目录

1. Probability

1.1. Sample Spaces and Events

Definition 1.1.1 (Sample Spaces)**:::>** The sample space Ω is the set of possible outcomes of an experiment. Points ω in Ω are called sample outcomes, realizations, or elements.

同一随机事件的样本空间的表达并不是唯一的.

以扔两枚硬币为例, 第一个样本空间为:$\{HH, HT, TH, TT\}$, 第二个样本空间为:{两正, 两反, 一正一反}, 区别在于, 第一个样本空间是一个等可能样本空间, 而第二个样本空间中的基本事件概率不相等.

Definition 1.1.2 (Events)**:::>** Subsets of the sample space Ω are called events.

我们这里说事件是样本空间的子集, 但并不是说样本空间的子集一定是事件. 而这涉及到测度论中的不可测集的问题. 对于事件的更严谨的定义: 事件是能定义概率的样本空间的子集.

Definition 1.1.3 (σ-域)**:::>** 这里引入 σ-域与 σ-代数的定义, 是为了更清晰地分析样本空间 ω 的哪些子集是事件.

定义事件族 $\mathscr{F}$, $\mathscr{F}$ 是集合的集合, $\mathscr{F}$ 中每一个元素都是事件. 事件族 (Ω 的子集) $\mathscr{F}$ 称为 σ-域 (也称为 σ-代数或事件体), 如果它满足如下三个条件:

- $\Omega \in \mathscr{F}$
- 若 $A \in \mathscr{F}$, 则 $A^c \in \mathscr{F}$

- 若 $A_1, A_2, \cdots \in \mathscr{F}$, 则 $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$. 注意前面的 A 是要达到无穷的, 即可数的事件之并也是事件.

从以上条件我们可以看出, σ-域对补运算, 可数和运算均封闭. 那第一第二性质组合, 我们发现空集也属于 σ-域, 进而再与第三个条件组合, 我们发现 σ-域对有限和也封闭 (将无穷-有限的部分全部设置为空集). 进而, 根据德摩根率, 有限交也封闭.

由定义可知, Ω 上最小的 σ-域为 $\{\Omega, \emptyset\}$, 最大的 σ-域为 Ω 全体子集组成的集合, 称为幂集 (也是最简单的 σ-域, 比如按概率的角度来讲每一个都是事件)

若 $A \subset B$, 则将 $\mathscr{F}_A$ 记作由 A 生成的 σ-域, 也即包含 A 的 σ-域中最小的一个, 即 $\{A, A^c, \Omega, \emptyset\}$, 共 4 个元素.

在概率论上的意义即是, 我们发现一个事件不论怎样运算结果都还是事件.

根据以上概念, 我们可以引入事件的定义, 若 $A \in \mathscr{F}$, 则 A 是一个事件. 此外, 在数学上, 我们将集合 Ω, 集合 Ω 的 σ-域 $\mathscr{F}$ 组合在一起, 称为一个可测空间 $(\Omega, \mathscr{F})$, 若再加上一个测度 μ, 则可获得一个测度空间 $(\Omega, \mathscr{F}, \mu)$.

Definition 1.1.4 (随机试验)**:::>** 随机试验是研究概率论的方法. 没有随机试验的话, 就没有客观概率. 与客观概率对应的是主观概率, 如" 明天爆发第三次世界大战的概率是 1%." 研究主观概率的方法可以采用类似贝叶斯统计的方法, 先估计出一个先验概率, 再根据事情的变化更新后验概率.

满足如下两个条件的试验被称为随机试验:

1. 试验可以在相同的情况下重复多次进行
2. 知道试验所有的结果, 但不知道究竟会出现哪个结果.

Example 1.1.5:::> If we toss a coin forever, then the sample space is the infinite set:

$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \ldots,) : \omega_i \in \{H, T\}\}$$

Let E be the event that the first head appears on the third toss. then

$$E = \{(\omega_1, \omega_2, \omega_3, \ldots,) : \omega_1 = T, \omega_2 = T, \omega_3 = H, \omega_i \in \{H, T\} \text{ for } i > 3\}.$$

Given an event A, let $A^c = \{\omega \in \Omega : \omega \notin A\}$ denote the complement of A.

The union of events A and B is defined:

$$A \bigcup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{ both }\}$$

If $A_1, A_2, \ldots$ is a sequence of sets then:

$$\bigcup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for at least one i }\}.$$

The intersection of A and B is

$$A \bigcap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$$

Sometimes we write $A \cap B$ as AB or (A, B).

If $A_1, A_2, \ldots$ is a sequence of sets then

$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all i }\}.$$

The set difference is defined by $A - B = \{\omega : \omega \in A, \omega \notin B\}$.

If every element of A is also contained in B we write $A \subset B$ or, equivalently, $B \supset A$.

If A is a finite set, let $|A|$ denote the number of elements in A.

Definition 1.1.6 (Disjoint events)**:::>** We say that $A_1, A_2, \ldots$ are disjoint or are mutually exclusive if $A_i \bigcap A_j = \emptyset$ whenever $i \neq j$

Definition 1.1.7 (Partition of sample space)**:::>** A partition of Ω is a sequence of disjoint sets $A_1, A_2, \ldots$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$.

Definition 1.1.8 (Indicator function of event A)**:::>** Given an event A, define the indicator function of A by:

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

Definition 1.1.9 (Monotone increasing)**:::>** A sequence of sets $A_1, A_2, \ldots$ is monotone increasing if $A_1 \subset A_2 \subset \ldots$ and we define $\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$.

A sequence of sets $A_1, A_2, \ldots$ is monotone decreasing if $A_1 \supset A_2 \supset \cdots$ and then we define $\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$. In either case, we will write $A_n \to A$.

1.2. Probability

Definition 1.2.1 (Probability)**:::>** 概率有多种定义方式:

1. 最开始通过统计定义来定义概率, 即用频率来近似概率. 有些书上这样定义:" 频率收敛的那个数就是概率", 这是通过大数定理推出的, 但是这又是一个逻辑循环.
2. 古典定义, 适用于古典定义的概率为古典概型. 古典概型的特点: 基本事件是有限的, 且基本事件是等可能的.
3. 几何定义, 适用于几何定义的概率为几何概型. 几何概型的 Ω 为可求面积或体积区域. 之后讲到的均匀分布实际上就是几何概型.

Definition 1.2.2 (Probability distribution)**:::>** A function $\mathbb{P}$ that assigns a real number $\mathbb{P}(A)$ to each event A is a probability distribution or a probability measure if it satisfies the following three axioms:

Axiom 1.2.3:::> $\mathbb{P}(A) \geqslant 0$

Axiom 1.2.4:::> $\mathbb{P}(\Omega) = 1$

Axiom 1.2.5:::> If $A_1, A_2, \ldots$ are disjoint then:

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

In fact, it is not always possible to assign a probability to every event A if the sample space is large, such as the whole real line. Instead, we assign probabilities to a limited class of set called a σ-field.

There are many interpretations of $\mathbb{P}(A)$. The two common interpretations are frequencies and degrees of beliefs. In the frequency interpretation, $\mathbb{P}(A)$ is the long run proportion of times that A is true in repetitions. The degree-of-belief interpretation is that $\mathbb{P}(A)$ measures an observer's strength of belief that A is true. The difference in the interpretation will not matter much until we deal with statistical inference. There, the deffering interpretations lead to two schools of inference: the frequentist and the Bayesian schools.

Definition 1.2.6 (概率的公理化定义)**:::>**

A *probability space* is a triple $(\Omega, \mathscr{F}, P)$.

Ω is a set called the *sample space*, $\mathscr{F}$ is a collection of subsets of Ω, and $P : \mathscr{F} \to [0,1]$ is the *probability measure*

概率 (也称为概率测度)P 为 F 上的非负值函数, 即对每一事件 $A \in \mathscr{F}$, 都可定义一个数 $P(A)$, 满足下列条件:

1. $0 \leqslant P(A) \leqslant 1$ for all $A \in \mathscr{F}$
2. $P(\Omega) = 1$
3. For any countable collection of events $A_1, A_2, \cdots$ which are disjoint, i.e. $A_i \cap A_j = \emptyset$ for all i, j, we have

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

这个性质称作可数可加性或可列可加性.

If Ω is finite (or countable), we usually take $\mathscr{F}$ to be all the subsets of Ω, i.e. the power set of Ω. However, if Ω is, say, R, we have to be a bit more careful and only include nice subsets, or else we cannot have a well-defined P.

Often it is not helpful to specify the full function P. Instead, in discrete cases, we just specify the probabilities of each outcome, and use the third axiom to obtain the full P.

Proposition 1.2.7:::>

$$\mathbb{P}(\emptyset) = 0$$

$$A \subset B \quad \Longrightarrow \quad \mathbb{P}(A) \leqslant \mathbb{P}(B)$$

$$0 \leqslant \mathbb{P}(A) \leqslant 1$$

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

$$A \bigcap B = \emptyset \Longrightarrow \mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

$$P(A) + P(B) - 1 \leqslant P(AB) \leqslant min(P(A), P(B))$$

$$\mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$$

Proof.

$$\begin{aligned}
\mathbb{P}(A \bigcup B) &= \mathbb{P}\left((AB^c) \bigcup (AB) \bigcup (A^c B)\right) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^c B) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^c B) + \mathbb{P}(AB) - \mathbb{P}(AB) \\
&= P\left((AB^c) \bigcup (AB)\right) + \mathbb{P}\left((A^c B) \bigcup (AB)\right) - \mathbb{P}(AB) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB).
\end{aligned}$$

■

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{1\leqslant i<j\leqslant n} \left|A_i \cap A_j\right| + \sum_{1\leqslant i<j<k\leqslant n} \left|A_i \cap A_j \cap A_k\right| - \cdots + (-1)^{n-1} |A_1 \cap \cdots \cap A_n|.$$

$$P\left(\bigcup_{n=1}^{n} A_n\right) = 1 - P\left(\bigcap_{n=1}^{n} A_n\right)$$

Example 1.2.8:::> Let $1, 2, \cdots, n$ be randomly permuted to $\pi(1), \pi(2), \cdots, \pi(n)$. If $i \neq \pi(i)$ for all i, we say we have a *derangement*.

Let $A_i = [i = \pi(i)]$.

Then

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k} P(A_k) - \sum_{k_1<k_2} P(A_{k_1} \cap A_{k_2}) + \cdots \\ &= n \cdot \frac{1}{n} - \binom{n}{2} \frac{1}{n} \frac{1}{n-1} + \binom{n}{3} \frac{1}{n} \frac{1}{n-1} \frac{1}{n-2} + \cdots \\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1} \frac{1}{n!} \\ &\to e^{-1} \end{aligned}$$

So the probability of derangement is $1 - P(\bigcup A_k) \approx 1 - e^{-1} \approx 0.632$.

Theorem 1.2.9 (Bonferroni's inequalities)**:::>** For any events $A_1, A_2, \cdots, A_n$ and $1 \leqslant r \leqslant n$, if r is odd, then

$$\begin{aligned} P\left(\bigcup_{1}^{n} A_i\right) \leqslant &\sum_{i_1} P(A_{i_1}) - \sum_{i_1<i_2} P(A_{i_1}A_{i_2}) + \sum_{i_1<i_2<i_3} P(A_{i_1}A_{i_2}A_{i_3}) + \cdots \\ &+ \sum_{i_1<i_2<\cdots<i_r} P(A_{i_1}A_{i_2}A_{i_3}\cdots A_{i_r}). \end{aligned}$$

If r is even, then

$$\begin{aligned} P\left(\bigcup_{1}^{n} A_i\right) \geqslant &\sum_{i_1} P(A_{i_1}) - \sum_{i_1<i_2} P(A_{i_1}A_{i_2}) + \sum_{i_1<i_2<i_3} P(A_{i_1}A_{i_2}A_{i_3}) + \cdots \\ &- \sum_{i_1<i_2<\cdots<i_r} P(A_{i_1}A_{i_2}A_{i_3}\cdots A_{i_r}). \end{aligned}$$

Theorem 1.2.10 (Continuity of Prabability)**:::>** If $A_n \to A$ then

$$\mathbb{P}(A_n) \to \mathbb{P}(A)$$

as $n \to \infty$

Proof. Suppose that A_n is monotone increasing so that $A_1 \subset A_2 \subset \cdots$. Let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$. Define $B_1 = A_1, B_2 = \{\omega \in \Omega : \omega \in A_2, \omega \notin A_1\}$, $B_3 = \{\omega \in \Omega : \omega \in A_3, \omega \notin A_2, \omega \notin A_1\}, \ldots$ It can be shown that $B_1, B_2, \ldots$ are disjoint, $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$ for each n and $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$. Hence, we have:

$$\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \mathbb{P}(B_i)$$

And:

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}(A)$$

■

1.3. Probability on Finite Sample Spaces

Definition 1.3.1 (uniform probability distribution)**:::>** Suppose that the sample space $\Omega = \{\omega_1, \ldots, \omega_n\}$ is finite and if each outcomes is equally likely, then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

which is called the uniform probability distribution.

或者, 也可以记作:

$$P(A) = \frac{\#A}{\#\Omega}$$

其中 $\#A$ 也叫做有利场合数.

Example 1.3.2 (摸球问题)**:::>** 摸球问题非常重要, 因为它和统计里的抽样问题较为相似. 但是要区分是放回还是不放回, 放回是独立问题, 不放回是相关问题.

- 袋中装有 α 个白球及 β 个黑球, 从袋子中不放回地抽取 $a+b$ 次, 求恰好 a 个白球和 b 个黑球的概率.

  这样不放回地抽取, 和一次抽取 $a+b$ 个球的概率是一样的. 抽取的结果为:

  $$\frac{C_{\alpha}^{a}C_{\beta}^{b}}{C_{a+\beta}^{a+b}}$$

- 袋中装有 α 个白球及 β 个黑球, 从袋子中放回地抽取 n 次, 求恰好 k 个白球的概率.

  $$C_n^k(\frac{\alpha}{\alpha+\beta})^k(\frac{\beta}{\alpha+\beta})^{n-k}$$

- 袋中装有 α 个白球及 β 个黑球, 从中不放回地抽取 k 个球, 求第 k 次摸球恰好是白球的概率.

  这个问题有多种处理方法, 比如我们可以将其转化为编号的 $\alpha+\beta$ 个球放进 $\alpha+\beta$ 个盒子里, 故样本空间是一个全排列问题, 而将第 k 个放入一个白球, 剩下的球又是一个全排列问题. 结果为:

  $$\frac{a(a+b-1)!}{(a+b)!}=\frac{a}{a+b}$$

  类似地, 如果放回地抽取 k 个球, 第 k 次摸球恰好是白球的概率显然也是 $\frac{a}{a+b}$.

  这即是著名的抽签问题, 抽签结果和次序无关. 但是需要注意的是, 抽签结果和次序无关的前提条件是所有人的信息都是完全对称的.

- 将 3 个球随机地放入 1, 2, 3 ,4 号盒子中去 (每盒放球数不限), 求至少有一个球的盒子的最小号码是 3 的概率.

  解:

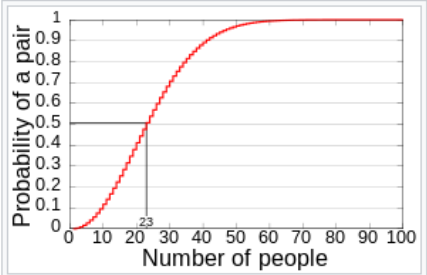
  $$P=\frac{2^3-1}{4^3}=\frac{7}{64}$$

  其中, 分子部分表示将三个球全部放在 3,4 号盒子里减去三个球全放在四号盒子里的情况.

  本题的关键是找出等可能的基本事件是什么.

错误的解题方法如下: 样本空间共有 $C_4^3 + C_4^3 \times 2 + C_4^1 = 20$ 种, 而所求事件共 $1+2=3$ 种, 这种解法错误的原因在于这些" 基本" 情况并非等可能的.

Example 1.3.3 (分房问题 (生日问题))**:::>** In probability theory, the birthday problem asks for the probability that, in a set of n randomly chosen people, at least two will share a birthday. The birthday paradox is that, counterintuitively, the probability of a shared birthday exceeds 50% in a group of only 23 people.



The computed probability of at least two people sharing a birthday versus the number of people

While it may seem surprising that only 23 individuals are required to reach a 50% probability of a shared birthday, this result is made more intuitive by considering that the comparisons of birthdays will be made between every possible pair of individuals. With 23 individuals, there are $(23 \times 22)/2 = 253$ pairs to consider, which is well over half the number of days in a year (182.5 or 183).

The goal is to compute $P(A)$, the probability that at least two people in the room have the same birthday. However, it is simpler to calculate $P(A')$, he probability that no two people in the room have the same birthday. Then, we have $P(A) = 1 - P(A')$.

则设 n 个人在房间内, n 个人生日均不同的概率为:

$$\overline{p(n,365)} = \frac{A_{365}^n}{365^n}$$

作为对比, 房间中 n 个其他人中与特定人 (比如自己) 有相同生日的概率

为:

$$q(n,365) = 1 - \left(\frac{364}{365}\right)^n$$

Example 1.3.4:::> Suppose we toss a fair coin until we get exatcly two heads. Describe the sample space S. What is the probability that exactly k tosses are required.

The sample space is a set of coin toss results sequences containing two heads, and ending in heads:

$$S = \left\{(r_1,\ldots,r_k) : r_i \in \{\text{ head, tails }\}, |\left\{r_j = \text{ head }\right\}| = 2, r_k = \text{ head }\right\}$$

The probability of stopping afte k tosses is the probability of obtaining exactly 1 head in the first $k-1$ tosses, in a procedure that would not stop after any number of tosses, followed by the probability of getting a head in the k-th toss. This value is:

$$\left((k-1)\left(\frac{1}{2}\right)^{k-1}\right)\left(\frac{1}{2}\right) = \frac{k-1}{2^k}$$

Note that, besides this combinatorial argument, we can verify that these probabilities do indeed add up to 1:

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$$

$$\frac{d}{dx}\frac{1}{1-x} = \sum_{k=0}^{\infty} \frac{d}{dx} x^k$$

$$\frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} k x^{k-1}$$

$$\frac{x}{(1-x)^2} = \sum_{k=0}^{\infty} k x^k$$

so, for $x = 1/2, \sum_{k=0}^{\infty} 2^{-k} k = 2$, and so

$$\sum_{k=0}^{\infty} \frac{k}{2^{k+1}} = 1$$

更进一步地, 我们可以计算: 平均需要抛掷多少硬币, 才会首次出现连续的 n 个正面?(当然这种情景跟上述不同, 上述情景只要" 出现" 即可, 而这个情景需要" 连续")

- 解法一: 递归. 先从退化情况开始: 扔硬币直到连续两次出现正面. 假设期望次数是 E, 我们开始扔, 有如下几种情况:
  - 扔到的是反面, 那么就要重新扔, 所以是 $\frac{E+1}{2}$
  - 扔到的是正面, 再扔一次又反面了, 则是 $\frac{E+2}{4}$
  - 扔到两次都是正面, 结束, 则是 $\frac{1}{4}\times 2$

  所以从递归的角度讲, $E=\frac{E+1}{2}+\frac{E+2}{4}+\frac{1}{4}\times 2$, 解得 $E=6$
- 解法二: 数列递推假设连续得到 k 次正面的数学期望为 x_k, 假设连续得到 $k+1$ 次正面的数学期望为 x_{k+1}, 在得到 k 次正面后, 为了得到 $k+1$ 次正面, 再抛下一次会有以下两种情况:
  - 抛出来的是正面, 概率为 $\frac{1}{2}$, 这时得到了连续 $k+1$ 次正面, 数学期望为 1
  - 抛出来的是反面, 概率为 $\frac{1}{2}$, 这时为了得到 $k+1$ 次连续正面, 前面抛的全部作废, 必须从头开始, 因此数学期望为 $1+x_{k+1}$

    综上得到等式:

    $$x_{k+1}=x_k+\frac{1}{2}\times 1+\frac{1}{2}\left(1+x_{k+1}\right)$$

    变形后得到一个数列递推式, 再由初始条件 $x_1=2$, 得到 $x_n=2^{n+1}-2$

Example 1.3.5:::> Let $\Omega=\{0,1,\dots,\}$. Prove that there does not exist a uniform distribution on Ω.

Proof. Assume that such a distribution exists, and let $\mathbb{P}(\{1\})=p$. Since the distribution is uniform, the probability associated with any set of size 1 is p, and the probability associated with any set of size n is np.

1. If $p>0$, then a finite set A of size $|A|=\lceil 2/p\rceil$ would have probability value $\mathbb{P}(A)=\lceil 2/p\rceil p\geqslant(2/p)p=2$, which is greater than 1 – a contradiction.
2. If $p=0$, then any finite set A must have $\mathbb{P}(A)=0$. But then $\mathbb{P}(\Omega)=\sum_i\mathbb{P}(\{i\})=\sum_i 0=0$, instead of 1— a contradiction.

■

Example 1.3.6 (match problem)**:::>** 一个屋子里面有 N 个人, 每个人有一顶帽子. 假如所有人把帽子扔到屋子中央, 然后每个人都随机选一顶帽子. 问:

1. 没有人捡到自己帽子的概率
2. 有 $k(k \leqslant N)$ 个人捡到自己帽子的概率

解决方法一

问题 1

先计算至少有一个人捡到自己帽子的概率. 设 $E_i, i = 1,2,\ldots,N$ 表示事件第 i 个人捡到了他自己的帽子. 则根据容斥原理, 至少一个人捡到自己帽子的概率 $P\left(\bigcup_{i=1}^{N} E_i\right)$ 就等于:

$$\begin{aligned} P\left(\bigcup_{i=1}^{N} E_i\right) = &\sum_{i=1}^{N} P(E_i) - \sum_{i_{1<i2}} P\left(E_{i_1} E_{i_2}\right) + \ldots \\ &+ (-1)^{n+1} \sum_{i_1 < i_2 \cdots i_n} P\left(E_{i_1} E_{i_2} \ldots E_{i_n}\right) \\ &+ \cdots + (-1)^{N+1} P(E_1 E_2 \ldots E_N) \end{aligned}$$

如果将实验结果看作是一个 N 维数组的话, 其中第 i 个元素表示被第 i 个人捡到的帽子的编号, 那么有 $N!$ 种可能的结果 (例如结果 $(1,2,3,\ldots,N)$ 表示所有人都拿到了自己的帽子). 进一步地, $E_{i_1} E_{i_2} \ldots E_{i_n}$ 表示事件是 $i_1, i_2, \ldots, i_n$ 这 n 个人拿到了自己的帽子, 这样的可能有 $(N-n)(N-n-1)\cdots 3\cdot 2\cdot 1 = (N-n)!$ 种, 因为剩下的 $N-n$ 个人随便选帽子. 因此:

$$P\left(E_{i_1} E_{i_2} \cdots E_{i_n}\right) = \frac{(N-n)!}{N!}$$

同时, 对于 $\sum_{i_1 < i_2 < \cdots < i_n} P\left(E_{i_1} E_{i_2} \ldots E_{i_n}\right)$ 总共有 $\binom{N}{n}$ (即 C_N^n) 种选法, 因此:

$$\sum_{i_1 < i_2 < \cdots < i_n} P\left(E_{i_1} E_{i_2} \ldots E_{i_n}\right) = \frac{N!(N-n)!}{(N-n)!n!N!} = \frac{1}{n!}$$

从而:

$$P\left(\bigcup_{i=1}^{N} E_i\right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{N+1}\frac{1}{N!}$$

所以, 没有人捡到自己帽子的概率就是:

$$1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^N}{N!}$$

当 N 非常大时, 此概率为 $e^{-1} \approx .36788$(e^x 的泰勒展开). 也就是说, N 很大时, 没有人捡到自己帽子的概率大约是 0.37, 而不是很多人认为的当 $N \to \infty$ 时概率会趋向于 1.

问题 b

设 E 表示事件 k 中每个人都拿到了自己的帽子, G 表示事件其它的 $N-k$ 个人中没有人拿到自己的帽子. 则:

$$P(EG) = P(E)P(G \mid E)$$

设 $F_i, i = 1, \ldots, k$, 表示事件第 i 个人拿到了自己的帽子. 则:

$$\begin{aligned} P(E) &= P(F_1F_2\cdots F_k) \\ &= P(F_1)P(F_2 \mid F_1)P(F_3 \mid F_2F_1)\cdots P\left(F_k \mid F_1\cdots F_{k-1}\right) \\ &= \frac{1}{N}\frac{1}{N-1}\frac{1}{N-2}\cdots\frac{1}{N-k+1} \\ &= \frac{(N-k)!}{N!} \end{aligned}$$

已知 k 人中的每个人都捡到了自己的帽子, 其余 $N-k$ 个人将随机地在他们的 $N-k$ 个帽子中选择, 因此其他人没有人捡到自己帽子的概率是:

$$P(G \mid E) = P_{N-k} = \sum_{i=0}^{N-k}(-1)^i/i!$$

因此, 指定的 k 人捡到自己帽子而其它人没有捡到自己帽子的概率是:

$$P(EG) = \frac{(N-k)!}{N!}P_{N-k}$$

因为选择 k 人有 $\binom{N}{k}$ 方法, 因此, 要求的概率就是:

$$P(\text{ 只有}k\text{ 个人捡到自己帽子}) = P_{N-k}/k! \approx e^{-1}/k! \text{ 当}k\text{ 很大时}$$

解决方法二

问题 a

设 E 表示事件没有匹配发生，此事件显然与 n 有关，记为 $P_n = P(E)$. 我们从第一个人是否选择到了自己的帽子开始, 分别记作 M 和 M^C. 则:

$$P_n = P(E) = P(E \mid M)P(M) + P\left(E \mid M^C\right)P\left(M^C\right)$$

显然, $P(E \mid M) = 0$, 因此

$$P_n = P\left(E \mid M^C\right)\frac{n-1}{n}$$

现在,$P\left(E \mid M^C\right)$ 为剩下 $n-1$ 个人都没有捡到自己帽子的概率 (且其中一人的帽子不在这些帽子中, 因为已经被第一个人捡走了). 此事发生可由两种独立事件组成: 那个人捡到了第一个人的帽子且其它人都没有捡到自己的帽子, 以及那个人捡到了除了第一个人和他自己的帽子以外的一顶帽子, 且其它人都没有捡到自己的帽子. 前者的概率是 $[1/(n-1)]P_{n-2}$, 由此可得:

$$P\left(E \mid M^C\right) = P_{n-1} + \frac{1}{n-1}P_{n-2}$$

因而, 我们可以得到:

$$P_n = \frac{n-1}{n}P_{n-1} + \frac{1}{n}P_{n-2}$$

或者, 等价地:

$$P_n - P_{n-1} = -\frac{1}{n}\left(P_{n-1} - P_{n-2}\right)$$

再由于 P_n 表示 n 个人中无匹配的概率, 因此:

$$P_1 = 0 \quad P_2 = \frac{1}{2}$$

可得:

$$P_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots + \frac{(-1)^n}{n!}$$

问题 b

为了计算只有 k 人得到自己帽子的概率, 我们考虑一个确定的 k 人组, 有且只有此 k 人捡到自己帽子的概率是:

$$\frac{1}{n}\frac{1}{n-1}\cdots\frac{1}{n-(k-1)}P_{n-k} = \frac{(n-k)!}{n!}P_{n-k}$$

其中 P_{n-k} 表示 $n-k$ 个人中没有捡到自己帽子的概率, 总共有 $\binom{n}{k}$ 种选法, 因此有且只有 k 个捡到自己帽子的概率是:

$$\frac{P_{n-k}}{k!} = \frac{\frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^{n-k}}{(n-k)!}}{k!}$$

在计算概率问题中, 我们时常可能遇到一些估计问题, 这里引入对 Stirling 公式的介绍.

Proposition 1.3.7:::> Before we state and prove Stirling's formula, we prove a weaker (but examinable) version:

$$\log n! \sim n \log n$$

Note that

$$\log n! = \sum_{k=1}^{n} \log k.$$

Now we claim that

$$\int_1^n \log x \, dx \leqslant \sum_1^n \log k \leqslant \int_1^{n+1} \log x \, dx.$$

We actually evaluate the integral to obtain

$$n \log n - n + 1 \leqslant \log n! \leqslant (n+1)\log(n+1) - n;$$

Divide both sides by $n \log n$ and let $n \to \infty$. Both sides tend to 1. So

$$\frac{\log n!}{n \log n} \to 1.$$

Theorem 1.3.8 (Stirling's formula)**:::>** As $n \to \infty$,

$$\log\left(\frac{n! e^n}{n^{n+\frac{1}{2}}}\right) = \log\sqrt{2\pi} + O\left(\frac{1}{n}\right)$$

Corollary 1.3.9:::>

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$$

Proof. (non-examinable) Define

$$d_n = \log\left(\frac{n!e^n}{n^{n+1/2}}\right) = \log n! - (n+1/2)\log n + n$$

Then

$$d_n - d_{n+1} = (n+1/2)\log\left(\frac{n+1}{n}\right) - 1.$$

Write $t = 1/(2n+1)$. Then

$$d_n - d_{n+1} = \frac{1}{2t}\log\left(\frac{1+t}{1-t}\right) - 1.$$

We can simplifying by noting that

$$\log(1+t) - t = -\frac{1}{2}t^2 + \frac{1}{3}t^3 - \frac{1}{4}t^4 + \cdots$$
$$\log(1-t) + t = -\frac{1}{2}t^2 - \frac{1}{3}t^3 - \frac{1}{4}t^4 - \cdots$$

Then if we subtract the equations and divide by $2t$, we obtain

$$\begin{aligned} d_n - d_{n+1} &= \frac{1}{3}t^2 + \frac{1}{5}t^4 + \frac{1}{7}t^6 + \cdots \\ &< \frac{1}{3}t^2 + \frac{1}{3}t^4 + \frac{1}{3}t^6 + \cdots \\ &= \frac{1}{3}\frac{t^2}{1-t^2} \\ &= \frac{1}{3}\frac{1}{(2n+1)^2 - 1} \\ &= \frac{1}{12}\left(\frac{1}{n} - \frac{1}{n+1}\right) \end{aligned}$$

By summing these bounds, we know that

$$d_1 - d_n < \frac{1}{12}\left(1 - \frac{1}{n}\right)$$

Then we know that d_n is bounded below by d_1+ something, and is decreasing since $d_n - d_{n+1}$ is positive. So it converges to a limit A. We know A is a lower bound for d_n since (d_n) is decreasing.

Suppose $m > n$. Then $d_n - d_m < \left(\frac{1}{n} - \frac{1}{m}\right)\frac{1}{12}$. So taking the limit as $m \to \infty$, we obtain an upper bound for d_n: $d_n < A + 1/(12n)$. Hence we know that

$$A < d_n < A + \frac{1}{12n}.$$

However, all these results are useless if we don't know what A is. To find A, we have a small detour to prove a formula:

Take $I_n = \int_0^{\pi/2} \sin^n \theta \, d\theta$. This is decreasing for increasing n as $\sin^n \theta$ gets smaller. We also know that

$$\begin{aligned} I_n &= \int_0^{\pi/2} \sin^n \theta \, d\theta \\ &= \left[-\cos\theta \sin^{n-1}\theta\right]_0^{\pi/2} + \int_0^{\pi/2} (n-1)\cos^2\theta \sin^{n-2}\theta \, d\theta \\ &= 0 + \int_0^{\pi/2} (n-1)(1-\sin^2\theta)\sin^{n-2}\theta \, d\theta \\ &= (n-1)(I_{n-2} - I_n) \end{aligned}$$

So

$$I_n = \frac{n-1}{n} I_{n-2}.$$

We can directly evaluate the integral to obtain $I_0 = \pi/2$, $I_1 = 1$. Then

$$\begin{aligned} I_{2n} &= \frac{1}{2}\cdot\frac{3}{4}\cdots\frac{2n-1}{2n}\pi/2 = \frac{(2n)!}{(2^n n!)^2}\frac{\pi}{2} \\ I_{2n+1} &= \frac{2}{3}\cdot\frac{4}{5}\cdots\frac{2n}{2n+1} = \frac{(2^n n!)^2}{(2n+1)!} \end{aligned}$$

So using the fact that I_n is decreasing, we know that

$$1 \leqslant \frac{I_{2n}}{I_{2n+1}} \leqslant \frac{I_{2n-1}}{I_{2n+1}} = 1 + \frac{1}{2n} \to 1.$$

Using the approximation $n! \sim n^{n+1/2} e^{-n+A}$, where A is the limit we want to find, we can approximate

$$\frac{I_{2n}}{I_{2n+1}} = \pi(2n+1)\left[\frac{((2n)!)^2}{2^{4n+1}(n!)^4}\right] \sim \pi(2n+1)\frac{1}{ne^{2A}} \to \frac{2\pi}{e^{2A}}.$$

Since the last expression is equal to 1, we know that $A = \log\sqrt{2\pi}$. Hooray for magic! ■

This approximation can be improved:

Proposition 1.3.10 (non-examinable)**:::>** We use the $1/12n$ term from the proof above to get a better approximation:

$$\sqrt{2\pi}n^{n+1/2}e^{-n+\frac{1}{12n+1}} \leqslant n! \leqslant \sqrt{2\pi}n^{n+1/2}e^{-n+\frac{1}{12n}}.$$

1.4. Geometric probability

Example 1.4.1 (约会问题)**:::>** 两人相约于晚 7 点到 8 点间在某地会面, 先到者等足 20 分钟便立即离去. 设两人的到达时刻在 7 点到 8 点间都是随机且等可能的. 求两人能会面的概率 p.

设 x 为男生到达的时间, y 为女生到达的时间, 0 为 7 点, 1 为 8 点,

则样本空间为:

$$\omega = \{(x,y) \,|\, 0 \leqslant x \leqslant 1, |\, 0 \leqslant y \leqslant 1\}$$

要求的事件为:

$$A = \left\{(x,y) \in \Omega : |x-y| \leqslant \frac{1}{3}\right\}$$

图形化求解, 在平面直角坐标系中画出面积, 可得概率为 $\frac{5}{9}$

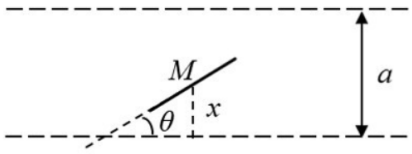
Example 1.4.2 (零概率事件)**:::>** 零概率事件不一定一定不发生, 测度为零的集合不一定是空集.

不可能事件的概率为零, 但概率为零的事件不一定为不可能事件.

Example 1.4.3 (Buffon 投针问题)**:::>** Suppose we have a floor made of parallel strips of wood, each the same width, and we drop a needle onto the floor. What is the probability that the needle will lie across a line between two strips? The solution for the sought probability p , in the case where the needle length l is not greater than the width a of the strips, is

$$p = \frac{2}{\pi}\frac{l}{a}$$

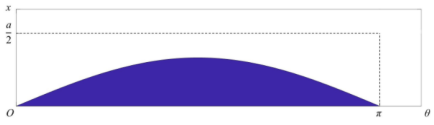
This can be used to design a Monte Carlo method for approximating the number π, although that was not the original motivation for de Buffon's question.



Buffon's needle problem

首先需要明确, 平面上刻画一个针的位置需要两个自由度. 假设针的中点为 M, 点 M 到最近的一条平行线的距离为 x, 针与平行线的夹角为 θ, 那么事件: 针与平行线相交可表述为:$x \leqslant \frac{l}{2}\sin\theta$, 而整个样本空间的测度表述是:$0 \leqslant x \leqslant \frac{a}{2}$, $0 \leqslant \theta \leqslant \pi$.

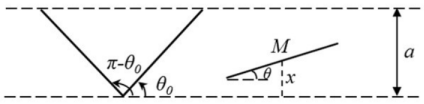
以 θ 为横坐标、x 为纵坐标作图即得



所以待求概率为蓝色部分面积除以矩形面积, 即:

$$P = \frac{\int_0^{\pi} \frac{l}{2}\sin\theta \mathrm{d}\theta}{\frac{a}{2}\pi} = \frac{2l}{a\pi}$$

而当 $l \geqslant a$ 时:



上图中临界角 $\theta_0 = \arcsin\frac{a}{l}$, 易知当 $\theta_0 \leqslant \theta \leqslant \pi - \theta_0$ 时细针必定与平行

线相交; 当 $0 \leqslant \theta \leqslant \theta_0$ 或 $\pi - \theta_0 \leqslant \theta \leqslant \pi$ 时, 只有当 $x \leqslant \frac{l}{2}\sin\theta$ 才与平行线相交综上作图有:

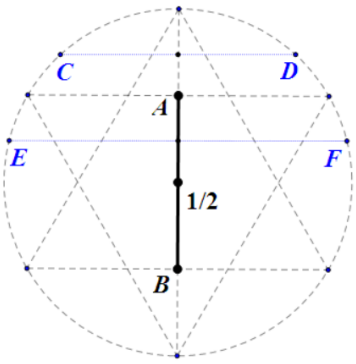
故针与平行线相交的概率为:

$$\begin{aligned} P &= \frac{2\left(\int_0^{\theta_0} \frac{l}{2}\sin\theta d\theta + \frac{a}{2}\left(\frac{\pi}{2} - \arcsin\frac{a}{l}\right)\right)}{\frac{a}{2}\pi} \\ &= 1 - \frac{1}{\pi}\left(2\arcsin\frac{a}{l} - \frac{2l}{a} + \frac{2}{a}\sqrt{l^2 - a^2}\right) \end{aligned}$$

Example 1.4.4 (Bertrand 悖论)**:::>** 在圆中“随机”画一根弦, 求其长度大于内接正三角形边长的概率. 由于问题的提法不确定, 分别以弦的端点、半径和中点的角度考虑, 可以得出三个不同的答案, 而针对各自的解释每个解法都是正确的.

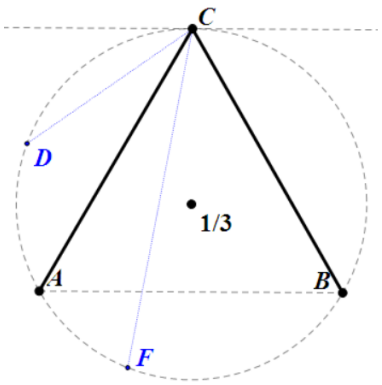
原因: 当随机试验有无穷多个可能结果时, 有时很难规定" 等可能" 这一概念.

- 几率为 $\frac{1}{2}$ 的情况: 假设中点在直径上等概率分布, AB 为直径长度的 $\frac{1}{2}$, 画弦时, CD 小于边长, EF 大于边长.



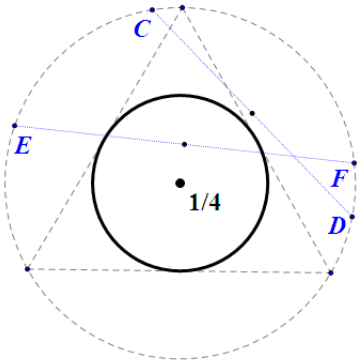
- 几率为 $\frac{1}{3}$ 的情况:

  如下图, ∠ACB 为 $\frac{\pi}{3}$, 画弦时, CD 小于边长, CF 大于边长. 需要弦的方向等概率分布, 即任意相邻两共端点的弦夹角相等, 由于这个夹角是圆周角, 圆周角相等意味着弦长相等, 弦长相等意味着内接正多边形. 所以这个要求可以等效为需要弦的端点在圆周上等概率分布.



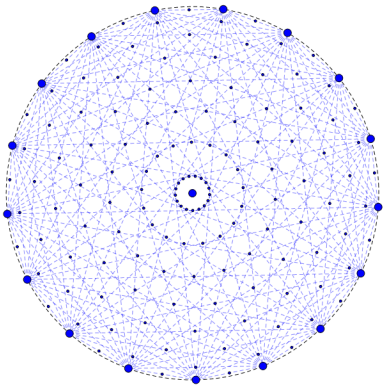
- 几率为 $\frac{1}{4}$ 的情况:

  如下图, 小圆面积为大圆面积的 1/4, 画弦时, CD 小于边长, EF 大于边长. 需要假设中点在圆内部等概率分布

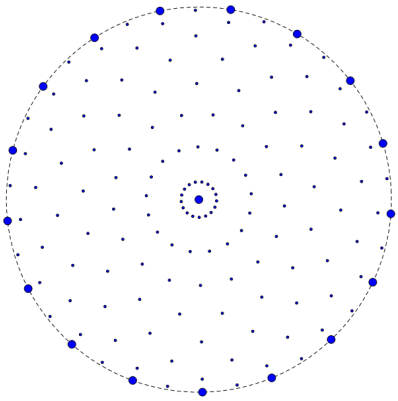


注意, “中点在直径上等概率分布”、“弦的端点在圆周上等概率分布”、“中点在圆内部等概率分布”这三个条件不是等价的, 用几何画板可以画正十七

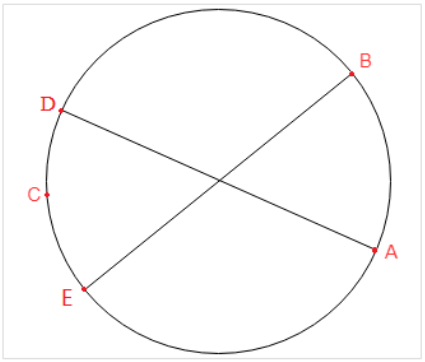
边形, 满足“弦的端点在圆周上等概率分布”, 那么一共有 136 根弦, 如下图:

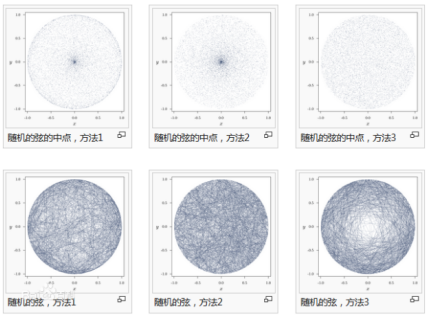


从下图看, 靠近圆心的位置弦的中点更密集, 出现的概率更大, 不满足“中点在圆内部等概率分布”



下面这张图更加直观:





为什么三个结果都是对的? 关键是对" 随机" 的理解不一样, 则貌似一个事件的随机试验是不同的, 因为样本空间不同 (在本例中, 对取弦的限制不同)

Example 1.4.5:::> 在一个圆环上随机取三个点, 求这三个点组成一个锐角三角形的概率. 取单位圆上任意两点 A 和 B, A 和 B 两点确定以后, 点 A, B, C 要构成锐角三角形, 点 C 必须在 DE 之间, 否则将构成锐角或钝角三角形. 设 AB 弧所对应的圆心角为 θ, 当且仅当 $\theta \in (0, \pi)$ 时有可能构成锐角三角形. θ 的概率密度是 $\frac{1}{\pi}$, 此时组成锐角三角形需要 C 点在 AB 对应的概率是 $\frac{\theta}{2\pi}$. 故在一个圆环上随机添加 3 点, 三个点组成一个锐角三角形的概率为:

$$\int_0^{\pi} \frac{1}{\pi} \cdot \frac{\theta}{2\pi} \mathrm{d}\theta = \left. \frac{\theta^2}{4\pi^2} \right|_0^{\pi} = \frac{1}{4}$$

1.5. Independent Events

Definition 1.5.1 (Independent Events)**:::>** Two events A and B are independent if

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$$

A set of events $\{A_i : i \in I\}$ is independent if

$$\mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \mathbb{P}(A_i)$$

for every finite subset J of I.

事件的独立性是概率论有别于测度论的重要特征.

Example 1.5.2:::> Suppose that A and B are disjoint events, each with positive probability. They are not independent. This follows since $\mathbb{P}(A)\mathbb{P}(B) > 0$ yet $\mathbb{P}(AB) = \mathbb{P}(\emptyset) = 0$ Except in this special case, there is no way to judge independence by looking at the sets in a Venn diagram. 也就是说, 独立和互斥是不同的两个概念.

Example 1.5.3:::> Suppose that $\mathbb{P}(A_i) = 1$ for each i. Prove that

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

Proof. 将证明转换为已知的条件.

首先利用德摩根率:

$$\mathbb{P}(\cap_{i=1}^{\infty} A_i) = 1 - \mathbb{P}((\cap_{i=1}^{\infty} A_i)^c) = 1 - \mathbb{P}(\cup_{i=1}^{\infty} A_i^c)$$

其次利用概率和的性质:

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i^c) \leqslant \sum_{i=1}^{\infty} \mathbb{P}(A_i^c) = \sum_{i=1}^{\infty} (1 - \mathbb{P}(A_i)) = \sum_{i=1}^{\infty} 0 = 0$$

由于概率值非负, 不等号方向不变, 命题得证. ■

Example 1.5.4:::> Suppose that A and B are independent events. Show that A^c and B^c are independent events.

Proof.

$$\begin{aligned}\mathbb{P}(A^cB^c) &= \mathbb{P}((A\cup B)^c) = 1-\mathbb{P}(A\cup B)\\ &= 1-(\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(AB))\\ &= 1-\mathbb{P}(A)-\mathbb{P}(B)+\mathbb{P}(A)\mathbb{P}(B)\\ &= 1-(1-\mathbb{P}(A^c))-(1-\mathbb{P}(B^c))+(1-\mathbb{P}(A^c))(1-\mathbb{P}(B^c))\\ &= \mathbb{P}(A^c)\mathbb{P}(B^c)\end{aligned}$$

■

Example 1.5.5:::> Show that if $\mathbb{P}(A)=0$ or $\mathbb{P}(A)=1$ then A is independent of every other event. Show that if A is independent of itself then $\mathbb{P}(A)$ is either 0 or 1 .

Proof. If $\mathbb{P}(A)=0$, then $\mathbb{P}(AB)=\mathbb{P}(A)-\mathbb{P}(A-B)=0-\mathbb{P}(A-B)\leqslant 0$, and since probabilities are non-negative we must have $\mathbb{P}(AB)=0$. Therefore $\mathbb{P}(AB)=\mathbb{P}(A)\mathbb{P}(B)=0$ for all events B, and A is independent of every other event.

If $\mathbb{P}(A)=1$, then $\mathbb{P}(A^c)=0$, and so A^c and B are independent for every other event B. Then, A is also independent from every other event B^c– which covers all potential events, since every event has a complement.

If A is independent of itself, $\mathbb{P}(AA)=\mathbb{P}(A)\mathbb{P}(A)$, so $\mathbb{P}(A)=\mathbb{P}(A)^2$ or $\mathbb{P}(A)(\mathbb{P}(A)-1)=0$. Therefore $\mathbb{P}(A)=0$ or $\mathbb{P}(A)=1$. ■

Example 1.5.6:::> A, B 相互独立, 即 $P(AB)=P(A)P(B)$

$$\begin{aligned}&\Leftrightarrow P(A\mid B)=P(A) && P(B)>0\\ &\Leftrightarrow P(A\mid \overline{B})=P(A) && P(\overline{B})>0\\ &\Leftrightarrow P(A\mid B)=P(A\mid \overline{B}) && 0<P(B)<1\end{aligned}$$

从而可得结论, 若四对事件 $\{A,B\},\{\overline{A},B\},\{A,\overline{B}\},\{\overline{A},\overline{B}\}$ 有一对是相互独立的, 则另外三对也都是相互独立的.

Definition 1.5.7 (多个事件的独立性及其实质)**:::>** 称 n 个事件 $A_1, A_2, \cdots, A_n$ 是相互独立的, 如果对任意自然数 $k(2 \leqslant k \leqslant n)$ 都有

$$P\left(A_{i_1} A_{i_2} \cdots A_{i_k}\right) = P\left(A_{i_1}\right) P\left(A_{i_2}\right) \cdots P\left(A_{i_k}\right)$$

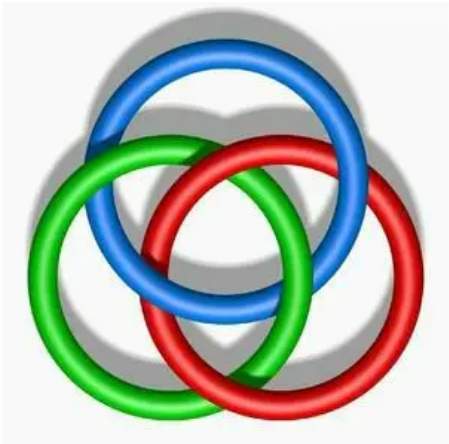
其中 $i_1, i_2, \cdots, i_k$ 是满足 $1 \leqslant i_1 < i_2 < \cdots < i_k \leqslant n$ 的任意 k 个自然数.

独立性的实质是指事件 σ 域互相独立.

Theorem 1.5.8:::> 如果事件 $A_1, A_2, \cdots, A_n$ 相互独立, 那么将该事件序列分成互不重叠的 m 组 $(m \leqslant n)$, 则这 m 组事件各自生成的 σ 域也是相互独立的. 后者的含义是指: 在此 m 个 σ-域任意各取一个事件, 则此 m 个事件总是相互独立的.

Example 1.5.9:::> 相互独立一定两两独立, 但两两独立不一定相互独立. 相互独立也叫联合独立, 要在两两独立的基础上满足 $p(abc) = p(a)p(b)p(c)$
举例: 扔一粒骰子, 扔出结果的事件是两两独立的.

一种叫做波罗梅奥环 (Borromean Rings) 的形状可以很好地说明这一点: 三个环两两独立, 但三个环无法分离.



Definition 1.5.10 (相关系数)**:::>** 在我们判断独立性的时候, 我们是通过判断 $P(AB) - P(A)P(B)$ 是否为零来判断其是否独立的.

但如果遇到极端事件 (某个事件概率为 1 或为 0), 这个公式就不能区分了. 而如果某个事件概率非常趋近于 0, 或者非常趋近于 1 的话, 我们为了规

范化处理, 可以定义相关系数:

$$\gamma_{A,B}=\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(A^C)P(B)P(B^C)}}$$

若 $\gamma_{A,B}>0$, 则 A,B 是相关的, 若 $\gamma_{A,B}=0$, 则两者是独立的, 若 $\gamma_{A,B}<0$, 则两者是负相关的.

性质:

$$\gamma_{AB}>0\Leftrightarrow P(A\mid B)>P(A)$$

即一个事件的发生让另一个事件的概率增大.

$$\gamma_{AB^C}=-\gamma_{AB}$$

$$\left|\gamma_{A,B}\right|\leqslant 1$$

$$\gamma_{A,B}=1\Leftrightarrow P(A\triangle B)=0$$

即 A, B 的对称差为 0, 也即 $P(A)=P(B)=P(AB)$

但是, 这样定义无法刻画三个事件的独立程度.

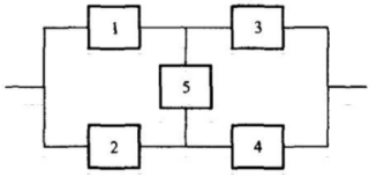
Example 1.5.11 (系统可靠性)**:::>** 我们称元件能正常工作的概率为该元件的可靠性. 而系统的可靠性就是该系统能正常工作的概率, 因此, 系统的可靠性就由各元件的可靠性决定.

串联系统: 由 n 个元件串联而成, 故只要有一个元件失效, 该系统就失效.

并联系统: 由 n 个元件并联而成, 故只要有一个元件正常工作, 则该系统就不会失效.

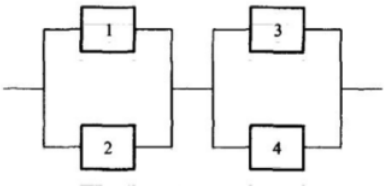
$$\begin{aligned}p_{并}&=P(A_1\cup\cdots\cup A_n)=1-P(A_1^c\cdots A_n^c)\\&=1-P(A_1^c)\cdots P(A_n^c)\\&=1-(1-p_1)\cdots(1-p_n)\end{aligned}$$

例: 求如下系统的可靠性:



此系统较为复杂, 我们通过考虑元件 5 是否正常工作, 把系统可靠性的计算, 化为两个比它简单的系统的可靠性的计算的组合.

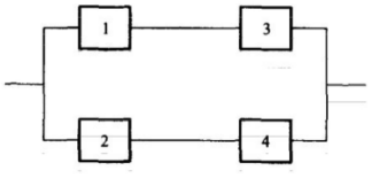
若元件 5 正常工作, 那么此时系统等效如下:



其可靠性为:

$$[1-(1-p_1)(1-p_2)][1-(1-p_3)(1-p_4)]$$

而若元件 5 失效, 那么此时系统等效如下:



其可靠性为:

$$1-(1-p_1p_3)(1-p_2p_4)$$

则由全概率公式知, 原系统的可靠性概率为:

$$\begin{aligned} & p_5[1-(1-p_1)(1-p_2)][1-(1-p_3)(1-p_4)] \\ & +(1-p_5)[1-(1-p_1p_3)(1-p_2p_4)] \end{aligned}$$

Example 1.5.12:::> 一架长机和两架僚机一同飞往某地进行轰炸, 但要到达目的地, 非要有无线电导航不可, 而只有长机具有此项设备. 一旦到达目的地, 各机将独立地进行轰炸且炸毁目标的概率为 0.3. 在到达目的地之前, 必须经过敌军的高射炮阵地上空, 此时任一架飞机被击落的概率为 0.2, 求目标被炸毁的概率.

Solution. 本题重要的是, 飞过去和炸毁这两个过程不能简单地直接用概率相乘, 因为轰炸时炸毁的概率与飞过去飞机的数量有关.

$$\begin{aligned} P &= 0.8\times 0.2\times 0.2\times 0.3 \\ &+2\times 0.8\times 0.8\times 0.2\times (1-0.7\times 0.7) \\ &+0.8\times 0.8\times 0.8\times (1-0.7\times 0.7\times 0.7) \\ &=0.476544 \end{aligned}$$

■

1.6. Conditional Probability

Definition 1.6.1 (Conditional Probability)**:::>** If $\mathbb{P}(B)>0$ then the conditional probability of A given B is

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

It is in general not true that $\mathbb{P}(A\mid B\bigcup C)=\mathbb{P}(A\mid B)+\mathbb{P}(A\mid C)$. The rules of probability apply to events on the left of the bar.

Example 1.6.2 (The presecutor's fallacy)**:::>** In general it is not the case that $\mathbb{P}(A\mid B)=\mathbb{P}(B\mid A)$. This mistake is made often enough in legal cases that it is sometimes called the prosecutor's fallacy.

For example, a medical test for a disease D has outcomes $+$ and $-$. The probabilities are:

	D	D^c
$+$	.009	.099
$-$	.001	.891

From the definition of conditional probability,

$$\mathbb{P}(+ \mid D) = \frac{\mathbb{P}(+ \bigcap D)}{\mathbb{P}(D)} = \frac{.009}{.009 + .001} = .9$$

$$\mathbb{P}(- \mid D^c) = \frac{\mathbb{P}\left(- \bigcap D^c\right)}{\mathbb{P}(D^c)} = \frac{.891}{.891 + .099} \approx .9$$

Apparently, the test is fairly accurate. Sick people yield a positive 90 percent of the time and healthy people yield a negative abouut 90 percent of the time.

Suppose you go for a test and get a positive. What is the probability you have the disease?

$$\mathbb{P}(D \mid +) = \frac{\mathbb{P}(+ \bigcap D)}{\mathbb{P}(+)} = \frac{.009}{.009 + .099} \approx .08$$

The lesson hear is that you need to compute the answer numerically. Don't trsut your intuition.

Another example: If a perpetrator were known to have the same blood type as a given defendant and 10% of the population to share that blood type, then one version of the prosecutor's fallacy would be to claim that, on that basis alone, the probability that the defendant is guilty is 90%. However, this conclusion is only close to correct if the defendant was selected as the main suspect based on robust evidence discovered prior to the blood test and unrelated to it (the blood match may then be an "unexpected coincidence"). Otherwise, the reasoning presented is flawed, as it overlooks the high prior probability (that is, prior to the blood test) that he is a random innocent person. Assume, for instance, that 1000 people live in the town where the murder occurred. This means that 100 people live there who have the perpetrator's blood type; therefore, the true probability that the defendant is guilty –based on the fact that his blood type matches that of the killer

–is only 1%, far less than the 90% argued by the prosecutor. The error is based on assuming that $P(A \mid B) = P(B \mid A)$, where A represents the event of finding evidence on the defendant, and B the event that the defendant is innocent. But this equality is not true: in fact, although $P(A \mid B)$ is usually very small,$P(B \mid A)$ may still be much higher.

Lemma 1.6.3:::> If A and B are independent events then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

Also, for any pair of events A and B,

$$\mathbb{P}(AB) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$$

That is to say, another interpretation of independence is that knowing B doesn't change the probability of A.

Example 1.6.4 (可靠性)**:::>** 在可靠性理论中, 某装置在时间段 $(0, t]$ 上无故障的概率 $g(t)$ 叫做可靠性函数 (简称可靠性), 而

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{h(t + \Delta t; t)}{\Delta t}$$

叫做故障强度函数 (死亡率或风险率), 其中 $h(t + \Delta t; t)$ 是该装置在 $(0, t]$ 上无故障的情况下, 在 $(t, t + \Delta t]$ 上出现故障的条件概率. 故障强度 $\lambda(t)$ 是装置可靠性的重要特征, 在实际问题中它可以通过试验来测定. 重要的是, 有了它如何确定可靠性函数. 因此其数学问题是: 设 $\lambda = \lambda(t)$ 已知, 试求装置的可靠性函数 $g(t)$ (其中设 $g(0) = 1$, 即假定设备在开始时刻都是无故障的)

Solution. 考虑下列事件

$$W_{(t,t+h]} = \text{“装置在}(t, t + \Delta t] \text{上出现故障"}$$

$$R_t = \text{“装置在}(0, t] \text{上无故障"}$$

根据乘法公式有:

$$g(t + \Delta t) = P\left(R_{t+\Delta t}\right) = P\left(R_t W^c_{(t,t+\Delta t]}\right) = P\left(R_t\right) P\left(W^c_{(t,t+\Delta t]} \mid R_t\right)$$

由于 $P\left(W_{(t,t+\Delta t]} \mid R_t\right)$ 是该装置在 $(0, t]$ 上无故障的情况下, 在 $(t, t + \Delta t]$ 上出现故障的条件概率, 它正是 $h(t + \Delta t; t)$, 所以

$$P\left(W_{(t,t+\Delta t)} \mid R_t\right) = h(t + \Delta t; t) = \lambda(t)\Delta t + o(\Delta t)$$

故

$$P\left(W^c_{(t,t+\Delta t]} \mid R_t\right) = 1 - P\left(W_{(t,t+\Delta t]} \mid R_t\right) = 1 - \lambda(t)\Delta t + o(\Delta t)$$

由此可见:

$$\frac{g(t+\Delta t) - g(t)}{\Delta t} = -\lambda(t)g(t) + o(1)$$

令 $\Delta t \to 0$, 得一阶变系数常微分方程:

$$g'(t) = -\lambda(t)g(t)$$

在初始条件 $g(0) = 1$ 下, 解此方程即得可靠性函数 $g(t)$ 的显式表示:

$$g(t) = e^{\int_a^t \lambda(s)d}$$

而此例中的可靠性函数 $g(t)$, 也可以解释为一个生物能够生存 t 年的概率.

■

1.7. Bayes' Theorem

Theorem 1.7.1 (The Law of Total Probability)**:::>** Let $A_1, \ldots, A_k$ be a partition of Ω. Then, for any event B,

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$$

Example 1.7.2 (赌徒输光问题)**:::>** 设甲有赌本 $i (i \geqslant 1)$ 元, 其对手乙有赌本 $a - i > 0$ 元. 每赌一次甲以概率 p 赢一元, 而以概率 $q = 1 - p$ 输一元. 假定不欠不借, 赌博一直到甲乙中有一人输光才结束. 因此, 两个人中的赢者最终有总赌资 a 元, 求甲输光的概率.

Solution. 一般地, 我们以 p_i 记甲有赌本 i 元而最终输光的概率. 求此概率的关键是给出下面的事件关系式, 而其方法也被称为首步分析法. 记事件 $A_i =$ " 甲有赌本 i 元而最终输光", $B =$ " 甲第一次赌赢", 则:

$$P(A_i \mid B) = P\left(A_{i+1}\right)$$

$$P(A_i \mid B^c) = P\left(A_{i-1}\right)$$

由上述关系式及全概率公式, 我们得到

$$\begin{aligned} p_i &= P(A_i) = P(A_i \mid B)P(B) + P(A_i \mid B^c)P(B^c) \\ &= P\left(A_{i+1}\right)p + P\left(A_{i-1}\right)q = p\,p_{i+1} + q\,p_{i-1} \end{aligned}$$

这是一个常系数二阶差分方程, 且满足两个边界条件:

$$p_0 = P(\text{“甲有赌本0 元而最终输光"}) = 1$$

$$p_a = P(\text{“甲有赌本}a\text{ 元而最终输光"}) = 0$$

为解这个方程, 注意到它等价于:

$$p\left(p_{i+1} - p_i\right) = q\left(p_i - p_{i-1}\right)$$

故当 $p > 0$ 且 $p \neq \frac{1}{2}$ 时, 由 $p_0 = 1$ 得到

$$\begin{aligned} p_i - p_1 &= \sum_{k=1}^{i-1} \left(\frac{q}{p}\right)^k (p_1 - 1) \\ &= \frac{\frac{q}{p} - \left(\frac{q}{p}\right)^i}{1 - \frac{q}{p}} (p_1 - 1). \end{aligned}$$

令 $i = a$, 再利用 $p_a = 0$ 可解出:

$$p_1 = \frac{\frac{q}{p} - \left(\frac{q}{p}\right)^a}{1 - \left(\frac{q}{p}\right)^a}$$

从而得到, 当 $p > 0$ 且 $p \neq \frac{1}{2}$ (即 $p \neq q$) 时有

$$p_i = \frac{\left(\frac{q}{p}\right)^i - \left(\frac{q}{p}\right)^a}{1 - \left(\frac{q}{p}\right)^a}$$

或者, 可以直接利用差分方程求解的一般方法来进行求解.

而当 $p = q = \frac{1}{2}$ 时, 我们有

$$p_i - p_1 = (i-1)(p_1 - 1)$$

令 $i = a$, 可得 $p_1 = \frac{a-1}{a}$. 从而有 $p_i = 1 - \frac{i}{a}$

由上面的结果可见: 在公平赌博中, 如果两个赌徒初始资本悬殊, 则赌本小的赌徒最终赢的概率是很小的.

上面的结果也适用于甲与资本无限的" 庄家" 赌博的情况 (例如, 设其赌本为 i 元, 而他希望赢到赌资总额为 a 就停止赌博). 此时, 从上面的计算结果还可以预见, 在公平赌博 (即 $p=\frac{1}{2}$ 时), 如果开始的赌本为 i 元, 且赌本相对于希望赢的钱很少, 那么他以几乎为 1 的概率是要以输光告终的. ■

Example 1.7.3 (首步分析法与末步分析法)**:::>** 全概率公式的应用场景常要用到首步分析法与末步分析法. 若后一个试验的结果与前一个试验的结果紧密相关, 前一个试验的全部可能结果就已经组成了 Ω 的一个划分, 故可利用全概率公式得到有关概率的一个递推方程, 解此方程即可获得我们所需的结论, 这种分析方法也称为末步分析法. 而赌徒问题中我们使用的则是首步分析法.

Example 1.7.4 (疯子问题)**:::>** 100 人坐飞机, 他们分别拿到了从 1 号到 100 号的座位, 这些乘客会按号码顺序登机并应当对号入座, 如果他们发现对应号座位被别人坐了, 就会在剩下空的座位随便挑一个坐. 现在假设 1 号乘客疯了 (其他人没疯), 他会在 100 个座位中随便选一个座位坐下, 问: 第 100 人正确坐到自己坐位的概率是多少? (也可推广到 n 名乘客 n 个座位的情况)

Solution. 设第 k 个人 ($2 \leqslant k \leqslant 100$) 找座位时, 他的座位被占的概率为 $P(k)$.

易见, $k=2$ 时, $P(k)=\frac{1}{100}$.

当 $k>2$ 时, 若第 k 个人座位被占, 则可能是第 $1 \sim k-1$ 个人占了他的座位, 把这些概率相加即可.

而第 i 个人 ($2 \leqslant i \leqslant k-1$) 占据第 k 个人的座位的先决条件是第 i 个人的座位被占.

