ISDS Week 1 to 3 Notes

Durham University

Contents

1 Basic Concepts

1.1 Some basic concepts

- **Data** consist of information coming from observations, counts, measurements, or responses.
- **Statistics** is the science of collecting, organising, analysing, and interpreting data in order to make decisions.
- A **population** is the collection of all outcomes, responses, measurements, or counts that are of interest. Populations may be finite or infinite. If a population of values consists of a fixed number of these values, the population is said to be finite. If, on the other hand, a population consists of an endless succession of values, the population is an infinite one.
- A **sample** is a subset of a population.
- A **parameter** is a numerical description of a population characteristic.
- A **statistic** is a numerical description of a sample characteristic.

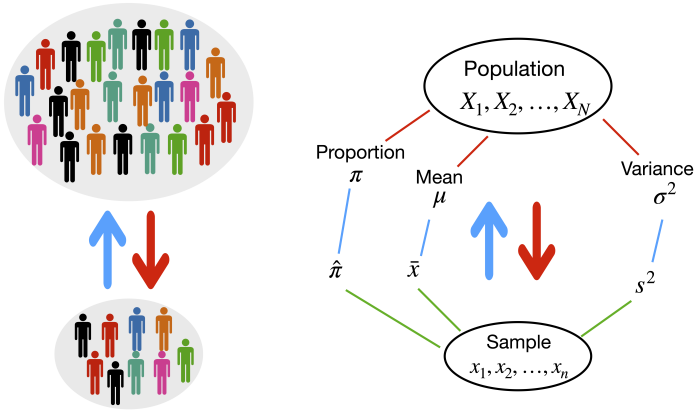
1.2 Branches of statistics

The study of statistics has two major branches - descriptive statistics and inferential statistics:

- **Descriptive statistics** is the branch of statistics that involves the organisation, summarisation, and display of data.
- **Inferential statistics** is the branch of statistics that involves using a sample to draw conclusions about a population, e.g. estimation and hypothesis testing.

1.3 What's the big idea?

There are many qualities of a population we might be interested in. These qualities are referred to as **parameters**. We can never know the value of these parameters in general. What we do instead is find corresponding values from a sample, and use these as estimates for the parameter values. These estimates we find from the sample are referred to as **statistics**.



1.4 Notation

Below is a table containing commonly-used notation for some of the parameters and statistics we will deal with most often.

	Population	Sample
Size	N	n
	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Proportion	π	$\hat{\pi}$
Correlation	ρ	r

2 Data Types in Statistics

2.1 Data collection methods (Traditional data)

There are several approaches we can take when collecting data:

- **Take a census:** a census is a count or measure of an entire population. Taking a census provides complete information, but it is often costly and difficult to perform.
- **Use sampling:** a sample is a count or measure of a part of a population. Statistics calculated from a sample are used to estimate population parameters.
- **Use a simulation:** collecting data often involves the use of computers. Simulations allow studying situations that are impractical or even dangerous to create in real life and often save time and money.
- **Perform an experiment:** e.g. to test the effect f imposing a new marketing strategy, one could perform an experiment by using the new marketing strategy in a certain region.

2.2 Data collection methods (Big data)

The characteristics of big data (the 4Vs):

- Volume: how much data is there?
- Variety: different types of data?
- Velocity: at what speed?
- Veracity: how accurate?

2.3 Types of data

Data sets can consist of two types of data:

- **Qualitative (categorical) data** consist of attributes, labels, or nonnumerical entries. e.g. name of cities, gender etc.
- **Quantitative data** consist of numerical measurements or counts. e.g. heights, weights, age. Quantitative data can be distinguished as:
  - **Discrete data** result when the number of possible values is either a finite number or a "countable" number. e.g. the number of phone calls you received in any given day.
  - **Continuous data** result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps. e.g. height, weight, sales and market shares.

2.4 Types of data (Econometrics)

- **Cross-sectional data:** Data on different entities (e.g. workers, consumers, firms, governmental units) for a single time period. For example, data on test scores in different school districts.
- **Time series data:** Data for a single entity (e.g. person, firm, country) collected at multiple time periods. For example, the rate of inflation or of unemployment for a country over the last 10 years.
- **Panel data:** Data for multiple entities in which each entity is observed at two or more time periods. For example, the daily prices of a number of stocks over two years.

2.5 Levels of measurement

- **Nominal:** Categories only, data cannot be arranged in an ordering scheme. (e.g. Marital status: single, married etc.)
- **Ordinal:** Categories are ordered, but differences cannot be determined or they are meaningless (e.g. a rating poor, average, good)
- **Interval:** differences between values are meaningful, but there is no natural starting point, ratios are meaningless (e.g. we cannot say that the temperature 80°F is twice as hot as 40°F)
- **Ratio:** Like interval level, but there is a natural zero starting point and rations are meaningful (e.g. £20 is twice as much as £10)

3 Descriptive Statistics

3.1 Measures of Central Tendency

Measures of central tendency provide numerical information about a 'typical' observation in the data.

- The **mean** (also called the average) of a data set is the sum of the data values divided by the number of observations.

$$\text{Sample mean:} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The **median** is the middle observation when the data set is sorted in ascending order. If the data set has an even number of observations, the median is the mean of the two middle observations.
- The **mode** is the data value that occurs with the greatest frequency. If no entry is repeated, the data set has no mode. If two (more than two) values occur with the same greatest frequency, each value is a mode and the data set is called bimodal (multimodal).

3.2 Measure of Variation (Dispersion)

The variation (dispersion) of a set of observations refers to the variability that they exhibit - how far from the average value do we expect individual values to be, in general?

- **Range** = maximum data value - minimum data value
- The **variance** measures the variability or spread of the observations from the mean.

$$\text{Sample variance:} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

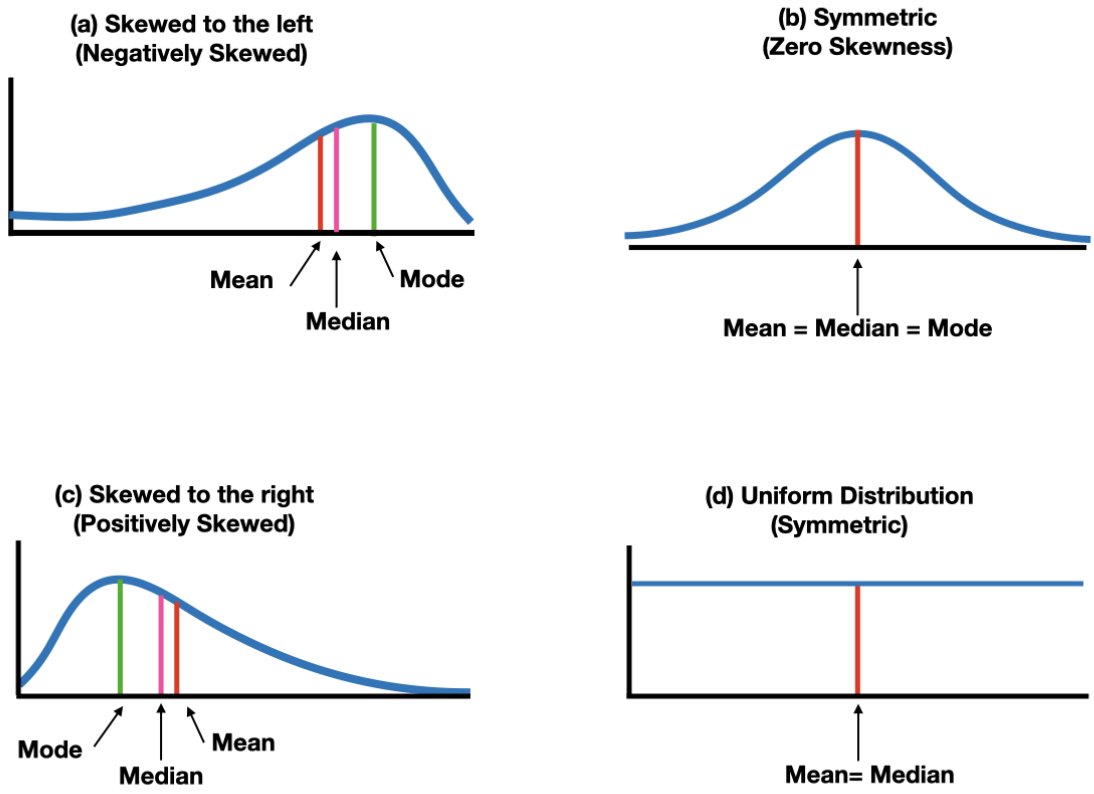
- Shortcut formula for sample variance is given by

$$\text{Sample variance:} \quad s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right\}$$

- The **standard deviation** (s) of a data set is the square root of the sample variance.

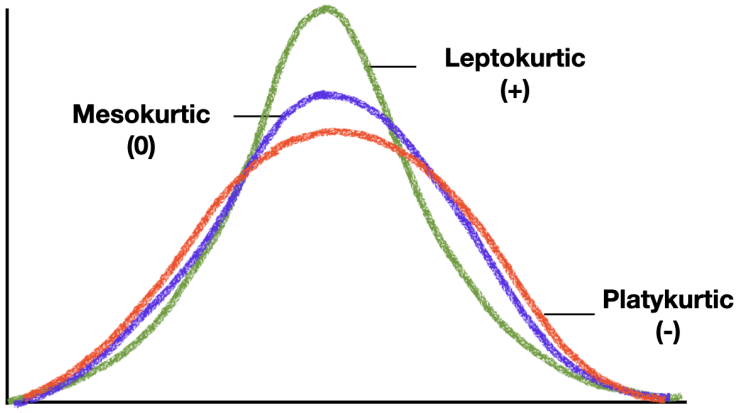
3.3 Shape of a distribution: Skewness

Skewness is a measure of the asymmetry of the distribution.



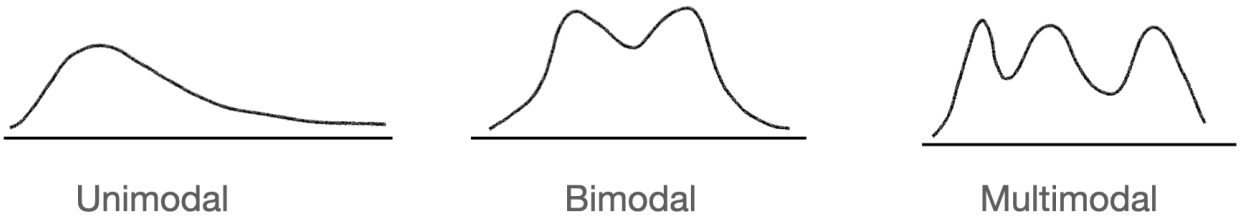
3.4 Shape of a distribution: Kurtosis

Kurtosis measures the degree of peakedness or flatness of the distribution.



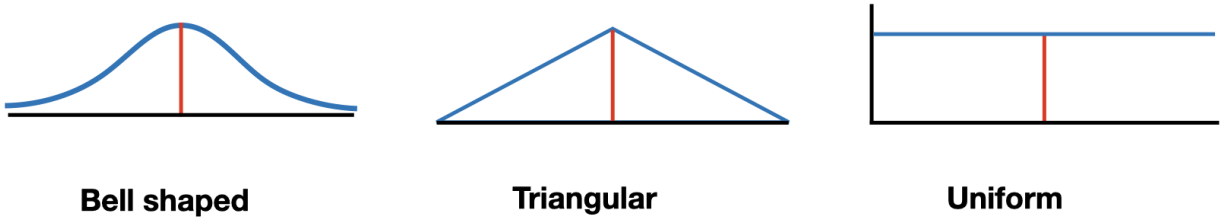
3.5 Modality

The number of highest points in a distribution gives us the modality. Note that in a situation in which a distribution has two or more "humps" which aren't equally high, we still describe the shape of the graph as "bimodal" or "multimodal", even though only the (equal) highest point of the curve represents the actual mode(s) of the data.



