Probability 概率论与数理统计

Probability Notes

---

1. **Sample space**

   ***Definition*: List of all possible outcomes of a random variable. Denoted by Ω.**

   **** Notice:*** The order of the event matters. Lets say we flip a coin and get the situation HT and TH . They are different since one is we get head first and another one is we get tail first

2. **Events**

   ***Definition*: The subset of sample space.**

3. **Event field & measurable space** 可测空间与事件域

   For these domains, we want it first to **include the empty set and entire set**, and also **be closed under the operation (operator) intersection, union, complement and difference**. What's more, we found that

   - Intersection can be represented by complement and union (De Morgan's Law) and ;
   - Difference can be represented by complement and intersection which is

   $$A - B = A \cap \overline{B}$$

   These means, in the algorithm of set (or here particular to probability), **Union** and **Complement** are the basic operator between sets. Therefore, the definition becomes

   Theorem 1. Let Ω be a sample space, $\mathscr{F}$ denotes the subset (or a collection) of events, then if

   - $\varnothing$ and $\Omega \in \mathscr{F}$
   - All algorithm are closed under **compliment** and **union**, which is

   $$A \in \mathscr{F} \implies A^c \in \mathscr{F}, \quad A_n \in \mathscr{F} \implies \bigcup_{n=1}^{\infty} A \in \mathscr{F}$$

   we say $\mathscr{F}$ is a valid domain of probability function and $(\Omega, \mathscr{F})$ is called **measurable field (or measurable space)**.

   For example, some domain can be $\mathscr{F} = \{\varnothing, A, \overline{A}, \Omega\}$.

   ** So, one thing need to be noticed is that **the concept 'probability' is actually based on $\mathscr{F}$ but not the sample space!** 可以理解成 **sample space** 中的元素组合之后变成了 **event field** 是一个集合的集合，概率是基于 **event space** 的.

   (a) *Some simple proposition*

4. **Permutation & Combination**

Before introducing the permutation and combination, there are two **basic principle of Counting**:

- If a experiment need k steps to complete, and there are $a, b, c, d, ...$ ways to finish each step respectively, then there are $a \cdot b \cdot c \cdot ....$ ways to achieve the experiments;
- If there are k different ways to complete one thing and there are also m_1, m_2, m_3 ways in each ways, then there are totally $m_1 + m_2 + m_3....$ ways to achieve.

Then here become the permutation & combination:

(a) *Permutation* How many different order of arrangement are there for k objects taken from a n population. The formula is

$$_nP_k = n(n-1)(n-2)(n-3)...(n-k+1) = \frac{n!}{(n-r)!}$$

The equation means, for the first position, we have n choice, second $(n-1)$ choices and so on.

(b) *Combination* Combination does not concerned about order of arrangement but just what is contained inside a group. The formula is

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

The formula can be divided into two part

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

The first part is the permutation of k objects from n population and then divided by the full permutation of k object so that there is no repeat. Because the ***full permutation*** 全排列 is the number of different arrangement a given series of stuff can be.

(c) *Combination with replacement* Taking an object k times form population with replacement to the population after taking. The formula is simply

$$_{(n+k-1)}C_k = \binom{n+k-1}{k} \; or \; \binom{n+k-1}{n-1}$$

Explanation: Let n space among $n+1$ '$\vert$' illustrate the size n population and we are going to take k times. Each time once we take that object, we add a '$\circ$' into the enumeration. Finally there will be k circle and still $n+1$ '$\vert$' for which the first and last '$\vert$' are fixed. Thus it become a combination of "position"— no matter we arrange the circles or the verticals the result will be the same because once we fixed one of it, the other one will be arranged automatically.

(d) *Multinomial theorem & Coefficient* Multinomial simply means how many different division are possible for dividing n distinct terms into r distinct groups of respective sizes $n_1, n_2, n_3, ..., n_r$. The notation is

$$\binom{n}{n_1, n_2, ..., n_r} = \frac{n!}{n_1!n_2!n_3!...n_r!}$$

where $n_1 + n_2 + n_3 + ... + n_r = n$. ****Notice:** There is no order inside each group but there is order between groups! Questions here

Example

How many different arrangement can be formed from the letters PEPPER?

$$\frac{6!}{3!2!}$$

First treat it to be 3 P and 3 other different letters. 6! is simply the full permutation. Then think we fix the other 3 digit and for each fixed permutation, we have 3*2*1 repeated arrangement because of 3 same p. Then divided by 2*1 there comes the result.

(e) *Inclusion and exclusion formula*

i. Start from case of two

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

reason for deducting the intersection is because A and B maybe overlap each other and the overlapped part is double counted;

ii. Case of three

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cup C) + P(A \cap B \cap C)$$

the first part, the summition of A, B and C is the "rough" total. Then reducing the intersection of each two of A, B and C. But the problem is that we also reduce the part that belongs to $A \cap B \cap C$, therefore we add it back.

then by analogy we can guess the general form of n size subset of the sample space which is

5. **Binomial theorem**

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

Proof

6. **Axiom definition of probability space**

Let Ω be the sample space, $\mathscr{F}$ be the event space/field (need to learn more about measurement). If for any event A, a real-valued function $P(A)$ satisfies:

- **Axiom 1** $0 \leq P(A) \leq 1$
- **Axiom 2** $P(\Omega) = 1$

- **Axiom 3** If $A_1, A_2, A_3, ...A_n, ...$ are mutually exclusive, then

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

Then we call the *triple* $(\Omega, \mathscr{F}, P)$ the ***Probability space***. 也由此可见, 概率基于可测空间

7. ***Conditional Probability**

The probability of an event under a given condition. Expression is

$$P(A \mid B)$$

means, the probability of event A given event B.

