概率论

作者：Kamden Wang

时间：December 15, 2022

主要参考资料：

《First Course in Probability》Sheldon Ross,

《概率论基础》李贤平,

《概率论指导书》李贤平,

台湾清华大学郑少为课程

目录

第 1 章 Combinatorial Analysis(组合分析)

1.1 Summary

1. 基本加法规则和乘法规则
2. 排列, 组合与多项式系数
3. 整数方程解的个数

Equations from Theoretical Exercises

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0} \tag{1.1}$$

$$\binom{2n}{n} = \sum_{k=0}^{n} \binom{n}{k}^2 \tag{1.2}$$

$$\sum_{k=1}^{n} \binom{n}{k} k^3 = 2^{n-3} n^2 (n+3) \tag{1.3}$$

$$\sum_{j=i}^{n} \binom{n}{j}\binom{j}{i} = \binom{n}{i} 2^{n-i} \quad i \le n \tag{1.4}$$

$$\sum_{j=i}^{n} \binom{n}{j}\binom{j}{i} (-1)^{n-j} = 0 \quad i < n \tag{1.5}$$

第 2 章 Axioms of Probability

2.1 Sample Space and events

定理 2.1 (DeMorgan)

$$(\bigcap_{i=1}^{n} E_i)^c = \bigcup_{i=1}^{n} (E_i)^c$$

♡

2.2 Axioms of Probability

公理 2.1

1. $$0 \le P(E) \le 1$$
2. $$P(S) = 1$$
3. $$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$$

♡

2.3 Some Simple propositions

命题 2.1

1. $$P(E^c) = 1 - P(E)$$
2. If $E \subset F$, then $P(E) \le P(F)$
3. $$P(E \cup F) = P(E) + P(F) - P(EF)$$
4. $$P(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i) - \sum_{i_1<i_2} P(E_{i_1}E_{i_2}) + \cdots + (-1)^{r+1} \sum_{i_1<i_2<\cdots i_r} P(E_{i_1} \cdots E_{i_r}) + \cdots + (-1)^{n+1} P(E_1 E_2 \cdots E_n)$$

♠

2.4 Sample Spaces having equally likely outcomes

$$P(E) = \frac{\text{\# outcomes in E}}{\text{\# outcomes in S}} \tag{2.1}$$

例题 2.1 If n people are present in a room, what is the probability that no two of them celebrate their birthday on the same day of the year? How large need n be so that this probability is less than 1/2?

解 由于每个人的生日可能在一年三百六十五天中任意一天, 那么一共就有 $(365)^n$ 种可能性结果 (这里忽略了闰年的情况) 假设每一种结果出现的可能性都是相同的, 那么所求的概率即为

$$\frac{(365)\times(364)\times\cdots\times(365-n+1)}{(365)^n}$$

事实是当 $n\leq 23$ 时, 这个概率就会小于 $\frac{1}{2}$ 也就是说只要有 23 个或以上的人在房间, 那么这个概率就会超过 $\frac{1}{2}$, 很多人在最初都会对这个结果感到惊讶, 因为 $23 << 365$, 事实上当有 50 个人的时候这个概率接近 97%, 当有 100 个人在房间里时这个概率几乎就是 1

例题 2.2 A deck of 52 playing cards is shuffled(洗), and the cards are turned up one at a time until the first ace appears. Is the next card, the card following the first ace more likely to be the ace of spades or the two of clubs?

解 要确定第一张 A 后面的牌是黑桃 A 的概率, 我们需要计算一副有 (52)! 排列方式的牌紧随第一个 A 之后是黑桃 A. 首先, 注意到 52 张牌的每个顺序都可以通过首先对不同于黑桃 A 的 51 张牌进行排序, 然后将黑桃 A 插入该顺序获得. 更进一步地,51 张牌排序 (51)! 种中只有第一个 A 后的一个空给黑桃 A . 因此

$$P\{\text{the ace of spades follow the first ace}\}=\frac{(51)!}{(52)!}=\frac{1}{52}$$

梅花二的情况同上, 也就是说 52 张牌被抽到的可能性是相同的. 这说明了抽签顺序与抽签结果无关.

例题 2.3 Suppose that each of N men at a party throws his hat into the center of the room.The hats are first mixed up, and then each man randomly selects a hat. What is the probability that none of the men selects his own hat?

解 我们首先计算相反的情况, 即至少有一个人选到了自己的帽子, 我们记事件 E_i $i=1,2,\cdots,N$ 为第 i 第 i 个人选到了自己的帽子. 根据**命题四**, 至少一个一个人选到自己的帽子的概率为

$$P(\bigcup_{i=1}^{N}E_i)=\sum_{i=1}^{N}P(E_i)-\sum_{i_1<i_2}P(E_{i_1}E_{i_2})+\cdots+(-1)^{n+1}\sum_{i_1<i_2<\cdots i_n}P(E_{i_1}\cdots E_{i_n})+\cdots$$
$$+(-1)^{N+1}P(E_1E_2\cdots E_N)$$

把结果看成一个 n 维向量, 例如 $(1,2,3,\cdots,N)$ 就代表每一个人选到了自己的帽子, 那么情况一共有 $N!$, 更进一步地, $E_{i_1}E_{i_2}\cdots E_{i_n}$ 就是有 n 个人接到了自己的帽子, 剩下的 $N-n$ 个人有 $(N-n)!$ 种结果

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

并且, 这样的选法一共有 $\begin{pmatrix}N\\n\end{pmatrix}$ 种, 因此

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

综上,

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

那么没人拿到自己帽子的概率就是

$$1-P(\bigcup_{i=1}^{N}E_i)$$

根据数学分析的知识可以知道当 N 充分大充分大的时候这个值接近于

$$e^{-1}\approx 0.37$$

2.5 Probability as continuous set function

定义 2.1

若事件序列 $\{E_n, n \geq 1\}$ 满足

$$E_1 \subset E_2 \subset \cdots \subset E_n \subset E_{n+1} \subset \cdots$$

那么该事件序列是增序列, 降序列同理

♣

定理 2.2

若一个事件序列是增/降序列, 那么

$$\lim_{n\to\infty} E_n = \bigcup_{i=1}^{\infty} E_i$$

同时

$$\lim_{n\to\infty} P(E_n) = P(\lim_{n\to\infty} E_n)$$

♡

例题 2.4 (Ross-Littlewood paradox)(球与花瓶悖论)
有一个无限大的花瓶和无穷多的球, 把球记作 $1, 2, 3, \cdots, n, \cdots$. 考虑一个实验, 在正午 12:00 前一分钟将 1-10 号球放入花瓶, 并把 10 号球取出来 (假设取出来不花费任何时间), 在正午 12:00 前 1/2 分钟将 11-20 号球放入花瓶, 并把 20 号球取出来, 这样一直做下来, 最后的问题是到正午 12:00 时花瓶里还有多少个球.

解 很多人认为这个问题的答案很明显，在中午 12 点，瓶中有无限多个球，因为任何编号不为 10n,n>=1，形式的球，都会被放入瓶中，并且不会被取出中午 12 点之前撤回因此，当按照所述进行实验时，问题就解决了。事实上概率为 0

We shall show that, with probability 1, the urn is empty at 12 P.M. Let us first consider ball number 1. Define En to be the event that ball number 1 is still in the urn after the first n withdrawals have been made. Clearly,

$$P(E_n) = \frac{9 \times 18 \times 27 \times (9n)}{10 \times 19 \times 28 \times (9n+1)}$$

[To understand this equation, just note that if ball number 1 is still to be in the urn after the first n withdrawals, the first ball withdrawn can be any one of 9, the second any one of 18 (there are 19 balls in the urn at the time of the second withdrawal, one of which must be ball number 1), and so on. The denominator is similarly obtained.]

Now, the event that ball number 1 is in the urn at 12 P.M. is just the event $\bigcap_{n=1}^{\infty} E_n$. Because the events E_n $n \geq 1$ are decreasing events, it follows from Theorem 2.2 that

$$P\{\text{ball number 1 is in the urn at 12 P.M.}\} = P(\bigcap_{n=1}^{\infty} E_n) = \lim_{n\to\infty} P(E_n) = \prod_{n=1}^{\infty} (\frac{9n}{9n+1})$$

We now show that $\prod_{n=1}^{\infty} (\frac{9n}{9n+1}) = 0$

Since $\prod_{n=1}^{\infty} (\frac{9n}{9n+1}) = [\prod_{n=1}^{\infty} (\frac{9n+1}{9n})]^{-1}$ this is equivalent to showing that

$$\prod_{n=1}^{\infty} (1 + \frac{1}{9n}) = \infty$$

Now, for all $m \geq 1$

$$\prod_{n=1}^{\infty} (1+\frac{1}{9n}) \geq \prod_{n=1}^{m} (1+\frac{1}{9n}) = (1+\frac{1}{9}) \times (1+\frac{1}{18}) \times \cdots \times (1+\frac{1}{9m}) > \frac{1}{9} + \frac{1}{18} + \cdots + \frac{1}{9m} = \frac{1}{9} \sum_{i=1}^{m} \frac{1}{i} \to \infty \quad (m \to \infty)$$

Thus, letting F_i denote the event that ball number i s in the urn at 12 P.M. we have shown that $P(F_1) = 0$. Similarly, we can show that $P(F_i) = 0$ for all i .

第 3 章 Conditional Probability and Independence

For an event, new information (i.e. some other event has occurred) could change its probability.
We call the altered probability a **conditional probability**.

3.1 Conditional Probabilities

定义 3.1 (条件概率)

$$P(E|F) = \frac{P(EF)}{P(F)}$$

♣

推论 3.1 (条件概率的乘法法则)

$$P(E_1E_2E_3\cdots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2)\cdots P(E_n|E_1\cdots E_{n-1})$$

♡

例题 3.1 在例题 2.3 中我们知道了 N 个人中没有人选到自己的帽子的概率是

$$P_N = \sum_{i=0}^{N}(-1)^i/i!$$

那么 N 个人中恰好有 k 个人选到了自己的帽子的概率是多少?

