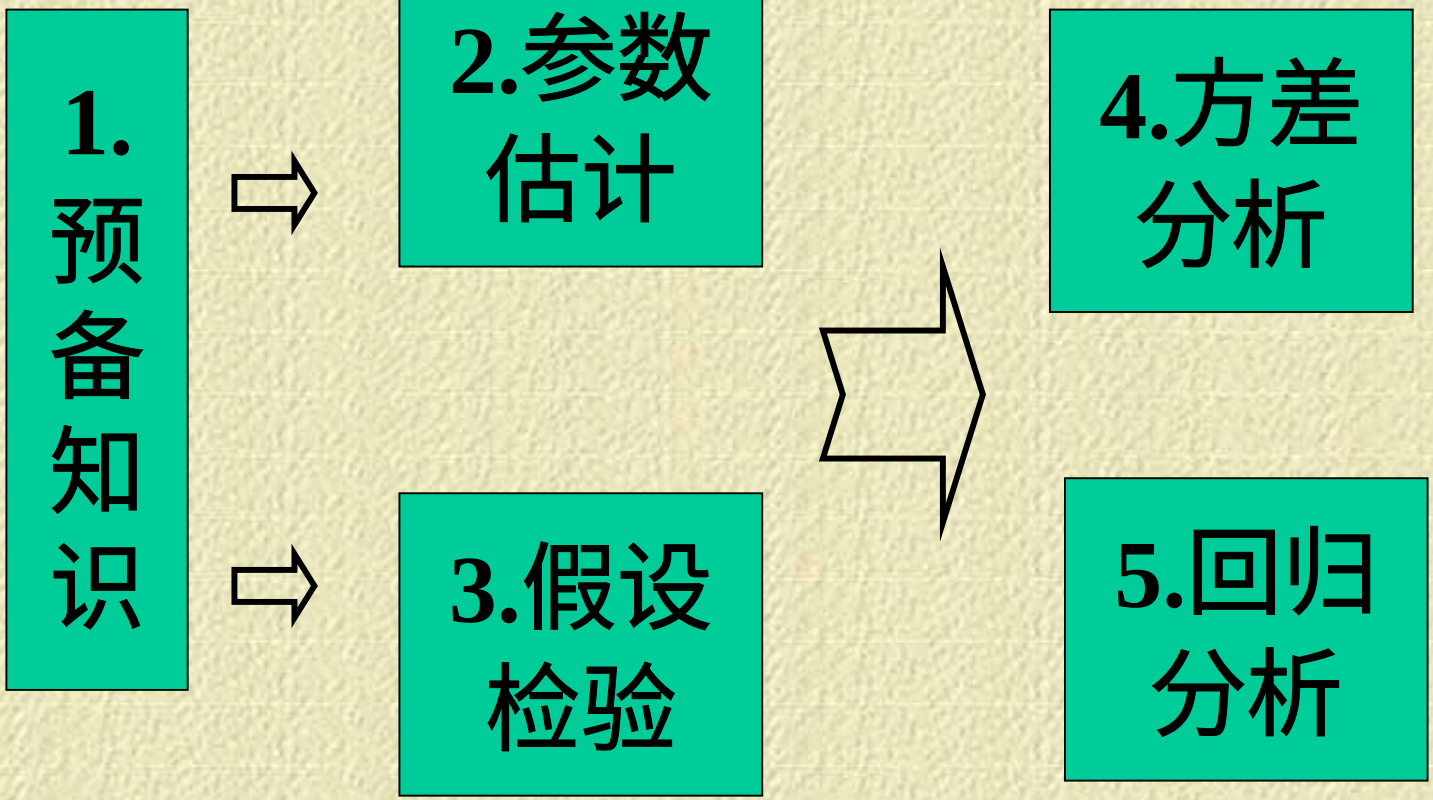
《应用数理统计》

常桂松
东北大学数学系

guisong_chang@126.com


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

参考教材

1. 《*Mathematical Statistics*》– R.V.Hogg
   (第四版中译本，世界图书，1979)

2. 《*Statistics -The Conceptual Approach*》– G. R. Iversen, ed
   (中译本，吴喜之等，高教-施普林格，2000)

3. 《*Mathematical Statistics and Data Analysis*》– J. A. Rice
   (第二版影印本，机械工业，2003)

4. 《数理统计引论》– 陈希孺 (科学，1999)

阅读材料

1. 《*Ideas of Statistics*》– J.L.Folks
   (中译本，魏宗舒等，上译出版，1987)

2. 《*The fascination of Statistics*》– R. J. Brook, ed
   (影印本，世界图书，1986)

3. 《*Applied Regression Analysis and other Multivariable Methods*》
   – D.G.Kleinbaum, ed (第三版影印,机械工业,2003)

4. 《*Analyzing Multivariate Data*》– J. M. Lattin, ed
   (影印本，机械工业，2003)

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

第1.4节　样本数据的收集

第1.5节　*EXCEL* 统计计算

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。

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。

1.1.1 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。

0 表示 抽烟者；1 表示 不抽烟者。

2. 样本 (*sample*)

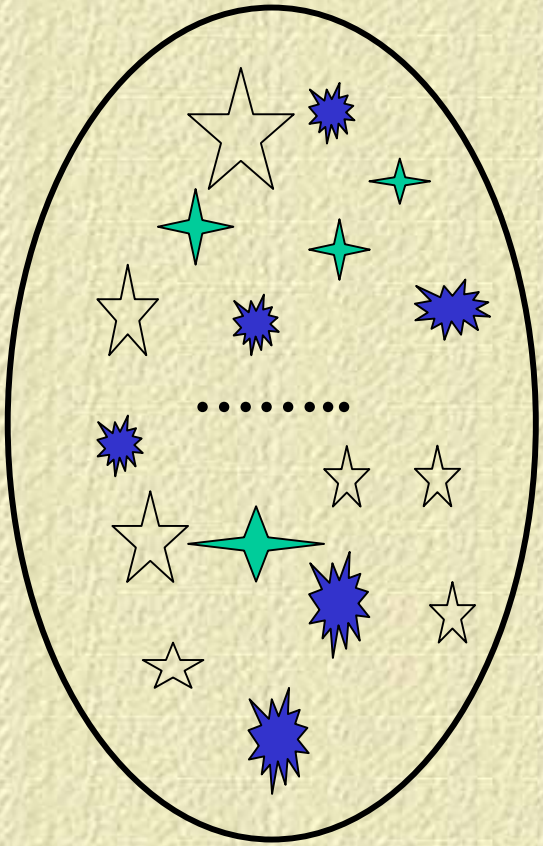
从总体中取出一个个体，称为从总体中得到一个样本。

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。

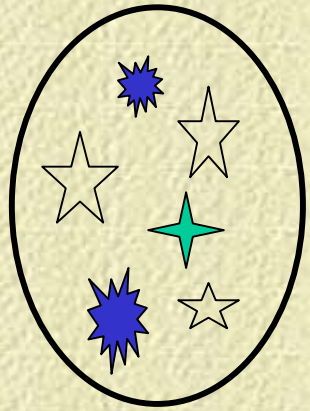
统计学的目的就是从样本去得出总体的信息。

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。 参数估计

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。

1.1.2 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： 希望了解某所高校学生月消费情况。

解决方法：从这所大学里随机地调查有代表性的一些学生，根据收集到的数据去得出这所大学学生每个月支出费用的有关信息。

1. 如何得到样本 ？ 抽样调查

不同家庭背景学生的比例应该各占多少？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。

这里总体是这所大学的学生月支出费用，我们不妨认为学生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 $X \sim N(\mu, \sigma^2)$，而这个学校相应的两个参数 μ 与 σ^2 是未知的。

