Notes on Statistics for Data Science
Spring 2021
Western Sydney University

Prof. Paul Hurley, Dr. Franco Ubaudi
(with thanks to prior lecturers)

2021-07-19
Version 1.0

Contents

Lecture 1

Introduction, motivation, puzzles and randomness

1.1 Organisation

Welcome to “Statistics for Data Science”, which encompasses the units “thinking about data 301108” and “nature of data 301114”.

Our teaching team is:

- Prof. Paul Hurley (lecturer, coordinator)
- Ryan Greenup (undergraduate: thinking about)
- Dr. Franco Ubaudi (postgrad: nature of)
- Shaira Viaje (undergraduate: thinking about)

We are here to help you learn, so please engage. This unit requires you to spend time getting to grips with statistical thinking and practising programming small pieces of code in R, a computer language for statistical analysis. Understanding *why* we are doing things makes all the difference to success in this unit.

For this unit all practicals and assessments will be through the means of a jupyter notebook, and you will have received by email an account on our server ds-stats.

Assessments: This unit will have 5 online quizzes, an assignment in two parts, and at the end a two hour practical exam.

Quizzes will be done on our moodle server – more information to come as we transition to that system.

Lectures are every Friday from 15:00 - 17:00. We record them and make them available for 7 days for going over again. Towards the end of the semester we will make all of them available again for revision.

The practical you attend depends on which unit you are doing. Thinking about data (undergrad) practicals are:

- Thursday 15:00 - 17:00 in PS-EA.LG.27
- Friday 15:00 - 17:00 in PS-EA.LG.18

If you are an offshore student, you should participate in the Thursday practical. To get a zoom link please send Paul an email. Key to success is attending and actively participating in these.

The practical sheets will appear, in advance of the lecture, in your home folder on the ds-stats server.

The notes you are reading now are an active document being updated weekly. From the 3rd week onwards these notes will appear at the latest on the Wednesday evening – we'll let you know when they are updated, together with a link. They will get even better with time!

Attendance at practicals and for the lecture are highly correlated with getting a good final result (and more important than that, learning a lot!). As you will learn later, *correlation is not causation.* It may only be that people who get good results might be the people who attend, though I do feel causation works both ways here :)

Not all of you start from the same point – some have experience with statistics already. Naturally for them the unit will be less challenging. We are here to help all of you succeed. Also, if you need additional help, mathematics education support or MESH (www.westernsydney.edu.au/mesh), a service offered by the university, is there to help.

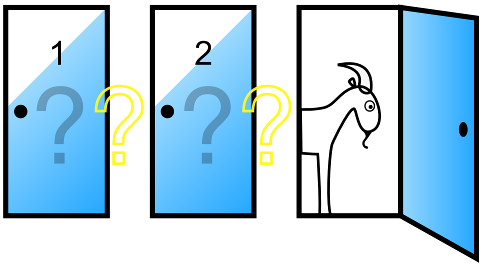
In this unit, we will use data-sets to guide us through some concepts, and we will use some so-called paradoxes or puzzles.

1.2 Some puzzles

Probability theory is very important as the language of the statistics. It gives us the modelling tools to create and understand statistical methods, and it gives us the ability to make precise concepts such as “randomness”. It is a tool to our understanding.

It also enables us to get past sometimes misleading intuition. Here are some probability questions we will come back to at various points throughout the unit.

Monty Hall problem



(credit for the image: wikipedia, released under public domain)

Monty Hall was a game show host in the US. Based on it, a probability puzzle arose.

It goes essentially as follows. Suppose you are presented by Monty with three doors. Behind two of them is a goat, but behind one of them is a car. You try to win the car. First, you pick a door. Monty doesn't open it. Instead he opens one of the other two doors to reveal a goat. And then he presents you with a choice: either keep your original door or pick the other door not yet open.

The question is this: which strategy is better, if any.

Most people think it doesn't make any difference: the chances that the car behind each door should remain the same. But Monty has given us extra information, and now the situation changes. It turns out that you double your chances by changing doors.

If you tear your hair out thinking about this you are not alone. This puzzle has generated much controversy, stubborn refusal to believe it makes any difference, and so on. The wikipedia article, `https://en.wikipedia.org/wiki/Monty_Hall_problem` has more of the details for those of you curious to learn more.

We are going to return to this puzzle: first through simulating the game many times (see your first practical worksheet) to convince ourselves of this result, and then through probability, proving that is the case.

The birthday problem

Suppose you have a number of people N in a room. Imagine, for example, a lecture hall: those mythical places where people gathered pre-COVID.

How many of them do you think share a birthday?

Let's pretend birthdays are distributed evenly throughout the year (making a uniform approximation), and let's not worry about leap years or time zones. Doing that is actually building a model that approximates reality. We can interpret our results given these conditions, which is standard in statistics (it's important not to forget the assumptions, and to appreciate when they break down).

Now, how high would N have to be to ensure a probability of 50% share the same birthday? Most people think that it would be high, perhaps $N-1$ in 365. This is because we think in terms of the probability of people in the room sharing a birthday with us. However, we only require *any two people* to share a birthday, and there are lots of pairs of people in a room with

It turns out that if $N \geq 23$, the probability of two people having the same birthday is over 50. If there are 70 people in the room, there is a 99.9% probability.

For more details, wikipedia is again our friend:
`https://en.wikipedia.org/wiki/Birthday_problem`

You might think this is just a parlour game, an experiment for fun. This counter-intuitive result though pops up all the time (hash clashing, doing multiple experiments and some sharing the same result), and failure to realise it means we can see patterns where there are none.

We will come back to this problem also by testing it through simulation, by checking the birthdays of the class, and then by deriving through probability theory the result.

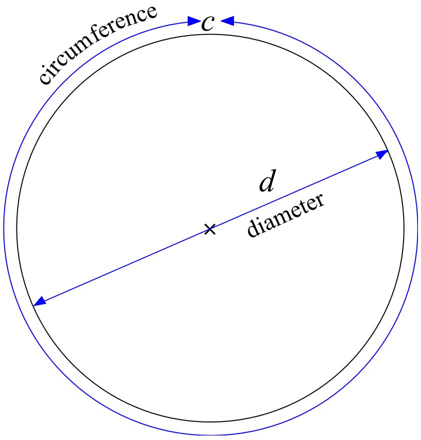
1.3 Statistical thinking

We are going to make concrete the concept of “randomness”: what it means for something to be random. Randomness is closely related to predictability.

Also important, and something we will keep emphasis is the difference between model and random experiment. So please keep that in mind and ask yourself each time if you grasped the concept.

As a jumping off point let’s look a number whose digits appear to be random (although it itself is deterministic).

π is the ratio of the circumference of a circle to its diameter.



Some properties:

- 3.141593 to 6 decimal places.
- π cannot be expressed as a fraction - it's irrational
- Its *decimal expansion* goes on forever.
- Given a sequence of digits, can we predict the next digit with any certainty? Or does the sequence seem *random*?

To 500 decimal places it looks like:

```
3.141592653589793238462643383279502884197169399375
10582097494459230781640628620899862803482534211706
79821480865132823066470938446095505822317253594081
28481117450284102701938521105559644622948954930381
96442881097566593344612847564823378678316527120190
91456485669234603486104543266482133936072602491412
73724587006606315588174881520920962829254091715364
36789259036001133053054882046652138414695194151160
94330572703657595919530921861173819326117931051185
48074462379962749567351885752724891227938183011949 16
```

That looks pretty random. But (i) what does it mean to be random? and (ii) is this random according to that definition?

The best mathematics works by a mixture of intuition and precision. The definitions inform us, and the definitions result from capturing intuition.

To 500 decimal places, the fraction 1/9 =:

```
0.111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111 111
```

We can also construct something whose numbers vary in a predictable way, such as $1/81 - 10/89999999991$ which is:

```
0.012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567
89012345678901234567890123456789012345678901234567 8901
```

This clearly is a repeating pattern.

While the constant $e =$

```
2.718281828459045235360287471352662497757247093699
95957496696762772407663035354759457138217852516642
74274663919320030599218174135966290435729003342952
60595630738132328627943490763233829880753195251019
01157383418793070215408914993488416750924476146066
80822648001684774118537423454424371075390777449920
69551702761838606261331384583000752044933826560297
60673711320070932870912744374704723069697720931014
16928368190255151086574637721112523897844250569536
96770785449969967946864454905987931636889230098793 11
```

e is a constant of great importance in mathematics, and it, like π is transcendental. That certainly has an air of randomness to it.

What it means to be random

There are generally two aspects to random.

1. The digits should occur the same proportion of the time (*uniformity*)
2. They should not be predictable, given part of the sequence/expansion we can't predict the next digit.

It can be shown that this is equivalent to

- single digits occur uniformly,
- pairs of digits occur uniformly,
- triples of digits occur uniformly
- … etc.

So is it true for π? Are the digits uniformly distributed? Or do some digits appear more often in?

For the first 500 digits, we get the following frequency table:

0	1	2	3	4	5	6	7	8	9
45	59	53	50	53	50	49	36	53	52

If they are uniform, then we should expect to see equal numbers of each digit.

Ten possible digits, and given the first 500, we *expect* to see 50 of each if they are uniform…

0	1	2	3	4	5	6	7	8	9
45	59	53	50	53	50	49	36	53	52

- Only two have exactly 50 occurrences (3 and 5).

- 7 occurs only 36 times.

What could explain this?

- They may not be uniform
- We only looked at 500 digits
- We only looked at the *first* 500 digits

Compare to randomly-generated digits

Now we reach one of the most important tools in our bag for testing randomness. The essential idea is to generate digits randomly chosen, with each digit equilikely[1].

Spend some time understanding why we do this. We are presented with some data (in this case digits from π)), and we want to see if they are consistent with numbers that are random. This is something we will keep coming back to: is something likely to be coincidence, or is there a pattern or effect? If a drug is tested, for example, we want to know if people got well just due to coincidence ("time heals all wounds") or because they took the drug.

Each row in the table below represents one frequency experiment of 500 random digits.

0	1	2	3	4	5	6	7	8	9
42	48	49	45	53	51	50	47	57	58
45	45	57	53	46	42	52	52	58	50
53	46	44	49	50	51	58	56	48	45
57	55	54	53	52	50	46	50	53	30
49	60	53	58	52	49	46	54	44	35
53	53	55	58	38	51	52	48	39	53
54	52	56	48	53	56	60	40	44	37
48	48	48	49	48	46	53	63	56	41
43	47	43	55	41	47	63	50	69	42
53	44	52	61	56	44	43	50	46	51

What do you notice? First, there are quite a few differences from the *expected* count of 50. The largest is 69 and the smallest 30.

Let's look at variation from 50 by subtraction.

0	1	2	3	4	5	6	7	8	9
-8	-2	-1	-5	3	1	0	-3	7	8
-5	-5	7	3	-4	-8	2	2	8	0
3	-4	-6	-1	0	1	8	6	-2	-5
7	5	4	3	2	0	-4	0	3	-20

[1] This means according to a uniform distribution (we come to that concept soon). Note that computers in general use pseudo-random sequences, but for our purposes these are a good enough approximation of "true" random numbers.

0	1	2	3	4	5	6	7	8	9
-1	10	3	8	2	-1	-4	4	-6	-15
3	3	5	8	-12	1	2	-2	-11	3
4	2	6	-2	3	6	10	-10	-6	-13
-2	-2	-2	-1	-2	-4	3	13	6	-9
-7	-3	-7	5	-9	-3	13	0	19	-8
3	-6	2	11	6	-6	-7	0	-4	1

This already easier to distinguish. But actually, we don't care too much if the differences are positive or negative, and actually it would be useful to magnify differences as they get further away from 50. So, let's make those positive and magnify by taking squares:

0	1	2	3	4	5	6	7	8	9	Total
64	4	1	25	9	1	0	9	49	64	226
25	25	49	9	16	64	4	4	64	0	260
9	16	36	1	0	1	64	36	4	25	192
49	25	16	9	4	0	16	0	9	400	528
1	100	9	64	4	1	16	16	36	225	472
9	9	25	64	144	1	4	4	121	9	390
16	4	36	4	9	36	100	100	36	169	510
4	4	4	1	4	16	9	169	36	81	328
49	9	49	25	81	9	169	0	361	64	816
9	36	4	121	36	36	49	0	16	1	308

The sum of squared differences is a measure of how different the counts are from the expected count (50). The expected value is 0.

This table is too big. We need to summarise this data in a meaningful way (what we are doing here is arriving at a "statistic"). So let's sum up each experiment into a single value: the total of the square differences.

Thus, the totals of the squared differences from the expected count of 50 for 10 sets of 500 random digits again are:

226	260	192	528	472	390	510	328	816	308

For the first 500 digits of π the equivalent total is 334. Is this consistent with the totals obtained for random digits?

The number seems small – counts of the digits of π seem to be similar to that for uniform random digits.

To get a feeling if this is small, we need to look at the other extreme: something we know has more structure. So look at the digits of the fraction 1/81.

```
0.01234567901234567901234567901234567901234567901 2
```

```
34567901234567901234567901234567901234567901234567
90123456790123456790123456790123456790123456790123
45679012345679012345679012345679012345679012345679
01234567901234567901234567901234567901234567901234
56790123456790123456790123456790123456790123456790
12345679012345679012345679012345679012345679012345
67901234567901234567901234567901234567901234567901
23456790123456790123456790123456790123456790123456
79012345679012345679012345679012345679012345679012 3456
```

Here is the count:

0	1	2	3	4	5	6	7	8	9
56	56	56	56	56	55	55	55	0	55

And then we see the sum of the squared differences is 2780.

In fact for 1/81 this total is around 2.5 times the largest value obtained when looking at random uniform digits (at least for these 10 sets).

sum of squared differences = 2780

Is 10 samples enough?

We took 10 examples of the sum of squared differences for sets of 500 random numbers. Perhaps we we need more. Let's have a look.

Number of sets	Maximum squared difference
10	816
100	1030
1000	1486
10000	1788

So for 1/81 the total squared difference is still larger than largest from 10000 sets of random digits. But for π, it is well under.

This suggests that π is *consistent* with random digits, whereas 1/81 is not.

Summary

Let's go over what we just discussed.

- π is irrational
- The expected digit count for 500 uniformly random digits is 50 of each.
- Measuring the distance of the first 500 digits from these expected counts gives a number that is not unusual for randomly generated digits.
- Doing the same for 1/81 gives a very unusual distance.

- 1/81 is not consistent with random digits, π (probably) is.
- We haven't considered pairs of digits, or triples etc.
- We have only looked at the first 500 digits of π.

In this lecture we looked at *one* aspect of whether the digits of π appear random: Uniformity of the distribution of single digits. We computed the distance of digit frequencies from an expected set of frequencies under uniformity, and compared that distance to distances similarly achieved using known random digits. If the distance is comparable, there is no evidence that the digits of π are not uniformly distributed.

All of this is rudimentary *hypothesis testing* which you may have seen before. Regardless, we are going to get to that in future lectures.

Lecture 2

Introduction to probability theory

Consider a single fair coin toss. Fair here refers to the fact that the coin has an equal chance of landing on heads or on tails.

We would say that the coin then has a 50% chance of landing on heads, and a 50% chance of landing on tails. Alternatively, we can say that the probability of heads is 0.5, and of tails is 0.5.

What can we say about probabilities? They represent the chance of an outcome occurring out of a set of outcomes. Throughout this lecture we will discuss examples of such systems, known as probability spaces. First though, we build up the concept slowly.

Let's get to some (hopefully) obvious facts about probability.

Probabilities lie between 0 and 1, 0 signifies that the outcome cannot occur, while 1 indicates it would always occur[1].

The sum of probabilities over all possible outcomes equals 1.

We often employ probability concepts in everyday without making those explicit. We are going to capture those intuitive properties in what is known as a probability space.

Here are some example events:

1. A single fair coin the possible outcomes are H/T, with probability 0.5 for either.
2. 3 coin tosses: the number of heads can be outcomes are 0, 1, 2, and 3.
3. A dice has possible outcomes are 1, 2, 3, 4, 5 and 6. If it's fair, then a good model is that all outcomes are equiprobable, with with probability of 1/6 for each number.

[1] To be pedantic, there are some quirky exceptions to this in infinite space – but we don't worry about that in this unit.

4. Suppose people get a disease at a rate of 10%. The possible outcomes are “get disease” and “no get disease”. One model is to say that the probabilities are 0.1 and 0.9 respectively.

The first one will be a Bernoulli random variable, the second a Binomial random variable. The fourth is also a Bernoulli random variable. So what do we mean by “random variable”?

2.1 Random Variables

A **random variable** is a mathematical concept that models a measurement or observation. It is a very powerful concept that allows us to distinguish between the theoretical abstraction of a model, and from measurements.

Random variables are usually denoted by capital letters, say X. So X could represent the number of heads out of 100 tosses. Then we would say the X has a Binomial distribution.

Observed values of X can be labelled $x_1, x_2, \ldots, x_N$.

Random variables can be described then by their probability distributions.

Bernoulli random variables: Think of a single fair coin toss modelled by a random variable X. Then

$$P(X = H) = 0.5$$
$$P(X = T) = 0.5.$$

If a coin were biased, say heads coming up 60% of the time, then we’d have that

$$P(X = H) = 0.6$$
$$P(X = T) = 0.4.$$

We can now generalise this to get a Bernoulli random variable:

$$P(X = H) = p$$
$$P(X = T) = 1 - p.$$

Note that p is a parameter of the Bernoulli random variable: change it and the distribution changes.

Binomial random variable: What happens when we want to count the outcomes of many coin tosses, how many heads or tails?

Suppose a fair coin is tossed 3 times. The eight equilikely possible outcomes are TTT, TTH, THT, HTT, THH, HTH, HHT, and HHH.

The possible number of heads outcomes are 0, 1, 2, and 3. There is only one way to get 0 or 3 heads, but there are 3 ways to get a single heads, and 3 ways to get 2 heads.

Head count	0	1	2	3
Combinations	1	3	3	1

If we toss a coin four times, the number of ways to get X heads is:

0	1	2	3	4
1	4	6	4	1

Since tossing a fair coin has a fifty-fifty chance of heads/tails each of the outcomes is equilikely. This means that for three coin tosses:

- the chance of 0 heads is 1/8,
- 1 head 3/8,
- 2 heads 3/8 and
- 3 heads 1/8, or

0	1	2	3
0.125	0.375	0.375	0.125

These numbers are **probabilities** of each outcome.

We'd like to generalise to any number of coin tosses (without having to simulate each single one). Calculating these tables each time is very time consuming. Fortunately, we can compute these counts when tossing a large number of coins directly. So we can calculate the probability distribution of the outcome of n coin tosses (more generally n Bernoulli trials). For simplicity, we will write down its probability distribution (for a proof, please see a standard textbook).

The probability of k heads from n coin tosses, where the coin has a probability of p of being heads is given by:

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

where

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

$\binom{n}{k}$ is from combinatorics, and is called "n choose k". $n!$ is the product of all the integers from 1 up to n.

A binomial model, allows us to answer questions such as:

- If a coin is tossed 10 times what would be the average **proportion** of heads?
- If a coin is tossed 100 times what would be the average **proportion** of heads?

- If a coin is tossed 10 times how likely is it you would see all heads?
- If a coin is tossed 100 times how likely is it you would see all heads?
- If a coin is tossed 100 times how likely is it you would see more (or less) than 10 heads?
- If each of 100 people have a 10% chance of getting a disease, how likely is it that no-one gets the disease?

A Binomial experiment has the following conditions:

- We have n independent events.
- Each event has the same probability p of success.
- We are interested in the number of successes from the n (Bernoulli) trials.

For example, suppose the are n individuals, and each has a probability of p of having a disease (or some other attribute). The probability that k out of the n individuals has the disease is found using the Binomial distribution.

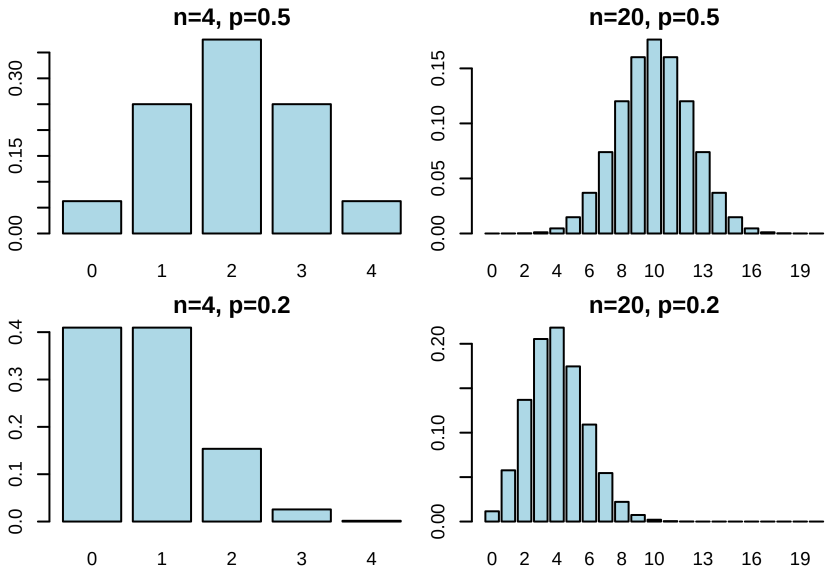
Let's look at a particular application.

Problem: Hard drives of a certain brand have a 10% failure rate after one year of use. If we run three of these hard drives for a year, what is the probability of 0, 1, 2, 3 and 4 hard drives failing after a year?

First we can see that the probability of k failures can be modelled by the Binomial distribution:

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

Using this distribution, we can look at the probability of failure, given different numbers of hard drives (varying n), and different individual probability of failures:



Notice that the probability of getting at least one hard disk failure can be higher than you might expect. For example if there is a 20% chance of any one drive failing, then there is a 60% chance that at least one drive will fail.

Note that this is a useful model, which makes some simplifying assumptions.

Lecture 3

Probability models

In the previous lecture we discussed coin tossing. How one coin toss can be seen as the outcome from a Bernoulli random variable, and how counting numbers of head (or tails) can be seen as the binomial distribution.

In this lecture, we will first define a probability space. Every random variable has a mean and variance. You'll have heard about these before most likely. It will just be a bit more formal here. Later, when we get to into statistics, we will look at empirical estimates of these parameters – that just means that we will use the data to calculate estimates of these parameters.

We will then talk about a few other useful distributions that are going to be used in this unit.

For those of you who wish to go into more depth, the following book is recommended: "Introduction to Probability Models", Sheldon Ross (2014).

I know you may be in a rush to get to data. If you are, bear with us because getting the models right first helps a lot with the data.

3.1 Probability spaces

A probability space consists of three elements:

1. a sample space Ω: all possible outcomes;
2. an event space $\mathcal{F}$: all possible events;
3. and a probability function P: a mapping of an event to a value in $[0, 1]$.

For example consider a single fair coin toss (Bernoulli with $p = 1/2$). The sample space is heads and tails: $\Omega = \{H, T\}$ (we can get either heads or tails).

The event space $\mathcal{F}$ includes H and T but also the event "heads or tails" written $H \cup T$. All together we get $\mathcal{F} = \{\{H\}, \{T\}, \{H \cup T\}\}$.

The probability measure maps those events
$P(H) = 0.5, P(T) = 0.5, P(H \cup T) = 1$.

For two coin tosses, the outcomes are $\Omega = \{HH, HT, TH, TT\}$. The events then are then the combinations of all possible outcomes. For example the event “2 heads or 2 tails” is given by $HH \cup TT$. And the probability function for that event is $P(HH \cup TT) = 1/2$.

Expected value and variance

Now we can define a random variable X a bit more formally. It’s the mapping from the sample space to a (real) number – $X : \Omega \mapsto \mathbb{R}$.

A random variable has a probability $P(X = e)$ for each event $e \in \mathcal{F}$.

We would like, in some sense, to define a central value of X. For example, for n coin tosses, what is the typical number of times we will see heads? To know what the average number of heads we’d see.

Let X then be the number of times heads comes up. We can then calculate the *expected value* of X, written $E[X]$. Suppose we had 3 coin tosses. X can take on values $0, 1, 2$ or 3.

$$E[X] = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3)$$

As we know the binomial distribution, we could calculate all of this (you will in a worksheet). Turns out one can show the expected value for any binomial random variable defined by p and n has

$$E[X] = np.$$

What this means is that if you did a lot of experiments with say n coins, and you looked at the average, it would converge to np.

Another important aspect is to calculate how much the random variable deviates from this central expected value. Hence we define the *variance* of a random variable as $E[(X - \mu)^2]$, where $\mu = E[X]$.

Without going into detail (we will revisit variance), the binomial variance is given by $np(1 - p)$.

3.2 Poisson distribution

The first thing to say is that poisson means fish in French. That’s not very relevant to its properties: it is actually named for Siméon Denis Poisson, a French mathematician. But it might help you remember the name. In the next section we will talk about being kicked to death by a horse. That may seem like an odd topic but it’s historical, and relevant.

The binomial distribution has non-zero probabilities for the integer numbers zero through n, where n is the number of (Bernoulli) trials (if I toss a coin five times, the probability of getting six or more heads is zero).