解 记事件 E 为 k 个人每个人都选到了自己的帽子, 设事件 G 为剩下的 $N-k$ 没有一个人选到了自己帽子, 那么根据条件概率公式有

$$P(EG) = P(E)P(G|E)$$

设事件 $F_i, i = 1, \cdots k$ 为第 i 个人选到了自己的帽子, 那么

$$P(E) = P(F_1F_2\cdots F_k) = P(F_1)P(F_2|F_1)P(F_3|F_1F_2)\cdots P(F_k|F_1\cdots F_{k-1}) = \frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-k+1} = \frac{(N-k)!}{N!}$$

对于剩下的 $N-k$ 个人, 我们只需要运用最开始的公式

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

考虑到 k 个人组合有 $\begin{pmatrix} N \\ k \end{pmatrix}$ 种, 那么

$$P(EG) = \frac{P_{N-k}}{k!} \approx \frac{e^{-1}}{k!} \quad (\text{当 N 足够大})$$

3.2 Bayes's Formula

设 E 和 F 两个事件, 我们可以把 E 表示为

$$E = EF \cup EF^c$$

由于分解出来的两个事件一定是互斥的, 根据第三条概率公理我们有

$$P(E) = P(EF) + P(EF^c) = P(E|F)P(F) + P(E|F^c)P(F^c) = P(E|F)P(F) + P(E|F^c)(1-P(F))$$

这个公式阐述了事件 E 的概率可以分为两个条件概率, 这个公式在概率论是最为常用的公式, 没有之一, 我们也称之为**全概率公式**.

定理 3.1 (全概率公式)

$$P(E)=\sum_i P(E|F_i)P(F_i)$$

♡

教材上给出了很多的例子, 可以自行参考

定义 3.2

(比率) 事件 A 的比率定义为

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

也就是说一个事件的比率告诉了我们事件 A 发生的可能性大还是不发生的可能性大.

♣

笔记 probability, likelihood, odds. 我理解三者的差别就是:probability 即这门课中用到的最多的概念, 由概率公理为基础得来, 一定是小于等于一的; likelihood 是很泛泛的概率, 即定性估计, 可能性大或者小, odds 就是比率可以大于一也可以小于一.

推论 3.2

$$\frac{P(H|E)}{P(H^c|E)}=\frac{P(H)}{P(H^c)}\frac{P(E|H)}{P(E|H^c)}$$

♡

下面我们来看 Bayes's 公式. 令 F_i 是互斥的事件, 假设现在事件 E 发生了, 现在我们对 F_i 中哪个发生了感兴趣, 根据全概率公式我们就能得出 Bayes's 公式.

定理 3.2 (Bayes)

$$P(F_j|E)=\frac{P(EF_j)}{P(E)}=\frac{P(E|F_j)P(F_j)}{\sum_{i=1}^n P(E|F_i)P(F_i)}$$

♡

3.3 Independent Events

条件概率中 $P(E|F)$ 通常不等于 $P(E)$, 现在我们研究特殊情况, 即它们两者相等.

定义 3.3 (独立)

我们称两个事件相互独立, 通俗来说, 事件 F 的发生不会影响事件 E 的发生. 根据条件概率我们可以推出两个事件独立的条件是

$$P(EF)=P(E)P(F)$$

♣

现在我们可以证明若 E 和 F 是独立的, 那么 E 和 F^c 也是独立的

命题 3.1

若 E,F 相互独立, 那么 E,F^c 相互独立.

♠

证明

$$E=EF\cup EF^c$$

$$P(E) = P(EF) + P(EF^c) = P(E)P(F) + P(EF^c) \Rightarrow P(EF^c) = P(E)(1 - P(F)) = P(E)P(F^c)$$

下面一个例子在概率理论史上占据了光辉的地位. 这是著名的点数问题 (problem of the points). 通常来讲, 这个问题是:两个玩家给出一定的筹码进行游戏, 筹码将全部给赢家. 但是一个意外使它们在决出胜负前必须暂停游戏, 此时两个玩家已经有了优势和劣势那么该如何分配筹码?

这个问题是由一个专业赌徒提出给法国数学家 Pascal 的, 为了攻克这个问题, Pascal 提出了这样一个重要的想法，即竞争对手应得的奖品比例应依靠他们各自的赢得胜利概率，如果该比赛要继续进行。帕斯卡（Pascal）制定了一些特殊案例，更重要的是，与著名的法国人皮埃尔 · 德 · 费马特（Pierre de Fermat）进行了联系，后者以数学家的身份享有很高的声誉。由此产生的想法交换不仅为要点问题提供了完整的解决方案，而且还为解决与机会游戏有关的许多其他问题的解决方案奠定了框架。这种庆祝的信件被某些人作为概率理论的出生日期，对于欧洲数学家对概率的兴趣也很重要，因为帕斯卡和费马都被认为是当时最重要的数学家之一。例如，在往来的短时间内，年轻的荷兰数学家克里斯蒂亚 · 霍根斯（Christiaan Huygens）来到巴黎讨论这些问题和解决方案，以及这个新领域的兴趣和活动迅速增长。

例题 3.2 (**The problem of the points**) ross p82
胜利的概率为 p, 那么 n 次胜利在 m 次失败前发生的概率为

$$P_{n,m} = \sum_{k=n}^{m+n-1} \binom{m+n-1}{k} p^k (1-p)^{m+n-1-k}$$

帕斯卡的解法:

$$P_{n,m} = p_{n-1,m} + (1-p)P_{n,m-1} \qquad n \geq 1, m \geq 1$$

费马的解法: 得到这个结果的充要条件是在前 $m+n-1$ 次试验中至少胜利了 n 次, 那么就可以由二项式定理解决.

下面来看两个赌博问题, 第一个有十分优雅的分析.
例题 3.3 ross p83
先看特殊情况, 有 n 个人, 每个人的初始点数都是 1, 那么每个人获胜的概率都是相同的, 把每个人分到有不同人数容量的小组, 那么这个小组获胜的概率就与人数有关, 再把这个小组看成一个点数与小组人数相同的人, 这样就能得到某个人获胜的概率, 也可以得出结果与人数是无关的.

例题 3.4 (**The gambler's ruin problem**) ross p83

3.4 $P(\cdot|F)$ is a Probability

条件概率满足概率的所有性质

例题 3.5 (**Laplace's rule of succession**) ross p95

例题 3.6 (**Updating information sequentially**) ross p96

第 4 章 Discrete Random Variables

4.1 Random Variables

标准的随机变量的定义建立在测度之上. 通俗说就是一个映射将样本空间中的变量映射上实数轴上.

4.2 Discrete Random Variables and its properties

定义 4.1 (Discrete Random Variable)

For a random variable X, let $\chi = \{X\omega : \omega \in \Omega\}$, be the range of X. Then X is called discrete if χ is a finite or countably infinite set.

♣

笔记 Commonly used tools to define the probability measures of discrete random variables:

1. Probability mass function
2. Cumulative distribution function
3. Moment generating function (Chapter 7)

When one of them is known, the remaining can be derived from it. The first one function can only be defined to discrete random variables. The remaining two can also define for continuous random variables.

定义 4.2 (Probability mass function)

If X is a discrete random variable. Then the probability mass function of X is defined by

$$f_X(x) = P_X(\{X = x\})$$

for $x \in R$

♣

Notice that the graph of the probability mass function only has some values in some point. And you can think of these values as the mass of the corresponding points.

定理 4.1 (The property of probability mass function)

If f_X is the probability mass function of a discrete random variable X with range χ, then

1. $f_X(x) \geq 0$, for all $x \in R$
2. $f_X(x) = 0$, for $x \notin \chi$
3. $\sum_{x\in\chi} f_X(x) = 1$
4. for $A \subset R$, $\quad P_X(X \in A) = \sum_{x\in A\cap\chi} f_X(x)$

♡

We can define probability mass function as any function that satisfies 1,2,3.

定义 4.3 (Cumulative distribution function)

A function $F_X : R \to R$ is called the cumulative distribution function of a random variable X if $F_X(x) \leq P_X(X \leq x), x \in R$.

♣

Notice that the cumulative distribution function has following properties.

定理 4.2 (The property of cumulative distribution function)

If $F_X x$ is the cumulative distribution function of a random variable X, then it must satisfy the following properties:

1. $0 \leq F_X(x) \leq 1$
2. $F_X(x)$ is nondecreasing, i.e. $F_X(A) \leq F_X(b)$ for $a \leq b$.
3. For any $x \in R$, $F_X(x)$ is continuous from the right, i.e. $F_X(x) = F_X(x^+) = \lim_{t \to x^+} F_X(t)$
4. $\lim_{x \to +\infty} F_X(x) = 1 \quad \lim_{x \to -\infty} F_X(x) = 0$
5. $P_X(X > x) = 1 - F_X(x)$ and $P_X(a < X \leq b) = F_X(b) - F_X(a)$
6. If X is discrete with probability mass function f_X, then for $x \in R$,
$$F_X(x) = \sum_{\substack{x_i \in \chi \\ x_i \leq x}} f_X(x_i), \quad f_X(x) = F_X(x) - F_X(x^-)$$
7. F_X has at most countably many discontinuity points.

♡

If a function F satisfies 2,3,4, then F is a cumulative distribution function of some random variable. We can think of the relationship between cumulative distribution function and probability mass function as the relationship between step function and delta function.

定理 4.3 (Transformation)

Let X be a discrete random variable with range χ and probability mass function f_X; let $Y = g(X)$, an another discrete random variable, and the probability mass function of Y is

$$f_Y(y) = \sum_{\substack{x \in \chi \\ g(x) = y}} f_X(x)$$

♡

例题 4.1 If $Y = X^2$, then $f_Y(y) = f_X(\sqrt{y}) + f_X(\sqrt{-y})$

4.3 Expected Value(Mean)

We often characterize a person by his/her height, weight, hair,.... How can we "roughly" characterize a distribution? We introduce the concept of mean.

定义 4.4

If X is a discrete random variable with probability mass function f_X and range χ, then the expectation (or expected value) of X is

$$E[x] = \sum_{x \in \chi} x f_X(x)$$

provided that the sum converges absolutely. i.e. $\sum |x| f_X(x) < \infty$

♣

例题 4.2 If all value in χ are equally likely, then $E[x]$ is simply the average of the possible values of X.

例题 4.3 (Indicator Function)

For an event $A \subset \Omega$, the indicator function of A is the random variable:

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

Its range χ is $\{0, 1\}$ and its probability mass function is

$$f(0) = P(A^c) = 1 - P(A) \quad f(1) = P(A)$$

for a probability measure P defined on Ω. So, $E[\mathbf{1}_A] = 0 \cdot [1 - P(A)] + 1 \cdot P(A) = P(A)$

Let's see the intuitive interpretation of Expectation

1. Expectation of a random variable parallels the notion of a weighted average, where more likely values are weighted higher than less likely values.
2. It is helpful to think of the expectation as the "center" of mass of the probability mass function. (center of gravity)
3. Expectation can be interpreted as a long-run average(because of Law of Large Number, Chapter 8)

定理 4.4 (Expectation of Transformation)

If X is discrete random variable with range χ and probability mass function f_X; let $Y = g(X)$, then

$$E[Y] = \sum_{x \in \chi} g(x) f_X(x)$$

provided that the sum converges absolutely.

