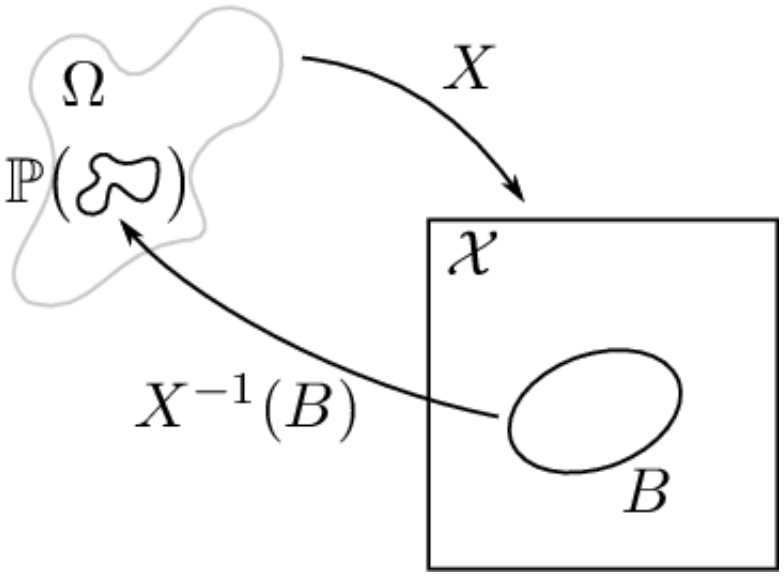
Probability, Statistics & Measure Theory

MASTER NOTES - VOLUME I



Sean Conlon
2021 -

"Probability is common sense reduced to calculation"
- Laplace

Contents

Topic I
Probability Theory

1 Sample Space & Probability Measure

1.1 Revision of Set Theory

We review some fundamental Set operations and theorems that appear frequently in our study of probability and measure.

Intersection & Union

Definition - Intersection and Union

The *intersection* of two sets S and T is given by

$$S \cap T = \{x : x \in S \text{ and } x \in T\}$$

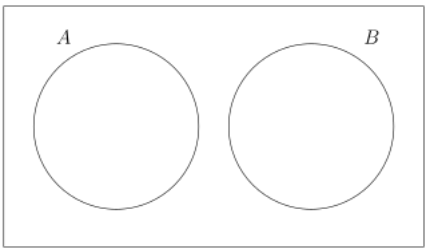
The *union* of two sets S and T is given by

$$S \cup T = \{x : x \in S \text{ or } x \in T\}$$

Disjoint Events

Definition - Disjoint/Mutually Exclusive Events

Two events A and B are said to be *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$



De Morgan's Laws

Theorem - De Morgan Laws

For any finite or infinite collection of sets S_n we have

$$\left(\bigcup_{n=1}^{\infty} S_n\right)^c = \bigcap_{n=1}^{\infty} S_n^c \qquad \text{and} \qquad \left(\bigcap_{n=1}^{\infty} S_n\right)^c = \bigcup_{n=1}^{\infty} S_n^c$$

1.2 Probability Axioms & Properties

Axioms of Probability

We contain below the axioms of probability under which all further theorems are constructed. For any outcome space Ω, probability measure $\mathbb{P}$ and disjoint events $A_n \subset \Omega$ we have

PA1 $\mathbb{P}(A_n) \geq 0$

PA2 $\mathbb{P}(\Omega) = 1$

PA3 $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$

The probability measure $\mathbb{P}$ has the following properties

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

Examples and Technique

Whilst it is tendency to think of DeMorgan's Laws as $n \longrightarrow \infty$, when solving set theory problems in the finite case, applying either of the laws is a good first step. Taking compliments of both sides is also a useful and common technique.

Example. Show that $(A^c \cap B^c)^c = A \cup B$ and $(A^c \cup B^c)^c = A \cap B$

Consider sets $A, B \subset \Omega$ and by DeMorgan's laws we have

$$A^c \cap B^c = (A \cup B)^c$$

Taking compliments of both sides we obtain the first of our desired equality's.

$$(A^c \cap B^c)^c = A \cup B$$

An identical technique will deliver the second of the equality's, only differing in that we apply the other of DeMorgan's laws.

1.3 Conditional Probability & Bayes Theorem

A key topic of probability theory; conditional probability considers the likelihood of an outcome *given that* some other outcome has already occurred. Conditional probability has great functional importance as it underpins much of applied probability theory.

Definition - Conditional Probability

The probability of event A given the occurrence of event B under the standard measure is given by

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

Note that we must have $\mathbb{P}(B) \neq 0$ Note further that $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability

Note that we may often rearrange the definition of $\mathbb{P}(A|B)$ to yield the following expression

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$$

Law of Total Probability

Total Probability Theorem

if $A_1, \ldots, A_n$ is a partition of Ω (that is, $\cup_{i=1}^{n} A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for any i and j) then we have

$$\mathbb{P}(B) = \mathbb{P}(B|A_1)\mathbb{P}(A_1) + \ldots + \mathbb{P}(B|A_n)\mathbb{P}(A_n)$$

Bayes' Theorem

Bayes' Theorem is one of the most important results of foundational probability. Bayes theorem see's wide application in prediction and forecasting.

Bayes' Theorem

If $A_1 \ldots, A_n$ is a partition of Ω then

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B|A_1)\mathbb{P}(A_1) + \ldots + \mathbb{P}(B|A_n)\mathbb{P}(A_n)}$$

1.4 Independence

Independence is the notion that two random outcomes occur separately from one another. Independence is a powerful assumption and yields useful results in probabilistic analysis.

Definition - Independence & Conditional Independence

We say that A and B are independent if

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

We say that A and B are *conditionally independent on* C if we have

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$$

This generalises to n outcomes/variables in which we have: $A_1, \ldots, A_n$ are independent with respect to measure $\mathbb{P}$ if we have

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

Notes and Corollaries

It is important to note that independence does not imply conditional idependence or vice-versa. However we do have the following

- A and B independent $\implies$ A and B^c are independent
- If $\mathbb{P}(B \cap C) \neq 0$ A and B are conditionally independent on C $\iff$ $\mathbb{P}(A|B \cap C) = \mathbb{P}(A|C)$
- If random variables X and Y are independent $\implies$ $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ and $Var(X + Y) = Var(X) + Var(Y)$

2 Discrete Random Variables

Given a discrete random variable X, the probability mass function is the measure that assigns the sample space to the output space. We write $p_X(x) = \mathbb{P}(X = x)$. All mass functions obey the following properties

Definition - Probability Mass Function

All probability mass functions p_X satisfy the following properties

PMF1 $p_X(x) \geq 0$ for all $x \in \mathbb{R}$.

PMF2 $p_X(x) > 0$ for at most a countable number of x. That is, $p_X(x)$ cannot be greater than 0 for all $x \in \mathbb{R}$

PMF3 $\sum_x p_X(x) = 1$

2.1 Expectation, Mean & Variance

Expectation

Definition - Expectation of a Discrete Random Variable

The *expected value* or mean of a discrete random variable X is given by

$$\mathbb{E}[X] = \sum_x x p_X(x)$$

Expectation is a linear operator. If $\alpha, \beta \in \mathbb{R}$ then

$$\mathbb{E}[\alpha X + \beta] = \alpha \mathbb{E}[X] + \beta$$

Variance

Definition - Variance of a Random Variable

The *variance* of a random variable X is given by

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

Variance **is not linear**. if $\alpha, \beta \in \mathbb{R}$ then

$$Var(\alpha X + \beta) = \alpha^2 Var(X)$$

In particular for a discrete random variable X we have the following expression

$$Var(X) = \sum_x (x - \mathbb{E}[X])\mathbb{P}(X = x)$$

2.2 Discrete Probability Distributions

2.2.1 Bernouli Random Variable

The Bernouli random variable models a *yes/no* random outcome. The distribution is parameterised by the likelihood of success p. If $X \sim \text{Bernoulli}(p)$ then the mass function is given by

$$p_X(x) = \begin{cases} p & x = 0 \\ 1 - p & x = 1 \end{cases}$$

with expectation and variance

$$\mathbb{E}[X] = p \qquad\qquad Var(X) = p(1-p)$$

Note that it is common to denote $q = 1 - p$

2.2.2 Binomial Random Variable

The Binomial random variable is used to model a number of successes in a sequence of n Bernoulli Trials. It models random process's that are repeated Bernoulli random variables. The binomial distribution is parameterised by the likelihood of success p and the number of trials n. If $X \sim \text{Binomial}(n, p)$ then the mass function is given by

