Random Variables

Damian Cox

March 3, 2019

Contents

1 Introduction

For many cases where we are looking at random events, the result of the experiment being done is quantifiable; that is, it is a measurable number. The event in turn involves a number. In such a case we are dealing with a variable which is random in nature, the subject of this section. In the case of a Random Variable:

- The experiment is producing and then measuring the variable.
- The event is the variable taking on a particular value, or range of values.

We will see that some of the examples we have already looked at can be recast as random variables.

1.1 Rolling Two Dice Revisited

In probability, we have studied the example of rolling two dice and the event of the difference being 2. The probability was found by counting the total number of outcomes in the sample space, counting those outcomes that satisfied the criteria of the event, then the proportion of the latter was the required probability, given that all outcomes are equally likely.

We will now treat this case as an example of a random variable, because, by calculating the difference, we are treating the symbols on the dice as numbers. In this case:

- The experiment is the rolling of the dice and seeing which numbers come up, then finding the difference.
- The outcomes are whatever numbers show up and the corresponding differences.
- The Random variable is the difference of the two numbers, which will be denoted as D.

- The event may now be written as $D = 2$.
- The business of calculating the probability proceeds as before, 8 outcomes satisfy the criterion for the event so $P[D = 2] = \frac{8}{36} = \frac{2}{9}$.

This case could be completely rewritten in terms of the random variable and changing the definition of the sample space.

- The experiment is the rolling of the dice and seeing which difference comes up.
- The sample space will be the numbers 0 to 5.
- The event may now be written as $D = 2$.
- The number of times the outcomes come up, out of 36, are:

0	1	2	3	4	5
6	10	8	6	4	2

- The probabilities of each of the events $D = i$, where i is a number from 0 to 5, are then the frequencies divided by 36.

2 Fundamental Ideas and Definitions

We will now look at the fundamental ideas needed to understand and use the concept of a random variable.

2.1 Discrete and Continuous

One of the first distinctions that must be made about a random variable is the type of number it may produce. This is done with the following two definitions.

2.1.1 Definition: Discrete Random Variable

A Random Variable is said to be discrete if it can only take on values from a finite or countably infinite set of distinct numbers, such as integers or a finite subset of numbers. Therefore the sample space of the random variable can be counted.

This type of random variable often arises as the result of a counting process. An example might be 'the number of students who turn up for a lecture.'

2.1.2 Definition: Continuous Random Variable

A Random Variable is said to be continuous if it can take on any real number as a value, even if within a certain range.

Continuous random variables are very often physical quantities, such as height, mass, distance, resistance, capacitance and times. In this case, it is no longer possible to talk about the chances of the variable being equal to particular values; since there are an infinite number of possible values, the chances of one particular value coming up are zero. Instead we talk about the probability of the variable being in a particular range. So if R is a discrete variable, the events can be of the form

$$R = 0; R = 1 \text{ or } R > 4.$$

With a continuous variable such as height H, the events are

$$H > 160cm \text{ or } H < 170cm.$$

2.2 Distributions

Random variables usually have a distribution; this is a law governing the probability of particular values of the variable coming up. In the case of a discrete random variable, this could simply be a list of probabilities for each value. More typically, it will be an equation giving the probability in terms of the value.

2.2.1 A Simple Example of a Distribution

Consider again the example of a random variable generated by rolling two dice and calculating the difference. This was a simple example of a discreet random variable, since the difference can only take on certain values, 0 up to 5. To calculate the probabilities for this variable we drew up the sample space and from this we wrote down the probability of all events based on counting the outcomes that satisfied its criterion. Here are the probabilities of the simple events of the form $D = i$, where i is a number from 0 to 5; they have been written as fractions of 18, the lowest common denominator:

$$P[D=0] = \frac{6}{36} = \frac{3}{18}, \ P[D=1] = \frac{10}{36} = \frac{5}{18}, \ P[D=2] = \frac{8}{36} = \frac{4}{18},$$
$$P[D=3] = \frac{6}{36} = \frac{3}{18}, \ P[D=4] = \frac{4}{36} = \frac{2}{18}, \ P[D=5] = \frac{2}{36} = \frac{1}{18}.$$

This list is complete; it gives the probability of every possible event formed from one outcome. It is therefore a distribution for the variable D, the difference between the two dice. The probability of any other event can be calculated from this list using the laws of probability.

2.2.2 A Uniform Distribution

Here is another example of a simple discrete distribution. Let X be a random variable which takes on one of the ten discrete values

$$1, 2, \ldots, 9, 10.$$

This list is the sample space for X.

The simple events for this variable will be of the form $X = i$, where i is a number from 1 to 10. Let us say that each value is equally likely to come up.

Then:

$$P[X = n] = 0.1, \text{ where } n = 1, 2, , \ldots, 9, 10.$$

This is an example of a very simple distribution called the discrete uniform distribution, where all the values for the variable have the same probability of coming up. If all probabilities of these events are added up, we get 1, as should be the case:

$$\sum_{1}^{10} P[X = i] = \sum_{1}^{10} 0.1 = 10 \times 0.1 = 1.$$

The sum is the same as the calculation of the probability of the event that X takes on a number between 1 and 10, which must be 1.

2.3 The Exponential Distribution

We will now look at a very important continuous distribution. Let X be a random variable which can take on any positive number and let λ be a positive parameter. We will describe this distribution by giving an equation for the probability the random variable X is less than a given value x:

$$P[X < x] = 1 - e^{-\lambda x}.$$

This distribution is called the Exponential Distribution. It is important that for x being any positive number, the probability goes from 0 to 1, as should be the case. In addition, the probability we get a number less than 0 has to be 0:

$$P[X < 0] = 1 - e^{-\lambda \times 0} = 1 - e^{0} = 1 - 1 = 0.$$

We can also see that as the possible value x gets larger and larger, the probability gets closer to 1, as the exponential term shrinks rapidly. In other words, as x gets higher, then getting a value less than x is more and more likely. This is a defining and very important property of this distribution.

The exponential distribution is vitally important for studying the probability of an occurrence happening within a certain time, where these occurrences happen per unit time at a fixed rate. This is particularly applicable to certain types of failure rates.

Consider a situation where a certain type of occurrence is happening in time and let λ be the fixed number of events happening per unit time. This number is itself independent of the time. Let T be the random variable of the time until the occurrence happens. Then the probability that the event will occur within a time t, in other words that $T < t$, is given by the Exponential Distribution:

$$P[T < t] = 1 - e^{-\lambda t}.$$

2.3.1 Example – Telephone Calls

Telephone calls are known to arrive in an exchange at a rate of 15 per minute. Calculate the probability that a call comes within 10 seconds.

If the calls are arriving at 15 per minute, then the rate (working with seconds as our unit of time) is $\lambda = \frac{1}{4}$. Then the required probability is:

$$P[T < 10] = 1 - e^{-\frac{1}{4}10} = 1 - e^{-2.5} = 1 - 0.082 = 0.918.$$

This result means that the probability of getting a call within the ten seconds is very high, 0.918. This is what we would expect, since the rate means that they arrive on average at one call every 4 seconds. Here is a graph of this probability, for the parameter $\lambda = \frac{1}{4}$.