♡

Let's see a special and useful transformation – linear transformation.

定理 4.5 (Linear Transformation of Expectation)

For $a, b \in R$, then

$$E[aX + b] = a \cdot E[X] + b$$

♡

4.4 Mean and Variance

定义 4.5 (Mean and Variance)

The expectation of X is also called the mean of X and/or f_X. The variance of X (and/or f_X) is defined by

$$\sigma_X{}^2 = Var[X] = E[(x - \mu_X)^2] = \sum_{x \in \chi} (x - \mu_X)^2 f_X(x)$$

μ_X denotes $E[x]$ and $\sigma_X{}^2$ denotes $Var[X]$. Also, σ_X is called the standard deviation of X.

♣

Note that μ_X and $\sigma_X{}^2$ only depends on f_X. They are fixed constants, not random numbers. And if X has units, then μ_X and σ_X have the same unit as X. and variance has unit squared.

Let's see the intuitive interpretation of Variance.

1. Variance is the weighted average value of the squared deviation of X from μ_X.
2. Variance is related to how the probability mass function is spread out.

性质 There are some properties of variance.

1. The variance of a random variable is always non-negative
2. The only random variable with variance equal 0 is a random variable which can only take on a singe value (μ_X). i.e. $Var[a] = 0$. And we call a degenerate random variable.

定理 4.6 (Linear Transformation of Variance)

For $a, b \in R$, then

$$Var[aX + b] = a^2 Var[X]$$

for standard deviation we have $\sigma_{aX+b} = |a|\sigma_X$

♡

定理 4.7 (Mean Square Error)

If X is a (discrete) random variable with mean μ_X, then for any $c \in R$,

$$E[(X-c)^2] = {\sigma_X}^2 + (c-\mu_X)^2$$

♡

推论 4.1

$E[(X-c)^2]$ is minimized by letting $c = \mu_X$, and the minimum value is ${\sigma_X}^2$

♡

We don't use the definition of the variance to compute the value of the variance. Instead we compute the variance by a very useful formula below.

推论 4.2

$$ {\sigma_X}^2 = E[X^2] - (E[X])^2$$

♡

We call $E[X^n]$ the $\underline{n^{th} moment}$ of X. Note that expectation is the first moment, and variance equals to the second moment minus the square of the first moment.

4.5 The Bernoulli and Binomial Distribution

Bernoulli distribution is also called "two point distribution", "degenerate binomial distribution". Let's first see the probability mass function.

Let $A_1, \cdots, A_n$ be independent events and $P(A_i) = p,\ i = 1, 2, \cdots, n$.
Let $X = \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}$.
Then, for $k = 0, 1, \cdots, n$,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The distribution of the random variable X is called the binomial distribution with parameters n and p. In particular, when $n = 1$, it is called the Bernoulli distribution with parameter p. Notice that a binomial random variable can be regarded as the sum of n independent Bernoulli random variables. The binomial distribution is called after the Binomial Theorem:

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

We want to study the mean and variance of the Binomial(n, p) distribution.

定理 4.8

The mean and variance of the Binomial(n, p) distribution are

$$\mu = np \qquad \sigma^2 = np(1-p)$$

♡

证明 We first look at the mean value.

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

Then, we study the variance, here we have a tricky method.

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

We know

$$Var[X] = E[x^2] - (E[x])^2 = (E[X^2] - E[x]) + E[X] - (E[x])^2 = n(n-1)p^2 + np + n^2p^2 = np(1-p)$$

4.6 The Poisson Distribution

It's very difficult to compute the binomial distribution when n has a big scale. So here comes Poisson distribution to approximately represent the binomial distribution. The distribution is named after Simeon Poisson, who derived the approximation of Poisson probability mass function to binomial probability mass function **When n large, $n >> k$, and $p_n \approx 0$**, we have

$$\binom{n}{k}p_n^k(1-p_n)^{n-k} \approx \frac{1}{k!}\lambda^k e^{-\lambda}$$

and the latter is exactly the probability mass function of Poisson distribution. And in this case, we take the parameter as follows

$$p_n \approx \frac{\lambda}{n} \Rightarrow \lambda \approx np_n = E[X_n]$$

The $\lambda(\approx np_n)$ can be interpreted as the average occurrence frequency.

As usual we study the mean value and variance of the Poisson distribution.

定理 4.9

The mean and variance of Poisson(λ) are

$$\mu = \lambda \qquad \sigma^2 = \lambda$$

♡

证明 We first look at the mean value using the same way in binomial distribution

$$E[x] = \sum_{n=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \lambda$$

For the variance, again we use the same tricky.

$$E[X(X-1)] = E[X^2 - X] = E[X^2] - E[X] = \lambda^2 \sum_{x=2}^{\infty} \frac{e^{\lambda}\lambda^{x-2}}{(x-2)!} = \lambda^2$$

And

$$Var[X] = E[x^2] - (E[x])^2 = (E[X^2] - E[x]) + E[X] - (E[x])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Poisson Process is a stochastic process related to Poisson distribution

定义 4.6 (Poisson Process)

A Poisson process with rate λ is a family of random variables N_t, $0 \leq t \leq \infty$, for which $N_0 = 0$ and $N_t - N_s \sim$Poisson $(\lambda(t-s))$, for $0 \leq s < t < \infty$, and

$$N_{t_i} - N_{s_i}, i = 1, 2, \cdots, m$$

are independent whenever

$$0 \leq s_1 < t_1 \leq s_2 < t_2 \leq \cdots \leq s_m < t_m$$

♣

4.7 Other Discrete Probability Distributions

We simply look at some other discrete probability distributions: Negative binomial distribution(Pascal distribution), Geometric distribution, Hypergeometric distribution.

Let's first see the first two distributions. The probability mass function of the negative binomial distribution is

$$P(Y_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$$

with parameters r, p. In particular, when $r = 1$, it is called geometric distribution with parameter p. And the geometric distribution has an interesting property – "memoryless" i.e. $P(Y_1 > s + t | Y_> s) = P(Y_1 > t)$. And a negative binomial random variable can be regarded as the sum of independent geometric random variables. The negative binomial distribution is called after the Negative Binomial Theorem:

$$\frac{1}{(1-t)^r} = \sum_{k=0}^{\infty} \binom{r+k-1}{k} t^k, \ |t| < 1$$

And the geometric distribution is called after the geometric sequence.

Let's look at the mean and variance.

定理 4.10

The mean and variance of negative binomial (r, p) are

$$\mu = \frac{r}{p} \qquad \sigma^2 = \frac{r(1-p)}{p^2}$$

♡

We move on to the Hypergeometric distribution. Consider a experiment – Draw a sample of $n (n \leq N)$ ball without replacement from a box containing R red balls and $N - R$ white balls.

Let X be the number of red balls in the sample. We want to know $P(X = k)$. Note that if drawn with replacement, then the distribution of X is a binomial distribution.

The probability mass function of the hypergeometric distribution is as follows

定理 4.11

For $k = 0, 1, 2, \cdots, n$,

$$P(X = k) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}$$

♡

The hypergeometric distribution is called after the hypergeometric identity:

$$\binom{a+b}{r} = \sum_{k=0}^{r} \binom{a}{k}\binom{b}{r-k}$$

定理 4.12

The mean and variance of hypergeometric (n, N, R) are

$$\mu = \frac{nR}{N} \qquad \sigma^2 = \frac{nR(N-R)(N-n)}{N^2(N-1)}$$

♡

The next theorem illustrates the relationship between hypergeometric distribution and the binomial distribution.

定理 4.13

Let $N_i \to \infty$ and $R_i \to \infty$ in such a way that $p_i = R_i/N_i \to p$ where $0 \le p \le 1$, then

$$\frac{\binom{R_i}{k}\binom{N_i - R_i}{n-k}}{\binom{N_i}{n}} \to \binom{n}{k} p^k (1-p)^{n-k}$$

♡

Intuition: When # of red and white balls are very large, n relatively small, without replacement $\approx$ with replacement.

In fact, these common discrete distribution is all derived from Bernoulli distribution.

第 5 章 Continuous Random Variables

5.1 Probability Density Function

For discrete random variables, only a finite or countably infinite number of possible values with positive probability.

Often, there is interest in random variables that can take (at least theoretically) on an uncountable number of possible values.

There are some properties about the continuous distribution

性质

1. $P_X(X = x) = P(x) = 0$, for any $x \in R$. i.e. Probability for X to take any single value is 0.
2. But, for $-\pi \le a < b \le \pi$, $P_X(X \in (a, b]) = P((a, b]) = \frac{b-a}{2\pi} > 0$. i.e. Positive probability is assigned to any $(a, b]$

Although we can't define a probability mass function in continuous random variables compared to discrete random variables, we can define a new function which is similar to probability mass function using calculus.

定义 5.1 (Probability Density Function)

A function $f : R \to R$ is called a probability density function if

1. $f(x) \ge 0$, for all $x \in (-\infty, +\infty)$
2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

♣

Note that f is not necessary to be a continuous function.

Now, we can define a continuous random variables using probability density function.

定义 5.2 (Continuous Random Variable)

A random variable X is called continuous if there exists a probability density function f such that for any set B of real numbers

$$P_X(X \in B) = \int_B f(x)\,dx$$

i.e. $P_X(a \le X \le b) = \int_a^b f(x)\,dx$

♣

We can get some properties of the probability density function

1. $P_X(X = x) = \int_x^x f(s)\,ds = 0$ for any $x \in R$
2. It does not matter whether the intervals are open or close, i.e.

$$P(X \in [a, b]) = P(X \in (a, b]) = P(X \in [a, b)) = P(X \in (a, b)).$$

3. It is important to remember that the value of a probability density $f(x)$ is NOT a probability itself. It just the "density" over there. $f(x)\,dx$ is a measure of how likely it is that X will be near x.
4. It is quite possible for a probability density function have greater than 1 while the probability mass function is always less or equal to 1.

Note that the probability density function is not unique, you can change the value of some countable points. But the probability mass function is unique, you can change the value of some point or you are changing the probability distribution.

We now look at the relationship between probability density function and its cumulative distribution function with

the following theorem.

定理 5.1

If F_x and f_X are the cumulative distribution function and the probability density function of a continuous random variable X, respectively, then

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)\,dy \qquad f_X(x) = F'(x) = \frac{d}{dx}F_X(x)$$

♡

The cumulative distribution function for continuous random variables has the same interpretation and property as discussed in the discrete case. The only difference is in plotting F_X. In discrete case, there are jumps (step function). In continuous case, F_X is a (absolutely) continuous non-decreasing function and it is differentiable almost everywhere.