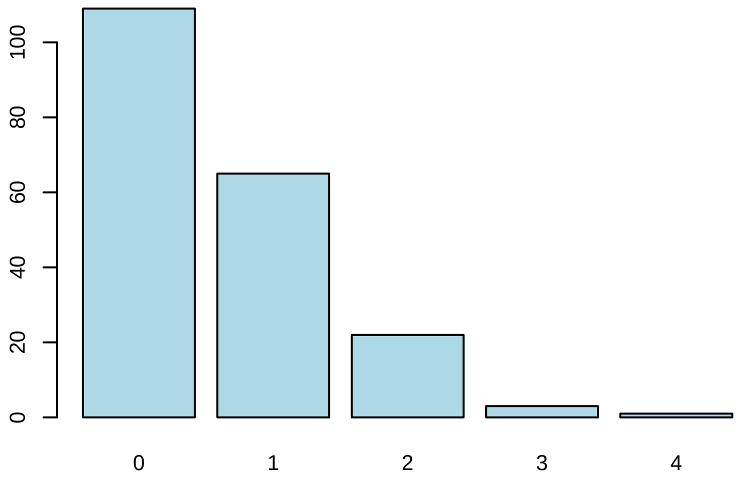
It is useful for examining the probability of a number of successes out of a given number of trials.

But what if instead we want to examine the count of occurrences of an event (e.g the number of cars that pass though an intersection, the number of goals scored in a football game)?

Being kicked by a horse to death

In 1898, von Bortkiewicz published the number of deaths from horse kicks in the Prussian army for 10 regiments over 20 years (200 observations).

Here's the data summarised in a bar chart:



It's divided up into number of deaths from horse kick for each regiment and year. These data are counts – they take values 0 through to 4 (in this case).

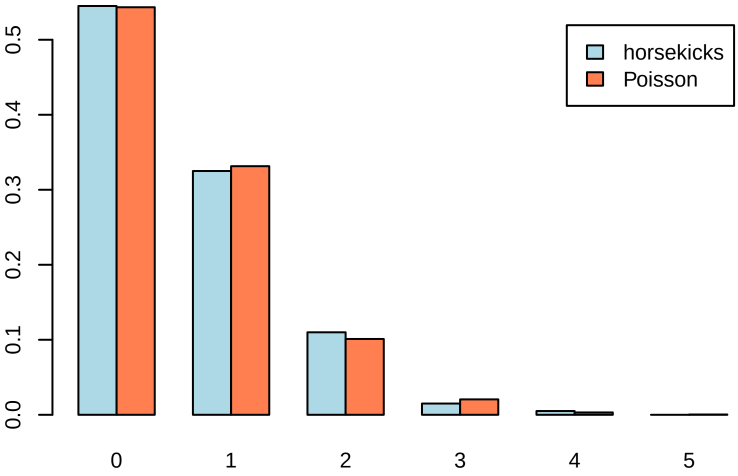
It doesn't lend itself to a binomial distribution. The data are not the result of a sequence of kicked/not kicked trials. There is no maximum number of occurrences either (no n).

For this data, von Bortkiewicz proposed the **Poisson** distribution. For this distribution the probability of observing k kicks is

$$P(k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

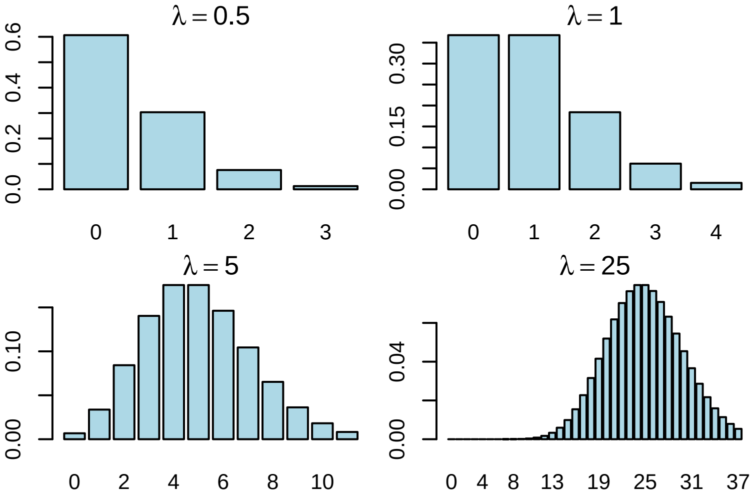
where λ is the expected number of kicks, and e is a mathematical constant ($e \approx 2.7182818$).

The number of deaths from horse kick compared to the Poisson distribution is shown here:



Poisson Probabilities

Examining the Poisson distribution:



The Poisson model has only one parameter, which we call λ.

Expected value of a Poisson variable The expected value of the Poisson model is λ (the mean if we could take a very large sample with the same λ).

Variance of a Poisson variable It turns out the **variance** is also λ.

Horse-kick data

For the horse kick data we observed 200 counts, tabulated:

0	1	2	3	4
109	65	22	3	1

.

So an estimate of the mean number of kicks per regiment per year is

$$(109 \times 0 + 65 \times 1 + 22 \times 2 + 3 \times 3 + 1 \times 4)/200 = 0.61.$$

3.3 Binomial and Poisson

Binomial Examples:

- Number of geminating seeds out of a batch.
- Number of people with disease.
- Number of insects killed with certain pesticide dose.

Expected value: np

Variance: $np(1-p)$

Poisson Examples

- Number of car insurance claims.
- Number of plants in an area.
- Number of people waiting in a queue.

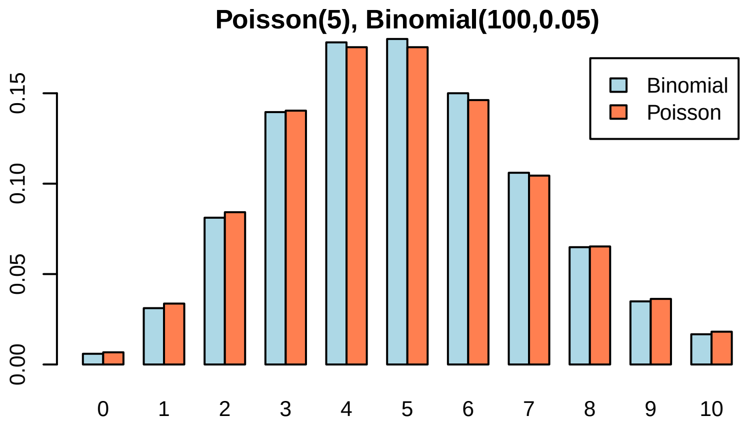
Expected value: λ

Variance: λ

Poisson approximation to binomial

It turns out that for large n and small p the binomial is close to a Poisson with $\lambda = np$. Note that $np(1-p) \approx np$, since $1-p$ is close to 1.

Let's compare binomial and Poisson distributions:



3.4 Normal distribution

Astronomy was one of the first sciences to make accurate measurements. Galileo and later Carl Friedrich Gauss, noted that errors in astronomical measurements were:

- Evenly positive and negative,
- More often close to zero,
- Declined in frequency the further away from zero they were.

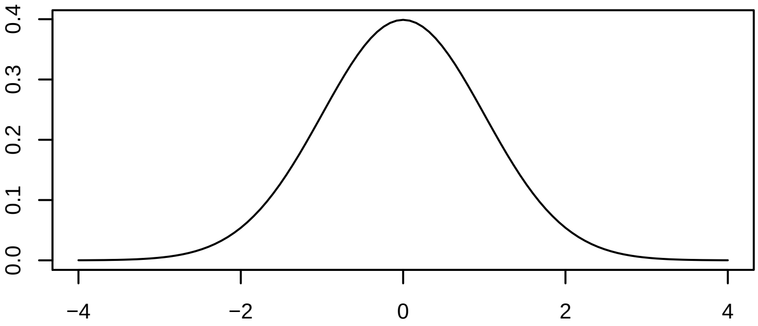
Gauss proposed the normal distribution as a **model** for these errors – hence also known as the Gaussian distribution.

X is a normal random variable with mean μ and variance σ^2, whose probability density function is

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

We then write that $X \sim N(\mu, \sigma^2)$.

With mean 0 and variance 1, you get the *standard normal* distribution:



The set of possible values for x are the real numbers, and continuous (contrast this with the binomial distribution which has on discrete values).

Thus there is an issue with interpreting $P(X = x)$ for any x. For example, suppose we model height by a normal distribution. It cannot, for example, be sensible to talk about the probability of having a height of exactly 178.42cm.

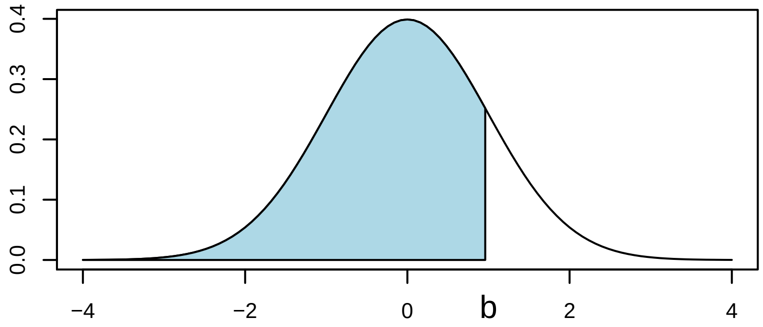
Instead, we can talk about the probability that $X > a$ or $X < b$. The probability that a student has height between 169 cm and 170 cm makes more interpretable sense.

Calculating normal probabilities in R

For a binomial we can talk about getting less than k heads etc. To do this, we add up the probabilities for all counts of heads less than k, since k is discrete. For a normal distribution this translates to areas.

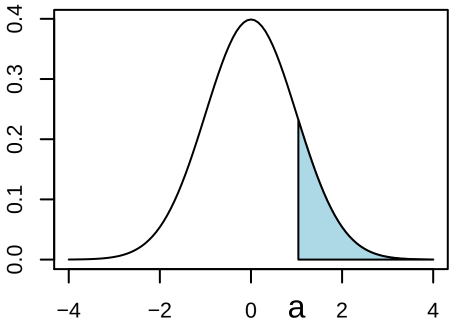
The probability that $X < b$ is the area under the normal density that is less than b.

Probability using area under the density:

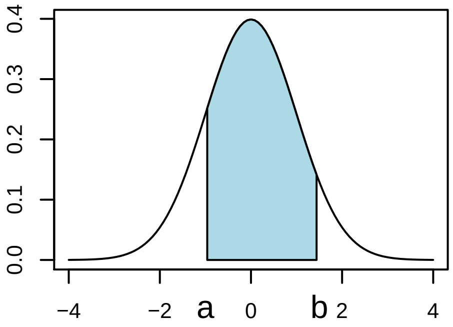


These probabilities can be tabulated for a range of x values. Also `R` computes them for us using `pnorm(b)`.

We can also compute probabilities like $P(X > a)$ and $P(a < X < b)$



$P(X > a) =$ `1-pnorm(a)`



$P(a < X < b) =$ pnorm(b) - pnorm(a).

3.5 Connection to binomial distribution

Suppose a fair coin is flipped 100 times. We'd like to know what the probability of getting 60 or more heads is.

Recall that the probability of seeing exactly k out of n "successes" (or heads etc) is given by the binomial distribution:

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

For a fair coin $p = 1/2$ this simplifies to:

$$P(X = k) = \binom{n}{k} \frac{1}{2^n}.$$

Try to see if you can show this as it will help with understanding.

So to calculate the probability of getting 60 or more heads from 100 flips we get

$$\frac{1}{2^n} \sum_{l=60}^{n} \binom{n}{k}.$$

de Moivre (18th century statistician) was often called upon to do calculations like this. He discovered a smooth curve very closely fits the binomial probabilities.

This was particularly significant at that time as calculating the above equation by hand was painful.

Here are a few examples:

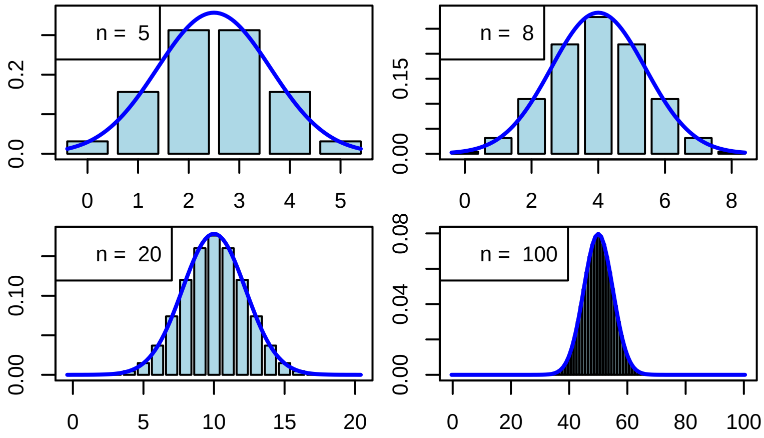


Figure 3.1: Approximate binomial distribution.

As n gets larger the approximation gets more exact. The curve that approximates it is the normal distribution.

Nowadays with computers it is less significant to approximate such sums. However, depending on the application, it can still be significant, and most importantly, the approximation tells us something about the centrality of the normal distribution, which we will discuss later.

The approximation works for any probability p –for example $p = 0.3$:

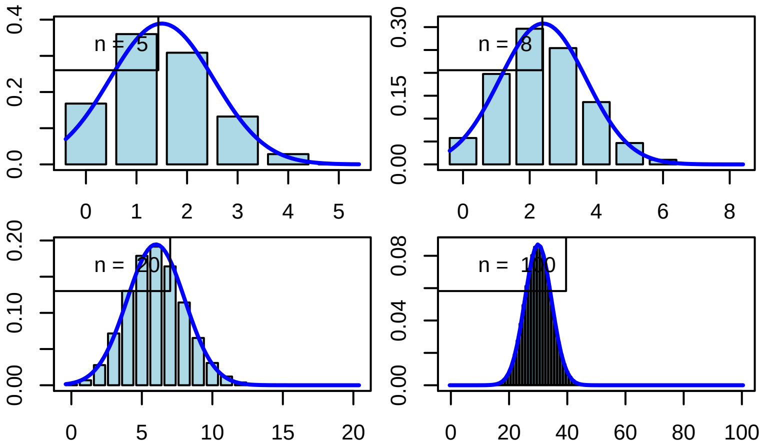


Figure 3.2: Approximate binomial distribution.

This means that calculations like the probability of 60 or more heads, can be converted to areas under the curves.

Let's see how to mesh the two distributions.

The set of values is continuous unlike in the binomial case.

For the 100 fair coin tosses, the mean $np = 100(0.5) = 50$. The variance is $np(1 - p) = 50(0.5) = 25$.

A good guess is to set the normal random variable we are using to approximate this binomial with the same mean and variance. This amounts to setting $\mu = np$ and $\sigma^2 = np(1 - p)$.

Let X be this approximation, we want to compute $P(X > 60)$.

We can then use R and the normal distribution to calculate the probability.

$$\texttt{pnorm(60, 50, 5, lower=FALSE)} = 0.023$$

It turns out the approximation can be improved by using a correction for the continuity. This involves approximating $P(B > k)$ with $P(X > k + 1/2)$, and $P(B < k)$ with $P(X < k - 1/2)$

R can also compute binomial probabilities with `pbinom`. So we can check how good our approximation is.

```
> n = c(5,10,100)
> x = floor(n*0.6)
> pb = 1-pbinom(x,size = n, prob = 0.5)
> pn = 1-pnorm((x+0.5-n/2)/sqrt(n/4))
```

n	x	pb	pn
5	3	0.1875	0.1855
10	6	0.1719	0.1714
100	60	0.0176	0.0179

Lecture 4

Independence, conditional probability, Monty Hall; and refugees

This lecturer introduces the concept of conditional probability. We then look specifically at how to obtain values for Monty Hall.

We also will look at parameter estimation from probability distributions.

First though we return to an important distribution that we implicitly talked about when looking at the digits of π.

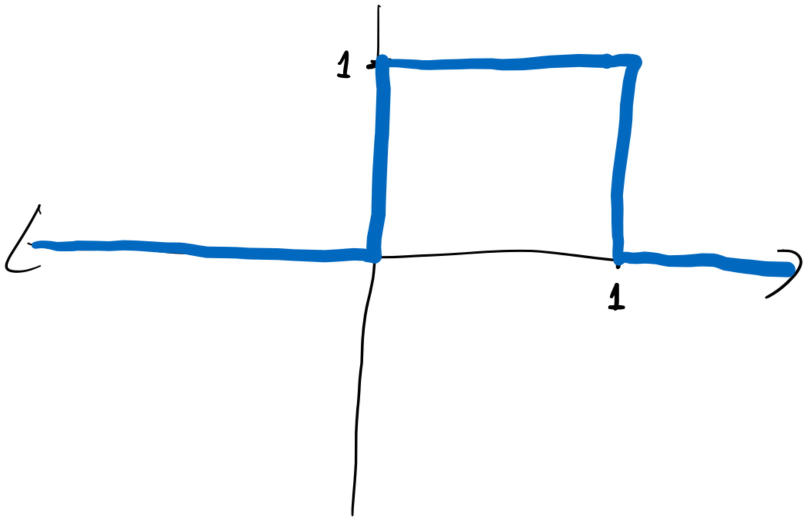
4.1 Uniform distribution

We talked about the uniform distribution in the first couple of lectures. We tried to ascertain if the digits of π were random: what we really meant is if the digits were *uniformly* distributed: each number $0, \ldots, 9$ being equilikely.

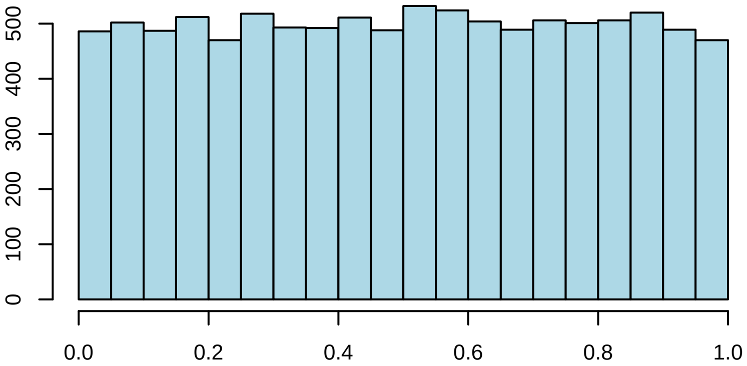
That was the *discrete uniform distribution* – discrete because it can only take on a few discrete values.

There is also the continuous version: when values are uniformly distributed over an interval. For simplicity, we can reduce that to look at the interval $[0, 1]$: any real number between 0 and 1, with both included. This is because any other uniform distribution between, say $[a, b]$, can always be scaled and translated back to the $[0, 1]$ case.

The uniform distribution has a probability density function:



We can simulate the uniform distribution using R:



That was a simulation of 10,000 random variables from a uniform distribution. Each of the bars in the histogram is roughly 500. There is random variation, but everything is pretty close. This is to be expected as 10,000 is a very large number and the variation in the results should be small.

4.2 Independent events/variables and conditionality

Suppose you have two events E and F from a probability space. The natural question to ask is what the probability of both events occurring is, and if one depends on the other: in other words what is $P(E \cap F)$?

For example, suppose we roll a (six-sided) dice twice. Normally a good model is that the probability of the first roll being a 6 is independent of whether you get a 5 in the second roll.

In general, we say E and F are independent if

$$P(E \cap F) = P(E)P(F).$$

If the dice is fair, with all rolls $1, \ldots, 6$ equilikely, then

$$\begin{aligned} P(5 \text{ in roll } 1, 6 \text{ in roll } 2) &= P(5 \text{ in roll } 1)P(6 \text{ in roll } 2) && (4.1)\\ &= 1/6 * 1/6 = 1/36. && (4.2) \end{aligned}$$

Notice that value, 1/36, is the probability of getting any particular combination of rolls: the number 5 and 6 don't play any role in the calculation. The values are uniformly distributed (discrete version).

The rolling of these dice can be written quite naturally using random variables. We say that X and Y are independent random variables if $P(X \cap Y) = P(X)P(Y)$ – in other words: the probability of X and Y occurring at the same time is simply the product of the respective probability of X and Y occurring.

And they are independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$.

Now suppose X = the number achieved in roll 1, and Y = the number achieved in roll 2. Then $P(X = i) = P(Y = i) = 1/6$, for any possible dice outcome $i = 1, \ldots, 6$.

Ok, that's independence. What's *dependence*?

A (popular?) gambling game is to bet based on getting particular sums of two dice: you first roll a 5, then a 6, and the total is 11. What's the probability of getting 11?

Let's attack the problem first a brute-force way. There are 36 combinations

$$(1,1),(1,2),(1,3),\ldots,(1,6);(2,1),\ldots,(2,6);\cdots;(6,1),\ldots,(6,6),$$

all just as likely.

These result in counts (check this yourself!) :

$$\begin{gathered} 2,3,4,5,6,7;3,4,5,6,7,8;4,5,6,7,8,9; \\ 5,6,7,8,9,10;6,7,8,9,10,11;7,8,9,10,11,12. \end{gathered}$$

There are only two cases of 11. So the probability of 11 a priori is $2/36 = 1/18$. What about 5? That appears 4 times. In general, it's harder to get higher numbers.

Come back to X, Y = number achieved in rolls 1 and 2 respectively. The total is given by $X + Y$ – that is another random variable, so let's give it a name. The total from two dice is $Z = X + Y$.

Observe that Z is not uniformly distributed: we just saw how 5 is more likely than 11 – and so on.

What's clear is that Z is neither independent of X nor of Y.

After the first roll you see some number $x \in \{1, \ldots, 6\}$. You then get to bet on whether you will get at least a certain total: for example 8. The amount you can win depends on the risk. Let's look at this risk.

Suppose the first roll is a 1. You've already no chance of getting 8. If it's a 2, you need a 6, so you've only a 1 in 6 chance. If you get a 3, your chances are getting better: a 5 or a 6 will get you over the hump. And so on.

So we formalise. As before, X = value on first roll, Y = value of second roll. We want to know probability $Z \geq z$ given that $X = x$, for some x and z.

This can be written as

$$P(Z \geq z \,|\, X = x)$$

and is said "the probability of Z being greater than or equal to z given that $X = x$. This is a *conditional probability.*

This sort of question is hopefully already getting you to see the relevance to the Monty Hall problem. Are the door picks independent? Or is there something else at play?

4.3 Conditional probability

Take two events E and F.

$$P(E \,|\, F) := \frac{P(E \cap F)}{P(F)}. \tag{4.3}$$

is the conditional probability that E occurs given that F has occurred.

Note that it is ill-defined if $P(F) = 0$.

If E and F are independent (see definition above) then

$$\begin{aligned} P(E \,|\, F) &= \frac{P(E)P(F)}{P(F)} \\ &= P(E). \end{aligned}$$

That makes sense: if then are independent events, then F occurring doesn't affect the probability of E occurring.

In the dice case, when we bet after the first roll, the probabilities of interest are $P(Z \geq z \,|\, X)$.

4.4 Independence vs uncorrelated

Correlation is an important topic we will cover later. It is concerned with estimating the expectation of products of random variables. If X and Y and random variables, then we would want to estimate:

$$E[XY].$$

X and Y are said to be uncorrelated if $E[XY] = E[X]E[Y]$.

This is a weaker condition than independence. As above, if X and Y are independent $P(X = x \cap Y = y) = P(X = x)P(Y = y)$, and it follows that $[XY] = E[X]E[Y]$.

The reverse is not true. They may be uncorrelated but dependent.

Many people confuse these concepts. I mention it now because it will be come clearer when we examine correlation: both the model, and estimating it using data.

4.5 Monty Hall: why you should change doors

Let's look, for the last time, at the Monty Hall game. If you forget it, refer back to the description in lecture 1.

Let the door containing the car be the random variable Y. The outcomes possible are $1, 2, 3$.

Y's distribution is uniform: $P(Y = 1) = P(Y = 2) = P(Y = 3) = 1/3$.

For simplicity, let's assume you choose door 1 at the beginning. As the car is distributed randomly through the doors, the analysis holds equally if it were 2 or 3.

Monty then opens door D, where the outcomes are door 2 or door 3, to show a goat behind it. The distribution of D is: $P(D = 2) = P(D = 3) = 1/2$.

It feels like the probability of you winning is 1/3 no matter what you do. The car was put behind the door at the start. However, we have already seen in worksheet 1 through simulation that we win 2/3 of the time by changing. Now we do the analysis from a theory point of view.

If $Y = 1$ (car is behind door 1), Monty will pick either door 2 or 3.

$$\begin{aligned} P(Y = 1 \cap D = 2) &= P(Y = 1)P(D = 2) \\ &= 1/3 * 1/2 \\ &= 1/6. \end{aligned}$$

Similarly

$$P(Y = 1 \cap D = 3) = 1/6.$$

However, if the car is behind door 2 or 3, Monty has one door that he can open, namely door 3 or door 2.

$$\begin{aligned} P(Y = 2 \cap D = 3) = 1/3 * 1 = 1/3 \\ P(Y = 3 \cap D = 2) = 1/3 * 1 = 1/3. \end{aligned}$$

Given that Monty opens door 3, the probability to win by keeping door 1 is the conditional probability:

$$\begin{aligned} P(Y = 1 \mid D = 3) &= \frac{P(Y = 1 \cap D = 3)}{P(D = 3)} \\ &= \frac{1/6}{1/2} \\ &= \frac{1}{3}. \end{aligned}$$

Thus the probability of losing in by keeping door 1 is $\frac{2}{3}$.

We could show the same argument if the host choose door 2.

It is therefore twice as likely to win by switching. In summary:

$$\begin{aligned} P(\text{keep door and win}) &= 1/3. \\ P(\text{keep and loose}) &= 2/3. \end{aligned}$$

4.6 Iraqi Refugees

Finally, some statistics you may be saying. In Uribe Guajardo et al. Int J Ment Health Syst (2016), the authors looked at the level of distress in 443 Iraqi refugees.

The distress level was measured by a psychological instrument (known as the K10) and is classified as either low, moderate, high or very high distress.

