MIT Intro to Probability 18.600

Ian Poon

November 2024

"*The utility of the measure-theoretic treatment of probability is that it unifies the discrete and the continuous cases, and makes the difference a question of which measure is used. Furthermore, it covers distributions that are neither discrete nor continuous nor mixtures of the two.*" ,Wikipedia. 18.600 is not a hard prerequisite of measure theoretic probaility theory(covered in 18.175 theory of probability) but is "*highly recommended*" apparently. It is essentially treatment of continuous distributions using only rienmann integrals(no lebesgue and measure theory etc). It satisfies most statistics prerequisties for pure mathematics statistics(except some graduate courses requiring 18.175) courses and *all* CS courses. It builds upon 6.042j too apparently. Of the 3 possible 1st courses to probability, 18.600 is the most rigorous and the preferred choice for pure mathematicians. The book used is `Sheldon Ross - A first course in probability 8th edition`

Contents

Remark 1. *Note that I have omitted the study of gamma, beta distributions(their iid sum, properties etc as in chapter 5,6). I feel it would be better to study the gamma,beta functions after a full treatment with* `MIT 18.112 Complex Analysis` *which you have only done halfway as of the date of creation of this document. Sadly, the proofs for some of the limit theorems that I was pretty excited for like CLT made assumptions on the continuity of certain functions. Again this is clearly not the most rigorous way to derive such theorems and I will wait till 18.175 Theory of Probability. I know for a fact the assumption required is "Lebesgue integrable".Measure theory go brr...*

1 Combinatorial Analysis(1)

Skipped as I have covered `MIT 18.211 Combinatorial Analysis` before which covers the topic with greater depth as well as graph theory

2 Axioms of Probability(2)

Definition 2

the set of all possible outcomes is called a **sample space**

Definition 3

We denote

- EF to be the intersection of set E and F
- $E \cup F$ the union
- E^c the complement

Proposition 4 (Basic Set Theory)

Note that

- Commutative laws: $E \cup F = F \cup E$
- Associative laws $(E \cup F) \cup G = E \cup (F \cup G) \quad (EF)G = E(FG)$
- Distributive laws $(E \cup F)G = EG \cup FG \quad EF \cup G = (E \cup G)(F \cup G)$

Proof. trivial and easy verifiable by venn diagrams

Proposition 5 (De Morgan)

$$\left(\bigcup_i E_i\right)^c = \bigcap_i E_i^c$$
$$\left(\bigcap_i E_i\right)^c = \bigcup_i E_i^c$$

Proof. Recall `Rudin:Intro to analysis`

Definition 6

Suppose we conduct an experiment in sample space S; The **probability** of an event E is defined by

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

where $n(E)$ is the number of times in the first n repetitions that the event E occurs

Fact 7 (Axioms of probability)

Consider

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. Fpr a sequence of mutually exlcusive events $E_1, E_2 \ldots$ that is $E_i E_j = \emptyset, i \neq j$ we have that

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

Axiom (3) essentially states that by assumption probability is a continuous function

Proposition 8

To make sense of this, we first assume certain axioms that allow such a definition to make sense and then explore the properties further

1. $P(E^c) = 1 - P(E)$
2. If $E \subset F$ then $P(E) \leq P(F)$
3. $P(E \cup F) = P(E) + P(F) - P(EF)$

Proof. For (1) consider by axiom (1) and (3) we have

$$P(E) + P(E^C) = 1$$

For (2) consider

$$F = E \cup E^C F$$

and that $P(E^C F) \geq 0$. For (3) consider that from axiom (3) we have

$$P(E \cup F) = P(E \cup E^c F) = P(E) + P(E^c F)$$

but since $F = EF \cup E^C F$ whe again obtain from axiom (3)

$$P(F) = P(EF) + P(E^c F)$$

on comparision of the 2 equatoins above, cancelling out the $P(E^c F)$ the proposition follows.

Proposition 9 (Inclusion Exlcusion)

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \ldots < i_r} P(E_{i_r} \ldots E_{i_r})$$

Proof. recall `MIT 18.211 Combinatory Analysis` which denoted "size" of set $|E|$ instead of $P(E)$. See probability as a *measure* to get this result(this will make more sense when you study theory of probability)

Definition 10

We say a sequence of events is an increasing sequence if

$$E_1 \subset E_2 \ldots \subset E_n \ldots$$

and decreasing if

$$E_1 \supset E_2 \ldots \supset E_n \ldots$$

Proposition 11

If $\{E_n, n \geq 1\}$ is either an increasing or decreasing sequence of events then

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

Proof. Notice that $\lim_{n \to \infty} E_n = \bigcup_{n=1}^{\infty} E_n$ and then apply axiom (3) in 7

3 conditional probability and independence(3)

Definition 12 (Conditional Probability)

If $P(F) > 0$ then

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

Corollary 13 (Multiplication rule)

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

Proof. Just expand it out. Notice in between successive fractions the denominator and numerator terms cancel out.

Theorem 14 (Baye's Formula)

Let E and F b events we may express E(see venn diagram as)

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

In which case we have

$$P(E) = P(E|F)P(F) + P(E|F^C)[1 - P(F)]$$

which actually make sense intuitively tbh

Proof. consider

$$\begin{aligned} P(E) &= P(EF) + P(EF^C) \\ &= P(E|F)P(F) + P(E|F^C)P(F^C) \\ &= P(E|F)P(F) + P(E|F^C)[1 - P(F)] \end{aligned}$$

Definition 15

The **odds** of an even A is defined by

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

The odds of F *given* an event E is then

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

Definition 16

Two events E and F are **independent** if

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

and dependent if not

notice that this implies for a pair of independent events

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

Definition 17

3 events are said to be independent if

$$P(EFG) = P(E)P(F)P(G)$$

and

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

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

$$P(FG) = P(F)P(G)$$

Example 18 (The problem of points)

see book

Example 19 (Gambler's Ruin)

see book

Proposition 20

Consider

1. $0 \leq P(E|F) \leq 1$
2. $P(S|F) = 1$
3. $P\left(\bigcup_{i=1}^{\infty} E_i|F\right) = \sum_{i=1}^{\infty} P(E_i|F)$

Proof. for (a) see that by defintition $0 \leq P(EF) \leq 1$ and $0 \leq P(F) \leq 1$. Then consider that since $EF \subset F$ we have $P(EF) \leq P(F)$ from 8

$$0 \leq P(EF)/P(F) \leq 1$$

for (b) see that

$$P(S|F) = \frac{P(SF)}{P(F)} = \frac{P(F)}{P(F)} = 1$$

(c) is obvious, the events still remain mutually exclusive

Example 21 (theory of runs)

see book

4 Random variables(4)

Frequently when an experiment is performed we are interested in some function of the outcome rather than the actual outcome itself. For instance we might be interested is the sum of 2 dice throws rather the outcomes $(1,6),(2,5)\ldots$ etc. These real valued functions/quantities of interest are what we call **random variables**

Example 22

Let Y denote the number of heats that appear then we could denote say

$$P\{Y=1\} = P\{(T,T,H),(T,H,T),(H,T,T)\} = \frac{3}{8}$$

Definition 23

For a random variable X the function F defined by

$$F(x) = P\{X \leq x\} \quad -\infty < x < \infty$$

is called the **cumulative distribution function** or simply the *distribution function*

Example 24 (Card Collection)

Suppose that there are N distinct types of coupons and that each time one obtains a coupon it is independently of previous selections,equally likely to be any one of the N types. The random variable of interest is T which the number of coupons that needs to be collected until one obtains a complete set of at least one of each type

Solution. Fix n and define events $A_1, A_2, \ldots, A_N$ as follows: A_j is the event where no type j coupon is contained among the first n coupons. Then

$$\begin{aligned} P\{T>n\} &= P\left\{\bigcup_{j=1}^{n} A_j\right\} \\ &= \sum_j P(A_j) - \sum\sum_{j_1<j_2} P(A_{j_1}A_{j_2}) + \ldots \\ &+ (-1)^{k+1}\sum\sum\sum_{j_1<j_2<\ldots<j_k} P(A_{j_1}A_{j_2}\ldots A_{j_k}) \ldots + (-1)^{N+1}P(A_1A_2\ldots A_N) \end{aligned}$$

the first equality follows because if within n draws and missing a distinct type then clearly need more than n draws to get the full set. The rest follows by **inclusion exlcusion identity**. Also see that

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

and that

$$P(A_{j_1}A_{j_2}) = \left\{\frac{N-2}{N}\right\}^n$$

and so on. Therefore we have

$$P\{T>n\} = \sum_i^{N-1} \binom{N}{i} \left\{\frac{N-i}{N}\right\}^n (-1)^{i+1}$$

and we can find $P\{T = n\}$ via

$$P\{T = n\} = P\{T > n-1\} - P\{T > n\}$$

Definition 25

A random variable can take on at most a countable number of possible values is said to be **discrete**.

Definition 26

For a **discrete random variable** X we define the **probability mass function** $p(a)$ of X by

$$p(a) = P(X = a)$$

where $p(a)$ is positive for at most a *countable* number of values of a that is

$$\begin{cases} p(x_i) \geq 0 & i = 1, 2 \dots \\ p(x) = 0 & \text{for all other values of } x \end{cases}$$

since X must take on one the values x_i we have

$$\sum_{n=1}^{\infty} p(x_i) = 1$$

4.1 expectation

Definition 27

the **expected value** is defined by

$$E(X) = \sum_{x;p(x)>0} xp(x)$$

Proposition 28

The expected value of random variable X is equal to the mean

Proof. Consider for the discrete case we have

$$E[X] = \sum_{i}^{n} x_i p(x_i) = \sum_{i}^{n} \frac{x_i}{n} = \mu$$

Similarly for the continuous case. The key idea is every value the random variable takes as an equal probability of occurring

Example 29

Let X denote a random variable that takes on any of the values $-1, 0$ and 1 with respective probabilities

$$P\{X=-1\}=.2 \quad P\{X=0\}=.5 \quad P\{X=1\}=.3$$

Compute $E[X^2]$

Solution. Let $Y=X^2$. Then we have

$$P\{Y=1\}=P\{X=-1\}+P\{X=1\}=.5 \tag{1}$$

$$P\{Y=0\}=P\{X=0\}=.5 \tag{2}$$

Hence

$$E[X^2]=1(.5)+0(.5)=.5$$

Proposition 30

If X is a discrete random variable that takes on one of the values $x_i, i \geq 1$ with respective probabilities $p(x_i)$ then for any real valued function g

$$E[g(X)]=\sum_i g(x_i)p(x_i)$$

Proof. Suppose that $y_j, j \geq 1$ represents the different values of $g(x_i), i \geq 1$ then grouping all the $g(x_i)$ having the same value gives

$$\begin{aligned}
\sum_i g(x_i)p(x_i) &= \sum_j \underbrace{\sum_{i;g(x_i)=y_j} g(x_i)p(x_i)}_{\text{grouping}} \\
&= \sum_j \sum_{i;g(x_i)=y_j} y_j p(x_i) \\
&= \sum_j y_j \sum_{i;g(x_i)=y_j} p(x_i) \\
&= \sum_j y_j P\{g(X)=y_j\} \\
&= E[g(x)]
\end{aligned}$$

For example for 29 we get

$$E\left\{X^2\right\}=(-1)^2(.2)+0^2(.5)+1^2(.3)=.5$$

which agrees with our result.

