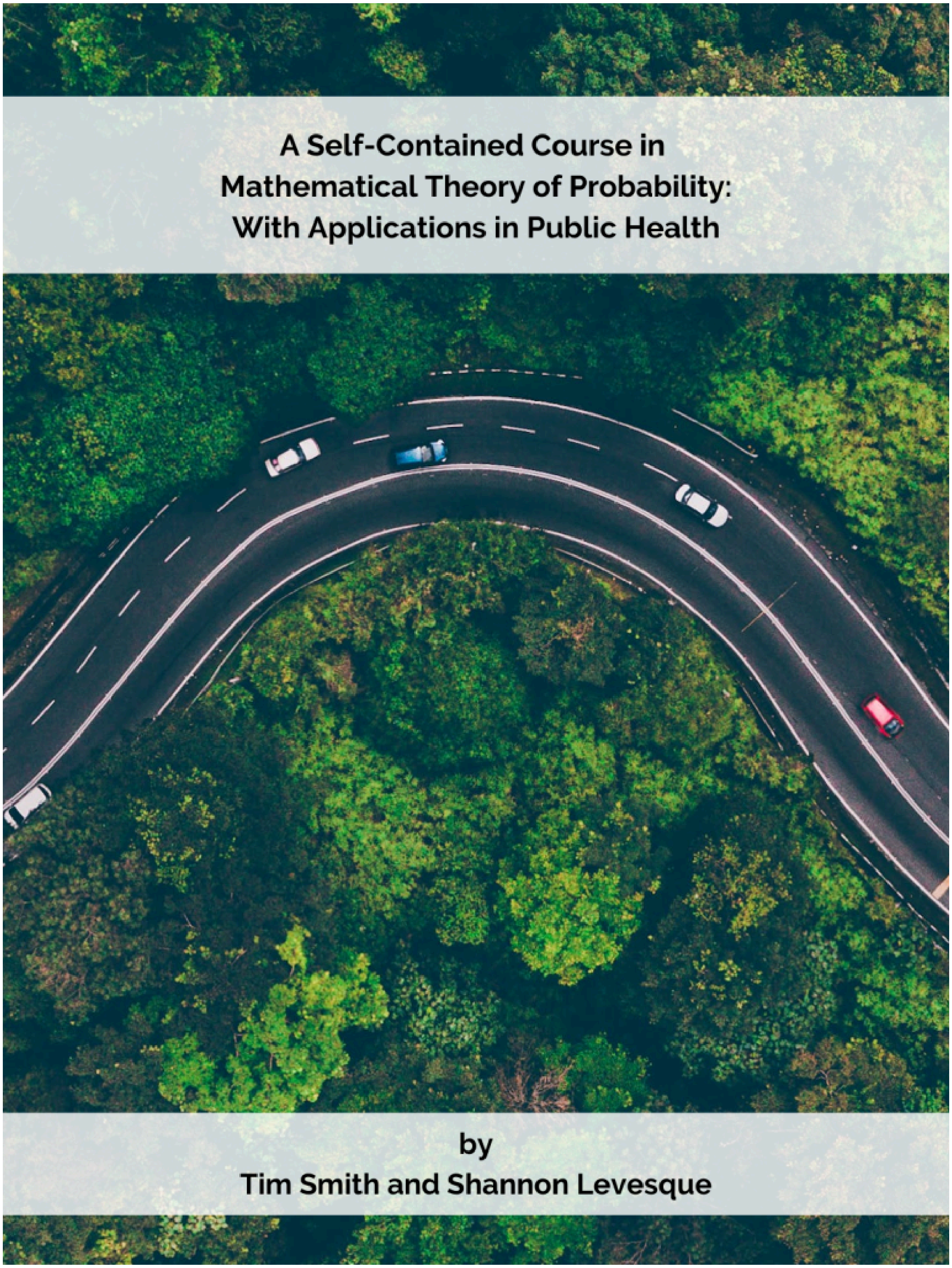
A Self-Contained Course in Mathematical Theory of Probability

A SELF-CONTAINED COURSE IN MATHEMATICAL THEORY OF PROBABILITY: WITH APPLICATIONS IN PUBLIC HEALTH

By Shannon Levesque & Tim Smith © 2022
Embry- Riddle Aeronautical University



A Self-Contained Course in Mathematical Theory of Probability: With Applications in Public Health

by

Tim Smith and Shannon Levesque

Contents

Preface & Acknowledgments

This textbook, which was written during the 2020 SARS2 (AKA COVID-19) lockdown, is designed for a higher-level undergraduate course for engineering or science students who are interested to gain knowledge of the underlying mathematical theory of probability. This textbook was designed from the course notes of a course that the author was teaching at Embry Riddle Aeronautical University during the academic terms of the crisis and became the primary reading material for the quickly adapted course in the new online modality. While there is no statistical prerequisite knowledge required to read this book, due to the fact that the study is designed for the reader to truly understand the underlying theory rather than just learn how to read computer output, it would be best read with some familiarity of elementary statistics. However, the book is self-contained, including the optional chapter zero review of descriptive statistics, and the only true prerequisite knowledge is a solid understanding of university level calculus. The intention for this textbook is for an elective type of course; however, the foundations are laid here for further mathematical study and this text could well serve as a transition for an interested student with little to no prior knowledge to then go on to study in the popular fields of data scientist, big data analysts, genetic algorithm designer or whatever the buzz words of the day may call it.

The author is very grateful for the opportunity to have taught the MA 412 course at his current institution and is very thankful to the many students who made corrections along the way. It is to those students, and the future students who will take MA 412, that this book is dedicated to.

1. Introduction

1.1 REVIEW OF DESCRIPTIVE STATISTICS

When working with a data set, we use the notation x_i for the i^{th} data point in a data set. For example, if we were working with the following data set of airplane's speeds

Data set 1.1.1 (airplane's speeds)

Speed (mph)
500
700
900

we would call $x_1 = 500$ and $x_2 = 700$ etc. Of course, in the "real world" it is common to work with extremely large data sets so it becomes necessary to calculate the descriptive statistics, which allow us to understand the various things a data set is telling us. These descriptive statistics are, generally speaking, divided into two categories: measures of central tendency *or* measures of dispersion. The first category, measures of central tendency, attempts to simply describe the average value or middle of the data set; namely, a few examples of the measures of central tendency are the median and the mean as given in definitions (Definition) 1.1.1 & 1.1.2. The second category, measures of dispersion, attempts to describe how spread out the data set is; namely, a few examples of the measures of dispersion are the range and the variance as given in Defintion 1.1.3 & 1.1.4.

It is common that many resources will attempt to describe a data set by graphical illustration. Although these illustrations are useful, it is essential to remember that as scientists we cannot rely on graphical analyses to draw conclusions. Rather, we require formal analytical mathematical statements. For example we know that ***for a data set to be considered a normal distribution*** the data set must have ***most of the data frequency near the middle*** with a ***symmetrical pattern and the frequency should be less the further away from the middle***. Hence, if a histogram is constructed it should look like this:

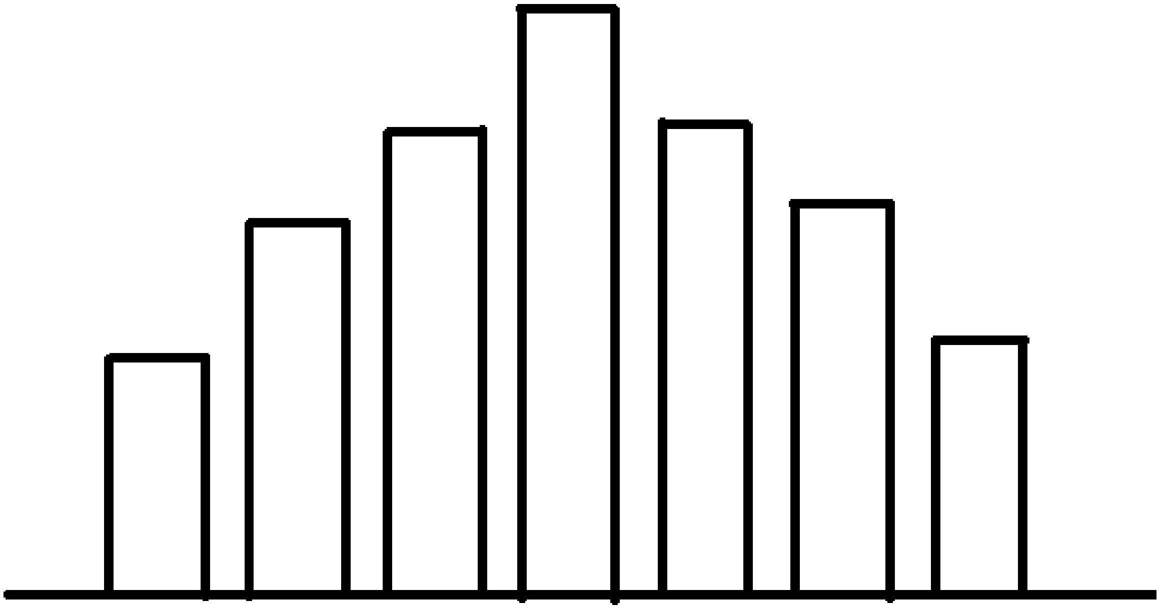


Figure 1. Histogram of a normal distribution

It is essential to understand that this graph alone does not prove nor reject the hypothesis that this distribution is normal. If one wanted to attempt to validate the hypothesis that this distribution is normal, then a formal "test of normality," including a formal computed analytical value to be compared to a formal analytical critical value, would be required. Prior to getting ahead of ourselves, let us summarize a few common descriptive statistics with proper analytical formulas.

Definition 1.1.1 – The *mean* (or arithmetic average) of a data set of n elements

$$\bar{X} = \frac{1}{n}\sum x_i$$

Example 1.1.1

Find the mean of data set 1.1.1

$$\bar{X} = \frac{1}{3}(500 + 700 + 900) = 700$$

Example 1.1.2

Find the mean of data set 1.1.2

Data set 1.1.2 (Car speeds)

Speed (mph)
20
30
40

$$\bar{X} = \frac{1}{3}(20 + 30 + 40) = 30$$

Definition 1.1.2 – The *median* of a data set of n elements

$\tilde{X} =$ the middle value of the data set when ranked (low – high order),
NOTE: if there’s a tie for the middle value, then the median is the average of the two middle values.

Example 1.1.3

Find the median of data set 1.1.1

Firstly, we must rank the data set, which in this case is already ordered, as X_1= 500 & X_2 = 700 & X_3 = 900.

Then, the median is simply found as the middle value. In this case

$$\tilde{X} = 700$$

It is important to note that the measures of central tendency alone do not completely describe the data set under consideration. For example, if we compute the mean & median of both data sets 1.1.3A & 1.1.3B, then we will find the results to be the same as 50. However, it is obvious that the data sets are quite different; namely, the first data set is very clustered together while the second data set is much more spread out.

Data set 1.1.3A

45
47
50
53
55

Data set 1.1.3B

30
35
50
65
70

Thus, we will need to consider measures of dispersion in addition to finding the mean or median. Now, a small value of dispersion would imply that the data set is closely clustered together while a large value of dispersion would mean the data set is more spread out; hence we would expect data set 1.1.3A to have smaller measures of dispersion than data set 1.1.3B. This is indeed true as we will find the variance of 1.1.3A is 17, while the variance of 1.1.3B is 312.50.

Definition 1.1.3 – The *range* of a data set of n elements

The distance between the largest & smallest value of the data set when ranked.

Example 1.1.4

Find the range of data set 1.1.1.

Firstly, we must rank the data set, which in this case is already ordered, as X_1= 500 & X_2 = 700 & X_3 = 900.

Then, the range is simply the largest value minus the smallest value X_3 – X_1 = 900 – 500 = 400.

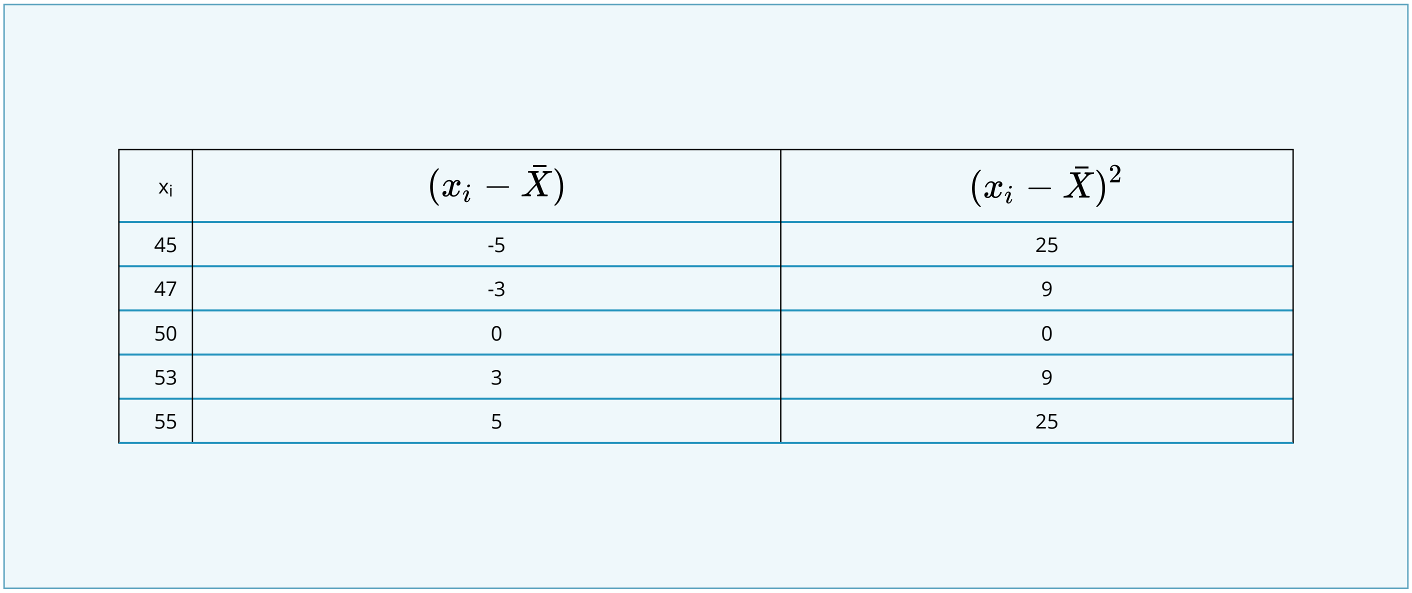
Definition 1.1.4 – The *variance* of a data set of n elements

$$s^2 = \frac{1}{n-1} \sum \left(x_i - \bar{X}\right)^2$$

Example 1.1.5

Find the variance of data set 1.1.3A.

First, we must find the mean, which in this case is 50. Next, it helps to use the following table to simplify the procedure for computing our formula.



x_i	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$
45	-5	25
47	-3	9
50	0	0
53	3	9
55	5	25

If we sum the last column we will find the sum of the squares, $\sum \left(x_i - \bar{X}\right)^2$

which in this example is 68. The variance is then found as this value divided by n-1. In this example we would divide by 4 to find the variance to be 68/4 = 17.

Definition 1.1.5 – The *standard deviation* of a data set of n elements

S = square root of variance

The standard deviation is essentially measuring the same thing as the variance did, however, by taking the square root we are bringing the measure back to the same dimension / units of the original data. For example, if we computed the variance of data set 2 we would find it to be 100. However, this would actually be in units of mph^2 which may not be the most practical in applications, yet the standard deviation would be in units of just mph.

Example 1.1.6

Find the standard deviation of data set 1.1.2.

First, we must find the variance which as noted above is 100 mph^2. Thus, by definition, the standard deviation is the square root of variance = $\sqrt{100mph^2} = 10mph.$

It is also very important to keep in mind "unit bias" when performing data analysis. For example if we were to compare our prior data set of car speeds

Data set 1.1.2 (Car speeds)

Speed (mph)
20
30
40

to our prior data set of plane speeds

Data set 1.1.1 (Aircraft speeds)

Speed (mph)
500
700
900

it should be apparent that a 200 mph difference of speed in one context is quite different than a 200 mph difference of speed in another (have you ever been passed by another vehicle on the freeway going 200 mph faster?). This “unit bias” can be eliminated by transforming the raw data to the “Z scores” which, as the next definition will outline, is done simply by dividing the individual data point’s deviation by the standard deviation. In fact, this standardization will show that the third car, which is going 10 mph above the mean of that data set, has the same “Z score” as the third plane, which is going 200 mph above the mean of that data set. Hence, one can infer that a 10mph deviation in car’s speed is essentially equivalent to a 200 mph deviation in an airplane’s speed.

Definition 1.1.6 – The *Z score* of data point X_i from a data set

$$Z = \frac{X_i - mean}{st\ dev}$$

Example 1.1.7

Compute the Z score for all data points in data sets 1.1.1 & 1.1.2.

x_i	$Z = \frac{x_i - \bar{X}}{s}$	x_i	$Z = \frac{x_i - \bar{X}}{s}$
20	-1	500	-1
30	0	700	0
40	1	900	1

1.2 INTRODUCTION TO CORRELATION & REGRESSION

To conclude this introductory chapter, in this section we will briefly introduce the idea of correlation between two data sets x and y which we will assume both contain an equal number of elements, namely n elements in each data set. The main idea with correlation, or perhaps the main question to ask, is this: is there a pattern between the two data sets? A common mistake that can be made is thinking that the only way to have a correlation between data set x and data set y is that the pattern between x and y must be linear, perhaps y = 2x or y =3x etc. However, this is not correct as there are many other correlation patterns which can occur between two data sets such as quadratic fits or exponential fits. Later on in the textbook we will study the concept of building a predictive model from an x,y data set pair known as a linear regression model, and this tool is one of the most widely applicable statistical models. Of course the linear regression model is a primary purpose of our study and for students of engineering or the sciences this linear predictive data model can be extremely useful. But, it is important to note that linear fits are not the only fit and just because data is correlated does not mean that a linear regression model will work. In mathematical terms one might say that a solid value of correlation is a necessary condition for a linear regression model to work, but it is not a sufficient condition!

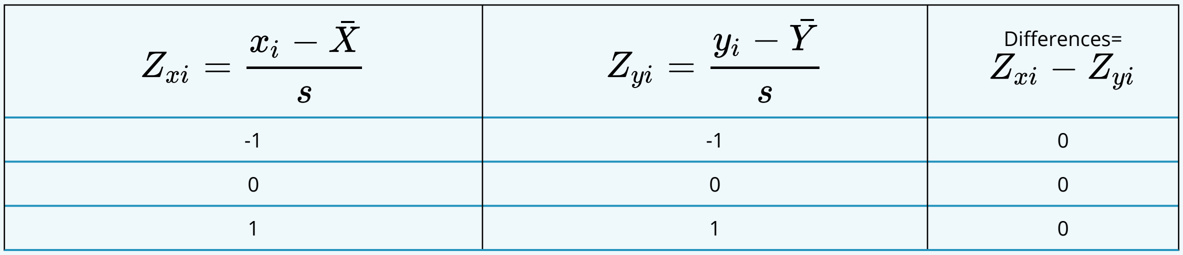
The correlation between two data sets is defined in terms of the deviation between the Z scores of the x data set and the Z scores of the y data set. No deviation between the data set's Z scores is defined as perfect correlation, while an extreme amount of deviation between the data set's Z scores is defined as a low correlation or near zero correlation. For example, if we were to revisit example 1.1.5 and call the data set of car's speeds the x data set and the data set of airplane's speeds the y data set and then compute the differences of those Z scores, we would essentially be studying the correlation pattern between the cars and airplanes. It is worthy to note, prior to working out the details of this example, that in this case we expect a perfect correlation as we have previously discovered that the car's speeds and airplane's speeds go up in a uniform pattern of 1 standard deviation each data point.

