目 录

第1章　概率的基本概念

汇总统计量

引理 1.1. ***//引理内容

推论 1.2. ***//推论内容

从数学角度，社会和自然现象可以划分为两类

• 确定现象(deterministic phenomenon)：在相同的条件下，其结果总是确定不变的。

• 随机现象(random phenomenon)：在相同的条件下，其结果未必相同，可能发生改变。

就像一枚硬币有两个面一样，随机现象也有两面性，一面是其偶然性，其发生存在随机性和偶然性，但是另一个面是必然性，大量的偶然性中存在必然的规律性，称之为统计规律性。概论和统计就是用来研究随机现象的数学工具，

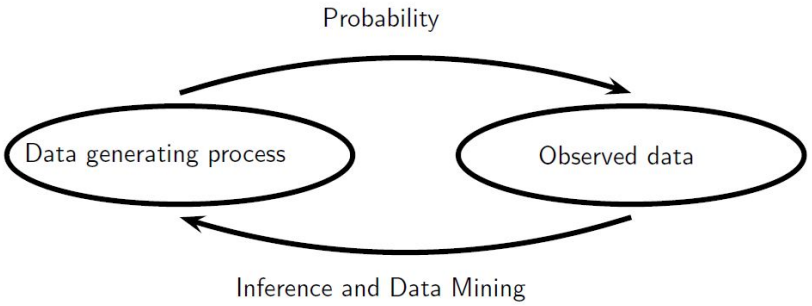

图 1.1 概率论和统计的区别

概论与统计推断是一个问题的两个方面

• 概论的基本问题：给定一个数据生成过程，研究输出的结果具有何种属性和规律，即根据概率分布或者概率密度，研究其常用统计量的性质。

• 统计的基本问题：与概论相反，给定输出的数据，研究和刻画生成数据的过程是什么，即根据数据，计算其统计量，并进一步估计其概率模型参数。

1.1 统计(Statistic)

Statistics（统计学）源自State（国家），原意是估计收集的国情资料。

高尔顿定律:高尔顿(F. Galton)于1889年在研究人类身高的亲子关系时发现的生物数量性状的“回归现象”，即平均来说，子代的表型值比亲代更接近于群体的平均值

clinical trial（临床实验）

统计学的目标在于基于从感兴趣的总体抽得的样本的测量信息，对该总体作出推断。

数据收集方案的设计，数据的概括和统计分析，解释研究结果

统计划分为两类

- 描述性统计
- 数理统计

数理统计是以数学和概率论为基础，研究

1. 如何采用有效的方式收集或者获取带有随机性的数据（调查和实验）
2. 在给定的模型（统计模型）下如何有效地使用和分析数据（统计分析）
3. 对所研究的问题进行推断（估计和检验）

数据随机性的来源

- 测量或者实验存在随机的误差，从而带来不确定性
- 总体的数据量很大或者获取总体的成本很高，无法全部加以研究，只能抽取部分数据或者实验部分情况，使得所得数据与实际数据之间存在偏差

获取数据的两种方式

- 调查（survay），抽样理论。不能控制条件活荷载因素，可以设计抽样方法
- 实验（experiment），实验设计。能够设计和控制各种影响结果的条件或因素

有效

- 建立一个在数学上可以处理并尽可能简单方便的模型来描述所得数据
- 数据要有代表性，包含尽可能多的、与所研究的问题有关的信息

估计（estimate）

- 点估计
- 区间估计

检验（test）

- 参数检验
- 非参数检验

总体（population）又称为统计总体，是指与所研究的问题有关的对象的全体所构成的集合，记为X。

总体中的每个对象称为个体。

样本（Sample）又称为样品或者子样，是指通过观测或者实验从总体中获取的一部分个体，从而得到的数据$X_1, \cdots, X_n$，这些数据的全体$\mathbf{X} = (X_1, \cdots, X_n)$就称为样本，其中n称为样本大小（sample size），又称为样本容量。数据$X_1, \cdots, X_n$中的每一个$X_i (1 \leq i \leq n)$在不引起混淆的情况下也称为样本。

从总体中按照一定规则抽取出一部分个体的行为，称为抽样（Sampling）

需要说明的是，个体所对应的记录可赙不止一个实数，可能是二维，甚至多维。

若某个集合$\mathscr{X}$包含了一切可能的样本值，则称$\mathscr{X}$为样本空间。

样本具有两重性

- 抽样方案实施之前：样本视为随机变量$\mathbf{X} = (X_1, \cdots, X_n)$
- 抽样方案实施之后：样本就是具体数值$\mathbf{x} = (x_1, \cdots, x_n)$，被称为样本值（观察值）

样本分布受到抽样方式的影响

- 有放回抽样
- 无放回抽样

一个问题的统计模型是指研究该问题时所抽取样本的分布，也被称为概率模型或者数学模型。

统计量（statistic）是为了刻画总体某个特征，而对于样本的一种加工。统计量是样本的函数，并且是一个完全有样本决定的量。

类似于样本，统计量也有两重特性。在没有抽样之前，统计量是一个随机变量，而抽样之后，即在给定样本$X_1, \cdots, X_n$情况，统计量$Y = g(X_1, \cdots, X_n)$是

一个确定的数，不含任何未知数。如果含有未知数，那么就不是统计量。

设$X_1, \cdots, X_n$为从某总体中抽出的样本。称

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.1}$$

为样本均值。称

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1.2}$$

为样本方差。称

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k \tag{1.3}$$

为k阶样本中心矩。称

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{1.4}$$

为k阶样本原点矩。

样本中位数

样本极差R

样本$X_1, \cdots, X_n$独立同分布，则$\prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$称为抽样分布,其中$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$

1.2 概率(Probability)的定义

概论是随机事件发生大小可能性的数学表征，在本质上概率是一个将随机事件映射到[0,1]区间的函数。

1.2.1 古典概论

古典概率定义：有限和等概论的结果（基本事件）。实验有许多可能的结果，每个结果叫做一个基本事件。

- 有限性：存在有限个基本事件
- 等概率：基本事件的概率相等

如果基本事件的个数为n，则事件A的概率定义为P(A)=m/n,其中m为事件A包含的基本事件个数。

古典概论的计算主要基于排列组合。基本的组合分析原理

- 乘法原理：事件分解多个步骤，每个步骤为一个子事件
- 加法原理：事件划分为不相交的子事件。

古典概论大部分问题都采用摸球模型/盒子模型来描述。

盒子模型（按箱分配质点或者小球）：将r个球放入n个不同的盒子中，并且$r \leq n$

- 球可分辨，每盒至多一球 P_n^r
- 球可分辨，每盒球数不限 n^r
- 球不可辨，每盒至多一球 $\binom{n}{r}$
- 球不可辨，每盒球数不限 $\binom{n+r-1}{r}$

摸球模型：从n个球依次地取出r个

- 球可分辨（球有序）v.t.球不可辨（球无序）
- 球不放回v.t.球有放回

表 1.1 是否可辨识（有序）与是否放回的组合

	有放回	不放回
可分辨(有序）	n^r	P_n^r
不可辨(无序）	$\binom{n}{r}$	$\binom{n+r-1}{r}$

其中

$$P_n^r = n(n-1)(n-2)\cdots(n-r+1)$$

$$\binom{n}{r} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!}$$

$\binom{n}{r}$被称为二项系数，因为是如下二项展示式的系数

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n=i}$$

probabilitycourse.com

- 不放回无序抽样（Unordered Sampling without Replacement）
- 不放回有序抽样（Ordered Sampling without Replacement）
- 放回无序样本（Unordered Sampling with Replacement）
- 放回有序样本（Ordered Sampling with Replacement）

摸球模型和盒子模型虽然描述不尽相同，但是本质上是一样的，各种情况能够一一对应。

- 有序抽样<–>可分辩

- 无序抽样<–>质点不可辩
- 放回抽样<–>箱中可容纳任意多质点
- 不放回抽样<–>箱中最多可容纳一个质点

关于二项系数的一些公式

令a=b=1，则(1.5)可以转换为

$$2^n = \sum_{i=0}^{n} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n}$$

由恒等式$(1+x)^{m+n} = (1+x)^m(1+x)^n$两边$x^k$项的系数相等，可以得到

$$\binom{m+n}{k} = \sum_{i=0}^{k} \binom{m}{i}\binom{n}{k-i}$$

根据组合的概念不难得出

$$\binom{n}{k} = \binom{n}{n-k}, 0 \le k \le n$$

例 1.1. (投球入格)设有n个球，每个球都能以相同的概率$\frac{1}{N}$落到N个格子($N \geq n$)的每一个格子中，试求：

- 某指定的n个格子中各有一个球的概率
- 任何n个格子中各有一个球的概率

生日问题

抽签与顺序无关

1.2.2 几何概论

几何概率，基于等长度/等面积/等体积，等概论，通过长度、面积和体积计算出来的概论。几何概率突破了古典概率事件有限性的约束。

若事件A_g可以等价为在区域Ω中随机地取一点，而该点落在区域g中，则其概论等于

$$P(A_g) = \frac{\text{g的测度}}{\Omega\text{的测度}}$$

1777年法国科学家蒲丰（Buffon， 1701～1788）提出了著名的蒲丰投针问题（Buffon's Needle Problem）

在平面上画着许多间距为a的平行线，取长度为l(l<a)的针，随机地投掷到此平面，则针与平线中任意一条相交的概论为$p = \frac{2l}{\pi a}$。如果投掷n次，针与直线相交m次。则可以通过p可以估计$\pi = \frac{2nl}{ma}$。

蒲丰投针问题开创了随机模拟法的先河。蒙特卡罗(Monte Carlo)方法。

贝特朗悖论（Bertrand's paradox）:在一给定圆内所有的弦中任选一条弦，求该弦的长度长于圆的内接正三角形边长的概率。

采用不同的等可能性假设，会导致同一事件有不同概率，因此为悖论。1899年贝特朗在巴黎出版《概率论》中对于几何概论提出了这个悖论，用于批评几何概论。

1.2.3 概论的统计定义

一事件出现的可能性大小，应由在多次重复实验中其出现的频繁程度去刻画。为此，提出提出了概率的统计定义。

定义 1.1. 对于随机事件A，若在n次独立重复的实验中出现m次，则称

$$F_n(A) = \frac{m}{n}$$

为随机事件A发生的频率，若n无限增大时，则F_n在某个值p附近摆动，则称事件A的概率P(A)=p

频率只是概论的估计而非概论本身。概论就是当实验次数无限增大时频率的极限。

1.2.4 概论的公理化定义

概率的公理化定义：定义概率必须满足的一般性质。

事件与集合能够建立一一对应关系

- 所有可能的结果<==>样本空间(sample space) Ω
- 基本事件或者实验结果<==>样本空间的元素或者样本点(sample point)
- 事件(Events)<==> 集合
- 事件的运算<==>集合的运算
- 事件的概论<==>集合的测度

1933年前苏联数学家柯尔莫哥洛夫（Andrei N. Kolmogorov，1903～1987）建立了概率论的公理化结构。

样本点(sample point）ω对应于随机实验的结果，而所有的样本点构成样本空间(sample space) Ω，

定义 1.2. 若$\mathscr{F}$是由样本空间Ω的一些子集构成的一个σ域，则称其为事件域(Event Field)，$\mathscr{F}$中的元素成为事件， Ω称为必然事件，$\varnothing$称为不可能事件。

定义 1.3. 定义在事件域$\mathscr{F}$上的一个集合函数P，如果其满足如下三个条件，就称其为概率

1. 非负性，对于一切$A \in \mathscr{F}$，$P(A) \leq 0$；
2. 规范性，$P(\Omega) = 1$；
3. 可列可加性，若$A_i \in \mathscr{F}(i = 1, 2, \cdots)$两两互不相容(相交)，则

$$P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

$(\Omega, \mathscr{F}, P)$称为概率空间，其中Ω是样本空间，$\mathscr{F}$是事件域，P是概率。

Galton Board

定理 1.3. 若干个互斥事件之和的概率，等于各事件的概率之和

$$P(A_1 + A_2 + \cdots) = P(A_1) + P(A_2) + \cdots \tag{1.5}$$

定理1.3也被称为概率加法定理

1.3 条件概率(Conditional Probability)

条件概率就是在附加一定条件后所得到的概率。例如，事件B发生后，对于事件A的概率可能产生如下影响

1. 有利于A的发生，即使得A的概率变大
2. 不利于A的发生，即使得A的概率变小
3. 不对A产生影响，即A的概率不会变化

定义 1.4 (Conditional Probability)**.** 设$(\Omega, \mathscr{F}, P)$为一个概率空间，$A \in \mathscr{F}$和$B \in \mathscr{F}$，并且$P(B) \neq 0$，则在给定B发生的条件下A的条件概率记为$P(A|B)$，定义为

$$P(A|B) = \frac{P(AB)}{P(B)}$$

根据定义1.4，可以得到

$$P(AB) = P(B)P(A|B)$$

可以把上式推广到任意n个事件的情况

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2|A_1)P(A_2|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$$

例 1.2 (波利亚坛子模型（Pólya urn model)). 坛子有a个白球和b个黑球，每次从罐中随机抽取一球，观察其颜色后，该球连同附加的c个同色球都被放回到坛子中，如此往复共抽取n次，问前面的n_1次出现黑球，后面的$n_2 = n - n_1$次出现红球的概率是多少?

解. □

注意这个答案只与黑球和红球出现次数有关，而与出现的顺序无关。

定义 1.5. 设有两事件A和B，若满足$P(AB) = P(A)P(B)$，则称A和B是统计独立的，简称独立的(independent)。

推论 1.4. *若事件A与B独立，并且$P(B) > 0$，则$P(A|B) = P(A)$。*

推论 1.5. *若事件A与B独立，则$\overline{A}$与B，A与$\overline{B}$，$\overline{A}$与$\overline{B}$分别是独立的。*

定义 1.6. 设$A_1, A_2, \cdots$为有限个或者无限个事件。如果从其中任意取出有限个事件$A_{i_1}, A_{i_2}, \cdots, A_{i_m}$都成立

$$P(A_{i_1}, A_{i_2}, \cdots, A_{i_m}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_m})$$

则称$A_1, A_2, \cdots$相互独立。

例 1.3 (伯恩斯坦反例). 一个均匀的正四面体，第一面染上红色，第二面染上白色，第三面染上黑色，第四面同时染上红，白，黑三种颜色。记事件A、B和C分别表示投一次均匀的正四面体出现红、白、黑颜色的事件。

解. $P(A) = P(B) = P(C) = \frac{1}{2}$

$P(AB) = P(BC) = P(AC) = \frac{1}{4}$

$P(ABC) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C)$ □

定理 1.6. 若干个独立事件$A_1, \cdots, A_n$之积的概率，等于各个事件概率的乘积：

$$P(A_1 \cdots A_n) = P(A_1) \cdots P(A_n)$$

定理1.6被称为乘法定理，其作用与加法定理1.3一样：把复杂事件的概率的计算归结为更简单的事件概率的计算，这当然有条件：相加是互斥，相乘是独立。

[相互独立事件至少发生其一的概率的计算] 若$A_1, A_2, \cdots, A_n$是n个相互独立事件，则由于

$$\overline{A_1 \cup A_2 \cup \cdots \cup A_n} = \overline{A_1 A_2} \cdots \overline{A_n}$$

因此

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= 1 - P(\overline{A_1 A_2} \cdots \overline{A_n}) \\ &= 1 - P(\overline{A_1}) P(\overline{A_2}) \cdots P(\overline{A_n}) \end{aligned}$$

定理 1.7 (The Law of Total Probability)**.** 设$(\Omega, \mathscr{F}, P)$为一个概率空间，事件域$\mathscr{F}$中的事件$B_1, B_2, \cdots$为样本空间Ω的一个划分，即当$i \neq j$时$B_i \cap B_j = \varnothing$并且$\bigcup B_i = \Omega$，那么对于任意一事件$A$，满足

$$\begin{aligned} P(A) &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots \\ &= \sum_{i=0}^{\infty} P(B_i)P(A|B_i) \end{aligned} \tag{1.6}$$

式(1.6)被称为“全概率公式”。顾名思义，“全部”概率P(A)被分解成了许多部分之和。它的理论和实用意义在于：在较复杂的情况下直接计算P(A)非常困难，但是A总是伴随某个B_i发生，适当构造一组B_i往往可以简化计算。这个公式还可以从另一个角度去理解，把B_i看作为导致事件A发生的一种可能途径或者原因。对于不同途径或者原因，A发生的概率即条件概率$P(A|B_i)$各个不同，而采取哪个途径或者原因却是随机的，遵循概率$P(B_i)$。

定理 1.8 (Bayes' Theorem)**.** 设$(\Omega, \mathscr{F}, P)$为一个概率空间，事件域$\mathscr{F}$中的事件$B_1, B_2, \cdots$为样本空间Ω的一个划分，那么对于任意一事件A，满足

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_j P(B_j)P(A|B_j)} \tag{1.7}$$

公式(1.7)被称为贝叶斯公式，首先出现在英国学者T. 贝叶斯（1702~1761）去世后的1763年的一本著作中。贝叶斯公式之所以著名，是因为其其哲理意义的解释：先看$P(B_1)$，$P(B_2)$，$\cdots$，其是在没有近一步的信息或者条件（不知道A是否发生）的情况下，对诸事件$B_1, B_2, \cdots$发生可能性大小的认识，因此$P(B_i)$也被称为**先验概率**。现在有了新的信息（知道了A发生），人们对于$B_1, B_2, \cdots$发生可能性大小有了新的估计,$P(B_i|A)$也被称为**后验概率**。

如果我们把事件A看作“结果”，把诸事件$B_1, B_2, \cdots$看成导致这结果的可能的“原因”，则可以形象地把全概率公式看作为“由原因推结果”，而贝叶斯公式则恰好相反，其作用在于“由结果推原因”。

1.4 随机变量(Random Variables)

定义 1.7 (Random Variables)**.** 随机变量为从样本空间到实数集的映射

$$X : \Omega \to \mathbb{R}$$

该映射对每个样本点ω赋予一个实数值$X(\omega)$。

[说明]概率P是样本集合（事件）的函数，其值域是[0,1]，而随机变量是样本点(实验结果)的函数，其值域是整个实数集$\mathbb{R}$。随机变量的反面是所谓确定性变量．随机事件是从静态的观点来研究随机现象，而随机变量则是一种动态的观点，一如数学分析中的常量与变量的区分。随机变量不仅仅将样本点映射到了实数，而且能够分析概率P，从而为采用统一的分析方法研究概率问题，构建了基础。

随机变量可以划分为离散型随机变量和连续型随机变量，但是在本文中并不做区分。

定义 1.8 (Distribution Function)**.** 分布函数$F : \mathbb{R} \to [0, 1]$，其定义为

$$F(x) = P(X \leq x)$$

分布函数是一个特殊样本集合（事件）的概率。通过分布函数，可以将对于随机变量的概率计算转化为对于分布函数的数值运算。

定理 1.9. *分布函数F(x)具有下列性质*

1. 单调性：若$x_1 < x_2$时，则$F(x_1) \leq F(x_2)$。

2. 规范性：$\lim_{x\to+\infty} F(x) = 1$，$\lim_{x\to-\infty} F(x) = 1$。

3. 右连续性[1]：$F(x+0) = F(x)$

根据实变函数论中关于单调函数的一般结果，可以推出分布函数具有如下性质

1. 分布函数至多只有可列个不连续点。

2. 对于分布函数F(x)有勒贝格分解

$$F(x) = c_1F_1(x) + c_2F_2(x) + c_3F_3(x) \tag{1.8}$$

其中$F_1(x)$是跳跃函数，而$F_2(x)$是绝对连续函数，$F_3(x)$是所谓奇异函数，它们都是分布函数，并且满足$0 \leq c_i \leq 1$，$i = 1, 2, 3$,和$c_1 + c_2 + c_3 = 1$。

离散型分布函数是跳跃函数，相当于在式(1.8)中，$c_1 = 1, c_2 = c_3 = 0$的场合，而连续型分布函数是绝对连续函数，相当于在式(1.8)中，$c_2 = 1, c_1 = c_3 = 0$的场合。$c_3 = 1, c_1 = c_2 = 0$的情况对应另一类分布函数，即奇异型分布函数，它是连续函数，但却不能表示为不定积分。

定义 1.9 (Probability Density Function)**.** 设X为定义在样本空间$Omega$上的随机变量，则存在可积函数$f(x)$满足

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y$$

称$f(x)$为X的概率密度函数。

概率密度函数反映了概率在x点处的“密集程度”

定义 1.10. 设$\mathbf{X} = (X_1, X_2, \cdots, X_n)$为一个n维向量，其每个向量$X_i(1 \leq i \leq n)$都是一维随机向量，则称$\mathbf{X}$是一个n维随机向量或n维随机变量。

多维随机变量(随机向量)

定义 1.11. 称n元函数

$$F(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n)$$

为随机向量$(X_1, X_2, \cdots, X_n)$的（联合）分布函数。

[1]右连续性与分布函数的定义密切相关，如果分布函数定义为$F(x) = P(X < x)$，则其满足左连续。

可以证明n维随机变量的分布函数$F(x_1, x_2, \cdots, x_n)$具有如下性质

1. 单调性：关于每个随机变量是单调不减函数。

2. 规范性：

$$F(x_1, \cdots, -\infty, \cdots, x_n) = 0$$

$$F(+\infty, +\infty, \cdots, +\infty) = 1$$

3. 右连续性：关于每个随机变量是右连续。

联合密度函数

边缘分布(Marginal Distribution)

