Probability, Statistics and Mathematical Reasoning

Aaron Manning

Version 0.0.3

Contents

1 The Field of Statistics **9**

2 General Descriptive Statistics **11**
- 2.1 Measures of Central Tendency . . . 11
  - 2.1.1 Mean . . . 11
  - 2.1.2 Median . . . 12
  - 2.1.3 Mode . . . 13
- 2.2 Measures of Spread . . . 13
  - 2.2.1 Range . . . 13
  - 2.2.2 Variance . . . 14
  - 2.2.3 Standard Deviation . . . 16
  - 2.2.4 Average Absolute Deviation . . . 16
  - 2.2.5 Selecting a Statistic to Measure Spread . . . 17
- 2.3 Terminology with Regard to Describing Shape of Uni-variate Data . . . 17
- 2.4 Further Descriptors of Data . . . 19
  - 2.4.1 Quartiles, Deciles and Percentiles . . . 19
  - 2.4.2 The Inter-Quartile Range and Outliers . . . 20
  - 2.4.3 Z-Scores . . . 21

3 Basic Set Theory **23**
- 3.1 What is a Set? . . . 23
- 3.2 Venn Diagrams . . . 23
- 3.3 Set Operations and Definitions . . . 24
  - 3.3.1 Set Union . . . 24
  - 3.3.2 Set Intersection . . . 24
  - 3.3.3 Set Difference . . . 25
  - 3.3.4 The Complement of a Set . . . 25
  - 3.3.5 Subsets . . . 25

4 Probability and Combinatorics **27**
- 4.1 Equal Odds Scenarios . . . 27
- 4.2 Graphical Representations of Probability . . . 28
- 4.3 Multiplication Rule of Probabilities . . . 29
- 4.4 Addition Rule for Probabilities . . . 31
- 4.5 Probability Trees . . . 32
- 4.6 Bayes Theorem . . . 33
- 4.7 Thinking Intuitively About Probability . . . 37
- 4.8 Fundamental Rule of Counting . . . 38
- 4.9 Permutations . . . 39
- 4.10 Multinomial Coefficient . . . 40

4.11 Combinations . . . 41
    4.11.1 Combinations Without Repetition . . . 41
    4.11.2 Combinations With Repetition: The Stars and Bars Approach . . . 42
4.12 Pascal's Triangle . . . 44
4.13 Binomial Expansion . . . 44

5 Discrete Random Variables 47

6 Continuous Random Variables 49
6.1 Representing Continuous Data . . . 49
6.2 Calculating Statistics from Continuous Distributions . . . 51
    6.2.1 Probability of Given Outcome . . . 52
    6.2.2 Expected Value . . . 52
    6.2.3 Mode . . . 53
    6.2.4 Variance and Standard Deviation . . . 54
6.3 Cumulative Distribution Function . . . 55
    6.3.1 Quantiles . . . 56
6.4 Generating Random Numbers from a Probability Density Function . . . 56

7 Normal Distribution 59
7.1 What is the Normal Distribution? . . . 59
7.2 Empirical Rule . . . 60
7.3 Z Tables . . . 61
7.4 Proof: Area Under the Normal Distribution . . . 64

8 Binomial Random Variables 67
8.1 Bernoulli Trial . . . 67
8.2 Binomial Distribution . . . 69
    8.2.1 Mean and Variance of the Binomial Distribution . . . 70

9 Further Descriptive Statistics 75
9.1 Moments . . . 75
    9.1.1 Raw and Crude Moments . . . 75
    9.1.2 Central Moments . . . 76
    9.1.3 Standardised Moments . . . 76
9.2 Pearson's Moment Coefficient of Skewness . . . 76
9.3 Other Methods of Quantifying Skewness . . . 78
9.4 Kurtosis . . . 79

10 Sampling Distributions 83
10.1 What is a Sampling Distribution? . . . 83
10.2 Sampling Distribution of the Sample Mean . . . 83
10.3 Sampling Distribution of the Sample Proportion . . . 84
    10.3.1 Perfect Sampling Distributions . . . 85
10.4 Central Limit Theorem . . . 85
10.5 Standard Error . . . 86

Future Changes and Updates

As is probably clear from the version number on the first page of this text, it is currently a work in progress and I am trying to regularly update it with new content and improvements. This chapter is just series of lists of planned changes, updates and new content.

Please note that at some points in the text, there are references made to later sections where instead of the section or chapter number, all that appears is ??. This means that that section will be in a future release but does not yet exist.

Also note that in addition to the planned chapters that have not yet been completed, an Appendix is planned for the end of the book which is referenced on occasion, but is not yet there.

The version numbers are 0.X.X for pre-release versions while version 1.0.0 will be the first release with all chapters until 'Principles of Hypothesis Testing and Research Theory' complete. Once that is complete, I will dedicate more time to improvements in existing content before a version 2.0.0 release to add more content beyond that final chapter.

Planned Sections for Existing Chapters

Further Descriptive Statistics:

- Moment Generating Functions

Sampling Distributions:

- Standard Error Example with Sample Proportions
- Proof of the Central Limit Theorem
- Sampling Distribution of the Sample Variance

Binomial Random Variables:

- Real World Calculations using the Normal Approximation
- Intuition for the Mean and Variance of a Binomial Distribution

Probability and Combinatorics:

- Permutations with Repetition and Smaller Selection Range
- Combinations With Repetition: The Stars and Bars Approach: Restrictions on the Number of Each Option

Planned Improvements to Existing Sections

Continuous Random Variables:

- Generating Data from Continuous Probability Function: More intuitive explanation as to why this works, change example used from previous to better suit example.

Probability and Combinatorics:

- Thinking Intuitively About Probability: Justification of current example with Bayes theorem, new examples (one modified version of current example, other the Monte Hall problem)

Further Descriptive Statistics

- Reworking beginning of the chapter to better show motivation behind the statistics that appear in that chapter

Planned Chapters and Current State of Completion

Within existing text:

- Discrete Random Variables 20%

At the end of the text:

- Confidence Intervals - 0%
- Statistical Estimation Theory - 30%
- Regression - 60%
- Principles of Hypothesis Testing and Research Theory - 60%

Preface

This textbook is designed to be a reference for many concepts in statistics, providing an in depth explanation as to the core principles behind many ideas in statistics and mathematical reasoning. In general, the content within any part of this text will build upon the content of the previous chapters. However, occasional diversions will be made, requiring content from later on in explanations alongside references to sections with additional detail, but this will be kept to a minimum in the interest of this test being a teaching tool more than anything. Despite this, many concepts will be introduced in a way that is very dense and complex, in the interest of providing a fully comprehensive understanding of each content and acting as a thorough reference.

As a consequence of this, the following design principles have been applied in the design of explanations for this text:

1. Compromises on depth should not be made in interest of covering content faster.
2. Start with the problem and motivation for the discovery of a concept or formula, and arrive at that formula as a conclusion, rather than presenting the concept to be taught and then justifying it.
3. It is more important to teach for intuition than proof.
4. Select the order of teaching concepts to minimise the need to present concepts for the reader to trust before proving later.
5. It is worthwhile to present common misconceptions and justify their incorrectness rather than just presenting the correct idea.

Throughout this text, *italics* will be used to denote the introduction of new concepts as well as to provide emphasis on important terminology as it relates to new concepts.

Chapter 1

The Field of Statistics

The term *statistics* is often used to describe the broad field of data analysis. However, this term has a specific meaning within the field. When studying data, a statistician can examine the entire group they are studying, called a *population*, or they can examine a subset of the population to make estimations about the nature of the entire population, called a *sample*. Characterises of a population are called *parameters* while characteristics of a sample are called *statistics*. Despite these definitions, the general field of study covering both of these areas has adopted the name *statistics*.

Data can come in many different forms, numerical data for example can be discrete in nature, where variables cannot be meaningfully subdivided, such as the number of leaves on each of a group of sunflowers, or continuous data, where variables can assume any value within a certain domain, such as temperatures. Categorical data refers to a type of data where a quantity is assigned to a variety of categories, to quantify the similarities and differences between them, such as the number of people who voted for any given candidate in an election. Data can also be collected with multiple variables that are related to each other, such as the way the value of a commodity changes over time.

Within the field of statistics, different methodologies are categorised as being a part of two groups, *descriptive statistics* and *inferential statistics*, of which, the former is used to describe or give characteristics of a data set, while the latter is used to influence decision making or provide answers to higher level questions or hypothesis, as well as to make predictions about unknown data.

Chapter 2

General Descriptive Statistics

2.1 Measures of Central Tendency

Measures of central tendency are statistics interested in quantifying the centre of a data set. This section will examine three measures of central tendency in terms of how to calculate them and when they are most useful.

2.1.1 Mean

The most commonly used measure of central tendency is the *arithmetic mean*, often just referred to as the *mean* or the *average*. This statistic is calculated as the sum of all data points divided by the quantity of them.

$$\mu = \frac{\displaystyle\sum_{i=1}^{N} x_i}{N}$$

Within this calculation, μ represents the population mean, N represents the population size (or number of measurements) and x represents a generic data point to be indexed by it's subscript.

The benefit of the mean as a measure of the centre of a data set is that the calculation is comprised of *all* data points.

The notation used when calculating the mean of a sample differs to that of a population to distinguish between the information being presented.

$$\overline{x} = \frac{\displaystyle\sum_{i=1}^{n} x_i}{n}$$

Other than the use of $\overline{x}$ for the sample mean, n is also used to describe the sample size as opposed to N for a population size.

One of the limitations of using the mean to measure centre is the influence of *outliers*, or data points that sit significantly further away from the rest of the data, covered further in 2.4.2.

Consider the following data set of average maximum temperatures per observation centre in Sydney throughout 2019 (Note: This data set will be used elsewhere within this chapter).

25.2 25.7 25.5 24.1 24.8 24.0 24.6 24.5 25.2 19.0 23.7 26.1 23.8 24.0 24.5 21.9 25.0 22.7

The calculation for the mean of this data set will be as follows:

$$\overline{x} = \frac{25.2+25.7+25.5+24.1+24.8+24.0+24.6+24.5+25.2+19.0+23.7+26.1+23.8+24.0+24.5+21.9+25.0+22.7}{18} \approx 24.128$$

In this calculation the mean assumes the units of the original data, in this case, degrees Celsius.

2.1.2 Median

Despite how *median* will often be described in entry level statistics, the median of a data set is the value (which is not necessarily a data point within the set) which divides the data set into two equal portions. There are many different ways to calculate the median of a data set, this section will cover the calculation with respect to a collected, specified data set of finite length. With respect to the above set of temperature data, the first step to calculating the median is to order the measurements in ascending order.

19.0 21.9 22.7 23.7 23.8 24.0 24.0 24.1 24.5 24.5 24.6 24.8 25.0 25.2 25.2 25.5 25.7 26.1

For a data set with an odd number of observations, the median is the middle observation. From a data set with an even number of observations, the median is the mean of the middle two observations.

To formalise this method, calculate a variable k, using $k = \frac{N}{2}$. In cases where k is an integer, the median, $\tilde{x}$, is calculated using:

$$\tilde{x} = \frac{x_k + x_{k+1}}{2}$$

In cases where k is not an integer the median is the data point at the index of the smallest integer greater than k or:

$$\tilde{x} = x_{\lceil k \rceil}$$

The key difference between the median and mean of a data set with respect to measuring centre is that the median is not influenced by the raw values of any data points except those in the middle of the set, thus significant outliers have no bearing on this calculation.

2.1.3 Mode

The *Mode*, sometimes denoted by *Mo*, of a data set is the most frequently occurring value within that set. Once again, calculating this with respect to the temperature data set, there are three modes, being 25.2, 24 and 24.5, each occurring twice.

There are many reasons why the mode is an inappropriate descriptor of this data set. The first of which being it's continuous nature, constrained by the accuracy of each measurement to one decimal place. Thus these measurements that appear to have occurred more than once only appear that way due to the limitations of the data collection. Additionally, given that there are three data points with the same number of occurrences, each only one higher than the rest, the mode is very loosely descriptive of the data. Here is an example of a generic data set in which the mode would be more descriptive and useful:

$$1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 3 \quad 4 \quad 6 \quad 7 \quad 9 \quad 12$$

Here the mode is 1, showing where the *plurality* of data points lie. This is useful in identifying the most likely outcome in collecting further samples. Despite mode being a useful descriptor in this scenario, it is very rare to use it in isolation. It can however act as a secondary descriptor to provide more details regarding the shape of a distribution.

2.2 Measures of Spread

Now that we have a variety of methods to describe the 'centre' of a data set, it is now important to describe where the rest of the data sits with respect to that data, of which there are also many different techniques called *measures of spread*.

2.2.1 Range

The most simple measure of spread is the *range* of a data set. The range is a measure of the difference between the greatest and least data point within a data set, or the size of the potential input space, and just as with the measures of central tendency covered above, it is measured in the units of the original data.

$$\text{Range} = x_{\mathsf{max}} - x_{\mathsf{min}}$$

It is important to note that when discussing *continuous probability functions*, often the term range is used to refer to the output space of the function while *domain* is used to describe the input space. While this is covered elsewhere within this text, the term range, when discussing a calculable statistic of a given data set, refers to the above definition and corresponding formula.

2.2.2 Variance

While range, as a measure of spread, communicates the size of the set of potential values within a data set, it makes no attempt to describe the shape of the data within that range. Furthermore, outliers, are the determining factor of the range as only two data points are used in the calculation. This is where the utility of *variance* becomes clear. Variance, also referred to as average square deviation, is a measure of the average squared distance of each data point from the mean and is thus calculated as follows:

$$\sigma^2 = \frac{\sum\limits_{i=1}^{N}(x_i - \mu)^2}{N}$$

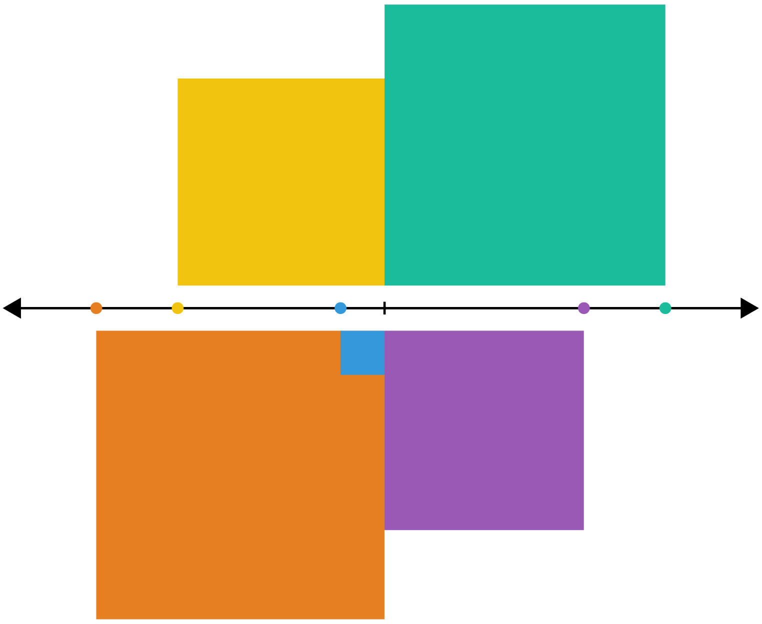
To gain an intuition for what this formula is calculating, let's examine the following small data set:

1 2 4 7 8

For which the variance would be calculated as follows:

$$\sigma^2 = \frac{(1-4.4)^2 + (2-4.4)^2 + (4-4.4)^2 + (7-4.4)^2 + (8-4.4)^2}{5} \approx 2.728$$

This number line shows each data point along with the mean, and the distance between each point and the mean marked with a line, each serving as one edge to a square.



Variance, is the mean area of the squares within this diagram. As a consequence of this, the units given to variance are the square of the units used within the data set. This is further covered in section 2.2.5.

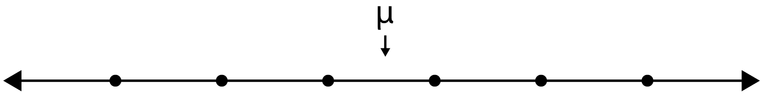
As a further note to calculation of variance, in the above formula, σ^2 is used to denote the variance of a *population* when the entire population is represented in the data set. A different calculation is used for sample variance. It is important to note that if the goal is to calculate the variance *of a sample*, the formula for population variance is still used, as in that case, the sample is the population being examined. However, when trying to *estimate* of the population variance *from a sample*, the below formula is used instead, where s^2 denotes a *sample variance*.

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

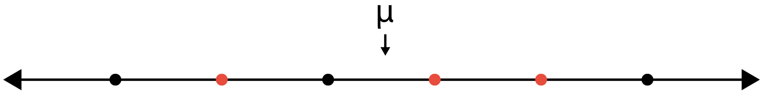
Here the calculation only changes with respect to the denominator of the fraction, using $n-1$ *degrees of freedom*, a term used to describe the number of parameters in a calculation that are free given the result is determined. When sample variance is calculated using n degrees of freedom, it is considered a *biased* estimate of population variance. While this change, called Bessel's correction, may seem counter-intuitive, in the vast majority of cases it provides a better estimate of population variance.

Intuition for Bessel's Correction

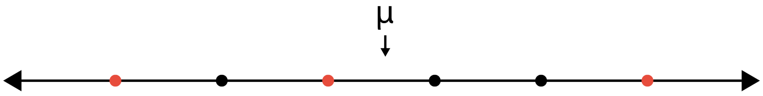
To illustrate why this is the case, we will use the following number line, with data points and the population mean marked along it.