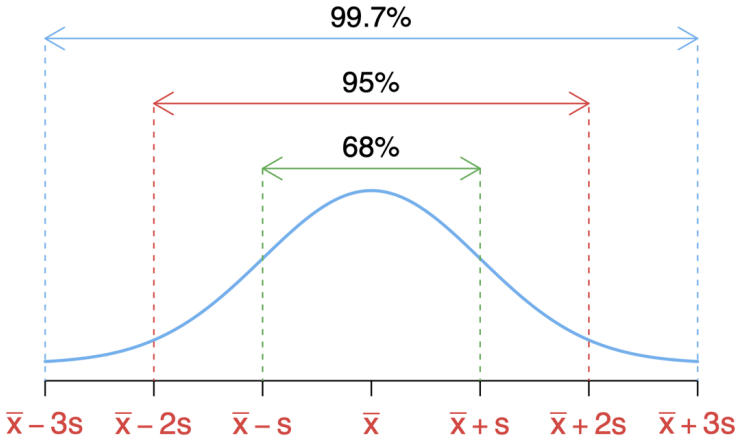
3.6 Symmetry

Below are three common forms of symmetrical distribution. Note that if a symmetrical distribution is also unimodal, the median, mode and mean will all be equal.



3.7 Empirical Rule

The empirical rule states (for a normally distributed data) that 68% of the data falls within one standard deviation; 95% of the data falls within two standard deviations; 99.7% of the data falls within three standard deviations from the mean.



3.8 Measure of Position: z-score

The **z-score** of an observation tells us the number of standard deviations that the observation is from the mean, that is, how far the observation is from the mean in units of standard deviation.

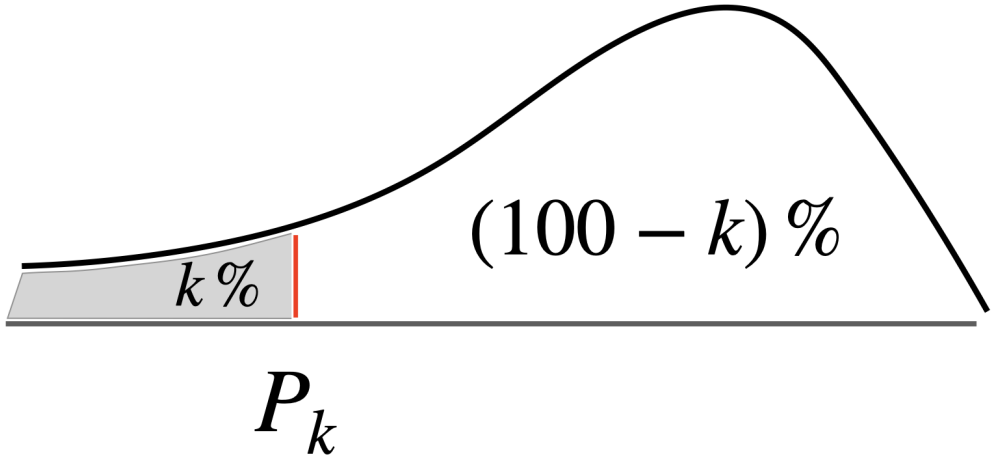
$$z = \frac{x - \bar{x}}{s}$$

As the z-score has no unit, it can be used to compare values from different data sets or to compare values within the same data set. The mean of z-scores is 0 and the standard deviation is 1.

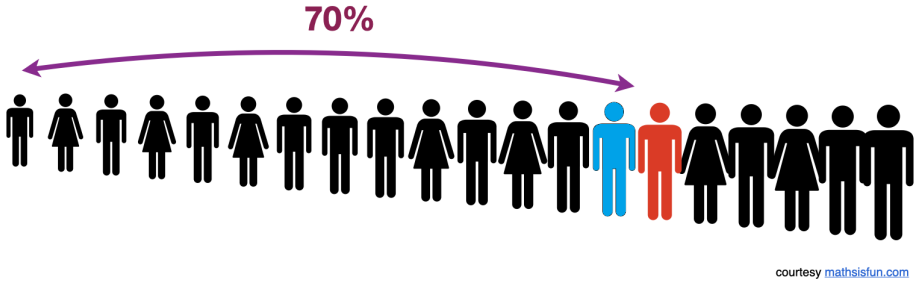
Note that $s > 0$ so if z is negative, the corresponding x-value is below the mean. If z is positive, the corresponding x-value is above the mean. And if $z = 0$, the corresponding x-value is equal to the mean.

3.9 Percentiles and Quartiles

- Given a set of observations, the kth percentile P_k is the value of X such that $k\%$ or less of the observations are less than P_k and $(100 - k)\%$ or less of the observations are greater than P_k:

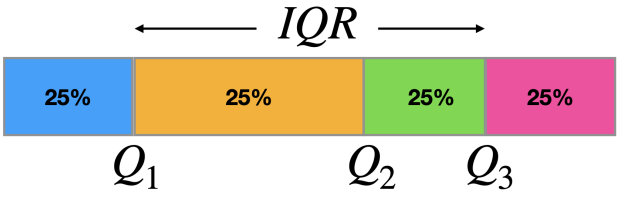


- The 25th percentile, Q_1, is often referred to as the first quartile.
- The 50th percentile (the median), Q_2, is referred to as the second or middle quartile.
- The 75th percentile, Q_3, is referred to as the third quartile
- Here is a quick example.



70% of people are shorter than the red figure. 30% of people are taller than the blue figure. The 70% percentile for height therefore lies between the blue and red figure.

- The four quartiles divide a data set into quarters (four equal parts). As the diagram below shows, the four equal parts do not necessarily have equal **lengths**, it is the number of data points which are the same within each part.

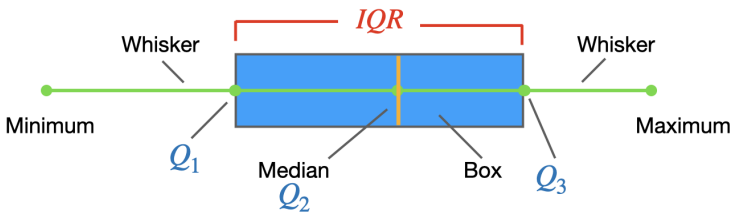


- The interquartile range (IQR) of a data set is the difference between the first and third quartiles ($IQR = Q_3 - Q_1$)
- The IQR is a measure of variation that gives you an idea of how much the middle 50% of the data varies.

3.10 Five-number summary & Boxplots

To draw a **boxplot** (also called a box-and-whisker plot), we need the following values (called the five-number summary):

- The minimum entry
- The first quartile Q_1
- The median (second quartile) Q_2
- The third quartile Q_3
- The maximum entry

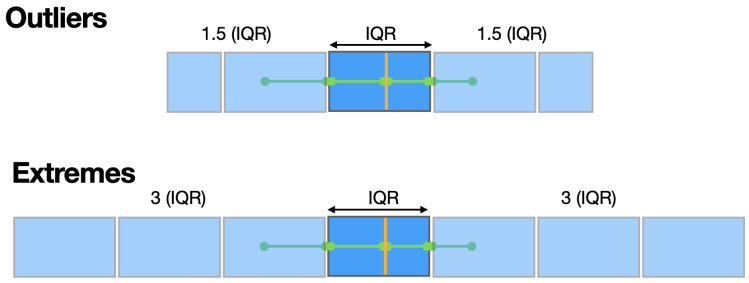


The box represents the interquartile range (IQR), which contains the middle 50% of values.

3.11 Outliers & Extremes values

Some data sets contain outliers or extremes values, observations that fall well outside the overall pattern of the data. Boxplots can help us to identify such values if some rules-of-thumb are used, e.g.:

- Outlier: Cases with values between 1.5 and 3 box lengths (the box length is the interquartile range) from the upper or lower edge of the box.
- Extremes: Cases with values more than 3 box lengths from the upper or lower edge of the box.



3.12 Descriptive statistics for qualitative variables

- Frequency distributions are tabular or graphical presentations of data that show each category for a variable and the frequency of the category's occurrence in the data set. Percentages for each category are often reported instead of, or in addition to, the frequencies.
- The mode can be used in this case as a measure of central tendency.
- Bar charts and pie charts are often used to display the results of categorical or qualitative variables. Pie charts can become cluttered and difficult to read if variables have many categories. Pie charts should always include information on the total number of data points.
- Bar charts can also be used to group together numerical values. Doing so loses the original values, however. An alternative is a stem-and-leaf plot, which makes the bars out of data values themselves.
- A dot plot can be used to quickly compare numerical values between multiple categories (clusted bar charts can also do this).

3.13 Example: Accounting final exam grades

The accounting final exam grades of 10 students are: 88, 51, 63, 85, 79, 65, 79, 70, 73, and 77. Their study programs, respectively, are: MA, MA, MBA, MBA, MBA, MBA, MBA, MSc, MSc, and MSc.

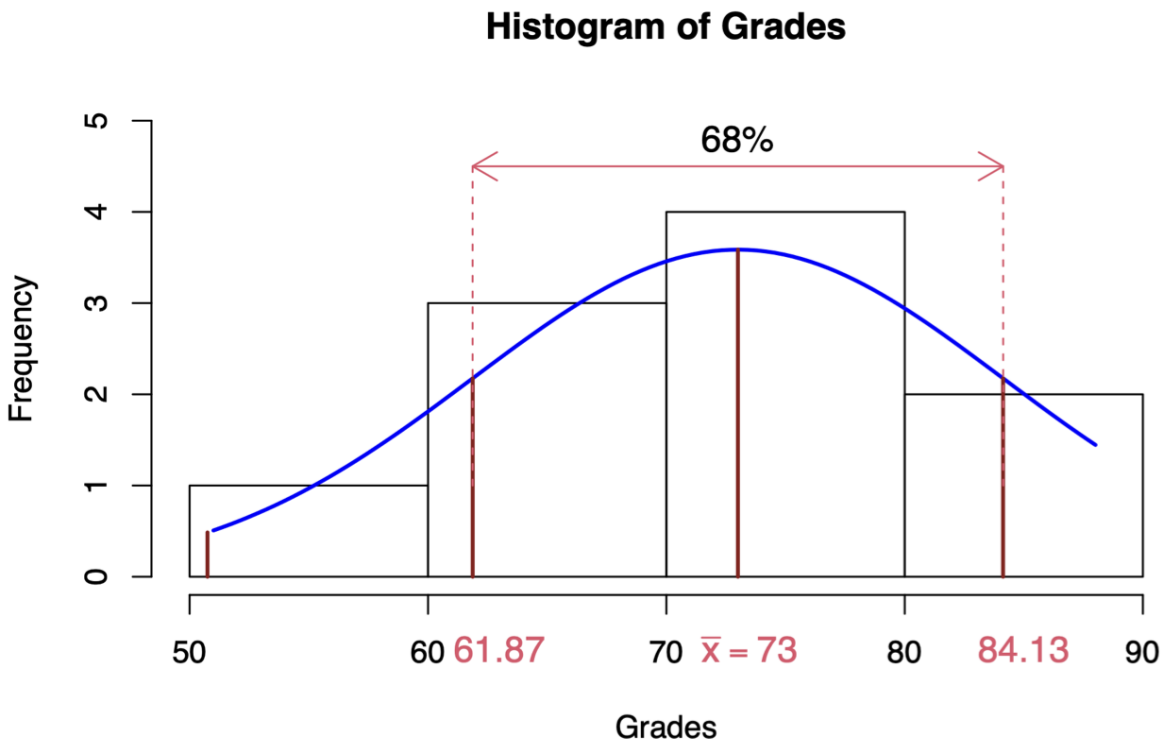
- The sample mean grade is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(88 + 51 + \ldots + 77) = 73$$

- Next we arrange the data from the lowest to the largest grade: 51, 63, 65, 70, **73**, **77**, 79, 79, 85, 88. The median grade is 75, which located midway between the 5th and 6th ordered data points $(73 + 77)/2 = 75$.
- The mode is 79 since it appears twice and all other grades appeared only once.
- The range is $88 - 51 = 37$.
- The sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}((88 - 73)^2 + \ldots + (77 - 73)^2) = 123.78$$

- The sample standard deviation: $s = \sqrt{123.78} = 11.13$
- The coefficient of variation: $CV = s/\bar{x} = 11.13/73 = 0.1525$
- Empirical rule: the empirical rule states (for normally distributed data) that 68% of the data falls within one standard deviation from the mean. In our example, this means that 68% of the grades fall between 61.87 and 84.13 (73 ± 11.12555)




```
# R codes for "Accounting final exam grades" example
# Data example
grades<-c(88,51,63,85,79,65,79,70,73,77)
program<-factor(c("MA","MA","MBA","MBA","MBA","MBA","MBA","MSc","MSc","MSc"))

# no of observations
length(grades)
```

```
## [1] 10
```

```
# Mean, Median, Variance, standard deviation, range, quantile
mean(grades)
```

```
## [1] 73
```

```
median(grades)
```

```
## [1] 75
```

```
var(grades)
```

```
## [1] 123.7778
```

```
sd(grades)
```

```
## [1] 11.12555
```

```
range(grades)
```

```
## [1] 51 88
```

```
quantile(grades,probs=c(0,0.25,0.5,0.75,1))
```

```
##    0%   25%   50%   75%  100%
## 51.00 66.25 75.00 79.00 88.00
```

```
# Summary
summary(grades)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.00   66.25   75.00   73.00   79.00   88.00
```