Let's look at the transformation. We want to find the distribution of $Y = g(x)$. Suppose that X is a continuous random variable with cumulative distribution function F_X and probability density function f_X. Consider $Y = g(X)$, where g is a strictly monotone(increasing or decreasing) function. Let R_Y be the range of g. Note that any strictly monotone function has an inverse function. The cumulative distribution function of Y is denoted by F_Y. Then

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

Here we suppose $g(x)$ is a strictly increasing function. And if $g(x)$ is a strictly decreasing function we can get

$$F_Y(y) = 1 - F_X(g^{-1}(x))$$

Next theorem tell us an interesting thing.

定理 5.2

Let X be a continuous random variable whose cumulative density function F_X possesses a unique inverse ${F_X}^{-1}$.
Let $Z = F_X(X)$, then Z has a uniform distribution on $[0, 1]$.
Let U be a uniform random variable on $[0, 1]$ and F is a cumulative distribution function which possesses a unique inverse F^{-1}. Let $X = F^{-1}(U)$, then the cumulative distribution function of X is F.

♡

The two part of the theorem are useful for pseudo-random number generation in computer simulation.

We move on to the transformation of the probability density function of Y and denote it by f_Y. Again we have following properties.

1. Suppose that g is a differentiable strictly increasing function. For $y \in R_Y$
$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy} = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$$
2. Suppose that g is a differentiable strictly decreasing function. For $y \in R_Y$
$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}(1 - F_X(g^{-1}(y))) = -f_X(g^{-1}(y))\frac{dg^{-1}(y)}{dy} = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$$

We can find there the $|\frac{dg^{-1}(y)}{dy}|$ here play a role for adjustment.

5.2 Expectation and Variance of Continuous Random Variables

Let study the expectation and variance of continuous random variables

定义 5.3 (Expectation)

If X has a probability density function f_X, then the expectation of X is defined by

$$E[x] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

provided that the integral converges absolutely.

♣

Then we see some properties of expectation

性质

1. Expectation of Transformation. If $Y = g(X)$, then
$$E[y] = \int_{-\infty}^{\infty} y \cdot f_Y(y)\, dy = \int_{-\infty}^{+\infty} g(x) \cdot f_X(x)\, dx$$
2. Expectation of Linear Function. For $a, b \in R$,
$$E[aX + b] = a \cdot E[X] + b$$

Let's move on to the variance

定义 5.4

If X has a probability density function f_X, then the expectation of X is also called the mean of X or f_X and denoted μ_X, so that $\mu_X = E[X]$. The variance of X is defined as

$$Var[X] = E[(X - \mu_X)^2] = \int_{-\infty}^{+\infty} (x - \mu_X)^2 \cdot f_X(x)\, dx$$

and denote by ${\sigma_X}^2$. The σ_X is called the standard deviation.

♣

Here are some properties of mean and variance.

性质

1. The mean and variance of continuous random variables have the same intuitive interpretation as in the discrete case
2. $Var[X] = E[X^2] - (E[X])^2$
3. Variance of Linear Function. For $a, b \in R$
$$Var[aX + b] = a^2 \cdot Var[X]$$

Now we talk about the relationship between expectation and cumulative distribution function.

定理 5.3

For a nonenegative continuous random variable X

$$E[X] = \int_0^{+\infty} 1 - F_X(x)\, dx = \int_0^{+\infty} P(X > x)\, dx$$

In fact

$$E[X] = \int_0^{+\infty} P(X > x)\, dx - \int_{-\infty}^{0} P(X < x)\, dx$$

♡

5.3 The Uniform Distribution

We can represent the uniform distribution as $X \sim Uniform(\alpha, \beta) \quad -\infty < \alpha < \beta < +\infty$

The probability density function of the uniform random variables:

$$f(x) = \begin{cases} \frac{1}{(\beta-\alpha)} & \alpha \le x \le \beta \\ 0 & otherwise \end{cases}$$

The cumulative distribution function of the uniform random variables:

$$F(x) = \begin{cases} 0 & x < \alpha \\ \frac{1}{\beta-\alpha}(x-\alpha) & \alpha \le x \le \beta \\ 1 & x > \beta \end{cases}$$

定理 5.4

The mean and variance of the uniform distribution are

$$\mu = \frac{\alpha+\beta}{2} \qquad \sigma^2 = \frac{(\beta-\alpha)^2}{12}$$

♡

5.4 Normal Distribution

Normal Distribution is also called Gaussian Distribution and it is almost the most important distribution.

For $\mu \in R$ and $\sigma > 0$, the function

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

is a probability density function since $f(x) \ge 0$ for all $x \in R$. And we put $y = \frac{x-\mu}{\sigma}$ which gives $x = \sigma y + \mu$ and $\frac{dx}{dy} = \sigma$ which gives $dx = \sigma dy$. This kind of linear transformation is called standardization. Hence

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy = \frac{I}{\sqrt{2\pi}} = 1$$

We can calculate I easily with the help of calculus.

The distribution of a random variable X with this probability density function is called the normal(Gaussian) distribution with parameter μ and σ, denoted by $N(\mu, \sigma^2)$.

The normal probability function is a bell-shaped curve. And it has following properties:

性质

1. It is symmetric about the point μ. i.e.

   $$f(\mu + \Delta) = f(\mu - \Delta)$$

   and falls off in the rate determined by σ.
2. The probability density function has a maximum at μ(can be shown by differentiation) and the maximum height is $\frac{1}{\sqrt{2\pi}\sigma}$
3. The probability is almost within the interval $[\mu - 3\sigma, \mu + 3\sigma](0.997)$. And it's called **3-$\sigma$ criterion**

The cumulative distribution function of normal distribution does not have a close form. Let's study the mean value and the variance.

定理 5.5

The mean and variance of a $N(\mu, \sigma^2)$ distribution are μ and σ^2 respectively.

♡

μ is called location parameter while σ is called scale or dispersion parameter.

证明 Use standardization.

Normal distribution is one of the most widely used distribution. It can be used to model the distribution of many natural phenomena because of the central limit theorem.

定理 5.6

Suppose that $X \sim N(\mu, \sigma^2)$. The random variable

$$Y = aX + b$$

where $a \neq 0$, is also normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$.

♡

A normal random variable with parameter $\mu = 0$ and $\sigma^2 = 1$ is called standard normal distribution. The $N(0, 1)$ distribution is very important since properties of any other normal distributions can be found from those of the standard normal.

The cumulative distribution function of the standard normal distribution is usually denoted by Φ.

定理 5.7

Suppose that $X \sim N(\mu, \sigma^2)$. The cumulative distribution function of X is

$$F_X(x) = \Phi(\frac{x-\mu}{\sigma})$$

And we have

$$\Phi(z) + \Phi(-z) = 1$$

♡

We can approximate the binomial distribution using the normal distribution. And this is an example of the central theorem.

定理 5.8

Suppose that $X_n \sim binomial(n, p)$. Define

$$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$$

Then, as $n \to \infty$, the distribution of Z_n converge to the $N(0, 1)$ distribution.

♡

And the size of n to achieve a good approximation depends on the value of p. And it is called Continuity Correction.

5.5 Exponential Distribution

For $\lambda > 0$, the function

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

is a probability density function that satisfies two properties.

The distribution of a random variable X with this probability density function is called the exponential distribution with parameter λ.

We can get the cumulative distribution function quickly

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

Now we look at the mean and variance of an exponential distribution with parameter λ.

定理 5.9

The mean and variance of an exponential distribution with parameter λ are

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

♡

证明 Note that $\Gamma(n) = (n-1)!$ then, $\int_0^{+\infty} xe^{-x}\,dx = \Gamma(2) = 1$ and $\int_0^{+\infty} x^2e^{-x}\,dx = \Gamma(3) = 2$

The exponential distribution is often used to model the length of waiting time until an event occurs or the lifetime of a product.

The parameter λ is call the **rate** and is the average number of events that occur in unit time. (This gives an intuitive interpretation of $E[X] = 1/\lambda$)

There is some relationship between exponential, gamma, and Poisson distribution. The rate parameter λ is the same for the Poisson exponential, and gamma random variables. And the exponential distribution can be thought of as the continuous analogue of the geometric distribution.

定理 5.10

The exponential distribution (like the geometric distribution) is **memoryless**, i.e. for $s, t \geq 0$

$$P(X > s + t | X > s) = P(X > t)$$

where $X \sim exponential(\lambda)$

♡

This means that the distribution of the waiting time to the next event remains the same regardless of how long we have already been waiting. And this only happens when events occur (or not) totally at random, i.e. independent of past history. Notice that it does not mean the two events $\{X > s + t\}$ and $\{X > s\}$ are independent.

5.6 Other Continuous Distributions

Before talk about the Gamma Distribution, we first look at the Gamma function

定义 5.5 (Gamma Function)

For $\alpha > 0$, the gamma function is defined as

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

♣

Here are some special properties of the gamma function:

性质

1. $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
2. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
3. $\Gamma(\alpha) = (\alpha - 1)!$ if α is an integer
4. $\Gamma(\frac{\alpha}{2}) = \frac{\sqrt{\pi}(\alpha-1)!}{2^{\alpha-1}(\frac{\alpha-1}{2})!}$ if α is an odd integer
5. Gamma Function is a generalization of the factorial functions

We move on to the Gamma distribution.

For $\alpha, \lambda > 0$, the function

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is a probability density function that satisfies two properties. And the distribution of a random variable X with this probability density function is called gamma distribution with parameter α and λ.

The cumulative distribution function can be expressed in terms of the incomplete gamma function.

$$F(x) = \begin{cases} \int_0^x \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}\, dy = \frac{1}{\Gamma(\alpha)} \cdot \gamma(\alpha, \lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Note that if α is an integer, we have

$$F(x) = 1 - \sum_0^{\alpha-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}$$

Notice that the the summation is exactly the Poisson distribution. Hence

$$P(X \leq x) = F_X(x) = 1 - F_Y(\alpha - 1)$$

The F_X is the gamma distribution and the F_Y is the cumulative distribution function of the Poisson random variable.

定理 5.11

The mean and variance of a gamma distribution with parameter α and λ are

$$\mu = \frac{\alpha}{\lambda} \qquad \sigma^2 = \frac{\alpha}{\lambda^2}$$

♡

Gamma distribution can be thought of as a continuous analogue of the negative binomial distribution.