于是有:

$$P(k)=\frac{1}{100}+\sum_{i=2}^{k-1} P(i)\frac{1}{101-i}$$

下面证明: $P(k)=\frac{1}{102-k}$

使用第二数学归纳法:

当 $k=2$ 时, 结论成立;

设 $2\leqslant k\leqslant t-1$ 时, 结论成立, 则当 $k=t$ 时:

$$P(t)=\frac{1}{100}+\sum_{i=2}^{t-1}\frac{1}{(102-i)(101-i)}=\frac{1}{100}+\sum_{i=2}^{t-1}\left(\frac{1}{101-i}-\frac{1}{102-i}\right)=\frac{1}{102-t}$$

结论亦成立!

综上, $P(k)=\frac{1}{102-k}$ 成立. 故 $k=100$ 时, 概率为 $P(100)=\frac{1}{2}$, 所以他占到自己座位的概率也为 $\frac{1}{2}$

更直观地理解这个 $\frac{1}{2}$ 的结论: 本问题等价于这个描述:2-99 号乘客登机后如果发现 1 号 (疯子) 坐在本属于自己的位子上, 就会请疯子离开, 然后疯子再随机找个空座. (就是平时大家在一票一座的交通工具上对号入座时的方式) 这样到 100 号登机时, 2-99 号都在自己座位上, 1 号疯子在自己座位上和 100 号乘客座位上概率相同, 所以是 1/2. 另外这个结果和总乘客数无关, 可以由 100 推广至任意 k.

■

这个问题还可以做一些有意思的扩展:

问: 平均有多少人没有坐到自己的位置?

Solution. 第一个人 (疯子) 没有做到自己位置上的概率是 $\frac{99}{100}$, 而第 k 个人登机时, 他的座位被占的概率是 $\frac{1}{102-k}$:

故可算期望:

$$\begin{aligned}&=\frac{99}{100}+\sum_{k=2}^{100}\frac{1}{102-k}\\&=1+\frac{1}{2}+\frac{1}{3}+\ldots\ldots+\frac{1}{99}\approx 4.6\end{aligned}$$

正是调和级数. ■

推广: n 个人的飞机中, 平均有 $\sum_{i=1}^{n-1}\frac{1}{i}\approx\ln(n-1)$ 个人没有坐到自己的位置.

Example 1.7.5 (Polya 模型)**:::>** 设袋中有 r 个红球, b 个黑球, 每次从中任取一球, 然后放回 (注意是要放回, 否则和一次抽取就没有区别了), 并再放回 $d(\geqslant -1)$ 个与之同色的球. 令 $B_i=$ " 第 i 次取到黑球", $A_k^{(n)}$ " 前 n 次取到 k 个黑球" 则" 前 n 次抽取中的第 n_1 次, $\cdots$, 第 n_k 次抽到黑球, 且第 $m_1,\cdots$, 第 m_{n-k} 次抽到红球" 的概率为 (这里 $\{1,\cdots,n\}=\{n_1,\cdots,n_k\}\cup\{m_1,\cdots,m_{n-k}\}$)

$$\begin{aligned}&P\left(B_{n_1}\cdots B_{n_k}B_{m_1}^c\cdots B_{m_{n-k}}^c\right)\\&=\frac{b(b+d)\cdots(b+[k-1]d)\cdot r(r+d)\cdots(r+[(n-k)-1]d)}{(b+r)(b+r+d)\cdots(b+r+[n-1]d)}\end{aligned}$$

所以

$$\begin{aligned}&P\left(A_k^{(n)}\right)\\=&C_n^k\frac{b(b+d)\cdots(b+[k-1]d)\cdot r(r+d)\cdots(r+[(n-k)-1]d)}{(b+r)(b+r+d)\cdots(b+r+[n-1]d)}\end{aligned}$$

$$P\left(B_{n+1}\,|\,A_k^{(n)}\right)=\frac{b+kd}{b+r+nd}$$

将它们代入全概率公式, 并令 $b'=b+d$, 就得到

$$\begin{aligned}P\left(B_{n+1}\right)&=\sum_{k=0}^{n}P\left(A_k^{(n)}\right)P\left(B_{n+1}\,|\,A_k^{(n)}\right)\\&=\frac{b}{b+r}\sum_{k=0}^{n}C_n^{\star}\frac{b'(b'+d)\cdots(b'+[k-1]d)\cdot r(r+d)\cdots(r+[(n-k)-1]d)}{(b'+r)(b'+r+d)\cdots(b'+r+[n-1]d)}\\&=\frac{b}{b+r}P(\Omega)=\frac{b}{b+r}.\end{aligned}$$

可见我们有

$$P\left(B_{n+1}\right)=P\left(B_1\right)$$

类似地我们可以证明, 对于任意 k 以及两两不相等的 $i_1,\cdots,i_k$ 有

$$P\left(B_{i_1}\cdots B_{i_t}\right)=P\left(B_1\cdots B_k\right)$$

这个特性称为 Polya 模型的可交换性.

事实上, 在摸球问题中我们用排列组合的方法也探讨过这个问题.

Example 1.7.6:::> 袋中装有 a 个白球, b 个黑球. 每次取出一球后, 总是放入一个白球. 这样进行了 n 次以后, 再从袋中取出一个球, 求它是白球的概率?

设第 n 次抽中白球的概率为 P_n, 可知此时的白球有 $(a+b)P_n$ 个, 故

$$P_{n+1}=\frac{P_n\cdot(a+b)P_n+(1-P_n)[(a+b)P_n+1]}{a+b}=\frac{(a+b)P_n+1-P_n}{a+b}=\left(1-\frac{1}{a+b}P_n\right)+\frac{1}{a+b}$$

故:

$$P_n=1-\frac{b}{a+b}\left(\frac{a+b-1}{a+b}\right)^n$$

Theorem 1.7.7 (Bayes' Theorem)**:::>** Let $A_1,\ldots,A_k$ be a partition of Ω such that $\mathbb{P}(A_i)>0$ for each i . If $\mathbb{P}(B)>0$ then, for each $i=1,\ldots,k$,

$$\mathbb{P}(A_i\mid B)=\frac{\mathbb{P}(B\mid A_i)\mathbb{P}(A_i)}{\sum_j\mathbb{P}\left(B\mid A_j\right)\mathbb{P}\left(A_j\right)}$$

We call $\mathbb{P}(A_i)$ the prior probability of A and $\mathbb{P}(A_i\mid B)$ the posterior probability of A.

Proof.

$$\mathbb{P}(A_i\mid B)=\frac{\mathbb{P}(A_iB)}{\mathbb{P}(B)}=\frac{\mathbb{P}(B\mid A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}=\frac{\mathbb{P}(B\mid A_i)\mathbb{P}(A_i)}{\sum_j\mathbb{P}\left(B\mid A_j\right)\mathbb{P}\left(A_j\right)}$$

■

Example 1.7.8 (假阳性问题)**:::>** 在之前, 我们介绍过" 检察官谬误", 现在我们可以通过贝叶斯公式给出更为准确的描述.

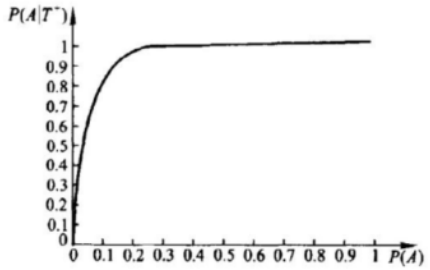
假设艾滋病普查血液试验的灵敏度 (即真有病的人试验结果呈阳性的概率) 为 95%, 因此, 进一步假设接受血液试验的不带艾滋病病毒的人中 99% 的试验结果为阴性, 且人群中每 1000 人中就有一人患艾滋病.

设 $A=$ " 被捡人带有艾滋病病毒", $T^+=$ " 试验结果呈阳性"

我们关心的是条件概率 $P(A \mid T^+)$

$$P(A \mid T^+) = \frac{P(T^+ \mid A)P(A)}{P(T^+ \mid A)P(A) + P(T^+ \mid A^c)P(A^c)}$$
$$= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.01 \times 0.999} \approx 0.087$$

$P(A \mid T^+)$ 的图形表明它是高度依赖于 $P(A)$ 的.



由于申请结婚登记的人群, 并不构成美国社会中艾滋病的" 高危" 人群, 它的发病率甚至比全社会还低, 因此, 在新婚夫妇中实行艾滋病普查的意义不大.

但是, 值得注意的是, 对处于感染艾滋病的" 高危" 人群, 这种普查方式将是很有效的.

那么, 对于假阳性的人群, 有什么能确定自己是否感染艾滋病毒的办法呢? 就是再测一次. 在我们上边的分析中, 已经把先验概率从 0.001 提升到了 0.087, 再测一次的话, " 发病率" 就成为了 0.087, 则准确率将会有显著提升.

Example 1.7.9:::> 从上例的分析可以看出, 贝叶斯统计一个显著的特征就是" 多次算", 这样准确性会大幅提升. 但是, 这就需要用计算机编程反复用新的结果去迭代更新. 所以, 如果先验和后验是同一个分布 (即共轭分布), 这样计算机就非常好算了.

Example 1.7.10 (Jailer's Paradox)**:::>** 也称三囚问题 three prisoners problem, 和 Monty Hall Problem 抽奖问题类似.

Three prisoners, A, B, and C, are in separate cells and sentenced to death.

The governor has selected one of them at random to be pardoned. The warden knows which one is pardoned, but is not allowed to tell. Prisoner A begs the warden to let him know the identity of one of the two who are going to be executed. "If B is to be pardoned, give me C's name. If C is to be pardoned, give me B's name. And if I'm to be pardoned, secretly flip a coin to decide whether to name B or C."

The warden tells A that B is to be executed. Prisoner A is pleased because he believes that his probability of surviving has gone up from 1/3 to 1/2, as it is now between him and C. Prisoner A secretly tells C the news, who reasons that A's chance of being pardoned is unchanged at 1/3, but he is pleased because his own chance has gone up to 2/3. Which prisoner is correct?

Call A, B and C the events that the corresponding prisoner will be pardoned, and b the event that the warden tells A that prisoner B is to be executed, then,

$$\begin{aligned} P(A \mid b) &= \frac{P(b \mid A)P(A)}{P(b \mid A)P(A)+P(b \mid B)P(B)+P(b \mid C)P(C)} \\ &= \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+0\times\frac{1}{3}+1\times\frac{1}{3}} = \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(C \mid b) &= \frac{P(b \mid C)P(C)}{P(b \mid A)P(A)+P(b \mid B)P(B)+P(b \mid C)P(C)} \\ &= \frac{1\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+0\times\frac{1}{3}+1\times\frac{1}{3}} = \frac{2}{3} \end{aligned}$$

The crucial difference making A and C unequal is that $P(b \mid A) = \frac{1}{2}$ but $P(b \mid C) = 1$.

这个问题与 Monty Hall 问题完全是对应的. 想象赦免对应门后有车, 被处决指的是羊, 因而" 更换" 的概率是三分之二.

更直观的解释: A 自己被赦免的概率是三分之一不变, 现在发现 B 被处决, 则被赦免的不是 A 就是 C, 所以 C 就是三分之二了.

Example 1.7.11:::> 设有四箱产品, 次品率分别为 0.1, 0.2, 0.3, 0.4. 现从任一箱中任取一件产品, 检查结果为合格品, 并将此产品放回原箱中, 试求

仍在这箱中任取一件是次品的概率.

设 B_i 为第一次取出的合格品来自 i 箱, B 为再从该箱中取出的产品为次品. 则:

$$P(B)=\sum_{i=1}^{4}(\mathrm{P}(\mathrm{B}_i)\mathrm{P}(\mathrm{B}\mid\mathrm{B}_i))$$

而由于 $P(B\mid B_1)=0.1, P(B\mid B_2)=0.2, P(B\mid B_3)=0.3, P(B\mid B_4)=0.4$ 从题设条件即可得知, 则问题的关键在于如何求 $P(B_i)$.

从贝叶斯的角度来理解这个问题, 当没有做取出合格放回这一动作时, $P(B_i)=0.25$ 对所有 i 均成立, 但是取出合格这一事件改变了概率分布.

设 $\mathrm{A}=\{$ 第一次取出的是合格品 $\}$, $\mathrm{A_i}=\{$ 第一次是从第 i 箱取出的 $\}$, 则:

$$\mathrm{P}(\mathrm{A}_1)=\mathrm{P}(\mathrm{A}_2)=\mathrm{P}(\mathrm{A}_3)=\mathrm{P}(\mathrm{A}_4)=0.25$$

$$\mathrm{P}(A\mid\mathrm{A}_1)=0.9, \mathrm{P}(A\mid\mathrm{A}_2)=0.8, \mathrm{P}(A\mid\mathrm{A}_3)=0.7, \mathrm{P}(A\mid\mathrm{A}_4)=0.6$$

$$P(A)=\sum_{i=1}^{4}(\mathrm{P}(\mathrm{A}_i)\mathrm{P}(A\mid\mathrm{A}_i))=0.75$$

$$P(B_1)=\mathrm{P}(A_1\mid\mathrm{A})=\frac{P(A_1)\mathrm{P}(A\mid\mathrm{A}_1)}{P(A)}=9/30$$

$$P(B_2)=\mathrm{P}(A_2\mid\mathrm{A})=\frac{P(A_2)\mathrm{P}(A\mid\mathrm{A}_2)}{P(A)}=8/30$$

$$P(B_3)=\mathrm{P}(A_3\mid\mathrm{A})=\frac{P(A_3)\mathrm{P}(A\mid\mathrm{A}_3)}{P(A)}=7/30$$

$$P(B_4)=\mathrm{P}(A_4\mid\mathrm{A})=\frac{P(A_4)\mathrm{P}(A\mid\mathrm{A}_4)}{P(A)}=6/30$$

故最后求得, 概率为 $\frac{7}{30}$

2. Random Variables

2.1. Introduction

Definition 2.1.1 (random variable)**:::>** A random variable is a mapping:

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

that assigns a real number $X(\omega)$ to each outcome ω

也即对随机试验的结果进行量化.

Example 2.1.2:::> Given a random variable X and a subset A of the real line, define $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$ and let

$$\mathbb{P}(X \in A) = \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}(\{\omega \in \Omega; X(\omega) \in A\})$$

$$\mathbb{P}(X = x) = \mathbb{P}\left(X^{-1}(x)\right) = \mathbb{P}(\{\omega \in \Omega; X(\omega) = x\})$$

Notice that X denotes the random variable and x denotes a particular value of X

2.2. Distribution Functions of Probability Functions

Definition 2.2.1 (Cumulative Distribution Function)**:::>** The cumulative distribution function, or CDF, is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leqslant x)$$

这个函数又被叫做 distribution function, 说明它对随机变量的研究非常重要.

CDF effectively contains all the information about the random variable. Sometimes we write the CDF as F instead of F_X.

CDF is right continuous(不一定左连续), non-decreasing, and that it is defined for all x.

从定义我们也能看出来, 任何一点随机变量取值的概率等于其分布函数在这一点" 跳跃" 的" 高度". 而连续型随机变量取单点值的概率为 0.

The following result shows that the CDF completely determines the distribution of a random variable.

Theorem 2.2.2:::> Let X have CDF F and let Y have CDF G. If $F(x) = G(x)$ for all x, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all A.

Theorem 2.2.3:::> A function F mapping the real line to $[0,1]$ is a CDF for some probability $\mathbb{P}$ if and only if F satisfies the following three conditions:

1. F is non-decreasing: $x_1 < x_2$ implies that $F(x_1) \leqslant F(x_2)$
2. F is normalized:
$$\lim_{x \to -\infty} F(x) = 0$$
and
$$\lim_{x \to \infty} F(x) = 1$$
3. F is right-continuous: $F(x) = F(x^+)$for all x, where
$$F(x^+) = \lim_{\substack{y \to x \\ y > x}} F(y)$$

Definition 2.2.4 (Countable)**:::>** A set is countable if it is finite or it can be put in a one-to-one correspondence with the integers. The even numbers, the odd numbers, and the rationals are countable; the set of real numbers between 0 and 1 is not countable.

Definition 2.2.5 (Discrete)**:::>** X is discrete if it takes countably many values $\{x_1, x_2, \ldots\}$ We define the probability function or probability mass function for X by $f_X(x) = \mathbb{P}(X = x)$

Thus, $f_X(x) \geqslant 0$ for all $x \in \mathbb{R}$ and $\sum_i f_X(x_i) = 1$. Sometimes we write f instead of f_X. The CDF of X is related to f_X by

$$F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{x_i \leqslant x} f_X(x_i)$$

Definition 2.2.6 (Probability density function)**:::>** A random variable X is continuous if there exists a function f_X such that $f_X(x) \geqslant 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for every $a \leqslant b$,

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$$

The function f_X is called the probability density function (PDF). We have that:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = F_X'(x)$ at all points x at which F_X is differentiable.

Sometimes we write $\int f(x)dx$ or $\int f$ to mean $\int_{-\infty}^{\infty} f(x)dx$

Continuous random variables can lead to confusion. Note that if X is continuous then $\mathbb{P}(X = x) = 0$ for every x. Don't try to think $f(x)$ as $\mathbb{P}(X = x)$. This only holds for discrete random variables. We get probabilities by integrating the PDF. 也就是说, 连续随机变量注重的是区间, 求的值是区间上的积分.

A PDF can be bigger than 1 (unlike a mass function). In fact, a PDF can be unbounded. For example, if $f(x) = (2/3)x^{-1/3}$ for $0 < x < 1$ and $f(x) = 0$ otherwise, then $\int f(x)dx = 1$ even though f is not bounded.

PDF 函数 $f(x)$ 不唯一, 可以差有数个点的取值.

Proposition 2.2.7 (连续型和离散型分布函数的联系与区别)**:::>** 对于连续型与离散型随机变量, 它们的密度或分布列与它们的分布函数是可以互相确定的, 具体如下:

1. 当随机变量 X 为离散型时 (其分布律为 $P(X=x_i)=p_i$), 按分布函数的定义有

$$F(x)=P(X\leqslant x)=\sum_{x_i\leqslant x}p_i$$

此时, $F(x)$ 是一个阶梯函数, 它只在 $x=x_i$ 点上有跳跃, 其跃度为 $p_i=F(x_i)-F\left(x_i^-\right)$. 反之, 如果一个随机变量的分布函数是阶梯函数, 那么此随机变量是离散型的, 此分布函数的所有可能跳跃处, 就是它所有可能的取值, 对应的跃度就是取该值的概率大小.

2. 当随机变量 X 为连续型时, 则有

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

此时 $F(x)$ 不仅是连续的, 而且在 $f(x)$ 的连续点 x 处, 还有导数

$$F'(x)=f(x)$$

反之, 如果一个随机变量的分布函数 $F(x)$ 是分段可微的, 那么, 此随机变量是连续型的, 此时导函数 $F'(x)$ 就是该随机变量的密度 (注意, 除连续点外, 它在其不连续点上的取值并不重要, 任意改动这些点上的取值并不影响概率的计算).

我们强调, 连续型的随机变量的分布函数一定是连续函数, 但若一个随机变量只具有连续的分布函数, 还不能说明它是连续型的随机变量, 因为此分布函数未必能表成一个可积函数的积分. 所以连续型随机变量与分布函数为连续函数的随机变量是两个完全不同的概念, 进一步探讨还需要用到实变函数的知识.

Example 2.2.8 (Uniform distribution)**:::>** Suppose that X has PDF

$$f_X(x)=\begin{cases}1 & \text{for } 0\leqslant x\leqslant 1\\ 0 & \text{otherwise}\end{cases}$$

Clearly, $f_X(x)\geqslant 0$ and $\int f_X(x)dx=1$. A random variable with this density is said to have a Uniform $(0,1)$ distribution. This is meant to capture the idea of

choosing a point at random between 0 and 1. The CDF is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leqslant x \leqslant 1 \\ 1 & x > 1 \end{cases}$$

Lemma 2.2.9:::> Let F be the CDF for a random variable X. Then:

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$where $F(x^-) = \lim_{y \uparrow x} F(y)$
2. $\mathbb{P}(x < X \leqslant y) = F(y) - F(x)$
3. $\mathbb{P}(X > x) = 1 - F(x)$
4. If X is continuous then

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leqslant X < b) \\ &= \mathbb{P}(a < X \leqslant b) = \mathbb{P}(a \leqslant X \leqslant b) \end{aligned}$$

Definition 2.2.10 (对称分布及其期望)**:::>** 对于一个随机变量, 若它概率分布的 PDF 是一个偶函数的话, 则它被称为一个对称分布. 若该随机变量的期望存在, 则期望为 0.

Definition 2.2.11 (Inverse CDF)**:::>** Let X be a random variable with CDF F. The inverse CDF or quantile function is defined by

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

for $q \in [0, 1]$. If F is strictly increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$.

We call $F^{-1}(1/4)$ the first quartile, $F^{-1}(1/2)$ the median (or second quartile), and $F^{-1}(3/4)$ the third quartile.

Definition 2.2.12 (Equal in distribution)**:::>** Two random variables X and Y are equal in distribution, written $X \stackrel{d}{=} Y$, if $F_X(x) = F_Y(x)$ for all x. This does not mean that X and Y are equal. Rather, it means that all probability statements about X and Y will the same.

For example, suppose that $\mathbb{P}(X=1)=\mathbb{P}(X=-1)=1/2$. Let $Y=-X$, then $\mathbb{P}(Y=1)=\mathbb{P}(Y=-1)=1/2$ and so $X \stackrel{d}{=} Y$. But X and Y are not equal. In fact, $\mathbb{P}(X=Y)=0$.

Definition 2.2.13 (随机变量的变异系数)**:::>** 对随机变量 X, 我们称其变异系数为:

$$\frac{\sqrt{DX}}{EX}$$

Definition 2.2.14 (随机变量的偏度)**:::>** 设随机变量 X 标准化后的结果为 $X*$, 我们有:

$$X* = \frac{X-EX}{\sqrt{DX}}$$

有性质:

$$EX* = 0, DX* = 1$$

而随机变量的偏度定义:

$$E\,(x*)^3$$

Definition 2.2.15 (随机变量的峰度)**:::>** 随机变量的峰度即为标准化后的四阶矩:

$$E[X*^4]$$

也有书上把它定义为

$$E[X*^4]-3$$

其中 3 是标准正态分布的峰度.

2.3. Some Important Discrete Random Variables

Example 2.3.1:::> It is traditional to write $X \sim F$ to indicate that X has distribution F. Read $X \sim F$ as " X has distribution F " not as "X is approximately F".

2.3.1 The point mass distribution

Definition 2.3.2 (The point mass distribution)**:::>** X has a point mass distribution at a, written $X \sim \delta_a$, if $\mathbb{P}(X = a) = 1$ in which case

$$F(x) = \begin{cases} 0 & x < a \\ 1 & x \geqslant a \end{cases}$$

The probability mass function is $f(x) = 1$ for $x = a$ and 0 otherwise.

2.3.2 The discrete uniform distribution

Definition 2.3.3 (The discrete uniform distribution)**:::>** Let $k > 1$ be a given integer. Suppose that X has probability mass function given by

$$f(x) = \begin{cases} 1/k & \text{for } x = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases}$$

We say that X has a uniform distribution on $\{1, \ldots, k\}$

2.3.3 The bernoulli distribution

Definition 2.3.4 (The bernoulli distribution)**:::>** Let X represent a binary coin flip. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say that X has a Bernoulli distribution written $X \sim \text{Bernoulli}(p)$. The probability function is $f(x) = p^x (1-p)^{1-x}$ for $x \in \{0, 1\}$.

如果是不标准的 Bernoulli 分布的话, 其形式如下: $\mathbb{P}(X = a) = p$ and $\mathbb{P}(X = b) = 1 - p$, 这时我们可以用 $X' = \frac{X-b}{a-b}$ 来表示, 将其化为标准的 Bernoulli 分布

Example 2.3.5 (n 重 Bernoulli 试验的三个模型)**:::>** n 重 Bernoulli 试验: 将 Bernoulli 试验重复 n 次.

1. n 重 Bernoulli 试验中, 恰好出现 k 次成功的概率为: $C_n^k p^k q^{n-k} \quad (k = 0, 1, \cdots, n)$ 若记 n 重 Bernoulli 试验中, 成功出现的次数为 X, 则:

$$P(X = k) = C_n^k p^k q^{n-k} \quad (k = 0, 1, \cdots, n) \leftrightarrow X \sim B(n, p)$$

二项分布.

2. 进行一系列独立重复的 Bernoulli 试验 (即 ∞ 重 Bernoulli 试验), 首次" 成功" 出现时, 恰好进行了 k 次试验的概率为:

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

若记 ∞ 重 Bernoulli 试验中, 首次" 成功" 出现所需的试验次数为 Y, 则:

$$P(Y = k) = q^{k-1} p \quad (k = 1, 2, \cdots) \leftrightarrow Y \sim Ge(p)$$

几何分布.

3. 进行一系列独立重复的 Bernoulli 试验 (即 ∞ 重 Bernoulli 试验), 第 r 次" 成功" 出现时, 恰好进行了 k 次试验的概率为:

$$C_{k-1}^{r-1} p^r q^{k-r} \quad (k = r, r+1, r+2, \cdots);$$

若记 ∞ 重 Bernoulli 试验中, 第 r 次" 成功" 出现时所需的试验次数为 Z, 则:

$$P(Z = k) = C_{k-1}^{r-1} p^r q^{k-r} \quad (k = r, r+1, r+2, \cdots) \leftrightarrow Z \sim NB(r, p)$$

负二项分布, Pascal 分布.

2.3.4 The binomial distribution

Definition 2.3.6 (The binomial distribution)**:::>** Suppose we have a coin which falls heads up with probability p for some $0 \leqslant p \leqslant 1$. Flip the coin n times and let X be the number of heads. Assume that the tosses are independent. Let $f(x) = \mathbb{P}(X = x)$ be the mass function. It can ben shown that:

$$f(x) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} & \text{for } x = 0, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

A random variable with this mass function is called a Binomial random variable and we write $X \sim \text{Binomial}(n, p)$. If $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$ then $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$.

X is a random variable; x denotes a particular value of the random variable; n and p are parameters, that is, fixed real numbers. The parameter p is usually unknown and must be estimated from data; that's what statistical inference is all about.

期望:

$$
\begin{aligned}
EX &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \\
&= np \left[\binom{n-1}{0} p^0 q^{n-1} + \binom{n-1}{1} p^1 q^{n-2} + \ldots + \binom{n-1}{n-1} p^{n-1} q^0 \right] \\
&= np
\end{aligned}
$$

Example 2.3.7:::> 设随机变量 $X \sim B(n, p)$, 令

$$
Y = \begin{cases} 1, & \text{若} X \text{ 为偶数} \\ -1, & \text{若} X \text{ 为奇数} \end{cases}
$$

求 EY.

本题即为求二项分布中偶项减奇项的概率.

记 p_n 为 X 取偶数的概率, 由于 $X \sim B(n, p)$, 记 $q = 1 - p$, 利用二项展开

公式:

$$(p+q)^n = C_n^0 q^n + C_n^1 p q^{n-1} + \cdots + C_n^n p^n$$

$$(q-p)^n = C_n^0 q^n - C_n^1 p q^{n-1} + \cdots + (-1)^n C_n^n p^n$$

因此, 可得 $p_n = \frac{1}{2}[(p+q)^n + (q-p)^n] = \frac{1}{2}[1+(1-2p)^n]$. 进而可得 $EY = (1-2p)^n$.

Example 2.3.8:::> 进行某种独立重复的试验, 每次试验成功的概率为 0.6, 失败的概率为 0.4 , 问在 3 次失败之前已经取得了 4 次成功的概率是多少?

可以这样理解: 一共进行了 7 次试验, 最后一次一定失败, 前六次中有 4 次成功, 所以是

$$C_6^2(0.4)^3 \times (0.6)^4 = 0.1244$$

Example 2.3.9:::> 二项分布 $B(n,p)$ 的最大可能值 $k*$ 存在, 且

$$k^* = \begin{cases} (n+1)p \text{ 或} (n+1)p-1, & \text{当} (n+1)p \text{ 为整数时,} \\ [(n+1)p], & \text{当} (n+1)p \text{ 为非整数时.} \end{cases}$$

证明: 解不等式

$$\begin{cases} P(x=k^*) \geqslant P(x=k^*+1) \\ P(x=k^*) \geqslant P(x=k^*-1) \end{cases}$$

得:

$$(n+1)p - 1 \leqslant k^* \leqslant (n+1)p$$

2.3.5 The geometric distribution

Definition 2.3.10 (The geometric distribution)**:::>** X has a geometric distribution with parameter $p \in (0,1)$, written $X \sim \text{Geom}(p)$, if

$$\mathbb{P}(X=k) = p(1-p)^{k-1}, \quad k \geqslant 1$$

We have that

$$\sum_{k=1}^{\infty} \mathbb{P}(X=k) = p\sum_{k=1}^{\infty}(1-p)^k = \frac{p}{1-(1-p)} = 1$$

Think of X as the number of flips needed until the first head when flipping a coin.

期望:

$$E(X)=\sum_{k=1}^{\infty}k(1-p)^{k-1}p=p\sum_{k=1}^{\infty}k(1-p)^{k-1}$$

令 $q=1-p, S=\sum_{k=1}^{\infty}kq^{k-1}$

$$S=1+2q+3q^2+4q^3+\ldots+kq^{k-1}, k=\infty$$
$$qS=q+2q^2+3q^3+4q^4+\ldots+kq^{k}, k=\infty$$

上式减下式得 (错位相减):

$$(1-q)S=1+q+q^2+q^3+\ldots+q^{k-1}-kq^k$$

$$S=\frac{1-q^k}{(1-q)^2}-\frac{kq^k}{1-q}=\frac{1}{(1-q)^2}=\frac{1}{p^2}$$

注意这里用到了 $k=\infty$ 则:

$$E(X)=p\frac{1}{p^2}=\frac{1}{p}$$

上面求解求和时错位相减也可以替换为 $\left(\sum_{i=1}^{\infty}x^k\right)'$ 来求解.

Theorem 2.3.11:::> 设 X 为只取正整数的随机变量, 则下列命题等价:

1. X 服从几何分布
2. $P(X>m+n\mid X>n)=P(X>m)\quad m,n=0,1,\cdots$ 这个特性被称为"无记忆性"
3. $P(X=m+n\mid X>n)=P(X=m)\quad m=1,2,\cdots,n=0,1,\cdots$

Proof. 1. $(1)\Rightarrow(2)$

由于 X 服从几何分布, 即 $P(X=k)=qp^{k-1}, k=1,2,\cdots$.

$$\begin{aligned}P(X>m+n\mid X>n)&=\frac{P(X>m+n, X>n)}{P(X>n)}\\&=\frac{P(X>m+n)}{P(X>n)}=\frac{\sum_{k=m+n+1}^{\infty}qp^{k-1}}{\sum_{k=n+1}^{\infty}qp^{k-1}}\\&=q^m=P(X>m)\end{aligned}$$

2. $(2)\Rightarrow(3)$

由

$$\frac{P(X>m+n)}{P(X>n)}=P(X>m)\quad m,n=0,1,\cdots$$

得

$$\frac{P(X>m-1+n)}{P(X>n)}=P(X>m-1)\quad m=1,2,\cdots;n=0,1,\cdots$$

注意到 X 只取正整数值, 故 $P(X=m)=P(X>m-1)-P(X>m)$, 后式减前式得

$$\frac{P(X=m+n)}{P(X>n)}=P(X=m)\quad m=1,2,\cdots;n=0,1,\cdots$$

即 $P(X=m+n\mid X>n)=P(X=m)\quad m=1,2,\cdots,n=0,1,\cdots$

3. $(3)\Rightarrow(1)$

由于 $\frac{P(X=m+n)}{P(X>n)}=P(X=m)\quad m=1,2,\cdots;n=0,1,\cdots$ 所以 $P(X=m+n)=P(X=m)P(X>n)$ 令 $G(m)=P(X>m), F(m)=P(X=m)\quad m=1,2,\cdots$ 则上式可化为

$$F(m+n)=G(n)F(m)\text{ 且}G(1)+F(1)=1$$

从而

$$P(X=k)=F(k)=G(1)F(k-1)=(G(1))^2F(k-2)=\cdots=(G(1))^{k-1}F(1)\quad k=1,2,\cdots$$

即 $X\sim Ge(F(1))$

■

Definition 2.3.12 (带截止的几何分布)**:::>** 背景: 做一系列独立重复的 Bernoulli 试验, 直到首次试验成功时或试验进行到第 m 次为止, 所需的试验的次数

$$P(X=k)=\begin{cases} q^{k-1}p, & k=1,2,\cdots,m-1 \\ q^{m-1}, & k=m \end{cases}$$

2.3.6 The Hypergeometric distribution

Definition 2.3.13 (Hypergeometric distribution)**:::>** In probability theory and statistics, the hypergeometric distribution is a discrete probability distribution that describes the probability of k successes (random draws for which the object drawn has a specified feature) in n draws , without replacement, from a finite population of size N that contains exactly K objects with that feature. , wherein each draw is either a success or a failure. In contrast, the binomial distribution describes the probability of k successes in n draws, with replacement.

$$p_X(k)=\frac{\begin{pmatrix} K \\ k \end{pmatrix}\begin{pmatrix} N-K \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

Example 2.3.14:::> 我们常提到可以用放回的情况 (即二项分布) 去近似不放回的情况 (超几何分布), 下面我们从数学上给以证明.

设 $T=M+N$, 且 X 服从超几何分布, 即

$$P(X=k)=\frac{C_M^k C_N^{n-k}}{C_{M+N}^n}, \quad k=0,1,2,\cdots,\min(M,n)$$

若 $\lim_{T\to\infty}\frac{M}{T}=p$, 则 $\lim_{T\to\infty}P(X=k)=C_n^k p^k(1-p)^{n-k}, k=0,1,2\cdots,n.$

𝔓roof. 由于当 $T \to \infty$ 时, $M \to \infty$, 故

$$\frac{C_M^k C_N^{n-k}}{C_{M+N}^n} =$$

$$C_n^k \left(\frac{M}{T}\right)^k \left(1-\frac{M}{T}\right)^{n-k} \frac{\left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)\cdots\left(1-\frac{k-1}{M}\right)\left(1-\frac{1}{N}\right)\left(1-\frac{2}{M}\right)\cdots\left(1-\frac{n-k-1}{n}\right)}{\left(1-\frac{1}{T}\right)\left(1-\frac{2}{T}\right)\cdots\left(1-\frac{n-1}{T}\right)}$$

当 $T \to \infty$ 时, 上式趋向于 $C_n^k p^k (1-p)^{n-k}$ ■

2.3.7 The Poisson distribution

𝔇efinition **2.3.15** (The Poisson distribution)**:::>** X has a Poisson distribution with parameter λ, written $X \sim \text{Poisson}(\lambda)$ if

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad x \geqslant 0$$

Note that

$$\sum_{x=0}^{\infty} f(x) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1$$

The Poisson is often used as a model for counts of rare events like radioactive decay and traffic accidents.

泊松分布是二项分布 n 很大而 p 很小时的一种极限形式. 二项分布是说, 已知某件事情发生的概率是 p, 那么做 n 次试验, 事情发生的次数就服从于二项分布. 而泊松分布是指某段连续时间内某件事情发生的次数. 而且“某件事情”发生所用的时间是可以忽略的. 例如, 在五分钟内, 电子元件遭受脉冲的次数, 就服从于泊松分布.

假如你把“连续的时间”分割成无数小份, 那么每个小份之间都是相互独立的。在每个很小的时间区间内, 电子元件都有可能“遭受到脉冲”或者“没有遭受到脉冲”, 这就可以被认为是一个 p 很小的二项分布. 而因为“连续的时间”被分割成无穷多份, 因此 n(试验次数) 很大. 所以, 泊松分布可以认为是二项分布的一种极限形式.

因为二项分布其实就是一个最最简单的“发生”与“不发生”的分布,

它可以描述非常多的随机的自然界现象, 因此其极限形式泊松分布自然也是非常有用的.

Theorem 2.3.16:::> If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$

Proof. 由于之前讲过的卷积性质:

$$P(X+Y) = P(X) * P(Y)$$

则:

$$M_{X+Y}(u) = M_X(u) \times M_Y(u) = e^{\lambda_1(e^u-1)} \times e^{\lambda_2(e^u-1)} = e^{(\lambda_1+\lambda_2)(e^u-1)}$$

■

Theorem 2.3.17:::> 若:

$$P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0, 1, 2, \cdots$$

则

$$EX = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}e^{-\lambda} = \lambda$$

最后一个等号运用了泰勒展开公式.

Theorem 2.3.18 (Poisson 分布的矩母函数)**:::>**

$$M_X(u) = E[e^{uX}] = \sum_{k=0}^{\infty} e^{uk}\frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{[\lambda e^u]^k}{k!}$$

由泰勒展开求和公式:

$$\sum_{k=0}^{\infty}\frac{X^k}{k!} = e^X$$

故:

$$M_X(u) = e^{-\lambda}e^{\lambda e^u} = e^{\lambda(e^u-1)}$$

Example 2.3.19 (泊松分布的应用)**:::>** 我们关心在一个小时之内到达某银行的客户数, 客户数为可数集, 样本空间为 $\Omega = \{0, 1, 2\ldots\}$. 取一个非常大的自然数 n, 我们可以把一个小时分解为等长的 n 段, 即 $\left(0, \frac{1}{n}\right], \left(\frac{1}{n}, \frac{2}{n}\right] \ldots \left(\frac{n-1}{n}, 1\right]$. 当 n 很大时, 一个区间段内有两个客户到达的概率几乎可以忽略不计. 假设

每段时间客户到达的概率相等, 且反比于 n, 不妨假设为 $\frac{\lambda}{n}$, 那么一小时内总的人数:

$$P^*(\{k\}) = \begin{pmatrix} n \\ k \end{pmatrix} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

令 $n \to \infty$, 则 $\frac{\begin{pmatrix} n \\ k \end{pmatrix}}{n^k} = \frac{n!}{k!\cdot(n-k)!\cdot n^k} \to \frac{1}{k!}$, $\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$, 因而

$$\mathscr{P}(\{k\}) = \frac{\lambda^k}{k!} e^{-\lambda} > 0$$

且 $\sum_{k=0}^{\infty} \mathscr{P}(\{k\}) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1$($e^{\lambda x}$ 在 $x = 0$ 处泰勒展开可得).

从以上推导我们可以看出, 这里有两个潜在的假设:

1. 每个小区间段是否有客人到达是独立的
2. 每个小区间段客人到达的概率反比于 n.

Proposition 2.3.20:::> 从上面的分析我们可以看出, 泊松分布适用于 $X \sim B(n,p)$ 而 n 很大, p 很小, np 不大也不小的情况. 若是 n 很大, p 不小呢? 我们之后可以用正态分布来解决.

Example 2.3.21 (随机分流的不变性)**:::>** 随机分流的不变性, 即 Poisson 分布在随机选择下的不变性.

假设某段时间里来百货公司的顾客数服从参数为 λ 的 Poisson 分布, 而在百货公司里每个顾客购买电视机的概率为 p, 且每个顾客是否购买电视机是独立的. 问在这段时间内, 百货公司售出 k 台电视机的概率有多大 (这里假定每人最多购买一台电视机).

记 X 为百货公司售出电视机的台数, 而 N 为这段时间内进入百货公司的人数.

记 X 为百货公司售出电视机的台数, 而 N 为这段时间内进入百货公司

的人数, 故由全概率公式知:

$$P(X=k)=\sum_{n=0}^{\infty}P(X=k\mid N=n)P(N=n)=\sum_{n=0}^{\infty}P(X=k\mid N=n)\frac{\lambda^n \mathrm{e}^{\lambda}}{n!}$$

由于在已知有 $N=n$ 名顾客进入百货公司的条件下, 百货公司售出电视机的台数服从参数为 n 和 p 的二项分布, 即

$$P(X=k\mid N=n)=\begin{cases} C_n^v p^k(1-p)^{nk}, & n\geqslant k \\ 0, & n<k \end{cases}$$

故

$$\begin{aligned} P(X=k)&=\sum_{n=k}^{\infty}C_n^k p^k(1-p)^{n-k}\frac{\lambda^n e^{-\lambda}}{n!} \\ &=\sum_{n=k}^{\infty}\frac{n!}{k!(n-k)!}\cdot\frac{(\lambda p)^k[\lambda(1-p)]^{n-k}\mathrm{e}^{-\lambda}}{n!} \\ &=\frac{(\lambda p)^k \mathrm{e}^{\lambda}}{k!}\sum_{n=k}^{\infty}\frac{[\lambda(1-p)]^{n-k}}{(n-k)!} \\ &=\frac{(\lambda p)^k \mathrm{e}^{\lambda}}{k!}\sum_{i=0}^{\infty}\frac{[\lambda(1-p)]^{i}}{i!} \\ &=\frac{(\lambda p)^k}{k!}\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda(1-p)}=\frac{(\lambda p)^k}{k!}\mathrm{e}^{-\lambda p}, \end{aligned}$$

即 X 服从参数为 λp 的 Poisson 分布.

由于这时间段里来百货公司的顾客数服从参数为 λ 的 Poisson 分布, 而顾客在百货公司里是否购买电视机是随机的, 其概率为 p, 则在这段时间内, 购买电视机的顾客数仍服从 Poisson 分布 (参数为 λp). 这一性质就是所谓的 Poisson 分布在随机选择下的不变性.

Theorem 2.3.22 (即 Poisson 分布在随机选择下的不变性定理)**:::>** 设 $X\sim \mathrm{Poisson}_{\lambda}$($X$ 表示某一群体的个数), 从此 X 个个体中逐个独立地以保留概率 p 筛选得到 Y 个个体, 则 $Y\sim \mathrm{Poisson}_{\lambda p}$.

Example 2.3.23:::> 在某一地区中, 雌昆虫的数目服从均值为 λ 的 Poisson 分布, 每只雌昆虫下卵的数目服从均值为 μ 的 Poisson 分布. 求该地区虫卵数目的分布.

设雌虫数目为 N, 则

$$P(N=i)=\frac{\lambda^i}{i!}e^{-\lambda}$$

而设一个雌虫下卵的数目为 k, 则

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

同时注意, i 个雌虫下卵的分布也符合泊松分布, 事实上是改变了泊松分布的 λ(根据 np 联想), 有:

$$P(X=k)=\frac{(i\mu)^k e^{-\mu i}}{k!}$$

由全概率公式, 我们有

$$P(Y=k)=\sum_{n=0}^{\infty}P(X=n)P(Y=k\mid X=n)$$

因而

$$\begin{aligned}P(X=k)&=\sum_{i=0}^{\infty}\frac{(i\mu)^k e^{-\mu i}}{k!}\frac{(\lambda)^i e^{-\lambda}}{i!}\\&=\frac{\mu^k e^{-\lambda}}{k!}\sum_{i=0}^{\infty}\frac{(\lambda e^{-\mu})^i (i)^k}{(i)!}\end{aligned}$$

Example 2.3.24:::> 设某个电脑公司销售了 200 台电脑, 该公司电脑的故障发生率为 0.01, 且每台电脑的故彰可由一个维修人员处理, 在公司的售后维修方案中, 公司考虑两种维修人员的安排方法:

1. 由 5 人来完成维修任务, 每人负责 40 个固定客户和电脑
2. 由 4 人来共同完成 200 台电脑的维修任务

试比较这两种安排哪个更妥当些.

问题的关键是要比较这两种安排下, 电脑发生故彰而得不到及时维修的概率的大小.

我们先来考虑第一种安排方案, 以 A_i 表示事件“第 i 个人负责的 40 台电脑中, 有电脑发生故障而得不到及时维修", 则该公司销售的 200 台电脑中有电脑发生故障而得不到及时维修的概率为 $P(A_1\cup A_2\cup A_3\cup A_4\cup A_5)$.

记 X 为" 第 1 个人负责的 40 台电脑中, 在某一小单位时间段内发生故障的台数" , 则随机变量 X 近似服从参数为 $\lambda=np=40\times 0.01=0.4$ 的 Poisson

分布, 于是

$$P(A_1)=P(X\geqslant 2)\approx\sum_{k=2}^{\infty}\frac{(0.4)^k\mathrm{e}^{0.4}}{k!}=1-\mathrm{e}^{0.4}-0.4\mathrm{e}^{-0.4}=0.062$$

故

$$P\left(\bigcup_{i=1}^{5}A_i\right)=1-P\left(\bigcap_{i=1}^{c}A_i^5\right)\approx 1-(1-0.062)^5\approx 0.274$$

对第二种安排方案, 记 Y 为这 200 台电脑在同一小单位时间段发生故障的台数, 则随机变量 Y 近似服从 $\lambda=np=2$ 的 Poisson 分布. 故有电脑发生故障而得不到及时维修的概率为

$$P(Y\geqslant 5)\approx\sum_{k=5}^{\infty}\frac{2^k\mathrm{e}^{-2}}{k!}=1-\sum_{k=0}^{4}\frac{2^k\mathrm{e}^{-2}}{k!}\approx 0.053\ll 0.274$$

这个结果正说明了协作的优点.

Example 2.3.25:::> 把一硬币抛掷 N 次, N 为随机的, 且服从参数为 λ 的 Poisson 分布. 硬币出现正面的概率为 p, 记 X 与 Y 分别为抛掷中正面和反面的次数.

1. 求 X 与 Y 的分布
2. 问 X 与 Y 是否相互独立, 说明其理由

第一题比较简单.

第二题需要用到一些技巧. 考虑:

$$P(X=x,Y=y)=\sum_{k=1}^{\infty}P(X=x,Y=y\mid N=k)P(N=k))$$

且只有当 $k=x+y$ 时, $P(X=r,Y=y\mid N=k)$ 才不为 0. 若两者独立, 两者分布律都不出现概率为 0 的情况, 因而相乘后也不会为 0. 和上述矛盾.

Example 2.3.26:::> 设 X_1,X_2 是独立同分布随机变量, 均服从参数为 $\lambda>0$ 的 Poisson 分布. 令

$$M=\max(X_1,X_2),N=\min(X_1,X_2)$$

分别求 M 与 N 的概率分布.

$P(M=m)$ 可以分为 $x_1=m, x_2<m$ 或 $x_1<m, x_2=m$ 或 $x_1=x_2=m$ 三种情况.

由于 $x_1, x_2 i.i.d$, 则:

$$P(M=m)=\frac{2e^{-\lambda}\lambda^m}{m!}\cdot\sum_{i=0}^{m-1}\frac{e^{-\lambda}\cdot\lambda^i}{i!}+\frac{\lambda^{2m}e^{-2\lambda}}{(m!)^2}$$

2.4. Bivariate Distributions

Definition 2.4.1 (Bivariate distributions)**:::>** Given a pair of discrete random variables X and Y, define the joint mass function by $f(x,y)=\mathbb{P}(X=xandY=y)$ We write f as $f_{X,Y}$ when we want to be more explicit.

Example 2.4.2:::> Here is a bivariate distribution for two random variables X and Y each taking values 0 or 1:

	$Y=0$	$Y=1$	
X = 0	1/9	2/9	1/3
X = 1	2/9	4/9	2/3
	1/3	2/3	1

Thus, $f(1,1)=\mathbb{P}(X=1,Y=1)=4/9$.

Definition 2.4.3 (联合分布函数 joint distribution function)**:::>** 对任意实数 x,y, 称函数 $F(x,y)=P(X\leqslant x, Y\leqslant y)$ 为二维随机变量 (X,Y) 的联合分布函数.

联合分布函数具有如下性质:

1. $F(x,y)$ 是 x 或 y 的不减函数
2. $0\leqslant F(x,y)\leqslant 1$, 且 $F(x,-\infty)=F(-\infty,y)=F(-\infty,-\infty)=0, F(+\infty,+\infty)=1$
3. $F(x,y)$ 关于 x 或 y 是右连续的

4. 对任意 $x_1 \leqslant x_2, y_1 \leqslant y_2$, $\quad F(x_2,y_2)-F(x_2,y_1)-F(x_1,y_2)+F(x_1,y_1) \geqslant 0$

Definition 2.4.4:::> In the continuous case, we call a function $f(x,y)$ a PDF for the random variables (X,Y) if

1. $f(x,y) \geqslant 0$ for all (x,y)
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$
3. for any set $A \subset \mathbb{R}\times\mathbb{R}, \mathbb{P}((X,Y)\in A) = \iint_A f(x,y)dxdy$

In the discrete or continuous case we define the joint CDF as $F_{X,Y}(x,y) = \mathbb{P}(X \leqslant x, Y \leqslant y)$

Example 2.4.5:::> Let (X,Y) have density

$$f(x,y) = \begin{cases} x+y & \text{if } 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned}\int_0^1\int_0^1 (x+y)dxdy &= \int_0^1\left[\int_0^1 xdx\right]dy + \int_0^1\left[\int_0^1 ydx\right]dy \\ &= \int_0^1 \frac{1}{2}dy + \int_0^1 ydy = \frac{1}{2}+\frac{1}{2} = 1\end{aligned}$$

which verifies that this is a PDF.

Example 2.4.6:::> 将一粒均匀骰子抛掷 n 次, 将所得的 n 个点数的最小值记为 X, 最大值记为 Y. 分别求出 X 与 Y 的分布列, 并求 $P(X=2,Y=5)$.

$$P(X=k) = \left(\frac{6-k+1}{6}\right)^n - \left(\frac{6-k}{6}\right)^n$$

$$P(Y=k) = \left(\frac{k}{6}\right)^n - \left(\frac{k-1}{6}\right)^n \quad k = 1,2,\cdots,6;$$

$$P(X=2,Y=5) = \frac{4^n - 2\cdot 3^n + 2^n}{6^n}$$

首先注意, X 与 Y 并不是独立的. 其次, 这里的运算非常类似集合的运算, 最后一个公式即类似容斥原理.

2.5. Marginal Distributions

Definition 2.5.1 (Marginal mass function)**:::>** If (X,Y) have joint distribution with mass function $f_{X,Y}$, then the marginal mass function for X is defined by

$$f_X(x)=\mathbb{P}(X=x)=\sum_y \mathbb{P}(X=x,Y=y)=\sum_y f(x,y)$$

and the marginal mass function for Y is defined by

$$f_Y(y)=\mathbb{P}(Y=y)=\sum_x \mathbb{P}(X=x,Y=y)=\sum_x f(x,y)$$

Definition 2.5.2 (边缘分布函数)**:::>**

$$F_X(x)\triangleq F(x,+\infty)=P(X\leqslant x)$$

称为 X 的边缘分布函数.

$$F_Y(y)\triangleq F(y,+\infty)=P(Y\leqslant y)$$

称为 Y 的边缘分布函数.

Definition 2.5.3:::> For continuous random variables, the marginal densities are

$$f_X(x)=\int f(x,y)dy$$

and

$$f_Y(y)=\int f(x,y)dx$$

The corresponding marginal distribution functions are denoted by F_X and F_Y.

2.6. Independent Random Variables

Definition 2.6.1:::> Two random variables X and Y are independent if, for every A and B,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

and we write $X \amalg Y$

In principle, to check whether X and Y are independent, we need to check the equation above for all subsets A and B. Fortunately, we have the following result which we state for continuous random variables though it is true for discrete random variables too.

Theorem 2.6.2:::> Let X and Y have joint PDF $f_{X,Y}$. Then X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all values x and y.

Example 2.6.3:::> 掷两颗均匀骰子, 记第一颗骰子出现的点数为 X, 而两颗骰子中点数的最大值为 Y, 求 (X,Y) 的联合分布律并判断它们是否独立.

X	1	2	3	4	5	6	$p_{i\cdot}$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
2	0	2/36	1/36	1/36	1/36	1/36	1/6
3	0	0	3/36	1/36	1/36	1/36	1/6
4	0	0	0	4/36	1/36	1/36	1/6
5	0	0	0	0	5/36	1/36	1/6
6	0	0	0	0	0	6/36	1/6
$p_{\cdot j}$	1/36	3/36	5/36	7/36	9/36	11/36	1

可知, X 与 Y 不相互独立.

Example 2.6.4:::> Suppose that X and Y are independent and both have the same density

$$f(x) = \begin{cases} 2x & \text{if } 0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

Let us find $\mathbb{P}(X+Y\leqslant 1)$. Using independence, the joint density is

$$f(x,y)=f_X(x)f_Y(y)=\begin{cases}4xy & \text{if } 0\leqslant x\leqslant 1,\quad 0\leqslant y\leqslant 1\\ 0 & \text{otherwise}\end{cases}$$

Now,

$$\begin{aligned}\mathbb{P}(X+Y\leqslant 1)&=\iint_{x+y\leqslant 1} f(x,y)dydx\\ &=4\int_0^1 x\left[\int_0^{1-x} ydy\right]dx\\ &=4\int_0^1 x\frac{(1-x)^2}{2}dx=\frac{1}{6}\end{aligned}$$

Theorem 2.6.5:::> Suppose that the range of X and Y is a (possibly infinite) rectangle. If $f(x,y)=g(x)h(y)$ for some functions g and h (not necessarily probability density functions) then X and Y are independent.

Example 2.6.6:::> Let X and Y have density

$$f(x,y)=\begin{cases}2e^{-(x+2y)} & \text{if } x>0 \text{ and } y>0\\ 0 & \text{otherwise.}\end{cases}$$

The range of X and Y is the rectangle $(0,\infty)\times(0,\infty)$. We can write $f(x,y)=g(x)h(y)$ where $g(x)=2e^{-x}$ and $h(y)=e^{-2y}$. Thus, X and Y are independent.

2.7. Expectation and Variance

2.7.1 Expectations of random variables

Definition 2.7.1 (Expectation)**:::>** The expected value, or mean, or first moment, of a random variable X is defined to be

$$\mathbb{E}(X)=\int x\,dF(x)=\begin{cases}\sum_x xf(x) & \text{if } X \text{ is discrete}\\ \int xf(x)dx & \text{if } X \text{ is continuous}\end{cases}$$

assuming that the sum(or integral) is well defined. We use the following notation to denote the expected value of X:

$$\mathbb{E}(X) = \mathbb{E}X = \int x dF(x) = \mu = \mu_X$$

The fact that $\mathbb{E}(X) \approx \sum_{i=1}^{n} X_i/n$ is actually more than a heuristic; it is a theorem called the law of large numbers.

The notation $\int x dF(x)$ is merely a convenient unifying notation so we don't have to write $\sum_x x f(x)$ for discrete random variables and $\int x f(x) dx$ for continuous random variables.

To ensure that $\mathbb{E}(X)$ is well defined, we say that $\mathbb{E}(X)$ exists if $\int_x |x| dF_X(x) < \infty$. Otherwise we say that the expectation does not exist.

Example 2.7.2:::> Recall that a random variable has a Cauchy distribution if it has density $f_X(x) = \{\pi(1+x^2)\}^{-1}$.

$$\int |x| dF(x) = \frac{2}{\pi} \int_0^\infty \frac{x dx}{1+x^2} = \left[x \tan^{-1}(x)\right]_0^\infty - \int_0^\infty \tan^{-1} x dx = \infty$$

so the means does not exist.

Theorem 2.7.3:::> 设连续随机变量 X 的分布函数为 $F(x)$, 且数学期望存在, 则:

$$E(X) = \int_0^\infty [1 - F(x)] dx - \int_{-\infty}^0 F(x) dx$$

也可写成:

$$EX = \int_0^\infty P(X > x) dx - \int_{-\infty}^0 P(X \leqslant x) dx$$

Proof.

$$E(X) = \int_{-\infty}^\infty x p(x) dx = \int_{-\infty}^0 x p(x) dx + \int_0^\infty x p(x) dx$$

将第一个积分写为二次积分, 然后改变积分次序, 有:

$$\begin{aligned}\int_{-\infty}^{0} x\,p(x)dx &= -\int_{-\infty}^{0}\left(\int_{x}^{0} dy\right)p(x)dx \\ &= -\int_{-\infty}^{0}\int_{-\infty}^{y} p(x)dxdy \\ &= -\int_{-\infty}^{0} F(y)dy\end{aligned}$$

第二个积分亦可写成二次积分, 然后改变积分次序, 可得:

$$\int_{0}^{\infty} x\,p(x)dx = \int_{0}^{\infty}[1-F(y)]dy$$

■

Theorem 2.7.4:::> 对于非负的连续型的随机变量有:

$$E(X)=\int_{0}^{\infty} P(X>x)dx$$

Theorem 2.7.5:::> 类似地, 对于非负的随机变量 $X\geqslant 0$, 我们有:

$$E\left(X^{k}\right)=\int_{0}^{+\infty} kx^{k-1}P(X>x)dx$$

证明的话也很类似, kx^{k-1} 视为积分, 然后交换积分

Theorem 2.7.6:::> 设 X 是仅取非负整数的离散型随机变量, 若期望存在, 则:

$$E(X)=\sum_{k=1}^{+\infty}P(X\geqslant k)$$

Proof.

$$\begin{aligned}E(X)=\sum_{k=0}^{+\infty}kP(X=k)&=\sum_{k=1}^{+\infty}\left[\sum_{i=1}^{k}P(X=k)\right] \\ =\sum_{i=1}^{+\infty}\left[\sum_{k=i}^{+\infty}P(X=k)\right]&=\sum_{k=1}^{+\infty}P(X\geqslant k)\end{aligned}$$

其中, 我们交换了求和顺序, 即对于一般式 $f(x)$, 我们有:

$$\sum_{k=1}^{+\infty}\left[\sum_{i=1}^{k}f(x)\right]=\sum_{i=1}^{+\infty}\left[\sum_{k=i}^{+\infty}f(x)\right]$$

这个交换求和顺序与我们上面交换二元积分的积分顺序有异曲同工之妙.

或者更直观地:

$$\begin{aligned} EX &= \sum_{n=1}^{\infty} nP(X=n) \\ &= P(X=1)+P(X=2)+\cdots+P(X=n)+\cdots \\ &\quad +P(X=2)+\cdots+P(X=n)+\cdots \\ &\quad +\cdots\cdots \\ &\quad +P(X=n)+\cdots \\ &\quad +\cdots \end{aligned}$$

最终结果为 $=\sum_{n=1}^{\infty} P(X \geqslant n)$ ■

Theorem 2.7.7:::>

$$\sum_{k=0}^{\infty} kP(X>k) = \frac{1}{2}\left[E\left(X^2\right)-E(X)\right]$$

Proof.

$$\begin{aligned} \sum_{k=0}^{\infty} kP(X>k) &= \sum_{k=0}^{\infty} k \sum_{i=k+1}^{\infty} P(X=i) \\ &= \sum_{i=1}^{\infty} P(X=i)\frac{(i-1)i}{2} \\ &= \frac{1}{2}\sum_{i=1}^{\infty} i^2 P(X=i) - \frac{1}{2}\sum_{i=1}^{\infty} iP(X=i) \\ &= \frac{1}{2}\left[E\left(X^2\right)-E(X)\right] \end{aligned}$$

■

Theorem 2.7.8 (The rule of the Lazy Statistician)**:::>** It is also known as the law of the unconscious statistician, or LOTUS.

Let $Y = r(X)$. Then

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x) dF_X(x)$$

Here is a special case. Let A be an event and let $r(x) = I_A(x)$ where $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$. Then

$$\mathbb{E}(I_A(X)) = \int I_A(x) f_X(x) dx = \int_A f_X(x) dx = \mathbb{P}(X \in A)$$

In other words, probability is a special case of expectation.

但是注意, 对于随机变量 X 来说, $r(X)$ 未必是随机变量, 需要满足函数 r 是可测的这一条件.

Example 2.7.9:::> Let $X \sim \text{Unif}(0,1)$. Let $Y = r(X) = e^X$. Then

$$\mathbb{E}(Y) = \int_0^1 e^x f(x) dx = \int_0^1 e^x dx = e - 1$$

Alternatively, you cold find $f_Y(y)$ which turns out to be $f_Y(y) = 1/y$ for $1 < y < e$. Then, $\mathbb{E}(Y) = \int_1^e y f(y) dy = e - 1$

注意, 这里要转换成 CDF 才能判断随机变量 Y 的表达式.

Example 2.7.10:::> Take a stick of unit length and break it at random. Let Y be the length of the longer piece. What is the mean of Y? If X is the break point then $X \sim \text{Unif}(0,1)$ and $Y = r(X) = \max\{X, 1-X\}$. Thus, $r(x) = 1 - x$ when $0 < x < 1/2$ and $r(x) = x$ when $1/2 \leqslant x < 1$. Hence,

$$\mathbb{E}(Y) = \int r(x) dF(x) = \int_0^{1/2} (1-x) dx + \int_{1/2}^1 x dx = \frac{3}{4}$$

Example 2.7.11:::> Functions of several variables are handled in a similar way. If $Z = r(X,Y)$, then

$$\mathbb{E}(Z) = \mathbb{E}(r(X,Y)) = \iint r(x,y) dF(x,y)$$

Let (X,Y) have a jointly uniform distribution on the unit square. Let $Z = r(X,Y) = X^2 + Y^2$. Then

$$\begin{aligned} \mathbb{E}(Z) &= \iint_0^1 r(x,y) dF(x,y) = \int_0^1 \int_0^1 \left(x^2 + y^2\right) dx dy \\ &= \int_0^1 x^2 dx + \int_0^1 y^2 dy = \frac{2}{3} \end{aligned}$$

Theorem 2.7.12:::> The k^{th} moment of X is defined to be $\mathbb{E}\left(X^k\right)$ assuming that $\mathbb{E}\left(|X|^k\right) < \infty$

这些" 矩" 对于刻画分布有着重要意义. 比如第一矩 $\mu = E(X)$ 均值, 第二矩 $E(X^2) - \mu^2$ 方差, 以及第三矩是分布的不对称性, 第四矩是尾巴有多沉、多厚等.

- k 阶 (原点) 矩: $E\left(X^k\right)$
- k 阶（中心）矩: $E\left((X - EX)^k\right)$
- X 和 Y 的 $k+l$ 阶混合矩: $E\left(X^k Y^l\right)$
- $k+l$ 阶混合中心矩: $E\left((X - EX)^k (Y - EY)^l\right)$

If the k^{th} moment exists and if $j < k$ then the j^{th} moment exists.

Proof. We have

$$\begin{aligned}
\mathbb{E}|X|^j &= \int_{-\infty}^{\infty} |x|^j f_X(x) dx \\
&= \int_{|x| \leqslant 1} |x|^j f_X(x) dx + \int_{|x|>1} |x|^j f_X(x) dx \\
&\leqslant \int_{|x| \leqslant 1} f_X(x) dx + \int_{|x|>1} |x|^k f_X(x) dx \\
&\leqslant 1 + \mathbb{E}\left(|X|^k\right) < \infty
\end{aligned}$$

The k^{th} central moment is defined to be $\mathbb{E}\left((X - \mu)^k\right)$ ■

Definition 2.7.13 (矩母函数 (矩生成函数)MGF)**:::>**

$$\text{MGF}_x(t) = \mathbb{E}[e^{tx}] = \begin{cases} \sum_x e^{tx} \cdot P(x) & x: \text{ discrete} \\ \int_x e^{tx} \cdot f(x) \mathrm{d}x & x: \text{ continuous} \end{cases}$$

矩母函数与分布函数 (CDF) 相互唯一确定.

获得矩母函数后可以十分方便得得到第 n 阶矩: 取 n 次 MGF 的导数并令 $t = 0$ 即可:

$$E(X^n) = \frac{d^n}{dt^n} MGF_x(t)\Big|_{t=0}$$

证明: 利用泰勒展开公式

$$e^{tx} = 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots + \frac{(tx)^n}{n!}$$

然后取期望值

$$\begin{aligned} E(e^{tx}) &= E\left(1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!} + \cdots + \frac{(tx)^n}{n!}\right) \\ &= E(1) + tE(x) + \frac{t^2}{2!}E\left(x^2\right) + \frac{t^3}{3!}E\left(x^3\right) + \cdots + \frac{t^n}{n!}E\left(x^n\right) \end{aligned}$$

现在, 对 t 取导数

$$\begin{aligned} \frac{d}{dt}E(e^{tx}) &= \frac{d}{dt}\left(E(1) + tE(x) + \frac{t^2}{2!}E\left(x^2\right) + \frac{t^3}{3!}E\left(x^3\right) + \cdots + \frac{t^n}{n!}E\left(x^n\right)\right) \\ &\quad \text{plug } t = 0 \text{ in } \cdots \\ &= 0 + E(x) + 0 + 0 + \cdots + 0 \\ &= \quad E(x) \end{aligned}$$

Theorem 2.7.14:::> 若 X, Y 相互独立, 则

$$M_{X+Y}(u) = M_X(u) \times M_Y(u)$$

即和的矩母函数等于矩母函数的乘积.

矩母函数其实类似于拉普拉斯变换. 在拉普拉斯变换中, 一个重要应用即是把卷积变为乘积, 这里同理.