（不同学校对应的这两个参数也就不相同）

Remark

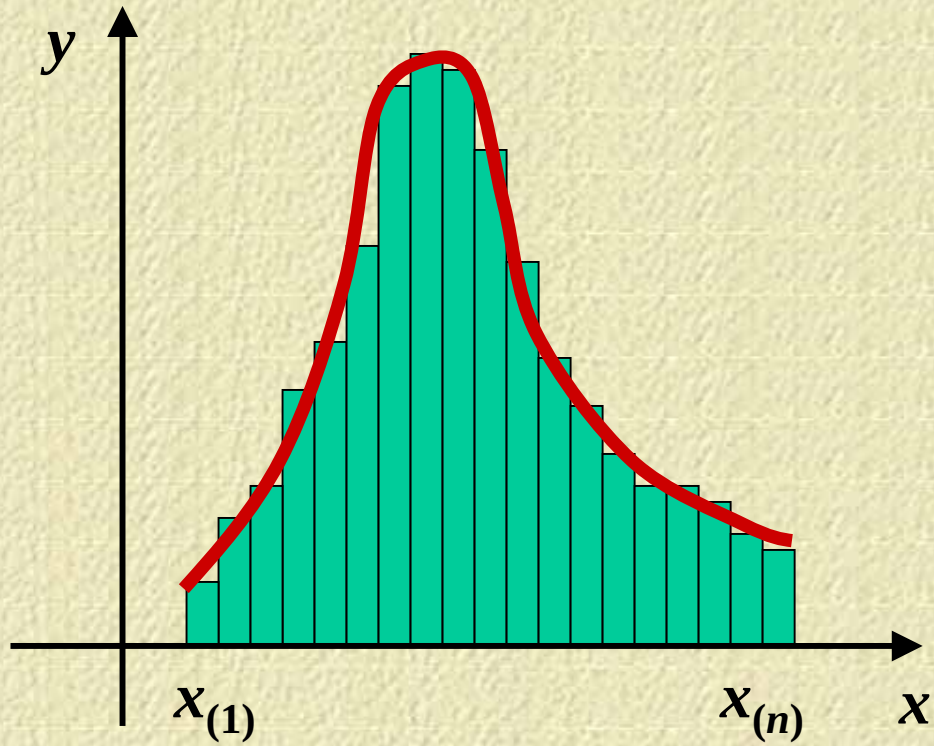
当不知道或者难以确定总体的分布类型时，在统计学中常常采用下面两种办法来近似得到总体分布的有关信息。

(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：$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$；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。

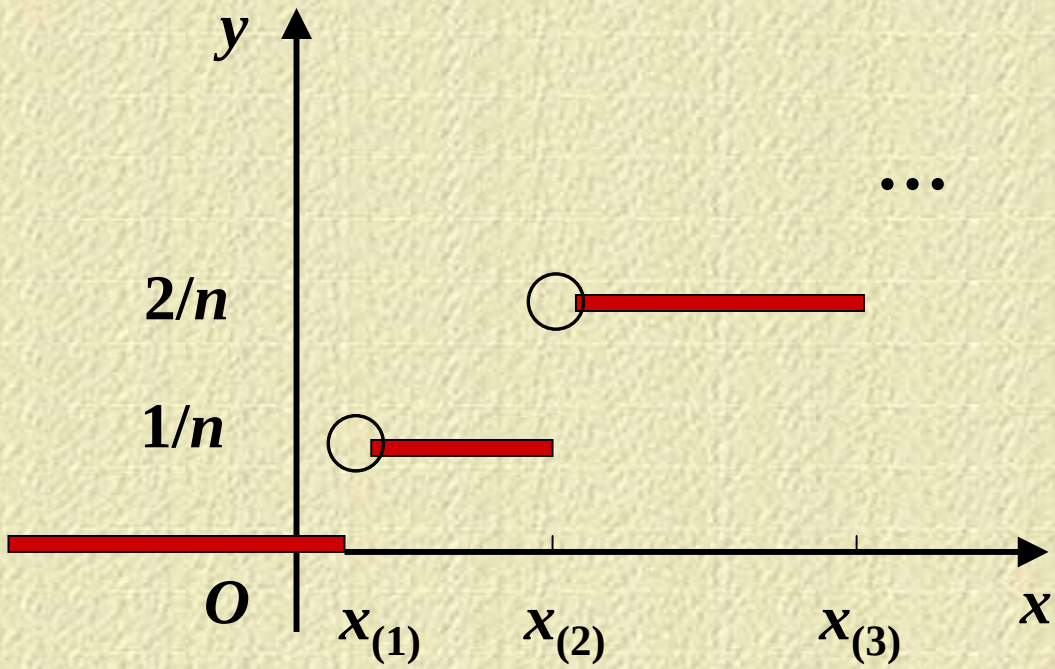
(2). 经验分布函数的方法

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

$F_n(x)=\{x_1$，…，x_n中小于 x 的个数$\}/n$



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 $F(x)$ 。

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。

调查一个学生得到了一个数据，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 个学生的月支出 $x_1, x_2, \ldots, x_n$ 的算术平均：

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

去估计这所学校学生的平均月支出费用 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

假定学校要制定相关一些政策，如奖学金、贷款、勤工俭学等；或者后勤服务、商业经营的价格等等。

共同关心的一些问题，比如说：

$$\mu > \mu_0 \ ?$$

这里 μ_0 是一个已知的常数。

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

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学校学生月平均支出 (即总体的参数 θ) 的估计值，自然就可以用它代替假设里的 θ 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：

小儿麻痹症、SARS疫苗的研制，

越战期间美国的征兵计划，

…

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，钢的去碳量与不同矿石、融化时间、炼钢炉体积等等是否有关？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

比如说产品质量与不同操作人员之间的关系。是否某些人生产出的产品质量偏高？如果偏高，这种差异是否是纯属偶然原因...

双因素方差分析

希望了解操作人员和设备这两个因素联合对质量的关系。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

1.1.3 统计数据的直观描述：图与表

为数据作图有两个目的：

帮助研究者从数据中提取信息，
很方便把信息传递给其它人。

统计数据的图主要有：

分类变量的饼图与条形图，
数值变量的点图、直方图与散点图

看懂图是21世纪的成年人必须具有的能力

1. 分类变量的图表示

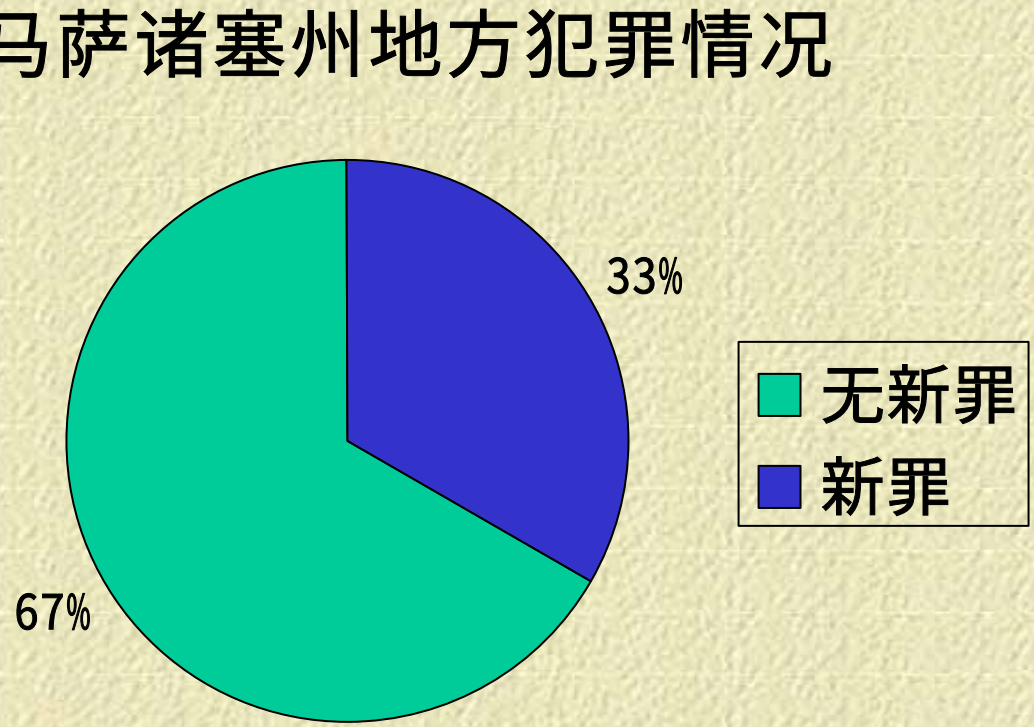
分类变量(*Categorical Variable*) 主要指这种变量的各个取值没有大小、顺序的区别，不能做数学运算。

