 = 9；
而分位点 $\chi_{0.05}^2(9) = 16.9$，$\chi_{0.15}^2(9) = 13.288$。

水平 0.05 下不能拒绝零假设，即认为粒子数服从泊松分布(拟合优度为0.1597)。

□

练习3.3.3　抛掷一枚骰子63次，结果如下

点数	1	2	3	4	5	6
次数	10	9	11	8	12	13

这颗骰子能否被认为是均匀的？

练习3.3.4

收集某种公开发行的彩票开奖结果，讨论中奖号码是否具有随机性。

练习3.3.5

用计算机生成 100 个随机数，或者查找一个随机数表取 50 个随机数，讨论所得结果。

二. Kolmogrov 检验

如果一组样本 X_1，…，X_n 来自分布 F，考虑检验问题：

$$H_0: F(x) = F_0(x) \Leftrightarrow H_1: F(x) \neq F_0(x)$$

首先计算经验分布函数：

$$F_n(x) = \begin{cases} 0, & x \le x_{(1)} \\ \dfrac{k}{n}, & x_{(k)} < x \le x_{(k+1)} \\ 1, & x > x_{(n)} \end{cases}$$

F_n 与 F_0 的一致距离定义为：

$$D_n = \sup | F_n(x) - F_0(x) |, \quad \text{对所有} x \in R^1$$

Glivenko-Cantelli 定理

如果 F_n 是分布函数 F_0 的经验分布函数，则 一致距离 D_n 几乎处处收敛到零，即

$$\mathrm{P}\{\lim D_n = 0\} = 1 \text{ 。}$$

Kolmogrov 定理

当 F_0 是一个连续分布时， D_n 的极限分布是

$$\mathrm{P}\{\ n^{1/2} D_n \le x\ \} \to Q(x) \text{ ，对所有} x \in \mathrm{R}^1$$

$$Q(x) = \sum_{k=-\infty}^{k=+\infty} (-1)^k \exp(-2k^2 x^2), \quad x > 0$$

D_n 的计算公式

对于 $1 \le k \le n$，定义 $d_k = \max\{d_{k1}, d_{k2}\}$:

$$d_{k1} = \left| \frac{k}{n} - F_0(x_{(k)}) \right|, \quad d_{k2} = \left| \frac{k-1}{n} - F_0(x_{(k)}) \right| 。$$

则有 $D_n = \max\{d_1, d_2, \ldots, d_n\}$;

并且当F_0 是连续分布时 D_n 的分布与 F_0 无关。

H_0: $F(x) = F_0(x)$ 的拒绝域应该是 $\{D_n \ge C\}$，而一般取拒绝域形式为 $\{n^{1/2} D_n \ge d\}$。当给定检验水平α 时即 $P\{n^{1/2} D_n \ge d\} = \alpha$，

具体的临界值 d 需要查表。

Kolmogrov 检验临界值($n \geq 40$)

$$P\{ n^{1/2} D_n \geq d \} = \alpha$$

α	0.9	0.75	0.50	0.25	0.10	0.05	0.01
d	0.575	0.678	0.830	1.02	1.23	1.36	1.63

在各种分布检验里Kolmogrov 检验效果比较好，灵敏度较高。但同时这种检验也有两点不足：

(1) 只适用于对连续分布的检验；

(2) 不能用来检验分布族(零假设分布不能含参数)。

练习3.3.6

从 1902 年 12月16日 开始到 1977 年 3月4日 一共记录了强烈地震 63 次 (*Richter Scale* ≥7.5 或者*Fatalities* ≥1000) ， 62 个间隔如下(天):

840	736	76	1336	304	129	638	157	584	710	335
375	9	937	145	887	46	1354	567	209	735	44
263	402	454	139	599	38	33	1901	194	36	780
83	365	121	695	759	667	203	832	92	150	294
319	40	436	328	82	280	562	460	556	30	246
220	434	721	40	99	384	1617				

检验相继两次强震的时间间隔是否来自参数为1/430 (即平均间隔为430 天) 的指数分布？

三. 卡方分析(列联表检验)

设 X 的可能值是 1，…，s，Y 的可能值是 1，…，t。现在进行了 n 次独立的观察，其中“X 取 i，Y 取 j”的次数是 n_{ij}，需要检验：

H_0：X、Y 相互独立。

X \ Y	1	…	j	…	t	
1	n_{11}	…	n_{1j}	…	n_{1t}	$n_{1\bullet}$
i	n_{i1}	…	n_{ij}	…	n_{it}	$n_{i\bullet}$
s	n_{s1}	…	n_{sj}	…	n_{st}	$n_{s\bullet}$
	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet t}$	n

记 $p_{ij} = \mathrm{P}(X = i,\ Y = j)$，

$p_i = \mathrm{P}(X = i)$，$q_j = \mathrm{P}(Y = j)$；需要检验的是

H_0：$p_{ij} = p_i \times q_j$ 对一切 i、j 成立

1. p_i、q_j 的极大似然估计

似然函数显然是 $\mathrm{L} = \prod_{i=1}^{s}\prod_{j=1}^{t}(p_i q_j)^{n_{ij}}$

因此对数似然函数为

$$\ln \mathrm{L} = \sum_{i=1}^{s}\sum_{j=1}^{t}(n_{ij}\ln p_i + n_{ij}\ln q_j)$$

下面分别对 p_i、q_j 求导建立似然方程组，

似然方程组

$$\frac{\partial \ln L}{\partial p_i} = 0, \quad \frac{\partial \ln L}{\partial q_j} = 0, \quad 1 \le i \le s-1, \quad 1 \le j \le t-1$$

而 $p_s = 1 - (p_1 + \ldots p_{s-1})$， $q_t = 1 - (q_1 + \ldots q_{t-1})$

得到 p_i、q_j 的极大似然估计为

$$\hat{p}_i = \frac{n_{i\bullet}}{n} \qquad \hat{q}_j = \frac{n_{\bullet j}}{n} \qquad 1 \le i \le s, \quad 1 \le j \le t$$

注意这里只估计了 $(s-1)+(t-1)$ 个未知参数

2. (分类变量) 独立性的检验

Pearson 的卡方统计量：

$$K^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{(n_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}$$

的极限分布是自由度为

$$st - (s-1) - (t-1) - 1 = (s-1)(t-1)$$

因此零假设(X、Y两个变量独立) 的水平 α 检验的拒绝域为 $K^2 > \chi_\alpha^2((s-1)(t-1))$。

2.1 列联表检验统计量的计算公式

$$K^2 = n\left[\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 1\right]$$

2.2 四格表检验统计量的计算公式

$$K^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}$$

四格表的零假设(X、Y两个变量独立) 的水平α检验的拒绝域为 $K^2 > \chi_\alpha^2(1)$。

$\chi_{0.01}^2(1) = 6.63$，　$\chi_{0.05}^2(1) = 3.84$

例3.3.7　P.G.Norton 与E.V.Dunn 1985年研究关于“打鼾”与“心脏病” 的关系，

	从来不打鼾	每晚都打鼾
有心脏病	24	30
没有心脏病	1355	224

解. 计算有关的K^2统计量，

$$K^2 = \frac{1633 \times (24 \times 224 - 30 \times 1355)^2}{54 \times 1579 \times 1379 \times 254} = 68$$

认为睡觉打鼾与心脏病有显著的统计关系。 □

例3.3.8　1969年美国征兵法案是否公平？

一年内的每一天被指定对应于一个1~366 的号码，按照抽取出的顺序作为服兵役的顺序。

显然被抽到一个小号码的概率不应该依赖于一个人的出生月份。但是最终公布的结果是

	1~6月	7~12月
1~183	73	110
183~366	109	74

是否应该认为下半年出生更可能先服兵役？

解. 计算有关的K^2统计量，

$$K^2 = \frac{366 \times (73 \times 74 - 110 \times 109)^2}{183 \times 183 \times 182 \times 184} = 14.16$$

p-值是 $p = \mathrm{P}(\chi^2(1) > 14.16) = 0.00017$

可以认为一个人的出生日期和他在1969年美国国防部公布的被征召入伍的顺序具有统计上的显著关系。

□

EXCEL 处理卡方分析

***CHITEST*(观察值域,理论值域)直接输出*p*-值**

其中观察值域$(s \times t)$为：

$$\begin{matrix} \cdots & & \\ \cdots & n_{ij} & \cdots \\ \cdots & & \\ \cdots & & \end{matrix}$$

其中理论值域$(s \times t)$为：

$$\begin{matrix} \cdots & & \\ \cdots & C_{ij} & \cdots \\ \cdots & & \end{matrix}$$

$$C_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

练习3.3.9

下表是美国人口普查局1992年高等教育年鉴的一组样本数据。分析这两组变量是否有关。

	亚裔	西班牙裔	白人
中学或更低	24	98	419
上过大学	27	34	310
专业人员或硕士	9	6	61

四. 秩检验

对于一组来自连续总体 F 的样本 X_1，…，X_n，相应顺序统计量为 $X_{(1)} \le \ldots \le X_{(n)}$，每个样本$X_i$的秩记为 R_i (即它在顺序统计量中的排序)，

$$R_i = \min\{k:\ 1 \le k \le n,\ X_i = X_{(k)}\}$$

不难证明随机向量 $R = (R_1, \ldots, R_n)$ 在集合 $A = \{a:\ a$ 是 1，2，…，n 的一个排列 $\}$ 上均匀分布，即对于任意一个固定的排列 a，都有：

$$\mathrm{P}\{R = a\} = \frac{1}{n!}$$

因此依赖于 R 的统计量(秩统计量) $T(R)$ 具有与X 的总体 F 无关、固定不变的分布

F. Wilcoxon 秩和检验

1945年 Wilcoxon 提出用来比较两个总体分布函数的非参数检验方法。

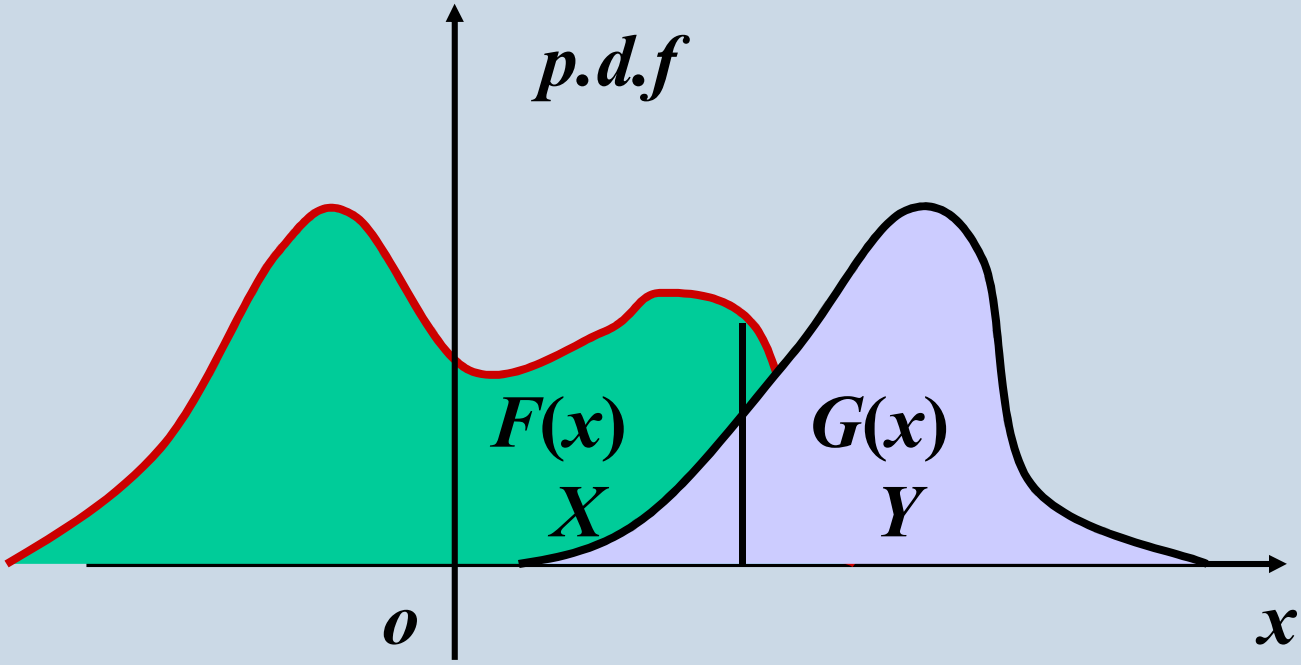
假定 $F(x)$ 与 $G(x)$ 是两个独立的连续总体，相应样本分别是 X_1，…，X_m 与 Y_1，…，Y_n。