For each refugee distress level, they had the following counts:

low	moderate	high	very high
123	70	93	157

Does this seem like a very distressed population? A priori you may think so. After all, we tend to think of refugees as being distressed: they had to leave their homeland, may have seen horrible atrocities at home, and are in a foreign land with an uncertain future. Remember though that the study seeks to find out if the statistics backs up these thoughts. We need evidence for this. Often we think certain things are related, when there is no evidence for it, and it may be false.

Secondly, when we look at the data without any reference point we're stuck. How do you know that this isn't how a typical population feels, with lots of business for psychiatrists ?

The answer lies in having a reference data set: something that represents "normality". For this study the used pre-obtained K10 information about 10,000 randomly chosen members of the Australian population. The percentages in each category were as follows:

low	moderate	high	very high
70.65	18.5	7.41	3.43

Now we have two data sets to compare. What we need now are the tools to do it. Some are visual tools to give us some indication. Ultimately, we will use a statistical tool to examine the data. The underlying question is now: does the distribution of refugees distress differ from that of the Australian population?

We are getting a bit more precise in our question, but we will need to formulate a bit tighter. First though, let's have a look at our dataset. Every good bit of data science starts with getting a feel for our data. In this unit you get a toolbox to perform data science, but their application is to some extent an art. Instead of paint brushes, you will learn statistical and visual techniques.

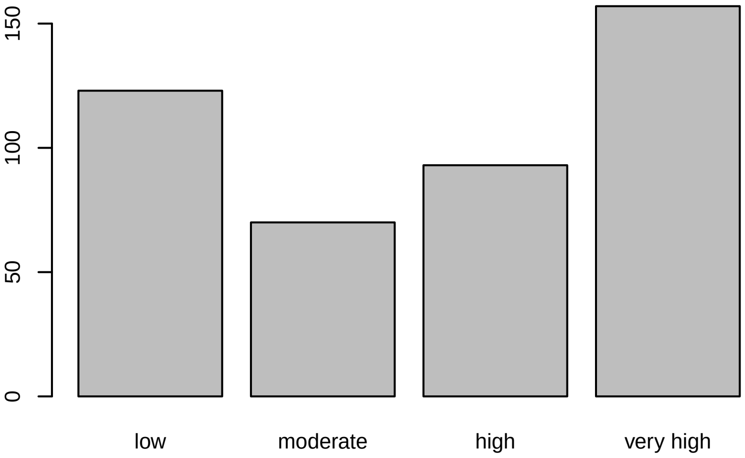
Bar plots

The first thing to do when confronted with data of this type is to appropriately plot it. Tables of numbers are harder to interpret in general. In this dataset it's not too bad, but if we had many categories it would quickly be very hard to get a feeling.

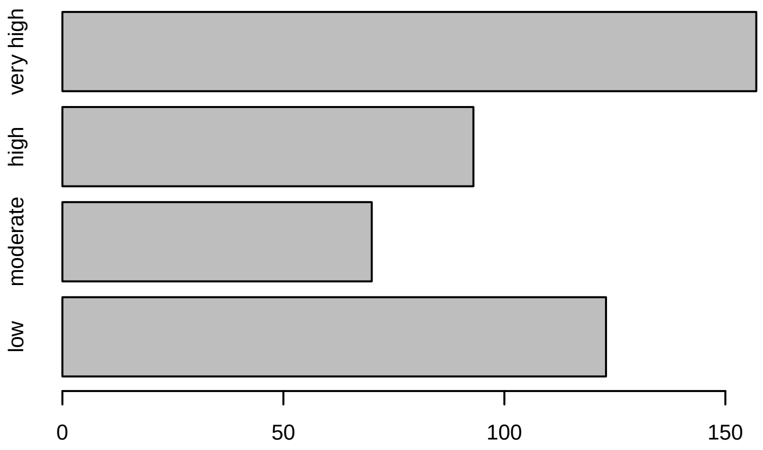
One useful plot here is a *bar plot*: we have categories, and numbers associated with each. In a bar plot, we draw per category a bar or box, whose height represents the count.

So, bar plots are ideal for visualising the counts associated to a set of categories. In this case, we have the count of people associated to each distress level category.

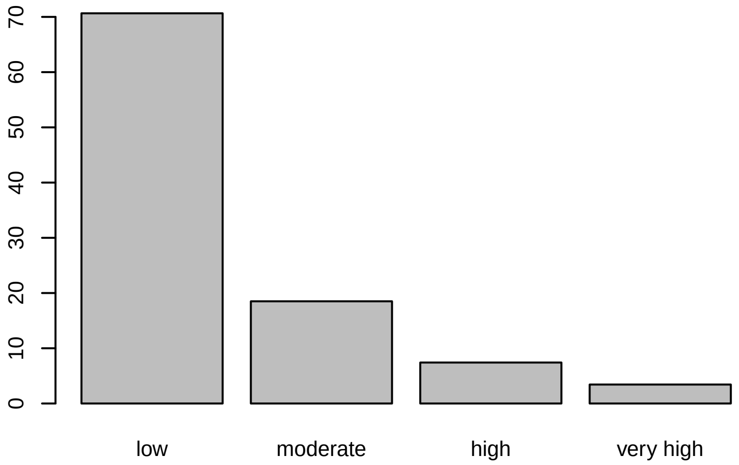
Here is the bar plot of the numbers of refugee per distress category:



Bar plots can also be drawn horizontally:



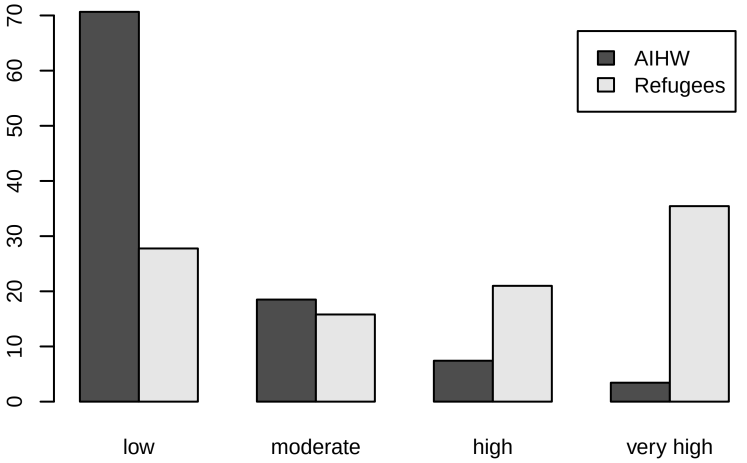
Here is bar plot for the AIHW data, the Australian reference population:



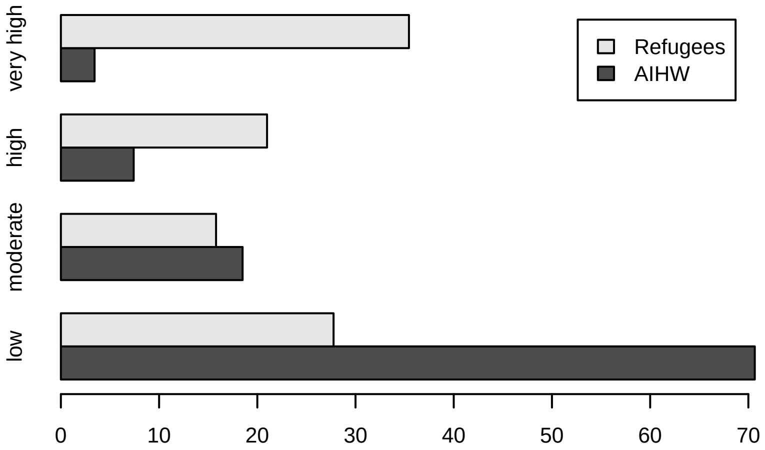
That helps a bit, but it would be useful to see them side-by-side. Do you already see an issue with that? The sample sizes are drastically different, and the AIHW reference data set has already been reduced to percentages.

Thus, we first need to convert the Iraqi refugee data set to percentages, by dividing by the total number of refugees whose distress was measured (443) and then multiplying by 100.

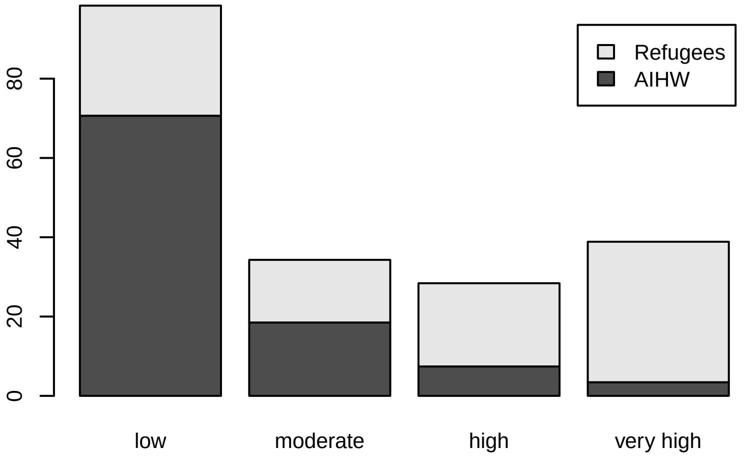
Now we can put both together, vertically so:



or also horizontally:



Stacked bars can also be used:



Let's get back now to the original question: does the distribution of refugee distress (significantly) differ from that of the Australian population?

Could we recycle how we did it for the digits of π? We would need to randomly allocate the 443 individuals to the four distress levels, then compare it to our sample. This would allow us to check if all distress levels are equally likely (the distribution is uniform).

But we are not interested in whether the distress level spread is uniform; we want to know if the distribution matches the AIHW percentages.

The method for examining the digits of π was not a bad start, but we need to generalise what we did there.

Overall the question can be re-written: what would the refugee distress distribution look like if it matched the AIHW distribution?

Instead of comparing to uniformly random distributions, we will use the AIHW reference data distribution.

Recall we took uniformly generated digits to see if the digits of π were consistent with it. This time we will simulate 443 individuals being allocated a category consistent with the AIHW percentages:

low	moderate	high	very high
70.65	18.5	7.41	3.43

For simplicity, we can round the numbers:

low	moderate	high	very high
71	19	7	3

We need a method to assign simulated individuals to these four categories in such a way that they reflect this distribution.

Suppose we have a uniform random variable Y over $[0,1]$ – that is, it can take any value in that interval uniformly.

Now we create a categorical random variable X from Y as follows:

$$X = \begin{cases} \text{low}, & Y < 71 \\ \text{medium}, & 71 \geq Y < 71 + 19 = 90 \\ \text{high}, & 90 \geq Y < 90 + 7 = 97 \\ \text{very high}, & \text{otherwise} \end{cases}.$$

This leads us to an algorithm. Let the computer give us a number y at random between 0 and 100 (this is a simulation of Y).

We then generate a category x from this number y by setting x to:

- low if x is less than 71

- medium if x is between 72 and 90 (=71+19) (inclusive)
- high if x is between 91 and 97 = (90+7)
- very high if x is greater than 97

When we have enough samples x, on average, 71 out of 100 will be low, 19 out of 100 will be medium etc.

Once we simulate the category counts, assuming that the AIHW percentages are true, we must ask "are the category counts from the sample (shown below) consistent with the simulated counts from the AIHW proportions?"

low	moderate	high	very high
123	70	93	157

Here are the simulation results repeated ten times:

low	moderate	high	very high
292	99	36	16
313	86	31	13
311	92	23	17
292	100	29	22
320	80	31	12
311	91	30	11
310	85	34	14
328	74	30	11
313	80	33	17
304	93	33	13

Notice as in the π case that these values do not exactly match the distribution but that there is some variation, consistent with randomness. It is important to emphasise this point: randomness allows variation. What we are trying to see is if some data is too "far away" from random to be just by chance.

We have 443 individuals in four categories, but they are not evenly spread. The AIHW has 70.65% in the low category. So the expected count in this category is $443 * 70.65/100 = 312.99$, since the sample size is 443.

Below are the remaining expected counts when using a sample size of 443:

	low	moderate	high	very high
percent	70.65	18.50	7.41	3.43
expected	312.99	81.96	32.84	15.20

So now we have generated a lot of simulated data. What we need now is a statistic to summarise the data, and see if a pattern emerges.

We will measure how much the category counts differ from their expected value. As we did for the π digits we can subtract the expected value from the actual counts and square and sum.

For the refugee data set we get = 59966.79. Look how this compares to the simulated data:

741.69	24.56	204.92	827.19	66.57	111.42	20.98	314.28	7.11	207.63

There is an issue though that did not occur with the π digits, because the simulated data had a uniform distribution from $0, 1, \ldots, 9$. Our square distance does not give enough **weight** to differences where there a small expected counts when compared to where there are large expected counts. This causes a distortion.

The fix for this is to take the count minus the expected count squared divided by the expected values (scaled), and then add up.

This leads us to:

$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

where O_i are the observed counts for each i, eg O_1 is the first observed count, and E_i is the corresponding expected count.

This formula is called the *chi-square distance* (or χ^2). In the next lecture we will look more closely at this distance.

4.7 Summary

In this lecture we looked at independence and conditional probability. We used these concepts to look at the Monty Hall problem. We also looked at our first big forage into statistics using the tools we are developing. In the next lecture, we will further develop those tools.

Lecture 5

Statistics, eels and refugees

In the last lecture, we saw how modifying the distance function used with the π digits to account for a non-uniform distribution lead to weighting each category differently, to a distance measure we call χ^2.

In this one, we are going go into more detail and come up with the χ^2 test.

We will spend some time discussing what a hypothesis is, look at a new dataset about eels and habitat, and introduce the (infamous) p-value.

We have been looking at categorical data up to now. At the end of the lecture, we'll look at our first quantitative data set.

5.1 The χ^2 test

In the previous lecture, we calculated the χ^2 statistic for comparing the refugee and Australian stress distributions.

We now want to use it to do what is known as a χ^2 test. To do this, let's first look at the χ^2 statistic (distance) from another vantage point.

Suppose we have a categorical random variable X which can be assigned to categories $c_1, c_2, \ldots, c_n$. For example, we had four categories for the AIHW (and refugee) dataset.

Let the probability of being in category c_i be given by p_i.

In the case of the π digits we knew the values p_i explicitly, as we used a uniform distribution as our reference point. The probability of getting any of the 10 digits was equal. That meant $p_1 = p_2 = \cdots = p_9 = p_{10} = 1/10$.

In general though, we may not know the precise values p_i, and need to estimate them from data.

Take the Australian (AIHW) dataset. This assigned the values $q_1 = 0.7065, q_2 = 0.185, q_3 = 0.741, q_4 = .0343$ for each of the four categories.

We now essentially use these numbers as estimates for p_i. This is justified because of the large dataset (10,000 samples)[1]

The hypothesis is that the Iraqi refugee population has a statistically different distribution to that of the Australian population. We ultimately want to know if our data is consistent with this hypothesis. In hypothesis testing, we call this the *alternative hypothesis*, labelled H_1.

To see if we can accept the alternative hypothesis, we look to see if evidence is there to be reject the *null hypothesis*, labelled H_0. For the data at hand, the null hypothesis is that the Iraqi refugee population is not statistically different to the stress distribution from the Australian reference population. We then try to see if the evidence from the data is strong enough to reject this null hypothesis, and thus accept the alternative one.

It can be helpful to think of this of like a court where a judge or jury tries to ascertain whether an individual is guilty of a crime. The data we have is the evidence. The null hypothesis in most codes of law is that of individual is innocent until proven guilty. If the evidence allows us to reject, beyond all “reasonable” doubt, then we accept the alternative hypothesis of guilty.

In hypothesis testing, we thus start off by assuming the null hypothesis is true, and see how it plays out.

We now have the tools to define the null hypothesis more formally. Say that the overall refugee population has a distribution into the four categories of r_1, r_2, r_3, r_4. Note that this is the model of the population as a whole, not just the sample of the 443 refugees whose stress levels were assessed.

Then the null hypothesis becomes:

$$H_0 : p_1 = r_1, p_2 = r_2, p_3 = r_3, p_4 = r_4$$

The alternative hypothesis becomes then:

$$H_1 : r_1, r_2, r_3, r_4 \text{are unrestricted.}$$

The χ^2 statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population.

The steps we take now are. Assume the null hypothesis. Measure the χ^2 statistic for the refugee dataset. The data again is:

	low	moderate	high	very high
refugees	123.00	70.00	93.00	157.0
expected	312.99	81.96	32.84	15.2

[1] More formally we could take into the account the accuracy of these as estimates too, but for simplicity we will take the model parameters directly as $p_i = q_i$.

and thus:

$$\frac{(123-312.99)^2}{312.99}+\frac{(70-81.96)^2}{81.96}+\frac{(93-32.84)^2}{32.84}+\frac{(157-15.2)^2}{15.2}=1550.08.$$

Then we perform, say 10,000, simulations to calculate 10,000 different χ^2 statistics. We now need to see how those compare to the refugee dataset distance of 1550.08. Simulation returns a lot of values using the base distribution from the AIHW set, thus assuming the null hypothesis were true. A χ^2 statistic will be small if the numbers returned are close to the expected values. Larger if they deviate a lot.

We are most interested in those that deviate, so see if the refugee value would fit in within those.

These are the maximum values for a series of ever-increasing numbers of simulation:

Number of sets	Maximum chi-squared difference
10	8.54
100	10.79
1000	16.72
10000	21.68

The distance for refugees is much larger than any of these. This indicates that Iraqi refugee distress levels are probably not the same as the Australian population. The null hypothesis can be safely be assumed to be very unlikely. We thus accept the alternative hypothesis: the refugee stress levels vary from the general population.

Note that here the case was pretty clear-cut. Even looking at the original histograms we could see how vastly different they seemed. Such strong conclusions are not always possible. Sometimes we need allow for some of the simulations being more extreme. We will come to that in our next example.

5.2 Eels - comparing two sets of counts

Consider two species of eels observed in three different habitats. The following counts were made:

	Border	Grass	Sand
G.moringa	264	127	99
G.vicinus	161	116	67

What we want to know is: is the distribution of eel species the same? Visualisation is of course helpful:

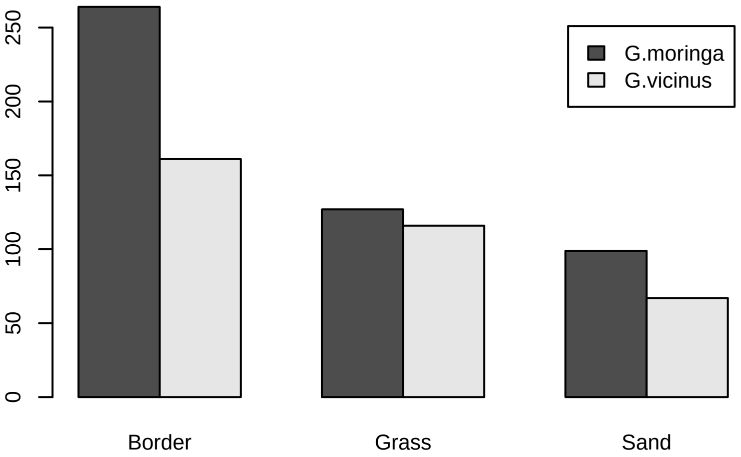


Figure 5.1: Eel counts at given habitat.

There are certainly fewer G.vicinus overall, and a lot fewer G.vinicus in the border habitat.

We actually do not care particularly about the overall number. That could just be some sampling bias. What is interesting is how they are spread between habitats: in other words the proportions in each habitat.

Proportion of sightings in each habitat

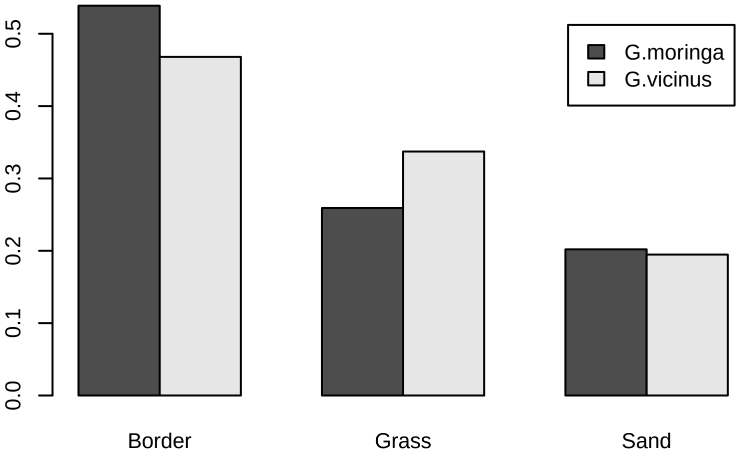


Figure 5.2: Eel proportions at given habitat.

Determining if the distribution of eel species could be the same isn't the same as in the Iraqi refugee case. There we asked if the distress level of refugees were different from the Australian population. And we essentially knew the average distress levels for the Australian population from a very large sample.

Here, neither set of counts is all that large. There is no reference distribution.

Instead, we will consider the aggregate of both species together, all 834 eels, and their habitat distribution. In other words, we will ignore the species label.

To work as before we have to find an expected value for each count and a way to simulate counts, under the assumption that the species distributions are the same.

If the species distributions are the same, then we can aggregate the counts across species to estimate an overall habitat distribution:

	Border	Grass	Sand
counts	425	243	166
proportion	0.5096	0.2914	0.199

Let the proportions be $p_1 = 0.5096$, $p_2 = 0.2914$ and $p_3 = 0.199$.

TODO: what would have happened if we took the same one at base value – what is H_0 etc.?

So if the distribution is the same we expect to see the same percentage of each species count in each habitat.

If there are $n_1 = 490$ eels of species G.moringa, then we should see $n_1 p_1$ in the Border habitat and $n_1 p_2$ in Grass where p_2 is the proportion in Grass etc.

For G.vinicus, these are $n_2 p_1, n_2 p_2$ and $n_2 p_3$ respectively.

The data is:

	Border	Grass	Sand	Total
G.moringa	264	127	99	490
G.vicinus	161	116	67	344

and thus the expected counts are:

	Border	Grass	Sand
G.moringa	249.7	142.77	97.53
G.vicinus	175.3	100.23	68.47

We need to simulate what the counts would look like if the two eel species shared the same distribution across habitats.

To do this, for each of the 834 eels, we sample a species using the proportions from the data:

G.moringa	G.vicinus
0.58753	0.41247

and a habitat using the proportions from the data:

Border	Grass	Sand
0.5095923	0.2913669	0.1990408

We then proceed by calculating the χ^2 distance from expected for our actual data and for a (large) number of simulations and compare.

Previously we were comparing to an (essentially) fixed set of expected values, but now we used the proportions in the data to compute them.

So we must recompute the expected values for every simulation

The null hypothesis is that "the distribution of eels across habitats does not differ". The converse, that these distribution do differ, is the alternative hypothesis.

χ^2 distance for simulated data

One simulation provided the counts below.

	Border	Grass	Sand
G.moringa	267	143	101
G.vicinus	173	86	64

In general, we do many simulations:

Number of sets	Maximum chi-squared difference
10	6.08
100	10.71
1000	14.23
10000	17.62

The chi-squared statistic for the actual eel counts is 6.26. That's bigger than the maximum from 10 simulation runs, but not for 100.

Does this mean that the null hypothesis cannot be rejected? Rather than being hasty, let's look at that in more detail.

After 1,000 simulations there are 46 greater than 6.26. The eels are not further than all simulations, but they are further away than most of them. It only happens 4.6% of the time.

This is called a *p*-value.

The *p*-value is the chance or proportion of the time that we would see a chi-squared distance as large or larger than the actual distance for the data, if we simulate assuming the null hypothesis is true.

We are going to define it more precisely in the next lecture.

5.3 Summary of hypothesis testing

So in summary, the hypothesis testing process has been so far:

1. Compute a summary statistic (mean or χ^2) of the sample data.
2. Generate many simulations of the data given that the null hypothesis is true.
3. Compute the summary statistic of each simulated data to obtain a distribution of summary statistics given that the null hypothesis is true.
4. Compare the data statistic to the simulated data statistic.
5. If they look different, then it is likely that the null hypothesis is false.

In a later lecture, we will show an alternative to simulation here, using what is known as the χ^2 distribution. For the moment let us stick to simulation and worry about that later.

5.4 Numerical data (maternal smoking)

All datasets we looked at so far were categorical: everything was assigned a category. We are now going to introduce a different sort of dataset, which is numerical. We will examine this case in detail in the next lecture.

Data (from the US) had the birth weights of more than 1,200 babies and the smoking status during pregnancy of the mother. An extract of it is:

bwt	smoke
3429	no
3229	no
3657	yes
3514	no
3086	yes
3886	no

The underlying question is does smoking affect the birth weight of infants?

This data consists of two **variables**:

- `bwt`: the birth weight of the infant in grams
- `smoke`: the smoking status of the mother

The variables are quite different in nature. One is a continuous measurement of weight, theoretically taking any positive value. The other takes only two values “yes” and “no”.

The qualitative variable `smoke` looks like:

no	yes
742	484

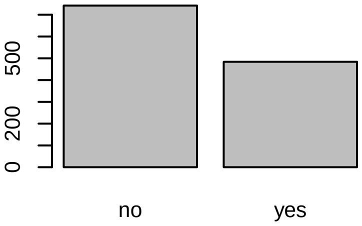


Figure 5.3: Count of smokers and non-smokers..

The quantitative variable bwt has a significant value: its mean. The mean birthweight is 3.466 Kg. We will look at the mean, and similar statistics in the next lecture.

Lecture 6

Quantitative data: Maternal Smoking Data

6.1 Probability vs statistics

Prompted by an observation from a student, I realised there might be some confusion about what is probability, statistics and the difference between the two[1]. So let's be explicit.

Probability:

1. Process of interest conceptualised as a probability model.
2. Use model to learn about probability of potential outcomes.