Clearly as the length of time increases, the probability of the occurrence happening, that is, receiving the telephone call, increases rapidly to 1.

2.4 Definitions for a Distribution

We will now look at some very important ideas concerning distributions for both Discrete and Continuous Random Variables. Much of our later discussions of the more important distributions will centre on the ideas we are going to cover here.

2.4.1 Cumulative Distribution Function

We have stated already that the concept of the distribution essentially means an equation giving the probability of a given value or range of values coming up for a random variable. This usually takes the form of a function giving us the probability the variable is less than a particular value.

Let X be a random variable, either continuous or discrete. Then the cumulative Distribution function is a function F such that

$$F(x) = P[X \leq x]$$

for a given value x. The function $F(x)$ gives the probability the variable X produces a value less than x. In practice, the function F is defining the variable, so we usually write

$$P[X \leq x] = F(x).$$

As an immediate consequence, we can write the probability of the variable X giving a value between two numbers a and b:

$$P[a < X \leq b] = F(b) - F(a).$$

To see why this is true, look at the three events:

$$X \leq a, X \leq b \text{ and } a < X \leq b.$$

We can see that the following relation between the three events is true:

$$[X \leq a] \text{ or } [a < X \leq b] = [X \leq b].$$

The two events on the LHS are mutually exclusive so that we can say

$$P[X \leq a] + P[a < X \leq b] = P[X \leq b].$$

Bring the first term across and we have

$$P[a < X \leq b] = P[X \leq b] - P[X \leq a].$$

We have shown that

$$P[a < X \leq b] = F(b) - F(a).$$

2.4.2 Probability Density Function

Related to the idea of a Cumulative Distribution Function is the Probability Density Function, the PDF. This is defined as follows for a continuous random variable X with Cumulative Distribution Function F, so:

$$F(x) = P[X \leq x].$$

Then the Probability Density Function f is the derivative of F with respect to x:

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}.$$

From this it follows that the CDF F is the integral of f. More precisely, the probability $F(x)$ is found by integrating the function f from the lower limit of the sample space, usually negative infinity, to x:

$$P[X \leq x] = \int_{-\infty}^{x} f(u)\mathrm{d}u$$

This idea is very important because in many contexts, when a distribution is being investigated either for the first time or when it arises in an analysis of a scientific or engineering question, the Probability density function is the more useful way to represent the variable. It also leads to the definition of some very important ideas as we will see.

With the definition of the density function, the probability of a random variable X taking on a value between two numbers can be defined in terms of the integral:

$$P[a < X < b] = \int_a^b f(u)\mathrm{d}u$$

With this idea, the PDF can be thought of as measuring the relative likelihood a variable takes on certain values.

For a discrete variable, there is no equivalent of differentiating the CDF, since the function does not depend on a continuous variable. Instead we talk about the Probability Mass Function, which is simply the function that gives probability the variable takes on a value. We saw an example of this with the analysis of the probability of the difference between two dice being rolled.

The following sections contain important examples of these ideas.

2.4.3 The Uniform Distribution

Let X be a random variable, which takes on one of the K discrete values $1, 2, \ldots, K$. This means that the simplest events involving this variable are:

$$X = n, \text{ where } n = 1, 2, \ldots, K.$$

For the Uniform Distribution, each value is equally likely to come up. Then for $n = 1, 2, \ldots, K$, set

$$P[X = n] = p,$$

where p is a probability. Then it must be the case that:

$$\sum_1^K P[X = i] = \sum_1^K p = Kp.$$

The sum is the same as the calculation of the probability of the event that X takes on a number between 1 and K, which must be 1. Therefore $Kp = 1$ and so $p = \frac{1}{K}$.

We have studied the example of the dice, which can be viewed as a discrete uniform random variable with $K = 6$ and $p = \frac{1}{6}$.

We will now write this formally as a Probability mass function for the variable X . It is:

$$P[X = n] = p, \text{ so } f(n) = p.$$

The CDF of this variable can be produced using the addition law of probability. The definition of the CDF is that

$$F(n) = P[X \leq n].$$

We find the probability by breaking it up into its constituent cases:

$$P[X \leq n] = P[X = 1 \text{ or } X = 2 \text{ or } \ldots \text{ or } X = n].$$

Since these are all mutually exclusive events, we can write this as

$$P[X \leq n] = P[X = 1] + P[X = 2] + \ldots + P[X = n].$$

Each of these probabilities is the same, the number p, so now:

$$P[X \leq n] = np.$$

This now means that the CDF of the uniform distribution described is

$$F(n) = np.$$

2.4.4 The Exponential Distribution

We will now find the CDF and PDF of a continuous random variable we have seen.

In a situation where a certain occurrence is happening in time with λ as the fixed number of occurrences happening per unit time. Let T be the random variable of the time until the next occurrence happens. Then the probability that it will occur within a time t, in other words that $T < t$, is given by the Exponential Distribution:

$$P[T < t] = 1 - e^{-\lambda t}.$$

We will identify the CDF and PDF of this distribution.

The CDF is

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

This function is differentiated to give the PDF:

$$f(t) = \frac{\mathrm{d}F}{\mathrm{d}t} = 0 - (-\lambda)e^{-\lambda t} = \lambda e^{-\lambda t}.$$

2.5 Expected Values

At this point we will introduce some ideas for random variables which give a formal version of some intuitive ideas we have around random variables. These are made possible with the ideas of the CDF and PDF.

2.5.1 Expected Values: Continuous Variables

Let X be a continuous random variable with probability density function f. Let R be the range of values which the variable can take on. The expected value of X, denoted by $E[X]$, is given by the equation:

$$E[X] = \int_R uf(u)\mathrm{d}u.$$

In this construction we are multiplying each possible value of the variable X, represented by u in the integral, and then multiplying it by how 'relatively' likely it is, then summing it all up with the integral. Therefore we are selecting what may be regarded as the middle value of the variable, in other words the mean value it can produce. In fact, this is a definition for the mean for a variable:

$$\mu = E[X] = \int_R uf(u)\mathrm{d}u.$$

In the same way, we can define how widely spread the values thrown up by a variable will be with a variance and so a standard deviation. We will denote the variance by Var$[X]$ and define it as

$$\mathrm{Var}[X] = E[(X-\mu)^2] = \int_R (u-\mu)^2 f(u)\mathrm{d}u.$$

Again, we are finding the squared distance of a value of the variable, u, from the mean μ, multiplying by the probability and then summing up this quantity with the integral. A bit of algebra on the integral shows that

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

This also forms the definition for the standard deviation:

$$\sigma^2 = E[(X-\mu)^2] = \int_R (u-\mu)^2 f(u)\mathrm{d}u = \int_R u^2 f(u)\mathrm{d}u - \mu^2.$$

These ideas of mean and standard deviation are analogues of the definitions of mean and variance for a list of numbers, in a very precise way. If we view a dataset as a set of n numbers generated by 'running the experiment' n times for a random variable , then it may now be viewed as a set of values of a variable. This is the idea of a sample. The values of the mean and standard deviation we calculate using the equations of descriptive statistics are now estimates of the true mean and standard deviation of the random variable generating the numbers.