把混合后的样本(X_1，…，X_m，Y_1，…，Y_n)排序，记 R_1，…，R_n 为 Y_1，…，Y_n 的秩。

Wilcoxon 提出，利用秩和：

$$W = R_1 + \ldots + R_n$$

作为检验统计量来比较这两个分布的大小。



由于 W 是 Y 样本(即$G(x)$) 在全体样本中的秩和，如果 $F(x) > G(x)$ 则 W 应该偏大；反之W应该偏小

1. 两个总体分布函数的秩和检验方法

H_0	H_1	拒绝域
$F(x) \le G(x)$	$F(x) > G(x)$	$W \ge d$
$F(x) \ge G(x)$	$F(x) < G(x)$	$W \le c$
$F(x) = G(x)$	$F(x) \ne G(x)$	$W \le c \cup W \ge d$

秩和统计量 $W = R_1 + \ldots + R_n$ 的计算公式

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} I[Y_j > X_i] + \frac{1}{2}n(n+1)$$

2. 非参数的均值检验

Wilcoxon 秩和检验也可以如下方式提出：假定 μ_1 与 μ_2 分别是两个非参数总体 $F(x)$ 与$G(x)$ 的均值，则

H_0	H_1	拒绝域
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$W \geq d$
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$W \leq c$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$W \leq c \cup W \geq d$

例3.3.10　一种产品可以用 A、B两种材料制造，随机抽取了12个产品按照性能从低到高排序：

B，B，A，B，B，A，A，B，A，A，A，A

解. 如果检验这两种材料产品性能是否有差异，

$$H_0: \mu_1 = \mu_2 \Leftrightarrow H_1: \mu_1 \neq \mu_2$$

0.05水平下的拒绝域为：$W \leq 21$ 或$W \geq 44$；

如果需要检验 A 材料的产品性能是否更好，

$$H_0: \mu_1 \geq \mu_2 \Leftrightarrow H_1: \mu_1 < \mu_2$$

0.05水平下的拒绝域为：$W \geq 43$ 。

而 B 材料产品的秩和为 $W = 1+2+4+5+8 = 20$。

□

五. 符号检验

使用简单，如果在重复观察时总体没有变化，则n 个样本中应该是正、负号基本上各占一半；

可以用二项分布计算出正号的 p-值：

$$p\text{-值} = BINOMDIST(k, n, p, 1)$$

在处理配对的数据时，符号检验的显著优点是相对比较安全

例3.3.11

Darwin 收集的异花与自花授粉植物高度

数据	1	2	3	4	5	6	7	8
异花	23.5	12.0	21.0	22.0	19.1	21.5	22.1	20.4
自花	17.4	20.4	20.0	20.0	18.4	18.6	18.6	15.3

数据	9	10	11	12	13	14	15
异花	18.3	21.6	23.3	21.0	22.1	23.0	12.0
自花	16.5	18.0	16.3	18.0	12.8	15.5	18.0

三种不同的统计分析方法

① 假定这些数据配对，需要检验它们的差值是否来自均值**0**的正态，***p*-值是0.0478**；

② 使用符号检验，***p*-值是0.0037**；

③ 假定这些数据来自不同总体，需要检验的是它们的均值是否相同，***p*-值是0.0203**。

都认为自花与异花授粉植株高度显著不同

□

习题 3.3

1-4. 教材 **176** 页

第 **14**、**16**、**18** 、**21** 题。

第3.4节 统计决策与Bayes理论

统计决策(又称统计判决)理论，通过引进各种优良性标准，把统计问题转化为数学最优化问题的求解。

Bayes 统计从 **1763** 年英国学者 **T.Bayes** 的论文《论有关机遇问题的求解》发展而来，实际上是一种归纳推理的理论。最核心(也是争议最大)的原则是其主观概率的思想，即先验分布的引入。

东北大学数学系

一. 统计决策(*Statistical decision*)的思想

1950年A. Wald 提出统计学应该是在不确定情况下作出决策的一门科学，包括参数估计、假设检验等都可以归结到“统计决策”的理论中。

1. 统计决策理论的三个要素

假定总体分布函数是$F(x, \theta)$，则未知参数θ全部可能值构成参数空间Θ；如果把我们在实际工作中可能采取的各种行为看作一个集合D，它称为是决策空间D；定义在$\Theta \times D$上的非负函数就称为损失函数$L(\theta, d)$，表示参数是θ时采取行为d而引起的损失。

2. 常见的两种损失函数

平方损失： $L(\theta, d) = (\theta - d)^2$,

绝对损失： $L(\theta, d) = |\theta - d|$

3. 风险函数 (*Risk fuction*)

由于我们采取的行为 d 一般都建立在一组样本的基础上即 $d = \delta(X_1, \ldots, X_n)$。有了样本以后，样本的函数 δ 就是一个决策，它的平均损失

$$R(\theta, \delta) = E\,[L(\theta, \delta(X_1, \ldots, X_n))]$$

就称为是决策 δ 的风险函数。

统计决策观点下的估计与检验

(1). 在参数点估计问题中常常取平方损失函数：

$$L(\theta, d) = (\theta - d)^2$$

此时风险函数就是一个估计量的均方误差

UMVUE 就是在无偏估计类中，
平方损失下风险最小

练习**3.4.1**

假定X_1，..，X_n来自参数p的两点分布，分别在平方损失与绝对损失下求p风险最小的点估计。

(2). 在假设检验问题中，取决策空间 $D = \{0, 1\}$，其中 k 表示接受假设 H_k，损失函数为 0-1 损失：

$$L(\theta, k) = \begin{cases} 0, & \text{如果 } \theta \in \Theta_k \\ 1, & \text{如果 } \theta \in \Theta_{1-k} \end{cases}$$

则一个检验的风险函数就是：

$$R(\theta, \phi) = \begin{cases} \beta_\phi(\theta), & \text{当 } \theta \in \Theta_0 \\ 1 - \beta_\phi(\theta), & \text{当 } \theta \in \Theta_1 \end{cases}$$

Neyman-Pearson 的假设检验理论，即指定一个 $0 < \alpha < 1$，把风险函数 $R(\theta, \phi)$ 在 Θ_0 上的取值控制为不超过 α，同时使得 $R(\theta, \phi)$ 在 Θ_1 上尽量小。

(3). 在区间估计理论中，取决策空间D为全部区间(a, b)的集合，损失函数一共有两个：

$$L_1(\theta, (a, b)) \text{ 与 } L_2(\theta, (a, b))$$

L_1 是 0-1 损失函数：

$$L_1(\theta, (a, b)) = \begin{cases} 0, & \text{当 } a < \theta < b \\ 1, & \text{否则} \end{cases}$$

L_2 反映了区间的精度，

例如 $L_2(\theta, (a, b)) = b - a$ 。

对于任意一个区间估计 $\delta(x) = (\varphi_1(x), \varphi_2(x))$，损失函数$L_1$ 导致风险函数：

$$R_1(\theta, \delta) = 1 - P\{\varphi_1(x) < \theta < \varphi_2(x)\}$$；

而损失 L_2 的风险函数是区间平均长度：

$$R_2(\theta, \delta) = E[\varphi_2(x) - \varphi_1(x)]$$

指定置信度为$1 - \alpha$ 就理解为 $R_1(\theta, \delta) \leq \alpha$，然后在这个条件下要求 $R_2(\theta, \delta)$ 尽可能地小。

4. 两个常用的决策原则

(1). Minimax 决策 (极小化极大方法)

对每一个决策都考虑最坏的可能状态，然后选择那个对于全部最坏状态而言是最佳的决策。

(2). Bayes 决策

对于参数θ定义一个概率分布(称为先验分布)，把风险函数 $R(\theta, d)$关于这个先验分布取数学期望(称为 **Bayes** 风险)，在决策空间D 中取使得**Bayes** 风险最小的那一个作为问题的解。

例3.4.2　汽车决策问题

一位统计学家准备买一辆汽车，计划使用两年行驶 **40,000** 英里。他把选择的范围缩小到两部：汽车*A* 价格 ＄**5,000**，平均每加仑汽油行驶 **20** 英里，汽车*B* 价格 ＄**6,700**，平均每加仑汽油行驶 **40** 英里；　假定汽油平均价格有三种，每加仑＄**1**，＄**2**，＄**3**，问这位统计学家应该选择哪一辆车？

解. 如果选择汽车 *A* 需要 **2,000** 加仑汽油；

如果选择汽车 *B* 需要 **1,000** 加仑汽油；

油价	$1	$2	$3
汽车*A* 总费用	$7,000	$9,000	$11,000
汽车*B* 总费用	$7,700	$8,700	$9,700

如果采用 **Minimax** 决策，*A* 车的最大总费用 $11,000，*B*车的最大总费用 $9,700；他应该选择汽车*B*。

如果采用 **Bayes** 决策，需要给出一组概率：

假定油价(主观概率)：	$1	$2	$3
	3/4	1/8	1/8

则*A* 车的平均总费用 $7,750，*B*车的平均总费用 $8,075；他应该选择汽车*A*。 □

练习3.4.3　收藏家问题

一位收藏家考虑购买一幅据说是著名画家的画，标价 **$5,000** 。如果是真品可值 **$10,000** ；如果赝品则一文不值。并且买了假画或者是没有买下真画都会损害她的名誉。损失表如下：

	真品	赝品
买	**-$5,000**	**$6,000**
不买	**$3,000**	**0**

她跑去征询一位鉴赏家的意见，而这位鉴赏家能够以 **0.8** 的概率判断一幅真画，同时以 **0.70** 的概率识别一幅假画。

现在有四种方案可供考虑：

决策**1**：一定要买；

决策**2**：坚决不买；

决策**3**：抛硬币决定是否买；

决策**4**：如果鉴赏家说是真品就买。

(1). 这位收藏家的 **Minimax** 决策是什么？

(2). 假定这幅画有**0.75**的先验概率是真画，**0.25** 的概率为假画，她的 **Bayes** 决策又应该是什么？

二． Bayes 统计理论

频率学派的观点

对于总体的未知参数θ，无论做点估计、区间估计或者假设检验，θ始终仅仅是一个常数。

我们抽取样本之前，对于 θ无任何了解，所有关于θ的信息都全部包含在样本之中。

Bayes 学派则认为，

在抽取样本之前，对θ已经有一些认识(先验知识)，这种知识可通过一个概率分布体现出来(先验分布$h(\theta)$)，它综合了我们在抽样之前对θ的全部信息。

1. Bayes 统计的基本思想

在随机向量(X, θ)中，我们只能通过对X的观察去间接地推导出θ。概率函数$f(x, \theta)$被看成样本x关于θ的条件分布$f(x|\theta)$，

因此抽取一组样本x，也就意味着对θ得到了新的信息，它将修正原来的观点(θ的先验分布)，这种对θ的新认识通过θ关于样本x的条件分布(后验分布)体现出来；

对θ的估计或者检验都依据后验分布来进行

Prior → Posterior

总体分布律或密度
$p(x, \theta)$

根据样本 ⇩

概率函数 $f(x|\theta)$
$= \prod p(x_i, \theta)$

先验分布 $h(\theta)$ ⇩

X 与 θ 的联合分布
$= h(\theta) \times \prod p(x_i, \theta)$

⇨

样本 X 的边缘分布
$= \int_{\Theta} h(\theta) \times \prod p(x_i, \theta)\, d\theta$

θ 关于样本 X 的后验分布律或者后验密度函数：

$$h(\theta|x) = \frac{h(\theta) \times \prod p(x_i, \theta)}{\int_{\Theta} h(\theta) \times \prod p(x_i, \theta)\, d\theta}$$