Corollary 31

If a and b ar constants then

$$E[aX+b]=aE[X]+b$$

Proof. Consider

$$\begin{aligned} E[aX+b] &= \sum_{x;p(x)>0} (ax+b)p(x) \\ &= a\sum xp(x) + b\sum p(x) \\ &= aE[X]+b \end{aligned}$$

Definition 32

The expected value of a random variable X as given by $E[X]$ is also referred to as the **mean** or the **first moment** of X. The quantity $E[X^n], n \geq 1$ is called the **nth moment** of X

See from the above that in general

$$E[X^n] = \sum_{x;p(x)>0} x^n p(x)$$

4.2 variance

So far we have $E(X)$ which gives the weighted average of our data. Suppose now we want a measure of the spread of data. So we define

Definition 33

If X is a random variable with mean μ then the **variance** of X denoted by $\text{Var}(X)$ is defined by

$$\text{Var}(X) = E[(X-\mu)^2]$$

Alternatively we have

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= \sum_x (x-\mu)^2 p(x) \\ &= \sum_x x^2 p(x) - 2\mu\sum_x xp(x) + \mu^2\sum_x p(x) \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= \boxed{E[X^2] - (E[X])^2} \end{aligned}$$

Corollary 34

See that we then have

$$\text{Var}(aX+b) = a^2\,\text{Var}(X)$$

Proof. consider

$$\begin{aligned}\operatorname{Var}(aX+b) &= E[(aX+b-E[aX+b])^2] \\ &= E[(aX+b-a\mu-b)^2] \\ &= E[a^2(X-\mu)^2] \\ &= a^2 E[(X-\mu)^2] \\ &= a^2 \operatorname{Var}(X)\end{aligned}$$

4.3 bernoulli

Definition 35

A random variable X is said to be a **Bernoulli random variable** if the probability mass function of X is given by

$$p(0) = P\{X=0\} = 1-p$$

$$p(1) = P\{X=1\} = p$$

The *bernoulli random variable* is fact part of a general type of random variables called binomial random variables.

Definition 36

Suppose now that n independent trials, each of which results in a success of proability p and failure $1-p$ then the **binomial random variable** having paramters (n,p) has probability mass function

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0,1,\ldots,n$$

which represents the probability of a sequnece of i successes and $n-i$ failures.

Also notice that

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = [p+(1-p)]^n = 1$$

by binomial expansion which shows that such a probability mass function is indeed well defined

Remark 37. *Notice that the* **Bernoulli random variable** *is simply the binomial random variable with parameters* $(1,p)$

Proposition 38

For a **binomial** random variable X with paramters n,p we have

$$E[X^k] = \sum_{i=1}^{n} i^k p(i) = \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i}$$

where i is the number of succeses

Using the identity

$$i\binom{n}{i} = n\binom{n-1}{i-1} \tag{3}$$

recall this result `MIT 18.211 Combinatory Analysis` we have

$$\begin{aligned} E[X^k] &= np\sum_{i=1}^{n} i^{k-1}\binom{n-1}{i-1}p^{i-1}(1-p)^{n-i} \\ &= np\sum_{j=0}^{n-1}(j+1)^{k-1}\binom{n-1}{j}p^{i}(1-p)^{n-1-j} \\ &= npE[(Y+1)^{k-1}] \end{aligned}$$

where the second line follows by change of varirable $j = i = 1$. Then in the special case where $k = 1$ we have

$$E[X] = np$$

To get the variance consider when $k = 2$. We then get

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

Since $E(X) = np$ we obtain

$$\text{Var}(X) = E[X^2] - (E(X))^2 = np[(n-1)p+1] - (np)^2 = np(1-p)$$

Theorem 39

In summary we have shown for a **binomial random variable**

$$E[X] = np$$

and

$$\text{Var}[X] = np(1-p)$$

Proposition 40

If X is a binomial random variable with paramters n, p where $0 < p < 1$ then as k goes from 0 to n, $P\{X = k\}$ first increases montonically then decreases monotonically reaching its largest value when k is the largest integer less than or equal to $(n+1)p$

Proof. Find

$$\begin{aligned} \frac{P\{X=k\}}{P\{X=k-1\}} &= \frac{\frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}}{\frac{n!}{(n-k+1)!(k-1)!}p^{k-1}(1-p)^{n-k+1}} \\ &= \frac{(n-k+1)p}{k(1-p)} \end{aligned}$$

Hence $P\{X = k\} \geq P\{X = k-1\}$ if and only if

$$(n-k+1)p \geq k(1-p)$$

or equivalently

$$k \geq (n+1)p$$

Example 41

For example

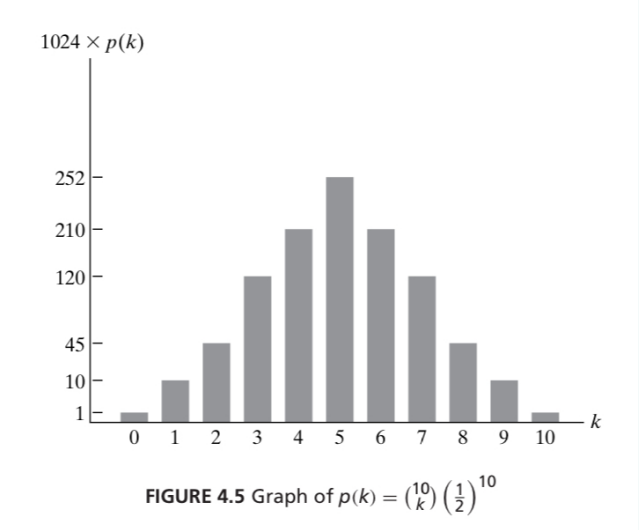


FIGURE 4.5 Graph of $p(k) = \binom{10}{k}\left(\frac{1}{2}\right)^{10}$

From 40 it provides a useful recursive formula for the binomial distribution function

$$P\{X \leq i\} = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}, \quad i = 0, 1 \ldots, n$$

via

$$P\{X = k+1\} = \frac{p}{1-p}\frac{n-k}{k+1} P\{X = k\}$$

4.4 poisson

Definition 42

A random variable X that takes on one of the values $0, 1, 2 \ldots$ is said to be a **poisson random variable** with paramter λ if for some $\lambda > 0$ we have

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

this then defines a well defined probability mass function since $\sum_{i=0}^{\infty} p(i) = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1$

Theorem 43

When n is large and p is small enough such that np is of moderate size then the binomial random variable may be approximated by the poisson random variable

Proof. To see this let $\lambda = np$ then consider

$$\begin{aligned} P\{X = i\} &= \frac{n!}{(n-1)!i!} p^i (1-p)^{n-i} \\ &= \frac{n(n-1)\ldots(n-i+1)}{n!}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \end{aligned}$$

but when n is large and λ is moderate we have

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \quad \frac{n(n-1)\dots(n-i+1)}{n^i} \approx 1, \quad \left(1-\frac{\lambda}{n}\right)^i \approx 1$$

so that we have

$$P\{X=i\} \approx e^{-\lambda}\frac{\lambda^i}{i!}$$

Proposition 44

Let X be poisson random variable then

$$E[X] = \text{Var}[X] = \lambda$$

Remark 45. *Intuitively you could see that this is true when you approximate it to be binomial variable using large n and moderate $\lambda = np$.In which case since $\text{Var}(x) = np(1-p)$ for binomial variable and that p is small we have $\text{Var}(x) \approx \lambda$. We now verify this result more rigourously below*

Consider

$$\begin{aligned} E[X] &= \sum_{i=0}^{\infty} \frac{ie^{-\lambda}\lambda^i}{i!} \\ &= \lambda \sum_{i=1}^{\infty} \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!} \\ &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \quad \text{letting } j = i-1 \\ &= \lambda \quad \text{since } \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda} \end{aligned}$$

Do similar calculations for $E(X^2)$ and use the formula $\text{Var}(X) = E[X^2] - (E[X])^2$ to obtain the result for the variance

Fact 46

Examples of random variables that generally obey the Poisson probability law are as follows

1. the number of misprints on a page of a book
2. the number of people in a community who survive to age 100
3. etc

They all are approximately Poisson for the same reason, that is by 43 since n large and p small binomial can be approximated by Poisson

Example 47

Consider an experiment counting the number of alpha particles given off in a 1 second interval by 1 gram of radioactive material. Suppose we know from the past that on average 3.2 alpha particles are given off. What is a good approximation to the probability that no more than 2 alpha particles will appear?

Solution. think of a gram of radioactive material as consisting of *large* number n atoms each of which has probability $3.2/n$ of decaying. Then we may approximate as poisson with $\lambda = 3.2$ to get

$$P\{X \leq 2\} = e^{-3.2} + 3.2e^{-3.2} \approx .3799$$

4.5 geometric

Definition 48

Suppose that independent trails each having a probability $p, 0 < p < 1$ of being a success are performed *until* a success occurs. If we let X equal the number of trials required then

$$P\{X = n\} = (1-p)^{n-1}p, \quad n = 1, 2, \ldots$$

Remark 49. *The keyword here is "until" which signals the use of geometric random variable*

Note that this a well defined probability density function since

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

where the second equality follows by sum of a convergent geometric series

Example 50

An urn contains N white and M black balls. Balls are randomly selected one at a time *until* a black one is obtained. If we assume that each ball selected is replaced before the next one is drawn what is the probability that

(a) exactly n draws are needed?

(b) at least k draws are needed?