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

With expectation and variance

$$\mathbb{E}[X] = np \qquad\qquad Var(X) = np(1-p)$$

2.2.3 Discrete Uniform Random Variable

The Discrete uniform random variable models situations where there is equal likelihood between any outcome in the closed interval $[m, n]$ The distribution is parameterised by the bounds of the interval, which must be integers. If $X \sim U[m, n]$ then the mass functions is given by

$$p_X(x) = \frac{1}{n - m + 1} \qquad x = m, \ldots, n$$

With expectation and variance

$$\mathbb{E}[X] = \frac{n+m}{2} \qquad\qquad Var(X) = \frac{(n-m+1)^2 - 1}{12}$$

2.2.4 Geometric Random Variable

A useful distribution; the geometric random variable models situations where we consider the amount of time before a success occurs. A common example is the random variable given by the number of coin flips before a heads occurs. The distribution is parameterised by the likelihood of success p. If $X \sim \text{Geometric}(p)$ then the mass function is given by

$$p_X(x) = p(1-p)^{x-1} \qquad x = 1, 2, \ldots$$

With expectation and variance

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

The intuition behind the probability mass function is that in order to have $X = x$, we must first have $x - 1$ successive and independently occurring failures, with probability $p - 1$, before we have a success with probability p

2.2.5 Poisson Random Variable

The poisson distribution expresses the possibility of a given number events occuring within a fixed time or space if these events occur with a known constant mean rate and independently of the time since the last event. The Poisson distribution is parameterised by $\lambda > 0$ which describes mean arrival rate. If $X \sim \text{Poisson}(\lambda)$ then the mass function is given by

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots$$

With expectation and variance

$$\mathbb{E}[X] = \lambda \qquad\qquad Var(X) = \lambda$$

Note that the Poisson distribution is one of the few such that $\mathbb{E}[X] = Var(X)$

Though not assessable in most courses, the intuition for the probability mass function comes from defining a random variable $X_n \sim \textit{Binomial}(n, \lambda/n)$. In taking $n \longrightarrow \infty$ the density of X_n approach's to that of the Poisson random variable. This in fact forms the basis of the so-called *Poisson approximation* in which binomial random variables with large n and small p can be approximated using a $\lambda = pn$ Poisson distribution. Doing so avoids the intensive computation of the factorials in a Binomial distribution's mass function.

Exercises on the Poisson random variable often involve recognising the taylor expansion of e^x which is

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

2.2.6 Indicator Function

The indicator function defined on a set X *indicates* membership of a an element to a subset of X. We often denote indicator functions by $\mathcal{I}_{x \in A}$ where

$$\mathcal{I}_{x \in A} = \begin{cases} 0 & x \notin A \\ 1 & x \in A \end{cases}$$

When the even A is random, indicator functions act as Bernoulli random variables. Suppose $\mathbb{P}(x \in A) = p$ then $\mathbb{E}[\mathcal{I}_{x \in A}] = p$ and $\mathbb{I}_{x \in A} \sim \text{Bernoulli}(p)$

2.3 Functions of Discrete Random Variables

In practice, we often wish to consider some function of a random variable X. This can be a linear function, summation, product ect cetera. In such circumstances, the function of the random variable $g(X)$ is itself a random variable. We write

$$Y = g(X)$$

where the PMF of Y can be deduced as

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x)$$

The expected value of Y is given by the so called *law of the unconscious statistician*;

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$$

It is important to note as well

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$$

Examples

We consider a famous example of this in the so-called *'St Petersburg Paradox'*. The problems goes, consider a game in which we flip a coin until it arrives at a tails. Let X denote the number of such flips. We then receive $\$2^X$. How much would one expect to win?

Indeed, $X \sim$ Geometric$(p = 1/2)$ and thus $\mathbb{E}[X] = 2$. It would be **erroneous** however to assume that the expected win would be given by $2^{\mathbb{E}[X]} = 2^2$ as $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$.

To approach this problem correctly, we shall define P as the pay-off of the game. Using the law of the unconscious statistician

$$\begin{aligned}
\mathbb{E}[P] &= \sum_{n=1}^{\infty} P(X = n)\mathbb{P}(X = n) \\
&= \sum_{n=1}^{\infty} 2^n \left(\frac{1}{2}\right)^n \\
&= \sum_{n=1} \frac{2^n}{2^n} \\
&= \sum_{n=1} 1
\end{aligned}$$

Hence, we arrive at the *paradox* of this problem, in that we have proven it has theoretically infinite pay-off!

2.4 Joint Densities of Multiple Discrete Random Variables

How does one deal with two (or more) random variables? Take for example the random variables S be the sum of two dice rolls and M the maximum of the two faces. How could we find $\mathbb{P}(S = 9, M = 5)$?

For instances of multiple random variables, say X and Y we define the *join density* of X and Y to be

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

The *marginal densities* p_X, p_Y can be retrieved from the joint density by considering

$$p_X(x) = \sum_y p_{X,Y}(x, y) \qquad p_Y(y) = \sum_x p_{X,Y}(x, y)$$

Joint densities also exhibit the following properties. If $Z = g(X, Y)$ then

$$p_Z(z) = \sum_{g(x,y)=z} p_{X,Y}(x, y)$$

$$\mathbb{E}[Z] = \mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y)$$

Examples

2.5 Conditioning on an Event or Random Variable

We can condition any random variable X on a event A or another random variable Y. To condition X on A we have

$$p_{X|A}(x) = \mathbb{P}(X = x|A) = \frac{\mathbb{P}(X = x, A)}{\mathbb{P}(A)}$$

More commonly, given a random variable $Y = y$ we can condition X on $Y = y$ as follows

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

which in its own sense is a random variable with exepectation

$$\mathbb{E}[X|Y = y] = \sum_x x\mathbb{P}(X = x|Y = y) = \sum_x x\frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Note that conditional probability mass functions are *'true'* pmf's in the sense that $p_{X|A}(x) \geq 0$ and

$$\sum_x p_{X|A}(x) = 1$$

Examples

2.6 Independence of Random Variables

Definition - Independence of Random Variables

Random variables X, Y are independent if for all pairs (x, y) we have

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \qquad \forall x, y$$

Properties of Independent Random Variables

If random variables X, Y are independent then for $g(X)$ and $h(Y)$ are independent for any functions h and g. Furthermore we have

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

$$Var(X + Y) = Var(X) + Var(Y)$$

3 Continuous Random Variables

Given a continuous random variable X, the probability density function f_X is the measure that assigned the sample space to the output space. For any continuous random variable X we write $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx$

Definition - Continuous Random Variable, Probability Density Function

All continuous random variables X and associated probability density functions f_X satisfy the following properties

PDF1 $\int_{-\infty}^{\infty} f(x)dx = 1$.

PDF2 $f(x) \geq 0$ for all $x \in \mathbb{R}$

Continuous random variables also obey the following properties

CRV1 for any particular x we have $\mathbb{P}(X = x) = 0$.

CRV2 for a very small choice of δ we have $\mathbb{P}(x \leq X \leq x + \delta) \approx f(x)\delta$

3.1 Expectation & Variance

Expectation of a continuous random variable is defined in exactly the same way as the discrete case, just with the use of an integral as opposed to a summation. All familiar properties of expectation and variance still hold.

Definition - Expectation & Variance of a Continuous Random Variable

Given a continuous random variable X we define the mean or expectation of X as

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

Which in turn defines the variance as

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \left(\int_{-\infty}^{\infty} x f_X(x)dx\right)^2$$

for any function $Y = g(X)$ we also have

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

3.2 Cumulative Density Function

Definition - Cumulative Density Function

The cumulative density function (CDF) of a random variable X is given by

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

All CDF's share the following properties

CDF1 F_X is non decreasing. That is $y \leq z \implies F_X(y) \leq F_X(z)$

CDF2 $\lim_{x \longrightarrow -\infty} F_X(x) = 0$ and $\lim_{x \longrightarrow \infty} F_X(x) = 1$

CDF3 F_X is right-continuous

CDF4 if X is a continuous random variable then $f_X(x) = F_X'(x)$

CDF5 if X is a discrete random variable then $p_X(x) = F_X(x) - F_X(x-1)$

Examples & Technique

Consider the common problem of finding the associated density function $f_Y(y)$ of random variable $Y = g(X)$ where X is some known, continuous random variable and g is an invertible function. A useful technique is to consider the CDF of Y and then differentiate. Doing so, we consider

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \leq g^{-1}(y)) \\ &= \int_{-\infty}^{g^{-1}(y)} f_X(x)dx \end{aligned}$$

Using the fundamental theorem of calculus and **CDF4** we can deduce the probability density function as