⇧

对 θ 积分或求和

2. Bayes 统计推断理论

得到θ关于样本X的后验分布$h(\theta|X)$以后，其余的如样本观察值、样本分布、先验分布等都不再有用，对θ的处理全部依赖于后验分布。

(1) **Bayes 点估计**：根据损失函数取成后验分布的中位数或者数学期望；

(2) **Bayes 区间估计**：直接由后验分布构造；

(3) **Bayes 检验**：由后验分布计算θ落在零假设或者对立假设中的概率，哪个大就接受哪个。

3. 如何确定先验分布

(1) 过去的经验或资料

(2) 同等无知原则 **(Bayes 假定)**

(3) 共轭先验等

Remark

采用同等无知原则时，先验分布可以不再是一个分布律或密度函数。如 $h(\theta) \equiv 1$，$\theta \in \mathbf{R}^1$，只要能保证样本 X 的边缘分布 $\int_{\Theta} h(\theta) \times \prod p(x_i, \theta)\, d\theta$ 有限从而 $h(\theta|X)$ 存在即可，此时称为广义先验分布。

4. Kernel 技巧

没有必要对 $h(\theta)\times f(x|\theta)$ 进行θ积分或求和求出样本的边缘分布…，而是直接看出后验分布的分布类型(与精确分布只相差一个常数)

Remark：

在平方损失函数下，“**Bayes** 估计”与“无偏估计”的交集是空集

例3.4.4　假定随机事件 A 发生的概率是 p，做了 n 次独立重复试验，根据试验结果估计 p 。

解. 总体是参数 p 的两点分布，n 个样本显然是：

$$X_i = \begin{cases} 0\text{，第 } i \text{ 次试验时} A \text{不发生；} \\ 1\text{，第 } i \text{ 次试验时} A \text{ 发生。} \end{cases}$$

考虑样本的概率函数：

$$f(x, p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

根据“同等无知”观点，取先验分布 $p \sim U(0, 1)$

因此 X 与 θ 的联合分布，即联合密度为：

$$h(\theta)\times f(x|\theta)=p^{\sum x_i}(1-p)^{n-\sum x_i},\ 0<p<1$$

所以样本 x 的边缘分布为一个 **Beta** 积分：

$$f(x)=\int_0^1 p^{\sum x_i}(1-p)^{n-\sum x_i}\,dp=\mathrm{B}\left(\sum x_i+1,n-\sum x_i+1\right)$$

$$=\frac{\Gamma(\sum x_i+1)\Gamma(n-\sum x_i+1)}{\Gamma(n+2)}$$

则参数 θ 关于样本 x 的后验密度为：

$$h(\theta \mid x) = \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{\mathrm{B}\left(\sum x_i + 1, n - \sum x_i + 1\right)}, \quad 0 < p < 1$$

如果取损失为平方损失函数，则参数 p 的 **Bayes** 估计应该是后验分布的均值，

$$\hat{p} = \int_0^1 p \times \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{\mathrm{B}\left(\sum x_i + 1, n - \sum x_i + 1\right)} dp$$

$$= \frac{\mathrm{B}\left(\sum x_i + 2, n - \sum x_i + 1\right)}{\mathrm{B}\left(\sum x_i + 1, n - \sum x_i + 1\right)} = \frac{\sum x_i + 1}{n + 2}$$

更简单的做法是：

使用kernel技巧，因为

$$h(\theta)\times f(x|\theta) = p^{\sum x_i}(1-p)^{n-\sum x_i}\ , 0<p<1$$

所以 Posterior ~ $Beta(\sum x_i + 1, n - \sum x_i + 1)$

而Beta(a,b)的数学期望为 $a/(a+b)$

因此在平方损失下，Bayes 估计显然是

$$\hat{p} = \frac{\sum x_i + 1}{n+2}$$

n 较小时，Bayes 估计要比样本均值更合理

□

例3.4.5

假定总体 $X \sim B(1, p)$，只有三个样本 X_1, X_2, X_3，考虑：$H_0: p=0.4 \Leftrightarrow H_1: p=0.5$。记决策 d_k 为“接受假设 H_k”，损失函数是：$L(d_0,0.4) = L(d_1,0.5) = 0$; $L(d_0,0.5) = 1, L(d_1,0.4) = 2$;

再假定参数 p 的先验分布是：

p	0.4	0.5
	1/3	2/3

求这个假设的Bayes 检验。

Bayes 检验为：

样本和等于0 或者1 时接受零假设 $p = 0.4$，

当样本和等于2 或者3 时接受对立假设 $p = 0.5$。

□

习题 3.4

1-3. 教材 **100** 页

第 **25**、**28**、**36** 题。

第4章 方差分析

方差分析 (***Analysis of Variance，ANOVA***) :

研究一个(或多个)分类自变量如何影响一个数值因变量的统计分析方法。

方差分析针对方差相同的多个正态总体，检验它们的均值是否相同。 即，

同时判断多组数据均值之间差异是否显著

东北大学数学系

方差分析的目的

①. 判断某些因素对于我们感兴趣的因变量是否具有“显著”的影响，

②. 如果因素间有交互效应，寻找最佳搭配方案。

方差分析的特点

①　方差分析与一般的假设检验不同

要比较均值是否相同，可以使用第三章假设检验的方法，但是只能处理两个均值。

方差分析处理的是多个均值的情况。

② 方差分析与回归、相关分析不同

回归与相关处理的是两个数值变量的问题，相应的散点在 x 轴上具有顺序(从小到大)，而方差分析的数据在 x 轴上可以任意交换位置。

常见的方差分析主要有：

单因素方差分析，双因素方差分析，

多因素方差分析。

一. 方差分析的数学模型

响应变量(因变量)：

进行随机试验所考察的数值指标；

因素或因子(自变量)：

影响因变量的各不同分类原因；

水平：

各个因素所构成的组或者类型。

Fisher的农业试验

考察小麦产量(y) 对于品种和施肥量的关系。

选择了：两个不同的小麦品种，
三个不同的施肥等级；
一共 $2\times3=6$ 种搭配做试验，建立模型。

$$\begin{cases} y_{11} = \theta_0 + \alpha_1 + \beta_1 + \varepsilon_{11} \\ y_{12} = \theta_0 + \alpha_1 + \beta_2 + \varepsilon_{12} \\ y_{13} = \theta_0 + \alpha_1 + \beta_3 + \varepsilon_{13} \\ y_{21} = \theta_0 + \alpha_2 + \beta_1 + \varepsilon_{21} \\ y_{22} = \theta_0 + \alpha_2 + \beta_2 + \varepsilon_{22} \\ y_{23} = \theta_0 + \alpha_2 + \beta_3 + \varepsilon_{23} \end{cases}$$

y_{ij} 是小麦产量，
α_1、α_2 是品种效应，
β_1、β_2、β_3 是施肥等级的效应，
θ_0 是其它因素的平均效应。

ε_{ij} 是随机误差，$\mathbf{i.i.d} \sim N(0, \sigma^2)$

品种是否对产量有影响 $\Leftrightarrow$ $\mathbf{H_{01}}$：$\alpha_1 = \alpha_2$

施肥量是否对产量有影响 $\Leftrightarrow$ $\mathbf{H_{02}}$：$\beta_1 = \beta_2 = \beta_3$

把这个模型写成矩阵的形式：$\mathbf{Y}=\mathbf{X}\beta+\varepsilon$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_0 \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{pmatrix}$$

在方差分析中，同一个因素的不同水平看成是模型里的不同变量，而不能看成是同一个自变量在不同试验里的取值。(否则需要 y 对 x 有线性相依关系)

二. 单因素方差分析

1. 数据的结构

自变量水平	试验指标观察值				组内平均
1	y_{11}	y_{12}	…	y_{1n1}	$\bar{y}_1$
2	y_{21}	y_{22}	…	y_{2n2}	$\bar{y}_2$
	…	…		…	
r	y_{r1}	y_{r2}	…	$y_{r\,nr}$	$\bar{y}_r$

影响 y 的只有一个因素，它有 r 个水平(组)，在第 i 个水平下针对 y 做了 n_i 次试验或观察，得到因变量的观察数据为 y_{i1}，…， $y_{i\,ni}$。

可以假定：

$$y_{ij} = \beta_i + \varepsilon_{ij} \quad ，\ 1 \le j \le n_i \ 、\ 1 \le i \le r$$

这里 ε_{ij} 对所有 i、j 都独立同分布于 $N(0, \sigma^2)$

单因素方差分析的主要任务：

1. 检验假设：H_0： $\beta_1 = \beta_2 = \ldots = \beta_r$；
2. 作出未知参数 β_1，…， β_r 以及 σ^2 的估计

2. 因子效应与误差方差的估计

按照模型的假定，因变量的观察值来自 r 个不同的正态总体：

$y_{11}, \cdots, y_{1n1}$ 来自总体 $N(\beta_1, \sigma^2)$；

$y_{21}, \cdots, y_{2n2}$ 来自总体 $N(\beta_2, \sigma^2)$；

…

$y_{r1}, \cdots, y_{rnr}$ 来自总体 $N(\beta_r, \sigma^2)$。

未知参数 $\beta_1, \cdots, \beta_r$ 的估计就采用各个总体的样本均值。

定理 4.1 方差分析中未知参数估计及分布

1. 因素各水平效应的估计采用各个组内平均，

$$\hat{\beta}_i = \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

相应的分布显然是：$\hat{\beta}_i \sim N(\beta_i, \frac{\sigma^2}{n_i})$ $1 \le i \le r$

2. 误差方差 σ^2 的估计利用残差平方和，

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-r} = \frac{1}{n-r}\sum_{i=1}^{r}\sum_{j=1}^{n_s}(y_{ij} - \bar{y}_i)^2$$

相应的分布是：$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-r)$

3. $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_r, \hat{\sigma}^2$ 之间相互独立。

3. 方差分析平方和分解公式

(1) 总平方和 $\mathbf{TSS} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$

它是观察到的每个数据与总平均的差异总和，表示因变量总的变化。

TSS衡量了全部 y_{ij} 的差异，它越大则说明y_{ij}之间的差异越大。产生TSS的原因只有两个：

(1). 因子不同的水平，即 β_1，…，β_r 的差异；
(2). 随机误差。

(2) 自变量平方和 CSS $= \sum_{i=1}^{r} n_i (\bar{y}_i - \bar{y})^2$

它是因为自变量不同的类型而产生的差异，表示自变量在因变量的变化中所占的份额。

∑ (每组平均 - 总平均)² 来刻化

(3) 残差平方和 RSS $= \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$

表示由其它原因引起的因变量变化，利用

∑ (观察值- 每组平均)² 来刻化

方差分析平方和分解

$$TSS = CSS + RSS$$

利用**Cochren**定理可以证明，一定条件下自变量平方和 **CSS** 与残差平方和 **RSS** 相互独立。

4. 单因素方差分析的检验

如果零假设H_0：$\beta_1 = \beta_2 = \ldots = \beta_r$ 成立，则有

$$\frac{CSS}{\sigma^2} \sim \chi^2(r-1)$$

由定理4.1.1 构造检验统计量

$$F\text{ 比}= \frac{n-r}{r-1}\frac{CSS}{RSS} \sim F(r-1, n-r)$$

因此这些分类自变量中每组均值都相同的一个水平α的拒绝域为：$\{F \geq F_\alpha(r-1, n-r)\}$

检验的p-值是 $P\{F(r-1, n-r) > F\text{ 比}\}$

单因素方差分析表

方差来源	平方和	自由度	均方	*F*-比	*p*-值
分类变量	**CSS**	***r*-1**	**CMS**		
残差变量	**RSS**	***n*-*r***	**RMS**		
总计	**TSS**	***n*-1**			