Solution. Let X denote the number of draws needed to select a black ball then X satisfies the above equation with $p = M/(M+N)$. Hence for (a)

$$P\{X = n\} = \left(\frac{N}{M+N}\right)^{n-1}\frac{M}{M+N} = \frac{MN^{n-1}}{(M+N)^n}$$

and for (b)

$$\begin{aligned} P\{X \geq k\} &= \frac{M}{M+N}\sum_{n=k}^{\infty}\left(\frac{N}{M+N}\right)^{n-1} \\ &= \left(\frac{M}{M+N}\right)\left(\frac{N}{M+N}\right)^{k-1} / \left[1 - \frac{N}{M+N}\right] \\ &= \left(\frac{N}{M+N}\right)^{k-1} \end{aligned}$$

Corollary 51

From the above example is it is clear to see that for a geometric fandom variable X with probability of success p we have

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

Proposition 52

The expectation and variance of a geometric random variable is given by

$$E[X] = \frac{1}{p}$$

and

$$\text{Var}(X) = \frac{1-p}{p^2}$$

respectively

Proof. With $q = 1 - p$ we have

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} i q^{i-1} p \\ &= \sum_{i=1}^{\infty} (i-1+1) q^{i-1} p \\ &= \sum_{i=1}^{\infty} (i-1) q^{i-1} p + \sum_{i=1}^{\infty} q^{i-1} p \\ &= \sum_{j=0}^{\infty} j q^{j} p + 1 \\ &= q \sum_{j=1}^{\infty} j q^{j-1} p + 1 \\ &= qE[X] + 1 \end{aligned}$$

then upon rearrangement the proposition is proved for the expectation. As for the variance just do the same steps, find the expression for $E[X^2]$ then use $\text{Var}(X) = E[X^2] - (E[X])^2$

4.6 Negative binomial random variable

It turns out that the geometric random variable belongs to more general type of random variables known as negative binomial random variables.

Definition 53 (Negative Binomial Random Variable)

Suppose that independent trials each having probability $p, 0 < p < 1$ of being a success are perforemed until a total of r successes is accumulated. If we let X equal the number of trials required then

$$P\{X = n\} = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \quad n = r, r+1 \dots$$

We say X whose probability density is given by the above is a **negative binomial random variable** with paramters (r, p)

Remark 54. *Clearly the* **geometric random variable** *is the negative binomial random variable with paramters* $(1, p)$

Firstly this makes sense because of the word "until" we know the nth trail must be a success. Then to have a total of r successes then there must be $r-1$ successes in the first $n-1$ trials.

Proposition 55

This is a well defined probability distribution function as

$$\sum_{n=r}^{\infty} P\{X = n\} = \sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} = 1$$

Proof. Rewrite the sum in terms of a new index $k = n - r$, which shifts the index so that the summation starts from $k = 0$:

$$\sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} = p^r \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} (1-p)^k.$$

Now, our goal is to evaluate the sum:

$$\sum_{k=0}^{\infty} \binom{r+k-1}{r-1} (1-p)^k.$$

The expression $\binom{r+k-1}{r-1}$ is the number of ways to distribute k indistinguishable items into r distinguishable bins, which suggests using generating functions. Specifically, there is a known generating function for this type of sum:

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

which converges for $|x| < 1$. Recall this indentity from `MIT 18.211 combinatory analysis` under *generalized binomial theorem*

Substitute $x = 1 - p$ into the generating function:

$$\sum_{k=0}^{\infty} \binom{r+k-1}{r-1} (1-p)^k = \frac{1}{(1-(1-p))^r} = \frac{1}{p^r}.$$

Now we can substitute back into our original expression:

$$p^r \sum_{k=0}^{\infty} \binom{r+k-1}{r-1} (1-p)^k = p^r \cdot \frac{1}{p^r} = 1.$$

which completes the proof analytically.

Problem 56

Compute the expected value and the variance of a negative binomial random variable with parameters r and p

Solution. refer to book. Hint: its very similar to 38. Doing so you will get

$$E[X] = \frac{r}{p}$$

and

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

5 Continuous random variables(5)

Definition 57

We say that X is a **continuous random variable** if there exists a nonnegative function f defined for all real $x \in (-\infty, \infty)$ such that for any set B of real numbers

$$P\{X \in B\} = \int_B f(x)dx$$

the function f is called the **probability density** function of the random variable X. That means for it to be well defined we must have

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx$$

and so

$$P\{X < a\} = P\{X \leq a\} = F(a) = \int_{-\infty}^{a} f(x)dx$$

Definition 58

The expectation of a continuous random variable X is

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Fact 59 (Density is the derivative of the cumulative function)

Note that the relationship between the cumulative distribution F and the probability density f is expressed by

$$F(a) = P\{X \in (-\infty, a]\} = \int_{-\infty}^{a} f(x)dx$$

in which case differentiating both sides yields

$$\frac{d}{da}F(a) = f(a)$$

Example 60

The density function of X is given by

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

Find $E[e^X]$

Solution. Let $Y = e^X$. Then we have for the culmulative probability function to x,

$$F_Y(x) = P\{Y \leq x\} = P\left\{e^X \leq x\right\} = P\{X \leq \log(x)\} = \int_0^{\log(x)} f(y)dy = \log(x)$$

By 59 we have the probability density function $f_Y(x)$ by differentiating the above

$$f_Y(x) = \frac{1}{x}$$

hence

$$E[e^X] = E[Y] = \int_{-\infty}^{\infty} x f_Y(x)dx = \int_1^e dx = e - 1$$

where the range $[1, e]$ of integration comes from $0 \leq x \leq 1$ being the only range of values of x where $f(x)$ is nonzero. In which case f_Y...to be continued

Proposition 61

If X is a continuous random variable with probability density function $f(x)$ then for any real valued function g

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Proof. We first consider the following lemma

Lemma 62

For a nonnegative random variable Y we have

$$E[Y] = \int_0^{\infty} P\{Y > y\}\, dy$$

Proof. Consider

$$\int_0^{\infty} P\{Y > y\}\, dy = \int_0^{\infty} \int_y^{\infty} f_Y(x)dxdy$$

where we have used the fact that $\int_y^{\infty} f_Y(x)dx = P\{Y > y\}$. Then interchaning the order of integration(by fubini theorem) we have

$$\int_0^{\infty} P\{Y > y\}\, dy = \int_0^{\infty} \left(\int_0^x dy\right) f_Y(x)dx = \int_0^{\infty} x f_Y(x)dx = E[Y]$$

To see how the change in order of integration occured via fubini theorem, specifically $\int_0^{\infty}\int_y^{\infty} dxdy = \int_0^{\infty}\int_0^x dydx$ consider that we are integrating over the region

$$x \geq y \quad \text{and} \quad y \geq 0$$

then it is clear to see the integrals describe both valid ways to integrate over the same region. The LHS integrates over x first then y while the RHS does y first then x.

Remark 63. *Notice this it he "continous" analagou for the discrete version we did previously(the one where we grouped the terms thingy)*

Now going back to prove 61 we have

$$\begin{aligned} E[g(x)] &= \int_0^\infty P\{g(X) > y\}\, dy \\ &= \int_0^\infty \int_{x:g(x)>y} f(x)dxdy \\ &= \int_{x;g(x)>0} \int_0^{g(x)} dy f(x)dx \\ &= \int_{x;g(x)>0} g(x)f(x)dx \end{aligned}$$

as desired □

Example 64

See that 61 supports the result in 60 since

$$E[e^X] = \int_0^1 e^x dx = e - 1$$

Proposition 65

For a continuous random variable X we have

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

and

$$E[aX + b]aE[X] + b$$

Proof. same result and same kind of proof as the discrete version

5.1 uniform

Definition 66

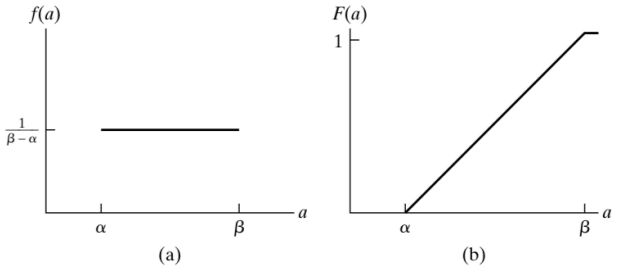
A random variable is said to be **uniformly distributed** over the itnerval $(0, 1)$ if its probability density function is given by

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

This is clearly well defined as $\int_{-\infty}^{\infty} f(x)dx = \int_0^1 dx = 1$ then we also have

$$P\{a \le X \le b\} = \int_a^b f(x) = b - a$$

for $0 < a < b < 1$.



Fact 67

In general we say that X is a uniform random variable on the interval (α, β) if the probability density function is given by

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

Since $F(a) = \int_{-\infty}^{a} f(x)dx$ it follows that

$$F(a) = \begin{cases} 0 & a \leq \alpha \\ \frac{a-\alpha}{\beta-\alpha} & \alpha < a < \beta \\ 1 & a \geq \beta \end{cases}$$

Example 68

Let X be uniformly distributed over (α, β). Find (a) $E[X]$ and (b) $\text{Var}(X)$

Solution. For (a) consider

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{\alpha}^{\beta} \frac{x}{\beta-\alpha} dx \\ &= \frac{\beta^2-\alpha^2}{2(\alpha-\beta)} \\ &= \frac{\beta+\alpha}{2} \end{aligned}$$

For (b) similary frind $E[X^2]$ then find

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

5.2 normal

Definition 69

We say that X is a **normal random variable** or simply that X is normally distributed with paramters μ and σ^2 if the density of X is given by

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

Its graph is what we know as the "bell curve". To show that this is well defined we have to prove that

Proposition 70

consider

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

Proof. First make the substitution $y = (x-\mu)/\sigma$ and see that

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

but we know that $\int_{-\infty}^{\infty} e^{-(y)^2/2}dy = \sqrt{2\pi}$(recall integration of gaussian integral in your `complex analysis notes` via polar coordinates). In which case the proposition is proved □

Proposition 71

If X is normally distributed with parameters μ and σ^2 then $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$

Proof. Consider

$$\begin{aligned} F_Y &= P\{Y \leq x\} \\ &= P\{aX + b \leq x\} \\ &= P\left\{X \leq \frac{x-b}{a}\right\} \\ &= F_X\left(\frac{x-b}{a}\right) \end{aligned}$$

and since the normal random variable is continuous by differentiation w have

$$\begin{aligned} f_Y(x) &= \frac{1}{a}f_X\left(\frac{x-b}{a}\right) \\ &= \frac{1}{\sqrt{2\pi}a\sigma}\exp\left\{-\left(\frac{x-b}{a}-\mu\right)^2/2\sigma^2\right\} \\ &= \frac{1}{\sqrt{2\pi}a\sigma}\exp\left\{-(x-b-a\mu)^2/2(a\sigma)^2\right\} \end{aligned}$$

Theorem 72

The parameters μ and σ^2 of a normal distribution represent the expected value and variance respectively

Proof. Let $Z = (X-\mu)/\sigma$. Then we have

$$\begin{aligned} E[z] &= \int_{-\infty}^{\infty} x f_Z(x)dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x e^{-x^2/2}dx \\ &= -\frac{1}{\sqrt{2\pi}}e^{-x^2/2}|_{-\infty}^{\infty} = 0 \end{aligned}$$

Using 65 we have that

$$\text{Var}(Z) = E[Z^2] - (E(Z))^2 = E[Z^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx$$

by integration by parts we have

$$\text{Var}(z) = \frac{1}{\sqrt{2\pi}}(-xe^{-x^2/2}|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2} dx) = 1$$

Now applying 71 since $X = \mu + \sigma Z$ we have

$$E[X] = \mu + \sigma E[Z] = \mu$$

and

$$\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2$$

□

An important implication of this is that is allows us to

Definition 73

Define $Z = (X - \mu)/\sigma$ to be the **standard/unit normal random variable** in which case has parameters $\mu = 0$ and $\sigma^2 = 1$

This bell curve/normal distribution graph is cleary symmetrical about $\mu = 0$.