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\int_{-\infty}^{g^{-1}(y)} f_X(x)dx \\ = g^{-1\prime}(y)f(g^{-1}(y)) \end{aligned}$$

3.3 Continuous Probability Distributions

3.3.1 Continuous Uniform Random Variable

The Continuous uniform random variable models situations where there is equal likelihood between any outcome in the closed interval [a, b] The distribution is parameterised by the bounds of the interval, which must be real numbers. If $X \sim U[a, b]$ then the density function of X is given by

$$f_X(x) = \frac{1}{b-a} \qquad a \leq x \leq b$$

With mean and variance

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad Var(X) = \frac{(b-a)^2}{12}$$

In particular, the $U[0, 1]$ Distribution has the property that $\mathbb{P}(X > z) = 1 - z$

3.3.2 Exponential Random Variable

The exponential distribution models the time between events in a Poisson point process; a process in which events occur continuously and independently at a constant average rate. It is the continuous analog of the geometric distribution and has the key property of being *memoryless*. The distribution is parameterised by the constant average rate of arrival; λ. If $X \sim$ *Exponential*(λ) then the density function is given by

$$f_X(x) = \lambda e^{-\lambda x}$$

With mean and variance

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

3.3.3 Cauchy Random Variable

The Cauchy distribution is often exemplified as a *pathological* distribution in the sense that is poorly behaved and offers little room for analysis. In practice, the random variable never arises and is more an object of study. The distribution is governed by a *shape parameter* α, which describes the width of the bell curve. If $X \sim \text{Cauchy}(\alpha)$ then the density function is given by

$$f_X(x) = \frac{\alpha}{\pi(x^2 + \alpha^2)}$$

X is pathological in the sense that it has undefined moments. Consequently, the mean and variance of the distribution do not exist.

3.3.4 Normal Random Variable

Arguably the most important amongst all distributions, the normal random variable models any *bell-curved* shape distribution. The notion of it being *normal* arises from the fact that it appears frequently in nature and application, and that random samples of size n tend to tend to approach a bell-curve as $n \longrightarrow \infty$. The distribution is parameterised by its expectation and variance. if $X \sim N(\mu, \sigma^2)$ then the density function is given by

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

3.3.5 Gamma Random Variable

The Gamma distribution is a two-parameter random variable that describes a family of distributions. Exponential and χ^2 are particular instances of the Gamma distribution. The random variable takes a shape parameter α and scale parameter λ. if $X \sim \text{Gamma}(\alpha, \lambda)$ then the density function is given by

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

With mean and variance

$$\mathbb{E}[X] = \frac{\alpha}{\lambda} \qquad\qquad Var(X) = \frac{\alpha}{\lambda^2}$$

3.4 Gamma Function & Gamma Distribution

We have presented the gamma random variable but are yet to define the gamma function $\Gamma(\cdot)$ and explore its properties.

Defintion - Gamma Function $\Gamma(\cdot)$

The gamma function is defined on $z \in \mathbb{R}^+$ (though can be extended to regions of the complex plane) and is given by

$$\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} du$$

The gamma function has the following properties

G.F.1 $\Gamma(1/2) = \sqrt{\pi}$

G.F.2 $\Gamma(z+1) = z\Gamma(z)$

G.F.3 $\Gamma(z) = (z-1)!$

Particular instances of the Gamma Distribution

Recall that the gamma distribution is governed by two parameters, its shape α and scale λ. Recall further that the density of the gamma distribution is given by

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

We obtain the Exponential(λ) distribution if we take $\alpha = 1$

$$f_X(x|\alpha = 1, \lambda) = \frac{\lambda}{\Gamma(1)} x^0 e^{-\lambda x} = \lambda e^{-\lambda x}$$

Useful results and Integrals of the Gamma Function

In proofs of the properties of the gamma function, particular results of integrals arise that are useful in exploring other density functions. We keep a catalogue of them below

3.5 Normal & Standard Normal Distribution

The normal distribution is paramount to the studies of probability and statistics, so much in fact that an entire subsection is dedicated to it here.

Definition - Normal Distribution

If $X \sim N(\mu, \sigma^2)$ then X has the following density function on $\mathbb{R}$

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

A particular case of the normal distribution is the $Z \sim N(0, 1)$ which has the following density function

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

Note that it is common to denote the above density function as $\phi(z)$ and distribution function (CDF) as $\Phi(z)$

Technique - Standardisation

It is common technique in proofs and applications to 'standardise' any normal distribution we are working with. Given that $X \sim N(\mu, \sigma^2)$, then the change of variable given by

$$Z = \frac{X - \mu}{\sigma}$$

Results in a $Z \sim N(0, 1)$ distribution. Note that conversely, we have

$$X = \sigma Z + \mu$$

resulting in a $X \sim N(\mu, \sigma^2)$ distribution.

3.6 Joint Normal Distribution

Definition - Joint Normal Distribution

The pair of continuous random variables (X, Y) are said to be *jointly normal* or *Bivariate* if for all $\alpha, \beta \in \mathbb{R}$ the linear combination

$$\alpha X + \beta Y \sim N(\mu, \sigma^2)$$

Theorem - Independence of (X, Y)

If (X, Y) are bivariate normal and are uncorrelated, then they are independent.

Theorem - Independence of Standard Bivariate Normal

If (X, Y) are standard Bivariate normal then $X \perp Y \iff r = 0$ where

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1 - r^2}} \exp\left\{-\frac{x^2 - 2rxy + y^2}{2(1 - r^2)}\right\}$$

3.7 Conditioning on an Event & Memoryless Property

Memorylessness is a property of certain probability distributions. It usually refers to the cases when the distribution of a "waiting time" until a certain event does not depend on how much time has elapsed already. To model memoryless situations accurately, we must constantly 'forget' which state the system is in: the probabilities would not be influenced by the history of the process. Only two kinds of distributions are memoryless: geometric distributions of non-negative integers and the exponential distributions of non-negative real numbers.

Definition - Conditional PDF

The conditional probability density function of a continuous random variable X given the event $\{X \in B\}$ with $\mathbb{P}(X \in B) > 0$ is given by

$$f_{X|B}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}(X \in B)} & x \in B \\ 0 & x \notin B \end{cases}$$

Definition - Memoryless Property & Examples

A random variable X is said to be *memoryless* if for any $x > 0$ and $a > 0$ we have

$$\mathbb{P}(X > x + a | X > a) = \mathbb{P}(X > x)$$

Examples

Let $X \sim$ *Exponential*(λ) find the density of $f_{X|X>a}$ for some $a > 0$

Solution. We can easily determine that $\mathbb{P}(X > a) = e^{-\lambda a}$ so using our definition above we have

$$\begin{aligned} f_{X|X>a} &= \frac{f_X(x)}{\mathbb{P}(X > a)} \\ &= e^{\lambda a} \lambda e^{-\lambda x} \\ &= \lambda e^{-\lambda(x-a)} \end{aligned}$$

Thus we conclude that

$$f_{X|X>a}(x) = \begin{cases} \lambda e^{-\lambda(x-a)} & x > a \\ 0 & x \leq a \end{cases}$$

Further Properties

It is important to keep in mind that $X|A$ is itself a random variable, with its own mean and variance defined as follows

Definition - Mean and Variance of $X|A$

Similar to any other continuous random variable we have

$$\mathbb{E}[g(X)|A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x) dx$$

The law of total probability theorem also extends rather naturally to the conditional distribution and expectation. This is formalised in the following theorem.

Theorem - Law of Total Probability (Density Version)

Let $A_1, \ldots, A_n$ be a partition of the space. We then have

$$f_X(x) = \sum_{i=1}^{n} f_{X|A_i}(x)\mathbb{P}(A_i)$$

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i]\mathbb{P}(A_i)$$

3.8 Joint Densities of Continuous Random Variables

Two random variables are said to be *jointly continuous* if there exists a non-negative function $f_{X,Y}(x,y)$ such that

$$\mathbb{P}(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x,y)dydx$$

If there exists such a function $f_{X,Y}(x,y)$ then we know that X and Y are both continuous random variables, with *marginal densities*

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

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

All our usual theorems on expectation extend to the joint density function

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

The conditional PDF $f_{X|Y}(x|y)$ of X *given* Y is defined for whenever $f_Y(y) > 0$ as

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

Furthermore, Baye's Theorem holds for continuous random varibles

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_Y(y)}{f_X(x)} = \frac{f_{Y|X}(y|x)f_Y(y)}{\int_{\mathbb{R}} f_{X|Y}(x|v)f_Y(v)dv}$$

Of particular importance is the topic of Independence. Two random variables X and Y are said to be independent if

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

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

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

We also have the following characterising theorem for independence of continuous random variables.