Example 2.7.15 (二项分布矩母函数及其应用)**:::>** 若 $X \sim B(n, p)$, 求 $E(X^3)$

由公式

$$E\left(X^3\right) = \sum_{k=0}^{n} k^3 C_n^k p^k q^{n-k}$$

显然太过复杂.

我们尝试用矩母函数解决. 首先我们可以求出, 两点分布的矩母函数为:

$$q + pe^u$$

而二项分布可以看作 n 个独立的两点分布的叠加, 故其矩母函数为

$$[q+pe^u]^n$$

则三阶矩可方便地求出.

Definition 2.7.16 (随机变量的中位数)**:::>** 一个数 x 若满足

$$P(X\leqslant x)\geqslant p, P(X\geqslant x)\geqslant 1-p,\quad 0<p<1$$

则称 x 为随机变量 X 的 $p-$ 分位数, 特别地, 当 $p=\frac{1}{2}$ 时, 称之为 X 的中位数.

显然, x 为随机变量 X 的中位数 $\Leftrightarrow \frac{1}{2}\leqslant F(x)\leqslant \frac{1}{2}+P(X=x)$, 且如果随机变量 X 的分布是对称的, 那么其对称中心显然是它的分布的中位数

Example 2.7.17:::> 随机变量 X 的分布为

$$X\sim\begin{pmatrix} -2 & 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{3} & \frac{1}{6} \end{pmatrix}$$

则 $EX=\frac{1}{6}$. 其中位数等于多少?

由于 $P(X\leqslant 0)=\frac{1}{2}, P(X\geqslant 0)=\frac{3}{4}>\frac{1}{2}$, 显然 0 是它的一个中位数. 而当 $0<x<1$ 时, $P(X\leqslant x)=P(X=-2)+P(X=0)=\frac{1}{2}$, 且 $P(X\geqslant x)=P(X=1)+P(X=2)=\frac{1}{2}$. 故每个 $x, 0\leqslant x<1$, 都是 X 的中位数.(有无穷个中位数)

2.7.2 Properties of Expectations

Theorem 2.7.18:::> If $X_1,\ldots,X_n$ are random variables and $a_1,\ldots,a_n$ are constants, then

$$\mathbb{E}\left(\sum_i a_i X_i\right)=\sum_i a_i \mathbb{E}(X_i)$$

这个结论对随机变量 X 没有任何要求.

这个结论非常重要, 它可在随机变量分解法中得到应用. 如"0-1" 分布的期望等于 p, 二项分布可以分解为 n 个相互独立的"0-1" 分布的和, 故其期望为 np. 超几何分布的期望是 $n\frac{a}{a+b}$.

注意其中 n 是有限的. 若 n 无穷大, 则这个结论不一定对. 这里要用实分析中的单调收敛定理深入分析.

Example 2.7.19:::> 设有 n 张卡片, 上面分别标有号码 $1,2,\cdots,n$. 从中任取 1 张记录它的号码, 然后放回, 然后再取第 2 张记录其号码, ..., 一直到有 k 张卡片被取出. 记 X 为这 k 张卡片的号码之和, 证明 $EX=\frac{k(n+1)}{2}$. 又问若不放回, 而 $k\leqslant n$, 那么结果又为如何?

结果仍为 $EX=\frac{k(n+1)}{2}$. 但是两者机制不同.

用随机变量分解法来理解. 前者 (放回) 的情况是: 每一回合得到的分数为随机变量, k 次过程为 k 个随机变量的和, 期望也随之相加. 后者 (不放回) 的情况是, 每一张牌" 贡献" 的分数是一个随机变量, 最后的分数为这 n 个随机变量的和.

Example 2.7.20:::> Let $X\sim$ Binomial(n,p). What is the mean of X? We could try to appeal to the definition:

$$\mathbb{E}(X)=\int x dF_X(x)=\sum_x x f_X(x)=\sum_{x=0}^{n} x\left(\begin{array}{c} n \\ x \end{array}\right)p^x(1-p)^{n-x}$$

but this is not an easy sum to evaluate.

Instead, note that $X=\sum_{i=1}^{n}X_i$ where $X_i=1$ if the i^{th} toss is heads and $X_i=0$ otherwise. Then $\mathbb{E}(X_i)=(p\times 1)+((1-p)\times 0)=p$ and $\mathbb{E}(X)=\mathbb{E}(\sum_i X_i)=\sum_i \mathbb{E}(X_i)=np$

Theorem 2.7.21:::> 如果存在 a,b, 使得 $P(a\leqslant X\leqslant b)=1$, 则 $E(X)$ 存在, 而且 $a\leqslant E(X)\leqslant b$.

Theorem 2.7.22:::>

$$E\left(X^2\right)\geqslant 0, \text{且} E\left(X^2\right)=0\Leftrightarrow P(X=0)=1$$

Theorem 2.7.23:::> Let $X_1,\ldots,X_n$ be independent random variables. Then,

$$\mathbb{E}\left(\prod_{i=1}^{n}X_i\right)=\prod_i \mathbb{E}(X_i)$$

Notice that the summation rule does not require independence but the multiplication rule does.

$$E(\frac{X}{Y}) = E(X)E(\frac{1}{(Y)}) \neq \frac{E(X)}{E(Y)}$$

Example 2.7.24 (St. Petersburg paradox)**:::>** Suppose we play a game in which we keep tossing a coin until you get a tail. If you get a tail on the ith round, then I pay you 2^i. The expected value is

$$E[X] = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \cdots = \infty.$$

This means that on average, you can expect to get an infinite amount of money! In real life, though, people would hardly be willing to pay $20 to play this game. There are many ways to resolve this paradox, such as taking into account the fact that the host of the game has only finitely many money and thus your real expected gain is much smaller.

解释这种现象需要用到期望效用理论, 这也正是丹尼尔·伯努利为解决圣彼得堡悖论所提出的. 他提出的理论主要包含两条原理: 边际效用递减原理: 一个人对于财富的占有多多益善, 即效用函数一阶导数大于零; 随着财富的增加, 满足程度的增加速度不断下降, 效用函数二阶导数小于零. 最大效用原理: 在风险和不确定条件下, 个人的决策行为准则是为了获得最大期望效用值而非最大期望金额值.

设赌徒资金为 x, 效用函数为对数效用 $U(x) = \ln x$, c 为肯付的门票费, 则此彩票的期望效用为有限

$$EU = \sum_{n=1}^{\infty} \frac{\ln(x + 2^{n-1} - c) - \ln x}{2^n} < +\infty$$

而如果赌场只有有限资金 W, 此时彩票的期望值为

$$\begin{aligned} EV &= \sum_{n=1}^{\infty} \min\left(2^{n-1}, W\right) \frac{1}{2^n} = \sum_{n=1}^{L} 2^{n-1} \frac{1}{2^n} + \sum_{n=L+1}^{\infty} W \frac{1}{2^n} \\ &= \frac{L}{2} + \frac{W}{2^L} \end{aligned}$$

其中 $L = [\log_2 W]$ 实际为赌场在无力支付时赌博最多可能的局数, 如果 $W =$ 10 亿美元, 则 $EV = 15.93$ 美元.

Example 2.7.25:::> We calculate the expected values of different distributions:

1. Poisson $P(\lambda)$. Let $X \sim P(\lambda)$. Then

$$P_X(r) = \frac{\lambda^r e^{-\lambda}}{r!}.$$

So

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

最后一步用到了泰勒展开式:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!}$$

2. Let $X \sim B(n,p)$. Then

$$\begin{aligned} E[x] &= \sum_{0}^{n} rP(x=r) \\ &= \sum_{0}^{n} r\binom{n}{r} p^r (1-p)^{n-r} \\ &= \sum_{0}^{n} r\frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \\ &= np \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)![(n-1)-(r-1)]!} p^{r-1} (1-p)^{(n-1)-(r-1)} \\ &= np \sum_{0}^{n-1} \binom{n-1}{r} p^r (1-p)^{n-1-r} \\ &= np. \end{aligned}$$

Theorem 2.7.26:::> $E[x]$ is a constant that minimizes $E[(X-c)^2]$ over c.
这个定理的含义是, 如果用一个常量代替随机变量的话, 使得其均方误差最小的量是期望.

$$\begin{aligned}E[(X-c)^2] &= E[(X-E[x]+E[x]-c)^2] \\ &= E[(X-E[x])^2+2(E[x]-c)(X-E[x])+(E[x]-c)^2] \\ &= E(X-E[x])^2+0+(E[x]-c)^2.\end{aligned}$$

This is clearly minimized when $c=E[x]$. Note that we obtained the zero in the middle because $E[X-E[x]]=E[x]-E[x]=0$.

这也正是数学期望的统计学意义, 且从表达式中我们可以看出, 这个最小平均平方误差正是我们说的方差.

Theorem 2.7.27 (Cauchy-Schwartz 不等式)**:::>** 若 $E(X^2)<+\infty$,$E(Y^2)<+\infty$, 则

$$(E(XY))^2 \leqslant E\left(X^2\right)E\left(Y^2\right)$$

Proof.

$$E(tX+Y)^2 = E(X)^2t^2+E(Y)^2+2E(tX)E(Y)$$

这个式子必大于零, 故其判别式小于等于 0, 则 Cauchy-Schwartz 不等式得证 ■

Theorem 2.7.28 (Markov's inequality)**:::>** 若 X 为一非负随机变量, 则

$$\mathrm{P}(X \geqslant a) \leqslant \frac{\mathrm{E}(X)}{a}$$

Proof.

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x) dx \\
&= \int_{0}^{\infty} x f(x) dx \\
&\geqslant \int_{a}^{\infty} x f(x) dx \\
&\geqslant \int_{a}^{\infty} a f(x) dx \\
&= a \int_{a}^{\infty} f(x) dx \\
&= a\mathrm{P}(X \geqslant a)
\end{aligned}$$

■

Theorem 2.7.29:::> 如果用一个常量代替随机变量的话, 使得其绝对值误差 $E\,|X-c\,|$ 最小的量是中位数.

Proof. 设随机变量 X 有连续的概率密度函数 $f(x)$, 令 $h(a)=E|X-a|$, 则:

$$h(a)=\int_{-\infty}^{a}(a-x)f(x)dx+\int_{a}^{\infty}(x-a)f(x)dx$$

利用变上限积分求导可得:

$$h'(a)=2F(a)-1;\quad h''(a)=2f(a)\geqslant 0$$

因而证毕. ■

Definition 2.7.30 (Indicator function)**:::>** The *indicator function* or *indicator variable* $I[A]$ (or I_A) of an event $A \subseteq \Omega$ is

$$I[A](\omega)=\begin{cases}1 & \omega \in A \\ 0 & \omega \notin A\end{cases}$$

This indicator random variable is not interesting by itself. However, it is a rather useful tool to prove results.

It has the following properties:

Proposition 2.7.31:::>

- $E[I[A]] = \sum_\omega p(\omega) I[A](\omega) = P(A)$.
- $I[A^C] = 1 - I[A]$.
- $I[A \cap B] = I[A] I[B]$.
- $I[A \cup B] = I[A] + I[B] - I[A] I[B]$.
- $I[A]^2 = I[A]$.

These are easy to prove from definition. In particular, the last property comes from the fact that $I[A]$ is either 0 and 1, and $0^2 = 0, 1^2 = 1$.

2.7.3 Variance

Definition 2.7.32 (Variance and standard deviation)**:::>** The *variance* of a random variable X is defined as

$$var(X) = \mathbb{E}[(X - \mathbb{E}[x])^2].$$

The *standard deviation* is the square root of the variance, $\sqrt{var(X)}$.

This is a measure of how "dispersed" the random variable X is. If we have a low variance, then the value of X is very likely to be close to $\mathbb{E}[x]$.

Theorem 2.7.33:::>

1. $varX \geqslant 0$. If $varX = 0$, then $P(X = E[x]) = 1$.
2. $var(a + bX) = b^2 var(X)$. This can be proved by expanding the definition and using the linearity of the expected value.
3. $var(X) = E[X^2] - E[x]^2$, also proven by expanding the definition.
4. If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{V}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathbb{V}(X_i)$$

Definition 2.7.34:::> If $X_1, \ldots, X_n$ are random variables then we define the sample mean to be

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

and the sample variance to be

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

Theorem 2.7.35:::> Let $X_1, \ldots, X_n$ be IID and let $\mu = \mathbb{E}(X_i), \sigma^2 = \mathbb{V}(X_i)$. Then

$$\mathbb{E}\left(\overline{X}_n\right) = \mu, \quad \mathbb{V}\left(\overline{X}_n\right) = \frac{\sigma^2}{n} \quad \text{and} \quad \mathbb{E}\left(S_n^2\right) = \sigma^2$$

Theorem 2.7.36 (切比雪夫不等式)**:::>**

$$\Pr(|X - \mathrm{E}(X)| \geqslant a) \leqslant \frac{\mathrm{Var}(X)}{a^2}$$

证明可以以Markov's inequality为基础. 考虑随机变量 $(X - \mathrm{E}(X))^2$, 根据马尔科夫不等式, 则可得:

$$\Pr\left((X - \mathrm{E}(X))^2 \geqslant a^2\right) \leqslant \frac{\mathrm{Var}(X)}{a^2}$$

Example 2.7.37:::> Suppose we test a prediction method, a neural net for example, on a set of n new test cases. Let $X_i = 1$ if the predictor is wrong and $X_i = 0$ if the predictor is right. Then $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$ is the observed error rate. Each X_i may be regarded as a Bernoulli with unknown mean p. We would like to know the true - but unknown - error rate p. Intuitively, we expect that $\overline{X}_n$ should be close to p. How likely is $\overline{X}_n$ to not be within ε of p ? We have that $\mathbb{V}\left(\overline{X}_n\right) = \mathbb{V}(X_1)/n = p(1-p)/n$ and

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leqslant \frac{\mathbb{V}\left(\overline{X}_n\right)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}$$

since $p(1-p) \leqslant \frac{1}{4}$ for all p. For $\varepsilon = .2$ and $n = 100$ the bound is .0625

Theorem 2.7.38 (Hoeffding's Inequality)**:::>** Let $Y_1, \ldots, Y_n$ be independent observations such that $\mathbb{E}(Y_i) = 0$ and $a_i \leqslant Y_i \leqslant b_i$. Let $\varepsilon > 0$. Then, for any $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geqslant \varepsilon\right) \leqslant e^{-t\varepsilon}\prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}$$

Theorem 2.7.39:::> Let $X_1,\ldots,X_n \sim \text{Bernoulli}(p)$. Then, for any $\varepsilon > 0$

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2}$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$

Example 2.7.40:::> Let $X_1,\ldots,X_n \sim \text{Bernoulli}(p)$. Let $n = 100$ and $\varepsilon = .2$. We saw that Chebyshev's inequality yielded

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > \varepsilon\right) \leqslant .0625$$

According to Hoeffding's inequality,

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| > .2\right) \leqslant 2e^{-2(100)(.2)^2} = .00067$$

which is much smaller than .0625.

Example 2.7.41 (Binomial distribution)**:::>** Let $X \sim B(n,p)$ be a binomial distribution. Then $E[x] = np$. We also have

$$\begin{aligned} E[X(X-1)] &= \sum_{r=0}^{n} r(r-1)\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} \\ &= n(n-1)p^2\sum_{r=2}^{n}\binom{n-2}{r-2}p^{r-2}(1-p)^{(n-2)-(r-2)} \\ &= n(n-1)p^2. \end{aligned}$$

The sum goes to 1 since it is the sum of all probabilities of a binomial $N(n-2,p)$ So $E[X^2] = n(n-1)p^2 + E[x] = n(n-1)p^2 + np$. So

$$var(X) = E[X^2] - (E[x])^2 = np(1-p) = npq.$$

或者, 利用 Bernoulli 分布的方差为 pq, 我们可得二项分布的方差为 npq

Example 2.7.42 (Poisson distribution)**:::>** If $X \sim P(\lambda)$, then $E[x] = \lambda$, and $var(X) = \lambda$, since $P(\lambda)$ is $B(n,p)$ with $n \to \infty, p \to 0, np \to \lambda$.

Example 2.7.43 (Geometric distribution)**:::>** Suppose $P(X = r) = q^r p$

for $r = 0, 1, 2, \cdots$. Then

$$\begin{aligned} E[x] &= \sum_0^\infty r\, pq^r \\ &= pq \sum_0^\infty r q^{r-1} \\ &= pq \sum_0^\infty \frac{d}{dq} q^r \\ &= pq \frac{d}{dq} \sum_0^\infty q^r \\ &= pq \frac{d}{dq} \frac{1}{1-q} \\ &= \frac{pq}{(1-q)^2} \\ &= \frac{q}{p}. \end{aligned}$$

Then

$$\begin{aligned} E[X(X-1)] &= \sum_0^\infty r(r-1) pq^r \\ &= pq^2 \sum_0^\infty r(r-1) q^{r-2} \\ &= pq^2 \frac{d^2}{dq^2} \frac{1}{1-q} \\ &= \frac{2pq^2}{(1-q)^3} \end{aligned}$$

So the variance is

$$var(X) = \frac{2pq^2}{(1-q)^3} + \frac{q}{p} - \frac{q^2}{p^2} = \frac{q}{p^2}.$$

Example 2.7.44 (Hypergeometric Distribution)**:::>** 超几何分布

$$P(X = x) = \frac{\begin{pmatrix} N_1 \\ x \end{pmatrix} \begin{pmatrix} N_2 \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

期望: 注意到 $\begin{pmatrix} N \\ n \end{pmatrix} = \frac{N}{n} \begin{pmatrix} N-1 \\ n-1 \end{pmatrix}$

则

$$E(X)=\frac{n}{N}\sum_{x=1}^{N_1}x\frac{\begin{pmatrix}N_1\\x\end{pmatrix}\begin{pmatrix}N_2\\n-x\end{pmatrix}}{\begin{pmatrix}N-1\\n-1\end{pmatrix}}=n\frac{N_1}{N}\sum_{x=1}^{N_1}\frac{\begin{pmatrix}N_1-1\\x-1\end{pmatrix}\begin{pmatrix}N_2\\n-x\end{pmatrix}}{\begin{pmatrix}N-1\\n-1\end{pmatrix}}$$

由于 $\begin{pmatrix}N_2\\n-x\end{pmatrix}=\begin{pmatrix}N_2\\(n-1)-(x-1)\end{pmatrix}$ 则

$$\text{原式}=n\frac{N_1}{N}\sum_{x=1}^{N_1}\frac{\begin{pmatrix}N_1-1\\x-1\end{pmatrix}\begin{pmatrix}N_2\\(n-1)-(x-1)\end{pmatrix}}{\begin{pmatrix}N-1\\n-1\end{pmatrix}}=n\frac{N_1}{N}\sum_{x=1}^{N_1}P(X=x-1)$$

由于 $\sum_{x=1}^{N_1}P(X=x-1)=1$ 则

$$\text{原式}=n\frac{N_1}{N}$$

方差: 注意到

$$\begin{pmatrix}N\\n\end{pmatrix}=\frac{N(N-1)}{n(n-1)}\begin{pmatrix}N-2\\n-2\end{pmatrix}$$

那么

$$E(X(X-1))=\frac{n(n-1)}{N(N-1)}\sum_{x=2}^{N_1}x(x-1)\frac{\begin{pmatrix}N_1\\x\end{pmatrix}\begin{pmatrix}N_2\\n-x\end{pmatrix}}{\begin{pmatrix}N-2\\n-2\end{pmatrix}}$$

继续化简,

$$=\frac{nN_1(n-1)(N_1-1)}{N(N-1)}\sum_{x=2}^{N_1}\frac{\begin{pmatrix}N_1-2\\x-2\end{pmatrix}\begin{pmatrix}N_2\\n-x\end{pmatrix}}{\begin{pmatrix}N-2\\n-2\end{pmatrix}}=\frac{nN_1(n-1)(N_1-1)}{N(N-1)}\sum_{x=2}^{N_1}P(X=x-2)$$

由于 $\sum_{x=2}^{N_1}P(X=x-2)=1$, 则

$$E(X(X-1))=\frac{nN_1(n-1)(N_1-1)}{N(N-1)}$$

因此

$$\sigma^2 = E(X(X-1)) + E(X) - E^2(X) = n\frac{N_1(N-N_1)(N-n)}{N^2(N-1)}$$

Example 2.7.45:::> Let $2n$ people (n husbands and n wives, with $n > 2$) sit alternate man-woman around the table randomly. Let N be the number of couples sitting next to each other.

Let $A_i = [i\text{th couple sits together}]$. Then

$$N = \sum_{i=1}^{n} I[A_i].$$

Then

$$E[N] = E\left[\sum I[A_i]\right] = \sum_{1}^{n} E\left[I[A_i]\right] = nE\left[I[A_1]\right] = nP(A_i) = n \cdot \frac{2}{n} = 2.$$

We also have

$$\begin{aligned} E[N^2] &= E\left[\left(\sum I[A_i]\right)^2\right] \\ &= E\left[\sum_i I[A_i]^2 + 2\sum_{i<j} I[A_i]I[A_j]\right] \\ &= nE\left[I[A_i]\right] + n(n-1)E\left[I[A_1]I[A_2]\right] \end{aligned}$$

We have $E[I[A_1]I[A_2]] = P(A_1 \cap A_2) = \frac{2}{n}\left(\frac{1}{n-1}\frac{1}{n-1} + \frac{n-2}{n-1}\frac{2}{n-1}\right)$(第二部分计算是按照第一对夫妻做好后另一位是否坐在第一对夫妻旁边分两类讨论的). Plugging in, we ultimately obtain $var(N) = \frac{2(n-2)}{n-1}$.

In fact, as $n \to \infty$, $N \sim P(2)$.

Example 2.7.46 (共同基金减少风险)**:::>** Cinematics 是一个电影院的连锁链, 而 RTV 是电视剧制作中心的制作商. 投资于 Cinematics 与 RTV 所得到的年收益分别为随机变量 X 和 Y, 假定它们的平均年收益相等:

$$EX = EY = 1.1$$

而标准差分别为:

$$\sigma_X = 0.08, \sigma_Y = 0.04$$

并且它们的相关系数为 $r_{X,Y} = -0.75$, 这里负的相关可以解释为: 看很多电视节目的人较少去电影院. 如果我们将 α 部分 $(0 \leqslant \alpha \leqslant 1)$ 的钱投资于 Cinematics, 而将 $1-\alpha$ 部分的钱投资于 RTV, 请问如何确定 α, 使投资风险最小?

Solution. 令 $Z = \alpha X + (1-a)Y$. 那么 $EZ = 1.1$, 它不依赖 a. 而其方差为:

$$\begin{aligned}\sigma_Z^2 &= D(\alpha X + (1-\alpha)Y) \\ &= \alpha^2 D(X) + 2\alpha(1-\alpha)\mathrm{Cov}(X.Y) + (1-\alpha)^2 D(Y) \\ &= 0.0064\alpha^2 + 2\alpha(1-\alpha) \times 0.08 \times 0.04 \times (-0.75) + 0.0016(1-\alpha)^2 \\ &= 0.0128\alpha^2 - 0.0080\alpha + 0.0016.\end{aligned}$$

$\alpha = \frac{0.0080}{0.0256} = 0.3125$ 时方差最小, 风险最小. ■

2.7.4 Covariance

Definition 2.7.47:::> If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the covariance between X and Y by

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y))$$

and the correlation between X and Y by

$$\rho = \rho_{X,Y} = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

相关系数也被记为 $r_{X,Y}$, 其用意为将协方差标准化, 去除量纲的影响.

Theorem 2.7.48:::> The covariance satisfies:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Theorem 2.7.49:::> The correlation satisfies:

$$-1 \leqslant \rho(X,Y) \leqslant 1$$

称之为 X 和 Y 的相关系数 (确切地说, 应该称为统计意义的线性相关系数)

Theorem 2.7.50:::> If $Y = aX+b$ for some constants a and b then $\rho(X,Y) = 1$ if $a > 0$ and $\rho(X,Y) = -1$ if $a < 0$. If X and Y are independent, then $\mathrm{Cov}(X,Y) = \rho = 0$. The converse is not true in general.

Theorem 2.7.51:::> 若 X, Y 不相关, 则 $\rho(X,Y) = 0$. 独立一定不相关, 但不相关不一定独立.

统计里的" 不相关" 只是说这两个随机变量没有线性关系, $Y = aX + b$ 的概率永远等于 0, 但可能有非线性关系.

Example 2.7.52:::> 用示性函数Indicator function举例: 若 $X = I(A)$, $Y = I(B)$, 则

$$\rho_{X,Y} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(A^c)P(B)P(B^c)}} = \rho_{A,B}$$

这也就是之前所说的事件的相关系数的公式.

Theorem 2.7.53:::>

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\,\mathrm{Cov}(X,Y)$$

$$\mathbb{V}(X-Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\,\mathrm{Cov}(X,Y)$$

More generally, for random variables $X_1, \ldots, X_n$

$$\mathbb{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \mathbb{V}(X_i) + 2\sum\sum_{i<j} a_i a_j \,\mathrm{Cov}\left(X_i, X_j\right)$$

Definition 2.7.54:::> 对于一个随机向量 $(X_1, \cdots, X_n)$ 协方差矩阵和相关系数矩阵: 协方差矩阵:

$$\Sigma = \left(\mathrm{Cov}\left(X_i, X_j\right)\right)_{i,j \leqslant n}$$

我们可以证明协方差矩阵是非负定的:

$$(a_1, \ldots a_n)\left(\mathrm{Cov}\left(X_i, X_j\right)\right)_{n\times n} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}^n = \sum_i \sum_j a_i a_j Cov\left(X_i, X_j\right)$$

而上式即为:

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{1} a_j x_j\right)$$

当然可改写为

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{1} a_i x_i\right)$$

也即为 $D(\sum_{i=1}^{n} a_i X_i)$, 既然是方差, 故必大于等于 0, 故矩阵非负定.

而之前我们介绍的公式:

$$D\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i DX_i + 2\sum_{i\neq j} a_i a_j Cov\left(X_i, X_j\right).$$

也可以通过以上矩阵的形式直观地证明出来.(区分在对角线上的和不在对角线上的即可)

相关系数矩阵:

$$R = \left(r_{X_i, X_j}\right)_{i,j\leqslant n}$$

它们都是对称矩阵.

Proposition 2.7.55 (协方差与相关系数的性质)**:::>** $\mathrm{Cov}(X,Y)$ 是对称的双线性函数, 即有:

- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- 若 a, b 是两个任意常数, 则 $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y)$
- 更一般地

$$\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \,\mathrm{Cov}\left(X_i, Y_j\right)$$

Theorem 2.7.56 (最佳线性预测)**:::>**

$$E[X - (\widehat{a}Y + \widehat{b})]^2 = \min_{a,b} E[X - (aY + b)]^2 = DX\left(1 - r_{X,Y}^2\right)$$

其中

$$\widehat{a} = \frac{\mathrm{Cov}(X,Y)}{DY} = r_{X,Y}\frac{\sqrt{DX}}{\sqrt{DY}}$$

$$EX = \widehat{a}EY + \widehat{b}$$

证明时用偏导数为 0 即可.

从这个式子也能十分方便地得出 $-1 \leqslant \rho(X,Y) \leqslant 1$ 的结论.

Theorem 2.7.57:::> 预测值 $\widehat{X} = (\widehat{a}Y + \widehat{b})$ 与残差 $X - \widehat{X}$ 不相关. 即

$$Cov(X - \widehat{X}, \widehat{X}) = 0$$

证明时将各项代入并利用协方差的性质即可.

Example 2.7.58:::> 我们记随机变量 X 为抛 n 次硬币后正面朝上的次数, Y 为反面向上的次数, 求 $\rho(X,Y)$.

由于 $X \sim B\left(n, \frac{1}{2}\right)$, $Y \sim B\left(n, \frac{1}{2}\right)$, 我们当然可以算期望并代入公式后得, 但还有更直观的解法: 由于 $X + Y = n$, 故有明确的线性关系:

$$Y = -X + n$$

则可知 $\rho(X,Y) = -1$

Theorem 2.7.59:::>

$$\rho_{aX+b,cY+d} = \frac{ac}{|ac|}\rho_{X,Y}$$

即取决于 a 与 c 同号还是异号. 从直观的角度上也非常好想象, 因为相关系数刻画的是线性关系.

Definition 2.7.60 (随机变量的标准化)**:::>**

$$X^* = \frac{X - EX}{\sqrt{DX}}$$

标准化后

$$E(X^*) = 0$$

$$D(X^*) = 1$$

$$\rho_{X^*,Y^*} = \rho_{X,Y}$$

易见对于 X,Y 的标准化随机变量 $X^* = \frac{X-EX}{\sqrt{DX}}, Y^* = \frac{Y-EY}{\sqrt{DY}}$ 有:

$$\mathrm{Cov}(X^*, Y^*) = r_{X,Y}$$

Example 2.7.61:::> 设 $\xi_1,\xi_2,\cdots,\xi_{n+m}(n>m)$ 相互独立同分布且具有有限方差, 试求 $X=\sum_{k=1}^{n}\xi_k$ 与 $Y=\sum_{k=1}^{n}\xi_{m+k}$ 的相关系数.

这里利用协方差的双线性性来解决:

$$r(x,y)=\frac{\sum_{k=1}^{n}\operatorname{Cov}(\xi_k,Y)}{\sqrt{DXDY}}=\frac{\sum_{k=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}\left(\xi_k,\xi_{j+m}\right)}{\sqrt{DXDY}}$$

由于相互独立同分布, 故不同 ξ 之间的协方差为 0, 且 X 与 Y 的方差可以直接相加得到.

故结果为:

$$\frac{n-m}{n}$$

Example 2.7.62:::> 将一颗骰子抛 n 次, 以 X 表示 1 点出现的次数, Y 表示 6 点出现的次数, 求 X 与 Y 的协方差 $\operatorname{Cov}(X,Y)$ 及相关系数.

令 $X_i=\begin{cases}1, & \text{若第}i\text{次出1点},\\ 0, & \text{否则};\end{cases}$, $Y_i=\begin{cases}1, & \text{若第}i\text{次出6点},\\ 0, & \text{否则},\end{cases}$ 则 $X=\sum_{i=1}^{n}X_i, Y=\sum_{i=1}^{n}Y_i$, 故 $EX=\frac{n}{6}, EY=\frac{n}{6}$. 而 $EX_iY_j=P\left(X_i=1,Y_j=1\right)=\begin{cases}0, & i=j\\ \frac{1}{6^2}, & i\neq j\end{cases}$, 故 $EXY=\sum_{i\neq j}\frac{1}{6^2}=\frac{n(n-1)}{36}$, 从而协方差 $\operatorname{Cov}(X,Y)$ 及相关系数分别为 $-\frac{n}{36}$ 与 $-\frac{1}{5}$

2.7.5 Conditional Expectation

Definition 2.7.63:::> The conditional expectation of X given $Y=y$ is

$$\mathbb{E}(X\mid Y=y)=\begin{cases}\sum x f_{X|Y}(x\mid y)dx & \text{discrete case}\\ \int x f_{X|Y}(x\mid y)dx & \text{continuous case}\end{cases}$$

If $r(x,y)$ is a function if x and y then

$$\mathbb{E}(r(X,Y)\mid Y=y)=\begin{cases}\sum r(x,y) f_{X|Y}(x\mid y)dx & \text{discrete case}\\ \int r(x,y) f_{X|Y}(x\mid y)dx & \text{continuous case}\end{cases}$$

$\mathbb{E}(X\mid Y=y)$ is a function of y. In other words, $\mathbb{E}(X\mid Y)$ is the random variable whose value is $\mathbb{E}(X\mid Y=y)$ when $Y=y$.

Example 2.7.64:::> Suppose we draw $X \sim \text{Unif}(0,1)$. After we observe $X = x$, we draw $Y \mid X = x \sim \text{Unif}(x,1)$. Intuitivewly, we expect that $\mathbb{E}(Y \mid X = x) = (1+x)/2$. In fact, $f_{Y|X}(y \mid x) = 1/(1-x)$ for $x < y < 1$ and

$$\mathbb{E}(Y \mid X = x) = \int_x^1 y f_{Y|X}(y \mid x) dy = \frac{1}{1-x} \int_x^1 y dy = \frac{1+x}{2}$$

as expected. Thus, $\mathbb{E}(Y \mid X) = (1+X)/2$.

Theorem 2.7.65:::> 设离散型随机变量 X 与 Y 相互独立, 则 $\mathbb{E}(X \mid Y) = \mathbb{E}X$

Proof. 由于 X 与 Y 相互独立, 故由

$$\begin{aligned}\mathbb{E}(X \mid Y = y_1) &= \sum_1 x_i P\left(X = x_i \mid Y = y_j\right) = \sum_i x_i P\left(X = x_i\right) \\ &= \mathbb{E}X,\end{aligned}$$

可得:

$$\mathbb{E}(X \mid Y) = \mathbb{E}X$$

■

但是反过来的话未必成立.

若

$$\mathbb{E}(Y \mid X) = \mathbb{E}Y$$

我们可以得到 $Cov(X,Y) = 0$, 即 X 与 Y 不相关.

Proof.

$$\mathbb{E}(XY) = \mathbb{E}\mathbb{E}(XY \mid X) = \mathbb{E}[X\mathbb{E}(Y \mid X)] = \mathbb{E}X\mathbb{E}Y$$

■

Theorem 2.7.66 (The rule of iterated expectations 全期望公式)**:::>** For random variables X and Y, assuming the expectations exist, we have that

$$\mathbb{E}[\mathbb{E}(Y \mid X)] = \mathbb{E}(Y)$$

and

$$\mathbb{E}[\mathbb{E}(X \mid Y)] = \mathbb{E}(X)$$

More generally, for any function $r(x,y)$ we have

$$\mathbb{E}[\mathbb{E}(r(X,Y) \mid X)] = \mathbb{E}(r(X,Y))$$

证明:

$$\begin{aligned}\mathbb{E}[\mathbb{E}(Y \mid X)] &= \int \mathbb{E}(Y \mid X = x) f_X(x) dx = \iint y f(y \mid x) dy f(x) dx \\ &= \iint y f(y \mid x) f(x) dx dy = \iint y f(x,y) dx dy = \mathbb{E}(Y)\end{aligned}$$

Theorem 2.7.67:::>

$$E\left(\sum_{i=1}^{n} a_i Y_i \mid X\right) = \sum_{i=1}^{n} a_i E(Y_i \mid X)$$

即条件期望具有线性性. 证明非常简单, 对每一个 $X = x_i$ 时, 期望都具有线性性. 再对 x 求和可得条件期望也存在线性性.

Example 2.7.68:::> 设随机向量 (X,Y) 的联合分布列为:

XY	1	2	3
1	$\frac{2}{27}$	$\frac{5}{27}$	$\frac{1}{27}$
2	$\frac{4}{27}$	$\frac{7}{27}$	$\frac{2}{27}$
3	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{2}{27}$

试求 $E(X \mid Y)$ 的分布列, 并利用此分布列求 EX

$E(X \mid Y)$	$\frac{13}{7}$	$\frac{28}{15}$	$\frac{11}{5}$
P	$\frac{7}{27}$	$\frac{15}{27}$	$\frac{5}{27}$

注意, 分布列永远是随机变量的值 + 随机变量出现的概率. 这里随机变量出现的概率实际上是 Y 变量出现的概率.

Example 2.7.69 (巴格达窃贼问题)**:::>** 一只小猫不幸陷进一个有三个门洞的山洞中. 第一个门洞通到一条通道, 沿此通道走两小时后可到达地面. 第二个门洞通到另一个通道, 沿它走三小时会使小猫重新回到这大山洞中.

第三个门洞通到第三个通道, 沿它走五小时后, 也会便小猫又回到这大山洞中. 假定这只小猫并无记忆, 且总是等可能地在三个门洞中任意选择一个, 我们来计算这只小猫到达地面的时间的数学期望.

设 Y 表示这只小猫最初选择的门洞 (我们把门洞分别编号为 1, 2, 3) 则

$$EX=\frac{1}{3}[E(X\mid Y=1)+E(X\mid Y=2)+E(X\mid Y=3)]$$

由于 $E(X\mid Y=1)=2$, 且 $Y=2$ 时, 可得 $X=3+X'$, 其中 X' 是之后小猫重新回到大山洞后所需的时间. 注意到 X 与 X' 同分布, 则有 $E(X\mid Y=2)=E(3+X')=3+EX'=3+EX$. 故

$$EX=\frac{1}{3}\cdot 2+\frac{1}{3}(3+EX)+\frac{1}{3}(5+EX)$$

可得 $EX=10$ 小时.

Example 2.7.70:::> 在生物遗传问题中, 我们经常要了解某物种的种群数目. 假设该物种在第 0 代 (祖先) 只有一个个体, 它所繁衍的第 1 代子女数 N 为一随机变量. 假设每个第 1 代子女又可独立地再繁衍子孙, 其繁衍的子孙数也服从同一分布. 因此, 该物种的第 2 代子孙的子孙的总数就为随机变量 Y:

$$Y=\sum_{i=1}^{N}X_i$$

其中的 $X_i(i\geqslant 1)$ 为祖先的第 i 个子女所繁衍的子孙数, 故 $\{X_i, i\geqslant 1\}$ 是独立同分布随机变量序列, N 为一正整数随机变量, 且与序列 $\{X_i, i\geqslant 1\}$ 相互独立.

由于 N 是随机变量, Y 是随机多个 (即 N 个) 随机变量的和, 基本想法是先算 N 取值固定的条件下, 分别求有关条件概率条件期望条件方差或条件协方差, 然后再对所有可能的情形用全概率公式或全期望公式

$$EY=E(E(Y\mid N))=\sum_{n}E(Y\mid N=n)P(N=n)$$

而由 N 与 $X_1,\cdots,X_n$ 之间的独立性, 我们有

$$\begin{aligned}E(Y\mid N=n)&=E\left(\sum_{i=1}^{N}X_i\mid N=n\right)=E\left(\sum_{i=1}^{n}X_i\mid N=n\right)\\&=E\left(\sum_{i=1}^{n}X_i\right)=nEX_1\end{aligned}$$

其中去掉了条件后的限制 $N=n$ 是因为两者独立. 因而

$$EY=\sum_n nEX_1P(N=n)=EX_1\sum_n nP(N=n)=EX_1\cdot EN$$

因此若设 $EN=EX_1=\mu$, 则 $EY=\mu\cdot\mu=\mu^2$.

进一步如设 Y_n 为第 n 代的子孙数, 数学归纳法易得 $EY_n=\mu^n$.

Example 2.7.71:::> 假设一条自动流水线正常工作时, 所生产的产品的 1 等品率为 p_1, 2 等品率为 p_2, 3 等品率为 p_3,$(p_1+p_2+p_3=1)$. 为保障产品质量, 厂方规定当生产出 1 件等外品时, 该流水线停工检修一次.

1. 在已知首次检修之前共生产了 n 件产品时, 求 n 件产品中 1 等品件数的数学期望.

   由于满足二项分布, n 件中一等品件数的期望为

   $$(n-1)\frac{P_1}{P_1+P_2}$$

2. 求首次检修之前生产的产品中 1 等品的数学期望

   $$Ex=\sum_{k=1}^{\infty}P_kP(x\mid Y=k)$$

   也等于

   $$\sum_{k=1}^{\infty}(1-P_3)^{k-1}\cdot P_3(k-1)\frac{P_1}{P_1+P_2}$$

   注意这里概率的值, 要求最后一次一定是一件等外品.

   化简为

   $$\frac{P_1P_3}{P_1+P_2}\cdot\sum_{j=0}^{\infty}j\,(1-P_3)^j$$

   利用错位相减法或是求导求和顺序互换法,

   $$\frac{P_1}{P_3}$$

Example 2.7.72:::> 抛一颗骰子直到 6 点出现为止, 记 X 为抛掷过程中 3 点出现的次数, 求 EX.

本题与上题类似, 但一定要注意, 当 $Y=n$ 时, $E(X\mid Y=n)=\frac{n-1}{5}$, 而不是 $\frac{n-1}{6}$ 或者是 $\frac{n}{6}$. 因为前 n-1 次的结果不能出现 6.

Definition 2.7.73:::> The conditional variance is defined as

$$\mathbb{V}(Y \mid X = x) = \int (y - \mu(x))^2 f(y \mid x) dy$$

where $\mu(x) = \mathbb{E}(Y \mid X = x)$

对于离散的情况: 在给定 Y 的条件下, X 的条件方差定义为:

$$D(X \mid Y) = E\left[(X - E(X \mid Y))^2 \mid Y\right]$$

也可写成:

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

Theorem 2.7.74:::> For random variables X and Y,

$$\mathbb{V}(Y) = \mathbb{E}\mathbb{V}(Y \mid X) + \mathbb{V}\mathbb{E}(Y \mid X)$$

Example 2.7.75:::> 条件期望重要性在于它起到了一个" 降维" 的作用.

若 X, Y 两个随机变量独立同分布服从几何分布, 求 $P(X > Y)$ 的概率:

$$P(X > Y) = \sum_{k=1}^{\infty} P(X > Y \mid Y = k) P(Y = k)$$

由于 X, Y 独立, 故上式等效于

$$P(X > Y) = \sum_{k=1}^{\infty} P(X > Y) P(Y = k)$$

由几何分布大于某数的概率公式 ($P(X > r) = q^r$), 可得原式为

$$\sum_{k=1}^{\infty} q^k \cdot q^{k-1} p.$$

当然, 这道题还有另一种更简便的做法, 即利用对称性. 随机变量 X, Y 的取值一共有三种可能的关系, 分别为 $X > Y, X < Y, X = Y$. 由对称性可得 $P(X > Y) = P(X < Y)$, 故只需计算 $P(X = Y)$ 即可.

Example 2.7.76:::> 在中我们看到了对称性的妙用, 类似地, 若 X_1, X_2, X_3 独立同分布, 求

$$E(2X_1 + 3X_2 \mid X_1 + X_2 + X_3)$$

由对称性,

$$E(X_1 \mid X_1+X_2+X_3)=E(X_2 \mid X_1+X_2+X_3)=E(X_3 \mid X_1+X_2+X_3)$$

由条件期望的定义 (固定竖线后面的值获得取值的随机变量) 得:

$$E(X_1+X_2+X_3 \mid X_1+X_2+X_3)=X_1+X_2+X_3$$

故可得, 原式 =

$$\frac{5}{3}(X_1+X_2+X_3)$$

Theorem 2.7.77 (最佳预测)**:::>** 设 X 与 Y 的平方的数学期望均存在, 则

$$E[X-E(X \mid Y)]^2=\min_{\varphi} E[X-\varphi(Y)]^2$$

直接从数学期望的概率意义, 可以直观地理解这个结果: 它的概率意义是, 按条件概率 $P(\cdot \mid Y=y_i)$, 在平均平方误差最小的意义下 X 的最佳估计. 于是, 作为随机变量的条件数学期望 $g(Y(\omega))=E(X \mid Y)$ 的概率意义就是: 它是用 Y 的任意函数 (在直观上, 这些函数组成的集合就是 Y 能提供的信息的全体) 来近似 X 时, 在平均平方意义下的最佳近似.

2.8. Entropy

Definition 2.8.1 (分布的熵)**:::>** 对于离散分布 $\boldsymbol{p}=(p_1,\cdots,p_n,\cdots)$, 我们定义它的熵为:

$$H(\boldsymbol{p})=-\sum_i p_i \ln p_i$$

再定义分布 p 关于分布 $\boldsymbol{q}=(q_1,\cdots,q_n,\cdots)$ 的 Kullback-Leibler 相对熵为

$$h(p,q)=\sum_i p_i \ln \frac{p_i}{q_i}$$

Theorem 2.8.2:::> 相对熵 $h(\boldsymbol{p},\boldsymbol{q})\geqslant 0$, 而且 $h(\boldsymbol{p},\boldsymbol{q})=0$ 当且仅当 $\boldsymbol{p}=\boldsymbol{q}$ 成立

Proof. 在 $[0,\infty)$ 上函数 $g(t)\equiv t-1-\ln t$ 在 $t\neq 1$ 时恒正, 且 $g(1)=0$.

取 $t=\frac{q_i}{p_i}$, 于是 $\ln\frac{q_i}{p_i}\leqslant\frac{q_i}{p_i}$, 即 $\ln\frac{p_i}{q_i}\geqslant 1-\frac{q_i}{p_i}$, 而且等号仅当 $\frac{p_i}{q_i}=1$ 时成立, 从而有

$$\sum_i p_i\ln\frac{p_i}{q_i}\geqslant\sum_i p_i\left(1-\frac{q_i}{p_i}\right)=0$$

故结论成立. ■

这个命题表明, 相对熵在相当程度上表达了 p 与 q 的差别: 当 $p=q$ 时, $h(p,q)=0$. 而当所有的 p_i 都与 q_i 接近时, $h(p,q)$ 就很小, 从而 $h(\boldsymbol{p},\boldsymbol{q})$ 可以看成 p 与 q 之间的一种" 准距离"(这里我们之所以称它为准距离, 是因为它既不对称 (即 $h(\boldsymbol{p},\boldsymbol{q})\neq h(\boldsymbol{q},\boldsymbol{p})$), 也不满足三角形不等式, 所以不满足数学中的距离公理)

Theorem 2.8.3 (有限个值的分布的熵)**:::>** 分布 $\boldsymbol{p}=(p_1,\cdots,p_N)$ 的熵满足

$$H(p)=-\sum_i p_i\ln p_i\leqslant\ln N$$

且等号当且仅当分布在 N 个值均匀时成立, 即以相同概率取 N 个值的分布的熵最大.

Proof. 记分布 $n=\left(\frac{1}{N},\cdots,\frac{1}{N}\right)$. 于是本结论是相对熵不等式 $h(p,n)\geqslant 0$ 的变形. ■

Example 2.8.4 (数学期望固定条件下的离散的最大熵分布)**:::>** 假定存在实数 α, 使 $x_i\geqslant\alpha(i\geqslant 1)$. 对于固定的 $(x_1,\cdots,x_n,\cdots)$ 与 μ, 在满足

$$P(\xi=x_i)=p_i, E\xi=\mu$$

的离散分布 $p=(p_1,\cdots,p_n,\cdots)$ 中, 具有最大熵的分布为 p^*:

$$p_i^*=Ce^{-\lambda_0 x_i}\quad\left(C=\left(\sum_i e^{-\lambda_0 x_i}\right)^{-1}\right)$$

其中 λ_0 满足条件

$$\sum_i e^{-\lambda_0 x_i}<\infty,\sum_i x_i e^{-\lambda_0 x_i}<\infty,\mu\sum_i e^{-\lambda_0 x_i}=\sum_i x_i e^{-\lambda_n x_i}$$

这个最大熵为 $\mu\cdot\lambda_0-\ln C$

Proof. 对于任意 $\lambda \neq 0$, 只要 $\sum e^{-\lambda x_i} < \infty$, 相对熵不等式

$$\begin{aligned} 0 \leqslant h(\boldsymbol{p}, \boldsymbol{p}^*) &= \sum_i p_i \ln \frac{p_i}{C_e x_i} \\ &= \sum_i p_i \ln p_i + \lambda \sum_i p_i x_i - \ln C \end{aligned}$$

也就是

$$H(p) = -\sum_i p_i \ln p_i \geqslant \lambda \mu - \ln C$$

所以

$$H(p^*) = -\sum_i C e^{-\lambda_0} \ln\left[C e^{-\lambda_0 x_i}\right] = \lambda_0 \mu - \ln C \geqslant H(\boldsymbol{p})$$

■

Example 2.8.5 (数学期望和方差都固定时的最大熵):::> 在均值为 μ, 方差为 σ^2, 且取值为 $x_k (k = 1, 2, \cdots)$ 的离散分布中, 分布 $\boldsymbol{p}^*$:

$$p_k^* = C e^{-\frac{(x_k - \alpha)^2}{\beta^2}}, C = \left(\sum_k e^{-\frac{(x_k \cdot a)^2}{\beta^2}} \right)^{-1}$$

的熵最大, 此最大熵为 $\frac{\sigma^2 + (\mu - \alpha)^2}{\beta^2} - \ln C$, 其中 α, β 满足:

$$\sum_k (x_k)^l e^{-\frac{(x_k - a)^2}{\beta^2}} < \infty (l = 0, 1, 2), \mu \sum_k e^{-\frac{(x_k - \alpha)^2}{\beta^2}} = \sum_k x_k e^{-\frac{(x_k - \alpha)^2}{\beta^2}}$$

$$\sigma^2 \left(\sum_k e \frac{(x_k \cdot a)^2}{\beta^2} \right)^2 = \left[\sum_k e^{\frac{(x_k - a)^2}{\beta^2}} \right] \left[\sum_k x_k^2 e^{-\frac{(x_k - a)^2}{\beta^2}} \right] - \left(\sum_k x_k e^{-\frac{(x_k - a)^2}{\beta^2}} \right)^2$$

2.9. Some important continuous random variable

2.9.1 The Uniform Distribution

Definition 2.9.1 (The Uniform Distribution):::> X has a Uniform (a, b) distribution, written $X \sim \text{Uniform}(a, b)$, if

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

where $a < b$. The distribution function is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b. \end{cases}$$

我们可算出:

$$EX = \frac{a+b}{2}, DX = \frac{(b-a)^2}{12}$$

Example 2.9.2:::> 均匀分布的重要性体现在计算机编程模拟分布上. 世界上的分布有无穷多个, 让计算机存储这无穷多个分布的数据非常困难. 故只储存 $U(0,1)$ 的数据, 称之为随机数.

在获得了这个数据后, 有多种方法模拟不同的随机变量, 我们介绍逆函数法. 假定计算机可以模拟 $[0,1]$ 上均匀分布的随机变量 U,U 称为随机数. 事实上, 该随机数是伪随机数, 不具有随机性, 只不过和随机数比较没有显著性差异而已.

那么我们如何解决任意一个随机变量的模拟呢? 我们的想法为: 给定一个分布 μ, 是否存在一个函数 $g(x)$, 使得 $g(U)$ 的分布为 μ?

从理论上来讲, 这是完全可以做到的.

假设 $F(x)$ 为 μ 的分布函数, $F^{-1}(u) = \inf\{x : F(x) \geqslant u\}$. 令 $g(u) = F^{-1}(u)$. 可以证明 $g(U)$ 的分布函数为 $F(x)$.

即:

$$F(F^{-1}(x) = x)$$

以指数分布为例, 直观的结果如下:

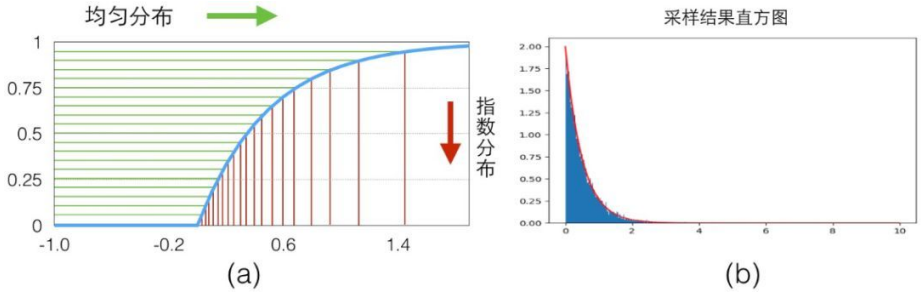


图 2.1: 指数分布逆函数法采样

2.9.2 Exponential Distribution

Definition 2.9.3 (Exponential Distribution)**:::>** X has an Exponential distribution with parameter λ, denoted by $X \sim \text{Exp}(\lambda)$, if

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

where $\lambda > 0$.

此外, 也有书上把指数分布 $E(\lambda)$ 定义为:

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

我们可以看到, 指数分布是 Gamma 分布的一种特殊情况, 其中 $\alpha = 1$, $\lambda = \frac{1}{\beta}$

指数分布的尾部部分:

$$P(X > x) = e^{-\lambda x}$$

Theorem 2.9.4:::> 若 $X \sim \text{E}(\lambda)$, 则

$$EX = \int_0^\infty x \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \Gamma(2) = \frac{1}{\lambda}$$

$$EX^2 = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2} \Gamma(3) = \frac{2}{\lambda^2}$$

因而可得 $DX = \frac{1}{\lambda^2}$

我们也可求得指数分布的矩母函数:

$$M_X(u) = \int_0^{\infty} e^{ux} \lambda e^{-\lambda x} du = \frac{\lambda}{\lambda - u} \quad u < \lambda$$

Theorem 2.9.5:::> 若 X 是连续型非负实值随机变量, 则下列叙述等价:

1. $X \sim \mathrm{E}(\lambda)$
2. X 具有" 无记忆性"
3. X 的死亡率 (风险率 hazard rate) 函数:

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} P(T \leqslant t + \Delta t \mid T > t) = \lambda$$

从第三个叙述可看出, 人类寿命并不符合指数分布, 因为人三岁的" 死亡速率" 和八十岁的" 死亡速率" 是不同的.

Example 2.9.6:::> 泊松分布, 指在单位时间里 (也可以是在时间 t 内), 事件出现的次数, 次数可以是 $x = 0, 1, 2, \cdots$ 直至可数无穷次.

指数分布指的是时间, 是我们关注的事件 A 第一次出现时, 等待的时间.

随机变量 X_1 表示事件 A 出现时我们等待的时间,

$$F_{X_1}(x) = P(X_1 \leqslant x) = 1 - P(X_1 > x)$$

$P(X_1 > x)$ 的概率即为在 $[0, x]$ 时间段内, 事件 A 出现 0 次的概率. 由

$$N(x) \sim \text{ Poisson } (\lambda x)$$

则

$$P(X_1 > x) = P(N(x) = 0) = \frac{e^{-\lambda x}(\lambda x)^0}{0!} = e^{-\lambda x}$$

则

$$F_{X_1}(x) = P(X_1 \leqslant x) = 1 - P(X_1 > x) = 1 - e^{-\lambda x}$$

$$f_{X_1}(x) = F'_{X_1}(x) = \lambda e^{-\lambda x} \sim E\left(\frac{1}{\lambda}\right), x > 0$$

即是一个参数为 $\frac{1}{\lambda}$ 的指数分布.

Example 2.9.7:::> 如果在 $(0,t]$ 发生的理赔次数服从强度为 λ 的 Poisson 过程, 如果记 N_t 为在 $(0,t]$ 发生的理赔次数, 则 N_t 就是强度为 λ 的 Poisson 过程. 记 τ_n 为第 n 次理赔发生的时刻 (其中 $0<\tau_1<\tau_2<\cdots<\tau_n<\cdots$), 则有:

$$\{N_t \geqslant n\} = \{\tau_n \leqslant t\}, n = 1,2,\cdots,$$

因此, τ_n 的分布函数为

$$\begin{aligned} F_{\tau_n}(t) &= P(\tau_n \leqslant t) = 1 - P(N_t \leqslant n-1) \\ &= 1 - \mathrm{e}^{\lambda t}\left[1 + \lambda t + \cdots + \frac{(\lambda t)^n}{(n-1)!}\right] \quad (n \geqslant 1, t > 0) \end{aligned}$$

两端对 t 求导, 可得 τ_n 的分布密度:

$$f_{\tau_n}(t) = \begin{cases} \lambda \mathrm{e}^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}\Big], & t \geqslant 0 \\ 0, & t < 0 \end{cases}$$

从而容易求得 τ_n 的数学期望为

$$E_{\tau_n} = \int_0^{\infty} t f_{t_n}(t)\mathrm{d}t = \frac{n}{\lambda}$$

2.9.3 Gamma distribution

Theorem 2.9.8 (Gamma 函数的性质)**:::>** 若 $\alpha > 0$, 则 Gamma 函数定义为:

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy$$

Gamma 函数具有如下性质:

1. $\Gamma(\alpha+1) = \alpha \times \Gamma(\alpha)$
2. $\Gamma(n+1) = n!$
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Definition 2.9.9 (Gamma distribution)**:::>** For $\alpha > 0$, the Gamma function is defined by $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy$. X has a Gamma distribution with parameters α and β, denoted by $X \sim \text{Gamma}(\alpha,\beta)$, if

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

where $\alpha, \beta > 0$. The exponential distribution is just a Gamma$(1, \beta)$ distribution. If $X_i \sim \text{Gamma}(\alpha_i, \beta)$ are independent, then $\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$. 可加性的证明可以从矩母函数得到.

Theorem 2.9.10:::> Γ 分布的矩母函数:

$$M_X(u) == (\frac{\lambda}{\lambda - u})^{\alpha} \quad u < \lambda$$

2.9.4 Beta distribution

Definition 2.9.11 (Beta 函数)**:::>** 定义 Beta 函数为

$$f(x) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Definition 2.9.12 (Beta distribution)**:::>** X has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$, denoted by $X \sim \text{Beta}(\alpha, \beta)$, if

$$f(x) = x^{\alpha-1}(1-x)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad 0 < x < 1$$

当 $\alpha = 1, \beta = 1$ 时, X 是均匀分布.

当 α 取自然数时, 退化为 Erlang Distribution.

Example 2.9.13 (Gamma 分布与 Beta 分布关系)**:::>** 若 $U \sim \Gamma(\alpha_1, \lambda)$, $V \sim \Gamma(\alpha_2, \lambda)$, 则

$$\frac{U}{U+V} \sim \text{Beta}(\alpha_1, \alpha_2)$$

2.9.5 t and Cauchy Distribution

Definition 2.9.14 (t and Cauchy Distribution)**:::>** X has a t distribution with ν degrees of freedom – written $X \sim t_\nu$ – if

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}$$

The t distribution is similar to a Normal but it has thicker tails. In fact, the Normal corresponds to a t with $\nu = \infty$.

The Cauchy distribution is a speccial case of the t distribution corresponding to $\nu = 1$. The density is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

To see that this is indeed a density:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{dx}{1+x^2} = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{d\tan^{-1}(x)}{dx} \\ &= \frac{1}{\pi}\left[\tan^{-1}(\infty) - \tan^{-1}(-\infty)\right] = \frac{1}{\pi}\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1\end{aligned}$$

Theorem 2.9.15 (柯西分布的期望不存在):::>

$$\int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{xdx}{1+x^2}$$

但这个积分应该被表示为

$$\lim_{A=-\infty,B=\infty}\frac{1}{\pi}\int_{A}^{B}\frac{xdx}{1+x^2}$$

该积分不存在, 故期望也不存在.

注意, 这个积分并不能被表示为柯西主值积分的形式:

$$\lim_{A=\infty}\frac{1}{\pi}\int_{-A}^{A}\frac{xdx}{1+x^2} = 0$$

之前我们在对称分布及其期望中提到, 对称分布期望若存在一定为零, 这里与柯西分布期望不存在因而不为 0 的结果并不矛盾.

2.9.6 The χ^2 distribution

Definition 2.9.16 (The χ^2 distribution):::> X has a χ^2 distribution with p degrees of freedom – written $X \sim \chi^2_p$ – if

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

If $Z_1, \ldots, Z_p$ are independent standard Normal random variables then $\sum_{i=1}^{p} Z_i^2 \sim \chi_p^2$

2.9.7 Gaussian Distribution

Definition 2.9.17 (Normal(Gaussian))**:::>** X has a Normal (or Gaussian) distribution with parameters μ and σ, denoted by $X \sim N(\mu, \sigma^2)$, if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. The parameter μ is the "center" (or mean) of the distribution and σ is the "spread"(or standard deviation) of the distribution.

We say that X has a standard Normal distribution if $\mu = 0$ and $\sigma = 1$. Tradition dictates that a standard Normal random variable is denoted by Z. The PDF and CDF of a standard Normal are denoted by $\varphi(z)$ and $\Phi(z)$.

Here are some useful facts:

1. If $X \sim N(\mu, \sigma^2)$, then $Z = (X-\mu)/\sigma \sim N(0,1)$ 这个例子在实际应用中具有重要的意义, 也即如果我们在遇到正态分布时, 第一步应该进行的工作便是将其标准化. 这个结论的证明可以将转换前后 CDF 的表达式计算出来即可得.

2. If $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$

3. If $X_i \sim N\left(\mu_i, \sigma_i^2\right), i = 1, \ldots, n$ are independent, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

这个看起来很不直观, 为什么几个独立的高斯分布加起来还是高斯分布呢?

给定两个独立正态分布 $X_1 \sim N(\mu_1, \sigma_1^2) X_2 \sim N(\mu_2, \sigma_2^2)$, 其概率密度函数分别为 f_1, f_2. 设随机变量 $Z = X_1 + X_2$, 那么 Z 的概率密度函数 $f_Z(z)$ 是什么呢?

我们的直观反应是 $f_Z(z)=f_1(z)+f_2(z)$, 这里就出错了. 举个例子, 当 $Z=1$ 时, 可以是 $X_1=0,X_2=1$, 也可以是 $X_1=1,X_2=0$, 总之 $X_1+X_2=Z$.

所以, 随机变量 Z 的概率密度函数其实是

$$f_Z(z)=\int_{-\infty}^{+\infty} f_1(x)f_2(z-x)dx$$

这个正是卷积形式.

$$\begin{aligned}
f_Z(z)&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{(z-x-\mu_Y)^2}{2\sigma_Y^2}\right]\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left[-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right]dx\\
&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y}\exp\left[-\frac{\sigma_X^2(z-x-\mu_Y)^2+\sigma_Y^2(x-\mu_X)^2}{2\sigma_X^2\sigma_Y^2}\right]dx\\
&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y}\\
&\exp\left[-\frac{\sigma_X^2\left(z^2+x^2+\mu_Y^2-2xz-2z\mu_Y+2x\mu_Y\right)+\sigma_Y^2\left(x^2+\mu_X^2-2x\mu_X\right)}{2\sigma_Y^2\sigma_X^2}\right]dx\\
&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sqrt{2\pi}\sigma_X\sigma_Y}\\
&\exp\left[-\frac{x^2\left(\sigma_X^2+\sigma_Y^2\right)-2x\left(\sigma_X^2(z-\mu_Y)+\sigma_Y^2\mu_X\right)+\sigma_X^2\left(z^2+\mu_Y^2-2z\mu_Y\right)+\sigma_Y^2\mu_X^2}{2\sigma_Y^2\sigma_X^2}\right]dx
\end{aligned}$$

我们定义 $\sigma_Z = \sqrt{\sigma_X^2 + \sigma_Y^2}$

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_Z} \frac{1}{\sqrt{2\pi}\frac{\sigma_X\sigma_Y}{\sigma_Z}}$$
$$\exp\left[-\frac{x^2 - 2x\frac{\sigma_X^2(z-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_Z^2} + \frac{\sigma_X^2\left(z^2+\mu_Y^2-2z\mu_Y\right)+\sigma_Y^2\mu_X^2}{\sigma_Z^2}}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_Z}\right)^2}\right] dx$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_Z} \frac{1}{\sqrt{2\pi}\frac{\sigma_X\sigma_Y}{\sigma_Z}}$$
$$\exp\left[-\frac{\left(x - \frac{\sigma_X^2(z-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_Z^2}\right)^2 - \left(\frac{\sigma_X^2(z-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_Z^2}\right)^2 + \frac{\sigma_X^2(z-\mu_Y)^2+\sigma_Y^2\mu_X^2}{\sigma_Z^2}}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_Z}\right)^2}\right] dx$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_Z}$$
$$\exp\left[-\frac{\sigma_Z^2\left(\sigma_X^2(z-\mu_Y)^2 + \sigma_Y^2\mu_X^2\right) - \left(\sigma_X^2(z-\mu_Y) + \sigma_Y^2\mu_X\right)^2}{2\sigma_Z^2\left(\sigma_X\sigma_Y\right)^2}\right]$$
$$\frac{1}{\sqrt{2\pi}\frac{\sigma_X\sigma_Y}{\sigma_Z}} \exp\left[-\frac{\left(x - \frac{\sigma_X^2(z-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_Z^2}\right]}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_Z}\right)^2}\right] dx$$
$$= \frac{1}{\sqrt{2\pi}\sigma_Z} \exp\left[-\frac{(z-(\mu_X+\mu_Y))^2}{2\sigma_Z^2}\right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\frac{\sigma_X\sigma_Y}{\sigma_Z}} \exp\left[-\frac{\left(x - \frac{\sigma_X^2(z-\mu_Y)+\sigma_Y^2\mu_X}{\sigma_Z^2}\right)^2}{2\left(\frac{\sigma_X\sigma_Y}{\sigma_Z}\right)^2}\right] dx$$

我们可以看到积分符号右边就是一个正态分布的密度函数形式, 所以这个积分结果为 1, 然后我们得到最终形式

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma_Z} \exp\left[-\frac{(z-(\mu_X+\mu_Y))^2}{2\sigma_Z^2}\right]$$

4. It follows that if $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned}\mathbb{P}(a < X < b) &= \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\end{aligned}$$

Thus we can compute any probabilities we want as long as we can compute the CDF $\Phi(z)$ of a standard Normal.