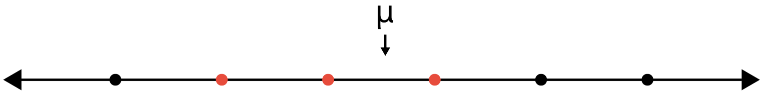


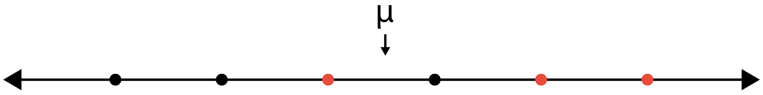
Now we will take a random selection of three data points as a sample.

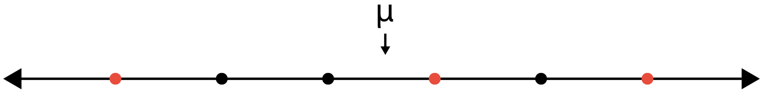


If we make a calculation of variance using the selected sample, the *population mean* and n as the divisor, the variance should be a perfectly unbiased estimate of the population variance. Remember, this is *not* how variance is calculated, as when taking a sample it is presumed the population is unknown. To understand why this method of calculating variance provides a good result, examine these other possible selections of samples.









As you can see from the above selections, while some predictions of variance using this method will be greater or smaller than the actual population variance, on average, this type of calculation will predict the population variance almost perfectly.

However, due to the fact that the population mean is unknown, this cannot be used in the calculation for variance. Consider any possible stand in for the mean within this calculation and how it would effect the variance. As this stand in assumes more extreme values further from the data set, the variance increases significantly. So the value to use which will provide the lowest variance, sits within the sample. In fact, the sample mean, shown on the below number line, is the possible value to use for the mean which gives the smallest possible result for variance. Thus, it is very rare for the sample mean to overestimate the population mean, only occurring where the sample is a selection of only the most extreme values, or the sample mean happens to be very close to the population mean. This is accounted for by dividing by $n-1$, increasing the result. A specific proof for the use of division by $n-1$ specifically can be found in section **??**.

Appendix A provides R code to demonstrate this effect through running simulations.

2.2.3 Standard Deviation

You may have noticed that within the calculations for population and sample variance, the symbols σ^2 and s^2 respectively, each have an exponent of 2. This is due to the use of σ and s to denote standard deviation. In accordance with the notation, population standard deviation is calculated as the square root of population variance.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

The same is true for calculations with respect to a sample.

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

There are many reasons why the standard deviation may perform better or worse than variance when measuring spread, these are covered in section 2.2.5.

2.2.4 Average Absolute Deviation

Average absolute deviation, also known as the *mean deviation from the mean* or *mean absolute deviation*, is calculated in a similar way to variance. However, instead of squaring the difference between each data point and the mean, the absolute value of this difference is taken such that each term is calculating the distance between these two points, without influencing their values.

$$\mathsf{MAD}(X) = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

This method of measuring spread is not very common, despite perhaps being the most intuitive and immediately understandable. The following section (2.2.5) examines why any given measure would be chosen over another.

2.2.5 Selecting a Statistic to Measure Spread

While this section is here for reference reasons, much of the terminology used is first covered elsewhere in this book. As a consequence, it may be worthwhile to return to this later.

One reason why standard deviation is often preferred to variance is that the given result assumes the units specified within the data set. For example, if the height of a group of individuals is collected in metres, and the variance calculated, the units for this variance are square metres, whereas the standard deviation is in metres. This however, should not cause confusion as to the way it changes the mathematics function, as the application of the square root function does not undo the process of squaring each distance within the variance. Due to the non-linear nature of the square root function, standard deviation is considered a *biased* measure of spread whereas variance is considered *unbiased*. On the other hand, the variance has many useful mathematical properties when used with respect to random variables covered in section **??**.

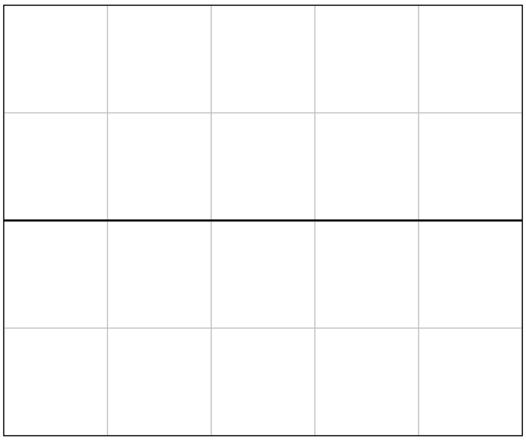
With respect to average absolute deviation, standard deviation has become a preferred option over this due to the fact that it shows reduced deviation throughout the each sample within a sampling distribution. The standard deviation of the sampling distribution of the sample mean will on average be lower than the standard deviation of the sampling distribution of the mean absolute deviation. We can infer from this that the standard deviation of a sample gives more consistent results with respect to estimating a population's standard deviation. Despite this conclusion, there are many reasons why average absolute deviation is a better alternative when it comes to non-ideal distributions.

In actuality, the frequency of use of these different statistics in the calculations of other statistics is a consequence of historical reasons, and the ongoing debates about their usefulness and usage.

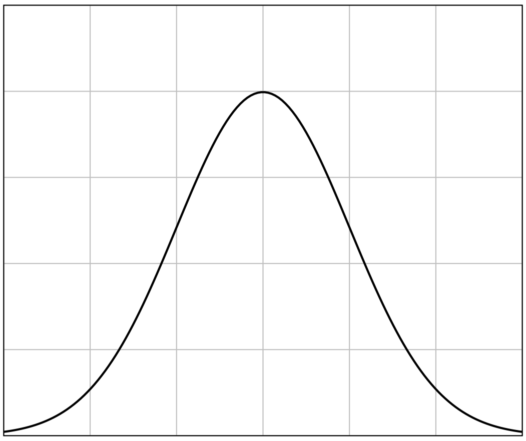
2.3 Terminology with Regard to Describing Shape of Univariate Data

When providing detail about a distribution of data, it is common to describe it's *shape*. This description of shape is a description of the form that the graphical representation of the data takes. While continuous functions will be covered later in Chapter 6, for the sake of this section, understand that the following graphs are a smooth representation of each score on the x axis and the probability on the y axis.

This is a uniform distribution, where all outcomes are equally represented.

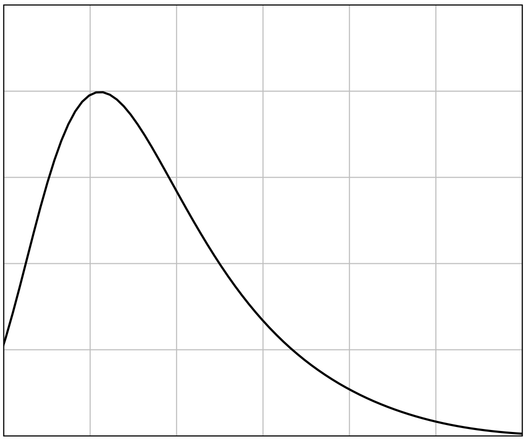


This is what is called a normal distribution. It is worthwhile to note that the normal distribution has a numerical definition, which will be discussed further in Chapter 7.

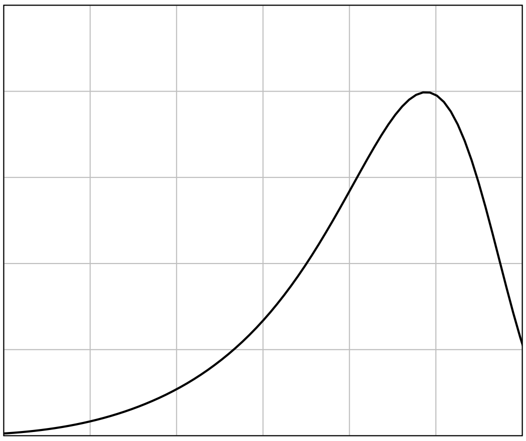


Beyond these simple distributions, many other terms are used to describe the shape of a distribution.

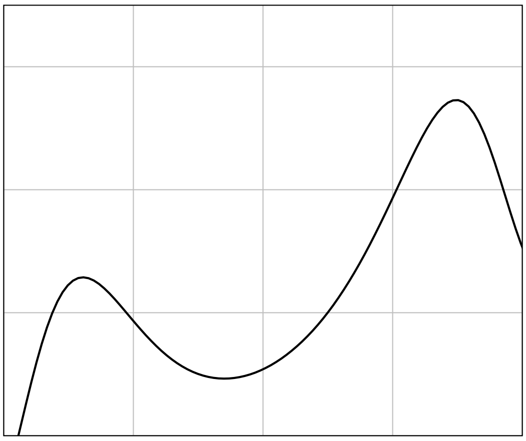
This distribution would be described as positively skewed, where most data sits on the left and the most extreme scores sit on the right.



The opposite is true for a negatively skewed distribution.



This distribution would be described as *bimodal*. While the name would imply that the given distribution has two modes, this term, along with the term *multimodal*, are generally used to describe the number of 'peaks' that the distribution has.



2.4 Further Descriptors of Data

2.4.1 Quartiles, Deciles and Percentiles

Just as the median is a tool to calculate the value which divides a given data set in half, quartile markers, of which the second is the median, are designed to divide the data set into four sections, such that one quarter of the data points sit below Q_1, one half sits below Q_2 and three quarters sit below Q_3.

The method to calculate these markers is similar to that of the calculation for the median outlined in section 2.1.2, the only difference being the calculation of k, for which the coefficient of N is not necessarily $\frac{1}{2}$ but rather $\frac{1}{4}$ for Q_1 and $\frac{3}{4}$ for Q_3.

Consider the following data set of the number of leaves per sunflower across 16 examples.

1 1 2 2 2 2 3 3 4 4 4 4 5 8 10 12

In the case of Q_1, $\frac{N}{4} = 4$, thus $Q_1 = \frac{x_4+x_5}{2} = 2$. Using the same rationale for the other quartiles, $Q_2 = 3.5$ and $Q_3 = 4.5$.

This same principle applies to deciles, which divide a data set into 10 sections and percentiles, which divide a data set into 100 sections.

In many entry level statistics resources, quartiles are taught to be calculated by listing all scores, crossing off the median and then finding the median of the remaining two sections. Here the ambiguity is introduced as to whether or not the global median is included in finding these local medians. Both of these methods are equivalent for data sets with an even number of observations. The method described above, is equivalent to that of including the median. This method, when not including the quartile markers, provides a perfect split for data sets of size $4n+1$ where $n \in \mathbb{Z}^+$. This is not true for the method that discounts the global median. In cases where the data set is of size $4n+3$, neither method provides a perfect 1:3 split but the method described above provides a result closer to this optimum ratio. However, it is important to know that when provided with quartiles for a data set, or calculating them with a computer program, there is no consistent standard.

2.4.2 The Inter-Quartile Range and Outliers

The *inter-quartile range* is defined as the range between the Q_1 and Q_3 markers.

$$\mathrm{IQR} = Q_3 - Q_1$$

With respect the above set of data on sunflower leaves, the $\mathrm{IQR} = 4.5 - 2 = 2.5$.

Within this data set, the highest score of 12 is considered an outlier, as it is significantly greater than the rest of the data set. It is common to remove or 'clean' outliers from a data set before calculating other statistics from it, as it is believed that the outlier was a cause of an improperly conducted experiment, or an instance of the randomness in the specific instance of collection of data.

The following numerical method is used to calculate the *upper bound* and *lower bound* of the *inner fences* of a data set respectively.

$$\mathrm{IF1} = Q_1 - 1.5 \times (IQR)$$

$$\mathrm{IF2} = Q_3 + 1.5 \times (IQR)$$

If a data point is less than the lower boundary of the inner fence, or greater than the upper boundary of the inner fence, it is often considered an outlier.

There is also a numerical definition of what is called an *extreme outlier*, which falls outside of the bounds of the *outer fences*. This is shown below.

$$\mathrm{OF1} = Q_1 - 3 \times (\mathrm{IQR})$$

$$\mathrm{OF2} = Q_3 + 3 \times (\mathrm{IQR})$$

This begs the question as to how data points that fall on the fences should be handled. While this case is exceedingly rare given most practical applications of the method, the majority of literature on the topic, despite not directly addressing the issue, describe outliers as falling *outside* of the fences, being less than Q_1 or greater than Q_3. One such example of this is 'The Basic Practice of Statistics' in which David S. Moore writes, 'Call an observation a suspected outlier if it falls more than 1.5*IQR above the third quartile or below the first quartile.'

Using these methods on the aforementioned data set, the $\mathrm{IF1} = 2 - 1.5 \times (2.5) = -1.75$, hence, there are no scores that sit below the lower boundary. Completing the operation for the upper boundary, $\mathrm{IF2} = 4.5 + 1.5 \times (2.5) = 8.25$. Thus both the scores of 12 and 10 are outliers.

When ‘cleaning’ data of outliers, it is crucial to be honest about the discrepancy between the data that was collected and that which is being presented. This includes keeping the quartile markers at their original values despite the fact that taking out data may change these. In the case of this data set, recalculating the IQR on the modified data set would then deem the next highest score an outlier, despite not fitting the original definition to be considered so.

It is important to note, that while this definition of outliers acts as a rule of thumb to be used on any given data set, many fields of science are not quantitatively precise about the definition of an outlier, only removing an outlier if there is a clear reason with respect to the poor collection of data to do so.

2.4.3 Z-Scores

Z-scores, sometimes referred to as *standard scores*, are a quantifier of the number of standard deviations that a given score sits, above or below the mean of the data set.

$$z_x = \frac{x - \overline{x}}{\sigma_x}$$

Z-scores are most useful in comparing results between different data sets which quantify the same or similar data. Presume Peter receives the same mark on his Mathematics test as he does his Science test. Both data sets are shown below with Peter’s score underlined.

Mathematics:

5 5 5 6 6 6 7 7 7 8 8 8 9 9 9 10 10 10 10 11 11 11 12 12 12 13 13 13 14 14
14 15 15 15

Science:

5 6 6 7 7 8 8 8 9 9 9 9 10 10 10 10 10 10 10 10 10 10 11 11 11 11 12 12 12
13 13 14 14 15

Both of these data sets have a mean of 10 and both have the same minimum and maximum values. Since Peter achieved the same score, he is the same distance from the mean in each case. Despite this, due to the differing shapes of the distribution (with Mathematics test data being more uniform and Science having a more prominent peak), his z-score for the Mathematics test was ≈ 0.651 while it was ≈ 0.877 for the Science test, indicating he performed better on the latter with respect to the cohort.

This is because the z-score can quantify how extreme a data point is with respect to *all* the others and is to a certain extend, directly based on the actual values of the data points, in a way that a measure like percentiles is not. Due to the difference from the mean potentially providing negative values, z-scores also indicate which side of the mean the data point sits on.

Z-scores will also show up as a parameter for further statistical calculations, for example within linear regression calculations.

Chapter 3

Basic Set Theory

This chapter is not designed as a thorough explanation of even very simple concepts in set theory, and ignores even basic identities with respect to set operations. It is here as simply an introduction to the idea of a set, and the language used in describing sets, such that the concepts make sense as they appear within later chapters. However, set theory is by no means a primary subject of this text.

3.1 What is a Set?

In Mathematics, a set is simply an *unordered* collection of *distinct* elements. When defining a set, it must be *unambiguous* whether any selected Mathematical object is in the set or not.

A set is often specified in one of the following ways:

1. A comma separated list of values, for example: $\{1, 2, 3\}$.
2. A variable with conditions on its value, for example: $\{x : x > 3\}$.

Sets can also be defined in a more abstract way.

As a consequence of our definition specifying that the elements are unordered and distinct, the following sets are all equal.

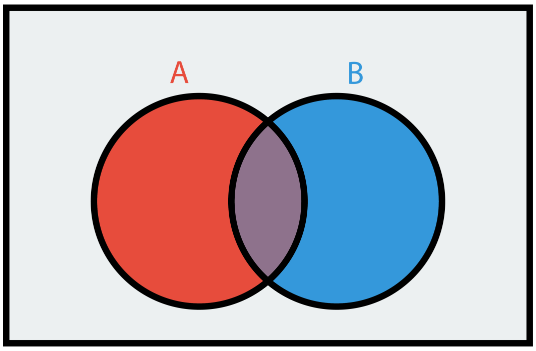
$$\{a, b, c\} = \{a, c, b\} = \{a, a, b, c\}$$

In the context of sets, the claim that two sets are *equal* is equivalent to saying that every element of the first set is within the second, and every element of the second set is

We also use the symbol $\emptyset$ to denote the empty set. Similarly we often denote the universal set by U, where the universal set is the set of everything relevant in the specific context.

3.2 Venn Diagrams

A common, abstract way to represent sets is with a Venn diagram. A Venn diagram represents each set with a circle, such that elements can be placed within a specific region based on which sets they are contained in. Here is a simple Venn diagram showing two sets, A and B.



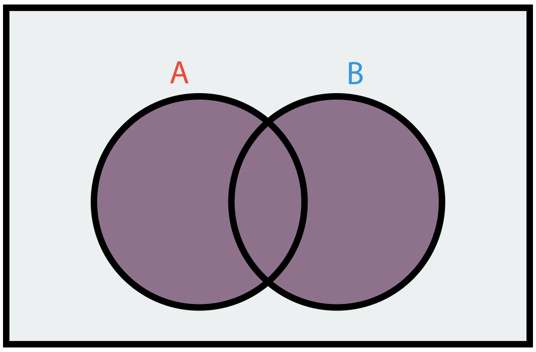
The following section will use this diagram as a base for representing the result of many set operations by colouring the corresponding regions.

3.3 Set Operations and Definitions

The following are definitions of mathematical operations that can be applied to sets, as well as descriptors for the relationship between sets.