其中

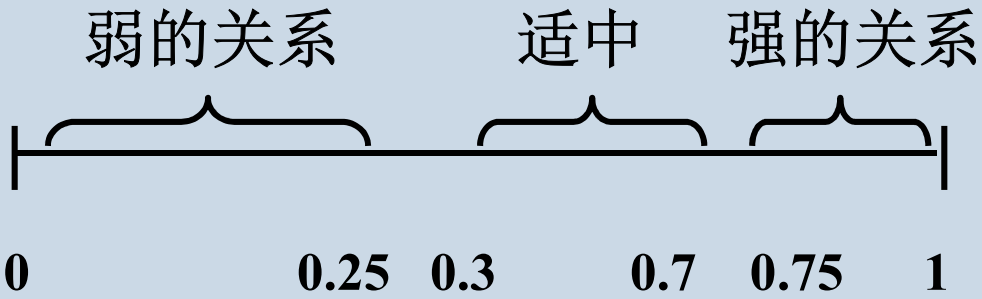
$$\mathbf{CMS} = \frac{\mathbf{CSS}}{r\text{-}1},\ \mathbf{RMS} = \frac{\mathbf{RSS}}{n\text{-}r},\ F\text{-比} = \frac{\mathbf{CMS}}{\mathbf{RMS}}$$

5. 变量关系的强度

$$R^2 = \frac{\text{自变量平方和}}{\text{总平方和}} = \frac{\text{CSS}}{\text{TSS}}$$

R^2 反映了在因变量全部的变化中，分类自变量产生影响所占的比例；

因此用 R (取正值)来衡量分类自变量与数值因变量的关系强度：



EXCEL 函数处理单因素方差分析

直接调用函数 ***DEVSQ***

计算出总平方和以及残差平方和。

$$DEVSQ(x_1,\cdots,x_n)=\sum_{k=1}^{n}(x_k-\bar{x})^2$$

(1) 没有必要计算总均值以及各组均值；

(2) 先计算各组偏差平方和 **RS1, …, RSr**；

再全部相加得到残差平方和 **RSS**。

例4.1　灯丝配料方案的优选

灯丝	使用寿命(小时)
甲	1600，1610，1650，1680，1700，1720，1800；
乙	1580，1640，1640，1700，1750；
丙	1460，1550，1600，1640，1660，1740， 1820， 1820；
丁	1510，1520，1530，1570，1600，1680；

解. $n = 26$，$r = 4$ 。

方差来源	平方和	自由度	均方	F-比	p-值
分类变量	47399.17	3	15799.7	1.9327	0.1538
残差变量	179850.8	22	8175.04		
总计	227250	25			

水平 0.05 下认为灯丝配料对寿命没有显著影响。

思考1

这里的R^2 =0.209 应该如何理解？

□

例**4.2**

下表是 ***FBI*** 给出的**1986~1992**年美国 **48** 个大陆洲暴力犯罪率**(** 次**/10** 万人**)**，按地理位置分成 **7** 个区。

新英格兰	147 140 149 557 336 426
中大西洋	986 572 359
中西部	423 308 800 804 258 285
	235 263 578 51 125 369
南方	427 833 306 164 476 675 588
	1036 334 540 558 274 395 758
西南	436 659 658 726
落基山区	293 524 157 222 267 719
太平洋岸	437 550 920

作方差分析判断犯罪率是否与地区有关?

解.

***p*-值是 0.07224**，因此在水平**0.05**下不能拒绝零假设，即应该认为在统计上没有显著的证据表明犯罪率和地区有关(各地犯罪率没有显著差异)。

R^2 **=0.236** 说明在影响犯罪率的所有因素中，地理位置占了将近四分之一的比例。

□

ANOVA (*r*=2) ⇔ 假设*1.4a*

$$\mathrm{CMS} = n_1(\bar{y}_1 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n})^2 + n_2(\bar{y}_2 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n})^2$$

$$= \frac{n_1 n_2}{n}(\bar{y}_1 - \bar{y}_2)^2 = (\bar{y}_1 - \bar{y}_2)^2 \times \frac{1}{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\mathrm{RMS} = \frac{1}{n-2} \times \{ \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_1}(y_{2j} - \bar{y}_2)^2 \}$$

$$= \frac{1}{n_1 + n_2 - 2} \times \{ (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 \}$$

故F 比$= \dfrac{\mathrm{CMS}}{\mathrm{RMS}} \sim F(1, n-2) = t^2(n_1 + n_2 - 2)$

三. 双因素方差分析

假定影响 y 的有两个因素A、B，各有r、s 个水平：$A_1, ..., A_r$, $B_1, ..., B_s$, 对这些水平所有的搭配 $A_i B_j$ 同时都做 l 次($l > 1$)试验得到模型：

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} ,$$

$$1 \le i \le r, 1 \le j \le s, 1 \le k \le l, \varepsilon_{ijk} \text{ i.i.d} \sim N(0, \sigma^2)$$

这里 μ 是响应变量 y 的总平均；

α_i :A 的主效应，A 在第i 个水平单独对 y 的效果；

β_j :B 的主效应，B 在第j 个水平单独对 y 的效果；

γ_{ij} :交互效应，因素A 在 i 、因素B 在 j 水平上联合对 y 的效果。

一般为了便于分析，还做如下假定：

$$\sum_{i=1}^{r}\alpha_i=\sum_{j=1}^{s}\beta_j=0=\sum_{i=1}^{r}\gamma_{ij}=\sum_{j=1}^{s}\gamma_{ij}$$

双因素方差分析需要讨论：

1. 因子的主效应是否显著；即检验：
   H_{01}：$\alpha_1=\alpha_2=\ldots=\alpha_r$，以及$H_{02}$：$\beta_1=\beta_2=\ldots=\beta_s$
2. 交互效应是否显著：H_{03}：$\gamma_{11}=\gamma_{12}=\ldots=\gamma_{rs}$

与单因素方差分析不同的是，如果拒绝了 H_{03}，还应该寻找最佳搭配。

处理思路类似单因素方差分析，把总平方和分解成若干平方和：有两个因素主效应产生的，有交互效应产生的，还有随机误差产生的，最后构造恰当的 F 统计量来检验 H_{01}、H_{02}、H_{03}。

引进如下符号：

总平均：

$$\bar{y} = \frac{1}{rsl}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{l} y_{ijk}$$

误差平均：

$$\bar{y}_{ij\bullet} = \frac{1}{l}\sum_{k=1}^{l} y_{ijk}$$

*A*因素平均：

$$\bar{y}_{i\bullet\bullet} = \frac{1}{s}\sum_{j=1}^{s} \bar{y}_{ij\bullet}$$

*B*因素平均：

$$\bar{y}_{\bullet j\bullet} = \frac{1}{r}\sum_{i=1}^{r} \bar{y}_{ij\bullet}$$

构造如下相应的五个平方和：

总平方和 $$\mathbf{TSS}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{l}(y_{ijk}-\bar{y})^2$$

*A*因子主效应平方和 $$\mathbf{SSA}=sl\sum_{i=1}^{r}(\bar{y}_{i\bullet\bullet}-\bar{y})^2$$

*B*因子主效应平方和 $$\mathbf{SSB}=rl\sum_{j=1}^{s}(\bar{y}_{\bullet j\bullet}-\bar{y})^2$$

交互效应平方和 $$\mathbf{SSAB}=l\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet\bullet}-\bar{y}_{\bullet j\bullet}+\bar{y})^2$$

随机误差平方和 $$\mathbf{RSS}=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{l}(y_{ijk}-\bar{y}_{ij\bullet})^2$$

仍然可以利用Cochren 定理证明，

$$TSS = SSA + SSB + SSAB + RSS$$

而且等式右边的四个平方和相互独立

(1) $\frac{RSS}{\sigma^2} \sim \chi^2(rs(l-1))$；

(2) 当H_{01} 成立时，$\frac{SSA}{\sigma^2} \sim \chi^2(r-1)$；

(3) 当H_{02} 成立时，$\frac{SSB}{\sigma^2} \sim \chi^2(s-1)$；

(4) 当H_{03} 成立时，$\frac{SSAB}{\sigma^2} \sim \chi^2((r-1)(s-1))$。

因此可以构造出三个零假设的 *F* 检验，

1. 对于零假设 H_{01}：$\alpha_1=\alpha_2=\ldots=\alpha_r$，构造

检验统计量 $F_A=\dfrac{rs(l-1)}{r-1}\dfrac{SSA}{RSS}\sim F(r\text{-}1,\ rs(l\text{-}1))$

双因素方差分析中A因素的主效应不显著的一个检验水平α的拒绝域为：

$$\{F_A\geq F_\alpha(r-1,\ rs(l-1))\}$$

2. 对于零假设 H_{02}：$\beta_1=\beta_2=\ldots=\beta_s$，构造

检验统计量 $F_B=\dfrac{rs(l-1)}{s-1}\dfrac{SSB}{RSS}\sim F(s\text{-}1,\ rs(l\text{-}1))$

双因素方差分析中B因素的主效应不显著的一个检验水平α的拒绝域为：

$$\{F_B\geq F_\alpha(s-1,\ rs(l-1))\}$$

3. 对于零假设 H_{03}： $\gamma_{11} = \gamma_{12} = \ldots = \gamma_{rs}$，
构造检验统计量 F_{AB}：

$$F_{AB} = \frac{rs(l-1)}{(r-1)(s-1)} \frac{\text{SSAB}}{\text{RSS}} \sim F((r\text{-}1)(s\text{-}1),\ rs(l\text{-}1))$$

双因素方差分析中A、B两个因素的交互效应不显著的一个检验水平α拒绝域为：

$$\{F_{AB} \geq F_{\alpha}((r\text{-}1)(s\text{-}1),\ rs(l\text{-}1))\}$$

双因素方差分析表

方差来源	平方和	自由度	平均平方和	F 值
因素A	SSA	$r-1$	$\frac{\text{SSA}}{r-1}$	$\frac{rs(l-1)}{r-1}\frac{\text{SSA}}{\text{RSS}}$
因素B	SSB	$s-1$	$\frac{\text{SSB}}{s-1}$	$\frac{rs(l-1)}{s-1}\frac{\text{SSB}}{\text{RSS}}$
$A\times B$	SSAB	$(r-1)(s-1)$	$\frac{\text{SSAB}}{(r-1)(s-1)}$	$\frac{rs(l-1)}{(r-1)(s-1)}\frac{\text{SSAB}}{\text{RSS}}$
残差	RSS	$rs(l-1)$	$\frac{\text{RSS}}{rs(l-1)}$	
总和	TSS	$rsl-1$		

EXCEL 函数处理双因素方差分析

① TSS = *DEVSQ*(全部试验数据)；

② *A* 因子主效应 SSA 的计算：

把 *A* 因子每个水平所在组的数据相加，得到 *r* 个和，计算DEVSQ(*A*的每组数据之和)，

$$SSA = \frac{DEVSQ(A\text{的每组数据之和})}{s\ l}$$

③ ***B*** 因子主效应 **SSB** 的计算：

把 ***B*** 因子每个水平所在组的数据相加，得到 ***s*** 个和，计算**DEVSQ(*B*的每组数据之和)**，

$$\mathbf{SSB} = \frac{DEVSQ(B\text{的每组数据之和})}{r\,l}$$

④ 残差平方和 **RSS** 的计算：

计算每一对搭配 $i \times j$ 的 l 个重复试验数据的偏差平方和，再对所有这些 $r\,s$ 个 ***DEVSQ*** 求和得到 **RSS**。

例4.3　橡胶配方中考虑3 种促进剂(A) 、4种氧化锌份量(B)各组合两次进行试验，测得24 组300％的定伸强力数据，对这组数据作双因素方差分析。

	B_1	B_2	B_3	B_4
A_1	31,33	34,36	35,36	39,38
A_2	33,34	36,37	37,39	38,41
A_3	35,37	37,38	39,40	42,44

解. 根据模型的定义，$r=3$，$s=4$，$l=2$

方差来源	平方和	自由度	均方	F-值
A(促进剂)	56.59	2	28.29	$F_A = 19.4$
B(氧化锌)	132.125	3	44.04	$F_B = 30.2$
交互 $A\times B$	4.75	6	0.7917	$F_{AB} = 0.5429$
误差	17.5	12	1.4583	
总和	210.9583	23		

相应的 F 分布上0.05 分位点是：

$F_{0.05}(3,12) = 3.49$, $F_{0.05}(2,12) = 3.89$, $F_{0.05}(6,12) = 3.00$

所以在**0.05**的显著水平下，$\mathbf{H_{01}}$、 $\mathbf{H_{02}}$被否定，即不同的促进剂或不同的氧化锌分量对橡胶定伸强力具有显著的影响；