Example 1.2.1

Compute the difference between the Z score for data points sets 1.1.1 & 1.1.2.

x_i	$Z_{xi} = \dfrac{x_i - \bar{X}}{s}$	y_i	$Z_{yi} = \dfrac{y_i - \bar{Y}}{s}$
20	-1	500	-1
30	0	700	0
40	1	900	1

To begin we recall from Ex 1.1.7 the Z score which we previously computed and then label accordingly as done above. To complete this example we then simply compute the differences



$Z_{xi} = \dfrac{x_i - \bar{X}}{s}$	$Z_{yi} = \dfrac{y_i - \bar{Y}}{s}$	Differences= $Z_{xi} - Z_{yi}$
-1	-1	0
0	0	0
1	1	0

We observed, as expected, that the total differences are zero which shows a perfect correlation between our data sets.

Now, for a formal definition of correlation we use the following definition which has the interpretation similar to a percent: r near 1 is near perfect correlation (analogous to 100% being near perfect chance) while r near 0 is low correlation (analogous to 0% being near no chance).

Definition 1.2.1 – The ***correlation*** between a data pair set x and y both of n elements

$$r = 1 - \frac{1}{2\,(n-1)} \sum_{i=1}^{n} (Z_{xi} - Z_{yi})^2$$

Example 1.2.2

Compute the correlation for data pairs from data sets 1.1.1 & 1.1.2.

To begin, we recall from Ex 2.1.1 the differences in the Z score which we previously computed and then we must compute the squared differences

$Z_{xi} = \frac{x_i - \bar{X}}{s}$	$Z_{yi} = \frac{y_i - \bar{Y}}{s}$	Differences= $Z_{xi} - Z_{yi}$	$(Z_{xi} - Z_{yi})^2$
-1	-1	0	0^2
0	0	0	0^2
1	1	0	0^2

Now, to complete the problem we utilize the definition

$$r = 1 - \frac{1}{2\,(n-1)} \sum_{i=1}^{n} (Z_{xi} - Z_{yi})^2$$

with n= 3. Doing so this yields the solution

$$r = 1 - \frac{1}{2\,(3-1)} \left(0^2 + 0^2 + 0^2\right) = 1 - 0 = 1$$

We observed, as expected, that the correlation here is a perfect correlation = 1, which again we can informally view analogously to a percent so in an informal sense we can think this data set is 100% correlated.

Example 1.2.3

Compute the correlation for data pairs from the data set 1.2.1 below.

Data set 1.2.1

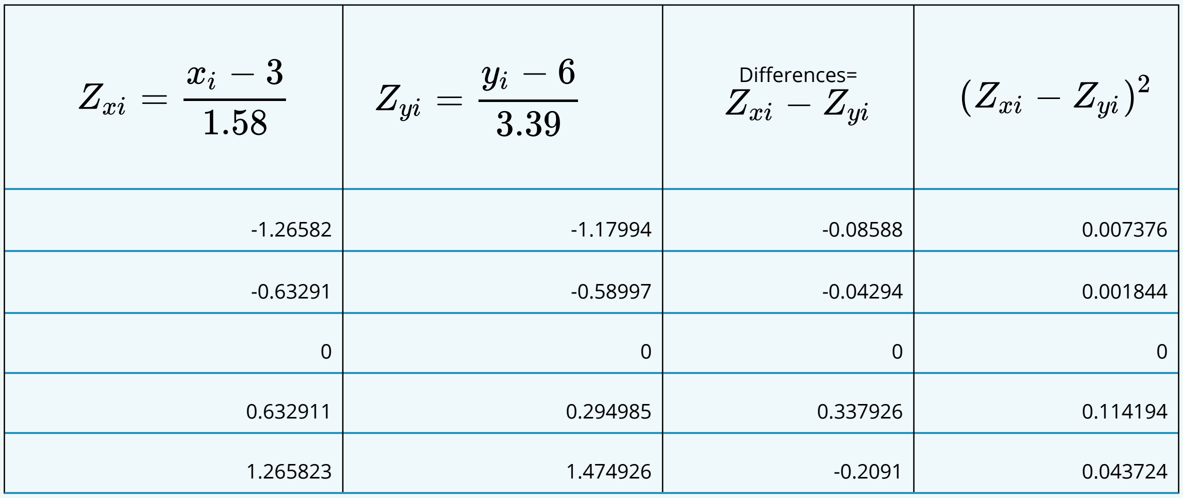
X	Y
1	2
2	4
3	6
4	7
5	11

Where it is given that the mean of x is 3 and y is 6, while the standard deviation of x is 1.58 and of y is 3.39.

To begin, we must compute the Z scores in for x and y separately

$Z_{xi} = \frac{x_i - \bar{X}}{s} = \frac{x_i - 3}{1.58}$	$Z_{yi} = \frac{y_i - \bar{Y}}{s} = \frac{y_i - 6}{3.39}$
$\frac{1-3}{1.58}$	$\frac{2-6}{3.39}$
$\frac{2-3}{1.58}$	$\frac{4-6}{3.39}$
$\frac{3-3}{1.58}$	$\frac{6-6}{3.39}$
$\frac{4-3}{1.58}$	$\frac{7-6}{3.39}$
$\frac{5-3}{1.58}$	$\frac{11-6}{3.39}$

Now, the differences in the Z scores must be computed and then their squares



$Z_{xi} = \frac{x_i - 3}{1.58}$	$Z_{yi} = \frac{y_i - 6}{3.39}$	Differences= $Z_{xi} - Z_{yi}$	$(Z_{xi} - Z_{yi})^2$
-1.26582	-1.17994	-0.08588	0.007376
-0.63291	-0.58997	-0.04294	0.001844
0	0	0	0
0.632911	0.294985	0.337926	0.114194
1.265823	1.474926	-0.2091	0.043724

Finally, to complete the problem we utilize the definition

$$r = 1 - \frac{1}{2\left(n-1\right)} \sum_{i=1}^{n} \left(Z_{xi} - Z_{yi}\right)^2$$

with n= 5. Doing so yields the solution

$$r = 1 - \frac{1}{2\left(5-1\right)} \left(0.007376 + 0.001844 + \ldots\right) = 1 - 0.02 = 0.98$$

We observed, as expected, that the correlation here is a very high correlation = 0.98 as expected because in the data set we can observe the pattern Y being approximately 2x. Again we can informally view analogously to a percent so in an informal sense we can think this data set is 90% correlated.

It is worthy to note that while the prior definition is the theoretically correct and the original definition it is not always the commonly used one. By putting in the definitions of Z scores and performing some algebraic manipulation the following alternate definition for correlation can be obtained and is useful since it is all in terms of values from the data set, hence it is not needed to first compute the Z scores. Also, for mathematical interest the correlation can be written as the covariance of X and Y divided by the products of their standard deviations. Namely, we can write

$$r = \frac{cov\,(x, y)}{s_x\, s_y}$$

Definition 1.2.2 – The *correlation* between a data set x and set y both of n elements

$$r = \frac{\sum((\bar{X} - x_i)(\bar{Y} - y_i))}{\sqrt{\sum(\bar{X} - x_i)^2 \cdot \sum(\bar{Y} - y_i)^2}}$$

The above definition will yield the exact same value as the correlation definition provided in the prior definition 1.2.1, and is actually derived from that prior definition, but this new formula is often preferred when coding formulas as it can be computed directly from raw data as opposed to needing the normalized “z values.”

One of the most useful applications of correlation for an (x,y) pair data set is to build a predictive model to predict the y variable in terms of the x variable as input. Namely, it is desired to create an equation of the form

$\hat{y} = mx + b,$

where the hat notation is utilized to distinguish it as being a predicted value; perhaps a future or forward data point.

Definition 1.2.3 – The *linear regression line* of a data set x and set y

$\hat{y} = mx + \beta,$

where

$$m = r\left(\frac{s_y}{s_z}\right)$$

and

$\beta = \bar{y} - m\bar{x}$

Example 1.2.4

Compute the linear regression line for data pairs from the data set 1.2.1.

To begin, we recall from the prior solution that

$s_y = 1.52, s_x = 0.71, \bar{y} = 6, \bar{x} = 3 \; and \; r = 0.9$

Hence, we can compute

$m = r\left(\frac{s_y}{s_z}\right) = 0.9\left(\frac{1.52}{0.71}\right) = 1.93$

and

$\beta = \bar{y} - m\bar{x} = 6 - 1.93 * 3 = 0.21$

Which yields our solution as the predictive model of our linear regression line

$\hat{y} = 1.93x + 0.21.$

The linear regression line has far reaching applications in various fields such as engineering or science and finance applications. For example, our data ended at the value of x being 5 so one could use the linear regression line to expand beyond that, perhaps to find the predicted y value associated with a future x of 6 as

$\hat{y}\left(6\right) = 1.93\left(6\right) + 0.21 = 11.79.$

For another example, our data set contained only integer values of x being 1 then x being 2 etc., and one could use the linear regression line to fill in between that, perhaps to find the predicted y value associated for a half way value x of 1.5 as

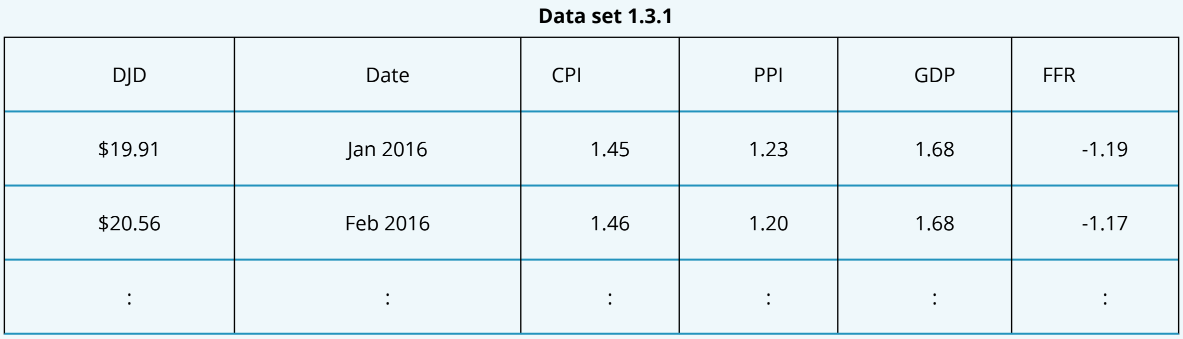
$\hat{y}\left(1.5\right) = 1.93\left(1.5\right) + 0.21 = 3.11.$

There are many other applications, and of course restrictions to, the linear regression line and this will be a central theme of the later chapters of this textbook. However, prior to continuing with our development it is necessary to overview, or perhaps receive for the informed reader, some key principles from the mathematical theory of probability which are contained in Chapter 2. Due to the fact that this text is designed as a self-contained resource, these principals in Chapter 2 are developed “from the ground up” and the informed reader may be able to jump forward at this point. Any reader who has the knowledge equivalent to a Junior or Senior year university level course in mathematical statistics and/or introductory probability theory can most likely jump to Chapter 3. Regardless of that progression, it is important to close this Chapter with one essential principal regarding regression: while it is logical that it only makes sense to use a linear regression model for a data set which is highly correlated, that does not ensure that the linear regression model will work or be statistically valid. Namely, it is vital to understand, as previously stated, that from a mathematical point of view one can say that a solid value of correlation is a necessary condition for a linear regression model to work, but it is not a sufficient condition!

1.3 A BRIEF COMMENT ABOUT MULTIVARIABLE DATA

While the purpose of this textbook is for engineering students to obtain the knowledge needed to understand the underlying mathematical theory of probability, as opposed to an introduction to the many methods of inferential statistics & data analysis, it is worthy to include here a quick discussion about how to deal with data sets of more than two variables. Namely, one of the most common questions that arises is: “now that we have the definition of correlation between x & y, how do we generalize that to a data set that has more than two variables, for example x & y & z?” And, the answer to this question is that generally we do not have a “multi correlation,” but in the following example a method is outlined as to how such a measure can be computed.

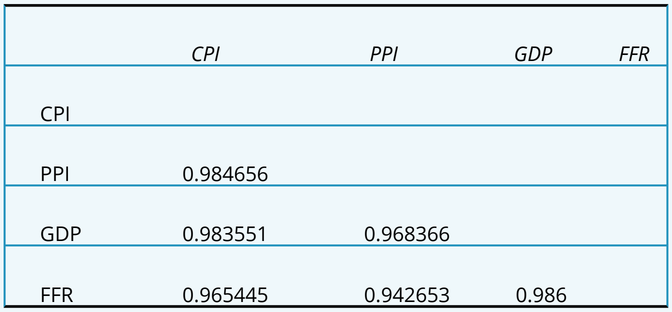
The data set below was used in a research study to predict the value of an exchange traded fund called DJD, which is a very popular investment instrument that is designed to track the famous United States DOW JONES index; however, this index only includes a subset of the index of companies that have either maintained or raised their dividends over the last year, hence it is often preferred by investors looking for a somewhat conservative and safe investment. Now, the goal of the research study was to determine a statistical model that utilized macroeconomic predictor variables to make a mathematical prediction of the fair value of DJD. A data set of monthly values from 2016 to 2109 was obtained, and a subset of that data is listed below for illustration where the predictor variables have been normalized. For those interested in the details the values used here were: Consumer Price Index (AKA “CPI” or the government’s measure of consumer inflation), Producer Price Index (AKA “PPI” or the government’s measure of product cost inflation), Gross Domestic Product (AKA “GDP”), and the Federal Funds Rate (AKA “FFR”), with all data being publicly available freely from official government websites.



Data set 1.3.1

DJD	Date	CPI	PPI	GDP	FFR
$19.91	Jan 2016	1.45	1.23	1.68	-1.19
$20.56	Feb 2016	1.46	1.20	1.68	-1.17
:	:	:	:	:	:

Now, in the analysis one of the worst things that can be done when creating such models is to have high amounts of inter correlation within the predictor variables; this applies regardless if the model being used is regular regression like was done in this research or more advanced modern models such as machine learning methods (of course technically regression is the nicest example of supervised machine learning). Often a first investigative step is to run a correlation matrix, which will yield results that look something like



	CPI	*PPI*	*GDP*	*FFR*
CPI				
PPI	0.984656			
GDP	0.983551	0.968366		
FFR	0.965445	0.942653	0.986	

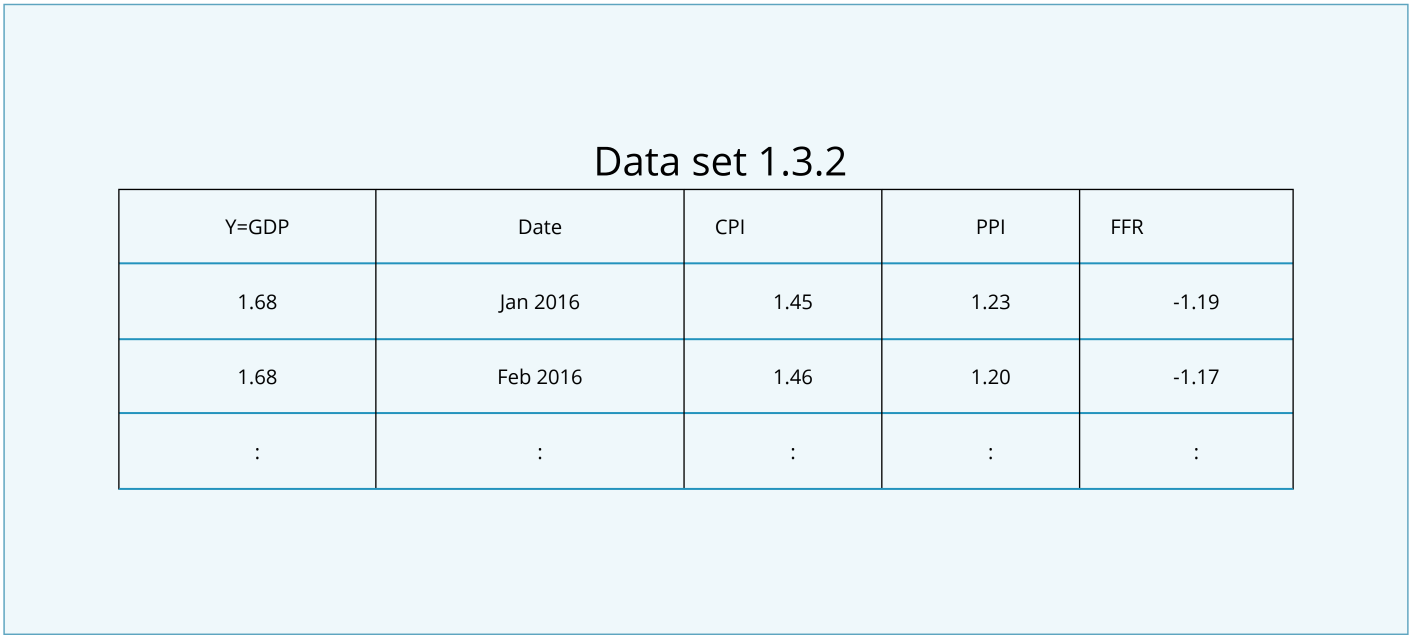
The issue with this procedure is that it is not really a multi variable correlation, rather it is a list of multiple individual correlations. For example, the value 0.984656 is the correlation between CPI & PPI in isolation, while the value of 0.968366 is the correlation between PPI & GDP in isolation etc.

The correlation matrix does provide information, but it does not really answer the question of how multi correlated these variables are. For example, the largest correlation between any two variables is seen between FFR and GDP, as that pair has the largest value of correlation being 0.986. However, when this was computed the computation was done in isolated. Namely, when calculating this the same correlation formula from our prior section

$$r = 1 - \frac{1}{2\,(n-1)} \sum_{i=1}^{n} \left(Z_{xi} - Z_{yi}\right)^2$$

was used with the x variable being FFR and the y variable being GDP, but none of the other variables were considered in the computation (i.e. it was done in isolation). Thus, this measure does not really define how one of the variables, say GDP, is correlated to all of the other variables. One informal resolution to this problem is to add up all of the individual correlations, e.g. 0.986 + 0.968366 + 0.983551. Doing so we would obtain the total for GDP to be 2.937917. Then similarly the total for CPI to be 2.933651, the total for PPI to be 2.895676 and the total for FFR to be 2.894098. While this does provide the information that GDP does appear to have the worst inter correlation, with CPI not so far behind, this is not exactly a mathematically proper definition.