Example 2.9.18:::> 正态分布的线性组合不一定是正态的: 若 $X \sim N(0,1)$, 则 $Y=-X \sim N(0,1)$, 然而, $X+Y=0$ 恒成立, 显然不是正态的.

上边所证明的内容是: 独立的正态分布的线性组合一定是正态的.

Theorem 2.9.19 (高斯分布的期望与方差)**:::>** 概率密度函数:

$$f(x)=\frac{1}{\sqrt{2\pi}\delta}e^{-\frac{(x-u)^2}{2\delta^2}}$$

证明积分结果为 1:

$$\begin{aligned}F&=\int_{-\infty}^{+\infty}f(x)dx=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\delta}e^{-\frac{(x-u)^2}{2\delta^2}}dx\\&=\frac{1}{\sqrt{2\pi}\delta}\int_{-\infty}^{+\infty}e^{-\frac{(x-u)^2}{2\delta^2}}d(x-u)\\&=\frac{1}{\sqrt{2\pi}\delta}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2\delta^2}}dx\end{aligned}$$

利用双重积分方法进行转化:

$$\begin{aligned}F^2&=\frac{1}{\sqrt{2\pi}\delta}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2\delta^2}}dx\frac{1}{\sqrt{2\pi}\delta}\int_{-\infty}^{+\infty}e^{-\frac{y^2}{2\delta^2}}dy\\&=\frac{1}{2\pi\delta^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}e^{-\frac{x^2+y^2}{2\delta^2}}dxdy\end{aligned}$$

令 $x=r\sin\theta, y=r\cos\theta$, 坐标系转换到极坐标系进行积分:

$$\begin{aligned}F^2&=\frac{1}{2\pi\delta^2}\int_0^{2\pi}\int_0^{+\infty}e^{-\frac{r^2}{2\delta^2}}rdrd\theta\\&=\frac{1}{2\pi\delta^2}\int_0^{2\pi}d\theta\int_0^{+\infty}e^{-\frac{r^2}{2\delta^2}}rdr\\&=\frac{1}{\delta^2}\int_0^{+\infty}e^{-\frac{r^2}{2\delta^2}}rdr\\&=\int_0^{+\infty}e^{-\frac{r^2}{2\delta^2}}d\left(\frac{r^2}{2\delta^2}\right)\\&=\int_0^{+\infty}e^{-m}dm\\&=1\end{aligned}$$

期望:

$$\begin{aligned}
E(x) &= \int_{-\infty}^{+\infty} x f(x) dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(x-u)^2}{2\delta^2}} x dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{x^2}{2\delta^2}} (x+u) dx \\
&= \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{x^2}{2\delta^2}} x dx}_{0} + u \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{x^2}{2\delta^2}} dx}_{1} \\
&= u
\end{aligned}$$

其中, 由于 $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{x^2}{2\delta^2}} x dx$ 为奇函数, 积分为 0, 因此可得高斯分布的期望: $E(x) = u$

方差: 可以直接用分部积分推导得:

$$\begin{aligned}
V &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(x-u)^2}{2\delta^2}} (x-u)^2 dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{x^2}{2\delta^2}} x^2 dx \\
&= -\frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x d\left(e^{-\frac{x^2}{2\delta^2}}\right) \\
&= -\frac{\delta}{\sqrt{2\pi}} (\underbrace{x e^{-\frac{x^2}{2\delta^2}} \Big|_{-\infty}^{+\infty}}_{0} - \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\delta^2}} dx) \\
&= \frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\delta^2}} dx \\
&= \delta^2 \cdot \frac{1}{\sqrt{2\pi}\delta} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2\delta^2}} dx \\
&= \delta^2
\end{aligned}$$

也可以利用:

$$\begin{aligned}
V(X) &= E\left(x^2\right) - E^2(x) \\
&= E\left(x^2\right) - u^2
\end{aligned}$$

而用与上类似的方法可以类似地得到:

$$E\left(x^2\right) = \delta^2 + u^2$$

Theorem 2.9.20 (正态分布的矩母函数)**:::>**

$$\begin{aligned} M_X(t) &= \mathrm{E}\left(e^{tX}\right) \\ &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} e^{tx} dx \\ &= e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)} \end{aligned}$$

推导时可以通过在指数函数内配方得到.

阶数	EX^k	中心矩
0	1	0
1	μ	0
2	$\mu^2+\sigma^2$	σ^2
3	$\mu^3+3\mu\sigma^2$	0
4	$\mu^4+6\mu^2\sigma^2+3\sigma^4$	$3\sigma^4$

而若是标准正态分布, 我们有结论:

$$E\,(x^*)^n = \begin{cases} (2k-1)!! & n=2k \\ 0, & n=2k+1 \end{cases}$$

Theorem 2.9.21 (高斯分布的导出)**:::>** 高斯分布的导出: 实际的测量中, 若 μ 是分布均值的真实值, 当然真实值我们永远都不可能知道, 因为我们活在一个误差的世界, 然后现在希望根据观测值 $x_1, x_2, \cdots, x_n$ 尽可能的去估计它. 首先我们记观察误差 $x_i - \mu$ 的分布密度函数为 $p\,(x_i-\mu)$, 然后给以下假设

1. $p(x)$ 关于 $x=0$ 对称, 且对于一切 x 成立 $p(x)>0$
2. $p(x)$ 具有连续的导函数.

由于我们的观察误差 $x_i-\mu$ 的分布密度函数为 $p\,(x_i-\mu)$, 那么此时的似然函数就是

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

实际上, 这个似然函数刻画了这组观测值落在真实均值 μ 附近的可能性大小.

当然此处高斯还给出了一个重要的假设: 观察值的平均值 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 作为未知参数 μ 的估计值时使得似然 $L(\mu)$ 最大.

若 $\overline{x}$ 使得似然函数似然 $L(\mu)$ 最大, 则它的必要条件是关于参数 μ 的偏导数在 $\mu=\overline{x}$ 处为 0.

即:

$$\left.\frac{d\ln L(\mu)}{d\mu}\right|_{\mu=\overline{x}}=0$$

此时, 我们记 $\frac{d\ln p(x)}{dx}=g(x)$, 由链式法则, 则 $g(x)=\frac{p'(x)}{p(x)}$, 同时根据连续函数假设, 我们有:

$$\begin{aligned}\left.\frac{d\ln L(\mu)}{d\mu}\right|_{\mu=\overline{x}} &= \left.\frac{d\ln\prod_{i=1}^{n}p(x_i-\mu)}{d\mu}\right|_{\mu=\overline{x}}\\ &= \left.\frac{d\sum_{i=1}^{n}\ln p(x_i-\mu)}{d\mu}\right|_{\mu=\overline{x}}\\ &= \sum_{i=1}^{n}\left.\frac{d\ln p(x_i-\mu)}{d\mu}\right|_{\mu=\overline{x}}\\ &= -\sum_{i=1}^{n}\left.g(x_i-\mu)\right|_{\mu=\overline{x}}\\ &= -\sum_{i=1}^{n}g(x_i-\overline{x})\\ &= 0\end{aligned}$$

最后, 我们可以得到:

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

若此时简化考虑, 令 $n=2$, 则上式可进一步化简为:

$$g(x_1-\overline{x})+g(x_2-\overline{x})=0$$

由于 $x_1-\overline{x}=-(x_2-\overline{x})$ 以及 x_1,x_2 的任意性, 我们可以得到 $g(x)$ 是个中心对称函数, 即 $g(x)=-g(-x)$ 对一切实数 x 成立.

此外, 当 $n=3$ 时, 上述方程可以简化得到

$$g(x_1-\overline{x})+g(x_2-\overline{x})+g(x_3-\overline{x})=0$$

由于 $x_1-\overline{x}=-[(x_2-\overline{x})+(x_3-\overline{x})]$ 以及 x_1,x_2,x_3 的任意性, 我们可以得到对一切实数 x,y 成立

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

这也是个大名鼎鼎的方程, 叫柯西函数方程, 此类方程在有理数范围内, 可以得到唯一的通解 $g(x)=bx$ 则:

$$\frac{p'(x)}{p(x)}=bx\Rightarrow\frac{\mathrm{d}p(x)}{p(x)}=bx\cdot\mathrm{d}x$$

则可得:

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

由于 $p(x)$ 是密度函数, 因此我们需要令 $b<0$, 否则指数函数对应的概率很容易就大于 1, 之后积分就无穷了. 不妨我们记 $b=-\frac{1}{\sigma^2}$, 则

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

且我们需要 $\int_{-\infty}^{\infty}p(x)dx=1$, 则可得 $K=\frac{1}{\sqrt{2\pi\sigma^2}}$.

故有

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

Theorem 2.9.22:::> 高斯分布函数的一些性质

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},\quad x\in\mathbb{R}$$

1. 高斯分布函数的光滑性非常好, 无穷可微
2. 高斯分布函数的最大值为 $\frac{1}{\sigma\sqrt{2\pi}}$, 在 μ 处取得
3. 高斯分布函数的驻点为 $x=\mu$
4. 高斯分布函数的拐点 (二阶导为零) 为 $x_1=\mu-\sigma$, $x_2=\mu+\sigma$

Example 2.9.23:::> Suppose that $X \sim N(3,5)$. Find $\mathbb{P}(X > 1)$. The solution is

$$\mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(Z < \frac{1-3}{\sqrt{5}}\right) = 1 - \Phi(-0.8944) = 0.81$$

Now find $q = \Phi^{-1}(0.2)$. This means we have to find q such that $\mathbb{P}(X < q) = 0.2$

We solve this by writing

$$0.2 = \mathbb{P}(X < q) = \mathbb{P}\left(Z < \frac{q-\mu}{\sigma}\right) = \Phi\left(\frac{q-\mu}{\sigma}\right)$$

From the Normal table, $\Phi(-0.8416) = 0.2$. Therefore,

$$-0.8416 = \frac{q-\mu}{\sigma} = \frac{q-3}{\sqrt{5}}$$

and hence $q = 3 - 0.8416\sqrt{5} = 1.1181$

Theorem 2.9.24 (高斯分布不等式)**:::>** 由切比雪夫不等式我们知道

$$P(|X - EX| < 3\sqrt{DX} \geqslant 1 - \frac{DX}{9DX}) = \frac{8}{9}$$

注意这个结论对任何随机变量均成立.

而工程上常用的 3σ 原则, 我们可以这样推导:

$$P(|X - EX| < 3\sqrt{DX} = p\left(|\frac{x-\mu}{\sigma}| < 3\right) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1$$

然后查表, 求得值为 99.7%

Theorem 2.9.25 (Mill's Inequality)**:::>** Let $Z \sim N(0,1)$. Then

$$\mathbb{P}(|Z| > t) \leqslant \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}$$

2.10. Multivariate Distributions and IID Samples

Definition 2.10.1:::> Let $X = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are random variables. We call X a random vector. Let $f(x_1, \ldots, x_n)$ denote the PDF. It is

possible to define their marginals, conditionals etc. We say that $X_1,\ldots,X_n$ are independent if, for every $A_1,\ldots,A_n$,

$$\mathbb{P}(X_1 \in A_1,\ldots,X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

It suffices to check that $f(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$

Definition 2.10.2:::> If $X_1,\ldots,X_n$ are independent and each has the same marginal distribution with CDF F, we say that $X_1,\ldots,X_n$ are IID(independent and identically districuted) and we write

$$X_1,\ldots X_n \sim F$$

If F has density f we also write $X_1,\ldots X_n \sim f$. We also call $X_1,\ldots,X_n$ a random sample of size n from F.

独立同分布具有相同的概率性质, 但是不能由此认为 $P(X=Y)=1$.

举一个反例:

X/Y	0	1	
0	1/4	1/4	1/2
1	1/4	1/4	1/2
	1/2	1/2	1

这两个情况下 X 和 Y 是独立同分布, 然而 $P(X=Y)=\frac{1}{2}$.

Example 2.10.3:::> 若 X_1,X_2,X_3 独立同分布, 求

$$E\left(\frac{2X_1+3X_2}{X_1+X_2+X_3}\right)$$

求解此式需要用到对称性. 由对称性可得,

$$E\left(\frac{X_1}{X_1+X_2+X_3}\right) = E\left(\frac{X_2}{X_1+X_2+X_3}\right) = E\left(\frac{X_3}{X_1+X_2+X_3}\right) = \frac{1}{3}$$

则相加可得结果为 $\frac{5}{3}$

Example 2.10.4:::> 若 $X_1,\ldots,X_n$ 为正的独立随机变量, 服从相同分布, 证明:

$$\mathrm{E}\left(\frac{\mathrm{X_1+X_2+\cdots+X_k}}{\mathrm{X_1+X_2+\cdots+X_n}}\right) = \frac{\mathrm{k}}{\mathrm{n}}$$

证明: 有对称性 $\frac{X_1}{X_1+\ldots+X_n}, \frac{X_2}{X_1+\ldots+X_n}, \cdots \frac{X_n}{X_1+\ldots+X_n}$ 同分布, 故

$$E\left(\frac{X_1}{X_1+\ldots+X_n}\right)=E\left(\frac{X_2}{X_1+\ldots+X_n}\right)=\cdots=E\left(\frac{X_n}{X_1+\ldots+X_n}\right)$$

而

$$E\left(\frac{X_1+\ldots+X_n}{X_1+\ldots+X_n}\right)=1$$

故

$$E\left(\frac{X_i}{X_1+\ldots+X_n}\right)=\frac{1}{n}$$

因此由期望的线性可得,

$$E\left(\frac{X_1+\ldots+X_k}{X_1+\ldots+X_n}\right)=\frac{k}{n}$$

Definition 2.10.5 (Multinomial)**:::>** The multivariate version of a Binomial is called a Multinomial. Consider drawing a ball from an urn which has balls with k different colors labled "color 1, color 2, . . . , color k." Let $p=(p_1,\ldots,p_k)$ where $p_j \geqslant 0$ and $\sum_{j=1}^{k} p_j = 1$ and suppose that p_j is the probability of drawing a ball of color j. Draw n times (independent draws with replacement) and let $X=(X_1,\ldots,X_k)$ where X_j is the number of times that color j appears. Hence, $n=\sum_{j=1}^{k} X_j$. We say that X has a Multinomial (n,p) distribution written $X \sim \text{Multinomial}(n,p)$. The probability function is

$$f(x)=\binom{n}{x_1\ldots x_k} p_1^{x_1}\cdots p_k^{x_k}$$

where

$$\binom{n}{x_1\ldots x_k}=\frac{n!}{x_1!\cdots x_k!}$$

Definition 2.10.6 (多维连续随机变量)**:::>** 定义 (X,Y) 为二维连续型随机变量, 如果存在非负可积函数 $f(x,y)$, 使得 $\forall x,y \in R$ 有

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv$$

我们称 $f(x,y)$ 为 (X,Y) 的联合分布密度函数.

Definition 2.10.7 (联合分布密度函数的性质)**:::>**

1. $f(x,y)$ 不唯一
2. $f(x,y)$ 为联合分布密度 $\Leftrightarrow \begin{cases} f(x,y)\geqslant 0 \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1 \end{cases}$
3. 若 $f(x,y)$ 在点 (x,y) 处连续, 则 $f(x,y)=\frac{\partial^2 F(x,y)}{\partial x \partial y}$
4. $P((X,Y)\in A)=\iint_A f(x,y)dxdy$
5. $$\mathrm{E}g(X,Y)=\iint_{R^2} g(x,y)f(x,y)dxdy$$

   要求后者绝对收敛. 这其实就是 Lotus 定理. 特别地,

   $$EX=\iint_{R^2} xf(x,y)dxdy$$

   $$EX^2=\iint_{R^2} x^2 f(x,y)dxdy$$

   从而可以计算 DX

   $$EXY=\iint_{R^2} xyf(x,y)dxdy$$

   从而可以计算 $\mathrm{Cov}(X,Y), r_{X,Y}$
6. 边缘分布密度的求法

   $$f_X(x)=\int_{-\infty}^{\infty} f(x,y)dy, \quad f_Y(y)=\int_{-\infty}^{\infty} f(x,y)dx$$

7. X 与 Y 相互独立 $\Leftrightarrow F(x,y)=F_X(x)F_Y(y) \quad \forall x,y$

   $$\Leftrightarrow p_{ij}=p_i\cdot p_j \forall i,j \quad \text{（离散型）}$$
   $$\Leftrightarrow f(x,y)=f_X(x)f_Y(y) \quad \forall x,y \text{ a.e.}$$
   $$\Leftrightarrow f_{X|Y}(x\mid y)=f_X(x) \quad \forall x \text{ a.e.}$$

   注意其中的 *a.e.*, 即对于连续型随机变量来言, 以上的等式不必每一点上都成立, 可以在几个点或一条线上不成立, 但不能有一块面积大于 0 的区域不成立. *a.e.* 的意思是 almost everywhere.

8. 若

$$f(x,y)=h(x)g(y)$$

其中 $a\leqslant x\leqslant b$ 和 $c\leqslant y\leqslant d$ 则可知, X,Y 独立, 且

$$f_X(x)=C_1\cdot h(x)$$

$$f_Y(y)=C_2\cdot g(y)$$

即矩形区域, 可分离变量的情况下, 则两者是独立的, 且边缘分布方便可得, 注意要估计到归一性即积分结果为 1.

Example 2.10.8:::> 设随机向量 (X,Y) 的联合密度函数为

$$f(x,y)=\frac{1}{2x^2y},\quad 1\leqslant x<\infty,\quad \frac{1}{x}<y<x$$

判断 X 与 Y 是否相互独立.

在求边缘分布前, 最重要的是在平面上画出 x,y 分布存在的区域, 之后再进行积分. 不然极易出现积分范围错误的情况.

求得

$$f_X(x)=\begin{cases}\frac{1}{x^2}\ln x, & x\geqslant 1,\\ 0, & \text{否则},\end{cases}$$

$$f_Y(y)=\begin{cases}\frac{1}{2}, & 0<y<1,\\ \frac{1}{2y^2}, & 1\leqslant y<+\infty,\\ 0, & \text{其他}.\end{cases}$$

故两者不相互独立.

Definition 2.10.9 (二维均匀分布)**:::>** 称随机向量 $(X,Y)\sim U(D)$, 若其联合概率密度函数为

$$f(x,y)=\begin{cases}\frac{1}{|D|}, & (x,y)\in D\\ 0 & \text{其它}\end{cases}$$

显然有

$$P((X,Y)\in A)=\frac{|A\cap D|}{|D|},\quad A\subset R^2$$

Theorem 2.10.10:::> 若 $(X,Y)\sim[\{[a,b]\times[c,d]\})$, 则 X,Y 独立, 且

$$X\sim U[a,b],\quad Y\sim U[c,d]$$

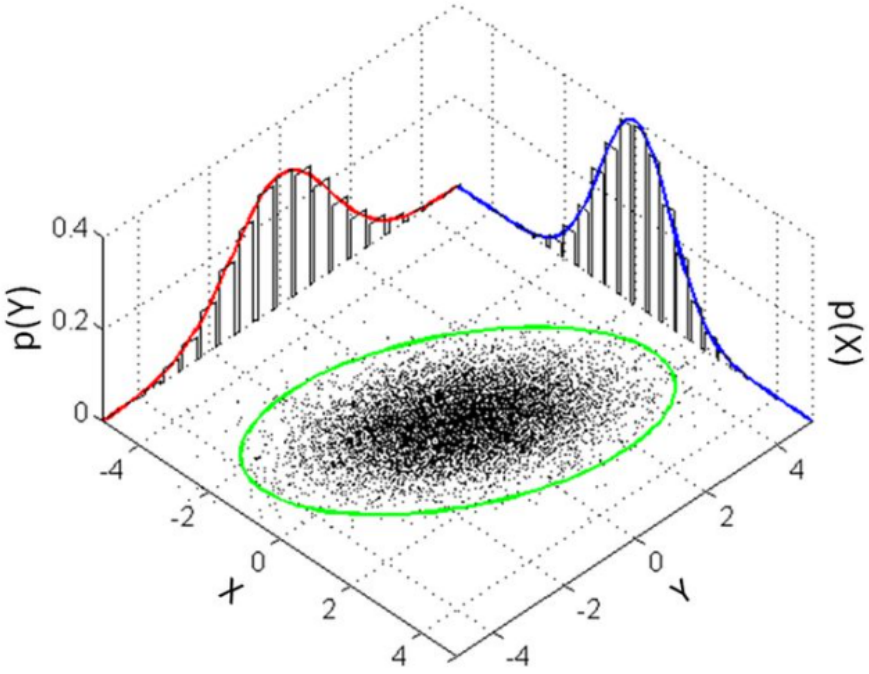
Definition 2.10.11 (二维正态分布)**:::>** 称随机向量 (X,Y) 服从参数为 $\mu_1,\mu_2,\rho,\sigma_1^2,\sigma_2^2$ 的二元正态分布 (记为 $(X,Y)\sim N(\mu_1,\mu_2,\rho,\sigma_1^2,\sigma_2^2)$), 若其联合概率密度函数为

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}Q(x,y)\right\}$$

其中 $\sigma_1,\sigma_2>0, Q(x,y)$ 是二次型

$$Q(x,y)=\frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\quad(|\rho|<1)$$

特别地, 称 $N(0,0,\rho,1,1)$ 为二元 ρ— 标准正态分布.



二维正态分布的图像

即每一个维度都是一个正态分布.

将函数 Q 用类似配方法的方式进行转化并将含 x 的式子与含 y 的式子进行分拆后进行双重积分, 可得结果为 1.

设随机向量 (X,Y) 服从二元 ρ - 标准正态分布. 我们求 Y 的边缘密度, 就是计算 $f(x,y)$ 关于 x 的积分. 为此注意在把 y 作为参数时, 被积函数 $f(x,y)$ 作为积分变量 x 的函数, 在指数位置上是负号后面一个二次函数. 因此可以利用对二次函数配方计算此积分. 即

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(x-\rho y)^2\right)\cdot\exp\left(-\frac{y^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{y^2}{2}}\cdot\frac{1}{\sqrt{2\pi(1-\rho^2)}}e^{\frac{(x-\rho y)^2}{2(1-\rho^2)}} \end{aligned}$$

其中第一个因子不含 x, 第二个因子作为 x 的函数恰是 $N(\rho y, 1-\rho^2)$ 的概率密度函数, 它对 x 积分是 1 , 于是我们得到 Y 的边缘密度

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}x = \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}$$

这正是标准正态分布的密度, 即 $N(0,1)$ 的密度. 对称地, X 的边缘密度也是 $N(0,1)$.

Definition 2.10.12 (二维正态分布的标准化)**:::>** 若 $(X,Y)\sim N(\mu_1,\mu_2,\rho,\sigma_1^2,\sigma_2^2)$, 令 $X^*=\frac{X-\mu_1}{\sigma_1}, Y^*=\frac{Y-\mu_2}{\sigma_2}$, 则

$$(X^*,Y^*)\sim N(0,0,\rho,1,1)$$

Proposition 2.10.13 (二维正态分布的性质)**:::>** 若 $(X,Y)\sim N(\mu_1,\mu_2,\rho,\sigma_1^2,\sigma_2^2)$, 则有

1. 二维正态分布的边缘是正态:

$$X\sim N\left(\mu_1,\sigma_1^2\right), Y\sim N\left(\mu_2,\sigma_2^2\right)$$

注意此结论反过来并不成立, 即若 $X\sim N(\mu_1,\sigma_1^2), Y\sim N(\mu_2,\sigma_2^2)$, 则 (X,Y) 不一定是二维正态分布. 而如果 X,Y 是独立的, 则此结论成立.

2.

$$Y|X=x\sim N\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1),\sigma_2^2\left(1-\rho^2\right)\right)$$

进而可得:

$$E(Y|X)=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(X-\mu_1)$$

$$V(Y|X)=\sigma_2^2\left(1-\rho^2\right)$$

这是一个非常重要的结果, 因为它的意义是: 若 X 和 Y 符合二维正态分布, 则用 X 预测 Y 的最佳预测是线性的.

3. 相关系数: 在计算相关系数时, 为简便可直接计算标准化后的相关系数, 即

$$r_{X,Y}=r_{X^*,Y^*}=E(X^*Y^*)$$

而由上结论可得:

$$E(Y^*|X^*)=\rho X^*$$

而

$$E(X^*Y^*)=EE(X^*Y^*|X^*)=E(X^*E(Y^*|X^*))=\rho E((X^*)^2)$$

又因为

$$X^*\sim N(0,1)$$

则原式的结果即为 ρ, 所以 ρ 就是 X,Y 的相关系数.

4. X,Y 相互独立 $\Leftrightarrow X^*,Y^*$ 相互独立 $\Leftrightarrow \rho=0$

5. 若 X,Y 服从二维正态分布, 则 $aX+bY$ 也服从二维正态分布:

$$(aX+bY)\sim N(a\mu_1+b\mu_2,a^2\sigma_1^2+b^2\sigma_2^2+2ab\rho\sigma_1\sigma_2)$$

Example 2.10.14:::> 上述期望与方差的计算具有借鉴意义, 如计算 $E(X^2Y^2)$ 时, 可以先转化为

$$EE(X^2Y^2|X)=E(X^2E(Y^2|X))$$

而期望与方差我们已求得, 即二阶矩已求得, 则上式可得.

Example 2.10.15:::> 上述最后一条性质是说, 若 X,Y 服从二维正态分布, 则独立和不相关是等价的. 这是因为在矩形区域中, 若 $\rho=0$, 则可以分离变量, 于是即为独立.

此外, 除二维正态分布外, 若 X 是一个二值变量, Y 也是一个二值变量, 则 X,Y 独立与不相关也是等价的.

Definition 2.10.16 (n 维正态分布)**:::>**

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\sum)}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \sum^{-1} (\vec{x}-\vec{\mu})\right)$$

其中, $\vec{x} = (x_1, x_2, \ldots, x_n)^T$, x_i 均为一维正态随机变量, x_i 的期望为 μ_i . $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$, $\sum$ 是 $\vec{x}$ 的 covariance matrix (设定为 positive semidefinite matrix):

$$\sum = \begin{bmatrix} \operatorname{cov}(x_1, x_1) & \operatorname{cov}(x_1, x_2) & \ldots & \operatorname{cov}(x_1, x_n) \\ \operatorname{cov}(x_2, x_1) & \operatorname{cov}(x_2, x_2) & \ldots & \operatorname{cov}(x_2, x_n) \\ \vdots & \vdots & \ddots & v\,\text{dots} \\ \operatorname{cov}(x_n, x_1) & \operatorname{cov}(x_n, x_2) & \ldots & \operatorname{cov}(x_n, x_n) \end{bmatrix}$$

与一维正态分布的 PDF 对比一下:

$$f(\vec{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

大致可以观察出来这样一个对应关系:

$$\sum \Leftrightarrow \sigma, \quad \vec{\mu} \Leftrightarrow \mu$$

在多维 Gauss 分布及其特征函数中还会借助特征函数的工具进行进一步的研究.

Definition 2.10.17 (高斯向量)**:::>** 对于二维高斯向量:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

2.11. Conditional Distributions

Definition 2.11.1:::> If X and Y are discrete, then we can compute the conditional distribution of X given that we have observed $Y = y$. Specially, $\mathbb{P}(X = x \mid Y = y) = \mathbb{P}(X = x, Y = y)/\mathbb{P}(Y = y)$.

Definition 2.11.2 (conditional probability mass function)**:::>**

$$f_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

if $f_Y(y) > 0$

Definition 2.11.3:::> For continuous random variables, the conditional probability density function is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

assuming that $f_Y(y) > 0$. Then,

$$\mathbb{P}(X \in A \mid Y = y) = \int_A f_{X|Y}(x \mid y)dx$$

From the definition of the conditional density, we see that $f_{X,Y}(x,y) = f_{X|Y}(x \mid y)f_Y(y) = f_{Y|X}(y \mid x)f_X(x)$

Example 2.11.4:::> 在 $\{1,2,3,4\}$ 中任取一数, 记为 X, 再从 $\{1,2,\cdots,X\}$ 中任取一数, 记为 Y, 求 (X,Y) 的联合分布律以及关于 Y 的边缘分布.

$$P(X=i,Y=j) = P(X=i)P(Y=j \mid X=i) = \begin{cases} \frac{1}{4}\times\frac{1}{i}, & 1 \leqslant j \leqslant i \\ 0, & j > i \end{cases}$$

Y	1	2	3	4	$p_{i.}$
1	1/4	0	0	0	1/4
2	1/8	1/8	0	0	1/4
3	1/12	1/12	1/12	0	1/4
4	1/16	1/16	1/16	1/16	1/4
$p_{\cdot j}$	25/48	13/48	5/36	7/36	1

它们独立吗? 一个很简单的判断方法, 表格里出现了 0, 专门去判断那一处.

Example 2.11.5:::> Let

$$f(x,y) = \begin{cases} x+y & \text{if } 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

Let us find $\mathbb{P}(X < 1/4 \mid Y = 1/3)$.

We have

$$f_Y(y) = \int_0^1 (x+y)dx = \int_0^1 x\,dx + \int_0^1 y\,dx = \frac{1}{2} + y$$

Hence,

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y}{y+\frac{1}{2}}$$

So,

$$\begin{aligned} P\left(X < \frac{1}{4} \mid Y = \frac{1}{3}\right) &= \int_0^{1/4} f_{X|Y}\left(x \mid \frac{1}{3}\right) dx \\ &= \int_0^{1/4} \frac{x+\frac{1}{3}}{\frac{1}{3}+\frac{1}{2}} dx = \frac{\frac{1}{32}+\frac{1}{12}}{\frac{1}{3}+\frac{1}{2}} = \frac{11}{80} \end{aligned}$$

Example 2.11.6:::> Suppose that $X \sim \text{Uniform}(0,1)$. After obtaining a value of X we generate $Y \mid X = x \sim \text{Uniform}(x,1)$. What is the marginal distribution of Y?

First note that,

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leqslant x \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

And

$$f_{Y|X}(y \mid x) = \begin{cases} \frac{1}{1-x} & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

So,

$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x) = \begin{cases} \frac{1}{1-x} & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal for Y is

$$f_Y(y) = \int_0^y f_{X,Y}(x,y) dx = \int_0^y \frac{dx}{1-x} = -\int_1^{1-y} \frac{du}{u} = -\log(1-y)$$

for $0 < y < 1$.

Definition 2.11.7 (条件密度与条件期望)**:::>** 设 (X,Y) 为二元连续型随机向量, 其联合概率密度函数为 $f(x,y)$. 对 $A \subset \boldsymbol{R}$, 若 $P(X \in A) > 0$, 我们可以定义在已知 $X \in A$ 的条件下, Y 的条件概率

$$P(a < Y \leqslant b \mid X \in A) = \frac{P(X \in A, a < Y \leqslant b)}{P(X \in A)}$$

如果对任意 $y \in \boldsymbol{R}$, 存在非负函数 $f_{Y|X\in A}(y)$, 满足

$$P(a < Y \leqslant b \mid X \in A) = \int_a^b f_{Y|X\in A}(y)dy$$

则称 $f_{Y|X\in A}(y)$ 为在给定 $X \in A$ 条件下, Y 的条件概率密度函数. 而称 $F_{Y|X\in A}(y) \equiv P(Y \leqslant y \mid X \in A)$ 为在给定 $X \in A$ 条件下, Y 的条件分布函数.

给定 $X \in A$ 条件下, Y 的条件期望定义为

$$E(Y \mid X \in A) = \int_{-\infty}^{\infty} y f_{Y|X\in A}(y)dy$$

和之前一样, 本式要求绝对收敛.

Example 2.11.8:::> 设 $X \sim N(0,1)$, 求 $E(X \mid X > 0)$

显然, 当 $x \leqslant 0$ 时, 有

$$F_{X|X>0}(x) = P(X \leqslant x \mid X > 0) = 0$$

而当 $x > 0$ 时, 我们有

$$F_{X|X>0}(x) = P(X \leqslant x \mid X > 0) = \frac{P(0 < X \leqslant x)}{P(X > 0)} = 2\left[\Phi(x) - \frac{1}{2}\right]$$

故

$$f_{X|X>0}(x) = \begin{cases} 2\varphi(x) & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

从而

$$E(X \mid X > 0) = 2\int_0^{+\infty} x\varphi(x)dx = \sqrt{\frac{2}{\pi}}$$

Definition 2.11.9 (条件分布密度另一种定义)**:::>** 对连续型随机向量 (X, Y), 若 $f_X(x) > 0$, 称 $\frac{f(x,y)}{f_X(x)}$ 为在给定 $X = x$ 的条件下, Y 的条件分布密度, 记为 $f_{Y|X}(y \mid x)$.

同理, 若 $f_Y(y) > 0$, 则称

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$$

为在给定 $Y=y$ 的条件下, X 的条件分布密度. 此时, 给定 $Y=y$ 的条件下, X 的条件期望为

$$E(X \mid Y=y)=\int_{-\infty}^{+\infty} x f_{X|Y}(x \mid y) dx \text{ (要求绝对收敛)}$$

Example 2.11.10:::> 设二维随机向量 (X,Y) 的联合概率密度函数为

$$f(x,y)=\frac{1}{x} \quad 0 \leqslant y \leqslant x \leqslant 1$$

则 X 的边缘概率密度函数为

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y) dy=\int_{-\infty}^{\infty} \frac{1}{x} I_{[0,x]}(y) dy=\int_0^x \frac{1}{x} dy=1 \quad (0 \leqslant x \leqslant 1)$$

而在 $x \notin [0,1]$ 时, 则有 $f_X(x)=0$. 故当 $0 \leqslant x \leqslant 1$ 时有

$$f_{Y|X}(y \mid x)=\frac{f(x,y)}{f_X(x)}=\frac{1}{x} \quad (0 \leqslant y \leqslant x)$$

即 $X \sim U[0,1]$. 且在 $X=x(0<x \leqslant 1)$ 的条件下, $Y \sim U[0,x]$. 从而

$$E(Y \mid X=x)=\frac{x}{2}, \quad 0<x \leqslant 1$$

Theorem 2.11.11 (条件期望的线性性质)**:::>**

$$E([aY+bZ+c] \mid X)=aE(Y \mid X)+bE(Z \mid X)+c$$

Example 2.11.12:::> 假设甲、乙两种电器产品的使用寿命 X 与 Y 分别服从参数为 λ, μ 的指数分布, 且假定它们的寿命分布是相互独立的, 试问产品甲的寿命比产品乙的短的概率为多大?

$$\begin{aligned}
P(X<Y) &= \int_{-\infty}^{+\infty} P(X<Y \mid Y=y) f_Y(y) \mathrm{d}y \\
&= \int_0^{+\infty} P(X<Y \mid Y=y) \mu \mathrm{e}^{-\mu y} \, \mathrm{d}y \\
&= \int_0^{+\infty} P(X<y) \mu \mathrm{e}^{\mu y} \, \mathrm{d}y \\
&= \int_0^{+\infty} \left(1-\mathrm{e}^{-\lambda y}\right) \mu \mathrm{e}^{\mu y} \, \mathrm{d}y \\
&= \frac{\lambda}{\lambda+\mu}.
\end{aligned}$$

2.12. 随机变量的函数的分布

Definition 2.12.1 (一维情形下函数分布)**:::>** 设 X 为具有概率密度函数 $f(x)$ 的连续型随机变量, 为计算 X 的函数 $Y=g(X)$ 的概率密度函数, 一般地, 我们先考虑 $Y=g(X)$ 的分布函数, 即

$$\begin{aligned}P(Y\leqslant y)&=P(g(X)\leqslant y)=P(g(X)\in(-\infty,y])\\&=P\left(X\in g^{-1}(-\infty,y]\right)=\int_{g^{-1}(-\infty,y]}f(x)dx\end{aligned}$$

然后, 对上式关于 y 求导, 便可得 Y 的密度函数。这里, $g^{-1}(-\infty,y]$ 是指集合 $(-\infty,y]$ 关于 g 的原象集.

若求密度函数, 可直接利用变限积分函数求导公式:

如果函数 $f(x)$ 连续, $\varphi(x)$ 和 $\varphi(x)$ 可导, 那么变限积分函数的求导 Q 公式可表示为

$$\Phi'(x)=\frac{d}{dx}\int_{\varphi(x)}^{\varphi(x)}f(t)dt=f[\varphi(x)]\varphi'(x)-f[\varphi(x)]\varphi'(x)$$

类似地, 设 $a(x),b(x)$ 可微, $g(x,y)$ 对 x 有连续的偏导数, 则

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}x}\left[\int_{a(x)}^{b(x)}g(x,y)\mathrm{d}y\right]\\&=b'(x)g(x,b(x))-a'(x)g(x,a(x))+\int_{a(x)}^{b(x)}\frac{\partial g(x,y)}{\partial x}\,\mathrm{d}y\end{aligned}$$

若 $Y=g(X)$, 且 $g(X)$ 分段单调, 则

$$f_Y(y)=\sum_i f_X(g_i^{-1}(y))\left|\frac{d}{dy}g_i^{-1}(y)\right|$$

其中 i 指的是第 i 支反函数, 将结果加和是因为不同的反函数在概率上是"或" 的关系.

通过以上公式, 我们直接可以得到:

$$f_{aX+b}=f_X(\frac{y-b}{a})\frac{1}{|a|}$$

Example 2.12.2:::> 设 $X\sim N(0,1)$, 且 $g(x)=x^2$, 则 $Y=g(X)=X^2$ 的分布函数为

$$\begin{aligned}P(Y\leqslant y)&=P\left(X^2\leqslant y\right)=P(-\sqrt{y}\leqslant X\leqslant\sqrt{y})\\&=\Phi(\sqrt{y})-\Phi(-\sqrt{y})=2\Phi(\sqrt{y})-1\quad 若y\geqslant 0\end{aligned}$$

这里, 我们用到了 $\Phi(x)=1-\Phi(-x)$ 这一事实, 再对上式关于 y 求导便可得 Y 的密度函数

$$f_Y(y)=2\frac{d}{dy}\Phi(\sqrt{y})=y^{-\frac{1}{2}}\Phi'(\sqrt{y})=\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}\quad(y\geqslant 0)$$

这是 $[0,\infty)$ 上的分布密度, 也可以写成在实数轴上的分布密度

$$f_Y(y)=\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}\cdot I_{[0,\infty)}(y)$$

我们发现 Y 的分布服从 $\Gamma(\frac{1}{2},\frac{1}{2})$, 也即服从 $\chi^2(1)$ 分布.

也即, 若 $X_i\sim N(0,1)$, 则

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

其中我们利用了 Gamma 分布的可加性.Gamma distribution

自由度为 n 的 χ^2 分布是 n 个独立的标准二项分布的和.

Example 2.12.3:::> 若 $X\sim E(\frac{1}{5})$, $Y=min(X,2)$, 求 Y 服从的分布.

$$F_Y(y)=P(Y\leqslant y)=P(min(X,2)<y)=\begin{cases}0, & y<0\\ P(X\leqslant y)=1-e^{-\frac{1}{5}}, & 0\leqslant y<2\\ 1, & y\geqslant 2\end{cases}$$

其中尤其要注意第二部分是不能取到 2 的.

则可得:

$$P(Y=a)=\begin{cases}0 & a\neq 2\\ e^{-\frac{1}{5}} & a=2\end{cases}$$

我们发现这是一个离散型随机变量. 连续型随机变量的函数, 不一定是连续型的.

注意这个是没有密度函数的, 因为它是离散型随机变量, 并不是连续型的.

Definition 2.12.4 (二维情形下单个函数分布)**:::>** 设 (X,Y) 为具有概率密度函数 $f(x,y)$ 的连续型随机变量, 为计算函数 $Z=g(X,Y)$ 的概率密度函数, 一般地, 我们也先考虑 $Z=g(X,Y)$ 的分布函数, 即

$$P(Z\leqslant z)=P(g(X,Y)\leqslant z)=\iint_{g(x,y)\leqslant z} f(x,y)dxdy$$

然后, 对上式关于 z 求导, 便可得 Z 的密度函数.

实质上, 整个过程也即是改变了积分区域, 进行积分时将新的限制加在 D_f 上进行两重积分即可.

Definition 2.12.5 (二维情形下两个函数分布)**:::>** 若随机变量 X_1 和 X_2 的联合概率密度函数 $f(x_1,x_2)$. 那么, 随机变量 $Y_1=y_1(X_1,X_2)$ 与 $Y_2=y_2(X_1,X_2)$ 的联合概率密度函数又是什么呢?

在积分中有变量替换公式. 设 $y_1=y_1(x_1,x_2), y_2=y_2(x_1,x_2)$ 是如下的一个双射 (一一映射) $T:(x_1,x_2)\to(y_1,y_2)$, 其定义域为 $D\subseteq\boldsymbol{R}^2$, 值域为 $R\subseteq\boldsymbol{R}^2$, 对此我们有如下的引理: 设 $g:\boldsymbol{R}^2\to\boldsymbol{R}$ 且光滑映射 T 将集合 $A\subseteq D$ 映射到集合 $B\subseteq R$, 则

$$\iint_A g(x_1,x_2)dx_1dx_2=\iint_B g(x_1(y_1,y_2),x_2(y_1,y_2))|J(y_1,y_2)|dy_1dy_2$$

其中 $J(y_1,y_2)$ 为 Jacob 行列式:

$$J(y_1,y_2)=\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}=\frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2}-\frac{\partial x_1}{\partial y_2}\frac{\partial x_2}{\partial y_1}$$

从而, 若随机变量 X_1 和 X_2 的联合概率密度函数为 $f(x_1,x_2)$, 那么, 有双射 $(Y_1,Y_2)=T(X_1,X_2)$ (记 T 的定义域为 D, 值域为 R) 给出的随机变量 Y_1 和 Y_2 的联合概率密度函数为

$$f_{Y_1,Y_2}(y_1,y_2)=f(x_1(y_1,y_2),x_2(y_1,y_2))|J(y_1,y_2)|I_R(y_1,y_2)$$

其中 $(x_1(y_1,y_2),x_2(y_1,y_2))$ 是 T 的逆映射, $I_R(y_1,y_2)$ 是平面集合 R 的示性函数.

Example 2.12.6:::> (X,Y) 的联合密度 $f(x,y)$, 求 $Z=aX+bY$ 的概率密度, 这里 $a,b>0$

1. 原理法: 通过 $P(Z\leqslant z)=P(aX+bY)\leqslant z)=\iint_{ax+by\leqslant z}f(x,y)dxdy$
2. 变换公式法: 做变换

$$\begin{cases} Z=aX+bY \\ W=Y(\text{ 辅助变换}) \end{cases}\Rightarrow\begin{cases} X=\frac{z-bW}{a} \\ Y=W \end{cases}\Rightarrow\frac{\partial(x,y)}{\partial(z,w)}=\begin{vmatrix} \frac{1}{a} & 0 \\ -\frac{b}{a} & 1 \end{vmatrix}=\frac{1}{a}$$

则有:

$$f_{Z,W}(z,w)=\frac{1}{a}f\left(\frac{z-bw}{a},w\right)$$

最后得:

$$\begin{aligned} f_Z(z)&=\int_{-\infty}^{\infty}f_{Z,W}(z,w)dw \\ &=\int_{-\infty}^{\infty}\frac{1}{a}f\left(\frac{z-by}{a},y\right)dy \end{aligned}$$

Theorem 2.12.7 (和差积商公式)**:::>** 设 (X,Y) 的联合密度 $f(x,y)$, 则

$$f_{aX+bY}(z)=\int_{-\infty}^{\infty}\frac{1}{|b|}f\left(x,\frac{z-ax}{b}\right)dx=\int_{-\infty}^{\infty}\frac{1}{|a|}f\left(\frac{z-by}{a},y\right)dy\quad(a,b\text{ 不同时为}0);$$

$$f_{XY}(z)=\int_{-\infty}^{\infty}\frac{1}{|x|}f\left(x,\frac{z}{x}\right)dx$$

$$f_{\frac{X}{Y}}(z)=\int_{-\infty}^{\infty}|y|f(zy,y)dx$$

特别地有, 卷积公式:

若随机变量 X 与 Y 相互独立, 则

$$f_{X+Y}(z)=\int_{-\infty}^{+\infty}f_X(x)f_Y(z-x)dx=\int_{-\infty}^{+\infty}f_X(z-y)f_Y(y)dy.$$

我们称 f_{X+Y} 为 f_X 和 f_Y 的卷积, 并记为

$$f_{X+Y}=f_X*f_Y=f_Y*f_X$$

Example 2.12.8:::> Let $X \sim \text{Gamma}(\alpha,1)$ and $Y \sim \text{Gamma}(\beta,1)$ and X,Y are independent Then

$$\frac{X}{X+Y} \sim \text{Beta}(\alpha,\beta)$$

To prove this, write the joint pdf

$$f_{X,Y}(x,y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} y^{\beta-1} e^{-(x+y)}$$

Make the transformation $U = \frac{X}{X+Y}$ and $V = X+Y$.

The Jacobian of the transformation $X = VU, Y = V(1-U)$ is equal to V so the joint distribution of U and V has pdf

$$\frac{v}{\Gamma(\alpha)\Gamma(\beta)}(vu)^{\alpha-1}(v(1-u))^{\beta-1}e^{-v} = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha+\beta-1} e^{-v} u^{\alpha-1}(1-u)^{\beta-1}$$

(on $\mathbb{R}_+ \times [0,1]$) and hence U and V are independent (because the pdf factors over u and v) with $V \sim \text{Gamma}(\alpha+\beta,1)$ and $U \sim \text{Beta}(\alpha,\beta)$ which is apparent from the terms $v^{\alpha+\beta-1}e^{-v}$ and $u^{\alpha-1}(1-u)^{\beta-1}$ respectively.

Example 2.12.9 (最大值与最小值的分布)**:::>** 即求 $M = \max_{1\leqslant i\leqslant n} X_i, N = \min_{1\leqslant i\leqslant n} X_i$ 的分布. 这里我们采用原理法:

$$F_N(u) = P(N \leqslant u) = P(\min_{1\leqslant i\leqslant n} X_i \leqslant u)$$

也即最小的一个要比 u 小, 故也等于

$$1 - P(\min_{1\leqslant i\leqslant n} X_i > u) = 1 - P(X_1 > u, X_2 > u, \cdots, X_n > u)$$

如果 X_i 不独立, 则 n 重积分; 若独立, 则拆成连乘形式. 若 $X_i$iid, 则上式则为

$$1 - (1-F(u))^n$$

再求导即可.

而 $F_M(u)$ 类似可得.

Example 2.12.10 (连续离散组合的分布)**:::>** 设随机变量 X 和 Y 独立, 其中 X 的概率分布为

$$X \sim \begin{pmatrix} 1 & 2 \\ 0.3 & 0.7 \end{pmatrix}$$

而 Y 的概率密度为 $f(y)$, 求随机变量 $U=X+Y$ 的概率密度 $g(u)$

$$\begin{aligned}F_U(u)&=P(U\leqslant u)=P(X+Y\leqslant u)\\&=P(X+Y\leqslant u\mid X=1)P(X=1)+P(X+Y\leqslant u\mid X=2)P(X=2)\\&=0.3P(Y\leqslant u-1\mid X=1)+0.7P(Y\leqslant u-2\mid X=2)\end{aligned}$$

由于 X 和 Y 独立, 可见

$$F_U(u)=0.3P(Y\leqslant u-1)+0.7P(Y\leqslant u-2)=0.3F(u-1)+0.7F(u-2)$$

由此, 得 U 的概率密度为

$$g(u)=F_U'(u)=0.3F'(u-1)+0.7F'(u-2)=0.3f(u-1)+0.7f(u-2).$$

即: 离散和连续的组合, 只需把离散的固定, 用全概率公式即可.

Example 2.12.11:::> 以上对于二维随机变量的讨论对于条件期望之类问题具有启示意义.

如, 求条件期望:

$$E(X|X+3Y=t)$$

则我们可以令

$$U=X,V=X+3Y$$

之后问题转化为:

$$E(U|V=t)$$

2.13. Characteristic function

2.13.1 随机变量的特征函数

Definition 2.13.1 (随机变量的特征函数)**:::>** 对于一个任意的随机变量 X, 称参数 θ 的函数

$$\varphi(\theta)=Ee^{i\theta X}=E(\mathrm{Cos}\,\theta X)+iE(\mathrm{Sin}\,\theta X)\quad(-\infty<\theta<\infty)$$

为 X 的特征函数, 其中 $i=\sqrt{-1}$.

1. 若 X 为离散型随机变量, 其分布为 $P(X=x_i)=p_i$, 则 X 的特征函数为

$$\varphi(\theta)=Ee^{i\theta x}=\sum_k e^{i\theta k_k}p_k$$

2. 若 X 为连续型随机变量, 其分布密度为 $f(x)$, 则 X 的特征函数为

$$\varphi(\theta)=Ee^{i\theta X}=\int_{-\infty}^{\infty}e^{i\theta t}f(x)dx$$

这正是傅里叶变换!

Example 2.13.2:::> 1. 0-1 分布的特征函数为 $\varphi(\theta)=q+pe^{i\theta}$

2. 若 $X\sim B(n,p)$, 则 $\varphi(\theta)=(q+pe^{i\theta})^n$

3. Poisson 分布 $P(\lambda)$ 的特征函数为

$$\varphi(\theta)=\sum_{k=0}^{\infty}e^{i\varepsilon k}\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}e^{\lambda e^{i\theta}}=e^{\lambda\left(e^{i\theta}-1\right)}$$

4. 正态分布 $N(\mu,\sigma^2)$ 的特征函数为

$$\varphi(\theta)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}e^{i\theta x-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}dx=e^{i\mu\theta-\frac{1}{2}\sigma^2\theta^2}$$

这个公式跟正态分布的 PDF 一样重要.

5. 若 $X\sim U[-1,1]$, 则

$$\varphi(\theta)=\int\frac{1}{2}e^{i\theta x}dx=\frac{sin\theta}{\theta}$$

在计算特征函数时, 我们有偷懒的方法, 即把矩母函数中 u 的地方用 $i\theta$ 代替.

Theorem 2.13.3 (特征函数的分析性质)**:::>** 设连续随机变量 X 的特征函数为 $\varphi_X(\theta)$, 则

1. $|\varphi_X(\theta)|\leqslant 1,\quad \varphi_X(0)=1$ 这是因为 $|\varphi(\theta)|=\left|Ee^{i\theta X}\right|\leqslant E\left|e^{i\theta x}\right|=1$

2. $\varphi_X(\theta)$ 为 R 上的一致连续函数

3. $\varphi_X(\theta)$ 的复共轭 $\overline{\varphi_X(\theta)} = \varphi_X(-\theta)$
4. $\varphi_X(\theta)$ 是非负定函数, 也即, 第 i 行第 j 列的元素为 $\varphi_X(\theta_i - \theta_j)$ 的矩阵为非负定矩阵.

Theorem 2.13.4 (特征函数的概率性质)**:::>**

1. 设 a, b 为常数, 则 $\varphi_{a+bX}(\theta) = e^{i\theta a}\varphi_X(b\theta)$. 从而常数 a 作为随机变量的特征函数为 $e^{i\theta a}$
2. 独立和: 设随机变量 X, Y 相互独立, 则 $\varphi_{X+Y}(\theta) = \varphi_X(\theta)\varphi_Y(\theta)$
3. 矩性质若 $E\left(\left|X^k\right|\right) < \infty$, 则在 θ 很小时有

$$\varphi(\theta) = \sum_{j=0}^{k} \frac{(i\theta)^j}{j!} EX^j + o\left(\theta^k\right), \quad (\theta \to 0)$$

   且对 $j = 0, 1, 2, \cdots, k$, 有 $EX^j = i^{-j}\varphi^{(j)}(0) \quad (j = 0, 1, 2, \cdots, k)$ 成立.
4. 唯一性定理: 随机变量的分布函数 $F(x)$ 由其特征函数 $\varphi(\theta)$ 唯一决定, 反之亦然.
5. 连续性定理: 设随机变量列 X_n 及随机变量 X 的分布函数分别为 $F_{X_n}(x), F_X(x)$, 特征函数分别为 $\varphi_{X_n}(\theta), \varphi_X(\theta)$. 那么以下述两种说法彼此等价:
   (a) 在函数 $F_X(x)$ 的一切连续点 x 上有 $F_{X_n}(x) \to F_X(x)$, 称为随机变量序列 X_n 依分布收敛到分布函数 F_X, 记为 $X_n \xrightarrow{D} F_X$. 也称为随机变量序列 X_n 依分布收敛到随机变量 X, 记为 $X_n \xrightarrow{D} X$
   (b) 对于任意 θ 有 $\varphi_{X_n}(\theta) \to \varphi_X(\theta)$

   连续性定理的重要性在于它强调了, 特征函数点点收敛只保证了分布函数的连续点上点点收敛, 不连续点就不保证了.

Example 2.13.5:::> 证明: 标准化后的 Poisson 分布, 在 $\lambda \to \infty$ 时, 是依分布收敛到标准正态分布的. 即

$$\frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow{D} N(0, 1)$$

证明需通过特征函数证明.

$$\varphi_{\frac{X-\lambda}{\sqrt{\lambda}}}(\theta)=e^{-i\sqrt{\lambda}\theta}\varphi_X(\frac{\theta}{\sqrt{\lambda}})$$

代入 Poisson 分布的特征函数, 将式子 Taylor 展开后, 发现在 $\lambda\to\infty$ 时, 得到 $e^{-\frac{\theta^2}{2}}$

$$\begin{aligned}
&\varphi_{\frac{X_\lambda-\lambda}{\sqrt{\lambda}}}(\theta)=e^{-i\sqrt{\lambda}\cdot\theta}\varphi_{X_\lambda}\left(\frac{\theta}{\sqrt{\lambda}}\right)\\
&=e^{-i\sqrt{\lambda}\cdot\theta}\exp\left(\lambda\left(e^{i\frac{\theta}{\sqrt{\lambda}}}-1\right)\right)\\
&=e^{-i\sqrt{\lambda}\cdot\theta}\exp\left(\lambda\left(i\frac{\theta}{\sqrt{\lambda}}-\frac{1}{2}\frac{\theta^2}{\lambda}+o\left(\frac{1}{\lambda}\right)\right)\right)\quad(\lambda\to\infty)\\
&=\exp\left(-\frac{\theta^2}{2}+o(1)\right)\to e^{-\frac{\theta^2}{2}}.
\end{aligned}$$

即: 当 $\lambda\to\infty$ 时, 其极限是 $N(0,1)$ 的特征函数. 故由唯一性定理和连续性定理可知

$$P\left(\frac{X_\lambda-\lambda}{\sqrt{\lambda}}\leqslant x\right)\to\Phi(x)\quad(\lambda\to\infty)$$

其中 $\Phi(x)$ 是标准正态随机变量的分布函数

Example 2.13.6:::> 设随机变量 $X_1,X_2,\cdots,X_n$ 相互独立, 分别服从参数为 $\lambda_k(1\leqslant k\leqslant n)$ 的 Poisson 分布, 求 $\mathrm{X}=\sum_{k=1}^n X_k$ 的分布

此处我们用特征函数的方法证明. 由于对于每一个 k,X_k 服从参数为 λ_k 的 Poisson 分布, 故其特征函数为

$$\varphi_{x_k}(\theta)=e^{\lambda_k\left(e^{i\theta-1}\right)}$$

则

$$\varphi_X(\theta)=\varphi_{\sum_{k=1}^n X_k}(\theta)=\prod_{k=1}^n\varphi_{X_k}(\theta)=\exp\left\{\left(\sum_{k=1}^n\lambda_k\right)\left(e^{i\theta}-1\right)\right\}$$

而 $\varphi_X(\theta)$ 显然是参数为 $\sum_{k=1}^n\lambda_k$ 的 Poisson 分布的特征函数, 故由唯一性定理即知 X 服从参数为 $\sum_{k=1}^n\lambda_k$ 的 Poisson 分布.

Example 2.13.7:::> 设 $X_i\sim E(\lambda), N\sim Ge(p)$, 两者独立, 求 $\sum_{i=1}^N X_i$ 符合的分布.

我们用特征函数解决:

$$\varphi_{\sum_{i=1}^{N} X_i}(\theta) = E e^{i\theta \sum_{i=1}^{N} X_i}$$

而根据我们之前介绍过的处理原则, 对于一个复合式中较难处理的随机变量, 我们处理的思路一般是先将这个随机变量固定下来.

则

$$\varphi_{\sum_{i=1}^{N} X_i}(\theta) = \sum_{n=1}^{\infty} E(\prod_{i=1}^{n} e^{i\theta X_i}) q^{n-1} p = \frac{p}{q} \sum_{n=1}^{\infty} (\frac{\lambda}{\lambda - i\theta})^n q^n$$

最后我们发现, 这正是参量为 $p\lambda$ 的指数分布的特征函数.

2.13.2 随机向量的特征函数

Definition 2.13.8 (随机向量的特征函数)**:::>** 设 $\boldsymbol{X} = (X_1, X_2, \cdots, X_m)^T$ 是一个 m 维随机向量, $\theta = (\theta_1, \theta_2, \cdots, \theta_m)^T$ 表示 R^m 中的实向量, 则 m 元函数

$$\varphi(\theta) = \varphi(\theta_1, \theta_2, \cdots, \theta_m) = E e^{i\theta^T \boldsymbol{x}}$$

称为 $\boldsymbol{X}$ 的特征函数, 或称为 $X_1, X_2, \cdots, X_m$ 的联合特征函数.

正如随机变量的特征函数对应傅里叶变换一样, 随机向量的特征函数对应的正是分析中的多元 Fourier 变换.

Proposition 2.13.9 (随机向量特征函数的性质)**:::>**

1. 求混合矩的公式: 若 $E\left|X_1^{k_1} X_2^{k_2} \cdots X_m^{k_m}\right| < \infty$, 则

$$E\left(X_1^{k_1} X_2^{k_2} \cdots X_m^{k_m}\right) = (-i)^{k_1 + \cdots + k_m} \frac{\partial^{k_1 + \cdots + k_m}}{\partial^{k_1} x_1 \cdots \partial^{k_m} x_m} \varphi(0, \cdots, 0)$$

2. 线性变换的特征函数公式: 设 $\boldsymbol{B}$ 是一个 $l \times m$ 矩阵, $\boldsymbol{b}$ 为 l 维列向量, $\theta = (\theta_1, \cdots, \theta_l)^T$ 为 l 维列向量, 那么 $\boldsymbol{BX} + \boldsymbol{b}$ 的特征函数 (l 维) 为 $\varphi_{BX+b}(\theta) = e^{i\theta^T b} \varphi_X\left(\boldsymbol{B}^T \theta\right)$.

3. 独立性判断定理: 设 $\varphi(\theta_1,\theta_2,\cdots,\theta_m)$ 为 m 维随机向量 $(X_1,X_2,\cdots,X_m)^T$ 的特征函数, $\varphi_{X_i}(\theta_i)$ 为 X_i 的特征函数, $i=1,2,\cdots,m$. 那么 $X_1,X_2,\cdots,X_m$ 相互独立的充要条件为对一切 $\theta_1,\theta_2,\cdots,\theta_m$, 有

$$\varphi(\theta_1,\theta_2,\cdots,\theta_m)=\varphi_{X_1}(\theta_1)\varphi_{X_2}(\theta_2)\cdots\varphi_{X_m}(\theta_m)$$

4. 上条性质可推广到随机向量的独立性判断: 设 $\varphi_X(\theta_1,\theta_2,\cdots,\theta_m)$ 为 m 维随机向量 $(X_1,X_2,\cdots,X_m)^T$ 的特征函数, $\varphi_Y\left(\theta_{m+1},\theta_{m+2},\cdots,\theta_{m+n}\right)$ 为 n 维随机向量 $(Y_1,Y_2,\cdots,Y_n)^T$ 的特征函数, $\varphi\left(\theta_1,\theta_2,\cdots,\theta_{m+n}\right)$ 为 $n+m$ 维随机向量 $(X_1,X_2,\cdots,X_m,Y_1,Y_2,\cdots Y_n)^T$ 的特征函数. 那么随机向量 $(X_1,X_2,\cdots,X_m)$ 与随机向量 $(Y_1,Y_2,\cdots,Y_n)$ 相互独立的充要条件为对一切 $\theta_1,\theta_2,\cdots,\theta_{m+n}$, 有

$$\varphi\left(\theta_1,\theta_2,\cdots,\theta_{m+n}\right)=\varphi_X(\theta_1,\cdots,\theta_m)\varphi_Y\left(\theta_{m+1},\cdots,\theta_{m+n}\right)$$

2.13.3 多维 Gauss 分布及其特征函数

Definition 2.13.10 (多维 Gauss 分布)**:::>** 设 n 个相互独立的随机变量 $Z_1,Z_2,\cdots,Z_n$ 均服从标准正态分布, 如果存在常数 $a_{ij}(1\leqslant i\leqslant m,1\leqslant j\leqslant n)$ 与 $\mu_i(1\leqslant i\leqslant m)$ 使得

$$\begin{aligned}
X_1&=a_{11}Z_1+\cdots+a_{1n}Z_n+\mu_1\\
X_2&=a_{21}Z_1+\cdots+a_{2n}Z_n+\mu_2\\
&\vdots\\
X_i&=a_{i1}Z_1+\cdots+a_{in}Z_n+\mu_i\\
&\vdots\\
X_m&=a_{m1}Z_1+\cdots+a_{mn}Z_n+\mu_m
\end{aligned}$$

即

$$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{Z}+\mu\ (\text{这里}\boldsymbol{A}=\left(a_{ij}\right)_{m\times n})$$

称由 $X_1,X_2,\cdots,X_m$ 构成的 m 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_m)^T$ 服从 m 维 (或 m 元) Gauss 分布。特别地, 当矩阵 $\boldsymbol{A}$ 的秩为 m 时 (即 $m\leqslant n$, 且 $\boldsymbol{A}$ 为满秩的情形), 称 m 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_m)^T$ 服从 m 维 (或 m 元) 联合正态分布, 简称 m 维正态分布.

𝔎emark. 1. m 维 Gauss 分布的一维边缘分布为一维 Gauss 分布 (或为一维正态, 或为常数)

对于一维 Gauss 分布来说, 若行不满秩, 则系数全为 0, 此时随机变量是一个常数.

2. 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_m)^T$ 为 Gauss 随机向量, 则有期望向量

$$E\boldsymbol{X}=(EX_1,EX_2,\cdots,EX_m)^T=(\mu_1,\cdots,\mu_m)^T=\mu$$

协方差矩阵为 $\Sigma=\left(\mathrm{Cov}\left(X_i,X_j\right)\right)_{m\times m}$, 其中

$$\begin{aligned}\mathrm{Cov}\left(X_i,X_j\right)&=\mathrm{Cov}\left(\sum_{k=1}^{n}a_{ik}\mathrm{Z}_k+\mu_i,\sum_{l=1}^{n}a_{jl}\mathrm{Z}_l+\mu_j\right)\\&=\mathrm{Cov}\left(\sum_{k=1}^{n}a_{ik}\mathrm{Z}_k,\sum_{l=1}^{n}a_{jl}\mathrm{Z}_l\right)\\&=\sum_{k,l}a_{ik}a_{jl}\,\mathrm{Cov}(\mathrm{Z}_k,\mathrm{Z}_l)=\sum_{k=1}^{n}a_{ik}a_{jk}\end{aligned}$$

即 $\Sigma=\left(\mathrm{Cov}\left(X_i,X_j\right)\right)_{i,j}=\boldsymbol{A}\boldsymbol{A}^T$

记此 m 维 Gauss 分布为 $\boldsymbol{X}\sim N(\mu,\Sigma)$ (对应于一维正态).

3. Gauss 分布 $\boldsymbol{X}\sim N(\mu,\Sigma)$ 是正态分布的充要条件是矩阵 Σ 的行列式非 0, 即不退化 (即 Σ 正定时.

■

𝔇efinition **2.13.11** (m 维 Gauss 随机变量的特征函数)**:::>** m 维 Gauss 分布 $N(\mu,\Sigma)$ 的 (m 维) 特征函数为

$$\varphi(\theta)=\exp\left\{i\theta^T\mu-\frac{1}{2}\theta^T\Sigma\theta\right)$$

反之亦成立.

𝔇efinition **2.13.12** (多维 Gauss 分布的等价定义)**:::>** m 维随机向量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_m)^T$ 称为服从 Gauss 分布, 如果它的特征函数有如下形式

$$\varphi(\theta)=\varphi(\theta_1,\theta_2,\cdots,\theta_m)=\exp\left\{i\theta^T\mu-\frac{1}{2}\theta^T\Sigma\theta\right)\right\}$$

其中 μ 为 m 维列向量, Σ 为 $m\times m$ 的非负定矩阵. 特别, 当 Σ 是正定矩阵时, 称它服从 m 维正态分布.