但是接受$\mathbf{H_{03}}$，即交互作用不显著的。

从***p*-**值的角度，

$\mathbf{H_{01}}$ 的 ***p*-**值是 **0.000174** ；

$\mathbf{H_{02}}$ 的 ***p*-**值是 **0.000007** ；

$\mathbf{H_{03}}$ 的 ***p*-**值有 **0.766517**。 □

例4.4

对一种火箭使用4种燃料，3种推进器进行射程试验，每种燃料与推进器各组合两次，一共试验了24次。

(单位：海里)

燃料	推进器 B_1	推进器 B_2	推进器 B_3
A_1	58.2	56.2	65.3
	52.6	41.2	60.8
A_2	49.1	54.1	51.6
	42.8	50.5	48.4
A_3	60.1	70.9	39.2
	58.3	73.2	40.7
A_4	75.8	58.2	48.7
	71.5	51.0	41.4

对这组数据作双因素方差分析。

解. 根据模型的定义，$r=4$，$s=3$，$l=2$

方差来源	平方和	自由度	平均平方和	F 值
A(燃料)	261.6750	3	87.2250	$F_A = 4.42$
B(推进器)	370.9808	2	185.4904	$F_B = 9.39$
交互 $A\times B$	1768.6925	6	294.7821	$F_{AB} = 14.9$
误差	236.95	12	19.7458	
总和	2638.2983	23		

相应的 F 分布上0.05 分位点是：

$F_{0.05}(3,12) = 3.49$, $F_{0.05}(2,12) = 3.89$, $F_{0.05}(6,12) = 3.00$

所以在**0.05**的显著水平下，这三个零假设 $\mathbf{H_{01}}$、 $\mathbf{H_{02}}$、 $\mathbf{H_{03}}$ 都被否定。即，不同的燃料或不同的推进器对射程有统计上的显著影响；而且交互作用是高度显著的($A_4 \times B_1$ 或 $A_3 \times B_2$)

从*p*-值的角度，

$\mathbf{H_{01}}$ 的 *p*-值是 **0.025969** ；

$\mathbf{H_{02}}$ 的 *p*-值是 **0.003506** ；

$\mathbf{H_{03}}$ 的 *p*-值只有 **0.000062**。 □

习题 4

1-2. 教材 **200** 页

第 **1**、**2** 题。

第5章 线性回归模型

第5.1节 线性模型理论

第5.2节 一元回归与相关分析

第5.3节 多元回归分析

东北大学数学系

第5.1节 线性模型理论

一. 线性模型的定义

定义5.1.1 y 是可观察的随机变量，x_1，…，x_m 是可观察的分类或数值变量，β_0，…，β_k 是未知参数，ε 是不可观察随机误差($\varepsilon \sim N(0, \sigma^2)$)。

$$y = \beta_0 + \sum_{i=1}^{k} f_i(x_1, \cdots, x_m)\beta_i + \varepsilon$$

称为是线性模型。

Remark

① 线性模型中“线性”是针对未知参数 β 而言，许多表面上的非线性模型本质也是线性的:

$$y=\alpha e^{\beta x}\times\varepsilon,\quad y=\alpha x^{\beta}\times\varepsilon,\quad \ln\varepsilon\sim N(0,\sigma^2);$$

而有些模型是实质上的非线性模型:

$$y=\alpha e^{\beta x}+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2);$$

以及 ***Logistic*** 模型:

$$y=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}+\varepsilon$$

② 一些统计学家喜欢把线性模型表示成：

$$E\,y = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k$$

含义是：线性模型就是一个随机变量的数学期望具有未知参数线性结构的统计模型。

在这种意义下 x_1，…，x_k 很自然就被称为“自变量”，y 也就被称作“因变量”。

自变量与因变量关系是一种统计上的关系，即因变量的均值是自变量的函数，而决不能认为因变量是自变量的函数。

③ 把模型 $E\,y = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k$ 改写成：

$$y = \beta_0 + x_1\beta_1 + \ldots + x_k\beta_k + \varepsilon,\quad E\,\varepsilon = 0$$

为方便处理，进一步假定 $\varepsilon \sim N(0, \sigma^2)$ 。

要处理这 $k+1$ 个未知参数 β_0，…，β_k，需要至少做 n 次独立试验 $(n > k+1)$；这些试验都在不同自变量取值下进行，其它条件都保持不变，最后写成矩阵的表达式，就是：

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,\quad E\,\varepsilon = 0$$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon\,,\quad E\,\varepsilon = 0$$

这里$\mathbf{Y} = (y_1,\ \ldots,\ y_n)^{\mathrm{T}}$表示可观察的因变量;

$\beta = (\beta_0,\ \ldots,\ \beta_k)^{\mathrm{T}}$ 表示待估计或检验的未知参数;

$\varepsilon = (\varepsilon_1,\ \ldots,\ \varepsilon_n)^{\mathrm{T}}$ 是随机误差,

一般假定 $\varepsilon_i \sim N(0, \sigma^2)$。

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \cdot & \cdot & \ldots & \cdot \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}$$

表示自变量,

$n > k+1$

二. 线性模型参数的估计

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

1. 未知参数 β 的估计

采用最小二乘的标准，即寻找 $\hat{\beta}$，使得：

$$\|\mathbf{Y}-\mathbf{X}\hat{\beta}\|^2 = \inf \|\mathbf{Y}-\mathbf{X}\beta\|^2 \text{ 对所有 } \beta \in \mathbf{R}^{k+1}$$

这样得到的 β 的估计称为是最小二乘估计(***LSE***)

LSE 的求解思路：平方和分解

$$\|Y-X\beta\|^2 = \|Y-X\hat{\beta} + X(\hat{\beta}-\beta)\|^2$$

$$= \|Y-X\hat{\beta}\|^2 + \|X(\hat{\beta}-\beta)\|^2 + 2(\hat{\beta}-\beta)^T X^T(Y-X\hat{\beta})$$

因此要使得对一切 $\beta \in R^{k+1}$ 都有

$\|Y-X\beta\|^2 \geq \|Y-X\hat{\beta}\|^2$，充分必要条件是：

$(\hat{\beta}-\beta)^T X^T(Y-X\hat{\beta}) = 0$ 对一切 $\beta \in R^{k+1}$ 都成立

由于 β 是 R^{k+1} 中任意一个向量，

所以 $X^T(Y-X\hat{\beta})$ 必须是一个 $k+1$ 维零向量，即：

$$(X^TX)\hat{\beta} = X^TY \quad \text{(正规方程)}$$

如果**X** 是满秩矩阵即 $\mathbf{rk(X)} = k+1$ 时，正规方程的解：

$$\hat{\beta} = (X^TX)^{-1}X^TY \equiv S^{-1}X^TY$$

就称为是线性模型 $Y = X\beta + \varepsilon$ 中参数向量 β 的最小二乘估计 (*LSE*)

$X\hat{\beta}$ 称为是经验回归函数，

$Y = X\hat{\beta}$ 是经验回归方程。

2. 误差方差σ^2的估计

把线性模型 $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$ 改写成如下形式：

$$y_i=\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}+\varepsilon_i,\quad 1\le i\le n$$

定义“残差” (*Residual*)

$$e_i=y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}-\ldots-\hat{\beta}_k x_{ik},\quad 1\le i\le n$$

作为随机误差ε_i的“估计”，则残差平方和：

$$Q_e=e_1^2+e_2^2+\ldots+e_n^2$$

$$=\|\mathbf{Y}-\mathbf{X}\hat{\beta}\|^2=\mathbf{Y}^T(\mathbf{I}_n-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^T)\mathbf{Y}$$

可以作为 σ^2 的估计(注意需要修正！)

定理5.1.1　线性模型的最小二乘估计

(1) 对于模型 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$，　β 的*LSE* 是

$$\hat{\beta} = \mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

(2) σ^2的*LSE*是

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}$$

思考

既然n 个观察数据 y_1，…，y_n 的方差都是σ^2，为什么不使用这组数据的样本方差, 而是要用残差平方和修正以后去估计σ^2？

最小二乘估计的矩阵代数含义

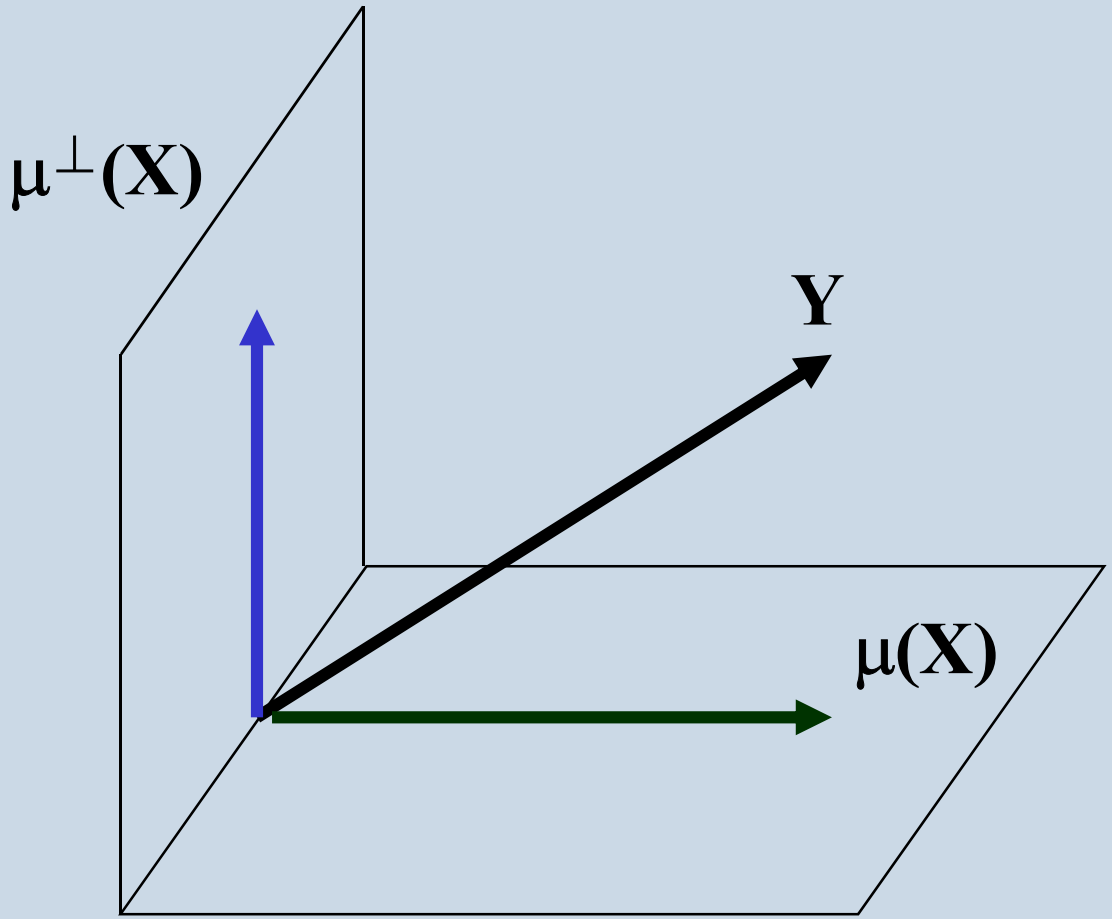
考虑矩阵 $\mathbf{X}$ 的$k+1$ 个n 维列向量生成的$\mathbf{R}^n$ 中的线性子空间 $\mu(\mathbf{X})$，不难证明 $\mu(\mathbf{X}) = \mu(\mathbf{X}\mathbf{X}^{\mathrm{T}})$。

由于$\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}$ 是一个对称、幂等的n 阶方阵，即它是一个正投影阵，恰好是$\mu(\mathbf{X})$的投影矩阵；

而 $\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}$ 是 $\mu(\mathbf{X})$ 的正交子空间 $\mu^{\perp}(\mathbf{X})$ 的投影矩阵，因此