Statistics:

1. Process of interest conceptualised as a probability model.
2. Data viewed as observed outcomes from model.
3. Use outcomes to learn about the model.

So we can see clear important differences here.

Somebody working in probability can be seen as doing this job:

Job of the probabilist

Given a probability model P on a space Ω find the probability $P(A)$ that the outcome of the experiment is A.

while the statistician job has a different flavour:

[1] Credit to Victor Panaretos for the explanation here.

Job of the statistician

Given an outcome of $A \subset \Omega$ – the data – of a probability experiment on Ω, tell me something "interesting" about the (unknown) probability model P that generated it.

Some interesting questions are:

1. Are the data more more consistent with one or another model?
2. Given a family of models, can we determine which model generated the data?
3. What range of models are consistent with a given set of data?
4. How to best answer these questions? (is there even a best way?)

Suppose we let $X_1, X_2, \dots X_{10}$ be random variables representing ten individual coin flips.

An appropriate model is that $X_i \overset{iid}{\sim} Bernoulli(\theta)$. For example, if it's a fair coin model, θ would be 0.5.

One outcome recorded is then:

$$(0, 0, 0, 1, 0, 1, 1, 1, 1, 1).$$

The probabilist asks:

- Probability of outcome as function of θ
- Probability of k-long run?
- If keep tossing, how many k-long runs?
- What about the sum of observations? How does it behave? How does it scale?

while the statistician has different questions in mind:

- Is the coin fair?
- What is a good guess of the value of θ on the basis of the observations?
- What range of θ is plausible on the basis of the observations?
- How much error do we make when trying to decide the above from the observations?
- How does our answer change if the observations are perturbed?
- Is there a "best" solution to the above problems?
- How sensitive are our answers to departures from the model?
- How do our "answers" behave as num. of tosses go to ∞?
- How many tosses would we need until we can get "accurate answers"?

In short, the statistician is presented with the data and estimates a model, makes a hypothesis. The probabilist is concerned with the consequences of a model.

6.2 Quantitative dataset

Recall from the end of the previous lecture the maternal smoking dataset:

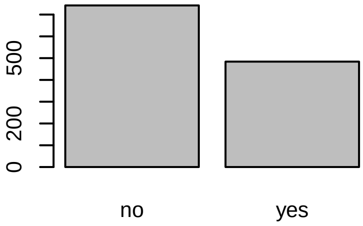


Figure 6.1: Count of smokers and non-smokers.

The underlying question is: does the data support the hypothesis that smoking affects infant birth weight?

The variable `bwt`, the infant birth weight, is quantitative. It is (essentially a) continuous measurement of weight, taking on theoretically any positive value.

In contrast, the variable `smoke`, the smoking status of the mother only takes on the value “yes” or “no”. It’s a qualitative variable.

The first thing to consider is how to look at the data.

6.3 Summarising data

We first examine the qualitative variable `smoke`. It is a qualitative (not quantitative) so the sensible thing is to tabulate the counts.

no	yes
742	484

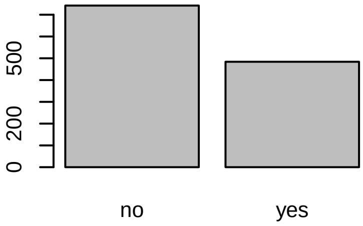


Figure 6.2: Count of smokers and non-smokers..

Numerical summaries of data

We next examine birth weight (bwt). When dealing with quantitative data, we should first get a feeling for the centrality of the data: a number (statistic) or numbers to represent effectively the data. This value is usually called the **average**.

The term average is used in many contexts and has no precise meaning. There are two more precise words **mean** and **median**.

Mean

The *sample* mean of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is given by

$$\frac{x_1 + x_2 + \cdots + x_n}{n}.$$

For example, the mean of $\{4.8, 5.2, 3.9, 5.3, 3.8\}$ is $23/5 = 4.6$.

The mean is (usually) in the middle of the data. It is called a measure of location since it is used to determine where the centre of the data lies.

To be precise we talk about the sample mean as opposed to the population mean (more on that later). It is also sometimes called the empirical mean.

The mean birth weight from our data is 3,466.83g.

Another measure of location is the median.

Median

The median of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ for n odd is a number such that half the data is greater and half less.

For $\{4.8, 5.2, 3.9, 5.3, 3.8\}$, the median is 4.8, 2 are above and 2 below. It is the middle observation.

If the number of observations n is even, it is a little trickier. There are two contenders for the middle point. By convention we defined the median then as half way between the two middle observations. So for example, given the data $\{14.2, 9.1, 10.1, 8.1, 8.7, 12.2\}$ the median is 9.6.

The median birth weight is 3,471.5g. In this case, there isn't too much difference between median and mean. However, in general the mean can be affected by **outliers**, or unusual points.

If we change the largest value in $\{4.8, 5.2, 3.9, 5.3, 3.8\}$ to get $\{4.8, 50.2, 3.9, 5.3, 3.8\}$ the mean goes to 13.6, while the median is still 4.8.

Connection with expected value

Let's come back to expected value $E[X]$ of a random variable X. It seems similar to what we just defined as mean. We talked at the beginning of this lecture of the difference between probability and statistics, a distinction that is important here. The random variable represents the model. Suppose we have outcomes from the random variable labelled

$$x_1, x_2, x_3, \ldots, x_n.$$

Then we take the (sample) mean of these values to get

$$m_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

What we expect[2] is that, with high probability, m_n will be close to $E[X]$. Or more precisely show that as $n \to \infty$ then $m_n \to E[X]$.

Turning this around, we can say that m_n is an estimate for $E[X]$. How precise an estimate that is (how much confidence we have in that estimate) is something we will return to.

Spread of data

After a rough idea of the **location** of the data, the next level of quantifying how the data is is how spread out it is. The simplest way over how many values the data spans:

Range

The range of $\{x_1, x_2, \ldots, x_n\}$ is the largest value minus the smallest value or:

$$\text{range} = \max_i x_i - \min_i x_i.$$

The range of $\{4.8, 5.2, 3.9, 5.3, 3.8\}$ is $5.3 - 3.8 = 1.5$. The range is heavily effected by outliers: the range of $\{4.8, 50.2, 3.9, 5.3, 3.8\}$ is $50.2 - 3.8 = 46.4$.

The range of the birth weights is 800.

Variance

A usually superior measure of spread (in that it captures more insight into the data), is how far typically the data is away from the mean:

Variance

The variance of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ is the average of the square distance of the observations to the mean:

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

The first oddity I hope you noticed is that we divide by $n - 1$, not n. This is called Bessel's correction. The reason for it has to do with the ability to estimate $E[(X - \mu)^2]$ the variance of a random variable of X (hopefully you recall this value from a previous lecture!).

If we use n rather than $n - 1$ it turns out to be a biased estimate for $E[(X - \mu)^2]$. Of course, for large n it clearly makes little to no difference. It's significance is for smaller data sets. For those curious you can read more on `https://en.wikipedia.org/wiki/Bessel%27s_correction`.

[2] and could formally prove, though we leave that out for simplicity

The variance is not on the same scale as the data. If the weights are in grams the variance is in grams-squared. So often, we use the square root of the variance as a measure of spread:

Standard Deviation

The standard deviation s = the square root of its variance.

The larger the variance the more spread out the data is. Imagine the data was always the same – $3, 3, 3, 3$ for example. Then the variance would be zero. If the data was not concentrated around the mean, the variance gets larger.

The standard deviation of birth weights is 288.83

One alternative to variance is to take the absolute value (remove the sign) and average:

Mean Absolute Deviation (MAD)

$$\text{MAD} = \frac{1}{n}\sum_i |x_i - \bar{x}|$$

This doesn't "punish" large values as much, and can be helpful depending on the dataset. It's also an excuse to have a MAD statistic.

Relating this back to the model, this is an estimate for $E[|X - \mu|]$.

Quartiles

The median has 50% of the data above and 50% below. It's in the middle of the data.

The central 50% of the data lies between the lower quartile Q1 and the upper quartile Q3. The lower quartile has 25% of the data below, and 75% above. The upper quartile has 75% of the data below, and 25% above, best illustrated here:

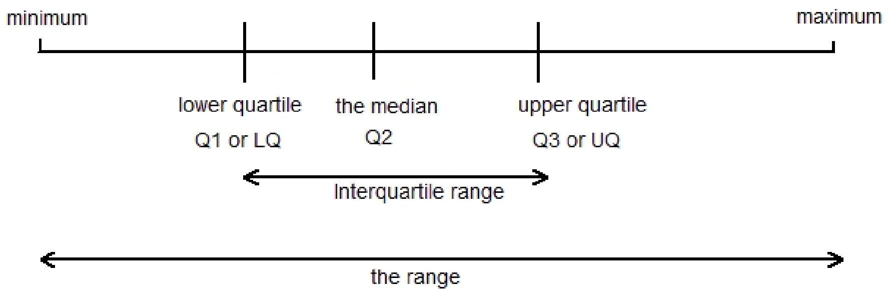


Figure 6.3: Quartiles of numeric data.

The inter-quartile range (IQR) is simply the distance between Q1 and Q3. It is (yet) another measure of spread.

For the birth weight data, the first quartile Q1 = 3279, the third quartile Q3 = 3621.25, while the IQR is 342.25.

6.4 Histograms

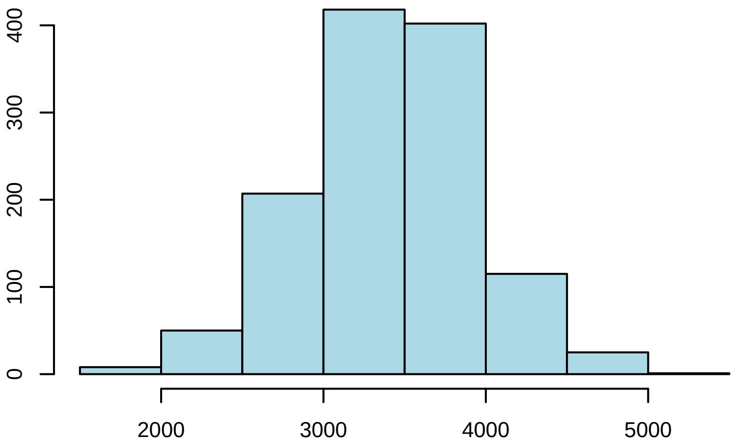
For qualitative data we drew bar plots. For quantitative data we do something similar, but we must place bars in order, to reflect that data has this given order.

To create a histogram:

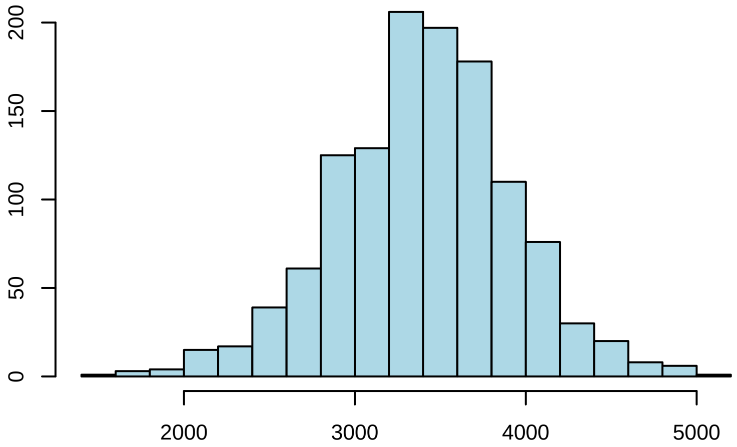
1. Split the range into a **good** number of bins.
2. Assign each observation to one bin, increasing the bin count by 1.
3. Plot of the count of observations assigned to each bin.

The histogram looks like a bar plot without gaps between the bars in order to show that the data is continuous.

Here is the histogram of birth weights, for a particular bin size:



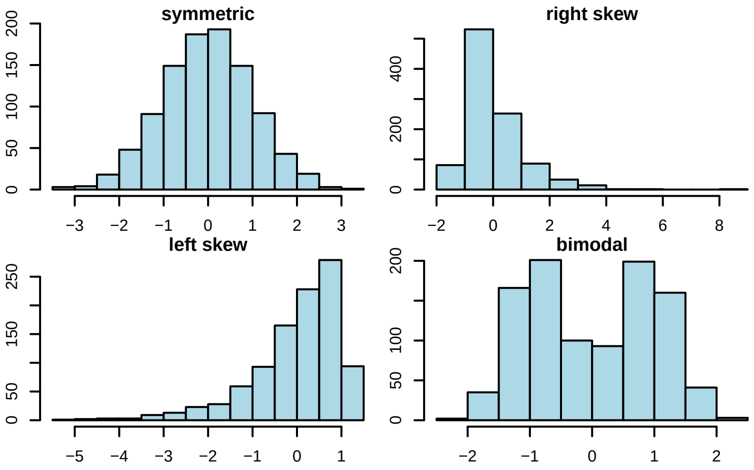
If we increase the bin size this is what happens:



A histogram can be seen as a view into or estimate of the distribution of the data.

Note the trade-off in choosing a “good” number of bins. Now we have finer details as we split the bigger bins into smaller ones. Depending though on the data, it may not be illustrative. Imagine all values in our data were different, and we went to such small intervals that only one data point landed in that bin. That means lots of bin, each having either 1 or 0 items in it. That wouldn’t yield too much insight.

Here are some common distribution shapes:



6.5 Back to birth weight and smoking

Let’s come back to our question about if smoking during pregnancy affects birth weight. To go in that direction, we can first compute the summaries and histograms for each group separately.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
no	1571	3229	3514	3515.639	3829	5029
yes	1657	2914	3286	3260.285	3600	4657

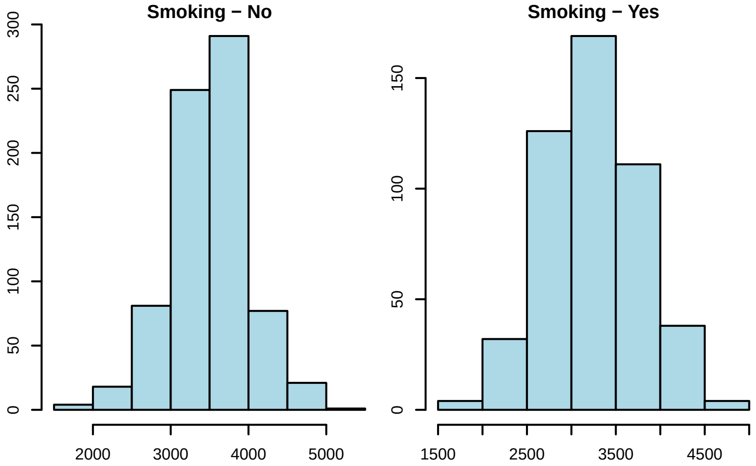


Figure 6.4: Histogram of birthweight.

Box plots

Side-by-side histograms are hard to interpret. Box plots display the summary information in a graphical way.

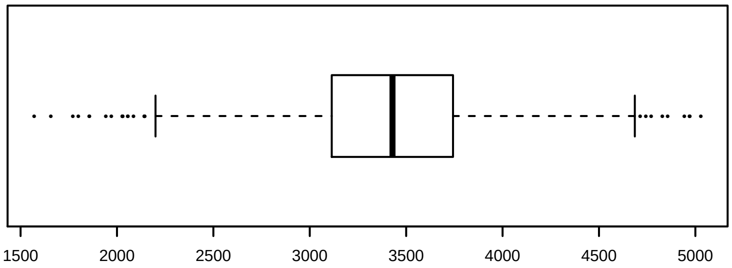


Figure 6.5: Box plot of birth weight.

The central box is defined by the lower and upper quartiles. It contains 50% of the data.

The thick line is the median.

The dotted lines (whiskers) extend to the furthest data point that is no more than 1.5 times the IQR from the box (in this version).

The data further than this are plotted as points.

We can use box plots by groups:

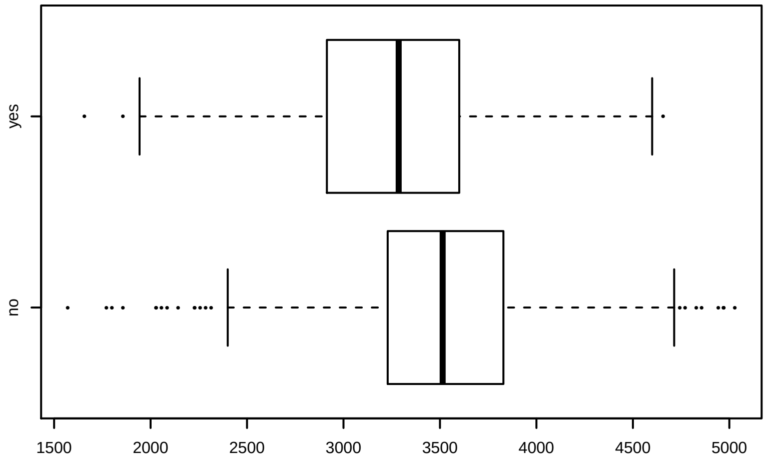


Figure 6.6: Box plot of birthweight given maternal smoking status.

What we see is that the “no” group seems to have a higher median birth weight, that the spread of the two groups seems about the same, and that the no group seems to have more points outside the 1.5 times the IQR of the box.

6.6 Comparing the two

We need to make the question “Does smoking effect maternal birth weight?” more concrete by asking a specific question about a measurement, and from that forming a hypothesis.

One formulation would be if the infants of smoking mothers have lower **average** birth weight. Then we need to choose between *mean* and *median.*

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
no	1571	3229	3514	3515.639	3829	5029
yes	1657	2914	3286	3260.285	3600	4657

The median and mean both seem lower for smoking mothers. We have to though be careful to rule out that this might be an artefact of the sample. We want statistical significant difference.

For the smoking data the difference in means is 255.35.

We must examine if this difference could be obtained by chance if the means really were the same, for samples of size 742 no, and 484 yes.

Let us denote the mean birth weight for babies from the smoking pregnant women sample be denoted $m_1 = 3260.285$, and for the non-smoking sample by $m_2 = 3515.639$.

Now we can formulate a hypothesis. Let the population mean for all smoking pregnant women be μ_1, and all those pregnant women who don't smoke be μ_2. Note that these are model parameters, not the measurements from the data.

Thus our (alternative) hypothesis can be:

$$H_1 : \mu_1 < \mu_2,$$

while the null hypothesis then becomes that they are equal:

$$H_0 : \mu_1 = \mu_2.$$

A couple of notes here. First, an alternative alternative hypothesis (sorry couldn't resist) could be

$$H_1 : \mu_1 \neq \mu_2.$$

This is an alternative hypothesis requiring potentially less evidence than for the stronger one $\mu_1 < \mu_2$. The difference between the two will become clear when we look at different p-value estimates.

Second, the sample means m_1 and m_2 are estimates for μ_1 and μ_2 respectively.

Now it remains to figure out how to see if we can reject the null hypothesis.

Suppose it were (namely that $\mu_1 = \mu_2$). Then the smoking labels – yes or no – do not affect the birth weight. This gives us a way to simulate alternatives in which the labels are ignored.

We essentially **randomly allocate the smoking labels** and get a new difference in means that is consistent with the means being equal. This shuffling the labels assumes that the labels have been randomly assigned and thus do not make a difference to the mean.

Card piles

Let's connect this to piles of cards.

Suppose you are presented with two piles of cards. The question you want to answer is: does it look like the card piles have been specially crafted, or that they have been randomly dealt out?

As is our wont, we first choose a statistic of interest. We first measure the difference in means of numbers in each pile.

We then randomise the data to determine what the statistic would look like when the piles are random. How we do that is as follows:

1. Shuffle both piles together.
2. Deal out two piles of the same size as the original piles.
3. Compute the statistic based on these new piles (the difference in means).
4. Repeat to obtain 1000 differences in means.

If the difference in means of the original pile is not like any of the difference in means of the randomised piles, then it is not likely that original two piles were randomly created.

The table below shows the first ten observations of the birth weight data. The row `smoke.random` is a random allocation of the smoke label to the birth weights.

This can be applied to the entire data set.

	1	2	3	4	5	6	7	8	9	10
bwt	3429	3229	3657	3514	3086	3886	3943	3771	3429	4086
smoke	no	no	yes	no	yes	no	no	no	no	yes
smoke.random	no	yes	no	no	no	no	yes	yes	no	no

We recorded the difference in means of birth weight for each of the 10,000 label shuffles. The summary is shown below. This represents the difference in means if we assume that the smoking has no effect.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-117.8866	-19.99175	0.2129742	0.3055107	20.87941	122.2384

The difference in means when using the label assignment from the data is 255.35. This value is much bigger than any of the 10,000 randomised values.

Our null hypothesis looks shaky, and we can indeed reject it.

6.7 Sales vs. office location

Suppose one wishes to compare the sales of the east and west offices of a particular company in a city (it's a very competitive company.

The data consists of sales amounts ($1000s) of 48 sales persons in the west office and 52 in the east. A sales manager is interested in whether there is evidence that the average sales per sales person between the two offices are different. She isn't just interested in the sample mean difference, but to truly know if statistically they can be separated from one another. There are bonuses and/or firings depending on it, so it's quite important.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
west	102.3928	140.7893	158.6182	154.0425	166.5258	193.9367
east	110.0611	144.6482	166.9310	162.6992	175.9951	206.5422

The difference in mean sales between the east and west is 8.66.

If we assume that the location does not effect the sales (our null hypothesis), we can shuffle the labels and compute the difference in means. Repeating this process 10,000 times provides 10,000 differences in means with the following summary.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-16.32221	-3.133544	-0.0633669	-0.0485031	3.077553	15.55492

289 of the 10,000 simulated differences in means are greater than the data difference in means 8.66 (2.89 percent).

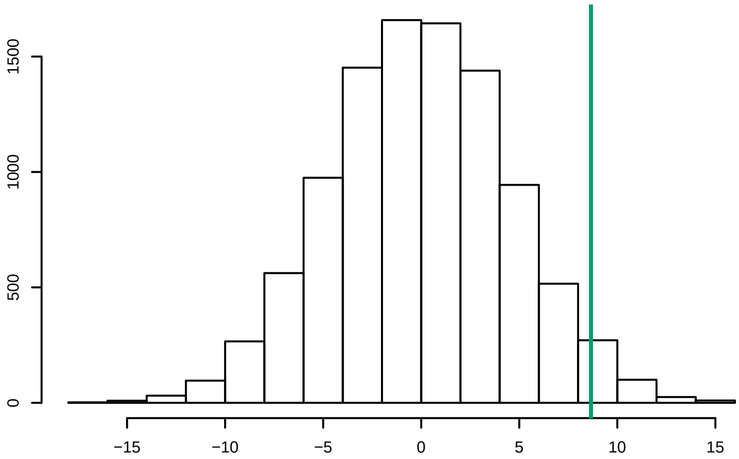


Figure 6.7: Difference in mean sales from randomly allocated sales offices.

The data difference in means looks extreme, but here we have another issue. For birth weight we were interested if smoking gave a **lower** birth weight. So if $H_1 : \mu_1 < \mu_2$.

Here we are interested if sales are just **different**. So here $H_1 : \mu_1 \neq \mu_2$.

In fact, there are 296 randomisations that give differences in means larger than 8.66 but with the opposite sign:

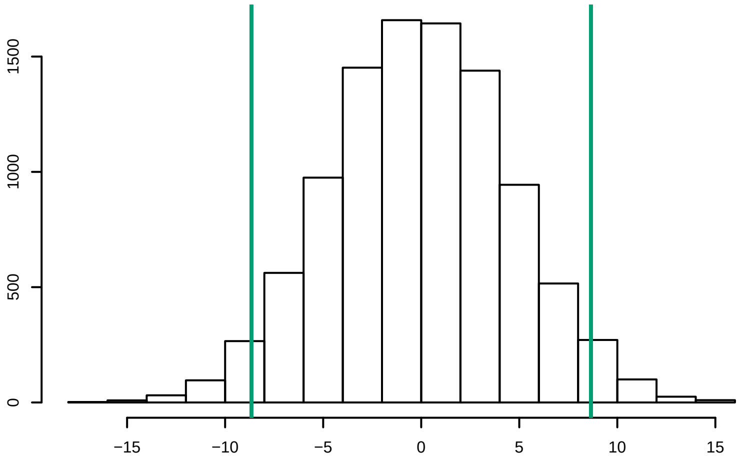


Figure 6.8: Difference in mean sales from randomly allocated sales offices.

Two-sided tests

When we are interested in just difference and not less than or greater than, we have a **two-sided test**. Then we need to count the randomisations on both sides. For the sales data there are 585, yielding a **p-value** of 0.0585.

Here we see the p-value again. A p-value is the proportion of samples from the randomisation that provide a statistic that is more extreme than the original sample, *assuming the null hypothesis.*

For office sales, what we are asking is what the chance that the randomisation can produce a difference in means is as large or larger than the original sample.

Using Medians

The above hypotheses both involved means, but it is equally possible to use medians.

For the sales data 10,000 differences in means using shuffled location provides:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-21.21642	-3.1677	-0.2059212	0.0226339	3.222071	21.00838

The difference in medians for data (west-east) is 8.3128681. Note that 670 of the randomised differences in means were greater than 8.3128681 and 624 were less than -8.3128681.

In the next worksheet, you will calculate the p-value of the above test.

Recap — Hypothesis testing

These questions are examples of hypothesis testing, where we compare a measurement on two groups of observations.

We propose a null hypothesis H_0 of no difference (in mean or median).

The alternative hypothesis H_1 is whether

- the means are different (**two-sided**),
- or one is greater than or one is less than other (**one-sided**).

This choice of alternative affects the way a p-value is computed.

To simulate the null hypothesis, we generate randomisations of the groups (shuffle) and compute the difference (in mean or median).