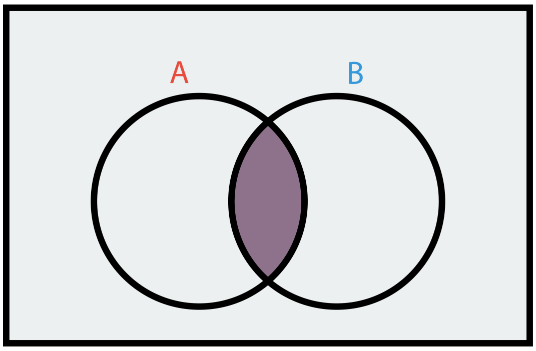
3.3.1 Set Union

The union of two sets, represented by $A \cup B$, is *the set of all elements present in either set*.



3.3.2 Set Intersection

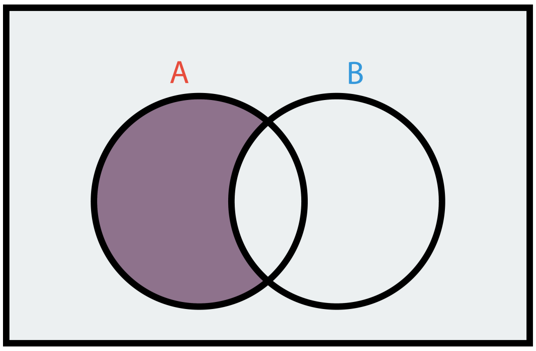
The intersection of two sets, represented by $A \cap B$, is *the set of all elements present in both sets*.



Two sets are considered *mutually exclusive* if their set intersection is the empty set.

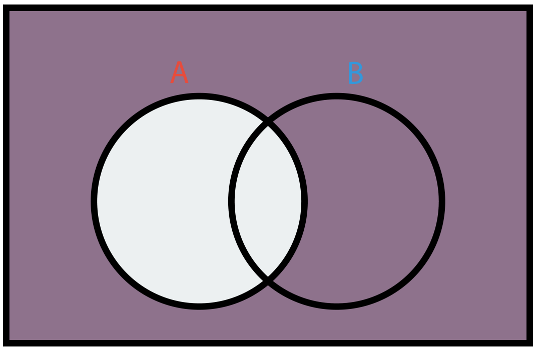
3.3.3 Set Difference

The difference between two sets, represented by $A - B$ or $A \setminus B$, is *the set of all elements present in the first, that aren't in the second*.



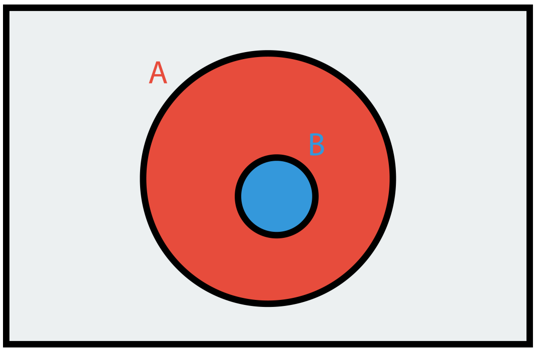
3.3.4 The Complement of a Set

The complement of a set, represented by A^c or $\overline{A}$, is *the set of everything not within that set*.



3.3.5 Subsets

When one set is contained within another it is considered a *subset* of the other. Strictly speaking, this means that if every member of set B is also a member of set A, then B is a subset of A, which is denoted by the expression $B \subseteq A$, and shown in the diagram below:



Note that this does not preclude the case where A and B are identical sets. It is possible to exclude this case by describing B as a *proper subset* of A, denoted by the expression $B \subset A$, which expresses that B is a subset of A and not equal to A.

Chapter 4

Probability and Combinatorics

4.1 Equal Odds Scenarios

The notion of using Mathematical objects to represent probabilities dates back to the the mid 17th century with Pierre de Fermat and Blaise Pascal, who attempted to use the idea of probability to analyse the expected outcome in games of chance. Despite using different approaches, their methods were in agreement with respect to the key underlying principles used to justify further studies of probability. These ideas being that:

1. Regardless of what the probabilities are that are involved in a situation, they stem from the notion of an equal odds scenario, where all outcomes are equally likely. This is because situations in the real world in which different outcomes are not equally likely are very difficult to assign probabilities to, whereas those that have equal odds for all outcomes can be justified *by the symmetry of the situation*. For example when tossing a coin, because of the symmetry of the coin, both the heads and tails sides can be examined using identical logic and reasoning, hence justifying why both outcomes are equally likely.

2. The notion of assigning a *theoretical probability* to given events exists as a way of attempting

to predict outcomes in real world events, but they do not exactly predict such events.

From these ideas comes the fundamental construction of how probabilities are represented Mathematically as follows:

Consider a set of possible outcomes such as rolling a 6 sided die. The entire set of outcomes that are possible within the examined situation is called the *sample space*. In this case, our sample space contains 6 outcomes. Hence we can formalise the idea of an *outcome* as an indivisible possibility in a probabilistic situation. A set of outcomes comprising a subset of the sample space is called an *event*. One such event could be rolling an even number. In general, we use $P(E)$ is used to denote the *probability* of an event E occurring.

In an even odds scenario, the probability of a given event can be calculated as:

$$P(E) = \frac{\text{number of desirable outcomes}}{\text{total number of outcomes}}$$

This formula, although very simple, stems from our observation about the probability of equal events needing to be equal, and the decision that the probability of the sample space should be 1. This decision, if not immediately intuitive, can be viewed as somewhat arbitrary. However, its usefulness will become clear. Regardless, from this we determine that the probability of a given event occurring must be on the interval $[0, 1]$.

Independent Events

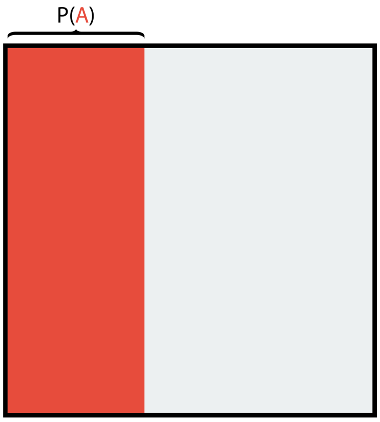
Events are considered independent if neither of them effects each other. The tossing of multiple coins are considered independent as the result of one toss does not influence the other.

Dependent Events

Events are considered dependent if the outcome of one event effects the probability of the other. For example, consider the selection of a playing card from a full deck of 52. The probability of selecting a red card from this deck is $\frac{26}{52}$. Presuming this card is not returned to the deck, the probability of the selecting another red card is $\frac{25}{51}$. Therefore, the second event is *dependent* on the first.

4.2 Graphical Representations of Probability

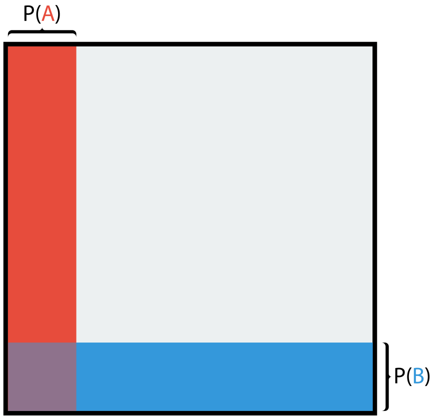
When working with probabilities, many textbooks use Venn diagrams and set notation to denote the dependence and compatibility of events. While this has proven useful as an introduction to set operations, this textbook will opt for what a graphical representation of probabilities that uses area within a square to denote the chance of an event occurring.



Examining this 1 by 1 square, the probability of the given event A is equal to the portion of the area it assumes in the square, which is equal to it's absolute area given the assigned dimensions of the square, and the width of this rectangle due to the length being 1. The applications of this method of displaying probabilities will become clear in the following sections.

4.3 Multiplication Rule of Probabilities

Consider the probability of two non-mutually exclusive events occurring together, where both probabilities are independent of each other. Representing event A on the left hand side of our square in red and event B on the bottom of the square in blue, the probability of event A and event B occurring is the area of their *intersection*, hence the notation used $P(A \cap B)$ is used.



The intersection of the two events in purple has area of the product of its length and width or $P(A)P(B)$.

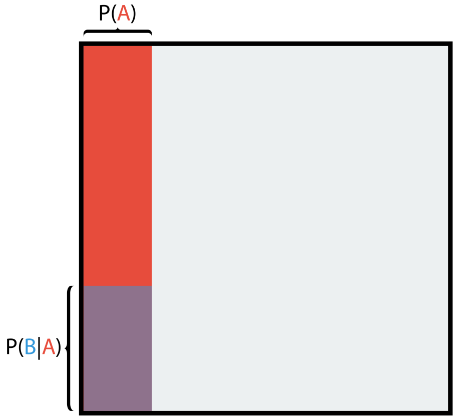
Therefore the probability of the intersection of two independent, non-mutually exclusive events is equal to the product of their probabilities or:

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

However, this formula relies on the above assumptions that the two events are in fact independent and non-mutually exclusive. In the context in which these concepts are often introduced, this is adequate, as many examples are given in terms of probabilities that are truly independent. In actuality, in many real world situations events are not independent. Consider the probability that you and your sibling are diagnosed with a specific disease, given that 1 in 5 people in the population receive such a diagnosis.

Using the above method this result would be $\frac{1}{25}$. However, if someone genetically related to you tests positive for the disease, the probability of you being diagnosed is higher than the expected 1 in 5 and vice versa.

Therefore we must find a way to generalise the formula for the probability of the intersection of two events. Let's consider a situation in which the event B is a subset of A.



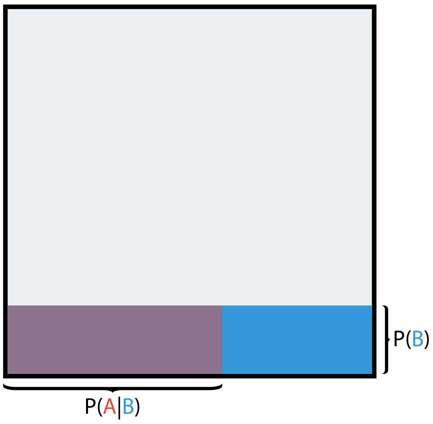
In the above diagram the vertical bar | means *assuming* or *given that*.

In this case the probability of the intersection, equal to the area of the intersection, is equal to:

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

Calculated simply as the product of the length and width of the smaller rectangle.

Note that while we have assumed that B is a subset of A in the above situation, this is done for simplicity and causes no loss of generality since any other cases where B occurs occur outside of the intersection we are interested in. We also have no loss of generality in terms of which event we treat as our assumption.



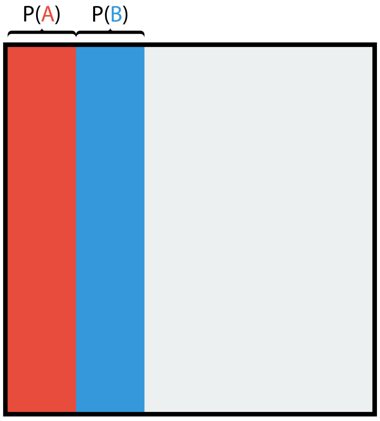
Therefore we can conclude in general that:

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

Note also that if the events are independent as in our original situation, the probability of A occurring given that B has, is simply equal to the product of the probabilities since the condition on the second term does not have any effect on its value.

4.4 Addition Rule for Probabilities

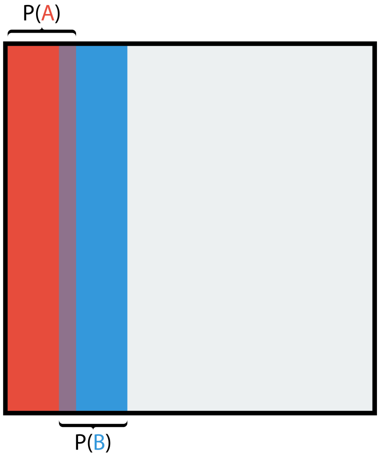
For cases where you desire to calculate the probability of one event *or* another event occurring, the probabilities are additive. This is because the set of possible outcomes is greater than that of either event. The following area shows this case where the events do not intersect.



In cases like this, where events are non-mutually exclusive, the probability of an event in the wider space or *union* of the two events is given by:

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

However, this formula does not account for the possibility of the two events occurring together as shown below.



If calculating the probability of either event occurring using the above method, the intersection of the events is counted twice, once as part of event A and again as part of event B. To correct for this we must subtract the intersection of the events. Therefore:

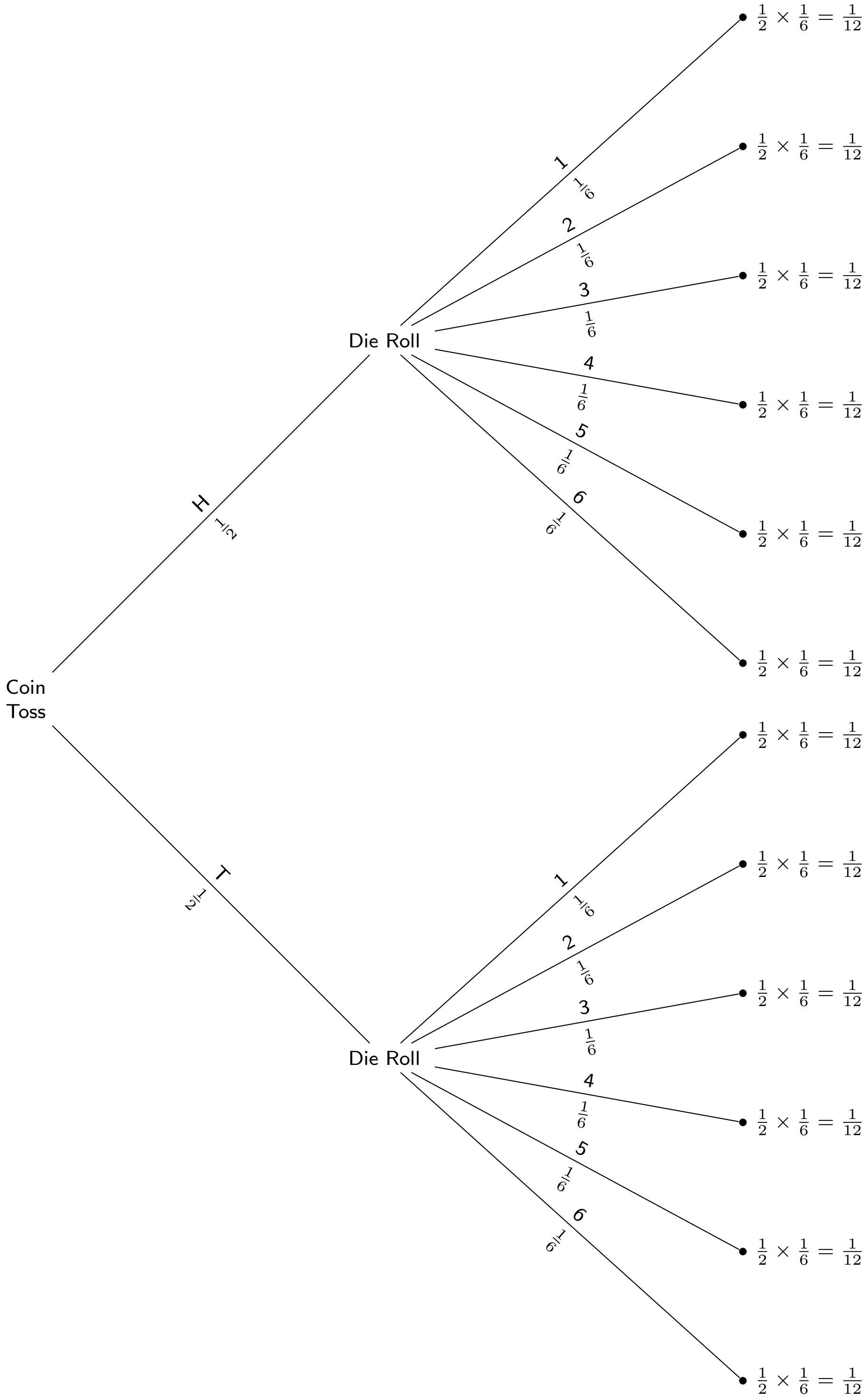
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Since in the first case there was no intersection of the two events, $P(A \cap B) = 0$. Therefore this formula works regardless of if the events are mutually exclusive or not.

4.5 Probability Trees

Probability trees are a graphical way of representing multi-stage events, allowing the calculations of combinations of events to be done intuitively.

Consider a two stage event of flipping a coin and subsequently rolling a die. These two events are independent and therefore the probability of any given outcome on the second set of branches are the same in each connection to the first set.


Coin Toss
H
1/2
T
1/2
Die Roll
1
1/6
2
1/6
3
1/6
4
1/6
5
1/6
6
1/6
$\frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$
Die Roll
1
1/6
2
1/6
3
1/6
4
1/6
5
1/6
6
1/6
$\frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$

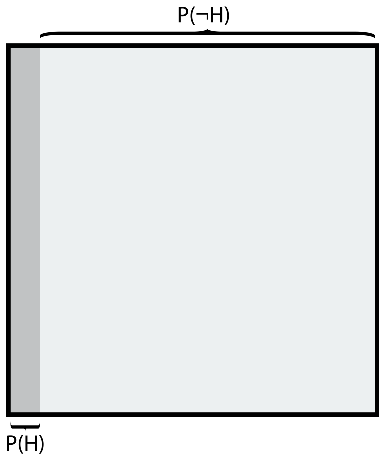
Given that these two events are different, all final outcomes are entirely unique. If the two events were the same, such as the tossing of two coins, the cases of heads first then tails and tails first then heads would need to be examined separately.