Proposition 74

We denote the cumulative distribution function of a standard normal random variable by $\Phi(x)$ that is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

then we have

$$\phi(-x) = 1 - \phi(x)$$

Proof. Consider that integration here is finding the area under the bell curve/normal distribution(which is symmetrical about $x = 0$ as mentioned earlier). Therefore seeing that

$$1 - \Phi(x) = \int_{-\infty}^{\infty} + \int_{x}^{-\infty} = \int_{x}^{\infty} = \int_{-\infty}^{-x} = \Phi(-X)$$

where the 3rd equality follows by symmetry

Corollary 75

Then we have for a normal random variable Z

$$P\{Z \leq -x\} = P\{Z > x\}, \quad -\infty < x < \infty$$

5.3 exponential

Definition 76

A continuous random variable whose probability density function is given for some $\lambda > 0$ by

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

is said to be an **exponential random variable**

Note that this is well defined as consider the culmulative distributoin function $F(a)$ of an exponential variable

$$\begin{aligned} F(a) &= P\{X \leq a\} \\ &= \int_0^a \lambda e^{-\lambda x} dx \\ &= -e^{-\lambda x}|_0^a \\ &= 1 - e^{-\lambda a}, \quad a \geq 0 \end{aligned}$$

and clearly $F(\infty) = 1$

Problem 77

Let X be an exponential random variable with paramter λ. Calculate

(a) $E[X]$

(b) $\text{Var}(X)$

Solution. Notice that

$$E[X^n] = \int_0^\infty x^n \lambda e^{-\lambda x} dx$$

by integration by parts we have

$$\begin{aligned} E[X^n] &= -x^n e^{-\lambda x}|_0^\infty + \int_0^\infty e^{-\lambda x} n x^{n-1} dx \\ &= 0 + \frac{n}{\lambda} \int_0^\infty \lambda e^{-\lambda x} x^{n-1} dx \\ &= \frac{n}{\lambda} E[X^{n-1}] \end{aligned}$$

Now letting $n = 1$ and $n = 2$ gives

$$E[X] = \frac{1}{\lambda}$$

and

$$E[X^2] = \frac{2}{\lambda} E[X] = \frac{2}{\lambda^2}$$

rspectively. Hence

$$\text{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

5.4 gamma

Definition 78

A random variable is said to have a **gamma distribution** with paramter $(\alpha, \lambda), \lambda > 0, \alpha > 0$ if its density function is given by

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

where $\Gamma(a)$ is called the **gamma function** is defined as

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy$$

Consider that integration of $\Gamma(\alpha)$ by parts yields

$$\begin{aligned} \Gamma(\alpha) &= -e^{-y} y^{\alpha-1}|_0^\infty + \int_0^\infty e^{-y}(\alpha-1) y^{\alpha-2} dy \\ &= (\alpha-1) \int_0^\infty e^{-y} y^{\alpha-2} dy \\ &= (\alpha-1)\Gamma(\alpha-1) \end{aligned}$$

Since $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$ we observe that

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

6 Jointly Distributed Random Variables(6)

so far we have concerned ourselves only with probability distributions for single random variables however we are often also interested in those with multiple.

Definition 79

We define for any two random variables X and Y the **joint cumulative probability distribution function** by

$$F(a, b) = P\{X \leq a, Y \leq b\}, \quad -\infty < a, b < \infty$$

where

$$P\{X \leq a, Y \leq b\} = P\{X \leq a \cap Y \leq b\}$$

The distribution of X can be derived from the joint distribution of X and Y as follows

$$\begin{aligned} F_X(a) &= P\{X \leq a\} \\ &= P\{X \leq a, X < \infty\} \\ &= P\left(\lim_{b\to\infty} X \leq a, Y \leq b\right) \\ &= \lim_{b\to\infty} P\{X \leq a, Y \leq b\} \\ &= \lim_{b\to\infty} F(a, b) \\ &= F(a, \infty) \end{aligned}$$

where line (4) made of of the fact that probability function is continuous by assumption as stated in 7. Similarly we have

$$F_Y(b) = P\{Y \leq b\} = \lim_{a\to\infty} F(a,b) = F(\infty, b)$$

Also note that

$$\begin{aligned} P\{X > a, Y > b\} &= 1 - P(\{X > a, Y > b\}^C) \\ &= 1 - P(\{X > a\}^C \cup \{Y > b\}^C) \\ &= 1 - P(\{X \leq a\} \cup \{Y \leq b\}) \\ &= 1 - [P\{X \leq a\} + P\{Y \leq b\} - P\{X \leq a, Y \leq b\}] \\ &= 1 - F_X(a) - F_Y(b) + F(a,b) \end{aligned}$$

where the second line follows by de morgan 5 specifically $(a \cap b)^c = a^c \cup b^c$

Definition 80

We say that X and Y are **jointly continuous** if there exists a function $f(x,y)$ having the property that for every set C of pairs of real numbers we have

$$P\{(X,Y) \in C\} = \int \int_{(x,y)\in C} f(x,y)dxyy$$

if A and B are an sets of real numbers then by defining $C = \{(x,y) : x \in A, y \in B\}$ then

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x,y)dxdy$$

since

$$F(a,b) = P\{X \in (-\infty, a], Y \in (-\infty, b]\} = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)dxdy$$

which upon differentiation gets

$$f(a,b) = \frac{\partial^2}{\partial a \partial b} F(a,b)$$

Also note that

$$\begin{aligned} P\{X \in A\} &= P\{X \in A, Y \in (-\infty, \infty)\} \\ &= \int_A \int_{-\infty}^{\infty} f(x,y)dydx \\ &= \int_A f_X(x)dx \end{aligned}$$

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \tag{4}$$

note that this may by generalized to n variables and that this is simply the case of $n = 2$

Example 81

The join density function of X and Y is given by

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Compute (a) $P\{X > 1, Y < 1\}$,(b)$P\{X < Y\}$ and (c)$P\{X < a\}$

Solution. (a)

$$= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y}dxdy$$

(b)

$$= \int_0^\infty \int_0^y 2e^{-x}e^{-2y}dxdy$$

see that the integral for dx ranges from 0 to y while dy ranges from 0 to ∞ which is all possible non zero values.

(c)

$$= \int_0^a \int_0^\infty 2e^{-x}e^{-2y}dydx$$

Example 82 (multinomial)

Suppose a sequence of n independent and identical experimentes are performed. Suppose that each experiment can result in any one of the r possible outcomes with respective probabilities $p_1, \ldots, p_r$ and $\sum_{i=1}^r p_i = 1$.If we let X_i denote the number among the n experiments that result in outcomes number i then

$$P\{X_1 = n_1, X_2 = n_1, \ldots, X_r = n_r\} = \frac{n!}{n_1!n_2!\ldots n_r!}p_1^{n_1}p_2^{n_2}\ldots p_r^{n_r}$$

where $\sum_{i=1}^r n_i = n$

This is easily seen when you see that $\frac{n!}{n_1!n_2!\ldots n_r!}$ is the number of possible outcomes corresponding to $\{X_1 = n_1, X_2 = n_1, \ldots, X_r = n_r\}$. Now in each outcome, the probability is clearly given by $p_1^{n_1}p_2^{n_2}\ldots p_r^{n_r}$

Proposition 83

The random variables X and Y are **independent** if for any two sets of real numbers A and B we have

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

Proof. By condition of independence when X and Y are independent random variables we have

$$p(x,y) = p_X(x)p_Y(y) \quad \forall x, y$$

then see that

$$\begin{aligned} P\{X \in A, Y \in B\} &= \sum_{y \in B} \sum_{x \in A} p(x,y) \\ &= \sum_{y} \sum_{x} p_X(x) p_Y(y) \\ &= \sum_{y} p_Y(y) \sum_{x} p_X(x) \\ &= P\{Y \in B\} P\{X \in A\} \end{aligned}$$

Example 84

Suppose that the number of people who enter a post office on a given day is a Poisson random variable with parameter λ. Show that if each person who enters the post office is a male with probability p and a female with probability $1-p$ then the number of males and females entering the post office are independent Poisson random variables with respective parameters λp and $\lambda(1-p)$

see book. hint: uses baye's theorem

Example 85 (Buffon's needle problem)

see book

Proposition 86

The continuous/discrete random variables X and Y are independent if and only if their joint probability desnity/mass function can be expressed as

$$f_{X,Y}(x,y) = h(x)g(y), \quad -\infty < x < \infty, -\infty < y < \infty$$

Proof. We will prove it for the continuous case. For the discrete just replace with sums. For the forward direction by definition of indpendence

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

the proposition is trivially true. Now for the backward direction suppose

$$f_{X,Y}(x,y) = h(x)g(x)$$

Since this is a well defined probability desnity function we must have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} h(x) dx \int_{-\infty}^{\infty} g(y) dy \\ &= C_1 C_2 \end{aligned}$$

where we have let $C_1 = \int h(x)dx$ and $C_2 = \int g(y)dy$. But from 4 we have

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = C_2 h(x)$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = C_1 g(y)$$

Since $C_1 C_2 = 1$ it follows that

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

□

Fact 87

to calculate the distribution $X+Y$ when they are independent. Suppose they are continuous then the cumulative distribution function is given by

$$\begin{aligned} F_{X+Y}(a) &= P\{X+Y \le a\} \\ &= \iint_{x+y\le a} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x) dx f_Y(y) dy \\ &= \int_{-\infty}^{\infty} F_X(a-y) f_Y(y) dy \end{aligned}$$

where the form in the last line is known as the **convolution** of the distributive functions F_X and F_Y. Since continuous we may differentiate get the probability density function from the cumulative distribution like so

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da} F_X(a-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy \end{aligned}$$

where the 2nd line is an application of **lebiniz integral rule**(recall `MIT 18.101 Munkres` notes)

We now consider the sums of independent random variables of various types.