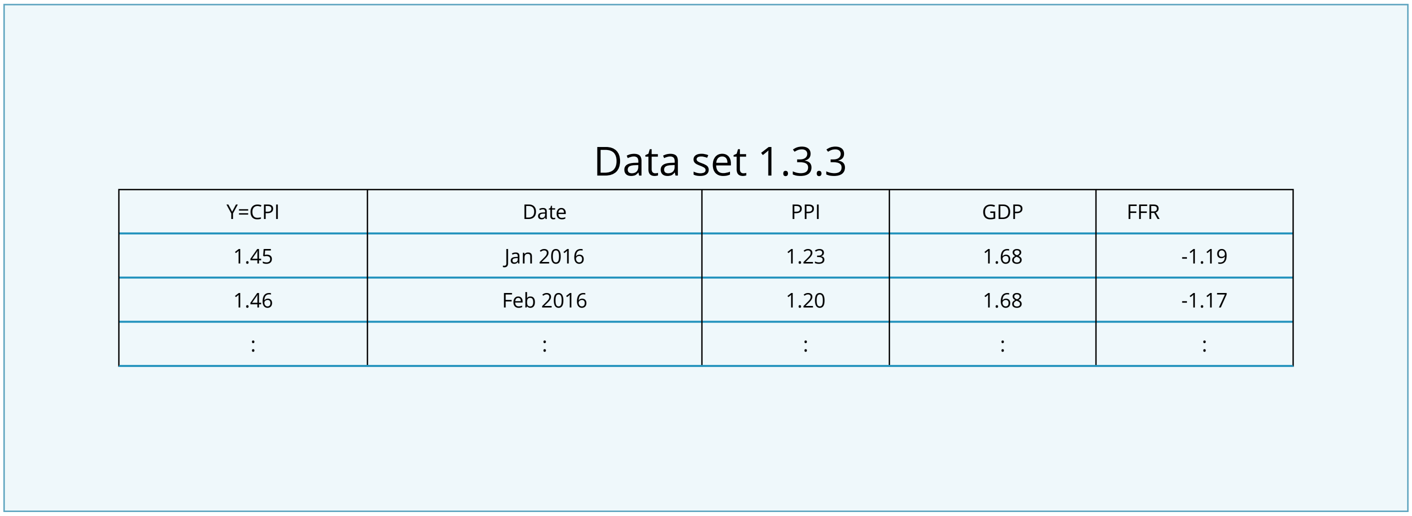
In order to measure how correlated one variable is related to multiple other predictor variables the following procedure is often applied. Firstly, take the variable of concern and move it into the column of the response variable; for example if we were to do this for GDP our data set would now look like



Data set 1.3.2

Y=GDP	Date	CPI	PPI	FFR
1.68	Jan 2016	1.45	1.23	-1.19
1.68	Feb 2016	1.46	1.20	-1.17
:	:	:	:	:

or if we were to do this for CPI our data set would now look like



Data set 1.3.3

Y=CPI	Date	PPI	GDP	FFR
1.45	Jan 2016	1.23	1.68	-1.19
1.46	Feb 2016	1.20	1.68	-1.17
:	:	:	:	:

Now, the next step is to run a multivariable regression model with the remaining predictors being used as normal, but the Y variable not being the original data desired but instead replaced by the predictor variable under consideration. Then after running this multivariable regression model from the output the value of the so called coefficient of determination (AKA R squared) should be noted. For example doing so for data set 1.32, which has GDP as the response, above yields R^2 = 0.987513087343, so this is essentially a measurement of how GDP is totally correlated (technically predicted by) the predictor variables CPI, PPI and FFR. Likewise, doing so for data set 1.33, which has CPI as the response, above yields R^2 = 0.984192914938173, so this is essentially a measurement of how GDP is totally correlated (again, technically predicted by) the predictor variables PPI, GDP and FFR.

Definition 1.3.1 – The *variance inflation factor (VIF)* for a predictor variable X_i from a set of predictor variables X_1,X_2,...

$$VIF = \frac{1}{1 - R_i^2}$$

where the

$$R_i^2$$

is the coefficient of determination obtained from the multi variable regression model predicting variable X_i in terms of the remaining predictor variables.

Example 1.3.1

Compute the VIF for both GDP and CPI from data set 1.3.1, and using the previously obtained information.

To solve this it is needed to obtained the coefficients of determination, but those have already been provided. Thus, we can quickly compute

$$VIF\,(GDP) = \frac{1}{1 - \ 0.987513087343} = 80.08$$

and

$$VIF\,(CPI) = \frac{1}{1 - 0.984192914938173} = 63.26$$

Now, while these computations are very interesting and provide information, the big question is do they really answer the question as to how “multi correlated” one variable is to all of the others, and the answer is sort of! The coefficient of determination does tell us what percentage of the variance in the variable under consideration is explained by a regression model created from the other variables, and while this is not exactly a correlation it does provide similar information. However, it is important to remember the main question in applications is not really to see how correlated one variable is to the others, but will putting both of them into the model cause an issue? The general rule of thumb is if a variable has a VIF > 10 then it would be advised to proceed with caution as to using the variable, and if a variable has a VIF closer to 100 then it absolutely should not be included. Moreover, what is interesting to observe is how the variance inflation factor sniffs out minor details. Namely, in this example we found, using the not so formal method of adding up all the individual correlations, that CPI and GDP were about the same, 2.933651 and 2.937917 respectively. However, the VIF method is much more accurate as it shows that GDP had a much higher value, 80 compared to 60, thus it provides a clear solution that of these variables GDP is the most correlated, formally what is referred to as multicolinearity.

It is important for the reader of this text, who again should be more focused on understanding the mathematical theory of probability, that the big take away is that while we do not really have a definition of multi correlation, this method of VIF is an important tool which can be used to narrow down a data set; hence, avoiding redundancy within the data set which can be an issue if one predictor variable is extremely correlated to many of the other predictor variables. Lastly, to close this chapter the formal definition of the coefficient of determination is provided for mathematical completeness, but no examples as those are generally discussed in courses such as inferential statistics or regression analysis.

Definition 1.3.2 – The ***coefficient of determination (*** R^2***)*** for a regression model

$$R^2 = 1 - FUV$$

where the term FUV is known as the fraction of unexplained variance, which can be computed in terms of the known data value,

$$y_i$$

along with the predicted data value,

$$\hat{y}_i$$

along with the mean as

$$\frac{\sum\left(y_i - \widehat{y_i}\right)^2}{\sum\left(y_i - \bar{y}\right)^2}$$

Chapter 1 Exercises

1. Compute the variances for the following two data sets:
    - Data Set 1: 5, 15, 20, 25, 35
    - Data Set 2: 15, 17, 20, 23, 25
2. Using the results from the previous problem, describe what variance tells us about the data set.
3. If a normally distributed data set has a mean of 75 and a standard deviation of 3, find the interval that contains 95% of the data.
4. If a normally distributed data set has a mean of 100 and a variance of 25, find the maximum value to be considered in a normal range (99%).

2. Definition of Probability

2.1 INTRODUCTION FROM EXAMPLES

When studying probability theory, it is very important to consider the perspective we have when investigating problems. As engineers or scientists, it is expected to have solution values that predict exactly when or exactly where some event will occur, i.e. deterministic solutions. However, in probability we do not have such solutions or problems; rather, we define the likelihood of outcomes. This begins with how we define our variables; namely we define a **random variable** (RV) as a number whose value depends on the outcome of a random experiment. The key point here is that the outcome of the experiment are random and not deterministic. A good example is the lottery: the odds say it is extremely unlikely to win but that does not mean you will not win. We do not know the outcome until the experiment of the lottery numbers being drawn is conducted.

There are, generally speaking, two “kinds” of variables: discrete variables and continuous variables. One of the simplest illustrations to demonstrate the difference between these two kinds of variables can be illustrated from a typical classroom situation; namely, the number of students in the class is a discrete variable while the time of the class is a continuous variable. A **discrete variable** is one that is finite and countable. For example, no matter how large the class is the number of students is countable! We also notice that the number of students is a finite discrete value, identified by a positive integer, as you can think when new students enter the class there is either 1 student or 2 students, but no in-between value such as a half of a student. On the other hand, a **continuous variable** is one that is infinite. For example, time is an infinite continuum. A student who is studying theoretical physics will be very interested to dialog about the matter of time as a variable and observable measurements. However, for simplification of the idea let us just look at one interesting property of continuous real numbers. There is a famous mathematical axiom that states between any two real numbers there is always at least one more value. Hence, any interval on the real number line contains an infinite number of values. Now, this idea can be illustrated by just considering two moments in time. Let us have a starting time of t=1 second and an ending time of t=2 seconds, and in doing so we see that there is a halfway point:

$$= \tfrac{1}{2}(1\ +\ 2\ ) = 1.5$$

Repeating this process using the original starting time of t=1 second but a new ending time of t=1.5 seconds we see that there is a new halfway point:

$$= \tfrac{1}{2}(1\ +\ 1.5\ ) = 1.25$$

Repeating this process once more using again the original starting time of t=1 second but a new ending time of t = 1.25 seconds we see that there is a new halfway point:

$$= \tfrac{1}{2}(1\ + 1.\ 25\ ) = 1.125$$

As you can see, this process could go on indefinitely, hence proving that between any two values of a continuous variable there are infinitely many points.

The prior result is very interesting from a pure mathematical number theoretical point of view alone, but it also yields one very interesting probability result for us to take note of: the probability of our RV being any one single value is exactly equal to zero. While this will be developed more formally later on, we can see the idea as follows using the classical definition of probability when it is applied to the sample space being any interval from the real number line:

$$P\left(A\right) = \frac{size\ of\ \ (AKA\ number\ of\ elements\ in)\ \ sample\ A}{size\ of\ sample\ space\ \Omega} = \frac{1}{\infty} = 0.$$

2.2 THEORETICAL VS EXPERIMENTAL DEFINITION

Prior to beginning our formalization of probability, we must first summarize some key terminology.

Definition 2.2.1 – A ***simple event*** of an experiment under consideration in an application of probability

a single outcome of the experiment under consideration.

Definition 2.2.2 – An *event* of an experiment under consideration in an application of probability

a collection of one or more *simple events.*

For example, if you were considering the experiment of drawing a card from a deck of 52 cards, a simple event would be the ace of spades from this deck of cards, as that is a single outcome (card). However, an event, which is a collection of one of more simple events, could be drawing an ace card, as the outcome of an ace technically consist of four simple events.

Definition 2.2.3 – The *sample space* of an experiment under consideration in an application of probability

All possible outcomes. The symbol � is often utilized to identify the sample space.

Now that we have defined the events and sample space, we can proceed to formalize our definition of probability!

Definition 2.2.4 – The *theoretical definition of probability* related to an experiment under consideration

$$P\left(event\right) = \frac{number\ of\ favorable\ outcomes}{total\ number\ of\ outcomes\ in\ sample\ space}$$

Which is often written as

$$P\left(event\right) = \frac{number\ of\ simple\ events\ in\ E}{size\ of\ \Omega}$$

For example, if we were playing a game of cards and wanted to find the probability of drawing an ace on a random trial, we would identify that there are 4 simple events in E, and the size of the sample space is 52. Hence, we can find

$$P\left(event\right) = \frac{number\ of\ simple\ events\ in\ E}{size\ of\ \Omega} = \frac{4}{52} = 0.0769..\quad \left(7\%\right).$$

It is very important to note here that we should not round up the solution, often a common practice is to cut, not round, solutions at the second decimal place and report the solution as a whole probability. Moreover, it may be desirable to keep more accuracy, perhaps 7.6% or 7.69% etc, but we should never round the solution up! While rounding 7.69% up to 7.7% (or perhaps rounding 7.6% up to 8%) may seem like an insignificant detail, in some real world applications such details can have serious consequences! Thus, for consistency in this text we will always follow the rule of “being a conservative statistician,” and always report our solutions as whole percentages, which are cut, not rounded, at the second decimal from our numerical results obtained.

Definition 2.2.5 – The ***empirical (or experimental) definition of probability*** related to an experiment under consideration

$$P\left(event\right) = \frac{number\ of\ sucess}{numer\ of\ trials}$$

Which is often written as

$$P\left(event\right) = \frac{number\ of\ simple\ events\ in\ E}{size\ of\ \Omega}$$

For example, if we were playing a game of cards that had four players which exhausted the deck, hence each player got 13 cards, and we got the hand: ace of hearts, jack of spades, 9 of hearts, 8 of clubs, 7 of spades, 6 of hearts, 5 of clubs, 4 of spades, 3 of hearts, 2 of clubs, king of spades, queen of spades, and the ace of spades. Then we could compute the probability of getting an ace as

$$P\left(event\right) = \frac{number\ of\ simple\ events\ in\ E}{size\ of\ \Omega} = \frac{2}{13} = 0.153.. \quad \left(15\%\right).$$

Now, at this point students often find these results a bit confusing. A common question that arises is “why are they not the same?”. The general answer is the theoretical probability tells you what should happen, and the empirical probability tells you what did occur. Usually they are closer, but the bottom line is we cannot predict the future and there is always something that is left to chance; hence the reason why you often see people buying the lottery at the gas station. While the probability of actually winning is minuscule, there is still a chance. The point here is that the outcome of each event is random! The theoretical definition of probability tells us that for every 52 cards we should get 4 aces, i.e. for every 13 cards we should get 1. However, that does not tell us this will happen, but rather it just defines a likelihood of it occurring. The truth here – and the important point which separates probability theory from applications of deterministic mathematical models such as differential equations – is that outcomes of experiments under consideration in probability theory are random: a person could play this game of cards all day and never actually get an ace but another person could play this game of cards once and get a great hand like the one outlined here with two aces. However, there is one important “fine print” to keep in mind: the law of large numbers states that as the number of trials gets larger, the empirical definition will be approximately the same as theoretical definition. For example, if we draw 13 cards and get 2 aces, it is a big deal that we got one extra ace than expected. However, if we draw 1300 cards and get 101 aces, it is NOT a big deal that we got one extra ace than expected.

In general, we will use the theoretical definition of probability when working out application problems here. In general, the results are often presented in a “PDF” probability chart. For example, the outcomes of a single card drawn could be presented as

X	P(X)
Ace	4/52 ≈ 0.07
Other Face Card (King, Queen or Jack)	12/52 ≈ 0.23
Other Card (regular number card)	36/52 ≈ 0.69

It is very important to note that not every probability chart is valid. In order for a probability chart (or the function used to create the probabilities) to be valid, it must satisfy the following:

Definition 2.2.6 – Valid Probability Function Requirements

- All P(X) values must be valid probability values: 0 < P < 1
- The sum of all P(X) values must be 100%: ∑P= 1

It is worthy to note that in many applications the condition (ii) will not exactly be one, but it should be extremely close. For example, looking at our probability function for the game of cards the sum would be ≈1 as 0.07 + 0.23 + 0.69 = 0.99, but this is of course just due to truncating.

2.3 MATHEMATICAL FORMALISMS

Suppose that $\Omega = \{e_1, e_2, \ldots, e_N\}$

is a finite sample space. Then the probability of an event E is the sum of the probabilities of all of the simple events contained in E. In symbols, this yields the definition

Definition 2.3.1 – The ***theoretical definition of probability*** related to an experiment under consideration

$$P\left(E\right) = \sum_{e_k \ in \ E} \quad P_k$$

Now, it is interesting to notice that if all the events

e_k

are equally likely, then every $P_k$is the same, namely $P_k = \frac{1}{N}$.

Thus, in this case the definition becomes

$$\sum_{e_k \ in \ E} \quad P_k = \sum_{e_k \ in \ E} \quad \frac{k}{N}$$

which, by calling k the number of events, yield

$$P\left(E\right) = \frac{number \ of \ simple \ events \ in \ E}{N}.$$

It is interesting to compare this definition to our prior definition in the last section with the main difference being this result only applies if all of the events are equally likely to occur. Moreover, on the latter definition, if we notice that N is the same as the size of the sample space, then the definition 2.3.1 is exactly the same as those from the prior section. However, if the events are not equally likely, then only the first definition can be applied, which will be illustrated in the next two examples.

Example 2.3.1

Use proper mathematical notation to both compute and set up the corresponding formulas needed to find the probability of drawing a heart card and separately a club card into a two card hand where each draw is taken from a separate 52 card deck, hence we can assume independence.

To solve this, we need to first define the events. Let us call the first event, the event of drawing a heart card, e_1. Doing so we can identify p_1 to be 1/4 which is equivalent to 13/52 or the number of heart cards divided by N. Likewise, we call the second event, the event of drawing a heart card, e_2. Doing so we can identify p_2 to be 1/4 which is equivalent to 13/52 or the number of heart cards divided by N. Now, the formal definition yields

$$\sum_{e_k \ in \ E} P_k = P_1 + P_2 = \frac{1}{4} + \frac{1}{4} = 0.5 \ \ (50\%)$$

Alternatively, this could have been computed using the prior definition

$$P\left(event\right) = \frac{number \ of \ hears + number \ of \ clubs}{size \ of \ \Omega} = \frac{26}{52} = 0.5 \ \ (50\%)\,.$$

Example 2.3.2

Use proper mathematical notation to both compute and set up the corresponding formulas needed to find the probability of drawing a blackjack on the second card, assuming you are holding a King, and then separately winning in a single roll on a 38 slot roulette wheel.

To solve this, we must first define the events. Let us call the first event, the event of winning the blackjack, e_1.This would actually be to draw an Ace from the remaining 51 cards, hence we can identify p_1 to be 4/51. Now, we call the second event, the event of winning on the roulette wheel, e_2. Doing so we can identify p_2 to be 1/38 since the only way to win is if the ball falls into the single slot chosen. Now, the formal definition yields

$$\sum_{e_k\ in\ E} P_k = P_1 + P_2 = \frac{4}{51} + \frac{1}{38} \approx 0.1\ \ (10\%)$$

This cannot be rewritten using the prior definition, as the two events are not equal likely; moreover, it is important to note that what is computed here is not the “and probability”. The interpretation of this result is just a probability sum of both probabilities and it is not representing the probability that someone would win both games in sequence. Moreover, while we will not be covering the definitions in full detail, nor discussing examples, the following formulas can be used to calculate the probability of sequential events given that the probability of the first event, which we will call A, is known in addition to the probability of the second event, which we will call B.

Definition 2.3.2 – *Addition Rule* and *Multiplication Rule*

The ***addition and multiplication*** rules state that the probability of event A or event B occurring is found to be

$$P\,(A\ or\ B) = P\,(A) + P\,(B) - P\,(A\ and\ B)$$

while, under the assumption of independence, the probability of event A and event B in sequence is found to be

$$P\,(A\ and\ B) = P\,(A) \cdot P\,(B)$$

It is worthy to note that in some examples the union & intersection notations are utilized, hence the probability of event A or event B occurring is often rewritten as

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

likewise the probability of event A and event B in sequence, under the assumption that the events are independent, is written as

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

Also, it is often common to see the notation

$$P\,(A')$$

which reference to the complement A, e.g

$$P\,(A') = 1 - P\,(A)$$

For example, if event A is drawing an ace from a standard deck of cards then

$$P\,(A') = 1 - P\,(A) = 1 - \frac{4}{52} = \frac{48}{52}$$

which is the probability of drawing any card other than an ace from a standard deck of cards.