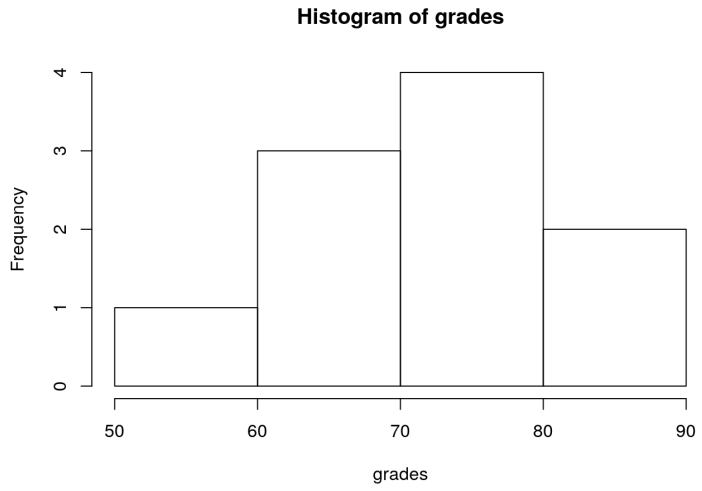
```
# Calculate z-score
(grades-mean(grades))/sd(grades)
```

```
##  [1]  1.3482484 -1.9774310 -0.8988323  1.0785987  0.5392994 -0.7190658
##  [7]  0.5392994 -0.2696497  0.0000000  0.3595329
```

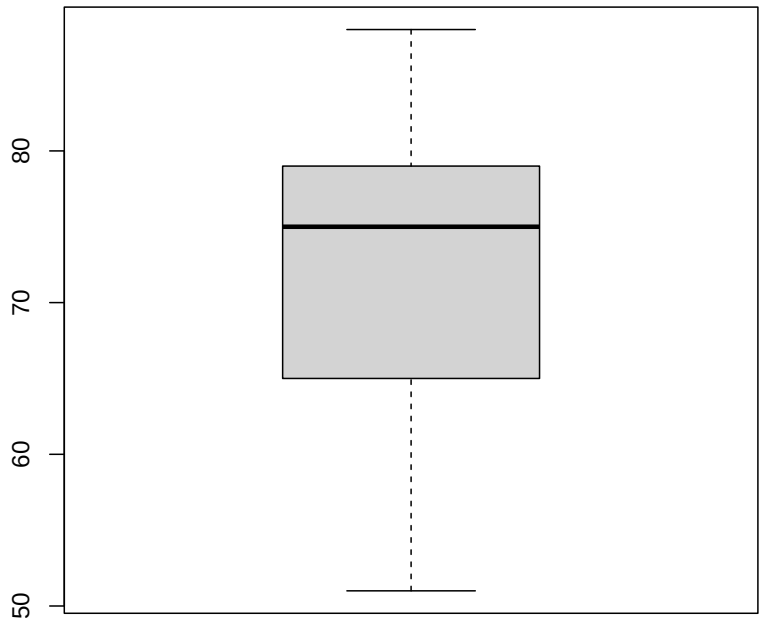
```
scale(grades)
```

```
##             [,1]
##  [1,]  1.3482484
##  [2,] -1.9774310
##  [3,] -0.8988323
##  [4,]  1.0785987
##  [5,]  0.5392994
##  [6,] -0.7190658
##  [7,]  0.5392994
##  [8,] -0.2696497
##  [9,]  0.0000000
## [10,]  0.3595329
## attr(,"scaled:center")
## [1] 73
## attr(,"scaled:scale")
## [1] 11.12555
```

```
#  Histograms present frequencies for values grouped into interval.
hist(grades,xlab="grades", main="Histogram of grades")
```




```
# Boxplot
boxplot(grades,xlab="grades")
```



grades

In a stem-and-leaf plot: each score on a variable is divided into two parts, the stem gives the leading digits and the leaf shows the trailing digits.

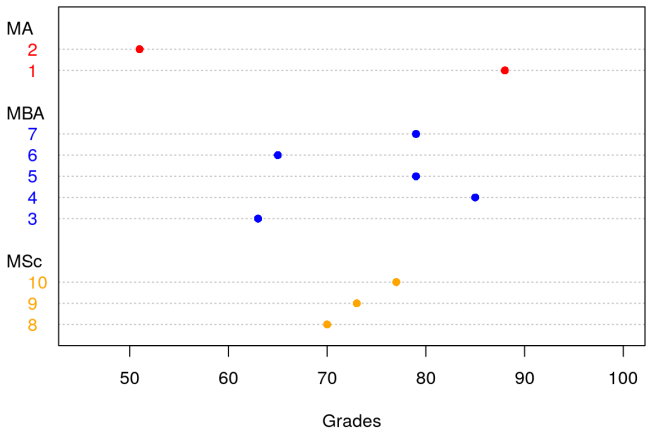
The accounting final exam grades (arranged from the lowest to the largest grade) are: 51, 63, 65, 70, 73, 77, 79, 79, 85, 88.

```
# Stem-and-leaf plot.
stem(grades)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 1
##   6 | 35
##   7 | 03799
##   8 | 58
```

A dot plot is a simple graph to show the relative positions of the data points.

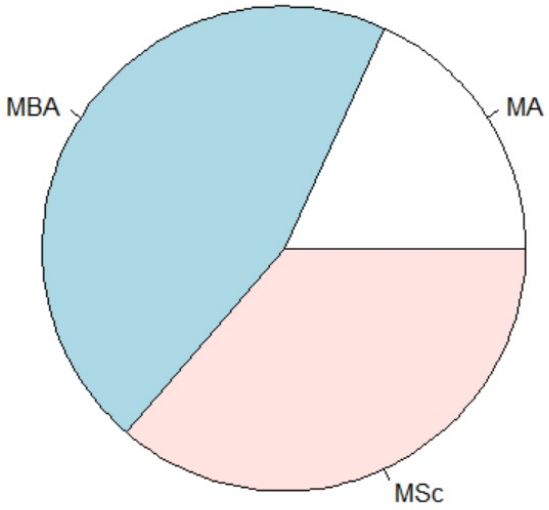
```
col2<-as.character(factor(program,labels=c("red","blue","orange")))
dotchart(grades, labels=factor(1:10), groups=program, pch=16, col=col2, xlab="Grades",xlim=c(45,100))
```



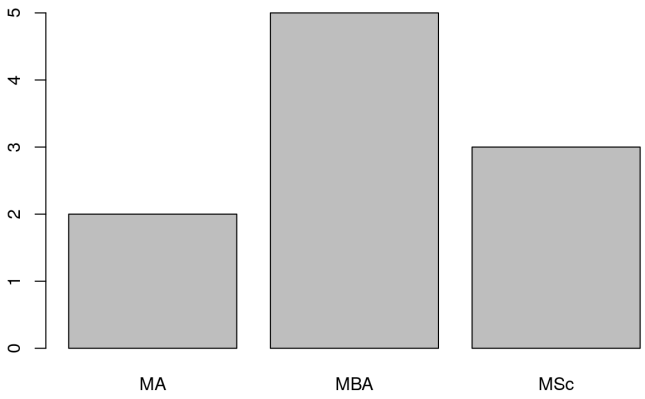

```
# Frequency table
table(program)
```

```
## program
##  MA MBA MSc
##   2   5   3
```

```
# Pie and Bar charts
pie(table(program))
```




```
barplot(table(program))
```



4 Probability

4.1 The Basic Idea

Probability is a measurement, a way we express how likely something is to happen. What makes probability so interesting as a measurement is a) it has no units, and b) there are lots of different ways to to conceive of and calculate what a probability could and "should" be.

Despite there being many different interpretations and philosophies of probability (a topic we will mostly be steering clear of in this module, but which will become very important next term if you take any module relating to the idea of **Bayesian statistics**), there are a few basic ideas everyone agrees on, which are:

1. An impossible event has probability 0, and no probability can ever be lower than 0.
2. A certain event has probability 1, and no probability can ever be higher than 1.
3. An event which is neither impossible nor certain has a probability between 0 and 1.
4. When comparing multiple events, an event with a higher probability is more likely to happen than an event with a lower probability. Events with the same probability are equally likely to happen.

One common way to think about probabilities is as **functions** - we input an event, and our output is a number telling us how likely that event is to happen. We'll talk more about this idea of probabilities as functions later in these notes.

Even this very quick, simple and broad summary throws up more questions, though, starting with one we'll answer in the next subsection: what **is** an **event**?

4.2 Outcomes and Events

Before I can define an event, I need to define two other aspects of probability theory: **outcomes** and the **outcome space**. For any situation in which we are uncertain about what is going to happen:

- An **outcome** is something that could happen, and which cannot happen in more than one **relevant** way.
- The **outcome space** is the collection of all the outcomes for the situation. Whatever the result of the situation, that result must correspond to one **and only one** outcome in the outcome space.

Note my use of the word "relevant" here - this is very important. What we consider relevant in any given situation is up to us, so it's possible two different people might define their outcomes for the same situtation

in different ways. As long as each of them are fully clear on their choice of outcomes and outcome space, this is entirely fine. As I say in the videos, though, if in doubt, define "relevant" as widely as you can. If information collected turns out to not be useful, you can disregard it. If information **not** collected turns out to be relevant, then you could be in real trouble!

Example I am about to roll a six-sided dice, and I want to express the probability of each one of the six numbers on the dice being the one that lands face-up. One way to define the outcomes here would be each of the six possible numbers, 1 to 6. If those are my choice of outcomes, my outcome space would be each of those numbers, $\{1, 2, 3, 4, 5, 6\}$.

Alternatively, if I wished to, I coul define the outcomes as "an even number" and "an odd number", in which case I might write my outcome space as, say, $\{O, E\}$. We normally wouldn't do this, because we can break each of those outcomes up into three separate results, but if we considered the specific number we roll to be **irrelevant** for our purposes, then nothing here about $\{O, E\}$ as an outcome space is in any sense incorrect or invalid.

Choosing your outcomes to be, say, "at least three" and "no more than three" **would** be incorrect, however, because if you roll a three, that would mean more than one outcome has occurred at the same time, which is not permissable. Similarly, you couldn't choose "less than three" and "more than three" as your only outcomes, as rolling a three would mean no outcome had occurred.

Example I am going to play Bizzfin, a game in which each turn I roll a fair six-sided dice and take a card from a standard western deck of playing card, containing 52 cards. I gain or lose points in the game depending on the combination of dice and card. What is the outcome space for the game?

In this case, I have two things to keep track of, the dice score, of which there are 6 possible values, and the card drawn, of which there are 52 possible values. The outcome space is every possibe combination of dice score and card, of which there are $6 \times 52 = 312$. I won't list them all here, but the outcomes could be expressed as paired values such as $(1, 7C)$, representing a score of 1 on the dice and drawing the Seven of Clubs. The outcome space could hen be represented as, say, $\{(1, 1C), (2, 1C), \ldots, (6, 1C), (1, 2C), \ldots, (6, KC), (1, 1D), \ldots, (6, AS)\}$. There are other ways we could represent all this, what matters is making our intent clear, and being consistent in whatever approach we're using.

We can now define an **event**. An event is either an outcome, or a combination of outcomes. For our previous example, if $\{1, 2, 3, 4, 5, 6\}$ is our outcome space, then any element of combination of elements from that set is an event. "1" is an event, "4 or more" is an event, "not prime" is an event, and so on.

Note that this means all outcomes are events, indeed we call them **simple** events. Not all events our outcomes, though; if an event comprises more than one outcome, it is called a **compound event**.

One last event we need to consider is the **empty event**. This is the event that no outcome occurs. This is impossible, as we must always have one outcome occur. As a result, the empty event has probability 0. It might seem odd to insist on this idea of an impossible event, which doesn't contain any outcomes and therefore has probability 0. It's very useful in terms of the set theory that mathematicians use to make probability work, though, which is why it's important to consider it.

4.3 Probabilities As Proportions

So how do we calculate a probability? Again, there are a number of different ways to answer that question, and again, we're going to essentially ignore that fact during this module.