Theorem 2.13.13:::> 多维 Gauss 随机向量经过线性变换仍然是 Gauss 随机向量, 即若 $\boldsymbol{X} \sim N(\mu, \Sigma)$, $\boldsymbol{B}$ 为 $n \times m$ 矩阵, b 为 n 维列向量, 那么

$$\boldsymbol{B}\boldsymbol{X} + \boldsymbol{b} \sim N\left(\boldsymbol{B}\mu + \boldsymbol{b}, \boldsymbol{B}\Sigma\boldsymbol{B}^T\right)$$

证明: 由于

$$\boldsymbol{X} = \boldsymbol{A}Z + \mu \text{ (这里} \boldsymbol{A} = \left(a_{ij}\right)_{m\times n} \text{)}$$

则

$$\boldsymbol{B}\boldsymbol{X} + \boldsymbol{b} = \boldsymbol{B}\boldsymbol{A}Z + \boldsymbol{B}\mu + \boldsymbol{b}$$

故方差为 $\boldsymbol{B}\mu + \boldsymbol{b}$, 而协方差为 $\boldsymbol{B}\boldsymbol{A}(\boldsymbol{B}\boldsymbol{A})^T$

Definition 2.13.14 (多维 Gauss 分布的等价定义)**:::>** m 维随机向量 $\boldsymbol{X} = (X_1, X_2, \cdots, X_m)^T$ 称为Gauss的, 如果对于任意一个 m 维向量 $\boldsymbol{a} = (a_1, a_2, \cdots, a_m)^T$, 一维随机变量 $\boldsymbol{a}^T\boldsymbol{X}$ 要么是正态的, 要么是常数.

对于一维随机向量 $a^T\boldsymbol{X}$ 来说, $a^T\boldsymbol{X} \sim N(a^T\mu, a^T\Sigma a)$. 由于 Σ 是非负定矩阵, 则二次型 $a^T\Sigma a$ 必然大于等于 0, 若非零, 则为正态; 若为零, 则退化为常数.

Definition 2.13.15 (多维正态分布的等价定义)**:::>** 正态分布具有分布密度. 即若 $\boldsymbol{X} \sim N(\mu, \Sigma)$, 其行列式 $|\Sigma| > 0$, 则对任意 $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$, $\boldsymbol{X}$ 的联合概率密度函数为

$$f(\boldsymbol{x}) = f(x_1, x_2, \cdots, x_m) = \left(\frac{1}{2\pi}\right)^{\frac{m}{2}} \frac{1}{\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \mu)^T\Sigma^{-1}(\boldsymbol{x} - \mu)\right\}$$

称为 m 维正态密度.

证明: 由于 $\boldsymbol{\Sigma}$ 非退化, 即它是正定的, 由线性代数知道必存在一个可逆矩阵 $\boldsymbol{\Lambda}$, 使 $\boldsymbol{\Sigma} = \Lambda\Lambda^T$. 作变量替换

$$\boldsymbol{y}(= \boldsymbol{y}(\boldsymbol{x})) = \Lambda^{-1}(x - \mu)$$

这个过程可对应于一维高斯分布中标准化的过程. 其 Jacobian 是 $\frac{\partial y}{\partial \boldsymbol{x}} = \boldsymbol{\Lambda}^{-1}$, 由于 $|\Sigma| = |\Lambda|^2$, 故其行列式的绝对值为 $\left|\boldsymbol{\Lambda}^{-1}\right| = \frac{1}{\sqrt{|\Sigma|}}$

而随机向量 $\boldsymbol{X}$ 的线性变换 $Y = \Lambda^{-1}(\boldsymbol{X} - \mu)$, 则

$$\boldsymbol{Y} \sim N\left(0, \Lambda^{-1}\Sigma\left(\Lambda^{-1}\right)^T\right) = N(0, I)$$

其中 I 是 $n\times n$ 的单位矩阵. 即它的分量是相互独立的标准正态随机变量, 故它有分布密度

$$f_Y(y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{y_1^2}{2}}\cdots\frac{1}{\sqrt{2\pi}}e^{-\frac{y_m^2}{2}}=\left(\frac{1}{2\pi}\right)^{\frac{m}{2}}e^{-\frac{1}{2}yy}$$

所以 X 有分布密度 (直观地: $f_Y(y)dy=f(x)dx$, 即 $f(x)=f_Y(y(x))\left|\frac{\partial y}{\partial x}\right|$)

$$f(x)=\left(\frac{1}{2\pi}\right)^{\frac{m}{2}}\frac{1}{\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right\}$$

Proposition 2.13.16:::> 由于多维高斯分布协方差矩阵中的元素是相关系数, 因此, 如果从一个 n 维高斯分布推出其中 m 个变量构成的 m 维高斯分布, 只需从 n 维高斯分布的协方差矩阵选取对应位置的相关系数构成新的协方差矩阵即可.

Theorem 2.13.17 (分量独立问题)**:::>** 设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_m)^T$ 为 m 维 Gauss 随机向量, 则 $X_1,X_2,\cdots,X_m$ 相互独立的充要条件为 $\mathrm{Cov}\left(X_i,X_j\right)=0\quad 1\leqslant i\neq j\leqslant m$, 即协方差矩阵是对角形的.

这里用到了中介绍的正态分布独立不相关是等价的的知识.

Theorem 2.13.18:::> 若 $\left(X_1,\cdots,X_{m+n}\right)^T$ 服从 Gauss 分布, 则 $(X_1,\cdots,X_n)^T$ 与 $\left(X_{n+1},\cdots,X_{n+m}\right)^T$ 相互独立的充要条件是

$$\sigma_{ij}=0\quad(i\leqslant n,j>n)$$

其中

$$\Sigma=(\sigma_{kl})\quad(k,l=1,\cdots,n+m)$$

是 $\left(X_1,\cdots,X_{m+n}\right)^T$ 的协方差矩阵. 而以上条件的含义是说, 协方差矩阵是准对角型的

Example 2.13.19:::> 设 $X\sim N(\mu,\Sigma)$, 则 AX 与 BX 独立当且仅当 $A\Sigma B^T=0$. (A 为 $(m-p)\times m, B$ 为 $p\times m$)

3. Stochastic process

3.1. Stochastic process

Definition 3.1.1 (Stochastic process)**:::>** 设 $(\Omega, \mathscr{F}, P)$ 是一个概率空间, 一族依赖于参数 $t(t \in T)$ 的随机变量 $\boldsymbol{X} = \{X_t : t \in T\}$ 称为一个随机过程. 其中 T 称为指标集, 对 T 中的每个 t, X_t 是一个随机变量 $X_t(\omega)$. 对每个固定的基本事件 $\omega, \{X_t(\omega) : t \in T\}$ 是一个定义在 T 上的实值函数, 称之为随机过程 $\boldsymbol{X}$ 的一条 (样本) 轨道.

一般来说, 指标集 T 是一维的, 因此我们可以将它理解为时间.

Proposition 3.1.2 (随机过程的分类)**:::>** 按过程的指标集和状态空间分类, 有:

1. 离散时间, 离散状态: 链. 若链具有马尔可夫性, 则就是马尔可夫链. 以及随机徘徊等.
2. 离散时间, 连续状态
3. 连续时间, 离散状态: 如 Poisson 过程
4. 连续时间, 连续状态: 如 Brown 运动.

Definition 3.1.3 (随机过程的有限维分布族)**:::>** 对任一随机过程 $\{X_t : t \in T\}$, 定义:

1. $F_{t_l}(x_1) \equiv P\left(X_{t_l} \leqslant x_1\right)$

2. $F_{t_1,t_2}(x_1,x_2)\equiv P\left(X_{t_1}\leqslant x_1,X_{t_2}\leqslant x_2\right)$
3. ……
4. $F_{t_1,\cdots,t_k}(x_1,\cdots,x_k)\equiv P\left(X_{t_1}\leqslant x_1,\cdots,X_{t_k}\leqslant x_k\right)$

为过程的有限维分布族, 它满足:

1. 对称性: 若 $\left\{i_1,\cdots,i_j\right\}$ 为 $\{1,\cdots,j\}$ 的一个排列, 则

$$F_{t_{i_1},\cdots,t_{i_j}}\left(x_{i_l},\cdots,x_{i_j}\right)=F_{t_l,\cdots,t_j}\left(x_1,\cdots,x_j\right)$$

2. 相容性: 对任意 $i<j$, 有

$$F_{t_1,\cdots,t_i,t_{i+1},\cdots,t_j}(x_1,\cdots,x_i,\infty,\cdots,\infty)=F_{t_1,\cdots,t_i}(x_1,\cdots,x_i)$$

有限维分布族是随机过程中最重要的一个量, 名字中带" 族" 是因为这是随着维数变化的.

Theorem 3.1.4 (Kolmogorov 随机过程存在性定理)**:::>** 设 $\left\{F_{t_1,\cdots,t_j}\left(x_1,\cdots,x_j\right)\right\}$ 是满足对称性和相容性条件的一函数族, 则存在一概率空间 (Ω,F,P) 和定义于其上的随机过程 $\{X_t:t\in T\}$, 使 X_t 的有限维分布族为

$$\left\{F_{t_1,\cdots,t_j}\left(x_1,\cdots,x_j\right)\right\}$$

.

Definition 3.1.5 (独立增量过程)**:::>** 设 $\boldsymbol{X}=\{X_t\}$ 是一个随机过程, 如果它在任意 s 个互不相交的区间上的增量 $X_{m_1}-X_{n_1},X_{m_2}-X_{n_2},\ldots,X_{m_s}-X_{n_s}$ 都相互独立, 则称随机过程 $\boldsymbol{X}$ 为一个独立增量过程. 又如果对任意的 $n>0$, 都有 $X_{m+n}-X_m(n>0)$ 对一切 m 同分布, 则称 $\boldsymbol{X}$ 为一个时齐的独立增量过程.

更加直观地理解, 独立增量过程就是增量相互独立的随机过程. 而时齐的独立增量过程, 增量只跟时间段的长度 (区间长度) 有关.

Theorem 3.1.6:::> 独立增量过程的有限维分布由其增量的分布和过程的初始分布唯一决定; 时齐的独立增量过程的有限维分布由其增量的 (一维) 分布和初始分布唯一决定.

Proof. 为方便起见, 不妨考虑离散状态的随机过程, 且设 $X_0 = x_0$, $t_1 < t_2 < \cdots < t_k$, 过程的有限维分布为:

$$P\left(X_{t_l} = x_1, \cdots, X_{t_k} = x_k\right) = P\left(X_0 = x_0, X_{t_1} - X_0 = x_1 - x_0, \cdots, X_{t_k} - X_{t_{k-1}} = x_k - x_{k-1}\right)$$
$$= P(X_0 = x_0) P\left(X_{t_l} - X_0 = x_1 - x_0\right) \cdots P\left(X_{t_k} - X_{t_{k-1}} = x_k - x_{k-1}\right)$$

若为时齐的独立增量过程, 则:

$$P\left(X_{t_k} - X_{t_{k-1}} = x_k - x_{k-1}\right) = P\left(X_{t_k - t_{k-1}} - X_0 = x_k - x_{k-1}\right) = P\left(X_{t_k - t_{k-1}} = x_k - x_{k-1} + x_0\right)$$

■

Definition 3.1.7 (Markov 过程)::> 马尔可夫性质是概率论中的一个概念. 当一个随机过程在给定现在状态及所有过去状态情况下, 其未来状态的条件概率分布仅依赖于当前状态; 换句话说, 在给定现在状态时, 它与过去状态（即该过程的历史路径）是条件独立的, 那么此随机过程即具有马尔可夫性质. 具有马尔可夫性质的过程通常称之为马尔可夫过程. 即" 已知现在, 将来与过去独立", 工程上又叫做无后效性.

数学上, 如果 $X(t), t > 0$ 是一个随机过程, 则马尔可夫性质就是指

$$\Pr[X(t+h) = y \mid X(s) = x(s), s \leqslant t] = \Pr[X(t+h) = y \mid X(t) = x(t)], \quad \forall h > 0$$

马尔可夫过程被称为是 (时间) 齐次的, 如果它满足

$$\Pr[X(t+h) = y \mid X(t) = x] = \Pr[X(h) = y \mid X(0) = x], \quad \forall t, h > 0$$

Theorem 3.1.8::> 独立增量过程是一个马尔可夫过程.

Proof.

$$\begin{aligned}
&P\left(X_t = x_t \mid X_{t_n} = x_n, X_{t_{n-1}} = x_{n-1}, \cdots X_{t_1} = x_1\right) \\
&= \frac{P\left(X_t = x_t, X_{t_n} = x_n, X_{t_{n-1}} = x_{n-1}, \cdots X_{t_1} = x_1\right)}{P\left(X_{t_n} = x_n, X_{t_{n-1}} = x_{n-1}, \cdots X_{t_1} = x_1\right)} \\
&= \frac{\prod_{i=1}^{n} P(X_{t_i} - X_{t_{i-1}} = x_i - x_{i-1}) \times P(X_t - X_{t_n} = x_t - x_{t_n})}{\prod_{i=1}^{n} P(X_{t_i} - X_{t_{i-1}} = x_i - x_{i-1})} \\
&= P(X_t - X_{t_n} = x_t - x_{t_n})
\end{aligned}$$

■

3.2. Random Walk

Definition 3.2.1 (Random Walk)**:::>** 设 (Ω, F, P) 是一个概率空间, 其上的独立同分布的随机变量序列 $\{Z_n\}$ 满足

$$Z_n \sim \begin{pmatrix} -1 & 1 \\ q & p \end{pmatrix} \quad q \equiv 1-p \quad (n=1,2,\ldots)$$

令

$$X_n = X_0 + \sum_{i=1}^{n} Z_i$$

则我们称随机变量序列 $\boldsymbol{X} = \{X_n, n \geqslant 1\}$ 为 (1-维) 简单随机徘徊, 其中 X_0 是任意一个取整数值的随机变量, 通常取 X_0 为 x. 特别地, 当 $p = \frac{1}{2}$ 时, 称之为 (1-维) 对称的简单随机徘徊.

Theorem 3.2.2:::> 简单随机徘徊是时齐的独立增量过程, 也即为一个 Markov 过程, 其一维分布为

$$P(X_n = k) = \begin{cases} C_n^{\frac{n+k-x}{2}} p^{\frac{n+k-x}{2}} (1-p)^{\frac{n-k+x}{2}}, & (n \geqslant |k-x|, n \text{ 与} k-x \text{ 同奇偶}) \\ 0 & (\text{其他情形}) \end{cases}$$

(直观地理解, 这个概率就是走了 n 步处于 k 位置的概率, 我们将起点设为 0)

Proof. 由于 $Z_n \sim \begin{pmatrix} -1 & 1 \\ q & p \end{pmatrix}$, 将其化为标准 0-1 分布, 令 $Y_n = \frac{Z_n+1}{2}$, 则 $Y_n \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$, 从而

$$X_n = X_0 + \sum_{i=1}^{n} (2Y_i - 1) = x + 2\sum_{i=1}^{n} Y_i - n$$

由于 $\sum_{i=1}^{n} Y_i \sim B(n,p)$, 故

$$P(X_n = k) = P\left(x + 2\sum_{i=1}^{n} Y_i - n = k\right) = P\left(\sum_{i=1}^{n} Y_i = \frac{k-x+n}{2}\right)$$

故

$$P(X_n = k) = \begin{cases} C_n^{\frac{n+k-x}{2}} p^{\frac{n+k-x}{2}} (1-p)^{\frac{n-k+x}{2}}, & (n \geqslant |k-x|, n \text{ 与} k-x \text{ 同奇偶}) \\ 0 & , \quad (\text{其他情形}) \end{cases}$$

对于有限维分布, 当 $X_0=0$ 时为

$$P\left(X_{n_1}=s_1,X_{n_2}=s_2,\ldots,X_{n_k}=s_k\right)=P\left(\sum_{i=1}^{n_1}Z_i=s_1,\sum_{i=n_1+1}^{n_2}Z_i=s_2-s_1,\ldots,\sum_{i=n_{k-1}+1}^{n_k}Z_i=s_k-s_{k-1}\right)$$

$$=\prod_{l=1}^{k}C_{m_l}^{\eta}p^{\eta}(1-p)^{m_l-\eta}\quad\left(m_l=n_l-n_{l-1},r_l=\frac{1}{2}\left(n_l-n_{l-1}+s_l-s_{l-1}\right)\right)$$

■

Example 3.2.3:::> 设 $\{X_n,n\geqslant 0\}$ 为一简单随机徘徊. 证明对任意正整数 n 与任意整数 $i_0,i_1,\cdots,i_{n+1}$ 有:

$$P\left(X_{n+1}=i_{n+1},X_{n-1}=i_{n-1}\mid X_n=i_n\right)$$
$$=P\left(X_{n+1}=i_{n+1}\mid X_n=i\right)P\left(X_{n-1}=i_{n-1}\mid X_n=i_n\right)$$

利用独立增量过程性质即可证明:

$$P\left(X_{n+1}=i_{n+1},X_{n-1}=i_{n-1}\mid X_n=i_n\right)$$
$$=P\left(X_{n+1}-X_n=i_{n+1}-i_n,X_n-X_{n-1}=i_n-i_{n-1}\mid X_n=i_n\right)$$
$$=P\left(x_{n+1}-x_n=i_{n+1}-i_n\mid x_n=i_n\right)\cdot P\left(x_n-x_{n-1}=i_{n-in-1}\mid x_n=i_n\right)$$
$$=P\left(x_{n+1}=i_{n+1}\mid x_n=i_n\right)P\left(x_{n-1}=i_{n-1}\mid x_n=i_n\right)$$

Theorem 3.2.4:::> 设 $X_0=x$, 则有

1.
$$EX_n^x=x+E\left(\sum_{i=1}^{n}Z_i\right)=x+nEZ_i=x+n(p-q)$$

2.
$$D(X_n^x)=D\left(x+\sum_{i=1}^{n}Z_i\right)=nDZ_i=4npq$$

3.
$$\mathrm{Cov}(X_n^x,X_m^x)=4(n\wedge m)pq$$

这个量表征了随机过程两个不同时刻的关联程度. 对于一切独立增量过程, 若 $n<m$, 则

$$\mathrm{Cov}(X_n^x,X_m^x)=D(X_n)$$

Example 3.2.5:::> A gambler starts with z, with $z < a$, and plays a game in which he wins 1 or loses 1 at each turn with probabilities p and q respectively. What are

$$p_z = P(\text{random walk hits } a \text{ before } 0 \mid \text{starts at } z),$$

and

$$q_z = P(\text{random walk hits } 0 \text{ before } a \mid \text{starts at } z)?$$

He either wins his first game, with probability p, or loses with probability q. So

$$p_z = q p_{z-1} + p p_{z+1},$$

for $0 < z < a$, and $p_0 = 0, p_a = 1$.

Try $p_z = t^z$. Then

$$pt^2 - t + q = (pt - q)(t - 1) = 0,$$

noting that $p = 1 - q$. If $p \neq q$, then

$$p_z = A1^z + B\left(\frac{q}{p}\right)^z.$$

Since $p_0 = 0$, we get $A = -B$. Since $p_a = 1$, we obtain

$$p_z = \frac{1-(q/p)^z}{1-(q/p)^a}.$$

If $p = q$, then $p_z = A + Bz = z/a$.

Similarly, (or perform the substitutions $p \mapsto q$, $q \mapsto p$ and $z \mapsto a - z$)

$$q_z = \frac{(q/p)^a - (q/p)^z}{(q/p)^a - 1}$$

if $p \neq q$, and

$$q_z = \frac{a - z}{a}$$

if $p = q$. Since $p_z + q_z = 1$, we know that the game will eventually end.

What if $a \to \infty$? What is the probability of going bankrupt?

$$\begin{aligned} P(\text{path hits } 0 \text{ ever}) &= P\left(\bigcup_{a=z+1}^{\infty} [\text{path hits } 0 \text{ before } a]\right) \\ &= \lim_{a\to\infty} P(\text{path hits } 0 \text{ before } a) \\ &= \lim_{a\to\infty} q_z \\ &= \begin{cases} (q/p)^z & p > q \\ 1 & p \leqslant q. \end{cases} \end{aligned}$$

So if the odds are against you (i.e. the probability of losing is greater than the probability of winning), then no matter how small the difference is, you are bound to going bankrupt eventually.

Example 3.2.6 (Duration of the game)**:::>** Let D_z = expected time until the random walk hits 0 or a, starting from z. We first show that this is bounded. We know that there is one simple way to hit 0 or a: get $+1$ or -1 for a times in a row. This happens with probability $p^a + q^a$, and takes a steps. So even if this were the only way to hit 0 or a, the expected duration would be $\frac{a}{p^a+q^a}$. So we must have

$$D_z \leqslant \frac{a}{p^a + q^a}$$

This is a *very* crude bound, but it is sufficient to show that it is bounded, and we can meaningfully apply formulas to this finite quantity.

We have

$$\begin{aligned} D_z &= E[\text{duration}] \\ &= E[E[\text{duration} \mid X_1]] \\ &= pE[\text{duration} \mid X_1 = 1] + qE[\text{duration} \mid X_1 = -1] \\ &= p(1 + D_{z+1}) + q(1 + D_{z-1}) \end{aligned}$$

So

$$D_z = 1 + pD_{z+1} + qD_{z-1},$$

subject to $D_0 = D_a = 0$.

We first find a particular solution by trying $D_z = Cz$. Then

$$Cz = 1 + pC(z+1) + qC(z-1).$$

So

$$C = \frac{1}{q-p},$$

for $p \neq q$. Then we find the complementary solution. Try $D_z = t^z$.

$$pt^2 - t + q = 0,$$

which has roots 1 and q/p. So the general solution is

$$D_z = A + B\left(\frac{q}{p}\right)^z + \frac{z}{q-p}.$$

Putting in the boundary conditions,

$$D_z = \frac{z}{q-p} - \frac{a}{q-p} \cdot \frac{1-(q/p)^z}{1-(q/p)^a}.$$

If $p = q$, then to find the particular solution, we have to try

$$D_z = Cz^2.$$

Then we find $C = -1$. So

$$D_z = -z^2 + A + Bz.$$

Then the boundary conditions give

$$D_z = z(a-z).$$

3.3. Poisson Process

Definition 3.3.1 (泊松过程的物理背景)**:::>** 将 $(0, t]$ 时间中到来的理赔次数记为 N_t. 假定

1. 在互不相交的各时段内, 理赔的发生是相互独立的. 即满足独立增量性.

2. 在相同长度的时间段中, 理赔发生的统计规律是相同的 (如果还要考虑季节差别等因素的影响, 那么这个假定只能在相对地不太长的时段内近似成立). 即时齐性.

3. 在每个小时间段 $(t, t+h]$ 内有两次或更多次的理赔发生的概率

$$P\left(\omega : N_{t+h}(\omega) - N_t(\omega) \geqslant 2\right) = o(h)$$

而

$$P\left(\omega : N_{t+h}(\omega) - N_t(\omega) = 1\right) = \lambda h + o(h)$$

因此, 在 $(t, t+h]$ 内无理赔发生的概率 $P\left(\omega : N_{t+h}(\omega) - N_t(\omega) = 0\right) = 1 - (\lambda h + o(h))$. 这个性质又叫做普通性质.

满足这三个性质的随机过程即为泊松分布.

Definition 3.3.2:::> N_t 是一个时齐的独立增量过程, 从而要知道其有限维分布, 只需要求 $P(N_t = k)$. 记 $p_k(t) = P(\omega : N_t(\omega) = k)$.

我们有

$$\begin{aligned} p_k(t+h) = P\left(\omega : N_{t+h}(\omega) = k\right) &= P\left(\omega : N_t(\omega) = k, N_{t+h}(\omega) - N_t(\omega) = 0\right) \\ &+ P\left(\omega : N_t(\omega) = k-1, N_{t+h}(\omega) - N_t(\omega) = 1\right) \\ &+ P\left(\omega : N_t(\omega) = k-m, N_{t+h}(\omega) - N_t(\omega) = m \geqslant 2\right) \end{aligned}$$

由于它是独立增量过程, 则第一项可写成

$$p_k(t)(1 - \lambda h + o(h))$$

第二项可写成

$$p_{k-1}(t)(\lambda h + o(h))$$

第三项可写成

$$o(h)$$

从而

$$p_k(t+h) - p_k(t) = -\lambda h\, p_k(t) + \lambda h\, p_{k-1}(t) + o(h). \quad (k \geqslant 1)$$

在上式等号两边同除以 h, 并令 $h \to 0$, 就得到如下的无穷个常微分方程组

$$p_k'(t) = -\lambda p_k(t) + \lambda p_{k-1}(t) \quad (k \geqslant 1)$$

类似地, 对 $k=0$ 我们还有

$$p_0(t+h)-p_0(t)=-\lambda p_0(t)h+o(h))$$

即

$$p_0'(t)=-\lambda p_0(t)$$

由 N_t 的定义, 有 $N_0=0$, 即我们有初始条件 $p_0(0)=1, p_k(0)=0$.

解常微分方程组, 可得

$$p_0(t)=e^{-\lambda t}$$

进而:

$$p_1'(t)=-\lambda p_1(t)+\lambda e^{-\lambda t}$$

可得

$$p_1(t)=\lambda t e^{-\lambda t}$$

进一步地,

$$p_2'(t)=-\lambda p_2(t)+\lambda^2 t e^{-\lambda t}$$

可得

$$p_2(t)=\frac{(\lambda t)^2}{2!}e^{-\lambda t}$$

一般地, 用数学归纳法, 得到

$$p_k'(t)=-\lambda p_k(t)+\lambda p_{k-1}(t)=-\lambda p_k(t)+\frac{\lambda(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda t}$$

这样我们就得到了随机变量 N_t 的分布:

$$P(\omega:N_t(\omega)=k)=p_k(t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

即

$$N_t\sim\begin{pmatrix} 0 & 1 & \cdots & n & \cdots \\ e^{-\lambda t} & \lambda t e^{-\lambda t} & \cdots & \frac{(\lambda t)^n}{n!}e^{-\lambda t} & \cdots \end{pmatrix}$$

我们也可以利用另一种方法. 在工程上, 解微分方程常用傅里叶变换或拉普拉斯变换将其转换为代数方程, 解出结果后再将其逆变换回原结果的做法. 对应到概率论中, 就是我们要想办法利用其矩母函数的性质.

引入一个参数 z, 并定义

$$M(t,z)=Ee^{zN_t}=\sum_{k=0}^{\infty}e^{zk}p_k(t)$$

那么我们有:

$$\begin{aligned}\frac{\partial M(t,z)}{\partial t}&=\sum_{k=0}^{\infty}e^{zk}p_k'(t)=-\lambda\sum_{k=0}^{\infty}e^{zk}p_k(t)+\lambda\sum_{k=1}^{\infty}e^{zk}p_{k-1}(t)\\&=-\lambda M(t,z)+\lambda e^z M(t,z)=\lambda(e^z-1)M(t,z).\end{aligned}$$

这是一个 z 为参数的关于自变量 t 的常系数常微分方程, 且满足初始条件 $M(0,z)=p_0(0)=1$. 此方程的解为

$$M(t,z)=e^{\lambda t(e^z-1)}=\sum_{k=0}^{\infty}\frac{(\lambda t)^k}{k!}e^{-\lambda t}e^{zk}$$

按矩母函数的定义, 便得到表达式

$$p_k(t)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

这种矩母函数方法, 适用于只取自然数值的随机变量过程.

Definition 3.3.3 (随机变量的矩母函数)**:::>** 随机变量 X 的矩母函数定义为 $m_X(u)=E\left(e^{uX}\right)$, $u\in R$ (如果后者的期望存在的话), 矩母函数 $m_X(u)$ 具有如下性质:

1. 若 X 的矩母函数 $m_X(u)$ 在 $u=0$ 的某一开领域内存在, 则 $E(X^n)=m_X^{(n)}(0)$
2. 若随机变量 X 与 Y 独立, 则 $m_{X+Y}(u)=m_X(u)m_Y(u)$
3. 当矩母函数存在时, 它可以唯一决定随机变量的分布

Example 3.3.4 (Poisson' 分布的矩母函数)**:::>** 若 $X\sim P(\lambda)$, 则

$$\begin{aligned}m_X(u)&=Ee^{uX}=\sum_{k=0}^{\infty}e^{uk}P(X=k)=\sum_{k=0}^{\infty}e^{uk}e^{-\lambda}\frac{\lambda^k}{k!}\\&=e^{\lambda(e^u-1)}\end{aligned}$$

从而 $EX=m_X'(0)=\lambda; EX^2=m_X''(0)=\lambda+\lambda^2$, 故 $DX=\lambda$

Definition 3.3.5 (Poisson 过程的数学定义)**:::>** 一个整数值随机过程 $N=\{N_t, t\geqslant 0\}$ 满足下述三个条件, 就称它为强度为 λ 的 Poisson 过程:

1. $N_t=0$
2. N_t 是独立增量过程 (即对任意的 $0\leqslant t_1<t_2<\cdots<t_n, N_t$ 在区间 $\left(t_i, t_{i+1}\right.$] 上的增量 $N_{t_{i+1}}-N_{t_i}$ 是相互独立的, $i=1,2,\cdots n-1$)
3. 对任意 $t>0, s\geqslant 0$ 增量 $N_{s+t}-N_s\sim P(\lambda t)$, 即

$$P\left(N_{s+t}-N_s=k\right)=\frac{(\lambda t)^k e^{-\lambda t}}{k!}\quad k=0,1,2,\cdots$$

Proposition 3.3.6 (Poisson 过程的有限维分布)**:::>** 设 $\{N_t, t\geqslant 0\}$ 为是度为 $\lambda>0$ 的 Poisson 过程, 则 N_t 的有限维分布为对任意 n, 任意 $0<t_1<t_2<\cdots<t_n$ 及任意非负整数 $k_1, k_2, \cdots, k_n$, 有: 对任意 n, 任意 $0<t_1<t_2<\cdots<t_n$ 及任意非负整数 $k_1, k_2, \cdots, k_n$, 有

$$\begin{aligned}&P\left(N_{t_1}=k_1, N_{t_2}=k_2, \cdots, N_{t_n}=k_n\right)\\&=\frac{\lambda^{k_n} e^{-\lambda t_n} t_1^{k_1}\left(t_2-t_1\right)^{k_2-k_1}\cdots\left(t_n-t_{n-1}\right)^{k_n-k_{n-1}}}{k_1!(k_2-k_1)!\cdots\left(k_n-k_{n-1}\right)!}\end{aligned}$$

Proposition 3.3.7:::>

$$EN_t=DN_t=\lambda t;\quad \operatorname{Cov}(N_s, N_t)=\lambda(s\wedge t)$$

类似的性质在中已有证明.

Proposition 3.3.8 (Poisson 过程的可加性)**:::>** 独立 Poisson 过程的和仍为 Poisson 过程, 且其强度恰为各强度之和.

Theorem 3.3.9:::> $\{X_t, t\geqslant 0\}$ 与 $\{Y_t, t\geqslant 0\}$ 分别为强度为 λ_1 和 λ_2 的 Poisson 过程, 且相互独立. 则

$$P\left(Y_t=k \mid X_t+Y_t=n\right)=\begin{cases} C_n^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-k}, & k=0,1,2,\cdots,n\\ 0, & \text{otherwise}\end{cases}$$

证明过程可以直接将条件概率上下的公式写出来, 相消后即获得此结果.

从而

$$E\left(Y_t \mid X_t+Y_t\right)=\left(X_t+Y_t\right)\frac{\lambda_2}{\lambda_1+\lambda_2}$$

.

$$D(Y_t \mid X_t + Y_t) = (X_t + Y_t)\frac{\lambda_2}{\lambda_1+\lambda_2}\frac{\lambda_1}{\lambda_1+\lambda_2}$$

Theorem 3.3.10:::> 设 $\{N_t, t \geqslant 0\}$ 为强度为 $\lambda > 0$ 的 Poisson 过程, 则对 $0 < u < t, 0 \leqslant k \leqslant n$, 有

$$P(N_u = k \mid N_t = n) = C_n^k \left(\frac{u}{t}\right)^k \left(1 - \frac{u}{t}\right)^{n-k}$$

即在条件概率 $P(* \mid N_t = n)$ 下, $N_u (u < t)$ 的条件分布为二项分布 $B\left(n, \frac{u}{t}\right)$.

Proof. 由 Poisson 过程的独立增量性知

$$\begin{aligned}
P(N_u = k \mid N_t = n) &= \frac{P(N_u = k, N_t = n)}{P(N_t = n)} \\
&= \frac{P(N_u = k, N_t - N_u = n - k)}{P(N_t = n)} \\
&= \frac{\left\{\frac{e^{-\lambda u}(\lambda u)^k}{k!}\right\} \cdot \left\{\frac{e^{-\lambda(t-u)}[\lambda(t-u)]^{n-k}}{(n-k)!}\right\}}{\frac{e^{-\lambda t}(\lambda t)^n}{n!}} \\
&= \frac{n!}{k!(n-k)!}\frac{u^k(t-u)^{nk}}{t^n} = C_n^k \left(\frac{u}{t}\right)^k \left(1 - \frac{u}{t}\right)^{n-k}
\end{aligned}$$

■

Example 3.3.11:::> $\{N(t)\}$ 为强度为 2 的 Poisson 过程, 求条件期望 $E[N(5)|N(2)]$ 以及 $E[N(2)|N(5)]$.

首先, 要明确泊松分布的物理意义. 由于这是同一个泊松分布, 所以可以理解为这是一个在人流的过程中已知某时间人的数量求之前或之后某个时间点人的数量的一个条件期望.

由于

$$\begin{aligned}
P(N_5 = k \mid N_2 = n) &= \frac{P(N_5 = k, N_2 = n)}{P(N_2 = n)} \\
&= \frac{P(N_5 - N_2 = k - n)P(N_2 = n)}{P(N_2 = n)} \\
&= P(N_3 = k - n)
\end{aligned}$$

根据泊松过程的性质, 个数与 (时间间隔长度 * 强度) 成泊松分布, 则:

$$E(N(5)\mid N(2))=\sum_{k=n}^{\infty}\frac{k\cdot 6^{k-n}e^{-6}}{(k-n)!}=\sum_{k=n}^{\infty}\frac{(k-n)\cdot 6^{k-n}e^{-6}}{(k-n)!}+\sum_{k=n}^{\infty}\frac{n\cdot 6^{k-n}e^{-6}}{(k-n)!}$$

套用泰勒公式, 则结果为:

$$6+n=6+N_2$$

而

$$\begin{aligned}&P(N_2=k\mid N_5=n)\\&=C_n^k\cdot\left(\frac{2}{5}\right)^k\cdot\left(\frac{3}{5}\right)^{n-k}\end{aligned}$$

故:

$$E[N(2)\mid N(5)]=\frac{2}{5}n=\frac{2}{5}N_5$$

Theorem 3.3.12:::> $\{X_t,t\geqslant 0\}$ 与 $\{Y_t,t\geqslant 0\}$ 分别为强度为 λ_1 和 λ_2 的 Poisson 过程, 且相互独立. 则

$$P(Y_t=k\mid X_t+Y_t=n)=\begin{cases}C_n^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-k}, & k=0,1,2,\cdots,n\\ 0, & \text{otherwise}\end{cases}$$

从而

$$E(Y_t\mid X_t+Y_t)=(X_t+Y_t)\frac{\lambda_2}{\lambda_1+\lambda_2}$$

这一定理从直观上非常容易理解, 概率的值和泊松分布其实并没有关系.

Example 3.3.13 (随机分流性)**:::>** 假设顾客到达商场的人数是强度为 λ 的 Poisson 过程, 且每一个到达商场的顾客是男性还是女性的概率分别为 p 和 $q(p+q=1)$, 若 $N_t^{(1)}$ 和 $N_t^{(2)}$ 分别为 $(0,t]$ 内到达商场的男性顾客和女性顾客数, 则 $N_t^{(1)}$ 和 $N_t^{(2)}$ 分别是强度为 λp 和 $\lambda(1-p)$ 的 Poisson 过程, 且相互独立.

证明思路: 首先, 由于顾客人数为泊松过程, 服从独立增量性; 且我们已知泊松分布的随机分流性, 故可证得以上结论.

我们先来计算 $N_t^{(1)}$ 和 $N_t^{(2)}$ 的联合分布. 由于 N_t 的分布是知道的, 因此我们自然想到要利用这一已知信息, 故可由全概率公式得

$$\begin{aligned}&P\left(N_t^{(1)}=n,N_t^{(2)}=m\right)\\&=\sum_{k=0}^{\infty}P\left(N_t^{(1)}=n,N^{(2)}=m\mid N(t)=k\right)P(N(t)=k)\end{aligned}$$

由于当 $k \neq n+m$ 时

$$P\left(N_t^{(1)}=n, N_t^{(2)}=m \mid N(t)=k\right)=0$$

故

$$\begin{aligned}
&P\left(N_t^{(1)}=n, N_t^{(2)}=m\right) \\
&=P\left(N_t^{(1)}=n, N_t^{(2)}=m \mid N(t)=n+m\right) P(N(t)=n+m) \\
&=P\left(N_t^{(1)}=n, N_t^{(2)}=m \mid N(t)=n+m\right) \frac{(\lambda t)^{n+m} \mathrm{e}^{\lambda t}}{(n+m)!}.
\end{aligned}$$

由于每一到这商场的顾客是男性或女性的概率分别为 p 和 q, 因此, 在已知有 $n+m$ 个人到达商场的条件下, 其中恰有 n 个男性顾客及 m 个女性顾客的条件分布正好为二项分布, 即

$$P\left(N_t^{(1)}=n, N_t^{(2)}=m \mid N(t)=n+m\right)=C_{n+m}^n p^n(1-p)^m$$

因此

$$\begin{aligned}
P\left(N_t^{(1)}=n, N_t^{(2)}=m\right) &= C_{n+m}^n p^n(1-p)^m \frac{(\lambda t)^{n+m} \mathrm{e}^{-\lambda t}}{(n+m)!} \\
&= \frac{(\lambda p t)^n \mathrm{e}^{-\lambda p t}}{n!} \cdot \frac{(\lambda(1-p) t)^m \mathrm{e}^{-\lambda(1-p) t}}{m!}
\end{aligned}$$

从而有

$$\begin{aligned}
&P\left(N_t^{(1)}=n\right)=\sum_{m=0}^{\infty} P\left(N_t^{(1)}=n, N_t^{(2)}=m\right) \\
&=\frac{(\lambda p t)^n}{n!} \mathrm{e}^{-\lambda p t} \sum_{m=0}^{\infty} \frac{(\lambda(1-p) t)^m}{m!} \mathrm{e}^{-\lambda(1-p) t}=\frac{(\lambda p t)^n}{n!} \mathrm{e}^{-\lambda p t}
\end{aligned}$$

类似地

$$P\left(N_t^{(2)}=m\right)=\frac{(\lambda(1-p) t)^m}{m!} \mathrm{e}^{-\lambda(1-p) t}$$

也就是说, $N_t^{(1)}$ 和 $N_t^{(2)}$ 分别服从强度为 λp 和 $\lambda(1-p)$ 的 Poisson 过程.

又因为对所有的 n, m 和 t, 有

$$P\left(N_t^{(1)}=n, N_t^{(2)}=m\right)=P\left(N_t^{(1)}=n\right) P\left(N_t^{(2)}=m\right)$$

故在 $(0, t]$ 内到达商场的男性顾客数 $N_t^{(1)}$ 和女性顾客数 $N_t^{(2)}$ 是相互独立的.

但是, 证明这两个随机过程相互独立远远没有结束, 我们现在才证明随机过程的一维分布相互独立. 总之, 经过冗长但不难的推导, 我们最后总能证得.

Definition 3.3.14 (复合泊松过程)**:::>** 设 $\{N_t, t \geqslant 0\}$ 为强度为 $\lambda > 0$ 的 Poisson 过程, $\{Y_i, i \geqslant 1\}$ 是与 $\{N_t, t \geqslant 0\}$ 独立的随机变量序列, 且 Y_i i.i.d.. 定义

$$X_t = \sum_{i=1}^{N_t} Y_i$$

为复合泊松过程. 注意若 $Y_i = 1$ 恒成立, 则 X_t 退化为泊松过程.

Theorem 3.3.15:::> 复合泊松过程 $\{Y_t : t \geqslant 0\}$ 为时齐的独立增量过程, 且矩母函数 $M_{Y_t}(u)$ 为:

$$\begin{aligned}
M_t(u) &= E\left(e^{uX_t}\right) = \sum_{n=0}^{\infty} E\left(e^{uX_t} \mid N_t = n\right) P(N_t = n) \\
&= \sum_{n=0}^{\infty} E\left(e^{u(Y_1+\cdots+Y_n)} \mid N_t = n\right) P(N_t = n) \\
&= \sum_{n=0}^{\infty} E\left(e^{u(Y_1+\cdots+Y_n)}\right) e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\
&= \sum_{n=0}^{\infty} (m_Y(u))^n \, e^{-\lambda t} \frac{(\lambda t)^n}{n!} \\
&= e^{\lambda t (m_Y(u) - 1)}
\end{aligned}$$

注意其中 N_t 是复合泊松过程复合的个数. 在公式的推导中利用了 Y_t 与 N_t 的独立性.

从而可得:

$$EX_t = \lambda t E Y_1$$
$$DX_t = \lambda t E Y_1^2$$

Theorem 3.3.16:::> 假设强度为 λ 的 Poisson 过程的每个事件被分成类型 $1, 2, \cdots, k$. 在时刻 s 一类事件发生与否与其他类事件发生与否相互独立, 并以概率 $P_i(s)$ 划归 i 型 $(i = 1, 2, k)$, $\sum_{i=1}^{k} p_i(s) = 1$. 以 $N_t^{(i)}$ 记在 $(0, t]$ 内 i 型事件到来的个数. 那么, 随机过程 $\mathrm{N}^{(i)} = \left\{N_t^{(i)}, t \geqslant 0\right\} (i = 1, 2, \cdots, k)$ 是相互独立的强度函数为 $\lambda \int_0^t p_i(s)\mathrm{d}s$ 的非时齐的 Poisson 过程.

Example 3.3.17:::> 已知一列顾客以泊松流到来, 记第一个到达的顾客到达时间为 τ, 求 τ 的分布:

我们通过 CDF 来表示这个分布, 即

$$F_\tau(t) = P(\tau \leqslant t) = 1 - P(\tau > t).$$

也即为:

$$1-P(N_t=0)=1-e^{-\lambda t}$$

之后我们会介绍, 这个分布被称为指数分布.

类似地, 若记第 i 名顾客到达的时间为 τ_i, 则 $\tau_i-\tau_{i-1}$ 也服从指数分布, 且它们之间是 iid 的.

3.4. Brownian motion

Definition 3.4.1 (Einstein 的模型)**:::>** 作随机运动的粒子在时间 $[0,t]$ 上的位移为 $\{B_s:0\leqslant s\leqslant t\}$ （初始位置 $B_0=0$）

1. 粒子位移的各分量都相互独立.(i.e. 各分量为独立增量过程) 三维布朗运动可以看作三个一维布朗运动的叠加;
2. 运动的统计规律对空间是对称的, i.e. $EB_t=0$;
3. 增量 $B_{t+h}-B_h$ 的分布与 h 无关 (i.e. 时齐), 且 $\sigma(t)\equiv E\left(B_{t+h}-B_h\right)^2$ 存在, 而且是 t 的连续函数. ($\Rightarrow\sigma(t+s)=\sigma(t)+\sigma(s)\Rightarrow\sigma(t)=Dt$)

结论: $\{B_t:t\geqslant 0\}$ 为时齐的独立增量过程, 其一维分布为 $N(0,Dt)$

理由: 如记 B_t 的特征函数为 $\varphi(t,\theta)=Ee^{i\theta B_t}(-\infty<\theta<\infty)$, 则

$$\begin{aligned}\varphi(t+s,\theta)-\varphi(t,\theta)&=Ee^{i\theta B_{t+s}}-Ee^{i\theta B_t}\\&=E\left[e^{i\theta B_t}\left(e^{i\theta\left(B_{t+s}-B_t\right)}-1\right)\right]=E\left[e^{i\theta(B_t-B_0)}\left(e^{i\theta\left(B_{t+s}-B_t\right)}-1\right)\right]\\&=E\left(e^{i\theta(B_t-B_0)}\right)E\left(e^{i\theta\left(B_{t+s}-B_t\right)}-1\right)=E\left(e^{i\theta B_t}\right)E\left(e^{i\theta(B_s-B_0)}-1\right)\\&=\varphi(t,\theta)E\left(e^{i\theta B_s}-1\right).\end{aligned}$$

则

$$\frac{\partial\varphi}{\partial t}(t,\theta)=-\frac{1}{2}D\theta^2\varphi(t,\theta)\text{泰勒展开}$$
$$\varphi(0,\theta)=1$$

解得

$$\varphi(t,\theta)=e^{-\frac{1}{2}\theta^2tD}$$

Definition 3.4.2 (Brown 运动的定义)**:::>** Brown 运动定义为满足以下条件的一个随机过程 $\boldsymbol{B}=\{B_t, t\geqslant 0\}$

1. $\boldsymbol{B}$ 是独立增量过程, 即对任意互不相交的区间 $(s_1,t_1],(s_2,t_2],\cdots,(s_n,t_n]$, 其上的增量 $B_{t_1}-B_{s_1}, B_{t_2}-B_{s_2},\cdots,B_{t_n}-B_{s_n}$ 都相互独立
2. 对于任意 $s\geqslant 0, t>0$, 增量 $B_{s+t}-B_s\sim N(0,Dt)$ (不依赖 s)
3. 对每一个固定的 $\omega, B_t(\omega)$ 是 t 的连续函数（此条件不是必需的)

特别地, 当 $D=1$ 时, 我们称之为标准 Brown 运动. 以下研究的 Brown 运动均为标准 Brown 运动.

注:

1. Markov 性: 已知现在 B_s 的条件下, 过去 $B_u(0\leqslant u<s)$ 与将来 B_{t+s} 是相互独立的;
2. $\boldsymbol{B}$ 的任意有限维分布为

$$P\left(\omega: B_{t_1}\leqslant x_1,\cdots,B_{t_n}\leqslant x_n\right)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f(u_1,u_2,\cdots,u_n)\,du_1\cdots du_n$$

$$=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}\frac{\exp\left\{-\left(\frac{u_1^2}{2t_1}+\frac{(u_2-u_1)^2}{2(t_2t_1)}+\cdots+\frac{\left(u_n-u_{n-1}\right)^2}{2\left(t_n-t_{n-1}\right)}\right\}}{(2\pi)^{\frac{n}{2}}\sqrt{t_1(t_2-t_1)\cdots\left(t_n-t_{n-1}\right)}}du_1\cdots du_n$$

3. Brown 运动是一个 Gauss 过程

Definition 3.4.3 (Gauss 过程)**:::>** 一个实值连续时间过程 X 称为 Gauss 过程, 如果每一有限维向量 $(X(t_1),X(t_2),\cdots,X(t_n))^T$ 均服从 Gauss 分布 $N(\boldsymbol{\mu}(\boldsymbol{t}),\boldsymbol{R}(\boldsymbol{t}))$, 其中的均值向量 μ 和协方差矩阵 $\boldsymbol{R}$ 均依赖于 $\boldsymbol{t}=(t_1,t_2,\cdots,t_n)$. 即这个随机过程任意时刻有限维的联合分布都是一个高斯分布.

Theorem 3.4.4:::> 设随机过程 $\{B_t: t\geqslant 0\}$ 满足 $B_0=0$. 那么 $\{B_t: t\geqslant 0\}$ 是 Brown 运动当且仅当它是 Gauss 过程, 而且满足 $EB_t\equiv 0, Cov(B_sB_t)=\min(s,t)$

Proof. 必要性: 若 $\{B_t : t \geqslant 0\}$ 是 Brown 运动 $(B_0 = 0)$, 则 $B_t \sim N(0,t)$, 故 $EB_t = 0$. 当 $s \leqslant t$ 时, 我们有

$$\begin{aligned}\mathrm{Cov}(B_s, B_t) &= \mathrm{Cov}(B_s, B_s + B_t - B_s)\\ &= \mathrm{Cov}(B_s, B_s) + \mathrm{Cov}(B_s, B_t - B_s) = s\end{aligned}$$

可见一般地我们有

$$\mathrm{Cov}(B_s, B_t) = E(B_s B_t) = s \wedge t \quad (\text{即} \min(s,t))$$

充分性: 若 $\{B_t : t \geqslant 0\}$ 是 Gauss 过程, 而且满足 $EB_t \equiv 0, E(B_s B_t) = \min(s,t)$ 则 $\forall s, t > 0$, 有 $E(B_t - B_s) = 0$

$$E(B_t - B_s)^2 = EB_t^2 + EB_s^2 - 2E(B_t B_s) = t + s - 2(s \wedge t) = |t - s|$$

而 $\forall \quad 0 \leqslant s_1 < t_1 \leqslant s_2 < t_2$, 有

$$Cov\left[\left(B_{t_1} - B_{s_1}\right)\left(B_{t_2} - B_{s_2}\right)\right] = Cov\left(B_{t_1} B_{t_2}\right) - Cov\left(B_{t_1} B_{s_2}\right) - Cov\left(B_{s_1} B_{t_2}\right) + Cov\left(B_{s_1} B_{s_2}\right)$$

$$= t_1 - t_1 - s_1 + s_1 = 0$$

即它们的协方差为 0 . 由于 Gauss 分布中分量相互独立等价于协方差为对角阵, 故可知 $\{B_t : t \geqslant 0\}$ 为独立增量过程, 且 $B_t - B_s \sim N(0, |t-s|)$, 故 $\{B_t : t \geqslant 0\}$ 为 Brown 运动. ■

Example 3.4.5::: 比如, 我们可以写出三个布朗运动的联合分布:

$$\begin{pmatrix} B_1 \\ B_2 \\ B_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 4 \end{pmatrix}\right)$$

其中, 协方差矩阵对角线的位置为方差, 而每个布朗运动的方差即是下标. 非对角线位置是两个布朗运动的协方差, 根据上述结论即是两者下标中较小的那个.

若求

$$P(B_1 + 2B_2 + 5B_4 \leqslant 1)$$

则可利用:

$$B_1 + 2B_2 + 5B_4 \sim N(0, 1 + 4 \times 2 + 25 \times 4 + 4 \times 1 + 20 \times 2 + 10 \times 1)$$

再用正态分布查表即可.

Example 3.4.6:::>

$$E(B_4|B_1)=E(B_4-B_1+B_1|B_1)=E(B_4-B_1|B_1)+B_1$$

而 B_4-B_1 与 B_1 独立, 且为 Gauss 分布, 故期望为零. 则

$$E(B_4|B_1)=B_1$$

而 $E(B_1|B_4)$ 就无法用上述方法解决.

但是我们有:

$$\begin{pmatrix} B_1 \\ B_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}\right)$$

是一个二维正态分布.

代入二维正态分布的性质中所介绍的条件分布公式即可得.

Theorem 3.4.7:::> 设 $\{B_t : t \geqslant 0\}$ 为 Brown 运动, 则

1. (平移不变性) $\{B_{t+a}-B_a, t \geqslant 0\}$ ($a \geqslant 0$ 为常数) 为 Brown 运动;
2. (尺度不变性) $\left\{\frac{B_{ct}}{\sqrt{c}}, t \geqslant 0\right\}$ ($c>0$ 为常数) 也是 Brown 运动;
3. $\left\{tB_{\frac{1}{t}}, t \geqslant 0\right\}$ (定义 $\left\{tB_{\frac{1}{t}}\right\}\Big|_{t=0}=0$) 为 Brown 运动

证明: 可利用上条定理证明: $\{B_t : t \geqslant 0\}$ 为 Brown 运动, 从而为 Gauss 过程, 因此 $\{B_{t+a}-B_a, t \geqslant 0\}$ 仍为 Gauss 过程, 且 $B_{0+a}-B_a=0$; $E\left(B_{t+a}-B_a\right)=0 \quad \forall t>0$. 而

$$\begin{aligned} &Cov\left[\left(B_{t+a}-B_a\right)\left(B_{s+a}-B_a\right)\right]=E\left(B_{t+a}B_{s+a}\right)-a-a+a \\ &=[(t+a)\wedge(s+a)]-a \\ &=s\wedge t+a-a=s\wedge t \end{aligned}$$

故 $\{B_{t+a}-B_a, t \geqslant 0\}$ ($a \geqslant 0$ 为常数) 为 Brown 运动.

Theorem 3.4.8 (Brown 运动的轨道性质)**:::>**

1. Brown 运动 $\{B_t, t \geqslant 0\}$ 的轨道处处连续处处不可导

2. Brown 运动 $\{B_t, t \geqslant 0\}$ 的首达 a 的首达时 $T_a = \min\{t > 0 : B_t = a\}$ 的分布为
$$P(T_a \leqslant t) = \sqrt{\frac{2}{\pi}} \int_{\frac{|a|}{\sqrt{t}}}^{\infty} e^{-\frac{x^2}{2}} dx = 2\left(1 - \Phi\left(\frac{|a|}{\sqrt{t}}\right)\right)$$
从而, T_a 几乎处处有限 (即 $P(T_a < \infty) = 1$) 且 $ET_a = +\infty$.

3. $\max_{0 \leqslant s \leqslant t} B_s$ 的密度函数为 $f_{\max}(a) = \frac{2}{\sqrt{2\pi t}} e^{-\frac{a^2}{2t}}, \quad a \geqslant 0$

4. 设 $\{B_t^x, t \geqslant 0\} = \{B_t + x, t \geqslant 0\}$ 为始于 x 的 Brown, 则 B_t^x 在 $(0, t)$ 中至少有一个零点的概率为 $\frac{|x|}{\sqrt{2\pi}} \int_0^t u^{-\frac{3}{2}} e^{-\frac{x^2}{2u}} du$
证明时: (注意到 $P(B_t^x$ 在 $(0,t)$ 中至少有一个零点 $) = P\left(\max_{0 \leqslant s \leqslant t} B_s^x \geqslant 0\right)$ 即可

5. 反正弦律:
$$P\left(B_t^x \text{ 在}(a,b) \text{ 中没有零点}\right) = \frac{2}{\pi} \arcsin \sqrt{\frac{a}{b}}$$

3.5. Markov Chains

3.5.1 The Markov Property and Transition Probability

Definition 3.5.1 (Markov chain)**:::>** Let $X = (X_0, X_1, \cdots)$ be a sequence of random variables taking values in some set S, the *state space*. We assume that S is countable (which could be finite).

We say X has the *Markov property* if for all $n \geqslant 0, i_0, \cdots, i_{n+1} \in S$, we have
$$P(X_{n+1} = i_{n+1} \mid X_0 = i_0, \cdots, X_n = i_n) = P(X_{n+1} = i_{n+1} \mid X_n = i_n).$$
If X has the Markov property, we call it a *Markov chain*.

We say that a Markov chain X is *homogeneous* if the conditional probabilities $P(X_{n+1} = j \mid X_n = i)$ do not depend on n.

All our chains X will be Markov and homogeneous unless otherwise specified.

Since the state space S is countable, we usually label the states by integers $i \in N$.

Example 3.5.2:::>

1. A random walk is a Markov chain.
2. The branching process is a Markov chain.

In general, to fully specify a (homogeneous) Markov chain, we will need two items:

1. The initial distribution $\lambda_i = P(X_0 = i)$. We can write this as a vector $\lambda = (\lambda_i : i \in S)$.
2. The transition probabilities $p_{i,j} = P(X_{n+1} = j \mid X_n = i)$. We can write this as a matrix $P = (p_{i,j})_{i,j \in S}$.

Proposition 3.5.3:::>

1. λ is a *distribution*, i.e. $\lambda_i \geqslant 0, \sum_i \lambda_i = 1$.
2. P is a *stochastic matrix*, i.e. $p_{i,j} \geqslant 0$ and $\sum_j p_{i,j} = 1$ for all i.

Proof.

1. Obvious since λ is a probability distribution.
2. $p_{i,j} \geqslant 0$ since p_{ij} is a probability. We also have

$$\sum_j p_{i,j} = \sum_j P(X_1 = j \mid X_0 = i) = 1$$

since $P(X_1 = \cdot \mid X_0 = i)$ is a probability distribution function. ■

Note that we only require the row sum to be 1, and the column sum need not be.

Theorem 3.5.4:::> Let λ be a distribution (on S) and P a stochastic matrix. The sequence $X = (X_0, X_1, \cdots)$ is a Markov chain with initial distribution λ and transition matrix P iff

$$P(X_0 = i_0, X_1 = i_1, \cdots, X_n = i_n) = \lambda_{i_0} p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},i_n} \qquad (*)$$

for all $n, i_0, \cdots, i_n$

Proof. Let A_k be the event $X_k = i_k$. Then we can write $(*)$ as

$$P(A_0 \cap A_1 \cap \cdots \cap A_n) = \lambda_{i_0} p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},i_n}. \qquad (*)$$

We first assume that X is a Markov chain. We prove $(*)$ by induction on n.

When $n = 0$, $(*)$ says $P(A_0) = \lambda_{i_0}$. This is true by definition of λ.

Assume that it is true for all $n < N$. Then

$$\begin{aligned} P(A_0 \cap A_1 \cap \cdots \cap A_N) &= P(A_0, \cdots, A_{N-1}) P(A_0, \cdots, A_N \mid A_0, \cdots, A_{N-1}) \\ &= \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{N-2},i_{N-1}} P(A_N \mid A_0, \cdots, A_{N-1}) \\ &= \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{N-2},i_{N-1}} P(A_N \mid A_{N-1}) \\ &= \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{N-2},i_{N-1}} p_{i_{N-1},i_N}. \end{aligned}$$

So it is true for N as well. Hence we are done by induction.

Conversely, suppose that $(*)$ holds. Then for $n = 0$, we have $P(X_0 = i_0) = \lambda_{i_0}$. Otherwise,

$$\begin{aligned} P(X_n = i_n \mid X_0 = i_0, \cdots, X_{n-1} = i_{n-1}) &= P(A_n \mid A_0 \cap \cdots \cap A_{n-1}) \\ &= \frac{P(A_0 \cap \cdots \cap A_n)}{P(A_0 \cap \cdots \cap A_{n-1})} \\ &= p_{i_{n-1},i_n}, \end{aligned}$$

which is independent of $i_0, \cdots, i_{n-2}$. So this is Markov. ■

Theorem 3.5.5 (Extended Markov property)**:::>** Let X be a Markov chain. For $n \geqslant 0$, any H given in terms of the past $\{X_i : i < n\}$, and any F given in terms of the future $\{X_i : i > n\}$, we have

$$P(F \mid X_n = i, H) = P(F \mid X_n = i).$$

Definition 3.5.6 (n-step transition probability)**:::>** The n-step transition probability from i to j is

$$p_{i,j}(n) = P(X_n = j \mid X_0 = i).$$

How do we compute these probabilities? The idea is to break this down into smaller parts. We want to express $p_{i,j}(m+n)$ in terms of n-step and m-step transition probabilities. Then we can iteratively break down an arbitrary $p_{i,j}(n)$ into expressions involving the one-step transition probabilities only.

To compute $p_{i,j}(m+n)$, we can think of this as a two-step process. We first go from i to some unknown point k in m steps, and then travel from k to j in n more steps. To find the probability to get from i to j, we consider all possible routes from i to j, and sum up all the probability of the paths. We have

$$\begin{aligned} p_{ij}(m+n) &= P(X_{m+n} \mid X_0 = i) \\ &= \sum_k P(X_{m+n} = j \mid X_m = k, X_0 = i) P(X_m = k \mid X_0 = i) \\ &= \sum_k P(X_{m+n} = j \mid X_m = k) P(X_m = k \mid X_0 = i) \\ &= \sum_k p_{i,k}(m) p_{k,j}(n). \end{aligned}$$

Theorem 3.5.7 (Chapman-Kolmogorov equation)**:::>**

$$p_{i,j}(m+n) = \sum_{k \in S} p_{i,k}(m) p_{k,j}(n).$$

This formula is suspiciously familiar. It is just matrix multiplication!

Example 3.5.8:::> Let $S = \{1,2\}$, with

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

We assume $0 < \alpha, \beta < 1$. We want to find the n-step transition probability.

We can achieve this via diagonalization. We can write P as

$$P = U^{-1} \begin{pmatrix} \varkappa_1 & 0 \\ 0 & \varkappa_2 \end{pmatrix} U,$$

where the $\varkappa_i$ are eigenvalues of P, and U is composed of the eigenvectors.

To find the eigenvalues, we calculate

$$\det(P-\lambda I)=(1-\alpha-\lambda)(1-\beta-\lambda)-\alpha\beta=0.$$

We solve this to obtain

$$\varkappa_1=1,\quad \varkappa_2=1-\alpha-\beta.$$

Usually, the next thing to do would be to find the eigenvectors to obtain U. However, here we can cheat a bit and not do that. Using the diagonalization of P, we have

$$P^n=U^{-1}\begin{pmatrix}\varkappa_1^n & 0\\ 0 & \varkappa_2^n\end{pmatrix}U.$$

We can now attempt to compute $p_{1,2}$. We know that it must be of the form

$$p_{1,2}=A\varkappa_1^n+B\varkappa_2^n=A+B(1-\alpha-\beta)^n$$

where A and B are constants coming from U and U^{-1}. However, we know well that

$$p_{1,2}(0)=0,\quad p_{12}(1)=\alpha.$$

So we obtain

$$\begin{aligned}A+B&=0\\ A+B(1-\alpha-\beta)&=\alpha.\end{aligned}$$

This is something we can solve, and obtain

$$p_{1,2}(n)=\frac{\alpha}{\alpha+\beta}(1-(1-\alpha-\beta)^n)=1-p_{1,1}(n).$$

How about $p_{2,1}$ and $p_{2,2}$? Well we don't need additional work. We can obtain these simply by interchanging α and β. So we obtain

$$P^n=\frac{1}{\alpha+\beta}\begin{pmatrix}\beta+\alpha(1-\alpha-\beta)^n & \alpha-\alpha(1-\alpha-\beta)^n\\ \beta-\beta(1-\beta-\alpha)^n & \alpha+\beta(1-\beta-\alpha)^n\end{pmatrix}$$

What happens as $n\to\infty$? We can take the limit and obtain

$$P^n\to\frac{1}{\alpha+\beta}\begin{pmatrix}\beta & \alpha\\ \beta & \alpha\end{pmatrix}$$

We see that the two rows are the same. This means that as time goes on, where we end up does not depend on where we started. We will later (near the end of the course) see that this is generally true for most Markov chains.

Alternatively, we can solve this by a difference equation. The recurrence relation is given by

$$p_{1,1}(n+1) = p_{1,1}(n)(1-\alpha) + p_{1,2}(n)\beta.$$

Writing in terms of p_{11} only, we have

$$p_{1,1}(n+1) = p_{1,1}(n)(1-\alpha) + (1-p_{1,1}(n))\beta.$$

We can solve this as we have done in IA Differential Equations.

3.5.2 Classification of chains and states

Definition 3.5.9 (Leading to and communicate)**:::>** Suppose we have two states $i, j \in S$. We write $i \to j$ (*i leads to j*) if there is some $n \geqslant 0$ such that $p_{i,j}(n) > 0$, i.e. it is possible for us to get from i to j (in multiple steps). Note that we allow $n = 0$. So we always have $i \to i$.

We write $i \leftrightarrow j$ if $i \to j$ and $j \to i$. If $i \leftrightarrow j$, we say i and j *communicate*.

Proposition 3.5.10:::> $\leftrightarrow$ is an equivalence relation.

Proof.

1. Reflexive: we have $i \leftrightarrow i$ since $p_{i,i}(0) = 1$.
2. Symmetric: trivial by definition.
3. Transitive: suppose $i \to j$ and $j \to k$. Since $i \to j$, there is some $m > 0$ such that $p_{i,j}(m) > 0$. Since $j \to k$, there is some n such that $p_{j,k}(n) > 0$. Then $p_{i,k}(m+n) = \sum_r p_{i,r}(m)p_{rk}(n) \geqslant p_{i,j}(m)p_{j,k}(n) > 0$. So $i \to k$.

   Similarly, if $j \to i$ and $k \to j$, then $k \to i$. So $i \leftrightarrow j$ and $j \leftrightarrow k$ implies that $i \leftrightarrow k$. ∎

So we have an equivalence relation, and we know what to do with equivalence relations. We form equivalence classes!