Since the total probability of the whole sample space must be equal to 1, the sum of all probabilities at the end must equal 1 and the sum of all connections to any given branch must equal 1.

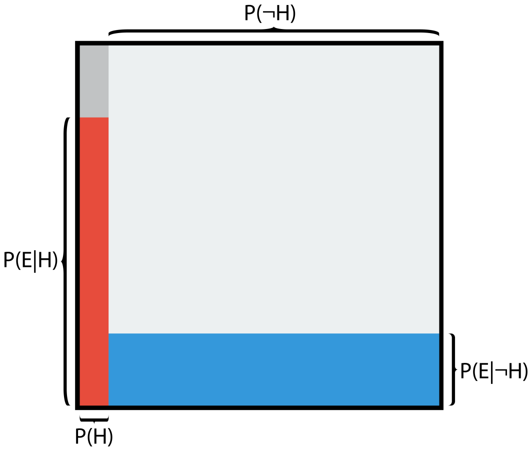
4.6 Bayes Theorem

A patient is given a diagnosis for a disease that effects 8% of the population. The patient aims to figure out the probability that they have the disease, by accounting for the margin of error in the test. The doctor notes that of all the people tested with the disease, 80% test positive and 20% test negative. Of those who do not have the disease, 20% are given a false positive diagnosis while 80% are correctly identified as to not having the disease. The question being examined here is, given this information, what is the probability that the patient has the disease? The intuitive and common but incorrect answer here is 80%, the probability that the test is correct. This situation provides the use case for Bayes theorem which provides the probability of a hypothesis being true given a new piece of evidence *and the previous probability of the hypothesis being true*. This is the part of the calculation which is often forgotten. In the above example, many people neglect to recognise that *the vast majority of people, regardless of diagnosis, don't have the disease*.

Let's derive a solution to the above problem constructively. Consider the following area divided up to represent the probability of the hypothesis that a randomly selected individual has the disease. Hence, our square is broken up into portions of size 8% and 92%, or the proportions of those infected and not infected in the population.



The probabilities of receiving a positive test result in the case where the patient has the disease and the case where the patient doesn't are 80% and 20% respectively. Shading these corresponding areas on the diagram, labelling them accordingly as $P(E|H)$ and $P(E|\neg H)$ or the probability of the *evidence* occurring (a positive test result) given both categories, that where the hypothesis is true and that where it is not.

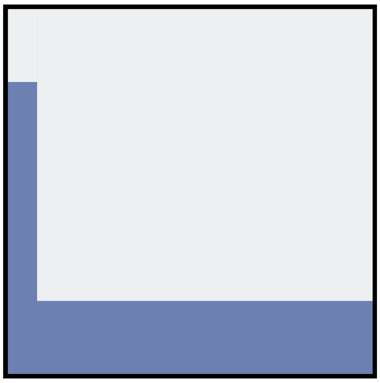


Given this information, the probability of an individual being infected given they've received a positive test result is the area of the red section (representing those with a positive test result who have the disease) divided by the sum of areas of the red and blue sections (representing everyone with a positive test result).

In a general sense, this method can be expressed as:

$$P(\text{having the disease given a positive test}) = \frac{\text{red area}}{\text{red area} + \text{blue area}}$$
$$P(H|E) = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)}$$

Examining the numerator of this fraction, the probability of the hypothesis being true and the evidence being true, plus the probability of the hypothesis not being true and the evidence being true, covers all situations in which the evidence is true and is therefore just the probability of the test being true. To visualise this consider the probabilities of the evidence being true and the evidence not being true as two separate areas.



Therefore the formula can be simplified as follows:

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

This is the general form for Bayes theorem as it is most commonly expressed.

In the case of the above example, the probability of the hypothesis being true that the patient has the disease can be calculated as:

$$
\begin{aligned}
P(H|E) &= \frac{P(H)P(E|H)}{P(E)} \\
&= \frac{0.08 \times 0.8}{0.8 \times 0.08 + 0.2 \times 0.92} \\
&\approx 0.258
\end{aligned}
$$

Moving outside of the geometric representation of these probabilities, for a simple algebraic proof for this formula consider the probability of two events occurring together, covered in section 4.3, where, in this case, the generic events A and B have been replaced with H and E.

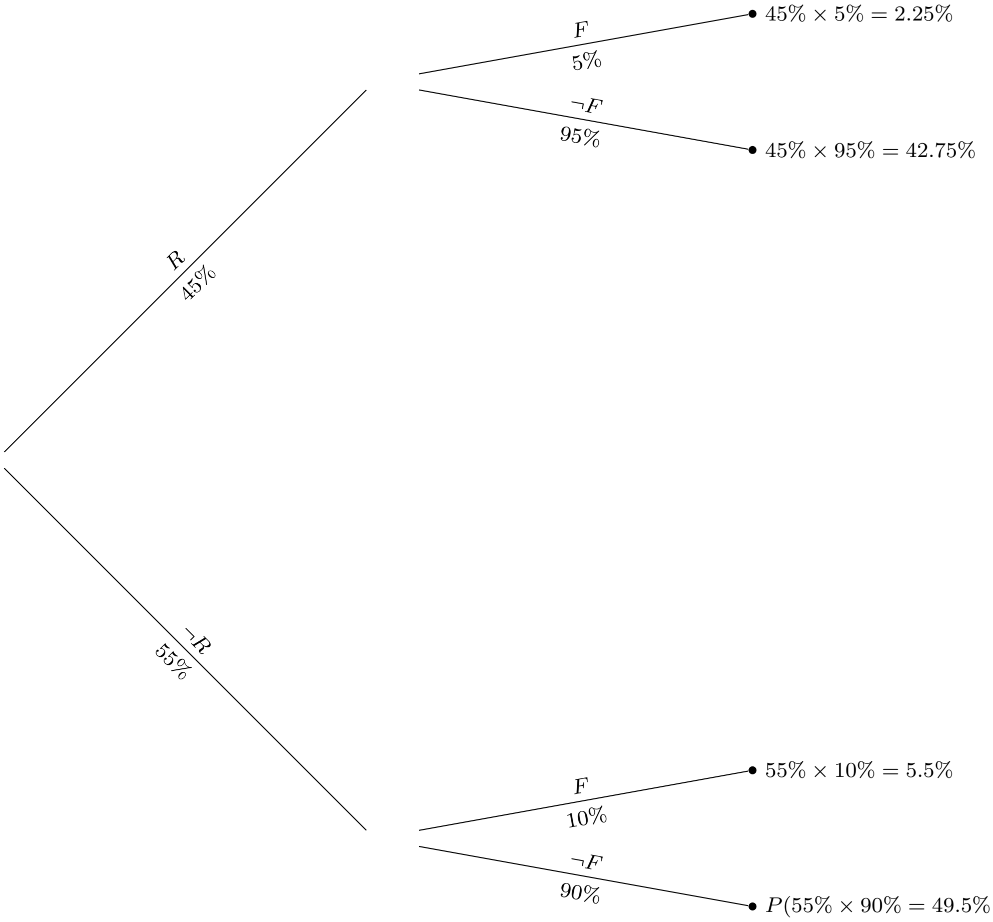
$$P(H \cap E) = P(E)P(H|E) = P(H)P(E|H)$$

Expressing the latter two terms of this expression as one equation and rearranging for $P(H|E)$ arrives us to the result of Bayes theorem.

$$
\begin{aligned}
P(E)P(H|E) &= P(H)P(E|H) \\
P(H|E) &= \frac{P(H)P(E|H)}{P(E)}
\end{aligned}
$$

The applications of this theorem are incredibly broad and general. Bayes theorem is useful in describing any situation under which new evidence is used to inform probabilities with which existing information is known. Another application of Bayes theorem is to more generally reverse the order of dependent and independent variables in a probability tree.

Assume that the probability for an individual in their profession receiving a pay raise in any given month is 45%. If this person receives a pay raise, their probability of being fired that month is 5%, otherwise it is 10%. Representing this with a probability tree:



Bayes theorem allows us to use the fact that someone was fired in the previous month to estimate the probability that they received a pay raise earlier that same month.

Lets consider the information that we already have, starting with the probabilities of receiving a raise from the first two branches:

- $P(R) = 45\%$
- $P(\neg R) = 55\%$

On the first branch we make the assumption that the individual has received a raise, and therefore we have the corresponding probabilities of being fired under this assumption:

- $P(F|R) = 5\%$
- $P(\neg F|R) = 95\%$

Similarly, on the second branch we make the assumption that the individual has not received a raise, and therefore we have the corresponding probabilities of being fired under this assumption:

- $P(F|\neg R) = 10\%$
- $P(\neg F|\neg R) = 90\%$

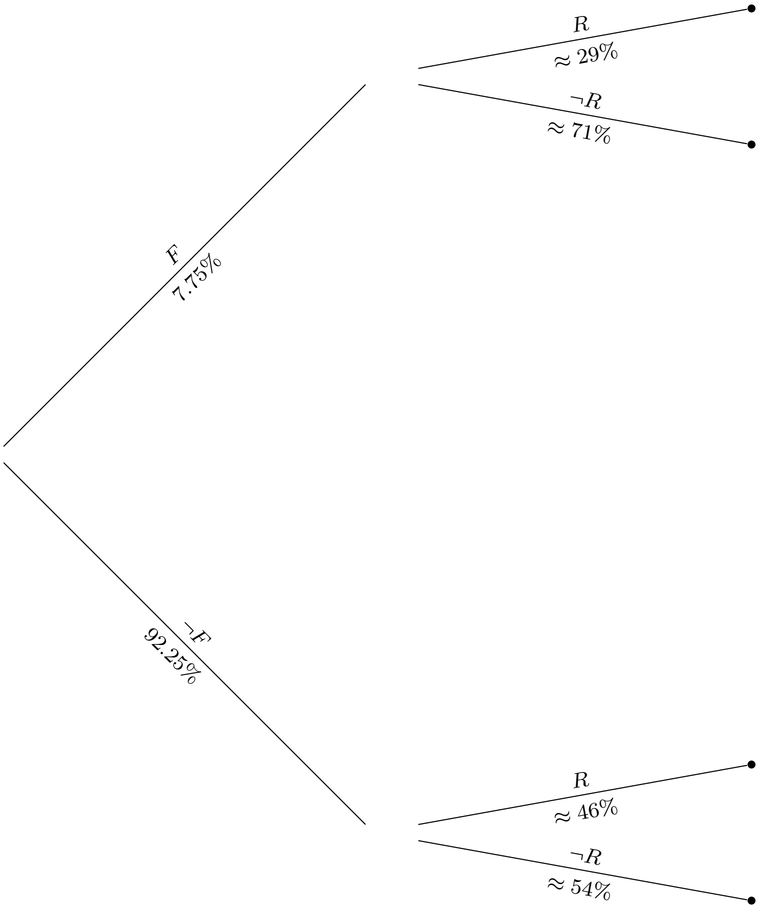
The probability of being fired or not fired is equal to the sum of events where that conditions is true (check the values at the end of the probability tree to make sure you understand where each of these come from):

- $P(F) = 2.25\% + 5.5\% = 7.75\%$
- $P(\neg F) = 42.75\% + 49.5\% = 92.25\%$

To find the probability of a pay raise given a subsequent firing, apply Bayes theorem:

$$
\begin{aligned}
P(R|F) &= \frac{P(R)P(F|R)}{P(F)} \\
&= \frac{45\% \times 5\%}{7.75\%} \approx 29\%
\end{aligned}
$$

Making these calculations for each possible combination a new probability tree can be drawn which reverses these conditions:



Try calculating the probabilities at the end of this tree for yourself and compare them to those from the first tree. What do you notice?

4.7 Thinking Intuitively About Probability

There are many aspects to probability problems that often seem counter intuitive, where two seemingly valid explanations yield different responses. This section will purely use examples to show how to approach these problems in a logical way, and how to apply the previously covered rules of probability to verify these solutions.

Two Coins

Many probability problems come down to a relationship between the order of events, and the information that is available. This scenario is a simple permutation of a variety of problems contingent on the same type of rationale.

Consider there are two coins, one is a standard coin with heads on one side and tails on the other, the other coin is atypical in that it has two heads sides. Consider that you close your eyes and randomly select a coin, covering it after it lands and removing the other coin. After looking at it you see that the coin landed heads. What is the probability that the other side of the coin is also heads?

There are two common approaches to this question with different results:

- The probability is $\frac{1}{2}$, as the coin that landed is one of two options, and this dictates the other side.
- The probability is $\frac{2}{3}$, as a head was observed, and two of the three heads are on the coin with which the other side must be heads.

I would encourage any readers at this point to sit and ponder the two scenarios, one of which is correct, and try and justify why one answer is correct and the other isn't.

The important trick to this problem is remembering that while the coin didn't land tails to start with, it was a possible result from the start that we happen to be uninterested in. Consider that there are four outcomes for the top of the coin, where each side of each coin is one of the four outcomes. Given that the coin selection is 50/50, and the coin toss is 50/50, the probability of each of these four outcomes being to topside of the coin that is observed is 0.25, and therefore the probability of observing heads on the top of the coin is 75%. Since, within this outcome, more heads are on the coin with two heads, the fact that you see heads on the top of the coin inherently gives you information about the likely-hood of it being a specific coin. The claim that the probability of the opposing side being heads is $\frac{1}{2}$ implies that the likely-hood of getting either coin is the same, which is true in the abstract, but not when you consider that had the top of the coin shown tails, the result wouldn't be counted.

4.8 Fundamental Rule of Counting

The fundamental rule of counting is a simple rule at the foundation of many counting problems. Consider that when choosing articles of clothing in the morning, you have 7 options for a shirt to wear and 9 options for a hat. How many possible unique shirt and hat pairs are possible? To many readers the answer to this question may seem very obvious, and if that is the case, I encourage you to, before reading on, justify to yourself why you arrived at the answer that you did, in terms of the fundamental ideas that you can.

Let's first ignore the options for hats, and therefore note that a possible of 7 different shirts could be chosen. Since each of these cases are unique, we can examine them independently. Given the selection of a given shirt, there are then 9 possible choices for hats. This is true regardless of which shirt was originally chosen, with the hat options being identical in every case. Hence, for each of the 7 possible selections of shirt, there are 9 possible selections for a hat, and therefore there are a total of $7 \times 9 = 63$ possible unique pairs.

This is an example of the fundamental rule of counting, which more generally states that if there are n ways to do one thing, and m ways to do another thing, there are $n\ timesm$ ways of doing both. This may look familiar to the rule in probability for the intersection of two independent events occurring.

4.9 Permutations

A *permutation* is an arrangement of a set of objects where the resulting order *is important*.

Permutations With Replacement

Consider a combination lock, in which digits from 0 to 9 can be selected and a total of 4 digits are required. By viewing this as a series of four slots, each which can be filled with any 1 of 10 different digits, the total number of options for each slot is 10 and thus the total number of options across the four slots is 10^4. Generalising this method, we can calculate the number of permutations as n^r where n is the number of objects in the set and r is the number of times an object is being selected.

Permutations Without Replacement

Now consider a situation in which the selection of an object removes it from the set of available options before the second selection. Consider a bag containing n distinct cards where cards are being drawn in sequence. The number of possible permutations in this situation is:

$$\begin{aligned} {}^{n}P_n = P(n,n) &= n(n-1)(n-2)...1 \\ &= n! \end{aligned}$$

For now don't worry about the notation used to represent this calculation. Here ! represents the factorial operation, which is defined only for *natural numbers* (positive numbers and 0) and is equal to the product of itself and all positive integers less than it for positive numbers and 1 for 0. The latter is because there is only 1 way to arrange 0 objects. The application of this operation is useful as after each selection from the bag, the number of possible options is reduced by 1.

Defining this recursively, consider the following function:

$$f(x) = \begin{cases} x \times f(x-1) & \text{for } x \in \mathbb{Z},\ x > 0 \\ 1 & \text{for } x = 0 \end{cases}$$

Therefore, applying this function to 5:

$$\begin{aligned} f(5) &= 5 \times f(4) \\ &= 5 \times 4 \times f(3) \\ &= 5 \times 4 \times 3 \times f(2) \\ &= 5 \times 4 \times 3 \times 2 \times f(1) \\ &= 5 \times 4 \times 3 \times 2 \times 1 \times f(0) \\ &= 5 \times 4 \times 3 \times 2 \times 1 \times 1 \\ &= 120 \end{aligned}$$

Smaller Selection Ranges

With respect to the above scenario, a modification can be made to this calculation to allow for situations in which not all of the cards are selected from the bag. Consider a selection of k cards from a bag containing n cards. Here the number of possible permutations is calculated as the product of the number of options and each integer below it until the number of integers multiplied is equal to the number of elements being selected.

$$
\begin{aligned}
{}^{n}P_k = P(n,k) &= n(n-1)(n-2)...(n-k+1) \\
&= \frac{n!}{(n-k)!}
\end{aligned}
$$

In this situation the notation now makes sense, with the preceding superscript denoting the number of elements in the set and the succeeding subscript denoting the number of elements being selected. As you can see two different types of notation are used, so both have been shown here in the introduction to this concept.

4.10 Multinomial Coefficient

Consider the multiset of characters:

$$[A, B, B, C, C, C]$$

where our goal is to calculate the number of distinct ways of arranging these characters or 'anagrams' of the 'word' ABBCCC. A multiset is just a set where multiple elements are allowed and constitute a unique set.