For our purposes, it suffices to think of probabilities as being **proportions**. The probability of an outcome is defined as the proportion of times an outcome **does** happen, out of all the times that outcome **could** have happened.

$$P(\text{Outcome}) = \frac{\text{Number of times outcome happens}}{\text{Number of times outcome could have happened}}$$

Example I am about to roll a four-sided dice, which I know to be fair (each number is equally likely to be rolled). I define my outcome space as $\{1, 2, 3, 4\}$. What is the probability I roll a 4?

Under the definition above, the probability of rolling a 4 is the number of times a 4 is rolled on a fair four-sided dice, divided by the number of times the dice is rolled, because a four showing is something that could happen on each roll.

Because the dice is fair, I will roll a 4 one fourth of the time I roll the dice.

It's important to note in the above example that I used a theoretical property of the dice - it is "fair". If I **actually** roll the dice multiple times, there is no guarantee I will get a 4 precisely one fourth of the time - indeed this is impossible if I don't throw the dice a number of times which is divisible by four! We will come back to this in the next subsection.

We find the probability of an **event** by adding together the probabilities of each outcome making the event up (remember an event by definition is made up of one or more outcomes). Alternatively, we can just tweak our previous definition - the probability of an event is defined as the proportion of times an event **does** happen, out of all the times that event **could** have happened. Thanks to the laws of maths, these two definitions are actually equivalent.

$$P(\text{Event}) = \frac{\text{Number of times event happens}}{\text{Number of times event could have happened}}$$

Example I am about to roll a four-sided dice, which I know to be fair (each number is equally likely to be rolled). I define my outcome space as $\{1, 2, 3, 4\}$. What is the probability I roll less than 4?

We can find this in two ways. Firstly, I could add up the probabilities for each of the three outcomes (1, 2 and 3) which make up the event "less than 4". Due to the dice being fair, each of these outcomes has the same probability as the outcome of rolling a 4, so they sum to three quarters. Alternatively, if I roll a fair four-sided dice, I will roll a 1, 2, or 3 three quarters of the time.

The above example shows us two important general ideas. Firstly, if we wanted to calculate the probability of rolling a 1, 2, 3 or 4, then that probability would equal $0.25 + 0.25 + 0.25 + 0.25 = 1$. This makes sense, though, because the event I'm considering now is one which contains every outcome. Therefore that event **must** happen.

Secondly, in a situation in which you have, say, n outcomes, each of which is equally likely, the probability of each outcome must be $1/n$, since they all have to have the same probability (otherwise they're not equally likely!) and adding all n of them together must result in a value of 1.

4.4 Relative Frequencies

Finding probabilities is actually very simple, then, so long as all outcomes are equally likely. Unfortunately, that's very often not the case. Often we don't know how much more or less likely one outcome is than another. Perhaps a coin is clipped, or a dice is loaded, or we're in any one of the billions of other circumstances where we can't assume all possible outcomes are just as likely as each other.

In such circumstances, we tend to resort to experimentation. In what follows we assume it's possible to run multiple experiments, each in identical or nearly identical circumstances.

The **relative frequency** of an event is our estimate of that event's probability. It is equal to the number of times we ran an experiment in which the event happened, divided by the total number of experiments we ran.

$$P(\text{Event}) \approx \text{Relative frequency of event} = \frac{\text{Number of experiments in which event happens}}{\text{Total number of experiments run}}$$

Note that if the event never happens in our experiments, the relative frequency is 0 (suggesting the event could be impossible). If the event always happens in or experiments, the relative frequency is 1 (suggesting the event might always happen). The more times we run the experiment, the more accurate we expect our relative frequencies to be (we might have to do this a **lot** if we're looking for an estimate of the probability of a rare event, and we don't want that estimate to just be 0). This is hopefully not surprising, since we can think of a probability as being the proportion of times an event occurred during an **infinite** sequence of events.

Example 4.5 I use `R` to simulate tosses of a fair coin. I ask `R` to do this 10 times, getting 2 results of Heads. I then ask `R` to do this 10,000 times, getting 5,056 results of Heads. Finally, I ask `R` to do this 10,000,000 times, getting 4,997,386 results of Heads. The relative frequencies I get are shown in the table below.

Outcome	10 tosses	10,000 tosses	10,000,000 tosses
Heads	0.2	0.5056	0.4997386
Tails	0.8	0.4944	0.5002614

We can see here how the relative frequencies are approaching the true probability of Heads, which is 0.5.

4.5 Independence

The concept of **independence** is absolutely critical to both probability and statistics. It is also very commonly misunderstood.

There are a number of different ways of thinking about independence. Some get a little technical, and we'll return to those later in the module. For now, though, I'll give you a definition in plain English. Two **events** are **independent** if learning whether one event has happened gives **no additional information** about whether the other event will happen.

For instance, say I roll two fair six-sided dice. The probability of rolling a one on such a dice is 1/6. If I tell you I rolled a one on the first dice, this would not cause you to rethink what the probability of rolling a one on the second dice is.

This kind of example is very commonly used to explain independence. Unfortunately, it can lead to people mistakenly believing two events are independent if the process which produce those events do not in any way interact. **This isn't actually the case!** Let's consider another example.

Example 4.6 I roll a fair six-sided dice with one hand, and roll a fair twenty-sided dice with the other hand. Let event A be "I roll a one on the dice in my left hand", and let event B be "I roll a one on the dice in my right hand". I roll the dice in such a way that they never touch, and so the number each ends up showing can't have any effect on what the number the other dice shows is. Are A and B independent events?

We know the probability of rolling a one on the six-sided dice is 1/6, and the probability of rolling a one on the twenty-sided dice is 1/20. Let's say the probability the six-sided dice is in my left hand is 1/2. We shall soon see how to find the probability of event A happening; it's $(1/2) \times (1/6) + (1/2) \times (1/20) = 13/120$, with event B having the same probability.

Now, suppose I roll the dice in my left hand, and I get a 20. This immediately tells us two things. First, event A didn't happen. Second, and much more importantly, the probability of event B **can't** be 13/20 any more. That's because we now know that the dice in my right hand must have been the one with six sides. Therefore, the probability of event B is now 1/6.

Note that if I told you ahead of time which dice was in each hand, events A and B **would** be independent, because this trick of learning about which dice is which from one roll no longer works - if we already knew I had the twenty-sided dice in my left hand, learning that I got a 20 from that roll would have no effect on our beliefs about the other roll.

This highlights the extremely important idea that what we can say about the probabilities of events depends on how we are defining our events, not just whatever's going on in whatever process we're interested in.

4.5.1 The Multiplicative Rule

The property of independence has an absolutely colossal number of uses within statistics. Perhaps the most fundamental is what I shall call **the multiplicative rule**.

The multiplicative rule: If events A and B are independent, then

$$P(A \text{ and } B) = P(A) \times P(B).$$

Example 4.7 In a game of Bizzfin (see **Example 4.2**), find the probability that I roll a 5 or 6 on the dice, while also drawing a red card.

In this example, our two events - roll a 5 or a 6 on the dice (call this event A), and draw a red card (call this event B) - are independent, because knowing what I rolled on the dice gives me no information about the card I drew, and vice versa.

Hence, we can use the multiplicative rule. The probability of rolling a 5 or a 6 on fair six-sided dice is $(1/6) + (1/6) = 1/3$. Half the cards in a western deck are red (the other half are black), so the probability of drawing a red card is $1/2$.

$$P(A \text{ and } B) = P(A) \times P(B) = \frac{1}{3} \times \frac{1}{2} = \frac{1}{6}.$$

So why does the multiplicative rule work? In Section 4.3, we talked about how in this module, we think of a probability of an event as being the proportion of times the event does happen, out of all the times it could. If A and B are independent events, then B is no more or less likely to happen if A happens than if A doesn't happen. Therefore, A happens with a proportion $P(A)$, and B happens with a proportion $P(B)$ of the times A has happened (and proportion P(B) of the times A hasn't happened, but we're not considering that right now). This means they both happen for a proportion $P(B)$ of the proportion $P(A)$, which is an overall proportion $P(B) \times P(A)$.

In other words, all **Example 4.7** is doing is using the fact that half the time I'll draw a red card, and one third **of those times**, I will roll a 5 or a 6.

Note that we cannot perform a similar trick when we **don't** have independence, because the proportion of times B happens will be **different** depending on whether A happens or doesn't happen; those two cases have to be considered separately. We will see soon how we can go about doing this.

4.6 Mutual Exclusivity

If independence is at one end of a scale regarding how much influence learning about one event has on our thinking about the other event, then **mutual exclusivity** exists at the opposite end of that scale. Two events are mutually exclusive if one of them happening means the other **cannot possibly happen**. Despite this being as far from independence as you can get - learning about one event can give you total information about the other - the two are very commonly mixed up. It's even more common to mix up when we can use the multiplicative rule which we've seen already, and when we can use the **additive rule** (or at least a special case of it), which we can use when we have mutually exclusive events.

4.6.1 Additive Rule (Mutually Exclusive Results)

The additive rule for mutually exclusive events: If events A and B are mutually exclusive, then

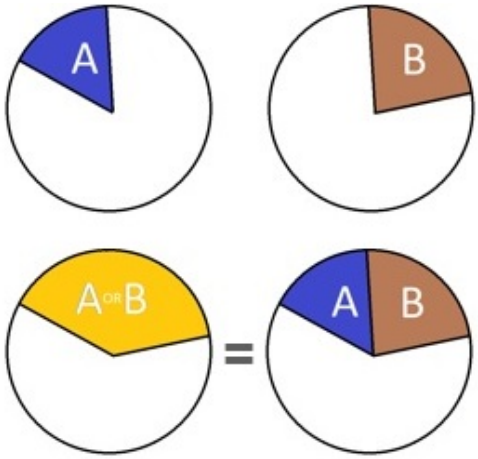
$$P(A \text{ or } B) = P(A) + P(B).$$

Example 4.8 My friend's two sons Alfred and Martin love to compete in triathlons. He decides, based on their previous times, that in the next triathalon they run, Alfred has a 0.07 probability of winning the triathlon, and Martin has a 0.09 probability of winning the triathlon. What is the probability one of my friend's sons wins the triathlon?

Let us define A as being the event Alfred wins the race, and define M as the event Martin wins the race. These events are **mutually exclusive** - they cannot both win the race. Therefore we can use the special case of the additive rule.

$$P(A \text{ or } M) = P(A) + P(M) = 0.07 + 0.09 = 0.16.$$

So why does the additive rule work? There are two ways to think about this. One is to think about the proportions involved. There is a certain proportion of times that A happens, and a certain proportion of times that B happens. There is never an occasion when both happens. As a result, the proportion of time where A or B happens must be the sum of the proportions when each happens, as shown in the diagrams below.



Alternatively, we can think back to how we define probabilities. If we run an infinite series of experiments, the number of times A or B happens must equal the number of times A happens plus the number of times B happens, since they can never both happen at once. The additive rule therefore just tells us it doesn't matter if we add the number of times A happens to the number of times B happens **before** dividing by the total number of times, or after.

Note that we can combine the multiplicative and additive rule where appropriate. That's part of how, in **Example 4.6**, I calculated the probability of rolling a one on a dice (call this R1) that had a probability 0.5 of being six-sided, and 0.5 of being twenty-sided. I used the fact that the event "this dice has six sides" (call this S) is mutually exclusive to the event "this dice has twenty sides" (call this T), and so

$$P(R1) = P(R1 \text{ and } S) + P(R1 \text{ and } T)$$

We still need to do a bit more work to see how this calculation worked overall; we'll come back to this in Subsection 4.9.

4.6.2 Additive Rule (General)

We can't use the additive rule in the same way when A and B are not mutually exclusive, because in that situation, there is certain proportion of times when A and B both happen.

This isn't difficult to deal with, though. Rather than add the proportions for A to the proportion for B, we could instead:

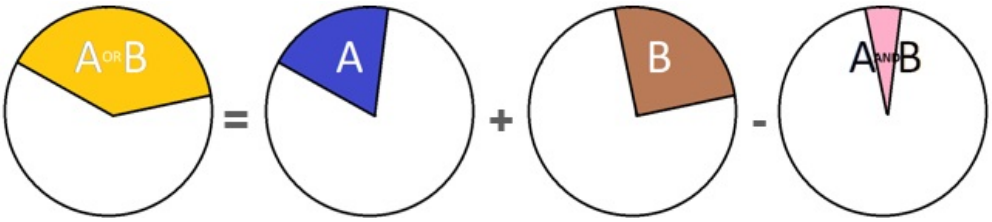
1. add the proportion for A to the proportion of B **but not** A;
2. add the proportion for B to the proportion of A **but not** B;
3. add the proportion for B **but not** A to the proportion of A **but not** B, and then add the proportion of A and B.
4. add the proportion for A to the proportion of B, but then subtract the proportion of A and B (because that proportion has been counted twice).