$\mathbf{X}\hat{\beta}$ 是 $\mathbf{Y}$ 到子空间 $\mu(\mathbf{X})$ 中的投影，

σ^2的LSE只和 $\mathbf{Y}$ 在 $\mu^{\perp}(\mathbf{X})$ 的投影向量有关。


$\mu^{\perp}(X)$
Y
$\mu(X)$

3. 最小二乘估计的无偏性质

引理5.1.2　随机向量的期望与方差公式

(1) 如果Y 是*n* 维随机向量，A 是*n* 阶对称矩阵

则 $E(\mathbf{Y}^{\mathrm{T}}\mathbf{AY}) = (E\mathbf{Y})^{\mathrm{T}}\mathbf{A}(E\mathbf{Y}) + \mathrm{tr}\{\mathbf{A}[\mathrm{Var}(\mathbf{Y})]\}$；

(2) 如果Y 是*n* 维随机向量，B 是$m \times n$ 阶矩阵

则 $\mathrm{Var}(\mathbf{BY}) = \mathbf{B}[\mathrm{Var}(\mathbf{Y})]\mathbf{B}^{\mathrm{T}}$

Remark

Var (Y)是Y 的协方差矩阵 $(Cov(y_i, y_j))_{n\times n}$；

迹 (*Trace*)具有如下性质：

$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$，$\mathrm{tr}(\mathbf{A}-\mathbf{B}) = \mathrm{tr}(\mathbf{A}) - \mathrm{tr}(\mathbf{B})$

注意到线性模型的形式 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$，因此

$$E\mathbf{Y} = \mathbf{X}\beta,\quad \mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}_n$$

说明 β 的*LSE* $\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ 是无偏估计。

根据引理**5.1.2**，残差平方和的数学期望是：

$$E(Q_e) = E[\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}]$$

$$= \beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X}\beta + \mathrm{tr}[(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\sigma^2\mathbf{I}_n]$$

$$= 0 + \sigma^2\,\mathrm{tr}(\mathbf{I}_n - \mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}) = \sigma^2[n - \mathrm{tr}(\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})]$$

$$= \sigma^2[n - \mathrm{tr}(\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X})] = (n - k - 1)\sigma^2 \quad 。$$

四. 估计量的分布

对于线性模型 $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$，$\beta\in \mathbf{R}^{k+1}$，

$\mathbf{X}$ 是$n\times(k+1)$ 满秩矩阵，$\varepsilon\sim N(0,\sigma^2\mathbf{I}_n)$

定理**5.1.3**

(1) β 的最小二乘估计服从 $k+1$ 维正态分布，

$$\hat{\beta}=\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\sim N(\beta,\sigma^2\mathbf{S}^{-1});$$

(2) σ^2的估计量服从卡方分布，即

$$\frac{n-k-1}{\sigma^2}\hat{\sigma}^2=\frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{X}\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}\sim\chi^2(n-k-1);$$

(3) $\hat{\beta}$ 与$\hat{\sigma}^2$相互独立。

证明.

首先，根据多维正态分布的一个性质：

> 如果$\mathbf{Y}$ 服从 n 维正态分布 $\sim N(\mu, \sum)$，$\mathbf{A}$ 是任意一个 $m \times n$ 矩阵，则$\mathbf{AY} \sim N(\mathbf{A}\mu, \mathbf{A}\sum\mathbf{A}^{\mathrm{T}})$

因为 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$， $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$，得到 $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n)$，现在已知 $\hat{\beta} = \mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$，而 $\mathbf{S}$ 是一个对称矩阵，因此显然有β 的***LSE***

$$\hat{\beta} \sim N(\mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta,\ \mathbf{S}^{-1}\mathbf{X}^{\mathrm{T}}(\sigma^2\mathbf{I}_n)\mathbf{X}\mathbf{S}^{-1})$$

$$\sim N(\beta, \sigma^2\mathbf{S}^{-1});$$

其次，注意到表达式：

$$\hat{\beta} = S^{-1}X^{T}Y = \beta + S^{-1}X^{T}\varepsilon ,$$

$$\hat{\sigma}^{2} = \frac{1}{n-k-1} Y^{T}(I_{n} - XS^{-1}X^{T})Y$$

$$= \frac{1}{n-k-1} \varepsilon^{T}(I_{n} - XS^{-1}X^{T}) \varepsilon$$

所以，

$$[\frac{X(\hat{\beta}-\beta)}{\sigma}]^{T}[\frac{X(\hat{\beta}-\beta)}{\sigma}] = (\frac{\varepsilon}{\sigma})^{T} XS^{-1}X^{T} (\frac{\varepsilon}{\sigma})$$，以及

$$(\frac{n-k-1}{\sigma^{2}})\hat{\sigma}^{2} = (\frac{\varepsilon}{\sigma})^{T}(I_{n} - XS^{-1}X^{T}) (\frac{\varepsilon}{\sigma})$$

都是 n 个标准正态的二次型，根据**Cochren**定理定理**5.1.3** 成立。

习题 5.1

对于线性模型：

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \text{其中} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

定义如下平方和：

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\mathrm{RegSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

证明必然成立分解关系 $\mathrm{TSS} = \mathrm{RegSS} + \mathrm{RSS}$

第5.2节 一元回归与相关分析

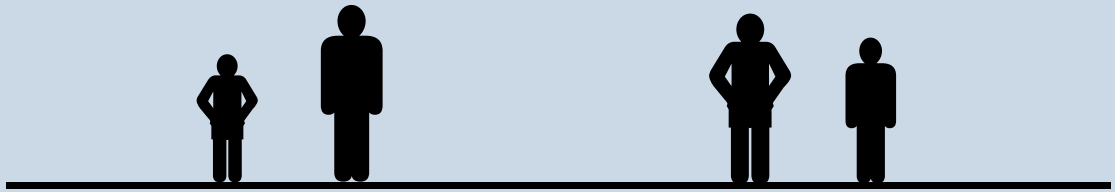
回归与相关分析是用于讨论数值变量之间关系的统计分析方法。

回归分析研究一个(或多个)自变量的变化如何影响因变量，

相关分析研究这两个数值变量的相关程度。

东北大学数学系

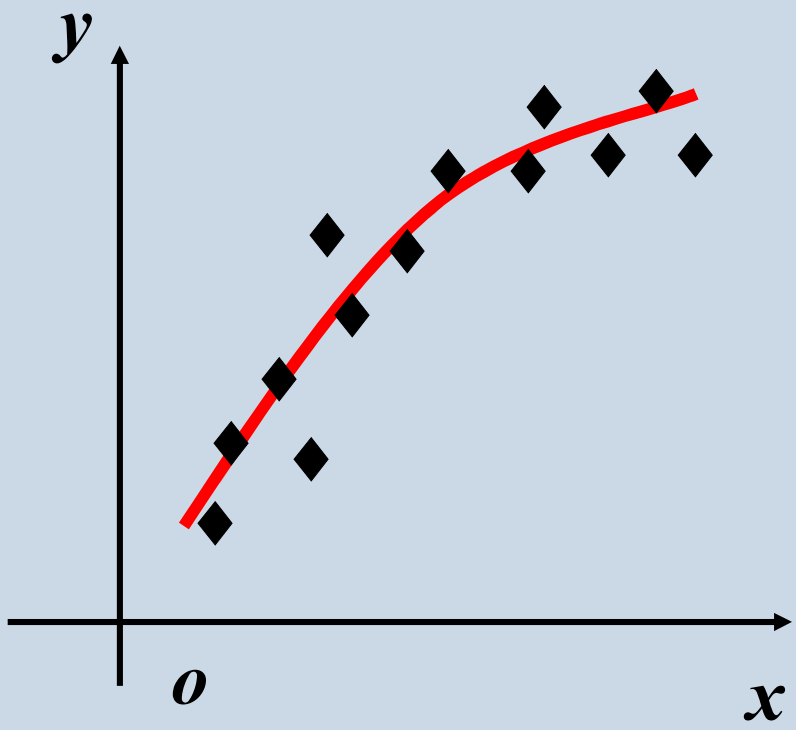
Regression



$y = 33.73 + 0.516\,x$　(单位：英寸)

“回归”的含义

直观上在一个总体中有两个特征(X, Y)，观察了n次得到平面上的n个点(x_1, y_1)，…，(x_n, y_n)。



如果一条曲线$y=f(x)$基本上通过这些点，或者这些点的大多数与这条曲线偏离很小，则称曲线是对观察值的拟合曲线，或者称为是y对于x的回归曲线。

如果 $f(x)$ 就是 x 的线性函数，即：

$$f(x)=\beta_0+x_1\beta_1+\ldots+x_k\beta_k,$$

线性回归模型就定义成：

$$y_i=\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}+\varepsilon_i,\quad 1\le i\le n$$

ε_i 独立同分布于 $N(0,\sigma^2)$

这时不再把 x 看成是随机变量 X 的观察值，而看成是一般的数量变量，因此线性回归模型也是一种线性模型：$\mathbf{Y}=\mathbf{X}\beta+\varepsilon$，$E\,\varepsilon=0$

$y=\beta_0+x_1\beta_1+\ldots+x_k\beta_k$ 就称为是回归方程

一. 一元线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i，\ 1 \le i \le n$$

未知参数 β_0、β_1、σ^2 的估计以及估计量的性质根据定理**5.1.1**、定理**5.1.3** 决定。

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{L_{xy}}{L_{xx}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}(L_{yy} - \hat{\beta}_1 L_{xy})$$

EXCEL 计算回归方程

自变量 X	x_1	x_2	…	x_n
因变量 Y	y_1	y_2	…	y_n

建立回归方程：　y = 截距 + 斜率 × x

截距 = ***INTERCEPT***(因变量数据, 自变量数据)

斜率 = ***SLOPE*** (因变量数据, 自变量数据)

$\hat{\sigma}$ = ***STEYX*** (因变量数据, 自变量数据)

或者直接使用函数***LINEST*** 得到回归方程

例**5.2.1**　食物中脂肪(克) 与所含热量(卡)的关系

随机选取**16**种食品，以脂肪含量作自变量 x ，因变量 y 是热量。

x	4	6	6	8	19	11	12	12	26
y	110	120	120	164	430	192	175	236	429
x	21	11	16	14	9	9	5		
y	318	249	281	160	147	210	120		