2.5.2 Mean and Variance of the Exponential Distribution

We will calculate the mean and variance (therefore the standard deviation) of the exponential function with parameter λ. The CDF and PDF are

$$F(t) = 1 - e^{-\lambda t},\ f(t) = \lambda e^{-\lambda t}.$$

The expected value is then

$$\mu = E[X] = \int_R u f(u)\mathrm{d}u = \int_0^\infty u\lambda e^{-\lambda u}\mathrm{d}u = \lambda\int_0^\infty u e^{-\lambda u}\mathrm{d}u.$$

Using integration by parts, this integral is

$$\int_0^\infty e^{-\lambda u}u \text{ d}u = \left[\left(-\frac{1}{\lambda}e^{-\lambda u}\right)u\right]_0^\infty - \int_0^\infty \left(-\frac{1}{\lambda}e^{-\lambda u}\right).1 \text{ d}u.$$

When we go to evaluate the first term and tidy up the remaining integral, we see that if $\lambda > 0$, the value of the exponential at ∞ is 0, giving a finite answer of 0. So while t grows linearly to $+\infty$, the exponential term multiplying it shrinks away to 0 much faster, so that we can formally say that

$$\lim_{u\to\infty} e^{-\lambda u}u = 0.$$

Therefore we can put in the limits of the integral as finite numbers:

$$0 - 0 + \frac{1}{\lambda}\int_0^\infty e^{-\lambda u} \text{ d}u = \frac{1}{\lambda}\int_0^\infty e^{-\lambda u} \text{ d}u.$$

Now carry out the remaining integration, keeping in mind that $\lambda > 0$:

$$\int_0^\infty e^{-\lambda u} \text{ d}u = \left[\frac{1}{-\lambda}e^{-\lambda u}\right]_0^\infty = 0 - \left(-\frac{1}{\lambda}.1\right) = \frac{1}{\lambda}.$$

The result for the overall integral is now:

$$\int_0^\infty e^{-\lambda u}u \text{ d}u = \frac{1}{\lambda^2},$$

so the expected value is

$$\mu = E[X] = \lambda\frac{1}{\lambda^2} = \frac{1}{\lambda}.$$

So the mean value for the exponential distribution is the inverse of the parameter λ, the rate at which the occurrences happen. In other words the mean value is the mean time between occurrences. The exponential distribution is often written using this parameter instead:

$$P[T \le t] = 1 - e^{-\lambda t} = 1 - \exp\left(-\frac{t}{\mu}\right).$$

Now we will find the variance and standard deviation.

$$\sigma^2 = E[(X-\mu)^2] = \int_R (u-\mu)^2 f(u)\text{d}u = \int_R u^2 f(u)\text{d}u - \mu^2.$$

This means we need to evaluate

$$E[X^2] = \int_0^\infty u^2 \lambda e^{-\lambda u} \mathrm{d}u.$$

To evaluate the integral

$$\int_0^\infty u^2 e^{-\lambda u} \; \mathrm{d}u$$

we use integration by parts:

$$\int_0^\infty u^2 e^{-\lambda u} \mathrm{d}u = \left[\frac{1}{-\lambda} e^{-\lambda u} u^2 \right]_0^\infty - \int_0^\infty \left(\frac{1}{-\lambda} e^{-\lambda u} \right) 2u \; \mathrm{d}u.$$

Putting in the limits in the first term, it is 0, so

$$\int_0^\infty u^2 e^{-\lambda u} \mathrm{d}u = \frac{2}{\lambda} \int_0^\infty e^{-\lambda u} u \; \mathrm{d}u.$$

We have just calculated the remaining integral, so the result here is

$$\int_0^\infty u^2 e^{-\lambda u} \mathrm{d}u = \frac{2}{\lambda^3}.$$

With this result,

$$\sigma^2 = \int_0^\infty u^2 f(u) \mathrm{d}u - \mu^2 = \lambda \frac{2}{\lambda^3} - \mu^2 = \frac{2}{\lambda^2} - \mu^2.$$

Recall our result that $\mu = \frac{1}{\lambda}$; then

$$\sigma^2 = \frac{2}{\lambda^2} - \mu^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

This means that the standard deviation is the same as the mean:

$$\sigma = \frac{1}{\lambda}.$$

2.5.3 Expected Values: Discrete Variables

There are corresponding definitions for ideas for discrete variables. The distribution for a discrete variable will be the probability mass function, a function giving the probability for each value of the random variable that can come up.

$$P[X = X_i] = f(X_i), \text{ for } i = 1, 2, \ldots$$

The expected value for the variable is then

$$\mu = E[X] = \sum_i X_i P[X = X_i] = \sum_i X_i f(X_i),$$

in other words, the sum of the given values times the probability of getting that value. The variance and standard deviation are given by

$$\sigma^2 = E[X^2] - \mu^2 = \sum_i X_i^2 P[X = X_i] - \mu^2 = \sum_i X_i^2 f(X_i) - \mu^2.$$

So far, the main discrete distribution we have seen is the uniform distribution; let X be a random variable, which takes on one of the K discrete values $1, 2, \ldots, K$. Each value is equally likely to come up:

$$P[X = n] = p,$$

for $n = 1, 2, \ldots, K$, and $p = \frac{1}{K}$. The expected value is:

$$E[X] = \sum_i X_i P[X = X_i] = \sum_i i \frac{1}{K} = \frac{1}{K} \sum_i i.$$

The sum of the numbers 1 to K is

$$\sum_i i = \frac{K(K+1)}{2}$$

so that

$$E[X] = \frac{1}{K} \frac{K(K+1)}{2} = \frac{K+1}{2}.$$

The standard deviation is now

$$\sigma^2 = \sum_i X_i^2 P[X = X_i] - \mu^2 = \frac{1}{K} \sum_i i^2 - \mu^2.$$

The sum of the squares of the numbers from 1 up to K are:

$$\sum_i i^2 = \frac{K(K+1)(2K+1)}{6}$$

so the standard deviation is

$$\sigma^2 = \frac{1}{K}\frac{K(K+1)(2K+1)}{6} - \left(\frac{K+1}{2}\right)^2.$$

Multiply out the terms above the line:

$$\sigma^2 = \frac{2K^2+3K+1}{6} - \frac{K^2+2K+1}{4}.$$

Add the fractions and simplify this expression down:

$$\sigma^2 = \frac{2(2K^2+3K+1)}{12} - \frac{3(K^2+2K+1)}{12} = \frac{K^2-1}{12}.$$

3 Two Important Discrete Distributions

We will now introduce two important discrete distributions. The second governs a random variable where we can see the origin of the equation defining the distribution. It is also the case that the first is derived from the second, though this is beyond the scope of our module.