Then we need to prove it is a probability space, under the fixed condition given. The first two proof are trivial, so we just focus on the third axiom. Prove,

$$P(\bigcup_{i=1}^{\infty} A_i \mid B) = \sum_{i=1}^{\infty} P(A_i \mid B).$$

Proof By the definition of conditional probability, we have

$$P(\bigcup_{i=1}^{\infty} A_i \mid B) = \frac{P(\{\bigcup_{i=1}^{\infty} A_i\} \cap B)}{P(B)}$$

By distributive law, we can get

$$= \frac{P(\bigcup_{i=1}^{\infty} A_i \cap B)}{P(B)}$$

here, since A_i are all mutually exclusive, then we can know $A_i \cap B$ are all mutually exclusive. Then we can write

$$= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)}$$

$$= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)}$$

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

(a) **Properties of** $P(A \mid B)$

- $P(E^c \mid F) = 1 - P(E \mid F)$

  Prove:

$$P(E^c \mid F) = \frac{P(E^c \cap F)}{P(F)} = \frac{P(F) - P(E \cap F)}{P(F)} = 1 - P(E \mid F)$$

- $P(E \mid F) \neq P(F \mid E)$

8. **Bayes's Theorem**

(a) **Total probability**

Definition Let $B_1, B_2, ..., B_n$ are n partition, which mean they are mutually exclusive and union is the whole set. If $P(B_i) > 0, i = 1, 2, ..., n,$, then for any event A we have

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

Proof Since

$$A = A \cap \Omega = A \cap (\bigcup_{i=1}^{n} B_i) = \bigcup_{i=1}^{n} (A \cap B_i)$$

and all $(A \cap B_i)$are mutually exclusive $\implies$ then for probability they can be added together which is

$$P(A) = P(\bigcup_{i=1}^{n} (A \cap B_i)) = \sum_{i=1}^{n} P(A \cap B_i)$$

***Notice:*

- The simplest form of total probability is

$$P(A) = P(B)P(A \mid B) + P(\overline{B})P(A \mid \overline{B})$$

- A $\in \Omega = \bigcup_{i=1}^{n} B_i$.

(b) **Bayes's Theorem**

Definition Let $B_1, B_2, ..., B_n$ are n partition, then

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

Proof

Bayes's formula is based on total probability,

$$P(B_i \mid A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i)P(A \mid B_i)}{P(A)}$$

then substitute the denominator by total probability

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

Done.

(c) **Sequential Bayes's formula**

Need to be made up

9. **Odds of Event**

The odds of event E, α, is defined as the ratio of probability of E and its compliment which is

$$\alpha = \frac{P(E)}{P(\overline{E})} = \frac{P(E)}{1 - P(E)}$$

then by looking at the ratio and compare it with 1 we can know which of E and its compliment is more likely to happen.

10. **Independence of events**

(a) Three ways to determine independence

$$P(B \mid A) = P(B)$$

$$P(A \mid B) = P(A)$$

$$P(A \cap B) = P(A)P(B)$$

The proof of third one is the rearrangement of the conditional probability formula.

**Notice: the basic logic is that the formula $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ is always true, and we know if A and B are independent then $P(A \mid B) = P(A)$.* Also the third one can be extended for n events.

(b) **Specify the difference and relationship with mutually exclusive events**:

In short, the conclusion is

$$Mutually\ exclusive \implies Dependent$$

$$Independent \implies Not\ mutually\ exclusive \implies Have\ intersection$$

One direction only!! Mutually exclusive means there is **no intersection** of two event which means when one event happens the other one can not happen at the same time. They can not happen simultaneously. It also means one event has influence on the probability of the other one happens.

But in order to understand independence, do not lay on Vinen diagram too much.

(c) **Conditional Independence**

Two events A and B are said to be conditionally independent if

$$P(B \mid A \cap E) = P(B)$$

$$P(A \mid B \cap E) = P(A)$$

$$P(A \cap B \mid E) = P(A \mid E)P(B \mid E)$$

Notice:

- Conditional and unconditional independence does not imply each other
- Conditional independence given G doe not implies conditional independence given G^c.

11. **Axioms of being a Cumulative distribution function**

Theorem 2. Let $\mathcal{F}$ denotes the cumulative distribution function.

- Monotone 单调性 $\mathcal{F}$ is an **non-decreasing** function defined on the whole $\mathbb{R}$ which is $(-\infty, +\infty)$, or precisely

$$\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2 \implies \mathcal{F}(x_1) \leq \mathcal{F}(x_2)$$

- Bounded 有界性　The value of $\mathcal{F}$ is bounded which is, $\forall\, x \in \mathbb{R}$, $0 \leq \mathcal{F}(x) \leq 1$,

$$\mathcal{F}(-\infty) = \lim_{x \to -\infty} \mathcal{F}(x) = 0, \;\; \mathcal{F}(+\infty) = \lim_{x \to +\infty} \mathcal{F}(x) = 1$$

- Right continuous 右连续性　$\mathcal{F}(x)$ is a right continuous function, which is

$$\lim_{x \to x_0^+} \mathcal{F}(x) = \mathcal{F}(x_0) \;\; or \;\; \mathcal{F}(x_0 + 0) = \mathcal{F}(x_0)$$

这三个条件是判断一个函数是否能成为分布函数 **(CDF)** 的充要条件.