边缘密度函数

G是若干光滑曲线围成的区域

$$P((x, y) \in G) = \iint_{(x,y)\in G} f(x, y)\mathrm{d}x\mathrm{d}y$$

事件独立和随机变量的独立

随机变量X和Y独立是指与X有关的任一事件发生与否跟与Y有关的任一事件发生与否无关。

设有两个随机变量或向量X_1和X_2，在给定了X_2取某个或者某些值的条件下，去求X_1的条件分布。根据条件概率的定义，可以得到

$$\begin{aligned} P(X_1 \le x_1 | X_2 = x_2) &= \lim_{\Delta x \to 0} P(X_1 <= x_1 | x_2 < X_2 \le x_2 + \Delta x) \\ &= \lim_{\Delta x \to 0} \frac{P(X_1 <= x_1, x < X_2 \le x_2 + \Delta x)}{P(x_2 < X_2 \le x_2 + \Delta x)} \\ &= \lim_{\Delta x \to 0} \frac{F(x_1, x_2 + \Delta x) - F(x_1, x_2)}{F(+\infty, x_2 + \Delta x) - F(+\infty, x_2)} \end{aligned} \tag{1.9}$$

对于连续密度函数的情况

$$P(X_1 \le x_1 | X_2 = x_2) = \lim_{\Delta x \to 0} \frac{\int_{x_2}^{x_2+\Delta x} \int_{-\infty}^{x_1} f(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2}{\int_{x_2}^{x_2+\Delta x} \int_{-\infty}^{+\infty} f(u_1, u_2)\mathrm{d}u_1\mathrm{d}u_2}$$

若将上式的分子分母分别除以Δx，在对分子和分母分别取极限，则可以得到

$$P(X_1 \le x_1 | X_2 = x_2) = \frac{\int_{-\infty}^{x_1} f(u_1, x_2)\mathrm{d}u_1}{\int_{-\infty}^{+\infty} f(u_1, x_2)\mathrm{d}u_1} = \int_{-\infty}^{x_1} \frac{f(u_1, x_2)}{f_2(x_2)}\mathrm{d}u_1$$

其中

$$f_2(x_2) = \int_{-\infty}^{+\infty} f(u_1, x_2)\mathrm{d}u_1$$

为X_2的边缘概率密度。因此，在$X_2 = x_2$时X_1的条件概率密度函数为

$$f_1(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$$

由上式可得到两个随机变量X_1和X_2的联合概论密度，等于其中之一的概论密度乘以在给定这一个之下另一个的条件概论密度。

$$f(x_1, x_2) = f_2(x_2)f_1(x_1|x_2) = f_1(x_1)f_2(x_2|x_1)$$

这些公式反映的实质可以推广到任意多个随机变量的场合：设有n维随机向量$(X_1, \cdots, X_n)$，其概论密度函数为$f(x_1, \cdots, x_n)$，则

$$f(x_1, \cdots, x_n) = g(x_1, \cdots, x_k)h(x_{k+1}, \cdots, x_n|x_1, \cdots, x_k)$$

其中，g是$(X_1, \cdots, X_k)$的概论密度，h是在给定$X_1 = x_1, \cdots, X_k = x_k$的条件下，$(X_{k+1}, \cdots, X_n)$的条件概论密度。

$$f_1(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2)\mathrm{d}x_2 = \int_{-\infty}^{+\infty} f_1(x_1|x_2)f(x_2)\mathrm{d}x_2$$

上式可以看作是全概率公式在概论密度这种情况下的表现形式。

联合概论密度、边缘概论密度与条件概论密度

一般而言，$f_1(x_1|x_2)$是随着x_2的变化而变化的。

定义 1.12 (Independent Random Variables)**.** 设n维随机向量$(X_1, X_2, \cdots, X_n)$的联合密度函数为$f(x_1, x_2, \cdots, x_n)$，而X_i的边缘密度函数为$f_i(x_i)$，$i = 1, 2, \cdots, n$。如果

$$f(x_1, x_2, \cdots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n)$$

则称随机变量$X_1, X_2, \cdots, X_n$相互独立或简称独立。

考察n个事件

$$A_1 = \{a_1 \le X_1 \le b_1\}, \cdots, A_n = \{a_n \le X_n \le b_n\} \tag{1.10}$$

定理 1.10. 如果随机变量$X_1, X_2, \cdots, X_n$相互独立，则对任何$a_i < b_i, i = 1, 2, \cdots, n$，由式*(1.10)*定义的$n$个事件$A_1, A_2, \cdots, A_n$也独立。

反之，如果对任何$a_i < b_i, i = 1, 2, \cdots, n$，事件$A_1, A_2, \cdots, A_n$相互独立，则随机变量$X_1, X_2, \cdots, X_n$也独立。

定理 1.11. 若连续型随机向量$(X_1, X_2, \cdots, X_n)$的概论密度函数$f(x_1, x_2, \cdots, x_n)$可以表示为n个函数$g_1, g_2, \cdots, g_n$之积，其中g_i只依赖于x_i，即

$$f(x_1, x_2, \cdots, x_n) = g_1(x_1)g_2(x_2)\cdots g_n(x_n)$$

则$X_1, X_2, \cdots, X_n$相互独立，并且X_i的边缘密度函数$f_i(x_i)$与$g_i(x_i)$只相差一个常数因子。

定理 1.12. 若$X_1, X_2, \cdots, X_n$相互独立，而

$$Y_1 = g_1(X_1, \cdots, X_m), Y_2 = g_2(X_{m+1}, \cdots, X_n)$$

则Y_1和Y_2独立。

1.5 工具(Tools)

定义 1.13 (Quantile Function)**.** 设X为一个随机变量，其概率分布函数为F(x)，则逆分布函数或者分位数函数等于为

$$F^{-1}(q) = \inf\{x : F(x) > q\}$$

其中，$q \in [0, 1]$。如果F严格递增并且连续，则$F^{-1}(q)$是满足F(x)=q的唯一实数x。

The most commonly used such constants are measures of central tendency (mean, median, and mode), and measures of dispersion (variance and mean deviation). Two other important measures are the coefficient of skewness and the coefficient of kurtosis. The coefficient of skewness measures the degree of asymmetry of the distribution, whereas the coefficient of kurtosis measures the degree of flatness of the distribution.

生成函数法（Generating Function）

定义 1.14. 设X为随机变量，c为常数，k为正整数，则$E[(X - c)^k]$成为关于c点的k阶矩。

比较重要的有两种情况

- c=0。这时$\alpha_k = E(X^k)$称为X的k阶原点矩。
- c=E(X)。这时$\mu_k = E[(X - E(X))^k]$称为X的k阶中心矩。

如果$\mu_3 > 0$，则称分布为正偏或右偏。如果$\mu_3 < 0$，则称分布为负偏或左偏。对于正态分布，$\mu_3 = 0$。

$$\beta_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

称为X或者分布的“偏度系数”（Skewness）。

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

称为X或者分布的“峰度系数”（Kurtosis）。

尖峰厚尾

中心矩(central moment)

原点矩(moment about the origin)

矩母函数(moment generating function):The moment generating function of a random variable X is defined by

$$M_X(t) = E(e^{tX})$$

The moment generating function is useful to derive the moments of X

$$E(X^k) = \left.\frac{\partial^k E(e^{tx})}{\partial t^k}\right|_{t=0}, k = 1, 2, \cdots$$

Characteristic Function: The characteristic function of a random variable X is defined by

$$\phi_X(t) = E(e^{itX})$$

生成函数

生成函数方法也被称为母函数，其基本思想是把离散的数列同多项式或者幂级数一一对应起来，从而将离散数列间的结合关系转换为多项式或幂级数之间的运算。

有限差算子 (finite difference)，也被称为向前差分 (forward difference)是一种非常古老的数学工具了，

引理 1.13 (柯西)**.** *若f(x)是连续函数(或单调函数)，且对一切x，y（或者一切*$x \geq 0$，$y \geq 0$*成立*

$$f(x)f(y) = f(x+y)$$

则

$$f(x) = a^x$$

其中$a \geq 0$，*是某一常数。*

证. 由条件可可知，对于任意整数n以及实数x有

$$f(nx) = f(x + \cdots + x) = [f(x)]^n$$

在上式中去$x = \frac{1}{n}$，得

$$f(1) = \left[f\left(\frac{1}{n}\right)\right]^n$$

记$a = f(1) \geq 0$，则

$$f\left(\frac{1}{n}\right) = a^{\frac{1}{n}}$$

因此，对任意正整数m以及n，成立

$$f\left(\frac{m}{n}\right) = f\left(\frac{1}{n}\right)^m = a^{\frac{m}{n}}$$

任意一个有理数都可以表示为m/n的形式，因此这样已经证明对于一切有理数x，$f(x) = a^x$成立，在利用连续性或者单调性可以证明对于无理数也成立，从而证明了引理。 □

1.6 随机变量的数字特征

随机变量的概率分布是随机变量的概率性质最完整的刻画，而随机变量的数字特征则是其分布所决定的常数，它刻画了随机变量（概率分布）某一方面的性质。

取一个数来表示随机变量取值规律的一个特征，这类数字称为随机变量的数字特征。

随机变量数字特征的分类

- 位置参数：
  - 期望（均值）

– 中位数，n满足$P(x \geq n) = 1/2$和 $P(x \leq n) = 1/2$

– 众数（mode）

• 刻度参数

– 方差

– 标准差

– 极差

数学期望（均值）与中位数

定义 1.15 (Expectation)**.** 设X为概率空间$(\Omega, \mathscr{F}, P)$中的随机变量，并且有概率密度函数f(x)。如果

$$\int_{\Omega} |x| f(x) \mathrm{d}x < \infty$$

则称

$$E(X) = \int_{\Omega} x f(x) \mathrm{d}x$$

为随机变量X的数学期望

E(X)是一个数，不是随机变量。E(X)在本质上是概率密度的加权平均。

$\int_{\Omega} |x| f(x) \mathrm{d}x < \infty$条件不仅确保数学期望存在，而且保证了数学期望的唯一性。在实际中所遇到的绝大多数的分布都是存在数学期望，但是也的确存在有意义的分布，不存在数学期望，例如柯西（Cauchy）分布。

有数学期望的定义，可以推导出如下的性质。

推论 1.14. *常数的数学期望等于该常数，即若c为常数，则$E(c) = c$。*

推论 1.15. *若c为常数，X为一个随机变量，则*

$$E(X + c) = E(X) + c$$

常数c可以看作一个特殊的随机变量X，满足P(X=c)=1。

推论 1.16. *若c为常数，X为一个随机变量，则*

$$E(cX) = c \cdot E(X)$$

推论 1.17. *若X和Y为两个随机变量，则*

$$E(X + Y) = E(X) + E(Y)$$

推论 1.18. *若随机变量X和Y相互独立，则*

$$E(XY) = E(X)E(Y)$$

定理 1.19 (The rule of the lazy statistician)**.** *若随机变量X的概率密度函数为f(x)，则Y=g(X)的数学期望为*

$$E(Y) = E\left(g(x)\right) = \int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x$$

定义 1.16 (Conditional Expectation)**.** 在给定X=x的情况下，Y的条件密度函数为$f(y|x)$，则Y的条件期望为

$$E(Y|X=x) = \int_{\Omega} yf(y|x)\mathrm{d}y \tag{1.11}$$

条件期望反映了随着X取值的变化，Y的平均值变化的情况。在统计学上，常把条件期望$E(Y|X=x)$作为x的函数，称为Y对于X的“回归函数”，而“回归分析”，即关于回归函数的统计研究，构成统计学的一个重要分支。

以f(x,y)记(X,Y)的联合密度函数，则X和Y的边缘密度函数分别为

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y, \quad f_2(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x$$

则可以得到

$$\begin{aligned} E(Y) &= \int_{-\infty}^{+\infty} yf_2(y)\mathrm{d}y \\ &= \int_{-\infty}^{+\infty} y\int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} yf(x,y)\mathrm{d}y\right]\mathrm{d}x \end{aligned}$$

又由于

$$E(Y|X=x) = \int_{-\infty}^{+\infty} y\frac{f(x,y)}{f_1(x)}\mathrm{d}y = \frac{\int_{-\infty}^{+\infty} yf(x,y)\mathrm{d}y}{f_1(x)}$$

因此

$$E(Y) = \int_{-\infty}^{+\infty} E(Y|X=x)f_1(x)\mathrm{d}x = E[E(Y|X=x)] \tag{1.12}$$

式(1.12)被称为条件期望的平滑公式，此公式可适用于更为一般的情况。如果X为n维随机向量$(X_1, X_2, \cdots, X_n)$，有概率密度$f(x_1, x_2, \cdots, x_n)$，则

$$E(Y) = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} E(Y|X_1=x_1, \cdots, X_n=x_n)f(x_1, \cdots, x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n \tag{1.13}$$

定义 1.17 (Median)**.** 设连续型随机变量X的分布函数为F(x)，则满足条件

$$P(X \leq m) = F(m) = 1/2 \tag{1.14}$$

的数m称为X或者分布F的中位数。

中位数可能不唯一，而是一个区间。

定义 1.18 (Variance)**.** 设X为随机变量，分布为F，则

$$Var(X) = E(X - E(X))^2 \tag{1.15}$$

称为X（或者分布F）的方差，其平方根$\sqrt{Var(X)}$（取正值）称为X（或分布F）的标准差(Stardard Deviation)，记为$\sigma(X)$。

方差和标准差描述了随机变量对于其数学期望的偏离程度（Dispersion）。均值-方差理论。

标准差与对应的随机变量有相同的量纲，更便于应用，但方差具有较好的数学性质，更经常使用。

E(X)为常数，则由方差的定义可以得到

$$\begin{aligned} Var(X) &= E(X^2 - 2E(X)X + (E(X))^2 \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - E^2(X) \end{aligned} \tag{1.16}$$

推论 1.20. *常数的方差为0。*

推论 1.21. *若c为常数，X为一个随机变量，则*

$$Var(X + c) = Var(X)$$

推论 1.22. *若c为常数，X为一个随机变量，则*

$$Var(cX) = c^2 Var(X)$$

定理 1.23. *独立随机变量之和的方差，等于各变量的方差之和，即*

$$Var(X + Y) = Var(X) + Var(Y) \tag{1.17}$$

设X为一个随机变量，E(X)=μ而Var(X)=σ^2。如果Y=$(X-\mu)/\sigma$，则E(Y)=0，而Var(Y)=1。这样，对X作一线性变换后，得到一个具有均值0、方差1的变量Y，称为Y是X的“标准化”。

定义 1.19 (Covariance). 若随机变量X和Y的期望分别为μ_1和μ_2，则称$E[(X-\mu_1)(Y-\mu_2)]$为X和Y的协方差，记为$Cov(X,Y)$。

推论 1.24. *若a_1，b_1，a_2，b_2都为常数，则*

$$Cov(a_1X+b_1,a_2Y+b_2)=a_1a_2Cov(X,Y)$$

推论 1.25. *若a_1，b_1，a_2，b_2都为常数，则*

$$Cov(a_1X+b_1Y,a_2X+b_2Y)=\begin{pmatrix}a_1 & b_1\end{pmatrix}\begin{pmatrix}Var(X) & Cov(X,Y)\\ Cov(X,Y) & Var(Y)\end{pmatrix}\begin{pmatrix}a_2\\ b_2\end{pmatrix}$$

推论 1.26. *若随机变量X和Y的期望分别为μ_1和μ_2，则*

$$Cov(X,Y)=E(X,Y)-\mu_1\mu_2$$

定理 1.27. *若X和Y相互独立，则$Cov(X,Y)=0$。*

定理 1.28. *$[Cov(X,Y)]^2\le\sigma_1^2\sigma_2^2$。等号成立当且仅当$X$和$Y$之间有严格线性关系(即存在常数$a$，$b$使得$Y=X+bX$)时成立。*

证. 设X和Y的期望分别为μ_1和μ_2，标准差分别为σ_1和σ_2，则

$$E\left[(X-\mu_1)t+(Y-\mu_2)\right]^2=\sigma_1^2t^2+2Cov(X,Y)t+\sigma_2^2$$

由于上式为非负，因此

$$4\sigma_1^2\sigma_2^2\ge[2Cov(X,Y)]^2$$

故$[Cov(X,Y)]^2\le\sigma_1^2\sigma_2^2$成立，并且当等号成立时

$$Cov(X,Y)=\pm\sigma_1\sigma_2$$

因此

$$E\left[(X-\mu_1)t+(Y-\mu_2)\right]^2=(\sigma_1t\pm\sigma_2)^2=0$$

当$t_0=\mp\sigma_2/\sigma_1$时，

$$E\left[(X-\mu_1)t_0+(Y-\mu_2)\right]^2=0$$

这意味着

$$(X-\mu_1)t_0+(Y-\mu_2)=0$$

因而

$$Y=-t_0X+\mu_1t_0+\mu_2$$

即X和Y之间有严格线性关系。

反之，若X和Y之间有严格的线性关系Y=aX+b，则

$$Cov(X,Y)=Cov(X,aX+b)=aVar(X)$$

又因为

$$Var(Y)=Var(aX+b)=a^2Var(X)$$

因此

$$[Cov(X,Y)]^2=a^2Var(X)Var(X)=Var(Y)Var(X)$$

即等号成立。 □

协方差与随机变量X和Y的量纲有关，即不同的量纲可能会得出不同的协方差。为了消除量纲的影响，提出了相关系数。

定义 1.20 (Correlation Coefficient)**.** 若随机变量X和Y的标准差分别为σ_1和σ_2，则称

$$Cov(X,Y)/(\sigma_1\sigma_2) \tag{1.18}$$

为X和Y的相关系数，记为$Corr(X,Y)$。

定理 1.29. *若X和Y独立，则$Corr(X,Y)=0$。*

定理 1.30. *$-1\le Corr(X,Y)\le 1$，或者$|Corr(X,Y)|\le 1$，等号成立当且仅当X和Y之间有严格线性关系。*

当$Corr(X,Y)=0$（或$Cov(X,Y)=0$）时，称两个随机变量X和Y不相关。需要说明的是由$Corr(X,Y)=0$不一定有X和Y独立。当$Corr(X,Y)>0$时，称两个随机变量X和Y正相关，而当$Corr(X,Y)<0$时则称负相关。$Corr(X,Y)=1$和$Corr(X,Y)=-1$分布被称为完全正相关和完全负相关。

相关系数也常被称为“线性相关系数”。这是因为，实际上相关系数并不是刻画了X和Y之间一般关系的程度，而只是“线性”关系的程度。

还可以从最小二乘法的角度解释“线性相关”：设有两个随机变量X和Y，现在想用X的某一线性函数aX+b来逼近Y，问要选择怎样的常数a和b，才能使得逼近的程度高？

$$\begin{aligned} E\left[(Y-aX-b)^2\right] &= E\left[(Y-\mu_2)-a(X-\mu_1)-c\right]^2 \\ &= \sigma_2^2+a^2\sigma_1^2-2aCov(X,Y)+c^2 \end{aligned}$$

其中

$$c=b-(\mu_2-a\mu_1)$$

为了使得上式达到最小，必须取

$$\begin{aligned} c &= 0 \\ a &= Cov(X,Y)/\sigma_1^2=\sigma_1^{-1}\sigma_2 Corr(X,Y) \end{aligned}$$

这样求出最佳线性逼近为

$$L(X)=\mu_2-\sigma_1^{-1}\sigma_2\rho\mu_1+\sigma_1^{-1}\sigma_2\rho X \tag{1.19}$$

其中$\rho=Corr(X,Y)$。这一逼近的残差为

$$\begin{aligned} E\left[(Y-L(X))^2\right] &= \sigma_2^2+a^2\sigma_1^2-2aCov(X,Y) \\ &= \sigma_2^2+(\sigma_1^{-1}\sigma_2\rho)^2\sigma_1^2-2(\sigma_1^{-1}\sigma_2\rho)\sigma_1\sigma_2\rho \\ &= \sigma_2^2(1-\rho^2) \end{aligned} \tag{1.20}$$

如果$\rho=\pm1$，则$E\left[(Y-L(X))^2\right]=0$，而$Y=L(X)$。这时Y与X有严格线性关系。

相关系数只能刻画线性关系的程度，而不能刻画一般的函数相依关系的程度。然而，对于二维正态分布而言，相关系数是X，Y的相关性的一个完美刻画，没有上面的缺点。

1. 若（X,Y)维尔二维正态，则即使允许你用户任何函数M(X)去逼近Y，并且仍以$E[(Y-M(X))^2]$最小为准则，那么你所得到的最佳逼近仍然是有式(1.19)所得到的L(X)。故在这个场合，只须考虑线性逼近足以，而这种逼近的程度完全有相关系数决定。

2. 当(X,Y)为二维正态时，由$Corr(X,Y)=0$，能够推出X，Y独立。即在这一特殊场合，独立与不相关是一回事。

定理 1.31. *若(X,Y)有二维正态分布*$N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，*则*$Corr(X,Y)=\rho$。

证.

$$\begin{aligned}Cov(X,Y)&=E\left[(X-\mu_1)(Y-\mu_2)\right]\\&=\left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1}\iint_{-\infty}^{+\infty}(x-\mu_1)(y-\mu_2)\cdot\\&\quad\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right]\mathrm{d}x\mathrm{d}y\end{aligned}$$

因为

$$\begin{aligned}&\frac{(x-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\\=&\left(\frac{x-\mu_1}{\sigma_1}-\frac{\rho(y-\mu_2)}{\sigma_2}\right)^2+\left(\sqrt{1-\rho^2}\frac{y-\mu_2}{\sigma_2}\right)^2\end{aligned}$$

作变量替换

$$u=\frac{1}{\sqrt{1-\rho^2}}\left(\frac{x-\mu_1}{\sigma_1}-\frac{\rho(y-\mu_2)}{\sigma_2}\right),v=\frac{y-\mu_2}{\sigma_2}$$

可以将上面的重积分转化为

$$Cov(X,Y)=\frac{1}{2\pi}\iint_{-\infty}^{+\infty}\left[\sqrt{1-\rho^2}\sigma_1u+\sigma_1\rho v\right]\sigma_2v\exp\left[-\frac{u^2+v^2}{2}\right]\mathrm{d}u\mathrm{d}v$$