Now, at this point we will not dive into the details, it is worthy to note that the formula given above for the "AND" probability is only valid under the assumption that event A is independent of event B, e.g. the outcome of one of the events does not have any impact on the outcome of the other. This is not always the case. For example, consider the case of drawing two cards from a deck of 52 cards. Let's call the first draw event A and the second draw event B. Say we wanted to find the probability that the second card was a King, it is reasonable to conclude that

$$P\left(B\right)=\frac{4}{51}$$

since there would only be 51 cards remaining, but this would be on the assumption that we knew the first draw was not a king. If the first draw was a king then we would conclude that

$$P\left(B\right)=\frac{3}{51}$$

The situation under consideration here is referred to as conditional probability, and the truth is that it is not really practical to define the probability of B until we know what happened on the first draw. However, we define the conditional probability

$$P\left(B\mid A\right)$$

as the probability of B occurring given that A already did. Moreover, in the case of dependent probabilities we redefine the multiplication rule to be

$$P\left(A\ and\ B\right)=P\left(A\right)\cdot P\left(B\mid A\right)$$

This of course reverts to the prior rule if event B is truly an independent event, as in that case the conditional probability would just be the same as

$$P\left(B\right)$$

due to the fact that, in this case event B would not have any dependence on event A.

It is common to see the dependent conditional probabilities formula rewritten as

$$P\ \left(B\mid A\right)=\frac{P\ \left(A\ and\ B\right)}{P\left(A\right)}$$

as often in applications, the conditional probability value is obtained by counting outcomes, which would have the same format as the right hand side here. Moreover, an interesting result, known as Bayes' theorem, comes if we begin by thinking the labels for A and B are just labels. Now, by reversing them the above could be rewritten as

$$P\ (A \mid B) = \frac{P\ (B\ and\ A)}{P\,(B)}$$

It is logical to conclude that the probability of A and B is exactly the same as the probability of B and A, so solving the first equation we find

$$P(A\ and\ B) = P(A) \cdot P(B|A)$$

Likewise, solving the second equation we find

$$P(B\ and\ A) = P(B) \cdot P(A|B)$$

And, equating the two then solving for the conditional probability $P\ (A \mid B)$, we obtain the famous Baye's theorem

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

The above formula has many applications, especially in business, as it allows one to find a conditional probability in another direction given the other. For example, consider the situation where event A is the event that the future economy will be good, so the market will go up, while event B is the event that an economist gives a good forecast. The conditional probability $P\,(B \mid A)$ would be a probability that could be obtained from prior data, namely the probability that the economist gave a good forecast when the economy was good. However, the conditional probability $P\,(A \mid B)$ would not be possible to obtain as nobody knows what the future of the economy will be, but the trick is by using Bayes' theorem one can approximate this value!

Chapter 2 Exercises

1. Jamie is joining a movie club. As part of her introductory package, she can choose from 12 action selections, 10 comedy selections, 7 fantasy selections and 5 horror selections. If Jamie chooses one selection from each category, how many ways can she choose her introductory package?
2. How many different four-letter secret codes can be formed if the first letter must be an A or B?

3. In a contest in which 15 contestants are entered, in how many ways can the 4 *distinct* prizes be awarded? (Meaning there is a different prize for 1st, 2nd, 3rd, and 4th.)
4. For the following problems, consider a group of 50 students. There are 8 Computer Engineering (CE) majors, 12 Computer Science (CS) majors, 20 Electrical Engineering (EE) majors, and 10 Software Engineering (SE) majors. There are no dual major students.
    - The department chair will pay for 16 students to go to a conference. In how many ways can the 16 students be selected if exactly 4 are selected from each major?
    - 8 of the students are lined up from left to right. In how many ways can this be done when we consider their individual names, not their majors?
    - 8 of the students are lined up from left to right. In how many ways can this be done if we consider only their majors, and not their names?
5. Amy, Jean, Keith, Tom, Susan, and Dave have all been invited to a birthday party. They arrive randomly and each person arrives at a different time. Find the probability that Jean will arrive first and Keith will arrive last.
6. A committee consisting of 6 people is to be selected from eight parents and four teachers. Find the probability that the selected group will consist of all parents.
7. You are dealt one card from a 52-card deck. Find the probability that you are NOT dealt a jack.
8. The physics department of a college has 15 male professors, 11 female professors, 7 male teaching assistants, and 5 female teaching assistants. If a person is selected at random from the group, find the probability that the selected person is a teaching assistant or a female.
9. A card is drawn from a 52-card deck and a fair coin is flipped. What is the probability of drawing a heart and flipping heads?
10. There are 45 chocolates in a box, all identically shaped. There are 16 filled with nuts, 15 with caramel, and 14 are solid chocolate. You randomly select one piece, eat it, and then select a second piece. Find the probability of selecting 2 solid chocolates in a row.
11. Numbered disks are placed in a box and one disk is selected at random. If there are 8 red disks numbered 1 through 8, and 2 yellow disks numbered 9 through 10, find the prob-

ability of selecting a red disk, given that an even-numbered disk is selected.

12. The two-way frequency table below shows the preference of sports to watch among males and females of a sample of 150 people.

	Hockey (H)	Basketball (B)	Tennis (T)	Total
Male (M)	41	23	15	79
Female (F)	10	16	45	71
Total	51	39	60	150

Find the following probabilities. Write each answer as a simplified fraction:

- $\mathrm{P}\left(\mathrm{T}\right) \ =$
- $\mathrm{P}\left(\mathrm{F}\right) \ =$
- $\mathrm{P}\left(\mathrm{F} \cap \mathrm{T}\right) =$
- $\mathrm{P}\left(\mathrm{F} \cup \mathrm{T}\right) =>$

3. Overview of Discrete Random Variables

3.1 DEFINITION OF A RANDOM VARIABLE

In deterministic scientific theory a variable, commonly x for location or t for time, is used to make a prediction through a mathematical model. For example, using newton's law one can solve a differential equation which will tell you the exact location of a rocket launched into space from Cape Kennedy after, t=1 minute to t=2 minutes etc. The location will be exactly determined from the solution, based, on initial conditions of the system. There is no error in the measurements. This is classical scientific theory, and how it works. On the other hand, statistical analysis is a bit different. For example, if a pitcher in a baseball game throws a fastball every time, but it is a little different each time: sometimes the pitcher throws it as possible, other times just not quite as fast or other times puts a spin on it which causes it to sink. For the batter the pitch would not exactly known from any solution, rather it would be a bit random, or what one might call a random outcome of a statistical experiment. Each time the batter stands at the plate, the pitch coming is random. Sure it might be one of a known set – fast fastball, not quite as fast fastball, slow sinking fastball – but each event is random. Interestingly these events should be random of each other, just because the last pitch was a fast fastball doesn't mean the next one won't be.

In statistical analysis we define a random variable, RV, to be a mathematical formalization of an outcome of a statistical experiment which depends on random events. The value of x is commonly used, and if the corresponding probability density is defined on real numbers then

$x \epsilon\ R.$

3.2 DISCRETE PROBABILITY DISTRIBUTIONS & EXAMPLES

It is common for a random variable x, which has n possible outcomes

$x_1, x_2, x_3, \cdots$

that have the corresponding probabilities

$p_1, p_2, p_3, \cdots$

e.g .

$p(x_i) = p_i$

to be presented in a chart as

$$\begin{matrix} x_1 & p_1 \\ x_2 & p_2 \\ : & : \end{matrix}$$

Definition 3.2.1 – The ***Expected Value*** of a probability distribution chart

The expected value of a probability distribution chart

\begin{matrix}x_1&p_1\\x_2&p_2\\:&:\\\end{matrix}

is defined to be

$$E(x) = \sum_{i=1}^{n} x_i \bullet p_i$$

If the outcomes of an experiment, which is the purchasing of a lottery ticket, which has only three outcomes:

$x_1 = loss\ of\ the\ \$10\ paid,\ x_2 = winnings\ of\ \$100\ AKA\ \$90\ profit$

or

$x_3 = winnings\ of\ \$100\ AKA\ \$990\ profit$

with the corresponding probabilities chart as

$$\begin{matrix} -10 & 0.95 \\ 90 & 0.04999 \\ 990 & 0.00001 \end{matrix}$$

Example 3.2.1

Find the expected value of the

$$\begin{matrix}-10 & 0.95\\ 90 & 0.04999\\ 990 & 0.00001\end{matrix}$$

and interpret what this result can tell about the experiment.

The above formula

Definition 3.2.2 – The *variance* of a probability distribution chart

The variance of a probability distribution chart
\begin{matrix}x_1&p_1\\x_2&p_2\\:&:\\\end{matrix}
is defined to be

$$VAR = \sum_{i=1}^{n} (x_i - \mu)^2 \bullet p_i$$

where μ is the numerical result obtained as the expected value.

Rather than discuss many repetitive examples of this, let us now consider examples of one of the most useful discrete probability distributions, the binomial.

Definition 3.2.3 – The *binomial distribution function*

$$\left(\frac{n!}{r! \bullet (n-r)!}\right) \bullet\ p^x \bullet (1-p)^{n-x}$$

where the random variable considered in the experiment has only two possible outcome, a success with associated probability = p, or a failure with associated probability = 1-p. Moreover, the experiment under consideration is repeated n times, with the trails being truly independent so that the result of the last trial has no effect on the result nor probabilities for the current trial.

$$VAR = \sum_{i=1}^{n} (x_i - \mu)^2 \bullet p_i$$

where μ is the numerical result obtained as the expected value.

It is worthy to note here that the factorials term out front, often called nCx or “n choose x,” is often done in a separate computation, for example using an online calculator, so it is more common to see the binomial written as

$nCx \bullet\ p^x \bullet (1-p)^{n-x}$

Example 3.2.1

Use a binomial probability distribution to find the probability of getting 7 answers correct from 10 total questions on a multiple choice test where each question has four choices, e.g. the probability of a correct guess is 1 out of 4,

$p = \frac{1}{4} = 0.25\ (25\%)$.

Now, the solution here would be obtained from the binomial

$nCx \bullet\ p^x \bullet (1-p)^{n-x}$

plugging in p as 0.25 and n = 10 which yields our density function

${}_{10}C_x \bullet 0.25^x \bullet (0.75)^{10-x}$.

The desired solution is obtained by plugging in x as 3 which yields, noting that $_{10}C_3$ is found to be 120 from the calculator, the solution

${}_{10}C_3 \bullet 0.25^7 \bullet (0.75)^3 \approx 0.003$

or 0.3%.

It is worthy to note that the solution of the prior example, 0.3%, tells us the probability to get exactly 7 right from 10 guess. If passing the test is defined as getting seven or more right, our solution is not the probability of passing. Rather to find such a value we would need to first use the formula again to find the probability of getting 8 right, p_8, and then find the probability of getting 9 right, p_9, and then probability of getting them all right, p_{10}, hence

$P\,(win) = p_7 p_8 p_9 p_{10}$

It is worthy to note that in practice one would not prefer to perform this calculation, and since it is possible to approximate our binomial with a regular normal, having mean =μ, and variance= σ^2, the same solution there could be computed as $P(7 < x < 10)$ using the normal density.

Definition 3.2.4 – The *mean*, μ, and the *variance*, σ^2 , of the binomial distribution function

$nCx \bullet\ p^x \bullet (1-p)^{n-x}$

are:

$\mu = np$

and

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

Now, while the emphasis of this text is on continuous probability distributions, which will be introduced in the next chapter, and most lecture examples commonly used for discrete probability distribution functions utilize the binomial, due to is wide range of applications, it is important to understand that it is not the only discrete probability function. Moreover, there are many other discrete probability functions and once the logic of the process is understood all that is needed to work with a new discrete probability function is the function's expression along with its interpretation. For example, the Poisson distribution, which expresses the probability a given number of events occurring in a fixed interval of time, provided that these events occur with a known constant mean, λ , and are independently of the time since the last event has the probability density function

$$P\left(X\right) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Once this function is defined for the probability experiment under consideration, then the remaining computations are logically the same as those outlined in the prior examples using the binomial. For example, if it is historically known that house on the intercostal river floods once every 50 years on average, then lambda would be 1 and we could set up our function as

$$P\left(X\right) = \frac{1^x e^{-1}}{x!}$$

Then, we could use it to compute the probability of no floods in the next 50 years, i.e. put x as zero, to be

P\left({X = 0} \right) = \frac{{{1^0}{e^{ - 1}}}}{{0!}} \approx 37\%

Likewise if we knew at our local airport historically a flight, which flew daily Monday through Friday, arrived more than 15 minutes late 2 times out of the week, then lambda would be 2 and we could set up our function as

$$P\left(X\right) = \frac{2^{x}e^{-2}}{x!}$$

which we could use compute various probabilities. In these applications, along with many other probability applications, it is very important to understand the implications of the phrase “are independently of the time since,” which is basically saying that each day is a new day. A good example is the river example, let us say that the river flooded last year and ponder the question if that has any affect on the likelihood of it flooding this year. While our common sense may make us think, well if it flooded last year then it most likely will not flood this year, this does not agree with what probability tells us. Using the Poisson probability density from our river example, and plugging in x as two, i.e. finding the probability of two floods in fifty years, we find the probability to be

$$P\left(X = 2\right) = \frac{1^{2}e^{-1}}{2!} \approx 18\%$$

This tells us that there is an eighteen percent chance that this river will flood again in the next forty-nine years, but it does not tell us anything about when this will occur. It is equal likely to occur this year as it is occur next year or the following year, and so forth. While this concept may not agree with our common sense, it is how probability works when we have the assumption of independence. Of course, not all probability problems have the assumption of independence and there is a procedure called conditional probability that addresses problems where the likelihood of the next event occurring does depend on results of prior outcomes.

Chapter 3 Exercises

1. Sarah is looking to buy a larger home for her family. She is only going to consider homes that have 3 or more bedrooms and more than 2500 square feet.

    - Is the number of bedrooms in houses that she considers a discrete or continuous random variable?
    - Is the square footage of houses she considers a discrete or continuous variable?

3. For a finite (discrete) random variable, state the two requirements for p_k to be a valid probability distribution.
4. For an infinite (continuous) random variable, state the two requirements for f(x) to be a valid probability density function (PDF).
5. Complete a probability distribution for the following scenarios and determine if it is a valid probability distribution
    - $P\left(x=1\right)=40$, $P\left(x=2\right)=10$, $P\left(x=3\right)=30$, $P\left(x=5\right)=20$
    - $P\left(x=0\right)=30$, $P\left(x=1\right)=20$, $P\left(x=2\right)=40$, $P\left(x=3\right)=20$
6. Consider the following probability distribution:

x	0	1	2	3
p(x)	.12	.38	.4	?

Assuming 0, 1, 2, and 3 are all the possible values of x, find p(3).

- What value of x is most probable?
- P(x < 1 or x ≥ 2) = _______
- P(x > 0) = _______

1. A bank branch collected data from customers regarding the number of credit cards they have. The probability distribution is displayed below.

x	0	1	2	3	4
p(x)	.13	.39	.28	.15	.05

Find the following (round to the nearest hundredth):

- $\mu =$__________
- $\sigma =$__________

2. On a ten question multiple choice test, you must get at least 7 questions correct to pass. Each question has five possibilities; hence, the probability of a correct guess is 20%.

   - If you guessed on all of the questions, what is the probability that you got exactly 7 correct?
   - What is the expected number of correct answers if you guessed on all 10 questions?

3. Determine if the following are appropriate binomial experiments. If so, solve using MatLab or binomial formulas. If not, explain why it is not a binomial experiment.

   - A plane is landing at LGA and there is a 75% chance that it will land on time. If every day of the week, this plane flies the same route and the weather, air traffic, etc is the same everyday and I fly one day a month for a year, what is the probability that I am on time at least 10 times.
   - I drive home 40 miles everyday and some days it rains and some days it does not. On the days it does not rain, the probability I make it on time is 80%. If next week I work 4 days and it only rained once what is the probability that I make it home on time everyday.

4. Stephen Curry was unanimously voted the MVP of the NBA for the 2015-2016 season. He has one of the highest free throw percentages at about 91%. So let's say the probability of

Curry making a free throw shot is 91%. Consider if Curry attempted 80 free throws. Let x represent the number of free throws made.

- Can the probability distribution of x be approximated by the binomial distribution?
- $E(x) =$______________________
  $\sigma^2 =$________________________
  $\sigma =$________________________
- Find the probability that Curry makes exactly 75 of the free throws
- Find the probability that Curry makes at least 60 of the free throws
- Find the probability that Steph Curry makes less than 10 of the free throws

4. Introduction Continuous Probability Theory

4.1 CONTINUOUS RANDOM VARIABLES

In the following section we will define one of the most important topics in the mathematical theory of probability, the continuous probability density function. It is from this density function that many results such as probability solutions and expected values will be derived. However, prior to doing so, it is important to note that for simplification through the remainder of this text, which is designed to for a first course in probability theory, we will only be considering examples of independent continuous random variables, hence all computations will involve single variable functions and their corresponding calculus computations

Definition 4.1.1 – Continuous random variables

A random variable X is a continuous random variable if the variable is defined on a scale as a regular "continuous" numerical values, e.g. X is a real number. Recall we define X to be a random variable if it is a mathematical formalization of a outcome which depends on random events.

At this level we will not attempt to develop the derivation or underlying motivation for our probability density functions; rather, we will define a density function f(x) for a random variable X to be the function that creates the probability as

$$P\left(A < x \leq B\right) = \int_{A}^{B} f\left(x\right) dx.$$

Now, this probability density function must meet two basic properties, which are in line with the axioms of probability:

Definition 4.1.2 – Requirements to be a density of a random variable X

- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)\, dx = 1.$

NOTE: if you have an example of a function desired to be used for a density that meets criteria (i) but not (ii) then you can create a valid density by the normalization process (similar to that of normalizing a vector) by dividing the constant $K = \int_{-\infty}^{+\infty} f(x)\, dx$, as one will find the function

$$\frac{1}{K} f(x)$$

will be a valid density function.

Example 4.1.1

Find the normalized density for a density of the form

e^{-Ax}

defined for x > 0 and zero elsewhere.

To begin we note that we must satisfy properties (i) and (ii) of definition 4.12. Now, it is first observed that property (i) is met because the exponential function is a strictly positive function. However, property (ii) is not met because

$\int_{-\infty}^{+\infty} f(x)\, dx = 1/A.$

Thus, using the logic from above we take the constant

$K = \frac{1}{A}$

to create the normalized density to be the so called *exponential density*

$f(x) = Ae^{-Ax}.$

Example 4.1.2

Find the value of C so that the function, which is defined as
$C \bullet\ X \bullet (1 - X)\ \ if\, 0 < X < 1$
or 0 otherwise, will be a valid density.

To begin we note that we must satisfy properties (i) and (ii) of definition 4.12. Now, it is first observed that property (i) is met because this parabolic function will be above the x axis, with zeros at x=1 and x=0, provided that the value of C is positive. Now, property (ii) is not met until we specify the value of C, thus we compute

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{0} 0dX\ + \int_{0}^{1} C \bullet X \bullet (1 - X)\,dX + \int_{1}^{\infty} 0\,dX = C\left(\frac{1}{2} - \frac{1}{3}\right) = \frac{C}{6}.$$

Now, in order to make this a valid density we must choose C=6.

Example 4.1.3

Verify that the *uniform density*

$f(x) = \dfrac{1}{R - L}\ \ if\ \ L < X < R$

or 0 otherwise, is a valid probability density function.

To begin we note that we must satisfy properties (i) and (ii) of definition 4.12. Now, it is observed that property (i) is met because the function is a constant positive value. In verifying property (ii), we obtain

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{\infty}^{L} 0dx + \int_{L}^{R} \frac{1}{R - L}dx + \int_{R}^{\infty} 0dx = \frac{R - L}{R - L} = 1.$$

Example 4.1.4

Verify that the *standard normal density*

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

is a valid probability density function.