12. **Discrete Random Variable**

Definition: The **Random Variable** is a **function**(or map) that maps a outcome in a sample space to a numerical quantity. We use uppercase letter to denote a random variable.

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

So X is based on the sample space. That means the independent variable can be anything, it does not matter it is a number or not but the value of X(ω) must be a real number.

Definition: Discrete random variable is when the value taken on by the random variable is finite or from a set of countably infinite set.

(a) **Probability mass function (PMF)**

$$f(a) = P(X = a)$$

- X is the random variable
- $0 \leq p(x) \leq 1$ for all x and $\sum_x p(x) = 1$.
- The counterpart for continuous random variable is called probability density function (PDF).

(b) **Cumulative distribution function (CDF)**

$$F(a) = P(X \geq a) = \sum_{all\ x \leq a} P(X)$$

Simply the accumulation of probability till a certain point. Notice it is a non-decreasing step function.

(c) **Expectation of Discrect Random variable X**

Denoted by $E(X)$ & μ, synonyms as the expectation of X, the mean/average of X which is

$$E(X) = \sum_{all\ X} x f(x)$$

- The E(X) is the weighted sum of x value, the P(X) are weights.
- The expectation is the long-run average of x value taken on by the random variable X if the experiments are to be repeated a large number of times.

To generalize the expectation, let $g(X)$ be a real function of X, then the expectation of $g(X)$ is

$$E(g(X)) = \sum_{all\ X} g(x)f(x)$$

- The linear property of expectation is: let a, b be real constant, X, X_1, X_2 be random variables and g_1, g_2 be real valued function, then we have

$$E(aX + b) = aE(X) + b$$

$$E(ag_1(X_1) + bg_2(X_2)) = aE(g_1(X_1)) + bE(g_2(X_2))$$

(d) **Variance and standard deviation of Discrete X**

Variance of X is denoted by Vn which is

$$V(X) = E\{(x - \mu)^2\} = \sum_{all\ x} (x - \mu)^2 f(x)$$

$$V(X) = E(X^2) - [E(X)]^2$$

Proof:

$$V(X) = E\{(x - \mu)^2\} = E(x^2 - 2\mu x + \mu^2)$$

$$= E(x^2) - 2\mu E(x) + \mu^2$$

$$= E(x^2) = 2E^2(x) + E^2(x)$$

$$= E(x^2) - E^2(x)$$

Done.

Standard deviation of X is

$$SD(X) = \sqrt{V(X)}$$

Similar to Expectation, $f(x)$ is the weight.

The linearity of variance is

$$V(aX + b) = a^2 V(X)$$

Proof is in Chapter4.

13. **Typical discrete distribution**

(a) **Binomial distribution 二项分布: $\mathcal{X} \sim Bin(n, p)$**

Random variable X is the number of success event A happens in n times Bernoulli experiments. Let the probability of event A happens be p, then

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

The cumulative probability function is

$$F(X) = \sum_{k=0}^{y} \binom{n}{k} p^k * (1 - p)^{n-k}$$

The expectation and variance are

$$E(Y) = np, \;\; Var(X) = np(1-p)$$

proof needed.

(b) **Negative binomial distribution** 负二项分布: $\mathcal{X} \sim Neg.Bin.(n, p)$

Let the random variable X be the number of experiments when the rth event A happen. The PMF is

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

The CPF is

$$F(X) = \sum_{k=r}^{y} \binom{k-1}{r-1} p^r * (1-p)^{k-r}$$

Starting from r is because event A has already happened r times so the total number of experiments should be at least r. The expectation and variance are

$$E(X) = \frac{r}{p}, \;\; Var(X) = \frac{r(1-p)}{p^2}$$

proof needed. One thing should be noticed here is, the last trail/experiment, event A must happen. Always, r is fixed, so $r-1$ is also fixed.

Actually, we want the event A to be happened r times (this is our purpose). So we stop repeating experiment once we get the number of happening of A which is r. 即一得到第 r 次发生就停止实验.

(c) **Geometric distribution** 几何分布: $\mathcal{X} \sim Geom.(p)$

Used when our interest is the probability of sth *fist* happens. Let the random variable X be the number of experiments that have been executed till we see the first success of event A. Then X is a Geometric random variable with parameter p, the probability of event A happens. The PMF is

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

the CPF is

$$F(X) = \sum_{k=1}^{y} p(1-p)^{k-1} = 1 - (1-p)^y$$

The expectation and variance are

$$E(X) = \frac{1}{p}, \;\; Var(X) = \frac{1-p}{p^2}$$

proof needed.

(d) **Hyper-geometric distribution** 超几何分布: $\mathcal{X} \sim Hypergeom(N, n, m)$

We have a small set of totally N objects, m of them are one kind and other $N-m$ are another kind. A total of $\#n$ are drawn form N *without replacement*. The distribution has three parameter which are

$$X \backsim Hypergeom(N, n, m)$$

It means the probability that k out of n are type A. The probability mass function is

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

where, $r=min\{m,n\}$.

The cumulative probability function is

$$F(X)=\sum_{k=0}^{y}\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

The expectation and variance are

$$E(X)=n\frac{m}{N},\ \ {}^{*}Var=n\frac{m}{N}(1-\frac{m}{N})\frac{N-n}{N-1}$$

proof needed.