因为

$$\iint_{-\infty}^{+\infty}uv\exp\left(-\frac{u^2+v^2}{2}\right)\mathrm{d}u\mathrm{d}v=\int_{-\infty}^{+\infty}u\exp\left(-\frac{u^2}{2}\right)\mathrm{d}u\int_{-\infty}^{+\infty}v\exp\left(-\frac{v^2}{2}\right)\mathrm{d}v=0$$

又因为

$$\iint_{-\infty}^{+\infty}v^2\exp\left(-\frac{u^2+v^2}{2}\right)\mathrm{d}u\mathrm{d}v=\int_{-\infty}^{+\infty}\exp\left(-\frac{u^2}{2}\right)\mathrm{d}u\int_{-\infty}^{+\infty}v^2\exp\left(-\frac{v^2}{2}\right)\mathrm{d}v=2\pi$$

得到$Cov(X,Y)=\rho\sigma_1\sigma_2$。于是

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sigma_1\sigma_2}=\rho$$

得证。 □

协方差和相关系数描述了两个随机变量关系的一种度量。

1.7 大数定理和中心极限定理

如果$X_1, X_2, \cdots, X_n$是随机变量，则n个随机变量和$X_1 + X_2 + \cdots + X_n$的分布，除了个别情况外，算起来都很复杂。因而自然的会提出问题：可否利用极限的方法来进行近似计算？在很一般的情况下，和的极限分布就是正态分布。

弱大数定理

$\bar{X}_n$依概率收敛于a。

引理 1.32 (马尔可夫不等式)**.** *若Y为只取非负值的随机变量，则对任意常数$\epsilon > 0$有*

$$P(Y \geq \epsilon) \leq E(Y)/\epsilon \tag{1.21}$$

证明. 设Y为连续型变量，概率密度函数为f(y)。因为Y支取非负值，所以当y<0时，f(y)=0。故

$$E(Y) = \int_0^\infty yf(y)\mathrm{d}y \geq \int_\epsilon^\infty yf(y)\mathrm{d}y$$

因为在$[\epsilon, \infty)$内总有$y \geq \epsilon$，并且

$$\int_\epsilon^\infty yf(y)\mathrm{d}y = P(Y \geq \epsilon)$$

因此

$$E(Y) \geq \int_\epsilon^\infty yf(y)\mathrm{d}y \geq \epsilon \int_\epsilon^\infty f(y)\mathrm{d}y = \epsilon P(y \geq \epsilon)$$

即$P(Y \geq \epsilon) \leq E(Y)/\epsilon$。 □

引理 1.33 (切比雪夫不等式)**.** *若$Var(Y)$存在，则*

$$P(|Y - E(Y)| \geq \epsilon) \leq Var(Y)/\epsilon^2 \tag{1.22}$$

证明. 设$X = (Y - E(Y))^2$，根据马尔可夫不等式，可以得到$P(X \leq \epsilon^2) \leq E(X)/\epsilon^2$。又因为$P((Y - E(Y))^2 \geq \epsilon^2) = P(|Y - E(Y)| \geq \epsilon)$，因此将Y带入即可得到不等式成立。 □

定义 1.21. 若$X_1, X_2, \cdots, X_n, \cdots$为随机变量，则定义

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

定理 1.34. 设$X_1, X_2, \cdots, X_n, \cdots$是独立同分布的随机变量，若它们的均值为$\mu$，并且它们的方差存在，记为$\sigma^2$，则对任意给定的$\epsilon > 0$有

$$\lim_{n\to\infty} P\left(|\bar{X}_n - a| \geq \epsilon\right) = 0 \tag{1.23}$$

证明. 根据定义可以得到

$$E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

因此由切比雪夫不等式可得

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq Var(\bar{X}_n)/\epsilon^2 \tag{1.24}$$

又因为$X_1, X_2, \cdots, X_n$独立，有

$$Var(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

将协方差带入可得

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0, \text{当} n \to \infty$$

得证。 □

定理 1.35 (中心极限定理(Central Limit Theorem)). 设$X_1, X_2, \cdots, X_n, \cdots$为独立同分布的随机变量，$E(X_i) = \mu$，$Var(X_i) = \sigma^2$，$0 < \sigma^2 < \infty$，则对任何实数$x$，有

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{n}\sigma}(X_1 + \cdots + X_n - n\mu) \leq x\right) = \Phi(x) \tag{1.25}$$

其中$\Phi(x)$是标准正态分布*N(0,1)*的分布函数。

注意$X_1 + \cdots + X_n$有均值$n\mu$，方差$n\sigma^2$。故

$$\frac{1}{\sqrt{n}\sigma}(X_1 + \cdots + X_n - n\mu)$$

就是$X_1 + \cdots + X_n$的标准化，使其均值变为0和方差变为1，以符合N(0,1)。

定理1.35也被称为林德伯格定理或者林德伯格-莱维定理，是这两位学者在20世纪20年代证明了此定理。历史上最早的中心极限定理是定理1.35的一个特例。

定理 1.36. 设$X_1, X_2, \cdots, X_n, \cdots$为独立同分布的随机变量，$X_i \sim Be(p)$，则对任何实数$x$，有

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{np(1-p)}}(X_1 + \cdots + X_n - np) \le x\right) = \Phi(x) \tag{1.26}$$

定理1.36称为棣莫弗-拉普拉斯定理，是历史上最早的中心极限定理。1716年棣莫弗讨论了p=1/2的情况，而拉普拉斯则把其推广到一般p的情况。

第2章　常用的概率分布

2.1　伯努利分布(Bernoulli Distribution)

如果事件域$\mathscr{F} = \{\Omega, A, \overline{A}, \varnothing\}$，则可以称出现A为“成功”，出现$\overline{A}$为“失败”。这种只有两种可能结果的实验称为伯努利试验(Bernoulli Experiment)。若随机变量X，在成功时取值1，概率为p，而失败时取值0，概率为1=p，则称X服从伯努利分布或者两点分布，记作$X \sim Be(p)$，其概率密度可以表示为

$$f(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{其他} \end{cases} \tag{2.1}$$

虽然伯努利试验简单，但是可以由伯努利试验生成或者衍生出很多其他的分布。若每隔Δt进行一次实验，则伯努利实验也可以看作一个随时间而变化的过程。

在伯努利实验中，到时刻$n\Delta t$为止，共进行n次实验，其中成功次数服从二项分布。而在泊松过程中，到时刻t为止发生的事件数目服从泊松分布。

为等待第一次成功，伯努利实验中的等待事件服从几何分布，而泊松过程中则服从指数分布，并且几何分布和指数分布都有无记忆。

为等待第r次成功，伯努利实验中的等待时间服从负二项分布，而在泊松过程则服从埃尔朗分布。

推论 2.1. *若随机变量$X \sim Be(p)$，则$E(X) = p$，$Var(X) = p(1-p)$。*

证.

$$E(X) = 0 \times (1-p) + 1 \times p = p$$

$$Var(X) = E(X^2) - (E(X))^2 = 0^2 \times (1-p) + 1^2 \times p - p^2 = p(1-p)$$

□

显然，当$p = \frac{1}{2}$时，方差最大。

2.2 几何分布(Geometric Distribution)

如果在伯努利试验中前X-1次皆失败，而第X次成功，也就是说，直到试验第一次成功为止的实验次数为X。显然，X为随机变量，并且当X=k时的概率为

$$g(k;p) = q^{k-1}p, k = 1, 2, \cdots$$

其中q=1-p。g(k;p)是几何级数$p, pq, pq^2, \cdots$的一般项，所以这个概率被称为几何分布，记为$X \sim G(p)$。

[**几何分布的无记忆性**]在伯努利实验中，等待首次成功的次数X服从几何分布。假设已知在前m次实验中没有出现成功，那么为了达到首次成功所需要等待的实验次数Y的概率分布为

$$\begin{aligned} P(Y = k) &= P(X = m + k | X > m) \\ &= \frac{P(X = m + k, X > m)}{P(X > m)} = \frac{P(X = m + k)}{P(X > m)} = \frac{q^{m+k-1}p}{q^m} \\ &= q^{k-1}p, \qquad k = 1, 2, \cdots \end{aligned}$$

还是服从几何分布，并且与之前失败次数m无关。

在离散型分布中，只有几何分布才具有无记忆性。

定理 2.2. *若X为只取正整数值的随机变量，在已知X>k的情况下，X=k+1的概率为p，并且与k无关，那么$X \sim G(p)$。*

证. 记$q_k = P(X > k)$和$p_k = P(X = k)$，那么$p_{k+1} = q_k - q_{k+1}$，并且在已知$X > k$的条件下，X=k+1的条件概率$P(X = k + 1 | X > k) = \frac{p_{k+1}}{q_k}$。因此

$$\frac{p_{k+1}}{q_k} = p$$

即

$$\frac{q_{k+1}}{q_k} = 1 - p$$

注意到q_0=1，那么$q_k = (1 - p)^k$，因此

$$p_k = (1 - p)^{k-1}p, \qquad k = 1, 2, \cdots$$

显然，上式为几何分布。 □

2.3 二项分布(Binomial Distribution)

二项分布(Binomial Distribution)记作$X \sim B(n,p)$，用于描述在相同条件下独立地重复n次伯努利试验中成功的次数X。二项分布是由n次伯努利试验组成的随机现象，必须满足以下四个条件

1. 重复进行n次随机试验；
2. 各次试验是相互独立；
3. 每次试验仅有两个可能结果，分别标记为0或者1；
4. 各次实验的条件是稳定的，每次试验中对应1和0的概率分别为概率为p和为1-p。

当X=m时的概率概率密度函数为

$$b(m;n,p) = Pr(X = m) = \binom{n}{m} p^m (1-p)^{n-m}$$

其中

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

设$Z_1, Z_2, \cdots, Z_n$是独立同分布的伯努利随机变量Be(p)。如果

$$X \triangleq Z_1 + Z_2 + \cdots + Z_n$$

则$X \sim B(n,p)$，即二项分布$B(n,p)$可以表示为n个独立的贝努力分布$Be(p)$之和。

之所以被称为二项分布，是因为其概论与二项式公式（二项式定理）具有相似的形式。

$$(x+y)^n = \sum_{m=0}^{n} \binom{n}{m} x^m y^{n-m}$$

[有放回抽样]有A和B两种不同的产品，数量分别为a和b，有放回的抽取n次，把可能的排列全体作为样本，总数为$(a+b)^n$，其中A样品的次数为m的数目为$\binom{n}{k} a^k b^{n-m}$，则正好有m个A产品的概率为

$$b(m) = \frac{\binom{n}{m} a^m b^{n-m}}{(a+b)^n} = \binom{n}{m} \left(\frac{a}{a+b}\right)^m \left(\frac{b}{a+b}\right)^{n-m}$$

如果设$p = \frac{a}{a+b}$，则上式可以简化为

$$b(m) = \binom{n}{m} p^m (1-p)^{n-m}$$

显示，上式即为标准的二项分布密度函数。

[直线上的随机游动]考虑x轴上一个质点，假定它只能位于整数点。在时刻t=0，其处于位置为a（a是整数），以后每隔单位时间，它总受到一个外力的随机作用，使位置发生变化，分别以概率p及概率q=1-p向正的或负的方向移动一个单位。我们关心的是质点在时刻t=n的位置。

若质点可以在整个数轴的整数点上移动，则称这种随机游动为无限随机游动。若在某点d设有一个吸收壁，质点一旦到达这个点即被吸收而不再游动，因此整个游动也就结束了，这种随机游动称为在d点有吸收壁的随机移动。此外，还可以考虑带有反射壁及弹性壁的随机游动。在一个随机游动中还可以具有不止一个壁。

两端有吸收壁的随机游动与概率论发展史上有名的赌徒输光问题有密切联系。

[赌徒输光问题]甲乙进行赌博，其赌本分别为a和b。若每局赌注为1，而甲乙在每局中赢的概率分别为p和1-p。试求甲（或者乙）把赌本输光的概率。

推论 2.3. *若$X \sim B(n,p)$，则$E(X)=np$，Var(X)=np(1-p)。*

证. 如果随机变量$X \sim B(n,p)$，则X可以表示为n个伯努利分布之和，即

$$X = Y_1 + Y_2 + \cdots + Y_n, \quad Y_i \sim Be(p), i = 1, 2, \cdots, n$$

因此

$$E(X) = E(Y_1) + E(Y_2) + \cdots + E(Y_n) = p + p + \cdots + p = np$$

$$Var(X) = Var(Y_1) + Var(Y_2) + \cdots + Var(Y_n) = np(1-p)$$

□

2.4 负二项分布(Negative Binomial Distribution)

负二项分布(Negative Binomial Distribution)也被称作帕斯卡分布(Pascal Distribution)。假设一系列独立的、同分布的伯努利实验，每次实验成功和失败的概率分别是p和1-p。实验将会一直重复下去，直到成功了r次。定义全部实验过程中失败的次数为随机变量X，则X服从负二项分布，记为$X \sim NB(r,p)$。考察$\{X=i\}$这个事件为使这个事件发生，需要同时发生如下两个事件

1. 在前i+r-1次实验中，恰恰成功r-1次；

2. 第i+r次实验成功；

这个两个事件的概论分别b(r-1;i+r-1,p)和p，由独立性可得

$$nb(i;r,p) = b(r-1;i+r-1,p)p = \binom{i+r-1}{r-1}p^r(1-p)^i, i = 0,1,2,\cdots$$

显然，当r=1时，负二项分布退化为几何分布，也就是说，几何分布是负二项分布当r=1时的特例。

之所以被称为负二项分布，是因为其概率类似于负指数二项展开式的系数。

$$\begin{aligned}(1-x)^{-r} &= \sum_{k=0}^{\infty}\binom{-r}{i}(-x)^k \\ &= \sum_{k=0}^{\infty}\frac{-r(-r-1)(-r-2)\cdots(-r-k+1)}{k!}(-x)^k \\ &= (-1)^{2k}\frac{r(r+1)\cdots(r+k-1)}{k!}x^k \\ &= \sum_{k=0}^{\infty}\binom{r+k-1}{k}x^k \\ &= \sum_{k=0}^{\infty}\binom{r+k-1}{r-1}x^k\end{aligned}$$

例 2.1. [巴拿赫火柴盒问题]某数学家有两盒火柴，每盒都有n根火柴，每次用火柴时他随机地在两盒中任取一盒并从中抽出一根。求该数学家用完一盒时另一盒还有r根火柴的概率？

例 2.2. [分赌注问题]甲乙两个赌徒按照某种方式赌博，约定先胜t局者赢得全部赌注，在其中的某一时刻，两赌徒不得不中止赌博，此时甲胜r局，乙B胜s局，并且r<s<t，应该如何合理分配赌注?

2.5 泊松分布(Poisson Distribution)

泊松分布(Poisson distribution)记作$P(\lambda)$，主要用于适合于描述单位时间（或空间）内随机事件发生的次数，其满足如下三个条件

1. 平稳性。事件发生概率是稳定的，在$[t_0, t_0+t)$中发生的概率与时间间隔长度t有关，而与时间起点t_0无关。

2. 独立增长。事件发生的概率相互独立且互不影响，即在$[t_0, t_0 + t)$中发生的概率与时刻t_0之前发生的事件无关。

3. 小概率性。在很小的事件范围内发生两次或两次以上事件的概率趋于0。

如果随机变量X服从Poisson分布，则事件$\{X = k\}$的概率为

$$p(k;\lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

为了方便起见，设所观察的这段时间为$[0, 1)$，取一个很大的自然数N，把时间段$[0, 1)$划分为等长的N段：

$$l_1 = \left[0, \frac{1}{N}\right), l_2 = \left[\frac{1}{N}, \frac{2}{N}\right), \cdots, l_N = \left[\frac{N-1}{N}, 1\right)$$

做如下两个假定

1. 在每段l_i内，恰发生一个事件的概率近似地与这段时间的长度1/N 成正比，即可取为λ/N。当N很大时， l_i很小，在l_i这一段时间内要发生两次或者更多次事件的概率是0。因此在l_i这段时间内不发生事故的概率为$1 - \lambda/N$。

2. $l_1 \cdots l_N$各段时间范围内是否发生时间是相互独立的

把在$[0, 1)$整段时间内所发生的事件数记为X，视作在N个划分之后的小时段内有事件的时段数，则按照上述两个假定，X应服从二项分布。于是

$$P(X = n) = \binom{N}{n}\left(\frac{\lambda}{N}\right)^n\left(1 - \frac{\lambda}{N}\right)^{N-n}$$

当$N \to \infty$时

$$\frac{\binom{N}{n}}{N^n} \to \frac{1}{n!}, \quad \left(1 - \frac{\lambda}{N}\right)^{N-n} \to e^{-\lambda}$$

$$P(X = n) = \frac{e^{-\lambda}\lambda^n}{n!}$$

从上述推导可以看出：泊松分布可作为二项分布的极限而得到。

Possion还可以从如下方式推导而来。设X_t为在时间间隔(0,t]发生的事件次数，则X_t为非负整数$\{0, 1, 2, \cdots\}$。对于任意非负整数k，定义$g(k, t) = P(X_t = k)$

1. $g(1, h) = \lambda h + o(h)$，$\lambda$为常数并且$\lambda > 0$

2. $\sum_{k=2}^{\infty} g(t, h) = o(h)$

3. 在不重叠时间间隔你发生的事件数相互独立

其中$o(h)$表示h的高阶无穷小，即满足

$$\lim_{h\to 0}\frac{o(h)}{h}=0$$

取$g(k,t)$的边界条件$g(0,0)=1$是合理的。在事件间隔(0,t+h]没有发生事件（k=0）当且仅当在两个时间间隔(0,t]和(t,t+h]都没有事件发生。在时间间隔(t,t+h]没有发生事件的概率为1− $=\lambda h+o(h)$。由于在两个互不重叠时间间隔发生的事件次数下相互独立，从而可以得到

$$g(0,t+h)=g(0,t)[1-\lambda h+o(h)]$$

进一步可以得到

$$\frac{g(0,t+h)-g(0,t)}{h}=-\lambda g(0,t)+\frac{g(0,t)o(h)}{h}\to -\lambda g(0,t)当h\to 0$$

因此，$g(0,t)$满足微分方程

$$\frac{d_t g(0,t)}{g(0,t)}=-\lambda$$

进而可以求得$g(0,t)$

$$\log g(0,t)=-\lambda t+c或g(0,t)=ce^{-\lambda t}$$

根据边界条件$g(0,0)=1$，可以得到

$$g(0,t)=e^{-\lambda t}$$

$$g(k+1,t+h)=g(k+1,t)[1-\lambda h+o(h)]+g(k,t)[\lambda h+o(h)]$$

可以得到

$$\frac{g(k+1,t+h)-g(k+1,t)}{h}=-\lambda g(k+1,t)+\lambda g(k,t)+[g(k+1,t)+g(k,t)]\frac{o(h)}{h}$$

因此可以得到微分方差

$$\frac{\mathrm{d}}{\mathrm{d}t}g(k+1,t)=-\lambda g(k+1,t)+\lambda g(k,t)$$

推论 2.4. *若$X\sim P(\lambda)$，则$E(X)=\lambda$，Var(X)=λ。*

证. 根据期望的定义直接得到

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} k \cdot e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda \end{aligned}$$

用类似的方法可求

$$\begin{aligned} E[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1) e^{\lambda} \frac{\lambda^k}{k!} \\ &= e^{\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \\ &= \lambda^2 \end{aligned}$$

故

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 = E(X^2) - \lambda^2 \\ &= E(X) + \lambda^2 - \lambda^2 \\ &= \lambda \end{aligned}$$

□

2.6 超几何分布(Hypergeometric Distribution)

超几何分布(Hypergeometric Distribution)记为$H(M,N,n)$，其概率表示为

$$h(m;n,N,M) = \binom{M}{m}\binom{N-M}{n-m} \Big/ \binom{N}{n}$$

其中m的取值范围为

$$\max(0, M+n-N) \le m \le \max(M,n)$$

[不放回抽样]有N个样品，可以划分为两类，分别为类型一和类型二，对应的数量分别为M和N-M，那么以不放回方式从中抽出n个样品，其中包含m个类型一的概论。

$$P(X=m) = \binom{M}{m}\binom{N-M}{n-m} \Big/ \binom{N}{n}$$

当n固定，M/N=p固定，当$N \to \infty$时，X近似地服从二项分布B(n,p)，也就是说，当产品总数很大而抽样的样品数不大时，采用有放回抽样（二项分布）和不放回抽样（超几何分布），差别不大。

估算某鱼塘中鱼的条数N。设第一次从鱼塘中捕得M条鱼，做了标记之后放回鱼塘中。经过适当的时间后，第二次从鱼塘中捕n条鱼，而其中有标记的有m条，则根据M、n和m估计N的值。

根据超几何分布，第二次捕到k条有标记与的概率为

$$p_m(N) = \binom{M}{m}\binom{N-M}{n-m} \Big/ \binom{N}{n}$$

$p_m(N)$可以看作未知参数N的函数，称为似然函数（likelihood function），通过求其最大值而得到N的估计，即所谓的最大似然估计发。由于

$$\frac{p_m(N)}{p_m(N-1)} = \frac{(N-M)(N-n)}{(M-M-n+m)N} = \frac{N^2-(M+n)N+Mn}{N^2-(M+n)N+mN} = \rho$$

当$mN < Mn$时，$\rho > 1$，而当当$mN > Mn$时，$\rho < 1$。因此，当

$$\hat{N} = \left[\frac{Mn}{m}\right] \tag{2.2}$$

2.7 均匀分布(Uniform Distribution)

设随机变量X有概论密度函数

$$f(x) = \begin{cases} 1/(b-a), \text{when } a \le x \le b \\ 0, \quad \text{else} \end{cases}$$