Definition 3.5.11 (Communicating classes)**:::>** The equivalence classes of $\leftrightarrow$ are *communicating classes*.

We have to be careful with these communicating classes. Different communicating classes are not completely isolated. Within a communicating class A, of course we can move between any two vertices. However, it is also possible that we can escape from a class A to a different class B. It is just that after going to B, we cannot return to class A. From B, we might be able to get to another space C. We can jump around all the time, but (if there are finitely many communicating classes) eventually we have to stop when we have visited every class. Then we are bound to stay in that class.

Since we are eventually going to be stuck in that class anyway, often, we can just consider this final communicating class and ignore the others. So wlog we can assume that the chain only has one communicating class.

Definition 3.5.12 (Irreducible chain)**:::>** A Markov chain is *irreducible* if there is a unique communication class.

From now on, we will mostly care about irreducible chains only.

More generally, we call a subset *closed* if we cannot escape from it.

Definition 3.5.13 (Closed)**:::>** A subset $C \subseteq S$ is *closed* if $p_{i,j} = 0$ for all $i \in C, j \notin C$.

Proposition 3.5.14:::> A subset C is closed iff "$i \in C, i \to j$ implies $j \in C$".

Proof. Assume C is closed. Let $i \in C, i \to j$. Since $i \to j$, there is some m such that $p_{i,j}(m) > 0$. Expanding the Chapman-Kolmogorov equation, we have

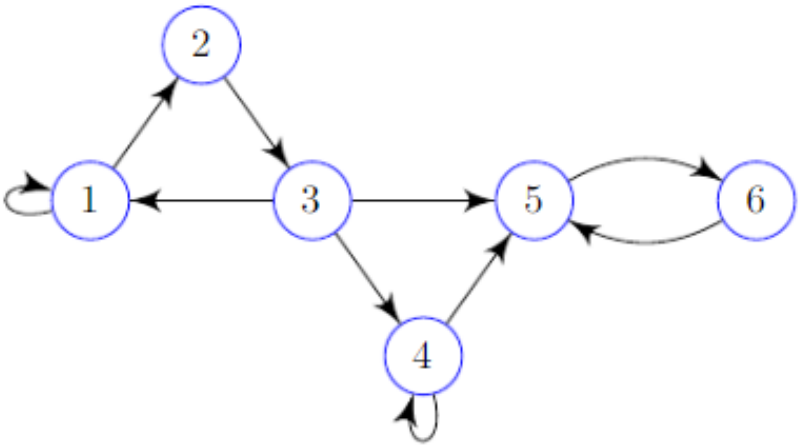
$$p_{i,j}(m) = \sum_{i_1,\cdots,i_{m-1}} p_{i,i_1} p_{i_1,i_2}, \cdots, p_{i_{m-1},j} > 0.$$

So there is some route $i, i_1, \cdots, i_{m-1}, j$ such that $p_{i,i_1}, p_{i_1,i_2}, \cdots, p_{i_{m-1},j} > 0$. Since $p_{i,i_1} > 0$, we have $i_1 \in C$. Since $p_{i_1,i_2} > 0$, we have $i_2 \in C$. By induction, we get that $j \in C$.

To prove the other direction, assume that "$i \in C, i \to j$ implies $j \in C$". So for any $i \in C, j \notin C$, then $i \not\to j$. So in particular $p_{i,j} = 0$. ■

Example 3.5.15:::> Consider $S = \{1,2,3,4,5,6\}$ with transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$



We see that the communicating classes are $\{1,2,3\}$, $\{4\}$, $\{5,6\}$, where $\{5,6\}$ is closed.

Definition 3.5.16:::> For convenience, we will introduce some notations. We write

$$P_i(A) = P(A \mid X_0 = i),$$

and

$$E_i(Z) = E(Z \mid X_0 = i).$$

Definition 3.5.17 (First passage time and probability)**:::>** Suppose we start from i, and randomly wander around. Eventually, we may or may not get to j. If we do, there is a time at which we first reach j. We call this the *first passage time*.

The *first passage time* of $j \in S$ starting from i is

$$T_j = \min\{n \geqslant 1 : X_n = j\}.$$

Note that this implicitly depends on i. Here we require $n \geqslant 1$. Otherwise T_i would always be 0.

The *first passage probability* is

$$f_{ij}(n) = P_i(T_j = n).$$

Definition 3.5.18 (Recurrent state)**:::>** A state $i \in S$ is *recurrent* (or *persistent*) if

$$P_i(T_i < \infty) = 1,$$

i.e. we will eventually get back to the state. Otherwise, we call the state *transient*.

Note that transient does *not* mean we don't get back. It's just that we are not sure that we will get back. We can show that if a state is recurrent, then the probability that we return to i infinitely many times is also 1.

4. Central limit theorem

4.1. Law of large numbers

Definition 4.1.1 (依概率收敛)**:::>** 若随机变量序列 $\{X_n\}$ 及随机变量 X 满足：对任意 $\varepsilon > 0$ 有

$$\lim_{n\to\infty} P(|X_n - X| \geqslant \varepsilon) = 0$$

Example 4.1.2:::> 设独立同分布的随机变量序列 $\{\xi_n\}$ 均服从 $[0,a]$ 上的均匀分布. 则

$$\max(\xi_1, \xi_2, \cdots \xi_n) \xrightarrow{P} a$$

Proof. 即证:

$$\forall 0 \leqslant \varepsilon \leqslant 1, P(|\max(\xi_1, \xi_2, \cdots \xi_n) - a| \geqslant \varepsilon) = 0$$

而

$$P(|\max(\xi_1, \xi_2, \cdots \xi_n) - a| \geqslant \varepsilon) = P(|\max(\xi_1, \xi_2, \cdots \xi_n) - a| \leqslant a - \varepsilon)$$

最大值小于 $a-\varepsilon$, 也即每个都小于 $a-\varepsilon$, 则上述概率为

$$(\frac{a-\varepsilon}{a})^n$$

故极限趋于 0. ■

Proposition 4.1.3:::>

1. (比较微积分的结论): 若 $\xi_n \xrightarrow{P} \xi, g(x)$ 是连续函数, 则 $g(\xi_n) \xrightarrow{P} g(\xi)$
   又若 $\eta_n \xrightarrow{P} \eta, a, b$ 为常数, 则

$$\xi_n + a \xrightarrow{P} \xi + a, \quad a\xi_n \pm b\eta_n \xrightarrow{P} a\xi \pm b\eta, \quad \xi_n\eta_n \xrightarrow{P} \xi\eta$$

   进一步, 如果还有 $\eta_n, \eta > 0$, 则 $\frac{\xi_n}{\eta_n} \xrightarrow{P} \frac{\xi}{\eta}$

2. 如果 $X_n \xrightarrow{P} X$, 那么它也是依分布收敛的, 即 $X_n \xrightarrow{D} X$

3. $X_n \xrightarrow{P} a$ (常数) 当且仅当 $X_n \xrightarrow{D} a$ (或 F_a); 所以也当且仅当特征函数 $\varphi_{X_n}(\theta) \to e^{ia\theta}$

Example 4.1.4 (依分布收敛不能推出依概率收敛)**:::>** Let $X \sim N(0,1)$. Let $X_n = -X$ for $n = 1,2,3,\ldots$; hence $X_n \sim N(0,1)$, X_n has the same distribution function as X for all n so, trivially,$\lim_n F_n(x) = F(x)$ for all x. Therefore, $X_n \rightsquigarrow X$. But $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|2X| > \varepsilon) = \mathbb{P}(|X| > \varepsilon/2) \neq 0$. So X_n does not converge to X in probability.

Example 4.1.5:::> Warning! One might conjecture that if $X_n \xrightarrow{\mathrm{P}} b$, then $\mathbb{E}(X_n) \to b$. This is not true. Let X_n be a random variable defined by

$$\mathbb{P}\left(X_n = n^2\right) = 1/n$$

and

$$\mathbb{P}(X_n = 0) = 1 - (1/n)$$

Now,

$$\mathbb{P}(|X_n| < \varepsilon) = \mathbb{P}(X_n = 0) = 1 - (1/n) \to 1$$

Hence,

$$X_n \xrightarrow{\mathrm{P}} 0$$

However,

$$\mathbb{E}(X_n) = \left[n^2 \times (1/n)\right] + [0 \times (1 - (1/n))] = n$$

Thus,

$$\mathbb{E}(X_n) \to \infty$$

Theorem 4.1.6 (Slutsky 定理)**:::>** $X_n \xrightarrow{D} X, Y_n \xrightarrow{P} a$, 则 $X_n + Y_n \xrightarrow{D} X + a$

Slusky 定理对四则运算都成立.

注意, Slusky 定理中要求一个随机变量收敛到常数, 不是常数的话可能不成立!

$\mathfrak{Proof}$. 只需证明 $X_n \xrightarrow{D} X, Y_n \xrightarrow{P} 0$, 则 $X_n + Y_n \xrightarrow{D} X$

1. 法一 (定义)

   $\forall x \in R, \quad \forall \varepsilon > 0$

   $$F_{X_n+Y_n}(x) = P(X_n + Y_n \leqslant x) = P(X_n + Y_n \leqslant x, |Y_n| \leqslant \varepsilon) + P(X_n + Y_n \leqslant x, |Y_n| > \varepsilon)$$
   $$\leqslant P(X_n + Y_n \leqslant x, |Y_n| \leqslant \varepsilon) + P(|Y_n| > \varepsilon)$$
   $$\leqslant P(X_n \leqslant x + \varepsilon) + P(|Y_n| > \varepsilon)$$

   则有

   $$\overline{\lim}_{n\to\infty} F_{X_n+Y_n}(x) \leqslant F_X(x+\varepsilon)$$

   同理,

   $$P(X_n \leqslant x - \varepsilon) = P(X_n \leqslant x - \varepsilon, |Y_n| \leqslant \varepsilon) + P(X_n \leqslant x - \varepsilon, |Y_n| > \varepsilon)$$
   $$\leqslant P(X_n + Y_n \leqslant x) + P(|Y_n| > \varepsilon)$$

   则有

   $$\underline{\lim}_{n\to\infty} F_{X_n+Y_n}(x) \geqslant F_X(x-\varepsilon)$$

   令 $\varepsilon \downarrow 0$, 则对 F_X 的连续点 x, 有 $F_{X_n+Y_n}(x) \to F_X(x)$, 即 $X_n + Y_n \xrightarrow{D} X$

2. 法二 (特征函数)

   由于 $Y_n \xrightarrow{P} 0 \Leftrightarrow Y_n \xrightarrow{D} 0$, 故 $Ee^{i\theta Y_n} \to 1$, 从而

   $$\left| E\left[e^{i\theta X_n} \left(e^{i\theta Y_n} - 1 \right) \right] \right| \leqslant E\left[\left| e^{i\theta Y_n} - 1 \right| \right] \to 0$$

   故

   $$\varphi_{X_n+Y_n}(\theta) = Ee^{i\theta(X_n+Y_n)} = Ee^{i\theta X_n} + E\left[e^{i\theta X_n} \left(e^{i\theta Y_n} - 1 \right) \right] \to Ee^{i\theta X} = \varphi_X(\theta)$$

   即 $X_n + Y_n \xrightarrow{D} X$

■

Theorem 4.1.7 (大数定律)**:::>** $\{X_n\}$ 满足大数定律:

$$\forall \varepsilon > 0, \text{有} \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| < \varepsilon\right) = 1$$

也即

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i \to 0$$

大数定理在微积分中的对应: Cauchy 定理.

即: 若 $a_n \to a$, 则

$$\frac{1}{n}\sum_{i=1}^{n} a_i \to a$$

此定理的证明需用到 $\varepsilon - N$ 语言.

Corollary 4.1.8 (Chebyshev 不等式)**:::>** 对具有有限方差的随机变量 X, 及任意的 $\varepsilon > 0$ 有

$$P(|X - EX| \geqslant \varepsilon) \leqslant \frac{DX}{\varepsilon^2}$$

Theorem 4.1.9 (Chebyshev 大数定律)**:::>** 若 $\{X_n\}$ 满足两两不相关, 且方差一致有界 (即 $\exists C, DX_i \leqslant C < +\infty, \forall i$), 则 $\{X_n\}$ 满足大数定律.

证明需要用到 Chebyshev 不等式:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| \geqslant \varepsilon\right) \leqslant \frac{D(\frac{1}{n}\sum^{n} x_i)}{\varepsilon^2} = \frac{D(\sum^{n} x_i)}{n^2\varepsilon^2}$$

由于 $\{X_n\}$ 满足两两不相关, 且方差一致有界, 故上式小于等于

$$\frac{c}{n^2\varepsilon^2} \to 0$$

Corollary 4.1.10 (Markov 大数定理)**:::>** 若随机变量序列满足 Markov 条件:

$$\frac{1}{n^2} D\left(\sum_{i=1}^{n} X_i\right) \to 0$$

则满足大数定理.

Markov 大数定理是 Chebyshev 大数定律的推论, 实质上是给出了 Chebyshev 大数定律的证明第一步后的条件.

Theorem 4.1.11 (Khintchine 大数定理)**:::>** 设 $\{X_n\}$ 是相互独立同分布的随机变量序列, 且其数学期望为 μ, 则 $\{X_n\}$ 服从大数定律, 即 $\frac{X_1+\cdots+X_n}{n} \xrightarrow{P} \mu \quad (n \to \infty)$

Proof. 利用独立同分布性质, 对于特征函数我们有

$$\begin{aligned}\varphi_{\frac{X_1+\cdots+X_n}{n}}(\theta) &= \varphi_{X_1+\cdots+X_n}\left(\frac{\theta}{n}\right) = \left[\varphi_{X_1}\left(\frac{\theta}{n}\right)\right]^n \\ &= \left[1+i\mu\frac{\theta}{n}+o\left(\frac{\theta}{n}\right)\right]^n \to e^{i\mu\theta}.\end{aligned}$$

■

Khintchine 大数定理为我们通过观察大量取样后获得的样本值来推断总体的数学期望提供了理论依据.

Corollary 4.1.12 (Bernoulli 大数定律)**:::>** 在事件 A 发生的概率为 p 的 n 次重复独立试验的 Bernoulli 概型中, 令 μ_n 为 n 次重复中试验 A 发生的次数. 我们有: $\frac{\mu_n}{n} \xrightarrow{P} p$

4.2. Central limit theory

Definition 4.2.1 (中心极限定理)**:::>** $\{X_n\}$ 满足中心极限定理:

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n EX_i}{\sqrt{D(\sum_{i=1}^n X_i)}} < x\right) = \Phi(x) \quad \forall x \in R$$

也即

$$S_n^* = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n EX_i}{\sqrt{D(\sum_{i=1}^n X_i)}} \to N(0,1)$$

等价于 $\sum_{i=1}^n X_i$ 近似服从正态分布:

$$\sum_{i=1}^n X_i \sim N(E(\sum_{i=1}^n X_i), D(\sum_{i=1}^n X_i))$$

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$
$$\overline{X}_n - \mu \approx N\left(0, \frac{\sigma^2}{n}\right)$$
$$\sqrt{n}\left(\overline{X}_n - \mu\right) \approx N\left(0, \sigma^2\right)$$
$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \approx N(0,1).$$

Theorem 4.2.2 (Levy- Lindeberg 中心极限定理)**:::>** 设独立同分布随机变量序列 $\{X_n\}$ 具有有限的数学期望 $EX_i = \mu$ 和方差 $DX_i = \sigma^2 \neq 0$, 则 $\{X_n\}$ 满足中心极限定理, 即当 $n \to \infty$ 时有 $S_n^* \xrightarrow{D} N(0,1)$

即:

$$\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n EX_i}{\sqrt{D\left(\sum_{i=1}^n X_i\right)}} \to N(0,1)$$

也即:

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$

我们有: 在 $n \to \infty$ 时, 等式左端分子分母均趋向于 0, 等式右端为一常数. 故此时我们可知分子分母趋向于 0 的速度是同阶的.

Proof. 记 $X_k - \mu$ 的特征函数为 $\varphi(t)$, 令 $\zeta_n = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, 则其特征函数为 $\left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$.

由于 $E(X_k) = \mu, D(X_k) = \sigma^2$ 故 $\varphi'(0) = 0, \varphi''(0) = -\sigma^2$, 因此

$$\varphi(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o\left(t^2\right)$$

所以

$$\left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = \left[1 - \frac{1}{2n}t^2 + o\left(\frac{t^2}{n}\right)\right]^n \to e^{-t^2/2}$$

由于 $e^{-t^2/2}$ 是连续函数, 它对应似分布函数为 $\Phi(Z)$, 因此由逆极限定理知

$$\lim_{n\to\infty} P(\zeta_n \leqslant z) \to \Phi(z)$$

■

本定理只假设 $\{X_n\}$ 独立同分布, 方差存在, 不管原来的分布是什么, 只要 n 充分大, 就可以用正态分布去逼近随机变量和的分布, 所以这个定理有着广泛的应用.

Theorem 4.2.3 (Polya 定理)**:::>** 设随机变量 X 分布函数 $F_X(x)$ 是连续函数, 又随机变量序列 $X_n \xrightarrow{D} F_X$. 那么分布函数 $F_{X_n}(x)$ 一致收敛到 $F_X(x)$

Theorem 4.2.4:::> 若随机变量序列 X_n 为独立同分布, $EX_n = \mu$ 且 $DX_n = \sigma^2 (n = 1, 2, \cdots)$, 则

$$\sup_{-\infty<<<+\infty} \left| P(X_1 + \cdots + X_n \leqslant x) - \frac{1}{\sigma\sqrt{2\pi \cdot n}} \int_{-\infty}^{x} e^{-\frac{(u-n\mu)^2}{2n\sigma^2}} du \right|^{n\to\infty} \to 0$$

Theorem 4.2.5 (De Moivre-Laplace 定理)**:::>** 设 n 次独立的重复试验中, 事件 A 在每次试验中出现的概率 $p(0 < p < 1)$, 随机变量 μ_A 为此 n 次试验中事件 A 出现的次数 (即 $\mu_A \sim B(n, p)$). 那么, 对于任意 $x \in \boldsymbol{R}$ 有

$$\lim_{n\to\infty} P\left(\frac{\mu_A - np}{\sqrt{np(1-p)}} \leqslant x \right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz$$

而且对于 x 是一致收敛.

此即为二项分布的正态逼近. 注意要与二项分布的泊松逼近区分开, Poisson 定理也是适用于 n 很大的情况下, 不过 Poisson 定理还需要额外的条件:n 很大, p 很小, np 不大不小.

我们有结论:n 很大, p 很小, np 不大不小时, Poisson 逼近精度比二项分布精确; 其它情况下二项分布更加精确.

Example 4.2.6:::> 已知 n 重贝努利试验中参数 $p = 0.75$, 问至少应该做多少次试验, 才能使试验成功的频率在 0.74 和 0.76 之间的概率不低于 0.95?

由题设, 求 n 使

$$P\left(0.74 < \frac{\mu_n}{n} < 0.76\right) = P\left(\left|\frac{\mu_n}{n} - 0.75\right| < 0.01\right) \geqslant 0.95$$

注意 $\mu_n \sim \mathrm{B}(n, p)$, 故由 De Moivre-Laplace 定理得

$$P\left(\left|\frac{\mu_n}{n} - 0.75\right| < 0.01\right) \approx 2\Phi\left(0.01\sqrt{\frac{n}{0.75(0.25)}}\right) - 1 \geqslant 0.95$$

即求 n 使

$$2\Phi\left(0.01\sqrt{\frac{n}{0.75(0.25)}}\right)\geqslant 0.975$$

即至少应该取 n 使

$$0.01\sqrt{\frac{n}{0.75(0.25)}}\approx 1.96$$

解得 $n=1.96^2\times 0.1875\approx 7203$.

Remark. 因为二项分布是离散分布, 而正态分布是连续分布, 所以用正态分布作为二项分布的近似计算中, 作些修正可以提高精度. 若 $k_1<k_2$ 均为整数, 一般先作如下修正后再用正态近似:

$$P(k_1\leqslant s_n\leqslant k_2)=P(k_1-0.5<s_n<k_2+0.5)$$

■

Theorem 4.2.7 (Liapunov 定理)::> 设独立随机变量序列 $\{X_n\}$ 的数学期望和方差分别为 $EX_i=\mu_i$ 和 $DX_i=\sigma_i^2(i=1,2,\cdots)$(这里是不同分布, 期望方差都不一样), 记 ${B_n}^2=\sum_{i=1}^n\sigma_i^2$, 如果满足以下的 Liapunov 条件: 存在 $\delta>0$, 使得

$$\lim_{n\to\infty}\frac{1}{{B_n}^{2+\delta}}\sum_{i=1}^{n}E(|X_i-\mu_i|)^{2+\delta}=0$$

则对任意的 $x\in\boldsymbol{R}$ 有

$$\lim_{n\to\infty}P\left(\frac{1}{B_n}\sum_{i=1}^{n}(X_i-\mu_i)\leqslant x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{z^2}{2}}dz$$

Example 4.2.8::> 一份考卷由 99 个题目组成, 并按由易到难顺序排列. 某学生答对第 1 题的概率为 0.99, 答对第 2 题的概率为 0.98, 一般地, 他答对第 i 题的概率为 $1-i/100, i=1,2,\cdots$. 假设该学生回答各题目是相互独立的, 并且要正确回答其中 60 个以上 (包括 60 个) 题目才算通过考试. 试计算该学生通过考试的可能性多大?

Solution. 设

$$X_i=\begin{cases}1, & \text{若学生答对第}i\text{ 题},\\ 0, & \text{若学生答错第}i\text{ 题}.\end{cases}$$

于是诸 X_i 相互独立, 且服从不同的两点分布:

$$P(X_i=1)=p_i=1-\frac{i}{100},\quad P(X_i=0)=1-p_i=\frac{i}{100},$$

而我们要求的是:

$$P\left(\sum_{i=1}^{99}X_i\geqslant 60\right)$$

为使用中心极限定理, 我们可以设想从 X_{100} 开始的随机变量都与 X_{99} 同分布, 且相互独立. 下面我们用 $\delta=1$ 来验证随机变量序列 $\{X_n\}$ 满足李雅普诺夫条件, 因为

$$B_n=\sqrt{\sum_{i=1}^{n}\mathrm{Var}(X_i)}=\sqrt{\sum_{i=1}^{n}p_i(1-p_i)}\to\infty(n\to\infty)$$

$$E\left(|X_i-p_i|^3\right)=(1-p_i)^3p_i+p_i^3(1-p_i)\leqslant p_i(1-p_i)$$

于是

$$\frac{1}{B_n^3}\sum_{i=1}^{n}E\left(|X_i-p_i|^3\right)\leqslant\frac{1}{\left[\sum_{i=1}^{n}p_i(1-p_i)\right]^{1/2}}\to 0(n\to\infty)$$

即 $\{X_n\}$ 满足李雅谱诺夫条件, 所以可以使用中心极限定理.

又因为在 $n=99$ 时,

$$E\left(\sum_{i=1}^{99}X_i\right)=\sum_{i=1}^{99}p_i=\sum_{i=1}^{99}\left(1-\frac{i}{100}\right)=49.5$$

$$B_{99}^2=\sum_{i=1}^{99}\mathrm{Var}(X_i)=\sum_{i=1}^{99}\left(1-\frac{i}{100}\right)\left(\frac{i}{100}\right)=16.665$$

所以该学生通过考试的可能性为

$$P\left(\sum_{i=1}^{99}X_i\geqslant 60\right)=P\left(\frac{\sum_{i=1}^{99}X_i-49.5}{\sqrt{16.665}}\geqslant\frac{60-49.5}{\sqrt{16.665}}\right)$$
$$\approx 1-\Phi(2.57)=0.005.$$

由此看出: 此学生通过考试的可能性很小, 大约只有千分之五. ■

Theorem 4.2.9 (Lindeberg 中心极限定理)**:::>** 设独立随机变量序列 $\{X_n\}, X_n\sim f_{X_n}(x)$ 的数学期望和方差分别为 $EX_i=\mu_i$ 和 $DX_i=\sigma_i^2(i=$

$1,2,\cdots$), 记 $B_n{}^2=\sum_{i=1}^n \sigma_i{}^2$, 如果 $\{X_n\}$ 满足以下的 Lindeberg 条件: $\forall \tau>0$ 有

$$\lim_{n\to\infty}\frac{1}{\tau^2 B_n^2}\sum_{i=1}^n\int_{|x-\mu_i|>\tau B_n}(x-\mu_i)^2 f_{X_i}(x)dx=0$$

则对任意的 $x\in \boldsymbol{R}$ 有

$$\lim_{n\to\infty}P\left(\frac{1}{B_n}\sum_{i=1}^n(X_i-\mu_i)\leqslant x\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{z^2}{2}}dz$$

Example 4.2.10:::> 若 $X_i$iid, $EX_1=0, DX_1=1$, 则

$$\frac{X_1+X_2+\cdots+X_n}{\sqrt{X_1^2+X_2^2+\cdots+X_n^2}}\xrightarrow{D}N(0,1)$$

Proof. 由Levy- Lindeberg 中心极限定理可得,

$$\frac{\sum_{i=1}nX_i}{\sqrt{n}}\xrightarrow{D}N(0,1)$$

而由大数定律延伸的矩估计定理可得:

$$\frac{1}{n}\sum_{i=1}^n X_i^2\xrightarrow{P}E(X_i)^2=1$$

再由Slutsky 定理中除法形式即可得. ■

Example 4.2.11 (随机模拟算法 Monte Carlo 方法)**:::>** 随机模拟方法的一个应用是计算积分, 其基本思想是将积分看成某个随机变量的期望. 取一个在 $f(x)$ 非零处恒正的分布密度 $g(x)$, 那么积分 $\int_a^b f(x)\mathrm{d}x$ 可以写为 $\int_a^b \frac{f(x)}{g(x)}g(x)\mathrm{d}x$. 于是对于密度为 $g(x)$ 的随机数 (称为 "$g-$ 随机数") ζ, 我们有

$$\int_a^b f(x)\mathrm{d}x=\int_a^b\frac{f(x)}{g(x)}g(x)\mathrm{d}x=E\left[\frac{f(\zeta)}{g(\zeta)}\right]$$

我们取 N 个独立的 $g-$随机数 $\zeta_1,\cdots,\zeta_N$. 由大数定律, 当 $N\to\infty$ 时, $\frac{1}{N}\left[\frac{f(\zeta_1)}{g(\zeta_1)}+\cdots+\frac{f(\zeta_N)}{g(\zeta_N)}\right]$ 按概率收敛到积分 $\int_a^b f(x)\mathrm{d}x$, 从而可以作为此积分的概率估计.

Theorem 4.2.12:::> The central limit theorem tells us that $Z_n=\sqrt{n}\left(\overline{X}_n-\mu\right)/\sigma$ is approximately N$(0,1)$. However, we rarely know σ. Later, we will see that we can estimate σ^2 from $X_1,\ldots,X_n$ by

$$S_n^2=\frac{1}{n-1}\sum_{i=1}^n\left(X_i-\overline{X}_n\right)^2$$

Assume the same conditions as the CLT. Then

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \rightsquigarrow N(0,1)$$

Theorem 4.2.13 (The Berry-Essèen Inequality)**:::>** This inequality is used to mesure the accuracy of the normal approximation.

$$\sup_z |\mathbb{P}(Z_n \leqslant z) - \Phi(z)| \leqslant \frac{33}{4} \frac{\mathbb{E}|X_1 - \mu|^3}{\sqrt{n}\sigma^3}$$

Theorem 4.2.14 (Multivariate central limit theorem)**:::>** Let $X_1, \ldots, X_n$ be IID random vectors where

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix}$$

with mean

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_{1i}) \\ \mathbb{E}(X_{2i}) \\ \vdots \\ \mathbb{E}(X_{ki}) \end{pmatrix}$$

and variance matrix Σ. Let

$$\overline{X} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_k \end{pmatrix}$$

where $\overline{X}_j = n^{-1} \sum_{i=1}^n X_{ji}$. Then,

$$\sqrt{n}(\overline{X} - \mu) \rightsquigarrow N(0, \Sigma)$$

Theorem 4.2.15 (The Delta Method)**:::>** Suppose that

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$$

and that g is a differentiable function such that $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0,1)$$

In other words,

$$Y_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{implies that} \quad g(Y_n) \approx N\left(g(\mu), \left(g'(\mu)\right)^2 \frac{\sigma^2}{n}\right)$$

注意这里的 μ 只是一个常数, 并不代表随机变量 Y_n 的均值.

Theorem 4.2.16 (The Multivariate Delta Method)**:::>** Suppose that $Y_n = (Y_{n1}, \ldots, Y_{nk})$ is a sequence of random random vectors such that

$$\sqrt{n}(Y_n - \mu) \sim N(0, \Sigma)$$

Let $g : \mathbb{R}^k \to \mathbb{R}$ and let

$$\nabla(y) = \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix}$$

Let ∇_μ denote $\nabla g(y)$ evaluated at $y = \mu$ and assume that the elements of ∇_μ are nonzero. Then

$$\sqrt{n}(g(Y_n) - g(\mu)) \rightsquigarrow N\left(0, \nabla_\mu^T \Sigma \nabla_\mu\right)$$

5. Fundamental Concepts in Statistical Inference

5.1. 样本与总体

A typical statistical inference question is:

Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F ?

In some cases, we may want to infer only some feature of F such as its mean.

Definition 5.1.1 (样本与总体)**:::>** 统计的最重要观点是用样本推断总体.

总体指的是研究对象某方面的数据指标, 而样本指的是对总体抽样. 在实际研究中, 一般选用的样本是简单随机样本, 具有随机性与独立性.

Definition 5.1.2 (样本的二重性)**:::>** 样本具有所谓的二重性: 一方面, 由于样本是从总体中随机抽取的, 抽取前无法预知它们的数值, 因此, 样本是随机变量, 用大写字母 $X_1, X_2, \cdots, X_n$ 表示; 另一方面, 样本在抽取以后经观测就有确定的观测值, 因此, 样本又是一组数值. 此时用小写字母 $x_1, x_2, \cdots, x_n$ 表示是恰当的.

Definition 5.1.3 (样本的分布函数)**:::>** 总体为 $F(x;\theta)$ 的样本 (容量为 n)$X_1, X_2, \cdots, X_n$ 的观测值为 $x_1, \cdots, x_n$, 则此样本的分布函数为 $F(x_1, \cdots, x_n) = \prod_{i=1}^{n} F(x_i;\theta)$

Definition 5.1.4 (测量问题)**:::>** 一个测量者对一个物理量 μ 进行重复测量, 此时一切可能的测量结果是 $(-\infty, \infty)$, 因此总体是一个取值于 $(-\infty, \infty)$

的随机变量 X. 测量结果 X 可以看作物理量 μ 与测量误差 ε 的叠加, 即:

$$X = \mu + \varepsilon$$

这里 μ 是一个确定的但未知的量, 我们称之为参数, 于是关于总体分布的假定主要是关于 ε 的分布的假定. 如下几种假定各在一些场合是合理的.

1. 由中心极限定理, 最常见的是假定随机误差 $\varepsilon \sim N(0, \sigma^2)$, 于是测量值的总体就是一个正态分布, 即 $X \sim N(\mu, \sigma^2)$, 这里总体中有两个未知参数 μ, σ^2. 如何确定 μ 与 σ^2 是统计学要研究的问题.

2. 假如我们不仅知道误差服从正态分布, 还知道分布的方差 (这通常可由测量系统本身的精度决定), 于是, 就可假定 $\varepsilon \sim N(0, \sigma_0^2)$, 其中 σ_0 是一个已知的常数, 如此, 总体仍是一个正态分布族 $N(\mu, \sigma_0^2)$, 但总体中只有一个未知参数 μ. 如何确定 μ 是统计学要研究的问题.

3. 假如我们并没有理由认定误差服从正态分布, 但可以认为误差的分布是关于 0 对称的, 则总体分布就变为一个分布类型未知但带有某种限制的分布, 通常它不能被有限个参数所描述, 常称为非参数分布.

5.2. 统计量

Definition 5.2.1 (统计量)**:::>** 简单随机样本 $X_1, X_2, \cdots, X_n$ 的函数 $g(X_1, \ldots, X_n)$ 称为样本的统计量. 函数 $g(x)$ 不含任何未知参数, 是一个完全已知的函数. 必须指出的是: 尽管统计量不依赖于未知参数, 但是它的分布一般还是依赖于未知参数的.

常见的统计量:

1. 样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $E\overline{X} = \mu = EX$, $D(\overline{X}) = \frac{\sigma^2}{n}$

2. 样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, $ES^2 = \sigma^2 = D(X)$

3. 样本修正方差 $S^{*2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = M_2 - M_1^2, E(S^{*2}) = \frac{n-1}{n}\sigma^2$

Definition 5.2.2 (样本方差)**:::>** 在无偏方差 $s^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\overline{x})^2$ 的定义中, n 为样本量, $\sum_{i=1}^n(x_i-\overline{x})^2$ 称为偏差平方和, $n-1$ 称为偏差平方和的自由度. 其含义是: 在 $\overline{x}$ 确定后, n 个偏差 $x_1-\overline{x},x_2-\overline{x},\cdots,x_n-\overline{x}$ 中只有 $n-1$ 个偏差可以自由变动, 而第 n 个则不能自由取值, 因为 $\sum(x_i-\overline{x})=0$.

样本偏差平方和有三个不同的表达式:

$$\sum(x_i-\overline{x})^2=\sum x_i^2-\frac{(\sum x_i)^2}{n}=\sum x_i^2-n\overline{x}^2.$$

Theorem 5.2.3:::> 设总体 X 具有二阶矩, 即 $E(X)=\mu$,$\mathrm{Var}(X)=\sigma^2<\infty$, $x_1,x_2,\cdots,x_n$ 为从该总体得到的样本, $\overline{x}$ 和 s^2 分别是样本均值和样本方差, 则

$$E(\overline{x})=\mu,\quad \mathrm{Var}(\overline{x})=\sigma^2/n$$

$$E\left(s^2\right)=\sigma^2$$

Proof. 由于

$$\mathbb{E}(\overline{x})=\frac{1}{n}E\left(\sum_{i=1}^n x_i\right)=\frac{n\mu}{n}=\mu$$

$$\mathrm{Var}(\overline{x})=\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n x_i\right)=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

可证得关于样本均值的结果.

而注意到

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

而

$$Ex_i^2=(Ex_i)^2+\mathrm{Var}(x_i)=\mu^2+\sigma^2$$

$$E\left(\overline{x}^2\right)=(E\overline{x})^2+\mathrm{Var}(\overline{x})=\mu^2+\sigma^2/n$$

于是

$$E\left(\sum_{i=1}^n(x_i-\overline{x})^2\right)=n\left(\mu^2+\sigma^2\right)-n\left(\mu^2+\sigma^2/n\right)=(n-1)\sigma^2$$

两边各除 $n-1$, 可得结论. ■

Definition 5.2.4 (顺序统计量)**:::>** 把样本 $X_1,\cdots,X_n$ 按观测值从小到大的顺序将它们重新排列成 $X_{(1)}\leqslant X_{(2)}\leqslant\cdots\leqslant X_{(n)}$, 称统计量 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 为顺序统计量, 称

$$R=X_{(n)}-X_{(1)}\equiv\max\{X_1,\cdots,X_n\}-\min\{X_1,\cdots,X_n\}$$

为样本极差, 称

$$X_{\text{med}}=\begin{cases}X_{(\frac{n+1}{2})}, & n\text{ 为奇数}\\ \frac{1}{2}\left[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}\right], & n\text{ 为偶数}\end{cases}$$

为样本中位数.

Definition 5.2.5 (经验分布函数)**:::>** $F_n(x)=\frac{V_n(x)}{n}=\frac{1}{n}\{X_1,\cdots,X_n$ 中小于或等于 x 的个数 $\}$

$$=\begin{cases}0, & \text{当}x\leqslant X_{(1)},\\ \frac{k}{n}, & \text{当}X_{(k)}<x\leqslant X_{(k+1)},k=1,2,\cdots,n-1,\\ 1, & \text{当}X_{(n)}<x\end{cases}$$

由于 $V_n(x)$ 是 iid 的 $0-1$ 分布的和 (一个数只可能大于 x 或小于 x), 故 $V_n(x)\sim B(n,F(x))$, 从而

$$E(F_n(x))=F(x)$$

$$D(F_n(x))=\frac{F(x)(1-F(x))}{n}$$

由中心极限定理:

$$\sqrt{n}[F_n(x)-F(x)]\xrightarrow{D}N(0,F(x)(1-F(x)))$$

所以可得,

$$F_n(x)\xrightarrow{p}F(x)$$

故经验分布函数是原分布函数的一个非常好的近似.

更进一步地, 有 Glivenko 定理:

$$\sup_{x\in R}|F_n(x)-F(x)|\to 0.\quad \text{a.s.}$$

经验分布函数的大致思路是, 如果想知道某个随机变量 X 的分布 F, 一般情况下当然无法准确知道, 不过如果我们手上有它的一些独立同分布的样本, 则可以把这些样本的" 频率" 近似为随机变量的" 概率". 经典统计学中一切统计推断都以样为依据, 其理由就在于此.

5.3. 抽样分布

5.3.1 抽样分布

Definition 5.3.1 (抽样分布)**:::>** 抽样分布即为统计量的分布. 对于样本 $X_1, X_2, \cdots, X_n$ 来说, 其统计量 $g(X_1, \cdots, X_n)$ 的分布, 称为抽样分布.

Definition 5.3.2 (顺序统计量的分布)**:::>** 虽然每个样本随机变量是 iid 的, 但是顺序统计量 (即第 k 小之类的问题) 则是相关的. 假设总体的分布密度为 $f(x)$, 分布函数为 $F(x)$, 则

$$X_{(k)} \sim \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$$

这个公式可以从概率论的角度来理解, 其实是一个三项分布. 要求第 k 小的样本落在 $x, x+\delta x$ 的区间内, 可将 n 个样本分成三块, 前 $k-1$ 个样本落在对应 $F(x)$ 的区域, 后 $n-k$ 个样本落在对应 $1-F(x)$ 的区域, 而第 k 小的样本落在 $f(x)$ 的区域.

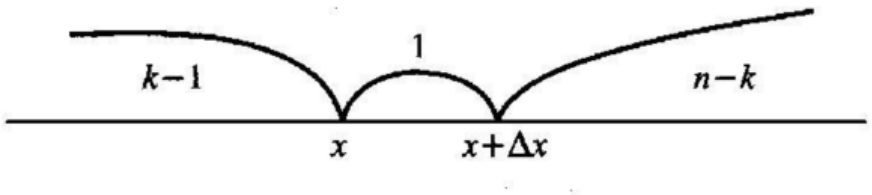


图 5.1

样本的每一个分量小于等于 x 的概率为 $F(x)$, 落人区间 $(x, x+\Delta x]$ 的概率为 $F(x+\Delta x)-F(x)$, 大于 $x+\Delta x$ 的概率为 $1-F(x+\Delta x)$, 而将 n 个分量

分成这样的三组, 总的分法有 $\frac{n!}{(k-1)!1!(n-k)!}$ 种. 于是, 若以 $F_k(x)$ 记 $x_{(k)}$ 的分布函数, 则由多项分布可得

$$F_k(x+\Delta x)-F_k(x)\approx \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}(F(x+\Delta x)-F(x))(1-F(x+\Delta x))^{n-k}$$

两边除以 Δx, 并令 $\Delta x\to 0$, 即有

$$\begin{aligned} p_k(x) &= \lim_{\Delta x\to 0}\frac{F_k(x+\Delta x)-F_k(x)}{\Delta x} \\ &= \frac{n!}{(k-1)!(n-k)!}(F(x))^{k-1}p(x)(1-F(x))^{n-k} \end{aligned}$$

Example 5.3.3:::> 设总体密度函数为

$$p(x)=3x^2,\quad 0<x<1$$

现从该总体抽得一个容量为 5 的样本, 试计算 $P\left(x_{(2)}<1/2\right)$

Solution. 我们首先应求出 $x_{(2)}$ 的分布.

由总体密度函数不难求出总体分布函数为

$$F(x)=\begin{cases} 0, & x\leqslant 0, \\ x^3, & 0<x<1, \\ 1, & x\geqslant 1. \end{cases}$$

则:

$$\begin{aligned} p_2(x) &= \frac{5!}{(2-1)!(5-2)!}(F(x))^{2-1}p(x)(1-F(x))^{5-2} \\ &= 20\cdot x^3\cdot 3x^2\cdot\left(1-x^3\right)^3 \\ &= 60x^5\left(1-x^3\right)^3,\quad 0<x<1, \end{aligned}$$

于是

$$\begin{aligned} P\left(x_{(2)}<1/2\right) &= \int_0^{1/2} 60x^5\left(1-x^3\right)^3\,\mathrm{d}x \\ &= \int_0^{1/8} 20y(1-y)^3\,\mathrm{d}y = \int_{7/8}^{1} 20\left(z^3-z^4\right)\mathrm{d}z \\ &= 5\left(1-\left(\frac{7}{8}\right)^4\right)-4\left(1-\left(\frac{7}{8}\right)^5\right)=0.1207 \end{aligned}$$

■

Definition 5.3.4 (多个顺序统计量的联合分布)**:::>**

$$\left(X_{(i)},X_{(j)}\right)\sim$$

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(x)]^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j}f(x)f(y),x\leqslant y$$

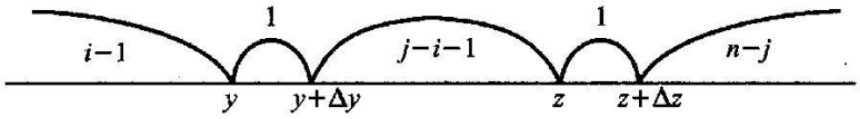


图 5.2

特别地, 有:

$$\left(X_{(1)},X_{(n)}\right)\sim n(n-1)[F(y)-F(x)]^{n-2}f(x)f(y),x\leqslant y$$

$$\left(X_{(1)},X_{(2)},\cdots,X_{(n)}\right)\sim$$

$$f_{X_{(1)},X_{(2)},\cdots,X_{(n)}}\left(x_{(1)},x_{(2)},\cdots,x_{(n)}\right)=n!\prod_{i=1}^{n}f\left(x_{(i)}\right),\forall x_{(1)}<x_{(2)}<\cdots<x_{(n)}$$

Example 5.3.5 (样本极差的分布密度)**:::>**

$$R=X_{(n)}-X_{(1)}$$

则

$$f_R(r)=\int_0^\infty f_{(X_{(1)},X_{(n)})}(u,r+u)du$$
$$=\int_0^\infty n(n-1)[F(r+u)-F(u)]^{n-2}f(u)f(r+u)du, r>0$$

Theorem 5.3.6:::> 设 $x_1,x_2,\cdots,x_n$ 是来自某个总体的样本, $\overline{x}$ 为样本均值

1. 若总体分布为 $N(\mu,\sigma^2)$, 则 $\overline{x}$ 的精确分布为 $N(\mu,\sigma^2/n)$

   Proof. 利用卷积公式, 可得知 $\sum_{i=1}^n x_i \sim N(n\mu,n\sigma^2)$, 由此可知 $\overline{x} \sim N(\mu,\sigma^2/n)$ ■

2. 若总体分布未知或不是正态分布, 但 $E(x)=\mu, \mathrm{Var}(x)=\sigma^2$, 则 n 较大时 $\overline{x}$ 的渐近分布为 $N(\mu,\sigma^2/n)$, 常记为 $\overline{x} \doteq N(\mu,\sigma^2/n)$. 这里渐近分布是指 n 较大时的近似分布.

   Proof. 由中心极限定理, $\sqrt{n}(\overline{x}-\mu)/\sigma \xrightarrow{L} N(0,1)$, 这表明 n 较大时 $\overline{x}$ 的渐近分布为 $N(\mu,\sigma^2/n)$, 证明完成. ■

Theorem 5.3.7:::> 设总体 X 具有二阶矩, 即 $E(X)=\mu, \mathrm{Var}(X)=\sigma^2<\infty, x_1,x_2,\cdots,x_n$ 为从该总体得到的样本, $\overline{x}$ 和 s^2 分别是样本均值和样本方差, 则

$$E(\overline{x})=\mu, \quad \mathrm{Var}(\overline{x})=\sigma^2/n$$
$$E\left(s^2\right)=\sigma^2$$

此定理表明, 样本均值的期望与总体均值相同, 而样本均值的方差是总体方差的 $1/n$.

Proof. 由于

$$E(\overline{x})=\frac{1}{n}E\left(\sum_{i=1}^n x_i\right)=\frac{n\mu}{n}=\mu$$
$$\mathrm{Var}(\overline{x})=\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n x_i\right)=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

故均值的性质成立.

又有:

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

而

$$Ex_i^2=(Ex_i)^2+\mathrm{Var}(x_i)=\mu^2+\sigma^2$$
$$E\left(\overline{x}^2\right)=(E\overline{x})^2+\mathrm{Var}(\overline{x})=\mu^2+\sigma^2/n$$

于是

$$E\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)=n\left(\mu^2+\sigma^2\right)-n\left(\mu^2+\sigma^2/n\right)=(n-1)\sigma^2$$

两边各除以 $n-1$ 可得. ■

Definition 5.3.8 (样本中位数的渐近分布)**:::>** 若总体的密度函数为 $f(x)$, 其中位数为 x_{med}, 样本中位数记为

$$X_{med}=\begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n\text{ 为奇数} \\ \frac{1}{2}\left[X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right], & n\text{ 为偶数}\end{cases}$$

则

$$X_{med}\sim N\left(x_{\mathrm{med}},\frac{1}{4n\cdot f^2(x_{\mathrm{med}})}\right)$$

样本 p 分位数

$$X_p=\begin{cases} X_{([np+1])}, & np\text{ 为非整数} \\ \frac{1}{2}\left[X_{(np)}+X_{(np+1)}\right], & np\text{ 为整数}\end{cases}$$

则

$$X_p\sim N\left(x_p,\frac{p(1-p)}{n\bullet f^2\left(x_p\right)}\right)$$

Example 5.3.9:::> 设总体为柯西分布, 密度函数为

$$p(x;\theta)=\frac{1}{\pi(1+(x-\theta))^2},\quad -\infty<x<\infty$$

其分布函数为

$$F(x;\theta)=\frac{1}{2}+\frac{1}{\pi}\arctan(x-\theta)$$

不难看出 θ 是该总体的中位数, 即 $x_{0.5}=\theta$. 设 $x_1,\cdots,x_n$ 是来自该总体的样本, 当样本量 n 较大时, 样本中位数 $m_{0.5}$ 的渐近分布为

$$m_{0.5}\div N\left(\theta,\frac{\pi^2}{4n}\right)$$

5.3.2　χ^2 分布、t 分布和 F 分布

Definition 5.3.10 (三大抽样分布)**:::>**　我们将 χ^2 分布、t 分布和 F 分布称为三大抽样分布, 这是统计学中除了正态分布外最重要的三个分布.

统计量的构造	抽样分布密度函数	期望	方差
$\chi^2=x_1^2+x_2^2+\cdots+x_n^2$	$p(y)=\frac{1}{\Gamma\left(\frac{n}{2}\right)2^{n/2}}y^{\frac{n}{2}-1}\mathrm{e}^{-\frac{y}{2}}(y>0)$	n	$2n$
$F=\frac{\left(y_1^2+\cdots+y_m^2\right)/m}{\left(x_1^2+\cdots+x_n^2\right)/n}$	$p(y)=\frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{m/2}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}y^{\frac{m}{2}-1}\left(1+\frac{m}{n}y\right)^{-\frac{m+n}{2}}$	$\frac{n}{n-2}(n>2)$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}(n>4)$
$t=\frac{y_1}{\sqrt{\left(x_1^2+\cdots+x_n^2\right)/n}}$	$p(y)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}$	$0(n>1)$	$\frac{n}{n-2}(n>2)$

Definition 5.3.11 (χ^2 分布)**:::>**

$$U\sim\chi^2(n)\Leftrightarrow U=\sum_{i=1}^{n}X_i^2$$

其中

$$X_i(i=1,2,\cdots,n)\ \text{i.i.d.}\ \sim N(0,1)$$

由Gamma distribution中的结论, 我们有:

$$X\sim N(0,1)\to X^2\sim\Gamma(\frac{1}{2},\frac{1}{2})$$

则我们有:

$$U\sim\chi^2(n)=\Gamma\left(\frac{n}{2},\frac{1}{2}\right)$$

且可得 $\chi^2(n)$ 分布的密度函数为:

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

其密度函数图像如下:

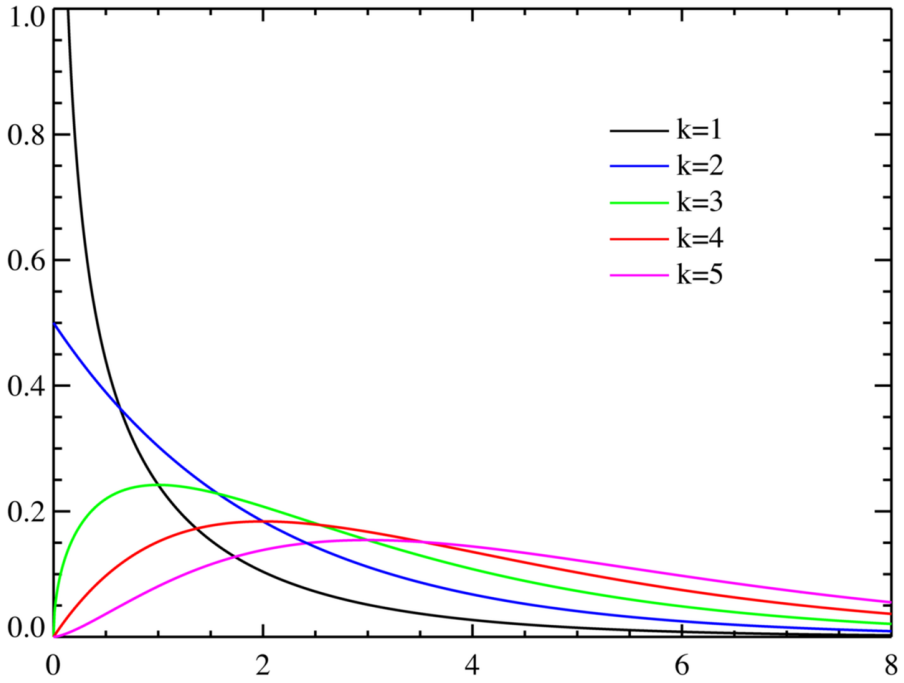


图 5.3: χ^2 分布的密度函数

期望方差均与自由度有关:

$$E(U)=n, D(U)=2n$$

Theorem 5.3.12 (χ^2 分布具有可加性)**:::>** 若 $X\sim\chi^2(n)$, $Y\sim\chi^2(m)$, 则 $X+Y\sim\chi^2(m+n)$.

χ^2 分布具有可加性的根源是 Γ 分布具有可加性.

Definition 5.3.13 (t 分布)**:::>**

$$T\sim t(n)\Leftrightarrow T=\frac{X}{\sqrt{\frac{Y}{n}}}$$

其中

$$X\sim N(0,1), Y\sim\chi^2(n), X, Y \text{ i.d}$$

$$f_T(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, -\infty<x<+\infty$$

t 分布的密度函数图像如下:

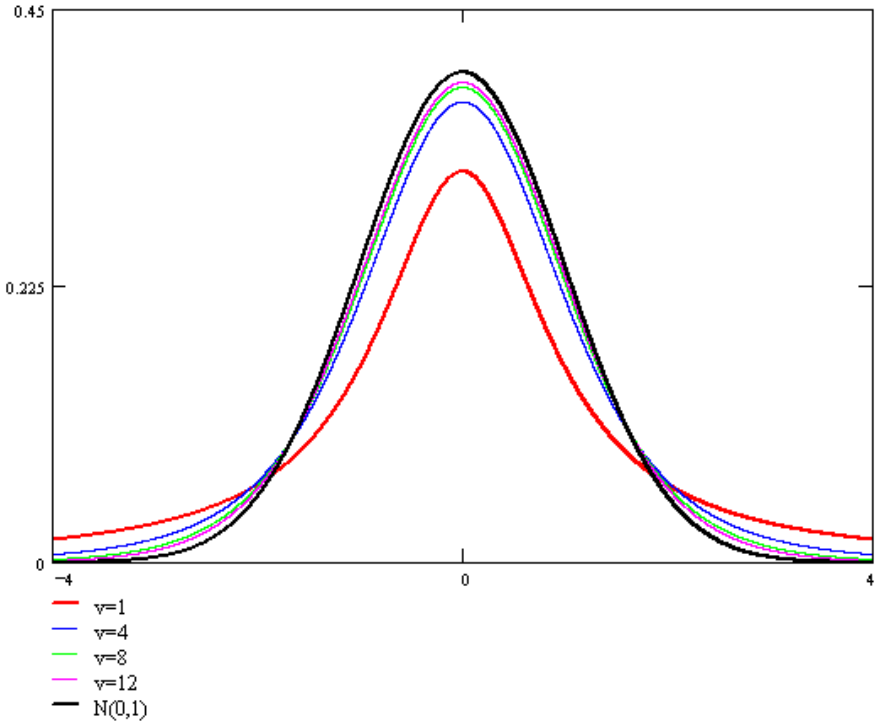


图 5.4: t 分布的密度函数

Theorem 5.3.14 (t 分布的性质)**:::>**

1. $n=1,\quad T\sim t(1)=C(1,0)$ 是一个标准 Cauchy 分布 t and Cauchy Distribution, 期望不存在.
2. $n>1,\quad ET=0$
3. $n>2,\quad DT=\frac{n}{n-2}$
4. $n\to\infty$ 时, 近似分布为标准正态分布 $N(0,1)$

Definition 5.3.15 (F 分布)**:::>**

$$F\sim F(m,n)\Leftrightarrow F=\frac{\frac{X}{m}}{\frac{Y}{n}}$$

其中

$$X \sim \chi^2(m), Y \sim \chi^2(n), X, Yi.d$$

有结论:

$$F \sim F(m,n) \Leftrightarrow \frac{1}{F} \sim F(n,m)$$

密度函数:

$$f_F(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} x^{\frac{m}{2}-1}\left(1+\frac{m}{n}x\right)^{-\frac{m+n}{2}}, x>0$$

其密度函数图像如下:

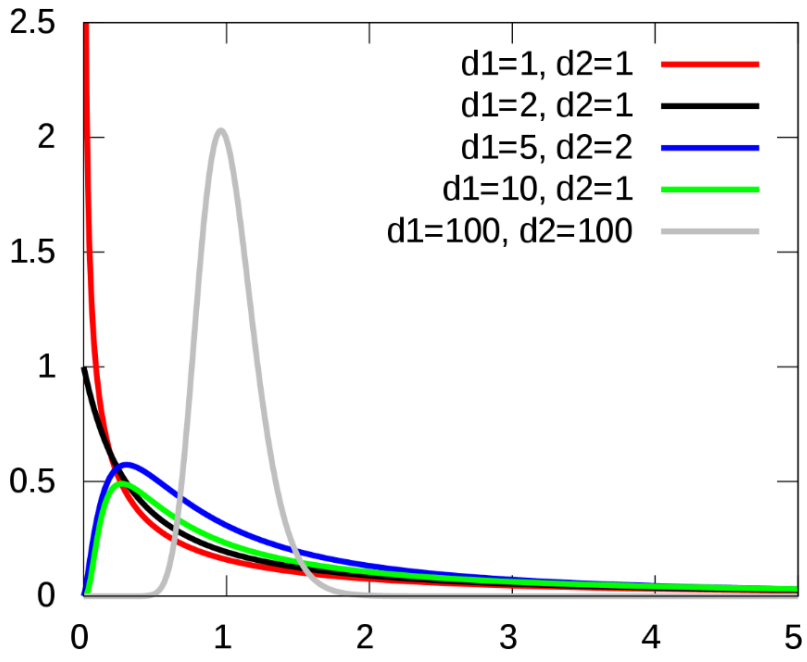


图 5.5: F 分布的密度函数

Example 5.3.16:::> 若 $X \sim F(m,n)$, 问随机变量 $Y = \frac{1}{1+\frac{m}{n}X}$ 服从的分布.

Solution. 由于 $X \sim F(m,n)$, 则

$$X = \frac{\frac{U}{m}}{\frac{V}{n}} = \frac{n}{m}\frac{U}{V}$$

则

$$Y = \frac{1}{1+\frac{m}{n}X} = \frac{V}{U+V}$$

由于

$$U \sim \chi^2(m), V \sim \chi^2(n)$$

因此

$$U \sim \Gamma(\frac{m}{2}, \frac{1}{2}), V \sim \Gamma(\frac{n}{2}, \frac{1}{2})$$

由Gamma 分布与 Beta 分布关系中的结果, 我们有

$$Y \sim \text{Beta}(\frac{n}{2}, \frac{m}{2})$$

■

Example 5.3.17:::> 若 $T \sim t(n)$, 问 T^2 服从的分布.

Solution. 由于 $T \sim t(n)$, 即

$$T \sim t(n) \Leftrightarrow T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

其中

$$X \sim N(0,1), Y \sim \chi^2(n), X, Y \text{ i.d}$$

则我们有

$$X^2 \sim \chi^2(1), Y \sim \chi^2(n), X^2, Y \text{ i.d}$$

则

$$T^2 \sim \frac{\frac{X^2}{1}}{\frac{Y}{n}} \sim F(1,n)$$

■

5.3.3 正态总体样本均值方差的分布

Theorem 5.3.18 (样本均值的分布)**:::>** 若 $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$, 则

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

从而

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

进而

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Theorem 5.3.19:::>

$$\chi^2 \equiv \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

且样本均值 $\overline{X}$ 与样本方差 S^2 相互独立.

这个结论非常重要, 一般来说, 样本均值 $\overline{X}$ 与样本方差 S^2 不是相互独立的, 也只有正态分布中这两者相互独立.

Proof. 证明需要用到下述两引理: 5.3.20, 5.3.21.

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}+\overline{X}-\mu}{\sigma}\right)^2$$

将后式用平方和公式展开.

又由于 $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$ 与 $\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1)$ 相互独立, 则原式等于:

$$n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 + \frac{(n-1)S^2}{\sigma^2}$$

由于两个独立的 χ^2 分布之差（前者的自由度大于后者）仍为 χ^2 分布, 故

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

■

Lemma 5.3.20:::> $X_1, X_2, \cdots, X_n$ 是来自总体 $X \sim N(\mu, \sigma^2)$ 的简单随机样本, 则 $\overline{X}$ 与 $\left(X_1 - \overline{X}, X_2 - \overline{X}, \cdots, X_n - \overline{X}\right)$ 相互独立.

Proof. 由于 $n+1$ 维向量 $\left(\overline{X}, X_1 - \overline{X}, X_2 - \overline{X}, \cdots, X_n - \overline{X}\right)^T$ 是正态分布的样本的线性组合, 故其服从 Gauss 分布

$$N\left(\begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

其中 $\mu^{(1)} = \mu, \mu^{(2)} = (0, 0, \cdots, 0)^T$

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 & 0 & \cdots & 0 \\ 0 & \frac{(n-1)\sigma^2}{n} & -\frac{\sigma^2}{n} & \cdots & -\frac{\sigma^2}{n} \\ 0 & -\frac{\sigma^2}{n} & \frac{(n-1)\sigma^2}{n} & \cdots & -\frac{\sigma^2}{n} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & -\frac{\sigma^2}{n} & \cdots & -\frac{\sigma^2}{n} & \frac{(n-1)\sigma^2}{n} \end{pmatrix}$$

为分块对角矩阵.

其中

$$\mathrm{Cov}\left(\overline{X}, X_i - \overline{X}\right) = \mathrm{Cov}\left(\overline{X}, X_i\right) - D\overline{X} = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0, i = 1, 2, \cdots, n$$

$$\begin{aligned} \mathrm{Cov}\left(X_i - \overline{X}, X_j - \overline{X}\right) &= \mathrm{Cov}\left(X_i, X_j\right) - \mathrm{Cov}\left(X_i, \overline{X}\right) - \mathrm{Cov}\left(\overline{X}, X_i\right) + D\overline{X} \\ &= \begin{cases} \frac{(n-1)\sigma^2}{n}, & i = j \\ -\frac{\sigma^2}{n}, & i \neq j \end{cases} \end{aligned}$$

由分块对角矩阵, 从而可知 $\overline{X}$ 和 S^2 相互独立. ■

Lemma 5.3.21:::> 若 X, Y i.d., $X \sim \chi^2(m), X + Y \sim \chi^2(m+n)$, 则 $Y \sim \chi^2(n)$

Proof. 由 χ^2 分布的特征函数, 得:

$$\varphi_X(\theta) = \left(\frac{\frac{1}{2}}{\frac{1}{2} - i\theta}\right)^{\frac{m}{2}}$$

$$\varphi_{X+Y}(\theta)=\left(\frac{\frac{1}{2}}{\frac{1}{2}-i\theta}\right)^{\frac{m+n}{2}}$$

由于

$$\varphi_{X+Y}(\theta)=\varphi_X(\theta)\varphi_Y(\theta)$$

故

$$\varphi_Y(\theta)=\left(\frac{\frac{1}{2}}{\frac{1}{2}-i\theta}\right)^{\frac{n}{2}}$$

■

Theorem 5.3.22 (双正态总体)**:::>** 若从 $X\sim N(\mu_1,\sigma_1^2)$ 中产生样本 $X_1,X_2,\cdots,X_n$, 从 $Y\sim N(\mu_2,\sigma_2^2)$ 中产生样本 $Y_1,Y_2,\cdots,Y_n$, 且样本 $X_1,X_2,\cdots,X_n$ 与 $Y_1,Y_2,\cdots,Y_n$ 相互独立, 则我们有:

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

2. 若 σ_1 和 σ_2 未知, 则
$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$
分布未知.

   但有一种特殊情况, 若 σ_1 和 σ_2 未知, 但 $\sigma_1^2=\sigma_2^2$ 时, 我们定义样本混合方差:
$$S_w^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$
   此时我们有:
$$T_{n_1,n_2}\equiv\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

$\mathfrak{Proof.}$ 由于

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$$

$$\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

在 $\sigma^2 \triangleq \sigma_1^2 = \sigma_2^2$ 时,

$$V = \frac{1}{\sigma^2}\left[(n_1-1)S_1^2 + (n_2-1)S_2^2\right] \sim \chi^2(n_1+n_2-2)$$

且

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

易知 U, V 独立, 则

$$T_{n_1,n_2} = \frac{U}{\sqrt{\frac{V}{n_1+n_2+-2}}} \sim t\,(n_1+n_2-2)$$

■

3.

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

$\mathfrak{Proof.}$ 由两样本独立可知, s_x^2 与 s_y^2 相互独立. 由有:

$$\frac{(n_1-1)s_x^2}{\sigma_1^2} \sim \chi^2(n_1-1), \quad \frac{(n_2-1)s_y^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

由 F 分布定义可知 $F_{n_1,n_2} \sim F(n_1-1, n_2-1)$ ■

$\mathfrak{Theorem}$ **5.3.23:::>** 若 X 的均值为 μ 且方差为 σ^2(注意这里并不一定要求是正态样本), 则下述等式永远成立

1. 样本均值是总体均值的无偏估计

$$E(\overline{X}) = \mu$$

2.

$$D(\overline{X}) = \frac{\sigma^2}{n}$$

3. 样本方差是总体方差的无偏估计:

$$E(S^2)=\sigma^2$$

Example 5.3.24:::> 设总体 $X\sim N(\mu,\sigma^2)$, 则

$$E\left(S^2\right)=\sigma^2$$

$$D\left(S^2\right)=\frac{2\sigma^4}{n-1}$$

Proof. 第二项的证明不能直接计算, 要利用

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

则我们有

$$D\left(\frac{(n-1)S^2}{\sigma^2}\right)=2(n-1)$$

则可得. ■

Example 5.3.25:::> 设 $X_1,X_2,\cdots,X_{20}$ 是来自总体 $X\sim N(0,\sigma^2)$ 的简单随机样本, 则

$$\frac{\sum_{i=1}^{10}(-1)^iX_i}{\sqrt{\sum_{i=11}^{20}X_i^2}}\sim t(10)$$

Proof.

$$\sum_{i=1}^{10}(-1)^iX_i\sim N(0,10\sigma^2)$$

标准化:

$$U=\frac{\sum_{i=1}^{10}(-1)^iX_i}{\sqrt{10}\sigma}\sim N(0,1)$$

而

$$\frac{X_i}{\sigma}\sim N(0,1)$$

$$V=\sum_{i=11}^{20}\left(\frac{X_i}{\sigma}\right)^2\sim\chi^2(10)$$

我们有 U,V 独立, 且

$$\frac{U}{\sqrt{\frac{V}{10}}}\sim t(10)$$

■

Theorem 5.3.26:::> 设 $x_1, x_2, \cdots, x_n$ 是来自正态分布 $N(\mu, \sigma^2)$ 的一个样本, $\overline{x}$ 与 s^2 分别是该样本的样本均值与样本方差, 则有

$$t = \frac{\sqrt{n}(\overline{x} - \mu)}{s} \sim t(n-1)$$

事实上, 这正是 t 分布发现时最重要的定理.