Characterisation of Independence

Two jointly continuous random variables X and Y are independent if and only if

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

Law of Total Expectation for Joint Densities

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} \mathbb{E}[XY|X=x]f_X(x)dx$$

3.9 Multivariate Normal Distribution

The multivariate normal distribution is a generalisation of the uni-variate normal to higher dimensions. It is of particular importance to latter study of probability in its relation to the mulit-variable central limit theorem.

We say $\mathbf{X} = (X_1, \ldots, X_n)$ is multivariate normal, or $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is the mean vector $\boldsymbol{\mu} = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])$ and $\boldsymbol{\Sigma}$ is a covariance matrix $\boldsymbol{\Sigma}_{ij} = Cov(X_i, X_j)$

Density

In the most general case, the n-dimensional multivariate normal normal distribution as the following density function. If $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp \frac{-1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Where we denote $|\boldsymbol{\Sigma}| = \det \boldsymbol{\Sigma}$. We will now specifically consider the two dimensional case of the distribution of $\mathbf{X} = (X, Y)$ where X and Y follow normal distributions.

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp - \frac{1}{2\,(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right]$$

where ρ is the correlation between X and Y and where $\sigma_X > 0$ and $\sigma_Y > 0$. In this case,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Theorems

Characterisation of Independence In the Bivariate Case

If (X, Y) are bivariate normal then X and Y are independent if and only if $\rho = 0$

Construction of a Bivariate Normal

If U and Z are two independent standard normal random variables and $V = \rho U + \sqrt{1-\rho^2} Z \quad (|\rho| < 1)$ then (U, V) has a standard bivariate normal distribution with correlation ρ

Conditional Densities of the Bivariate Normal

If (U, V) are bivaraite normal with correlation coefficient ρ then the random variable

$$U|V = v \sim N(\rho v, 1 - \rho^2)$$

4 Further Topics on Random Variables

4.1 Total Expectation and Variance Theorems

A tool useful for Probability theorists is the law of total expectation defined as follows. A similar, but distinct theorem also applies for the variance of a random variable.

Law of Total Expectation

If X is a random variable such that its expectation is defined and Y is a random variable on the same probability space, then

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

If X is discrete then

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) = \sum_y \mathbb{E}(X|Y=y)\mathbb{P}(Y=y)$$

If X is continuous

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

Law of Total Variance

If X and Y are variables on the same probability space and the variance of X is finite then

$$Var(X) = \mathbb{E}(Var(X|Y)) + Var(\mathbb{E}(X|Y))$$

Examples

Suppose that only two factories supply light bulbs to the market. Factory X's bulbs work for an average of 5000 hours, whereas factory Y's bulbs work for an average of 4000 hours. It is known that factory X supplies 60% of the total bulbs available. What is the expected length of time that a purchased bulb will work for?

Define the variable of interest $L :=$ lifetime of a light bulb. Then by the law of total expectation

$$\begin{aligned}\mathbb{E}(L) &= \mathbb{E}(L|X)\mathbb{P}(X) + \mathbb{E}(L|Y)\mathbb{P}(Y) \\ &= 5000(0.6) + 4000(0.4) \\ &= 4600\end{aligned}$$

Another question is given a fair coin, what is the expected number of flips necessary in order to see the sequence HHH?

4.2 Covariance & Correlation

Covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values (that is, the variables tend to show similar behavior), the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, (that is, the variables tend to show opposite behavior), the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables

Definition - Covariance of Random Variables (X, Y)

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

An equivelant definition which is useful for proofs is as follows

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

Alternative Identity for Covariance

$$\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[XY|X]] = \mathbb{E}[X\mathbb{E}[Y|X]]$$

The correlation coefficient of two random variables $\rho(X, Y)$ standardises the covariance to the interval $[-1, 1]$ and is defined as follows

Definition - Correlation Coefficient of Random Variables (X, Y)

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

Note: a Correlation coefficient of 0 does not always imply independence.

4.3 Simulations

Simulations play a critical role in computational and applied probability. The ability to draw numbers from a particular distribution underpins many applications such as Monte Carlo and Las Vegas algorithms. In this section we explore two rudimentary algorithms for simulating a random variable.

Inverse Transform Algorithm

The inverse transform method is a relatively simple algorithm for sampling a continuous random variable X with known distribution function (CDF) F_X. The first step of the algorithm is to find the inverse of $F_X(x)$, which makes it impractical for Distributions with difficult-to-invert CDF's, such as the normal random variable, and entirely useless for non-invertiable Density Functions.

Input: CDF $F_X(x)$ of random variable X we wish to sample;
1. Determine the Inverse of the distribution function F_X^{-1};
2. Generate $U \sim U[0,1]$;
3. Let $X = F_X^{-1}(U)$;

Then X follows the distribution governed by F_X
Algorithm 1: Inverse Transform for Continuous Distributions

Example

Suppose we wish to sample from $X \sim \text{Exponential}(\lambda)$. We recall that $F_X(x) = 1 - e^{-\lambda x}$. The first step is to find the inverse of F_X

$$x = 1 - e^{-\lambda F^{-1}(x)}$$
$$\therefore F_X^{-1}(x) = \frac{-1}{\lambda} \ln(1-x)$$

To sample from X we now take a sample $u = U[0,1]$ and then take $X = F_X^{-1}(u)$.

Intuition

We want to generate X with distribution function $F_X(x)$. Intuitively, $F_X(x)$ is monotone increasing. We also assume we have some method for sampling from a $U[0,1]$

To sample from the distribution of X we want to find some monote increasing transformation T such that $T(U) \stackrel{d}{=} X$. We will have

$$\begin{aligned} F_X(x) = \mathbb{P}(X \leq x) &= \mathbb{P}(T(U) \leq x) \\ &= \mathbb{P}(U \leq T^{-1}(x)) \\ &= T^{-1}(x) \end{aligned}$$

Where we obtain the final line using the property of the $U[0,1]$ distribution. So we can sample $F_X(x) = T^{-1}(x)$, which implies we can sample X by using $T^{-1} = F_X^{-1}$.

4.4 Moment Generating Functions & Moment Generating Functions of Common Distributions

Recall that the n^{th} moment of the random variable X is defined to be $\mathbb{E}[X^n]$. The moment generating function gives an efficient method for computing such expectations. The definition and basic properties are outlined below.

Defintion - Moment Generating Function

Given a random variable X the moment generating function $M_X(t)$ is given by $M_X(t) = \mathbb{E}[e^{tX}]$ in the discrete case that is

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_x e^{tx} p_X(x)$$

and in the continuous case that is

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$$

All Moment Generating Functions $M_X(t)$ obey the following properites

MGF1 $M_X(0) = 1$ however $M_X(t)$ may not be defined for all t

MGF2 $\mathbb{E}[X^n] = M_X^n(0)$ where M^n denotes the n^{th} derivative

4.4.1 Bernoulli Random Variable

if $X \sim \text{Bernoulli}(p)$ with $X \in \{1, 0\}$ then

$$M_X(t) = (1-p) + pe^t$$

4.4.2 Binomial Random Variable

if $XS \sim \text{Binomial}(n, p)$ then

$$M_X(t) = ((1-p) + pe^t)^n$$

Note that this is (intuitively) the moment generating function of a Bernoulli random variable, raised to the n^{th} power

4.4.3 Geometric Random Random Variable

if $X \sim \text{Geometric}(p)$ then

$$M_X(t) = \begin{cases} pe^t \frac{1}{1-(e^t(1-p))} & \text{if } (1-p) + pe^t < 1 \implies t < \log\left(\frac{1}{1-p}\right) \\ +\infty & \text{otherwise} \end{cases}$$

4.4.4 Poisson Random Variable

if $X \sim \text{Poisson}(\lambda)$ then

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

4.4.5 (Continuous) Uniform Random Variable

If $X \sim U[a, b]$ then

$$M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

4.4.6 Exponential Random Variable

If $X \sim \text{Exponential}(\lambda)$ then

$$M_X(t) = \lambda\left(\frac{1}{\lambda - t}\right) \quad \text{if } t > \lambda \text{ otherwise } +\infty$$

4.4.7 Gama Random Variable

if $X \sim \text{Gamma}(\alpha, \lambda)$ then

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}$$

Note that this is the exponential moment generating function raised to the power of α

4.4.8 Normal Random Variable

if $X \sim N(\mu, \sigma^2)$ then

$$M_X(t) = \exp\left\{\frac{\sigma^2 t^2}{2} + t\mu\right\}$$

Theorem - Equivalency of Distributions

if $M_X(t) = M_Y(t)$ for all $t \in \mathcal{O}$ where $\mathcal{O}$ is an open interval containing 0 then $X \overset{d}{=} Y$. Note that proving this theorem requires heavy use of measure theory and hence the assumptions on $\mathcal{O}$.