称为X服从区间[a,b]上的均匀分布，其中a，b都是常数并且$-\infty < a < b < +\infty$。均匀分布记为$X \sim U(a,b)$。

均匀分布的概论分布函数为

$$F(x) = \begin{cases} 0, & x \le a \\ (x-a)/(b-a), & a < x < b \\ 1, & x \ge b \end{cases}$$

若随机变量X的分布函数为F(x)，则根据定义1.13，设$F^{-1}(y)$为其分位数函数。

随机变量$Y = F(X)$，若F(x)是连续函数，对于$0 \le y \le 1$,

$$P(Y \le y) = P(F(X) \le y) = P\left(X \le F^{-1}(y)\right) = F\left(F^{-1}(y)\right) = y \tag{2.3}$$

也就是说$Y = F(X)$服从[0,1]均匀分布。

若随机变量Y服从[0,1]均匀分布，则对于任意$y \in [0,1]$成立

$$P(Y \le y) = y$$

若F(x)为任意一个随机变量的分布函数，令$X = F^{-1}(Y)$，则

$$P(X \le x) = P\left(F^{-1}(Y) \le x\right) = P\left(Y \le F(x)\right) = F(x) \tag{2.4}$$

因此，X是服从分布函数F(x)的随机变量。

通过式(2.4)，我们根据[0,1]均匀分布的随机变量的样本，生成分布函数为F(x)的随机变量的样本。

推论 2.5. 若随机变量$X \sim U(a,b)$，则$E(X) = \frac{b+a}{2}$，$Var(X) = \frac{(b-a)^2}{12}$。

证.

$$E(X) = \int_a^b x\frac{1}{b-a}\mathrm{d}x == \frac{b+a}{2}$$

$$E(X^2) = \int_a^b x^2\frac{1}{b-a}\mathrm{d}x == \frac{b^2+ab+a^2}{3}$$

因而

$$Var(X) = E(X^2) - (E(X))^2 = \frac{(b-a)^2}{12}$$

□

2.8 指数分布(Exponential Distribution)

若随机变量X具有如下的概论密度函数，则称其服从指数分布(Exponential Distribution，并记为$X \sim EXP(\lambda)$

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

指数分布的概论分布函数

$$F(x)=\begin{cases}1-e^{-\lambda x}, x\geq 0\\ 0, \quad x<0\end{cases}$$

指数分布最常见的一个场合是寿命分布，例如电子元件的寿命、电话的通话时间和伺机系统的服务时间等。如果用随机变量X表示上述例子中的寿命、通话时间和服务时间，则X的分布为指数分布。

指数分布的一个重要特征是无记忆性（Memoryless Property），也被称为遗失记忆性，用概论表达为

$$P(X>x+h|X>x)=P(X>h)\quad x>0,h>0$$

由于$P(X>x)=1-F(x)=e^{-\lambda x}$，并且

$$\{X>x\}\cap\{X\geq x+h\}=\{X\geq x+h\}$$

则可以得到

$$P(X>x+h|X>x)=\frac{P(X\geq x+h)}{P(X>x)}=\frac{e^{-\lambda(x+h)}}{e^{-\lambda x}}=e^{-\lambda h}=P(X>h)$$

指数分布是唯一具有无记忆性的连续型分布。设分布函数为F(x)，记

$$G(x)=P(X>x)$$

则无记忆可以表示为如下关系

$$G(x+h)=G(x)G(h),x\geq 0,h\geq 0$$

由引理1.13可知

$$G(x)=a^x,\quad x\geq 0$$

又因为G(x)关于x单调减小，根据G(x)的定义，可以知0<a<1，因此$\ln a<0$。设$\lambda=-\ln a$，则

$$F(x)=1-G(x)=1-e^{-\lambda x},\quad x\geq 0$$

问题得证。

失效率(Failure Rate)是指在某一个时刻x尚未失效的产品，在随后单位时间T内发生失效的概论，即从x时刻开始单位时间T内失效的产品数与工作到x+T时刻尚未失效的产品数之比，即失效率曲线。

指数概率分布从技术上要假设“无老化”，也就是说：元件在时刻x尚能正常工作的条件下，其失效率中保持为某个常数$\lambda > 0$，与x无关。失效率定义为单位长度时间内失效的概论。用条件概论的形式，上述假定可表示为

$$\lim_{h\to 0}\frac{P(x \le X \le x+h|X>x)}{h} = \lambda,$$

由于

$$\{X > x\} \cap \{x \le X \le x+h\} = \{x < X \le x+h\}$$

则

$$\begin{aligned}\lim_{h\to 0}\frac{P(x \le X \le x+h|X>x)}{h} &= \lim_{h\to 0}\frac{P(x < X \le x+h)}{h(1-F(x))} \\ &= \lim_{h\to 0}\frac{\dfrac{P(x < X \le x+h)}{h}}{1-F(x)} \\ &= \lim_{h\to 0}\frac{\dfrac{F(x+h)-F(x)}{h}}{1-F(x)} \\ &= \frac{F'(x)}{1-F(x)} = \lambda\end{aligned}$$

这个微分方程的通解为$F(x) = 1-Ce^{-\lambda x}$，其中x>0。常数C可以由初始条件F(0)=0计算得出为1。

由上式可知，在指数分布中λ即为失效率。失效率越高，平均寿命越短，而λ^{-1}就是平均寿命。

在泊松分布和指数分布之间存在有趣的联系。泊松分布刻画了单位时间内独立随机事件出现或者发生的次数，而指数分布描述了这些随机事件发生的时间间隔。假设一个泊松过程，在单位时间内以平均速率λ发生事件，那么在每个单位时间内平均会发生λ个事件。在时间间隔x内平均发生λx次，因此在时间间隔x内事件发生次数K服从泊松分布

$$P(K=k) = \frac{e^{-\lambda x}(\lambda x)^k}{k!}$$

采用如下定义

- K_t：截至到时刻t为止所发生的事件总次数。
- X_t：在时刻t发生事件后再次发生事件的间隔。

根据定义，显然在时刻t发生一个事件后，在时间区间(t,t+x]中没有事件发生等价于在时刻t和t+x所发生的事件个数没有变化，即如下两个事件等价

$$\{X_t > x\} \equiv \{K_t = K_{t+x}\}$$

进一步可以得到

$$P(X_t \leq x) = 1 - P(X_t > x) = 1 - P(K_{t+x} - K_t = 0)$$

K_t满足泊松分布，因此

$$P(K_{t+x} - K_t = 0) = P(K_x - K_0 = 0)$$

又由于$K_0 = 0$，因此

$$P(X_t \leq x) = 1 - P(K_x = 0) = 1 - e^{-\lambda x}$$

$P(K_x = 0)$即为在时间间隔x内没有事件发生的概率，可得最终结论

$$P(X_t \leq x) = 1 - e^{-\lambda x}$$

指数分布用来表示泊松过程中独立随机事件发生的时间间隔，即泊松过程第k次随机事件与第k+1 次随机事件出现的时间间隔服从指数分布。

推论 2.6. 若随机变量$X \sim EXP(\lambda)$，则$E(X) = \frac{1}{\lambda}$，$Var(X) = \frac{1}{\lambda^2}$。

证.

$$\begin{aligned} E(X) &= \int_0^{\infty} x\lambda e^{-\lambda x} \mathrm{d}x \\ &= -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x} \mathrm{d}x \\ &= \frac{1}{\lambda} \end{aligned} \tag{2.5}$$

又因为

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2\lambda e^{-\lambda x} \mathrm{d}x \\ &= -x^2 e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x} \mathrm{d}x \\ &= \frac{2}{\lambda}\int_0^{\infty} x\lambda e^{-\lambda x} \mathrm{d}x \\ &= \frac{2}{\lambda^2} \end{aligned}$$

因此

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{1}{\lambda^2}$$

□

2.9 威布尔分布(Weibull Distribution)

指数分布描述了无老化情况下的寿命分布，即在失效率是一个常值λ，其与x无关。但是实际中很多产品存在老化现象，也就是说，随着寿命的增加，失效率也会随之上升。如果降失效率设置为λx^m，其中$\lambda > 0$，$m > 0$。在这种条件下，寿命分布F(x)满足微分方程

$$\frac{F'}{1 - F(x)} = \lambda x^m$$

结合初始条件F(0)=0，可以得到

$$F(x) = 1 - e^{-(\lambda/m+1)x^{m+1}}$$

取$\alpha = m + 1$，$\lambda/(m + 1) = 1/\beta$，并且把x记为$x - \delta$，其中$\alpha > 1, \beta > 0, \delta \geq 0$，则可以得到

$$F(x) = \begin{cases} 1 - \exp\left(-\dfrac{(x-\delta)^\alpha}{\beta}\right), & x \geq \delta \\ 0, & x < \delta \end{cases} \tag{2.6}$$

X的密度函数为

$$f(x) = \begin{cases} \dfrac{\alpha}{\beta}(x-\delta)^{\alpha-1} \exp\left(-\dfrac{(x-\delta)^\alpha}{\beta}\right), & x \geq \delta \\ 0, & x < \delta \end{cases} \tag{2.7}$$

若随机变量X满足上述分布密度，则称X服从威布尔分布(Weibull Distribution)或者韦布尔分布，记为$X \sim W(\alpha, \beta, \delta)$。

威布尔分布与很多分布都有关系。如，当a=1，它是指数分布；a=2且时，是瑞利分布（Rayleigh Distribution）。

2.10 埃尔朗分布(Erlang Distribution)

若$X(t)$是参数为λt的泊松过程，以W_r记它的第r个事件发生的时刻。事件$\{W_r < t\}$代表第r个事件发生在时刻t之前，事件$\{X(t) \geq r\}$代表时刻t已经发生r个

事件，因此

$$\{W_r < t\} = \{X(t) \geq r\}$$

W_r的分布函数可以表示为

$$\begin{aligned} F(t) &= P(W_r < t) = P(X(t) \geq r) \\ &= \sum_{k=r}^{\infty} \frac{(\lambda t)^k e^{-\lambda t}}{k!} \\ &= 1 - \sum_{k=0}^{r-1} \frac{(\lambda t)^k e^{-\lambda t}}{k!} \end{aligned} \tag{2.8}$$

相应的概率密度函数为

$$\begin{aligned} f(t) &= F'(t) \\ &= -\left[\sum_{k=0}^{r-1} \frac{(\lambda t)^k e^{-\lambda t}(-\lambda)}{k!} + \sum_{k=0}^{r-1} \frac{k(\lambda t)^{k-1}(\lambda) e^{-\lambda t}}{k!}\right] \\ &= \lambda e^{-\lambda t} \sum_{k=0}^{r-1} \frac{(\lambda t)^k}{k!} - \lambda e^{-\lambda t} \sum_{k=1}^{r-1} \frac{(\lambda t)^{k-1}}{(k-1)!} \\ &= \frac{\lambda(\lambda t)^{r-1}}{(r-1)!} e^{-\lambda t} \\ &= \frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t} \end{aligned} \tag{2.9}$$

对于任意正整数r以及实数$\lambda > 0$

$$\int_0^{\infty} \frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t} \mathrm{d}x = 1 \tag{2.10}$$

因此，式(2.9)是一个密度函数，被称为埃尔朗（Erlang）分布。

泊松过程中第r个事件发生的时刻W_r服从埃尔朗分布。当r=1时，埃尔朗分布简化为指数分布。若记

$$\begin{aligned} \tau_1 &= W_1 \\ \tau_r &= W_r - W_{r-1}, \quad r = 2, 3, \cdots \end{aligned}$$

则τ_1表示泊松过程第r-1个事件与第r个事件之间的间隔。通过事件间隔，可以将W_r表示为

$$W_r = \tau_1 + \tau_2 + \cdots + \tau_r \tag{2.11}$$

可以证明$\tau_1, \tau_2, \cdots, \tau_r$均服从参数为$\lambda$的指数分布，且相互独立。

2.11 正态分布(Normal Distribution)

若随机变量X的概率分布密度函数为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<+\infty \tag{2.12}$$

则称X服从正态分布(Normal Distribution)，记为$X\sim N(\mu,\sigma^2)$。正态分布有时也被称为高斯分布(Gaussian Distribution)，在台湾地区也被称为常态分布。当$\mu=0,\sigma=1$时， N(0, 1)被称为标准正态分布，其概论分布函数记为$\Phi(x)$，而对应的概论密度函数记为$\phi(x)$。

正态分布之所以被称为高斯分布，是因为正态分布最初是由德国著名数学家高斯在研究误差理论时发现的。然而，在此之前就已经有了与正态密度函数相关的初步研究。

正态分布的初步形式最早是由棣莫弗（Abraham de Moivre，1667-1754）在1733年引入，作为p为1/2而n很大情况下二项分布计算的渐进公式。拉普拉斯在1812年发表的《分析概率论》（Theorie Analytique des Probabilites）中对棣莫佛的结论作了扩展，发展成系统的理论，现在这一结论通常被称为棣莫佛－拉普拉斯定理。但是将它作为一个分布来研究则归功于高斯（Gauss，1777-1855），他在19世纪初在研究测量误差函数，给出了严格的证明，后被高尔顿命名为正态分布。这项研究又是当代统计学中重要思想——最大似然法的源头。

如果$X_n\sim B(n,p)$，则根据$E(X_n)=np$，$Varr(X_n)=np(1-p)$，可以得到

$$E\left(\frac{X_n}{n}\right)=p,\quad Varr\left(\frac{X_n}{n}\right)=\frac{p(1-p)}{n}$$

因此当$n\to\infty$时，频率$\frac{X_n}{n}$的数学期望保持不变，而方差趋于0，也就是说其为常数p。

在测量中，若μ为真值，x_i为观察值，而误差$x_i-\mu$的分布密度函数为$p(x_i-\mu)$。经验表明$p(x)$关于x=0对称，而且对于一切x有p(x)>0。为推导方便起见，还假设p(x)具有连续导函数。

如果有独立同分布的观察值$x_1,x_2,\cdots,x_n$，则其似然函数为

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

其表征了这组观察值落在μ附件的可能性的大小。高斯的假定是：观察值的平均值$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)$作为未知参数$\mu$的估值使得$L(\mu)$达到最大。

若$\bar{x}$使似然函数$L(\mu)$达到最大值，则

$$\frac{\mathrm{d}\ln L(\mu)}{\mathrm{d}\mu}\Big|_{\mu=\bar{x}}=0 \tag{2.13}$$

记$y=x-\mu$和$g(y)=\dfrac{\mathrm{d}\ln p(y)}{\mathrm{d}y}$，则$g(y)=\dfrac{p'(y)}{p(y)}$，由假设知道它好定义而且是连续函数。这是式(2.13)变成

$$\sum_{i=1}^{n}g(x_i-\bar{x})=0$$

又因为

$$\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=0$$

显然对于一切实数x，y成立

$$g(x)+g(y)=g(x+y)$$

这是柯西函数方程，很容易证明其解必为$g(y)=by$。

事实上，若记$f(y)=e^{g(y)}$，则方程(2.14)转化为

$$f(x)f(y)=f(x+y)$$

这方程对于一切x，y成立，且$f(y)$是连续函数，因此由引理1.13可知$f(y)=a^y,a\geq 0$，从而可得$g(y)=by$。

因此

$$\ln p(y)=\frac{b}{2}y^2+c$$

即

$$p(y)=e^{\frac{b}{2}y^2+c},-\infty<x<+\infty$$

$p(y)$为密度函数，因此$b<0$，记$b=-\frac{1}{\sigma^2}$，则

$$p(y)=Ke^{-\frac{y^2}{2\sigma^2}},-\infty<y<+\infty$$

有概率密度函数的规范化条件$\int_{-\infty}^{+\infty}p(y)\mathrm{d}y=1$，设$x=t\sigma$,则可以得到

$$K\sigma\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}\mathrm{d}t=1$$

考虑

$$\left[\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\mathrm{d}t\right]^2 = \int_{-\infty}^{+\infty} e^{-x^2/2}\mathrm{d}x \int_{-\infty}^{+\infty} e^{-y^2/2}\mathrm{d}y = \iint_{-\infty}^{+\infty} e^{-(x^2+y^2)/2}\mathrm{d}x\mathrm{d}y$$

转化成极坐标$x = r\cos\theta$，$y = r\sin\theta$

$$\left[\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}\mathrm{d}t\right]^2 = \int_0^{2\pi}\int_{-\infty}^{+\infty} e^{-r^2} r\mathrm{d}r\mathrm{d}\theta = \int_0^{2\pi}\left[\int_0^{+\infty} e^{-r^2}\mathrm{d}r^2\right]\mathrm{d}\theta = 2\pi$$

可以得到$K = \frac{1}{\sqrt{2\pi}\sigma}$，故

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}}, -\infty < y < +\infty$$

将$x - \mu$带入可以得到

$$p(x-\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$$

正态分布是最常见的一种分布。一般来说，若影响某一个数量指标的随机因素很多，这些因素为加性的并且每个因素所起的作用又不大，则这个指标服从正态分布。

It appears both as the limit of additive small effects and as a representation of symmetric phenomena without long tails

推论 2.7. 若$X \sim N(\mu, \sigma^2)$，则$Y = (X-\mu)/\sigma \sim N(0,1)$。

$$\begin{aligned} P(Y \le x) &= P((X-\mu)/\sigma \le x) \\ &= P(X \le \mu + \sigma x) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\mu+\sigma x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right)\mathrm{d}u \end{aligned}$$

推论 2.8. 若$X \sim N(\mu, \sigma^2)$，则$E(X)=\mu$，$Var(X)=\sigma^2$。

证. 作变量替换$x = \mu + \sigma t$，化为

$$\begin{aligned} E(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\mu + \sigma t) e^{-t^2/2}\mathrm{d}t \\ &= \mu\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-t^2/2}\mathrm{d}t + \sigma\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} t e^{-t^2/2}\mathrm{d}t \end{aligned}$$

上式右边第一项为μ，第二项为0。因此

$$E(X) = \mu \tag{2.14}$$

由方差的定义可以得到

$$Var(X) = E(X-\mu)^2 = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(x-\mu)^2\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x$$

作变量替换$x = \mu + \sigma t$，得

$$Var(X) = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t^2\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t = \frac{2\sigma^2}{\sqrt{2\pi}}\int_{0}^{+\infty}t^2\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t$$

又由于

$$\begin{aligned}\int_0^{+\infty}t^2\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t &= -\int_0^{+\infty}t\exp\left(-\frac{t^2}{2}\right)\mathrm{d}\left(-\frac{t^2}{2}\right)\\ &= -t\exp\left(-\frac{t^2}{2}\right)\Big|_0^{+\infty} + \int_0^{+\infty}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t\\ &= \frac{1}{2}\int_{-\infty}^{+\infty}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t = \frac{\sqrt{2\pi}}{2}\end{aligned}$$

因此

$$Var(X) = \sigma^2 \tag{2.15}$$

□

二维正态分布记为$N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，概论密度函数有形式

$$\begin{aligned}f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\Bigg\{-\frac{1}{2(1-\rho^2)}\Bigg[&\frac{(x_1-\mu_1)^2}{\sigma_1^2} -\\ &\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\Bigg]\Bigg\}\end{aligned} \tag{2.16}$$

令$\boldsymbol{\Sigma} = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$（$\boldsymbol{\Sigma}$为正定矩阵），则二维正态分布密度函数可以表示如下更加简约的形式

$$f(x_1,x_2) = \frac{1}{2\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

其中

$$\mathbf{x} = \begin{pmatrix}x_1\\ x_2\end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$$

由公式(2.16)作变量替换

$$u=(1-\rho^2)^{-1/2}\frac{x_1-\mu_1}{\sigma_1},v=(1-\rho^2)^{-1/2}\frac{x_2-\mu_2}{\sigma_2}$$

得

$$\iint_{-\infty}^{+\infty}f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_1=\frac{1}{2\pi}\sqrt{1-\rho^2}\iint_{-\infty}^{+\infty}\exp\left(-\frac{u^2-2\rho uv+v^2}{2}\right)\mathrm{d}u\mathrm{d}v$$

再作变量替换$t_1=u-\rho v$，$t_2=\sqrt{1-\rho^2}v$。注意到

$$u^2-2\rho uv+v^2=(u-\rho v)^2+(1-\rho^2)v^2=t_1^2+t_2^2$$

对应的雅克比行列式为

$$\begin{vmatrix}\partial t_1/\partial u & \partial t_1/\partial v\\ \partial t_2/\partial u & \partial t_2/\partial v\end{vmatrix}=\begin{vmatrix}1 & -\rho\\ 0 & \sqrt{1-\rho^2}\end{vmatrix}=\sqrt{1-\rho^2}$$

从而得到

$$\begin{aligned}&\iint_{-\infty}^{+\infty}f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_1\\&=\frac{1}{2\pi}\sqrt{1-\rho^2}\left(\sqrt{1-\rho^2}\right)^{-1}\iint_{-\infty}^{+\infty}\exp\left(\frac{t_1^2+t_2^2}{2}\right)\mathrm{d}t_1\mathrm{d}t_1\\&=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-t_1^2/2}\mathrm{d}t_1\int_{-\infty}^{+\infty}e^{-t_2^2/2}\mathrm{d}t_2\\&=\frac{1}{2\pi}\sqrt{2\pi}\sqrt{2\pi}=1\end{aligned}$$

对于二维正态分布$N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$可以通过配方法，表示为两个一维正态分布函数乘积的形式，其中一个一维正态分布函数仅仅与一个随机变量X_1（或者X_2）有关，而另一个一维正态分布函数的μ依赖于随机变量X_1（或者X_2）。

式(2.16)的指数部分可以分解为

$$\begin{aligned}&\frac{(x_1-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\\=&(1-\rho^2)\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\left(\frac{x_2-\mu_2}{\sigma_2}-\rho\frac{x_1-\mu_1}{\sigma_1}\right)^2\\=&(1-\rho^2)\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{\left[x_2-\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right)\right]^2}{\sigma_2^2}\end{aligned}$$