Proof. 由于

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

将结论左端改写为

$$\frac{\sqrt{n}(\overline{x} - \mu)}{s} = \frac{\frac{\overline{x} - \mu}{\sigma / \sqrt{n}}}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}}$$

由于分子是标准正态变量, 分母的根号里是自由度为 $n-1$ 的 χ^2 变量除以它的自由度, 且分子与分母相互独立, 由 t 分布定义可知 $t \sim t(n-1)$. ■

5.4. 充分统计量

Definition 5.4.1 (充分统计量)**:::>** 样本中包含的关于总体的信息可分为两部分: 其一是关于总体结构的信息, 即反映总体分布的结构; 其二是关于总体中未知参数的信息. 统计量具有压缩数据功能, 一个好的统计量应该能将样本中包含未知参数的全部信息提取出来, 这种不损失未知参数的信息的统计量就是我们要介绍的充分统计量.

充分统计量的概念是 Fisher 在 1922 年提出的, 其思想源于他与天文学家 Eddington 有关估计标准差的争论中. 设 $x_1, x_2, \cdots, x_n$ 为来自 $N(\mu, \sigma^2)$ 的独立同分布样本, 现要估计 σ. Fisher 主张用样本标准差 s, 而 Eddington 则主张用如下的平均绝对偏差:

$$d = \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|$$

Fisher 认为: 在 s 中包含了样本中有关 σ 的全部信息, 而 d 则否, 这就是充分统计量的思想.

设 $X_1, X_2, \cdots, X_n$ 是总体 $\boldsymbol{X}$ 的简单样本, 总体分布函数为 $F(x,\theta)$, 称统计量 $T = T(X_1, X_2, \cdots, X_n)$ 为 θ 的充分统计量, 如果在给定 T 的取值后, $X_1, X_2, \cdots, X_n$ 的条件分布与 θ 无关.

用离散的情形叙述, 即 $P(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$ 与 θ 无关.

事实上, " 无关" 即说明条件 “$T = t$ ” 的出现使得从样本分布 $F_\theta(X)$ 到条件分布 $F_\theta(X \mid T = t)$ 有关 θ 的信息消失了, 这说明有关 θ 的信息都含在统计量 T 之中.

Theorem 5.4.2 (充分统计量的充分定理)**:::>** 若样本 $X_1, X_2, \cdots, X_n$ 的联合分布密度 (或 p.m.f) 为 $f(x_1, x_2, \ldots, x_n \mid \theta)$, 统计量 $T = T(X_1, X_2, \ldots, X_n)$ 的分布密度 (或 p.m.f) 为 $q(t \mid \theta)$, 则若 $\frac{f(x_1, x_2, \ldots, x_n \mid \theta)}{q(T(x_1, x_2, \ldots, x_n) \mid \theta)}$ 与 θ 无关, 则 $T = T(X_1, X_2, \ldots, X_n)$ 为充分统计量.

Example 5.4.3 (离散情形)**:::>** 设 $X_1, X_2, \cdots, X_n$ 是总体 $X \sim B(1, \theta), 0 < \theta < 1$(即 Bernoulli 分布) 的一个样本, 证明 $T = \sum_{i=1}^n X_i$ 为 θ 的充分统计量.

Proof. 这里可以直接用定义证明, 即只需证明 $P(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$ 与 θ 无关, 这里 $x_i = 0, 1$, $t = \sum_{i=1}^n x_i$.

由于 $T = \sum_{i=1}^n X_i \sim B(n, \theta)$,

$$
\begin{aligned}
P(X_1 = x_1, \ldots, X_n = x_n \mid T = t) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)} \\
&= \frac{P\left(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n-1} x_i\right)}{P(T = t)} \\
&= \frac{\prod_{i=1}^{n-1} [\theta^{x_i}(1-\theta)^{1-x_i}] \cdot \theta^{t - \sum_{i=1}^{n-1} x_i}(1-\theta)^{1-\left(t - \sum_{i=1}^{n-1} x_i\right)}}{C_n^t \theta^t (1-\theta)^{n-t}} = \frac{1}{C_n^t}
\end{aligned}
$$

■

Example 5.4.4 (连续情形)**:::>** 设 $X_1, X_2, \cdots, X_n$ 是总体 $X \sim N(\mu, 1)$ 的一个样本, 证明 $T = \overline{X}$ 为 μ 的充分统计量

Proof. 这里利用上述充分性定理证明.

$X_1, X_2, \cdots, X_n$ 的联合分布密度为

$$\begin{aligned} & f(x_1, x_2, \ldots, x_n \mid \mu) \\ &= \prod_{i=1}^{n} f(x_i; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}} \end{aligned}$$

而 $T = \overline{X} \sim N\left(\mu, \frac{1}{n}\right)$ 的分布密度为 $q(t \mid \mu) = \frac{1}{\sqrt{\frac{2\pi}{n}}} e^{-\frac{(t-\mu)^2}{2\cdot\frac{1}{n}}}$, 故

$$\begin{aligned} & \frac{f(x_1, x_2, \ldots, x_n \mid \mu)}{q(T(x_1, x_2, \ldots, x_n) \mid \mu)} \\ &= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}}}{\sqrt{\frac{2\pi}{n}} e^{-\frac{(\overline{x}-\mu)^2}{2\cdot\frac{1}{n}}}} \\ &= n^{-\frac{1}{2}} (2\pi)^{-\frac{n-1}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{2}} \end{aligned}$$

与 μ 无关, 故 $T = \overline{X}$ 为 μ 的充分统计量.

■

Example 5.4.5:::> 设 X_1, X_2 是总体 $X \sim P(\lambda)$ 的一个样本, 证明 $T_1 = X_1 + X_2$ 为 λ 的充分统计量, 而 $T_2 = X_1 + 2X_2$ 不是 λ 的充分统计量.

Proof.

$$\begin{aligned} & P(X_1 = x_1, X_2 = x_2 \mid T_1 = t) \\ &= \begin{cases} \frac{P(X_1=x_1, X_2=t-x_1)}{P(T_1=t)} = \frac{\frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \frac{\lambda^{t-x_1} e^{-\lambda}}{(t-x_1)!}}{\frac{(2\lambda)^t e^{-2\lambda}}{t!}} = \frac{C_t^{x_1}}{2^t}, & t = x_1 + x_2, x_1, x_2 = 0, 1, 2, \cdots \\ 0, & otherwise \end{cases} \end{aligned}$$

与 λ 无关, 故 $T_1 = X_1 + X_2$ 为 λ 的充分统计量.

$$
\begin{aligned}
&P(X_1=0,X_2=1\mid T_2=2)=P(X_1=0,X_2=1\mid X_1+2X_2=2)\\
&=\frac{P(X_1=0,X_2=1)}{P(X_1+2X_2=2)}\\
&=\frac{P(X_1=0,X_2=1)}{P(X_1=0,X_2=1)+P(X_1=2,X_2=0)}\\
&=\frac{e^{-\lambda}\lambda e^{-\lambda}}{e^{-\lambda}\lambda e^{-\lambda}+\frac{\lambda^2e^{-\lambda}}{2!}e^{-\lambda}}=\frac{2}{2+\lambda}
\end{aligned}
$$

依赖于 λ, 故 $T_2=X_1+2X_2$ 不是 λ 的充分统计量. ■

Theorem 5.4.6 (因子分解定理)**:::>** 因子分解定理为我们寻找充分统计量指明了方向.

总体概率函数 $f(x,\theta)$, 样本 $(X_1,\ldots,X_n)$, 统计量 $T(X_1,\ldots,X_n)$ 为未知参数 θ 的充分统计量的充要条件是样本在已发生情形 $(X_1,\ldots,X_n)\to(x_1,\ldots,x_n)$ 处的联合概率函数可以分解为 $T(x_1,\ldots,x_n)=t$ 和 θ 的函数 $g(t,\theta)$ 与样本观察值的函数 $h(x_1,\ldots,x_n)$ 的乘积.

即 T 是 θ 的充分统计量 $\Leftrightarrow$

$$
L(\theta)=\prod_{i=1}^{n}f(x_i;\theta)=h(x_1,x_2,\ldots,x_n)\,g\left(T(x_1,x_2,\ldots,x_n);\theta\right)\text{且h非负}
$$

Proof. 只证明离散情形:

1. 必要性: $P(X_1=x_1,\ldots,X_n=x_n\mid T=t)\overset{T(x_1,\ldots,x_n)=t}{=}h(x_1,\ldots,x_n)$ 与 θ 无关

   由于 $\{T=t\}\supset\{X_1=x_1,\cdots,X_n=x_n\}$, 也即 $P(X_1=x_1,\ldots,X_n=x_n,T=t)=P(X_1=x_1,\ldots,X_n=x_n)$,(这个式子在下边两个方向上的证明都有应用.)
   因此

$$
\begin{aligned}
P(X_1=x_1,\ldots,X_n=x_n)&=P(X_1=x_1,\ldots,X_n=x_n\mid T=t)P(T=t)\\
&=h(x_1,\ldots,x_n)P(T=t)=g(t,\theta)h(x_1,\ldots,x_n)
\end{aligned}
$$

2. 充分性: 现在已知 $P(X_1=x_1,\ldots,X_n=x_n)=g(t,\theta)h(x_1,\ldots,x_n)$

$$P(T=t)=\sum_{T(x_1,\cdots,x_n)=t}P(X_1=x_1,\ldots,X_n=x_n)$$
$$=\sum_{T(x_1,\cdots,x_n)=t}g(t,\theta)h(x_1,\ldots,x_n)=g(t,\theta)\sum_{T(x_1,\cdots,x_n)=t}h(x_1,\ldots,x_n)$$

因此

$$P(X_1=x_1,\ldots,X_n=x_n\mid T=t)\overset{x_1+\ldots+x_n=t}{=}\frac{P(X_1=x_1,\ldots,X_n=x_n,T=t)}{P(T=t)}$$
$$=\frac{P(X_1=x_1,\ldots,X_n=x_n,T=t)}{g(t,\theta)\sum_{T(x_1,\ldots,x_n)=t}h(x_1,\ldots,x_n)}=\frac{P(X_1=x_1,\ldots,X_n=x_n)}{g(t,\theta)\sum_{T(x_1,\ldots,x_n)=t}h(x_1,\ldots,x_n)}$$
$$=\frac{g(t,\theta)h(x_1,\ldots,x_n)}{g(t,\theta)\sum_{T(x_1,\ldots,x_n)=t}h(x_1,\ldots,x_n)}=\frac{h(x_1,\ldots,x_n)}{\sum_{T(x_1,\ldots,x_n)=t}h(x_1,\ldots,x_n)}\text{ 与}\theta\text{ 无关}$$

■

Example 5.4.7:::> 设 $X_1,X_2,\cdots,X_n$ 是总体 $X\sim B(1,\theta),0<\theta<1$(即 Bernoulli 分布), 证明 $T=\sum_{i=1}^n X_i$ 为 θ 的充分统计量

Proof. 用因子分解定理证明.

样本的分布为:

$$P(X_1=x_1,\ldots,X_n=x_n)=\prod_{i=1}^n\left[\theta^{x_i}(1-\theta)^{1-x_i}\right]=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

取

$$h(x_1,x_2,\ldots,x_n)=1$$
$$g(T(x_1,x_2,\ldots,x_n);\theta)=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

故 $T=\sum_{i=1}^n X_i$ 为 θ 的充分统计量. ■

Corollary 5.4.8:::> 设统计量 T 为 θ 的充分统计量, 统计量 S 与 T 一一对应, 则 S 也为 θ 的充分统计量.

Example 5.4.9:::> 设 $X_1,X_2,\cdots,X_n$ 是总体 $X\sim N(\mu,\sigma^2)$ 的一个样本, μ,σ^2 均未知, 求 (μ,σ^2) 的充分统计量.

Solution. 样本的分布密度为

$$\begin{aligned} f_{\mu,\sigma^2}(x_1,x_2,\ldots,x_n) &= \prod_{i=1}^{n} f\left(x_i;\mu,\sigma^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}+\frac{2\mu\sum_{i=1}^n x_i}{2\sigma^2}-\frac{n\mu^2}{2\sigma^2}} \end{aligned}$$

故 $T = T(X_1,X_2,\ldots,X_n) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ 为 (μ,σ^2) 的充分统计量. ■

Remark. 显然 $T_1 = \left(\overline{X}, S^2\right)$ 为 (μ,σ^2) 的充分统计量, 这个一般更常用. 若 σ^2 未知, $\overline{X}$ 不是 μ 的充分统计量; μ 未知时, S^2 也不是 σ^2 的充分统计量. 但若 σ^2 已知, $\overline{X}$ 是 μ 的充分统计量, $\mu = \mu_0$ 已知, $\sum_{i=1}^n (X_i-\mu_0)^2$ 是 σ^2 的充分统计量. ■

Example 5.4.10:::> 设 $X_1,X_2,\cdots,X_n$ 是总体 $X \sim U\left(-\frac{\theta}{2},\frac{\theta}{2}\right)$ 的一个样本, 求参数 θ 的充分统计量.

Solution. 样本的分布密度为

$$\begin{aligned} f(x_1,x_2,\ldots,x_n) &= \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n} 1_{\left\{-\frac{\theta}{2}<x_1,\cdots,x_n<\frac{\theta}{2}\right\}} \\ &= \frac{1}{\theta^n} 1_{\left\{-\frac{\theta}{2}<x_{(1)}\leqslant x_{(n)}<\frac{\theta}{2}\right\}} \end{aligned}$$

故 $\left(X_{(1)},X_{(n)}\right)$ 为 θ 的充分统计量.

当然全样本 $(X_1,X_2,\ldots,X_n)$ 与顺序统计量 $\left(X_{(1)},X_{(2)},\ldots,X_{(n)}\right)$ 也是 θ 的充分统计量. ■

Example 5.4.11 (Pareto distribution)**:::>** 若 $X \sim f(x;\theta) = \theta \cdot \frac{1}{x^{\theta+1}} (x>1)$, 我们称 X 服从 Pareto distribution, 这是一个在金融与统计等领域应用广泛的分布, 可用于研究二八定律等. 我们研究它的充分统计量:

Solution.

$$\begin{aligned} f(x_1\cdot x_n;\theta) &= \prod_{i=1}^{n}\left(\theta \frac{1}{x_i^{\theta+1}} 1_{\{x_i>1\}}\right) \\ &= \theta^n \mathrm{e}^{-(\theta+1)\sum_{i=1}^n \ln x_i} 1_{\{x_i>1\}} \end{aligned}$$

则 $T=\sum_{i=1}^{n}\ln X_i$ 即为一个充分统计量. ■

$\mathfrak{Remark}$. 我们有结论, 若 X 服从 Pareto distribution, 则 $Y=\ln X$ 服从参量为 θ 的指数分布, $\mathfrak{id}\ \mathfrak{est} Y=\ln X\sim E(\theta)\sim\Gamma(1,\theta)$. 证明借助随机变量的函数的分布中的方法即可.

因而

$$T\sim\Gamma(n,\theta)$$

■

6. 参数估计

6.1. Parametric and Nonparametric Models

Definition 6.1.1 (Statistical Model)**:::>** A statistical model $\mathfrak{F}$ is a set of distributions (or densities or regression functions). A parametric model is a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad \mu \in \mathbb{R}, \sigma > 0 \right\}$$

This is a two-parameter model. We have written the density as $f(x;\mu,\sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters.

In general, a parametric model takes the form

$$\mathfrak{F} = \{ f(x;\theta) : \theta \in \Theta \}$$

where θ is an unknown parameter (or vector of parameters) that can take values in the parameter space Θ. If θ is a vector but we are only interested in one component of θ, we call the remaining parameters nuisance parameters. A nonparametric model is a set $\mathfrak{F}$ that cannot be parameterized by a finite number of parameters. For example, $\mathfrak{F}_{\text{ALL}} = \{$ all CDF' $s\}$ is nonparametric. The distinction between parametric and nonparametric is more subtle than this but we don't need a rigorous definition for our purposes.

Example 6.1.2 (One-dimensional Parametric Estimation)**:::>** Let $X_1, \ldots, X_n$ be independent Bernoulli (p) observations. The problem is to estimate the parameter p.

Example 6.1.3 (Two-dimensional Parametric Estimation)**:::>** Suppose that $X_1,\ldots X_n \sim F$ and we assume that the PDF $f \in \mathfrak{F}$ where $\mathfrak{F}$ is given in

$$\mathfrak{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \mu \in \mathbb{R}, \sigma > 0 \right\}$$

In this case there are two parameters, μ and σ. The goal is to estimate the parameters from the data. If we are only interested in estimating μ, then μ is the parameter of interest and σ is a nuisance parameter.

Example 6.1.4 (Nonparametric estimation of the cdf)**:::>** Let $X_1,\ldots,X_n$ be independent observations from a CDF F. The problem is to estimate F assuming only that $F \in \mathfrak{F}_{\text{ALL}} = \{ \text{ all CDF}'s \}$

Example 6.1.5 (Nonparametric estimation of the CDF)**:::>** Let $X_1,\ldots,X_n$ be independent observations from a CDF F and let $f = F'$ be the PDF. Suppose we want to estimate the PDF f. It is not possible to estimate f assuming only that $F \in \mathfrak{F}_{\text{ALL}}$. We need to assume some smoothness on f. For example, we might assume that $f \in \mathfrak{F} = \mathfrak{F}_{\text{DENS}} \cap \mathfrak{F}_{\text{SOB}}$ where $\mathfrak{F}_{\text{DENS}}$ is the set of all probability density functions and

$$\mathfrak{F}_{\text{SOB}} = \left\{ f : \int \left(f''(x)\right)^2 dx < \infty \right\}$$

The class $\mathfrak{F}_{SOB}$ is called a Sobolev space; it is the set of functions that are not "too wiggly."

6.2. 参数的点估计

Definition 6.2.1 (点估计)**:::>** 若总体服从 $X \sim F(x;\theta)$, 其中 F 为形式已知的分布, θ 为待估参数, 点估计即为把估计看成 n 维空间里的一个点, 即用样本 $X_1,\cdots,X_n$ 等来估计 θ. 常用的点估计方法有两种, 分别为矩估计与极大似然估计.

点估计可以与区间估计形成对比: 这种区间估计通常是在频率论推断的情况下的置信区间, 或在贝叶斯推断的情况下的可信区间.

6.2.1 矩估计

Definition 6.2.2 (矩估计)**:::>** 矩估计的思想是用样本的矩作为总体矩的估计.

其原理为Khintchine 大数定理的推论: 若总体 k 阶矩存在, 则

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mathbb{E}X^k = \mu_k, k \geqslant 1$$

其方法为:

1. 求出总体的 k 阶原点矩: $a_k = \mathbb{E}X^k = \int_{-\infty}^{+\infty} x^k \,\mathrm{d}F(x;\theta_1,\theta_2,\ldots,\theta_m)$
2. 解方程组 $a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k (k = 1,2,...,m)$, 得 $\theta_k = \theta_k(X_1,X_2,\ldots,X_n)$ 即为所求

Example 6.2.3:::> 随机变量 $X \sim P(\lambda)$, 其中 $\lambda > 0$ 是一个未知参量. 求问 λ 的矩估计 $\widehat{\lambda}$.

Solution. 由于只有一个参数, 所以我们只需要求出总体的一阶矩即可.

由于

$$\mathbb{E}X = \lambda$$

我们用样本矩替代总体矩, 则

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

■

Example 6.2.4:::> 设

$$X \sim f(x) = \begin{cases} \frac{6x}{\theta^3}(\theta - x) & 0 < x < \theta \\ 0 & \text{else} \end{cases}$$

求

1. θ 的矩估计量 $\widehat{\theta}$

2. $\widehat{\theta}$ 的方差

Solution. 1. 由于只有一个参数, 故我们只需求出总体的一阶矩即可.

$$\mathbb{E}x = \int_0^\theta x \cdot \frac{6x}{\theta^3}(\theta - x)dx = \frac{\theta}{2}$$

因而

$$\widehat{\theta} = 2\overline{X}$$

■

Example 6.2.5:::> 若 $X \sim U\left[-\frac{\theta}{2}, \frac{\theta}{2}\right]$, 求 θ 的矩估计量 $\widehat{\theta}$.

Solution. 我们发现

$$\mathbb{E}X = 0$$

虽然只有一个参数, 但我们无法使用一阶矩来求解, 但不要慌, 我们之前即是 k 从小到大列方程组, 列到足以求解问题为止, 因此求二阶矩:

$$\mathbb{E}X^2 = DX + (\mathbb{E}X)^2 = \frac{\theta^2}{12}$$

用样本二阶矩代替总体二阶矩:

$$\frac{1}{n}\sum_{i=1}^n X_i^2 = \frac{\theta^2}{12}$$

可得

$$\widehat{\theta} = \sqrt{\frac{12}{n}\sum_{i=1}^n X_i^2}$$

■

Example 6.2.6:::> $X \sim N(\mu, \sigma^2), X_1, X_2, \ldots, X_n$ 为其样本, 求 μ 的矩估计量 $\widehat{\mu}$ 与 σ^2 的矩估计量 $\widehat{\sigma^2}$.

Solution. 令

$$\frac{1}{n}\sum_{i=1}^n X_i = \mathbb{E}X = \mu$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2=\mathbb{E}X^2=\sigma^2+\mu^2$$

可解得

$$\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

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

■

其中的 S^2 即为样本方差2, 而 $\frac{n-1}{n}S^2$ 正是样本修正方差3

6.2.2 极大似然估计 MLE

Definition 6.2.7 (极大似然估计的原理)**:::>** 极大似然估计的原理即为概率论的一个基本思想: 概率越大的事件在一次试验中越容易发生.

我们定义似然函数为样本取到样本值的概率:

$$L(x,\dots x_n,\theta)=\begin{cases}P(X_1=x_1,\cdots X_n=x_n,\theta) & \text{离散}\\ \prod_{i=1}^{n}f_X(x_i,\theta) & \text{连续}\end{cases}$$

我们令

$$\widehat{\theta}_{MLE}=\arg\max_{\theta}L(x_1,\cdots x_n,\theta)$$

由于似然函数是以连乘的形式出现的, 一般情况下 (当然不是绝对的) 我们会通过求对数似然函数最大值来处理.

步骤:

1. 写出似然函数 $L(\theta)=\prod_{i=1}^{n}f(x_i;\theta)$, 求出 $\ln\mathrm{L}$ 及似然方程 $\frac{\partial \ln L}{\partial \theta_i}=0$
2. 解似然方程得到 $\theta_i(x_1,x_2,\dots,x_n)$, 即极大似然估计 $\theta_i(X_1,X_2,\dots,X_n)$

Remark. 似然方程无解时, 求出 θ 的定义域中使得似然函数最大的值, 即为最大似然估计. 判断时求导再根据驻点, 边界点, 间断点来判断即可. ■

Example 6.2.8:::> 设总体 X 的概率分布为

$$X \sim \begin{pmatrix} 0 & 1 & 2 & 3 \\ \theta^2 & 2\theta(1-\theta) & \theta^2 & 1-2\theta \end{pmatrix}$$

其中 $\theta\left(0<\theta<\frac{1}{2}\right)$ 是未知参数, 利用总体 X 的如下样本值

$$3,1,3,0,3,1,2,3,$$

求 θ 的矩估计值和最大似然估计值.

Solution. 矩估计值与矩估计量计算较为简单:

$$\mathbb{E}X = -4\theta + 3$$

则矩估计量为 $\frac{3-\overline{X}}{4}$, 代入样本数据 $\overline{X}=2$ 得矩估计值为 2.

似然函数

$$L(\theta) = P(X_1 = 3, \cdots X_8 = 3, \theta) = (1-2\theta)^4 \times \theta^2 \times [(2\theta(1-\theta))]^2 \times \theta^2$$

求对数似然函数最大值

$$\ln L(\theta) = \ln 4 + 6\ln\theta + 4\ln(1-2\theta) + 2\ln(1-\theta)$$

求导

$$\frac{\partial \ln L}{\partial \theta} = \frac{6}{\theta} - 2\frac{4}{1-2\theta} - \frac{2}{1-\theta}$$

令其得 0 , 解得

$$\widehat{\theta}_{MLE} = \frac{7 \pm \sqrt{13}}{12}$$

由于题目中要求 $0<\theta<\frac{1}{2}$, 故最终结果为 $\widehat{\theta}_{MLE} = \frac{7-\sqrt{13}}{12}$ ■

Example 6.2.9:::> 已知 $X \sim N(\mu, \sigma^2)$, 求最大似然估计 $\widehat{\mu}_{MLE}$ 和 $\widehat{\sigma^2_{MLE}}$

Solution.

$$L\left(x_1, \cdots, x_n, \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2}\pi}\right)^n \left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

对数似然:

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

求偏导:

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

同理, 将 σ^2 看成一个整体求偏导 $\frac{\partial \ln L}{\partial(\sigma^2)}$.

解得:

$$\begin{cases} \widehat{\mu}_{MLE} = \overline{x} \\ \widehat{\sigma}^2_{MLE} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2 = \dfrac{n-1}{n}s^2 \end{cases}$$

我们发现和矩估计结果一样.

■

Example 6.2.10:::> 设总体 $X \sim U[0,\theta], \theta(>0)$ 未知, $X_1, X_2, \cdots, X_n$ 是一个样本, 试求 θ 的最大似然估计量 $\widehat{\theta}$.

Solution. 似然函数

$$L(\theta) = \prod_{i=1}^{n} f(x_i,\theta) = \frac{1}{\theta^n}1_{\{0\leqslant x_1\cdots x_n\leqslant\theta\}}$$

这里不能盲目地程式化地动手求导, 而应思考一下我们运算要达到的目的 (求导的话会发现对数似然函数求导后结果恒为 $-n$, 而这个结果只能说明它对似然函数单减). 从似然函数可以看出, 要使 $L(x_1,\cdots,x_n;\theta)$ 最大, 就要使 θ 尽可能地小, 但 θ 又不能小于 $x_{(n)} = \max\{x_1,\cdots,x_n\}$, 所以, θ 取 $x_{(n)} = \max\{x_1,\cdots,x_n\}$ 时就使 $L(x_1,\cdots,x_n;\theta)$ 达到最大, 故 $\widehat{\theta}_L = X_{(n)}$. ■

6.2.3 点估计的评价标准

Definition 6.2.11 (点估计的评价标准)**:::>** 由于点估计种类很多, 因此我们要有方法对不同点估计进行评价. 一般来说, 从无偏性, 有效性, 相合 (一致) 性三个方面来评判点估计.

Definition 6.2.12 (无偏性)**:::>** θ 的估计量 $\widehat{\theta}(x_1,\cdots,x_n)$ 满足 $\mathbb{E}\widehat{\theta}<+\infty$, 且 $\forall\theta\in\Theta$, 有 $\mathbb{E}\widehat{\theta}=\theta$, 称 $\widehat{\theta}$ 为 θ 的无偏估计量.

Definition 6.2.13 (估计的均方误差)**:::>** 我们用均方误差 MSE 来定量衡量无偏性. θ 的均方误差:

$$\mathrm{MSE}(\theta,\theta)=\mathbb{E}(\theta-\theta)^2=D\widehat{\theta}+(\mathbb{E}\theta-\theta)^2=D(\widehat{\theta})+(\mathrm{Bias}(\widehat{\theta}))^2$$

若 θ 是无偏估计, 则 $\mathrm{MSE}(\theta,\theta)=D\theta$.

Example 6.2.14:::> 设 $\boldsymbol{X}\sim U[0,\theta]$, 参数 θ 未知, $X_1,X_2,\ldots,X_n$ 是其大小为 $\boldsymbol{n}$ 的样本. 则

1. 矩估计量 $\widehat{\theta}_M=2\overline{X}$ 是无偏的;

   Proof.

   $$\mathbb{E}\widehat{\theta}_{\text{矩}}=2\mathbb{E}(\overline{X})=2\cdot\frac{\theta}{2}=\theta$$

   是无偏的. ■

2. 似然估计 $\widehat{\theta}_L=X_{(n)}=\max\{X_1,X_2,\cdots,X_n\}$, 不是参数 θ 的无偏估计

   Proof. 我们需要求 $\mathbb{E}(\theta_{\mathrm{MLE}})$. 定义 $Z=\max_{1\leqslant i\leqslant n}X_i$, 由于

   $$P(Z\leqslant z)=P(X_1\leqslant z,X_2\leqslant z,\cdots,X_n\leqslant z)$$

   则

   $$F_Z(z)=[F_X(x)]^n=\begin{cases}0 & z<0\\ \left(\frac{z}{\theta}\right)^n & 0\leqslant z<\theta\\ 1 & z\geqslant\theta\end{cases}$$

   则

   $$\mathbb{E}Z=\mathbb{E}(\widehat{\theta_{\mathrm{MLE}}})=\int_0^\theta z\cdot\mathrm{d}F_Z(z)=\int_0^\theta z\,\mathrm{d}\left(\frac{z}{\theta}\right)^n=\frac{n}{n+1}\theta$$

   结果不为 θ, 故不是参数 θ 的无偏估计.

   进而, 根据结果我们发现修正后的 $\frac{n+1}{n}X_{(n)}$ 为无偏估计.

   ■

3. $\widehat{\theta}_M = 2\overline{X}$ 与 $\frac{n+1}{n}X_{(n)}$ 均为 θ 的无偏估计. 而 $\frac{n+1}{n}\widehat{\theta}_L = \frac{n+1}{n}X_{(n)}$ 是比 $\widehat{\theta}_M = 2\overline{X}$ 有效的估计量.

$\mathfrak{Proof}$.

$$D\left(\widehat{\theta}_1\right) = 4D(\overline{X}) = 4 \cdot \frac{DX_1}{n} = \frac{4}{n} \cdot \frac{\theta^2}{12}$$

而由上述分布式可得:

$$DZ = \mathbb{E}Z^2 - (\mathbb{E}Z)^2 = \frac{n}{n+2}\theta^2 - (\frac{n}{n+1}\theta)^2$$

经比较可得

$$D\left(\widehat{\theta}_2\right) \leqslant D\left(\widehat{\theta}_1\right)$$

■

$\mathfrak{Definition}$ **6.2.15** (有效性)**:::>** 有效性是在无偏的估计下完成的. 对于 θ 的无偏估计量 θ_1, θ_2, 若对于任意的 $\theta \in \Theta$ 有 $D(\theta_1) \leqslant D(\theta_2)$, 且至少有一个 $\theta \in \Theta$ 使得上述不等式严格成立, 则称 θ_1 比 θ_2 有效.

$\mathfrak{Theorem}$ **6.2.16:::>** 设总体 X 的 $\boldsymbol{k}$ 阶矩 $\mu_k = \mathbb{E}X^k, k \geqslant 1$ 存在. 设 $X_1, X_2, \ldots, X_n$ 是 X 的一个样本, 则不论总体服从什么分布, $\boldsymbol{k}$ 阶样本矩 $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 一定是 $\boldsymbol{k}$ 阶总体矩 μ_k 的无偏估计, 例如样本方差 S^2 一定是总体方差的无偏估计.

$\mathfrak{Theorem}$ **6.2.17:::>** 设总体 X 的 $\boldsymbol{k}$ 阶矩 $\mu_k = \mathbb{E}X^k, k \geqslant 1$ 存在. 设 $X_1, X_2, \ldots, X_n$ 是 X 的一个样本, 则不论总体服从什么分布, $\alpha_1 X_1 + \cdots + \alpha_n X_n$ 为 $\mathbb{E}X$ 的无偏估计, 其中 $\sum \alpha_i = 1$ 且 $\alpha_i \geqslant 0$.

而在这类无偏估计中, 最有效的是 $\alpha_i = \frac{1}{n}$.

$\mathfrak{Definition}$ **6.2.18** (相合性)**:::>** $\forall \theta \in \Theta, \widehat{\theta}_n(x_1, \cdots, x_n) \xrightarrow{P} \theta$, 则称 θ_n 为 θ 的相合估计 (一致估计). 相合性是一个大样本性质. 相合性要求是最基本的要求, 不满足相合性的估计一般不予考虑.

验证相合性一般会用到大数定律或是依概率收敛的性质来判断.

$\mathfrak{Theorem}$ **6.2.19:::>** 若 $\lim_{n\to\infty} \mathbb{E}\theta_n = \theta$ (渐近无偏估计) 且 $\lim_{n\to\infty} D\theta_n = 0$, 则 θ_n 为 θ 的相合估计.

Proof. 只需证明, $\forall \varepsilon > 0, P\left(\left|\widehat{\theta}_n - \theta\right| \geqslant \varepsilon\right) \to 0$.e 而

$$P\left(\left|\widehat{\theta}_n - \theta\right| \geqslant \varepsilon\right) = P\left(\left|\widehat{\theta}_n - \mathbb{E}\widehat{\theta}_n + \mathbb{E}\widehat{\theta}_n - \theta\right| \geqslant \varepsilon\right)$$

而

$$P\left(\left|\widehat{\theta}_n - E\widehat{\theta}_n\right| \geqslant \frac{\varepsilon}{2}\right) \leqslant \frac{D\left(\widehat{\theta}_n\right)}{\left(\frac{\varepsilon}{2}\right)^2} \to 0$$

最后一步是切比雪夫不等式切比雪夫不等式. ■

Theorem 6.2.20:::> 若 $\widehat{\theta}_n^1, \widehat{\theta}_n^2, \cdots, \widehat{\theta}_n^k$ 分别为 $\theta_1, \theta_2, \cdots, \theta_k$ 的相合估计, $\eta = g(\theta_1, \theta_2, \cdots, \theta_k)$ 是 $\theta_1, \theta_2, \cdots, \theta_k$ 的连续函数, 则 $g\left(\widehat{\theta}_n^1, \widehat{\theta}_n^2, \cdots, \widehat{\theta}_n^k\right)$ 为 $g(\theta_1, \theta_2, \cdots, \theta_k)$ 的相合估计.

Example 6.2.21:::> 设 $X \sim U[0, \theta]$, 参数 θ 未知, $X_1, X_2, \ldots, X_n$ 是其容量为 n 的样本. 则其最大似然估计 $\widehat{\theta}_L = X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$ 是 θ 的相合估计.

Proof.

$$\mathbb{E}\left(\widehat{\theta}_L\right) = \mathbb{E}[\max\{X_1, X_2, \cdots, X_n\}] = \frac{n}{n+1}\theta \to \theta$$

$$D\left(\widehat{\theta}_L\right) = \frac{n}{n+2}\theta^2 - \left[\frac{n}{n+1}\theta\right]^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 \to 0$$

故 $\widehat{\theta}_L = X_{(n)} = \max\{X_1, X_2, \cdots, X_n\}$ 是 θ 的相合估计.

■

6.2.4 UMVUE 与有效估计

Definition 6.2.22 (UMVUE):::> θ^* 是 θ 的无偏估计, 若对于 θ 的任意一个无偏估计量 θ, 有 $D\theta^* \leqslant D\theta$, 则 θ^* 是 θ 的一致最小方差无偏估计, 记为 MVUE 或 UMVUE(uniformly minimum-variance unbiased estimator).

Remark. 这里一致是指 $\forall \theta \in \Theta$. ■

Theorem 6.2.23 (Rao-Blackwell 定理)**:::>** 随机变量 X 的方差存在, 令 $\varphi(Y)=\mathbb{E}(X\mid Y)$, 则

$$\mathbb{E}[\varphi(Y)]=\mathbb{E}X, D[\varphi(Y)]\leqslant DX$$

且等号成立当且仅当 $P(X=\varphi(Y))=1$.

Proof. 不妨设 $\mathbb{E}X=0$, 则

$$\begin{aligned}
&D[\varphi(Y)]=\mathbb{E}[\varphi(Y)]^2\\
&=\mathbb{E}[\mathbb{E}(X\mid Y)]^2\\
&=\mathbb{E}\{[\mathbb{E}(X\mid Y)][\mathbb{E}(X\mid Y)]\}\\
&=\mathbb{E}[\mathbb{E}(X\mathbb{E}(X\mid Y)\mid Y)]\\
&=\mathbb{E}(X\mathbb{E}(X\mid Y))\\
&\leqslant\sqrt{\mathbb{E}X^2\mathbb{E}[\mathbb{E}(X\mid Y)]^2}
\end{aligned}$$

从而

$$D[\varphi(Y)]=\mathbb{E}[\mathbb{E}(X\mid Y)]^2\leqslant\mathbb{E}X^2=DX$$

■

Remark. 后面部分证明的时候也可以直接利用

$$DX=D[\mathbb{E}(X\mid Y)]+\mathbb{E}[D(X\mid Y)]$$

■

Corollary 6.2.24 (Rao-Blackwell 定理推论)**:::>** 这个推论比Rao-Blackwell 定理还重要, 甚至有时它被称为 Rao-Blackwell 定理.

若 $T=T(X_1,X_2,\cdots,X_n)$ 为 θ 的充分统计量, $\widehat{\theta}$ 为 θ 的任一无偏估计量, 则 $\widetilde{\theta}=\mathbb{E}(\widehat{\theta}\mid T)$ 也为 θ 的无偏估计量, 且 $D(\widetilde{\theta})\leqslant D\widehat{\theta}$

Example 6.2.25:::> 设 $X_1,X_2,\cdots,X_n$ 是总体 $X\sim B(1,p), 0<p<1$ (即 Bernoulli 分布) 的一个样本, 显然估计量 X_1 是 p 的无偏估计. 我们用 Rao-Blackwell 定理求 p 的改进的无偏估计量.

Solution. 由于 $T=\sum_{i=1}^{n}X_i$ 为 p 的充分统计量, 由于 X_1 只取 0 和 1 两个值, 故

$$\begin{aligned}
&\mathbb{E}[X_1 \mid T=t]=P(X_1=1 \mid T=t)\\
&=\frac{P(X_1=1,T=t)}{P(T=t)}\\
&=\frac{P(X_1=1,X_2+X_3+\cdots+X_n=t-1)}{P(T=t)}\\
&=\frac{P(X_1=1)P(X_2+X_3+\cdots+X_n=t-1)}{P(T=t)}\\
&=\frac{pC_{n-1}^{t-1}p^{t-1}(1-p)^{n-1-(t-1)}}{C_n^t p^t(1-p)^{n-t}}\\
&=\frac{t}{n}
\end{aligned}$$

故由 Rao-Blackwell 定理知:$\mathbb{E}[X_1 \mid T]=\frac{T}{n}(=\overline{X})$ 是 p 的改进的无偏估计量. 所以我们从 X_1 出发找到了一个比 X_1 更好的无偏估计.

■

Remark. 也可以利用对称性直接得,

$$E[X_1 \mid T]=E\left[X_1 \mid \sum_{i=1}^{n}X_i\right]=\frac{\sum_{i=1}^{n}X_i}{n}$$

■

Remark. 再来一次

$$\mathbb{E}\left(\frac{T}{n} \mid T\right)=\frac{T}{n}$$

这即是一个 UMVUE. ■

Definition 6.2.26 (零无偏估计量)**:::>** 零无偏估计量是指期望为 0 的统计量.

Theorem 6.2.27 (零无偏估计定理)**:::>** 这是 UMVUE 的一个判别准则. 如果 UMVUE 存在, 则由 Rao-Blackwell 定理的推论知, 它一定是充分统计量的函数.

如果 $\widehat{\theta}$ 是 θ 的方差有限的无偏估计量, 且对任何零无偏估计量 $\varphi = \varphi(X_1,\cdots,X_n)$, 都有

$$\mathrm{Cov}_\theta(\widehat{\theta},\varphi)=0,\forall\theta\in\Theta$$

则 $\widehat{\theta}$ 是 θ 的 UMVUE.

Proof. 对 θ 的任意一个无偏估计量 $\widetilde{\theta}$, 显然 $\varphi\triangleq\widetilde{\theta}-\widehat{\theta}$ 为零无偏估计量, 且

$$\begin{aligned}&D(\widetilde{\theta})=D(\widehat{\theta}+\varphi)=D(\widehat{\theta})+D(\varphi)+2\mathrm{Cov}(\widehat{\theta},\varphi)\\&=D(\widehat{\theta})+D(\varphi)\geqslant D(\widehat{\theta}),\forall\theta\in\Theta\end{aligned}$$

■

Example 6.2.28:::> 设 $X\sim E\left(\frac{1}{\theta}\right)$, 参数 θ 未知, $X_1,X_2,\ldots,X_n$ 是其容量为 n 的样本. 则 $\overline{X}$ 为 θ 的 UMVUE.

Proof. 由因子分解定理, $T=T(X_1,X_2,\ldots,X_n)=\sum_{i=1}^n X_i$ 为 θ 的充分统计量, $\overline{X}=\frac{T}{n}$ 为 θ 的无偏估计量.

设任何零无偏估计量 $\varphi=\varphi(X_1,\cdots,X_n)$, 由于

$$\mathbb{E}\varphi=\int_0^\infty\cdots\int_0^\infty\varphi(x_1,\cdots,x_n)\prod_{i=1}^n\left[\frac{e^{-\frac{x_i}{\theta}}}{\theta}\right]\mathrm{d}x_1\cdots\mathrm{d}x_n=0$$

即

$$\int_0^\infty\cdots\int_0^\infty\varphi(x_1,\cdots,x_n)e^{-\sum_{i=1}^n\frac{x_i}{\theta}}\,\mathrm{d}x_1\cdots\mathrm{d}x_n=0$$

两边对 θ 求导得

$$\int_0^\infty\cdots\int_0^\infty\frac{n\overline{x}}{\theta^2}\varphi(x_1,\cdots,x_n)e^{-\sum_{i=1}^n\frac{x_i}{\theta}}\,\mathrm{d}x_1\cdots\mathrm{d}x_n=0$$

即 $\mathbb{E}(\overline{X}\varphi)=0$, 即 $\mathrm{Cov}(\overline{X},\varphi)=0$, 故 $\overline{X}$ 为 θ 的 UMVUE. ■

Definition 6.2.29 (Fisher 信息量 (信息函数))**:::>** 设总体的密度函数 (或 pmf)$f(x;\theta),\theta\in\Theta$ 满足下列五条条件 (这五条条件也被称为 C-R 正则条件):

1. 参数空间 Θ 是直线上的一个开区间;
2. 支撑 $S=\{x;f(x;\theta)>0\}$ 与 θ 无关;

3. 导数 $\frac{\partial}{\partial\theta}f(x;\theta)$ 对一切 $\theta\in\Theta$ 都存在;

4. 对 $f(x;\theta)$, 积分与微分运算可交换次序, 即

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}f(x;\theta)dx=\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x;\theta)dx$$

5. 期望 $I(\theta)=\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(X;\theta)\right]^2$ 存在

则称该期望 $I(\theta)$ 为总体分布的 Fisher 信息量.

Remark. 如果二阶导数对一切 $\theta\in\Theta$ 都存在, 则 $I(\theta)$ 还可以用下式计算

$$I(\theta)=-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right]$$

■

Remark. 这五个性质里积分号与可导号可交换才是最为本质的条件.

■

Remark. $I(\theta)$ 信息量的意义: 当 $D\eta=\frac{1}{nI(\theta)}I(\theta)$ 越大, $D(\eta)$ 越小估计精度高, 而 $I(\theta)$ 大, 则认为模型本身所含的信息量较多, 或者说 θ 易认识, 所以可视 $I(\theta)$ 反应了模型中含有信息的量 ■

Example 6.2.30 (Fisher Information 的直观阐释)**:::>** 来源: 知乎回答

首先我们看一下 Fisher Information 的定义:

假设你观察到 iid 的数据 $X_1,X_2,\ldots X_n$ 服从一个概率分布 $f(X;\theta),\theta$ 是你的目标参数 (for simplicity, 这里 θ 是个标量, 且不考虑 nuissance parameter), 那么你的似然函数 (likelihood) 就是:

$$L(\mathrm{X};\theta)=\prod_{i=1}^{n}f(X_i;\theta)$$

为了解得 Maximum Likelihood Estimate(MLE), 我们要让 log likelihood 的一阶导数得 0, 然后解这个方程, 得到 $\widehat{\theta}_{MLE}$. 这个 log likelihood 的一阶导数也叫

score function:

$$S(\mathrm{X};\theta)=\sum_{i=1}^{n}\frac{\partial \log f\left(X_i;\theta\right)}{\partial \theta}$$

那么 Fisher Information $I(\theta)$ 的定义就是这个 Score function 的二阶矩 (second moment):

$$I(\theta)=\mathbb{E}\left[S(X;\theta)^2\right]$$

一般情况下, 可以容易地证明, $\mathbb{E}[S(\mathrm{X};\theta)]=0$, 从而得到:

$$I(\theta)=\mathbb{E}\left[S(X;\theta)^2\right]-\mathbb{E}[S(X;\theta)]^2=\mathrm{Var}[S(X;\theta)]$$

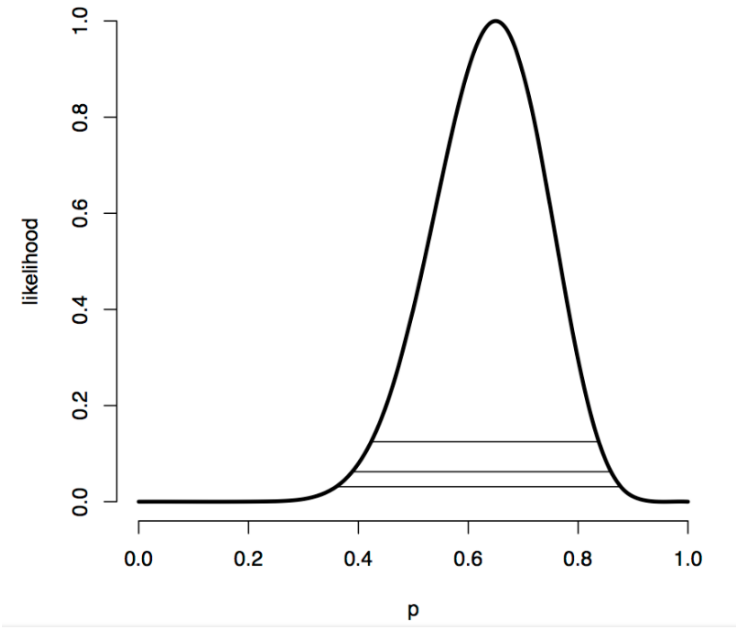
于是得到了 Fisher Information 的第一条数学意义: 就是用来估计 MLE 的方程的方差. 它的直观表述就是, 随着收集的数据越来越多, 这个方差由于是一个 Independent sum 的形式, 也就变的越来越大, 也就象征着得到的信息越来越多.

而且, 如果 log likelihood 二阶可导, 在一般情况下可以容易地证明:

$$\mathbb{E}\left[S(\mathrm{X};\theta)^2\right]=-\mathbb{E}\left(\frac{\partial^2}{\partial \theta^2}\log L(\mathrm{X};\theta)\right)$$

于是得到了 Fisher Information 的第二条数学意义: log likelihood 在参数真实值处的负二阶导数的期望。这个意义好像很抽象, 但其实超级好懂.

首先看一下一个 normalized Bernoulli log likelihood 长啥样:



对于这样的一个 log likelihood function, 它越平而宽, 就代表我们对于参数估计的能力越差, 它高而窄, 就代表我们对于参数估计的能力越好, 也就是信息量越大.

而这个 log likelihood 在参数真实值处的负二阶导数, 就反应了这个 log likelihood 在顶点处的弯曲程度, 弯曲程度越大, 整个 log likelihood 的形状就越偏向于高而窄, 也就代表掌握的信息越多.

然后, 在一般情况下, 通过对 score function 在真实值除泰勒展开, 然后应用中心极限定理, 弱大数定律, 依概率一致收敛, 以及 Slutsky 定理, 可以证明 MLE 的渐进分布的方差是 $I^{-1}(\theta)$, 即

$$\mathrm{Var}\left(\widehat{\theta}_{MLE}\right) = I^{-1}(\theta)$$

这也就是 Fisher Information 的第三条数学意义. 不过这样说不严谨, 严格地说, 应该是

$$\sqrt{n}\left(\widehat{\theta}_{MLE} - \theta\right) \xrightarrow{\nu} N\left(0, I^{*}(\theta)^{-1}\right)$$

这里 $I^{*}(\theta)$ 是当只观察到一个 X 值时的 Fisher Information, 当有 n 个 i.i.d 观测值时 $I^{*}(\theta) = I(\theta)/n$. 所以这时的直观解释就是, Fisher Information 反映了

我们对参数估计的准确度, 它越大, 对参数估计的准确度越高, 即代表了越多的信息.

Theorem 6.2.31 (Cremer-Rao 不等式)**:::>** 设 $X_1, X_2, \cdots, X_n$ 为取自具有 pdf (或 pmf) $f(x;\theta), \theta \in \Theta = \{\theta : a < \theta < b\}$ 的总体 X 的一个样本, a, b 为已知常数, a 可以取 $-\infty$, b 可以取 $+\infty$. 又有 $\eta = \eta(X_1, X_2, \cdots, X_n)$ 是 $g(\theta)$ 的一个无偏估计, 且满足 C-R 正则条件:

1. 集合 $\{x : f(x;\theta) > 0\}$ 与 θ 无关;
2. $g'(\theta)$ 与 $\frac{\partial f(x;\theta)}{\partial \theta}$ 存在, 且对一切 $\theta \in \Theta$

$$\frac{\partial}{\partial \theta} \int f(x;\theta) dx = \int \frac{\partial f(x;\theta)}{\partial \theta} dx$$

$$\begin{aligned} &\frac{\partial}{\partial \theta} \iint \cdots \int \eta(x_1, x_2, \cdots, x_n) f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta) dx_1 dx_2 \cdots dx_n \\ &= \iint \cdots \int \eta(x_1, x_2, \cdots, x_n) \frac{\partial}{\partial \theta} \left[\prod_{i=1}^{n} f(x_i;\theta) \right] dx_1 dx_2 \cdots dx_n \end{aligned}$$

即

$$\frac{\partial}{\partial \theta} \mathbb{E}(\eta(\mathrm{X})) = \mathbb{E}\left(\frac{\partial}{\partial \theta} \eta(\mathrm{X}) \right)$$

令 $I(\theta) = E_\theta \left(\frac{\partial \ln f(X;\theta)}{\partial \theta} \right)^2 > 0$, 则这个无偏估计的方差有一个下界:

$$D_\theta \eta \geqslant \frac{[g'(\theta)]^2}{nI(\theta)} \tag{*}$$

并且存在一个有可能依赖于 θ 但不依赖于 $X_1, X_2, \cdots, X_n$ 的数 K, 使得等式 $\sum_{i=1}^{n} \frac{\partial \ln f(X_i;\theta)}{\partial \theta} = K(\eta - g(\theta))$ 以概率 1 成立, 以上这个条件为式 (*) 中等式成立的充要条件. 特别地, 当 $g(\theta) = \theta$ 时, 不等式 (*) 化为:

$$D_\theta \eta \geqslant \frac{1}{nI(\theta)}$$

Remark. 若 ξ 是离散型随机变量, $f(x;\theta)$ 则表示为 $P(\xi = x;\theta)$, 相应的积分号改为求和号. ■

Remark. 在使用 R-C 不等式时不必验证 2 是否成立, 因为在一般情况下, 当 1 成立时 2 自动满足 ■

Remark. 满足上述条件中 1, 2 假定的估计量称为正则估计 ■

Remark. R-C 下界 $\frac{\left[g'(\theta)\right]^2}{nI(\theta)}$ 不是所有无偏估计的下界, 而是无偏估计类中一个子集: 正则无偏估计的方差下界 ■

Remark. 达到R-C下界的估计量一定是UMVUE, 但反之不然, UMVUE不一定达到RC下界. ■

Definition 6.2.32 (有效估计)**:::>** 若 C-R 不等式的等号成立 (达到 C-R 下界), 则称 $\eta = \eta(X_1, X_2, \cdots, X_n)$ 是 $g(\theta)$ 的有效估计, 有效估计一定是 UMVUE, 但 UMVUE 不一定是有效估计.

Definition 6.2.33 (无偏估计的效率)**:::>** 设 T_1, T_2 为 θ 的两个无偏估计, 其方差均存在, 称

$$\mathrm{eff}_\theta(T_1 \mid T_2) \triangleq \frac{D_\theta(T_2)}{D_\theta(T_1)}$$

为 T_1 关于 T_2 的效率.

若 $\mathrm{eff}_\theta(T_1 \mid T_2) > 1$, 则称 T_1 比 T_2 有效.

若 T 为 θ 的有效估计 (即 $D_\theta(T) = \frac{1}{nI(\theta)}$), 则定义 θ 的无偏估计 T_1 的效率为

$$\mathrm{eff}_\theta(T_1) = eff_\theta(T_1 \mid T) = \frac{D_\theta(T)}{D_\theta(T_1)} = \frac{1}{nI(\theta)D_\theta(T_1)}$$

故有效估计的效率为 1, 任一无偏估计的效率不超过 1.

Example 6.2.34:::> 设样本 $X_1, X_2, \ldots, X_n$ 是来自 $B(1, p)$ 两点分布的样本, 求 p 的有效分布与 UMVUE.

Solution. 设总体分布为 $f(x;p)=p^x(1-p)^{1-x},\quad x=0,1,0<p<1$, 易证 $\{f(x;p)\mid p\in p(0,1)\}$ 满足 C-R 正则条件, 又因为

$$I(p)=\mathbb{E}_p\left[\frac{\partial}{\partial p}\ln f(X,p)\right]^2=\mathbb{E}_p\left[\frac{X-p}{p(1-p)}\right]^2=\frac{1}{p^2(1-p)^2}\mathbb{E}_p(X-p)^2=\frac{1}{p(1-p)}$$

故 $\frac{1}{nI(p)}=\frac{p(1-p)}{n}$ 为 p 的无偏估计的 C - R 下界.

而之前在6.2.25证明过, $\overline{X}$ 是 p 的无偏估计, 有:

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

可知 $\overline{X}$ 达到了 C-R 下界, 故 $\overline{X}$ 为 p 的 UMVUE. ■

Example 6.2.35:::> 设总体 X 的密度函数为

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

其中未知参数 $\lambda>0$. 总体的一个样本为 $(X_1,X_2,\cdots,X_n)$, 证明 $\overline{X}$ 是 $\frac{1}{\lambda}$ 的一个 UMVUE.

Proof. 由指数分布的总体满足正则条件可得:

$$I(\lambda)=-\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\ln f(X;\lambda)\right]=-\mathbb{E}\left(\frac{-1}{\lambda^2}\right)=\frac{1}{\lambda^2}$$

则

$$\frac{\left[\left(\frac{1}{\lambda}\right)'\right]^2}{nI(\lambda)}=\frac{\left[\frac{-1}{\lambda^2}\right]^2}{n\frac{1}{\lambda^2}}=\frac{1}{n\lambda^2}$$

为 $\frac{1}{\lambda}$ 的无偏估计方差的 C-R 下界, 故 $\overline{X}$ 是 $\frac{1}{\lambda}$ 的一个 UMVUE. ■

6.2.5 完备统计量与 Lemann-Scheffe 定理

Definition 6.2.36 (完备统计量):::> 若 $\forall\theta\in\Theta$,

$$\mathbb{E}_\theta(g(X))=0\Rightarrow P_\theta(g(X)=0)=1$$

则称分布族 $\{f_\theta(x):\theta\in\Theta\}$ 是完备的.

若 T 的分布族是完备的, 则称统计量 $T=T(X_1,X_2,\cdots,X_n)$ 为完备的.

Remark. 完备分布族条件下, $E_\theta(g_1(X))=E_\theta(g_2(X))$, 则有

$$P_\theta(g_1(X)=g_2(X))=1$$

■

Example 6.2.37:::> 二项分布族 $\{B(n,p):0<p<1\}$ (n 已知) 是完备分布族.

Proof. 若

$$\mathbb{E}_p(g(X))=\sum_{k=0}^{n}g(k)C_n^k p^k(1-p)^{n-k}=0,\quad 0<p<1$$

则

$$\sum_{k=0}^{n}g(k)C_n^k\left(\frac{p}{1-p}\right)^k=0,\quad 0<p<1$$

等式左面是 $\frac{p}{1-p}$ 的 n 次多项式, 故 $g(k)=0,k=0,1,2,\cdots,n$, 证毕. ■

Theorem 6.2.38:::> 若似然函数可以表示为如下形式:

$$\prod_{i=1}^{n}f(x_i;\theta)=C(\theta)h(x_1,x_2,\ldots,x_n)\exp\{b(\theta)T(x_1,x_2,\ldots,x_n)\}$$

则 $T=T(X_1,X_2,\ldots,X_n)$ 是 θ 的充分完备统计量.

而若:

$$\prod_{i=1}^{n}f(x_i;\theta)=C(\theta)h(x_1,x_2,\ldots,x_n)\exp\{b_1(\theta)T_1(x_1,x_2,\ldots,x_n)+b_2(\theta)T_2(x_1,x_2,\ldots,x_n)\}$$

则 (T_1,T_2) 是 $\theta=(\theta_1,\theta_2)$ 的充分完备统计量.

Theorem 6.2.39 (Lehmann-Scheffe 定理):::> 若 T 是 θ 的充分完备统计量, $\varphi(X_1,\cdots,X_n)$ 是 $g(\theta)$ 的一个无偏估计, 则 $E[\varphi(X_1,\cdots,X_n)\mid T]$ 为 $g(\theta)$ 的惟一的 UMVUE.

Example 6.2.40:::> UMVUE 的求解步骤:

1. 求出参数 θ 的充分完备统计量 T
2. 求出 $ET = g(\theta)$, 则 $\theta = g^{-1}(T)$ 是 θ 的一个无偏估计, 或求出一个无偏估计, 然后改写成为用 T 表示的函数.
3. 综合, $E[g^{-1}(T) \mid T] = g^{-1}(T)$ 是 θ 的 UMVUE

或者, 求出 θ 的矩估计或 ML 估计, 再求效率, 为 1 则必为 UMVUE.

Example 6.2.41:::> 设样本 $X_1, X_2, \ldots, X_n$ 是来自 $B(1,p)$, 求 p 的 UMVUE.

Solution. $T = \sum_{i=1}^{n} X_i \sim B(n,p)$ 为充分完备统计量, 且 X_1 为 p 的无偏估计. 则由 Lehmann - Scheffe 定理, p 唯一的 UMVUE 为

$$E(X_1 \mid T) = E\left(X_1 \mid \sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

■

Example 6.2.42:::> 设样本 $X_1, X_2, \ldots, X_n$ 是来自 $N(\theta,1)$, 求 θ 的 UMVUE

Solution. $T = \overline{X} \sim N(\theta,1)$ 是充分完备统计量 (指数族分布), 且 X_1 为 θ 的无偏估计. 由 Lehmann-Scheffe 定理, θ 的 UMVUE 为 $E(X_1 \mid T)$.

由于

$$\begin{pmatrix} X_1 \\ \overline{X} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ \theta \end{pmatrix}, \begin{pmatrix} 1 & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} \end{pmatrix}\right)$$

利用多维高斯分布条件期望的结论多维 Gauss 分布及其特征函数, 可得

$$\begin{aligned} E(X_1 \mid T) &= E\left(X_1 \mid \overline{X}\right) = \theta + \frac{\operatorname{Cov}\left(X_1, \overline{X}\right)}{D\overline{X}}(\overline{X} - \theta) \\ &= \theta + (\overline{X} - \theta) = \overline{X} \end{aligned}$$

即 $\overline{X}$ 为 θ 的 UMVUE. ■

Example 6.2.43:::> $X_1 \cdots X_n (n > 4)$ 是总体 $X \sim P(\lambda)$ 的样本, 求 $e^{-3\lambda}$ 的 UMVUE.

Solution. $T=\sum_{i=1}^{n}X_i\sim p(n\lambda)$ 是 λ 的充分完备统计量. 对于 $U=1_{\{X_1=0,X_2=0,X_3=0\}}$, $\mathbb{E}U=P(X_1=0,X_2=0,X_3=0)=e^{-3\lambda}$

由 Lehmann-Scheffe 定理,

$$\begin{aligned}\mathbb{E}(U\mid T)&=P(X_1=0,X_2=0,X_3=0\mid T=t)\\&=\frac{P(X_1=0,X_2=0,X_3=0,\sum_{i=4}^{n}X_i=t)}{p(T=t)}\\&=\frac{\left(e^{-\lambda}\right)^3\cdot\frac{((n-3)\lambda)^t e^{-(n-3)\lambda}}{t!}}{\frac{(n\lambda)^t e^{-n\lambda}}{t!}}\\&=\left(1-\frac{3}{n}\right)^t\\&=\left(1-\frac{3}{n}\right)^{\sum_{i=1}^{n}X_i}\end{aligned}$$

■

6.3. 参数的区间估计

Definition 6.3.1 (参数的区间估计):::> 点估计是用一个统计量来作为参数的估计, 区间估计是找两个统计量

$$\widehat{\theta}_L=\widehat{\theta}_L(X_1,X_2,\cdots,X_n)$$

和

$$\widehat{\theta}_U=\widehat{\theta}_U(X_1,X_2,\cdots,X_n)$$

其中对任何样本观测值都有 $\widehat{\theta}_L\leqslant\widehat{\theta}_U$, 并用区间 $\left[\widehat{\theta}_L,\widehat{\theta}_U\right]$ 作为参数的区间估计 (量).

Remark. 注意这里闭区间不是必要的, 可以是开区间或左开右闭等. ■

Example 6.3.2:::> $X_1,X_2,\cdots,X_9$ 是取自正态总体 $N(\mu,1)$ 的一个样本, 那么 $[\overline{X}-1,\overline{X}+1]$ 就是参数 μ 的一个区间估计. 在点估计部分, 我们用样本均值 $\overline{X}$ 作为 μ 的点估计量, 在很多场合下, 我们需要得到 μ 的一个 (随机) 区间估计.

$$
\begin{aligned}
&P(\mu \in [\overline{X}-1, \overline{X}+1]) = P(-1 \leqslant \overline{X}-\mu \leqslant 1) \\
&= P\left(-3 \leqslant \frac{\overline{X}-\mu}{\sqrt{\frac{1}{9}}} \leqslant 3\right) \\
&= 2\Phi(3)-1 \approx 0.997
\end{aligned}
$$

也就是说, 随机区间 $[\overline{X}-1, \overline{X}+1]$ 涵盖参数真值的概率是 0.997.

Definition 6.3.3 (置信度)**:::>** 随机区间涵盖参数真值的概率称为置信度, 即下面这个概率

$$P_\theta\left(\theta \in \left[\widehat{\theta}_L(X_1, X_2, \cdots, X_n), \widehat{\theta}_U(X_1, X_2, \cdots, X_n)\right]\right)$$

这个概率一般来说与 θ 有关, 希望最小的置信度 (置信水平或系数) 也比较大, 即 $\inf_{\theta\in\Theta} P_\theta\left(\theta \in \left[\widehat{\theta}_L, \widehat{\theta}_U\right]\right)$ 比较大.

Definition 6.3.4 (置信区间平均长度)**:::>** 除了置信系数的要求, 当然我们希望置信区间的平均长度 $\mathbb{E}_\theta\left(\widehat{\theta}_U - \widehat{\theta}_L\right)$ 越小越好. 即置信水平给定时, 置信区间的平均长度越小越好, 即置信度越高.

Definition 6.3.5 (Neyman 准则)**:::>** 在保证置信水平的前提下, 要求置信区间平均长度越小越好.

Definition 6.3.6 (置信水平为 $1-\alpha$ (双侧) 置信区间)**:::>** 若

$$P_\theta\left(\widehat{\theta}_L(X_1, \cdots, X_n) < \theta < \widehat{\theta}_U(X_1, \cdots, X_n)\right) \geqslant 1-\alpha, \forall \theta \in \Theta$$

则称 $\left[\widehat{\theta}_L(X_1, X_2, \cdots, X_n), \widehat{\theta}_U(X_1, X_2, \cdots, X_n)\right]$ 为 θ 的置信水平为 $1-\alpha$ 的置信区间. 其中 $\widehat{\theta}_L(X)$ 称为 (双侧) 置信下限, $\widehat{\theta}_U(X)$ 称为 (双侧) 置信上限.