6.1 sums of uniform iid

Example 88 (Sum of two independent uniform random variables)

If X and Y are independent random variables both uniformly distributed on $(0,1)$ calculate the probability density of $X+Y$

Solution. From 87 we have that

$$f_X(a) = f_Y(a) = \begin{cases} 1 & 0 < a < 1 \\ 0 & \text{otherwise} \end{cases}$$

so we obtain

$$f_{X+Y}(a) = \int_0^1 f_X(a-y) dy$$

The range of the integration $0 < y < 1$ implies that $a < a - y < a - 1$. But by definition we also know $0 < a - y < 1$ is the only non zero range of values for $f_X(a - y)$. For $0 < a < 1$ this yields

$$f_{X+Y}(a) = \int_0^a dy = a$$

Therefore for $0 \leq a \leq 1$ this yields

$$f_{X+Y}(a) = \int_0^a dy = a$$

the bottom range is not $a - 1$ because the lowest possible bottom endpoint is 0. likewise for $1 < a < 2$ we get

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2 - a$$

where the upper range is not a because the highest possible top endpoint is 1. Hence overall we have triangular shaped probability density function

Proposition 89

Suppose that $X_1, X_2, \ldots, X_n$ are independent uniform (0,1) random variables and let the cumulative $F_n(x)$ denote

$$F_n(x) = P\{X_1 + \ldots + X_n \leq x\}$$

then we have that

$$F_n(x) = x^n/n!, \quad 0 \leq x \leq 1$$

Proof. We will prove this by induction. That is supposing

$$F_{n-1}(x) = x^{n-1}/(n-1)!$$

we aim to use this to show that the F_n case is true. From 87 we have that

$$\begin{aligned} F_n(x) &= \int_0^1 F_{n-1}(x-y) f_{X_n}(y) dy \\ &= \frac{1}{(n-1)!} \int_0^x (x-y)^{n-1} dy \\ &= x^n/n! \end{aligned}$$

6.2 sums of normal iid

Proposition 90

If $X_i, i = 1, \ldots, n$ are independent random variables that are normally distrbuted with respective paramters $\mu_i, \sigma_i^2, i = 1, \ldots, n$ then $\sum_{i=1}^n X_i$ is normally distributed with paramters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$

Proof. Suppose X and Y are independent normal random variables with X having mean 0 and variance σ^2 and Y having mean 0 and variance 1. Firstly consider

Lemma 91

Let X and Y be independent normal random variables with X having mean 0 and variance σ^2 and Y having mean 0 and variance 1, then $X + Y$ is normal with mean 0 and variance $1 + \sigma^2$

Proof. Let $c = \frac{1}{2\sigma^2} + \frac{1}{2} = \frac{1+\sigma^2}{2\sigma^2}$ then collecting y terms for the integrand of the convolution we have

$$\begin{aligned} f_X(a-y)f_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(a-y)^2}{2\sigma^2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{-y^2}{2}\right\} \\ &= \frac{1}{2\pi\sigma} \exp\left\{\frac{-a^2}{2\sigma^2}\right\} \exp\left\{-c\left(y^2 - 2y\frac{a}{1+\sigma^2}\right)\right\} \end{aligned}$$

Then completing the square for y then taking integral to find the convolution we have

$$\begin{aligned} f_{X+Y}(a) &= \frac{1}{2\pi\sigma} \exp\left\{-\frac{a^2}{2\sigma^2}\right\} \exp\left\{\frac{a^2}{2\sigma^2(1+\sigma^2)}\right\} \\ &\quad \times \int_{-\infty}^{\infty} \exp\left\{-c\left(y - \frac{a}{1+\sigma^2}\right)^2\right\} dy \\ &= \frac{1}{2\pi\sigma} \exp\left\{-\frac{a^2}{2(1+\sigma^2)}\right\} \int_{-\infty}^{\infty} \exp\left\{-cx^2\right\} dx \\ &= C \exp\left\{-\frac{a^2}{2(1+\sigma^2)}\right\} \end{aligned}$$

where C is some constant independent of a. On comparison this implies the lemma as desired and the 2nd line applied a change of variable $x = \left(y - \frac{a}{1+\sigma^2}\right)$. □

Returning back to our proposition we have by expressing $X_1 + X_2$ in terms of their means and variances,

$$X_1 + X_2 = \sigma_2\left(\frac{X_1 - \mu_1}{\sigma_2} + \frac{X_2 - \mu_2}{\sigma_2}\right) + \mu_1 + \mu_2$$

Notice that the 1st term in the brackets has mean 0 and variance $\frac{\sigma_1^2}{\sigma_2^2}$ while the 2nd term has mean 0 and variance 1. To see this consider that the 2nd term is the standard normal variable. As for the 1st term consider

$$\text{Var}\left[\left(\frac{\sigma_1}{\sigma_2}\right)\left(\frac{X_1 - \mu}{\sigma_1}\right)\right] = \left(\frac{\sigma_1}{\sigma_2}\right)^2 \text{Var}\left[\left(\frac{X_1 - \mu}{\sigma_1}\right)\right] = \left(\frac{\sigma_1}{\sigma_2}\right)^2$$

Then by 31 and 34 the proposition follows for the case $n = 2$ which serves as our base case for our inductive proof. Now by induction suppose

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n$$

and by induction hypothesis and the proposition follows.

6.3 sums Poisson and binomial iid

Example 92 (Sums of independent Poisson random variables)
If X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2 compute the distribution $X + Y$

Solution. Consider

$$\begin{aligned} P\{X+Y=n\} &= \sum_{k=0}^{n} P\{X=k, Y=n-k\} \\ &= \sum_{k=0}^{n} P\{X=k\} P\{Y=n-k\} \\ &= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!} \lambda_1^k \lambda_2^{n-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1+\lambda_2)^n \end{aligned}$$

where line 2 is a direct application of 83

Example 93 (Sums of independent binomial random variables)
Let X and Y be independent binomial random variables with respective parameters (n, p) and (m, p). Calculate the distribution of $X+Y$. Show that $X+Y$ is binomial with parameters $(n+m, p)$

Solution. Intuitively this already makes sense but to show this rigorously consider

$$\begin{aligned} P\{X+Y=n\} &= \sum_{k=0}^{n} P\{X=k, Y=n-k\} \\ &= \sum_{k=0}^{n} P\{X=k\} P\{Y=n-k\} \\ &= \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \binom{m}{k-i} p^{k-i} q^{m-k+i} \end{aligned}$$

where $q = 1-p$ and where $\binom{r}{j} = 0$ when $j < 0$. Thus

$$P\{X+Y=k\} = p^k q^{n+m-k} \sum_{i=0}^{n} \binom{n}{i} \binom{m}{k-i}$$

and the conclusion follows upon application of the identity

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

recall this from `MIT 18.211 Combinatorial Analysis`

6.4 sums geometric iid

Proposition 94 (p_i same version)
Let $X_1 \ldots, X_n$ be independent geometric random variables with X_i having the parameter p_i for $i = 1, \ldots, n$. If all the p_i are the *same* then for $k \geq n$

Proof. Denote $p = p_i, \forall i$. Then consider for example n coins each with probability p of flipping to a head. Then flip

coin 1 until it shows a head. The number of flips required is X_1. Do so for the rest of the other coins. Suppose $\sum_i X_i = k$. Notice that in all possible cases the only condition is that the kth flip corresponding to that of last/nth coin to be flipped must be a head(again due to the word "until"). The rest of the n-1 heads can be any of the previous k-1 flips. This precisely describes the negative binomial random variable which you recall by 53 we thus have

$$P\{S_n = k\} = \binom{k-1}{n-1} p^n (1-p)^{k-n}, k \geq n$$

Proposition 95 (p_i distinct version)

Let $X_1 \ldots, X_n$ be independent geometric random variables with X_i having the parameter p_i for $i = 1, \ldots, n$. If all the p_i are *distinct* then for $k \geq n$

$$P\{S_n = k\} = \sum_{i=1}^{n} p_i q_i^{k-1} \prod_{j \neq i} \frac{p_j}{p_j - p_i}$$

Proof. Do induction on $n + k$. See book for more. Also clarify the proof on two variables used in the book with someone. Ask if it is correct? Seems a little strange to me for some reason

6.5 conditional distributions

Recall from the definition 16 of conditional probability that

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

We do so analogously for probability mass functions.

Definition 96 (Discrete version)

We define the **conditional probability mass function** of X given $Y = y$ by

$$\begin{aligned} p_{X|Y}(x|y) &= P\{X = x|Y = y\} \\ &= \frac{P\{X = x, Y = y\}}{P\{Y = y\}} \\ &= \frac{p(x,y)}{p_Y(y)} \end{aligned}$$

Notice that

$$F_{X|Y}(x|y) = P\{X \leq x|Y = y\} = \sum_{a \leq x} p_{X|Y}(a|y)$$

Analogous to 16 once again which shows that if E is independent $P(E|F) = P(E)$, if X is independent we have

$$\begin{aligned} p_{X|Y}(x|y) &= P\{X = x|Y = y\} \\ &= \frac{P\{X = x, Y = y\}}{P\{Y = y\}} \\ &= \frac{P\{X = x\} P\{Y = y\}}{P\{Y = y\}} \\ &= P\{X = x\} \end{aligned}$$

where line 3 is a direct application of 83. Now let us extend this to the continuous case

Definition 97 (Continuous version)

If X and Y are jointly continuous then for any set A we have

$$P\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)dx$$

and so the conditional *culmulative* distributive function is related by

$$F_{X|Y}(a|y) = P\{X \leq a|Y = y\} = \int_{-\infty}^{a} f_{X|Y}(x|y)dx$$

7 properties of expectation(7)

Proposition 98

if $P\{a \leq x \leq b\} = 1$ then

$$a \leq E[X] \leq b$$

Proof. Easy. Recall that $E[X]$ is just a weighted average of the possible values of X.