3.1 The Poisson Distribution

We used the exponential distribution to study the probability of an occurrence happening within a certain time, where the number of occurrences happening per unit time is a fixed number, that is, independent of the time. A second distribution arises in the same context. Consider a situation where a certain type of occurrence is happening in time and let λ be the fixed number of events happening per unit time. Let T be a given interval of time. Let N be the random variable of the number of these events which occur in a time T. Then the random variable N has probability mass function

$$P[N=n] = e^{-\lambda T}\frac{(\lambda T)^n}{n!}.$$

This PMF is often written by setting $\mu = \lambda T$ and so the equation becomes:

$$P[N=n] = e^{-\mu}\frac{\mu^n}{n!}.$$

In fact the quantity μ is the mean for the distribution:

$$\mu = E[N] = \lambda T.$$

It is also the case that the variance and standard deviation are given by

$$\sigma^2 = \mu^2.$$

Here is an example of a calculation for this distribution, in the context of the example of telephone calls we studied above. Be mindful of the difference between this distribution and the exponential.

3.1.1 Example - Telephone Calls

Telephone calls are known to arrive in an exchange at a rate of 15 per minute. Calculate the probability that 3 calls come in a ten second period.

If the calls are arriving at 15 per minute, then the rate (working with seconds as our unit if time) is $\lambda = \frac{1}{4}$. Then with $T = 10$ the parameter μ is given by $\mu = \frac{1}{4} \times 10 = 2.5$. Then the probability of getting 3 calls in this time period is:

$$P[N = 3] = e^{-2.5}\frac{2.5^3}{3!} = 0.082 \times \frac{15.625}{6} = 0.2135.$$

The following example will show the link between the exponential and Poisson distributions. Always be mindful of the fact that the exponential distribution deals with time, so is a continuous variable, whereas the Poisson distribution deals with a count of incidents, so is a discrete variable.

3.1.2 Example - A Toll booth

A company operates a tollgate on an Autoroute in the Pyrénées. It has been established that for the month of October, cars arrive at this booth at a rate of one every 2 minutes. Carry out the following three calculations.

1. Calculate the probability that a car arrives within the next 4 minutes from a given starting time.

2. Calculate the probability that 3 cars arrive in a ten-minute period for the toll booth.

3. For another toll-booth, the rate is one car every 4 minutes. Calculate the probability that another car arrives within 8 minutes there.

The solutions are presented in turn:

1. This is the exponential distribution. The rate information is '1 car every 2 minutes', so this means $\lambda = \frac{1}{2}$ cars per minute. The period of time is $T < 4$ minutes. So if T is the random variable of the amount of time till the next event, then

$$P[T < 4] = 1 - e^{-\lambda 4} = 1 - e^{-2} = 1 - 0.1353 = 0.8647.$$

2. This is now the Poisson distribution, with parameter $\mu = \lambda T = 5$.

$$P[N = 3] = e^{-5}\frac{5^3}{3!} = 0.00674 \times \frac{125}{6} = 0.1404.$$

While we are here, we will calculate the figures for $N = 2$ and $N = 4$:

$$P[N = 4] = e^{-5}\frac{5^4}{4!} = 0.00674 \times \frac{625}{24} = 0.174.$$

$$P[N = 2] = e^{-5}\frac{5^2}{2!} = 0.00674 \times \frac{25}{2} = 0.0842.$$

Clearly the probability for $N = 4$ is the highest since it is closest to the expected value $\mu = 5$.

3. The rate information now is '1 car every 4 minutes', so this is $\frac{1}{2}$ cars per minute. The period of time is 8 minutes. So if T is the random variable of the amount of time till the next event, then

$$P[T < 8] = 1 - e^{-\lambda 8} = 1 - e^{-2} = 1 - 0.1353 = 0.8647.$$

This is exactly the same as the fist calculation because in both cases the time in question was double the mean time between incidents.

3.2 The Binomial Distribution

Consider the following problem - it is known that 11% of the Irish population is left-handed. The following question is posed: in a randomly selected group of 12 people, find the probability of finding 3 people who are left-handed. To answer this question, look at one case - calculate the probability of the first 3 of the 12 being left-handed. It was said at the start that the group have been randomly selected; this means that we can assume there is no connection between them, so each person being left or right-handed is an independent event. The probabilities of these events for each person are therefore multiplied to find the overall probability, so the probability of the first 3 being left-handed and therefore the remaining 9 being right- handed, is

$$0.11^3 0.89^9.$$

This is just one way of having three people being left-handed. Any selection of 3 from the 12 could be the relevant left-handed people. For each one, the probability of that selection coming up is the same as the value shown above: These possible outcomes are all mutually exclusive, so to find the overall probability, add on the probability above for each combination, in other words, multiply by ${}^{12}C_3$. The probability of 3 left-handed people in a group of 12 is therefore

$${}^{12}C_3 0.11^3 0.89^9.$$

This is an example of what is called the Binomial distribution.

3.2.1 Definition - The Binomial Distribution

A trial is being repeated, with a possible result A. The following is known:

- Each time the trial is done, the probability of result A turning up is p.
- The trial is repeated n times.

Let X be the random variable of the number of times result A comes up. The probability of getting r results from n trials is:

$$P[X = r] =^n C_r p^r (1-p)^{n-r}.$$

Take note of the important feature that the probability of getting result A each time the trial is run is the same, the value p.

The case of the number of left-handed people fits into this pattern.

- The trial being repeated is checking whether or not a person is left-handed. The probability of it occurring for each individual 'tested' is $p = 0.11$. This number comes from the proportion of 0.11 of the wider population being left-handed and then from our definition of a probability.
- In the example we studied, this is being repeated 12 times, so this means $n = 12$.
- Let L be the random variable of the number of left-handers in the group. The probability we are looking at is then $P[L = 3]$. Using the binomial distribution,

  $$n = 12, p = 0.11, 1 - p = 0.89.$$

  So

  $$P[L = 3] =^{12} C_3 0.11^3 0.89^9 = 220 \times 0.00133 \times 0.35036 = 0.1026.$$

For a Binomial Distribution with parameters n and p, it can be proved that the mean and standard deviation are given by

$$\mu = np \text{ and } \sigma^2 = np(1-p).$$

3.2.2 Example: Defective Components

A factory is producing components, of which 1.5% are defective. They are packed in boxes, each containing 20 components. Calculate the probability that a box has 2 defective components.

This is a case of the binomial distribution.

- The trial being repeated is checking whether or not a component is defective. The probability of it occurring for each individual 'tested' is $p = 0.015$. This number comes from the proportion of 1.5 % in the information we have and then from our definition of a probability.
- There are 20 components in each box, so this means $n = 20$.
- let X be the random variable of the number of defective components in a given box. The probability we are looking at is then $P[X = 2]$. Using the binomial distribution,

  $$n = 20, p = 0.015, 1 - p = 0.985.$$

  So

  $$P[X = 2] =^{20} C_2 0.015^2 0.985^{18} = 190 \times 0.0152 \times 0.98518 = 0.033.$$

Thus of every 1000 boxes coming out of the factory, 33 would have two defectives.