(e) **Poisson distribution 泊松分布:** $\mathcal{X}\sim Poisson(\lambda)$

There are three conditions must be satisfied when using the poisson process:

- n is large
- p is small
- $np=\lambda$ is a constant

i. *Poisson limit*

The poisson limit is typically a approximation of binomial distribution under the above three conditions are satisfied.

Suppose $\mathcal{X}$ is a binomial random variable, where

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\ k=0,\ 1,\ ...,\ n$$

If $n\to\infty$ and $p\to 0$ in such a way that $\lambda=np$ remains constant, then

$$\lim_{n\to\infty,p\to 0,np=\lambda}P(X=k)=\lim_{n\to\infty,p\to 0,np=\lambda}\binom{n}{k}p^k(1-p)^{n-k}=\frac{e^{-np}(np)^k}{k!}$$

proof needed (technical point here is the definition of e^x which is $\lim_{n\to\infty}(1+\frac{x}{n})^n=e^x$)

ii. *Poisson distribution and Poisson process*

The probability mass function of poisson distribution is just the Poisson limit.

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

**The proof is on page227 of Book An introduction to Mathematical statistics and Its Application which is simply checking whether the function satisfy the axioms of being a probability function.*

another way to approach the poisson distribution is needed, which is in terms of average rate.

14. **Continuous distribution**

(a) **Normal distribution 正态分布: $\mathbf{X} \sim \mathbf{N}(\mu, \sigma^2)$**

Normal distribution has two parameter, one is μ and one is σ. μ is called 位置参数 and σ is called 尺度参数. 改变 μ, 即改变中轴对称线的位置因此图像左右移动而形状不变, 若改变 σ 图像扁平程度发生改变——σ 愈大则愈扁平, 愈小则愈瘦尖. 另外, $\mu \pm \sigma$ 是图像的拐点 (inflection point).

- **Probability density function**

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

- **Cumulative distribution function**

$$\mathcal{F}(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

- **Expectation:**

$$E(X) = \mu$$

Proof. Let $X \sim N(\mu, \sigma^2)$. Then by Theorem 3 we have $U = \frac{x-\mu}{\sigma} \sim N(0, 1)$, the Standard normal distribution. Then, by definition of expectation of continuous variable we have

$$E(U) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cdot e^{-\frac{u^2}{2}} du$$

Obviously the expectation is a odd function, so the integral equals **zero**. Then we have

$$U = \frac{x-\mu}{\sigma} \implies x = \mu + \sigma u$$

Then operate the expectation on both side we get

$$E(X) = E(\mu + \sigma u) \implies E(X) = \mu + \sigma E(\mu) = \mu$$

□

- **Variance:**

$$Var(X) = \sigma^2$$

Proof. Similarly, let X and U be such normal distribution. Then

$$Var(U) = E(U^2) - E^2(U)$$

since $E^2(U) = 0$, then

$$Var(U) = E(U^2)$$

By the definition of expectation, we have

$$E(U^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 \cdot e^{-\frac{u^2}{2}} du$$

then Integral by substitution. Let

$$-\frac{1}{2}u^2 = y \implies -u\ du = dx$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \left(\int_{-\infty}^{\infty} -2y \cdot e^y dy \right)$$

□

i. ***Standard normal distribution:*** $X \sim N(0,1)$

Specifically, one of the normal distribution, $X \sim N(0,1)$ is called **Standard normal distribution**.

Theorem 3. If $X \sim N(\mu, \sigma^2)$, then $U = \frac{x-\mu}{\sigma} \sim N(0,1)$

Proof Let $\mathcal{F}_X(x) and \mathcal{F}_U(u)$ denotes the CDF of random variable X and U. Then by the definition of CDF we can get

$$\mathcal{F}_U(u) = P(U \leq u) = P(\frac{x-\mu}{\sigma} \leq u) \implies = P(x \leq \mu + \sigma u) = \mathcal{F}_X(\mu + \sigma u)$$

Since the CDF are continuous and differentiable every where, then the PDF of U is

$$p_U(u) = \frac{d}{du}\mathcal{F}_U(u) = \frac{d}{du}\mathcal{F}_X(\mu + \sigma u)$$

$$= p_X(\mu + \sigma u) \cdot \sigma \implies \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \sigma$$

plug into $x = \mu + \sigma u$, finally we get

$$p_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}}$$

which is exactly the PDF of standard normal distribution.
标准化正态分布的目的是可以用标准正态分布表获得需要的概率。

ii. ***3 σ law for normal distribution***

68.26% $\sim$ 95.45% $\sim$ 99.73%, each with a σ separation.

(b) **Uniform distribution 均匀分布: $\mathbf{X} \sim$ Uni(α, β)**

The two parameter are the lower and upper bound of the PDF of uniform distribution. The PDF are the same every where. The PDF is

$$p(x) = \begin{cases} \frac{1}{\beta - \alpha}, & x \in [\alpha, \beta]; \\ 0\ , & \text{otherwise} \end{cases}$$

the CDF is

$$F(x) = \begin{cases} 0\ , & x \in (-\infty, \alpha] \\ \frac{x-\alpha}{\beta-\alpha}\ , & x \in [\alpha, \beta) \\ 1\ , & x \in [\beta, \infty) \end{cases}$$

An example for uniform distribution is, the abrasion level （磨损程度）of tires （轮胎）。Since the probability of being worn are the same to all the points on the tire. So, the position on the tire worn follows uniform distribution, which is $X \sim (0, 2\pi r)$.