α is called the shape parameter and λ scale parameter.

A special case of the gamma case of the gamma distribution occurs when $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$ for some positive integer n. This is known as the Chi-square distribution with n degrees of freedom (Chapter 6).

We talk about the Beta distribution with the same way.

定义 5.6 (Beta Function)

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

♣

For $\alpha, \beta > 0$, the function

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

is a probability density function.

The distribution of a random variable X with this probability density function is called the beta distribution with parameters α and β.

The cumulative distribution function of beta distribution can be expressed in terms of the incomplete beta function, i.e. $F(x) = 0$ for $x < 0$, $F(x) = 1$ for $x > 1$, and for $0 \leq x \leq 1$,

$$F(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x y^{\alpha-1}(1-y)^{\beta-1}\, dy = \frac{B(x;\alpha,\beta)}{B(\alpha,\beta)}$$
$$= \sum_{i=\alpha}^{\alpha+\beta-1} \frac{(\alpha+\beta-1)!}{i!(\alpha+\beta-1-i)!} x^i (1-x)^{\alpha+\beta-1-i} = \sum_{i=\alpha}^{\alpha+\beta-1} \binom{\alpha+\beta-1}{i} x^i (1-x)^{(\alpha+\beta-1)-i}$$

定理 5.12

The mean and variance of a beta distribution with parameter α and β are

$$\mu = \frac{\alpha}{\alpha+\beta} \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

♡

Some properties of Beta Distribution:

1. When $\alpha = \beta = 1$, the beta distribution is the same as the uniform$(0, 1)$
2. Whenever $\alpha = \beta$, the beta distribution is symmetric about $x = 0.5$, i.e.
$$f(0.5 - \Delta) = f(0.5 + \Delta)$$
3. As the common value of α and β increases, the distribution becomes more peaked at $x = 0.5$ and there is less probability outside of the central portion.
4. When $\beta > \alpha$, values close to 0 become more likely than those close to 1; when $\beta < \alpha$, values close to 1 are more likely than those close to 0.(Skewed)

Weibull Distribution

For $\alpha, \beta > 0$ and $\nu \in R$, the function

$$f(x) = \begin{cases} \frac{\beta}{\alpha}(\frac{x-\nu}{\alpha})^{\beta-1} e^{-(\frac{x-\nu}{\alpha})^{\beta}} & x \geq \nu \\ 0 & x < \nu \end{cases}$$

is a probability density function. We can prove $\int_{-\infty}^{+\infty} f(x) = 1$ with Δ transformation.

The distribution of a random variable X with this probability density function is called the Weibull distribution with parameter α, β and ν.

The cumulative distribution function of Weibull distribution is

$$F(x) = \begin{cases} 1 - e^{-(\frac{x-\nu}{\alpha}^{\beta})} & x \geq \nu \\ 0 & x < \nu \end{cases}$$

定理 5.13

The mean and variance of a Weibull distribution with parameter α, β and ν are

$$\mu = \alpha\Gamma(1 + \frac{1}{\beta}) + \nu$$
$$\sigma^2 = \alpha^2\{\Gamma(1 + \frac{2}{\beta}) - [\Gamma(1 + \frac{1}{\beta})]^2\}$$

♡

Some properties of Weibull distribution:

1. Weibull distribution is widely used to model lifetime
2. α is called scale parameter and β is called shape parameter, ν is called location parameter.

定理 5.14

If $X \sim$ exponential(λ), then

$$Y = \alpha(\lambda X)^{\frac{1}{\beta}} + \nu$$

is distributed as Weibull with parameter α, β, ν.

♡

Cauchy Distribution

For $\mu \in R$ and $\sigma > 0$, the function

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

is a probability density function. We can prove that by substitute $\frac{(x-\mu)^2}{\sigma^2}$ with y.

The distribution of a random variable X with this probability density function is called the Cauchy distribution with parameter μ, σ, denoted by Cauchy(μ, σ).

The cumulative distribution function of Cauchy distribution is

$$F(x) = \int_{-\infty}^{x} \frac{\sigma}{\pi} \frac{1}{\sigma^2 + (y-\mu)^2}\, dy = \frac{1}{2} + \frac{1}{\pi} \arctan(\frac{x-\mu}{\sigma}) \quad -\infty < x < +\infty$$

The mean and variance of Cauchy distribution do not exist because the integral does not converge absolutely.

Some properties of Cauchy distribution:

1. Cauchy distribution is a heavy tail distribution (integral goes to infinity)
2. μ is called location parameter and σ is called scale parameter

定理 5.15

If $X \sim$ Cauchy(μ, σ), then $aX + b \sim$ Cauchy $(a\mu + b, |a|\sigma)$.

第 6 章 Jointly Distributed Random Variables

6.1 Joint Distribution Functions

In previous chapter we focus on univariate random variable. However, often a single experiment will have more than one random variables which are of interest.

定义 6.1

Given a sample space Ω and a probability measure P defined on the subset of Ω, random variables

$$X_1, X_2, \cdots, X_n : \Omega \to R$$

are said to be jointly distributed.

♣

笔记 Note that $X_1, \cdots, X_n$ must be maps defined on "same" sample space Ω. Ω is critical in discussing issues.

We can regard n jointly distributed random variables as a random vector

$$X = (X_1, X_2, \cdots, X_n) : \Omega \to R^n$$

定义 6.2

The probability measure of X is called the joint distribution of $X_1, \cdots, X_n$. The probability measure of X_i is called the marginal distribution of X_i

♣

We can deduce that :

1. When joint distribution is given, its corresponding marginal distributions are known
2. Joint distribution offers more information

We can characterize the joint distribution of X in terms of its

1. Joint Cumulative Distribution Function
2. Joint Probability Mass(Density) Function
3. Joint Moment Generating Function (Chapter 7)

定义 6.3 (Joint Cumulative Distribution Function)

The joint cumulative distribution function of $X = (X_1, \cdots, X_n)$ is defined as

$$F_X(x_1, \cdots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \cdots, X_n \le x_n)$$

♣

定理 6.1

Suppose that F_X is a joint cumulative distribution function. Then

1. $0 \le F_X(x_1, \cdots, x_n) \le 1$, for $-\infty < x_i < +\infty$
2. $\lim_{x_1, \cdots, x_n \to +\infty} F_X(x_1, \cdots, x_n) = 1$
3. For any $i \in \{1, \cdots, n\}$

$$\lim_{x_i \to -\infty} F_X(x_1, \cdots, x_n) = 0$$

4. F_X is continuous from the right with respect to each of the coordinates, or any subset of them jointly
5. If $x_i \le x_i', i = 1, \cdots, n$, then

$$F_X(x_1, \cdots, x_n) \le F_X(t_1, \cdots, t_n) \le F_X(x_1', \cdots, x_n')$$

where $t_i \in \{x_i, x_i'\}, i = 1, 2, \cdots, n$. When $n = 2$, we have

$$F_{X_1,X_2}(x_1, x_2) \leq F_{X_1,X_2}(x_1', x_2) \leq F_{X_1,X_2}(x_1', x_2')$$

6. If $x_1 \leq x_1'$ and $x_2 \leq x_2'$, then

$$P(x_1 < X_1 \leq x_1', x_2 < X_2 \leq x_2') = F_{X_1,X_2}(x_1', x_2') - F_{X_1,X_2}(x_1, x_2') - F_{X_1,X_2}(x_1', x_2) + F_{X_1,X_2}(x_1, x_2)$$

In particular, let $x_1' \to \infty$ and $x_2' \to \infty$, we get

$$P(x_1 < X_1 < \infty, x_2 < X_2 < \infty) = 1 - F_{X_1}(x_1) - F_{X_2}(x_2) + F_{X_1,X_2}(x_1, x_2)$$

7. The joint cumulative distribution function of $X_1, \cdots, X_k, k < n$ is

$$\begin{aligned} F_{X_1,\cdots,X_k}(x_1, \cdots, x_k) &= P(X_1 \leq x_1, \cdots, X_k \leq x_k) \\ &= P(X_1 \leq x_1, \cdots, X_k \leq x_k, -\infty < X_{k+1} < +\infty, \cdots, -\infty < X_n < +\infty) \end{aligned}$$

In particular, the marginal cumulative distribution function of X_1 is

$$F_{X_1}(x) = P(X_1 \leq x) = \lim_{x_2,\cdots,x_n \to +\infty} F_X(x, x_2, \cdots, x_n)$$

♡

定理 6.2

A function $F_X(x_1, \cdots, x_n)$ can be a joint cumulative function if F_X satisfies 1-5 in Theorem 6.1.

♡

Let's see the joint probability mass function.

定义 6.4

Suppose that $X_1, \cdots, X_n$ are discrete random variables. The joint probability mass function of $X = (X_1, \cdots, X_n)$ is defined as

$$p_X(x_1, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n)$$

♣

定理 6.3

Suppose that p_X is a joint probability mass function. Then,

1. $p_X(x_1, \cdots, x_n) \geq 0$, for $-\infty < x_i < +\infty,\ i = 1, 2, \cdots, n$
2. There exists a finite or countably infinite set $\chi \subset R^n$ such that $p_X(x_1, \cdots, x_n) = 0$ for $(x_1, \cdots, x_n) \notin \chi$
3. $\sum_{x \in \chi} p_X(x) = 1$, where $x = (x_1, x_2, \cdots, x_n)$
4. For $A \subset R^n, P(X \in A) = \sum_{x \in A \cap \chi} p_X(x)$
5. The joint probability mass function of $X_1, \cdots, X_k$ is

$$p_{X_1,\cdots,X_k}(x_1, \cdots, x_k) = P(X_1 = x_1, \cdots, X_k = x_k) = \sum_{\substack{(x_1,\cdots,x_n) \in \chi \\ -\infty < x_{k+1} < +\infty, \cdots, -\infty < x_n < +\infty}} p_X(x_1, \cdots, x_k, \cdots, x_n)$$

♡

定理 6.4

A function $p_X(x_1, \cdots, x_n)$ can be a joint probability mass function if p_X satisfies 1-3 in Theorem 6.3.

♡

Next theorem shows the relationship between the joint cumulative distribution function and the probability mass function of X.

定理 6.5

If F_X and p_X are the joint cumulative distribution function and joint probability mass function of X, then

$$F_X(x_1, \cdots, x_n) = \sum_{\substack{(t_1,\cdots,t_n)\in\chi \\ t_1\le x_1,\cdots,t_n\le x_n}} p_X(t_1, \cdots, t_n)$$

And

$$p_X(x) = F_X(x) - F_X(x^-)$$

where $x = (x_{10}, \cdots, x_{n0})$

♡

Let's see the joint probability density function.