Proposition 99 (Expectation of joint probability)

If X and Y have a joint probability mass function $p(x, y)$ then

$$E[g(X, Y)] = \sum_y \sum_x g(x, y)p(x, y)$$

Likewise for the continuous case for the joint probability *density* function $f(x, y)$ we have

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f(x, y)dxdy$$

Proof. By 62(see that we have replaced Y with $g(x, y)$) we have that

$$E[g(x, y)] = \int_0^{\infty} P\{g(X, Y) > t\}\, dt$$

but we know that

$$P\{g(X, y) > t\} = \iint_{(x,y):g(x,y)>t} f(x, y)dydx$$

then combining the two we have

$$E[g(X, Y)] = \int_0^{\infty} \iint_{(x,y):g(x,y)>t} f(x, y)dydxdt$$

exchanging the order of integration via fubini theorem similar to what we did in 62 we have

$$E[g(X, Y)] = \int_x \int_y \int_{t=0}^{g(x,y)} f(x, y)dtdydx = \int_x \int_y g(x, y)f(x, y)dydx$$

An important application of the proposition above is to

Proposition 100

Prove that in general(continuous/discrete) we have

$$E[X+Y] = E[X] + E[Y]$$

Proof. Let $g(X,Y) = X+Y$ then in the continuous case we have

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dydx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dxdy \\ &= \int_{-\infty}^{\infty}xf_X(x) + \int_{-\infty}^{\infty}yf_Y(y)dy \\ &= E[X] + E[Y] \end{aligned}$$

□

Corollary 101

It follows from above that $E[X] \geq E[Y]$ if $X \geq Y$

Definition 102 (Sample Mean)

Let $X_1, \ldots, X_n$ be iid random variables having a distribution function F and expected value μ. Such a sequence of random variables is said to constitute a **sample** from the distribution F. The quantity

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

is called the **sample mean**

Also note that

$$\begin{aligned} E[\overline{X}] &= E\left[\sum_{i=1}^{n}\frac{X_i}{n}\right] \\ &= \frac{1}{n}E\left[\sum_{i=1}^{n}X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}E[X_i] \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned}$$

Theorem 103 (Boole's Inequality)

Let $A_1, \ldots, A_n$ denote events(may or may not be independent) then

$$E\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Proof. define the indicator variables X_i, $i = 1, \ldots, n$ by

$$X_i = \begin{cases} 1 & \text{if } A_i \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Let

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

and let

$$Y = \begin{cases} 1 & X \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

that is Y is equal 1 if at least one of the A_i occurs and is 0 otherwise. Then we have

$$\begin{aligned} X &\geq Y \\ E[X] &\geq E[Y] \\ E[X] &= \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} 1 \times P(A_i) + 0 \times P(A_i^c) = \sum_{i=1}^{n} P(A_i) \\ E[Y] &= P\left\{\text{at least one of the } A_i \text{ occur}\right\} = P\left(\bigcup_{i=1}^{n} A_i\right) \end{aligned}$$

Remark 104. *Notice we have shown that there is a one to one correspondence between expectations and probabilities for indicator functions*

then Boole's inequality follows

7.1 moment of expectation

...to be continued

7.2 Covariance,Variance of sums and correlations

Proposition 105

If X and Y are independent then for any functions h and g

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Proof. Suppose that X and Y are jointly continuous with joint density $f(x,y)$ then

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx \\ &= E[h(Y)]E[g(X)] \end{aligned}$$

the proof the discrete case is similar □

Definition 106

the **covariance** between X and Y denoted by $\text{Cov}(X,Y)$ is defined by

$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

upon expansion we get

$$\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$$

which is clearly 0 when X and Y are independent by 105

Proposition 107

Consider

(i) $\text{Cov}(X,Y) = \text{Cov}(Y,X)$

(ii) $\text{Cov}(X,X) = \text{Var}(X)$

(iii) $\text{Cov}(aX,Y) = a\,\text{Cov}(Y,X)$

(iv) $\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i,Y_i)$

Proof. (i) and (ii) follow immediately from definition. For (iii) consider just like how you expanded out the expression for convariance above, now replace X with aX and see that $\text{Cov}(aX,Y) = E[aXY] - E[aX]E[Y] = a\,\text{Cov}(X,Y)$. Needless this applies to if aY as well showing the covariance is a **billinear** function.As for (iv), it immediately follows by billinearity of the covariance function.

Corollary 108

Prove

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

Proof. From (ii) and (iv) in 107 we have

$$\begin{aligned}\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \text{Var}(X_i) + \sum\sum_{i\neq j} \text{Cov}(X_i, X_j)\end{aligned}$$

where the last line essentially separates the summands into cases where $i = j$ and $i \neq j$. Then upon rearrangement the corollary is obtained

Remark 109. *Notice that* $\sum\sum_{i\neq j} \text{Cov}(X_i, X_j) = 2\sum\sum_{i<j} \text{Cov}(X_i, X_j) = 0$ *when* $X_i, \forall i$ *independent so we get*

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

Let us consider some example applications of this

Example 110 (sample variance)
Let $X_1, \ldots, X_n$ be idd random variables having expected value μ and variance σ^2. Then let $\overline{X}$ be the *sample mean*(recall 102). The quantities $X_i - \overline{X}, i = 1, \ldots, n$ are called **deviations**. Then the random variable

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

is called the **sample variance**. Find (a) $\text{Var}(\overline{X})$ and (b) $E[S^2]$

Solution. (a)

$$\begin{aligned}\text{Var}(\overline{X}) &= \left(\frac{1}{n}\right)^2 \text{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \text{Var}(X_i) \quad \text{by independence recall 109} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

As for (b) we start with the following algebraic identity

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n} (X_i - \mu + \mu + \overline{X})^2 \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 + \sum_{i=1}^{n} (\overline{X} - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n} (X_i - \mu) \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\overline{X} - \mu)^2 - 2(\overline{X} - \mu)n(X_i - \mu) \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2\end{aligned}$$

Taking expectations we then have

$$\begin{aligned}(n-1)E[S^2] &= \sum_{i=1}^{n}[E(X_i-\mu)^2] - nE[(\overline{X}-\mu)^2] \\ &= n\sigma^2 - n\,\mathrm{Var}(\overline{X}) \\ &= (n-1)\sigma^2\end{aligned}$$

so we conclude that

$$E[S^2] = \sigma^2$$

Example 111 (Sampling from a finite population)
Consider a set of N people and a real a number v which represents the input of each person in a given poll. We let v_i be the input of person i for $i = 1, \ldots, N$. Suppose that all the v_i are unknown and to gather information, a group of n of the N people are "randomly chosen" in the sense that $\binom{N}{n}$ subsets are equally likely to be chosen. These n people are then surveyed. If S denotes the sum of the n sampled values, determine its mean and variance

Solution. For each person $i, i = 1, \ldots, N$ define an indicator variable I_i to indicate whether or not that person is included in the sample. That is

$$I_i = \begin{cases} 1 & \text{person i is in the random sample} \\ 0 & \text{otherwise} \end{cases}$$

Now S can be expressed by

$$S = \sum_{i=1}^{n} v_i I_i$$

so taking expectations we have

$$E[S] = \sum_{i=1}^{N} v_i E[I_i]$$

And

$$\begin{aligned}\mathrm{Var}(S) &= \sum_{i=1}^{N}\mathrm{Var}(v_i I_i) + 2\sum\sum_{i<j}\mathrm{Cov}(v_i I_i, v_j I_j) \\ &= \sum_{i=1}^{N} v_i^2\,\mathrm{Var}(I_i) + 2\sum\sum_{i<j} v_i v_j\,\mathrm{Cov}(I_i, I_j)\end{aligned}$$

Because

$$\begin{aligned}E[I_i] &= \frac{n}{N} \\ E[I_i I_j] &= \frac{n}{N}\frac{n-1}{N-1}\end{aligned}$$

where the 1st line follows recall 104 that there is a 1 to 1 correspondence between probabilities and expectations of indicator functions and the probability of being in the sample is clearly n/N. it follows that

$$\begin{aligned}\mathrm{Var}(I_i) &= E[I_i^2] - (E[I_i])^2 = \frac{n}{N} - \left(\frac{n}{N}\right)^2 = \left(1 - \frac{n}{N}\right) \\ \mathrm{Cov}(I_i, I_j) &= E[I_i I_j] - E[I_i]E[I_j] = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2\end{aligned}$$

the second line follows from 106.

Remark 112. *since I_i is an indicator variable so $E[I_i^2] = \sum_{I_i \neq 0} 1^2 \frac{1}{N} = E[I_i]$*

Then plugging in the expressions for $\text{Var}(I_i)$ and $\text{Cov}(I_i, I_j)$ into $\text{Var}(S)$ we have

$$\text{Var}(S) = \frac{n}{N}\left(\frac{N-n}{N}\right)\sum_{i=1}^{N} v_i^2 - \frac{2n(N-n)}{N^2(N-1)}\sum\sum_{i<j} v_i v_j$$

which upon simplication gets

$$\boxed{\text{Var}(S) = \frac{n(N-n)}{N-1}\left(\frac{\sum_{i=1}^{N} v_i^2}{N}\right) - \overline{v}^2}$$

also for $E[S]$ recalling the definition of sample mean $\overline{v}$ we have

$$\boxed{E[S] = \sum_{i=1}^{N} v_i \left(\frac{n}{N}\right) = n\overline{v}}$$

Definition 113 (Correlation)

The **correlation** of two random variables X and Y denoted by $p(X, Y)$ is defined as long as $\text{Var}(X)\,\text{Var}(Y)$ is positive by

$$p(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Fact 114

The correlation coefficient is a measure of the degree of linearity between X and Y. In particular a value of $p(X, Y)$ near $+1$ and -1 indicates a high degree of linearity between X and Y

- a positive $p(X, Y)$ indicates that Y tends to increase when X increases
- a negative $p(X, Y)$ indicates that Y tends to decrease when X increases

while a value near 0 indicates that such linearity is absent

- when $p(X, Y) = 0$ we say X and Y are **uncorrelated**

We will now make sense of this. First consider

Proposition 115

show that

$$-1 \leq p(X, Y) \leq 1$$

Proof. Suppose that X and Y have variances given by σ_x^2 and σ_y^2 respectively then on one hand since variance ≥ 0

$$\begin{aligned} 0 &\leq \text{Var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}\right) \\ &= \frac{\text{Var}(Y)}{\sigma_y^2} + \frac{\text{Var}(X)}{\sigma_x^2} + \frac{2\,\text{Cov}(X, Y)}{\sigma_x \sigma_y} \quad \text{recall 108} \\ &= 2[1 + p(X, Y)] \end{aligned}$$

where the last line follows as $\sigma_x = \sqrt{\text{Var}(X)}$ and $\sigma_y = \sqrt{\text{Var}(Y)}$. This implies that

$$-1 \leq p(X, y)$$

on the other hand

$$\begin{aligned} 0 &\leq \text{Var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}\right) \\ &= \frac{\text{Var}(X)}{\sigma_X^2} + \frac{\text{Var}(Y)}{(-\sigma_y)^2} - \frac{2\,\text{Cov}(X,Y)}{\sigma_x \sigma_y} \quad \text{recall 108,denominator and minus sign follows by bilinearity of Cov} \\ &= 2[1 - p(X,Y)] \end{aligned}$$

this implies

$$p(X, Y) \leq 1$$

□

We now show that the sample mean and a deviation from the sample mean are uncorrelated