We then compare the actual difference (from the data) to randomised differences to get a p-value.

p-value

The p-value is the chance of seeing a difference as extreme as the actual difference in means, given that null hypothesis is true.

It is used as a measure of evidence against the null, in favour of the chosen alternative.

Using medians gave a different p-value (in the sales case).

The number of randomisations used determines *how accurately we estimate the p-value.*

Summary

- Numerical data can be summarised using its mean, median, variance, quartiles and interquartile range.
- Numerical data can be visualised using a histogram.
- Boxplots allow us to compare numeric data from two or more groups.
- Shuffling labels allows us to test if the mean depends on the label.
- We can also test the difference in medians.
- These tests are hypothesis tests, that require a null hypothesis H_0 (that we can simulate), and an alternative Hypothesis H_1 (that we conclude if we reject the Null Hypothesis).
- The p-value is the probability of obtaining the statistic (difference in mean or median) value, or a more extreme value, assuming the null hypothesis is true.
- If the alternative hypothesis refers to **difference**, then the p-value uses both sides of the randomisation distribution.

6.8 What is data?

All fields of endeavour these days collect information about individuals or objects of interest. This might be collected for an operational purpose such as details of customers, transaction details or telephone call details.

Sometimes it collected for a specific purpose such as customer surveys, environmental monitoring or scientific experiments.

All of this is data, which can take many forms.

- Quantitative - numerical measurements or counts eg. bank balance, item cost, …
- Qualitative - non-numerical, categorical data eg. hair colour, group membership (treatment versus control)
- Unstructured Data - text, audio, images, video etc. eg. tweets, accident descriptions, audio transcripts, Instagram,…

Populations and samples

When data is collected it usually takes the form of a number of **observations** on one or more **variables**. Often this data can be arranged as a **data matrix** (in `R` a `data.frame`) such as this one:

id	gender	race	ses	schtyp	prog	read	write	math	science	socst
90	female	white	high	public	academic	42	54	50	50	52
27	male	asian	middle	public	academic	53	61	61	57	56
96	female	white	high	public	academic	65	54	61	58	56
22	male	hispanic	middle	public	vocation	42	39	39	56	46
82	female	white	high	public	academic	68	62	65	69	61
56	male	white	middle	public	vocation	55	45	46	58	51
173	female	white	low	public	general	50	62	61	63	51
121	female	white	middle	public	vocation	68	59	53	63	61
146	male	white	high	public	academic	55	62	64	63	66
179	female	white	middle	private	academic	47	65	60	50	56

Data is rarely collected on all individuals of interest. Usually we have access to a small subset, such as a survey of 1000 residents in Sydney or the yields for 10 plots of some crop.

When data is collected on all individuals of interest it is called a **census**. Even then, the census may represent only one time period[3]

The set of all individuals of interest is called the **population**. The data we collect is called a **sample**, a subset of the population.

Inference and Estimation

We are usually only specifically interested in the what the sample can tell us about the population.

For the birth weight and smoking data we are interested if smoking is associated with lower birth weight in the general population.

For the eels we want to know if the two species populations **generally** live in different habitats (not the samples we happened to observe).

The sample is used to either *infer* something about the population, or *estimate* something in the population.

Inference means to answer a specific question about the population using a sample. **Estimation** means to estimate something in the population using a sample.

Example inferences include if the mean birth weight for smoking mothers less than that for non-smoking mothers or if the habitat distribution of G.moringa differs from that of G.vicinus.

Example estimation includes arriving at a figure of the average difference in birth weight for smoking and nonsmoking pregnant women, and what the

[3] In insurance data, we have data on *all* current policy-holders, but we are interested in predicting the costs for *next year*).

average number of deaths by horse kick per regiment per year in the Prussian army is.

Types of Studies

One major distinction in the way data is collected is between observation and experiments. In observational studies, no intervention is made to the individuals in the population, data is simply observed.

Observational studies include surveys where a number of individuals are deliberately asked a specific set of questions. For example, environmental monitoring where the contaminants of the air (for example) are measured.

Experiments on the other hand involve an intervention in individuals sampled from a population - such as applying fertiliser to a crop or trialling a new drug on cancer patients.

Random Sampling

Generating a sample, where possible, from a population is a critical component of any study. The sample should be in some sense *representative* of the population of interest.

It would be problematic to sample a cohort of students by asking only those in the front row. Similarly, if a conservative newspaper wanted to gauge the wider popularity of a conservative candidate, sampling its readers would not be representative.

In particular, having a representative sample allows us to **generalise** to the population. The simplest way to ensure a sample is representative is to draw members of the population at **random** to form the sample. This is called **random sampling**.

Experiments - Random Allocation

The key idea in **experiments** is that even with a random sample, we must then still allocate the treatment randomly.

If interested in comparing the effect of two fertilisers, we can choose seed or plants at random from the population, then allocate the two fertilisers randomly.

In comparing a new drug to a standard treatment we can, we can draw patients at random, then allocate the drugs randomly.

In contrast, if we want to compare the effect of a single drug on males and females, we can draw males and females at random, but we cannot allocate gender randomly…

Random allocation allows us to draw conclusions about *causality* - the treatment *causes* the effect. Without random allocation we can only find *associations.*

Studies where we cannot feasibly allocate **treatments** at random are observational studies. For example, comparing responses to a survey by gender. We cannot allocate gender.

When comparing contaminated sites to pristine ones, we cannot allocate which sites are to be contaminated (usually/hopefully).

Surveys are observational Studies are surveys — questionnaires sent to groups of people to determine attitudes and relationships to demographic factors. Most environmental studies are observational.

Prospective versus retrospective studies Prospective studies set up the sample and follow the outcomes as they occur (for example, clinical trials are usually prospective).

Retrospective studies look at already collected data and try and associate risk factors with outcomes (public health research using hospital records is often retrospective).

Study design

Bias, Variance and Confounding

There are several issues that can arise when designing a study.

Bias is caused when the sample is not representative.

Estimates may not be representative of the population but vary systematically depending on the sample. For example, estimating the size of a population of crabs, by measuring those caught in a particular trap, might be biased due to larger crabs having a higher chance of being caught.

Or responder/non-responder bias: in surveys response is usually voluntary. If the survey is asking for opinions only those with strong positive or negative opinions may reply.

Another example could be studying the cholesterol levels of people, by sampling in a fast food car park.

Regression to the mean

One type of biased sampling is called regression to the mean. When measuring a group of individuals it may be that some are extreme, but when measured again they are not.

For example, suppose a new blood pressure treatment is to be tested. People with high blood pressure (as measured in a screening test) are selected for the treatment. After treatment the blood pressure is measured again. Even if the treatment does nothing it is possible the high measurements will revert to the mean.

Of course, real effect of treatment could exist, but sampling the highest blood pressures after one measurement may not be best.

Variability

All data measurement has variability, but sometimes some of it can be controlled.

Think about assessing a weight loss program, which of the following designs would you think is best?

1. Taking two groups of people, putting one group on the program and the others not, and weighing (only) after they have been on the program.
2. Taking two groups of people, matching so that for each person in the program group there is a person of similar weight in the non-program group. Weighing (only) after they have been on the program.
3. Taking two groups of people, putting one group on the program and the others not. Then weighing before the program and after, and comparing the changes in weight.

It is likely that (given big enough samples) all three could distinguish any difference, but the third design eliminates uninteresting variability by considering changes.

There are usually multiple sources of variability. Think of a measurement, it is probably made up of several components added (or multiplied?) together.

Population average: eucalyptus trees would have an average height for example.

Natural variation: Each tree just naturally has a different height due to genetics and/or environment.

Temporal variation: A single tree may have different heights in its life cycle.

Measurement error: Measuring the height of tree is not an error free process. Measuring the same tree at the same time, may give different heights at different measurement attempts. This might include investigator error.

Confounding

Confounding occurs when sub-populations of interest happen to coincide with features not of primary interest. It can occur by chance, or by bad design.

For example, does more lichen grow on north or south facing sides of trees? If we look a the north side of trees on Hawkesbury campus, and the south side of trees on Parramatta campus, then “side” and “campus” are confounded. We don’t know if differences are due to “side” or “campus”

Confounding can usually be avoided by careful design (how we take a sample).

Placebo effect

One kind of confounding can be due to the **placebo effect**.

If individuals are offered a treatment, and outcomes compared to untreated individuals, it is possible the act of offering *any* treatment has the effect.

So often new treatments are compared to a control (existing well known treatment), or a placebo, which could be:

- an inactive treatment that mimics the active treatment in all ways except the component being tested;
- eg. chalk pill versus active pill;
- saline injection versus active drug injection;
- or "cup of tea and a chat" versus formalised counselling session.

Sampling methods

In order to eliminate bias, appropriate sampling methods must be used. The simplest form of sampling is **simple random sampling**. In this case, every member of the population is given the same random chance of being chosen for a sample.

R (and other software) can help with this by giving an identifier for every member of the population and randomly choosing a sample. It is often assumed that a sample is a simple random sample without much justification.

Sometimes simple random sampling is expensive. For example, interviewing members of a population may involve visiting a wide variety of locations. **Cluster sampling** is a cheaper, sometimes easier, alternative.

If the population forms natural groups or clusters, these can be randomly sampled instead. For example, if the population is all residents of a region, the clusters might be the towns in the region. Sampling the towns (and then maybe the people within the towns) would potentially involve less travel.

Done correctly, cluster sampling does not introduce bias.

Stratified sampling attempts to account for a potentially confounding variable. If you want to survey people on internet attitudes, you might first divide the population into age groups, and split the sample across the groups. This makes sure that random sampling does not miss an age group by chance.

Unless the age groups are sampled in proportion to their size, adjustments will need to be made.

Blocking is similar to stratified sampling. Suppose you want to compare the growth of plants under four treatments.

There are four fields that can be used. We could randomly allocate treatments to fields. But it would be better if we could divide each field into four plots, and use each treatment once in each field. Then the fields are blocks and we can allow for any random variation due to the fields alone.

Other "sampling" methods

Here are some questionable sampling methods.

Convenience samples involves using a sample of individuals because they are close at hand or convenient. You sample political opinions by walking around a shopping centre early one morning or estimating student debt by asking those in the current class. Research using convenience samples must justify how the sample is

Snowball sampling sounds cooler than it is. It just means that subjects are drawn from the contacts of the first round of participants.

Responder bias — asking a wide range of people to take part in a study, but allowing participants to self select

Principles of study design

In general there are a few key principles of study design:

1. Randomisation: eliminates bias and allows generalisation.
2. control variation: allows more subtle effects to be detected.
3. replication: helps to overcome variation due to unknown sources.
4. blocking: helps to overcome variation due to known sources

Summary

- Data comes in many forms. In its simplest form it is quantitative or qualitative.
- The population contains all items of interest. A sample is a randomly selected subset of these.
- We can infer answers to questions about a population, or estimate parameters.
- We can control the variables of an experiment, but only observe the variables of an observational study.
- Studies must be designed to account for different sources of variability.

Lecture 7

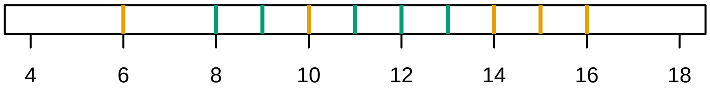
Confidence, bootstrapping and mapping disease

7.1 Wilcoxon-Mann-Whitney test

We focussed on comparing the average birth weight between the groups. There is a different way to compare groups where we think of a shift in distribution rather than a difference in average.

The Wilcoxon-Mann-Whitney test is based on the idea that if we are interested if one group (group B say) generally has high values than the other (group A) then, we can count the number of times an observation in group B exceeds an observation in group A.

To illustrate consider the data below:



The 5 green lines represent the values in group A, and the 5 orange ones those in group B.

The general idea is then, for each orange line count the number of green lines to the left of it (less than), and then add up the total for all light lines.

In this example we get 0, 2, 5, 5, 5 for a total of 17.

Wilcoxon-Mann-Whitney statistic

Suppose we have n in one group called x_i and m in another group called y_j. $U_{ij} = 1$ if $x_i < y_j$ and $U_{ij} = 0$ when $x_i > y_j$ (clever things are needed when they are equal). Then the sum of those is then the Wilcoxon-Mann-Whitney statistic

$$U = \sum_j \sum_i U_{ij}.$$

U has a maximum possible value nm when all y_j are bigger that all x_i and a minimum possible value of 0.

Note how this test ignores the value of each point and only considers the order.

Problem We want to test if the birth weight is lower when smoking status is "yes". Calculate the value of U for the below sample:

bwt	smoke
3429	no
3229	no
3657	yes
3514	no
3086	yes
3886	no

For the maternal smoking data $U = 231918$, the p-value can be shown to be essentially zero. For the sales data, $U = 1517$ and the p-value estimate is 0.064.

7.2 Population vs sample

It is really important to emphasise that questions such as "are the means of these two groups equal?" do not mean simply taking the means from the smoking and non-smoking mothers **in our data** and seeing if they are the same or not.

We are trying to make an **inference** about smoking and non-smoking mothers in general.

The complete set of all smoking/non-smoking mothers is our **population** of interest. The data we have is just a **sample**. The sample size does turn out to be a good one in this case.

Population

A **population** contains all individuals or objects of interest. Data are collected from a **sample**, which is a subset of the population.

If the mean birth weight in the population of infants born to non-smoking mothers is really μ_1 (and μ_2 for the smoking mothers) then the mean in the sample is an **estimate** of the mean in the population ($\bar{x}_1$ is an estimate of μ_1).

We are really interested in whether $\mu_1 = \mu_2$ or not. (equiv. $\mu_1 - \mu_2 = 0$).

We try and see if the **observed** difference in **sample means** is likely to have occurred **if** the population means are equal.

We have been asking if there is any evidence against the null hypothesis

$$H_0 : \mu_1 = \mu_2.$$

We have in mind a particular alternative, one of

$$\begin{array}{ll} H_1 : & \mu_1 \neq \mu_2 \\ H_1 : & \mu_1 > \mu_2 \\ H_1 : & \mu_1 < \mu_2 \end{array}$$

We then compute a **test statistic** and evaluate the chance of seeing something as extreme assuming the null hypothesis.

7.3 Confidence intervals

Answering a specific question about a hypothesis can tell us that there is a difference. It doesn't give us any information though about the actual reduction in birth weight associated with smoking. It's a binary thing.

From the data, we see the difference means $\bar{x}_1 - \bar{x}_2 = 255.35$.

This can be seen as an estimate of the population difference $(\mu_1 - \mu_2)$, but how good is it? It would be better if we could get a **range** that is likely to contain the population difference. Then we can make more scientific statements such as "the true population difference in mean birth weight lies between … and … with a certain confidence".

In contrast to a hypothesis test, it's not about assuming H_0, no difference in means, and seeing what the consequences are. It's in a sense about accepting a difference and estimating it. They are not unrelated however – a topic we will return to later.

The data is one sample from the population of interest. Suppose for a moment that we could take more samples from the population. Each difference in sample means would be an estimate of the difference in population means.

They will generally all be different. Some differences could be very large due to random chance. But most would be close to the population mean difference.

Consider the following simulation. One group has $n_1 = 50$ observations and a population mean of $\mu_1 = 10$. The other is $n_2 = 50$ observations and a population mean of $\mu_2 = 15$. There are 1000 sample differences. The true difference is 5.

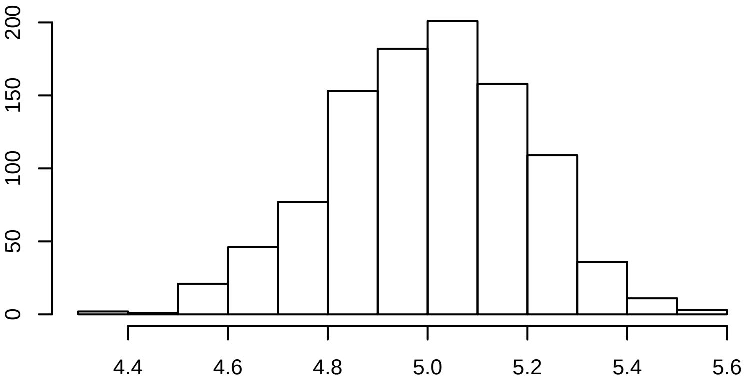


Figure 7.1: Difference in means from many samples from the same populations.

We observe that **most** (95% of them) of the differences lie between 4.61 and 5.36. The differences are centred on the true population difference. If we didn't know the true difference, we might say that there is a high chance that the true difference is between 4.61 and 5.36.

This allows us to say that the range 4.61 to 5.36 is a 95% **confidence interval** for the difference in means.

Pulling data up by the bootstraps

That was the world of simulation, where we can generate new samples at will. Unfortunately, we can almost never take multiple samples this way. It is usually an effort to obtain one sample of size n, so taking 1000 samples of size n is out of the question.

This brings us to the concept of bootstrapping: using the original data to simulate multiple samples. If we took samples of the same size taking each value only once, we would obviously end up getting the same sample as the original. So instead, we sample with replacement. This is called resampling.

Bootstrap Sampling

Bootstrap sampling is the process of resampling from our data to estimate the distribution of a sample measurement.

We can then compute the difference between the sample means for these resampled data sets, and use it to construct a range or confidence interval. This whole procedure is called bootstrapping.

Let's do this for the birth weight data:

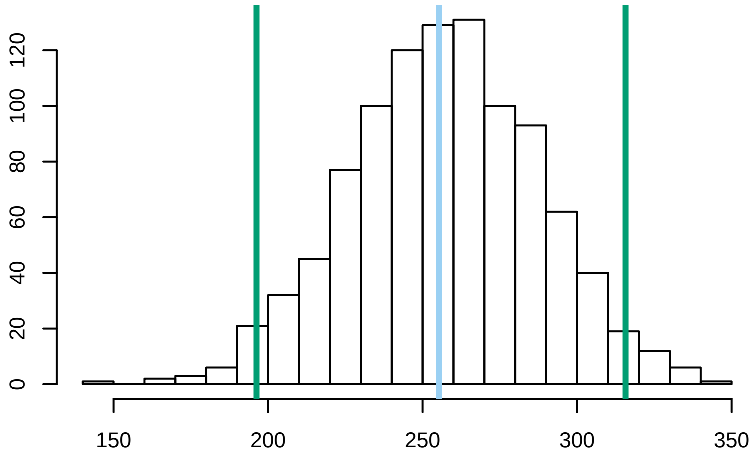


Figure 7.2: Bootstrap estimate of the distribution of the difference in sample means.

Observe that 95% of the bootstrap samples lie between 196.26 and 315.68 (the darker lines).

95% bootstrap confidence interval is between 196.26 and 315.68. The interval boundaries are chosen so that 2.5% of the bootstrap samples are less than 196.26 and 2.5% of the bootstrap samples are greater than 315.68.

This is done by taking the 1000 bootstrap sampled differences in means and finding the 25 largest and 25 smallest (25/1000 = 2.5%).

Note that there is nothing special about 95% except that its (reasonably) close to 100%. We could also compute the 90% bootstrap confidence interval between 206.29 and 306.38 (5% above and below).

Now let’s look at the sales data:

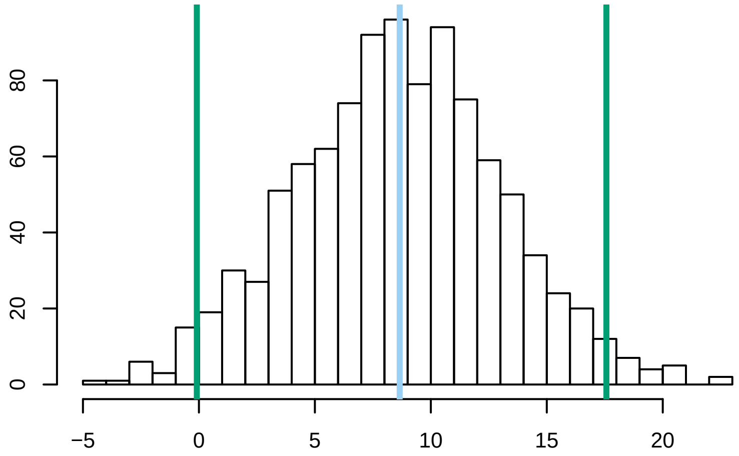


Figure 7.3: Bootstrap estimate of the distribution of the difference in sample means.

Here, 95% bootstrap confidence interval is between -0.09 and 17.57.

Hypothesis tests vs confidence intervals: In a sense hypothesis tests and confidence intervals are equivalent. If a 95% confidence interval for the difference in means does not contain the value zero, the the p-value for a two-sided test of whether the difference in means is zero **must be** less than 5% (0.05).

7.4 Mapping disease

The map below highlights the top 10% of US counties with the highest (adjusted) kidney cancer rates.

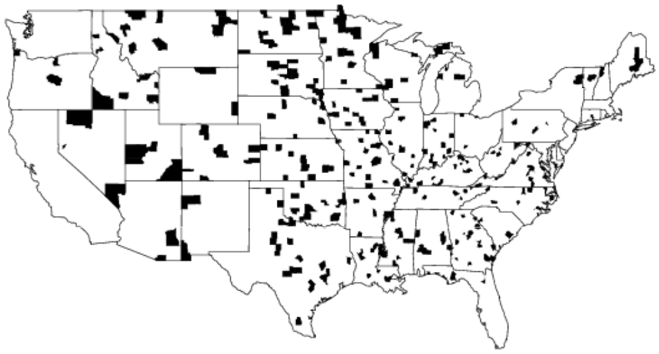


Figure 7.4: Highest Kidney Cancer Rates

Looking at this graph, you could wonder why there are so many cases in the mid-west.

Let us see what's happening with the bottom 10% of US counties with the lowest (adjusted) kidney cancer rates.

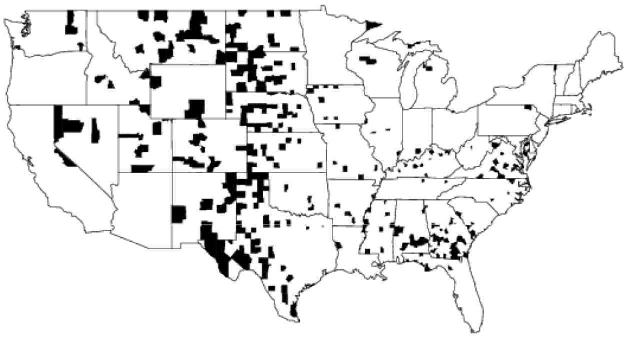


Figure 7.5: Lowest Kidney Cancer Rates

The highest and lowest rates are in very similar areas. What is going on?

To get some insight we are going to simulate, and see if the simulation can reproduce the results under its assumptions, and thereby gather insight.

Let us create some cities given the names A–J and simulate people within them getting a disease, by assigning a each individual a 10% chance of getting the disease. Cities A–E will have a population of 1,000 and F–J 100. To get an idea of the trend, we will repeat it 10 times.

The code used is as follows:

```
> set.seed(12345)
> popn = rep(c(100, 1000), each=5)
> for(i in 1:10) {
+     cnt = rbinom(length(popn), popn, 0.1)
+     rate = cnt/popn
+     names(rate) = LETTERS[1:10]
+     cols = rep("grey", 10)
+     cols[which.max(rate)] = "red"
+     cols[which.min(rate)] = "blue"
+     barplot(rate, ylim=c(0,0.2), col=cols)
+     Sys.sleep(1)
+ }
```

Repeated 1,000 times results in the typical bar plot:

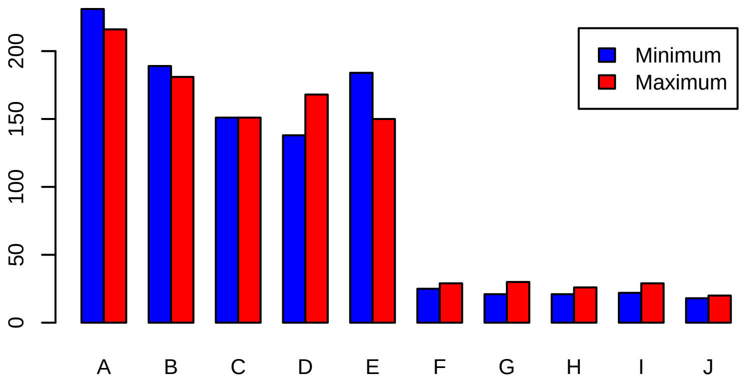
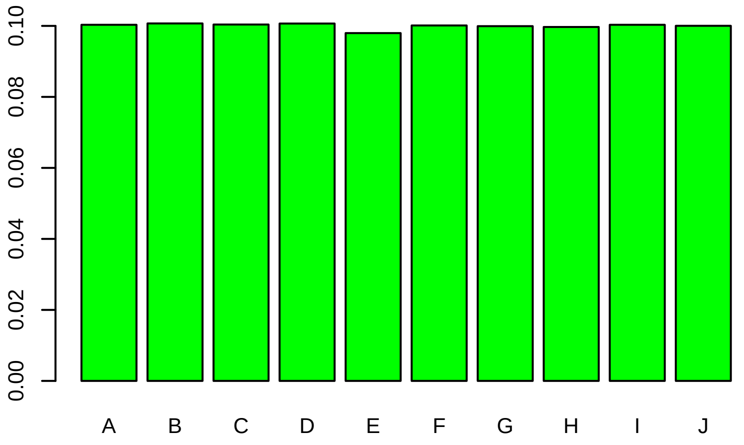


Figure 7.6: Location of maximum and minimum rates from simulation.

Just as in the dataset we started with, the minimum and maximum estimated rate tend to be in the left half of the cities.

Now, let's examine the average rate of disease for the 10 cities (over 1000 simulations):



To get some insight, we can model the how likely no-one is of getting a disease when people have a 10% chance of getting it using the binomial distribution.