定义 6.5

A function $f_X(x_1, \cdots, x_n)$ can be a joint probability density function if

1. $f_X(x_1, \cdots, x_n) \ge 0$, for $-\infty < x_i < +\infty, i = 1, 2, \cdots, n$
2. $\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(x_1, \cdots, x_n)\, dx_1 \cdots dx_n = 1$

♣

定义 6.6

Suppose that $X_1, \cdots, X_n$ are continuous random variables. The joint probability density function of $X = (X_1, \cdots, X_n)$ is a function $f_X(x_1, \cdots, x_n)$ satisfying 1. and 2. above, and for any event $A \subset R^n$

$$P(X \in A) = \int \cdots \int_A f_X(x_1, \cdots, x_n)\, dx_1, \cdots dx_n$$

♣

定理 6.6

Suppose that f_X is the joint probability density function of $X = (X_1, \cdots, X_n)$. Then, the joint probability density function of $X_1, \cdots, X_k,\ k < n$, is

$$f_{X_1,\cdots,X_k}(x_1, \cdots, x_k) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(x_1, \cdots, x_k, x_{k+1}, \cdots, x_n)\, dx_{k+1} \cdots dx_n$$

If particular, the marginal probability density function of X_1 is

$$f_{X_1}(x) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_X(x, x_2, \cdots, x_n)\, dx_2 \cdots dx_n$$

♡

定理 6.7

If F_X and f_X are the joint cumulative distribution function and joint probability density function of X, then

$$F_X(x_1, \cdots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_X(t_1, \cdots, t_n)\, dt_1 \cdots dt_n$$

and

$$f_X(x_1, \cdots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_X(x_1, \cdots, x_n).$$

at the continuity points of f_X.

♡

We now discuss a famous distribution in discrete random variable. – Multinomial Distribution.

Recall Partitions:

If $n \ge 1$ and $n_1, \cdots, n_m \ge 0$ are integers for which

$$n_1 + n_2 + \cdots + n_m = n$$

then a set of n elements may be partitioned into m subsets of size $n_1, \cdots, n_m$ in

$$\binom{n}{n_1, \cdots, n_m} = \frac{n!}{n_1! \times \cdots \times n_m!}$$

ways.

Consider a basic experiment which can result in one of m types of outcomes. Denote its sample space as

$$\Omega_0 = \{1, 2, \cdots, m\}$$

Let $p_i = P$ (outcome i appears in basic experiment), then

1. $p_1, \cdots, p_m \geq 0$
2. $p_1 + \cdots + p_m = 1$

Repeat the basic experiment n times. Then, the sample space for the n trials is

$$\Omega = \Omega_0 \times \cdots \times \Omega_0 = {\Omega_0}^n$$

Let $X_i =$ number of trials with outcomes i, $i = 1, 2, \cdots, m$, Then

1. $X_1, \cdots, X_m : \Omega \to R$
2. $X_1 + \cdots + X_m = n$

The joint probability mass function of $X_1, \cdots, X_m$ is

$$p_X(x_1, \cdots, x_m) = P(X_1 = x_1, \cdots, X_m = x_m) = \binom{n}{x_1, \cdots, x_m} {p_1}^{x_1} \times \cdots \times {p_m}^{x_m}$$

for $x_1, \cdots, x_m \geq 0$ and $x_1 + \cdots + x_m = n$.

The distribution of a random vector $X = (X_1, \cdots, X_m)$ with the above joint probability mass function is called the multinomial distribution with parameter n, m and $p_1, \cdots, p_m$, denoted by Multinomial$(n, m, p_1, \cdots, p_m)$.

The multinomial distribution is called after the multinomial theorem. And it is a generalization of the binomial distribution from 2 types of outcomes to m types of outcomes.

Here are some properties of the multinomial distribution:

1. Because $X_i = n - (X_1 + \cdots + X_{i-1} + X_{n+1} + \cdots + X_m)$, and $p_i = 1 - (p_1 + \cdots + p_{i-1} + p_{i+1} + \cdots + p_m)$, we can rewrite the random vector with $n - 1$ elements with a combination of these elements.
2. Marginal Distribution. Suppose that

   $$(X_1, \cdots, X_m) \sim Multinomial(n, m, p_1, \cdots, p_k, p_{k+1}, \cdots, p_m)$$

   For $1 \leq k < m$, the distribution of

   $$(X_1, \cdots X_k, X_{k+1} + \cdots + X_m)$$

   is Multinomial$(n, k + 1, p_1, \cdots, p_k, p_{k+1} + \cdots + p_m)$.
3. Mean and Variance.

   $$E[X_i] = np_i \qquad Var[X_i] = np(1 - p_i)$$

   for $i = 1, \cdots, m$.

6.2 Independent Random Variables

Recall

If joint distribution is given, marginal distributions are known.

The converse statement does not hold in general.

However, when random variables are independent, marginal distributions + independence $\Rightarrow$ joint distribution.

定义 6.7

The n jointly distributed random variables $X_1, \cdots, X_n$ are called (mutually) independent if and only if for any (measurable) sets $A_i \subset R,\ i = 1, \cdots, n$, the events

$$\{X_1 \in A_1\}, \cdots, \{X_n \in A_n\}$$

are (mutually) independent. That is,

$$P(X_{i_1} \in A_{i_1}, X_{i_2} \in A_{i_2}, \cdots, X_{i_k} \in A_{i_k}) = P(X_{i_1} \in A_{i_1}) \times P(X_{i_2} \in A_{i_2}) \times \cdots \times P(X_{i_k} \in A_{i_k})$$

for any $1 \leq i_1 < i_2 \cdots < i_k \leq n$,

If $X_1, \cdots, X_n$ are independent, for $1 \leq k < n$,

$$P(X_{k+1} \in A_{k+1}, \cdots, X_n \in A_n | X_1 \in A_1, \cdots, X_k \in A_k) = P(X_{k+1} \in A_{k+1}, \cdots, X_n \in A_n)$$

provided that $P(X_1 \in A_1, \cdots, X_k \in A_k) > 0$

In other words, the values of $X_1, \cdots, X_k$ do not carry any information about the distribution of $X_{k+1}, \cdots, X_n$.

定理 6.8 (Factorization Theorem)

The random variables $X = (X_1, \cdots, X_n)$ are independent if and only if one of the following conditions holds.

1. $F_X(x_1, \cdots, x_n) = F_{X_1}(x_1) \times \cdots \times F_{X_n}(x_n)$, where F_X is the joint cumulative distribution function of X and F_X is the marginal cumulative distribution function of X_i for $i = 1, \cdots, n$.
2. Suppose that $X_1, \cdots, X_n$ are discrete random variables. $p_X(x_1, \cdots, x_n) = p_{X_1}(x_1) \times \cdots \times p_{X_n}(x_n)$, where p_X is the joint probability mass function of X and p_{X_i} is the marginal probability mass function of X_i for $i = 1, \cdots, n$.
3. Suppose that $X_1, \cdots, X_n$ are continuous random variables. $f_X(x_1, \cdots, x_n) = f_{X_1}(x_1) \times \cdots \times f_{X_n}(x_n)$, where f_X is the joint probability density function of X and f_{X_i} is the marginal probability density function of X_i for $i = 1, \cdots, n$.

So you can use above test to verify whether the random variables are independent or not.

定理 6.9

If $X = (X_1, \cdots, X_n)$ are independent and

$$Y_i = g_i(X_i) \quad i = 1, \cdots, n$$

then $Y_1, \cdots, Y_n$ are independent.

Next theorem

定理 6.10

$X = (X_1, \cdots, X_n)$ are independent if and only if there exist univariate functions $g_i(x),\ i = 1, \cdots, n$ such that

1. when $X_1, \cdots, X_n$ are discrete random variables with joint probability mass function p_X,
$$p_X(x_1, \cdots, x_n) \propto g_1(x_1) \times \cdots \times g_n(x_n),\ -\infty < x_i < +\infty,\ i = 1, \cdots, n$$
2. when $X_1, \cdots, X_n$ are continuous random variables with joint probability density function f_X,
$$f_X(x_1, \cdots, x_n) \propto f_1(x_1) \times \cdots \times f_n(x_n),\ -\infty < x_i < +\infty,\ i = 1, \cdots, n$$

Note that if the region produced by random variables is not a "cross product set"(rectangle). Then these random variables are not independent.

We now move on to the transformation $R^n \to R^k$. The following methods are useful when deal with this topic:

1. Method of Events (probability mass function)
2. Method of Cumulative Distribution Function
3. Method of Probability Density Function
4. Method of Moment Generating Function.

We discuss the first three one by one.

定理 6.11

0 The distribution of Y is determined by the distribution of X as follows: for any event $B \subset R^k$

$$P_Y(Y \in B) = P_X(X \in A)$$

where $A = g^{-1}(B) \subset R^n$

♡

Now let's see the application of the transformation.

定理 6.12

If X, Y are independent, and $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$, then $Z = X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$

♡

推论 6.1

If $X_1, \cdots, X_n$ are independent, and $X_i \sim \text{Poisson}(\lambda_i),\ i = 1, \cdots, n$, then

$$X_1 + \cdots + X_n \sim Poisson(\lambda_1 + \cdots + \lambda_n)$$

♡

笔记 Note that here we use convolution.