To begin we note that we must satisfy properties (i) and (ii) of definition 4.12. Now, it is first observed that property (i) is met as the exponential function is a strictly positive function. However, we must verify property (ii), and in doing so we obtain

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}\,dx$$

This is a very difficult integral to do in closed form, but one can compute numerically and verify that this integral is indeed equal to 1, hence the provided density is a valid probability density function!

4.2 COMMONLY UTILIZED CONTINUOUS DENSITY FUNCTIONS

It is worthy to take note of a few of these common density functions as they will frequently be used in examples as we move forward and have many common real world applications! The following are the most likely examples that you will encounter are:

The ***exponential density*** is

$$f(x) = Ae^{-Ax}$$

which is defined for x > 0,

and the ***uniform density*** is

$$f(x) = \frac{1}{R-L}$$

which is defined for L < x < R,

where both of these densities serve useful for textbook illustrative examples due to the fact that the resulting integrals turn out to be doable without the need for complicated integration techniques.

The ***normal density*** is

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

which is defined for all x, and this density is by far one of the most applicable in real world modeling applications.

The ***standard normal density*** is

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

which is defined for all x, and is a special case of the normal density with mean μ=0 along with variance �=1, serves as the backbone of many theoretical mathematical statistical results such as the famous central limit theorem.

The *T density* is

$$f\left(x\right)=\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^{2}}{v}\right)^{-\frac{v+1}{2}}$$

which is defined for x > 0 with v being the degrees of freedom. This density is utilized in applications as an approximation for the normal density when some of the information of the population mean, μ, or variance, �^{2} , is unknown.

The ***chi squared density*** is

$$f\left(x\right)=\frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}x^{\frac{v}{2}-1}e^{-\frac{x}{2}}$$

which is defined for x > 0, with v being the degrees of freedom. This density is utilized in applications for error analysis when considering the sum of squares error and/or Goodness of fit error analysis.

The ***Fdensity*** is

$$f\left(x\right)=\frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{x\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}\sqrt{\frac{\left(d_1x\right)^{d_1}d_2^{d_2}}{\left(d_1x+d_2\right)^{d_1+d_2}}}$$

which is defined for x > 0, where d_1 and d_2 are the degrees of freedom, numerator and denominator respectively. This density is related to a ratio of two chi squared densities and is very useful in a great deal of applications. especially the analysis of linear regression.

The ***logistic density*** is

$$f\left(x\right)=\frac{e^{-(x-\mu)/s}}{s(1+e^{-\frac{x-\mu}{s}})^{2}}$$

which is defined for x > 0, with s representing the scale not standard deviation as one might expect. This density is very useful in the analysis of regression when applied to case when the response variable in the form of a categorical “1/0” variable (AKA logistic regression).

Some more generalized “abstract” examples are:

The ***Beta density*** is

$$f\left(x\right)=\frac{\Gamma\left(a+b\right)}{\Gamma\left(a\right)\Gamma\left(b\right)}x^{a-1}\left(1-x\right)^{b-1}$$

which is defined for x > 0.

The factor � is the so called "Gamma function," which normalizes the density. There is

a formal definition of this function valid for any values of n, but for our purposes it will suffice to use the definition:

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

for integer values

$\Gamma\left(\frac{n}{2}\right)=\sqrt{\pi}\frac{\left(n-2\right)!!}{2^{\frac{n-1}{2}}}$

for halves using odd n; note

$\left(n-2\right)$!!=(n-2)·(n-4)⋯

Now, we have several probability density functions let us look at some examples

Example 4.1.5

For the *exponential density*

$f\left(x\right)=e^{-x}$

Find the probability P(0<x<5).

To begin we know the density is as given above so we just need the probably integral

$P\left(0<x<5\right)=\displaystyle\int_0^5 f\left(x\right)dx$

Thus, we compute

$$\int_0^5 e^{-x}dx=\left[-e^{-x}\right]_{x=0}^{x=5}=1-e^{-5}\approx 0.993$$

Hence, we have computed the probability P (0<x<5) = 99.3%.

Example 4.1.6

For the *uniform density*

with R = 10 and L = 0 find the probability P (0<x<2).

To begin we note that our density will be

$f\left(x\right)=\frac{1}{10}$

and the probability integral will be

$$P\left(0<x<2\right)=\int_{0}^{2}f\left(x\right)dx.$$

Thus, we compute

$$\int_{0}^{2}\frac{1}{10}dx=\left[\frac{x}{10}\right]_{x=0}^{x=2}=0.2$$

Hence, we have computed the probability P(0<x<2) = 20%.

Example 4.1.7

For the ***standard normal density***

$$f\left(x\right)=\frac{1}{\sqrt{2\pi}}e^{\frac{-x^{2}}{2}}$$

find the probability P(0<x<2).

To begin we know the density is as given above so we just need the probably integral

$$P\left(0<x<2\right)=\int_{0}^{2}f\left(x\right)dx.$$

Thus, we compute

$$\int_{0}^{2}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^{2}}{2}}dx.$$

However, this integral, which is ultimately an integral of the form , is not solvable in closed form so numerical approximations will be required (which yield the solution of approximately 47%) . In the next chapter we will further discuss how to work with normal density, as it is one of the most important densities if not the most important, and we will look at some applications of nice function in MATLAB. For now, we will move on with developing further properties of probability distribution theory, namely the expected value and variance.

In the following section we will define two extremely useful properties of statistics the expectation and the variance. Generally speaking one can view these in an analogous manner as the expected value and variance are interpreted in elementary data analysis. Namely, the expectation (AKA expected value) can be viewed as the average value or “what we expect to get on average,” which is frequently just called the mean and often the symbol μ is utilized. And, the variance can be viewed as a measure of dispersion or "how spread out is the data," which is often notated by the symbol �^{2}. For simplification we will define, for a random variable x, the expectation as E(x) and moving forward write all expressions, definitions and so forth in terms of E(x) as not only is it good practice for consistency, but it is also the proper and formal way to define things!

4.3 EXPECTATION AND VARIANCE

Definition 4.3.1 – The ***expectation*** of a continuous random variable X with density f(x)

$$E\left(x\right)=\int_{\Omega}x\bullet f\left(x\right)dx$$

At this time we will focus on solving examples and address interpretations along with theoretical implications for later studies. However, it is good for the reader to understand the solution obtained is an expected value and not a probability, i.e. it does not have to be within the usual range of 0 to 1 rather the answer can be viewed as just a number!

Example 4.3.1

Find the expectation for the ***Standard Normal density***

$$f\left(x\right)=\frac{1}{\sqrt{2\pi}}e^{-0.5x^{2}}$$

To begin we recall the above definition of the expectation is .

$$E\left(x\right)=\int_{\Omega}x\bullet f\left(x\right)dx$$

and we compute

$$E\left(x\right)=\int_{-\infty}^{\infty}x\bullet\frac{1}{\sqrt{2\pi}}e^{-0.5x^{2}}dx=\left[-\frac{1}{\sqrt{2\pi}}e^{-0.5x^{2}}\right]_{-\infty}^{\infty}=0.$$

Example 4.3.2

Find the expectation for the particular case of the ***Beta density***

$6X \bullet (1 - X)\ \ if\, 0 < X < 1$, or 0 otherwise.

To begin we recall the above definition of the expectation is

$$E(x) = \int_{\Omega} x \bullet f(x)\, dx$$

and we compute

$$E(x) = \int_{0}^{1} X \bullet (6X \bullet (1 - X))\, dX = \frac{1}{2}.$$

Definition 4.3.2 – The ***variance*** of a continuous random variable X with density f(x)

$$VAR(x) = \int_{\Omega} (x - \mu)^2 \bullet f(x)\, dx$$

where the symbol μ is representing the value of the expectation for the density, as often the expectation is interpreted as a mean or average value. Again, at this time we will quickly observe solving an example and leave interpretations along with theoretical implications for later studies.

Example 4.3.3

Find the variance for the *Standard Normal density*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}$$

To begin we recall the above definition of the variance is

$$VAR(x) = \int_{\Omega} (x - \mu)^2 \bullet f(x)\, dx$$

Thus, we compute

$$VAR(x) = \int_{-\infty}^{\infty} (x - 0)^2 \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}\, dx = 1.$$

For the purpose of this textbook study, the preceding definitions will suffice to cover all forthcoming needed mathematical theory. However, we will close this

section with the following definitions and results as they can be very useful in applications to actually compute the expectation and/or variance for a density without

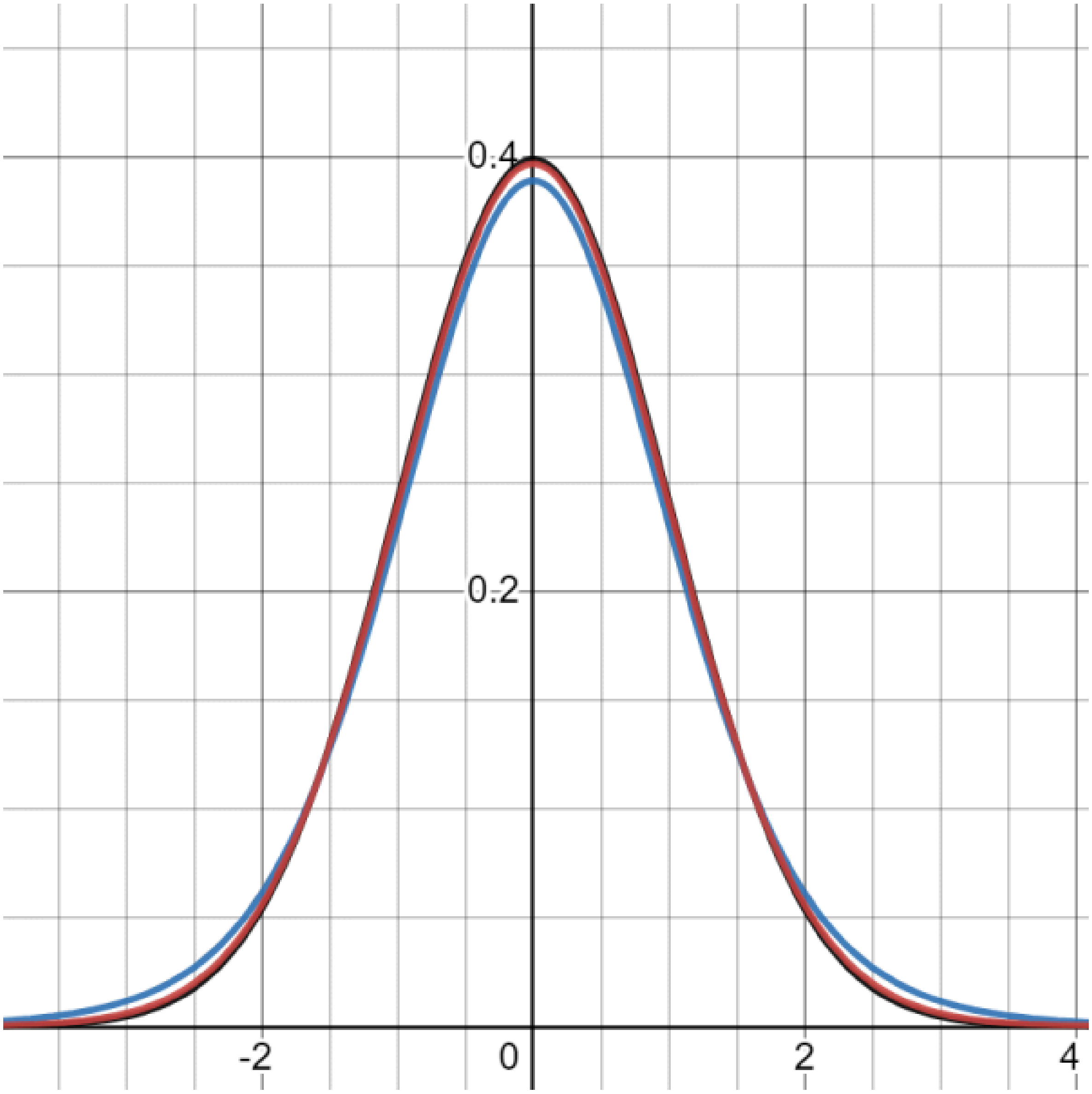
conducting the integrals from the prior definitions.

4.4 NORMAL DENSITY AND ITS SMALL SAMPLE APPROXIMATION

To begin our illustration we will actually steal the results from a couple of example forthcoming in the next chapter (examples 5.12 & 5.1.3), and while we will not yet go into the details of the calculation, for the result called a P value, it is interesting to note that the solution yielded the results of 0.0062 if the normal density was utilized or 0.0027 is the T density was used with v=5 degrees of freedom. Moreover, if we increase the degrees of freedom, say to 50 then 100, we obtain the results of this P value to be 0.008 then 0.007 respectively. Thus, one can conclude that as the degrees of freedom for the T distribution get larger the result gets closer to the normal distribution! Of course the question here often comes as to what exactly are degrees of freedom? The exact answer to this question depends on the exact application and/or statistical experiment under consideration, but in most cases one can think that the degrees of freedom are about the same as, or very closely related to, the sample size of the data set being used in the statistical study. For example, in the classical hypothesis testing, which is one of the most commonly applied techniques, if the T density is used as the density the degrees of will actually be one less than the sample size of the data set, e.g. v= n-1. Thus, it is often comment to think of the T density as a small sample size approximation of the normal density. Moreover, one can actually prove that

$$\lim_{n\to\infty}\left(\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}\right)=\frac{1}{\sqrt{2\pi}}e^{-0.5x^2}$$

This result can be viewed visually below in the graphs, where the BLUE is the T density with 5 degrees of freedom, and the RED is the T density with 50 while the OTHER is the normal distribution.



An interactive version of this graph is available at desmos.com by selecting the image.

It is worthy to note that in most applications it is not practical to actually find the population mean nor variance, hence the values needed for the true normal density; thus, in practice most applications will use the T density as the model. The T density actually provides results that are a little more conservative, which is always a good thing to do when doing statistical analysis!

4.5 EXAMPLES WITH APPLICATION TO ERROR ANALYSIS

To begin our first application we will assume an experiment has been done – either building a statistical model such as regression or a regular experiment such as trying a new method to make a part for an airplane – and it is desired to conduct analysis on the mean squared error, e.g. on the term

$(y_i - \hat{y})^2$

where the regular y represents the data and the $\hat{y}$ is used for either the approximated value from the model or the desired value obtained from the underlying engineering scientific theory. Moreover, if one can assume that each of the

y_i

terms are normally distributed, and each

y_i

is independent of all of the others, then the sum of their squares would be a chi-squared distribution

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\left(\frac{k}{2}\right)-1} e^{-\frac{x}{2}}$$

where in this example the notation k, which is referred to as degrees of freedom, is how many independent variables are added. Hence, one can conclude that the distribution

$$f(x) = \frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)} x^{\left(\frac{k}{2}\right)-1} e^{-\frac{x}{2}}$$

can be used when analyzing the term

$$\sum_{i=1}^{k} (y_i - \hat{y})^2.$$

In many applications it is desired to evaluate the Mean Squared Error “MSE,” which is defined as

$$MSE = \frac{1}{k}$$

$$\sum_{i=1}^{k} (y_i - \hat{y})^2$$

However, it is commonly not possible to compute this value on the whole data set; often it is not possible to obtain the full data set, so only a small data set is obtainable & analyzed, or only a subset is used for the computation initially as the other part of the data set is used to create the model (AKA training/testing data). If the full data set is of size n, and the total held back for testing is of size k, the MSE to compute would be

$$MSE = \frac{1}{k}\sum_{i=k+1}^{k+n}\left(y_i - \hat{y}\right)^2$$

A method known as statistical learning, which can be defined as “a framework for machine learning drawing from the fields of statistics and functional analysis, which deals with the statistical inference problem of finding a predictive function from a data set. For example, a method could be used on a data set that was split into this training/testing data set format in many different ways, and an algorithm could be written to obtain the outcome of all of the possible splits that would then identify the best model to use.

The MSE can also be written in terms of an expectation as

$$MSE\left(\hat{y}\right) = E\left[\left(\hat{y} - y\right)^2\right]$$

which is often rewritten as

$$MSE\left(\hat{y}\right) = VAR\left(\hat{y}\right) + BIAS(\hat{y}, y)^2$$

where the latter term is known as the bias. The BIAS term, formally known as the bias of an estimator, is defined as the difference between the estimator’s expected value and the true value of the parameter being estimated.

Another application of the chi squared density function is referred to as the goodness of fit. Technically this method requires to utilize the logic of hypothesis testing, which will not be introduced here until the next chapter, but the main idea can be understood if we think to compare an approximated value, $\hat{y}$, with a true data y. If the approximated is referred to as O, and the true data is E, the Pearson’s chi-squared statistics is defined as

$$\sum \frac{\left(O_i - E_i\right)^2}{E_i}$$

and is often used in practice to describe how well a statistical model under consideration fits a set of data. It is also used widely in test for normality of residuals.

4.6 EXAMPLES OF REAL WORLD APPLICATIONS

To begin our first application, it is worthy to mention the famous ***Black-Scholes partial differential equation***

$$\frac{\partial V}{\partial t}+\frac{1}{2}\sigma^2\frac{\partial^2 V}{\partial S^2}+rX\frac{\partial V}{\partial S}-rS=0$$

which was originally derived by Fisher Black and Myron Scholes utilizing probabilistic methods to calculate the fair price of a European call option, V, given the information of the initial stock, S, price along with the risk free rate, r, and the volatility, σ. The solution to this equation is give as the famous ***Black-Scholes equation***

$$S_0N\left(d_1\right)-Ke^{rt}N\left(d_2\right)$$

where S_0 is the initial price of the stock, K is the so called strike price, and N(#) is our standard normal cumulative distribution, i.e.

$$N(\#)=\int_{-\infty}^{\#}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\,dx.$$

In applications, the details of the numbers d_1 and d_2 can be obtained through two messy formulas involving the other known values, such as risk free rate and volatility, and often this formula is used as a starting point to obtain forward stock valuations.

Now, while the Black-Scholes equation is extremely useful in applications, it does have two concerns in regard to our desired purpose here, to use our probability theory to obtain stock price valuations. Firstly, it requires a measure known as volatility to be specified, which in practice is often extremely difficult, if not nearly impossible, to accurately define. Secondly, the result gives the fair price value of the option on a stock, rather than the price of the stock itself, and while options can be utilized to obtain information about the future price of a stock it would be preferred for our purposes here to obtain stock valuations directly.

To begin our brief study of stock valuations, we must first introduce the concept known as the time value of money. While this concept is a major core concept in advance studies in finance, in principal it is not overly complicated as at its core it simply says: “a dollar to be received in the future is worth less than a dollar to be received today.” To understand this idea, let us consider the following example: if your instructor would give you the option of $10,000 today or $10,500 at the end of the year, which is worth more? The answer really depends on various situations in the economy, and of course the needs of the individual, but for simplification let us determine how much the $10,000 given today could grow to if it was invested in a safe investment. Moreover, if we could invest the $10,000 in some sort of savings or bond, with a guaranteed rate of return of r%, then the $10,000 would become (1+r)*$1,000. At the time of writing this textbook, the United States Treasure was offering Series I Savings Bonds at a rate of 9.62%, hence if that investment was chose the $10,000 today would be worth $10,962. This is much higher than the other option of the $10,500 at the end of the year. It is this concept from where we obtain the time value of money, and discounting of future cash flows.