如：性别变量、属性变量等

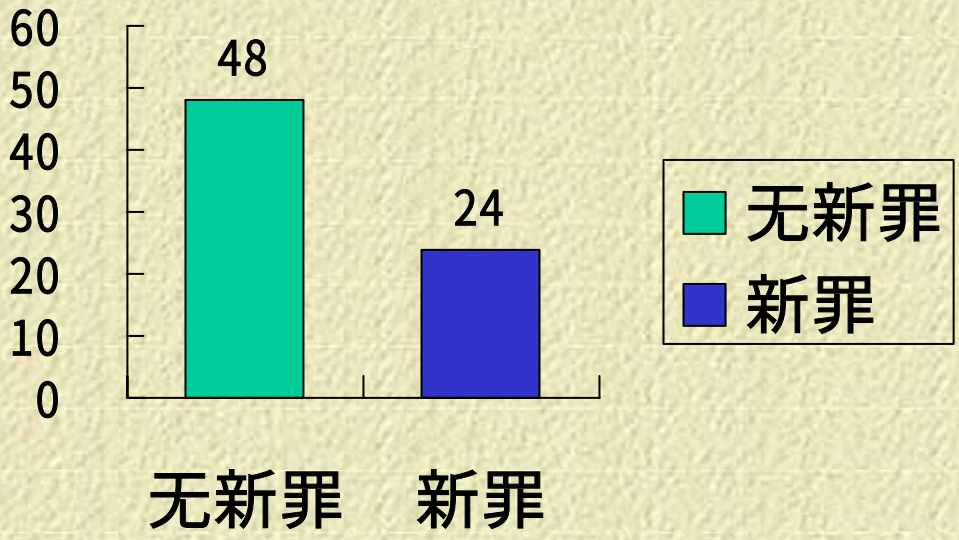
主要有饼图、条形图两种表示方法

例1.1.1

马萨诸塞州犯罪情况(1993年)



马萨诸塞州地方犯罪情况



2. 一个数值变量的图表示

数值变量又称为度量变量(***metric variable***)，这些变量的取值之间可以做数学运算。

不考虑区间或比例变量这些度量变量。

数值变量的图表示主要有；

点图(点线图、盒形图、茎叶图)，
直方图，散点图，时间序列图

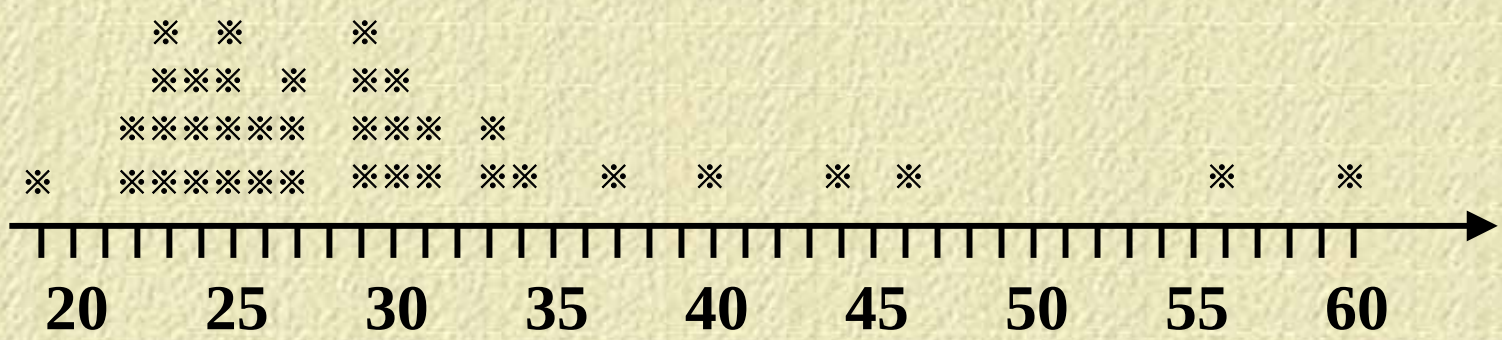
例1.1.2

下面是1995年美国一家地方报纸列出的一个星期内申请结婚的 37 位女性的年龄：

30	27	56	40	30	26	31	24
23	25	29	33	29	22	33	29
46	25	34	19	23	44	29	30
25	23	60	25	27	37	24	22
27	31	24	26	23			

最小的19岁，最大的60岁

(1). 点线图(*Lineplot*)



可以看出大多数数据集中在22岁~34岁

(2). 盒形图 (*Boxplot*)

与点线图具有相同的刻度，各占 1/4，盒子占一半。

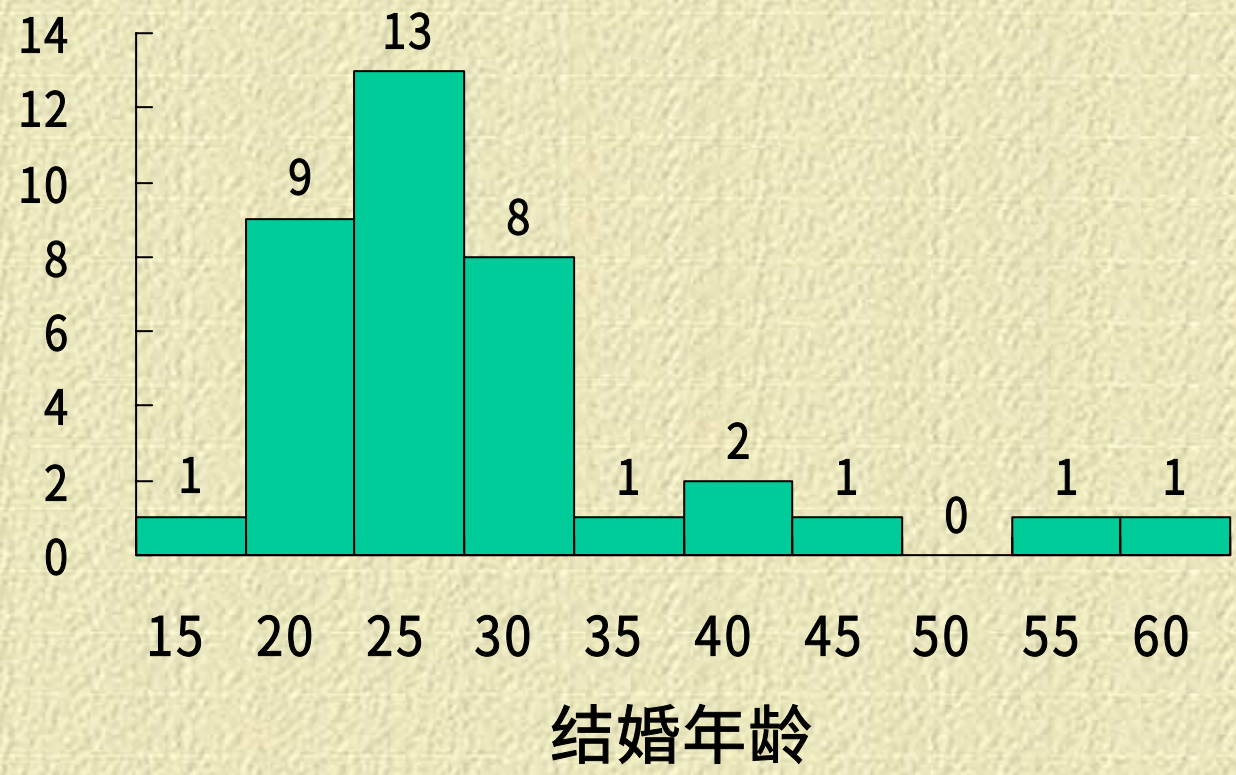
盒形图主要出现在专业文献中。

19 24 27 32 60

(3). 茎叶图 (*Stemplot*)

6	0
5+	6
5	
4+	6
4	0 4
3+	7
3	0 0 0 1 1 3 3 4
2+	5 5 5 5 6 6 7 7 7 9 9 9 9
2	2 2 3 3 3 3 4 4 4
1+	9

(4). 直方图 (*Histogram*)



3. 两个数值变量的图表示

(1). 散点图(*Scatterplot*)