根据这组样本数据讨论热量与脂肪含量的关系。

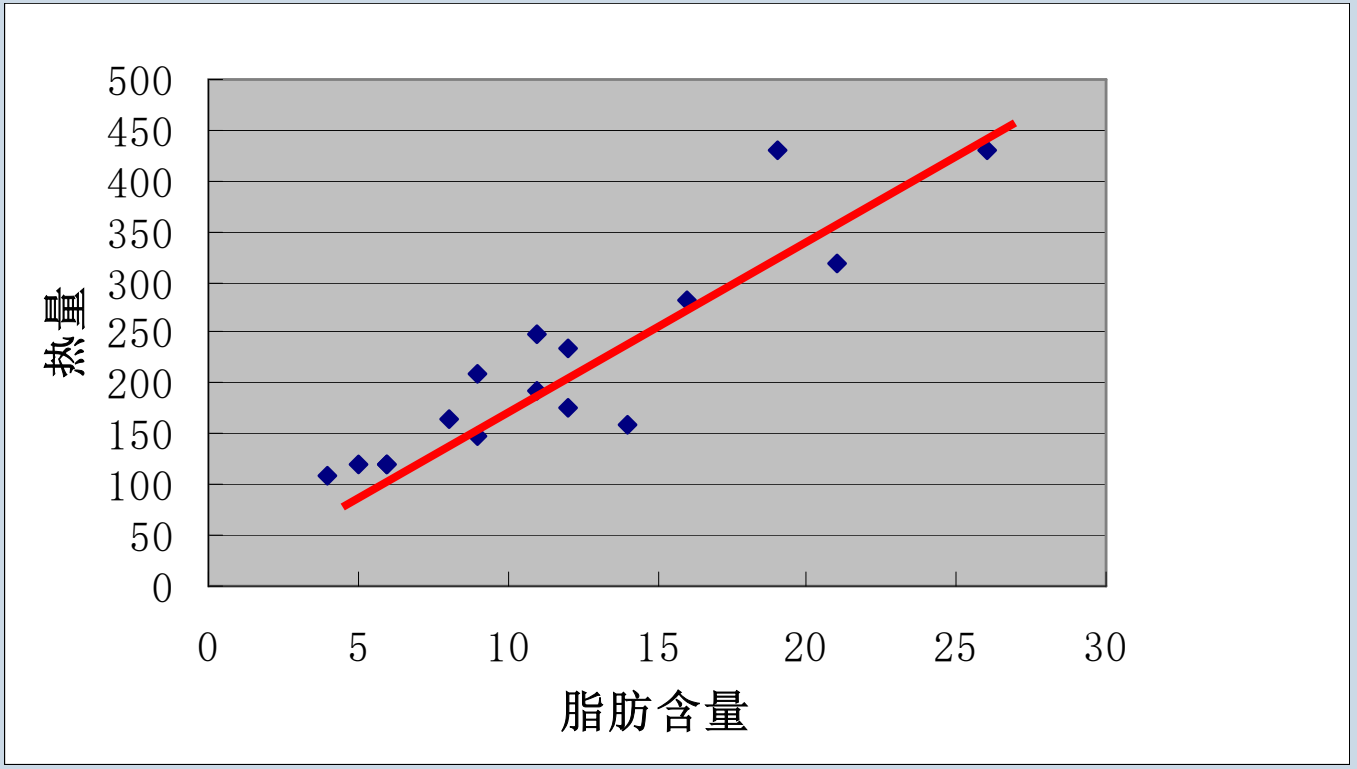
两个数值变量是否具有线性关系？

(1). 散点图的作用

对两个变量做统计分析之前首先要看散点图

如果具有线性关系，那么散点图中的点相应地应该落在某一条直线的附近。

如果散点图没有表示出线性关系，就不能直接做回归分析。


500
450
400
350
300
250
200
150
100
50
0
热量
0
5
10
15
20
25
30
脂肪含量

(2). 正相关与负相关

两个数值变量具有正相关关系，是指因变量将随着自变量的增加而增加，因此对应的直线从左下角到右上角(斜率为正)。

同理负相关是指因变量将随着自变量的增加而减小，对应直线从左上角到右下角(斜率为负)。

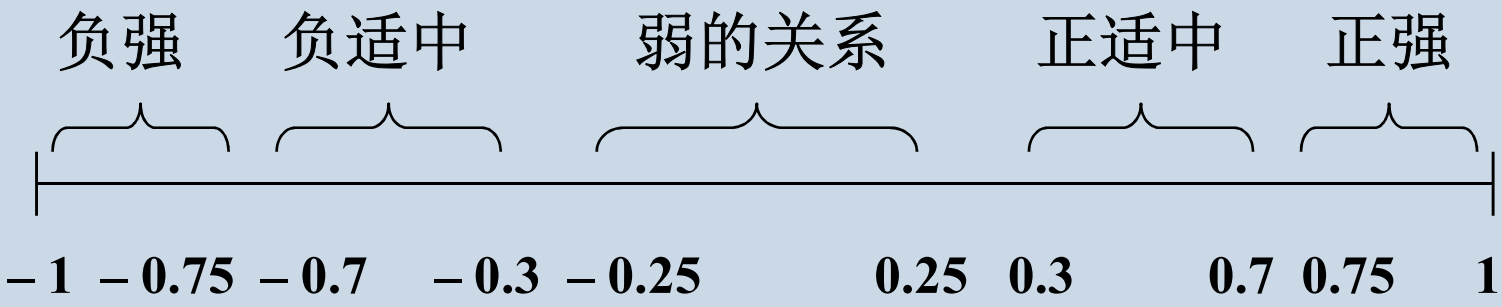
例题数据显示热量与脂肪含量具有正相关关系

两个变量的关系如何量化？

(1). 关系的强度

通过计算相关系数 *r* 来讨论

r 是介于 - 1 到 1 之间的小数，一般认为



EXCEL 计算相关系数

***r* = *CORREL* (自变量数据, 因变量数据)**

解析表达式：

$$CORREL = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(2). 异常点的处理

如果有 **5%** 的异常点，计算出的相关系数将和去掉异常点后得出的结果有显著差异。

因此根据公式计算相关系数时，应该注意散点图中散点的分布。

(3). 相关系数 *r* 的确切含义

在产生因变量变化的所有因素中，
自变量占据了其中$r^2 \times 100\%$ 的份额

脂肪与热量的相关系数 $r = 0.91$， $r^2 = 0.83$ □

例**5.2.2**　成年女性身高与腿长的关系

腿长关于身高的回归方程为：

腿长 ＝ **-16.073** ＋**0.7194** 身高；

反之，身高关于腿长的回归方程为：

身高 ＝ **31.7713** ＋**1.2903** 腿长。

□

二. 简单的相关分析

回归分析的平方和分解

对于 $y = \beta_0 + \beta_1 x + \varepsilon$ 甚至更一般的回归模型，根据练习**5.1.1**显然都有：

TSS = RegSS + RSS

TSS：<u>总(变差)平方和</u>，因变量 y 在其均值附近总的变化；

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

RegSS：回归平方和，
自变量 x 所引起的
因变量 y 的变化；

$$\mathrm{RegSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

RSS：残差平方和，
随机误差所引起的
因变量 y 的变化；

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

因为 **RegSS / TSS** 刻画了回归平方和在总平方和所占的比重，它越大也就说明自变量x对于因变量 y 的影响越大，即回归关系越显著。

所以相关系数 r ($r^2 = \mathbf{RegSS / TSS}$)
刻画了 y 与 x 线性关系程度的大小

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_j - \bar{y})^2}}$$

很容易证明这里的 r^2 就是 $L_{x}{}^{2}{}_{y} / L_{xx} L_{yy}$

三. 回归方程的检验与区间估计

1. 回归系数的假设检验

根据样本数据建立了两个数值变量间的回归方程后，首先应该检验这个方程是否成立，

即利用假设检验讨论是否有 $\mathbf{H_0}$： $\beta_1 = 0$ ？

这个已经建立的(线性)回归方程的好坏，取决于相应的假设检验的 ***p*-值**。

根据定理**5.1.3**，估计量的分布为

(1) $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n}+\frac{\bar{x}^2}{L_{xx}}))$

(2) $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{L_{xx}})$

(3) $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 不独立，协方差为

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{L_{xx}}$$

(4) $\hat{\sigma}^2$ 与 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都独立，并且

$$\frac{n-2}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(n-2)$$

要检验回归关系是否显著，可以利用t 分布：

$$\frac{\hat{\beta}_1}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sim t(n-2)$$

更多的是采用：$\frac{\hat{\beta}_1^2}{\hat{\sigma}^2}L_{xx}\sim F(1,n-2)$

即，$\frac{(n-2)L_{xy}^2}{L_{xx}L_{yy}-L_{xy}^2}\sim F(1,n-2)$

这个检验统计量恰好就是 $(n\text{-}2)r^2/(1\text{-}r^2)$，也就是 $(n\text{-}2)$ RegSS / RSS。

利用F统计量 $F=\dfrac{(n-2)r^2}{(1-r^2)}$

$H_0: \beta_1=0$ 的否定域是 $\{F > F_{0.05}(1, n\text{-}2)\}$，如果零假设被否定，即认为回归方程成立。

Remark

当零假设没有被拒绝，意味着这两个数值变量之间不存在前面建立的线性回归关系，但是它们之间可能存在着其它类型的关系

因此讨论两个数值变量的回归关系时，比较恰当的做法是计算有关的F检验的p-值，它越小说明所建立的回归方程越好。

2. 回归系数的区间估计

从 $\hat{\beta}_1 - \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} t_{\alpha/2}(n-2)$

到 $\hat{\beta}_1 + \dfrac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} t_{\alpha/2}(n-2)$

原因是 $\dfrac{\hat{\beta}_1-\beta_1}{\hat{\sigma}}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \sim t(n-2)$

例**5.2.3**　哺乳动物出生以后开始玩耍的时间 p 与开始行走的时间 w 似乎有关：

类别	走动w(天)	玩耍 p(天)
人类	**360**	**90**
大猩猩	**165**	**105**
猫	**21**	**21**
家犬	**23**	**26**
挪威鼠	**11**	**14**
乌鸫	**18**	**28**
混血猕猴	**18**	**21**
黑猩猩	**150**	**105**
松鼠猴	**45**	**68**
花鼠	**45**	**75**
白脸猴	**18**	**46**