Definition 4.6.1 – Present value of a future cash flow

The present value of a future cash flow in the amount of FV$ to be obtained in time T years is

$$= \frac{FV}{(1+r)^T}$$

where r is the risk free rate. Often this formula is referred to as discounting a future cash flow.

At this level we will not attempt to develop a full treaty on what exactly the risk free rate is, but rather we will just accept here the use of 7%.

Example 4.6.1

Find the present value of $10,000 to be given to a freshmen college student from their family after completing college, which is 4 years forward in time.

To begin we note here that the future cash flow is $10,000 and the value of T is 4, thus using the present value formula we obtain the present value of this cash flow to be

$$= \frac{\$10,000}{(1+r)^4}$$

Now, in practice the next step would be to decide what risk free rate should be applied, but as previously noted here we will just apply the value of r to be 7%, hence our solution is

$$= \frac{\$10,000}{(1.07)^4} = \$7,628.95$$

From the prior example we can clearly see that $10,000 now is not the same as $10,000 to be obtained four years in the future. In fact $10,000 in four years is theoretically equal to $7,268.95 today. Of course there is a lot more to such problems in the real world, for example if this was a business is there some chance that whoever is saying they will pay us $10,000 in four years, may not do so? Perhaps the business may default on some of their obligations, or in extreme cases actually go out of business. However, for our purposes let us just accept that the present value, of a future cash flow in the amount of FV$ to be obtained in time T years is

$$= \frac{FV}{(1.07)^T}$$

Now, how does this apply to stock valuation? While there is an entire industry, in addition to a very active academic research topic, dedicated to making predictions of stock market investments utilizing many modern methods, one of the most simple – yet very useful and frequently successfully utilized – is the 10 year free cash flow analysis. The long story short is, at the end of a year, after a business collects all of its revenue and pays all of its debts and obligations it will have some money left over, this is referred to as free cash flow. The 10 year free cash flow analysis of a company simply says that the current valuation of a business is the present value of next 10 years of a company's future free cash flows, and the value of this company's stock should be this value divided by the number of shares outstanding in the market.

Example 4.6.2

A large American company, that has been in business of over a hundred years, has a current cash flow of $1,000,000 and is expected to maintain roughly the same business operating procedures for the next several decades. Use the 10 year free cash flow analysis to determine the current value of this company.

To begin we note here that this year's cash flow is $1,000,000 and from the details provided in the question we can assume that this same cash flow will occur next year, and then the following year etc.

Now, the present value of this year's cash flow is $1,000,000,

But the present value of next year's cash flow is

$$= \frac{\$1,000,000}{(1.07)^1} = \$934,579.44$$

Likewise, the present value of following year's cash flow is

$$= \frac{\$1,000,000}{(1.07)^2} = \$873,438.73$$

And, continuing in this fashion for a total of 10 terms (note the last value of T will actually be T=9, as we are including this year's or starting at T=0), then summing we can obtain the value of this company to be $7,515,232=

$$PV = \$1,000,000 + \frac{\$1,000,000}{(1.07)^1} + + \frac{\$1,000,000}{(1.07)^2} + \cdots + \frac{\$1,000,000}{(1.07)^9}$$

Example 4.6.3

The same large American company described in the prior example is knows to have 400,000 shares outstanding in the market, determine the value of this company's stock price.

To begin we recall that the 10 year free cash flow analysis states "the current valuation of a business is the present value of next 10 years of a company's future free cash flows, and the value of this company's stock should be this value divided by the number of shares outstanding in the market." From the prior example we computed the current value of this company to be $7,515,232, and we know that there are 400,000 shares outstanding in the market, hence the price of this company's stock should be

$$\frac{PV}{number\ shares} = \frac{\$7,515,232}{400,000} = \$187.88$$

Now, a few very important points must be made at this point! Firstly, the value computed here is what the company's stock value should be, but it is very common that the stock's value in the market will be very different. This is due to the fact that what we view as the "stock market," is really the secondary market. Namely, when a company first goes public it sells a predefined number of shares through an initial public offering at a fixed price. The individuals who purchased those shares directly then can resell them to other investors in the open market place – AKA the secondary market – at any time for any price, and this is where the stock market price swings come: if there are more people wanting to buy a specific stock than there are people offering it for sale, then the price will rise and likewise in the other direction. The second point is regarding how to accurately know if the company will maintain its cash flows? The answer to this question is not a simple one, but is at the core of an investment philosophy! A wise investor will chose to invest in a company that is not only expected to maintain their cashflows, but rather grow them as the company grows with time. For example, if the company from the prior example was expected to moderately grow, perhaps its cashflows would increase by 10% year of year; thus, after one year the cashflow would be $1,100,000 and then after the second year it would be $1,210,000. In theory this growth could be at any rate, but other than being able to time travel into the future to investigate there is no way to really know. A wise investor will do a very detailed investigation of the company and its competitors, and then make predictions based on the information obtained. While mathematics may not be able to exactly model such uncertainty, it is possible to apply a probability density as

$$f\left(x\right) = \begin{matrix} p_1 \; if \; unchanged \\ p_2 \; if \; moderate \; growth \\ 1 - p_1 - p_2 \; if \; strong \; growth \end{matrix}$$

and then apply it to compute an expected valuation, where here the p1 and p2 values are subjectively created. To illustrate this let us revisit the prior example and introduce some growth.

Example 4.6.4

A large American company, that has been in business of over a hundred years, has a current cash flow of $1,000,000 and is expected to grow this cash flow by 10% year over year for the next several decades. Use the 10 year free cash flow analysis to determine the current value of this company. Then compute the value of this company's stock if 400,000 shares are outstanding in the market.

To begin we note here that this year's cash flow is $1,000,000 and from the details provided in the question we can assume that next year that this cash flow will grow by 10%, hence it will be $1,000,000 + 0.1*$1,000,000 = $1,100,000. Then the next year's cash flow will follow, hence it will be $1,100,000 + 0.1*$1,100,000 = $1,210,000, and this will continue up to the 9th year which will be $2,357,948

Now, the present value of this year's cash flow is $1,000,000,

But the present value of next year's cash flow is

$$= \frac{\$1,100,000}{(1.07)^1} = \$1,028,037.38$$

Likewise, the present value of following year's cash flow is

$$= \frac{\$1,210,000}{(1.07)^2} = \$1,056,860.86$$

Then, continuing in this fashion and summing we can obtain the value of this company to be $11,360,801=

$$PV = \$1,000,000 + \frac{\$1,100,000}{(1.07)^1} + \frac{\$1,210,000}{(1.07)^2} + \cdots + \frac{\$2,357,948}{(1.07)^9}$$

Lastly, we know that there are 400,000 shares outstanding in the market, hence the price of this company's stock should be

$$\frac{PV}{number\ shares} = \frac{\$11,360,801}{400,000} = \$284.02$$

As expected, the result obtained here with growth is quite a bit higher than the prior example of no growth, which was $187.88. So the question arises as to which is the most accurate, which is the most likely? Well the truth is it is nearly impossible to accurately tell what will occur in the future, but a common trick used in applications is to compute some sort of combination of the outcomes, perhaps a weighted average of the outcomes. For example, if we claimed that there is a low probability – say 33% – that the company will maintain, and there was a strong probability – say 67% – that the company will grow, would a valid price target be?

$$0.33 \bullet \$187.88 + 0.67 \bullet \$284.02$$

The answer, just like the answer to many questions in finance and stock investments, is maybe? While this book is a textbook in probability, not financial mathematics, and this is the end of our application section to stock valuation, the author will end with the one accepted fact in stock market investments and some sound advice: if an accurate stock valuation is obtained and if the current asset is selling either below (or above) then in the long run the price will revert, in a rational functioning market, to the correct valuation. Moreover, there is no magic scheme to win in the markets – the stock market is a net sum zero game meaning for every person that makes $1 someone gives $1 – but long term investing in solid companies will build wealth, especially if the investor can purchase at a discount! The life story of the legendary investor Warren Buffet is a great read to further understand this wisdom.

Chapter 4 Exercises

1. Let the exponential function f(x) = e^{-3x} , x ≥ 0, and assume the variable x represents time (in hours) after 11am that I arrive on campus.
   - Create a normalized and valid PDF.
   - Use the PDF created in part a to find the probability that I arrive to campus between 11am and 12:30pm.
   - Use the PDF created in part a to find the probability that I arrive to campus after 1pm.
   - Redo part c leaving your answer in terms of $\Gamma(\#)$ "the Gamma Function".
   - Find the value of X so that it is 95% likely I will arrive to campus before time X.
2. Use the uniform density function f(x) = 1/10 on the domain 0 ≤ X ≤ 10 and assuming the variable x represent time (in hours) when an event occurs.
   - Use this PDF to find the probability that the event occurs within 0 to 5 hours.
   - Use this PDF to find the probability that the event occurs after 10 hours.
3. Given a RV defined for the range of *x between 0 and 10* whose data suggests a quadratic model, use the data points to create a valid and normalized PDF. data points: $(0,3),\ (0.1,3.21),\ (0.5,4.25)$ Use y=Ax^2+Bx+C
   - Model equation after found values A, B, C is the equation: y=_______________
   - Valid normalized PDF is y=_______________________
4. Use the function

$$\begin{cases} x^2, & when\ 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$$

   - Create a normalized and valid PDF
   - Use the PDF found in part a to find the probability $P\,(0 \le x \le 1)$

- Compute the expectation for the PDF created in part a.
- Compute the variance for the PDF created in part a.

5. Use the normalized beta density

$$f(x) = \begin{cases} 6x(1-x), & when\, 0 \leq x \leq 1 \\ 0, & elsewhere \end{cases}$$

- Compute the expectation for the PDF
- Compute the variance for the PDF

5. Introduction of Advanced Analytical & Inferential Statistics

5.1 FORMAL DEFINITION OF ALPHA LEVELS AND P VALUES

In this chapter the idea we want to consider is this: let's say we have two data sets – call one the control or historical data and call the other the experiment – and we want to test, formally, if there is a statistically significant difference between them. The inference is whatever experiment, or effect we applied, worked, and we want to conclude that it caused difference. However, since this is a maths foundations textbook & course we do not dive into the details of such matters. It is worthy to mention that one should never make a conclusion from a single experiment data set, rather they should look for trends in the data across multiple data sets.

Definition 5.1.1 – *Statistically significant difference*

We say that an observed difference between two data sets is statistically significant if it is unlikely to have occurred by chance alone (often referred to as unlikely to have occurred given the null hypothesis).

Definition 5.1.2 – *Critical Value Z*

We define the critical value of an experiment, conducted at the level of $1 - \alpha$ percentage confidence, to be the solution, Z, of the equation where the integration to compute the probability would be utilizing the probability density function associated with the data. It is worthy to note that in many textbooks or educational websites one will see this definition written in the classical "two tailed format," but here we have rewritten this, using the argument of symmetry, by adding half of the alpha,

$1 - \alpha + \frac{\alpha}{2}$, from the left tail to make the equation in a cumulative format,

which is useful as many coding languages utilize the cumulative distribution function. For example NORMCDF(Z) in matlab will return the probability for a standard normal up to the value of the random variable equal to Z.

Example 5.1.1

Find the critical value for a 95% confidence, assuming the density function is a standard normal

To begin we note here that alpha is 0.05. Now, the solution to ensure 95% is within would be

$$P(x < Z) = 1 - 0.025$$

Or rewriting this in terms of the density function, it becomes

$$\int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-0.5x^2} dx = 0.975$$

and while this solution can not be obtained explicitly (the definite integral of

e^{u^2}

is not known in exact closed format) it is possible to create and run a numerical code which will yield the well known solution

$$Z_\alpha = 1.96$$

where the common notation, Z_α, of the solution is used here and henceforth.

An important comment to note here is on the rounding, say for example the program yielded the solution to the above to be 1.951, we could not round that down to 1.95. As while the solution would tell us that 1.951 is the value of our random variable so that 95% is below, the value of 1.95 would be to the left on the axis, so there would be less area. The value of 1.95 would possible only cover 94.9%, which is not 95% as we claim the level of confidence to be. It is always essential in statistical analysis to round in a way called "statistically conservative," this is rounding as to always ensure to we meet or exceed the level of confidence of area under the curve; namely round up critical value and round down test statistical values obtained from experimental data.

Definition 5.1.3 – *P value*

When given the test statistic, TS, of a statistical experiment governed by the probability density function, f(x), we define the P value as

$$P = \int_{TS}^{\infty} f(x)\, dx$$

where the calculation is essential the area (or doubt) that is remaining in the rejection region above Za. Hence, when comparing two models the one with the lower P value is preferred!

Example 5.1.2

Find the P value given a test stat of 2.5, assuming this experiment was done using a standard normal density.

To begin we note here that TS = 2.5 and the density is a standard normal,

$$\frac{1}{\sqrt{2\pi}} e^{-0.5x^2}$$

so our definition of P value

would become

$$P = \int_{2.5}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-0.5x^2}\, dx$$

which has the solution 0.0062 (or 0.6%)

Example 5.1.3

Find the P value, using the same test stat from the prior example, TS = 2.5, but assuming this experiment was done using a T density, with five degrees of freedom.

To begin we note here that alpha is TS = 2.5 and the density is a T with v=5,

$constant * \left(1 + \frac{x^2}{5}\right)^{-1.5}$, so our definition & computation of P value would be

$$P = \int_{TS}^{\infty} f\left(x\right) dx$$

with the density formula noted above, and has a solution of 0.00275 (or 0.27%).

Now, it is important to note here that the P value is different for the same test stat when using a different distribution; hence, why the P value is the simplest stat to look at. Think: the lower the P value the better.

It is very important to remember, as a data analyst reporting results showing two data sets are significantly different, is not the same as nor does it prove Cause and Effect! Moreover, even if a longitudinal trend is shown within the data this does not prove a scientific theory. Data Science may be used to investigate solutions and/or governing equations, but such things – especially within fields like engineering – can only truly be developed utilizing the classic mathematical framework, e.g. modeling through things such as Newton's Law.

Definition 5.1.4 – *Confidence Interval*

$$MEAN - Z_{\alpha}\left(StError\right) \;\; to \;\; MEAN + Z_{\alpha}\left(StError\right)$$

The term of Standard Error is most commonly taken to be the standard deviation divided by the square root of the same size./calculation is essential the area (or doubt) that is remaining in the rejection region above Z_{α}. Hence, when comparing two models the one with the lower P value is preferred!

The big idea, or definition, to understand in the applications of the confidence interval is this: if a confidence interval is built using historical data, then an experiment is conducted and the mean of this experimental data set is obtained. Then, if the experimental mean is outside of the confidence interval, we define the difference to be statistically significantly different; hence, we can infer, but not prove, that whatever treatment was applied to the experimental data set worked.

5.2 MATHEMATICAL DEVELOPMENT OF CONFIDENCE INTERVALS AND INTRODUCTION TO HYPOTHESIS TESTING

In the prior chapter the general theory of probability was summarized along with the main concepts of probability density functions, cumulative distributions, and moment generating functions. Now, in this chapter we will embark on the study of one of the most powerful and important real-world applications of mathematics: the theory of hypothesis testing! The general idea of hypothesis testing can be summarized as the process of obtaining some data from an experiment and then using probability theory to attempt to validate a claim. Moving forward, we will refer to the claim as the hypothesis and generally speaking the experiment will involve the implementation of something that wasn't utilized in the past which we will refer to as the treatment. While the idea will not be discussed in detail here many instructors effectively teach hypothesis testing through the parallel logic of a court case. Namely, in a court case the defendant is assumed innocent until proven guilty beyond a reasonable amount of doubt as decided by a jury. Likewise, in hypothesis testing we desire to validate our claim the hypothesis, but we will take the stance that it is not valid (AKA assumed innocent) until proven otherwise beyond a mathematical amount of certainty (AKA beyond reasonable doubt).

In general, the hypothesis procedure will consist of 4 steps:

First, the hypothesis is made as a mathematical statement.

Second, the so called "critical value" and "rejection region" are defined.

Third, calculation of the test statistic.

Fourth, conclusions are stated.

In the following derivation, we will assume that the hypothesis is being studied on the simple difference on a population mean after the application of a treatment. Namely, we will consider the so called "null hypothesis" as μ = population mean value as given. The idea of this hypothesis statement is that the symbol μ is in a sense representing the population mean moving forward in time with the treatment applied consistently in the future (i.e. this statement is saying that the mean does not change when the treatment is applied). In the same manner that a defendant is assumed innocent in a court case until proven otherwise, we will assume this null hypothesis is truthful until proven otherwise.

The null hypothesis will be rejected if our soon to be defined test statistics falls outside our mathematical region which is defined from our chosen level of statistical certainty. Namely, if we define our statistical certainty to be at a level of $(1-\alpha)$then the critical value z_α (AKA endpoints) of our mathematical region can be found from the probability statement: $P(-z_\alpha < X < z_\alpha) = 1-\alpha$. For example, if we take a 95% confidence level the critical value will solve the equation

$$\int_{-z_\alpha}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1-\alpha = 0.95$$

This example will yield the solution of $z_\alpha = 1.96$ which is a very important, if not “famous,” critical value and very much worth remembering! It is worthy to note that many authors will use the notation
$z_{\alpha/2}$ due to the fact that this example is an illustration of a so called “two tailed test.” A two tailed test is one where the allowable error is allowed either above the critical value or below the negative of the critical value, hence the error is split in half. A one tailed test it where the error is not split in half, hence only outside of the critical value in “one tail.” For now, we will restrict our study to the two tailed examples for simplification.

Looking back to the original claim, we see that we have defined two regions: the range within the region is where we expect things to be and the range outside of the region, which is to be viewed as an oddity. Thus, we can define the region outside to be the region to reject the null hypothesis. Namely, if our soon to be defined test statistic is either greater than z_α (or less than –z_α) we will reject the null hypothesis. Or, in a cleaner mathematical statement we can say:

Reject the null if | | test stat | | $>z_\alpha$.

Now, the only missing point is the so-called test statistic. We will formally define and prove where this value comes from shortly but let us first accept the definition so that we can view a few examples to illustrate this process of hypothesis testing.

Definition 5.2.1 – The *Test statistic* for a single sample hypothesis test of differences of mean

TS = $\dfrac{X-\mu}{\sigma/\sqrt{n}}$

Example 5.2.1

To begin solving, we recall that a full solution to a hypothesis test problem has four steps:

First, the hypothesis is made as a mathematical statement.

Second, the so called “critical value” and “rejection region” are defined.

Third, calculation of the test statistic.

Fourth, conclusions are stated.

For our present example, we will assume that the level of confidence is 95% and the test is a two tailed test (which would make sense as the researcher wanted to unbiasedly test for an effect on speed rather than specifically test for an increase). So, we already know the second step is with 1.96 being the critical value. Hence, all we really need to compute are the 1st and 3rd steps. To begin, we must define the desired hypothesis, which is what we really want to show and is often referred to as the alternate hypothesis. In this example, the researcher knows that the population mean is 73 and they are attempting to see if the drug has an effect on that speed. Thus, we set the alternate hypothesis to state that μ is different than 73. Then, we also must construct the null hypothesis (the logical opposite of the alternate hypothesis) which in this case will state that μ is equal to 73. Now, all that remains is to compute the test statistic from our formula and then use our results to conclude.

In doing so, we obtain the for step solution as:

First, null H: μ = 73.

Alt H: μ ≠ 73.

Second, assume the null is truthful and reject if || TS || > 1.96.