把两个变量分别作为横轴和纵轴描出散点

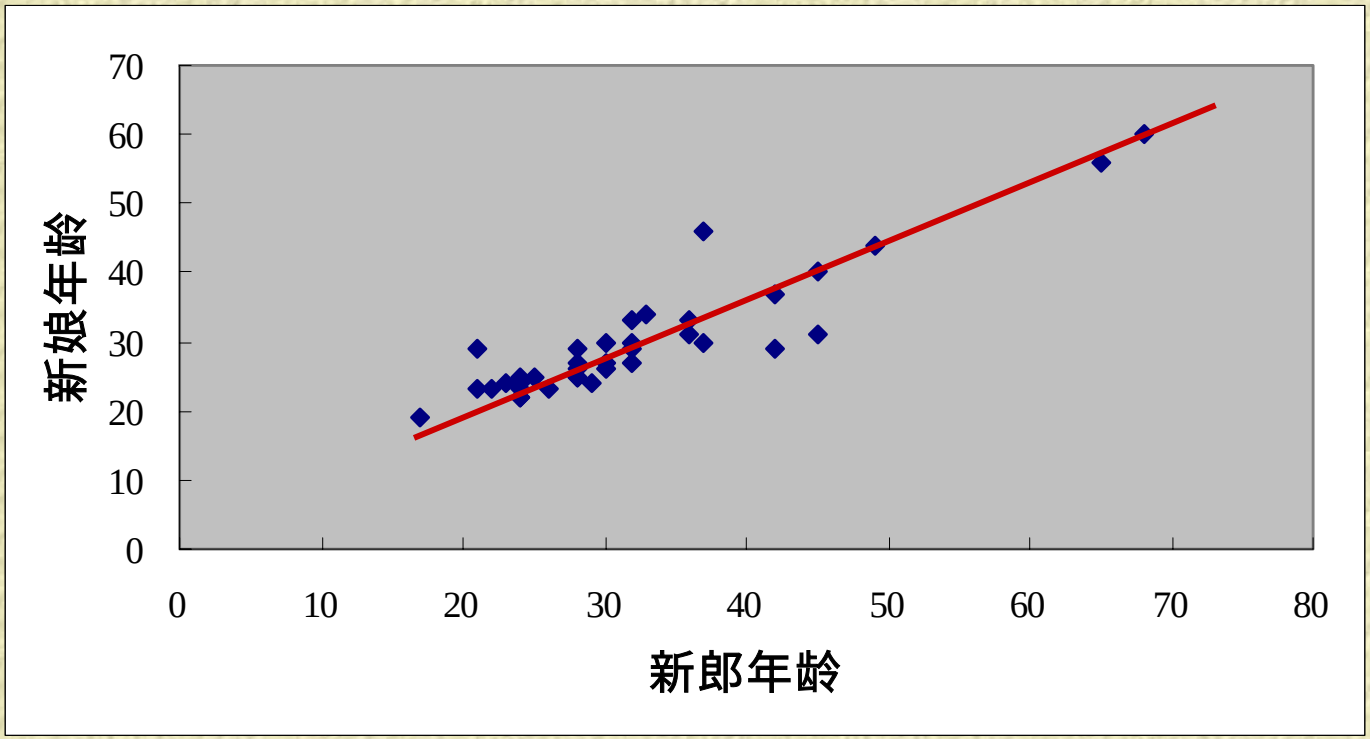
散点图在简化数据的同时，能够保留原始数据的信息。

例1.1.3 下面是37对新婚夫妻的年龄(新郎,新娘)

(37,30) (30,27) (65,56) (45,40) (32,30) (28,26)
(45,31) (29,24) (26,23) (28,25) (42,29) (36,33)
(32,29) (24,22) (32,33) (21,29) (37,46) (28,25)
(33,34) (17,19) (21,23) (24,23) (49,44) (28,29)
(30,30) (24,25) (22,23) (68,60) (25,25) (32,27)
(42,37) (24,24) (24,22) (28,27) (36,31) (23,24) (30,26)