Corollary: This theorem implies that the distribution of random variable X can be recognised by $M_X(t)$

Theorem - MGF of Sums of Independant Random Variables

Suppose $X \perp Y$, then the Moment Generating Function of $M_{X+Y}(t)$ is given by

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

4.5 Sums of Independent Random Variables & Convolution

To find densities of some transforms/functions of random variables we can apply the CDF technique, this cannot be done so for summations, In such instances, we need to apply a different array of techniques.

Moment Generating Function Approach

The distribution of the sum of two or more independent random variables can be done by first computing the moment generating function and then identifying its distribution. To illustrate this consider $Y = \sum_{i=1}^{n} X_i$ then by the Independence of X_i we have

$$M_Y(t) = M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

Example i: Suppose $X_i \sim \text{Poisson}(\lambda_i)$ and we wish to find the distribution of $Y = \sum_{i=1}^{n} X_i$ where X_i mutually independent. Using the above approach

$$\begin{aligned} M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) &= \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} \\ &= e^{\lambda_1(e^t-1)} \cdots e^{\lambda_n(e^t-1)} \\ &= e^{(e^t-1)\sum_{i=1}^{n}\lambda_i} \end{aligned}$$

Which is the moment generating function of a Poisson$(\sum_{i=1}^{n} \lambda_i)$ random variable and thus $Y \sim$ Poisson$(\sum_{i=1}^{n} \lambda_i)$

Example ii: Suppose $X_i \sim \text{Exponential}(\lambda)$ and we wish to find the distribution of $Y = \sum_{i=1}^{n} X_i$. Using the moment generating function approach

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$$

Which we recognise as the moment generating function of a $Y \sim \text{Gamma}(n, \lambda)$

Direct Approach (Convolution)

In the discrete case, consider discrete random variables X and Y and suppose we wish to find the distribution of $Z = X + Y$. Using the law of total probability we find

$$\begin{aligned}
\mathbb{P}(Z = z) = \mathbb{P}(X + Y = z) &= \sum_x \mathbb{P}(X + Y = z | X = x)\mathbb{P}(X = x) \\
&= \sum_x \mathbb{P}(x + Y = z | X = x)\mathbb{P}(X = x) \\
&= \sum_x \mathbb{P}(Y = z - x | X = x)\mathbb{P}(X = x) \\
&= \sum_x \mathbb{P}(Y = z - x)\mathbb{P}(X = x) \qquad \text{as } X \perp Y \\
&= \sum_x p_Y(z - x)p_X(x)
\end{aligned}$$

We can obtain a similar expression by conditioning on Y. In doing so we find that

$$p_Z(z) = \sum_x p_Y(z - x)p_X(x) = \sum_y p_X(z - y)p_Y(y)$$

This can be generalised to the case of continuous random variables as follows. Suppose X, Y are independent, continuous random variables. Then by the law of total probability

$$\begin{aligned}
\mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) &= \int_{-\infty}^{\infty} \mathbb{P}(X + Y \leq z | X = x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \mathbb{P}(x + Y \leq z | X = x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \mathbb{P}(Y \leq z - x | X = x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \mathbb{P}(Y \leq z - x) f_X(x) dx \qquad \text{as } X \perp Y \\
&= \int_{-\infty}^{\infty} F_Y(z - x) f_X(x) dx
\end{aligned}$$

Differentiating with respect to z and we find

$$\begin{aligned}
f_Z(z) = f_{X+Y}(z) &= \frac{d}{dz} \int_{-\infty}^{\infty} F_Y(z - x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} \frac{\partial}{\partial z} (F_Y(z - x) f_X(x)) dx \\
&= \int_{-\infty}^{\infty} f_Y(z - x) f_X(x) dx
\end{aligned}$$

Conditioning on Y instead of X yields the alternative identity

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z - x) f_X(x) dx = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy$$

Examples:

Suppose $X \sim U[0,2]$ and $Y \sim U[3,4]$ and $X \perp Y$. Find the density function of $Z = X + Y$.

Solution. We will takle this problem using the convolution technique. The *trick* to convolutions is just case analysis as we will see as follows. Clearly $Z = z \in [3,6]$ and we recall the convolution formula;

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x) dx$$

4.6 Variance of Sums of Random Variables

In the case of the mean, we can apply the linearity of expectation to readily deduce that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. However, what about variance?

Theorem - Variance of Sums of Random Variables

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Proof.

$$\begin{aligned} Var(X+Y) &= \mathbb{E}[(X+Y-(\mathbb{E}[X]+\mathbb{E}[Y]))^2] \\ &= \mathbb{E}[((X-\mathbb{E}[X])+(Y-\mathbb{E}[Y]))^2] \\ &= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= Var(X) + Var(Y) + 2Cov(X,Y) \end{aligned}$$

□

Theorem - Variance of Sums of n Random Variables

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum_{k=2}^{n} \sum_{l=1}^{k-1} Cov(X_k, X_l)$$

The intuition behind this formula is to think of a matrix of random variables, and the first sum covers the Covariance across the the diagonal and the latter double sum covers every other combination.

4.7 Random Sums of Random Variables

The motivation for studying the distribution of random sums of random variables can be found in applications, such as for Insurance companies modelling their risk exposure. Suppose that X_i are iid and N follows some integer distribution independent from X. We now wish to find the distribution

$$Y = \sum_{i=1}^{N} X_i$$

To do so, we follow the route of using moment generating function, in combination with the law of total probability

$$\begin{aligned} M_Y(t) = \mathbb{E}[e^{tY}] &= \sum_{n=1}^{\infty} \mathbb{E}[e^{t(\sum_{i=1}^{n} X_i)}]\mathbb{P}(N=n) \\ &= \sum_{n=1}^{\infty} \mathbb{E}[e^{tX_i}]^n \mathbb{P}(N=n) \end{aligned}$$

Examples:

4.8 Least Squares Estimation

Suppose we wish to estimate the value of a random variable X. Practically speaking, we wish to minimise the so-called *expected square error* or

$$\mathbb{E}[(X-c)^2]$$

where c is our predictor for X. For this, we have our first and perhaps obvious theorem

Best Least Squares Estimator for a Random Variable

$$\text{argmin } \mathbb{E}[(X-c)^2] = \mathbb{E}[X]$$

Conditional Best Estimator

As we see in the previous section, the natural best estimator for a random variable X is unsurprisingly its expectation, $\mathbb{E}[X]$. The less obvious case is estimating X, based on a known oberservation of a known, (hopefully) related random variable Y. That is, what is the best choice c in the sense that the following expectation is minimised

$$\mathbb{E}[(X-c)^2|Y=y]$$

This gives way to our second theorem on the topic

Best Conditional Least Squares Estimator for a Random Variable

$$\text{argmin } \mathbb{E}[(X-c)^2|Y=y] = \mathbb{E}[X|Y=y]$$

Examples

Suppose (X,Y) have the following probability density function. Find the value of c and determine the best least squares estimator of $Y|X=x$ Where

$$f_{X,Y} = c\exp\left(-\frac{2}{3}(x^2+xy+y^2)\right)$$

Solution. It immediately appears to a Bivariate Normal Distribtuion, which has density function

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-r^2}}\exp\left\{-\frac{x^2-2rxy+y^2}{2(1-r^2)}\right\}$$

To find the value of r, we need only find the value such that

$$x^2-2rxy+y^2 = x^2+xy+y^2$$

Which is easily deduced as $-(1/2)$. Thus we can calulcate the coefficient c as follows

$$c = \frac{1}{2\pi\sqrt{1-(-1/2)^2}} = \frac{1}{2\pi\sqrt{3/4}} = \frac{1}{\pi\sqrt{3}}$$

To now deduce the best estimator of $Y|X=x$, we recall this is given by $\mathbb{E}[X|Y=y] = \rho x = -(1/2)x$

5 Limit Theorems

5.1 Central Limit Theorem

Let $X_1, \ldots, X_n$ be independent and identically distributed such that $\mathbb{E}[X_i] = \mu$ and $var(X_i) = \sigma^2$ then if we define the *empirical mean*

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

we have $\mathbb{E}[\bar{X}_n] = \mu$ and $var(\bar{X}_n) = \sigma^2/n$. These are the so-called "*first order approximations*" of the sample data. The more powerful limit technique is the central limit theorem, which is referred to as the "*second order approximations*". The theorem is extremely powerful in that it allows us to treat $\bar{X}_n$ as a standard normal random variable for sufficiently large n.