We will continue working with this example and calculate the probabilities of the following events.

1. There are no defectives,
2. There is 1 defective.
3. There are 2 or more defectives.

4. There are less than 3 defectives,

Here are the calculations:

1. The probability of getting no defectives is

$$P[X=0] = ^{20} C_0 0.015^0 0.985^{20} = 1 \times 1 \times 0.98520 = 0.739.$$

   Bear in mind that $^{20}C_0 = 1$, this is because there is only one way of having no defectives in the box.

2. The probability of getting one defective is:

$$P[X=1] = ^{20} C_1 0.015^1 0.985^{19} = 20 \times 0.015 \times 0.98529 = 0.225.$$

3. To find the probability of getting less than three defectives, first look at exactly what this means. It means that

$$X = 0 \text{ or } X = 1 \text{ or } X = 2.$$

   The probability of getting less than three defectives is then

$$P[X < 3] = P[X = 0 \text{ or } X = 1 \text{ or } X = 2].$$

   Since these are mutually exclusive events, the probabilities can be added:

$$P[X < 3] = P[X = 0] + P[X = 1] + P[X = 2].$$

   These are all probabilities we have worked out. So this means

$$P[X < 3] = 0.739 + 0.225 + 0.033 = 0.997.$$

4. To calculate the probability that in a box of 20, two or more are defective, we could add the probabilities for $X = 2$, $X = 3$, $X = 4$ and so on. However, it would be simpler to observe that the event of getting two or more defectives is the complement of getting none or one. Thus

$$P[X \geq 2] = 1 - P[X = 0 \text{ or } X = 1].$$

Since these are mutually exclusive events on the right-hand side, the probabilities can be added:

$$P[X = 0 \text{ or } X = 1] = P[X = 0] + P[X = 1] = 0.739 + 0.225 = 0.964.$$

So this means

$$P[X \geq 2] = 1 - 0.964 = 0.036.$$

3.2.3 Multiple Choice Exams

A student, lacking in knowledge, takes a multiple-choice examination and answers the questions at random. The binomial distribution allows us to calculate the probability the student passes. Here is an example.

A multiple-choice exam has 20 questions, all with a choice of five answers. A student chooses their answers at random. Calculate the probability they pass the exam, with 40% as the pass mark.

- The exam has 20 questions, the trial 'chose the correct answer to this question?' is repeated 20 times; so $n = 20$.
- Each time a student answers a question, they have a choice of 5 answers, so there is a probability of 0.2 of getting it right. Then $p = 0.2$.

Let X be the random variable of the number of questions the student answers correctly. Then the probability of getting r questions correct is

$$P[X = r] =^{20} C_r 0.2^r 0.8^{20-r}$$

To pass the exam, a student needs 8 questions or more right. This event is:

$$X \geq 8 \equiv X = 8 \text{ or } X = 9 \text{ or } \ldots \text{ or } X = 20.$$

Since these are all mutually exclusive events:

$$P[X \geq 8] = P[X = 8] + P[X = 9] + \ldots + P[X = 20].$$

When calculating these probabilities, they will be quite low, and if a level of accuracy is chosen, say three decimal places, then the individual probabilities will quickly be effectively zero. It is therefore quite easy to calculate. The results are;

- $P[X = 8] =^{20} C_8 0.2^8 0.8^{12} = 0.022161.$
- $P[X = 9] =^{20} C_9 0.2^9 0.8^{11} = 0.0073866.$
- At $X = 12$ the numbers are very low: $P[X = 12] =^{20} C_{12} 0.2^{12} 0.8^8 = 0.000086.$

To three decimal places:

$$P[X \geq 8] = 0.0321.$$

It is notable that the expected value for this case of the distribution, the expected value of the number of correct questions, is 20 x 0.2 = 4, so we expect to get 4 questions right 'on average'.

3.2.4 Approximating the Binomial Distribution

A very useful approximation for calculating the Binomial distribution is shown here. If the number n is reasonably large, and the probability p is a small number, such as is shown in the previous example, then the Binomial probabilities can be written as Poisson probabilities, with $\mu = np$. In the last case the components have been packed in boxes of 20 and the probability for each individual component being defective was very low, 0.015. The mean figure for the Poisson is then

$$\mu = np = 20 \times 0.015 = 0.3.$$

This number is the expected number of defective components found in a box of 20 and, since the closest possible integer is 0, it is no surprise that $X = 0$ was the event with the highest probability.

4 The Normal Distribution

4.1 Introduction

The Normal distribution is one of the most important distributions of a continuous variable, arising in many contexts in nature. It has several very important properties which are precisely why it comes up so much. Because of its ubiquity, it is crucially important for statistical testing, as we shall see.

We will define it first and then investigate its properties.

4.1.1 Definition of the Normal Distribution

The Normal distribution is characterised by two real parameters a and b, with $b > 0$, and defined by the following probability density function.

$$f(x) = \frac{1}{b\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right).$$

The first notable thing about this distribution is that it is defined for all real numbers and so the variable can give any real number as a value.

From the PDF, the Cumulative Distribution function is given by

$$P[X < x] = \int_{-\infty}^{x} \frac{1}{b\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{u-a}{b}\right)^2\right) \, \mathrm{d}u$$

4.1.2 Parameters of the Distribution

Some work with the defining equations of this distribution will show that the mean and variance are given by

$$\mu = E[X] = a \text{ and } \sigma^2 = \mathrm{Var}[X] = b^2.$$

This means that in the original defining equation, the numbers a and b are the mean and standard deviation respectively. So therefore we write the PDF as:

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

and the cumulative distribution function is then

$$P[X < x] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right) \; \mathrm{d}u$$

Recall that for many cases of the use of the binomial distribution, the probability p of the event being counted often came from proportions of the population the sample came from. In a similar way, the mean and the standard deviation for the normal distribution may come from the analysis of a population. Thus the ideas of the mean and standard deviation are the same as those from the study of data presentation. We will, however, explore this idea in more detail under statistical testing.

If a variable X is normally distributed, this means that if a large number of values are generated of the variable, then they will be more likely to be close to the mean μ and unlikely to be far from it. Just how 'likely' or 'unlikely' is determined by the standard deviation σ.

The integral in the definition of this distribution can not in fact be done by normal algebraic or analytical methods. To be used practically, the integration is done by numerical integration. We will see ways to use this later on.