All four of these approaches give us the same answer, but its the fourth one we use to define the general additive rule.

The additive rule in general is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Note that this still works for mutually exclusive events, since in such a case, $P(A \text{ and } B) = 0$.



4.7 Complements

The concept of a **complement** is a little different to what we've looked at so far. We've focussed on when two events are independent, or when they're mutually exclusive (or when they're neither), but there's nothing stopping us thinking about any number of independent events, or any number of mutually exclusive events.

In contrast, each and every event has one and only one complement. That's because for an event A, the **complement** of A is the event that A **doesn't happen**. We denote this event as A^C.

Rules of complements

1. The complement of a complement is itself: $A^{C^C} = A$. In words, if A doesn't **not** happen, A happens.
2. Every outcome in the outcome belongs to one and only one of events A and A^C.
3. A and A^C are mutually exclusive events, because A must either happen or not happen, it cannot do both at the same time.
4. The event that A happens *or* A^C happens is certain.
5. The complement of the event including all outcomes is the empty event, and vice versa (this is why we need the concept of the empty event).

As a consequence of the 3rd and 4th rules, we also have

$$P(A \text{ or } A^C) = P(A) + P(A^C) = 1.$$

It is for this reason that we have the result, which you may have seen before, that

$$P(A \text{ happens}) = 1 - P(A \text{ doesn't happen}).$$

While each event A has only one complement, we can nevertheless extend what we've seen above by talking about **mutually exclusive and exhaustive events**. This is a set of m events $A_1, A_2, \ldots, A_m$, such that every outcome in the outcome space belongs to one and only one event. This in turn means A_i and A_j are mutually exclusive whenever $i \neq j$, and that one and only one A_i must occur.

When we have such events, we must have

$$\sum_{i=1}^{n} P(A_i) = 1 \quad (*).$$

One way to justify this result is that every event A_i is mutually exclusive to all other events, and therefore the sum of their probabilities must equal the probability that (A_1 or A_2 or $\ldots$ or A_m) happens. Since every outcome belongs to one of the events A_i, we must have

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_m) = 1$$

and hence

$$\sum_{i=1}^{n} P(A_i) = P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_m) = 1$$

The rule that $P(A) + P(A^C) = 1$ is actually a special case of rule $(*)$, because A and A^C are mutually exclusive and exhaustive events - they can't both happen, and one of them **must** happen.

Example 4.9 In the game of bizzfin (see **Example 4.2**), what is the probability that I either don't roll a 1, or that I don't draw a Heart card?

We can count all the outcomes which lie inside this event, but it might be faster to count the outcomes which lie inside its complement. There are 1×52 I can roll a one and draw a Heart card. The probability that I don't roll a 1 or don't draw a Heart card is therefore $1 - 52/312 = 260/312$.

Note that the probablity of not rolling a 1 or not drawing a Heart card is not the same as not rolling a 1 **and** not drawing a Heart card. The latter event here has a probability of 195/312 (try checking this for yourself).

4.8 An Introduction To Expectation

One of the most fundamental concepts in both probability and statistics is that of **expectation**. Indeed, it's so important we'll meet it more than once, in what might seem to be quite different contexts.

For now, we'll think about the relationship between the probability of an event, and how many times we expect that event to happen. If we have r opportunities for an event A to happen, and the probability of the event happening each time is $P(A)$, then the number of times we expect the event to happen is $r \times P(A)$.

Example 4.10 Suppose market research has been done on orders in the Palatine Cafe, and the research shows that the probability a customer's offer includes coffee is found to be 2/5. If the Palatine Cafe serves 300 people in a day, the expected number of orders including coffee will be $300 \times (2/5) = 120$.

4.9 Conditional Probability

4.9.1 Motivation

Sometimes we learn information that will make us want to reassess the probability of something happening. Imagine I've bought a National Lottery ticket, and chosen six numbers from 1 to 49. I will win the jackpot

if the same six numbers are chosen by the Lottery Machine at the end of the week. Assuming the Lottery Machine chooses those six numbers randomly and fairly, my probability of winning the jackpot is 0.0000072%.

Imagine further though that when the first number is announced, it matches one of the six I have chosen. I now need the Lottery Machine to choose the five numbers I have remaining in order to win the jackpot, which has a probability of 0.000058%. If the second number also matches one of mine, the probability of winning goes up to 0.00056%, and so on. If on the other hand any number comes out that doesn't match mine, then I know I've lost, and the probability drops to zero.

This is a very simple example of situations in which the probability of an event can change, based on additional information we have received. Because situations in which we want to rethink probabilities based on additional information are so common (especially in statistics), we need some way to update probabilities in a mathematically rigorous manner. This is done through what we call **conditional probabilities**.

4.9.2 Updating the outcome space

One useful way to think about how updating probabilities works is to consider the outcome space. As we saw in Section 4.2, the outcome space is the set of all the possible outcomes of whatever situation we are looking at, so that one and only one outcome can happen, and that no outcome listed can be further broken up into two or more relevant possible results.

Gaining information about a situation corresponds to learning that some outcome cannot have happened. To return to my lottery example, we can think of the outcomes being the 13,983,816 different ways to choose six different numbers from 1 to 49. Once I learn the first number chosen by the Lottery Machine matches one of mine, all the outcomes in which **none** of my numbers are chosen become impossible.

This then gives me a new outcome space, where each outcome remaining includes the number the Lottery Machine has picked first. Once we learn the second number picked, that gives us another new outcome space, which contains only the outcomes which include both the numbers picked, and so on.

In situations in which all outcomes are equally likely, updating probabilities can be quite quick - you throw away the outcomes which are no longer possible, and calculate the new probability using the outcomes you have left.

In such a case, the probability event A happens given event B has happened is found as follows

$$P(A \text{ happens, given } B \text{ has happened}) = \frac{\text{Number of outcomes in both } A \text{ and } B}{\text{Number of outcomes in } B}$$

We often write $P(A \text{ happens, given } B \text{ has happened})$ as $P(A|B)$, said out loud as "the probability of A given B", for ease of notation.

4.9.3 The conditional probability formula

Often we're not lucky enough to be in a situation in which all outcomes are equally likely, of course. Still, the approach we just looked at can be adapted to situations in which some outcomes are more likely than others. Remember that we define the probability that event A happens as

$$P(A) = \frac{\text{Number of times } A \text{ happens}}{\text{Number of times } A \text{ could have happened}}$$

Now that we know B has happened, then all the times B **didn't** happen, whether or not A happened as well, are no longer relevant. We're interested only in how many times B happened **and** A happened, out of the number of times B happened and A **could have** happened.

$$P(A|B) = \frac{\text{Number of times } B \text{ and } A \text{ happens}}{\text{Number of times } B \text{ happens and } A \text{ could have happened}} \quad (**)$$

We can take this a stage further, using the mathematical result that a fraction does not change its value if you divide both top and bottom by the same number.

$$\begin{aligned} P(A|B) &= \frac{\text{Number of times } B \text{ and } A \text{ happens/Number of times something could have happened}}{\text{Number of times } B \text{ happens and } A \text{ could have happened/Number of times something could have happened}} \\ &= \frac{P(A \text{ and } B)}{P(B)}. \end{aligned}$$

This alternative form of the formula is commonly considered to be "the" conditional probability formula.

Important: For events A and B, we can talk about $P(B|A)$ just as easily as $P(A|B)$. This sometimes seems odd, if event B either happens or doesn't happen before event A does or doesn't happen. Probability is all about our beliefs about events **we personally** don't know whether have happened yet. It is perfectly possible for an event to have happened or not without you **knowing** whether it has happened or not, and in such circumstances it is still perfectly reasonable to discuss probability.

It's also important to bear in mind that, in general, $P(A|B) \neq P(B|A)$. We can see from equation $(**)$ that $P(A|B) = P(B|A)$ would only be true when the number of times A could have happened = the number of times B could have happened.

Example 4.11 In a catch-and-release study, one hundred swallows in the UK are captured during summer, given leg rings (numbered 1 to 100), and released into the wild. Two months later, one of the swallows is spotted in Tunisia.

Assuming the leg ring on the spotted swallow is equally likely to be each of the numbers between 1 and 100, find:

1. $P(A)$, where A is the event that the leg ring number is 20 or lower;
2. $P(B)$ where B is the event that the leg ring number is prime;
3. $P(A|B)$;
4. $P(B|A)$.

Answers (we use $\{1, \ldots, 100\}$ as the outcome space here):

1. 20 of the 100 outcomes are 20 or lower, hence $P(A) = 20/100 = 1/5$.
2. There are 25 prime numbers between 1 and 100. Hence 25 of the 100 outcomes are prime, and $P(B) = 25/100 = 1/4$.
3. Since we know B happened, our outcome space for this calculation contains only the 25 prime numbers between 1 and 100. Of these, 7 are below 20 (20 itself is not prime, of course). Hence $P(A|B) = 7/25$.
4. Since we know A happened, our outcome space for this calculation contains only the numbers 1 to 20. Of these, 7 are prime. Hence $P(B|A) = 7/20$.

Note that this is one of many situations in which $P(A|B) \neq P(B|A)$.

4.9.4 Conditional probabilities and relative frequency

Of course, there are many situations in which don't know the proportions we would need in order to directly calculate $P(A|B)$. In such situations, we can once again make use of **relative frequencies**, as in Section 4.4.

$$P(A|B) \approx \text{Relative frequency of } A|B = \frac{\text{Number of experiments run in which } A \text{ and } B \text{ happens}}{\text{Total number of experiments run in which } B \text{ happens}}$$

4.9.5 Conditional probabilities and independence

Our working definition of independence, as discussed in Section 4.5, is that the events A and B are independent of learning whether event A (B) has happens does not cause us to change our beliefs about the chance of B (A) happening.

If A and B are independent, then, what does that imply about the probabilities $P(A|B)$ and $P(B|A)$? We can answer this question using the definition of independence, but instead, we'll consider it from the perspective of the relevant equations, and then discuss how the implications of those equations match up to our definition of independence.

We know that

$$P(A \text{ and } B) = P(A)P(B)$$

when A and B are independent, and that

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \Rightarrow P(A \text{ and } B) = P(A|B)P(B)$$

whether A and B are independent or not. When A and B are independent, then, we must have

$$P(A \text{ and } B) = P(A|B)P(B) = P(A)P(B)$$

and hence that $P(A|B) = P(A)$. A very similar argument gives us that when A and B are independent, $P(B|A) = P(B)$.

So why does this make sense? What we are seeing here is that A and B are independent, what we believe the probability for A is after learning B has happened **should be the same as** what we believe the probability for A is **without knowing** B had happened. If A and B are independent, it doesn't make any difference to our beliefs about A if we learn B happened; that's what independence **means**!

The same argument justifies why $P(B|A) = P(B)$; learning A has happened gives us absolutely no additional information about how likely an event independent of A is likely to happen.

This is a very useful result in statistics. The equation $P(A \text{ and } B) = P(A|B)P(B)$ also comes in handy an awful lot, and is sometimes called the **conditional multiplicative rule**.

We can now finally fully understand what happened in **Example 4.6**. I showed already that the probability of rolling a 1 with the dice in my left hand is

$$P(R1) = P(R1 \text{ and } S) + P(R1 \text{ and } T)$$

using the additive rule. We can now use the conditional multiplicative rule to get

$$P(R1) = P(R1|S)P(S) + P(R1|T)P(T) = \frac{1}{6} \times \frac{1}{2} + \frac{1}{/}20 \times \frac{1}{/}2$$

as stated.

4.10 Bayes Rule

One of the most important uses of the conditional multiplicative rule is that it leads us to **Bayes rule**. This is a result so important it birthed an entire new approach to statistics; the helpfully-named **Bayesian statistics** that most if not all of you will be learning more about in Epiphany term.

Bayes rule works by exploiting the fact that $P(A \text{ and } B)$ must equal $P(B \text{ and } A)$. Either both events happen, or they don't both happen; it can't matter which of the two events we name first. Applying the conditional multiplicative rule, then, we have

$$P(B|A)P(A) = P(B \text{ and } A) = P(A \text{ and } B) = P(A|B)P(B)$$
$$\Rightarrow P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

This simple little trick unlocks an amazingly powerful result - we have a way of swapping round the order of events in a conditional probability. We can go back in time!