Central Limit Theorem

$$\lim_{n\longrightarrow\infty} \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{d}{=} Z \sim N(0,1)$$

Or equivalently

$$\lim_{n\longrightarrow\infty} \mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = \Phi(x)$$

5.2 Markov & Chebyshev Inequalities

The Chebyshev inequality provides a useful bound on $\mathbb{P}(X > a)$. It is particularly useful when the density of X is unknown (as is often the case in application) and the value of $\mathbb{E}[X]$ is known. The Markov Inequality is also a particular case of the Chebyshev inequality - to see the connection, one need only look at the proof.

Chebyshev Inequality

If random variable $X \geq 0$ and $a > 0$ then

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}$$

Markov Inequality

If X has mean μ and variance σ^2 then for any $c > 0$

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

5.3 Weak Law of Large Numbers

We return to the setting of consider i.i.d X_i with mean μ and variance σ^2 and setting $\bar{X}_n = \sum_{i=1}^{n} X_i$ the weak law of large numbers provides a framework for proving and applying the central limit theorem.

Weak Law of Large Numbers

$$\lim_{n \longrightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0 \qquad \epsilon > 0$$

6 Appendix

Useful Identities

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \qquad \sum_{i=1}^{n} i^2 = \frac{n^3}{3} + \frac{n^2}{2} + n$$

Taylor Series & Power Series

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$$

Integration By Parts

$$\int G(x)f(x)dx = f(x)g(x) - \int F(x)g(x)dx$$

Chain Rule

$$\frac{d}{dx}e^{f(x)} = f'(x)e^{f(x)}$$

Card Probability

The expected number of cards one would see in an n card deck in order to see the first of k cards is given by

$$\frac{n+1}{k+1}$$

For example, to determine how many cards one would expect to turn in a regular deck in order to see the first ace is given by $53/5 = 10.6$

Topic II
Proof Portfolio

7 Set Theory, Probability Axioms & Probability Measure

Properties of the Probability Measure

From probability axioms, prove that for any event A we have $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

Proof. Consider an event space Ω partitioned into disjoint events A and A^c it then follows

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

A simple rearrangement will produce the desired equality. □

Conditional Probability

Show that events A, B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A|B^c)$

Proof. □

8 Discrete Random Variables

Binomial Random Variable

Let $X \sim$ Binomial(n, p) derive $\mathbb{E}[X] = np$ and Var(X)=npq

Proof. We first recall that $\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n$ To assist in deriving $\mathbb{E}[X]$ we first consider the following expression

$$\begin{aligned}
\sum_{k=0}^{n} k\binom{n}{k} a^k b^{n-k} &= 0 + \sum_{k=1}^{n} k\binom{n}{k} a^k b^{n-k} \\
&= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} a^k b^{n-k} \\
&= \sum_{k=1}^{n} \frac{n(n-1)!}{(k-1)!((n-1)-(k-1)!} a^k b^{(n-1)-(k-1)} \\
&= an \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!((n-1)-(k-1)!} a^{k-1} b^{(n-1)-(k-1)} \\
&= an \sum_{j=0}^{n-1} \binom{n-1}{j} a^j b^{n-j} \qquad \text{by defining } j = k-1 \\
&= an(a+b)^{n-1}
\end{aligned}$$

Applying this to the expectation of X we find

$$\mathbb{E}[X] = \sum_{x=0}^{n} k\binom{n}{x} a^x b^{n-x} = np(1+(1-p))^{n-1} = np$$

□

Poisson Random Variable

Let $X \sim$Poisson(λ), prove that $\mathbb{E}[X] = \lambda$ and Var(X)= λ

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!} \\
&= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \\
&= \lambda \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!} \qquad \text{by defining } y = x - 1 \\
&= \lambda
\end{aligned}$$

We obtain the last line by recognising that $\sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!} = 1$. To compute the variance, we can avoid direct computation of the second moment and instead consider $\mathbb{E}[X(X-1)]$ as follows.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=0}^{\infty} x(x-1) \frac{\lambda^x e^{-\lambda}}{x!} \\
&= \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2} e^{-\lambda}}{(x-2)!} \\
&= \lambda^2 \sum_{y=0}^{\infty} y \frac{\lambda^y e^{-\lambda}}{y!} \qquad \text{Using the same trick as before} \\
&= \lambda^2
\end{aligned}$$

We can now easily compute the second moment and variance by recognising $\mathbb{E}[X^2] = \lambda^2 + \mathbb{E}[X]$

□

9 Continuous Random Variables

Exponential Random Varible

Let $X \sim$Exponential(λ) prove that $f_X(x)$ satisfies the properties of a density function, $\mathbb{E}[X] = 1/\lambda$ and $\mathbb{E}[X^2] = 1/\lambda^2$

Proof. We recall that $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$ and $\lambda > 0$. Clearly, f_X is non-negative on its domain as $\lambda > 0$. To show that it satisfies the other property of a pdf

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(x)dx &= \int_0^{\infty} \lambda e^{-\lambda x} dx \\
&= -e^{-\lambda x}\Big|_{x=0}^{x \longrightarrow \infty} \\
&= \lim_{x \longrightarrow \infty} -e^{-\lambda x} - (-e^{\lambda(0)}) \\
&= 1 + \lim_{x \longrightarrow \infty} -e^{-\lambda x} \\
&= 1
\end{aligned}$$

We now deduce the expectation of X by applying integration by parts □

Gamma Function

Prove the following properties of the Gamma Function, $\Gamma(1/2) = \sqrt{\pi}$ *as well as* $\Gamma(z+1) = z\Gamma(z)$ *and* $\Gamma(z) = (z-1)!$

We begin by first proving that $\Gamma(1/2) = \sqrt{\pi}$

We now prove that $\Gamma(z+1) = z\Gamma(z)$

Proof. We have

$$\begin{aligned}\Gamma(z+1) = \int_0^\infty u^z e^{-u} du &= -\int_0^\infty u^z(-1)e^{-u} du \\ &= u^z e^{-u}|_{u=0}^{u\longrightarrow\infty} + \int_0^\infty z u^{z-1} e^{-u} du\end{aligned}$$

Here, we use integration by parts with $g(u) = u^z$ and $F(u) = -e^{-u}$. By L'Hopitals rule, the limit as $u \longrightarrow \infty$ of $u^z e^{-u} \longrightarrow 0$ and thus we are left with

$$\Gamma(z+1) = \int_0^\infty z u^{z-1} e^{-u} du = z\int_0^\infty u^{z-1} e^{-u} du = z\Gamma(z)$$

□

Finally we can prove for all positive integers n we have $\Gamma(n) = (n-1)!$

Proof. This is by far the easiest result to prove, and we do so by induction. If we proceed on the hypothesis that for all positive integers up to n we have $\Gamma(n) = (n-1)!$ then we have in the case of $n+1$

$$\begin{aligned}\Gamma(n+1) &= n\Gamma(n) \\ &= n\cdot(n-1)! \qquad \text{by induction hypothesis} \\ &= n!\end{aligned}$$

Hence completing the induction. □

Gamma Random Variable

Let $X \sim$Gamma(α, λ) prove that f_X is indeed a density function, show that $\mathbb{E}[X] = \alpha/\lambda$ and $Var(X) = \alpha/\lambda^2$

Proof. We recall that if $X \sim$Gamma(α, λ) then $\alpha, \lambda > 0$ and

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

Clearly, by the non-negativitiy of $\alpha, \lambda > 0$, $f_X(x) > 0$ for all x also. We show that the surface integrates to 1

$$\begin{aligned}\int_0^\infty f_X(x)dx &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \lambda^\alpha x^{\alpha-1} e^{\lambda x} dx\end{aligned}$$

We take the change of varibale $u = \lambda x$ and note that $dx = du/\lambda$

$$\begin{aligned}\frac{1}{\Gamma(\alpha)} \int_0^\infty \lambda^\alpha x^{\alpha-1} e^{-\lambda x} dx &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \lambda^{\alpha-1} x^{\alpha-1} e^{-u} du \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty (\lambda x)^{\alpha-1} e^{-u} du \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty (u)^{\alpha-1} e^{-u} du = \frac{1}{\Gamma(\alpha)} \Gamma(\alpha) = 1\end{aligned}$$

□

Normal Random Variable

Let $X \sim N(\mu, \sigma^2)$ prove that f_X is a density function, $\mathbb{E}[X] = \mu$ and Var(X)=σ^2

Proof. The main trick in all these proofs is to *'standardise'* the distribution by applying the change of variable $z = (x - \mu)/\sigma$ note that this implies $dz = dx/\sigma$