因此，二维正态分布密度密度可以分解为

$$f(x_1,x_2)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right)\times\frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{(x_2-\mu_2')^2}{2\sigma_2^2(1-\rho^2)}\right)\quad(2.17)$$

其中

$$\mu_2' = \mu_2 + \rho\sigma_1^{-1}\sigma_2(x_1 - \mu_1)$$

由于x_1和x_2的对称性，类似地可以得到

$$f(x_1, x_2) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right) \times \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left(-\frac{(x_1-\mu_1')^2}{2\sigma_1^2(1-\rho^2)}\right) \quad (2.18)$$

其中

$$\mu_1' = \mu_1 + \rho\sigma_1\sigma_2^{-1}(x_2 - \mu_2)$$

式(2.17)右边为两个正态分布密度函数的乘积，第一部分为$N(\mu_1, \sigma_1^2)$的密度函数，而第二部分为$N\left(\mu_2 + \rho\sigma_1^{-1}\sigma_2(x_1-\mu_1), (1-\rho^2)\sigma_2^2\right)$的密度函数。

类似于式(2.17)，在式(2.18)中第一部分为$N(\mu_2, \sigma_2^2)$的密度函数，而第二部分为$N\left(\mu_1 + \rho\sigma_1\sigma_2^{-1}(x_2-\mu_2), (1-\rho^2)\sigma_2^2\right)$的密度函数。

在式(2.17)中μ_2'与变量x_2无关，因此X_1的边缘密度函数为

$$\begin{aligned} f_1(x_1) &= \int_{-\infty}^{+\infty} f(x_1, x_2)\mathrm{d}x_2 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right) \times \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{(x_2-\mu_2')^2}{2\sigma_2^2(1-\rho^2)}\right)\mathrm{d}x_2 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right) \end{aligned}$$

类似地从式(2.18)，可以得到X_2的边缘密度函数为

$$f_2(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right)$$

因此，若(X_1, X_2)为二维正态分布$N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则$X_1$和$X_2$的边缘分布分别是一维正态分别$N(\mu_1, \sigma_1^2)$和$N(\mu_2, \sigma_2^2)$。也就是说，二维正态分布的边缘分布仍然为正态分布。显然，由二维正态分布的联合分布可以唯一的确定边缘分布，反之却不成立，即无法由边缘分布确定二维正态分布，因为边缘分布丢失了ρ。

若(X_1, X_2)为二维正态分布$N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，在给定$X_1 = x_1$的条件下，$X_2$的条件密度函数

$$\begin{aligned} f_2(x_2|x_1) &= \frac{f(x_1, x_2)}{f_1(x_1)} \\ &= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{(x_2 - (\mu_2 + \rho\sigma_1^{-1}\sigma_2(x_1-\mu_1)))^2}{2(1-\rho^2)\sigma_2^2}\right] \end{aligned}$$

这是正态分布$N(\mu_2+\rho\sigma_1^{-1}\sigma_2(x_1-\mu_1),\sigma_2^2(1-\rho^2))$的概论密度函数。在二维正态分布中，以一个随机变量为条件的分布仍为正态分布。因此，条件期望为

$$E(X_2|X_1=x_1)=\mu_2+\rho\sigma_1^{-1}\sigma_2(x_1-\mu_1)$$

显然，此条件期望是x_1的线性函数。如果$\rho>0$，则$E(X_2|X_1=x_1)$随着x_1的增加而增加，即X_2的“均值”有随着x_1增长而增长的趋势，因此被称为“正相关”。相反地，若$\rho<0$，被称为负相关。

当且仅当$\rho=0$时，联合概率密度函数$f(x_1,x_2)$才可以表示为两个边缘密度$f_1(x_1)$和$f_2(x_2)$之积。因此，当且仅当$\rho=0$时，X_1和X_2独立。

$$\begin{aligned}X_1,X_2 \text{ are independent} &\Leftrightarrow f(x_1,x_2)=f_1(x_1)f_2(x_2)\\&\Leftrightarrow \sqrt{1-\rho^2}=1,\rho=0\\&\Leftrightarrow \rho=0\end{aligned}$$

2.12 Γ分布(Gamma distribution)

Γ函数（读作Gamma函数）的定义

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}\mathrm{d}t,x>0$$

由于

$$\Gamma(x+1)=\int_0^\infty t^x e^{-t}\mathrm{d}t=-t^x e^{-t}\Big|_0^\infty+x\int_0^\infty t^{x-1}e^{-t}\mathrm{d}t=x\Gamma(x)$$

Γ函数有如下的递推公示

$$\Gamma(x+1)=x\Gamma(x)$$

如果x=1，那么可以直接计算$\Gamma(1)=\int_0^\infty e^{-t}dt=1$。因此，如果n是一个正整数，则

$$\Gamma(n)=(n-1)!$$

也就是说，Γ函数可以看作是阶乘在实数和复数域的推广，或者看作一个光滑曲线并且能够连接所有满足(n,(n-1)!)的所有整数。因此Γ函数有时也被称为阶乘函数(Factorial Function)。

$$\Gamma(1/2)=\int_0^\infty t^{-1/2}e^{-t}\mathrm{d}t$$

在做变量替换$t = u^2$后，可以得到

$$\begin{aligned}\Gamma(1/2) &= \int_0^{\infty} e^{-u^2} u^{-1} (2u\mathrm{d}u) \\ &= 2\int_0^{\infty} e^{-u^2}\mathrm{d}u = \int_{-\infty}^{+\infty} e^{-u^2}\mathrm{d}u\end{aligned}$$

令$u = v/\sqrt{2}$

$$\Gamma(1/2) = \frac{1}{\sqrt{2}}\int_{-\infty}^{+\infty} e^{-v^2/2}\mathrm{d}v = \frac{1}{\sqrt{2}}\sqrt{2\pi} = \sqrt{\pi}$$

定理 2.9 (Bohr-Mullerup)**.** 如果$f : (0,\infty) \to (0,\infty)$,并且满足

1. $f(1) = 1$
2. $f(x+1) = xf(x)$
3. $\log f(x)$是凸函数

，则$f(x) = \Gamma(x)$，即$\Gamma(x)$是唯一满足上述条件的函数

对于Γ函数稍加变形，即可以得到Γ分布

$$\int_0^{\infty} \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}dx = 1$$

简单形式的Γ分布

$$f(x;\alpha) = \begin{cases}\dfrac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & ,x > 0 \\ 0 & ,elsewhere\end{cases} \tag{2.19}$$

针对上式做一个简单变换$x = \beta t$，就可以得到更一般形式的Γ分布

$$f(x;\alpha,\beta) = \begin{cases}\dfrac{\beta^{\alpha}t^{\alpha-1}e^{-\beta t}}{\Gamma(\alpha)} & ,x > 0 \\ 0 & ,elsewhere\end{cases} \tag{2.20}$$

其中

- α称为shape parameter，主要决定分布曲线的形状
- β称为rate parameter或者inverse scale parameter（$\frac{1}{\beta}$称为scale parameter），主要决定曲线有多陡峭

上述一般形式的Γ分布记为$\Gamma(\alpha,\beta)$。

在简单形式的Γ分布中，取$\alpha = k+1$，可得

$$\Gamma(x|\alpha = k+1) = \frac{x^k e^{-x}}{\Gamma(k+1)}$$

Γ分布是Poisson分布在正实数集上的连续化版本

2.13 β分布(Beta distribution)

β函数（读作Beta函数）

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t, x>0, y>0$$

β函数具有对称性

$$B(x,y)=\beta(y,x)$$

β函数与Γ函数具有如下关系

$$B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

根据定义，两个独立的Γ函数的乘积可以写作 β函数与Γ函数具有如下关系

$$\Gamma(x)\Gamma(y)=\int_0^\infty u^{x-1}e^{-u}\mathrm{d}u\int_0^\infty v^{y-1}e^{-v}\mathrm{d}v=\int_0^\infty\int_0^\infty u^{x-1}v^{y-1}e^{-(u+v)}\mathrm{d}u\mathrm{d}v$$

应用替换$u=zt$和$v=z(1-t)$，可以将上式进一步转化为

$$\Gamma(x)\Gamma(y)=\int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t\int_0^\infty z^{x+y-1}e^{-z}\mathrm{d}z$$

使用β函数与Γ函数可得

$$\Gamma(x)\Gamma(y)=B(x,y)\Gamma(x+y)$$

如果m，n为正整数，则

$$B(m,n)=\frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}=\frac{(m-1)!(n-1)!}{(m+n-1)!}$$

若随机变量X的分布密度函数为($\alpha>0$，$\beta>0$)

$$be(x;\alpha,\beta)=\begin{cases}\dfrac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} & 0\le x\le 1\\ 0 & other\end{cases}\tag{2.21}$$

则成X服从参数为α和β的Beta分布，记作$X\sim BE(\alpha,\beta)$。

二项式分布的随机变量$X\sim B(n,p)$满足如下一个很奇妙的恒等式

$$P(X\le k)=\frac{n!}{k!(n-k-1)!}\int_p^1 t^k(1-t)^{n-k-1}dt$$

1. $X_1, X_2, \cdots, X_n \sim$ i.i.d. $Uniform(0,1)$
2. 把这n个随机变量排序后得到的顺序统计量$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$
3. 问$X_{(k)}$的分布是多少

$$
\begin{aligned}
&P(x \le X_{(k)} \le x + \Delta x) \\
&= n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x)
\end{aligned}
$$

$$
\begin{aligned}
f(x) &= \lim_{\Delta x \to 0} \frac{P(x \le X_{(k)} \le x + \Delta x)}{\Delta x} \\
&= n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k} \\
&= \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}
\end{aligned}
$$

利用Gamma函数，可以将$f(x)$表达为

$$
f(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-x)^{n-k}
$$

取$\alpha = k, \beta = n - k + 1$ 可以得到更加简洁的Beta分布表达式

$$
f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{k-1}(1-x)^{n-k}
$$

Beta分布中的参数α, β可以理解为物理计数，被称为伪计数(pseudo-count)

1. n个随机变量$X_1, X_2, \cdots, X_n \sim$ i.i.d. $Uniform(0,1)$,排序后得到的顺序统计量，我们假设$p = X_{(k)}$
2. $Y_1, Y_2, \cdots, Y_m \sim$ i.i.d. $Uniform(0,1)$,其中有m_1个比p小，m_2个比p大
3. 问$P(p|Y_1, Y_2, \cdots, Y_m)$的分布是什么

1. $X_1, X_2, \cdots, X_n \sim$ i.i.d. $Uniform(0,1)$
2. 把这n个随机变量排序后得到的顺序统计量$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$
3. 问$(X_{(k_1)}, X_{(k_1+k_2)})$的联合分布是多少

假设$k_1 - 1$个随机变量在$[0, x_1)$之间,$X_{(k_1)}$在$[x_1, x_1 + \Delta x)$之间,$k_2 - 1$个随机变量在$[x_1, x_1 + x_2)$之间,$X_{(k_2)}$在$[x_1 + x_2, x_1 + x_2 + \Delta x)$之间,$n - k_2 - k_1$个随机变量在$[x_1 + x_2 + \Delta x)$fi1)之间。

为了保持公式的简洁对称，取参数x_3，并满足$x_1+x_2+x_3=1$

$$P\left(X_{(k_1)}\in(x_1,x_1+\Delta x),X_{(k_1+k_2)}\in(x_2,x_2+\Delta x)\right)$$
$$=\frac{n!}{(k_1-1)!(k_2-1)!(n-k_1-k_2)!}x_1^{k_1-1}x_2^{k_2-1}x_3^{n-k_1-k_2}(\Delta x)^2$$

$(X_{(k_1)},X_{(k_1+k_2)})$的联合分布

$$f(x_1,x_2,x_3)=\frac{n!}{(k_1-1)!(k_2-1)!(n-k_1-k_2)!}x_1^{k_1-1}x_2^{k_2-1}x_3^{n-k_1-k_2}$$
$$=\frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}x_3^{\alpha_3-1}$$

其中，$\alpha_1=k_1,\alpha_2=k_2,\alpha_3=n-k_1-k_2+1$

Dirihlet分布是beta分布在高维度的推广

Dirichlet分布

$$f(x_1,\cdots,x_{K-1};\alpha_1,\cdots,\alpha_K)=\frac{1}{B(\alpha)}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$

$$x_1,\cdots,x_{K-1}>0$$
$$x_1+\cdots+x_{K-1}<1$$
$$x_K=1-x_1-\cdots-x_{K-1}$$

$$B(\alpha)=\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)},\alpha=(\alpha_1,\cdots,\alpha_K)$$

1. $X_1,X_2,\cdots,X_n\sim$ i.i.d. $Uniform(0,1)$,把这n个随机变量排序后得到的顺序统计量$X_{(1)},X_{(2)},\cdots,X_{(n)}$

2. 令$p_1=X_{(k_1)},p_2=X_{(k_2)},p_3=1-p_1-p_2,\overrightarrow{p}=(p_1,p_2,p_3)$

3. $Y_1,Y_2,\cdots,Y_m\sim$ i.i.d. $Uniform(0,1)$,其中Y_i落到$[0,p_1),[p_1,p_2),[p_2,1)$三个区间的个数分别为$m_1,m_2,m_3$，其中$m=m_1+m_2+m_3$

4. 问$P(\overrightarrow{p}|Y_1,Y_2,\cdots,Y_m)$的分布是什么

2.14 卡方分布(Chi-Squared Distribution)

若随机变量X的分布密度函数为(n>0)

$$k_n(x) = \begin{cases} \dfrac{1}{\Gamma\left(\dfrac{n}{2}\right) 2^{n/2}} e^{-x/2} x^{(n-2)/2}, & x > 0 \\ 0, & x \le 0 \end{cases} \tag{2.22}$$

则称X服从自由度为n的皮尔逊卡方分布(Chi-Squared Distribution)，记为$X \sim \chi_n^2$。

由$x = 2t$替换x

$$\int_0^\infty e^{-x/2} x^{(n-2)/2} \mathrm{d}x = 2^{n/2} \int_0^\infty e^{-t} t^{(n-2)/2} \mathrm{d}t = 2^{n/2} \Gamma\left(\frac{n}{2}\right)$$

因此

$$\int_0^\infty k_n(x) \mathrm{d}x = 1$$

即式(2.22)是概率密度函数。

设X有概论密度函数f(x)，如果$Y = X^2$，则

$$F(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f(t) \mathrm{d}t$$

对y求导，得到Y的概论密度函数

$$l(y) = \begin{cases} \dfrac{1}{2} y^{-1/2} \left[f(\sqrt{y}) + f(-\sqrt{y}) \right], y > 0 \\ 0, \quad y \le 0 \end{cases}$$

定理 2.10. *若$X_1, X_2, \cdots, X_n$相互独立，并且服从标准正态分布$N(0,1)$，则$Y = X_1^2 + X_2^2 \cdots, X_n^2$服从自由度$n$的卡方分布$\chi_n^2$。*

证1. 用数学归纳法证明

当n=1时，$Y = X^2$的概论密度可以表示为

$$l(y) = \frac{1}{2} y^{-1/2} \left[f(\sqrt{y}) + f(-\sqrt{y}) \right] = \frac{1}{2\sqrt{2\pi y}} \left(e^{-y/2} + e^{-y/2} \right) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}$$

注意到$\Gamma(1/2) = \sqrt{\pi}$，因此上式即为n=1时的卡方分布。

假设n-1时成立，则在n时Y可以表示为$Z + X_n^2$，其中$Z = X_1^2, \cdots, X_{n-1}^2$。由归纳假设可知，Z有概论密度函数$k_{n-1}(x)$，而$X_n^2$有概论密度函数$k_1(x)$。由定理1.12，可知Z和$X_n^2$相互独立，因此Y的概论密度函数为

$$l(y) = \int_{-\infty}^{\infty} k_{n-1}(x)k_1(y-x)\mathrm{d}x = \int_0^y k_{n-1}(x)k_1(y-x)\mathrm{d}x$$

后一式是因为$k_{n-1}(t)$和$k_1(t)$都只有在t>0时才不为0，故有效的积分区间为$0 \leq x \leq y$。带入式-2.22，可以得到

$$\begin{aligned} l(y) &= \int_0^y \left(\Gamma\left(\frac{n-1}{2}\right)2^{(n-1)/2}\right)^{-1} e^{-x/2}x^{(n-3)/2}\left(\Gamma\left(\frac{1}{2}\right)2^{1/2}\right)^{-1} e^{-(y-x)/2}(y-x)^{-1/2}\mathrm{d}x \\ &= \left(\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)2^{n/2}\right)^{-1} e^{-y/2}\int_0^y x^{(n-3)/2}(y-x)^{-1/2}\mathrm{d}x \end{aligned}$$

在积分中做变量代换x=yt，可以得到

$$\begin{aligned} \int_0^y x^{(n-3)/2}(y-x)^{-1/2}\mathrm{d}x &= y^{(n-2)/2}\int_0^1 t^{(n-3)/2}(1-t)^{-1/2}\mathrm{d}t \\ &= y^{(n-2)/2}\beta\left(\frac{n-1}{2},\frac{1}{2}\right) \\ &= y^{(n-2)/2}\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)\Big/\Gamma\left(\frac{n}{2}\right) \end{aligned}$$

以此带入上式，即得$l(y) = k_n(y)$，从而证明了结果对于n也成立。这就完成了归纳证明。 □

证2. 因为$X_1, X_2, \cdots, X_n$相互独立同分布，所以其联合概率密度函数可以表示为

$$g(x_1, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n x_i^2\right\}$$

据此，$Y = X_1^2 + X_2^2 \cdots + X_n^2$概率分函数可以表示为

$$F(y) = P(x_1^2 + \cdots + x_n^2 < y) = \frac{1}{(2\pi)^{n/2}}\int\cdots\int_{x_1^2+\cdots+x_n^2<y}\exp\left\{-\frac{1}{2}\sum_{i=1}^n x_i^2\right\}\mathrm{d}x_1\cdots\mathrm{d}x_n$$

做如下的球形变换

$$\begin{cases} x_1 = \rho\cos\theta_1\cos\theta_2\cdots\cos\theta_{n-1} \\ x_2 = \rho\cos\theta_1\cos\theta_2\cdots\sin\theta_{n-1} \\ \cdots \\ x_n = \rho\sin\theta_1 \end{cases}$$

此变换的Jacob矩阵行列式为

$$J = \frac{\partial(x_1, \cdots, x_n)}{\partial(\rho, \theta_1, \cdots, \theta_{n-1})} = \rho^{n-1} D(\theta_1, \cdots, \theta_{n-1})$$

其中$D(\theta_1, \cdots, \theta_{n-1})$与$\rho$无关。于是，

$$\begin{aligned} F(y) &= \frac{1}{(2\pi)^{n/2}} \int_0^{\sqrt{y}} e^{-\rho^2/2} \rho^{n-1} \mathrm{d}\rho \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cdots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} D(\theta_1, \cdots, \theta_{n-1}) \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{n-1} \\ &= C_n \frac{1}{(2\pi)^{n/2}} \int_0^{\sqrt{y}} e^{-\rho^2/2} \rho^{n-1} \mathrm{d}\rho \end{aligned}$$

其中

$$C_n = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cdots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} D(\theta_1, \cdots, \theta_{n-1}) \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_{n-1}$$

做变量替换$\rho = \sqrt{x}$，则

$$\begin{aligned} F(y) &= C_n \int_0^y e^{-x/2} x^{(n-1)/2} \frac{1}{2x^{1/2}} \mathrm{d}x \\ &= \frac{C_n}{2} \int_0^y e^{-x/2} x^{n/2-1} \mathrm{d}x \end{aligned}$$

又因为

$$1 = F(+\infty) = \frac{C_n}{2} \int_0^{\infty} e^{-x/2} x^{n/2-1} \mathrm{d}x \tag{2.23}$$

做变量替换$t = x/2$

$$1 = C_n 2^{n/2-1} \int_0^{\infty} e^{-t} t^{n/2-1} \mathrm{d}t = C_n 2^{n/2-1} \Gamma\left(\frac{n}{2}\right)$$

因此

$$C_n = \frac{1}{2^{n/2-1} \Gamma\left(\frac{n}{2}\right)}$$

从而得到概率分布函数

$$F(y) = \frac{1}{2^{n/2} \Gamma\left(\frac{n}{2}\right)} \int_0^y e^{-x/2} x^{n/2-1} \mathrm{d}$$

进而得到概率密度函数

$$F(y) = \frac{1}{2^{n/2} \Gamma\left(\frac{n}{2}\right)} e^{-y/2} y^{n/2-1}$$

得证。 □

自由度n：因为Y表示为n个独立变量$X_1, X_2, \cdots, X_n$的平方和，其中每个变量X_i都能随意变化，可以说它有一个自由度。这样的变量共有n个，因此有n个自由度。

命题 2.11 (卡方分布的可加性)**.** 设X_1，X_2相互独立，$X_1 \sim \chi_m^2$，$X_2 \sim \chi_n^2$，则$X_1 + X_2 \sim \chi_{m+n}^2$。

一种简单的证明是从卡方变量的表达式出发，设$Y_1, \cdots, Y_{m+n}$独立且都有分布$N(0,1)$。令$X_1 = Y_1^2 + \cdots + Y_m^2$，$X_2 = Y_{m+1}^2 + \cdots + Y_{m+n}^2$，则

$$X_1 \sim \chi_m^2, X_2 \sim \chi_n^2$$

而

$$X_1 + X_2 = Y_1^2 + \cdots + Y_{m+n}^2$$

为m+n个标准正态随机变量的平方和，因此其分布为χ_{m+n}^2。

命题 2.12. 若$X_1, X_2, \cdots, X_n$相互独立，且都服从指数分布$EP(\lambda)$，则

$$X = 2\lambda(X_1 + X_2 + \cdots + X_n) \sim \chi_{2n}^2$$

首先，由X_i的概论密度函数为$\lambda e^{-\lambda x}$，可知$2\lambda X_i$，当x>0时的概论密度函数为$\frac{1}{2}e^{-x/2}$，当$x \leq 0$时概论密度为0。在式(2.22)中令n=2，可知$\chi_2^2 = \frac{1}{2}e^{-x/2}$，因此$2\lambda X_i \sim \chi_2^2$。再由$X_1, X_2, \cdots, X_n$相互独立，利用上面的性质，即可以得出结论。