We covered the binomial model in Lecture 2. Recall we can use to estimate, for example, if a coin is tossed 10 times what the average **proportion** of heads would be. A binomial experiment requires n independent events and that each event has the same probability p of success. We are interested then in the number of successes from the n trials.

The probability of k successes from the n trials is a binomial distribution with probabilities:

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

where X is a binomial random variable.

For example, suppose thererr are n individuals, and each has a probability of p of having a disease (or some other attribute). The probability that k out of the n individuals has the disease is found using the binomial distribution.

From counts to rates

The binomial distribution provides the probability of obtaining a number of successes. That's not what we looked at with the kidney cancer data: there we saw however examined rates of success (kidney cancer).

Let X be a binomial random variable from n trials. We are really interested in the rate of success which is another random variable $Y = X/n$. (to convert from a count to a rate, we divide by the total possible number of counts).

The expected value of the rate Y is p. The variance[1] is $p(1-p)/n$.

The key point here is that the larger the population n the **smaller** the variance.

We simulated ten cities A–J with the same underlying rate of disease p, and most of the maxima and minima occur in the left half A–E.

As A–E had a (simulated) population of $n = 100$ and F–J a population of $n = 1000$. They all had the same expected value. But the second half have a much higher variability.

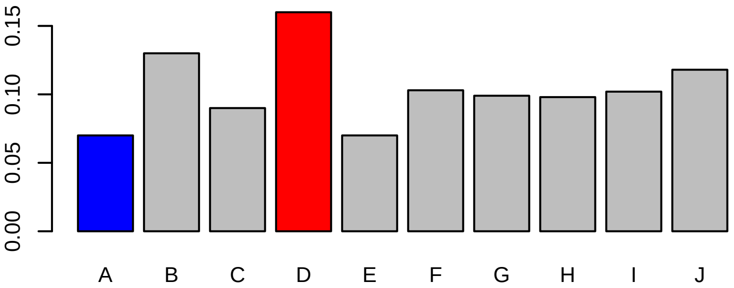
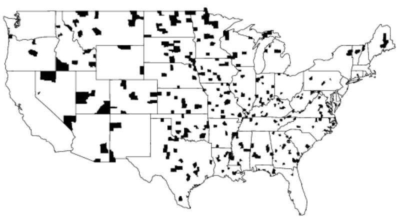


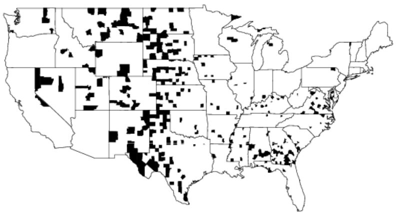
Figure 7.7: Maximum and minimum rate from simulation.

So the chances are the largest and smallest values will occur in A–E. The same is true of the cancer rates.

[1] For those interested in how we arrive at this:

$$\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = \frac{E[X^2]}{n^2} - \frac{E[X]^2}{n^2} = \frac{\mathrm{Var}(X)}{n^2} = \frac{n(1-p)p}{n^2} = \frac{(1-p)p}{n}.$$





The mid-west counties have smaller populations. The moral of the story: **be careful how you interpret graphs of rates**.

7.5 Poisson and binomial confidence intervals

Suppose we see 15 seeds germinate out of a plate of 20. Then we estimate the rate of germination as 15/20 or 0.75 or 75%.

If we assume a binomial model, (assuming seeds germinate or not with same rate), then this can be seen as an *estimate* of p.

But remember that the 0.75 is the proportion of our sample, not the population. It's a *point estimate.*

How can we compute a confidence interval for the population proportion of seeds that will germinate?

Bootstrap binomial confidence intervals

We compute bootstrapped binomial confidence intervals, by resampling from the **seeds** with replacement as follows.

First, we sample, with replacement, $n = 20$ seeds from the set with 15 copies of "germinate" and 5 copies of "not germinate", and count the number of germinates.

Then, we repeat this a large number of times to obtain a distribution of bootstrap proportions.

Finally, we find the interval that contains the middle 95% (or another chosen percentage).

Here's the result:

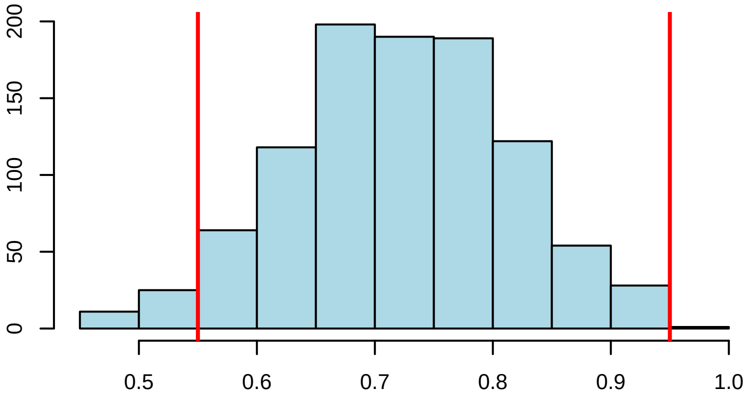


Figure 7.8: 95% bootstrap confidence interval for the proportion of germinating seeds.

The bootstrap 95% confidence interval is $[0.55, 0.95]$.

A problem with this is that we can only ever count an integer (number of germinating seeds). Each bootstrap sample provides a number of germinated seeds. So we can only get rate estimates of 0/20, 1/20, 2/20, ..., 20/20. This discreteness can sometimes be a problem, but it is reasonable here.

Calculate Poisson confidence intervals

For the horse kick data, we had 200 observations, and can bootstrap these as follows.

First, resample with replacement from the 200 observations and compute the mean Then, repeat many times (1000 say) and find an interval that contains 95% of the examples.

Here’s the result of one simulation:

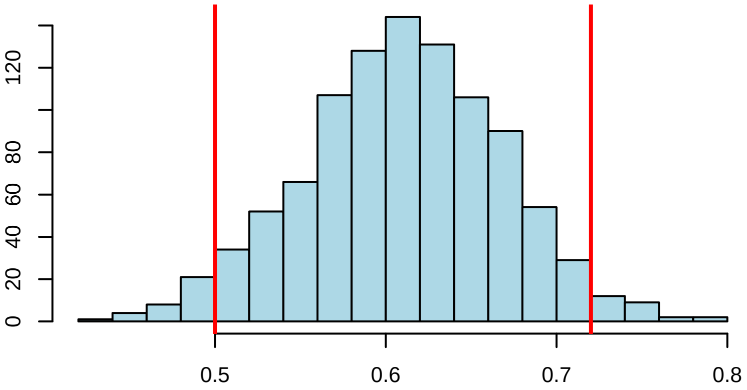


Figure 7.9: 95% bootstrap confidence interval for the mean number of deaths by horse kick per year.

The 95% confidence interval is here $[0.5, 0.72]$

Lecture 8

Correlation: do taller people earn more?

In previous lectures we looked at the relationship between two variables: for example between eel species and habitat, and at smoking status and birth weight.

In this lecture we are interested in the relationships between two variables. In particular, we are going to examine the relationship between height and income, as illustrated by a particular dataset. And as we are really into nature data, we shall look at crabs and their size before and after they moult.

The data we have is retrospective and observational. This means are looking for associations not causation.

The height and income relationship has been the subject of quite a bit of research in both economics and social science, See for example, Steckel, Richard H. 1995. "Stature and the Standard of Living." *Journal of Economic Literature.* Life is not fair and such associations can arise between beauty and income, and so on.

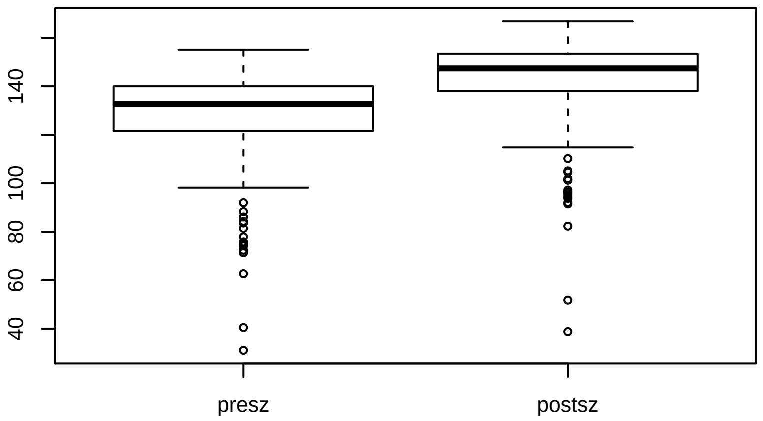
The data we have has 1376 observations of height and income. It has a few messy features. So first we will motivate using a simpler dataset, containing measurements of crabs before and after moulting.

8.1 Moulting crabs and scatter plots

This data consists of before and after moulting measurements on 472 female Dungeness[1] crabs.

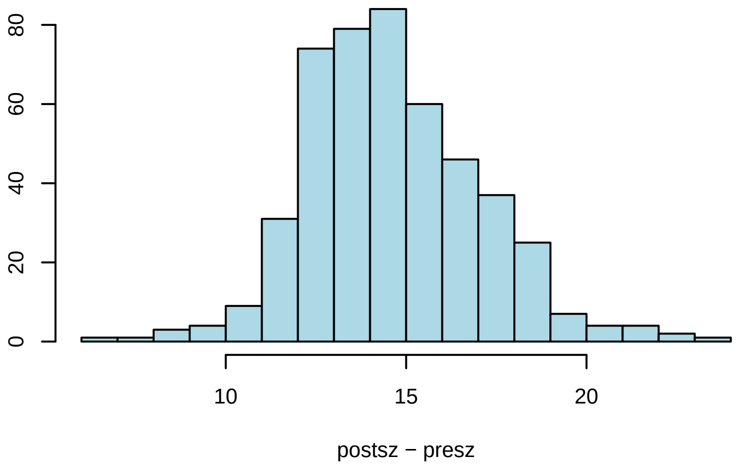
Crabs moult their shells periodically in order to grow. Here's the before and after crab moulting measurements as a box-plot:

[1] Dungeness is a headland on the coast of Kent, England.



The pre-moult size is denoted by `presz`), and the post-moult size by `postsz`.

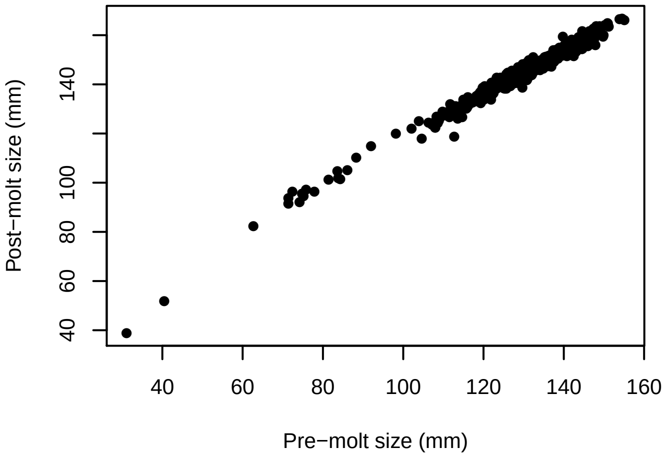
It's not too informative. Sure, the median is larger afterwards, and it may be that the postsz data is shifted-upwards. However, a crab pre- and post-moult size is related to one other. Neglecting the relationship between measurements is not so informative. So a more interesting plot would be to look at the difference in sizes:



The histogram shows that on average crabs grow by about 15mm when they moult (mean is around that). That doesn't tell us though if, for example, small crabs grow more or less than large crabs afterwards.

That would be a reasonable hypothesis: younger animals, humans included, tend to grow faster than older ones (until we start shrinking!).

The histogram of differences doesn't allow us to see how the pre-size affects the post-size. To examine the relationship between pre and post moulting size more closely let's look at a scatter plot:



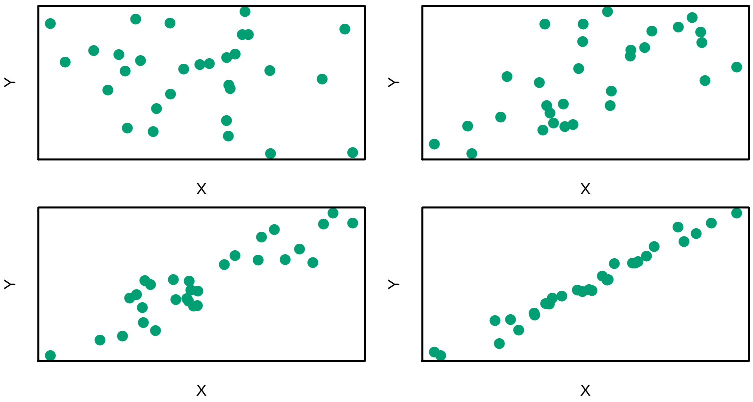
Scatter plots are a way of displaying the relationship between two quantitative variables. Box plots display only each variable separately.

One variable is chosen to be the horizontal or x-axis variable, and the other the vertical or y-axis variable. Each observation is then plotted as a point, vertically above its x-value and horizontally across from its y-value.

The question of interest determines which variable goes on which axis. We tend to think of the y-axis been what follows afterwards[2].

It shows clearly that the larger the crab pre-moulting, the larger the post-moulting. The relationship seems strong: knowing one gives a good indication of the other. What we are concerned with now is how strong this relationship is.

The strength of a relationship measures the information that one variable gives about the other. Consider the following scatter plots:



Which, do you think, exhibits the strongest x and y relationships?

All show that the **strength** of the relationship increases. As x increases so does y.

[2] As we shall see it is not critical when examining their relationship,

The top left panel seems to have no particular relationship. The top right has some, but it is weak. The bottom left is stronger, while the bottom right seems nearly perfect. How can we quantify this?

We'd like to have a statistic to capture how strong the relationship it is.

8.2 Covariance and correlation

First, some modelling. Let X, Y be random variables on some distribution.

Recall that X and Y are uncorrelated if $E[XY] = E[X]E[Y]$, where $E[XY]$ is the expected value of XY. As previously noted, they are independent if $P(XY) = P(X)P(Y)$, which is a stronger relationship than uncorrelated.

So in a sense we want a statistic that estimates, from a given data, how likely it is that they are correlated.

We first define the covariance between two variables.

Covariance

For random variables X, Y, their covariance is given by

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

Note that covariance is a generalisation of variance, as $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$. It goes from how a random variable varies with itself to how two do.

If X and Y are uncorrelated, $\mathrm{Cov}(X,Y) = 0$.

A sample estimator for $\mathrm{COV}(X,Y)$ given samples $\{x_1, \dots x_n\}$ and $\{y_1, \dots, y_n\}$ is then

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

where $\bar{x}$ and $\bar{y}$ denote the sample means of x and y respectively.

Estimating the covariance seems like a good statistic to have. Even better is to normalise it so as to account for variation (divide by $\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$). This results in the correlation coefficient:

Correlation coefficient

The correlation coefficient (or just the correlation) between random variables X and Y is given by

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

with $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$.

Note that $\mathrm{Corr}(X, X) = 1$, which is perfect correlation, as you'd expect. If we correlate X against its negative you get that $\mathrm{Corr}(X, -X) = -1$. Any other combination $\mathrm{Corr}(X, Y)$ will be in the range $[-1, 1]$.

So now we have a measure of how two random variables are correlated with each other. That's the model side.

Now let's get back to the data side. Suppose we have quantitative data representing samples from the random variables X and Y. It consists of n pairs of data (x_i, y_i). An estimate for $\mathrm{Corr}(X, Y)$ is given by the (Pearson) sample correlation coefficient:

(Pearson) Sample correlation coefficient

For dataset $\{(x_i, y_i) : i = 1, \dots, n\}$ the sample correlation coefficient is given by:

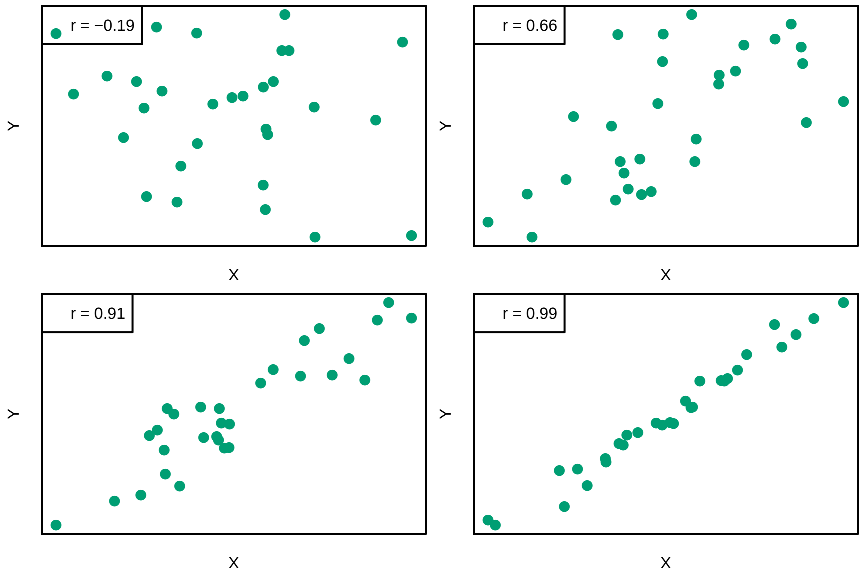
$$r = \frac{1}{n-1}\frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y},$$

where $\bar{x}$ and $\bar{y}$ denote their respective means, and s_x and s_y their respective standard deviations.

This estimator is unbiased, and, for those interested in knowing more, the maximum likelihood estimate of $\mathrm{Corr}(X, Y)$. As its formula is the same form, we will also

The coefficient r can only take values between -1 and 1. Negative correlations indicate that when X increases, Y tends to decrease.

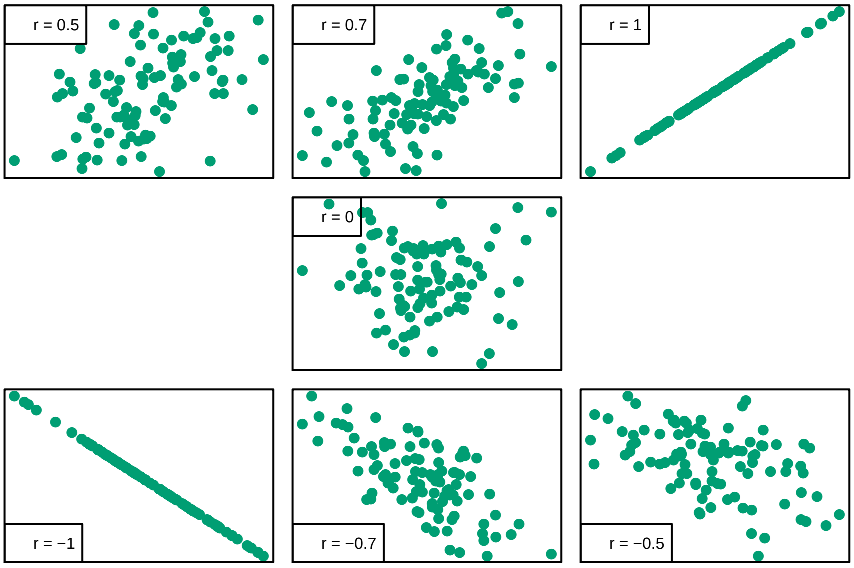
Here are some examples from data:



When r is small, the dots are scattered more, and the stronger the relationship, the more linear it gets. For a given estimate r of $\mathrm{Corr}(X, Y)$ a rough summary is given below:

- $r = +1.00$: estimated perfect increasing linear relationship
- $r = +0.70$: strong increasing linear relationship
- $r = +0.50$: some increasing linear relationship
- $r = 0.00$: no detectable relationship
- $r = -0.50$: some decreasing linear relationship
- $r = -0.70$: strong decreasing linear relationship
- $r = -1.00$: perfect decreasing linear relationship

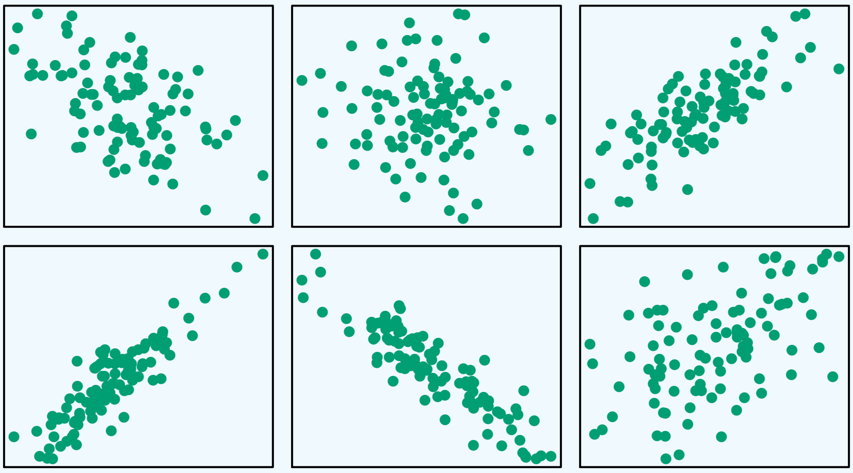
Some other examples are:



Note in particular how $r = -1$ and $r = 1$ are essentially rotated versions of each other.

Exercise

Order the the following plots from lowest to highest correlation.



Crabs moulting correlation

The sample correlation coefficient for pre and post moult size of the crabs is $r = 0.99$.

This is very high, and shows a strong linear relationship between pre and post moult size.

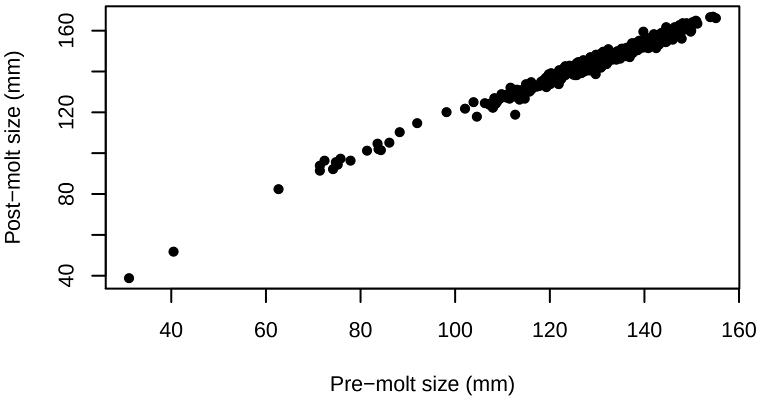
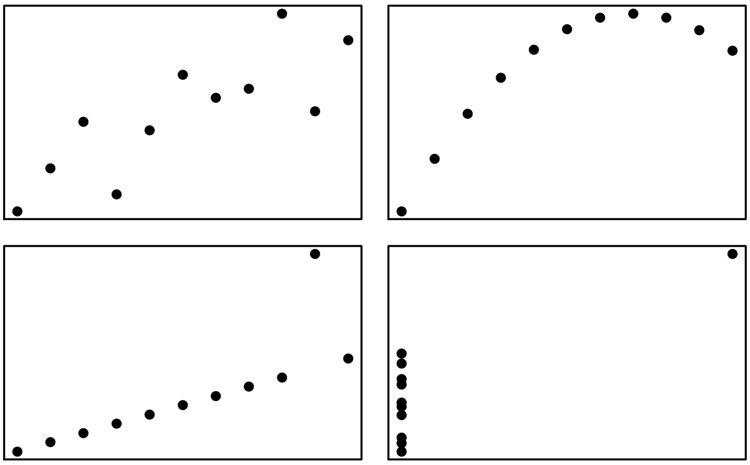


Figure 8.1: Scatter plot: pre vs post-moulting crabs.

8.3 Anscombe data sets and the Spearman correlation

Note that correlation only indicates a **linear** (straight line) relationship — all of the following have a correlation of 0.82:



The correlation we talked about so far is often called the Pearson correlation. One alternative is the Spearman correlation. It tries to alleviate any issues to do with unusual points or curved (non-linear) relationships.

Spearman correlation

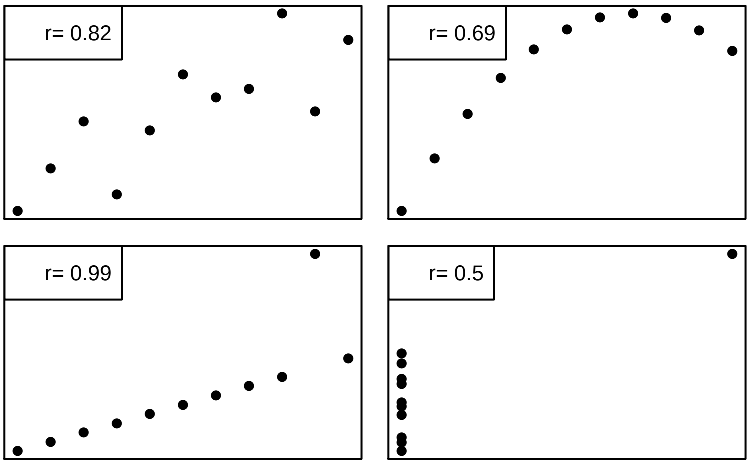
1. Variables are individually **ranked**: the smallest value is replaced with rank 1, the second smallest value with rank 2, etc.
2. Compute the Pearson sample correlation of the ranks.

Spearman correlation measures then the extent to which the ranks are linearly related.

The crab moult data has Pearson correlation coefficient 0.99, so either correlation statistic gives us the same value[3].

For the Anscombe data sets, the Spearman correlation coefficients are as follows:

[3] Though with potentially different confidence intervals – we discuss those later.