To prove that f_X is a density, we integrate over its domain and show that it equates to one. We have

$$\begin{aligned}
\int_{-\infty}^{\infty} f_X(x)dx &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \qquad \text{applying the change of variable } z = (x-\mu)/\sigma \\
&= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz + \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\
&= 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \qquad \text{by symmetry} \\
&= 2\int_{0}^{\infty} \frac{1}{\sqrt{\pi}} e^{-t^2} dt \qquad \text{by defining the change of variable } y = \frac{z}{\sqrt{2}} \\
&= \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} e^{-t^2} dt \\
&= \frac{2}{\sqrt{\pi}} \frac{\Gamma(1/2)}{2} = 1 \qquad \text{we recognise the above integral from the proof of } \Gamma(1/2) = \sqrt{\pi}
\end{aligned}$$

□

Proof. We now prove the expectation of X is given by μ. In doing so, we will continue to use the same trick of standardisation.

$$\begin{aligned}
\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx &= \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} (\sigma z + \mu) \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \qquad \text{applying the change of variable } z = (x-\mu)/\sigma \\
&= \sigma \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\
&= \mu + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{\frac{-z^2}{2}} dz \\
&= \mu + \frac{\sigma}{\sqrt{2\pi}} \left[-e^{\frac{-z^2}{2}} \right]_{z \longrightarrow -\infty}^{z \longrightarrow \infty} = \mu
\end{aligned}$$

We achieve the last line by recognising the anti derivative of $ze^{z^2/2}$ is given by $-e^{-z^2/2}$ and take the subsequent limit to 0 using L'Hoptial's rule.

□

Proof. Now, we advance to proving the variance is given by σ^2. We consider the second moment as follows

$$\begin{aligned}
\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} (\sigma z + \mu)^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \qquad \text{applying the change of variable } z = (x-\mu)/\sigma \\
&= \sigma^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz + 2\sigma\mu \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz + \mu^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\
&= \mu^2 + \sigma^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz
\end{aligned}$$

Where we achieve the last line by recognosing the integral for the expectation of $Z \sim N(0,1)$ random variable, and the integral over the domain of $Z \sim N(0,1)$ random variable. We are now left with the following integral

$$\sigma^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{\frac{-z^2}{2}} dz$$

□

10 Moment Generating Functions

Prove the justification for studying the moment generating functions of random variables $\mathbb{E}[X^n] = M_X^{(n)}(0)$

Proof.

$$\mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(tx)^n}{n!}\right]$$

□

11 Limit Theorems & Famous Inequalities

Chebyshev Inequality

For a positive random variable X and $a > 0$ we show

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

Proof. We start by defining $Y = \mathcal{I}_{X>a}$ and then proceeding with the expectation

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[XY] + \mathbb{E}[((1-Y)X)] \\
&\geq \mathbb{E}[XY] \\
&\geq \mathbb{E}[aY] \qquad \text{as } X \geq a \\
&= a\mathbb{P}(X > a)
\end{aligned}$$

Simple rearrangement yields the desired inequality. □

Markov Inequality

Let X be a random variable with mean μ and variance σ^2. We show that for any $c > 0$

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Proof.

$$\begin{aligned}
\mathbb{P}(|X-\mu| \geq c) &= \mathbb{P}((X-\mu)^2 \geq c^2) \\
&\leq \frac{\mathbb{E}[(X-\mu)^2]}{c^2} \qquad \text{applying Chebyshev inequality} \\
&= \frac{Var(X)}{c^2} \\
&= \frac{\sigma^2}{c^2}
\end{aligned}$$

□

Weak Law of Large Numbers

Let X_i be i.i.d with mean μ and variance σ^2. We show that

$$\lim_{n \longrightarrow \infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = 0$$

Proof. By the Markov inequality

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2}$$

We now easily deduce that

$$\begin{aligned} Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \qquad \text{as } X_i \perp X_j \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Thus we have

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Now using the fact that the probability measure $\mathbb{P}$ is non-negative we have

$$0 \leq \lim_{n\longrightarrow\infty} \mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \lim_{n\longrightarrow\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

□

A (Simplified) Proof of the Central Limit Theorem

Let X_i be i.i.d such that X_i has mean μ and variance σ^2. We define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ and prove that

$$\lim_{n\longrightarrow\infty} \mathbb{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$$

Proof. We define $Z_n = (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ we then create the 'standardised' variable

$$\tilde{X}_i = \frac{X_i - \mu}{\sigma}$$

This results in $\mathbb{E}[\tilde{X}_i] = 0$ and $var(\tilde{X}_i) = 1$ as well as

$$Z_n = \frac{\sum_{i=1}^{n} \tilde{X}_i}{\sqrt{n}}$$

We now proceed to consider the MGF of Z_n as follows

$$\begin{aligned} M_{Z_n}(t) = \mathbb{E}[e^{tZ_n}] &= \mathbb{E}\left[e^{\frac{t}{\sqrt{n}}\sum_{i=1}^{n}\tilde{X}_i}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{n} e^{\frac{t}{\sqrt{n}}\tilde{X}_i}\right] \\ &= \prod_{i=1}^{n} \mathbb{E}[e^{\frac{t}{\sqrt{n}}\tilde{X}_i}] \\ &= (\mathbb{E}[e^{\frac{t}{\sqrt{n}}\tilde{X}_i}])^n \\ &= (M_{\tilde{X}_i}(t/\sqrt{n}))^n \end{aligned}$$

We recall the Maclaurin expansion around x for a function f is

$$f(x) = f(0) + f^{'}(0)x + f^{''}(0)\frac{x^2}{2} + O(h^3)$$

Applying this to $M_{\tilde{X}_i}(t/\sqrt{n})$ we find

$$\begin{aligned} M_{\tilde{X}_i}(t/\sqrt{n}) &= M_{\tilde{X}_i}(0) + M^{'}_{\tilde{X}_i}(0)\frac{t}{\sqrt{n}} + M^{''}_{\tilde{X}_i}(0)\frac{t^2}{2n} \\ &= 1 + \frac{t^2}{2n} \end{aligned}$$

Therefore we have the MGF of Z_n to be

$$M_{Z_n}(t) = \left(1 + \frac{t^2}{2n}\right)^n$$

$$\therefore \quad \lim_{n \longrightarrow \infty} M_{Z_n}(t) = \lim_{n \longrightarrow \infty} \underbrace{\left(1 + \frac{t^2}{2n}\right)^n}_{\text{Definition of } e^a} = e^{\frac{t^2}{2}}$$

and we identify that is the MGF of a standard normal random variable. Hence Z_n approachs a standard normal random variable as $n \longrightarrow \infty$.

One might question the rigour of this proof, in particular that it requires on the technical condition of the MGF holding. This above proof is valid for any *well behaved* i.i.d X_i. All our favourite discrete and continuous distributions fall into this category. For a more rigorous proof (which is reserved for advanced courses in probability), we follow a similar argument though consider the so-called *characteristic function* $\mathbb{E}[e^{\mathbf{t}\mathbf{i}X}]$ as opposed to the MGF.

□

Topic III
Exercises & Solutions

12 Set Theory, Probability Axioms & Probability Measure

MTH2222 Tutorial Exercises & Homework

Tutorial 2 - Homework Question 1

Suppose that A and B are independent. Prove that A^c and B^c are also independent.

Proof.

$$\begin{aligned}\mathbb{P}(A^c \cap B^c) &= 1 - \mathbb{P}(A \cup B) \qquad \text{By DeMorgan Laws}\\ &= 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B))\\ &= 1 - \mathbb{P}(A) - \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B)\\ &= (1 - \mathbb{P}(A))(1 - \mathbb{P}(B))\\ &= \mathbb{P}(A^c)\mathbb{P}(B^c)\end{aligned}$$

□

Tutorial 2 - Question 4

Suppose that A is independent of itself. Prove that $\mathbb{P}(A) = 1$ or $\mathbb{P}(A) = 0$

Proof. Let A be independent of itself. Then we have

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$$

Thus we have

$$\mathbb{P}(A)^2 - \mathbb{P}(A) = 0$$

This is a qaudratic equation with roots $\mathbb{P}(A) = 1$ and $\mathbb{P}(A) = 0$

□

13 Discrete Random Variables

MTH2222 Tutorial Exercises & Homework

Tutorial 3 - Question 1

Fischer and Spassky play a sudden-death chess match whereby the first player to win a game wins the match. Each game is won by Fischer with probability p, by Spassky with probability q, and is a draw with probability $1 - p - q$. Find the probability that Fischer Wins the match, as well as the PMF, expectatoin and variance of the duration of the game.

Solution.