While it may be natural to use the formula derived in the previous section for calculating permutations with a smaller selection range, there is an interesting problem that arises. Due to the repeating characters in the original multiset, we will end up double counting some elements if we treat all characters as distinct. For example, if we take the characters in the order they are already in, and swap the two Bs, we end up with what we have counted as a new permutation, but of course they are the same word. One way we can avoid this problem is count the duplicates and then remove them after the fact. If we just examine the Bs, which we initially counted as distinct elements, there are two ways that these two Bs can be arranged. Therefore for each permutation we counted, there are two identical cases where the two Bs were swapped which we can remove by dividing by 2. A similar situation happens with the Cs where there are $3! = 6$ duplicates of each case as that's how many ways the Cs can be arranged when treating them as distinct.

Combining these two facts together gives us the number of permutations as:

$$\frac{5!}{2! \times 3!}$$

Which generalises to:

$$\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! \times k_2! \times \ldots \times k_m!}$$

This is called the multinomial coefficient and the equation above shows a shorthand notation used when writing it. Note that even when there is only 1 of each of the given elements it is still specified so that the sum of the numbers at the bottom of the vector is the number at the top. So for the original example, the multinomial coefficient notation would look like:

$$\binom{5}{3,2,1}$$

The reason for this name comes from its usefullness in calculating the coefficient in the expansion of a multinomial.

Consider the following multinomial expansion:

$$\begin{aligned}(a+b+c)^3 &= (a+b+c)(a+b+c)(a+b+c)\\ &= a^3+3a^2b+3a^2c+3ab^2+3ac^2+6abc+b^3+c^3+3bc^2+3b^2c\end{aligned}$$

Consider each term of the expansion as a selection of one term, a, b or c, from each set of parenthesis. For example, let's consider the case where an a, b and a are chosen. This could happen in that order, or in a variety of other orders, producing additional a^2b terms. The number of possible ways that these values can be arranged is the same as the number of ways the letters in the set $\{a,a,b\}$ can be arranged. Which, using the above formula is $\frac{3!}{2!}=3$, which agrees with the value in the above expression.

4.11 Combinations

A *combination* is similar to a permutation. However, the order of the selected elements is not relevant, therefore there is generally a smaller number of possible combinations for any set than there are permutations.

4.11.1 Combinations Without Repetition

Consider the set of characters:

$$\{A,B,C\}$$

When selecting 2 from this set there is a total of 6 permutations, these are as follows:

$$\{AB,BA,AC,CA,BC,CB\}$$

However, when calculating combinations, selections such as AB and BA are considered the same, for this reason only one of these is counted. Removing all other duplicates as well the total number of combinations in 3 and are as follows:

$$\{AB,AC,BC\}$$

Generalising this rule, the number of duplicates in any given selection will be equal to the number of ways that many objects can be arranged, or $k!$. Dividing this out in the formula for permutations of a certain selection range:

$${}^nC_k = C(n,k) = \binom{n}{k,(n-k)} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Note the column vector form looks very similar to that of the multinomial coefficient to its left, however the second value is omitted and inferred to be the remaining value due to the requirement that the sum of the values at the bottom is the value at the top. This is also why values of 1 are typically included in the multinomial coefficient.

There are many situations where this new approach may be useful such as counting hands in a game of cards, where the order within a hand doesn't change the hand or selecting members for a committee.

4.11.2 Combinations With Repetition: The Stars and Bars Approach

Consider a trip to a fruit shop, in which you are trying to select 8 pieces of fruit, out of a selection of apples, bananas, cherries and dragon fruit. We will examine how many possible unique selections of fruit can be made. It is safe to assume there is at least 8 of each time of fruit available.

Let's first observe that we have a situation where there are 8 selections to be made, each from the same 4 options. If this were a situation in which the selection of order were important, this would be the same as our situation of permutations with replacement, just like the number of possible combinations in a password. However, if we were to use this approach, which would give a result of 4^8, we would be over counting due to the fact that the order we end up with the fruit in doesn't make any difference. This allows us to describe this type of problem as a combination, due to the order being unimportant, with repetition, since any given type of fruit can occupy any number of our possible 8 choices.

While we could make an attempt to divide through by our duplicates in the above approach, I will suggest an alternate way of examining the problem that will make it easier.

Let's reverse the problem, and consider the 4 fruit as like four slots, that we may assign 8 tokens, shown by stars, into, where each star represents selecting that specific fruit. For example, let's consider a selection of 2 apples, 4 bananas, 1 cherry and 1 dragon fruit. We represent this selection in the above schema as follows, using the starting letter of each fruit to represent it:

A	B	C	D
$\star\star$	$\star\star\star\star$	$\star$	$\star$

While the fruit labels are given here, the positions from left to right are sufficient to identify a unique combination, so let's remove those.

$\star\star$	$\star\star\star\star$	$\star$	$\star$

In case it is not already clear, our goal is to reduce this problem to the simplest terms we possibly can as to help us generalise not only to different selections within this situation, but to attempt to generalise to different types of problems.

$$\star\star \mid \star\star\star\star \mid \star \mid \star$$

We are now left with 8 stars, and 3 bars dividing them. This diagram is sufficient to identify our selection, and furthermore every possible selection in the situation has a unique corresponding diagram. The reason that this is the case depends on a few facts:

- The lack of uniqueness of the stars, where changing the order of stars within a group gives the same diagram.

- The fact that the groups are ordered, and therefore each section separated by a bar represents a given option that can be selected.
- The fact that there is no limitation on where the stars can be placed, only that there are 8 and they must all be placed.

This last note also further justifies the removal of the 'dividers' on the edges, since those cannot be moved or changed.

Here are some examples of selections and the corresponding diagrams in the above scenario:

1 apple, 2 bananas, 3 cherries, 2 dragon fruit

$$\star \mid \star\star \mid \star\star\star \mid \star\star$$

0 apples, 0 bananas, 0 cherries, 8 dragon fruit

$$\mid \; \mid \; \mid \star\star\star\star\star\star\star\star$$

0 apples, 3 bananas, 5 cherries, 0 dragon fruit

$$\mid \star\star\star \mid \star\star\star\star\star \mid$$

Now we have the task of counting how many of these diagrams are possible. To do so, first notice that it is possible to view each diagram as a selection of where to put the 3 bars given a line of 8 stars, or where to put the 8 stars given a line of 3 bars. As a consequence of this symmetry, we can view this as a problem of how many ways we can uniquely arrange 3 bars and 8 stars, which may look familiar given our study of anagrams, where our multiset of 'letters' is $[\star, \star, \star, \star, \star, \star, \star, \star, |, |, |]$.

Hence the solution is the multinomial coefficient:

$$\binom{8+3}{8,3}$$

or binomial coefficient:

$$\binom{8+3}{8}$$

Now to generalise this to other types of problems, let's backtrack through where the numbers we used came from. The 8 stars was equal to the number of items we were selecting, and 3 bars was equal to one less than the number of options that were available. So to calculate the number of combinations with repetition (and no limitations on the number of each repeated element) where there are n options available and r are being chosen, the total number of combinations is equal to:

$$\binom{n+r-1}{r}$$

While this formula is handy for quick calculations, many combinations with repetition problems are hidden behind a more complicated scenario and it is often helpful to think through the stars and bars diagram for each problem that comes up, a task that becomes relatively quick with a little practice.

4.12 Pascal's Triangle

Pascal's triangle, named after French mathematician Blaise Pascal, is a triangular array of numbers in which any given number is the sum of the two above it.

$$\begin{array}{ccccccccc} & & & & 1 & & & & \\ & & & 1 & & 1 & & & \\ & & 1 & & 2 & & 1 & & \\ & 1 & & 3 & & 3 & & 1 & \\ 1 & & 4 & & 6 & & 4 & & 1 \end{array}$$

As as alternate method for understanding the calculation of each entry in the array, consider each entry's index within the array in the form (row, column) where the counting of each dimension starts at 0. From this, the entry at (row, column) = $^{\text{row}}C_{\text{column}}$.

From this, a symmetry in the results of calculating combinations can be seen. This generalises to the following statement.

$$^{n}C_{k} = {}^{n}C_{n-k}$$

This is because, given that order doesn't matter, the selection of combinations is simply the separation of objects of a set into two groups and thus the group considered the desired selection is an unimportant factor in the number of possible separations.

4.13 Binomial Expansion

Consider the following mathematical expression.

$$(x+y)^n$$

The term on the inside of the parenthesis is called a *binomial* as it is comprised of 2 distinct terms, each with a different integer exponent or different variable. The process of expanding the product of these terms across the external power is called *binomial expansion*.

As a first example, let's expand this set of brackets step by step where $n = 2$:

$$(x+y)^2$$

$$(x+y)(x+y)$$

$$x(x+y) + y(x+y)$$

$$x^2 + xy + xy + y^2$$

$$x^2 + 2xy + y^2$$

Notice the coefficients and exponents of each term as well as the number of terms.

Applying a similar process to a cubic expansion:

$$(x+y)^3$$

$$(x+y)(x+y)^2$$

$$(x+y)(x^2+2xy+y^2)$$

$$x(x^2+2xy+y^2)+y(x^2+2xy+y^2)$$

$$x^3+2x^2y+xy^2+x^2y+2xy^2+y^3$$

$$x^3+3x^2y+3xy^2+y^3$$

Notice how in both cases the exponents of the first term decrease from n to 0 across each term, while the exponents of the second term increase from 0 to n across each term. Also notice the pattern of the coefficients of each term, which follows the nth row of Pascal's triangle, presuming the index starts at 0. As a result of this pattern, we can generalise the method for binomial expansion as:

$$(a+b)^m = \sum_{y=0}^{m} \binom{m}{y} a^m b^{m-y}$$

Ultimately this is a special case of the multinomial coefficient, but this calculation in particular will be useful in future and therefore has been explicitly described.

Chapter 5

Discrete Random Variables

DUE TO THIS BOOK BEING IN PRE-RELEASE, THIS CHAPTER HAS NOT YET BEEN IMPLEMENTED. GO TO ‘FUTURE CHANGES AND UPDATES’ AT THE START OF THE TEXT FOR DETAILS.

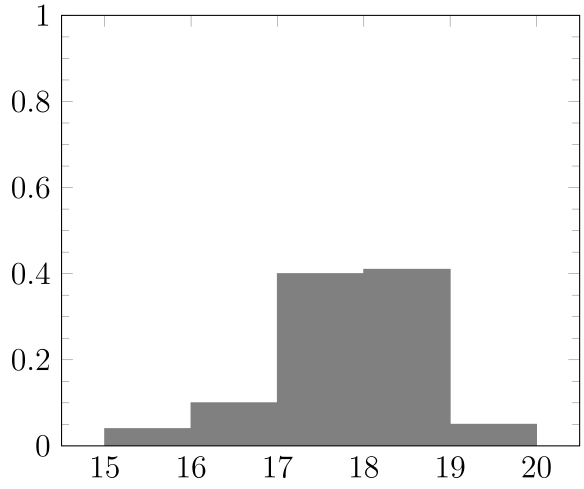
Chapter 6

Continuous Random Variables

6.1 Representing Continuous Data

When data is *continuous*, data points can each assume any possible value within a given domain, with no limitations on the smallest divisor. In these cases a *continuous probability distribution*, defined by a *probability density function (PDF)* is used. Continuous probability distributions are defined using a mathematical function which gives a relationship between the input space x, and the output space $P(x)$, within a specified *domain* of possible outcomes.

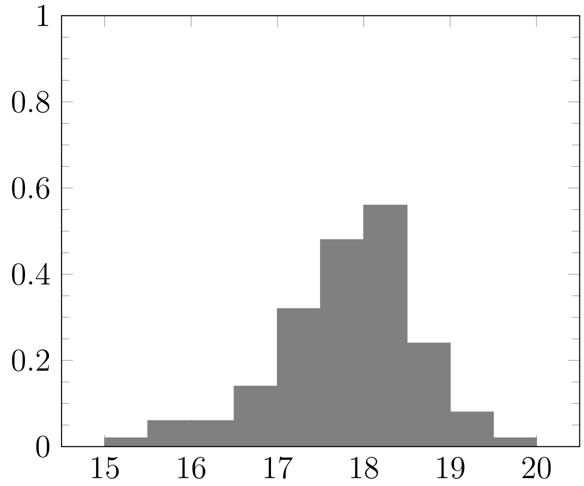
You may recall when working with *discrete probability functions* that finding the probability of any given outcome is as simple as reading it's corresponding value from the y axis. With continuous data, it is not this simple. To illustrate this, let's examine the following discrete probability distribution, which represents the heights of individuals in intervals of 1 decimetre (100mm). Take notice of the fact that unlike most discrete probability distributions, in this case instead of each column representing a discrete point, the x axis represents continuous data with each column representing the proportion of scores within that range, thus the data being measured is actually continuous but is currently being treated as if it is discrete. The reason for this will shortly become clear.



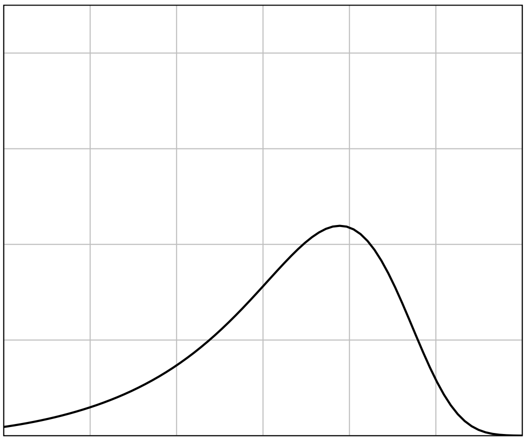
Notice now how the probability of any given category is equal to the area of that column, calculated as the distance in x multiplied by the distance in $P(X)$.

The data with which this distribution is based on, in terms of the values any measurement can assume, is continuous. Despite this, the limitations as a result of measurement force the data into categories, each covering a range of 1 decimetre.

The below distribution represents the same population. However, in this case measurements have been taken in intervals half the size. Therefore the sum of the heights of two adjacent columns must be equal to the twice height of the column that they are replacing. This is such that the data from the original graph is still preserved by *calculating areas* and the sum of the areas of all columns is still equal to 1.



Continuing this process of splitting each column produces a graph that is closer to a continuous distribution.

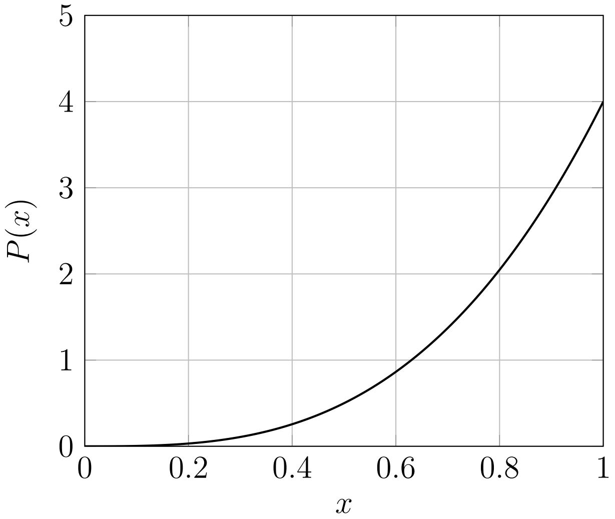


Thus, assuming we can establish a mathematical function for the continuous curve, the probability of an event occurring within any range can be calculated as area under the curve within that range, or the definite integral between the minimum and maximum value of that range. As a consequence of this representation the vertical axis of our graph no longer represents the probability of any specific score but rather a *probability density*.

This means that the probability of any given single score within the domain where the continuous probability function is defined, has a *probability of 0*, despite being a *possible* outcome. Think of the single discrete point being examined as being one outcome of an infinite number of possibilities due to the nature of continuous data. This does however introduce a paradox of a certain outcome being *possible*, but having a *probability of 0*. Furthermore, this is different to outcomes which are *impossible*, as to be categorised as such the score must sit *outside of the domain* of the function, or where the probability density is 0. The specific Mathematical resolution to this paradox is beyond the scope of this text. However it is most important at this point to simply develop an intuition as to why with truly continuous data, the question of asking what the probability of a single value is is ultimately meaningless, and in any real world situation the question of the probability of a specific point is actually dictated by a range defined by the error in measurement.

6.2 Calculating Statistics from Continuous Distributions

Consider the following continuous probability distribution, defined as $P(x) = 4x^3$ for $0 \leq x \leq 1$. Notice the use of *less than or equal to* signs rather than *less than* when defining the domain. As established in section 6.1, this is an unimportant difference as $P(0) = 0$ and $P(1) = 0$. However, both ways of representing the domain of continuous functions are used. The reason for this decision will be detailed in section 6.2.3.



From this function we can also now calculate many of the same descriptive statistics introduced in the first chapter.

6.2.1 Probability of Given Outcome

To calculate the probability of an outcome between values a and b, compute the definite integral within that range with respect to x.

$$P(a \leq X \leq b) = \int_{a}^{b} P(x)dx$$

6.2.2 Expected Value

Just as the expected value of a discrete random variable can be calculated as the sum of each score multiplied by it's relative frequency, for a continuous random variable the method is the same. However, this sum is the sum of an infinite number of infinitely small parts, or the integral. Therefore the expected value of a variable X with a probability density function $P(x)$ is:

$$E[X] = \int_{\text{min}}^{\text{max}} x \cdot P(x)dx$$

For the above distribution:

$$\begin{aligned} \mu &= \int_{0}^{1} x \cdot 4x^3 dx \\ &= \left[\frac{4x^5}{5}\right]_0^1 \\ &= \frac{4(1)^5}{5} - \frac{4(0)^5}{5} \\ &= \frac{4}{5} \end{aligned}$$

Therefore the mean or expected value is equal to $\frac{4}{5}$.