4.1.3 The Standard Normal Distribution

Consider the case where the mean is $\mu = 0$ and the standard deviation is $\sigma = 1$, the equation for the density function reduces to

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

This case is known as the standard Normal distribution and the variable is denoted by Z. The cumulative distribution function is then

$$P[Z < z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\tfrac{1}{2}u^2\right) \; \mathrm{d}u.$$

One result that can be proved using analytical geometry is:

$$\int_{-\infty}^{\infty} \exp\left(-\tfrac{1}{2}u^2\right) \; \mathrm{d}u = \sqrt{2\pi}.$$

This ensures that the following result is valid:

$$P[Z < \infty] = 1.$$

This would be expected for any proper CDF; in other words, the sum of probabilities for all the real numbers is 1.

An extremely useful result that can be proved by substitution in the equation defining the CDF is stated here. Let Z be the variable with the standard normal distribution and let X be a normally distributed random variable with mean μ and standard deviation σ. Then

$$P[X < a] = P\left[Z < \frac{a-\mu}{\sigma}\right] \text{ or } P[Z < b] = P[X < \mu + \sigma b].$$

This result essentially says that if a normally distributed random variable is re-scaled to the standard normal distribution then its probabilities do not change. It effectively means that the two variables are linked by the relation

$$Z = \frac{X-\mu}{\sigma} \text{ or } X = \mu + \sigma Z.$$

Thus every random variable that follows a normal distribution with a given mean and standard deviation may be viewed as a recaled version of the stnadard normal variable Z.

4.1.4 The Bell Curve

Because the integral defining the CDF for this important distribution relies on numerical integration or a spreadsheet for practical use, the normal distribution is often visualised using a graph of the curve for the density function. This is a very useful way to remember key properties of the distribution that allow its simple and easy use. The graph is shown here; it is the well-known Bell curve.

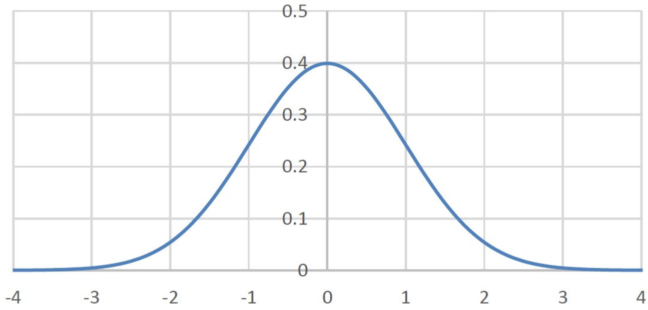

Figure 1: The PDF of the Standard Normal Distribution

Since the probability $P[X < a]$ for a distribution is the integration of the density function, then that probability can be viewed as the area under the curve of the PDF.

Recall that PDF for the standard normal distribution, which we will now denote by f_Z, is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}z^2\right).$$

Since the variable z appears in this function as a square, it means that it is symmetric about the vertical axis, as can be seen in the graph. The total area under the curve will be equal to 1.

We saw that any normally distributed random variable X with mean μ and standard deviation σ can be linked to the standard variable by the 're-scaling' equations

$$P[X < a] = P\left[Z < \frac{a-\mu}{\sigma}\right] \quad \text{or} \quad P[Z < b] = P[X < \mu + \sigma b].$$

This means a graph for the PDF of X can be constructed by relabelling the point $z = 0$ as $X = \mu$ and then $z = 1$ as $X = \mu + \sigma$, $z = 2$ as $X = \mu + 2\sigma$ and so on. The negative values are relabelled in a corresponding way.

We will now look at some probabilities for the standard normal variable Z and see how they are represented on the Bell curve. Consider the following probability with its integral representation:

$$P[Z < -1] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1} \exp\left(-\tfrac{1}{2}u^2\right) \, \mathrm{d}u.$$

This can be represented as an area under the Bell curve. It is the area shaded in blue shown in figure 2. Conversely, let us look at the event $Z > 1$. Then

$$P[Z > 1] = \frac{1}{\sqrt{2\pi}} \int_{1}^{\infty} \exp\left(-\tfrac{1}{2}u^2\right) \, \mathrm{d}u.$$

This will be the same as the previous probability, as illustrated as the area under the curve shaded green shown in figure 3.

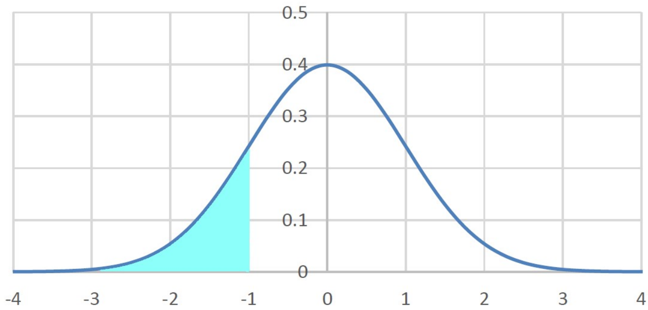

Figure 2: The Probability $Z < -1$

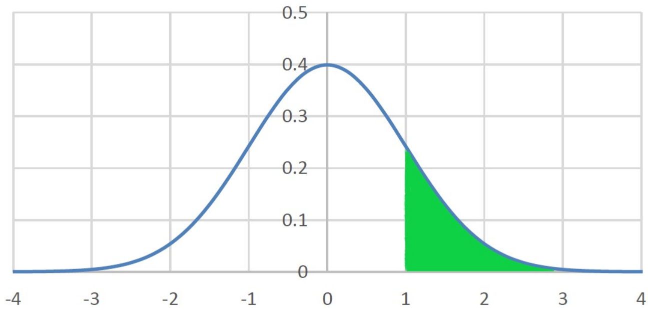

Figure 3: The Probability $Z > 1$

The symmetry of the PDF shown in these two figures convinces us that

$$P[Z > 1] = P[Z < -1].$$

The diagrams also illustrates that the un-shaded area in figure 3 represents the probability $P[Z < 1]$, because the laws of probability tells us that

$$P[Z < 1] + P[Z > 1] = 1$$

and the total area under the curve must be 1.

Observing the symmetry of the curve also reminds us that

$$P[Z < 0] = 0.5,\ P[Z > 0] = 0.5$$

and then

$$P[X < \mu] = 0.5,\ P[X > \mu] = 0.5.$$

4.2 Using the Normal Distribution

Having defined the distribution and seen how it is represented geometrically, we will now state the symmetry property of the distribution and proceed on to show how probabilities are calculated in practice for the distribution.

Let Z represent the standard normal variable and let f_Z represent its the PDF, as before.

4.2.1 Symmetry Property of Z

The standard normal distribution has the property that for any real number a then

$$P[Z > a] = P[Z < -a].$$

We can also use the simplest laws of probability in calculating values from any normally distributed variable:

$$P[Z < 1] = 1 - P[Z > 1] \text{ or } P[Z > 1] = 1 - P[Z < 1].$$

4.2.2 The Z-Tables

Before the advent of modern spreadsheets and the computing power to make them so useful in statistical applications, the main method of dealing with the non-integrable nature of the CDF of the normal distribution was by means of a tables. The steps involved were as follows.