Method of cumulative distribution function:

1. In the $(X_1, \cdots, X_n)$ space, find the region that corresponds to
$$\{Y_1 \le y_1, \cdots, Y_k \le y_k\}$$
2. Find $F_Y(y_1, \cdots, y_k) = P(Y_1 \le y_1, \cdots, Y_k \le y_k)$ by summing the joint probability mass function or integrating the joint probability density function of $X_1, \cdots, X_n$ over the region identified in 1.
3. (for continuous case) Find the joint probability density function of Y by differentiating $F_Y(y_1, \cdots, y_k)$, i.e.
$$f_Y(y_1, \cdots, y_k) = \frac{\partial^k}{\partial y_1 \cdots \partial y_k} F_Y(y_1, \cdots, y_k)$$

定理 6.13

If X, Y are independent, and $X \sim \text{Gamma}(\alpha_1, \lambda)$, $Y \sim \text{Gamma}(\alpha_2, \lambda)$, then

$$Z = X + Y \sim Gamma(\alpha_1 + \alpha_2, \lambda)$$

♡

推论 6.2

If $X_1, \cdots, X_n$ are independent, and $X_i \sim \text{Gamma}(\alpha_i, \lambda),\ i = 1, \cdots, n$, then

$$X_1 + \cdots + X_n \sim Gamma(\alpha_1 + \cdots + \alpha_n, \lambda)$$

♡

推论 6.3

If $X_1, \cdots, X_n$ are independent, and $X_i \sim \text{Exponential }(\lambda),\ i = 1, \cdots, n$, then

$$X_1 + \cdots + X_n \sim Gamma(n, \lambda)$$

♡

定理 6.14

If $X_1, \cdots, X_n$ are independent, and $X_i \sim \text{Normal}(\mu_i, {\sigma_i}^2),\ i = 1, \cdots, n$, then

$$X_1 + \cdots + X_n \sim Normal(\mu_1 + \cdots + \mu_n, {\sigma_1}^2 + \cdots + {\sigma_n}^2)$$

♡

Method of probability density function:

定理 6.15

Let $X = (X_1, \cdots, X_n)$ be continuous random variables with the joint probability density function $f_X(x_1, \cdots, x_n)$. Let

$$Y = (Y_1, \cdots, Y_n) = g(X)$$

where g is a bijection. so that its inverse exists and is denoted by

$$x = g^{-1}(y) = w(y) = (w_1(y), \cdots, w_n(y))$$

Assume w have continuous partial derivatives. Let the determinant of the Jacobian matrix

$$J(y_1, \cdots, y_n) = |[\frac{\partial w_i}{\partial y_i}]|$$

Then $f_Y(y) = f_X(g^{-1}(y)) \cdot Abs[J]$ for y s.t. $y = g(x)$ for some x, and $f_Y(y) = 0$, otherwise.

♡

定理 6.16

If X_1, X_2 are independent, and

$$X_1 \sim Gamma(\alpha_1, \lambda) \qquad X_2 \sim Gamma(\alpha_2, \lambda)$$

then $Y_1 = \frac{X_1}{X_1 + X_2} \sim \text{Beta}(\alpha_1, \alpha_2)$

♡

6.3 Order Statistics(顺序统计量)

quantile(分位数) in mathematical statistics.

定义 6.8

Let $X_1, \cdots, X_n$ be random variables. We sort the $X_i's$ and denote by

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

the order statistics. Using the notation, $X_i = i$ th-smallest value in $X_1, \cdots, X_n,\ i = 1, 2, \cdots, n$

X_1, X_n is the minimum and the maximum respectively and $R \equiv X_n - X_{(1)}$ is called range. $S_j \equiv X_{(j)} - X_{(j-1)},\ j = 2, \cdots, n$ are called jth spacing.

♣

笔记 Note that Order Statistics is an transformation. But it's inverse does not exist.

定义 6.9

$X_1, \cdots, X_n$ are called independent, identically distributed(i.i.d) with cumulative distribution function F/probability mass function p/probability density function f if the random variables $X_1, \cdots, X_n$ are independent and have a common marginal distribution with F/probability mass function p/probability density function f.

♣

笔记 In the discussion about order statistics, we only consider the case that $X_1, \cdots, X_n$ are i.i.d. Although $X_1, \cdots, X_n$ are independent, their order statistics $X_{(1)}, \cdots, X_{(n)}$ are not independent in general.

定理 6.17

Suppose that $X_1, \cdots, X_n$ are i.i.d. with cumulative distributive function F.

1. The cumulative distribution function of $X_{(1)}$ is $1-[1-F(x)]^n$, and the cumulative distribution function of $X_{(n)}$ is $[F(x)]^n$
2. If X are continuous and F has a probability density function f, then the probability density function of $X_{(1)}$ is $nf(x)[1-F(x)]^{n-1}$, and the probability density function of $X_{(n)}$ is $nf(x)[F(x)]^{n-1}$

♡

定理 6.18

Suppose that $X_1, \cdots, X_n$ are i.i.d. with probability mass function p or probability density function f. Then, the joint probability mass function or probability density function of $X_{(1)}, \cdots, X_{(n)}$ is

$$p_{X_{1},\cdots,X_{(n)}}(x_1, \cdots, x_n) = n! \times p(x_1) \times \cdots p(x_n)$$
$$f_{X_{1},\cdots,X_{(n)}}(x_1, \cdots, x_n) = n! \times f(x_1) \times \cdots f(x_n)$$

for $x_1 \leq x_2 \leq \cdots \leq x_n$, and 0 otherwise.

♡

定理 6.19

If $X_1, \cdots, X_n$ are i.i.d. with cumulative distribution function F and probability density function f, then

1. The probability density function of the k^{th} order statistic $X_{(k)}$ is
$$f_{X_{(k)}}(x) = \binom{n}{1,k-1,n-k} f(x)F(x)^{k-1}[1-F(x)]^{n-k}$$
2. The cumulative distribution function of $X_{(k)}$ is
$$F_{X_{(k)}}(x) = \sum_{m=k}^{n} \binom{n}{m} [F(x)]^m [1-F(x)]^{n-m}$$
note that in the discrete random variables' case, the second equation is still correct.

♡

定理 6.20

If $X_1, \cdots, X_n$ are i.i.d. with cumulative distribution function F and probability density function f, then

1. The joint probability density function of $X_{(1)}, X_{(n)}$ is
$$f_{X_{(1)},X_{(n)}}(s,t) = n(n-1)f(s)f(t)[F(t)-F(s)]^{n-2}$$
for $s \leq t$, and 0 otherwise.
2. The probability density function of the range $R = X_{(n)} - X_{(1)}$ is
$$f_R(r) = \int_{-\infty}^{+\infty} f_{X_{(1)},X_{(n)}}(u, u+r)\, du$$
for $r \geq 0$, and 0 otherwise.

♡

定理 6.21

If $X_1, \cdots, X_n$ are i.i.d. with cumulative distribution function F and probability density function f, then

1. The join probability density function of $X_{(i)}, X_{(j)}$, where $1 \leq i < j \leq n$ is
$$f_{X_{(i)},X_{(j)}}(s,t) = \binom{n}{1,1,(i-1),(j-i-1),(n-j)} f(s)f(t)[F(x)]^{i-1}[F(t)-F(s)]^{j-i-1}[1-F(t)]^{n-j}$$
for $s \leq t$ and 0 otherwise.
2. The probability density function of the j^{th} spacing $S_j = X_{(j)} - X_{(j-1)}$ is
$$f_{S_j}(s) = \int_{-\infty}^{+\infty} f_{X_{(j-1)},X_{(j)}}(u, u+s)\, du$$
for $s \geq 0$, and 0 otherwise.

♡

6.4 Conditional Distribution

定义 6.10

Let $X(\in R^n)$ and $Y(\in R^m)$ be discrete random vectors and (X, Y) have a joint probability mass function $p_{X,Y}(x, y)$, then the conditional joint probability mass function of Y given $X = x$ is defined as

$$p_{Y|X}(y|x) \equiv P(\{Y = y\}|\{X = x\}) = \frac{P(\{X = x, Y = y\})}{P(\{X = x\})} = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

if $p_X(x) > 0$. The probability is defined to be zero if $p_X(x) = 0$.

♣

笔记

1. For each fixed x, $p_{Y|X}(y|x)$ is a joint probability mass function for y, since
$$\sum_y p_{Y|X}(y|x) = 1$$
2. For an event B of Y, the probability that $Y \in B$ given $X = x$ is
$$P(Y \in B|X = x) = \sum_{u \in B} p_{Y|X}(u|x)$$
3. The conditional joint cumulative distribution function of Y given $X = x$ can be similarly defined from the conditional joint probability mass function $p_{Y|X}(y|x)$, i.e.
$$F_{Y|X}(y|x) = P(Y \le y|X = x) = \sum_{u \le y} p_{Y|X}(u|x)$$

定理 6.22

Let $X_1, \cdots, X_m$ be independent and $X \sim$ Poisson $(\lambda_i),\ i = 1, \cdots, m$ Let $Y = X_1 + \cdots + X_m$, then

$$(X_1, \cdots, X_m|Y = n) \sim Multinomial(n, m, p_1, \cdots, p_m)$$

where $p_i = \lambda_i/(\lambda_1 + \cdots + \lambda_m)$ for $i = 1, \cdots, m$

♡

定义 6.11

Let $X(\in R^n)$ and $Y(\in R^m)$ be continuous random vectors and (X, Y) have a joint probability density function $f_{X,Y}(x, y)$, then the conditional joint probability density function of Y given $X = x$ is defined as

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

if $f_X(x) > 0$, and 0 otherwise.

♣

笔记

1. $P(X = x) = 0$ for continuous random vector X.
2. The justification of $f_{Y|X}(y|x)$ comes from
$$P(Y \le y|x - (\Delta x/2) < X < x + (x + \Delta x/2)) \approx \int_{-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)}\, dv$$
3. For each fixed x, $f_{Y|X}(y|x)$ is a joint probability density function for y.
4. For an event B of Y, we can write
$$P(Y \in B|X = x) = \int_B f_{Y|X}(y|x)\, dy$$
5. The conditional joint cumulative function of Y given $X = x$ can be similarly defined from the conditional joint probability mass function $f_{Y|X}(y|x)$. i.e.
$$F_{Y|X}(y|x) = P(Y \le y|X = x) = \int_{-\infty}^{y} f_{Y|X}(t|x)\, dt$$

If we encounter a mixed joint distribution, then the definition of conditional distribution can be similarly generalized to the case in which some random variables are discrete and the others continuous.

定理 6.23 (Multiplication Law)

Let X, Y be random vectors and (X, Y) have a joint probability density function $f_{X,Y}(x, y)$ or probability mass function $p_{X,Y}(x, y)$, then

$$p_{X,Y}(x, y) = p_{Y|X}(y|x) \times p_X(x)$$

or

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) \times f_X(x)$$

♡

定理 6.24 (Law of Total Probability)

Let X, Y be random vectors and (X, Y) have a joint probability density function $f_{X,Y}(x, y)$ or probability mass function $p_{X,Y}(x, y)$, then

$$p_Y(y) = \sum_{x=-\infty}^{+\infty} p_{Y|X}(y|x)p_X(x)$$

or

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{Y|X}(y|x)f_X(x)\,dx$$

♡

定理 6.25 (Bayes Theorem)

Let X, Y be random vectors and (X, Y) have a joint probability density function $f_{X,Y}(x, y)$ or probability mass function $p_{X,Y}(x, y)$, then

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x=-\infty}^{+\infty} p_{Y|X}(y|x)p_X(x)}$$

or

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{+\infty} f_{Y|X}(y|x)f_X(x)\,dx}$$

♡

定理 6.26 (Conditional Distribution and Independent)

Let X, Y be random vectors and (X, Y) have a joint probability density function $f_{X,Y}(x, y)$ or probability mass function $p_{X,Y}(x, y)$. Then, X, Y are independent if and only if

$$p_{Y|X}(y|x) = p_Y(y)$$

or

$$f_{Y|X}(y|x) = f_Y(y)$$

♡

第 7 章 Properties of Expectation

7.1 Introduction

1. The expectation of the transformation of several random variables.
2. Riemann-Stieltjes Integral.

7.2 Expectation of Sums of Random Variables

1. We can change the process of taking expectation and the process of summing up. Such as the properties of absolute convergence.
2. We can change the process of taking expectation and the process of product when the random variables are independent. The division usually don't satisfy this property

7.3 Covariance, Variance of Sums, and Correlations

定义 7.1

Suppose that X, Y are two random variables with finite means μ_X, μ_Y and variances ${\sigma_X}^2, {\sigma_Y}^2$, respectively.