- **Expectation**

$$E(x) = \frac{\alpha + \beta}{2}$$

Proof By defination of expectation

$$E(X) = \int_{\alpha}^{\beta} x \cdot \frac{1}{\beta - \alpha} = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{\alpha + \beta}{2}$$

- **Variance**

$$Var(X) = \frac{(\beta - \alpha)^2}{12}$$

Proof By definition of variance,

$$Var(X) = E(X^2) - E^2(X) = \frac{\alpha^2 + \alpha \cdot \beta + \beta^2}{3} - \frac{(\alpha - \beta)}{4} = \frac{(\beta - \alpha)^2}{12}$$

(c) **Exponential distribution** 指数分布: $X \sim Expone(\lambda)$

Exponential distribution is related to Poisson distribution. It depicts the time between two consecutive happening events. Thus, let X be the random variable indicating the time between two events in the poisson distribution, then

- **Probability density function**

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

- **Cumulative distribution function**

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

指数分布是一种偏态分布，由于只能取非负实数，因此常被用于“寿命”分布。

- **Expectation**

$$E(X) = \frac{1}{\lambda}$$

Proof Simply take the integral of its *pdf* we get

$$E(X) = \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx = \lambda \int_0^{\infty} x e^{-\lambda x} dx$$

integral by parts, then

$$= \lambda(-\frac{1}{\lambda} e^{-\lambda x} \cdot x \Big|_0^{\infty} - \int_0^{\infty} -\frac{1}{\lambda} e^{-\lambda x}) = \frac{1}{\lambda}$$

- **Variance**

$$Var(X) = \frac{1}{\lambda^2}$$

Proof By definition of variance

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

$$= \int_0^\infty x^2 \cdot \lambda e^{-\lambda x} dx - \frac{1}{\lambda^2} = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

i. **Interpret exponential distribution from Poission distribution**

Let's say $X \sim Poission(\lambda)$. We want to generate the expression of exponential distribution under the Poisson distribution. Thus,

$$P\left(X > t\right) = P\left(\textit{Zero event happens in length of time } t\right)$$

or more formally, let $N \sim Poission(\lambda)$, then

$$P(X > t) = P(N(t) = 0) = \frac{(\lambda t)^0 \cdot e^{-\lambda t}}{0!} = e^{-\lambda t}$$

$$\implies P(X \leq t) = 1 - P(X > t) = 1 - e^{-\lambda t} = F_X(x)$$

then take the derivative of the CDF to get the PDF. 值得注意的是，此处在 e 的次数及底数上用 t 的原因是想 用 **t** 作为 **scalar** 来调节相应时间长度的 **rate**，因为从泊松分布及泊松极限的发生条件可以得知其发生的频率与时间长度成正比。

ii. **Memory-less property of exponential distribution** 指数分布的无记忆性

Theorem 4. Let the random variable $X \sim Expon(\lambda)$, then for any $s > 0$ and $t > 0$, we have

$$P(X > s + t \mid X > s) = P(X > t)$$

上式的含义是：记 X 为某种产品的使用寿命，若其服从指数分布，则，如果已知其使用了 s 时长未发生事故，则再能使用的时间 t 与已经使用的时间长度 s 无关，相当于从时刻 t 开始从心计算概率，简而言之即无记忆性。(此处说的已经使用 s 时长但却用的 $X > s$ 的原因是使用了 **s** 时长与使用时间大于 **S** 的概率等价，因为连续随机变量的分布中一个点无概率可言)。

Proof. Since X flows exponential distribution, then $P(X > s) = e^{-\lambda t}$ for which $s > 0$. Also obviously

$$\{X > s + t\} \subseteq \{X > s\} \implies \{X > s + t\} \cap \{X > s\} = \{X > s + t\}$$

then by definition of conditional probability

$$P(X > s + t \mid X > s) = \frac{P(\{X > s + t\} \cap \{X > s\})}{P(X > s)}$$

$$= \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$

□

(d) **Gamma distribution 伽马分布: $X \sim$ Gamma(α, λ)**

i. **Gamma Function** The function of form

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

is called **gamma function**. It has two main properties:

- $\Gamma(1) = 1$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
  Need to be made up The proof of the $\frac{1}{2}$ one need to use Gaussian Error Function
- $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$, easy to be prove by integral by parts. *Then by induction we have*

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) = n!$$

ii. **Gamma distribution**

- **PDF**

$$p(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\lambda x} & , x \geq 0 \\ 0 \ , & x < 0 \end{cases}$$

  where the parameter α is called the 形状参数 and the other parameter λ is called 尺度参数。

- **Expectation**

$$E(X) = \frac{\alpha}{\lambda}$$

  Proof. By definition of expectation we have

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

  it looks similar to the expression of gamma function, so by substitution, let $\lambda x = u$, then $\lambda\, dx = du$ and $x = \frac{u}{\lambda}$, plug in we get

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \int_0^\infty \frac{u^\alpha}{\lambda^{\alpha+1}} \cdot e^{-u}\, du$$

$$= \frac{1}{\Gamma(\alpha) \cdot \lambda} \cdot \Gamma(\alpha + 1) = \frac{\alpha \cdot \Gamma(\alpha)}{\Gamma(\alpha) \cdot \lambda} = \frac{\alpha}{\lambda}$$

  □

- **Variance**

$$Var(X) = \frac{\alpha}{\lambda^2}$$

  The proof is exactly the same as the expectation, omitted here.