上α分位点或者分位数记为$\chi_n^2(\alpha)$。

自由度，能够自由变换的自变量个数。

推论 2.13. 若随机变量$X \sim \chi_n^2$，则$E(X)=n$，$Var(X)=2n$。

证. X可以表示为$X_1^2 + X_2^2 + \cdots + X_n^2$，其中$X_1, X_2, \cdots, X_n$相互独立并且都服从标准正态分布N(0,1)。对于任意$0 \leq i \leq n$，有

$$E(X_i^2) = Var(X_i) + E(X_i) = 1 + 0 = 1$$

因此

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

因为

$$Var(X_i^2) = E(X_i^4) - \left[E(X_i^2)\right]^2 = E(X_i^4) - 1$$

而

$$E(X_i^4) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^4 e^{-x^2/2} \mathrm{d}x = \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} x^4 e^{-x^2/2} \mathrm{d}x$$

做变量替换$x = \sqrt{2t}$，有

$$\begin{aligned} E(X_i^4) &= \frac{2}{\sqrt{2\pi}} 2\sqrt{2} \int_0^{+\infty} t^{3/2} e^{-t} \mathrm{d}t = \frac{4}{\sqrt{\pi}} \Gamma\left(\frac{5}{2}\right) \\ &= \frac{4}{\pi} \frac{3}{2} \frac{1}{2} \sqrt{\pi} = 3 \end{aligned}$$

故$Var(X_i^4) = 3 - 1 = 2$，而

$$Var(X) = \sum_{i=0}^{n} Var(X_i) = 2n$$

□

2.15 二维均匀分布

设$G \subset R^2$，$|G|$表示函数面积，若

$$(x, y) \sim f(x, y) = f(n) = \begin{cases} \frac{1}{|G|}, & (x, y) \in G \\ 0, & \text{others} \end{cases} \tag{2.24}$$

则(x,y)称为G上的均匀分布，其概率密度函数也可以表示为

$$f(x, y) = \frac{1}{|G|} I_{(x,y) \in G}$$

2.16 柯西分布(Cauchy Distribution)

若随机变量X的分布密度函数为

$$C(x; \theta, \lambda) = \frac{1}{\pi\lambda \left[1 + \left(\frac{x - \theta}{\lambda}\right)^2\right]}, \lambda > 0, -\infty < \theta < +\infty, -\infty < x < +\infty \tag{2.25}$$

则称X服从参数为θ（位置参数）和λ（尺度参数）的柯西分布，记为$X \sim C(\theta, \lambda)$。

当$\theta = 0$和$\lambda = 1$时，称为标准柯西分布。

推论 2.14. 柯西分布的均值和方差均不存在。

定理2.15. 若X_1和X_2相互独立，并且都服从标准正态分布$N(0,1)$，则$Y = X_1/|X_2|$服从标准柯西分布$C(0,1)$。

定理 2.16. 若$X_1, X_2, \cdots, X_n$相互独立，并且$X_i \sim C(\theta_i, \lambda_i), i = 1, \cdots, n$，则$X = X_1 + \cdots + X_n \sim C(\theta_1 + \cdots + \theta_n, \lambda_1 + \cdots + \lambda_n)$。

2.17 多项分布(Multinomial Distribution)

多项分布(Multinomial Distribution)是二项分布的推广。假设每次实验结果有k种不同结果，$A_1, A_2, \cdots, A_k$，其中每种结果可能发生的概率分别为$p_1, p_2, \cdots, p_k$，并且$p_1 + p_2 + \cdots + p_k = 1$。进行N 次相互独立的实验，$A_1$共发生$n_1$次，$A_2$共发生$n_2$ 次，$\cdots$，A_k共发生n_k次的概率为

$$P(n_1, \cdots, n_k | N, p_1, \cdots, p_k) = \frac{N!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i}$$

其中$N = n_1 + n_2 + \cdots + n_k$

多项分布标记为$M(N; p_1, p_2, \cdots, p_k)$，其名称的来由是因多项展开式

$$(x_1 + \cdots + x_k)^N = \sum^{*} \frac{N!}{n_1! \cdots n_k!} x_1^{n_1} \cdots x_k^{n_k}$$

$\sum^{*}$表示求和的范围为：n_i为非负整数，$n_1 + \cdots + n_k = N$

多项 B 函数

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}, \quad \alpha = (\alpha_1, \cdots, \alpha_k)$$

表达为

$$P(n_1, \cdots, n_k | N, p_1, \cdots, p_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} p_i^{n_i}$$

设$X = (X_1, X_2, \cdots, X_k)$服从多项分布$M(N; p_1, p_2, \cdots, p_k)$，就$X_1$的边缘分布。

$$P(X_1 = n) = \sum_{n_2, \cdots, n_k}^{'} \frac{N!}{n_2! \cdots n_k!} p_2^{n_2} \cdots p_k^{n_k} \cdot p_1^n / n!$$

其中$\sum'_{n_2, \cdots, n_k}$表示求和的范围为：$n_2 \cdots n_k$都是非负整数，其和为$N - n$。令

$$p_2' = p_2/(1 - p_1), \cdots, p_k' = p_k/(1 - p_1)$$

则

$$p_2' + \cdots + p_k' = (p_2 + \cdots + p_k)/(1 - p_1) = (1 - p_1)/(1 - p_1)$$

上式可以改写为

$$P(X_1 = n) = \frac{N!}{n!(N-n)!} p_1^k (1 - p_1)^{N-n} \cdot \sum_{n_2, \cdots, n_k}{}' \frac{(N-n)!}{n_2! \cdots n_k!} {p_2'}^{n_2} \cdots {p_k'}^{n_k}$$

由于

$$\sum_{n_2, \cdots, n_k}{}' \frac{(N-n)!}{n_2! \cdots n_k!} {p_2'}^{n_2} \cdots {p_k'}^{n_k} = ({p_2'}^{n_2} + \cdots + {p_k'}^{n_k})^{N-n} = 1^{N-n} = 1$$

于是得到

$$P(X_1 = n) = \frac{N!}{n!(N-n)!} p_1^k (1 - p_1)^{N-n} = b(k; N, p_1)$$

正是二项分布$B(N, p_1)$。

2.18 随机变量函数的概论分布

已知某个或者某些随机变量$X_1, \cdots, X_n$的分布，另有一些随机变量$Y_1, \cdots, Y_m$，它们都是$X_1, \cdots, X_n$的函数：

$$Y_i = g_i(X_1, \cdots, X_n), \quad i = 1, \cdots, m \tag{2.26}$$

求$Y_1, \cdots, Y_m$的概率分布。

$X_1, \cdots, X_n$是原始的观察数据或实验数据，$Y_1, \cdots, Y_m$则是为某个目的而将这些原始数据“加工”得到的量，称为“统计量”。

考虑如下单个统计量的特殊情况

$$Y = g(X_1, \cdots, X_n)$$

则Y的概率分布为

$$F_y(y) = P(Y \le y) = \int \cdots \int_{g(x_1, \cdots, x_n) \le y} f(x_1, \cdots, x_n) \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

其中$f(x_1, \cdots, x_n)$为$X_1, \cdots, X_n$的联合概率密度。如果已经得到$F_y(y)$的表达式，对于$F_y(y)$求导可以得到Y的概率密度函数。

对于求$F_y(y)$的方法大体上可以划分为直接法和变换法两类。直接法通过把$g(x_1, \cdots, x_n) \le y$转化为关于$x_1, \cdots, x_n$的等价事件而求得Y的分布函数。

命题 2.17. 设随机变量X有概论密度函数$f(x)$。设$Y=g(X)$，g是一个严格单调（单调上升或单调下降）的函数。由于g的严格单调性，其反函数$X=h(Y)$存在且h的导数h'也存在。则Y的概论密度函数为

$$l(y) = f(h(y))|h'(y)|$$

上述结论可以推广到一般的情况(2.26)

$$\begin{aligned} F(y_1, \cdots, y_m) &= P(Y_1 \le y_1, \cdots, Y_m \le y_m) \\ &= \int \cdots \int_{g_1(x_1,\cdots,x_n)\le y_1, \cdots, g_m(x_1,\cdots,x_n)\le y_m} f(x_1, \cdots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n \end{aligned}$$

如果从X到Y的映射关系具有更加良好的性质，那么上式可以进一步的简化。设$(X_1, \cdots, X_n)$有概率密度函数$f(x_1, \cdots, x_n)$，并而

$$Y_i = g_i(X_1, \cdots, X_n), i = 1, \cdots, n$$

为从$(X_1, \cdots, X_n)$到$(Y_1, \cdots, Y_n)$的一一对应变换。因此，g_i存在逆变换，记为

$$X_i = h_i(Y_1, \cdots, Y_n), i = 1, \cdots, n$$

此变换的雅克比行列式为

$$J(y_1, \cdots, y_n) = \frac{\partial(h_1, \cdots, h_n)}{\partial(y_1, \cdots, y_n)} = \begin{vmatrix} \partial h_1/\partial y_1 & \cdots & \partial h_1/\partial y_n \\ \cdots & \cdots & \cdots \\ \partial h_n/\partial y_1 & \cdots & \partial h_n/\partial y_n \end{vmatrix}$$

则$(Y_1, \cdots, Y_n)$的概率分布为

$$F(y_1, \cdots, y_n) = \int_{t_1 \le y_1} \cdots \int_{t_n \le y_n} f\left(h_1(t_1, \cdots, t_n), \cdots, h_n(t_1, \cdots, t_n)\right) \cdot |J(t_1, \cdots, t_n)| \mathrm{d}t_1 \cdots \mathrm{d}t_n$$

分布对于$y_1, \cdots, y_n$求导，得到密度函数为

$$l(y_1, \cdots, y_n) = f\left(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots, y_n)\right) \cdot |J(y_1, \cdots, y_n)|$$

2.18.1 随机变量和的密度函数

[离散卷积公式]若X和Y是相互独立的随机变量，它们都取非负整数值，其概率分布分别为a_i和b_i，则$Y = X + Y$概率分布为

$$P(Z = z) = \sum_{i=0}^{z} a_i b_{z-i} \tag{2.27}$$

命题 2.18. 随机向量$(X_1, X_2, \cdots, X_n)$服从多形式分布$M(N, p_1, p_2, \cdots, p_n), n \geq 3$，则$Y = X_1 + X_2$服从二项式分布$B(N, p1 + p2)$。

命题 2.19. 设X_1和X_2相互独立，分别服从二项分布$B(n_1, p)$和$B(n_2, p)$，则$Y = X_1 + X_2$服从二项式分布$B(n_1 + n_2, p)$。

命题 2.20. 设X_1和X_2相互独立，分别服从*Poisson*分布$P(\lambda_1)$和$P(\lambda_2)$，则$Y = X_1 + X_2$服从*Poisson*分布$P(\lambda_1 + \lambda_2)$。

设(X_1, X_2)的联合密度函数为$f(x_1, x_2)$，则$Y = X_1 + X_2$的密度函数为$l(y) = \int_{-\infty}^{\infty} f(x, y - x)\mathrm{d}x = \int_{-\infty}^{\infty} f(y - x, x)\mathrm{d}x$

$$P(y \leq y) = P(X_1 + X_2 \leq y) = \iint_B f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2$$

其中B为集合$X_1 + X2 \leq y$。先固定x_1，并对x_2积分

$$P(y \leq y) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y - x_1} f(x_1, x_2)\mathrm{d}x_2 \right) \mathrm{d}x_1$$

对y就导数，即得Y的概论密度函数

$$l(y) = \int_{-\infty}^{\infty} f(x_1, y - x_1)\mathrm{d}x_1 = f(x, y - x)\mathrm{d}x \tag{2.28}$$

如果X_1，X_2相互独立，则$f(x_1, x_2) = f_1(x_1) f_2(x_2)$，则上式可以写成

$$\begin{aligned} l(y) &= \int_{-\infty}^{\infty} f_1(x) f_2(y - x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_1(y - x) f_2(x)\mathrm{d}x \\ &= f_1 * f_2 \end{aligned}$$

其中*表示卷积。

命题 2.21. 设X_1和X_2相互独立，分别服从正态分布$N(\mu_1, \sigma_1^2)$和$N(\mu_2, \sigma_2^2)$，则$Y = X_1 + X_2$的概论密度函数为$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$。

证. 因为X_1和X_2相互独立，所以$f(x_1, x_2) = f_1(x_1) f_2(x_2)$。又由式(2.28)可以得到

$$l(y) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} \left(\frac{(x - \mu_1)^2}{\sigma_1^2} + \frac{(y - x - \mu_2)^2}{\sigma_2^2} \right) \right] \mathrm{d}x$$

在上式中指数部分可以分拆为两个部分的和

$$
\begin{aligned}
&\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-x-\mu_2)^2}{\sigma_2^2}\\
&=\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)x^2-2\left(\frac{\mu_1}{\sigma_1^2}+\frac{(y-\mu_2)}{\sigma_2^2}\right)x+\frac{\mu_1^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\\
&=\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}x^2-2\frac{\mu_1\sigma_2^2+(y-\mu_2)\sigma_1^2}{\sigma_1^2\sigma_2^2}x+\frac{[\mu_1\sigma_2^2+(y-\mu_2)\sigma_1^2]^2}{\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)}\\
&\qquad-\frac{[\mu_1\sigma_2^2+(y-\mu_2)\sigma_1^2]^2}{\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)}+\frac{\mu_1^2\sigma_2^2+(y-\mu_2)^2\sigma_1^2}{\sigma_1^2\sigma_2^2}\\
&=\left(\frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1\sigma_2}x-\frac{\mu_1\sigma_2^2+(y-\mu_2)\sigma_1^2}{\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}\right)^2-\frac{(y-\mu_2)^2\sigma_1^4+2\mu_1(y-\mu_2)\sigma_1^2\sigma_2^2+\mu_1^2\sigma_2^4}{\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)}\\
&\qquad+\frac{(y-\mu_2)^2\sigma_1^4+(y-\mu_2)^2\sigma_1^2\sigma_2^2+\mu_1^2\sigma_1^2\sigma_2^2+\mu_1^2\sigma_2^4}{\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)}\\
&=\left(\frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1\sigma_2}x-\frac{\mu_1\sigma_2^2-(y-\mu_2)\sigma_1^2}{\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}\right)^2+\frac{(y-\mu_2)^2-2\mu_1(y-\mu_2)+\mu_1^2}{\sigma_1^2+\sigma_2^2}\\
&=(ax-b)^2+\frac{(y-\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}
\end{aligned}
$$

其中

$$
\begin{aligned}
a&=\frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1\sigma_2}\\
b&=\frac{\mu_1\sigma_2^2-(y-\mu_2)\sigma_1^2}{\sigma_1\sigma_2\sqrt{\sigma_1^2+\sigma_2^2}}
\end{aligned}
$$

带入上式可以得到

$$
l(y)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[\frac{(y-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}(ax-b)^2\right)\mathrm{d}x
$$

由于a，b都与x无关，做变量代换t=ax-b，并利用$\int_{-\infty}^{\infty}e^{-t^2/2}\mathrm{d}x=\sqrt{2\pi}$，得到

$$
l(y)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[\frac{(y-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]\frac{\sqrt{2\pi}}{a}=\frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left[\frac{(y-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]
$$

这正是正态分布$N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$的密度函数。由此可见，两个独立的正态变量的和仍服从正态分布，并且其期望和方差分布为两个正态随机变量期望和方差的和。 □

这个事实的逆命题也成立：如果Y服从正态分布，而Y表成两个独立随机变量X_1与X_2之和，则X_1和X_2必都服从正态分布。这个事实称为正态分布的“再生性”，也称为正态分布再生定理（Theorem of Reproduction of Normal Distribution）。即使X_1和X_2不独立，只要其联合分布为二维正态$N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，则$Y = X_1 + X_2$仍为正态：$Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$。

命题 2.22. *若$X_1, \cdots, X_n$相互独立，分别服从正态分布$N(\mu_1, \sigma_1^2), \cdots, N(\mu_n, \sigma_n^2)$，则$X_1 + \cdots + X_n$服从正态分布$N(\mu_1 + \cdots + \mu_n, \sigma_1^2, \cdots + \sigma_n^2)$。*

2.18.2 随机变量商的密度函数

设(X_1, X_2)有密度函数$f(x_1, x_2)$，要求$Y = X_2/X_1$的密度函数。为了简单起见，限制X_1只取正值的情况。

因为$X_1 > 0$，事件$\{Y \le y\} = \{X_2/X_1 \le y\}$可以写为$\{X_2 \le X_1 y\}$。记$B = \{x_2 \le x_1 y\}$，则通过化重积分为累积分可得

$$P(Y \le y) = \iint_B f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 = \int_0^{\infty} \int_{-\infty}^{x_1 y} f(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1$$

对于y求导数，得到Y的密度函数

$$l(y) = \int_0^{\infty} x_1 f(x_1, x_1 y)\mathrm{d}x_1$$

若X_1，X_2相互独立，则$f(x_1, x_2) = f_1(x_1)f_2(x_2)$，从而上式可以表示为

$$l(y) = \int_0^{\infty} x_1 f_1(x_1) f_2(x_1 y)\mathrm{d}x_1 \tag{2.29}$$

定理 2.23. *设X_1，X_2相互独立，$X_1 \sim \chi_n^2$，$X_2 \sim N(0, 1)$，而$Y = X_2/\sqrt{X_1/n}$，则Y服从“自由度n的t分布”的密度函数，简记为$Y \sim t_n$，其密度函数为*

$$t_n(y) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}} \tag{2.30}$$

证. 记$Z = \sqrt{X_1/n}$。先求出Z的概率密度函数$g(z)$。有

$$P(Z \le z) = P(\sqrt{X_1/n} \le z) = P(X_1 \le nz^2) = \int_0^{nz^2} k_n(x)\mathrm{d}x$$

两边对z求导，得Z的密度函数为

$$g(z) = 2nzk_n(nz^2)$$

以$f_1(x_1)=2nx_1k_n(nx_1^2)$和$f_2(x_2)=(\sqrt{2\pi})^{-1}e^{-x_2^2/2}$应用于公式(2.29)，得到Y的密度函数，记为$t_n(y)$

$$\begin{aligned}t_n(y)&=\frac{1}{\sqrt{2\pi}2^{n/2}\Gamma(n/2)}\int_0^\infty 2nx_1^2e^{-nx_1^2/2}(nx_1^2)^{(n-2)/2}\times e^{-(x_1y)^2/2}\mathrm{d}x_1\\&=\frac{2n^{n/2}}{\sqrt{2\pi}2^{n/2}\Gamma(n/2)}\int_0^\infty x_1^n\exp\left[-\frac{1}{2}(nx_1^2+x_1^2y^2)\right]\mathrm{d}x_1\end{aligned}\tag{2.31}$$

作变量替换$x_1=\sqrt{2/(n+y^2)}\sqrt{t}$，上面的积分变为

$$\begin{aligned}\int_0^\infty x_1^n\exp\left[-\frac{1}{2}(nx_1^2+x_1^2y^2)\right]\mathrm{d}x_1&=\frac{1}{2}\left(\frac{2}{n+y^2}\right)^{(n+1)/2}\int_0^\infty e^{-t}t^{(n-1)/2}\mathrm{d}t\\&=\frac{1}{2}\left(\frac{2}{n+y^2}\right)^{(n+1)/2}\Gamma\left(\frac{n+1}{2}\right)\end{aligned}$$

以此带入(2.31)，并略加整理，即得$Y=X_2/\sqrt{X_1/n}$的密度函数为式(2.30)。 □

这个分布是英国统计学家W.哥色特在1908年以“student”的笔名首先发表的。

定理 2.24. 设X_1，X_2相互独立，$X_1\sim\chi_n^2$，$X_2\sim\chi_m^2$，而$Y=m^{-1}X_2/n^{-1}X_1$，则Y服从“自由度m，n的F分布”，简记为$Y\sim F_{mn}$，其密度函数为

$$f_{mn}(y)=m^{m/2}n^{n/2}\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}y^{m/2-1}(my+n)^{-(m+n)/2},y>0\tag{2.32}$$

因为X_1和X_2相互独立，则$m^{-1}X_2$和$n^{-1}X_1$也相互独立，并且概率密度可以分别表示为$nk_n(nx_1)$和$mk_m(mx_2)$。以此导入式(2.29)，得到Y的概率密度函数

$$\begin{aligned}f_{mn}(y)&=mn\int_0^\infty x_1k_n(nx_1)k_m(mx_1y)\mathrm{d}x_1\\&=mn\left[2^{m/2}\Gamma\left(\frac{m}{2}\right)2^{n/2}\Gamma\left(\frac{n}{2}\right)\right]^{-1}\cdot\int_0^\infty x_1e^{-nx_1/2}(nx_1)^{n/2-1}e^{-mx_1y/2}(mx_1y)^{m/2-1}\mathrm{d}x_1\\&=\left[2^{(m+n)/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\right]^{-1}m^{m/2}n^{n/2}y^{m/2-1}\cdot\int_0^\infty e^{(my+n)x_1/2}x_1^{(m+n)/2-1}\mathrm{d}x_1\end{aligned}$$

做变量代换$t=(my+n)x_1/2$，上式的积分分为

$$\begin{aligned}\int_0^\infty e^{(my+n)x_1/2}x_1^{(m+n)/2-1}\mathrm{d}x_1&=2^{(m+n)/2}(my+n)^{-(m+n)/2}\int_0^\infty e^{-t}t^{(m+n)/2-1}\mathrm{d}t\\&=2^{(m+n)/2}(my+n)^{-(m+n/2}\Gamma\left(\frac{m+n}{2}\right)\end{aligned}$$