There are any number of situations in which this is extremely useful. Here's just one: imagine a GP wants to diagnose a sick patient. What she might want to do is figure out the probability of different diseases, given the symptoms being shown. That could be extremely hard to work out, though. It might be much easier to look up the probabilities of the **symptoms** given different diseases, and then use Bayes rule to swap that round into what she really wants. Let's dig into how that might work with an example.

Example 4.12 A doctor wants to calculate the probability a patient has the flu, given they have a cough, a sore throat, and aching arms. The doctor knows the following information:

a) Currently 3% of the UK population is suffering from the flu (based on the latest NHS data).

b) Currently 5% of the UK population is reporting they have all three of a cough, a sore throat, and open arms (again, based on the latest NHS data).

c) 90% of people with the flu report all three of a cough, a sore throat, and open arms (this has been demonstrated by previous medical research).

Using these values, the doctor can calculate the probability their patient has the flu. Denoting by F the event that a patient has the flu, and denoting by S the event that a patient has all three listed symptoms:

$$\begin{aligned} P(F|S) &= \frac{P(S|F)P(F)}{P(S)} \\ &= \frac{0.9 \times 0.03}{0.05} \\ &= 0.54. \end{aligned}$$

One way to intepret this result is that, before learning about the patient's symptoms, it would be sensible for the doctor to assume their probability of having the flue is 0.03. Once she learns about the symptoms, this probability increases by a multiple of 18.

5 Random Variables

We've now covered the basics of probability theory, upon which just about everything in the field of statistics is based. There's still a lot more to talk about, though. The kinds of examples we've been looking at have been very simple in terms of the situations involved (which isn't to say the maths might not have been challenging). These sorts of situations - dice rolls, coin tosses, lotteries - aren't really where we need to put our focus. If you can work out probabilities just through counting equally likely outcomes, you don't need to call in a statistician.

The kinds of situation where a statistician becomes useful are those where you **can't** just do some calculations based on the situation's properties. We get called in when conclusions - potentially quite complex and subtle ones - need to be made by studying the behaviour of a system, via the data that system produces.

This will require an understanding of how we think about the data we collect in terms of the underlying probabilities of the situation. This in turn will mean grasping the concept of a **random variable**, another idea which is absolutely foundational to the practice of statistics.

So what is a random variable? How do we define them, how do they work, and why are they useful? Answering all three of those questions requires we consider an even more fundamental question first: what is a variable?

5.1 What is a Variable?

A **variable** is a quantity that can take one of several possible values. There are all sorts of ways in which this can happen. A value can regularly change as we observe it (temperature in a room), a value might be changeable by us as people (temperature of our oven), a value might be unchanging but unknown to us, so that it has several possible values and we don't know which is true (number of people born in the year 10,000 BCE). All of these are variables - they're quantities that have more than one possible value. Actually, we can even define a fixed **and** known value as a variable (number of people born in my house last year - that would be zero). It's quite a boring thing to do, but it's possible to do it, and sometimes useful.

All of that is quite a broad definition, but that's not a bad thing. Indeed, given there are so many situations in which we might want to think about variables, it's actually quite important to have so wide a definition. We can get a little more technical, though, as follows: a variable X is associated with a set $\mathcal{X}$ of values, which describe the possible values X can take. The following would all be examples of variables:

- X, $\mathcal{X} = \mathbb{N}$ (representing all positive whole numbers)
- Y, $\mathcal{Y} = [0, 1]$ (representing all real numbers between 0 and 1);
- Z, $\mathcal{Z} = \{\text{Red, Blue, Green}\}$ (this means Z is a qualitative value);
- A, $\mathcal{A} = \{0\}$.

Note that variable A is of the kind earlier discussed, where it can only take one value, in this case 0. A little dull, but nothing in the notation I just gave you requires the set associated with a variable has to have more than one element.

5.2 Making Variables Random

So far, none of the variables I've defined are random, because they lack a crucial property of random variables. To be a random variable, a variable has to be associated not just with a set of possible values, but an **expression of belief** regarding how likely each of those values is to be the one taken.

For example, let's say I define a variable X, with $\mathcal{X} = \{H, T\}$. This variable could very well represent the result of a coin toss, which will be either Heads (H), or Tails (T). Coin tosses are a very common example of a random situation, but the variable X itself is not random yet. To become a random variable, I need to express my belief about how likely each of the values in $\mathcal{X}$ are to being the value X takes. I could do this by setting the probability of Heads at 0.5, $P(H) = 0.5$, representing a fair coin. I could set $P(H) = 0.3$, if I think the coin has been tampered with in some way so that Heads only come up 30% of the time. I could

even not assign a number at all, using algebra instead; $P(H) = p,\ p \in [0, 1]$ expresses my belief that the probability of getting Heads is p, where p is an unspecified number between 0 and 1 (note my use of the Greek letter epsilon, $\in$, to denote "belongs to" here).

Each of these expressions of belief turns X into a random variable, and turns $\mathcal{X}$ into the outcome space for X. We describe the value that X ends up taking as a **realisation**, often denoted x.

It's a common mistake to see non-random variables and random variables as somehow opposite concepts. I don't think this is a useful approach. I think a better way to consider what we're doing is in terms of a house. Randomness is something you build **on top of** the concept of variables. It makes no more sense to say non-random variables and random variables are opposites than it is to say one-storey houses and two-storey houses are opposites. In particular, a variable in which you express certainty about which value it will take is **still** a random variable, even though the outcome is fully known, because you've expressed that fact as a belief.

5.3 Probability Functions

5.3.1 Functions

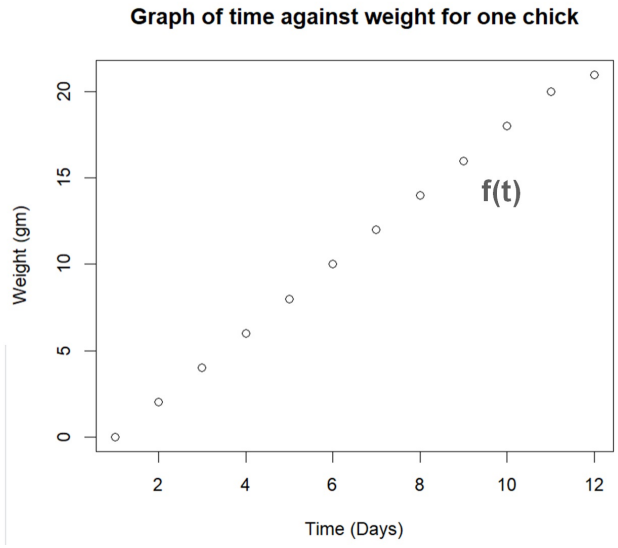
Functions, in the context of mathematics, are processes by which an input is converted into an output. The inputs, the processes, and the outputs can all be quite weird and complicated sometimes, but the basic idea is always the same.

Commonly in maths, we use letters to denote functions. The expression $f(x) = x^2$, for instance, defines a function f which takes an input x, and squares it.

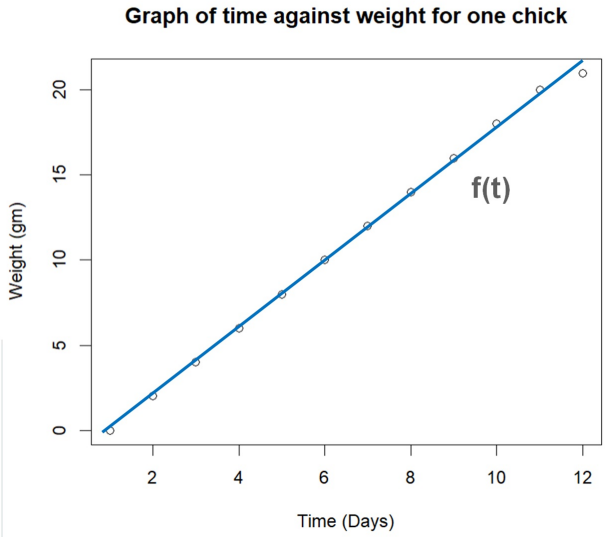
Functions also require what we call a **domain**, which is all the values you are allowed to use as inputs, and a **range**, which is all the values you can get as an output. What the range looks like will depend on what your domain looks like. If I say the function f has a domain of $\mathbb{R}$ (all the real numbers), then the range is the interval $[0, \infty)$, because for any non-negative number, there is always **some** number you can square to get to that number. You can't get negative numbers after squaring a real number, though, so negative numbers aren't in the range of f. If I instead use the domain $[2, \infty)$ for f, the range becomes $[4, \infty)$, because the smallest output I can get is now $2^2 = 4$.

We say a function is **closed form** if we always know what the output is for every input. Not all functions have this property, especially in statistics, where we often think of real-world situations as being powered by functions we can never find, merely try to estimate through our models.

Example 5.1 The data set `chicks` in `R` tracks the weights of 50 newly-hatched chicks, with each chick being weighed once a day. For any one chick, we could imagine a function, say f, for which the input t is the time in minutes since a chick hatched, and the output $f(t)$ is the chick's weight in grams.



We can draw a graph for this function, as shown below. Because we only have one measurement a day, though, the graph has long stretches of time in which the chick must have weighed **something**, but we cannot possibly say what. The function f therefore does not have closed form. Attempting to estimate the outputs for the values of t for which we didn't record the chick's weight, say by drawing a line of best fit (see below), is a classic example of what we do in statistics.



5.3.2 Probabilities as functions

We'll now combine what we understand about probabilities with what we understand about functions. I'm going to be quite precise in my terminology here, so make sure you're comfortable with that terminology before reading beyond it.

A **probability function**, which is also known as a **probability distribution**, is a **function**. The **input** of a probability distribution will either be one **outcome**, or an event made up of multiple outcomes. The **output** of a probability function will be a number between 0 and 1, representing the **probability** of the outcome or outcomes which make up the input.

Be careful to note then that an **outcome** is not an **output**, it is an **input**. Also note that sometimes,

statisticians may use the word "probability" to talk about the **probability function**, rather than the probabilities that the function gives us as outputs. It should usually be clear in context which one is being referred to, but keep on your toes!

Example 5.2 Consider a six-sided dice. Let the value I next roll on that dice be expressed as a variable X. The variable can take values from $\mathcal{X} = \{1, \ldots, 6\}$. To make this into a random variable, I need to add a belief statement. I shall assume the dice is fair, and that therefore each of the elements of $\mathcal{X}$ has a $1/6$ probability of being the one I roll.

It is common in mathematics to denote outcomes using a lower-case Greek omega. I shall therefore denote the outcome "I roll a one" as ω_1, the outcome "I roll a two" as ω_2, and so on, up to ω_6. I will now call my probability function P, and express the outputs as follows

$$P(\omega_i) = \frac{1}{6}, \text{ for every } i \in \mathcal{X}.$$

I could instead offer a different belief statement. Perhaps I know the dice has been tampered with, so that the probability of rolling a six is actually $1/2$, with the other five outcomes being as likely as each other, though all less likely than a six.

I'll define a new probability function $\tilde{P}$ for this belief:

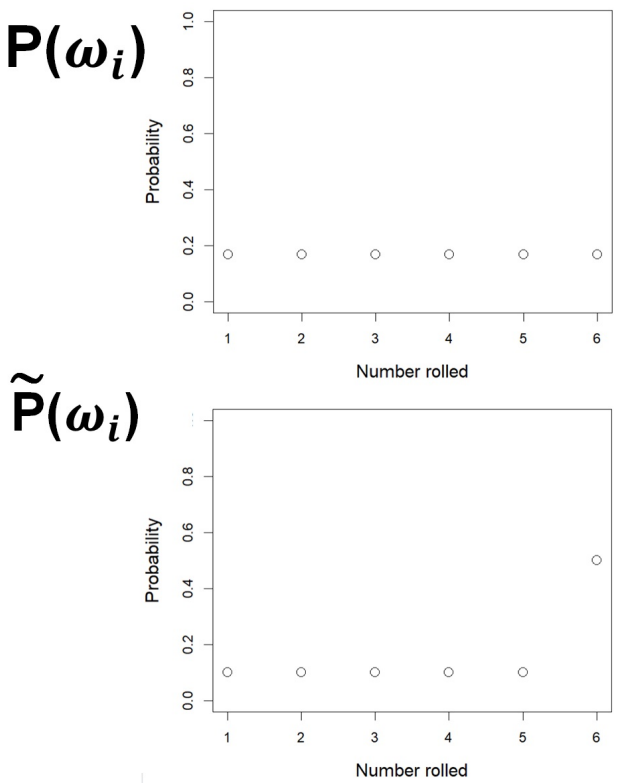
$$\begin{aligned} \tilde{P}(\omega_i) &= \frac{1}{10}, \text{ for every } i \in \{1, 2, 3, 4, 5\}, \\ \tilde{P}(\omega_6) &= \frac{1}{2}. \end{aligned}$$

5.4 Introduction to Discrete and Continuous Probability Functions

There are two broad forms of probability functions, both of which we'll consider in more detail later in the notes. One are **discrete distributions**. These are probability functions for which the outcomes (and hence the inputs) are either finite in number (such as in **Example 5.1**, or when dealing with, say, drawing cards from a deck, or pulling marbles from a bag), or they are what we call **countably infinite**. An outcome space is countably infinite if there's an infinite number of outcomes, but there are gaps between the outcomes which represent impossible events.