iii. **Two special cases for gamma distribution**

- When $\alpha = 1$, the gamma distribution is the exponential distribution
- When $\alpha = \frac{n}{2}$, $\lambda = \frac{1}{2}$, we call this kind of gamma distribution the **Chi square distribution under the degree of freedom n,** 即自由度为 n 的卡方分布。Need to be made up

iv. **Interpretation of Gamma distribution**

Definition: Let n be the number of consecutive events happens which follows Poisson distribution, or simply the Poisson process. And by definition, the time between two consecutive events happen follows **exponential distribution**. Then, we define (can prove) the ***time needed to wait until the nth event happens follows Gamma distribution.***

a proof from Poisson process is needed

建立路径：泊松过程（分布）$\longrightarrow$ 指数分布 $\longrightarrow$ Gamma 分布

(e) **Beta distribution: X $\sim$ Be(a,b)**

i. **Beta function** The function in form

$$B(a,b) = \int_0^1 x^{a-1} \cdot (1-x)^{b-1}\, dx$$

is called **Beta function**, it has the following properties:

- B(a,b) = B(b,a)

  Proof. Let $y = 1 - x$, then

  $$B(a,b) = \int_1^0 (1-y)^{b-1} \cdot y^{a-1} \cdot (-1)\, dy = \int_0^1 (1-y)^{b-1} \cdot y^{a-1} \cdot\, dy = B(b,a)$$

  simply integral by substitution □

- The relationship between Beta and Gamma function is

  $$B(a,b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}$$

  Proof.

  proof need to be made up (need Jacobi determinate)

  □

- **PDF**

  $$p(x) = \begin{cases} \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot x^{a-1}(1-x)^{b-1} \ , \ 0 < x < 1 \\ 0 \ , \ otherwise \end{cases}$$

  where both a and b are shape parameter 形状参数。An important property is, when $a = b$, then density function is symmetric by $x = \frac{1}{2}$. (easy to check).

- **Expectation**

  $$E(X) = \frac{a}{a+b}$$

  Proof. Proof needed □

- **Variance**

$$Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Proof. Proof needed □

ii. Interpretation of Beta distribution

It can be used as a distribution of probability of an event. More need to be made up!!!

15. **Multi-dimensional random variable & its distribution**

Definition 0.1. Multi-dimensional random variable If $X_1(w), X_2(w), ..., X_n(w)$ are n random variables defined on the **same** sample space $\Omega = \{w\}$, then we call

$$\mathbf{X}(w) = \big(X_1(w), X_2(w), ..., X_n(w)\big)$$

the ***n-dimensional* random variable** or ***random vector***.

**Notice: The key point is that the multi-dimension rv. is defined on the smae sample space. If the random variable is defined on different sample space, let's say Ω_1 & Ω_2, then we can only discuss it based on the product space $\Omega_1 \times \Omega_2 = \{(w_1, w_2)\,|w_1 \in \Omega_1, w_2 \in \Omega_2\}$.*

Definition 0.2. Joint distribution function Let $x_1, x_2, ..., x_n \in \mathcal{R}$, then the probability of simultaneously happening events $\{X_1 \le x_1\}$, $\{X_2 \le x_2\}, ..., \{X_n \le x_n\}$

$$F(x_1, x_2, ..., x_n) = P(X_1 \le x_1,\ X_2 \le x_2,\ ...,\ X_n \le x_n)$$

is called the **Joint distribution function**.

Theorem 5. Properties of **F**(x,y)

- **Monotone** 单调性 $F(x, y)$ 分别对 x, y 单调非减。即,

$$x_1 < x_2 \implies F(x_1, y) \le F(x_2, y),\ \ \&\ \ y_1 < y_2 \implies F(x, y_1) \le F(x, y_2)$$

- **Boundary** 有界性 For any x & y, $0 \le F(x, y) \le 1$, and

$$F(-\infty, y) = \lim_{x \to -\infty} F(x, y) = 0$$

$$F(x, -\infty) = \lim_{y \to -\infty} F(x, y) = 0$$

$$F(\infty, \infty) = \lim_{x \to \infty, y \to \infty} F(x, y) = 1$$

- **Right continuous** 右连续性 For both variable are right continuous

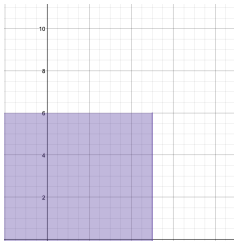
$$F(x + 0, y) = F(x, y)$$

$$F(x, y + 0) = F(X, y)$$

- **Non-negative 非负性** For any $a < b$, $c < d \in \mathbb{R}$,

$$P(a < x \leq b, c < y \leq d) = F(b,d) - F(a,d) - F(b,c) - F(a,c) \geq 0$$

the above line can be expressed as a domain of double integral/x-y plane



we can think the domain as the two dimension xy plane and the upper right corner is the position of any point (x, y). Then the CDF is just including and excluding the area to make sure not double counting.

Some proof may be needed

(a) **Joint distribution random variable (2 dimensional)**

Definition 0.3. (Discrete case) The probability mass function for joint distribution is

$$p_{ij} = P(X = x_i, Y = y_j), \quad i, j = 1, 2, ...$$

Definition 0.4. (Continuous case) The cumulative distribution function (CDF) of joint distribution is

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p(u, v)\, dudv$$

where p is a **non-negative** function. Then, we call the function

$$p(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

the **probability density function** of joint distribution.