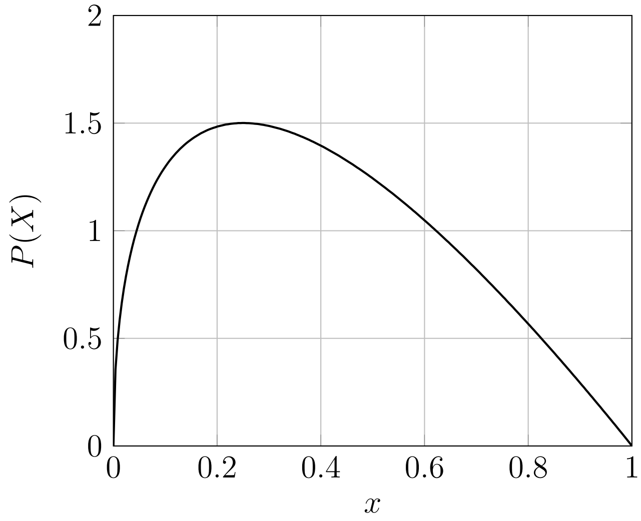
In a general sense, the expectation of any specific property of a PDF can be calculated as the infinite sum of the property multiplied by the value of the PDF. This will come in handy with variance just as it did with discrete random variables.

6.2.3 Mode

While the concept of a mode may not seem make any real sense with respect to continuous data, for practical reasons, the mode is defined as the input for the probability density function that provides the greatest output.

When trying to determine the mode of a probability density function, it can be very useful the view the function graphically first. In the case of the current example of $P(X) = 4x^3$, the mode sits at the edge of the domain, at $x = 1$. This why the convention of this textbook has been to use less than or equal signs, that way the extreme value theorem guarantees that a maximum or mode actually exists.

In cases where the mode is not at the edge of the domain, finding it requires a different method. Take the following probability function defined as $P(X) = 6(\sqrt{x} - x)$ for $0 \leq x \leq 1$.



In this case, we need to find the stationary point of the function, as this is where the function is at its maximum value. This stationary point is where the functions derivative $P'(x) = 0$.

$$P'(x) = \frac{3}{\sqrt{x}} - 6$$

$$0 = \frac{3}{\sqrt{x}} - 6$$

In this case, $P'(x) = 0$ when $x = \frac{1}{4}$. This is the mode of $P(x)$.

This method of finding the mode is simply an application of the extreme value theorem, generalised to the following process:

1. Note the values of the PDF at its endpoints.
2. Compute and note the values of the PDF at any turning points.

3. Compute and note the values of the PDF at any points where it is not differentiable.
4. Identify the greatest value of all calculated values.

6.2.4 Variance and Standard Deviation

You may recall from working with discrete probability distributions that the variance can be computed using the following method.

$$\mathrm{Var}(X) = \sum_{i=0}^{N} (x_i - \mu)^2 \times f(x_i)$$

Where $f(x)$ is the relative frequency of each score and i iterates between each distinct outcome or score. By the same rationale as this method, the variance of a continuous distribution can be calculated using:

$$\mathrm{Var}(X) = \int_{\text{min}}^{\text{max}} (x - \mu)^2 \cdot P(x)dx$$

While this is a perfectly valid method for calculating the variance of a PDF, by using a property derived in section **??** on discrete probability distributions, we can simplify this further. Consider that:

$$E[X - \mu]^2 = E[X^2] - E[X]^2$$

Given that $E[X]$ or the mean is a property with computation method outlined in section 6.2.2, we only now need to be able to compute $E[X^2]$, which, intuitively, can be found with the following:

$$E[X^2] = \int_{\text{min}}^{\text{max}} x^2 \cdot P(x)dx$$

Taking the sum of this value and the square of the mean, provides a simpler calculation for the variance of a PDF.

Considering our initial PDF of $P(x) = 4x^3$ defined from 0 to 1 where $E[x] = \frac{4}{5}$,

$$\begin{aligned} E[X^2] &= \int_{0}^{1} x^2 \cdot 4x^3 dx \\ &= \left[\frac{2x^6}{3}\right]_0^1 \\ &= \frac{2(1)^6}{3} - \frac{2(0)^6}{3} \\ &= \frac{2}{3} \end{aligned}$$

Therefore the variance of this distribution is:

$$\mathrm{Var}(X) = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75}$$

Given that $\mathsf{SD}(X) = \sqrt{\mathrm{Var}(X)}$, the standard deviation of this distribution is $\sqrt{\frac{2}{75}} \approx 0.163$.

6.3 Cumulative Distribution Function

A *cumulative distribution function (CDM)* is a function created from a PDF which *accumulates* the probabilities as the values in the input space increase. This means any given input to the CDM provides an output of the probability of any score less than the input.

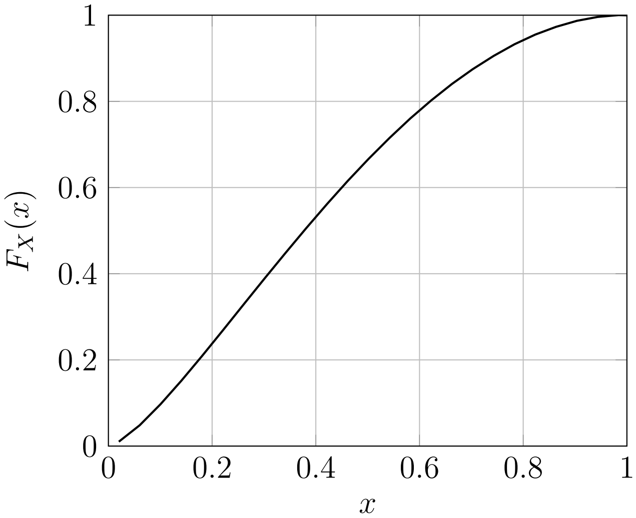
To find the cumulative distribution function calculate the definite integral from the the lowest part of the domain to a generic x of the probability density function.

$$\int_{\mathsf{min}}^{x} P(x)dx$$

As a result of this method of calculation, any cumulative distribution will be *monotonic increasing* within the domain it's defined, with the values increasing from 0 to 1. Here is the cumulative distribution function for the aforementioned probability density function of $P(x) = 6(\sqrt{x} - x)$.

$$\begin{aligned}
F_X(x) &= \int_0^x 6(\sqrt{x} - x)dx \\
&= \left[4\sqrt{x^3} - 3x^2\right]_0^x \\
&= (4\sqrt{x^3} - 3x^2) - (4\sqrt{0^3} - 3 \cdot 0^2) \\
&= 4\sqrt{x^3} - 3x^2
\end{aligned}$$

Here is a graphical representation of this cumulative distribution function:



Such a function is useful when calculating any range of probabilities as such a calculation is a matter of taking the difference of the values of the cumulative distribution function at the desired points, rather than calculating an integral every time.

6.3.1 Quantiles

Cumulative distribution functions are particularly useful when calculating quartiles, deciles and percentiles. As an example, to find the median of any probability density function, find the value for x where the cumulative density function equals 50%. With the above example:

$$0.5 = 4\sqrt{x^3} - 3x^2$$

$$x \approx 0.377, 1.556$$

Here since 1.556 is outside of the domain of the function, it is to be discarded, thus the median is 0.377. Due to a CDF always being monotonic increasing, there will only be one solution for x within the domain of the established function.

As a logical consequence of this method, quartiles, deciles and percentiles can be found by calculating x when the CDF is equal to the percentage of scores you wish to be less than the corresponding marker.

6.4 Generating Random Numbers from a Probability Density Function

While a PDF is useful in calculating statistics, it is not a function that can be directly used to take random samples. Despite this, the technique used to generate random numbers according to a given distribution is quite intuitive. Let's examine the PDF introduced in 6.2.3 of $P(x) = 6(\sqrt{x} - x)$ and it's corresponding CDF calculated in section 6.3 of F$(x) = 4\sqrt{x^3} - 3x^2$.

First we must find the inverse function (or relation) of the CDF:

$$y = F_X(x)$$

$$y = 4\sqrt{x^3} - 3x^2$$

Taking the inverse:

$$x = 4\sqrt{y^3} - 3y^2$$

Now, we can generate a random floating point number between 0 and 1, using it as the input for the inverse of the CDF to get a result of our randomized number. For example if our uniform distribution gives 0.634.

$$0.634 = 4\sqrt{y^3} - 3y^2$$

Calculating all values for y:

$$y \approx 0.476, 1.478$$

Given that the CDF is only defined in a specified domain, the range of the inverse function is the same as that original domain. Therefore if multiple values are found within this calculation, the desired output is the one domain of the original PDF and CDF, of which there will always only be one. In this case, 0.476 is the random number that has been generated from the PDF.

Chapter 7

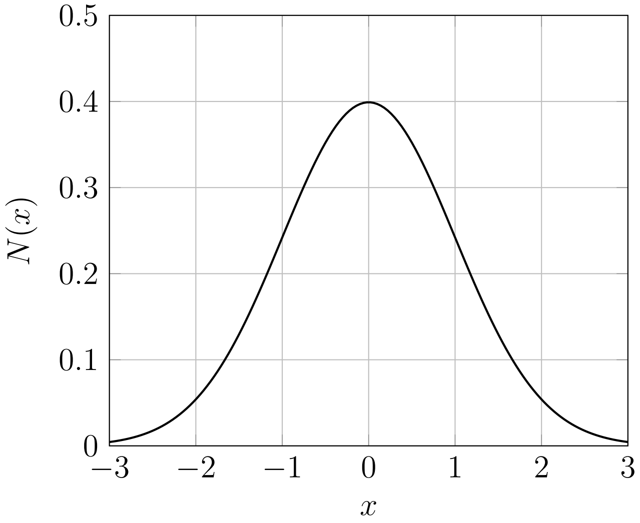
Normal Distribution

7.1 What is the Normal Distribution?

The *normal distribution*, also called the *bell curve*, is a continuous probability distribution parameterised by the mean and standard deviation.

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

Here is an example of a normal distribution with standard deviation of 1 and mean of 0, also called the *standard normal distribution*. This is a useful idea as any given score has a z score of equivalent value.

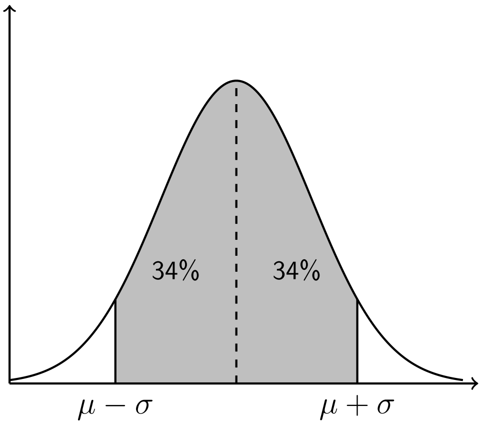


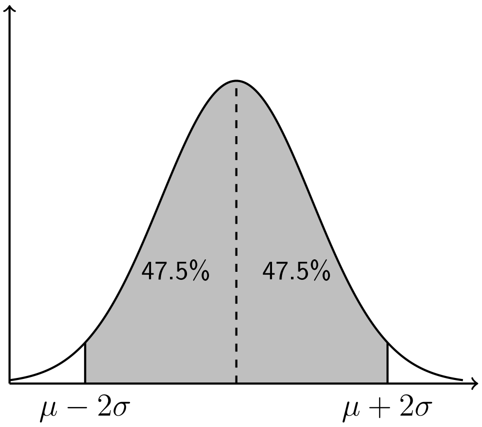
The normal distribution is a useful concept due to the number of statistics that are modelled according to it including the weight of newborn babies, height of NBA players, blood pressure and IQ scores. As a consequence of this many statistical tests rely on the assumption that the data being examined is normally distributed. This section will only act as a brief introduction to the definition of the normal distribution and some of its properties. However, its uses will become apparant through later sections of this text.

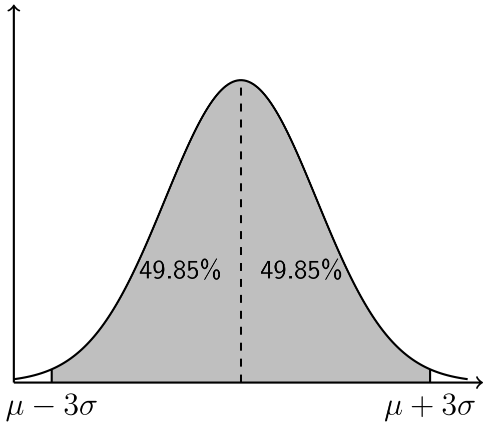
7.2 Empirical Rule

A unique property of the normal distribution function is that the percentage of scores that lie within any number of standard deviations is the same, regardless of the mean and standard deviation of the distribution.

The following diagrams show the number of scores within 1, 2 and 3 standard deviations of the mean respectively.







This is called the Empirical Rule or the 68–95–99.7 rule and specifies that 68% of scores sit within 1 standard deviation, 95% within 2 and 99.7% within 3 on either side of the mean.

7.3 Z Tables

Z tables are an index of the percentage of scores below any given z score for the normal distribution and are a more quantitatively precise representation of the Emprical Rule. It is common for z tables to be split into two tables, one for negative z scores and one for positive z scores.

z	0	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Due to the fact that the integral of the probability density function for the normal distribution cannot be expressed in terms of elementary functions, the actual values within the table are produced by methods of numerical approximation.

As an example of how z tables are used, let's calculate $P(-0.5 < X < 0.15)$ for a normal distribution with a mean of 0.1 and a standard deviation of 0.5.

First we use the transformation $Z = \frac{X-\mu}{\sigma}$ to calculate the z scores of the bounds of the range being examined.

$$z_{-0.5} = \frac{-0.5 - 0.1}{0.5} = -1.2$$
$$z_{0.15} = \frac{0.15 - 0.1}{0.5} = 0.1$$

Examining the z tables, for the percentage of scores less than a z score of -1.2, we need to find the row labeled -1.2 and column labelled .00, in this case, 0.1151. For the percentage of scores below a z score of 0.1, we need to look up the row labelled 0.10 and the column labelled .05, in this case 0.5596.

Now to find the portion of scores that lie between these points, subtract the greater value from the smaller one.

$$P(-0.5 < x < 0.15) \approx 0.5596 - 0.1151 = 0.4445 = 44.45\%$$

When the z scores calculated are of greater accuracy than that of the table, either round to the nearest interval in z or provide a range in the final calculation by rounding up for the lower score and down for the upper score to calculate the lower bound of the range, and vice versa to calculate the upper bound.

7.4 Proof: Area Under the Normal Distribution

While many attempts will be made to abstract and simplify the calculations involved, the following section assumes a high level of understanding of calculus including double integrals, vector calculus and non-linear matrix transformations.

Given that the normal distribution is defined for all x, with its value approaching 0 as $x \to \infty$ and as $x \to -\infty$, it can be difficult to verify that the area under this curve is equal to 1. The following is a proof that the area under the normal distribution approaches 1 as a greater domain is observed.

To calculate the area A, we must compute the definite integral from $-\infty$ to ∞ of the normal distribution with respect to x.

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

By letting $u = \frac{x-\mu}{\sigma}$, $du = \frac{dx}{\sigma}$ and $dx = \sigma du$.

$$A = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}u^2} \cdot \sigma du$$

Here the boundaries do not change since as $x \to -\infty$, $u \to -\infty$ and as $x \to \infty$, $u \to \infty$.

Dividing out σ:

$$A = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}u^2} \cdot du$$

Squaring both sides and changing the integrating variable of the second instance to v.

$$A^2 = \left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}u^2} \cdot du\right)\left(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}v^2} \cdot dv\right)$$

Combining the integrals into a double integral, moving the constant term to outside of the integral in each:

$$A^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(u^2+v^2)} \cdot dudv$$

Changing the integrating variables to polar form, $u = r \cdot \cos(\theta)$ and $v = r \cdot \sin(\theta)$. The domain of the new integrals are limited by their properties where $0 \leq \theta \leq 2\pi$ and $0 \leq r < \infty$.

When changing these variables, $dudv = |J|drd\theta$, where J represents the Jacobian matrix. For those unfamiliar with multivariate vector calculus, in its simplest form, the Jacobian matrix represents a non-linear transformation of space which in this case, maps the original integrating variables to the new ones.