利用*EXCEL*文件.xls，输入数据后生成散点图

EXCEL 生成的散点图



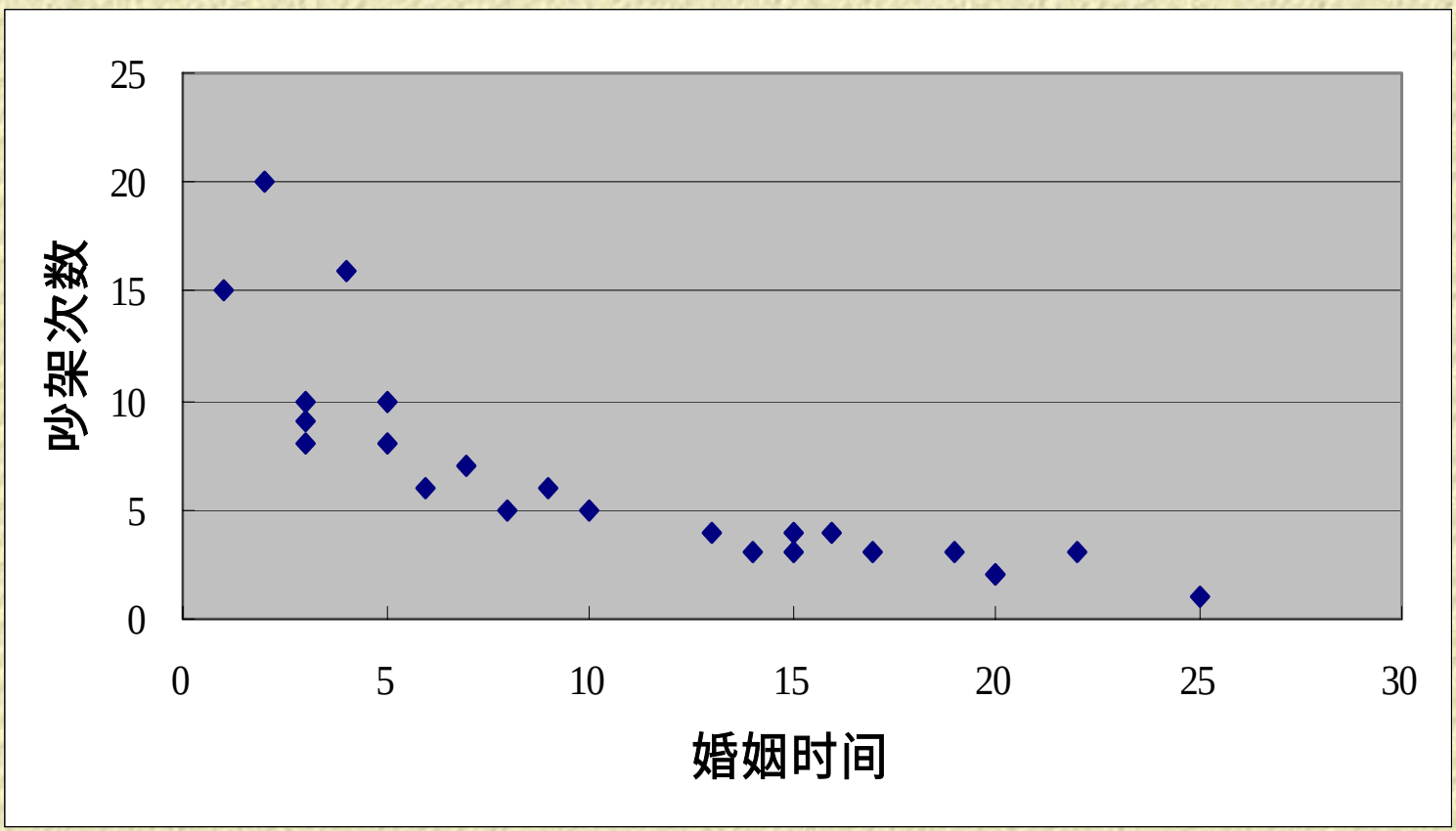
例1.1.4

下面是 24 对夫妻的数据，有两个变量：结婚时间和一年内的吵架次数。

结婚年数	5	2	4	1	3	6	5	8	3	7	3	9
争吵次数	10	20	16	15	9	6	8	5	10	7	8	6

结婚年数	10	15	13	20	16	25	22	14	15	19	17	20
争吵次数	5	3	4	2	4	1	3	3	4	3	3	2

结婚时间与吵架次数的散点图



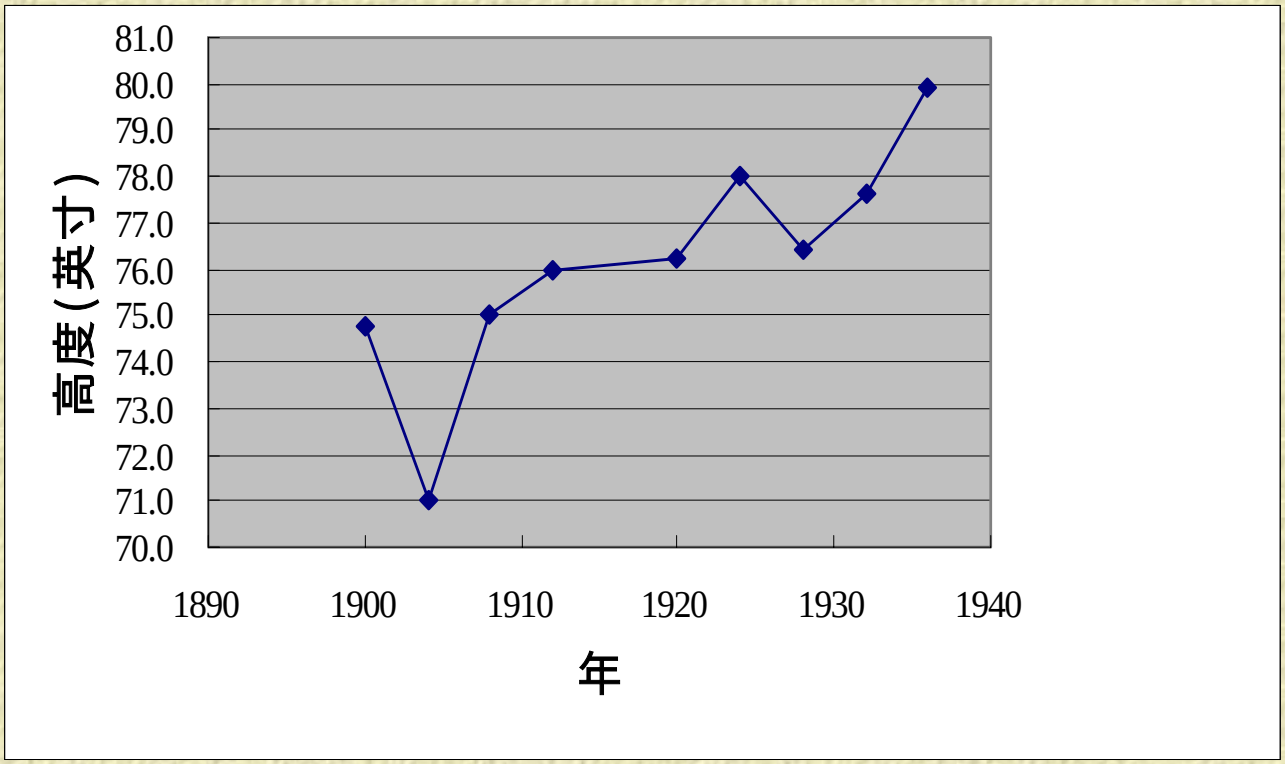
(2). 时间序列图

特殊散点图，以时间作为横轴的变量

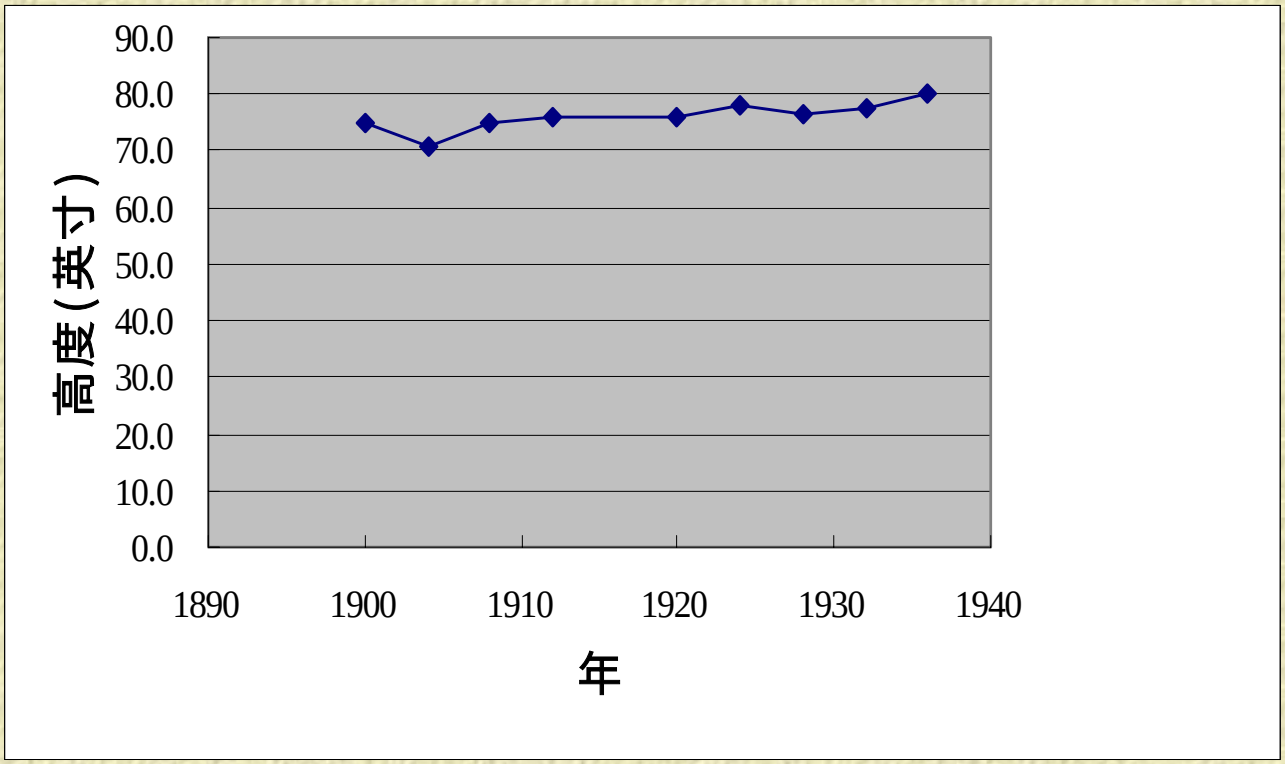
时间序列图能够反映出一个变量随着时间而变化的趋势。

需要注意的时，时间序列图的形状可能会给读者错觉。

例1.1.5 奥运会男子跳高冠军成绩



以 0 英寸作为纵轴起点的时间序列图



4. 统计表 (*table*)

美国人口普查局1992年高等教育年鉴的一组样本数据：

	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

5. 优秀统计图的标准

在最短的时间里，用最少的笔墨，使用最小的空间，给读者最多的信息

1.1.4 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

练习1.1.6

收集实际生活与学习、工作中的一些统计数据。

练习1.1.7

证明经验分布函数$F_n(x)$的收敛性。

第1.2节 概率论基础

1.2.1 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

1.2.2 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 $P(B|A)$

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

1.2.3 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

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：$P(X \leqslant x)$

离散随机变量的分布函数是阶梯型跳跃函数，

对满足$(x_k \leqslant x)$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，

对满足$(t \leqslant x)$的密度函数 $f(t)$ 积分得到。

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、分布律、期望、方差；
指数分布：背景、分布律、期望、方差；
正态分布：背景、分布律、期望、方差。

1.2.4 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 $f(x|\theta)$ 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 $f(x|\theta)$，则X 与θ 的联合分布是 $h(\theta)\times f(x|\theta)$。