Definition 0.5. Marginal mass function For random variable, their one mmf is

$$\sum_{j=1}^{\infty} P(X + x_i, Y = y_j) = P(X = x_i)$$

and similar for Y is

$$\sum_{i=1}^{\infty} P(X + x_i, Y = y_j) = P(Y = y_j)$$

Definition 0.6. Marginal density function Similarly as in the discrete case, mdf is

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

$$p_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx$$

May need some proof and reference to here Notice to clear the boundary of integration. 划清楚积分的边界。

(b) **Independence between random variables** 变量间独立性

Definition 0.7. Let n dimensional random vector $(X_1, X_2, ..., X_n)$ has the joint cumulative function $F(x_1, x_2, ..., x_n)$ and let $F_i(x_1)$ is the marginal cumulative function of X_i. Then, if

$$F(x_1, x_2, ..., x_n) = \prod F_i(x_i)$$

the the random variables $X_1, X_2, ..., X_n$ are independent.

Then, we can get a more useful conclusion: For simplicity, consider the case with only 2 random variable X, y, taking twice derivative on both side of the equation (CDF), for x first and then y (the order does not matter need to know in what cases the order of derivative does not matter). We get

$$\frac{\partial}{\partial x}F(x,y) = \frac{\partial}{\partial x}\Big(F_X(x)\cdot F_Y(y)\Big) = \frac{\partial}{\partial x}F_X(x) = \frac{\partial}{\partial x}F_X(x)\cdot F_Y(y) = f_X(x)\cdot F_Y(y)$$

The for Y we have

$$\frac{\partial^2}{\partial x \partial y}F(x,y) = f_X(x)\cdot\frac{\partial}{\partial y}F_Y(y) = f_X(x)\cdot f_Y(y)$$

which is exactly the same as

$$f_{XY}(x,y) = f_X(x)\cdot f_Y(y)$$

so we can use this to check the independence between variables. 即，边际密度函数乘积等于联合密度函数。

建立路径：如同之前检验独立性一样，需要两者概率乘积等于交集 **intersection** 的概率一样。在分布中能表示概率的函数是 **CDF**。因此我们用 **CDF** 表示独立性，进而通过推导得到密度函数乘积等于联合分布密度函数。

(c) **Properties of Expectation of joint distribution**

- if X and Y are independent, then $E(X+Y) = E(X) + E(Y)$ proof needed
- If X and Y are independent, then $E(XY) = E(X)\cdot E(Y)$ proof needed
- if X and Y are independent, then $Var(X \pm Y) = Var(X) \pm Var(Y)$ proof needed

(d) **Covariance in Joint distribution** 协方差

Definition 0.8. Let (X, Y) be a two dimensional random vector. If the expectation $E[(X - E(X))(Y - E(Y))]$ exists, then we call it the **Covariance** of X and Y, or the 相关（中心）矩, which is

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

It depicts the **correlation level** 相关程度 of the random variables. It has the following characteristics:

- Cov(X,Y) > 0, X and Y are positively related; 即有同时增加或减小的趋势 $\Longrightarrow$ **(x,y)** 分布于一三象限
- Cov(X,Y) < 0, X and Y are negatively related; 即有向不同方向增加或减小的趋势 $\Longrightarrow$ **(x,y)** 分布于二四象限

- Cov(X,Y) = 0, X and Y are either not related or there is a non-linear relation between the random variables.非线性关系参见课本 P178

i. **Properties**

- $Cov(X,Y) = E(XY) - E(X)E(Y)$ proof needed
- If X and Y are independent, then $Cov(X,Y) = 0$. 反之不然。注意独立与不相关在逻辑上是包含关系——即不相关包含独立。
- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$ proof needed.
  注意 cov(X,Y) 正负号与 Var(X + Y) 的正负号相同。
  Then generalize it to n dimensional cases,

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum\sum_{i<j} a_i a_j Cov(X_i, X_j)$$

- $Cov(X,Y) = Cov(Y,X)$. Travail by the definition of covariance.
- $Cov(X,a) = 0$ where a is a constant. proof needed.
- $Cov(aX,bY) = abCov(X,Y)$ proof needed.
- $Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$ proof needed.
- $Cov(X,X) = Var(X)$ proof needed.
- $Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, Y_j)$

ii. **Correlation coefficient** 相关系数

协方差有单位, 如米 (m), 秒 (s) etc. 为消除单位的影响, 引入新的概念 **correlation coefficient.**

Definition 0.9. Let (X,Y) be a two dimensional random vector, and $Var(X) = \sigma_X^2 > 0$, $Var(Y) = \sigma_Y^2 > 0$, then we call

$$Corr(X,Y) = \rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

the **(linear) correlation Coefficient** （线性）相关系数。

Another interpretation of correlation coefficient is, it is the **Covariance** of the **standardized** random variable, which is: Let the expectation of X and Y are μ_X and μ_Y respectively, and

$$X^* = \frac{X - \mu_X}{\sigma_X}, \quad Y^* = \frac{Y - \mu_Y}{\sigma_Y}$$

then we have

$$Cov(X^*, Y^*) = Cov(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho(X,Y)$$

some more consideration is needed here.