- A Table was produced giving the values of the probability $P[Z > a]$ or equivalently $P[Z < a]$, for values a.
- To find a probability, such as $P[X > a]$, involving a normally distributed random variable X with mean μ and standard deviation σ, it is re-scaled int one involving Z, using the equation

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

  This is called standardisation. In practice this means calculating the figure:

  $$\frac{a - \mu}{\sigma}.$$

- This figure found in the tables, if need be takimng note any negative signs and finding the absolute value, and so the probability is found:

  $$P\left[Z > \left|\frac{a - \mu}{\sigma}\right|\right].$$

- This then gives us the probability we want by means of symmetry and the laws of probability.

Here is a small section of the table of probabilities $P[Z > a]$:

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07 ...
0.0	0.5000	0.4960	0.4920	0.4880	0.484	0.4801	0.4761	0.4721
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423
1.1	0.1357	0.1335	**0.1314**	0.1292	0.1271	0.1251	0.123	0.1210
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192

The table is used as follows. It is a table for values $P[Z > a]$, so the number a can be found to two decimal places. The first two figures are found by working down the delimiting values on the left-hand side, the next figure is found working across the top. These then form coordinates for a cell in the table, where the probability is found. Thus the number shown in bold in the middle of the table is the probability for 1.1 and 0.02, so

$$P[Z > 1.12] = 0.1314.$$

We will now show a variety of examples of finding probabilities using standardisation and these tables.

4.2.3 Example: Ball-bearing Production

The diameter of the ball bearings being produced in a factory is a normally distributed random variable, D, with mean and standard deviation

$$\mu = 4.02\text{mm}, \ \sigma = 0.07\text{mm}.$$

Calculate the probability of the following events:

1. A ball-bearing chosen at random has a diameter greater than 4.1mm.

2. A ball bearing has a diameter less than 4.05 mm.

3. A ball bearing has a diameter less than 3.9 mm.

4. A ball bearing has a diameter more than 4mm.

The solutions are presented here.

The first question means we are calculating the probability

$$P[D > 4.1\text{mm}].$$

To find this probability, firstly work out the standardised number:

$$P[D > a] = P\left[Z > \frac{a - \mu}{\sigma}\right]$$

so

$$P[D > 4.1\text{mm}] = P\left[Z > \frac{4.1\text{mm} - 4.02\text{mm}}{0.07\text{mm}}\right] = P\left[Z > 1.14\right].$$

When this number is found in the tables, the corresponding probability is 0.8729:

$$P\left[Z > 1.14\right] = 0.1271.$$

The procedure here meant that we translated an event involving variable D to one involving Z. To summarise our result:

$$P[D > 4.1\text{mm}] = P\left[Z > 1.14\right] = 0.1271.$$

Take note of how the original question ended up as a 'greater than' problem in the Z variable, so no more work was needed. The unit for the diameters has also been included, so it is abundantly clear whether we are working with D or Z.

For the second question, we proceed as before. The step of rounding off the decimal places of the standardised number is shown:

$$P[D < 4.05\text{mm}] = P\left[Z < \frac{4.05\text{mm} - 4.02\text{mm}}{0.07\text{mm}}\right] = P\left[Z < 0.42857\ldots\right] =$$
$$= P\left[Z < 0.43\right].$$

Looking up 0.43 in the table gives 0.3336, which means that

$$P\left[Z > 0.43\right] = 0.3336.$$

The probability we are looking for is 'less than', so there is one more step:

$$P\left[Z < 0.43\right] = 1 - P\left[Z > 0.43\right] = 1 - 0.3336 = 0.6664.$$

The steps are summarised here:

$$P[D < 4.05\text{mm}] = P\left[Z < \frac{4.05\text{mm} - 4.02\text{mm}}{0.07\text{mm}}\right] = P\left[Z < 0.43\right] =$$
$$= 1 - P\left[Z > 0.43\right] = 1 - 0.3336 = 0.6664.$$

For the third part, the standardisation goes as follows:

$$P[D < 3.9\text{mm}] = P\left[Z < \frac{3.9\text{mm} - 4.02\text{mm}}{0.07\text{mm}}\right] = P\left[Z < -1.71\right].$$

We now have a negative number, so we use the symmetry property of the standard normal distribution:

$$P\left[Z < -1.71\right] = P\left[Z > 1.71\right] = 0.0436.$$

Using the symmetry property may be summarised by saying that the minus sign is dropped (or introduced) and the direction of the inequality is reversed. summarising this calculation:

$$P[D < 3.9\text{mm}] = P\left[Z < -1.71\right] = P\left[Z > 1.71\right] = 0.0436.$$

For the fourth part, the standardisation is:

$$P[D > 4.0\text{mm}] = P\left[Z > \frac{4.0\text{mm} - 4.02\text{mm}}{0.07\text{mm}}\right] = P\left[Z > -0.29\right].$$

With a negative number, use the symmetry property:

$$P\left[Z > -0.29\right] = P\left[Z < 0.29\right].$$

Finally this is the converse event to what is in the tables:

$$P\left[Z < 0.29\right] = 1 - P\left[Z > 0.29\right] = 1 - 0.3859 = 0.6141.$$

Summarising this calculation:

$$P[D > 4.0\text{mm}] = P\left[Z > -0.29\right] = P\left[Z < 0.29\right] =$$

$$1 - P\left[Z > 0.29\right] = 1 - 0.3859 = 0.6141.$$

4.2.4 Example: Heights

The height of men is normally distributed, with mean 1.71 metres and standard deviation 0.11 metres. Find the probability that the height of a man chosen at random is:

1. Greater than 1.76 metres.
2. Less than 1.74 metres.
3. Less than 1.64 metres.
4. Greater than 1.54 metres.
5. Between 1.74 metres and 1.76 metres.
6. Between 1.64 metres and 1.76 metres.

Let H be the random variable of height in men. The mean and standard deviation are

$$\mu = 1.71\text{m}, \ \sigma = 0.11\text{m}.$$

For the first question:

$$P[H > 1.76\text{m}] = P\left[Z > \frac{1.76\text{m} - 1.71\text{m}}{0.11\text{m}}\right] = P[Z > 0.45].$$

So with the tables,

$$P[H > 1.76\text{m}] = P[Z > 0.45] = 0.3264.$$

For the second part, standardising gives

$$P[H < 1.74\text{m}] = P[Z < 0.27] = 1 - P[Z > 0.27] = 1 - 0.3936 = 0.6064.$$

For the third part, standardising means that:

$$P[H < 1.64\text{m}] = P[Z < -0.64].$$

Then apply symmetry:

$$P[Z < -0.64] = P[Z > 0.64] = 0.2611.$$

For the last part:

$$P[H > 1.54\text{m}] = P[Z > -0.17] = P[Z < 0.17] = 1 - P[Z > 0.17] =$$

$$= 1 - 0.4325 = 0.5675.$$

For the fifth question, the probability is broken up as

$$P[1.74\text{m} < H < 1.76\text{m}] = P[H > 1.74\text{m}] - P[H > 1.76\text{m}].$$

since both of the values in this question were above the mean 1.71m, we know that both of these probabilities are calculated with direct 'greater than' use of the tables, so it was convenient to write them as 'greater than' for the order of subtraction. So :

$$P[H > 1.74\text{m}] = P[Z > 0.27] = 0.3936,$$

$$P[H > 1.76\text{m}] = P[Z > 0.45] = 0.3264.$$

The probability we need is

$$P[1.74\text{m} < H < 1.76\text{m}] = P[H > 1.74\text{m}] - P[H > 1.76\text{m}] =$$

$$= P[Z > 0.27] - P[Z > 0.45] = 0.3936 - 0.3264 = 0.0672.$$

For the last question,

$$P[1.64\text{m} < H < 1.76\text{m}] = P[H > 1.64\text{m}] - P[H > 1.76\text{m}].$$