To compute this determinant:

$$|J| = Det\left(\begin{bmatrix} \frac{du}{d\theta} & \frac{du}{dr} \\ \frac{du}{d\theta} & \frac{dv}{dr} \end{bmatrix}\right)$$

$$\begin{aligned}
|J| &= Det\left(\begin{bmatrix} \frac{du}{d\theta} & \frac{du}{dr} \\ \frac{du}{d\theta} & \frac{dv}{dr} \end{bmatrix}\right) \\
|J| &= Det\left(\begin{bmatrix} -r \cdot \sin(\theta) & \cos(\theta) \\ r \cdot \cos(\theta) & \sin(\theta) \end{bmatrix}\right) \\
|J| &= |-r \cdot \sin(\theta) \cdot \sin(\theta) - r \cdot \cos(\theta) \cdot \cos(\theta)| \\
|J| &= |-r(\sin^2(\theta) + \cos^2(\theta)| \\
|J| &= |-r| \\
|J| &= r
\end{aligned}$$

Hence,

$$\begin{aligned}
A^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}(r^2 \cdot \cos^2(\theta) + r^2 \cdot \sin^2(\theta))} \cdot rdrd\theta \\
&= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}(r^2(\cos^2(\theta) + \sin^2(\theta))} \cdot rdrd\theta
\end{aligned}$$

Since $\sin^2(\theta)+\cos^2(\theta)=1$.

$$\begin{aligned}
A^2 &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2}\cdot r dr d\theta \\
A^2 &= \frac{1}{2\pi}\int_0^{2\pi}\left(\int_0^{\infty} e^{-\frac{1}{2}r^2}\cdot r dr\right) d\theta
\end{aligned}$$

By letting $w=r^2$, $dw=2rdr$. As $r\to\infty$, $w\to\infty$, when $r=0$, $w=0$.

$$\begin{aligned}
A^2 &= \frac{1}{2\pi}\int_0^{2\pi}\frac{1}{2}\left(\int_0^{\infty} e^{-\frac{1}{2}w}\cdot 2rdr\right) d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi}\frac{1}{2}\left(\int_0^{\infty} e^{-\frac{1}{2}w}dw\right) d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} -\left[e^{-\frac{1}{2}w}\right]_0^{\infty} d\theta
\end{aligned}$$

As $w\to\infty$, $e^{-\frac{1}{2}w}\to 0$ and when $w=0$, $e^{-\frac{1}{2}w}=1$.

$$\begin{aligned}
A^2 &= \frac{1}{2\pi}\int_0^{2\pi} d\theta \\
&= \frac{1}{2\pi}\left[\theta\right]_0^{2\pi} \\
&= \frac{1}{2\pi}(2\pi-0) \\
&= 1 \\
A &= 1
\end{aligned}$$

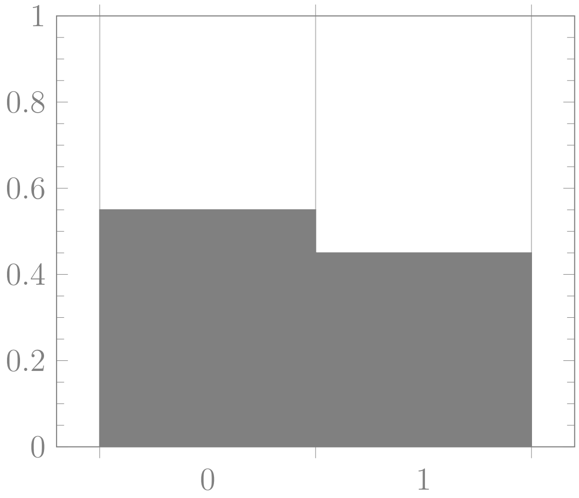
Therefore the area under a normal distribution of any kind is 1.

Chapter 8

Binomial Random Variables

8.1 Bernoulli Trial

A *Bernoulli trial* is an event where there are only 2 possible outcomes or scores and are represented by a distribution just like this one:



Here the event assigned to the value of 1 is the favourable outcome with a probability denoted by p. Therefore the complimentary event of this, where the favourable outcome does not occur, has a probability of $1 - p$.

Mean

The mean of a Bernoulli distribution is calculated just as with any other discrete probability distribution.

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

Expanding this definition using the characteristics of the distribution:

$$E[X] = 0(1 - p) \times 1(p)$$

$$= p$$

Therefore, as a consequence of the way values have been assigned to the different outcomes, the mean is equal to just p. This makes logical sense given the expected proportion of trials to succeed over many trials will be the probability of any given trial.

Variance and Standard Deviation

For the variance of a Bernoulli distribution:

$$\mathrm{Var}(X) = \sum_{i=1}^{n} (x_i - E[x])^2 \times P(x_i)$$

Expanding this definition using the characteristics of the distribution:

$$\mathrm{Var}(X) = p(1 - p)^2 + (0 - p)^2 \times (1 - p)$$

$$
\begin{aligned}
&= p(1 - 2p + p^2) + p^2(1 - p) \\
&= p - 2p^2 + p3 + p2 - p3 \\
&= p - p^2 \\
&= p(1 - p)
\end{aligned}
$$

From this the standard deviation can be calculated as:

$$
\begin{aligned}
\text{SD}(X) &= \sqrt{\mathrm{Var}(X)} \\
&= \sqrt{p(1 - p)}
\end{aligned}
$$

8.2 Binomial Distribution

A *Binomial distribution* is a discrete probability distribution comprised of a series of *Bernoulli trials* where each score is the probability of exactly that number of trials occurring from a specified total number of trials.

We will use k to denote the number of trials we wish to succeed out of n, and p for the probability of any given trial succeeding, just as we did with the Bernoulli trial.

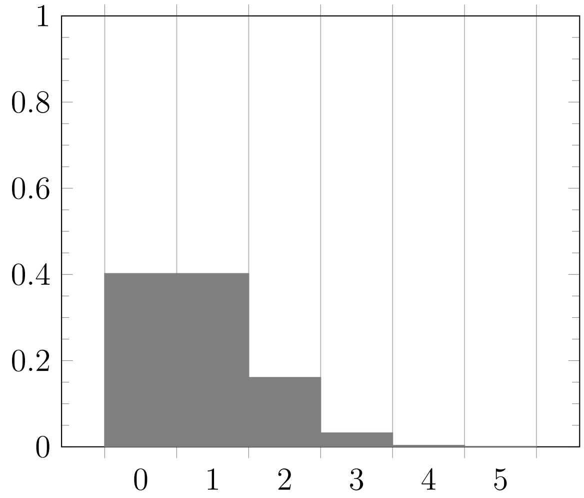
By using the multiplicative rule for probabilities, the probability of an event with a probability p occurring k times with k trials will be p^k. If we want k trials to succeed and the rest to fail out of n trials, we have to account for those terms as well, giving the overall probability as $p^k(1-p)^{n-k}$.

There is however an issue with this calculation, and that is the assumption that the failures and successes happen in a specific order. If we wish to include all cases with k successes, we have to multiply this term by the number of ways we could get the same number of successes and fails, but in unique orders. This is the same as the number of ways we can order the two options, success and fail, removing duplicates. This can be done using the multinomial coefficient or binomial coefficient, hence the name of this type of distribution.

Therefore we can use the following to calculate the probability of k trials succeeding out of n trials, where p is the probability of each trial succeeding:

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

Consider the Bernoulli trial of the rolling of a 6 sided die where an outcome of 1 is favourable. Let's generate a Binomial distribution for 5 rolls of this dice using the above formula.



Due to the probability of each event being less than 0.5 the distribution has a positive skew. In general terms:

- When $p < 0.5$, the distribution has a positive skew.
- When $p = 0.5$, the distribution is symmetrical.
- When $p > 0.5$, the distribution has a negative skew.

Therefore, the greater the probability of each trial succeeding the more likely it is for a greater number of trials to succeed.

The usefulness of the binomial distribution goes beyond situations like this, including many games of chance. Ultimately though a binomial distribution is most often applied when some more complex discrete or continuous data has a specific threshold at which the data is separated into two distinct categories and therefore is even useful in some simple statistical tests.

8.2.1 Mean and Variance of the Binomial Distribution

Verifying the Formula for the Mean

Recall that the general formula for the mean of a discrete probability distribution is:

$$E[X] = \sum xP(x)$$

Applying this to a Binomial distribution:

$$E[X] = \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x}$$

$$= \sum_{x=0}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Given that when $x = 0$, the entire term is equal to zero, the summation limits can be changed to $x = 1$ to n without effecting the value of the overall expression. This then allows further algebra to be done with respect to x.

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

Given that $n - x = (n-1) - (x-1)$,

$$E[X] = np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!((n-1)-(x-1))!} p^{x-1} (1-p)^{(n-1)-(x-1)}$$

Let $y = x - 1$ and $m = n - 1$. Changing the limits of summation, when $x = 1$, $y = 0$ and when $x = n$, $y = n - 1 = m$.

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

Recalling the general form of a Binomial expansion:

$$(a+b)^m = \sum_{y=0}^{m} \binom{m}{y} a^m b^{m-y}$$

The following substitution can be made:

$$
\begin{aligned}
E[X] &= np(p + (1-p))^m \\
&= np1^m \\
&= np
\end{aligned}
$$

Where n is the number of trials and p is the probability of success in a single trial.

Applying this to the previous example of rolling a 1 on a six sided die:

$$E[X] = \frac{1}{6} \times 5 = \frac{5}{6}$$

Verifying the formula for the Variance

When deriving the variance of a Binomial distribution, we will use the property that $E[x(x-1)] = E[x^2 - x] = E[x^2] - E[x]$, to then add $E[x]$ and use $E[(x-\mu]^2 = E[x^2] - E[x]$.

$$E[X^2 - X] = \sum_{i=1}^{n}(x^2 - x)P(x)$$

$$= \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Since the term being summed across is equal to 0 when $x = 0$ or $x = 1$, we can change the limits of summation.

$$E[X^2 - X] = \sum_{x=2}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

$$= \sum_{x=2}^{n} x(x-1)\frac{n!}{x(x-1)(x-2)!(n-x)!}p^x(1-p)^{n-x}$$

$$= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!}p^x(1-p)^{n-x}$$

$$= n(n-1)p^2\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!}p^{x-2}(1-p)^{n-x}$$

Given that $n - x = (n-2) - (x-2)$.

$$E[X^2 - X] = n(n-1)p^2\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!((n-2)-(x-2))!}p^{x-2}(1-p)^{(n-2)-(x-2)}$$

Let $y = x - 2$ and $m = n - 2$. Changing the limits of summation, when $x = 2$, y = 0 and when $x = n$, $y = n - 2 = m$.

$$E[X^2 - X] = n(n-1)p^2\sum_{y=0}^{m} \frac{m!}{y!(m-y)!}p^y(1-p)^{m-y}$$

Once again using Binomial expansion:

$$E[X^2 - X] = n(n-1)p^2(p + (1-p))^m$$

$$= n(n-1)p^2 1^m$$

$$= n(n-1)p^2$$

Since $E[X] = np$ and $E[X^2] - E[X] = n(n-1)p^2$,

$$\begin{aligned} E[X^2] &= (E[X^2] - E[X]) + E[X] \\ &= n(n-1)p^2 + np \end{aligned}$$

Applying this to calculate the variance:

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= np(np - p + 1 - np) \\ &= np(1-p) \end{aligned}$$

Where, once again, n is the number of trials and p is the probability of success in a single trial.

Applying this to the above example:

$$\mathrm{Var}(X) = 5 \times \frac{1}{6} \times \left(1 - \frac{1}{6}\right) = \frac{25}{36}$$

Chapter 9

Further Descriptive Statistics

9.1 Moments

In statistics, the term *moments* refers to a group of descriptive statistics, some of which have already been covered, that provide a measure for location or shape of data. A moment, in it's most basic form, is the average distance, raised to some power, between each data point and some reference point. This chapter will introduce the concept of moments and address where we have already applied them, such that future parts of this textbook reliant on them will make sense.

9.1.1 Raw and Crude Moments

Raw moments, also called *crude moments* or just *moments*, quantify the expected value of the distance between each data point and 0 raised to some power. The first moment is simply defined as:

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

Where each score is the respective distance between that score and 0.

Higher moments, raised to a higher power can be calculated similarly. Here is a generic method for calculating the nth standard moment.

$$\mu'_n = E[X^n] = \int_{\text{min}}^{\text{max}} x^n P(x) dx$$

Examining these calculations, the zeroth moment of any data set will be 1 and the first moment will be the mean.

9.1.2 Central Moments

Central moments are moments taken about the centre or the mean of a data set rather than 0. Therefore the nth central moment is calculated as.

$$\mu_n = E[(X - \mu)^n] = \int_{\text{min}}^{\text{max}} (X - \mu)^n P(x) dx$$

Once again, the zeroth central moment will always be 1 while the second central moment is the statistic of variance.

Therefore we can also define the standard deviation in terms of moments as:

$$\sigma = \sqrt{\mu_2}$$

9.1.3 Standardised Moments

From these definitions, we can define what are called *standardised central moments* as:

$$\tilde{\mu}_n = \frac{\mu_n}{\sigma^n}$$

Here, the zeroth and first standardised moments are 0, the second is 1 and the third and fourth are used to measure skewness and kurtosis which will be covered in the following sections.

9.2 Pearson's Moment Coefficient of Skewness

Skewness is a measure of *asymmetry* in a uni-variate data set is. The most common method of quantifying skewness is *Pearson's moment coefficient of skewness* and is the third standardised moment. As an alternative to calculating this in terms of central moments, this statistic can also be calculated using the following formula, which gives an intuition for the calculation arranging it such that it is clear that skewness is simply the average of the cubes of all the z scores.

$$\tilde{\mu}_3 = \frac{\sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^3}{N}$$

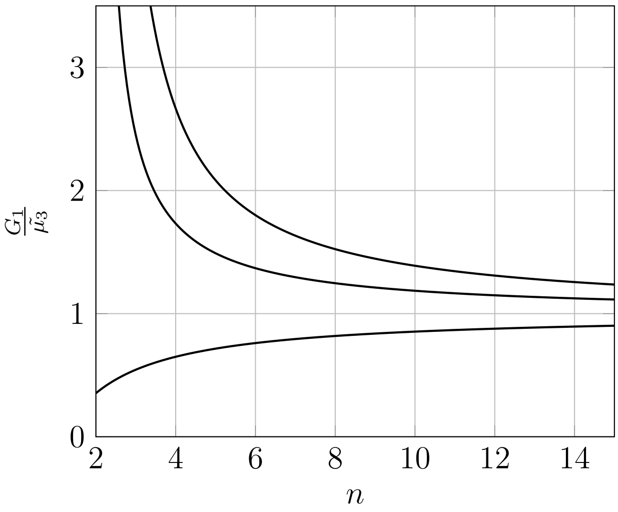
Just as with many other statistics, a slight modification is made when trying to estimate the parameter from a sample:

$$b_1 = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3}{n-1}$$

While these formulae are those originally developed by Pearson, many computer programs use a modification for the calculation of skewness that takes sample size further into account. The following calculations are all used within different statistical software:

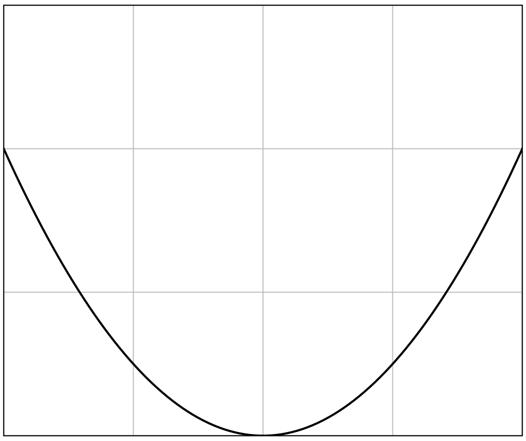
- $G_1 = \frac{N^2}{(N-1)(N-2)} \times \tilde{\mu}_3$
- $G_1 = \frac{\sqrt{N(N-1)}}{(N-2)} \times \tilde{\mu}_3$
- $G_1 = \left(\frac{n-1}{n}\right)^{\frac{3}{2}} \times \tilde{\mu}_3$

As the sample size n increases, these coefficients approach 1 and therefore have minimal effect on the value of the final calculation for large samples. Plotting these three coefficients for $1 \leq x \leq 15$:



Any data set that is perfectly symmetrical should have a skewness of 0 as each cubed z score should have a corresponding point with the opposite sign. Where data is asymmetrical, the magnitude of z scores on one side of the mean will be greater than the magnitude of z scores on the other side. While it may seem counter-intuitive, positively skewed data has a skewness of negative value and vice versa, as skewness quantifies how many and how extreme the points are with respect to the mean.

It is important to note that skewness quantifies *symmetricallity* not *normality* and bi-modal data can also have a skewness of 0 where the peaks are symmetrical about the mean. Hence, the below distribution would have also a skewness of 0.



9.3 Other Methods of Quantifying Skewness

While the above calculations of Pearson's third skewness coefficient and it's subsequent modifications have become the most common measurements of skewness. There are other less popular methods based on different statistics which also measure skewness.

Pearson's Mode Skewness

Pearson's Mode Skewness, also called *Pearson's first skewness coefficient* is a measure of the distance between the mean and mode in standard deviations.

$$SK_1 = \frac{\mu - \textit{MO}}{\sigma}$$

Pearson's Median Skewness

Pearson's Median Coefficient, also called *Pearson's second skewness coefficient* is a measure of the distance between the median and mode in standard deviations.

$$SK_2 = \frac{\mu - \tilde{x}}{\sigma}$$

Quantile Defined Skewness

There are a variety of measures of skewness that are defined using quantiles such as *Bowley's measure of skewness* (*Yule's coefficient*) and *Kelly's measure of skewness*. However, as described by Richard A. Groeneveld and Glen Meeden in their paper "Measuring Skewness and Kurtosis", many quantile based measures of skewness are reducible to the following function.

$$\gamma(u) = \frac{F^{-1}(u) + F^{-1}(1-u) - 2F^{-1}(0.5)}{F^{-1}(u) - F^{-1}(1-u)}$$

Where $F^{-1}()$ is the CDF of the examined function with which the skewness is to be determined. From this, Bowley's measure of skewness is $\gamma(0.75)$, while Kelley's measure of skewness can be expressed as $\gamma(0.1)$.