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. 独立相同类型随机变量的和

正态分布的可加性

二项分布的可加性

卡方分布的可加性

1.2.5 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：

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

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 $f(x)$。

1.2.6 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： **Bernoulli 定理**

依分布收敛：分布函数的收敛

几乎处处收敛：**Kolmogrov 定理**

2. 中心极限定理

De moivre-Laplace 定理

Lindeberg-Levy 定理

Lindeberg 定理

第1.3节 统计量与抽样分布

1.3.1 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，$g(\cdot)$ 是一个完全已知的函数，则称 $g(X_1, \ldots, X_n)$ 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 $g(X_1, \ldots, X_n)$ 的相应的观察值就是 $g(x_1, \ldots, x_n)$。

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？

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

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 $f(x,\theta)$

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 $g(\cdot)$，只要 $E_\theta g(T) = 0$ 就可以推出 $P_\theta\{ g(T) = 0 \} = 1$，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. 因为全体样本之和服从二项分布$B(n,p)$，

如果对所有的 $0<p<1$ 下式都成立：

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

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

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) $f(x,\theta)$ 可表示成：

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

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 $(\Sigma T_1(X_i),\ldots,\Sigma T_k(X_i))$ 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 $X_{(n)}$ 。

1.3.2 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 $M(X)$ ：

$$P\{X \le M(X)\} \ge 0.5 \le P\{X \ge M(X)\}$$ **；**

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，否则应该用中位数。

(3). 样本的众数适用于离散的总体

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。

注意：“变异系数” $v = S / \overline{X}$ 用来比较不同均值时的两组数据的离散程度

例如两组数据{ 2,4,6,8,10 } 与 { 20,40,60,80,100 }
样本均值分别是6与60，而S_1^2 =10，S_2^2 = 1000 ；
表面上看起来第二组数据更分散。

它们的变异系数都是 $v = \sqrt{10} / 6$

因此理解为这两组数据的离散程度相同。

思考2

举两个现实生活中考虑变异系数的例子。

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：

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

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

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 $f(x)$，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：

$$f(y_1,\dots,y_n) = n!\, f(y_1)\dots f(y_n), y_1 \le y_2 \le \dots \le y_n$$

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

$$f_k(y) = \frac{n!}{(k-1)!(n-k)!} f(y)F(y)^{k-1}[1-F(y)]^{n-k}$$

例1.3.9　极小统计量$X_{(1)}$的概率密度函数是：

$$f_1(y) = nf(y)[1 - F(y)]^{n-1}$$

极大统计量$X_{(n)}$的概率密度函数是：

$$f_n(y) = nf(y)[F(y)]^{n-1}$$

思考3

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($k < l$) 的
联合概率密度函数：

$$f_{k,l}(y_k, y_l) = \frac{n!}{(k-1)!(l-k-1)!(n-l)!} f(y_k) f(y_l)$$
$$\times F(y_k)^{k-1}[F(y_l)-F(y_k)]^{l-k-1}[1-F(y_l)]^{n-l}, \quad y_k < y_l$$

例1.3.10　极差的概率密度函数是：

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

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4. 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？

统计学家的脚

例1.3.11　乙同学毕业后求职于一家公司。总经理说公司平均月薪是 3000 元。一个月后乙同学得到工资1000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1000 元。应该如何理解乙同学的遭遇 ？

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 70	70 ~ 200	大于 200
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。

1.3.3 统计学中的三个分布

(A) 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 $2n$

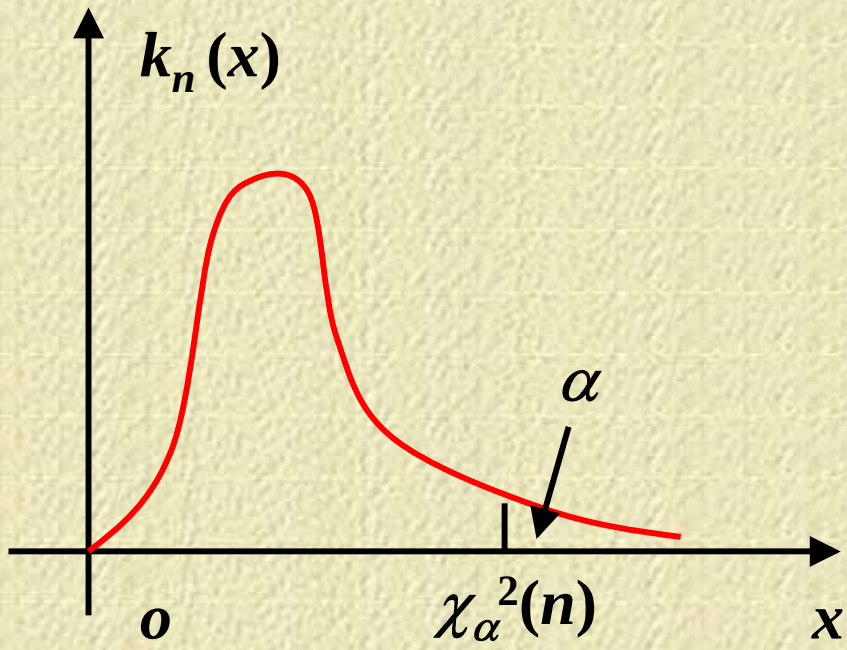
3. 卡方分布具有“可加性”