The thing it suggests more than *Covariance* is **if the correlation coefficient = ±1, it means X and Y are perfect linearly related (perfect positive or negative related).**

(e) **Conditional distribution and expectation** 条件分布与条件期望

i. **Conditional distribution for discrete cases**

First, define the joint distribution (X, Y) to be

$$p_{ij} = P(X = x_i, Y = y_j)$$

then, the definition goes here

Definition 0.10. For every y_j that makes $P(Y = y_i) = p_{\cdot j} = \sum_{i=1}^{\infty} p_{ij} > 0$, we call

$$p_{i|j} = P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{i\cdot}}$$

the conditional distribution (PMF) of X under the given $Y = y_j$. Same for Y given $X = x_i$.

Definition 0.11. The CDF of X given Y=y_j is defined as

$$F(x|y_i) = \sum_{x_i \le x} P(x_i|y + j) = \sum_{x_i \le x} p_{i|j}$$

same for Y given X.

值得注意的是，联合分布列只有一个，而条件分布列却有多个，富含大量信息。其数亩取决于变量组合数目的多少。

ii. **Conditional distribution for continuous cases**

The case for continuous is different. Since for each point the probability is 0. So it is reasonable to take the limit for a given value of y. First give the definition.

Definition 0.12. For any y that makes $p_Y(y) > 0$, the conditional distribution function and conditional density function under given x are

$$F(x \,|\, y) = \int_{-\infty}^{x} \frac{p(u, v)}{p_Y(y)}\, du, \quad p(x \,|\, y) = \frac{p(x, y)}{p_Y(y)}$$

Proof. By the definition of conditional probability,

$$\begin{aligned} P(X \le x \mid Y = y) &= \lim_{h \to 0} P(X \le x \mid y \le Y \le y + h) \\ &= \lim_{h \to 0} \frac{P(X \le x \mid y \le Y \le y + h)}{P(y \le Y \le y + h)} \\ &= \lim_{h \to 0} \frac{\displaystyle\int_{-\infty}^{x} \int_{y}^{y+h} p(u, v)\, dv\, du}{\displaystyle\int_{y}^{y+h} p_Y(v)\, dv} \end{aligned}$$

we apply as small trick here

$$= \lim_{h \to 0} \frac{\displaystyle\int_{-\infty}^{x} \left\{ \frac{1}{h} \int_{y}^{y+h} p(u, v)\, dv \right\} du}{\displaystyle\frac{1}{h} \int_{y}^{y+h} p_Y(v)\, dv}$$

the reason for multiplying by $\frac{1}{h}$ is to construct a scenario to use **mean value theorem of calculus.** Then, if $p_Y(y)$ and $p(x, y)$ are both continuous at y (*The condition must satisfy to use mean value theorem*), we can get

$$\lim_{h\to 0}\frac{1}{h}\int_y^{y+h} p(u,v)dv = \lim_{h\to 0}\frac{1}{h}\cdot h\cdot p(u,h') = p(u,y)$$

where $h' \in [y,\, y+h]$ and when h is approaching 0, h' is exactly the point y;

$$\lim_{h\to 0}\frac{1}{h}\int_y^{y+h} p_Y(v)\,dv = \lim_{h\to 0}\frac{1}{h}\cdot h\cdot p_Y(h') = p_Y(y)$$

finally combine them together we get the statement in definition. □

iii. **Conditional expectation** 条件期望

Definition 0.13. If the expectation of conditional distribution exists, it is called the conditional expectation. It is defined as

$$E(X\,|\,Y=y) = \begin{cases} \sum_i x_i P(X=x_i\,|\,Y=y) \text{ , (X,Y) are two dimensional discrete random vector,} \\ \int_{-\infty}^{\infty} xp(x\,|\,y)\,dx \text{ , (X,Y) are two dimensional continuous random vector.} \end{cases}$$

The case for Y given x is the same.

值得指出的是,在 **X given Y** 的情况下,对于 **X** 的条件期望是关于 **Y** 的函数。进一步,$E(X\,|\,Y=y)$ 本身可以写成 $E(X\,|\,Y=y)=g(y)$, 这样期望本身也可以看做一个随机变量。

Theorem 6. Expectation by conditioning (重期望公式) Let (X,Y) be a two dimensional random vector and $E(X)$ exists, then

$$E(X) = E_Y(E_{X\,|\,Y}(X\,|\,Y))$$

Proof. (For continuous case only) Set up the notation. Let $p(x, y)$ be the joint density distribution function. Then

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xp_X(x)\,dx = \int_{-\infty}^{\infty} x(\int_{-\infty}^{\infty} p(x,y)\,dy)dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)\,dx\,dy \end{aligned}$$

by definition of conditional probability we have

$$\begin{aligned} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\cdot p(x\,|\,y)\cdot p_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} p_Y(y)\Big\{\int_{-\infty}^{\infty} x\cdot p(x\,|\,y)\,dx\Big\}\,dy \end{aligned}$$

the stuff within brace is exactly the $E_{X|Y}(X\,|\,Y)$. Then substitute it by a random variable/ function of y $g(y)$. Then

$$= \int_{-\infty}^{\infty} g(y)\cdot p_Y(y)\,dy = E(g(Y)) = E(E(X\,|\,Y))$$

□

概率论中较为深刻的结论。其内在逻辑是 如果要求一个在很大范围内的随机变量的均值/期望, 而计算十分复杂 $\Longrightarrow$ 选取一个有关的随机变量 $\mathbf{Y}$, 以 $\mathbf{Y}$ 作为条件将 $\mathbf{X}$ 分成很多的小块, 求每个小块的期望 $\Longrightarrow$ 再对此类以 $\mathbf{Y}$ 的性质来加权平均