Proposition 116

Let $X_1, \ldots, X_n$ be iid random variables having variance σ^2 show that

$$\text{Cov}(X_i - \overline{X}, \overline{X}) = 0$$

Proof. Consider

$$\begin{aligned} \text{Cov}(X_i - \overline{X}, \overline{X}) &= \text{Cov}(X_i, \overline{X}) - \text{Cov}(\overline{X}, \overline{X}) \\ &= \text{Cov}\left(X_i, \frac{1}{n}\sum_{j=1}^{n} X_j\right) - \text{Var}(\overline{X}) \\ &= \frac{1}{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) - \frac{\sigma^2}{n} \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0 \end{aligned}$$

where the 1st line follows by bi-linearity as usual and the last line follows using

$$\text{Cov}(X_i, X_j) = \begin{cases} 0 & j \neq i \text{ by independence} \\ \sigma^2 & j = 1 \text{ since } \text{Var}(X_i) = \sigma^2 \end{cases}$$

7.3 conditional expectation

Recall

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

where $p_{X|Y}(x|y) = p_X(x)$ if X and Y are independent

Definition 117 (conditional expectation)

Analagously we define the **conditional expectation** of X given $Y = y$ for all values of y such that $p_Y(y) > 0$ by

$$\begin{aligned} E[X|Y = y] &= \sum_x xP\{X = x|Y = y\} \\ &= \sum_x xp_{X|Y}(x|y) \end{aligned}$$

we can do this similarly for Y too(swap the positions of X, x and Y, y above). Note that if X and Y are indepenent it is clear from above that $E[X|Y = y] = E[X]$

Example 118

If X and Y are independent binomia random variables with indentical parmters n and p calculate the conditional expected value of X given $X + Y = m$

Solution. Consider

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

So we have (see an example in the book for hypergeometric distribution expectation derivation)

$$E[X|X + Y = m] = \frac{m}{2}$$

Example 119

Suppose that the joint density of X and Y is given by

$$f(x, y) = \frac{e^{-x/y}e^{-y}}{y}, \quad 0 < x < \infty, 0 < y < \infty$$

Compute $E[X|Y = y]$

Solution. First find the conditional desnity

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \\ &= \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)dx} \\ &= \frac{1}{y}e^{-x/y} \end{aligned}$$

Then

$$E[X|Y=y] = \int_0^\infty \frac{x}{y} e^{-x/y} dx = y$$

Proposition 120

We denote $E[X|Y]$ as a function of the random variable Y whose value at $Y=y$ is $E[X|Y=y]$

$$E[X] = E[E[X|Y]]$$

Proof. Consider that the RHS implies for the discrete case

$$E[X] = \sum_y E[X|Y=y] P\{Y=y\}$$

and for the continuous case

$$E[X] = \sum_{-\infty}^{\infty} E[X|Y=y] f_Y(y) dy$$

then for the discrete case we have

$$\begin{aligned}
\sum_y E[X|Y=y] P\{Y=y\} &= \sum_y \sum_x x P\{X=x|Y=y\} P\{Y=y\} \\
&= \sum_y \sum_x x \frac{P\{X=x, Y=y\}}{P\{Y=y\}} P\{Y=y\} \\
&= \sum_y \sum_x x P\{X=x, Y=y\} \\
&= \sum_x x \sum_y P\{X=x, Y=y\} \\
&= \sum_x x P\{X=x\} \\
&= E[X]
\end{aligned}$$

similar for the continuous case □

Example 121

A miner is trapped in a mine containing 3 doors. The first leads to a tunnel to safety that takes 3 hours of travel. The 2nd and 3rd takes 5 and 7 hours of travel respectively but leads the miner back into the mine. What is the expected length of time until he reaches to safety

Solution. Consider that

$$\begin{aligned}
E[X|Y=1] &= 3 \\
E[X|Y=2] &= 5 + E[X] \\
E[X|Y=3] &= 7 + E[X]
\end{aligned}$$

Notice the last 2 equations follows because once you are returned back to the mine after travelling, situation restarts again. Oof real bad luck...

Then use

$$E[X] = \sum_i E[X|Y = i]P\{Y = i\}$$

to find that noting that $P\{Y = i\} = \frac{1}{3}$ for all i.

$$E[X] = 15$$

Example 122

The game of craps is began by rolling an ordinary pair of dice. If the sum of the dice is 2,3 or 12 the player loses. If it is 7 or 11 the player wins. If it is any number i, they player continues to roll the dice until the sum if either 7 or i. If it is 7 the player loses; if it is i the player wins. Let R denote the number of rolls of the dice in a game of craps. Find

(a) $E[R]$

(b) $E[R|\text{player wins}]$

(c) $E[R|\text{player loses}]$

Solution. If we let P_i denote the probability that the sum of the dice is i. Then

$$P_i = \frac{i-1}{36}$$

since we have a twin dice rolls will look like so $(1, i-1), (2, i-2), \ldots, (i-1, i)$(we cant have $(i, 0)$ as the dice has no face with 0). $36 = 6 \times 6$ corresponds to the entire sample space of twin dice rolls Also we have

$$E[R] = \sum_{i=2}^{12} E[R|S = i]P_i$$

However

$$E[R|S = i] = \begin{cases} 1 & i = 2, 3, 7, 11, 12 \\ 1 + \frac{1}{P_i + P_y} & \text{otherwise} \end{cases}$$

where the 2nd condition occurs because if the sum is a value i that does not end the game, then the dice will continue to be rolled until the sum is either i or 7 in which case it is a geometric random variable with paramter $P_i + P_7$ so its expectation follows like so. recall 52. So upon calculation we get for (a)

$$E[R] = 1 + \sum_{i=4}^{6} \frac{P_i}{P_i + P_7} + \sum_{i=8}^{10} \frac{P_i}{P_i + P_7} = 3.376$$

For (b) let p be the probability that the player wins.

$$p = \sum_{i=2}^{12} P\{\text{win}|S = i\} P_i = P_7 + P_{11} + \sum_{i=4}^{6} \frac{P_i}{P_i + P_7} + \sum_{i=8}^{10} \frac{P_i}{P_i + P_7} = 0.493$$

Now let $Q_i = P\{S = i|\text{win}\}$ then we have

$$Q_2 = Q_3 = Q_{12} = 0, \quad Q_7 = P_7/p, Q_{11} = P_{11}/p$$

and for $i = 4, 5, 6, 8, 9, 10$ we havee

$$\begin{aligned} Q_i &= \frac{P\{S = i, \text{win}\}}{P\{\text{win}\}} \\ &= \frac{P_i P\{\text{win}|S = i\}}{p} \\ &= \frac{P_i^2}{p(P_i + P_y)} \end{aligned}$$

And so

$$E[R|\text{win}] = \sum_i E[R|\text{win}, S = i]Q_i = 2.938$$

For (c) consider that

$$E[R] = E[R|\text{win}]p + E[R|\text{lose}](1 - p)$$

and so

$$E[R|\text{lose}] = \frac{E[R] - E[R|\text{win}]p}{1 - p} = 3.801$$

□

Not only can we obtain expectations by first conditioning on an appropriate random variable but we may also use this approach to compute probabilities. To see this let E denote an arbitruary event and define the indicator random variable by

$$X = \begin{cases} 1 & \text{E occurs} \\ 0 & \text{E does not occur} \end{cases}$$

It follows from the definition of X that

$$E[X] = P(E)$$
$$E[X|Y = y] = P(E|Y = y)$$

therefore we have

$$P(E) = \begin{cases} \sum_y P(E|Y = y)P(Y = y) & \text{Y is discrete} \\ \int_{-\infty}^{\infty} P(E|Y = y)f_Y(y)dy & \text{Y is continuous} \end{cases}$$

Example 123 (Best Prize Problem)

Suppose that we are to be presented with n distinct prizes in sequence. After being presented with a prize we must immediately decide whether to accept it or to reject it and consider the next best prize. The only information we are given when deciding whether to accept the prize is the relative rank of that prize compared to ones already seen. That is for instance when the fifth prize is presented we learn how it compares with the four prizes already seen. Suppose that once a prize is rejected it is lost and that our objective is to maximize the probability of obtaining the best prize. Assuming that all $n!$ of the prizes are equally likely how well can we do?

Definition 124

Just as have defined the conditional expectation of X for a given Y we can also define the conditional variance of X given that $Y = y$

$$\text{Var}(X|Y) = E[(X - E[X|Y])^2|Y]$$

Too see how this obtained consider that on expansion we have

$$\text{Var}(X|Y) = E[X^2|Y] - (E[X|Y])^2$$

With the same reasoning as 117 if X, Y are independent we have $\text{Var}(X|Y = y) = \text{Var}(X)$

Proposition 125

The conditional variance formula is given by

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y])$$

Proof. Firstly taking expectations of the above equation we have

$$E[\text{Var}(X|Y)] = E[E[X^2|Y]] - E[(E[X|Y])^2] = E[X^2] - E[(E[X|Y])^2]$$

but we also know

$$\text{Var}(E[X|Y]) = E[(E[X|Y])^2] - (E[E[X|Y]])^2 = E[(E[X|Y])^2] - (E[X])^2$$

so on combining the 2 equations yields the proposition

Example 126

Suppose that by any time t the number of people that have arrived at a train depot is a Poisson random variable with mean λt. If the initial train arrives at the depot at a time(independent of when the passengers arrive) that is uniformly distributed over $(0, T)$ what are the mean and variance of the number of passengers who enter the train?

Fact 127

Sometimes we may want to think of the expectation $E[Y]$ as a "best guess/predictor" of the value of Y. Best in the sense that among all the constants m the expectation $E[(Y - m)^2]$ is minimized when $m = E[Y]$

More precisely

Proposition 128

Let $g(x)$ be a function. Then $E[(Y - g(X))^2]$ is minimized when $g(X) = E[Y|X]$ that is

$$E[(Y - g(X))^2] \geq E[(Y - E[Y|X])^2]$$

Proof. Consider

$$\begin{aligned} E[(Y - g(X))^2|X] &= E[(Y - E[Y|X] + E[Y|X] - g(X))^2|X] \\ &= E[(Y - E[Y|X])^2|X] \\ &+ E[(E[Y|X] - g(X))^2|X] \\ &+ 2E[(Y - E[Y|X])(E[Y|X] - g(X))|X] \end{aligned}$$

However for a given X, $E[Y|X] - g(X)$ being a function of X can be trated as a constant. Thus we have

$$\begin{aligned} E[(Y - E[Y|X])(E[Y|X] - g(X))|X] &= (E[Y|X] - g(X))E[Y - E[Y|X]|X] \\ &= (E[Y|X] - g(X))(E[Y|X] - E[Y|X]) \\ &= 0 \end{aligned}$$

and so the proposition follows given expectations ≥ 0.