For instance, if we think about the number of insects on the planet, there's no easy way to think about the biggest possible number that could be, so we would consider the outcome space to be infinite. However, we would consider it impossible to have any number of insects that wasn't a whole number. Therefore there are gaps between outcomes which represent impossible events, and the outcome space is countably infinite.

Both of the probability functions we looked at in **Example 5.2** are discrete distributions, because there is only a finite number of rolls we could see. If we try to imagine a dice with infinite faces (or perhaps a finite number, but a finite number which is unknown and with no limit on how big it could be), we would still call any probability function related to that dice a discrete distribution, because non-whole number rolls would be impossible. Sometimes we call a discrete distribution a **probability mass function**, or PMF.



Continuous distributions, in contrast, are probability functions for which the outcome space includes infinite values. The uniform distribution $U[0,1]$ is a good example of such a function. This is a probability function which takes any value in the interval $[0,1]$ as an input (and there are an infinite number of such values), and which gives its outputs in such a way that all outcome values are considered equally likely.

This quickly gets a bit complicated, though, because since there are infinitely many outcomes, and they all have to have the same probability, the probability of getting any one specific outcome has to be zero. This is true of all continuous distributions, in fact. In order to deal with this, we have to express continuous distributions so they only give non-zero outputs when the inputs are *intervals*. Continuous distribusions are also referred to as **probability density functions**, or PDFs.

5.5 Discrete Random Variables

In this section we shall consider three of the most common discrete random variables. In each case, the set of realisations that the variable can take is either the natural numbers (0,1,2, etc.) or some subset of those numbers. As we've noted, the realisations of discrete random variable doesn't **have** to be limited to natural numbers (just imagine a six-sided die with the faces labelled "-1", "0,", "0.5", "$\sqrt{2}$", "π", and "e"), but they will do so here.

5.5.1 The Bernoulli distribution

The **Bernoulli distribution** is possibly the simplest distribution we can think of, other than a distribution where a random variable always takes the same value (we call that the **atomic distribution**).

For a Bernoulli random variable X, the corresponding outcome space $\mathcal{X} = \{0,1\}$. We denote the probability of a realisation equalling 1 as being p. This then means $P(X = 0) = 1 - p$.

We express this whole set-up through the notation $X \sim Ber(p)$. Here, p is the one and only **parameter** of the distribution.

You will be seeing that tilde sign $\sim$ a lot in this module (and possibly others). It is a short way of saying

"...is a random variable which has distribution...". Hence $X \sim Ber(p)$ can be thought of as saying "X is a random variable which has a Bernoulli distribution".

5.5.2 The binomial distribution

The **binomial distribution** can be thought of in two ways. The common way we express it is as follows: imagine we are about to observe the results of n trials. Each trial can only have two outcomes (often denoted 0 and 1, or "success" and "failure"). Let the random variable X represent the number of successful trials (so $\mathcal{X} = \{0, 1, \ldots, n\}$).

If each trial has the same probability p of success, and if every trial is independent of every other trial (so learning about the result of one trial gives you no information about whether any other trial will be successful), X will have a binomial distribution, $X \sim Bin(n, p)$. The two parameters here are n and p, illustrating the fact that the two ways to change the behaviour of the series of trials is to change how many trials are performed, and/or to change the probability of each trial being successful.

For given values of n and p, we can express the probability of each possible realisation within $\mathcal{X} = \{0, 1, \ldots, n\}$. In what follows, the notation $r!$, pronounced r **factorial**, denotes the product of all natural numbers between 1 and r, i.e. $r! = 1 \times 2 \times \ldots \times (r-1) \times r$. For book-keeping reasons, we set $0! = 1! = 1$.

$$P(X = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

We can break down this equation to see how it works. The term p^r represents the probability of observing r successful trials in a row (using the fact that the trials are independent). The term $(1-p)^{n-r}$ represents the probability of observing $n-r$ failed trails in a row. Combining these gives us that the probability of observing r success followed by $n-r$ failures is $p^r(1-p)^{n-r}$. Again under the assumption of independence, this must also equal, say the probability of observing $n-r$ failures followed by r successes, or the probability of observing r successes and $n-r$ failures **in any order at all**.

We can therefore find $P(X = r)$ by multiplying this probability of observing r successes and $n-r$ failures in one specific combination by the number of such combinations there are (this relies on the additive rule for mutually exclusive events, and in noting each specific order of r successes and $n-r$ failures must be mutually exclusive to any other specific order of r successes and $n-r$ failures).

The expression $\frac{n!}{r!(n-r)!}$ gives us that number of different combinations of r successes and $n-r$ failures. I won't explain how this works here, but I've put together a separate document with that explanation, which is also available in the Week 3 material on Ultra.

I mentioned above that there are two ways to think about the binomial distribution. The second is to recognise that each individual one of the n trials we're observing has the same probability p of success. Therefore, each individual trial can be represented by a Bernoulli random variable. The behaviour of the random variable $X \sim Bin(n, p)$ is therefore equivalent to the behaviour of the **sum** of n Bernoulli random variables $Y_1, \ldots, Y_n$, each of which is a Bernoulli random variable, $Y_i \sim Ber(p)$.

For this to work, each of these n Bernoulli random variables has to be independent of all the others. We refer to such groups of independent random variables, all with the same distribution, as **independent and identically distributed** random variables, or iid RVs for short.

5.5.3 The Poisson distribution

Both the Bernoulli and binomial distribution assume can only happen a certain number of times, and concern themselves with how many times it in fact does happen. The Poisson distribution is a little different. We use the Poisson distribution in situations where there is no specific limit to how something might happen. Rather than fix the number of events, then, the Poisson distribution works across a specific **interval** (commonly but not always an interval of time), in which some event can happen (in theory) any number of times. In order

for this to work, we assume that the events in question happen independently - that is, an event happening at a given point in the interval tells us nothing additional about when we might expect the next event to occur.

The Poisson distribution has a single parameter, λ, which is referred to as the *intensity*. The larger the value of λ, the more times we expect to see the event happen over the interval associated with the distribution. We express that the random variable X has a Poisson distribution with intensity λ by writing $X \sim Pois(\lambda)$.

The formula for finding the the probability of seeing exactly r events over the interval associated with $X \sim Pois(\lambda)$ is written below.

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}.$$

Note that the Poisson distribution is somewhat different to the Bernoulli and binomial distributions, in that there is no upper limit to the value of its realisations. There are therefore infinitely many possible values a Poisson random variable can take. Once past a certain point, though (which varies depending on the value of λ for the distribution), the probabilities get smaller and smaller as the realisation value gets larger and larger, and they do so in such a way that, even though there is an infinite number of them, these probabilities still all sum to 1.

5.6 Continuous Random Variables

- For a continuous random variable, the role of the probability mass function is taken by a density function, $f(x)$, which has the properties that $f(x) \geq 0$ and
$$\int_{-\infty}^{\infty} f(x)dx = 1$$
- For any $a < b$, the probability that X falls in the interval (a, b) is the area under the density function between a and b:
$$P(a < X < b) = \int_a^b f(x)dx$$
- Thus the probability that a continuous random variable X takes on any particular value is 0:
$$P(X = c) = \int_c^c f(x)dx = 0$$
%Although this may seem strange initially, it is really quite natural. If the uniform random variable of Example A had a positive probability of being any particular number, it should have the same probability for any number in $[0, 1]$, in which case the sum of the probabilities of any countably infinite subset of $[0, 1]$ (for example, the rational numbers) would be infinite.
- If X is a continuous random variable, then
$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$
Note that this is not true for a discrete random variable.

5.7 Cumulative distribution function

- The **cumulative distribution function** (cdf) of a continuous random variable X is defined as:
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$
- The cdf can be used to evaluate the probability that X falls in an interval:
$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

5.8 Characteristics of probability distributions

- If X is a continuous random variable with density $f(x)$, then

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

or in general, for any function g,

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

- The variance of X is

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

- The variance of X is the average value of the squared deviation of X from its mean.
- The variance of X can also be expressed as $Var(X) = E(X^2) - [E(X)]^2$.

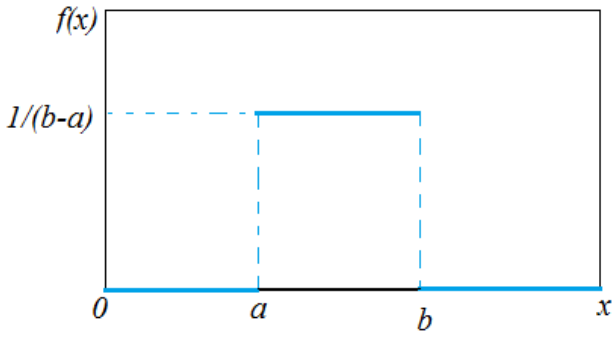
5.9 Some useful continuous distributions

5.9.1 Uniform distribution

- A random variable X with the density function

$$f(x) = \frac{1}{b-a},\ a \le x \le b$$

is called the uniform distribution on the interval $[a, b]$.



- The cumulative distribution function is

$$F(x) = \begin{cases} 0 & \text{for} \quad x < a \\ \frac{x-a}{b-a} & \text{for} \quad a \le x < b \\ 1 & \text{for} \quad x \ge b \end{cases}$$

- A special case, $f(x) = 1$ and $0 \le x \le 1$.

5.9.2 Exponential distribution

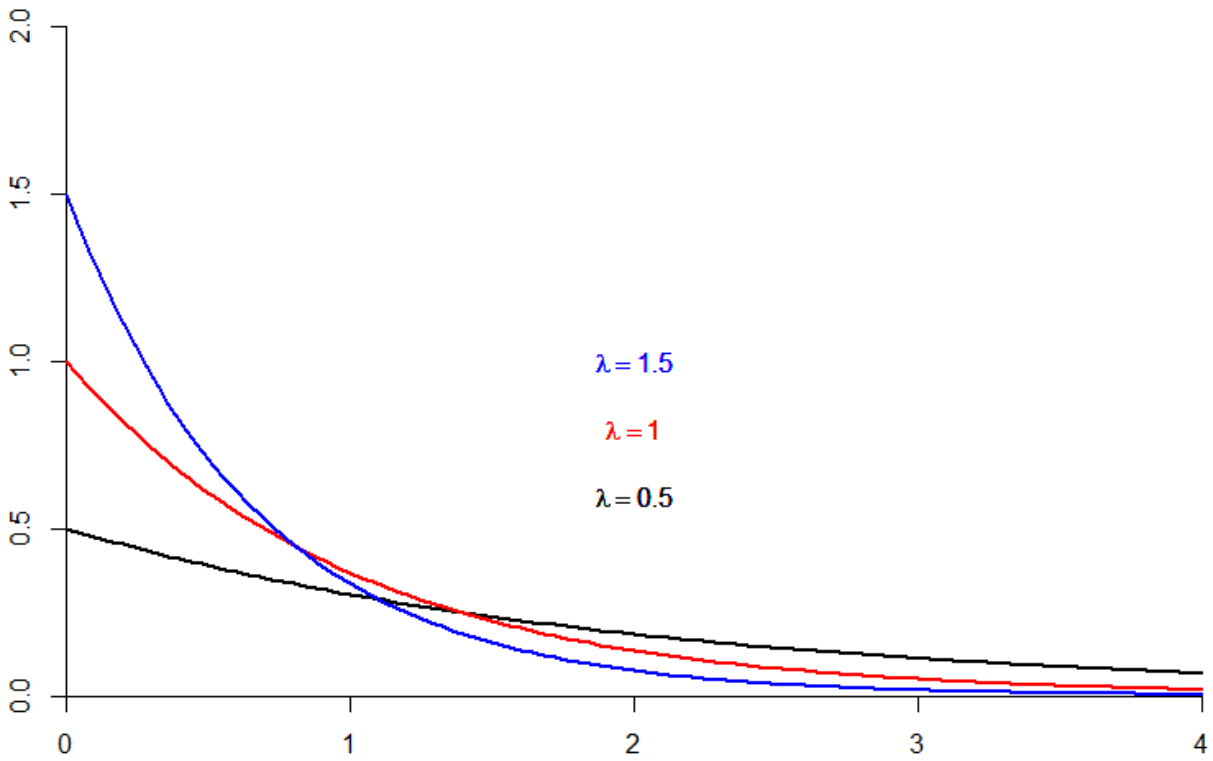
- The exponential density function is

$$f(x) = \lambda e^{-\lambda x},\ x \ge 0 \ \text{ and } \ \lambda > 0$$

- The cumulative distribution function is

$$F(x) = \int_{-\infty}^{x} f(u)du = 1 - e^{-\lambda x}$$

- The exponential distribution is often used to model lifetimes or waiting times data.