Standardise:

$$P[H > 1.64\text{m}] = P[Z > -0.64] = P[Z < 0.64] = 1 - P[Z > 0.64] =$$

$$= 1 - 0.2611 = 0.7389,$$

$$P[H > 1.76\text{m}] = P[Z > 0.45] = 0.3264.$$

So the required probability is :

$$P[1.74\text{m} < H < 1.76\text{m}] = P[H > 1.74\text{m}] - P[H > 1.76\text{m}] =$$

$$= 0.7389 - 0.3264 = 0.4125.$$

4.3 Working in Reverse

A petrol station finds that the number of litres of petrol sold in a week is a normally distributed random variable with mean 2,500L and standard deviation 200L. Calculate how much petrol they should stock so the probability of running out is 0.02.

For this problem, we are given the probability, we need a figure such that the probability that the sales exceed it is 0.02. If we set S to be the random variable of sales of petrol then the event we are looking at and its probability is

$$P[S > a] = 0.02.$$

The problem is to find the number a.

Finding the probability 0.02 within the body of the tables, we see that it came from the value 2.05. Mathematically,

$$P[Z > 2.05] = 0.02.$$

Thus we need the value of petrol sales that was standardised to give 2.05, then we will have our figure. In other words, we know that

$$\frac{a - \mu}{\sigma} = 2.05.$$

Solving this equation means that

$$a = 2.05\sigma + \mu = 2,500\text{L} + 2.05 \times 200\text{L} = 2,910\text{L}.$$

4.3.1 Example: Ball-bearing Production

Recall the diameter of the ball bearings being produced in a factory is a normally distributed random variable, D, with mean and standard deviation

$$\mu = 4.02\text{mm}, \ \sigma = 0.07\text{mm}.$$

Find a diameter such that there is only a 5% chance a ball bearing exceeds it. Repeat for 2.5%.

For the first question, we are looking for a number a, a diameter, such that

$$P[D > a] = 0.05.$$

If we find the probability 0.05 within the tables, we see it came from the number 1.65:

$$P[Z > 1.65] = 0.05.$$

So we must 'reverse standardise' this figure. This is simply

$$a = 1.65 \times 0.07\text{mm} + 4.02\text{mm} = 4.1355\text{mm}.$$

So then we know that

$$P[D > 4.1355\text{mm}] = 0.05.$$

If we now repeat this for the probability 0.025, we are looking for a number a, a diameter, such that

$$P[D > a] = 0.025.$$

From the tables

$$P[Z > 1.96] = 0.025.$$

Then a is:

$$a = 1.96 \times 0.07\text{mm} + 4.02\text{mm} = 4.1572\text{mm}.$$

So :

$$P[D > 4.1572\text{mm}] = 0.025.$$

4.3.2 Example: Heights

The height of men is normally distributed, with mean 1.71 metres and standard deviation 0.11 metres. Find a height such that the probability that a man chosen at random is not as tall is (i) 0.1 (ii) 0.9. Find a height such that the probability a man chosen at random is taller is 0.95.

Let H be the random variable of height in men. The mean and standard deviation are

$$\mu = 1.71\text{m}, \ \sigma = 0.11\text{m}.$$

For the first question, we are looking for a height a such that

$$P[H < a] = 0.9.$$

The probability 0.9 is not in the tables, so change the question. Equivalently, we are looking for a height a such that

$$P[H > a] = 0.1.$$

From the tables:

$$P[Z > 1.28] = 0.1.$$

Reverse standardise this figure:

$$a = 1.28 \times 0.11\text{m} + 1.71\text{m} = 1.8508\text{m}.$$

For the second part, we are looking for a height b such that:

$$P[H < b] = 0.1.$$

The probability 0.1 came from the number 1.28 again:

$$P[Z > 1.28] = 0.1.$$

The inequality goes the wrong way so use the symmetry property:

$$P[Z < -1.28] = 0.1.$$

So we reverse standardise this figure:

$$b = -1.28 \times 0.11\text{m} + 1.71\text{m} = 1.5692\text{m}.$$

In summary we have this result:

$$P[H < 1.5692\text{m}] = 0.1.$$

For the third question, we are looking for a height c such that

$$P[H > c] = 0.95.$$

This means that

$$P[H < c] = 0.05.$$

The information from the tables is that

$$P[Z > 1.65] = 0.05.$$

Turn around the inequality:

$$P[Z < -1.65] = 0.05.$$

Reverse standardise:

$$c = -1.65 \times 0.11\text{m} + 1.71\text{m} = 1.5285\text{m}.$$

4.3.3 Middle Ranges

For the case of height in men, find the middle 90% of heights. This means two values a and b such that

$$P[a < H < b] = 0.9,$$

but because it is the middle 90%, the probabilities of H being above or below the range are the same. This gives us a way of tackling the calculation, since we have two probabilities:

$$P[H < a] = 0.05 \text{ and } P[H > b] = 0.05.$$

The probabilities of these three events add up to 1. The 0.05 comes from $0.05 = (1 - 0.9)/2$.

These cases are just like the problems above and, and we will see, both involve the same z-value from the tables except for a difference in sign.

For the upper limit:

$$P[H > b] = 0.05; P[Z > 1.65] = 0.05$$

so then

$$b = 1.65 \times 0.11\text{m} + 1.71\text{m} = 1.8915\text{m}.$$

For the lower limit:

$$P[H < a] = 0.05; P[Z > 1.65] = 0.05 \Rightarrow P[Z < -1.65] = 0.05.$$

Reverse standardise:

$$a = -1.65 \times 0.11\text{m} + 1.71\text{m} = 1.5285\text{m}.$$

This shows that once we have found the z value for the upper limit, we have it for the lower limit with a change of sign.

We will establish the middle 95% of values; the probability for either 'tail' is $(1 - 0.95)/2 = 0.025$. From the tables

$$P[Z > 1.96] = 0.025.$$

Then the two limits are:

$$a = -1.96 \times 0.11\text{m} + 1.71\text{m} = 1.4944\text{m},$$

$$b = 1.96 \times 0.11\text{m} + 1.71\text{m} = 1.9256\text{m}.$$