解. 假定 p 对于 w 有线性回归关系，

回归直线为：$\hat{p} = 35.81 + 0.235\,w$ ，

对应检验统计量 $F = 9.50635$ ，而p-值只有

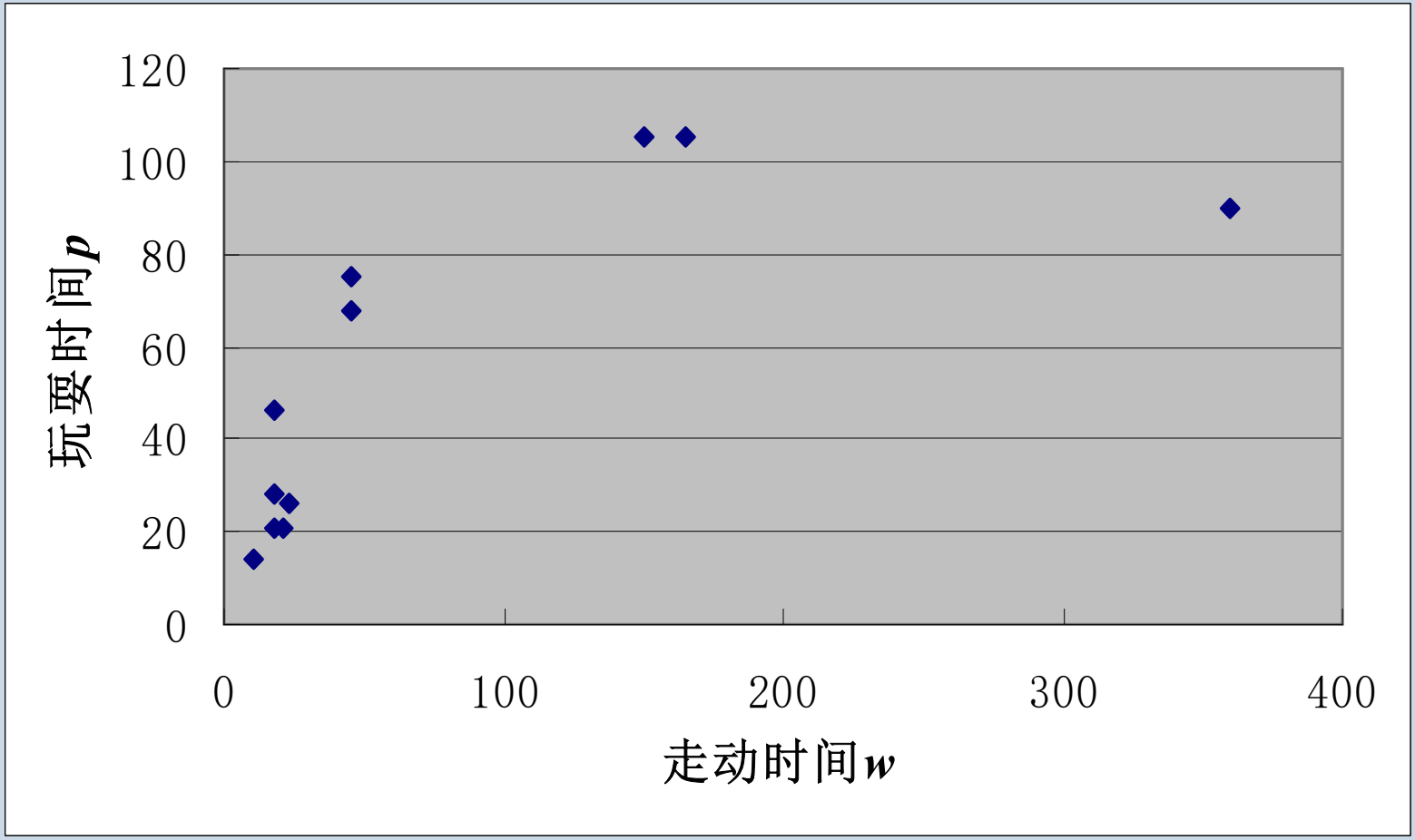
$$P(F(1,9) > 9.50635) = 0.01307,$$

即水平0.01下回归关系不显著，而在水平0.05下才有显著的回归关系。

比较恰当的解决方法：

画出散点图，数据似乎落在幂函数曲线：

$$p = a\,w^{b} \quad (a > 0,\ b > 0)$$ 的附近。


120
100
80
60
40
20
0
玩耍时间p
0
100
200
300
400
走动时间w

作变换：$y_i = \ln p_i$，　$x_i = \ln w_i$ ；

对数据 (x_i, y_i) 作回归拟合得到回归方程：

$$\hat{y} = 1.689 + 0.561\, x \text{ 。}$$

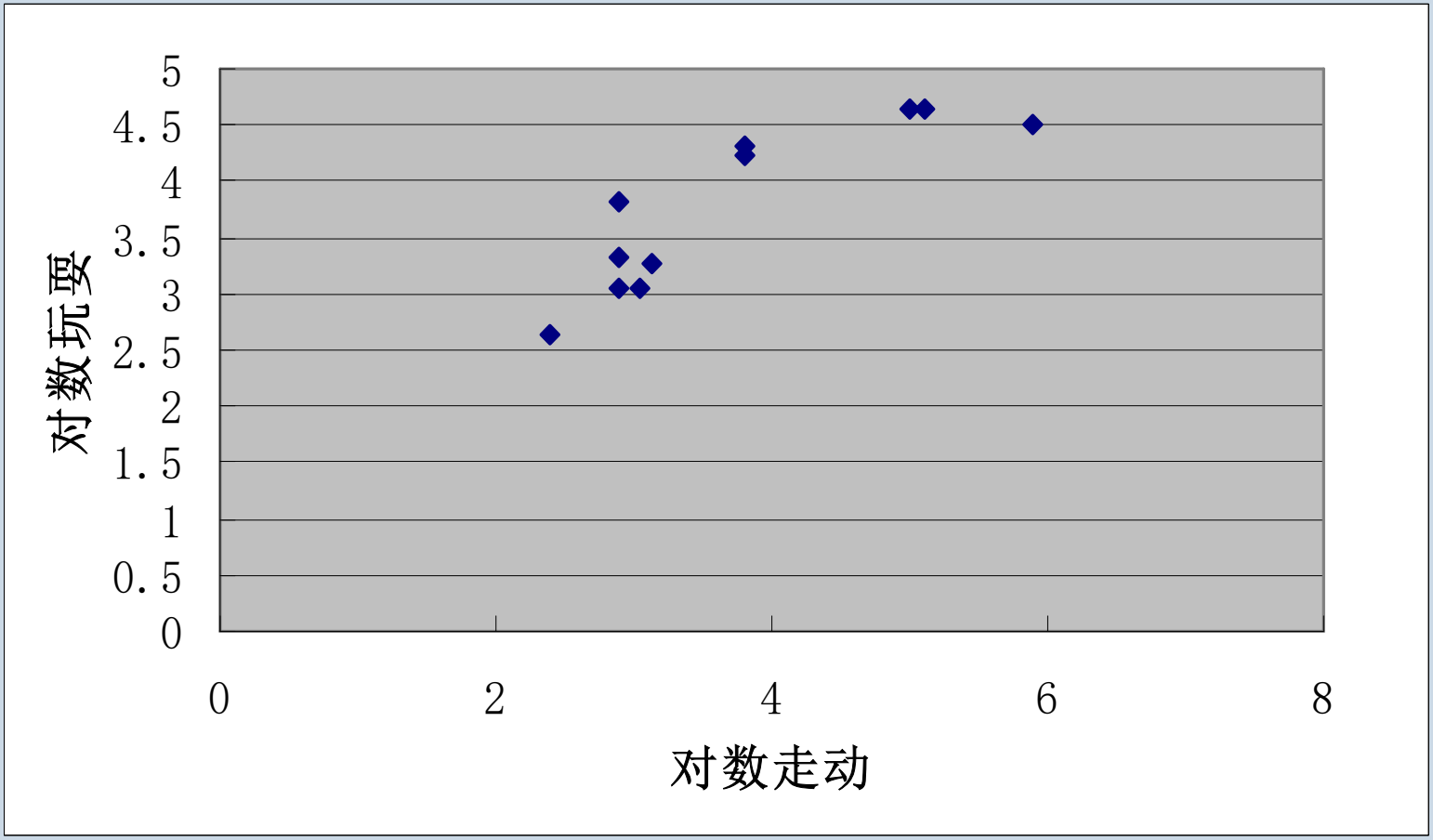
此时对应的检验统计量 $F = 27.438$ ， 而

p-值 $\mathrm{P}(F(1,9) > 27.438) = 0.00054$，

即在水平0.001下回归关系也是显著的。

哺乳动物出生以后开始玩耍的时间 p 关于开始走动的时间 w 的回归关系更可能应该是：

$$p = 5.42\, w^{0.561} \text{ 。}$$


对数玩耍
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
2
4
6
8
对数走动

四. 回归方程的预测与控制

假定对回归模型 $y = \beta_0 + \beta_1 x + \varepsilon$，我们已经观察到了一组数据$(x_i, y_i)$，$1 \le i \le n$。

现在希望了解 $x = x_0$ 时对应的 $y = y_0$ 的情况。

很自然的，应该有关系：

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$$

1. 回归方程的预测

如果只需要 y_0 的一个点估计，显然有：

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

如果需要预测 y_0 的一个范围，则应该求出一个区间估计，但必须知道与 y_0 有关的分布。

记 $y_0^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$ (可以计算出来)

这里 y_0^* 只可能和随机变量 ε_1，…，ε_n 有关，

并且根据定理 **5.1.3**，($\hat{\beta}_0$，$\hat{\beta}_1$) 服从二维正态分布，因此 y_0^* 是一个只与 ε_1，…，ε_n 有关的服从一维正态分布的随机变量。

现在 $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$，只和随机变量 ε_0 有关。因此，

$y_0 - y_0^*$ 是两个独立正态随机变量的差，仍然服从正态分布。

有 $$y_0 - y_0^* \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right)$$

把 $y_0 - y_0^*$ 的分布中心标准化为：

$$\frac{y_0 - y_0^*}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim N(0,1)$$

要消去未知参数σ，使用总体方差σ^2 的估计$\hat{\sigma}^2$；
它与 $\hat{\beta}_0$、$\hat{\beta}_1$ 都独立，自然也与 $y_0 - y_0^*$ 独立。

$$\frac{n-2}{\sigma^2}\hat{\sigma}^2 \sim \chi^2(n-2)$$

$$\frac{y_0 - y_0^*}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim t(n-2)$$

$$y_0^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

从而得到 y_0 的一个置信度 $1-\alpha$ 区间估计，或者说给出一个 x_0，则相应的因变量 y_0 以 $1-\alpha$ 的概率在如下的一个范围内变化：

从 $(\hat{\beta}_0+\hat{\beta}_1 x_0 - h)$ 到 $(\hat{\beta}_0+\hat{\beta}_1 x_0 + h)$

这里 $h = t_{\alpha/2}(n-2)\,\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

2. 回归方程的控制

这是预测问题的逆问题，即需要 y_0 以 $1-\alpha$ 的概率落在一个范围 (A, B) 内，问 x_0 的变化范围应该是什么？

只需要取 x_0 使得：

$A \le y_0^* - h$ 以及 $y_0^* + h \le B$

同时成立，即可解出 x_0 相应的变化范围

这里

$$y_0^* = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$h = t_{\alpha/2}(n-2)\,\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

例5.2.4　给出例5.2.2中成年女性当身高 $x_0=170$ 时腿长 y_0 的预测区间。

解：　回归方程的 p-值为 2.13×10^{-9}，即

腿长 ＝ -16.073 ＋0.7194 身高

具有高度显著的统计意义。

身高 $x_0=170$ 时腿长 y_0 的预测区间为：

(102.746 , 109.6882)

□

例5.2.5　对某种型号钢的抗拉强度 Y 与硬度 X 观察了20 个数据，建立回归方程、检验并预测 $x = 230$ 时的 y 值。

x	277	257	255	278	306	268	285	286	272	285
y	103	99.5	93	105	110	98	103.5	103	104	103
x	286	269	246	255	253	255	269	297	257	250
y	108	100	96.5	92	94	94	99	109	95.5	91

□

回归分析需要注意的几点

(1) 实际问题中回归模型的建立要依赖于专业知识，并且注意散点图的使用；

(2) 即使回归模型通过了检验也只能认为所研究的变量是统计相关的；

(3) 回归分析一般需要与相关分析结合起来；

(4) 异方差性、序列相关性、多重共线性问题。

习题 5.2

1-3. 教材 **234** 页

第 **3**、**4**、**17** 题。

第5.3节 多元回归分析

对于一般的多元回归模型：

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \beta \in \mathbf{R}^{k+1}$$

一. 未知参数的估计

常数项 β_0 、各回归系数 β_i 以及误差方差 σ^2 的估计自然就采用定理 **5.1.1** 及**5.1.3**中的估计。

东北大学数学系

二. 回归模型的检验

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

如果拒绝了零假设，则说明 y 的期望与自变量 $x_1, \ldots, x_k$ 具有显著的线性关系，回归方程成立；否则说明我们建立的回归关系不显著。

同理如练习**5.1.1**定义总平方和**TSS**、
回归平方和**RegSS** 以及 残差平方和**RSS**，

$$\mathrm{TSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$\mathrm{RegSS}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \qquad \mathrm{RSS}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

仍然可以证明 **RegSS** 与 **RSS** 独立，

根据定理**5.1.3**，

$$\frac{\mathrm{RSS}}{\sigma^2}\sim\chi^2(n-k-1)$$

并且零假设 H_0：$\beta_1 = \ldots = \beta_k = 0$ 成立时，有

$$\frac{\text{RegSS}}{\sigma^2} \sim \chi^2(k)$$

所以由 F 统计量的构造，

$$F = \frac{n-k-1}{k} \frac{\text{RegSS}}{\text{RSS}} \sim F(k, n-k-1)$$

故零假设(即回归模型不成立) 的一个水平 α 的拒绝域为 $\{ F \geq F_\alpha(k, n\text{-}k\text{-}1) \}$

检验的 p-值是 $P\{F(k, n\text{-}k\text{-}1) > F$ 比$\}$

EXCEL函数*LINEST*检验多元回归模型

以 $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \varepsilon$ 为例，

	$\hat{\beta}_3$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_0$
	$\sqrt{c_{33}}\hat{\sigma}$	$\sqrt{c_{22}}\hat{\sigma}$	$\sqrt{c_{11}}\hat{\sigma}$	$\sqrt{c_{00}}\hat{\sigma}$
	r^2	$\hat{\sigma}$		
	F 比	n-3-1		
	RegSS	RSS		

三. 回归因子的挑选

逐步回归的想法

讨论假设检验的问题：

$$\mathbf{H}_{0i}: \ \beta_i = 0 \quad \Leftrightarrow \quad \mathbf{H}_{1i}: \ \beta_i \neq 0$$

如果接受了这个零假设，就可以把因子 x_i 从模型中剔除。

(有时候也采用逐一添加回归因子的方法)

对于 H_{0i}： $\beta_i = 0$ ， $i = 1, 2, \ldots, k$

如果采用 t 检验，则检验统计量为

$$T_i = \frac{\hat{\beta}_i}{\sqrt{c_{ii}}\hat{\sigma}} \quad ，\text{自由度 } n\text{-}k\text{-}1$$

检验的p-值是 $P\{ t(n-k-1) > T_i \}$

如果采用 F 检验，则检验统计量为

$$F_i = \frac{\hat{\beta}_i^2}{c_{ii}\hat{\sigma}^2} \quad ,$$

检验的p-值是 $P\{ F(1, \ n-k-1) > F_i \}$

例**5.3.1**　钢的去碳量 **Y** 与两种矿石 X_1、X_2 以及溶化时间 X_3 有关，实际测量了**49**组数据，作多元回归分析。

解. 见相关数据文件。

□