9.4 Kurtosis

Kurtosis is a measure of the strength of the tails of a distribution relative to a normal distribution. It is also the *fourth standardized central moment of the probability model* or $\tilde{\mu}_4$. Just as with skewness, it can be calculated in terms of z scores:

$$\tilde{\mu}_4 = \frac{\sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^4}{N}$$

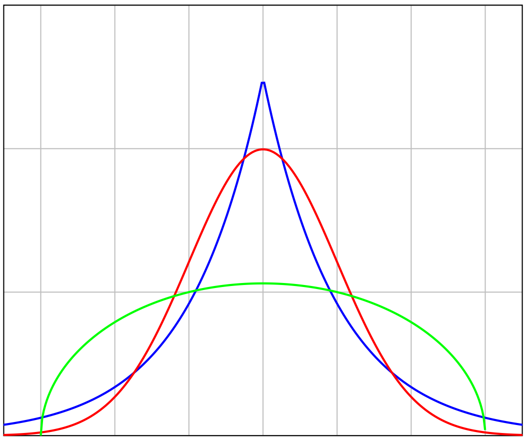
Unlike skewness, this calculation is not modified for estimating population parameters from a sample. However, a slight modification is often made such that the standard normal distribution has a kurtosis of 0, this is called *excess kurtosis* but is often confusingly just referred to as kurtosis.

$$\tilde{\mu}_4 - 3 = \frac{\sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^4}{N} - 3$$

After calculating excess kurtosis a given distribution can be assigned to one of the following categories.

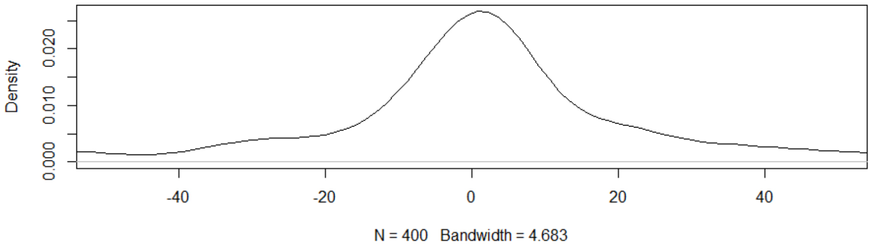
- **Mesokurtic**: Where the excess kurtosis is 0. Examples include a normal distribution of any kind and the binomial distribution when $p = \frac{1}{2} \pm \sqrt{\frac{1}{12}}$.
- **Leptokurtic**: Where the excess kurtosis is greater than 0. The poisson distribution and exponential distribution are leptokurtic.
- **Platykurtic**: Where the excess kurtosis is less than 0. The raised cosine distribution and uniform distribution are platykurtic.

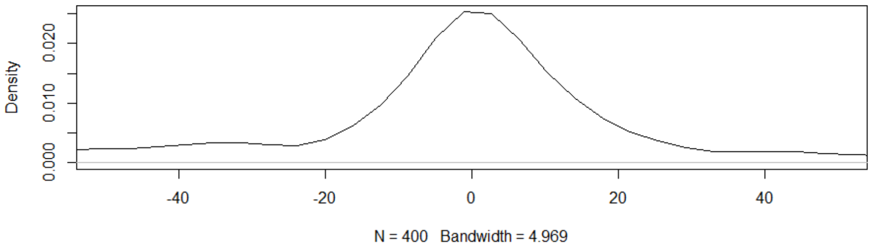
To gain a further intuition for these terms and the broader idea of kurtosis, consult the diagram below which shows extreme cases of these with the *mesokurtic distribution in red*, *leptokurtic distribution in blue* and *platykurtic distribution in green*.



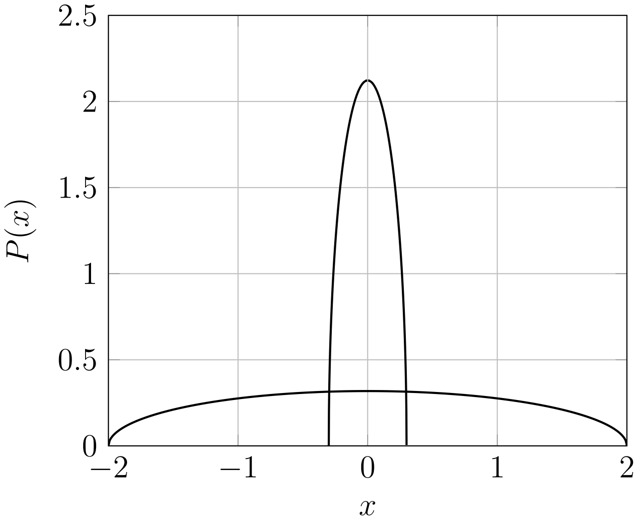
Notice how because kurtosis is dependent on the standard deviation yet measures tailedness, it can be difficult to intuitively guess the kurtosis of a distribution. When examining the above distributions, it can be easy to fall into the trap of assuming that kurtosis is a measure of ‘peakedness’ as this seems to logically follow given these distributions. A large part of this misconception comes from the creator of this measure, Karl Pearson, who commented on the matter saying that excess kurtosis measures the "degree of flat-toppedness which is greater or less than that of the normal curve. Given two frequency distributions which have the same variability as measured by the standard deviation, they may be relatively more or less flat-topped than the normal curve. If more flat-topped I term them platykurtic, if less flat-topped leptokurtic, and if equally flat-topped mesokurtic." While this would appear to make intuitive sense, Pearson was incorrect in this assertion. Given that Kurtosis is actually a measure of tailedness and not peakedenss, leptokurtic distributions generally have long and thin tails, while platykurtic distributions have thick and short tails.

If we define the shoulders of a distribution as being at $\mu \pm \sigma$, then a shift in the mass of a given distribution away from these shoulders outwards to the centre and the tails, such that the standard deviation remains unchanged, the kurtosis would increase. This is why it is common to describe kurtosis as a measure of peakedness. Providing a counterexample of this, consider the following two distributions with very similar shaped peaks, the first of which with a kurtosis of approximately 10.54 and the second with 47.29. This difference is due to the thickness of the tails on either side of the peak.





Another example is the Wigner semi-circle distribution, parameterised by the radius. Any distribution of this nature has a kurtosis of -1. Here are two instances of this distribution the taller of which with a radius of 0.3 and the shorter of which with a radius of 2.



It is however true that small kurtosis can be a sign of bimodality and thus the usage of this statistic is dependent on the nature of the data and should be selected carefully.

It is also valuable to mention that the kurtosis of any distribution is bounded by the skewness squared plus 1.

$$\tilde{\mu}_4 \geq (\tilde{\mu}_3)^2 + 1$$

Chapter 10

Sampling Distributions

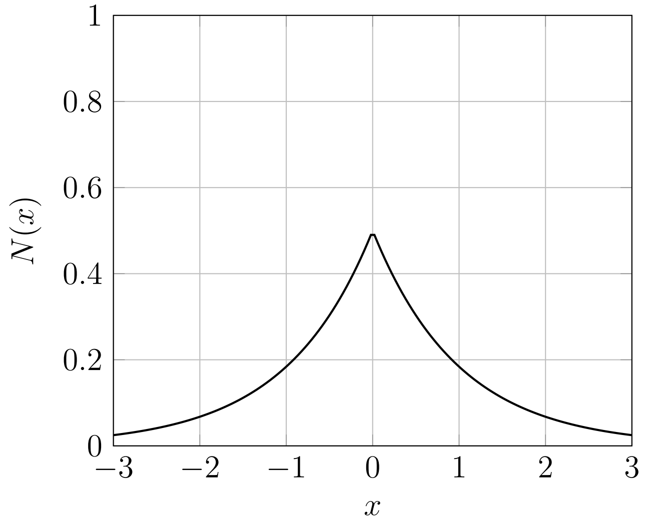
10.1 What is a Sampling Distribution?

The concept of *sampling distributions*, and the subsequent calculations and concepts that come from them, are often needlessly confusing as a result of ambiguity between terminology. To avoid this confusion, take note of the discrepancy between the phrases: *number of measurements per sample* and *number of samples*, both of which will be used to avoid the commonly confusing term of *sample size* which can be used to mean either.

A sampling distribution is a distribution created from a given statistic calculated from a large number of samples of predetermined size from some population or PDF.

10.2 Sampling Distribution of the Sample Mean

Consider the following PDF, defined between $-\infty$ and ∞, used to represent traits within a population.



Let's take a sample of with a given number of measurements $n = 5$ and calculate the mean.

0.414 0.436 1.598 4.605 1.637 $\overline{x} \approx 1.706$

Repeating this process:

0.126 0.277 0.181 0.412 1.452 $\overline{x} \approx 0.490$

1.103 0.119 0.421 2.283 0.017 $\overline{x} \approx 1.076$

0.970 4.200 0.970 0.570 0.703 $\overline{x} \approx 1.483$

...

The distribution formed by the sample means on the right is called the *sampling distribution of the sample mean*.

10.3 Sampling Distribution of the Sample Proportion

While sampling distributions can be calculated from sample means, they can also be created from many other statistics including sample proportions.

A *sampling distribution of the sample proportion* closely mirrors a sampling distribution of the sample mean however instead of calculating the mean from each sample, a proportion is calculated. Often this will be the proportion of individuals that exhibit some trait within a population. Consider a bag of marbles where the actual proportions of marble colours are 15% blue, 40% red, 20% yellow and %25 green. Taking repeated samples of marbles, each of size 5, and calculating the proportion $\hat{p}$ of marbles that are red, where $\hat{p}$ denotes a *prediction* for the actual proportion p.

Yellow Green Red Red Red $\hat{p} = 0.6$

Yellow Blue Red Green Red $\hat{p} = 0.4$

Red Yellow Yellow Yellow Blue $\hat{p} = 0.2$

Green Blue Red Green Yellow $\hat{p} = 0.2$

...

The sampling distribution of the sample proportion is the distribution of sample proportions listed to the right of each sample.

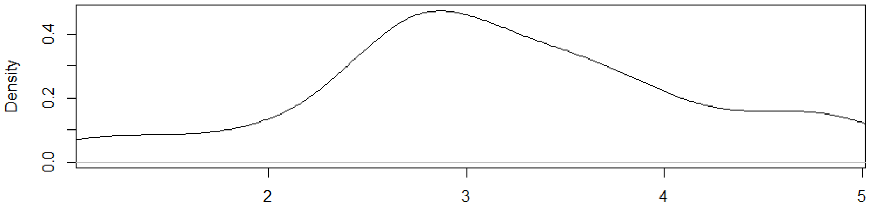
10.3.1 Perfect Sampling Distributions

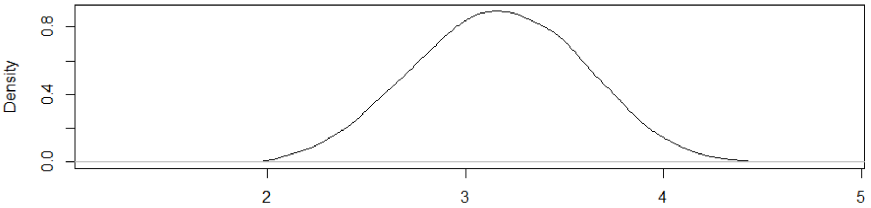
A *perfect sampling distribution* is a theoretical sampling distribution created from every possible combination of samples within a given population and can therefore only exist where there is a set population size, not a continuous function. This type of function, despite not being used with real data, is a useful idea as statistics calculated from each sample represent every possible statistic that could be calculating when sampling the population with the same number of measurements.

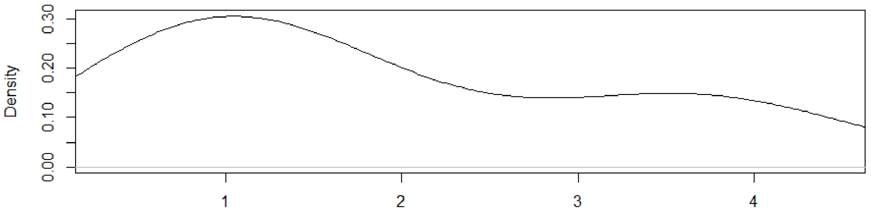
10.4 Central Limit Theorem

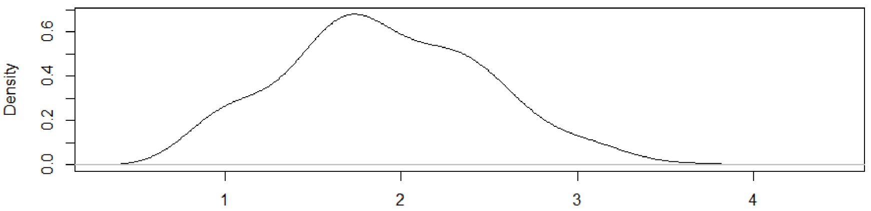
The *Central Limit Theorem* states that a sampling distribution of the sample mean or sample proportion, generated from *any* probability function, will approximate a *normal distribution*. Furthermore, the normality of the distribution will increase as the number of measurements per sample increases.

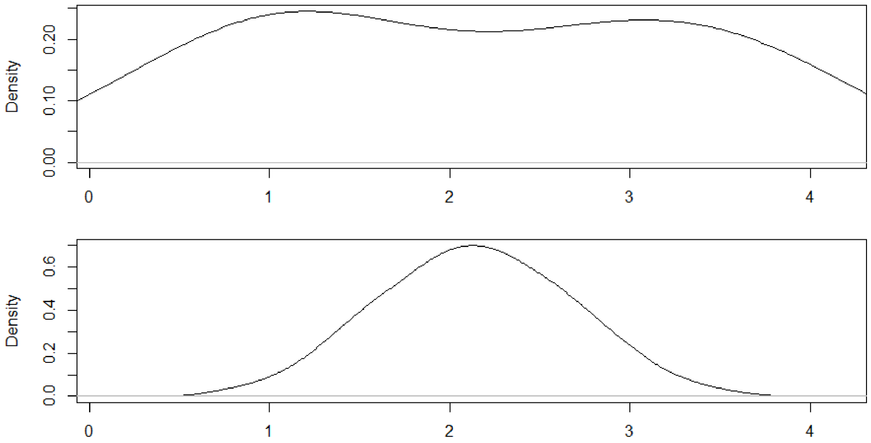
Here are graphs are results from the program found in Appendix B which generates a set of random numbers and calculates a perfect sampling distribution from this. In each set the first graph represents the population being examined and the second represents the sampling distribution. As you can see, regardless of this function, the sampling distribution is closer to a normal distribution.











Notice how each sampling distribution is defined in a smaller domain than the original function, due to the fact that the most extreme values of the original function cannot show up in the sampling distribution as a sample mean assuming there is more than 1 measurement per sample. Expanding this idea, the values towards the centre are likely to have the highest frequency in the sampling distribution.

To gain a further intuition for why this is, let's examine the extreme cases of 1 measurement per sample and the entire population in every sample. Where one measurement is taken per sample, each sample mean is the same as the corresponding data point and therefore the sampling distribution should be an approximation of the original population. Where the entire population is sampled each time the mean of each sample is the population mean and is therefore the same in every case. This creates a distribution with one data point of relative frequency 1 and all others with relative frequency of 0, which is a very narrow normal distribution. From the first case to the second, as the number of measurements per sample increases, the distribution approaches normality.

This concept of the central limit theorem becomes useful when trying to analyse data which has a less than favourable underlying distribution.

10.5 Standard Error

The *standard error* of a sample statistic is a measure of the standard deviation of a sampling distribution of that sample statistic, estimated from population parameters. The variance of a sampling distribution can be calculated using:

$$\sigma_{\overline{x}}^2 = \frac{\sigma^2}{n}$$

Where n is the number of measurements per sample in the sampling distribution and σ is the standard deviation of the population. The standard deviation of the sampling distribution or the standard error of the mean can be calculated by taking the square root of both sides of the above equation as follows.

$$\mathsf{SE}_{\overline{x}} = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

As proof for this formula consider the random variables X_i, where $1 \leq i \leq n$, each set having the same distribution with variance σ^2. We will take these to be the samples from an overall population.

Thus the standard error should be equal to the variance of the sample means or:

$$\operatorname{Var}(\bar{x}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

Using the identity that $\operatorname{Var}(kX) = k^2 \operatorname{Var}(X)$, it is true that:

$$\operatorname{Var}(\overline{x}) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right)$$

Given that these variables are independently sampled, it is true that:

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \operatorname{Var}(X_i)\\ &= \sum_{i=1}^{n} \sigma^n\\ &= n\sigma^2\end{aligned}$$

Hence,

$$\begin{aligned}\operatorname{Var}(\overline{x}) &= \frac{1}{n^2} n\sigma^2\\ &= \frac{\sigma^2}{n}\end{aligned}$$

Example: Standard Error with Sample Means

Eggs are sold in cartons of 12 which represents a random sampling of all available eggs. Given that the mean mass of eggs is 90 grams with a standard deviations of 5g. What is the probability that the total mass of eggs in one carton exceeds 1.1kg?

The probability of the total of 12 eggs being greater than 1.1kg is the same as the probability of that the average mass of each egg is greater than $\frac{1100}{12} = 91\frac{2}{3}$g. To find this probability given a sample of 12, calculate the standard error for a sampling distribution of the sample mean with 12 measurements per sample.

$$\mathsf{SE}_{\overline{x}} = \frac{5}{\sqrt{12}} = \frac{5\sqrt{3}}{6}$$

The z score of the desired value on the original distribution of egg mass is equal to:

$$z_{91\frac{2}{3}} = \frac{91\frac{2}{3} - 90}{\frac{5\sqrt{3}}{6}} = \frac{2\sqrt{3}}{3} \approx 1.547$$

Using a z table or computer program, the probability of a z score of 1.547 or greater on a normal distribution is approximately 12.4%.

While it may seem intuitive within this example to want to use the original probability density curve to calculate this probability, this provides an inaccurate result. When examining a single egg the original distribution can be used effectively to find this probability. The reason this, or an extension of this idea, will not work for larger samples, is that within the sample of 16, some eggs will be of less mass and some will be of greater mass, compensating for each other to create a mean similar to that of the population. Given that the sampling distribution of the sample mean is the distribution of means from samples of a certain size, by taking this distribution and stretching the x axis out by a factor 16 and compressing the frequency by the same factor, we now have a distribution for the probability of the total mass of each carton exceeding a certain value.