14 Continuous Random Variables

MTH2222 Tutorial Exercises & Homework

Tutorial 5 - Question 2

Find the n^{th} moment of an exponential random variable

Solution.

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

Expanding this recursive expression we find

$$\mathbb{E}[X^n] = \frac{n!}{\lambda^n}$$

Tutorial 6 - Question 2

Find the n^{th} moment of a standard normal random variable

Solution.

$$\begin{aligned}\mathbb{E}[Z^n] &= \int_{-\infty}^{\infty} z^n \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\ &= \int_{-\infty}^{\infty} z^{n-1} \frac{1}{\sqrt{2\pi}} z e^{\frac{-z^2}{2}} dz\end{aligned}$$

Applying integration by parts, noting that

$$\int z e^{\frac{-z^2}{2}} dz = -e^{\frac{-z^2}{2}}$$

We find

$$\begin{aligned}\int_{-\infty}^{\infty} z^{n-1} \frac{1}{\sqrt{2\pi}} z e^{\frac{-z^2}{2}} dz &= \frac{-1}{\sqrt{2\pi}} z^{n-1} e^{\frac{-z^2}{2}} |_{z \longrightarrow -\infty}^{z \longrightarrow \infty} + \int_{-\infty}^{\infty} (n-1) z^{n-2} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\ &= 0 + (n-1) \int_{-\infty}^{\infty} z^{n-2} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\ &= (n-1)\mathbb{E}[Z^{n-2}]\end{aligned}$$

Tutorial 6 - Question 3

Compute $\mathbb{E}[e^{tZ}]$ where Z is the standard normal random variable

Solution.

$$\begin{aligned}\mathbb{E}[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2} + tz} dz \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(z-t)^2}{2}} e^{\frac{t^2}{2}} dz \\ &= e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(z-t)^2}{2}} dz}_{N(t,1)} \\ &= e^{\frac{1}{2}t^2}\end{aligned}$$

Tutorial 6 - Question 6

Dino, the cook, has good days and bad days with equal frequency. On a good day, the time (in hours) it takes Dino to cook a souffle is described by the PDF $f_G(g) = 2$ if $1/2 < g < 2$ and on a bad day, it is goverened by the PDF $f_B(b) = 1$ if $1/2 < b < 3/2$. Find the conditional probability that today was a bad day, given that it took Dino less than three quarters of an hour to cook a souffle

Solution. Let T denote the time taken to cook a souffle. We are looking to find $\mathbb{P}(B|T \leq 3/4)$. To find this we use Bayes' theorem

$$\begin{aligned}\mathbb{P}(B|T \leq 3/4) &= \frac{\mathbb{P}(T \leq 3/4|B)\mathbb{P}(B)}{\mathbb{P}(T \leq 3/4)} \\ &= \frac{\mathbb{P}(T \leq 3/4|B)\mathbb{P}(B)}{\mathbb{P}(T \leq 3/4|B)\mathbb{P}(B) + \mathbb{P}(T \leq 3/4|G)\mathbb{P}(G)}\end{aligned}$$

Using the densities for $f_B(b)$ and $f_G(g)$ we can compute the above probability, along with $\mathbb{P}(G) = \mathbb{P}(B) = 1/2$. Substituting the above values we find $\mathbb{P}(B|T \leq 3/4) = 1/3$

15 Joint Random Variables

Tutorial 7 - Question 1

Suppose $X \sim$ *Exponential*(λ) *and* $\mathbb{E}[Y|X = x] = x^2$ *find* $\mathbb{E}[XY]$.

Solution. By law of total expectation for densities

$$\begin{aligned}\mathbb{E}[XY] = \int_{-\infty}^{\infty} E[XY|X = x] f_X(x)dx &= \int_{-\infty}^{\infty} E[xY|X = x] f_X(x)dx \\ &= \int_{-\infty}^{\infty} xE[Y|X = x] f_X(x)dx \\ &= \int_{-\infty}^{\infty} x^3 f_X(x)dx\end{aligned}$$

Through some fun integration we can find $\mathbb{E}[XY] = 6/\lambda^3$

Tutorial 7 - Question 2

The joint density function is given by $f_{X,Y}(x, y) = \lambda^3(y - x)e^{-\lambda y}$ *for* $0 < x < y$. *Find the marginal density functions.*

Solution. Note the clever change of variables to set up the terminals of integration nicely.

$$\begin{aligned}f_X(x) &= \int_x^{\infty} \lambda^3(y - x)e^{-\lambda y}dy \\ &= \int_0^{\infty} \lambda^3 z e^{-\lambda(z+x)}dz \qquad \text{Using the change of variables } z = y - x \\ &= \lambda^2 e^{-\lambda x} \int_0^{\infty} \lambda z e^{-\lambda z}dz \\ &= \lambda e^{-\lambda x}\end{aligned}$$

We'll omit the solution for finding $f_Y(y)$ as it is very simple integration.

16 Moment Generating Functions

Tutorial 8 - Question 3

The MGFs of two independent discrete random variables X and Y are

$$M_X(t) = \left(\frac{1}{2}e^{2t} + \frac{1}{2}e^{4t}\right)^7 \qquad M_Y(t) = e^{8(e^t-1)}$$

Find $p_X(15)$, $p_Y(5)$, $\mathbb{E}[X]$, $\mathbb{E}[Y^2]$ *and* $\mathbb{P}(X+Y=15)$

17 Limit Theorems

Additional Questions

Expected Rolls to see all sides - Facebook

What is the expected number of rolls to see all sides of a fair 6-sided die?

Let $X :=$ no. of rolls in order to see all sides of a fair die. On our first roll, we guaranteed to see a new face. let X_2 be defined as the no. of rolls in order to see a new face. Clearly $X_2 \sim Geometric(p = \frac{5}{6})$. Defining X_3 in similar fashion and so forth, we obtain

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{6} \mathbb{E}[X_i] \\ &= 1 + \frac{6}{5} + \frac{6}{4} + \cdots + \frac{6}{1} \\ &= 14.7\end{aligned}$$

Expected coin flips for a particular sequence - Two Sigma

What is the expected number of coin flips needed in order to observe the sequence HH?

Define $X =$ no. of flips until we see the sequence HH. We now apply the law of total expectation

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) = \sum_{y} \mathbb{E}(X|Y = y)\mathbb{P}(Y = y)$$

Where the random variable Y Denotes the outcome of the first two flips. We now have

$$\mathbb{E}(X) = \frac{1}{4}(2) + \frac{1}{2}(\mathbb{E}(X) + 1) + \frac{1}{4}(\mathbb{E}(X) + 2)$$

Solving this linear expression for $\mathbb{E}(X)$ we find that

$$\mathbb{E}(X) = 6$$

Topic IV
Mathematical Statistics

18 Further Distributions

Mathematical statistics requires and employs less conventional distributions than that of probability. We list some important ones below, as well as some intuition on their application. We first outline some common terminology used within these distributions and the field of applied statistics

19 Hypothesis Testing

The process of testing whether or not a sample of data supports a particular hypotheses is called hypothesis testing. Generally, hypotheses concern particular properties of interest for a given population, such as its parameters, like μ (for example, the mean conversion rate among a set of users). The steps in testing a hypothesis are as follows:

1. State the null hypothesis H_0 and alternative hypothesis H_1
2. Use a particular test statistic of the null-hypothesis to determine the associated p-value
3. Compare the p-value to the level of significance α

The null hypothesis H_0 is the baseline - it is assumed to be true and we conduct a hypothesis test to determine if the data provides significant enough evidence to reject it in favour of the alternative, H_1.

One Tailed & Two Tailed Tests

Hypothesis tests fall into one of two categories. The first is the **one-tailed**, in which we test the parameter μ under the assumptions that

$$H_0 : \mu = \mu_0$$

and we wish to test either of the following alternative hypothesis'

$$H_1 : \mu > \mu_0 \qquad \text{or} \qquad H_1 : \mu < \mu_0$$

Whereas a **two-tailed** test considers the alternative hypothesis

$$H_1 : \mu \neq \mu_0$$

19.1 Test Statistics

A test statistic is a numerical summary designed for the purpose of determining whether the null hypothesis or the alternative hypothesis should be accepted as correct. More specifically, it assumes that the parameter of interest follows a particular sampling distribution under the null hypothesis.

Z - Test

Z-tests are used when we wish to test the mean of a population μ, the sample size n is sufficiently large and the population variance σ^2 is known. The z-statistic is given by

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

t - Test

t-tests are also used when testing the mean of the population μ. Unlike Z-tests however, we opt to use this test when the sample size is small or large sampling is impractical and the population variance is unknown. The t-statistic is given by

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\hat{p}$ - Test for Population Proportion

20 Parameter Estimation