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 $\chi_\alpha^2(n)$。

(*B*) *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

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

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。

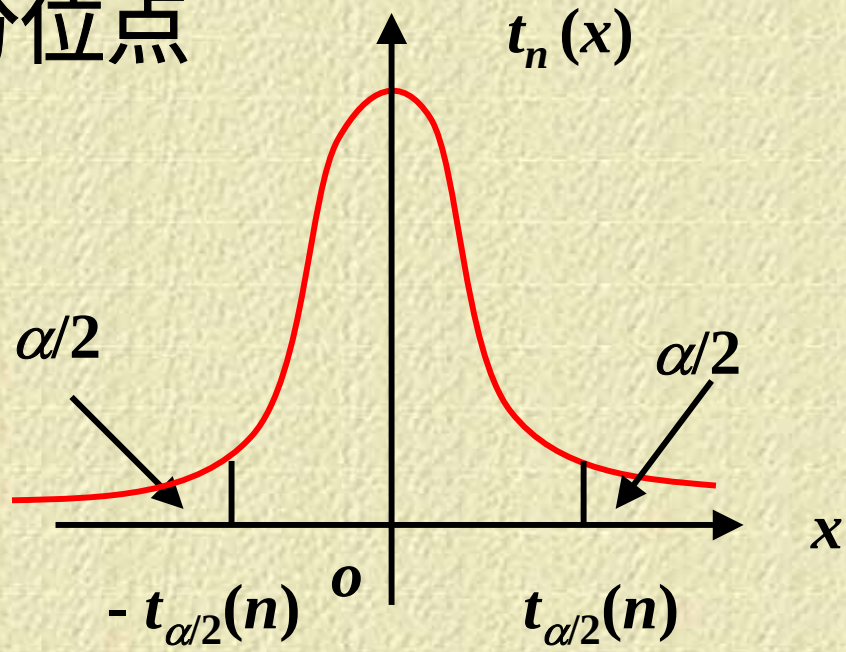
4. t 分布的双侧分位点

假定 $X \sim t(n)$，

给定：$0 < \alpha < 1$，

如果一个数 c 满足：

$P\{|X| > c\} = \alpha$，

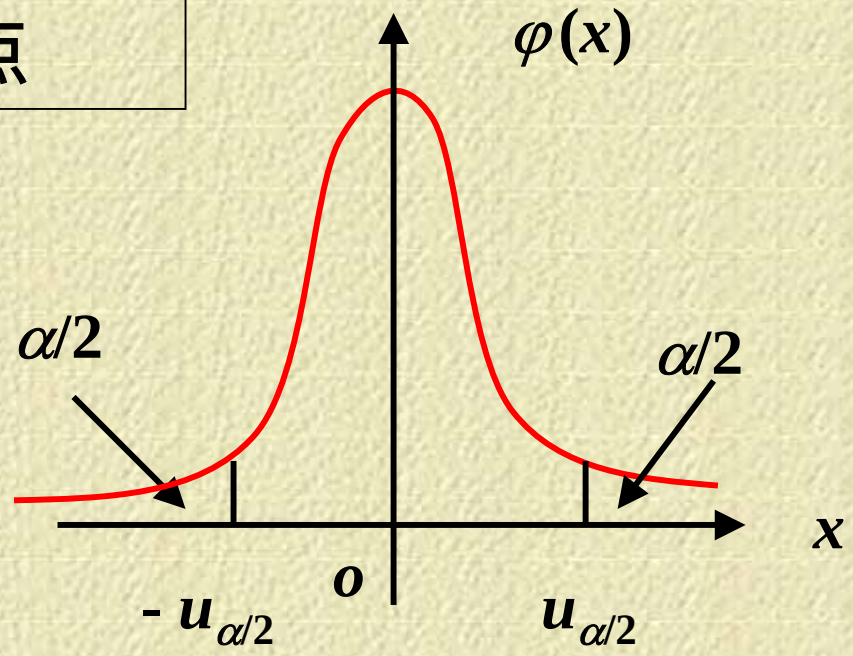


则称这个数 c 是自由度 n 的 t 分布的双侧 α 分位点 (数)，记成 $t_{\alpha/2}(n)$。

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 $u_{0.025} = 1.96$

(C) 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

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

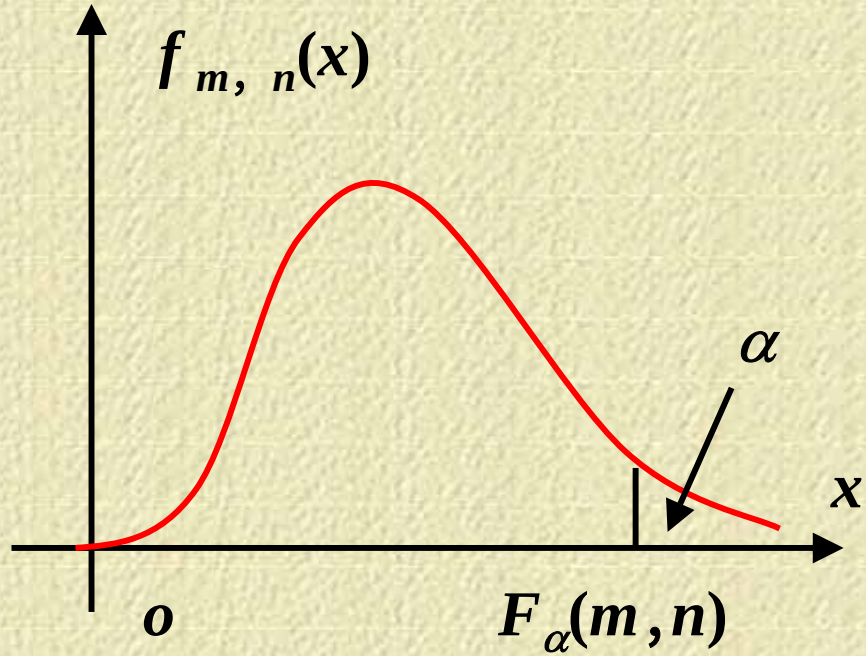
数学期望是 $n/(n-2)$ ($n \geq 3$)

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。

Gamma 分布 $G(\alpha,\beta)$

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

这里参数 $\alpha>0$，$\beta>0$ ；

$\Gamma(\alpha)$ 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 $G(1,\lambda)$。
2. 自由度 n 的卡方分布 $\chi^2(n)$ 就是 $G\left(\frac{n}{2},\frac{1}{2}\right)$。
3. *Gamma* 分布 $G(\alpha,\beta)$ 对于α 具有可加性；
   而且如果 $X\sim G(\alpha,\beta)$，则 $cX\sim G(\alpha,\beta/c)$

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

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

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：

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

数据中的信息

2001年美国家庭收入状况

(— 美联储2003年调查报告)

1. 美国家庭净资产(万美元)：

	中位数	平均值
富裕家庭(10%)	83.36	225.82
贫穷家庭(20%)	0.79	5.26
中等家庭	8.61	39.55
45-54岁主持的家庭	13.2	48.56

2. 美国家庭税前年收入(万美元)：

	中位数	平均值
富裕家庭(10%)	16.96	30.27
贫穷家庭(20%)	1.03	1.00
中等家庭	3.99	6.80
35-44岁主持的家庭	4.43	5.64
45-54岁主持的家庭	4.65	7.64

3. 美国家庭偿还债务能力

平均杠杆比率(负债/总资产)： 12.1%

平均家庭偿债占总收入： 12.5%

练习1.3.14

(1) 总体 $X \sim U(0,\theta)$ 时，计算$X_{(n)}$ 的数学期望；

(2) 总体 $X \sim U(\theta,\theta+1)$时计算期望 $E[X_{(1)}+X_{(n)}]$。

练习1.3.15

对于任意随机变量 X，证明：

$h_1(c) = E(X-c)^2$ 达到极小，当且仅当 $c = E\,X$；

$h_2(c) = E|X-c|$ 达到极小，当且仅当 $c = M(X)$。

练习1.3.16

证明*Gamma* 分布有关性质以及定理1.3.2。

第1.4节 样本数据的收集

收集数据主要有两种方法：

观测数据：研究人员从旁观者的角度观测得到；

试验数据：研究人员控制某些变量，然后测量每次试验的结果，以此来研究变量之间的因果关系。

研究问题前，要对研究的问题有一个明确、详细的定义，尽量减少非随机因素的干扰。

1.4.1 观测数据

1. 样本的选择标准 — “具有代表性”

要保证总体中每个个体以确定概率进入样本中，这样的样本被称为 “随机样本” (***Random Sample***) 。

随机样本的作用：把根据它们得出的当前数据的结论推广到原来的总体时，仍然是成立的。

2. 方便样本(*Convenience Sample*)

为了方便而随便得到的、常常是“不好”的样本。

从它们得到的结论不应该应用到总体。

例如，

心理学的一些研究结果 — 用自己的学生作样本，结论应该就只适用于选修心理课程的同学；

一些杂志的读者调查，……

3. 如何具体去抽取样本

“简单随机样本” (*Simple Random Sample*) ：

每个个体以相等的概率被选中进入样本。

最常用的方法：

把总体的每个个体编号，由计算机随机生成 抽样号码。

Remark

也可以有例外的情况 (“不等概率抽样”) 。

其它常用的抽样方法

(1) 分层抽样(*Stratified Sampling*)：

总体中存在若干明显不同的群体时，把总体划分为不同的层，在各个层内单独抽样。

(2) 整群抽样(*Cluster Sampling*)：

把总体随机划分成小群体，对抽到的小群体进行普查。

(3) 多阶抽样(*Multi-stage Sampling*)

(4) 系统抽样(*Systematic Sampling*)

1.4.2 观测数据的误差与错误

1. 抽样误差(*Sampling Error*)

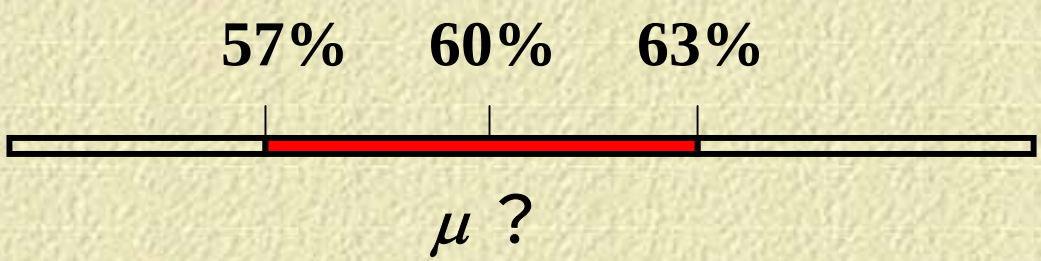
这是不可避免的随机误差，产生的原因在于：每次我们抽取的样本是总体中随机得到的一部分

对于同一个总体的两次不同抽样，由于包含的样本不完全相同，因此所得到结论也就有差别；其次，这些结论与总体的真实情况也有差别。

抽样误差理解为所有这些差异的最大界限

如果没有特别指明，默认的情况下，
统计学中的结论都是以 95% 的可能成立。

“样本比例是 60%，抽样误差为 3 个百分点”
—— 真实比例以 95% 的可能介于 (60±3) %



2. 主要的错误(非抽样误差) 及原因

2.1 未响应误差(*Nonresponse Error*)