以此带入上式可得

$$f_{mn}(y) = m^{m/2}n^{n/2}\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}y^{m/2-1}(my+n)^{-(m+n)/2}, y > 0 \tag{2.33}$$

当$y \leq 0$时，$f_{mn}(y) = 0$，因为Y只取正值。

式(2.32)被称为“自由度m，n的F分布”，记为$F_{m,n}$。注意分子的自由度在前。

χ^2，t和F这三个分布合称为“统计上的三大分布”。

费希尔引理

引理 2.25. 设$X_1, X_2, \cdots, X_n$ *i.i.d.*，$\sim N(\mu, \sigma^2)$。记$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，则

1. $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$
2. $\sum_{i=1}^{n}(X_i - \bar{X})^2\sigma^2 \sim \chi^2_{n-1}$
3. $\bar{X}$与$\sum_{i=1}^{n}(X_i - \bar{X})^2$独立

证. 证明正文。

□

罗纳德·艾尔默·费希尔(Ronald Aylmer Fisher, R. A. Fisher, 1890年2月17日-1962年7月29日)，英国统计学家、生物进化学家、数学家、遗传学家和优生学家。是现代统计科学的奠基人之一

命题 2.26. 设$X_1, \cdots, X_n$独立同分布，分布服从正态分布$N(\mu, \sigma^2)$。记

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

则

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1} \tag{2.34}$$

2.18.3 随机变量极值的密度函数

若$X_1, \cdots, X_n$相互独立，具有相同的分布函数$F(x)$和密度函数$f(x)$，则$Y = \min_{i=1}^{n}\{X_i\}$和$Z = \max_{i=1}^{n}\{X_i\}$的分布。

$$
\begin{aligned}
P(Y \geq y) &= P(\min_{i=1}^{n}\{X_i\} \geq y) \\
&= P(X_1 \geq y, X_2 \geq y, \cdots, X_n \geq y) \\
&= P(X_1 \geq y) \cdot P(X_2 \geq y) \cdots P(X_n \geq y) \\
&= [1 - F(y)]^n
\end{aligned}
$$

因此

$$
P(Y < y) = 1 - [1 - F(y)]^n \tag{2.35}
$$

而

$$
\begin{aligned}
P(Z < z) &= P(\max_{i=1}^{n}\{X_i\} < z) \\
&= P(X_1 < z, X_2 < z, \cdots, X_n < z) \\
&= P(X_1 < z) \cdot P(X_2 < z) \cdots P(X_n < z) \\
&= [F(z)]^n
\end{aligned} \tag{2.36}
$$

(Y, Z)的联合分布，记$G(y, z) = P(Y < y, Z < z)$。

若$y \geq z$,则

$$
G(y, z) = P(Y < y, Z < z) = P(Z < z) = [F(z)]^n \tag{2.37}
$$

若$y < z$,则

$$
\begin{aligned}
G(y, z) &= P(Y < y, Z < z) \\
&= P(Z < z) - P(Y \geq y, Z < z) \\
&= [F(z)]^n - [F(z) - F(y)]^n
\end{aligned} \tag{2.38}
$$

其联合概率密度函数为

$$
g(y, z) = \begin{cases} 0, y \geq z \\ n(n-1)\,[F(z) - F(y)]^{n-2} f(y) f(z), y < z \end{cases} \tag{2.39}
$$

我们求极差$R = Z - Y$的分布密度函数$f_R(r)$，显然当$r \leq 0$时，$f_R(r) = 0$，若$r > 0$，则

$$\begin{aligned} P(R < r) &= \iint_{z-y<r} g(y,z)\mathrm{d}y\mathrm{d}z \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{y+r} g(y,z)\mathrm{d}z \right] \mathrm{d}y \\ &= \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{r} g(y,y+t)\mathrm{d}t \right] \mathrm{d}y \\ &= \int_{-\infty}^{r} \left[\int_{-\infty}^{+\infty} g(y,y+t)\mathrm{d}y \right] \mathrm{d}t \end{aligned}$$

因此

$$f_R(r) = \int_{-\infty}^{+\infty} g(y,y+t)\mathrm{d}y \tag{2.40}$$

将式(2.39)带入可以得到

$$f_R(r) = n(n-1) \int_{-\infty}^{+\infty} [F(y+r) - F(y)]^{n-2} p(y)p(y+r)\mathrm{d}y \tag{2.41}$$

在实际应用中，极值分布与“百年一遇”等概念经常出现在灾害性天气预报中。

2.19 统计三大分布的数字特征

第3章 估计(Estimation)

3.1 点估计(Point Estimation)

总体$X \sim f(x, \boldsymbol{\theta})$，其中$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$属于参数空间$\boldsymbol{\Theta}$。也就是说，$\{f(x, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$是总体X的分布族，$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$是待定的未知参数。点估计的问题是：根据从总体中抽取的一个样本$X_1, \cdots, X_n$，如何估计未知参数$\boldsymbol{\theta}$，即构造设当的统计量$\hat{\theta}_i = g_i(X_1, \cdots, X_n)$（$i = 1, \cdots, k$）作为$\theta_i$的估计值。因此，$\hat{\theta}_i$称为$\theta_i$的估计量。之所以被称为点估计，是因为参数$\boldsymbol{\theta}$对应于坐标系中的一个点。

两种方法

- 矩估计
- 最大似然估计（MLE）

统计推断问题的解，往往可以从很多不同的方法和途径去考虑，并无一成不变的方法。不同方法固然有优劣之分，但这种优劣也是相对于一定的准则而言，并不是绝对的，例如估计甲在某一准则下优于已，而乙又在另一个准则下优于甲。

3.1.1 矩估计(Moment Estimation)

矩估计法是K.Pearson在19世纪末和20世纪初的一系列文章中引入的。这个方法的思想很简单：根据模型计算总体X的前k阶原点矩或者k阶中心矩，得出包含参数的矩表达式，然后利用样本矩替代总体矩，建立由k个表达式组成的方程组，并求解得到方程组的解。

由定义(1.14)，随机变量X的k阶原点矩和的k阶中心矩分别定义为

$$\alpha_k = E(X^k), \quad \mu_k = E[(X - E(X))^k]$$

对于分布族$f(x, \boldsymbol{\theta})$而言，其k阶原点矩$\alpha_k(\boldsymbol{\theta})$和k阶中心矩$\mu_k(\boldsymbol{\theta})$是以$\boldsymbol{\theta}$为参数的表达式。

定义 3.1. 对于一个样本$X_1, \cdots, X_n$和任意正整数k，定义

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

为样本k阶原点矩和k阶中心矩，其中$\bar{X} = a_1$为样本均值。

在样本大小n较大时，样本矩接近于对应的总体矩，从而得到包含k个未知数的方程组

$$\alpha_i(\boldsymbol{\theta}) = a_i(\text{或}\mu_i(\boldsymbol{\theta}) = m_i), i = 1, \cdots, k$$

由上述方程组可以求得解，记为$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)$，称为参数$\boldsymbol{\theta}$的矩估计。需要说明的是，虽然能够利用高阶矩建立方程组，但是在一般情况下采取如下的原则：尽可能地采用低阶矩处理。

3.1.2 最大似然估计(Maximum-Likelihood Estimation)

似然，英文为Likelihood，相似程度和可能程度。对于一个样本$\mathbf{X}$，样本的概率函数$f(\mathbf{X},)$与似然函数$L(\boldsymbol{\theta}, \mathbf{X})$本质上是相同的，仅仅是变量不同而已。前者将视$\boldsymbol{\theta}$为已知，x为样本空间上的变量值，而后者则视x为固定的抽样值，$\boldsymbol{\theta}$为参数空间$\boldsymbol{\Theta}$的变量值。

在统计学中，极大似然估计（MLE）是一种估计统计模型参数的方法，其在给定模型的情况下，能够应用抽样数据，对于模型参数进行估计。最大似然估计最早是由数学家Gauss在1821年针对于正态分布提出来的，之后Fisher在1912年的一篇论文中提出了一般分布下的最大似然。

定义 3.2 (Linkelihood Function)**.** 对于分布族$\{f(x, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$，在给定一个抽样$X_1, \cdots, X_n$的情况下，定义$L(\boldsymbol{\theta}, \mathbf{X})$为这n个样本$\mathbf{X} = (X_1, \cdots, X_n)$的联合概率分布，并称其为$\boldsymbol{\theta}$的似然函数，简称为似然函数。此外，称$\ln L(\boldsymbol{\theta}, \mathbf{X})$为对数似然函数，记为$l(\boldsymbol{\theta}, \mathbf{X})$或$l(\boldsymbol{\theta})$。

由定义可以看出，似然函数是以$\boldsymbol{\theta}$为未知参数而以$\mathbf{X}$为已知常数的样本联合概率分布。最大似然估计方法是在给定$\mathbf{X} = (X_1, \cdots, X_n)$下，从参数空间$\boldsymbol{\Theta}$中寻找使得最大似然函数$L(\boldsymbol{\theta}, \mathbf{X})$取得最大值的$\boldsymbol{\theta}$，即形式化的表示为

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}, \mathbf{X})$$

对数函数$\ln x$是单调增函数，因此当$f(x)$取得最大值时，$\ln f(x)$也取得极大值。因此上述求解可以等价的表示为

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l(\boldsymbol{\theta}, \mathbf{X})$$

之所以采用对数似然函数，是因为对于指数形式的分布函数，能够将指数乘积的形式转化为和的形式，从而简化求解过程。

如果似然函数$L(\boldsymbol{\theta}, \mathbf{X})$对于$\theta_i$（$i = 1, \cdots, k$）可微，则可以通过求解下面的方程求得估计值

$$\frac{\partial L(\boldsymbol{\theta}, \mathbf{X})}{\partial \theta_i}, i = 1, \cdots, k$$

或者求解等价的方程组

$$\frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \theta_i}, i = 1, \cdots, k$$

如果采用独立抽样，则$X_1, \cdots, X_n$为独立同分布（independent and identically distributed）的样本，其似然函数可以表示为

$$L(\boldsymbol{\theta}, \mathbf{X}) = \prod_{i=1}^{n} f(X_i|\boldsymbol{\theta})$$

类似地，对数似然函数可以表示为

$$l(\boldsymbol{\theta}, \mathbf{X}) = \sum_{i=1}^{n} \ln f(X_i|\boldsymbol{\theta})$$

3.1.3 点估计的评估准则

在考虑一个评估量得优劣时，必须从整体性能上去衡量它。这里所谓得“整体性能”

1. 无偏性
2. 方差小
3. 相合性

估计量的无偏性

设统计总体得分布包含未知参数$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$，而$\mathbf{X} = (X_1, \cdots, X_n)$是从该总体中抽出的样本，要根据样本来估计$g(\boldsymbol{\theta})$，其中g为一个已知函数。设$\hat{g}(\mathbf{X})$是一个估计量。如果对于任意可能得$\boldsymbol{\theta}$都有

$$E_{\boldsymbol{\theta}}[\hat{g}(\mathbf{X})] = g(\boldsymbol{\theta})$$

则称$\hat{g}$是g得一个无偏估计量。

估计量的无偏性有两个含义。第一个含义是没有系统性的偏差。虽然估计值$\hat{g}$总是随机地在实际值g上下波动，产生正负偏差，但是这些正负偏差在概率上的平均值为0。另一个含义就是如果抽样N次，抽样值分别为$(\mathbf{X}_1, \cdots, \mathbf{X}_N)$，对于任何$1 \leq i \leq N$，$\mathbf{X}_i = (X_1^{(i)}, \cdots, X_n^{(i)})$。根据大数定理，$N \to \infty$时，各次估计

值的平均，即$\sum_{i=1}^{N}\hat{g}(\mathbf{X}_i)/N$，依概率收到被估计值$g(\boldsymbol{\theta})$。如果没有无偏性，则无论使用多少次，其平均值都会与真值存在一定距离，这一距离就是系统误差。

最小方差无偏估计

设$\mathbf{X}=(X_1,\cdots,X_n)$是从某一个带参数$\boldsymbol{\theta}$的总体总抽取的样本，要估计$\boldsymbol{\theta}$。若采用估计量$\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}(\mathbf{X})$，则其误差为$\hat{\boldsymbol{\theta}}(\mathbf{X})-\boldsymbol{\theta}$。这误差随样本的具体值而定，也是随机的。定义

$$M_{\hat{\theta}}=E_{\theta} \tag{3.1}$$

均方误差

(Consistent Estimation)

3.2 区间估计(Interval Estimation)

为置信度或置信水平(confidence level).

精确度和可靠性

精度和置信度

样本大小的确定

第4章　假设检验(Hypothesis Test)

为显著性水平(significant level)

或零假设(null hypothesis).

对立假设或备选假设(alternative hypothesis)

Type I and Type II errors

critical region

one-sided hypotheses.

第5章　贝叶斯分析(Bayesian Analysis)

$p(\theta)$ 先验分布：顾名思义，先验分布是在没有观察数据情况下基于已有的经验对于参数概率分布的假设

$p(\theta|\mathbf{X})$ 或$p(\theta|x_1,\cdots,x_n)$ 后验分布：在获取观测数据之后重新对于参数分布的估计，因此称之为后验

$p(\mathbf{X})$或$p(\theta|x_1,\cdots,x_n)$ 观察数据的发生概率

$p(\mathbf{X}|\theta)$或$p(x_1,\cdots,x_n)|\theta)$ 似然函数，即为$L(\theta|\mathbf{X})$

后验分布 = 似然函数* 先验分布/ 观测数据

$$posterior = \frac{likelihood \cdot prior}{evidence}$$

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta) \cdot P(\theta)}{P(\mathbf{X})}$$

其中

$$P(\mathbf{X}) = \int_{\theta \in \Theta} P(\mathbf{X}|\theta)P(\theta)d\theta$$

在贝叶斯估计中，计算$P(\mathbf{X})$是最为困难的部分。

在预测问题中，贝叶斯方法扩展了最大后验估计

$$P(\tilde{x}|\mathbf{X}) = \int_{\theta \in \Theta} p(\tilde{x}|\theta)\frac{P(\mathbf{X}|\theta) \cdot P(\theta)}{P(\mathbf{X})}d\theta$$

先验是对于未来而言，已经得到的后验分布可以看作未来的先验分布。

贝叶斯公式和贝叶斯假设

贝叶斯假设：参数θ的无信息先验分布$\pi(\theta)$应在θ的取值范围内是“均匀”分布的。

后验分布$\pi(\theta|\mathbf{x})$是在样本$\mathbf{x}$给定下的θ的条件分布，基于后验分布的统计推断就意味着只考虑已出现的数据，而认为未出现的数据与推断无关，这种观点被称为“条件观点”。

经典统计学认为参数θ的无偏估计$\hat{\theta}$应该满足如下等式

$$E_{\hat{\theta}}(\mathbf{x}) = \int_x \hat{\theta}(\mathbf{x})p(\mathbf{x}|\theta)d\mathbf{x} = 0$$

贝叶斯推断中不用无偏性，而是采用条件方法。

5.0.0.1 共轭先验(conjugate priors)

定义 5.1 (共轭先验分布)**.** 设θ是总体分布中的参数(参数向量)，$\pi(\theta)$是θ的先验密度函数，假如由抽样信息算得的后验密度函数与$\pi(\theta)$ 有相同的函数形式，则称$\pi(\theta)$是θ的(自然)共轭先验分布

共轭先验在贝叶斯推理中具有重要意义，它使得后验分布和先验具有相同的函数形式。

共轭先验(conjugate priors) Conjugate prior

In Bayesian probability theory, if the posterior distributions $p(\theta|x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function.

表 5.1 共轭先验

Distribution	Conjugate Prior
Bernoulli	Beta distribution
Multinomial	Dirichlet distribution
Gaussian, Given variance, mean unknown	Gaussian distribution
Gaussian, Given mean, variance unknown	Gamma distribution
Gaussian, both mean and variance unknown	Gaussian-Gamma distribuation

在贝叶斯统计中，某一不确定量p的先验概率分布是在考虑"观测数据"前，能表达p不确定性的概率分布。它旨在描述这个不确定量的不确定程度，而不是这个不确定量的随机性。这个不确定量可以是一个参数，或者是一个隐含变量（latent variable）

贝叶斯学派和频率学派的区别之一是特别重视先验信息对于inference的影响，而引入先验信息的手段有“贝叶斯原则“（即把先验信息当着均匀分布）等四大类其中有重要影响的一类是：共轭先验

1. 将未知参数看成随机变量(随机向量)，记作θ，当θ已知时，样本$x_1, \cdots, x_n$对于θ的条件密度记为$p(x_1, \cdots, x_n|\theta)$，或者简写为$p(\mathbf{x}|\theta)$

2. 设法确定先验分布$\pi(\theta)$

3. 利用条件分布密度$p(x_1,\cdots,x_n|\theta)$和先验分布$\pi(\theta)$，可以求出$x_1,\cdots,x_n$与θ的联合分布和样本$x_1,\cdots,x_n$的分布，于是就可用它们求的θ对$x_1,\cdots,x_n$的条件分布密度，记后验分布密度$h(\theta|x_1,\cdots,x_n)$

4. 利用后验分布密度 $h(\theta|x_1,\cdots,x_n)$作出对于θ的推断

对于贝叶斯学派的匹配

- 参数θ看成随机变量是否妥当
- 先验分布是否存在？如何选取

无信息先验分布(Non-informative Priors)

二项分布与beta分布

二项分布中概率x的先验分布

$$p(x)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

似然概率为

$$p(y|x)=\binom{n}{k}x^k(1-x)^{n-k}$$

后验概率

$$p(x|y)=\frac{1}{B(\alpha+k,\beta+n-k)}x^{\alpha+k-1}(1-x)^{\beta+n-k-1}$$

独立实验序列，每次成功概率为θ，则实验n次成功k次的概率为

$$g(k|\theta)=\binom{n}{k}\theta^k(1-\theta)^{n-k}$$

上面的概率可以看作当θ已知时，k对于θ的条件概率。要从样本值k去估计θ。如果知道θ的边缘密度$q(\theta)$，则有贝叶斯公式，可以求出θ对于 k的条件密度

$$f(\theta|k)=\frac{q(\theta)\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\int_0^1 q(\theta)\binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta}$$

$q(\theta)$与实验结果无关，反应了在进行统计实验前，对于概率θ的知识，所以称之为先验分布。$f(\theta|k)$综合先验分布以及实验结果所带来的关于θ的信息，起与统计实验结果有关，称为后验分布。

如果对于实验一无所知，可以假设先验分布是$[0,1]$上的均匀分布

$$q(\theta)=\begin{cases}1, & \theta\in[0,1]\\ 0, & other\end{cases}$$

$$f(\theta|k) = \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}}{\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}d\theta}$$

如果用后验分布的期望值，即θ对于k条件期望去估计，就可以得到估计量

$$\begin{aligned}\hat{\theta} = E\theta|k &= \frac{1}{B(k+1,n-k+1)}\int_0^1 \theta.\theta^k(1-\theta)^{n-k}d\theta \\ &= \frac{B(k+2,n-k+1)}{B(k+1,n-k+1)} \\ &= \frac{k+1}{n+2}\end{aligned} \tag{5.1}$$

多项分布与Dirichlet分布

$\mathbf{x} = (x_1, x_2, \cdots, x_k)$的先验分布

$$p(x;\alpha_1,\alpha_2,\cdots,\alpha_k) = \frac{1}{B(\alpha)}\prod_{i=1}^{k} x_i^{\alpha_i-1}$$

似然函数

$$p(y|x) = \frac{n!}{n_1!n_2!\cdots n_k!}x_1^{n_1}\cdots x_k^{n_k}$$

后验概率

$$p(x|y) = \frac{1}{Z}\prod_{i=1}^{k} x_i^{\alpha_i+n_i-1}$$

其中后验概率

$$B(\alpha) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\Sigma_{i=1}^{k}\alpha_i)}$$

$$\Sigma_{i=1}^{k}x_i = 1$$

The selection of prior distributions by formal rules

In Bayesian statistical inference, a prior probability distribution, often called simply the prior, of an uncertain quantity p is the probability distribution that would express one's uncertainty about p before some evidence is taken into account.

In Bayesian statistics, the posterior probability of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence or background is taken into account.The posterior probability is the probability of the parameters θ given the evidence X: $p(\theta|x)$.

A prior probability is a marginal probability, interpreted as a description of what is known about a variable in the absence of some evidence. The posterior probability is

then the conditional probability of the variable taking the evidence into account. The posterior probability is computed from the prior and the likelihood function via Bayes' theorem.

先验概率是在缺乏某个事实的情况下描述一个变量；而后验概率是在考虑了一个事实之后的条件概率。

定理 5.1 (贝叶斯定理)**.** 假设$A \subseteq \bigcup_{i=1}^{n} B_i, P(A) > 0$,并且对于任何的$i \neq j$,满足$B_i \bigcap B_j = \varnothing$, 则有

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

先验概率，后验概率

定义 5.2 (随机变量)**.** 随机变量X是一个从样本空间到实数域集合的映射

$$X : \Omega \to \mathcal{R}$$

correlation does not imply causation

第6章　统计(statistics)

Sample Space S：Set of possible outcomes of a random experiment/survey .

Event: Every subset of is an Event.