Third, TS = $\frac{71-73}{\frac{21}{\sqrt{81}}} = -0.857.$

Fourth, since the TS does not fall in the rejection region, we fail to reject the null.

It is very important to note in this example that the result is just simply failure to reject the null hypothesis. This wording is very important, and it is essential to understand that this conclusion does not disprove anything, nor do we accept anything, rather we have just failed to reject the null hypothesis. Perhaps, one will find it useful to think that we have attempted to do something and failed to do so. Hence, our conclusion is that we did not do anything, or perhaps a more sophisticated way it to say we have “no conclusion!” Analogously, when a jury is tasked to find a defendant guilty beyond a reasonable doubt, if they do not find the evidence, then their formal result is to say “not guilty” or “no, we did not find sufficient evidence.”

Example 5.2.2

An instructor wants to see if group activity work increase test scores. Currently the school’s average math score is 85 with a standard deviation of 4. A sample of 36 students are assigned to do group work in class. Their average is 90.

Perform an appropriate hypothesis test.

Again, we note that a full solution to a hypothesis test problem has four steps, and to begin our present example we observe that the desired hypothesis is specifically to increase the test scores. It is known that the population mean score is 85, so the logical choice for the Alt H is : μ > 85. As in our last example, we will assume that the level of confidence is 95%, but this is a one tailed test. Therefore, the prior critical value of 1.96 would not be the correct critical value. To find the correct value we would need to go back to our probability density theory. In doing so, we obtain the desired equation to solve

$$\int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \alpha = 0.95$$

which will yield the solution of

$$z_\alpha = 1.65$$

We are now prepared to fully develop our hypothesis testing procedure:

First, null H: μ = 85.

Alt H: μ > 85.

Second, assume the null is truthful, and reject if: TS > 1.65.

Third, TS = $\dfrac{90-85}{\frac{4}{\sqrt{36}}} = 7.5.$

Fourth, since the TS does fall in the rejection region we reject the null.

We previously presented the definition of the test statistic formula without development nor proof. Let us now formally define and prove from where this formula comes. We assume that the population problem we are studying is modeled by a normal distribution with mean μ and standard deviation �, hence X~N(μ,�). Now, in regards to the sample we will need to utilize two Lemmas from a theorem of advanced probability theory known as the Central Limit Theorem. Namely, we will take as definition the following:

Definition 5.2.2 – The *mean* and *variance* of a sampling distribution

If X_1,X_2,..., X_n are random variables with

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

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

and if is the random variable of the sample means of all the simple random sample size n from a population with expected value E(X), and variance Var(X) then

$$E(\bar{X}) = E(X)$$

$$Var(\bar{X}) = \frac{1}{n}Var(X)$$

We now need to prove our main foundational result, which is illustrated in the following theorem definition.

Definition 5.2.3 – *Central Limit Theorem*

If X_1,X_2,..., X_n are normally distributed random variables with mean μ and standard deviation �, then

$$\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$

Proof: Let us begin by recalling a fact about random variables, namely if

$$X \sim N\left(\mu, \sigma\right)$$

and we consider the RV =aX, where a is a fixed constant, then we can show by some routine algebra on the cumulative distribution function that this

$$RV \sim N\left(a\mu, a\sigma\right).$$

A similar result is well known that if we have two random variables

$$X \sim N\left(\mu, \sigma\right)$$

and

$$Y \sim N\left(\nu, \varphi\right)$$

and if we consider the RV = X + Y, then we will find

$$X + Y \sim N\left(\mu + \nu, \sigma + \varphi\right).$$

Thus, we can draw the conclusion that our X_1,X_2,..., X_n are normally distributed random variables that

$$X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Now, we shall look at the expression

\frac{\bar X -μ}{σ/\sqrt{n}}=\frac{\sqrt{n}}{σ}\bar X -\frac{\sqrt{n}}{σ}μ.

From the results above we will see that this has a mean 0 and standard deviation 1, hence we have proved that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

which means the use of the standard normal distribution for our critical values are validated.

5.3 MOMENT GENERATING FUNCTION

The Moment Generating Function, as its name implies, is a function that we will create which can be used to find the moments of the probability distribution, and these moments are very useful in applications to find things such as the mean and variance. The good news is once we have obtained the MGF, we will be able to obtain these moments without computing lengthy integrations such as we encountered in prior sections when attempting to directly compute the variance. However, as we will see shortly the actual definition of the MGF is actually itself an integration, but in most applications one will be able to start from a known MGF rather than needing to compute, phew!

The formal definition of the Moment Generating Function, is the expected value of the function e^{tx}, where x is the random variable and t is a new variable which is not related to x, hence it can be treated as a constant in operations such as the expected value integration which is with respect to x. Thus, we have arrived at our main definition for this section:

Definition 5.3.1 – The *moment generating function* for density f(x)

$$MGF = E(e^{tx}) = \int_{\Omega} e^{tx} \bullet f(x)dx$$

where the result is a function of t.

Example 5.3.1

Find the moment generating function for the uniform density with L =1 and R = 5, i.e. the density

$f\left(x\right)=\frac{1}{4}.$

To begin our solution we note that we have the density defined as $f\left(x\right)=\frac{1}{4}$, and from our prior knowledge we recall that this function is defined on the sample space of 1<x<5. Hence, we can now compute its MGF as

$$MGF = E\left(e^{tx}\right) = \int_{\Omega} e^{tx} \bullet f(x)dx = \int_{1}^{5} e^{tx} \bullet \frac{1}{4}dx.$$

Now, to compute this integration it is first noted that the value of t, while it is officially a variable, can be treated as a constant within this integration, hence we can compute the integration as

$$MGF = \frac{1}{4}\left[e^{tx}\right]_{x=1}^{5} = \frac{1}{4}\left(e^{5t} - e^{t}\right)$$

Example 5.3.2

Find the moment generating function for the exponential density with A=2, i.e. the density

$f\left(x\right)=2e^{-2x}.$

To begin our solution we note that we have the density defined as $f\left(x\right)=2e^{-2x}$, and from our prior knowledge we recall that this function is defined on the sample space of x >0. Hence, we can now compute its MGF as

$$MGF = E\left(e^{tx}\right) = \int_{\Omega} e^{tx} \bullet f(x)dx = \int_{0}^{\infty} e^{tx} \bullet 2e^{-2x}dx.$$

Now, to compute this integration, again it is important to recall that while the value of t is officially a variable, here it can be treated as a constant within the integration. Also, the useful property of exponential functions, $e^{A} \bullet e^{B} = e^{A+B}$, is applied and doing so we find

$$MGF = 2\int_{0}^{\infty} e^{tx-2x}dx = 2\int_{0}^{\infty} e^{x(t-2)}dx$$

$$= \frac{2}{t+2}\left[e^{x(t-2)}\right]_{x=0}^{\infty} = \frac{2}{t-2}\left(e^{-\infty} - e^{0}\right) = \frac{2}{2-t}$$

It was assumed here that the value of t was chosen so that t-2<0, hence the value of the parameter t was chosen so that our integration would converge.

In a pure mathematical point of view one may state the conclusion from the last example that we have found the moment generating function to be $\phi\left(t\right) = \frac{2}{2-t}$ which is only defined for t < 2. However, for our purposes in this textbook we will not include such details as we will only be working with well-known Moment Generating Functions which are stable and defined at the value of t as needed in applications (generally it is needed to evaluate these function at t = 0), but the abstract concept of convergence it important to be aware of as not all density functions have a convergent MGF, such as the next example will illustrate.

Example 5.3.3

Find the moment generating function for T density, with v=2, i.e. the density

$$f\left(x\right) = \frac{\Gamma\left(\frac{3}{2}\right)}{\sqrt{2\pi}\Gamma\left(2\right)}\left(1+\frac{x^{2}}{2}\right)^{-\frac{3}{2}}$$

To begin our solution we note that we have the density defined as $f\left(x\right) = \frac{\Gamma\left(\frac{3}{2}\right)}{\sqrt{2\pi}\Gamma\left(1\right)}\left(1+\frac{x^{2}}{2}\right)^{-\frac{3}{2}}$ and from our prior knowledge we recall that this function is defined on the sample space of x >0. Hence, we can now compute its MGF as

$$MGF = E\left(e^{tx}\right) = \int_{\Omega} e^{tx} \bullet f(x)dx = \int_{0}^{\infty} e^{tx} \bullet \frac{\Gamma(\frac{3}{2})}{\sqrt{2\pi}\Gamma(1)}(1+\frac{x^{2}}{2})^{-\frac{3}{2}}dx.$$

Prior to computing this integration it is worthy to recall the known particular values of the gamma, namely that $\Gamma\left(\frac{3}{2}\right) = \frac{1}{2}\sqrt{\pi}$ and $\Gamma\left(1\right) = 1$, hence our integration simplifies to

$$= \frac{1}{2\sqrt{2}} \int_0^\infty e^{tx}\left(1 + \frac{x^2}{2}\right)^{-\frac{3}{2}} dx$$

Now, while no exact “closed form” of this integral (e.g. indefinite integral) is known, it is possible to compute the integration using series methods; however, doing so would result in a solution that involves all positive power terms of the form $(tx)^n$. This would then need to be evaluated at positive infinity, which would lead to a divergent result. Hence, the conclusion we obtain is that this MGF integration for the T density does not converge, so we conclude that the T density does not have an MGF.

Prior to continuing our development of Moment Generating Functions, along with their applications, it is worthy to revisit our list of common density function that are frequently used in examples, and now also state their Moment Generating Functions. The following are the most likely examples that you will encounter are:

The ***exponential density*** is $f\left(x\right) = Ae^{-Ax}$ which is defined for x 0, and its MGF is $\phi\left(t\right) = \frac{A}{A-t}$

The ***uniform density*** is $f(x) = \frac{1}{(R-L)}$ which is defined for L < x < R, and its MGF is $\phi(t) = \frac{1}{t(R-L)}(e^{Rt} - e^{Lt})$

The ***normal density*** is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ which is defined for all x, and its MGF is $\phi\left(t\right) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

The ***Tdensity*** is $f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma(\frac{v}{2})}\left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}$ which is defined for x > 0 with v being the degrees of freedom, and as noted in the last example from the prior section it does not have a MGF.

The *chi squareddensity* is $f\left(x\right)=\frac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}x^{\left(\frac{k}{2}\right)-1}e^{-\frac{x}{2}}$ which is defined for x > 0, with v being the degrees of freedom, and and its MGF is $\phi\left(t\right)=\left(1-2t\right)^{-v/2}$

As the proof of the prior section developed, we can now alternately find the nth moment,
$\mu_n=E\left(x^n\right),$
of a probability distribution without needing to actually computing the expectation integration.

Definition 5.3.1 – The *n'thmoment* of a density f(x)

$$\mu_n=E(x^n)=\frac{d^n}{dt^n}[MGF]_{t=0}.$$

This result can be extremely useful in applications, such as finding the variance which one can show is equal to $\mu_2-(\mu_1)^2$.

Prior to moving forward with examples, let us quickly outline the proof of the above result. To being, the nth moment is formally defined as
$E(x^n)$

Now, the moment generating function is defined as
$MGF=E\left(e^{tx}\right).$
And, expanding the exponential function from this definition in a power series, we obtain that the moment generating function can alternately be written as

$$MGF=E\left(\sum_{n=0}^{\infty}\frac{(tx)^n}{n!}\right)=1+t\bullet E\left(x\right)+\frac{t^2}{2}E\left(x^2\right)+\ldots$$

Then, if one takes the first derivative and evaluates that at t being zero, all terms except the expected value of x will vanish hence, we have established that

$$\frac{d}{dt}[MGF]_{t=0}=E\left(x\right)$$

Likewise, if one takes two derivative and evaluates that at t being zero, all terms except the expected value of x squared will vanish hence, we have established that

$$\frac{d^2}{dt^2}[MGF]_{t=0} = E\left(x^2\right)$$

And, this pattern can be continued on indefinitely to prove the result provided in definition 5.3.1. Moreover, if one plays a little algebra – recalling that the mean "μ" is a constant so it can be taken outside of the integration – with the prior definition of variance

$$VAR = E(x-\mu)^2 = \int_{\Omega}(x-\mu)^2 dx.$$

It can be established that

$$VAR = \int_{\Omega} x^2 dx - \mu^2$$

or in terms of moments

$$VAR = \mu_2 - (\mu_1)^2.$$

Then, recalling that μ_1 can be computed as the first derivative of the MGF, while μ_2 can be computed as the second derivative of the MGF, we can see it is possible to obtain both the mean & variance through this new method as

$$MEAN = \mu_1 \qquad VAR = \mu_2 - (\mu_1)^2$$

The method developed above can be used to compute the expectation and variance without computing any of the integrals, such as done in the prior examples. Let us now look to some examples for illustration.

Example 5.3.4

For the ***exponential density*** is $f(x) = 2e^{-2x}$ which is defined on the sample space for x>0, and has the MGF is $\phi(t) = \dfrac{2}{2-t}$ find the mean and variance, firstly by the classical "integration" method, then secondly by the MGF method.

To begin our solution we note that we have the density defined as $f(x) = 2e^{-2x}$ and from our prior knowledge we recall that this function is defined on the sample space of x >0. Hence, we can now compute its expectation as

$$E(x) = \int_{\Omega} x \bullet f(x) dx = \int_0^{\infty} xe^{-2x} dx$$

The solution to this integration is found to be = 0.5. Now, on the other hand we can compute the 1st moment as

$$\frac{d}{dt}[MGF]_{t=0} = \frac{d}{dt}\left[\frac{2}{2-t}\right]_{t=0} = \frac{1}{2}$$

Likewise, we can now compute the variance as expectation

$$VAR = E(x-\mu)^2 = \int_\Omega (x-\mu)^2 \bullet f(x)dx = \int_0^\infty \left(x-\frac{1}{2}\right)^2 e^{-2x}dx$$

The solution to this integration is found to be = 0.25. Now, on the other hand we can compute the variance as

$$\frac{d^2}{dt^2}[MGF]_{t=0} - \mu^2 = \frac{d^2}{dt^2}\left[\frac{2}{2-t}\right]_{t=0} - \left(\frac{1}{2}\right)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Example 5.3.5

Find the mean of for chi squared density, with v=2 degrees of freedom, firstly by the classical "integration" method, then secondly by the MGF method.

i.e. the density

$$f(x) = \frac{1}{2\Gamma(1)}x^{(1)-1}e^{-\frac{x}{2}} = \frac{1}{2}e^{-\frac{x}{2}}$$

with the associated MGF is $\phi(t) = (1-2t)^{-1}$

To begin our solution we note that we have the density defined as $f(x) = \frac{1}{2}x^{-\frac{x}{2}}$

and from our prior knowledge we recall that this function is defined on the sample space of x >0. Hence, we can now compute its expectation as

$$E(x) = \int_\Omega x \bullet f(x)dx = \int_0^\infty xe^{-\frac{x}{2}}dx.$$

The solution to this integration is found to be =2. Now, on the other hand we can compute the 1st moment as

$$\frac{d}{dt}[MGF]_{t=0} = \frac{d}{dt}[(1-2t)^{-1}]_{t=0} = 2$$

Example 5.3.6

A probability density function is under investigation, but it is not known explicitly, and using a set of 100 data points it is found to have the MGF e^{t+4t^2}.

Use this function to find the mean and variance, and then use those results to find the 95% two tailed confidence interval. Also, make a comment as to if it is possible to those results and/or the confidence interval results to find the actual probability density function.

Now, to being we can compute the 1st moment as

$$\frac{d}{dt}[MGF]_{t=0} = \frac{d}{dt}\left[e^{t+4t^2}\right]_{t=0} = 1$$

Then, we can compute the variance as

$$\frac{d^2}{dt^2}[MGF]_{t=0} - \mu^2 = \frac{d^2}{dt^2}\left[e^{t+4t^2}\right]_{t=0} - (1)^2 = 3 - 1 = 2$$

And, we can then set up the 95% confidence interval

$MEAN - 1.96 \bullet \sqrt{\frac{VAR}{n}}$ to $MEAN + 1.96 \bullet \sqrt{\frac{VAR}{n}}$,

by simply plugging in the value of the mean as 1 and the value of the variance "VAR" as 2, and n as 100; doing so yields the solution that we are "95% confident that the population mean is between 0.73 and 1.27." Lastly, to address the question as to if any of this information could be used to determine the actual probability density function, the answer to that question is a bit of a grey area; moreover, if we knew this was from a normal data set and the results we obtained were accurate representations of the population mean μ, and variance σ^2, then we could define the density as $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-2)^2}$. However, it is not clear from the information given – albeit that the given MGF does look familiar – that it would be safe to conclude it is a normal distribution. Moreover, it is important to remember that in practice one should never attempt to draw conclusions from a single data set! If we had multiple data sets and we saw a trend in the results and we had some real world context to suggest the phenomena was modeled by a normal distribution, then we could move closer to conclusion but it does not appear that can be done with the information given. It is important to remember that many results, such as the famous central limit theorem, are only valid for a sample of samples!

Chapter 5 Exercises

1. From the central limit theorem the term $\frac{\sigma}{\sqrt{n}}$ is often called the standard error term. What happens to this term as the sample size grows without bounds? (HINT: perhaps fix σ , say standard normal with $\sigma = 1$, and then try a few big then bigger samples n=10, then n=100, then $n = 1000 \ldots$)
2. For a sample of n=31, governed by the normal density, set up (**not compute**) the integral equation to find the
   Z critical value for the following two tailed confidence levels using mean=10& variance.
   - $\frac{\alpha}{2} = 0.1$
   - 95% confidence
   - 99% confidence
3. An experimental study is done to improve the cost to drive an electric car; it is found the average gas price is $2.69 which leads to a ten cent fuel cost per mile in cars with a standard deviation of one cent, use this information as the population. Now, in your experiment of 31 cars you worked with Elon Musk and created both a new battery & interstate charging system. This led to the average cost going down to nine cents per mile. At the 95% confidence level do you feel confident to say your experiment had an effect and/or is a *statistically significant difference*? NOTE: use $Z_{\frac{\alpha}{2}} = 1.96$
4. Compute E(x) for N~(3,1) {e.g. a normal with mean $\mu = 3$ and st dev $\sigma = 1$}
   - by integral definition (set up & simplify integral and use integration software)
   - by MGF
5. Compute P(0<x<1) for T_2 (set up & simplify integral and use integration software)
6. Compute E(x) for χ_2

  - by integral definition (do by hand)
  - by MGF

7. Use the exponential PDF with a=7, i.e. $f\left(x\right)=7e^{-7x}$, for x>0
   - Use the traditional definition $E\left(x\right)=\int_{\Omega}x\bullet f\left(x\right)dx$ to compute the expectation. (do by hand (IBP))
   - Use the traditional definition $\sigma^{2}=\int_{\Omega}\left(x-\mu\right)^{2}\bullet f\left(x\right)dx$ to compute variance. (do by hand (IBP))
   - Use table provided in class to write MGF for this PDF & use it to compute the 1st & 2nd moments.
   - Compute the variance as $\sigma^{2}=\mu_{2}-\left(\mu\right)^{2}$ and verify that it yields the same solution as part b.
   - Use the traditional definition $E\left(x^{2}\right)=\int_{\Omega}x^{2}\bullet f\left(x\right)dx$ to verify your part c solution for μ_{2}

8. Use the uniform PDF with L=3, i.e. $f\left(x\right)=\frac{1}{3}$, for 0<x<3
   - Use the traditional definition $\sigma^{2}=\int_{\Omega}\left(x-\mu\right)^{2}\bullet f\left(x\right)dx$ to compute variance (do by hand)
   - Use table provided to write MGF for this PDF & use it to compute the 1st & 2nd moments. Hint: you will need to expand e^{3t} using a Taylor series. Hint: Taylor series for
   $$e^{x}=\sum_{n=0}^{\infty}\left(\frac{x^{n}}{n!}\right)=1+x+\frac{1}{2!}x^{2}+\frac{1}{3!}x^{3}+\ldots$$
   - Compute the variance as $\sigma^{2}=\mu_{2}-\left(\mu\right)^{2}$ and verify that it yields the same solution as part a.