若条件改为等号:

$$P_\theta\left(\widehat{\theta}_L(X_1, \cdots, X_n) < \theta < \widehat{\theta}_U(X_1, \cdots, X_n)\right) = 1-\alpha, \forall \theta \in \Theta$$

则称为同等置信区间.

置信水平 $1-\alpha$ 有一个频率解释: 在大量重复使用 θ 的置信区间 $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ 时, 每次得到的样本观测值是不同的, 从而每次得到的区间也是不一样的. 对一次具体的观测值而言, θ 可能在 $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ 内, 也可能不在. 平均而言, 在这大量的区间估计观测值中, 至少有 $100(1-\alpha)\%$ 包含 θ.

Definition 6.3.7 (置信水平为 $1-\alpha$ 单侧置信区间)**:::>** 若

$$P_\theta\left(\theta < \widehat{\theta}_U(X_1,\cdots,X_n)\right) \geqslant 1-\alpha, \forall \theta \in \Theta$$

则我们称 $\widehat{\theta}_U(X)$ 为单侧置信上限.

若

$$P_\theta\left(\theta > \widehat{\theta}_L(X_1,\cdots,X_n)\right) \geqslant 1-\alpha, \forall \theta \in \Theta$$

则我们称 $\widehat{\theta}_L(X)$ 为单侧置信下限.

若改为等号, 则获得同等单侧置信上下限.

Theorem 6.3.8:::> 若 $\widehat{\theta}_L(X), \widehat{\theta}_U(X)$ 分别为参数 θ 的置信水平为 $1-\alpha_1$, $1-\alpha_2$ 的单侧置信上下限, 则 $\left[\widehat{\theta}_L(X), \widehat{\theta}_U(X)\right]$ 为 θ 的水平为 $1-(\alpha_1+\alpha+2)$ 的双侧置信区间.

Definition 6.3.9 (置信区间的一般求法: 枢轴变量法 (pivot))**:::>**

1. 先找一个与要估计的参数 θ (或 $g(\theta)$) 有关的统计量 $\boldsymbol{T}$, 一般是良好的点估计 (MLE 或充分统计量)
2. 寻找一样本 $X_1, X_2, \cdots, X_n$ 的函数, 即找出 $\boldsymbol{T}$ 和 θ (或 $g(\theta)$) 的某一函数:

   $$Z = \mathrm{Z}(X_1, X_2, \cdots, X_n, \theta) \equiv \mathrm{Z}(T, \theta)$$

   它只含待估参数和样本, 不含其它未知参数, 并且 Z 的分布（或渐近分布）G 已知, 且不依赖于任何未知参数 (也不依赖于待估参数 θ) (称 Z 为枢轴变量);
3. 对于给定的置信度 $1-\alpha$, 定出两常数 a, b, 使得

   $$P(a \leqslant \mathrm{Z}(X_1, X_2, \cdots, X_n, \theta) \leqslant b) = 1-\alpha$$

   其中 a, b 选用原则: $\mathbb{E}_\theta\left(\widehat{\theta}_U - \widehat{\theta}_L\right)$ 最小 (最优置信区间); 一般 a, b 选用分布 (或渐近分布) G 的 $\frac{\alpha}{2}$ 和 $1-\frac{\alpha}{2}$ 分位点 (等尾置信区间). 等尾置信区间不一定是最优的, 只有枢轴变量的分布为单峰对称分布时 (如标准正态分布), 则等尾置信区间是最优置信区间.

4. 若能从 $a \leqslant Z(X_1, X_2, \cdots, X_n, \theta) \leqslant b$, 得到等价的不等式

$$\widehat{\theta}_L(X_1, X_2, \cdots, X_n) \leqslant \theta \leqslant \widehat{\theta}_U(X_1, X_2, \cdots, X_n)$$

那么 $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ 就是 θ (或 $g(\theta)$) 的一个置信度为 $1-\alpha$ 的置信区间.

Example 6.3.10:::> 设样本 $X_1, \cdots, X_8$ 是来自 $U[0, \theta]$, 求 θ 的置信度为 $(1-\alpha)\%$ 的等尾置信区间与最优置信区间.

Solution. 第一步: 从充分统计量或点估计出发找枢轴量.

由6.2.10, $X_{(n)}$(样本最大值) 为 θ 的 MLE, 故是充分统计量. 而可知其分布函数 $X_{(n)} \sim [F(x)]^n$, 密度函数为 $X_{(n)} \sim n\left(\frac{x}{\theta}\right)^{n-1} \cdot \frac{1}{\theta} \quad 0 \leqslant x \leqslant \theta$

则枢轴量为:

$$Y = \frac{X_{(n)}}{\theta} \sim f_Y(y) = 8y^7 1_{\{0 \leqslant y \leqslant 1\}}$$

这个分布不依赖于包括 θ 在内的任何一个参数.

第二步: 确定常数

等尾置信区间:

$$P\left(y_{\frac{\alpha}{2}} \leqslant \frac{X_{(n)}}{\theta} \leqslant y_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

这是根据等尾区间定义写出的, 其中 $\int_0^{y_{\frac{\alpha}{2}}} 8y^7 dy = \frac{\alpha}{2}, \quad y_{\frac{\alpha}{2}} = \left(\frac{\alpha}{2}\right)^{\frac{1}{8}}$, 同理 $y_{1-\frac{\alpha}{2}} = \left(\frac{1-\alpha}{2}\right)^{\frac{1}{8}}$

第三步: 不等式改写

等尾置信区间 $\left[\frac{X_{(n)}}{\left(\frac{1-\alpha}{2}\right)^{\frac{1}{8}}}, \frac{X_{(n)}}{\left(\frac{\alpha}{2}\right)^{\frac{1}{8}}}\right]$

同上述步骤, 有最优置信区间:

$$P\left(a \leqslant \frac{X_{(n)}}{\theta} \leqslant b\right) = 1-\alpha$$

$$b^8-a^8=1-\alpha$$

则

$$\mathbb{E}_\theta\left(\widehat{\theta}_U-\widehat{\theta}_L\right)=\left(\frac{1}{a}-\frac{1}{b}\right)\mathbb{E}X_{(n)}$$

即求, 在满足 $b^8-a^8=1-\alpha, 0\leqslant a<b\leqslant 1$ 的条件下, $\left(\frac{1}{a}-\frac{1}{b}\right)\mathbb{E}X_{(n)}$ 的最小值. 求条件极值问题 (可忽略 $EX_{(n)}$), 解得 $b=1, a=\sqrt[n]{\alpha}\quad (n=8)$, 记最优置信区间 $\left[X_{(n)}, \frac{X_{(n)}}{\sqrt[n]{\alpha}}\right]$.

■

Example 6.3.11:::> 设样本 $X_1,\cdots,X_n$ 是来自 $E\left(\frac{1}{\theta}\right)$, 求 θ 的置信度为 $(1-\alpha)\%$ 的等尾置信下限

Solution. 第一步: 从充分统计量或点估计出发找枢轴量

由于 $T=\sum_{i=1}^n X_i$ 是 θ 的充分统计量, $T=\sum_{i=1}^n X_i\sim\Gamma\left(n,\frac{1}{\theta}\right)$. 枢轴量为:

$$\frac{2T}{\theta}\sim\Gamma\left(n,\frac{1}{2}\right)=\chi^2(2n)$$

注意到这是一个非对称分布.

第二步: 确定常数

$$P\left(\frac{2T}{\theta}\leqslant\chi^2_{1-\alpha}(2n)\right)=1-\alpha$$

第三步: 不等式改写

$$P\left(\theta\geqslant\frac{2T}{\chi^2_{1-\alpha}(2n)}\right)=1-\alpha$$

则 θ 的置信度为 $(1-\alpha)\%$ 的等尾置信下限为 $\widehat{\theta}_L=\frac{2T}{\chi^2_{1-\alpha}(2n)}$.

注意这里求的是单侧等尾置信下限, 而有一个技巧, 就是要把单侧等尾置信下限中的 α 换为 $\frac{\alpha}{2}$ 即得到双侧等尾置信下限. 即双侧等尾置信下限为:

$$\widehat{\theta}_L=\frac{2T}{\chi^2_{1-\frac{\alpha}{2}}(2n)}$$

■

Example 6.3.12:::> 设样本 $X_1,\cdots,X_n$ (大样本) 是来自 $B(1,p)$, 求 p 的置信度为 $(1-\alpha)\%$ 的等尾置信区间

Solution. 第一步: 从充分统计量或点估计出发找枢轴量

由于 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 是 θ 的充分统计量, 为简化问题, 找枢轴量需要用到中心极限定理:

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

(渐近正态)

则枢轴量为

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

第二步: 确定常数

$$P\left(u_{\frac{\alpha}{2}}\leqslant\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}}\leqslant u_{1-\frac{\alpha}{2}}\right)=1-\alpha$$

第三步: 不等式改写

$$p\in\frac{1}{1+\frac{\lambda}{n}}\left[\overline{X}+\frac{\lambda}{2n}\pm\sqrt{\frac{\overline{X}(1-\overline{X})}{n}\lambda+\left(\frac{\lambda}{2n}\right)^2}\right]\quad\left(\lambda=u_{1-\frac{\alpha}{2}}^2\right)$$

由于 n 很大, 故 θ 的置信度为 $(1-\alpha)\%$ 的等尾置信区间为

$$\left[\overline{X}+u_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right]$$

■

Definition 6.3.13 (正态总体 $X\sim N(\mu,\sigma^2)$ 参数的置信区间问题)**:::>** 此类问题可归于如下类别中:

1. 求 μ 的置信度为 1- α 的置信区间. 此类问题又有如下情况:
   (a) 设方差 σ^2 已知
   (b) 设方差 σ^2 未知
2. 方差 σ^2 的置信度 $1-\alpha$ 置信区间. 此类问题又有如下情况:
   (a) 设 μ 已知
   (b) 设 μ 未知
3. 两个总体均值差 $\mu_1-\mu_2$ 的置信区间. 此类问题又有如下情况:
   (a) 设 ${\sigma_1}^2, {\sigma_2}^2$ 均为已知
   (b) 设 $\sigma_1^2=\sigma_2^2=\sigma^2$, 但 σ^2 为未知
   (c) 设 σ_1^2, σ_2^2 均为未知, 且不知它们是否相等
4. 两个总体方差比 ${\sigma_1^2}/{\sigma_2}^2$ 的置信区间

Theorem 6.3.14 (正态总体均值的区间估计)**:::>**

1. 条件: 正态总体, 方差已知

   枢轴量:

   $$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

   双侧等尾置信区间 (也是最优的):

   $$\left(\overline{X}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\overline{X}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

   这里的 $u_{1-\alpha/2}$ 等式表示的是正态分布的分位点.

2. 条件: 正态总体, 方差未知

   枢轴量:

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

   双侧等尾置信区间:

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

而单侧置信上, 下限只需要替换就好了:

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

其中的 $t_{1-\alpha}(n-1)$ 查表即可.

Theorem 6.3.15 (正态总体方差的区间估计)**:::>**

1. 条件: 正态总体, 均值未知.

   由于

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

   则枢轴量即为 $\frac{(n-1)S^2}{\sigma^2}$.

   等尾置信区间:

   $$P\left(\chi^2_{\frac{\alpha}{2}}(n-1)\leqslant\frac{(n-1)S^2}{\sigma^2}\leqslant\chi^2_{1-\frac{\alpha}{2}}(n-1)\right)=1-\alpha$$

   为方便起见, 我们求其等尾置信区间:

   $$P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\leqslant\sigma^2\leqslant\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right)=1-\alpha$$

   于是得到 σ^2 的一个置信水平为 $1-\alpha$ 的置信区间

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

   此外还可以得到标准差 σ 的一个置信水平为 $1-\alpha$ 的置信区间

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

2. 条件: 正态总体, 均值已知.

   均值已知时, 如果我们此时仍用

   $$\frac{(n-1)S^2}{\sigma^2}=\frac{\sum_{i=1}^n\left(X_i-\overline{X}\right)^2}{\sigma^2}\sim\chi^2(n-1)$$

作为枢轴量, 其实不是不行, 但是这样就没有充分利用已知的信息, 会导致精度下降.

这时若写出联合密度, 得到充分统计量为 $\sum_{i=1}^{n}(X_i-\mu)^2$, 从它出发得到

$$\sum_{i=1}^{n}(X_i-\mu)^2/\sigma^2 \sim \chi^2(n)$$

这就是 μ 已知时合理的枢轴量.

$$\mathrm{P}_{\sigma^2}\left\{\chi^2_{1-\alpha/2}(n) \leqslant \sum_{i=1}^{n}(X_i-\mu)^2/\sigma^2 \leqslant \chi^2_{\alpha/2}(n)\right\} = 1-\alpha$$

同等置信区间为:

$$\left[\sum_{i=1}^{n}(X_i-\mu)^2/\chi^2_{\alpha/2}(n), \sum_{i=1}^{n}(X_i-\mu)^2/\chi^2_{1-\alpha/2}(n)\right]$$

Theorem 6.3.16 (正态总体 (μ,σ^2) 的置信域)**:::>** 这是一个二维问题, 但我们仍然可以用类似一维问题中固定的思路去解决.

第一步, 从充分统计量出发构造枢轴量. 已证明过 $\left(\overline{X},S^2\right)$ 是 (μ,σ^2) 的充分统计量, 且有

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

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

而且 $\left(\overline{X},S^2\right)$ 是独立的, 所以上述两个量也是独立的. 而每个量的分布都是已知的, 所以联合后分布也是完全已知的. 因此取枢轴量为

$$\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}, \frac{Q^2}{\sigma^2}\right)$$

第二步确定常数:

二维情形一般取长方形区域 $[c_1,c_2]\times[d_1,d_2]$, 则概率为

$$P\left(c_1 \leqslant \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leqslant c_2, d_1 \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant d_2\right) = 1-\alpha$$

根据独立性, 概率等于

$$P\left(c_1 \leqslant \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leqslant c_2\right) P\left(d_1 \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant d_2\right) = 1-\alpha$$

此时我们平均分配概率, 即这两个概率都是 $\sqrt{1-\alpha}$, 即

$$\mathrm{P}_{\mu,\sigma^2}\left\{\frac{|\overline{X}-\mu|}{\sigma/\sqrt{n}} \leqslant c\right\} = \sqrt{1-\alpha} \widehat{=} 1-\alpha_1$$

$$\mathrm{P}_{\mu,\sigma^2}\left\{d_1 \leqslant \frac{Q^2}{\sigma^2} \leqslant d_2\right\} = \sqrt{1-\alpha} = 1-\alpha_1$$

类似前面用分位数表示常数, 得

$$c = u_{\alpha_1/2}, d_1 = \chi^2_{1-\alpha_1/2}(n-1), d_2 = \chi^2_{\alpha_1/2}(n-1)$$

第三步, 等价改写概率不等式.

根据独立性有,

$$\mathrm{P}_{\mu,\sigma^2}\left\{(\overline{X}-\mu)^2 \leqslant c^2\sigma^2/n, \quad Q^2/d_2 \leqslant \sigma^2 \leqslant Q^2/d_1\right\} = 1-\alpha$$

因此置信域为

$$\left\{(\mu,\sigma^2): \quad (\overline{X}-\mu)^2 \leqslant c^2\sigma^2/n, \quad Q^2/d_2 \leqslant \sigma^2 \leqslant Q^2/d_1\right\}$$

Theorem 6.3.17 (双正态总体均值差的区间估计)**:::>** 两个总体, $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$, 且两个总体之间是独立的.

σ_1^2, σ_2^2 均已知的情况下,

枢轴量

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

双侧等尾置信区间

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm u_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\sigma_1^2=\sigma_2^2$, 但两者未知的情况下

枢轴量

$$T=\frac{\left(\overline{X}_1-\overline{X}_2\right)-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t\left(n_1+n_2-2\right)$$

其中

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

双侧等尾置信区间:

$$\left(\overline{X}_1-\overline{X}_2\right)\pm t_{1-\alpha/2}(n_1+n_2-2)\cdot S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

σ_1^2,σ_2^2 均未知, 且为大样本

枢轴量:

$$Z=\frac{\left(\overline{X}_1-\overline{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\sim AN(0,1)$$

这里的 AN 指的是渐近正态分布.

双侧等尾置信区间:

$$\left(\overline{X}_1-\overline{X}_2\right)\pm u_{1-\alpha/2}\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

$\sigma_1^2\neq\sigma_2^2$, 且为小样本

枢轴量:

$$T=\frac{\left(\overline{X}_1-\overline{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\sim t(f)$$

其中

$$f=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

双侧等尾置信区间:

$$\left(\overline{X}_1-\overline{X}_2\right)\pm t_{1-\alpha/2}(f)\cdot\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

Theorem 6.3.18 (双正态总体方差比 σ_1^2/σ_2^2 的区间估计)**:::>** 均值 μ_1,μ_2 未知的情况下:

$$\frac{(n_1-1)S_1^2}{\sigma_1^2}\sim\chi^2(n_1-1)$$

$$\frac{(n_2-1)S_2^2}{\sigma_2^2}\sim\chi^2(n_2-1)$$

且两者独立.

则

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

即

$$P\left(F_{\frac{\alpha}{2}}(n_1-1,n_2-1)\leqslant\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\leqslant F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)\right)=1-\alpha$$

也即

$$P\left(\frac{S_1^2/S_2^2}{F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)}leq\frac{\sigma_1^2}{\sigma_2^2}\leqslant\frac{S_1^2/S_2^2}{F_{\frac{\alpha}{2}}(n_1-1,n_2-1)}\right)=1-\alpha$$

而在均值已知的情况下, 从点估计量出发寻找枢轴量, 类似单个正态总体下均值已知时方差的区间估计, 而当时构造的枢轴量为

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}(X_i-\mu)^2/n\sim\chi^2(n)$$

对于两个正态总体的方差可以分别用点估计量替代, 这样分子分母都是卡方分布, 通过配凑系数就可以得到 F 分布, 即

$$\frac{\frac{1}{\sigma_1^2}\sum_{i=1}^{m}(X_i-\mu_1)^2/m}{\frac{1}{\sigma_2^2}\sum_{j=1}^{m}\left(Y_j-\mu_2\right)^2/n}\sim F(m,n)$$

Theorem 6.3.19 (总体未知, 均值已知的区间估计)**:::>** 一般方法: 充分统计量 (MLE)T 的分布已知: $T\sim F(t,\theta)$, 注意这是一个分布函数而不是密度函数.

用2.9.2中的结论,

$$F(T;\theta)\sim U(0,1)$$

我们发现 $F(T;\theta)$ 包含了统计量与未知参数, 同时不包含其它东西, 故我们将其定为枢轴量.

即解不等式:

$$P\left(\frac{\alpha}{2}\leqslant F(T,\theta)\leqslant 1-\frac{\alpha}{2}\right)=1-\alpha$$

即可获得区间.

Example 6.3.20:::> $X\sim(\mu,\sigma^2)$, σ^2 已知, 求 μ 的信度为 $1-\alpha$ 的置信区间.

Solution. 充分统计量

$$\overline{X}\sim N(\mu,\frac{\sigma^2}{n})$$

而其分布函数为

$$F(t,\mu)=\Phi\left(\frac{t-\mu}{\sigma/\sqrt{n}}\right)$$

根据总体未知, 均值已知的区间估计中, 取枢轴量时将充分统计量中 t 的部分用 T 来代替即可, 枢轴量为:

$$\Phi\left(\frac{t-\mu}{\sigma/\sqrt{n}}\right)\sim U[0,1]$$

则

$$P\left(\frac{\alpha}{2}\leqslant\overline{\Phi}\left(\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}\right)\leqslant 1-\frac{\alpha}{2}\right)=1-\alpha$$

解不等式发现与我们之前在正态总体均值的区间估计中算的结果一致.

■

Definition 6.3.21 (自助法 (Bootstrap))**:::>** 这一方法也称为重抽样法, 想法就是用经验分布函数作为总体分布的替代.

假设 $\widetilde{X}=(X_1,\cdots,X_n)$ 是总体的一组简单随机样本, $\widehat{\theta}=\widehat{\theta}(\widetilde{X})$ 是参数 θ 的点估计量, 现在考虑构造枢轴量为 $\widehat{\theta}-\theta$.

此时如果 $\widehat{\theta}-\theta$ 的分布是已知的, 那我们就通过

$$\mathrm{P}(c\leqslant\widehat{\theta}-\theta\leqslant d)=1-\alpha$$

得到两个常数, 然后等价改写不等式就可以得到区间估计了.

但如果 $\widehat{\theta}-\theta$ 的分布不是已知的, 那我们考虑如下方法得到它的近似分布.

首先 $\widehat{\theta}=\widehat{\theta}(\widetilde{X})$ 的分布是与总体分布有关的, 但是因为我们并不知道总体分布, 所以这个统计量的分布也就不知道了. 但是我们通过 $\widetilde{X}=(X_1,\cdots,X_n)$ 能够得到经验分布函数, 而经验分布函数又是总体分布很好的近似, 所以我们就用经验分布函数来代替总体分布.

现在我们利用这个近似已知的总体, 用模拟的方法从中抽取新的一组样本

$$\widetilde{X}^*=(X_1^*,\cdots,X_m^*)$$

利用这个新的样本, 用同样的方法构造 $\widehat{\theta}^*=\widehat{\theta}\left(\widetilde{X}^*\right)$, 那么可以用 $\widehat{\theta}^*$ 近似 $\widehat{\theta}$.

又注意到 $\widehat{\theta}=\widehat{\theta}(\widetilde{X})$ 是 θ 的估计, 所以 $\widehat{\theta}-\theta$ 的分布可以用 $\widehat{\theta}^*-\widehat{\theta}(\widetilde{x})$ 的分布近似, 而 $\widehat{\theta}^*-\widehat{\theta}(\widetilde{x})$ 的分布又是完全已知的, 因此我们通过 $\widehat{\theta}^*-\widehat{\theta}(\widetilde{x})$ 的分布得到常数 c,d

$$\mathrm{P}^*\left(c\leqslant\widehat{\theta}^*-\widehat{\theta}(\widetilde{x})\leqslant d\right)=1-\alpha$$

然后将不等式中间近似转换, 得到

$$\mathrm{P}^*(c\leqslant\widehat{\theta}-\theta\leqslant d)\approx 1-\alpha$$

再等价改写不等式得到近似的区间估计为:

$$[\widehat{\theta}(\widetilde{X})-d,\widehat{\theta}(\widetilde{X})-c]$$

Definition 6.3.22 (Bayes 区间估计)**:::>** Bayes 区间估计中的" 区间" 被称为可信区间. 在获得参数 θ 的后验分布后,

$$P(\widehat{\theta}_L \leqslant \theta \leqslant \widehat{\theta}_U) = 1-\alpha$$

即可获得可信区间.

6.4. 参数的假设检验

Definition 6.4.1 (假设检验基本思想)**:::>** 假设检验类似于反证法. 我们先提出假设, 如果在假设条件之下发生不太合理的现象, 我们就拒绝假设.

Definition 6.4.2 (假设)**:::>** 所要检验真假的称为原假设或零假设, 与之互斥的称为备择假设或对立假设.

对于参数分布族 $\mathscr{F} = \{p_\theta(x) : \theta \in \Theta\}$, 关于其未知参数的假设可以是

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

其中 Θ_0, Θ_1 不相交.

如果不是对于参数分布族的假设, 则称为非参数假设.

如果原假设中 Θ_0 只包含一个元素, 称为简单假设, 否则称为复杂假设.

Remark.

1. 原假设和备择假设不一定互补
2. 备择假设一般是由样本产生疑问, 想要研究的.
3. 不成文的规定: 原假设一般含有等号.

■

Definition 6.4.3 (检验)**:::>** 所谓的检验就是给出一个规则, 有了样本观测值后根据这个规则就能判断原假设是否为真.

这个规则实质就是对样本空间进行划分, 将样本空间分成拒绝域和接受域:

$$\Omega = D \cup \overline{D}$$

当样本落入拒绝域, 拒绝原假设; 否则保留原假设.

有时为了方便, 引入检验函数:

$$\varphi(\widetilde{x}) = \begin{cases} 1, & \text{当} \widetilde{x} \in D \\ 0, & \text{当} \widetilde{x} \in \overline{D} \end{cases}$$

检验函数与检验规则一一对应.

因为样本是多维的, 为了确定拒绝域, 通常需要构造检验统计量.

Definition 6.4.4 (两类错误)**:::>** 利用检验规则对原假设进行判断的时候, 会遇到两类错误.

1. 第一类错误, 也叫弃真错误. 即原假设成立, 但认为原假设为假. 犯这类错误的概率也称为弃真概率, 记为 $\alpha(\theta)$:

$$\alpha(\theta) = \mathrm{P}\left(\widetilde{X} \in D \mid H_0 \text{ 为真}\right) = \mathrm{P}_\theta(\widetilde{X} \in D), \quad \theta \in \Theta_0$$

2. 第二类错误, 也叫取伪错误. 即原假设不成立, 但认为原假设为真. 犯这类错误的概率也称为取伪概率, 记为 $\beta(\theta)$:

$$\beta(\theta) = \mathrm{P}\left(\widetilde{X} \in \overline{D} \mid H_1 \text{ 为真}\right) = \mathrm{P}_\theta(\widetilde{X} \in \overline{D}), \quad \theta \in \Theta_1$$

在样本容量固定的条件下, 弃真概率与取伪概率是相互制约的. 当其中一个概率减小, 势必会引起另外一个概率增大.

Definition 6.4.5 (势函数)**:::>** 检验的势函数定义为 $g(\theta) = \mathrm{P}_\theta(\widetilde{X} \in D), \quad \theta \in \Theta$, 表示样本落入拒绝域的概率, 即用检验否定原假设的概率. 它与两类错误有如下关系:

$$\mathrm{P}_\theta(\widetilde{X} \in D) = \begin{cases} \alpha(\theta) & \text{当} \theta \in \Theta_0 \\ 1 - \beta(\theta) & \text{当} \theta \in \Theta_1 \end{cases}$$

即在原假设的参数空间中, 它等于弃真概率; 在备择假设的参数空间中, 它等于 1-取伪概率.

Definition 6.4.6:::> 1. 双侧检验:$H_0:\theta=\theta_0 \longleftrightarrow H_1:\theta\neq\theta_\theta$

2. 右侧检验:$H_0:\theta\leqslant\theta_0 \longleftrightarrow H_1:\theta>\theta_\theta$

3. 左侧检验:$H_0:\theta\geqslant\theta_0 \longleftrightarrow H_1:\theta<\theta_\theta$

Definition 6.4.7 (Neyman-Pearson 原则与显著性水平为 α 的检验)**:::>** Neyman-Pearson 原则: 在控制犯第一类错误的前提下: $\alpha(\theta)\leqslant\alpha$ (α 即为显著性水平), 使犯第二类错误的概率尽可能的小.

一般取 $\alpha=\sup_{\theta\in\Theta_0}\alpha(\theta)$.

Definition 6.4.8 (假设检验的步骤)**:::>** 假设检验的一般步骤如下:

1. 根据实际问题中所关心的问题, 建立原假设 H_0 和备择假设 H_1
2. 选择一个合适的统计量 V (要求 H_0 为真时, V 的分布已知), 并根据其取值特点确定一个合适的拒绝域形式. 一般拒绝域的形式可以是单侧或双侧的, 要根据备择假设来选取.

   我们可以总结出规律, 拒绝域选取的形式和备择假设 H_1 的大于小于号方向相同.
3. 由给定的显著性水平 α, 利用关系式

$$P_{H_0}\left(\left\{V<V_{\frac{\alpha}{2}}\right\}\bigcup\left\{V>V_{1-\frac{\alpha}{2}}\right\}\right)=\alpha$$

$$P_{H_0}\left(\left\{V>V_{1-\alpha}\right\}\right)=\alpha$$

$$P_{H_0}\left(\left\{V<V_{\alpha}\right\}\right)=\alpha$$

   中的某一个, 求出水平为 α 的检验拒绝域.
4. 根据样本观察值, 算出 V 的观察值, 并根据它作出接受还是拒绝 H_0.

Theorem 6.4.9 (单正态总体的检验统计量与拒绝域)**:::>** 对于一个正态总体 $X\sim N(\mu,\sigma^2)$, 我们要选择一个合适的统计量, 要求 H_0 为真时, 其分布已知. 并根据统计量的特性选择合适的拒绝域. 我们可以将应选择的统计量与拒绝域总结如下:

1. σ^2 已知, 检验 $\mathrm{H}_0: \mu=\mu_0 \leftrightarrow H_1: \mu \neq \mu_0$(双侧检验), 统计量:

$$U=\frac{\overline{X}-\mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

拒绝域为:

$$W=\{(x_1, \cdots, x_n): |u|>C\}$$

而由于

$$P_{H_0}((X_1, \cdots, X_n) \in W)=P_{H_0}\left(\left\{U<u_{\frac{\alpha}{2}}\right\} \cup\left\{U>u_{1-\frac{\alpha}{2}}\right\}\right)=\alpha,$$

故

$$W=\left\{(x_1, \cdots, x_n): |u|>u_{1-\frac{\alpha}{2}}\right\}=\left\{u<u_{\frac{\alpha}{2}} \text{ or } u>u_{1-\frac{\alpha}{2}}\right\}$$

2. σ^2 已知, 检验 $H_0: \mu \geqslant \mu_0 \leftrightarrow H_1: \mu<\mu_0$(左边检验), 统计量:

$$U=\frac{\overline{X}-\mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

拒绝域为:

$$W=\{(x_1, \cdots, x_n): u(x_1, \cdots, x_n)<C\}$$

而

$$\begin{aligned}
&\alpha(\mu)=P\left\{(X_1, X_2, \cdots, X_n) \in W \mid H_0 \text{ 为真}\right\} \\
&=P(U<C \mid \mu \geqslant \mu_0)=P\left(\frac{\overline{X}-\mu_0}{\sigma / \sqrt{n}}<C \mid \mu \geqslant \mu_0\right) \\
&=P\left(\frac{\overline{X}-\mu+\mu-\mu_0}{\sigma / \sqrt{n}}<C \mid \mu \geqslant \mu_0\right) \\
&=\Phi\left(C+\frac{\mu_0-\mu}{\sigma / \sqrt{n}}\right) \quad (\mu \geqslant \mu_0)
\end{aligned}$$

$\alpha(\mu)$ 关于 μ 单减, $\sup_{\mu \geqslant \mu_0} \alpha(\mu)=\Phi(C)=\alpha$, 故 $C=u_\alpha$ 故在显著性水平 α 时, 拒绝域为 $W=\{(x_1, \cdots, x_n): u(x_1, \cdots, x_n)<u_\alpha\} \triangleq\{u<u_\alpha\}$

它与假设 $\mathrm{H}_0: \mu=\mu_0 \leftrightarrow H_1: \mu<\mu_0$ 的拒绝域是一样的.

3. σ^2 未知, 检验 $\mathrm{H}_0: \mu=\mu_0 \leftrightarrow H_1: \mu \neq \mu_0$, 统计量:

$$T=\frac{\overline{X}-\mu_0}{S_n^* / \sqrt{n}} \sim t(n-1)$$

拒绝域为:

$$W=\left\{(x_1, \cdots, x_n): |t|>t_{1-\frac{\alpha}{2}}\right\}=\left\{t<t_{\frac{\alpha}{2}} \text{ or } t>t_{1-\frac{\alpha}{2}}\right\}$$

4. μ 已知, 检验 $\mathrm{H}_0: \sigma^2=\sigma_0^2 \leftrightarrow H_1: \sigma^2 \neq \sigma_0^2$ 统计量:

$$\chi^2=\frac{\sum_{i=1}^n\left(X_i-\mu\right)^2}{\sigma_0^2} \sim \chi^2(n)$$

5. μ 未知, 检验 $\mathrm{H}_0: \sigma^2=\sigma_0^2 \leftrightarrow H_1: \sigma^2 \neq \sigma_0^2$ 统计量:

$$\chi^2=\frac{\sum_{i=1}^n\left(X_i-\overline{X}\right)^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Theorem 6.4.10:::> 将复杂假设中原假设的不等号改为等号, 使原假设成为简单假设, 两者拒绝域的临界值相同.

Example 6.4.11:::> 均值 μ 未知时, 假设 $H_0: \sigma^2 \leqslant \sigma_0^2 \leftrightarrow H_1: \sigma^2>\sigma_0^2$ 的拒绝域与假设 $H_0: \sigma^2=\sigma_0^2 \leftrightarrow H_1: \sigma^2>\sigma_0^2$ 的拒绝域相同, 均为

$$W=\left\{\left(x_1, \cdots, x_n\right): \frac{(n-1) s^2}{\sigma_0^2}>\chi^2(n-1)\right\}$$

Definition 6.4.12 (正态双总体的假设检验问题)**:::>** 两个总体 $X \sim N\left(\mu_1, \sigma_1^2\right), \quad Y \sim N\left(\mu_2, \sigma_2^2\right)$,

1. 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知时, 检验 $\mathrm{H}_0: \mu_1=\mu_2 \leftrightarrow \mathrm{H}_1: \mu_1 \neq \mu_2$

$$T=\frac{\overline{X}-\overline{Y}}{\sqrt{\left(n_1-1\right) S_{1 n_1}^{* 2}+\left(n_2-1\right) S_{2 n_2}^{* 2}}} \sqrt{\frac{n_1 n_2\left(n_1+n_2-2\right)}{n_1+n_2}} \sim t\left(n_1+n_2-2\right)$$

2. 当 μ_1, μ_2 未知时, 检验 $\mathrm{H}_0: \sigma_1^2=\sigma_2^2 \leftrightarrow H_1: \sigma_1^2 \neq \sigma_2^2$,

$$F=\frac{S_{1 n_1}^{* 2}}{S_{2 n_2}^{* 2}} \sim F\left(n_1-1, n_2-1\right)$$

Theorem 6.4.13 (区间估计与假设检验之间的关系)**:::>** 设 $X_1, \cdots, X_n$ 是总体 $X \sim N\left(\mu, \sigma^2\right)$ (σ^2 已知) 的一个样本, 检验 $H_0: \mu=\mu_0 \leftrightarrow H_1: \mu \neq \mu_0$, 检验统计量为 $U=\frac{\overline{X}-\mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$, 其拒绝域为

$$\begin{aligned} W & =\left\{\left(x_1, x_2, \cdots, x_n\right):|u|>u_{1-\frac{\alpha}{2}}\right\} \\ & =\left\{\left(x_1, x_2, \cdots, x_n\right):|\overline{x}-\mu_0|>u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right\} \end{aligned}$$

故接受域为

$$W^c = \left\{(x_1, x_2, \cdots, x_n) : |\overline{x} - \mu_0| \leqslant u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right\}$$
$$= \left\{(x_1, x_2, \cdots, x_n) : \overline{x} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leqslant \mu_0 \leqslant \overline{x} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right\}$$

我们发现, 将小写改为大写,μ_0 改为 μ, 那么接受域就为参数的置信度为 $1-\alpha$ 的置信区间, 即 $\left[\overline{X} - u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + u_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$.

一般地, 设 $A(\theta_0)$ 是 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$ 的接受域, 令

$$C(x_1, x_2, \cdots, x_n) = \{\theta : (x_1, x_2, \cdots, x_n) \in A(\theta_0)\}$$

则 $C(X_1, X_2, \cdots, X_n)$ 为参数 θ 的置信度为 $1-\alpha$ 的置信区间.

反之, 若 $C(X_1, X_2, \cdots, X_n)$ 为参数 θ 的置信度为 $1-\alpha$ 的置信区间, 令

$$A(\theta_0) = \{(x_1, x_2, \cdots, x_n) : \theta_0 \in C(x_1, x_2, \cdots, x_n)\}$$

则 $A(\theta_0)$ 是 $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$ 的置信水平为 α 的接受域.

Definition 6.4.14 (单参数指数型分布族假设检验问题)**:::>** 对于单参数指数型分布族, 若样本 $X_1, \cdots, X_n$ 的联合密度函数为

$$\prod_{i=1}^{n} f(x_i; \theta) = C(\theta) h(x_1, x_2, \ldots, x_n) \exp\{Q(\theta) T(x_1, x_2, \ldots, x_n)\}$$

其中 $Q(\theta)$ 关于 θ 严格递增, 我们一般取检验统计量为 $T(X_1, X_2, \cdots, X_n)$, 考察三类检验问题

1. $H_0 : \theta \geqslant \theta_0 \leftrightarrow H_1 : \theta < \theta_0$
2. $H_0 : \theta \leqslant \theta_0 \leftrightarrow H_1 : \theta > \theta_0$
3. $H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$

在确定拒绝域的形式时, 我们需要先求检验统计量的期望 $E_\theta T(X_1, X_2, \cdots, X_n)$, 如果它是 θ 的单增函数, 那么对于第一个假设而言, 当 $T(X_1, X_2, \cdots, X_n)$ 比较小的时候我们会拒绝原假设; 类似地可以得到上述三种检验问题的拒绝域形式分别为

1. $\{T(x_1,x_2,\cdots,x_n)<c\}$
2. $\{T(x_1,x_2,\cdots,x_n)>c\}$
3. $\{T(x_1,x_2,\cdots,x_n)<c_1\}\cup\{T(x_1,x_2,\cdots,x_n)>c_2\}$

计算临界值时, 我们要使得犯第一类错误的概率小于显著性水平 α, 即原假设成立的条件下样本落入拒绝域的概率的上确界小于显著性水平. 概率的上确界就是 $\theta=\theta_0$ 时的概率, 即

$$\sup_{\theta\subset\theta_0} P_\theta\{T(X_1,X_2,\cdots,X_n)<c\}=P_{\theta_0}\{T(X_1,X_2,\cdots,X_n)<c\}$$

若 $\theta=\theta_0$ 时分布完全确定, 令上式为 α 即可求出临界值.

Example 6.4.15:::> 设样本 $X_1,\cdots,X_n$ 是来自 $E\left(\frac{1}{\theta}\right)$, 检验假设

$$H_0:\theta\geqslant\theta_0\leftrightarrow H_1:\theta<\theta_0$$

Solution. 样本的分布为

$$f(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}f(x_i;\theta)=\left(\frac{1}{\theta}\right)^n e^{-\frac{1}{\theta}\sum_{i=1}^n x_i}1_{\{x_1,x_2,\cdots,x_n>0\}}$$

取检验统计量为 $T(X_1,X_2,\cdots,X_n)=\sum_{i=1}^n X_i$.

其期望为

$$E_\theta T(X_1,X_2,\cdots,X_n)=n\theta$$

是 θ 的单增函数, 又有

$$T=\sum_{i=1}^{n}X_i\sim\Gamma\left(n,\frac{1}{\theta}\right)$$

故拒绝域的形式为

$$\{T(x_1,x_2,\cdots,x_n)<c\}=\{n\overline{x}<c\}$$

犯第一类错误的概率的最大值为

$$\sup_{\theta\geqslant\theta_0} P_\theta\{n\overline{X}<c\}=P_{\theta_0}\{n\overline{X}<c\}=\alpha$$

由于

$$T=\sum_{i=1}^{n}X_i\sim\Gamma\left(n,\frac{1}{\theta}\right)$$

故

$$\frac{2n\overline{X}}{\theta}\sim\Gamma\left(n,\frac{1}{2}\right)=\chi^2(2n)$$

因此临界值为 $c=\frac{\theta_0}{2}\chi^2_\alpha(2n)$.

拒绝域为

$$\left\{(x_1,x_2,\cdots,x_n):n\overline{x}<\frac{\theta_0}{2}\chi^2_\alpha(2n)\right\}$$
$$=\left\{(x_1,x_2,\cdots,x_n):\frac{2n\overline{x}}{\theta_0}<\chi^2_\alpha(2n)\right\}$$

■

Definition 6.4.16 (假设检验的 p 值)**:::>** 定义假设检验的 p 值为:

1. 在原假设为真的前提下出现观察样本以及更极端情况的概率
2. 利用样本数据能拒绝原假设的最小显著性水平
3. 观察到的 (实例样本数据的) 显著性水平
4. 表示对原假设的支持程度, 是用于确定是否应该拒绝原假设的另一种方法

Definition 6.4.17 (p 值和控制犯第一类错误的概率 α 之间的关系)**:::>** 在经典统计 (即不考虑贝叶斯学派) 中, 从对这一问题的处理方式不同, 可分为两派:

1. Neyman-Pearson 派: α 是建立在 $H_0\leftrightarrow H_1$ 体系下的, 需要事先设定 H_0,H_1, 通过事先给定 α, 即犯第一类错误的概率来划定拒绝域的边界. Neyman-Pearson 派不会可以关注 p 值的具体值, 而是只看观察结果是否落在拒绝域之内.
2. Fisher 派: p 值, 是在原假设正确的前提下, 出现观察结果或比之前更极端情形的概率, 其计算根本不涉及备择假设. Fisher 派在计算出 p 值以后, 根据其大小判断是否拒绝原假设, 认为 p 值的界限可以取 0.05 或 0.01 等比较小的数, 并把这个界限称为显著性水平.

Definition 6.4.18 (p 值的计算)**:::>** 设 T 表示假设检验的检验的统计量, 当 H_0 为真时, 可由样本数据计算出该统计量的观测值 C, 根据检验统计量 T 的具体分布, 可求出 p 值.

1. 左侧检验的 p 值为: $p = P(T \leqslant C)$
2. 右侧检验的 p 值为: $p = P(T \geqslant C)$
3. 双侧检验的 p 值为: 检验统计量 T 落在样本统计量 C 为端点的尾部区域内的概率的两倍: 即

$$p = 2P(T \geqslant C) \text{ (当}C \text{ 位于分布曲线的右端时)}$$

$$p = 2P(T \leqslant C) \text{ (当}C \text{ 位于分布曲线的左端时)}$$

若 T 服从正态分布和 t 分布, 其分布曲线是关于纵轴对称的, 则 $p = P(|T| \geqslant C)$.

Definition 6.4.19 (利用 p 值进行检验)**:::>** 计算出 p 值以后, 将给定的显著性水平 α 与 p 值比较, 就可以作出检验的结论:

1. 如果 p 值 $< \alpha$, 则在显著性水平 α 下拒绝原假设.
2. 如果 p 值 $\geqslant \alpha$, 则在显著性水平 α 下不拒绝原假设

Definition 6.4.20 (最优势检验)**:::>** 设 $X_1, \cdots, X_n$ 是总体 $X \sim f(x, \theta)(\theta \in \Theta)$ 的一个样本, 考虑假设

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

$X_1, \cdots, X_n$ 的取值范围 $S_n = W \bigcup W^c$, 这里 W 为假设检验的拒绝域.

检验函数:

$$\varphi(x_1, \cdots, x_n) = \begin{cases} 1, & (x_1, \cdots, x_n) \in W \\ 0, & (x_1, \cdots, x_n) \notin W \end{cases} = 1_{\{(x_1, \cdots, x_n) x \in W\}}$$

检验的势函数为:

$$g(\theta) = E_\theta(\varphi(x_1, \cdots, x_n)) = \begin{cases} \alpha(\theta), & \theta \in \Theta_0 \\ 1 - \beta(\theta), & \theta \in \Theta_1 \end{cases}$$

定义: 当样本容量 n 固定时, 检验法 φ_1 与 φ_2 的势分别为 $g_1(\theta), g_2(\theta)$, 水平均为 α(i.e. $\sup_{\theta\in\Theta_0} g_1(\theta)\leqslant\alpha, \sup_{\theta\in\Theta_0} g_2(\theta)\leqslant\alpha$), 若 $\forall\theta\in\Theta$, 均有

$$g_1(\theta)\geqslant g_2(\theta)$$

则称 φ_1 比 φ_2 的更有效.

若水平为 α 的检验法中, φ_1 比一切水平为 α 的检验法都有效, 则称 φ_1 是水平为 α 的最优势检验.

一般而言, 寻找最优势检验比较困难. 一个比较好的例子是似然比检验.

Definition 6.4.21 (似然比检验 (简单假设对简单假设))**:::>** 设 $X_1,\cdots,X_n$ 是总体 $X\sim f(x,\theta)(\theta\in\Theta)$ 的一个样本, 考虑简单假设对简单假设的问题

$$H_0:\theta=\theta_0\leftrightarrow H_1:\theta=\theta_1$$

定义似然比

$$\lambda(x_1,x_2,\cdots,x_n)=\frac{L(x_1,x_2,\cdots,x_n,\theta_1)}{L(x_1,x_2,\cdots,x_n,\theta_0)}=\prod_{i=1}^{n}\frac{f(x_i,\theta_1)}{f(x_i,\theta_0)}$$

检验的拒绝域为:

$$W=\{(x_1,x_2,\cdots,x_n):\lambda(x_1,x_2,\cdots,x_n)\geqslant k\}(k\text{ 待定})$$

Theorem 6.4.22 (Neyman-Pearson 引理)**:::>** 设 $X_1,\cdots,X_n$ 是总体 $X\sim f(x,\theta)(\theta\in\Theta)$, 对

$$H_0:\theta=\theta_0\leftrightarrow H_1:\theta=\theta_1$$

φ 是其似然比检验法, 若检验法 φ 与 $\widetilde{\varphi}$ 的势分别为 $g(\theta),\widetilde{g}(\theta)$ (水平为 α), 且满足 $\widetilde{g}(\theta_0)\leqslant g(\theta_0)$, 则 $\widetilde{g}(\theta_1)\leqslant g(\theta_1)$.

即若似然比检验的水平为 α(id est$g(\theta_0)=\alpha$), 则对任一水平为 α 的检验 $\widetilde{\varphi}$, 若 $\widetilde{g}(\theta_0)\leqslant\alpha$, 则似然比检验 φ 为最优势检验.

Example 6.4.23:::> 设 $X_1,\cdots,X_n$ 是 $N(\mu,\sigma^2)$ (σ 已知) 的一个样本, 检验假设

$$H_0:\mu=\mu_0\leftrightarrow H_1:\mu=\mu_1$$

其中 $\mu_1>\mu_0$.

Solution. 似然比

$$\begin{aligned}\lambda(x_1,x_2,\cdots,x_n)&=\frac{L(x_1,x_2,\cdots,x_n,\mu_1)}{L(x_1,x_2,\cdots,x_n,\mu_0)}\\&=\frac{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\sum_{i=1}^n\frac{(x_i-\mu_1)^2}{2\sigma^2}\right\}}{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left\{-\sum_{i=1}^n\frac{(x_i-\mu_0)^2}{2\sigma^2}\right\}}\\&=\exp\left\{n\overline{x}\left(\frac{\mu_1-\mu_0}{\sigma^2}\right)-\frac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}\right\}\end{aligned}$$

当 H_0 为真时,

$$\begin{aligned}&\lambda(x_1,x_2,\cdots,x_n)\geqslant k\Rightarrow\overline{x}\geqslant\frac{\ln k+\frac{n\left(\mu_1^2-\mu_0^2\right)}{2\sigma^2}}{n\left(\frac{\mu_1-\mu_0}{\sigma^2}\right)}\\&\Rightarrow\frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\geqslant\frac{\sigma\ln k}{\sqrt{n}(\mu_1-\mu_0)}+\frac{\sqrt{n}(\mu_1-\mu_0)}{2\sigma}\triangleq C\end{aligned}$$

拒绝域为

$$\begin{aligned}W&=\{(x_1,x_2,\cdots,x_n):\lambda(x_1,x_2,\cdots,x_n)\geqslant k\}\\&=\left\{(x_1,x_2,\cdots,x_n):\frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\geqslant C\right\}\\&=\left\{(x_1,x_2,\cdots,x_n):\frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\geqslant u_{1-\alpha}\right\}\end{aligned}$$

是最优势检验. ■

Example 6.4.24::> 设 $X_1,\cdots,X_n$ 是 $E\left(\frac{1}{\theta}\right)$ (σ 已知) 的一个样本, 检验假设

$$H_0:\theta=\theta_0\leftrightarrow H_1:\theta=\theta_1$$

其中 $\theta_1<\theta_0$.

Solution. 似然比

$$\begin{aligned}\lambda(x_1,x_2,\cdots,x_n)&=\frac{L(x_1,x_2,\cdots,x_n,\theta_1)}{L(x_1,x_2,\cdots,x_n,\theta_0)}\\&=\left(\frac{\theta_0}{\theta_1}\right)^n\exp\left\{-\left(\frac{1}{\theta_1}-\frac{1}{\theta_0}\right)\sum_{i=1}^n x_i\right\}\end{aligned}$$

$$\begin{aligned}&\lambda(x_1,x_2,\cdots,x_n)\geqslant k\\&\Leftrightarrow\frac{2}{\theta_0}\sum_{i=1}^{n}x_i\leqslant\frac{2}{\theta_0}\left(\frac{1}{\theta_1}-\frac{1}{\theta_0}\right)^{-1}\left(n\ln\frac{\theta_0}{\theta_1}-\ln k\right)\triangleq C\end{aligned}$$

由于

$$T\triangleq\frac{2}{\theta_0}\sum_{i=1}^{n}X_i\sim\chi^2(2n)$$

则最优势检验的拒绝域为

$$W=\left\{(x_1,x_2,\cdots,x_n):T\leqslant\chi_\alpha^2(2n)\right\}$$

■

Definition 6.4.25 (广义的似然比检验 (复合假设))**:::>** 假设

$$H_0:\theta\in\Theta_0\leftrightarrow H_1:\theta\in\Theta_1=\Theta-\Theta_0$$

广义似然比

$$\lambda(x_1,x_2,\cdots,x_n)=\frac{\sup_{\theta\in\Theta}L(x_1,x_2,\cdots,x_n,\theta)}{\sup_{\theta\in\Theta_0}(x_1,x_2,\cdots,x_n,\theta)}$$

拒绝域形式:

$$\begin{aligned}&W=\{(x_1,x_2,\cdots,x_n):\lambda(x_1,x_2,\cdots,x_n)\geqslant C\}\\&P(\lambda(X_1,X_2,\cdots,X_n)\geqslant C)\leqslant\alpha,\forall\theta\in\Theta_0(\text{ under }H_0)\end{aligned}$$

7. Bayesian Inference

7.1. The Bayesian Philosophy

𝔇efinition 7.1.1 (The Bayesian Philosophy)**:::>** 统计学中有两个大的学派: 频率学派 (经典学派) 和 Bayes 学派.

经典统计使用两种信息: 总体信息和样本信息. Bayes 统计认为除了上述两种信息以外, 统计推断还应该使用第三种信息: 先验信息.

Bayes 统计的基本观点是: 任一未知量 θ 都可看作随机变量, 可用一个概率分布去描述, 这个分布称为先验分布; 在获得样本之后, 总体分布, 样本与先验分布通过 Bayes 公式结合起来得到一个关于未知量 θ 新的分布: 后验分布; 任何关于 θ 的统计推断都应该基于 θ 的后验分布进行.

7.2. Bayes 估计

𝔇efinition 7.2.1 (Bayes 公式的密度函数形式)**:::>** 总体依赖于参数 θ 的概率函数在经典统计中记为 $f(x;\theta)$, 它表示参数空间 $\mathbf{\Theta}$ 中不同的 θ 对应不同的分布. 而在 Bayes 统计中应记为 $f(x;\theta)$, 它表示在 θ 取某个给定随机变量值时总体的条件概率函数.

𝔇efinition 7.2.2 (从先验分布到后验分布的步骤)**:::>**

1. 根据参数 θ 的先验信息确定先验分布 $\pi(\theta)$

2. 产生样本. 而从 Bayes 的观点看, 样本 $X=(x_1,\cdots,x_n)$ 的产生要分两步进行, 首先设想从先验分布 $\pi(\theta)$ 产生一个样本 θ_0, 其次从 $f(x\mid\theta_0)$ 中产生一组样本.

   此时样本 $X=(x_1,\cdots,x_n)$ 的联合条件概率函数为

$$f(x\mid\theta_0)=f(x_1,\cdots,x_n\mid\theta_0)=\prod_{i=1}^{n}f(x_i\mid\theta_0)$$

   这个分布综合了总体信息与样本信息.

3. 由于 θ_0 是设想出来的, 仍然是未知的, 它是按先验分布 $\pi(\theta)$ 产生的.

   为把先验信息综合进去, 不能只考虑 θ_0, 对 θ 的其他值发生的可能性也要加以考虑, 故要用 $\pi(\theta)$ 进行综合. 这样一来, 样本 X 和参数 θ 的联合分布为

$$h(X,\theta)=\pi(\theta\mid X)m(X)$$

   其中 $m(X)$ 是 X 的边际概率函数, 即

$$m(X)=\int_{\Theta}h(X,\theta)d\theta=\int_{\Theta}f(X\mid\theta)\pi(\theta)d\theta$$

   它与 θ 无关, 或者说 $m(X)$ 中不含 θ 的任何信息.

   注意我们的目的是要对未知参数 θ 作统计推断. 在没有样本信息时, 我们只能依据先验分布对 θ 作出推断. 在有了样本观察值 $X=(x_1,\cdots,x_n)$ 之后, 我们应依据 $h(X,\theta)$ 对 θ 作出推断, 而又因为 $m(X)$ 中不含 θ 的任何信息的任何信息, 因此能用来对 θ 作出推断的仅是条件分布 $\pi(\theta\mid X)$, 它的计算公式是

$$\pi(\theta\mid X)=\frac{h(X,\theta)}{m(X)}=\frac{f(X\mid\theta)\pi(\theta)}{\int_{\Theta}f(X\mid\theta)\pi(\theta)d\theta}$$

$$\pi(\theta\mid X)\propto f(X\mid\theta)\pi(\theta)$$

   这里的正比关系是因为 X 的边际概率函数 $m(x)$ 与 θ 无关, 所以相当于一个常数, 我们并不关心它的取值.

   这个条件分布称为 θ 的后验分布, 它集中了总体, 样本和先验中有关 θ 的一切信息, 它是用总体和样本对先验分布 $\pi(\theta)$ 作调整的结果, 它要比 $\pi(\theta)$ 更接近 θ 的实际情况.

Definition 7.2.3 (Bayes 估计)**:::>** 由后验分布 $\pi(\theta \mid X)$ 估计 θ 有三种常用的方法:

1. 使用后验分布的密度函数最大值做为 θ 的点估计的最大后验估计
2. 使用后验分布的中位数作为 θ 的点估计的后验中位数估计
3. 使用后验分布的均值作为 θ 的点估计的后验期望估计

用的最多的是后验期望估计, 它一般也简称为 Bayes 估计, 记为 $\widehat{\theta}_B$.

Example 7.2.4:::> 设某事件 A 在一次试验中发生的概率为 θ, 为了估计 θ, 对试验进行了 n 次独立观测, 其中事件 A 发生了 X 次.

显然有 $X \mid \theta \sim b(n,\theta)$, 即

$$P(X=x \mid \theta)=\left(\begin{array}{c} n \\ x \end{array}\right)\theta^x(1-\theta)^{n-x}, \quad x=0,1,\cdots n$$

假若我们在试验前对事件 A 没有什么了解, 就使用均匀分布 $U(0,1)$ 作为 θ 的先验分布, 写出 $\boldsymbol{X}$ 和 θ 的联合分布:

$$L(x_1,\cdots x_n \mid \theta)=\prod_{i=1}^{m} f(x_i \mid \theta)=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

记总体发生次数为 x, 则

$$L(x,\theta)=\left(\begin{array}{c} n \\ x \end{array}\right)\theta^x(1-\theta)^{n-x}, \quad x=0,1;\cdots n, 0<\theta<1$$

可得后验分布:

$$\pi(\theta \mid X) \propto \theta^x(1-\theta)^{n-x}=\theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1}, \quad 0<\theta<1$$

我们敏锐地发现, 这个分布符合 Beta 分布的数学定义式:

$$\theta \mid x \sim \text{Beta}(x+1, n-x+1)$$

由于 $\theta \mid x \sim \mathrm{Be}(x+1, n-x+1)$, 其后验分布期望估计为:

$$\widehat{\theta}_B = E(\theta \mid x) = \frac{x+1}{n+2}$$

Remark. 我们可以通过求出 X 的边际概率函数来获得 θ 后验分布的准确数学表达式:

$$m(x) = \binom{n}{x} \int_0^1 \theta^x (1-\theta)^{n-x} d\theta = \binom{n}{x} \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$$

故 θ 的后验分布为

$$\pi(\theta \mid X) = \frac{h(X,\theta)}{m(X)} = \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} \theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1}, \quad 0 < \theta < 1$$

■

Remark. 假如不用先验信息, 只用总体信息和样本信息, 那么事件 A 发生的概率的最大似然估计为

$$\widehat{\theta}_M = \frac{x}{n}$$

它与 Bayes 估计是不同的两个估计.

某些场合, Bayes 估计要比最大似然估计更合理一点. ■

Example 7.2.5:::> 总体 $X \sim P(\lambda), \lambda > 0$, 先验分布 $\pi(\lambda) \sim E(1)$

后验分布

$$\begin{aligned}
\pi(\lambda \mid x_1, \cdots x_n) &\propto L(x_1, \cdots x_n \mid \lambda)\, \pi(\lambda) \\
&= \prod_{i=1}^{m} f(x_i \mid \lambda) \cdot \pi(\lambda) \\
&= \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \cdot e^{-\lambda} \\
&= \lambda^{\left(\sum_{i=1}^n x_i + 1\right) - 1} e^{-(n+1)\lambda}
\end{aligned}$$

而这正是一个Gamma distribution: $\mathrm{Gamma}(\sum_{i=1}^n x_i + 1, n+1)$. 故有:

$$\widehat{\lambda_B} = \frac{\sum_{i=1}^n x_i + 1}{n+1}$$

Example 7.2.6:::> 设 $x_1,\cdots,x_n$ 是来自正态分布 $N(\mu,\sigma_0^2)$ 的一个样本, 其中 σ_0^2 已知, μ 未知, 假设 μ 的先验分布亦为 $N(\theta,\tau^2)$, 其中先验均值 θ 和先验方差 τ^2 均已知, 试求 μ 的 Bayes 估计.

Solution. 样本 X 的分布和 μ 的先验分布分别为

$$f(X\mid\mu)=\left(2\pi\sigma_0^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)_i^2\right\}$$

$$\pi(\mu)=\left(2\pi\tau^2\right)^{-1/2}\exp\left\{-\frac{1}{2\tau^2}(\mu-\theta)^2\right\}$$

由此可以写出 X 与 μ 的联合分布

$$h(X,\mu)=k_1\cdot\exp\left\{-\frac{1}{2}\left[\frac{n\mu^2-2n\mu\overline{x}+\sum_{i=1}^{n}x_i^2}{\sigma_0^2}+\frac{\mu^2-2\theta\mu+\theta^2}{\tau^2}\right]\right\}$$

其中 $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $\quad k_1=(2\pi)^{-(n+1)/2\tau^{-1}\sigma_0^{-n}}$.

若记

$$A=\frac{n}{\sigma_0^2}+\frac{1}{\tau^2},\quad B=\frac{n\overline{x}}{\sigma_0^2}+\frac{\theta}{\tau^2},\quad C=\frac{\sum_{i=1}^{n}x_i^2}{\sigma_0^2}+\frac{\theta^2}{\tau^2}$$

则有

$$\begin{aligned}h(X,\mu)&=k_1\exp\left\{-\frac{1}{2}\left[A\mu^2-2B\mu+C\right]\right\}\\&=k_1\exp\left\{-\frac{(\mu-B/A)^2}{2/A}-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\end{aligned}$$

注意到 A,B,C 均与 μ 无关, 由此容易算得样本的边际密度函数

$$m(X)=\int_{-\infty}^{+\infty}h(X,\mu)d\mu=k_1\exp\left\{-\frac{1}{2}\left(C-B^2/A\right)\right\}(2\pi/A)^{1/2}$$

应用 Bayes 公式即可得到后验分布

$$\pi(\mu\mid X)=\frac{h(X,\mu)}{m(X)}=(2\pi/A)^{1/2}\exp\left\{-\frac{1}{2/A}(\mu-B/A)^2\right\}$$

这说明在样本给定后, μ 的后验分布为 $N(B/A,1/A)$, 即

$$\mu\mid M\sim N\left(\frac{n\overline{x}\sigma_0^{-2}+\theta\tau^{-2}}{n\sigma_0^{-2}+\tau^{-2}},\frac{1}{n\sigma_0^2+\tau^{-2}}\right)$$

后验均值即为其 Bayes 估计:

$$\widehat{\mu}=\frac{n/\sigma_0^2}{n/\sigma_0^2+1/\tau^2}\overline{x}+\frac{1/\tau^2}{n/\sigma_0^2+1/\tau^2}\theta$$

它是样本均值 $\overline{x}$ 与先验均值 θ 的加权平均. 当总体方差 σ_0^2 较小或样本量 n 较大时, 样本均值 $\overline{x}$ 的权重较大; 当先验方差 τ^2 较小时, 先验均值 θ 的权重较大, 这一综合很符合人们的经验, 也是可以接受的. ■

Definition 7.2.7 (共轭先验分布)**:::>** 设 θ 是总体参数, $\pi(\theta)$ 是其先验分布, 若对任意的样本观测值得到的后验分布 $\pi(\theta\mid X)$ 与 $\pi(\theta)$ 属于同一个分布族, 则称该分布族是 θ 的共轭先验分布 (族).

Theorem 7.2.8 (共轭先验分布列表)**:::>**

1. 总体分布 $f(x\mid\theta)$:

   $$N\left(\mu,\sigma^2\right)$$

   其中 σ^2 已知, $\theta=\mu$.

   先验分布 $\pi(\theta)$:

   $$\pi(\mu)\leftrightarrow N\left(\alpha,\tau^2\right)$$

   后验分布 $\pi(\theta\mid x)$:

   $$N\left(\frac{n\overline{x}\sigma^{-2}+\alpha\tau^{-2}}{n\sigma^{-2}+\tau^{-2}},\frac{1}{n\sigma^{-2}+\tau^{-2}}\right)$$

2. 总体分布 $f(x\mid\theta)$:

   $$\Gamma(\alpha,\beta)$$

   其中 α 已知, $\theta=\beta$.

   先验分布 $\pi(\theta)$:

   $$\pi(\beta)\leftrightarrow\Gamma(\alpha_0,\beta_0)$$

   后验分布 $\pi(\theta\mid x)$:

   $$\Gamma\left(\alpha_0+\alpha n,\beta_0+\sum_{i=1}^{n}x_i\right)$$

3. 总体分布 $f(x \mid \theta)$:

$$B(n_0, p)$$

其中 $\theta = p$.

先验分布 $\pi(\theta)$:

$$\pi(p) \leftrightarrow \text{Beta}(a, b)$$

后验分布 $\pi(\theta \mid x)$:

$$\text{Beta}\left(a + \sum_{i=1}^{n} x_i, b + nn_0 - \sum_{i=1}^{n} x_i\right)$$

4. 总体分布 $f(x \mid \theta)$:

$$P(\lambda)$$

其中 $\theta = \lambda$.

先验分布 $\pi(\theta)$:

$$\pi(\lambda) \leftrightarrow \Gamma(\alpha, \beta)$$

后验分布 $\pi(\theta \mid x)$:

$$\Gamma\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n\right)$$

5. 总体分布 $f(x \mid \theta)$:

$$NB(r, p)$$

其中 $\theta = p$.

先验分布 $\pi(\theta)$:

$$\pi(p) \leftrightarrow \text{Beta}(a, b)$$

后验分布 $\pi(\theta \mid x)$:

$$\text{Beta}\left(a + nr, b + \sum_{i=1}^{n} x_i - nr\right)$$

6. 总体分布 $f(x \mid \theta)$:

$$N\left(\mu, \sigma^2\right)$$

其中 μ 已知, $\theta = \sigma^2$.

先验分布 $\pi(\theta)$:

$$\pi\left(\sigma^2\right) \leftrightarrow IGa(\alpha, \lambda)$$

后验分布 $\pi(\theta \mid x)$:

$$IGa\left(\alpha + \frac{n}{2}, \lambda + \sum_{i=1}^{n}(x_i - \mu)^2\right)$$

其中 $X \sim IGa(\alpha, \lambda)$ 为" 逆 Gamma 分布":

若随机变量 ξ 的密度函数为:

$$f(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\left(\frac{1}{x}\right)^{\alpha+1} e^{-\frac{\lambda}{x}}, x > 0$$

则

$$\xi \sim \Gamma(\alpha, \lambda) \Leftrightarrow \xi^{-1} \sim IGa(\alpha, \lambda)$$