8.4 Hypothesis testing for correlation

We can via. the sample correlation now get a statistic. The naturally question to ask is if we can establish if the different variables are correlated. This translates into a hypothesis about whether the population correlation $\mathrm{Corr}(X, Y)$ is statistically different from 0. If it is equal to 0 then there is no (linear) relationship. The observed correlation may have occurred by chance. In practise, it is almost never the case that $r = 0$ – data will give us something. For short-hand let $\rho = \mathrm{Corr}(X, Y)$. What we want to know then is if $\rho = 0$.

This leads to null and alternative hypotheses:

$$\begin{aligned} H_0 &: \rho = 0, \\ H_1 &: \rho \neq 0. \end{aligned}$$

To test the hypothesis, we use as before a permutation-based simulation. We simulate the outcome given that the null hypothesis H_0 would hold.

The general idea is to say that the pairings (x_i, y_i) don't matter. So, we randomly reorder (permute) one of the variables and compute the sample correlation, breaking this relationship. Logically these new "pairings" are not in general correlated. For each pairing, we get a new sample correlation value.

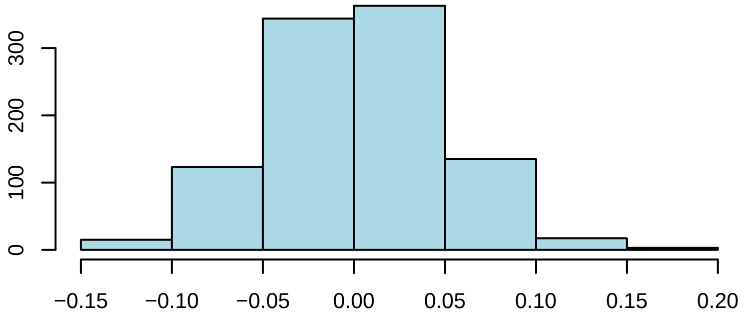
Repeating this many times will provide us with the distribution of sample correlations when the population correlation is zero.

Here's the R code:

```
> ## compute the data correlation
> obs.cor = cor(molt$presz, molt$postsz)
> x= replicate(1000, {
+   ## shuffle the post molt varaibles
+   post.perm = sample(molt$postsz)
+   ## compute the correlation of the pre and shuffled post molt values
+   cor(molt$presz, post.perm)
+ })
```

The function `cor` computes the Pearson sample correlation. Alternatively, we could swap out the`cor` line, and put in the Spearman correlation instead using `cor(molt$presz, molt$postsz, method="spearman")`

Below shows some correlations computed under the assumption of the null hypothesis:



Recall that the observed data correlation $r = 0.99$. From that dataset, we can compute the p-value numerically (as you will do in the worksheet). It is pretty obvious already though that none of those values are greater than 0.99, meaning the estimated p-value is 0.

8.5 Correlation confidence intervals

We can compute a confidence interval for a sample correlation coefficient using the same resampling (bootstrap) methods as before.

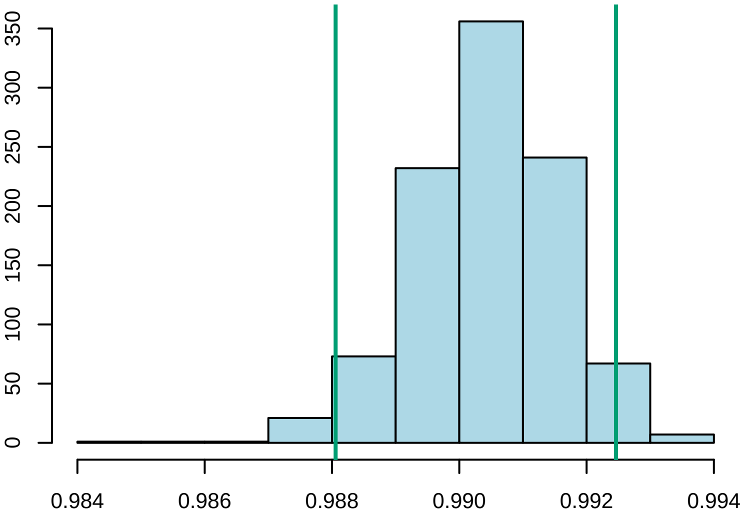
In this case we respect the pairs of observations. The easiest way to do this is to resample observation numbers.

```
> ## n is the number of observations
> n = nrow(molt)
> x= replicate(1000, {
+   ## bootstrap sample a set of observations.
+   samp = sample(1:n, replace=TRUE, size=n)
+   ## compute the correlation of the bootstrap sample.
+   cor(molt$presz[samp], molt$postsz[samp])
+ })
> ## provide the middle 95% of bootstrap correlations.
> quantile(x, c(0.025,0.975))
```

One run gives the confidence interval:

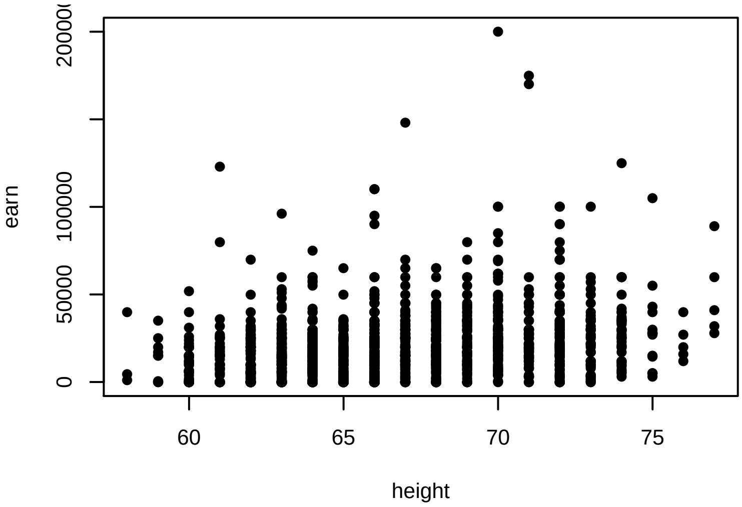
```
     2.5%     97.5%
0.9880568 0.9924622,
```

seen in the histogram:



8.6 Tall people and income once more

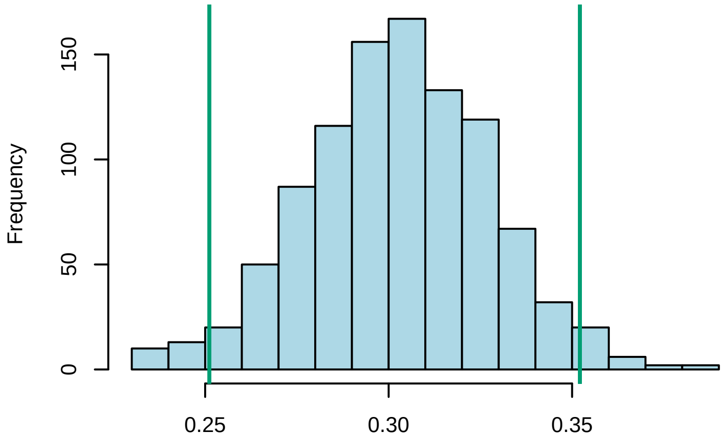
Let's return to the income versus height data set:



It's a little hard to see any relationship. Let's calculate anyway the sample correlation. It's $r = 0.302$, obviously much weaker than for the crab moulting data.

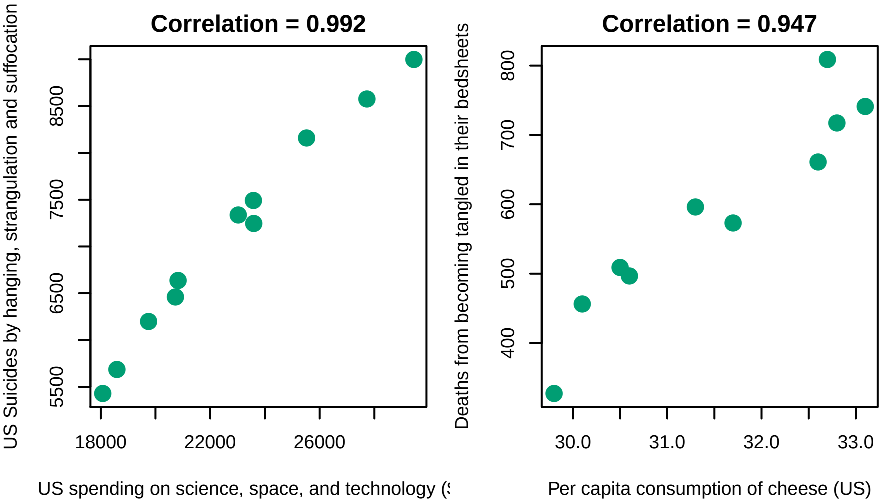
The question is then is this value consistent with the population correlation being different from 0?

We only have a sample to infer about the population, so we can use the bootstrap method. One simulation run obtained a 95% confidence interval for the true correlation of $[0.251, 0.352]$, with the simulation correlation values given by:



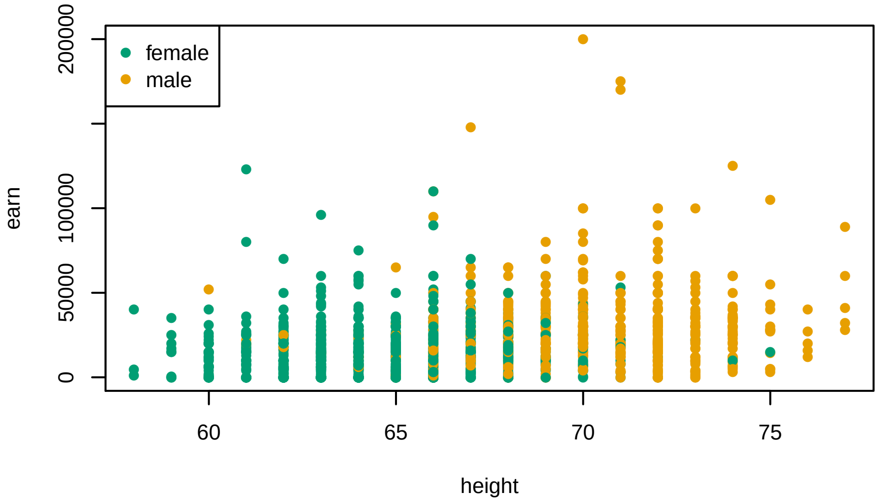
This suggests a weak correlation between the variables, as 0 is not in the confidence interval. However, we have to be careful in over-interpreting the result.

First, note that correlation does not imply causation. We only seem to have association. Consider the following relationships:



We laugh because it doesn't seem sensible that cheese production could affect people getting tangled in their bedsheets (although eating cheese before bed is said not to be a good idea). However, these associations can mislead in reasonable questions.

Let us look at the heights and income from a different viewpoint. Its sample correlation was 0.302. Men and women have different height distributions and those were combined together in the original analysis. If we split them up, visually we can already see that:



Then computing the individual sample correlations reveals 0.065 for females, and 0.097 for males. These small values that would not allow us to with the null hypotheses "female height is not related to income", and "male height is not related to income" (simulation as done above would confirm that).

So gender seems to be a confounding factor. Taking it into account reveals evidence is lacking to show the correlation is not zero.

8.7 Linear relationships between variables

We have seen that the Pearson sample correlation measures the extent of a linear relationship between two variables and that the Spearman correlation uses ranks to measure possibly non-linear relationships. Note that both work only with two variables.

Correlation doesn't tell us how to predict the post moult size from the pre-moult size, for example. Nor does it allow us to ask if post-moult size just pre-moult plus a constant.

If post-moult size (y) were just pre moult size (x) plus a constant then we could write the *rule* (or *equation*)

$$y = x + a,$$

and the sample correlation would be 1.

Also if post-moult size were pre-moult size plus 10% we could write

$$y = 1.1x$$

and the sample correlation would also be 1.00. Even if $y = 2x$ the sample correlation would also be 1. We need different tools to look at these relationships, all to be discussed in the next lecture.

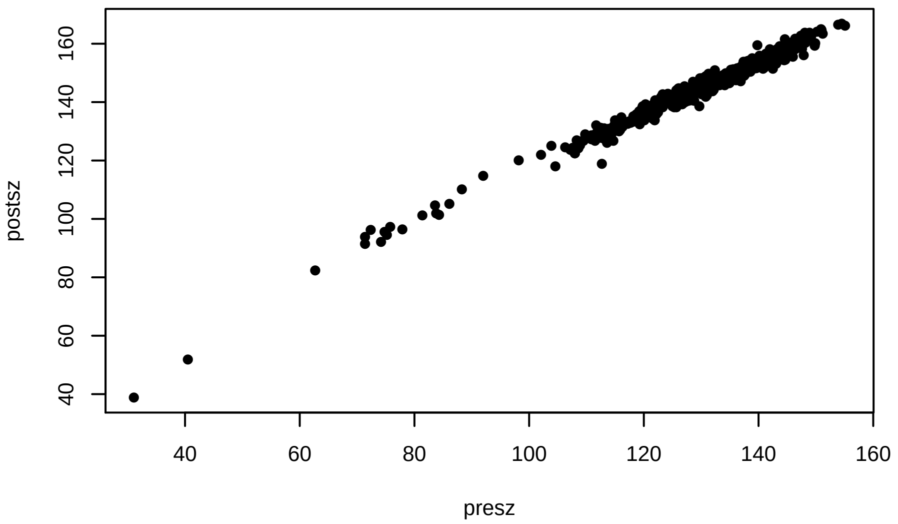
8.8 Summary

- Scatter plots examine the relationship between two quantitative variables.
- The strength of the relationship can be measured using correlation.
- Pearson correlation measures the linearity of the relationship, Spearman correlation the linearity of the ranks.
- The correlation statistic is between $[-1, 1]$, where 0 implies no correlation.
- We can test if the population correlation is not zero using permutations.
- We can compute the population correlation confidence interval using bootstrap samples.
- Correlation does not imply causation, and confounding variables can cloud the picture.

Lecture 9

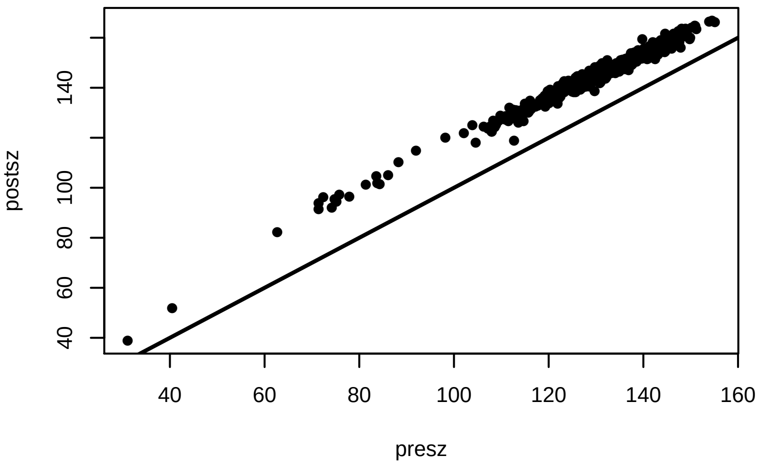
Simple linear regression

Recall that the moulting crabs scatter plot shows close to a straight line:



The correlation measured the strength of this linear relationship. What is is unclear from this plot is if the difference in before and after moulting size is a constant increase (e.g. increase of 5mm no matter what the before size is), a percentage increase (e.g. increase of 10% of the before size), or a combination of both.

If we add a line where pre-moult would equal post-moult:



It would appear that the points are just lifted up above the line by a constant amount.

This lecture is concerned with estimating the relationship between the variables, in particular a linear relationship. This will enable us to analyse if the above observation is a good explanation or not.

9.1 Straight Lines

You probably recall from high school mathematics the concept of a (straight) line connecting x and y, expressed by

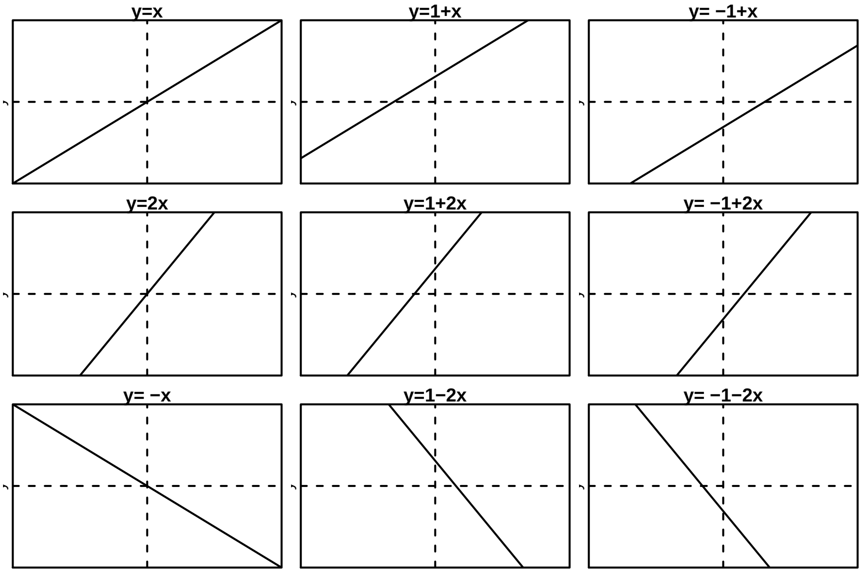
$$y = a + bx.$$

It involve two parameters: the slope b and the intercept a.

It captures that for every x, the corresponding y is b times x plus a. The parameters a and b are constants. They do not depend on x or y).

Most line can be described this way. A vertical line cannot however (as they are described by $x = d$ for some constant d).

Here are some example lines:



9.2 Simple linear regression

Suppose we are given a particular value x. Note that, x is given and not a random variable.

Let Y be random variable. In the crabs case, x is the pre-moult size, and Y the post-moult size.

In simple linear regression, we examine the relationship that

$$Y = a + bx + \varepsilon,$$

where $\varepsilon \sim N(0, \mu)$ is called noise. It is a normal random variable with mean 0 and variance μ.

The game is to predict a reasonable instance y given x. Given the randomness involved, we cannot hope to guess y perfectly. Instead, we want to predict a typical value given x. Namely, we will look at estimating the $E[Y \mid x]$. From here on in, let $y = E[Y \mid x]$.

Given some data, the goal is then to **estimate** good values of a and b.

y is called the dependent variable or response. x is called the independent variable or predictor. The line is a model that attempts to predict y from x.

Given pairs of data (x_i, y_i) for $i = 1, \ldots, n$, the data usually does not sit on a straight line, but might be close to it.

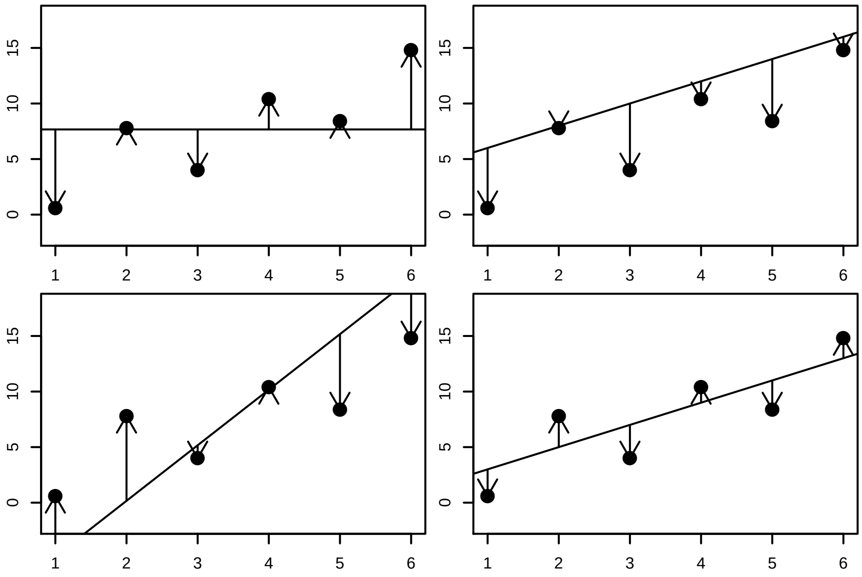
In practice, we need to allow for the **error** around the line.

$$y = a + bx + \varepsilon$$

or

$$y_i = a + bx_i + \varepsilon_i$$

There are many values of a and b that we could use that look close the data, So which line should we choose?



We want to determine the **best fit line**, for some definition of "best".

Any line has the equation $y = a + bx$ for constants a and b. So for **any** given constants a and b each x_i gives rise to a fitted value $\hat{y}_i = a + bx_i$ say.

We let $e_i = y_i - \hat{y}_i$ be the difference between the actual observed y_i and that **predicted** by the line $\hat{y}_i$.

e_i^2 measures the squared distance of the line defined by a and b from the observation y_i and $\sum_{i=1}^n e_i^2$ measures the squared distance of the line to all the observations.

The e_i are called **residuals** and the sum of them squared, $\sum_{i=1}^n e_i^2$ is the **residual sum of squares** or **RSS** for short.

So we can define the best fit line to be given by the choice of a and b the minimises the RSS

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

To minimise RSS over a and b, we need some calculus. The least squares estimates of a and b are given by

$$\hat{b} = \frac{\mathrm{cov}(x, y)}{\mathrm{cov}(x, x)} \text{ and } \hat{a} = \bar{y} - \hat{b}\bar{x}$$

where $\bar{x}$ and $\bar{y}$ denote the mean of x and y respectively.

We can thus see the relationship of the slope to the Pearson correlation r.

Moulting crabs

For the crab moulting data with `presz` as x, `postsz` as y

$$\begin{array}{llll} \bar{x} & = 129.21 & SS_{XX} & = 118542.69 \\ \bar{y} & = 143.9 & SS_{YY} & = 100957.55 \\ n & = 472 & SS_{XY} & = 108343.84 \end{array}$$

so that

$$\hat{b} = 108343.84/118542.69 = 0.914$$

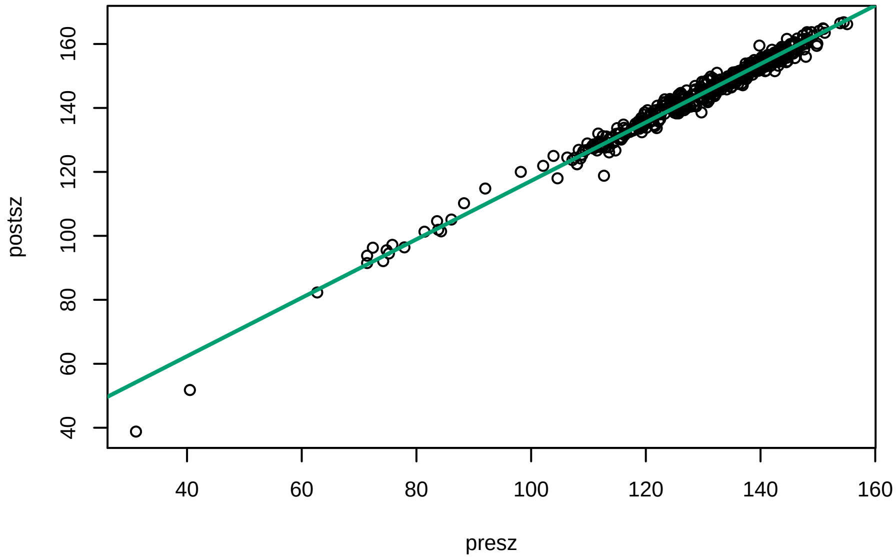
and

$$\hat{a} = 143.9 - 0.91 \times 129.21 = 25.803$$

As is invariably the case, `R` can do all the calculations for us.

```
> lm(postsz ~ presz, data=molt)
```

```
Call:
lm(formula = postsz ~ presz, data = molt)

Coefficients:
(Intercept)        presz
     25.803        0.914
```



9.3 Slope and Intercept

The slope and intercept have particular interpretations. The slope represents the amount by which y increases for every unit increase in x. For the crabs the slope is 0.914 so that (on average) for every mm the crab is larger (than average) pre moulting it is 0.914 mm larger (than average) post moulting

Another way to write the **regression** line is $(\hat{y}_i - \hat{y}_j) = \hat{b}(x_i - x_j)$

The intercept is the value of y when x is zero. Sometimes this is meaningless. For the crabs, a pre-moult size of zero is nonsense. However, the post- moult size of such a crab would be 25.803mm

In combination, the slope and intercept tell us how to compute the **expected** y value for a given x value. Using the estimated values $\hat{a}$ and $\hat{b}$, we can estimate the expected value of y given x.

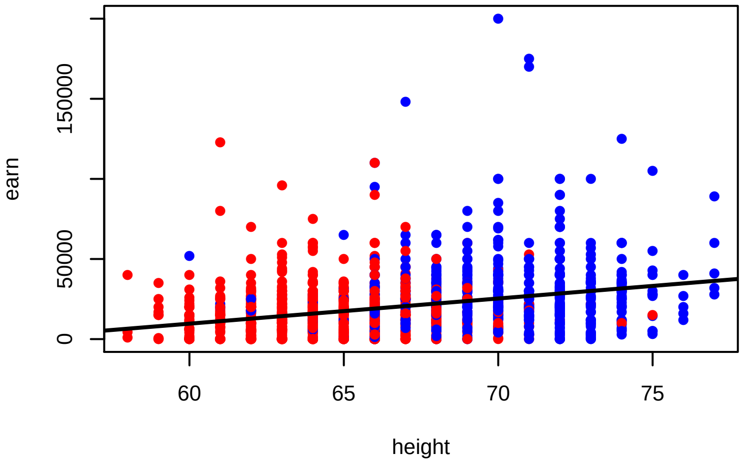
For example, using $\hat{a} = 25.803$mm and $\hat{b} = 0.914$, for a crab of pre-moult size $x = 120$mm the predicted post-moult size is
$\hat{y} = 25.803 + 0.914 \times 120 = 135.48$mm.