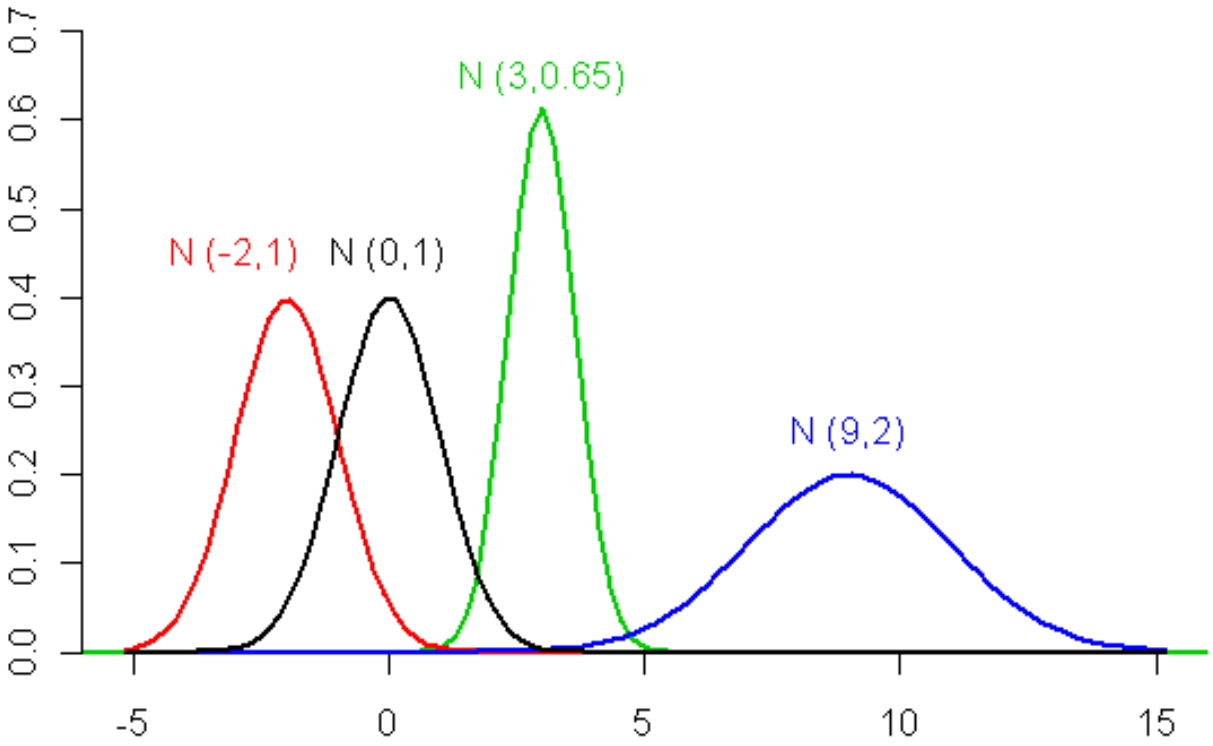


5.9.3 Normal distribution, $N(\mu, \sigma^2)$

- The normal (Gaussian) distribution plays a central role in probability and statistics, probably the most widely known and used of all distributions
- The normal distribution fits many natural phenomena, e.g. human's height, weight, IQ scores. In business, for example, the annual cost of household insurance, among others.
- The density function of the normal distribution depends on two parameters, μ and σ (where $-\infty < \mu < \infty$, $\sigma > 0$):

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

- The parameters μ and σ are the mean and standard deviation of the normal density.
- We write $X \sim N(\mu, \sigma^2)$ as short way of saying 'X follows a normal distribution with mean μ and variance σ^2'.



5.9.4 Standard normal distribution $N(\mu = 0, \sigma^2 = 1)$

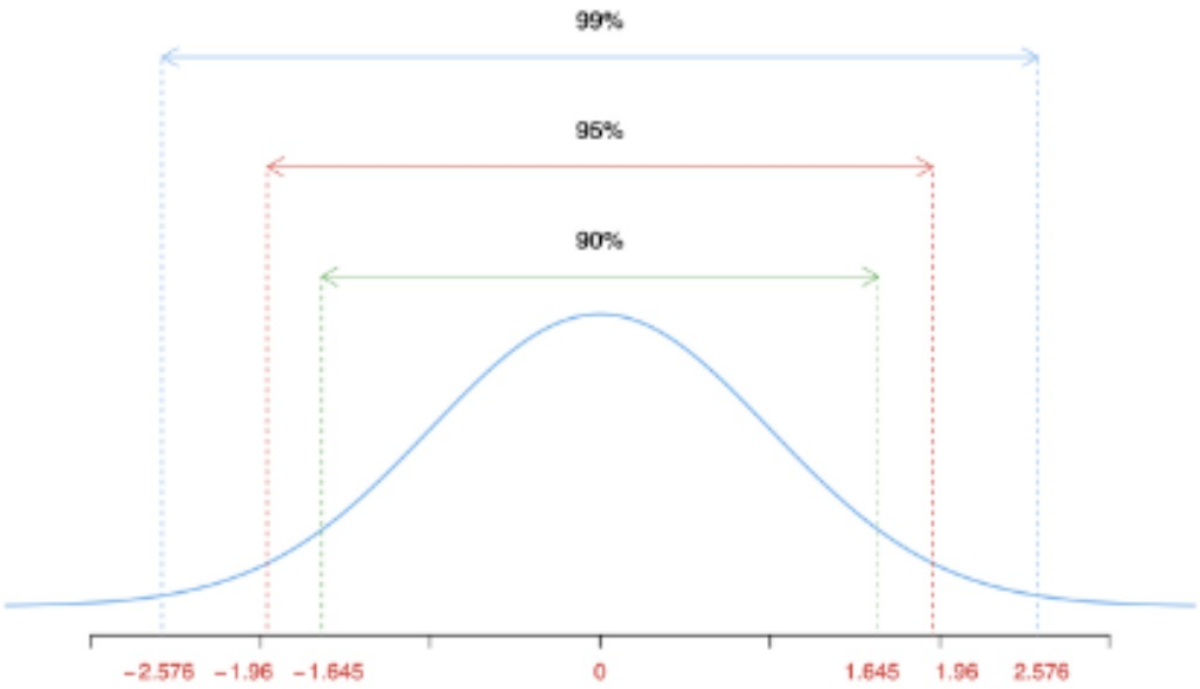
- The probability density function of the standardized normal distribution is given by:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, -\infty < z < \infty$$

- We write $Z \sim N(0, 1)$ as short way of saying 'Z follows a standard normal distribution with mean 0 and variance 1'.
- To standardize any variable X (into Z) we calculate Z as:

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

  The Z-score calculated above indicates how many standard deviations X is from the mean.



5.9.5 Example

- If f_X is a normal density function with parameters μ and σ, then

$$f_Y(y) = \frac{1}{a\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{y-b-a\mu}{a\sigma}\right)^2\right]$$

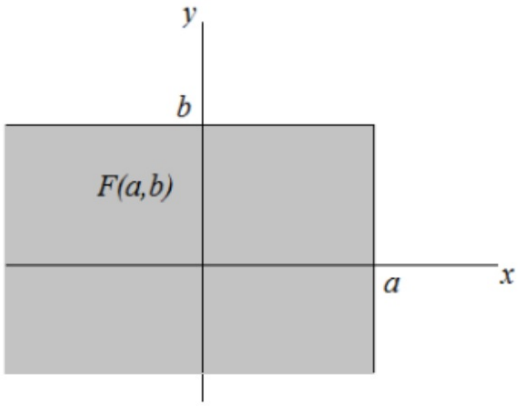
- Thus, $Y = aX + b$ follows a normal distribution with parameters $a\mu + b$ and $a\sigma$.
- If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$.
- Can you use this to show that $Z \sim N(0, 1)$?

5.10 Joint distributions

- The joint behaviour of two random variables, X and Y, is determined by the cumulative distribution function,

$$F(x, y) = P(X \leq x, Y \leq y)$$

regardless of whether X and Y are continuous or discrete. The cdf gives the probability that the point (X, Y) belongs to a semi-infinite rectangle in the plane.



- The joint density function $f(x, y)$ of two **continuous random variables** X and Y is such that

$$f(x, y) \geq 0$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$$

$$\int_{c}^{d} \int_{a}^{b} f(x, y)dxdy = P(a \leq X \leq b, c \leq Y \leq d)$$

  The marginal density function of X is

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

  Similarly, the marginal density function of Y is

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

- The **cdf** of two **continuous random variables** X and Y can be obtained as

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)dvdu$$

  and

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

  wherever the derivative is defined.

5.11 Conditional probability (density) function, PDF

- The conditional probability (density) functions may be obtained as follows:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f(y)} \quad \text{conditional PDF of } X$$

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f(x)} \quad \text{conditional PDF of } Y$$

- Two random variables X and Y are statistically independent if and only if

$$f(x, y) = f(x)f(y)$$

  That is, if the joint PDF can be expressed as the product of the marginal PDFs. So,

$$f_{X|Y}(x|y) = f(x) \text{ and } f_{Y|X}(y|x) = f(y)$$

5.12 Properties of Expected values and Variance

- The expected value of a constant is the constant itself, i.e. if c is a constant, $E(c) = c$.
- The variance of a constant is zero, i.e. if c is a constant, $Var(c) = 0$.
- If a and b are constants, and $Y = aX + b$, then $E(Y) = aE(X) + b$ and $Var(Y) = a^2 Var(X)$ (if $Var(X)$ exists).
- If X and Y are independent, then $E(XY) = E(X)E(Y)$ and
$$Var(X + Y) = Var(X) + Var(Y)$$
$$Var(X - Y) = Var(X) + Var(Y)$$
- If X and Y are independent random variables and g and h are fixed functions, then
$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

5.13 Covariance

- Let X and Y be two random variables with means μ_x and μ_y, respectively. Then the **covariance** between the two variables is defined as
$$cov(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - \mu_x\mu_y$$
- If X and Y are independent, then $cov(X, Y) = 0$.
- If two variables are uncorrelated, that does not in general imply that they are independent.
- $Var(X) = cov(X, X)$
- $cov(bX + a, dY + c) = bd\ cov(X, Y)$, where a, b, c, and d are constants.

5.14 Correlation Coefficient

- The (population) correlation coefficient ρ is defined as
$$\rho = \frac{cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{cov(X, Y)}{\sigma_x\sigma_y}$$
- Thus, ρ is a measure of **linear** association between two variables and lies between -1 (indicating perfect negative association) and $+1$ (indicating perfect positive association).
- $cov(X, Y) = \rho\ \sigma_x\sigma_y$
- Variances of correlated variables,
$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2cov(X, Y)$$
$$Var(X \pm Y) = Var(X) + Var(Y) \pm 2\rho\ \sigma_x\sigma_y$$

5.15 Conditional expectation and conditional variance

Let $f(x, y)$ be the joint PDF of random variables X and Y. The conditional expectation of X, given $Y = y$, is defined as
$$E(X|Y = y) = \sum_x x f_{X|Y}(x|Y = y) \quad \text{if } X \text{ is discrete}$$
$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|Y = y)dx \quad \text{if } X \text{ is continuous}$$
The conditional variance of X given $Y = y$ is defined as, if X is discrete,
$$Var(X|Y = y) = \sum_x [X - E(X|Y = y)]^2 f_{X|Y}(x|Y = y)$$
and if X is continuous,
$$Var(X|Y = y) = \int_{-\infty}^{\infty} [X - E(X|Y = y)]^2 f_{X|Y}(x|Y = y)dx$$