由于应该抽取哪些样本在观测数据前已经事先设计好，因此有可能无法收集到其中的某些样本，从而产生缺失数据。

大多数情况下，未响应者和响应者差别不大

2.2 响应误差(*Response Error*)

在社会调查过程中，因为提问的方式、问题的位置、调查员的影响等使得被调查者回答时产生的偏差。

例1.4.1 Rugg 试验：

A：您认为美国应该禁止反对民主的公开言论吗？

B：您认为美国应该允许反对民主的公开言论吗？

例1.4.2　1984年美国总统大选的模拟选票

民主党	共和党
□ 蒙代尔	□ 里根
与弗拉罗	与布什

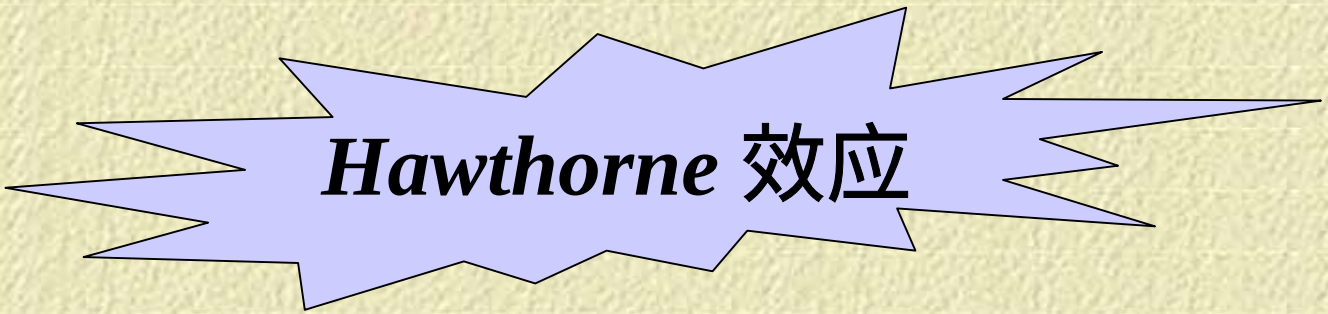
你对于这张选票有什么疑问没有？

1.4.3 试验数据：鉴别因果关系

通过在试验中控制实验对象收集到的变量数据

试验组： 被控制了某些变量的样本集合

对照组： 随机选择的一些样本



统计学家在试验中的任务

(1) 确定样本容量(即数据的规模)

在成本的要求下希望精度最高，

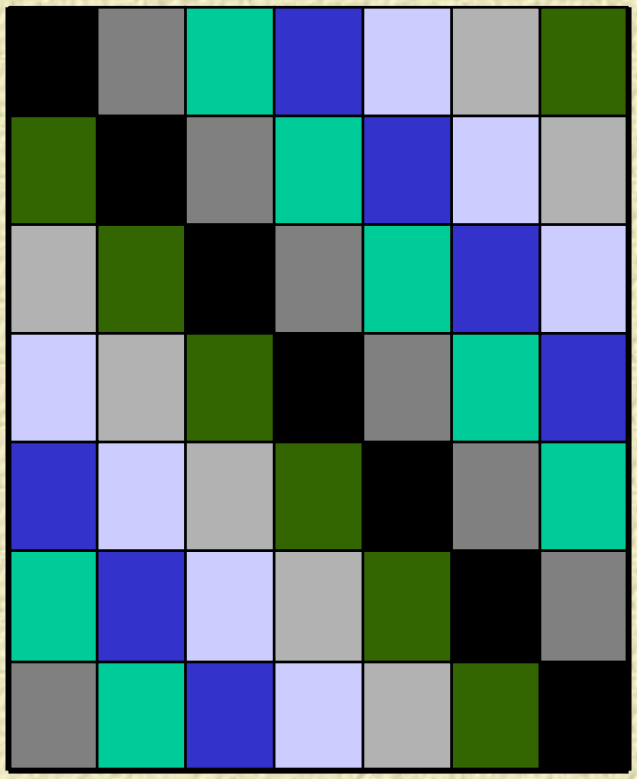
或者是 在精度的要求下希望成本最小

(2) 在试验开始前详细设计与分析

要使得得到的数据符合统计分析的标准

(3) 研究与发明新的试验分析方法

R.A.Fisher 窗户



1.4.4 一个抽样调查实例

1936年美国总统大选的预测

民主党：罗斯福 ←→ 共和党：兰登

当时著名的杂志《文学摘要》根据他们的200 万份调查表的结果预测兰登将以 57% 对 43% 获得大选。

Gallup 从《文学摘要》的名单中抽取 3000 份再调查，得到结果与《文学摘要》的结果几乎没有差别。而他自己抽样的5,000名选民的资料则显示罗斯福将获得大选。

《文学摘要》的“共和党偏差”

按照电话簿和俱乐部成员的名单发放调查表。

敏感问题

1. **Warner 模型**
2. **Simons 模型**

第1.5节　*EXCEL* 统计计算

1.5.1 样本统计量的计算

1. 与“平均”有关的统计量

均值　　*AVERAGE*(全部数据)

中位数　　*MEDIAN* (全部数据)

众数　　*MODE* (全部数据)

2. 与“变差”有关的统计量

方差　　*VAR* (全部数据)

标准差　　*STDEV* (全部数据))

极差　*MAX* (全部数据) - *MIN* (全部数据)

Remark

偏差平方和 *DEVSQ* 的使用

3. 与“顺序”有关的统计量

秩　　*RANK* (某数据, 全部数据, 1)

RANK (某数据, 全部数据, 0) 表示降序

Remark

数据重复出现时 *RANK* 输出的是排在最前面的位置。

1.5.2 有关分布的概率计算

统计学中关心的是 *p*-值，即我们能够观察到所抽取的样本或者更极端情况的概率。

标准正态分布的 *p*-值

*EXCEL*函数： *NORMSDIST*

①如果 $z_0 < 0$， *p*-值 = $NORMSDIST(z_0)$

②如果 $z_0 > 0$， *p*-值 = $1 - NORMSDIST(z_0)$

例1.5.1　假定 $X \sim N(0,1)$

则概率 $P(X \leq -0.8765)$

$= NORMSDIST(-0.8765) = 0.19038$

而概率 $P(X > 5.1234)$

$= 1 - NORMSDIST(5.1234) = 1.50304 \times 10^{-7}$

卡方分布的 *p*-值

EXCEL*函数：*CHIDIST

***p*-值 = *CHIDIST*(K_0^2, 自由度)**

例1.5.2　假定 $X \sim \chi^2(30)$

则概率 P(X ＞ 51.234)

= *CHIDIST*(51.234, 30) = 0.0092

t 分布的 p-值

*EXCEL*函数： *TDIST*

① 如果 $t_0 > 0$， p-值 = *TDIST*(t_0, 自由度, 1)

② 如果 $t_0 < 0$， p-值 = *TDIST*(ABS(t_0), 自由度, 1)

TDIST(t_0, 自由度, 2) 表示双边概率

例1.5.3　假定 $X \sim t(20)$

则概率 $P(X > 3.456)$ 或 $P(X \leq -3.456)$

$= TDIST(3.456, 20, 1) = 0.00125$

而概率 $P(|X| > 3.456)$

$= TDIST(3.456, 20, 2) = 0.00250$

F 分布的 *p*-值

EXCEL*函数： *FDIST

***p*-值 = *FDIST*(F_0, 分子自由度, 分母自由度)**

例1.5.4　假定 $X \sim F(21, 45)$

则概率 $P(X > 4.5678)$

$$= FDIST(4.5678, 21, 45) = 9.61377 \times 10^{-6}$$

二项分布的概率

EXCEL*函数： *BINOMDIST

***BINOMDIST* (成功次数,试验次数,成功概率,1或0)。**

1表示累积分布，0表示当前概率。

例1.5.5　抛均匀硬币500次，

正面恰好出现220次的概率是

$BINOMDIST\,(220,500,0.5,0)= 0.000973081$

正面最多220次(或至少280次)的概率是

$BINOMDIST\,(220,500,0.5,1)= 0.00413$

练习1.5.6

抛均匀骰子20次，一个六点也没有掷出的概率？如果抛300次，一点最多出现65次的概率？