7.4 Moment Generating Funcions

Definition 129

The moment generating function $M(t)$ of the random variable X is defined for all real values of t by

$$M(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} p(x) & \text{X is discrete with mass function } p(x) \\ \int_{-\infty}^{\infty} e^{tx} p(x) & \text{X is continuous with density } f(x) \end{cases}$$

We call $M(t)$ a moment generating function because all of the moments of X can be obtained successively differentiating $M(t)$ then evaluating the result at $t = 0$. That is we have assumed that differentiation like so is valid

$$M'(t) = \frac{d}{dt} E[e^{tX}] = E[\frac{d}{dt}(e^{tX})] = E[Xe^{tX}]$$

Remark 130. *in general the interchange of limits in and out of the expectation function requires it to be Lebesgue integrable. But for this course we will just assume so. We will look at this more in detail in theory of probability* `MIT 18.175`

Similary we have assumed for the discrete case

$$\frac{d}{dt}\left[\sum_x e^{tx} p(x)\right] = \sum_x \frac{d}{dt}[e^{tx} p(x)]$$

and for the continuous case(assuming able to apply lebiniz integral rule)

$$\frac{d}{dt}\left[\int e^{tx} p(x)\right] = \int \frac{d}{dt}[e^{tx} p(x)]$$

Example 131

Notice that

$$M'(0) = E[X] \quad \text{and} \quad M''(0) = E[\frac{d}{dt}(Xe^{tX})_{t=0}] = E[X^2]$$

Corollary 132

In particular it is easy to see by observation

$$M^n(t) = E[X^n e^{tX}] \quad n \geq 1$$

so

$$M^n(0) = E[X^n] \quad n \geq 1$$

Example 133

If X is a binomial random variable with paramters n and p hen

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\ &= (pe^t + 1 - p)^n \end{aligned}$$

where the last equality followed by binomial theorem

Example 134

If X is a poisson random variable with paramter λ then

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \sum_{n=0}^{\infty} \frac{e^{tn} e^{0\lambda} \lambda^n}{n!} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= \exp\left\{\lambda(e^t - 1)\right\} \end{aligned}$$

Example 135

If X is exponential random variable with paramter λ

$$\begin{aligned} M(t) &= E[e^t X] \\ &= \int_0^{\infty} e^{tx} e^{-\lambda x} dx \\ &= \frac{\lambda}{\lambda - 1} \end{aligned}$$

for $t < \lambda$

Example 136

Let Z be unit normal random variable with paramters $0, 1$. Then

$$\begin{aligned} M_Z(t) &= E[e^{tZ}] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{t^2/2} \end{aligned}$$

Example 137

Let X be an arbitrary normal variable then we have

$$\begin{aligned} E_X(t) &= E[e^{tX}] \\ &= E[e^{t(\mu+\sigma Z)}] \\ &= \exp\left\{\frac{\sigma^2 t^2}{2} + \mu t\right\} \end{aligned}$$

to see this you may verify that $E[\mu + \sigma Z] = \mu$ and $\text{Var}(\mu + \sigma Z) = \sigma^2$ given that Z is a unit normal variable

7.5 Additional properties of normal random variables

Definition 138

Let $Z_1, \ldots, Z_n$ be a set of independent *unit normal random variables*. If for some constants $a_{ij}, 1 \leq i \leq m, 1 \leq j \leq n$ and $\mu_i, 1 \leq i \leq m$ we have

$$\begin{aligned} X_1 &= a_{11}Z_1 + \ldots + a_{1n}Z_n + \mu_1 \\ X_2 &= a_{21}Z_1 + \ldots + a_{2n}Z_n + \mu_2 \\ &\vdots \\ X_i &= a_{i1}Z_1 + \ldots + a_{in}Z_n + \mu_i \\ &\vdots \\ X_m &= a_{m1}Z_1 + \ldots + a_{mn}Z_n + \mu_m \end{aligned}$$

Remark 139. *Notice that you can write this in matrix form to make it more compact*

Then we see that

$$\begin{aligned} E[X_i] &= \mu_i \\ \text{Var}(X_i) &= \sum_{j=1}^{n} a_{ij}^2 \end{aligned}$$

Theorem 140

If $X_1, \ldots, X_d$ are iid normal random variables with mean μ and variance σ^2 then the sample mean $\overline{X}$ and sample variance S^2 are independent. $\overline{X}$ is a normal random variable with mean μ and variance σ^2/n and $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $n-1$ degrees of freedom

8 limit theorems

8.1 chebyshev inequality and the weak law of large numbers

Theorem 141 (Markov Inequality)

If X is a random variable that takes only nonnegative values then for any $a > 0$

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Proof. For $a > 0$ let

$$I = \begin{cases} 1 & X \geq a \\ 0 & \text{otherwise} \end{cases}$$

therefore since $X \geq 0$ we have that

$$I \leq \frac{X}{a}$$

To see this $X \geq 0$ in general except $X \geq a > 0$ as indicated by $I = 0, 1$ respectively We can write these conditions together compactly as $X \geq Ia$. Now taking expectations of our inequality we have

$$E[I] \leq \frac{E[X]}{a}$$

but $E[I] = P\{X \geq a\}$(recall 104). Therefore the proposition follows

Theorem 142 (Chebyshev inequality)

If X is a random variable with finite mean μ and variance σ^2 then for any $k > 0$

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

Proof. Since $(X-\mu)^2$ is nonnegative random variable on application of *Markov's inequality* with $a = k^2$ we obtain

$$P\left\{(X-\mu)^2 \geq k^2\right\} \leq \frac{E[(X-\mu)^2]}{k^2}$$

but since $(X-\mu)^2 \geq k^2$ if and only if $|X-\mu| \geq k$ the above is equivalent to

$$P\{|X-\mu| \geq k\} \leq \frac{E[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

since $E[(X-\mu)] = E[X] - \mu = 0$(recall properties of expectation:translation by constants). Then recall that $\text{Var}(X - \mu) = E[(X-\mu)^2] - (E[(X-\mu)])^2 = \sigma^2$(variance invariant by translation by constants if you recall)

Theorem 143 (Weak Law of Large Numbers)

Let $X_1, X_2, \ldots$ be a sequence of iid random variables having finite mean $E[X_i] = \mu$. Then fr any $\varepsilon > 0$

$$P\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right\} \to 0$$

as $n \to \infty$

Proof. We prove this under the additional assumption tha the random variables have finite variance σ^2. Now since

$$E\left[\frac{X_1 + \ldots + X_n}{n}\right] = \mu \text{ and } \text{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

To see this simply recall sample mean and variance from earlier(110). Now by *chebyshev inequality* it is clear that

$$0 \leq P\left\{\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{n\varepsilon^2}$$

where the left inequality i just basically the axiom of probability.Therefore by taking the limit of $n \to \infty$ we immediately have our desired result by squeeze theorem □

8.2 central limit theorem

Theorem 144 (Central Limit Theorem)

Let $X_1, X_2, \ldots$ be a sequence of iid random variables each with mean μ and variance σ^2 then the distribution of

$$\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is for $-\infty < a < \infty$,

$$P\left\{\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

as $n \to \infty$

Proof. Do `18.175 Theory of Probability` for a proper treatment instead of just assuming random lemmas with no basis as in the book

8.3 strong law of large numbers

Theorem 145 (Strong Law of Large numbers)

Let $X_1, X_2, \ldots$ be a sequence of iid random normal variables each having a finite mean $\mu = E[X_i]$ then with probability 1

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu$$

as $n \to \infty$

Proof. Do `18.175 Theory of Probability` for a proper treatment instead of just assuming random lemmas with no basis as in the book

9 additional topics

9.1 other inequalites

Theorem 146 (Chernoff Bounds)

Prove

$$P\{X \geq a\} \leq e^{-ta}M(t) \quad \forall t > 0$$
$$P\{X \leq a\} \leq e^{-ta}M(t) \quad \forall t < 0$$

Proof. Recall that a moment generating function is defined by $M(t) = E[e^{tX}]$. Noting that the taking the exp of both sides of an inequality does not affect it we have

$$P\{X \geq a\} = P\left\{e^{tX} \geq e^{ta}\right\} \leq E[e^{tX}]e^{-ta}$$

where the final inequality follows by *markov's inequality*. This is applicable because e^{tX} is always nonnegative(recall exponential function graph). Do so similary for

$$P\{X \leq a\} = P\left\{e^{tX} \geq e^{ta}\right\} \leq E[e^{tX}]e^{ta}$$

where we had the same inequality order as for the above case because we are using $t < 0$. □

Before we proceed, we clarify that the definition of convexity(if you recall there are many equivalent ones depending on the context recall `MIT Nonlinear Optimization`) used will be the following

Definition 147

A twice differentiable real valued function $f(x)$ is said to be **convex** if $f''(x) \geq 0$ for all x and **concave** if $f''(x) \leq 0$

Theorem 148 (Jensen Inequality)

If $f(x)$ is a convex function(as defined above) then

$$E[f(X)] \geq f(E[X])$$

provided that the expectations exist and are finite

Proof. Expanding $f(x)$ in a taylor expansion about $\mu = E[X]$ yields

$$f(x) = f(\mu) + f'(\mu)(x - \mu) + \frac{f''(\xi)(x - \mu)^2}{2}$$

where ξ is some value between x and μ(recall mean value theorem). Since $f''(\xi) \geq 0$ we obtain

$$f(x) \geq f(\mu) + f'(\mu)(x - \mu)$$

hence

$$f(X) \geq f(\mu) + f'(\mu)(X - \mu)$$

then taking expectations we have

$$E[f(X)] \geq f(\mu) + f'(\mu)E[X - \mu] = f(\mu) = f(E[x])$$

again recall $E[X - \mu] = 0$ as mentioned in the proof for 142.

9.2 poisson processes

recall in 46 we described in what situations will modeling a distribution as some poisson distribution function $P\{N = k\}$ will be appropriate. What if instead of a single event, we want to model a sequence of events with respect to time as some poisson probability density function $P\{N(t) = k\}$? To do so we first need to ensure the following assumptions are true

1. The probability of exactly 1 event occurs in a given interval of length h is equal to $\lambda h + o(h)$...to be continued