6. Applications to Disease Prediction and Spread Prevention

6.1 SUMMARY OF SIR MODEL

To begin our brief summary of disease modeling, it is noted from Wikipedia article (XYZ1), that as of April 2020 the basic reproduction number “R_0” for SARS-CoV-2 was estimated to be between 1.4 and 3.9. The computation of this value is absolutely essential to any mathematical prediction of the disease spread, as this value basically explains how many people one infected individual will spread the virus to. However, this value is often extremely challenging to obtain in real time as not only is it very difficult to trace either where an individual was infected from and/or to whom they may have passed it to, but this value is also very much affected by the regional aspects. For example, an individual in a highly urbanized area such as New York City will generally interact with hundreds and hundreds of people during their usual everyday lives commuting on subways or walking on crowded sidewalks, hence interacts with a lot more hosts to potentially spread to; while an individual in a rural area will generally interact with only a few people during their everyday lives, hence interacting with only a few potential hosts to spread to.

Now, while there are many advance techniques to resolve such issues in data collection we will not address those here but rather we will focus on outlining the main steps in developing a mathematical model to model a spread in disease. Then in the following two sections a summary of two applications of hands on real world data analysis is illustrated to show how one can, in real time, obtain practical useful information to understand more in real time of how a disease is spreading.

One of the most commonly utilized infectious disease predictions methods is the so called SIR model This model attempts to predict, as time moves for-

ward, the number of **I**nfected and then **R**ecovered individuals from a population of **S**usceptible individuals. Hence, the SIR model name is referring to the flow $S \to I \to R$. To summarize the logic behind this model we first define the following functions
$S\left(t\right) =$ the number of not yet infected individuals susceptible of the disease at time t.
$I\left(t\right) =$ the number of not yet recorded individuals infected with the disease at time t.
$R\left(t\right) =$ the number of previously infected individual, who are now recovered at time t.
The goal of the SIR model is to create a mathematical model between these three functions, commonly starting from some initial data $S\left(0\right) = S_0$ and $I\left(0\right) = I_0$.

Now, while we will not discuss the intricate details, nor how they adjust the modeling, a few main assumptions we note here are that in our simplified modeling we assume that all of the individuals in $S\left(t\right)$ are equally likely to become infected the disease, and all of the individuals in $I\left(t\right)$ are equally likely to spread the disease. In addition, we assume that once an individual is infected, hence moves from $I\left(t\right)$ into $R\left(t\right)$, they can no longer spread nor be re infected by the disease. Furthermore, the function $R\left(t\right)$ is actually a compartment that collects all individuals after they leave $I\left(t\right)$, hence it includes both recovered individuals who survived then gained immunity, but it also includes individuals who died. In addition, we do not attempt to predict any methods to adjust these transitions, such as applying external factor like cure medicines, nor do we attempt to introduce any jump functions, i.e. $S\left(t\right) \to R\left(t\right)$, that result from applying external factors such as vaccinations. In short, the SIR model we construct here is just a good start for an initial model stage in the research.

The commonly accepted assumption (XYZ2) of the SIR model, is that the rate of change of change with respect to time of $S\left(t\right)$ is proportional to the ratio, from the total population, of the product of current number of susceptible times the infected. Thus, by defining the constant value of N as the sum $S\left(t\right) + I\left(t\right) + R\left(t\right)$, this yields a differential equation for $S\left(t\right)$ as

$$\frac{dS}{dt} = -\beta \frac{S\left(t\right) I\left(t\right)}{N}$$

where β is the to be determined proportionally constant, and it is worthy to note that this value of beta is also the probability of an individual within $S\left(t\right)$ to become infected with the disease of which is equal likely for all individuals. Also, it should be noted that the negative value is utilized in the differential equation to explain the fact that as time moves forward the size of $S\left(t\right)$ decrease since individuals move out of $S\left(t\right)$ into $I\left(t\right)$. In addition, if we define γ as the to be determined proportionally constant modeling the rate of change from $I\left(t\right)$ into $R\left(t\right)$ this yields a differential equation for $R\left(t\right)$ as

$$\frac{dR}{dt}=\gamma\, I\left(t\right)$$

It is worthy to note here that γ can also be viewed as the mean recovery/death rate which can be approximated from real data in real time.

If it assumed that this is a closed system then the rate of change of $I\left(t\right)$ can be computed applying simple in minus out logic, hence this yields a differential equation for $I\left(t\right)$ as

$$\frac{dI}{dt}=\beta\frac{S\left(t\right)I\left(t\right)}{N}-\gamma\, I\left(t\right).$$

Thus, a classic 3×3 system of three differential equations for three unknown functions has been obtained. While this system is a complex non linear differential equation, there are some methods of solution which can yield some extremely useful information.

Prior to outlining these solution methods, it is worthwhile to take a step back and consider what are the most important pieces of information to obtain in real time as a new disease is spreading in real time. Namely the basic production number, R_0, is one of the most desirable pieces of information to obtained. If the basic production number is known, the interpretation of its value can be used by doctors along with governmental officials to determine if a disease spread will be a minor event or an epidemic or in the worst cases become a global pandemic. Now, in a most simple cases one can determine the number of infected cases over time, in the early stages of a disease spreading, as simple exponential growth model with a logarithmic growth rate of

$$K=\frac{d}{dt}ln\left[I\left(t\right)\right].$$

Then, if from data, it is possible to estimate that after time, T_I, an individual infects exactly
R_0 new individuals then the value of K can alternately be computed

$$K = \frac{\ln(R_0)}{T_I}$$

and from this information both the values of initial growth rate along with the basic production number can be approximated. However, it is extremely difficult to actually obtain the information needed in real time and often by the time it is discovered that a disease is actively spreading in the real world the growth has moved much further along in its evolution than being modeled by such a simple model utilized for this rudimentary solution. Thus, a more advanced methodology is commonly required to estimate R_0 which is where our full 3×3 system of differential equations can be utilized.

The formal definition the basic production number is

$$R_0 = \beta\tau$$

where beta is as previously defined, but often unknown and very difficult to estimate in real time, and � is the mean infectious period which is often able to be estimated in real time by observation of real cases. Hence, if one can create a mathematical model from data in real time and then compare it to a mathematical model from our prior differential equation solution it maybe possible to extract the value of beta, and thus accurately estimate the value of R_0. To begin, we take our system of three differential equations

$$\frac{dS}{dt} = -\beta\frac{S\left(t\right)I\left(t\right)}{N},$$

$$\frac{dI}{dt} = \beta\frac{S\left(t\right)I\left(t\right)}{N} - \gamma\, I\left(t\right),$$

$$\frac{dR}{dt} = \gamma\, I\left(t\right).$$

And, we note that since the sum S+I+R is assumed to be a constant value of the total population, we have the fact of

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$$

And, due to the fact that $\tau = \frac{1}{\gamma}$ we note the basic reproduction number can be rewritten as

$$\frac{\beta}{\gamma} = R_0.$$

$$\frac{dI}{dt} = \left(R_0\frac{S}{N} - 1\right)\gamma\, I\left(t\right)$$

we can observe that the value within the parentheses tells a lot of practical information. First, it is worthy to note that the value of $\gamma\, I\,(t)$ will always be positive in sign. Thus we can conclude that if

$R_0 > \dfrac{N}{S}$ then the disease will spread rapidly, as the sign of $\dfrac{dI}{dt}$ will be positive, hence increasing. Thus, with initial data, a disease can be defined as one will spread to an epidemic outbreak, or in worst cases a pandemic, if

$$R_0 > \frac{N}{S(0)}$$

while it will not be expected to if

$$R_0 < \frac{N}{S\,(0)}$$

This information is one of the most powerful pieces of information that doctors and/or governmental officials can be provided with in real time when making policy decisions for a new disease. The only major issue with this is that by the time enough data has been collected to obtain this critical piece of information is obtained, in real time the disease has often spread so far that there is often not much that can be done to stop the spread of the disease, other than attempts to reduce the function $S\,(t)$ such as social distancing.

Now, to actually obtain a model solution for this system of equations, some algebraic manipulation is needed. To begin if the first is divided by the third it is obtained that

$$\frac{dS}{dR} = -R_0\frac{S\,(t)}{N}$$

and by routine variable separation for the first order differential equation this becomes

$$\frac{1}{S}dS = \left(\frac{-R_0}{N}\right)dR.$$

Then, by conducting a definite integration it is found that

$$Ln\,[S\,(t) - S\,(0)] = \frac{-R_0}{N}R\,(t) + \frac{R_0}{N}R\,(0)$$

From which the solution

$$S\,(t) = S\,(0)\,e^{-\frac{R_0}{N}(R(t)-R(0))}$$

is obtained. While this is not exact solution to our 3×3 systems of equations, as the value of $S\,(t)$ obtained depends on the value of $R\,(t)$, it is an extremely useful piece of information which can be used in real time. If, in real time, data is collected to measure the value of $R\left(t\right)$ and possibly even $I\left(t\right)$ this solution, which can be rewritten as

$$N-I\left(t\right)=R\left(t\right)+S\left(0\right)e^{-\frac{R_0}{N}\left(R(t)-R(0)\right)}$$

is an extremely powerful solution. From here further methods can be applied to either obtain actual individuals closed from solutions for $\left(t\right)$ and $I\left(t\right)$, or can yield an approximation for the basic reproduction number. Namely, if a nonlinear regression data fit (AKA using the command ~nls in R or python coding) a data fit solution can be created and then by comparing the two one can extract an approximation value for
R_0.
While this modeling is an extremely interesting mathematical model we will not continue further development on the topic here, but rather we will quickly look in the next section at one real world data example from New York City of the COVID-19 disease spread in to illustrate how the statistical methods learned can be applied to actually model a regression solution. Then we will end this textbook in the proceeding section to discuss an extremely interesting research question, how to determine what factors drive citizens to actively participate in social distancing measures. This is a very important topic to study as within the scientific community it accepted that once a disease is activity spreading in an epidemic, or even worse a pandemic, spread then the most effective way to stop the disease is social distancing. This is due to the fact that while these mathematical solutions are beautiful to study, the downfall is that in real life once a disease starts spreading from person to person there is absolutely nothing that can be done to stop it. The only tools we have available to us such as humans in such a battle against a virus are common sense preventative measures to spread the disease in daily life (e.g. wearing filtration masks and gloves or other personal protective equipment), or taking societal measures such as social distancing. While this is accepted by most in the community, we can now validate it mathematical as from our solution of the second equation we noted the disease will not become a pandemic if

$$R_0<\frac{N}{S\left(0\right)}$$

While we do not have any control over R naught, and the value of N is fixed, we can greatly reduce the value of $S\left(0\right)$ by implementing social distancing measures; in fact as that value approaches zero the right hand side of this bound will become infinite which ensures society will win the battle against the virus, or here mathematically a battle against the direction of an inequality symbol!

6.2 DATA ILLUSTRATIONS OF SEEKING THE EXPONENTIAL INFLECTION POINT

3/1/2020	0	1
3/3/2020	0	3
3/4/2020	0	8
3/5/2020	0	11
3/6/2020	0	18
3/7/2020	0	25
3/8/2020	0	46
3/9/2020	0	103
3/10/2020	0	173
3/11/2020	1	326
3/12/2020	2	681
3/13/2020	2	1299
3/14/2020	4	1941
3/15/2020	10	2969
3/16/2020	19	5085
3/17/2020	26	7532
3/18/2020	47	10481
3/19/2020	71	14159
3/20/2020	116	18144
3/21/2020	157	20744
3/22/2020	205	23288
3/23/2020	288	26790
3/24/2020	382	31180
3/25/2020	503(+31.7%)	35914
3/26/2020	688(+36.8%)	40840
3/27/2020	897(+30.4%)	45829
3/28/2020	1,162(+29.5%)	49214
3/29/2020	1,444(+24.3%)	52651
3/30/2020	1,757(+21.7%)	58666
3/31/2020	2,126(+21%)	63834
4/1/2020	2,545	68859
4/2/2020	3,001	74504
4/3/2020	3,465	80020
4/4/2020	3,942	83772
4/5/2020	4,475	87386

4/6/2020	5,024	93592
4/7/2020	5,599	99511
4/8/2020	6,118	104915
4/9/2020	6,638	109781
4/10/2020	7,137	114022
4/11/2020	7,645	117588
4/12/2020	8,172	120304
4/13/2020	8,699	123514
4/14/2020	9,181	127569
4/15/2020	9,606	131366
4/16/2020	9,987	134819
4/17/2020	10,334	138318
4/18/2020	10,682	140397
4/19/2020	11,031	142679
4/20/2020	11,355	146393
4/21/2020	11,640	149395
4/22/2020	11,924	152809
4/23/2020	12,208	155596
4/24/2020	12,480	157994
4/25/2020	12,691	159508
4/26/2020	12,899	160498
4/27/2020	13,114	162728
4/28/2020	13,292	165369
4/29/2020	13,449	167633
4/30/2020	13,590	169555

Now, prior to conducting the data analysis for this data set it is useful to make a note about how, in real time, it is preferred to look at a very simple measure which is the rate of change of cases not the actual increase in raw numbers. For example, the day over day change from March 24th to March 25th was 31.7% which is easily computed as number of cases reported on March 25th – cumulative number of cases up to March 24th, then to make this a percentage ratio the result is divided by the cumulative number.

$$\% = \frac{\#new}{cumulative} = \frac{total\ cumulative}{cumulative} = \frac{688 - 503}{503}$$

This very simple computation is one of the most important data values to watch in real time, as while the number of new cases may still be a large value it is the percentage change that really tells the story, really tells when the spread is slowing. Obviously, in the early days of the disease spread that value will be rapidly changing due to small numbers, but once the progression of the disease continues it is noted that this number becomes more stable. Namely, in the New York City data this percentage change was steadily growing up to March 26th , when it reached its maximum value of just under 37%. Then, the value steadily declined over the next few days, 30.4% on March 27th and 29.5% on March 28th . The value continued a steady decline reaching 19.7% on April 1st and then falling to 9.3% on April 8th staying below 20% for the reminder of the month. Thus, this numerical value is a measure of when the spread of the disease starts to slow, and as one can see within this timeline the peak corresponds to the time shortly after the strict social distancing measures were put in for place for NYC residence. Furthermore, it is debatable as to what measure is the most accurate to use, the number of infected or the number of deaths, and while each is now without some doubt it is generally accepted that the number of deaths is more practical in real time as in order to actually know the total number of infected persons each member of the population would be needed to be tested and it is not practical to do so in real time. However, to study the models in the prior section, related to data analysis methods, a data fit for $I\left(t\right)$ is desired, thus we will conduct one now but not address the actual validity of the data obtained nor any corrections to the data values due to advanced sampling methods that could be applied.

If a simple regression model was run on the natural log of the data of the number of cases during the month of March, the following result is obtained.

$$\ln\left(\hat{y}\right) = 0.37x + 1.62$$

Thus, one can back solve this to see the approximate exponential growth model for the number of infected cases as

$$I\left(t\right) \approx I_0 e^{0.37x}$$

And, from this one can compare to the either one of two things: the formal solution to the methods learning in the prior section, hence matching parameters, or a training/testing data set from current time data. Either way, the model should be effective at making short term predictions on the future spread of the data. It would be at interesting study to look at this same exercise, but at different times in the future; hence, one could conduct a post hoc type statistical analysis to look if measure taken by local authorities had any effect of the spread?

Exercise Answers

ANSWERS TO CHAPTER 1 EXERCISES

1. $s_1^2 = 125, s_2^2 = 17$
2. The larger the variance, the more spread the data.
3. 69<x<81
4. 85<x<115, max value 115

ANSWERS TO CHAPTER 2 EXERCISES

1. 4,200 choices
2. 35,152 codes
3. 32,760 ways
4. a. 35,254,642,500 ways, b. 21,646,947,168,000 ways, c. 65,536 ways
5. $\frac{1}{30}$
6. $\frac{1}{33}$
7. $\frac{12}{13}$
8. $\frac{23}{38}$
9. $\frac{1}{8}$

10. $\frac{91}{990}$
11. $\frac{4}{5}$
12. a. $\frac{2}{5}$, b. $\frac{71}{150}$, c. $\frac{37}{50}$, d. $\frac{3}{10}$, e. $\frac{43}{75}$

ANSWERS TO CHAPTER 3 EXERCISES

1. a. discrete, b. continuous
2. $p_k > 0, \quad sump_k = 1$
3. $f\left(x\right) \geq 0, \quad \int_{-\infty}^{\infty} f\left(x\right) dx = 1$
4. a.

x	1	2	3	5
p(x)	.4	.1	.3	.2

Valid Probability Distribution

b.

x	0	1	2	3
p(x)	.3	.2	.4	.2

Not a Valid Probability Distribution

1. a. $p\left(3\right) = 0.1$, b. x=2, c. 0.62, d. 0.88

1. a. 1.6, b. 1.05
2. a. 0.00078, b. 2
3. a. binomial; 0.39, b. not a binomial; probability of success is not the same for each trial. (answers may vary)
4. a. binomial, b. 72.8, , 6.552, 2.560 c. 0.120, d. 0.999≈ 1, e. ≈ 0

ANSWERS TO CHAPTER 4 EXERCISES

1. a. $3e^{-3x}$, b. 0.988, c. 1, d. $\Gamma(1)$, e. 0.9985 hrs after 11am (ie. by 12pm)
2. a. 0.5, b. 0
3. a. $y = x^2 + 2x + 3$, b. $y = \frac{3}{1390}\left(x^2 + 2x + 3\right)$
4. a. $\begin{cases} \frac{3}{8}x^2, & 0 \leq x \leq 2 \\ 0 & elsewhere \end{cases}$, b. 0.125, c. 1.5, d. 0.15
5. a. 0.5, b. 0.05

ANSWERS TO CHAPTER 5 EXERCISES

1. $\lim_{n\to\infty} \frac{\sigma}{\sqrt{n}} = 0$
2. a. $\int_{-z_{0.1}}^{z_{0.1}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 0.8$, b. $\int_{-z_{0.025}}^{z_{0.025}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 0.95$, c. $\int_{-z_{0.005}}^{z_{0.005}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 0.99$
3. Yes, there is a statistically significant difference.
4. a. 3, b. 3
5. 0.28868
6. a. 2, b. 2
7. a. $\frac{1}{7}$, b. $\frac{1}{49}$, c. $\frac{7}{7-t}$; $\frac{1}{7}$; $\frac{2}{49}$, d. $\frac{1}{49}$, e. $\frac{2}{49}$
8. a. $\frac{3}{4}$, b. $\frac{1}{3t}\left(e^{3t} - 1\right)$; $\frac{3}{2}$; 3 , c. $\frac{3}{4}$