Probability Set Function P

随机变量是将随机事件与实数联系起来的纽带，即将样本空间的事件映射为一个实数 A Random Variable X is a real valued function defined on S

$$X : S \to \mathcal{R}$$

Our Interest of Probabilities:

- Given $B \subset R$, want $P(x \in B)$.

$$\{X \in B\} = X^{-1}(B) = \{s \in S : X(s) \in B\} \Rightarrow P(X \in B) = P(\{s \in S : X(s) \in R\})$$

- Distribution function (df) of a random variable X

$$F(x) = P(X \leq x), x \in R$$

个人的理解：X^{-1}是一个从实数集到基本事件集集合的映射，即值域是一个基本事件的集合，其中基本事件是指不能分解的事件。

sample:样本

sampling: 抽样

probability sampling

representative sample

观测值$x = (x_1, \cdots, x_n)$的分布依赖于参数$\theta \in \Theta$

Fisher的初衷，统计应该是关于预测、解释和处理数据的学问。

正态分布成立的条件是大量的但每一个作用较小的因素的作用

关于统计推断如何做这个问题的主张和想法，划分为两大阵营

其一叫频率学派，其特征是把需要推断的参数θ视作固定且未知常数，而观测的样本X是独立和同分布的随机变量(independent and identically distributed,i.i.d)，

其着眼点在样本空间，有关的概率计算都是针对X的分布。基于理论平均性能，而贝叶斯学派认为平均性能不能很好地衡量统计过程的本质

另一派叫做贝叶斯学派，其认为独立同分布是在逻辑上是站不住脚的，虽然同样通过$p(X|\theta)$来描述概率的分布，但是将参数θ 视作随机变量，具有概率密度$p(\theta)$，其着眼点在参数空间，重视参数θ的分布，固定的操作模式是通过参数的先验分布结合样本信息得到参数的后验分布。统计推理需要基于实验并通过实验过程中所观测到的数据

$$p(\theta|x) = p(x|\theta)p(\theta)/p(x)$$

其中

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

为x的边缘分布，或称为预测分布(predictive distribution).

贝叶斯统计操作具有固定模式先验分布+样本信息$\Rightarrow$后验分布

贝叶斯学派描述给定x后，θ的不确定性，并计算关系

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

对于估计参数θ的问题，频率学派通常从如下两种方法中选择，用于生成估计器

- 针对于固定数目n的样本，优化一个基于风险的准则
- 优化当$n \to \infty$时渐进的性能度量

贝叶斯方法通过先于数据的概率特性来描述这个最初的不确定性，并把它和数据结合起来产业一个后于数据的修订了的概率。所有这些都是在概率运算的框架之内进行。

假定随机变量X的抽样分布密度为$p(x|\theta)$，其参数θ有先验分布$p(\theta)$,则后验分布为给了样本x之后的条件分布

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

这里分母

$$p(x) = \int p(x|\theta)p(\theta)d\theta$$

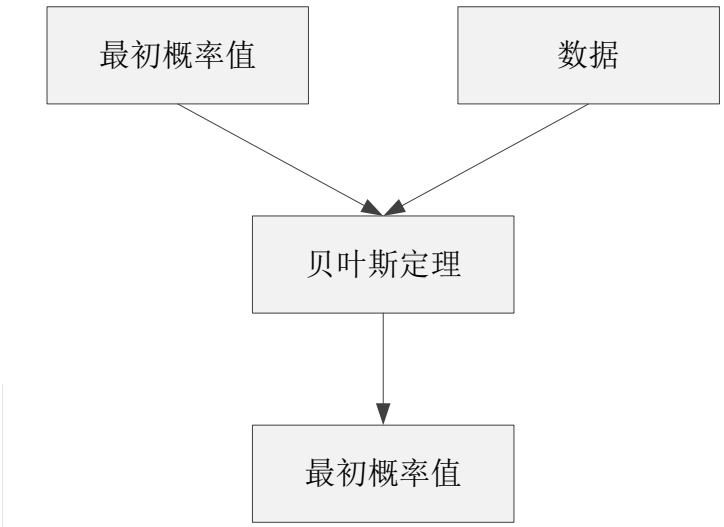

图 6.1 Bayesian Statistics

为x的边缘分布，或者为预测分布(predictive distribution)。分子为x和θ的联合分布$p(x|\theta)p(\theta) = p(x, \theta)$ 。

贝叶斯推断的步骤

1. 确定以参数θ识别的一个随机随机变量分布族
2. 根据过去额经验或其他信箱来确定参数的一个先验分布
3. 利用上面的公式求后验分布，并用后验分布进行所需的推断

贝叶斯推断的两个问题：

- 先验分布的选择
- 后验分布的计算

先验是对于未来而言，已经得到的后验分布可以看作未来的先验分布。

贝叶斯学派的最基本观点是：任一个未知量θ都可看作一个随机变量，应用一个概率分布取描述对于θ的未知状况。

在先验分布中经常也有未确定的参数，称为超参数(hyperparameter)。可以用所谓的第二类最大似然法来估计。

基于总体信息、样本信息和先验信息进行的统计推断被称为贝叶斯统计学，其与经典统计学的主要差别在于是否利用先验信息。

统计学的任务是有效地收集资料（即数据），并对之进行处理（整理、分析和推断等）。

收集资料有两种方式：观察和实验。与此相应，在统计学中产生了两个分支学科：

- 抽样调查或者抽样技术
- 实验设计

统计推断问题可以抽象如下数学问题：总体的概率分布$F_\theta(x)$包含了其值未知的参数θ(θ可以为向量)。从该总体随机抽样，得到样本$x_1, x_2, \cdots, x_n$，要通过抽样样本获得对θ的某些了解。统计推断的两个基本形式

- 估计问题。通过样本$x_1, x_2, \cdots, x_n$对于θ的值作出估计。由于估计的对象是参数，因此也称为参数估计。参数估计又分为两种基本形式
  - 点估计：用一个数值作为未知参数θ的估计值
  - 区间估计：用一个区间把θ估计在这个区间内
- 检查问题。有一个待判断的、与总体概率分布的参数有关的命题（在统计学上，称之为”假设检验“）。使用样本去判断一个假设是否成立，称为”假设检验“。假设检验问题的ut回答只有两种：接受假设或否定假设。

在检验一个假设时可能犯两种错误之一：一是假设本来对，但是被否定，称之为第一种错误；另一是假设本来不对，但被接受了，称之为第二种错误。

假阳性（false positives）和假阴性（false negatives）

Five Number Summary

- the upper quartile
- the lower quartile
- the median
- the extremes: the minimum and the maximum

Cramér–Rao bound (CRB) or Cramér–Rao lower bound (CRLB).The bound is also known as the Cramér–Rao inequality

In mathematical statistics, the **Fisher information** is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends. The probability function for X, which is also the likelihood function for θ, is a function $f(X;\theta)$; it is the probability mass (or probability density) of the random variable X conditional on the value of θ. The partial derivative with respect to θ of the natural logarithm of the likelihood function

is called the **score**.

Under certain regularity conditions, it can be shown that the first moment of the score (that is, its expected value) is 0:

$$E\left[\frac{\partial}{\partial\theta}\ln f(X;\theta)|\theta\right] = E\left[\frac{\partial f(X;\theta)}{f(X;\theta)}|\theta\right] = \int \frac{\partial f(x;\theta)}{f(x;\theta)} f(x;\theta)dx$$
$$= \int \frac{\partial}{\partial\theta} f(X;\theta)dx = \frac{\partial}{\partial\theta}\int f(X;\theta)dx = \frac{\partial}{\partial\theta}1 = 0$$

The second moment is called the Fisher information:

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta\right)^2 \mid \theta\right] = \int \left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 f(x;\theta)dx$$

Consider an unbiased estimator $\hat{\theta}(X)$. Mathematically, we write

$$E\left[\hat{\theta}(X) - \theta \mid \theta\right] = \int \left[\hat{\theta}(X) - \theta\right] f(x;\theta)dx = 0$$

The variance of any unbiased estimator $\hat{\theta}$ of θ is then bounded by the reciprocal of the Fisher information $I(\theta)$

$$var(\hat{\theta}) \geq \frac{1}{\mathcal{I}(\theta)}$$

测量尺度的四个水平

- 名义尺度(Categorieal/Qualitative:Nominal)
- 次序尺度(Categorieal/Qualitative:Ordinal)
- 区间尺度(Numeric/Quantitatve: Interval)
- 比例尺度(Numeric/Quantitatve: Ratio)

Two major types of statistics: descriptive statistics and inferential statistics.

Two methods other than a census for obtaining information are sampling and experimentation.

- simple random sampling
- Systematic Random Sampling
- Cluster Sampling
- Stratified Sampling
- Multistage Sampling

Principles of Experimental Design

• control: Two or more treatments should be compared

• randomization: The experimental units should be randomly divided into groups to avoid unintentional selection bias in constituting the groups.

• replication: A sufficient number of experimental units should be used to ensure that randomization creates groups that resemble each other closely and to increase the chances of detecting any differences among the treatments.

对照组(control group)和治疗组(treatment group)

Response variable: The characteristic of the experimental outcome that is to be measured or observed.

Factor: A variable whose effect on the response variable is of interest in the experiment.

Levels: The possible values of a factor.

Treatment: Each experimental condition. For one-factor experiments, the treatments are the levels of the single factor. For multifactor experiments, each treatment is a combination of levels of the factors.

In a **completely randomized design**, all the experimental units are assigned randomly among all the treatments

In a **randomized block design**, the experimental units are assigned randomly among all the treatments separately within each block.

第7章　统计决策论

θ成为参数，Θ成为参数空间。通常将决策叫做行为，特定的行为有a表示，所有的行为集合由$\mathscr{A}$表示。损失函数$L(\theta, a)$定义域为$\Theta \times \mathscr{A}$

定义 7.1 (贝叶斯期望损失)**.** 在做决策时，若$\pi(\theta)$为θ的可信概率分布，则行为a的贝叶斯期望损失为

$$\rho(\pi, a) = E^{\pi} L(\theta, a) = \int_{\Theta} L(\theta, a) dF^{\pi}(\theta) \tag{7.1}$$

频率派计算期望损失采用完全不同的方法，即对随机项$\mathbf{X}$取平均。经典推断方法没有包含先验和损失的信息。

用风险函数来选择一个决策法则是困难的，因为存在很多容许的决策法则(即决策法则之间不能在风险上总占优势，是交叉的)。为了选择一个具体的法则，必须引入其他附加原则。在经典统计学：最大似然性、无偏性、最小方差性和最小二乘原理。在决策论中的三个最重要的原则：贝叶斯风险原则、极小化极大原则和不变性原则。

称为“似然函数”的直观原因是，使$f(x|\theta)$大的θ比使$f(x|\theta)$小的θ 更“像是”θ的真实值。因为，如果$f(x|\theta)$ 大，x的出现就更有道理。

似然原理与另两个几乎被普遍接受的自然原理等价：第一个是充分性原理，第二是(弱)条件性原理

第8章 Regression Analysis

归分析（regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法. In statistics, regression analysis is a statistical process for estimating the relationships among variables.

Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).For Galton, regression had only this biological meaning,but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context.

A regression model relates Y to a function of $\mathbf{X}$ and β

$$Y \approx f(\mathbf{X}, \beta)$$

Regression models predict a value of the Y variable given known values of the $\mathbf{X}$ variables. Prediction within the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction outside this range of the data is known as extrapolation.

8.1 广义线性模型(generalized linear model)

Multiple Linear Regression (Ordinary Least Squares):Two Key Assumptions

- Y is Normally distributed random variable.
- Variance of Y is constant (homoscedastic).

$$y_i = z_i'\beta + \epsilon_i, \quad i = 1, \cdots, n$$

where z_i is the design vector which is an appropriate function of the covariate vector x_i.

The expectation μ_i is related to the linear predictor $\eta_i = z_i^{'}\beta$ by

$$\mu_i = h(\eta_i) = h(z_i^{'}\beta) \quad resp. \quad \eta_i = g(\mu_i)$$

Thus,a specific generalized linear model is fully characterized by three components:

- the type of the exponential family,
- the response or link function, and
- the design vector.

the densities of responses y_i, can always be written

$$f(y_i|\theta_i, \phi_i, \omega_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\omega_i + c(y_i, \phi_i, \omega_i)\right\}$$

其中

- θ_i称为自然参数(natural parameter)
- ϕ is an additional scala or dispersion parameter
- $b(\cdot)$ and $c(\cdot)$ are specific functions corresponding to the type of exponential family
- ω_i is a weight

如果$\omega_i = 1$,则可以简化为

$$f(y_i; \theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right\}$$

$$E[y] = \mu$$

$$\theta = \theta(\mu) = z^{'}\beta$$

$$\begin{aligned} P(Y = y) &= \mu^y(1-\mu)^{1-y} \\ &= (1-\mu)\left(\frac{\mu}{1-\mu}\right)^y \\ &= (1-\mu)\exp\left(y\ln\frac{\mu}{1-\mu}\right) \end{aligned} \tag{8.1}$$

逻辑函数为S形函数（Sigmoid function）联系函数(link function)

Response variables响应变量

response, outcome or dependent variable

explanatory variables(解释变量) or predictor variables or independent variables

A quantitative explanatory variable is sometimes called a covariate.

binary, dichotomous or binomial variables

polychotomous, polytomous or multinomial

exponential family of distributions

联系函数(link function).

指数型分布，指数族分布(exponential family of distributions.)

哑(或虚) 变量(dummy variable)

analysis of variance (ANOVA) Analysis of covariance (ANCOVA)

联合概率密度函数(joint probability density function)

One-way analyses 单因素分析

minimum bias procedures.1 These procedures impose a set of equations relating the observed data, the rating variables, and a set of parameters to be determined. An iterative procedure solves the system of equations by attempting to converge to the optimal solution

通过四个问题解释逻辑回归模型：1)为何弃用线性回归：多变量线性回归假设响应变量是方差恒定的正态分布，具有很大局限性；2)为何采用指数族分布：将一个分布密度函数转换为式指数族分布的形式，不仅能够适合Poisson、Binomial 和Gamma 等不同分布，而且易于计算多个独立响应变量的联合概率密度。3)为何使用sigmoid函数，逻辑回归针对的响应变量为Bernoulli分布，因此推导可得采用指数形式的Bernoulli分布，显然根据对于响应变量分布的不同假设，可以采用不同的函数形式；4)如何得到逻辑回归，假设有n个相互独立的解

释变量Z_i作为主要因素叠加影响Y，在Y=1 时得可以得到对应的逻辑回归模型。

$$
\begin{aligned}
f(y) &= \prod_{i=1}^{n} f(z_i;\beta_i) \\
&= \prod_{i=1} \exp\left[z_i\beta_i + \psi_i(\cdot)\right] \\
&= \exp\left[\sum_{i=1}^{n} z_i\beta_i + \sum_{i=1}^{n} \psi_i(\cdot)\right]
\end{aligned}
$$

$$
\log\frac{\pi}{1-\pi} = \sum_{i=1}^{n} z_i\beta_i = \boldsymbol{\beta}\cdot\mathbf{Z}
$$

• 应用更广泛：多变量线性回归假设响应变量是方差恒定的正态分布。指数族分布避免了线性回归对于响应变量的约束，能够适合Poisson、Binomial和Gamma等不同分布；

• 指数族分布的好处是易于计算多个独立响应变量的联合概率密度，例如在最大似然估计和多个解释参数的组合、log形式的似然函数

如下形式的线性模型

$$
E(Y) = \mu = \mathbf{x}^T\boldsymbol{\beta}, \qquad Y \sim N(\mu, \sigma^2)
$$

如下的函数g被称为联系函数(link function)

$$
g(\mu) = \mathbf{x}^T\boldsymbol{\beta}
$$

指数族分布的一般形式为

$$
f(y;\theta) = \exp\left[a(y)b(\theta) + c(\theta) + d(y)\right]
$$

如果$a(y) = y$,则该分布被称之为标准形式(Canonical Form)。$b(\theta)$有时被称为分布的自然参数(natural parameter)

指数族分布所满足的性质 1）根据概率密度函数的定义，曲线下面的面积为1

$$
f(y;\theta) =
$$

针对上式两边分别针对θ取导数，可得

$$
\frac{d}{d\theta}\int f(y;\theta)dx = \frac{d}{d\theta}\cdot 1 = 0
$$

如果响应变量Y，由n个解释变量$Z_i(i = 1, \cdots, n)$决定，则可以表示为

Bernoulli分布$\pi = P(Y = 1)$, 则$P(Y = 0) = 1 - \pi$,Bernoulli分布可以表示为指数形式

$$\begin{aligned} P(Y = y) &= \pi^y (1 - \pi)^{1-y} \\ &= (1 - \pi) \left(\frac{\pi}{1 - \pi} \right)^y \\ &= (1 - \pi) \exp \left(y \log \frac{\pi}{1 - \pi} \right) \\ &= exp \left(y \log \frac{\pi}{1 - \pi} + \log(1 - \pi) \right) \end{aligned}$$

令$\theta = \log \frac{\pi}{1-\pi}$，则$\pi = \frac{exp(\theta)}{1+exp(\theta)}$

$$P(Y = y) = exp \left(\theta y - \log(1 + e^{\theta}) \right)$$

$(z_1, \cdots, z_n)$的联合概率函数可以表示为

取Poisson 分布

$$\begin{aligned} P(Y = y) &= \frac{1}{y!} e^{-\lambda} \lambda^y \\ &= \frac{1}{y!} \exp \left(y \log \lambda - \lambda \right) \\ &= \exp \left(y \log \lambda - \lambda - \log(y!) \right) \end{aligned}$$

$$Y \sim N(\mu_y, \sigma_y^2)$$

$$y = \mathbf{X}^T \boldsymbol{\beta} + e$$

耐性分布(tolerance distribution)

logit函数(logit function)

$$log \left[\frac{\pi}{1 - \pi} \right]$$

8.2 点估计

8.2.1 Least Squares Estimation

与最大似然估计相比，最小二乘估计无需假设响应变量额分布，而最大似然估计需要指定联合概率分布

8.2.1.1 最大似然估计(Maximum likelihood estimation)

Consider a model which gives the probability density function of observable random variable X as a function of a parameter θ. Then for a specific value x of X, the function $L(\theta|x) = P(X = x|\theta)$ is a likelihood function of θ: it gives a measure of how "likely" any particular value of θ is, if we know that X has the value x.

The law of likelihood

$$\Lambda = \frac{L(a|X = x)}{L(b|X = x)} = \frac{P(X = x|a)}{P(X = x|b)}$$

In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters.

Suppose there is a sample $x_1, x_2, \cdots, x_n$ of n independent and identically distributed observations, coming from a distribution with an unknown probability density function $f_0(\cdot)$. It is however surmised that the function f_0 belongs to a certain family of distributions $\{f(\cdot|\theta), \theta \in \Theta\}$ (where θ is a vector of parameters for this family), called the parametric model, so that $f_0 = f(\cdot|\theta_0)$. The value θ_0 is unknown and is referred to as the true value of the parameter. It is desirable to find an estimator $\hat{\theta}$ which would be as close to the true value θ_0 as possible.

For an **independent and identically distributed** sample, this joint density function is

$$f(x_1, x_2, \cdots, x_n|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta$$

The likelihood

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

The log-likelihood

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

the average log-likelihood

$$\hat{l} = \frac{1}{n} \ln L$$

The method of maximum likelihood estimates θ_0 by finding a value of θ that maximizes $\hat{l}(\theta; x)$.This method of estimation defines a maximum-likelihood estimator (MLE)

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{l}(\theta; x_1, x_2, \cdots, x_n)$$

最大似然估计寻找使似然最大化的参数

$$L(\theta|\mathbf{X}) \triangleq= P(\mathbf{X}|\theta) = \bigcap_{x \in \mathbf{X}} P(X = x|\theta) = \prod_{x \in X} P(x|\theta)$$

为了简化，通常采用log形式的似然$\mathcal{L} \triangleq \log L$

$$\hat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{X}) = \arg\max_{\theta} \sum_{x \in \mathbf{X}} \log P(X = x|\theta)$$

8.2.2 最大后验估计(Maximum a posteriori estimation)

最大后验估计非常类似于最大似然估计，但是其允许通过先验分布$p(\theta)$从而在参数上包含一些先验知识

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\mathbf{X})$$

通过贝叶斯规则，能够将上式改写为

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta} \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})} \\
&= \arg\max_{\theta} P(\mathbf{X}|\theta)P(\theta) \\
&= \arg\max_{\theta} \{\mathcal{L}(\theta|\mathbf{X}) + \log p(\theta)\} \\
&= \arg\max_{\theta} \{\sum_{x \in \mathbf{X}} \log p(x|\theta) + \log p(\theta)\}
\end{aligned}$$

在实践中，先验概率不仅用于编码额外的知识，而且能够通过强制采用简单模型而防止过拟合

8.2.3 定理

Hoeffding不等式(集中不等式)适用于有界的随机变量。设有两两独立的一系列随机变量$X_1, \cdots, X_n$。假设对于所有的$1 \leq i \leq n$, X_i都是几乎有界的变量，即满足

$$P(X_i \in [a_i, b_i]) = 1$$

那么这n个随机变量的经验期望

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

满足以下的不等式

$$P(\overline{X} - E[\overline{X}] \geq t) \leq \exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$
$$P(|\overline{X} - E[\overline{X}]| \geq t) \leq 2\exp\left(-\frac{2t^2n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

如果$X_1, X_2, \cdots, X_n$为一组独立同分布的参数为p的伯努利分布随机变量, 上述不等式可以表示为

$$P(|\overline{X} - E[\overline{X}]| \geq \delta) \leq 2\exp\left(-2\delta^2 n\right)$$

参考文献

陈浩元. 著录文后参考文献的规则及注意事项 [J]. 编辑学报, 2005, 17(6): 413-415.

Lamport L. Document preparation system [M]. Addison-Wesley Reading, MA, 1986.

probabilitycourse.com. Unordered sampling with replacement [J/OL]. https://www.probabilitycourse.com/chapter2/2_1_4_unordered_with_replacement.php.

Walls S C, Barichivich W J, Brown M E. Drought, deluge and declines: the impact of precipitation extremes on amphibians in a changing climate [J/OL]. Biology, 2013, 2(1): 399-418[2013-11-04]. http://www.mdpi.com/2079-7737/2/1/399. DOI: 10.3390/biology2010399.

Wikibook. http://en.wikibooks.org/wiki/latex [M]. On-line Resources, 2014.