1. Let $g(x, y) = (x - \mu_X)(y - \mu_Y)$, then
$$Cov[X, Y] \equiv E[g(X, Y)] = E[(X - \mu_X)(Y - \mu_Y)]$$
is called the covariance between X and Y, denote by σ_{XY}
2. The correlation (coefficient) between X and Y is defined as
$$Cor[X, Y] = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$
and denoted by ρ_{XY}
3. X, Y are called uncorrelated if $\rho_{XY} = 0$ or $\sigma_{XY} = 0$

♣

笔记

1. Note that you need joint distribution to calculate the covariance or the correlation while you just need marginal distribution to calculate the mean and variance.
2. Covariance is a measure of the joint variability of X and Y, or their degree of association but not necessary a causal relationship.
3. Covariance can be influenced by the unit while Correlation Coefficient is unit free.
4. Correlation coefficient measures the strength of the linear relationship between X and Y

定理 7.1

$$Cov[X, Y] = E[XY] - \mu_X \mu_Y$$

♡

推论 7.1

If X, Y are independent, then $Cov[XY] = 0$, i.e. X, Y are uncorrelated.

♡

笔记 The converse statement is not necessarily true. If Y is a function of X, the correlation can be 0 but it does not mean there is no relationship between X and Y. It just tells that X and Y don't have a linear relationship.

推论 7.2

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

定理 7.2

$$Cov[a_0 + a_1X_1 + \cdots + a_nX_n, b_0 + b_1Y_1 + \cdots + b_nY_m] = \sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_jCov[X_i, Y_j]$$

定理 7.3 (Variance of Sum)

$$Var[a_0 + a_1X_1 + \cdots + a_nX_n] = \sum_{i=1}^{n} {a_i}^2 Var[X_i] + 2\sum_{1\le i<j\le n} a_ia_jCov[X_i]$$

推论 7.3

If $X_1, \cdots, X_n$ are uncorrelated, then

$$Var[a_0 + a_1X_1 + \cdots + a_nX_n] = \sum_{i=1}^{n} {a_i}^2 Var[X_i]$$

推论 7.4

If $X_1, \cdots, X_n$ are uncorrelated and

$$Var[X_1] = \cdots = Var[X_n] \equiv \sigma^2 < +\infty$$

then

$$Var[\frac{\sum_{i=1}^{n} X_i}{n}] = \frac{\sigma^2}{n}$$

which is the embodiment of Law of Large Number

推论 7.5

Suppose that $X_1, \cdots, X_n$ are uncorrelated and have same mean μ and variance σ^2. Let

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}{n-1}$$

then $E[S^2] = \sigma^2$. Here $\overline{X_n} = \sum_{i=1}^{n} X_i$

定理 7.4 (ρ of linear transformation)

$$Cor[a_0 + a_1X, b_0 + b_1Y] = sign(a_1b_1) \cdot Cor(X, Y)$$

and

$$|Cor[a_0 + a_1X, b_0 + b_1Y]| = |Cor[X, Y]|$$

i.e. $|\rho_{XY}|$ is invariant under location and scale changes.

定理 7.5 (some properties of ρ)

1. $-1 \leq \rho_{Xy} \leq 1 (\iff |Cov[X,Y]| \leq \sigma_X \sigma_Y)$
2. $\rho_{XY} = \pm 1$ if and only if there exist $a, b \in R$ such that $P(Y = aX + b) = 1$. i.e.$Y = aX + b$ almost surely
3. Furthermore, $\rho_{XY} = 1$, if $a > 0$ and $\rho_{XY} = -1$, if $a < 0$.

♡

7.4 Conditional Expectation

定义 7.2 (Conditional Expectation)

♣

定理 7.6 (Law of Total Expectation)

For two random vectors $X \in R^m$ and $Y \in R^n$,

$$E_X\{E_{Y|X}[h(Y)|X]\} = E_Y[h(y)]$$

In particular, let $h(Y) = Y_i$, we have

$$E_X[E_{Y|X}[Y_i|X]] = E_Y[Y_i]$$

♡

定理 7.7 (Variance Decomposition)

For two random vectors X and Y,

$$Var_Y[Y_i] = Var_X[E_{Y|X}[Y_i|X]] + E_X[Var_{Y|X}[Y_i|X]]$$

♡

7.5 Conditional Expectation and Prediction

7.6 Moment Generating Functions

定义 7.3 (Moment and Central Moment)

If a random variable X has a cumulative distribution function F_X, then

$$\mu_k \equiv E[X^k] = \int_{-\infty}^{+\infty} x^k \, dF_X(x), \quad k = 1, 2, \cdots,$$

are called the k^{th} **moments** of X provided that the integral converges absolutely, and

$$\mu_k{}' \equiv E[(X - \mu_X)^k] = \int_{-\infty}^{+\infty} (x - \mu_X)^k \, dF_X(x), \quad k = 2, 3, \cdots$$

are called k^{th} **moment about the mean** μ_X or **central moment** of X provided that the integral converges absolutely.

♣

笔记

1. Central moment is a linear transformation of moments.

$$\mu_k{}' = E[(x - \mu_X)^k] = E[\sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{n-i} X^i] = \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{n-i} E[X^i] = \sum_{i=0}^{k} \binom{k}{i} (-\mu_X)^{n-i} \mu_i$$

   And we define that $\mu_0 = 1$

2. Moment is again a linear transformation of the central moment.

$$\mu_k = E[X^k] = E[((X - \mu_X) + \mu_X)^k] = \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{n-i} E[(X - \mu_X)^i] = \sum_{i=0}^{k} \binom{k}{i} (\mu_X)^{n-i} \mu_i{}'$$

In particular,

$$E[X] = \mu_X = \mu_1$$

$$Var[X] = {\sigma_X}^2 = {\mu_2}' = \mu_2 - {\mu_1}^2 = E[X^2] - (E[X])^2$$

The (central) moments give a lot useful information about the distribution in addition to mean and variance, e.g.

1. Skewness(偏度)(a measure of the asymmetric): $\frac{{\mu_3}'}{\sigma^3} = E[(\frac{X-\mu}{\sigma})^3]$
2. Kurtosis(峰度)(a measure of the "heavy tails"): $\frac{{\mu_4}'}{\sigma^4} = E[(\frac{X-\mu}{\sigma})^4]$

定义 7.4 (Moment Generating Function)

If X is a random variable with the cumulative distribution function F_X, then

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{+\infty} e^{tx}\, dF_X(x)$$

is called the moment generating function (mgf) of X provided that the integral converges absolutely in some non-degenerate interval of t.

♣

笔记

1. The mgf is a function of the variable t. It's the Laplace transformation of the probability density function in continuous case.
2. The mgf may only exist for some particular values of t.
3. $M_X(t)$ always exists at $t = 0$ and $M_X(0) = 1$

You can get the mgf of some common distribution in the lecture notes.

定理 7.8 (Uniqueness Theorem)

Suppose that the mgfs $M_X(t)$ and $M_Y(t)$ of random variables X and Y exist for all $|t| < h$ for some $h > 0$. If

$$M_X(t) = M_Y(t)$$

for $|t| < h$, then

$$F_X(z) = F_Y(z)$$

for all $z \in R$, where F_X and F_Y are the cumulative distribution functions of X and Y, respectively.

♡

Application of the uniqueness theorem

1. When a moment generating function exists for all $|t| < h$ for some $h > 0$, there is a unique distribution corresponding to that moment generating function.
2. This allows us to use mgfs to find distributions of transformed random variables in some cases.
3. This technique is most commonly used for linear combinations of independent random variables $X_1, \cdots, X_n$.

The next theorem tells the relationship between MGF and Moment.

定理 7.9

If $M_X(t)$ exists for $|t| < h$ for some $h > 0$, then

$$M_X(0) = 1$$

and

$$M_X{}^{(k)}(0) = \mu_k \quad k = 1, 2, \cdots$$

μ_k is related to the coefficients in the Taylor expansion of $M_X(t)$ ♡

If you know all moments that means you can know the distribution.

定理 7.10 (MGF for linear transformation)

For constant a, b,

$$M_{a+bX}(t) = e^{at} M_X(bt)$$

♡

定理 7.11 (MGF of Sum of independent random variables)

If $X_1, \cdots, X_n$ are independent each with mgfs $M_{X_1}(t), \cdots, M_{X_n}(t)$, respectively, then the mgfs of $S = X_1 + \cdots + X_n$ is

$$M_S(t) = M_{X_1}(t) \times \cdots \times M_{X_n}(t)$$

♡

定义 7.5 (Joint Moment Generating Function)

For random variables $X_1, \cdots, X_n$, their joint moment generating function is defined as

$$M_{X_1,\cdots,X_n}(t_1, \cdots, t_n) = E_{X_1,\cdots,X_n}(e^{t_1 X_1 + \cdots + t_n X_n})$$

provided that the expectation exists.

♣

Let's see some properties of joint mgf

1. $M_{X_1}(t) = M_{X_1,\cdots,X_n}(t, 0, \cdots, 0)$.
2. yielding uniqueness theorem.
3. $X_1, \cdots, X_n$ are independent if and only if

$$M_{X_1,\cdots,X_n}(t_1, \cdots, t_n) = M_{X_1}(t) \times \cdots \times M_{X_n}(t_n)$$

4.

$$\frac{\partial^{k_1+\cdots+k_n}}{\partial {t_1}^{k_1} \times \cdots \times \partial {t_n}^{k_n}} M_{X_1,\cdots,X_n}(0, 0, \cdots, 0) = E_{X_1,\cdots,X_n}({X_1}^{k_1} \times \cdots \times {X_n}^{k_n})$$

-END. 2023/3/11 22:01 流感