The fact that the slope here is less than 1, suggests that larger crabs actually grow by a smaller amount on average than the smaller crabs. Is this a sampling issue or true of the population?

Using the permutation approach we can easily look for evidence against the hypothesis $b = 0$ (slope is zero). For the crabs moulting data it is more interesting look for evidence against the hypothesis $b = 1$ (slope is one). We will return to this.

For the heights and earnings data though we are interested if the slope is zero. A slope of zero would mean heights do not affect earnings. For this data the least squares slope is $\hat{b} = 1571.05$ and the intercept is $\hat{a} =$ -84633.92.

Here is how the earnings slope looks:



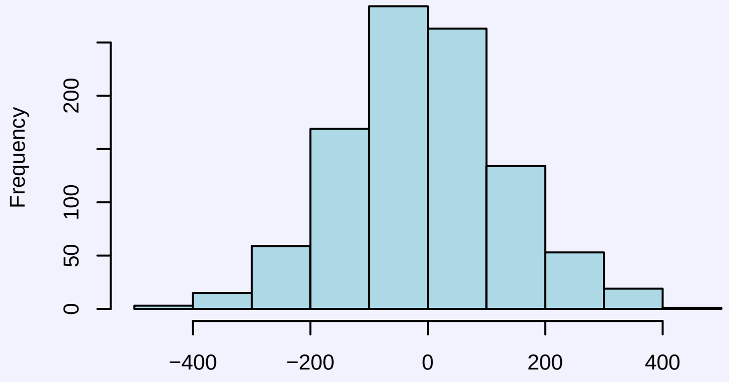
The slope of the line is 1571.05 and intercept is -84633.92. This means that for each inch taller, earnings are, on average, $1571 greater.

It also means oddly enough that someone of zero height earns minus $84,634 (of course that is a false interpretation!). That is the sample slope. Is the population slope though different from zero? For that, we can use the same permutation strategy as for correlation.

```
> ## Fit the linear model to the data.
> m = lm(earn~height, data=heights)
> ## Compute the slope of the model (coefficient 2).
> slope = coef(m)[2]
> x= replicate(1000, {
+   ## shuffle the height variable to simulate b = 0
+   height.perm = sample(heights$height)
+   ## Fit the linear model to the shuffled data.
+   m = lm(earn~height.perm, data=heights)
+   ## Compute the slope of the model (the second coefficient).
+   coef(m)[2]
+ })
```

Exercise

Below is the distribution of $\hat{b}$ when $b = 0$.



Recall that the data slope is $\hat{b} = 1571.05$. What is the p value and the conclusion of the test $(H_0 : b = 0,\ H_A : b \neq 0)$?

Back to crabs moulting

We are interested in testing if the population slope equals one. There are several ways to tackle this. The easiest is to modify the equation

$$postsz = a + b \times presz$$

to ressemble:

$$(postsz - presz) = a + (b-1)presz.$$

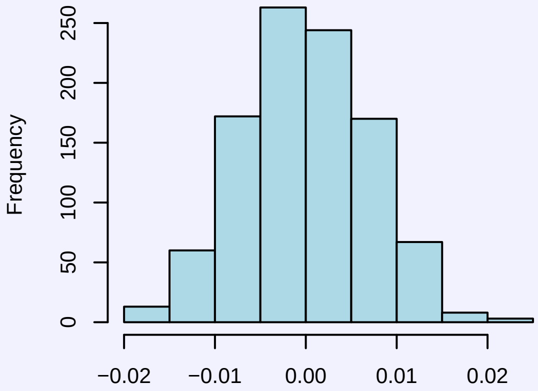
We can then fit y =(postsz-presz) against x =presz and test its slope against zero.

```
> slope = coef(lm((postsz-presz)~presz, data=molt))[2]
> x= replicate(1000, {
+   presz.perm = sample(molt$presz)
+   coef(lm((postsz-presz)~presz.perm, data=molt))[2]
+ })
```

The observed slope was -0.086.

Problem

Below shows the distribution of $\hat{b} - 1$, when $b = 1$.



Given that $\hat{b} - 1 = -0.086$ in the data, what is the p value and the conclusion of the hypothesis test ($H_0 : b - 1 = 0$, $H_A : b - 1 \neq 0$)?

Confidence Intervals

The hypothesis test showed strong evidence that the change in size during moult is not simply a constant. The evidence suggests that larger crabs grow by a smaller amount. To find a confidence interval for the slope we can use the bootstrap idea again.

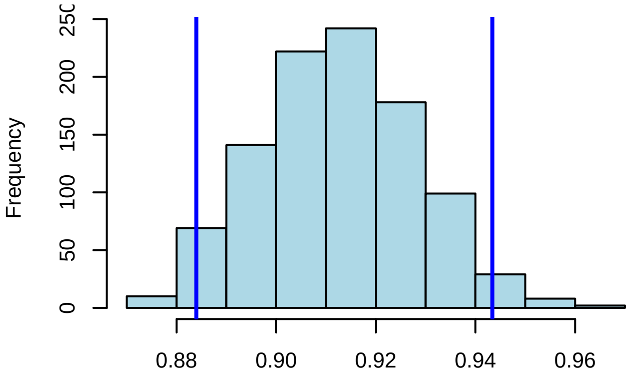
It is important that we sample pairs (x_i, y_i) of points to keep the relationship intact.

So simply, sample with replacement from the pairs of points, compute the slope and repeat, using the bootstrapped slopes to find a confidence interval in the usual way.

For the crab moult data:

```
> n = nrow(molt)
> slope = coef(lm(postsz~presz, data=molt))[2]
> x= replicate(1000, {
+   ## bootstrap sample the data rows.
+   samp = sample(1:n, replace=TRUE, size=n)
+   ## compute the slope of the bootstrapped rows.
+   coef(lm(postsz~presz, data=molt[samp,]))[2]
+ })
```

The observed slope is 0.914. The 95% confidence interval was [0.884, 0.943], as can be seen here:



Confidence intervals and hypothesis testing can be done in a similar way for the intercept. However, the slope is usually the main interest.

9.4 Residuals

When fitting a straight line there is usually a difference between the fitted value $\hat{y}_i$ and the actual data y_i. Remember we modelled the response y_i as a function of the x_i using

$$y_i = a + bx_i + \varepsilon_i.$$

The ε_i is the unknown bit not explained by the line. Once we have estimated $\hat{a}$ and $\hat{b}$ we can compute **fitted values** for each pair of data points.

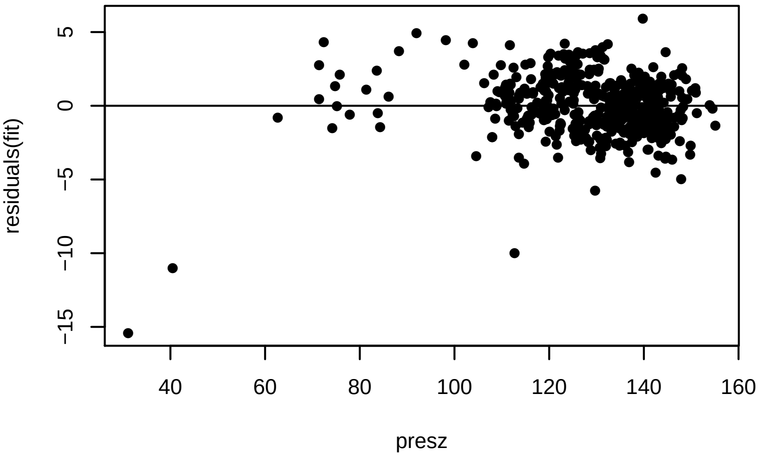
$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

The difference between the fitted values and the observed value is called the residual

$$e_i = y_i - \hat{y}_i$$

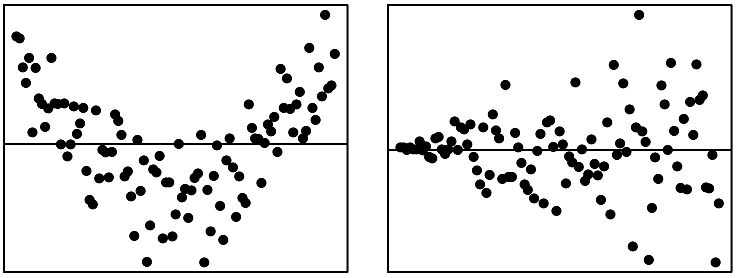
The residuals should be a more or less random scatter of points. To check the appropriateness of the model we can simply plot the residuals against their corresponding x values.

We are looking for any systematic variation. Here is the crab moult residual plot:



There are a few rather large negative residuals. Otherwise, there is no particular pattern.

Let's look at some typical problems found in residuals:

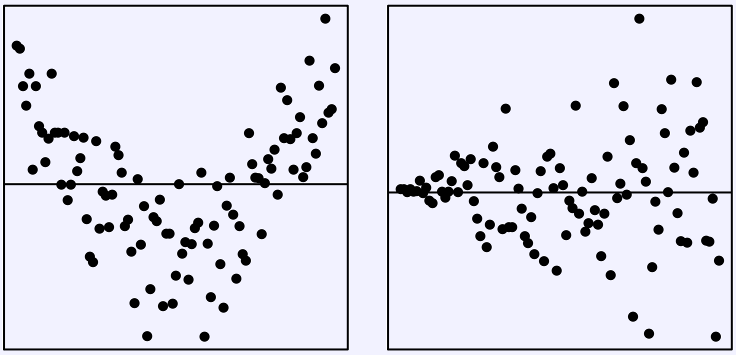


The left panel shows a set of residuals where the true model is NOT a straight line. The line under-predicts at the left and right, and over predicts in the middle.

The right panel shows a set of residuals that **fan out** to the right. This indicates that the variability depends on the x value, and simple least squares is not appropriate

Problem

Consider the following residuals:



Given an example of data that would provide such residuals, one for the left set and one for the right. Note that the horizontal line shows where the residuals are zero.

Residual Sum of Squares

The sum of the residuals squared is called the **residual sum of squares** or RSS for short.

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

This is what we minimised to estimate the slope and intercept.

The variance of the residuals can be estimated using the formula

$$s^2 = \frac{RSS}{n-2}.$$

R-squared

Another important (but often mis-used) quantity is the R-squared — often written R^2. It represents the proportion of variation in the y variable explained by the regression on the x variable.

The total variation in y is variation ignoring x, also called the total sum of squares, is

$$\text{cov}(y, y) = SS_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

The RSS is the residual variation, so

$$R^2 = 1 - \frac{RSS}{SS_{Total}}.$$

In simple linear regression, $R^2 = r^2$ where r is the Pearson correlation. Let's calculate this for the crab moult data. The sample size is $n = 472$, the RSS is 1935.09, and the variance of the residuals is $s^2 = 4.1172$. Then the total sum of squares is $SS_{Total} = 100957.55$, and thus $R^2 = 0.9808$.

R^2 is often expressed as a percentage. People often interpret the R^2 as a measure of quality of the model. But a model with a low R^2 may still be useful. If the slope is significantly different from zero, the regression contains some predictive ability.

9.5 Prediction

As we wrote before, given an estimate slope and intercept, $\hat{a}$ and $\hat{b}$, we can compute a fitted $\hat{y}$ for any x.

$$\hat{y} = \hat{a} + \hat{b}x.$$

x does not have to be **in** the original data, but it could be. $\hat{y}$ is a *prediction* of the expected value of Y at that value of x.

Predicted post moult sizes at pre moult sizes of 120, 140 and 160mm.

$$\begin{aligned} 25.803 + 0.914 \times 120 &= 135.48mm \\ 25.803 + 0.914 \times 140 &= 153.76mm \\ 25.803 + 0.914 \times 160 &= 172.04mm \end{aligned}$$

Confidence Interval for the mean of a predicted value

Again we can use our bootstrap technique to find a confidence interval for the predicted mean, and proceed as follows:

1. Generate a bootstrap sample of pairs of data
2. Fit the regression
3. Make the prediction
4. Repeat many times and construct an interval

For the crab moulting data at 120 the actual prediction is 135.48mm. A 95% confidence interval is 135.09, 135.88 mm.

9.6 Summary

- Simple linear regression fits the model $y = a + bx$ to the data by computing estimates of coefficients a and b.
- $\hat{b}$ is the estimate of the model slope (increase in y as x increases by 1).
- $\hat{a}$ is the estimate of the model intercept (expected value of y when $x = 0$).
- The best line is determined by least squares between the model line and the data.
- We can test if the slope $b = 0$ or any other value (e.g. $b = 1$).
- We can also compute the confidence interval for b.
- Examining residuals shows if the model is appropriate.

- R^2 measures the goodness of fit of the model.
- We can use the model to compute the expected y for a given x and provide a confidence interval.

Lecture 10

T-tests

In all previous lectures we used simulation, permutation tests and bootstrapping to obtain estimated p-values, confidence intervals etc.

Throughout this unit we have looked at computational ways (using randomisation) to conduct hypothesis tests and estimate confidence intervals in various situations. We will come at these now from a normal theory stand-point.

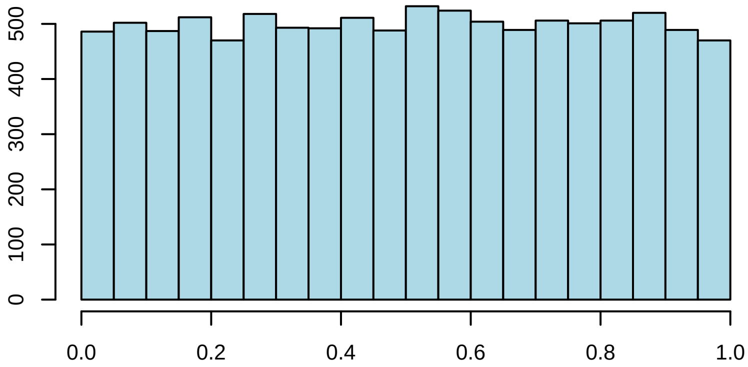
Simulation for hypothesis testing and confidence intervals is a relatively new technique in statistics. Traditionally, when computers could not be used, normal theory was the mainstay of statistical inference, and remains so for good reasons. Simulation does not return the same value twice while normal theory does, and in some cases in it is more robust, in particular for low sample sizes. It is also the case that a lot of variable on a population are approximately modelled by normal distribution – such as height in the general population. And, as we shall see, even when an original random variable is not normal, the average of a combination of them can be well modelled by a normal distribution.

So for these final lectures we will focus primarily on normal theory. In this one, we will first get to understand what the central limit theorem tells us, then look at a popular statistic called the t-test, before look at the relationship of the χ^2 statistic to its corresponding χ^2 statistic. We shall in particular look at paired data (connected data sets), and its use of the paired t-test.

10.1 Central Limit theorem

As we have previously seen, the normal distribution is useful for approximating probabilities of the binomial distribution. That was an even more powerful result before computers came along. The normal distribution has a much more central fundamental role, as we now illustrate.

Recall the continuous uniform distribution on $[0, 1]$. Here is a histogram of sampling from such a uniform distribution:



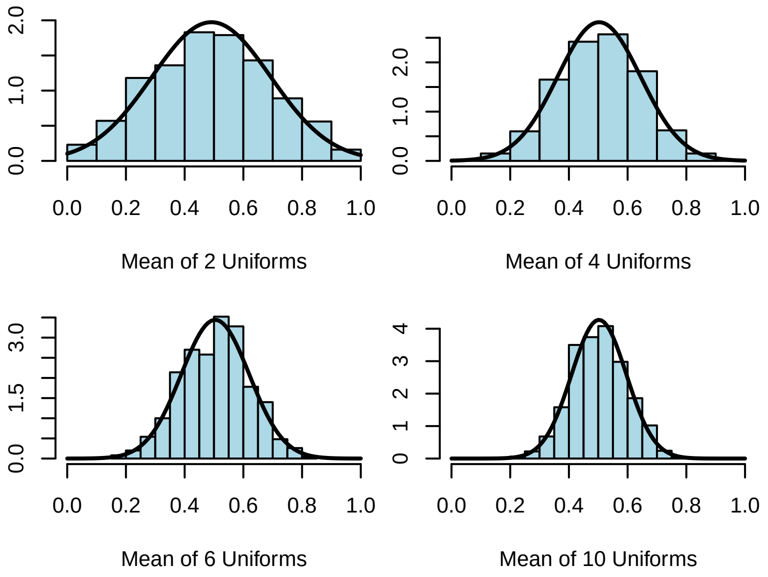
Now suppose we have n independent, identically distributed (written iid for short) random variables

$$X_1, X_2, \ldots, X_n \sim \mathcal{U}(0,1)$$

(in other words n independent uniform random variables on $[0,1]$), and we average them:

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

Let's do a simulation, for different values of n:

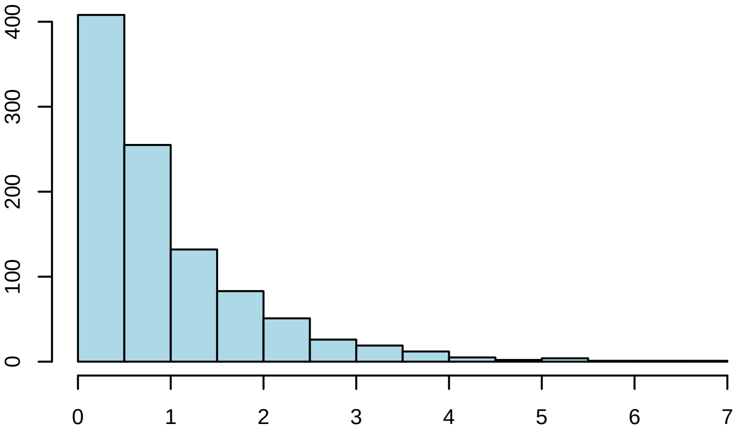


As our number of samples get larger we start to see a trend. To convey this trend, we have superimposed a black curve on each histogram. It is converging to what I hope you recognise as something resembling a normal curve.

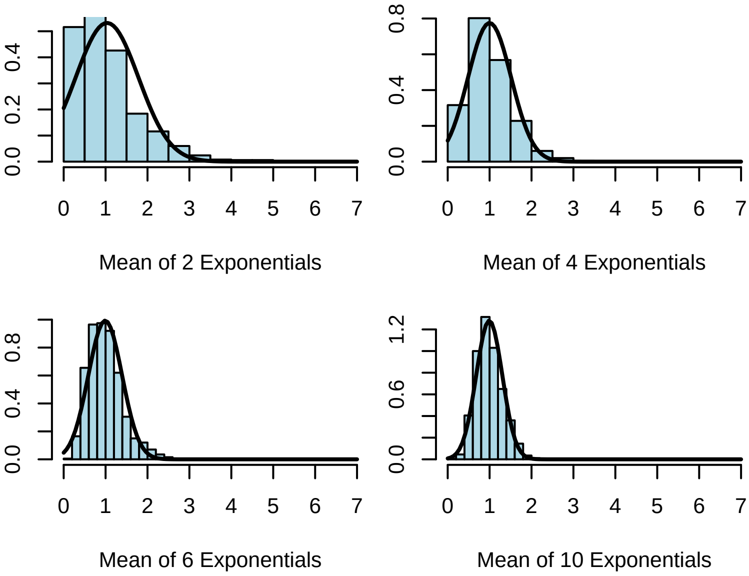
Could this be a fluke? Let us look at a somewhat trickier distribution, the exponential distribution $\mathcal{E}(\lambda)$ with pdf:

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

So we simulate to have an idea of its structure:



As you can see it is highly skewed. Again let's look at the sum of n iid exponential random variables for various values of n:



As n gets bigger, with the exception of the left-hand side (non-negative values are not possible from the exponential distribution), the mean gets closer to resembling a normal curve.

It turns out this tendency towards a normal distribution holds in general, and can be formalised.

First, let's define what a sampling distribution is.

Sampling distribution

Let S be a statistic considered a random variable. Consider now n instances $S_1, S_2, \dots, S_n$. Then its *sampling distribution* is how those S_i are distributed.

Note that the sampling distribution is a function of the n times we sample it. As n gets large it gets closer to the population distribution of the statistic.

Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be iid[a] with mean μ and variance σ^2, and let

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

Then $\sqrt{n}(\bar{X_n} - \mu)$ converges (in distribution) to a Normal distribution as $n \to \infty$.

[a] independent and identically distributed

We are not going to look at how this can be proved, but focus on its significance. It means that for large enough number of samples n from any distribution, the mean will approximately follow a normal distribution.

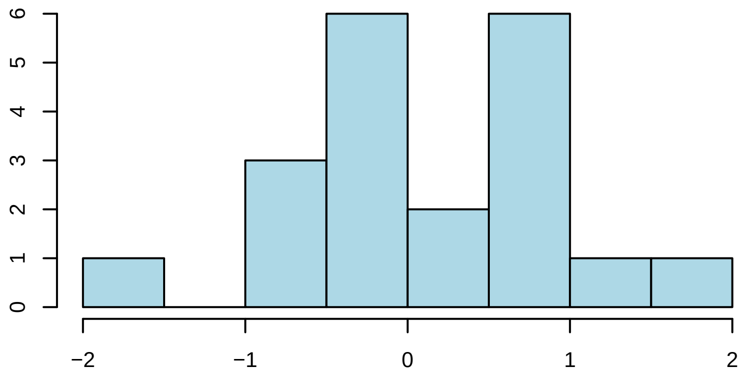
The theorem says that it will eventually converge but not the rate of convergence. There is a lot hidden in the concept of “large enough n”, and that value depends on the distribution of original X_i. For symmetric smooth distributions, it can already be a good approximation for quite small n, as low as 10. $n = 10$).

For skewed, bimodal or discrete distributions n needs to be larger. Often $n > 30$ is used as a rule of thumb.

10.2 Testing for normality

There are times when when we need to determine if data is normally distributed. Formal testing procedures exist for this. For example, the Kolmogorov-Smirnov or Shapiro-Wilk test. We will not cover those in this unit, and instead look at a graphical check.

The first thing that I hope occurred to you was to plot a histogram of the data. Something like this:



As you can see, the information is quite limited. Comparing a histogram to a normal density is pretty difficult unless you have a lot of data and small bins. A better approach is to use a *quantile-quantile plot*, also known as a *QQ-plot*.

The idea is to sort the data and plot it against what you should see if the data were exactly normally distributed.

Obtaining a QQ-plot

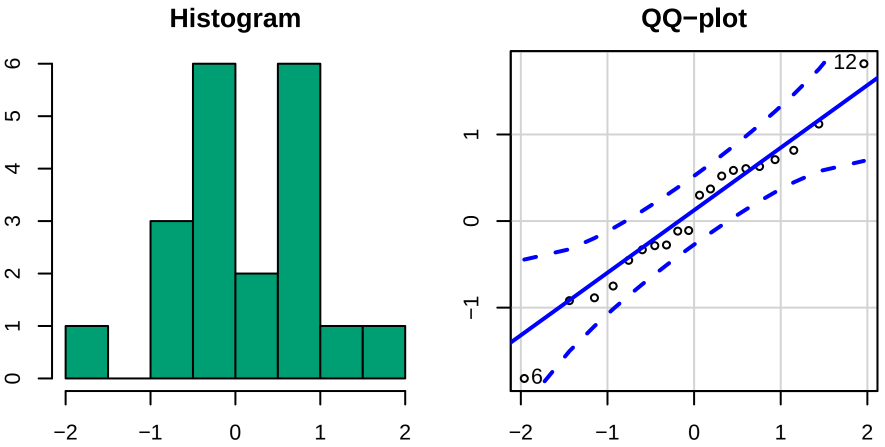
1. Let the data $y_1, \ldots, y_n$ be pre-sorted in increasing order. These divide the line into $n+1$ intervals: $n-1$ between each pair of ys and 2 at the ends.
2. Calculate a set of points $z_1, \ldots, z_n$ that divide the line into $n+1$ intervals based on equal normal probabilities such that by

$$P(Z < z_i) = \frac{i}{n+1}$$

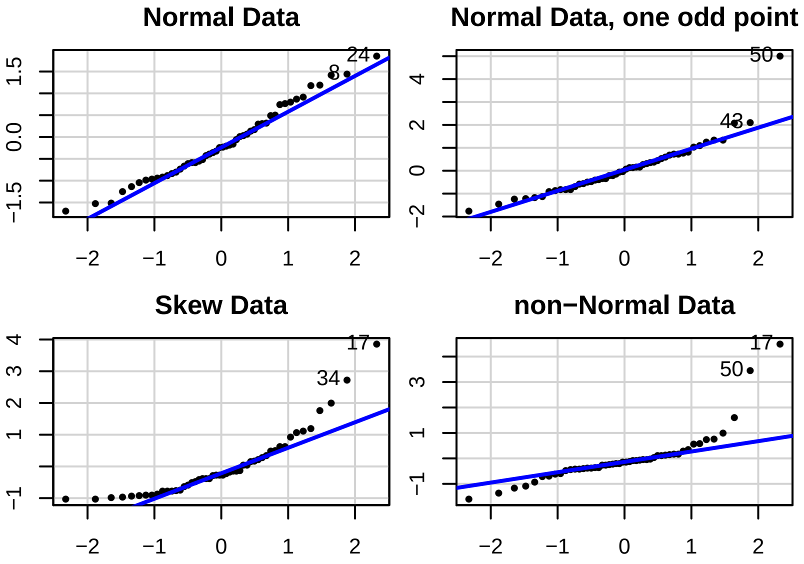
   where Z is standard normal.
3. Plot each y_i against the corresponding z_i.

If the data are normal it should be close to a straight line.

Here's a data set with a corresponding histogram and its corresponding QQ-plot:



Let’s look at some other examples:



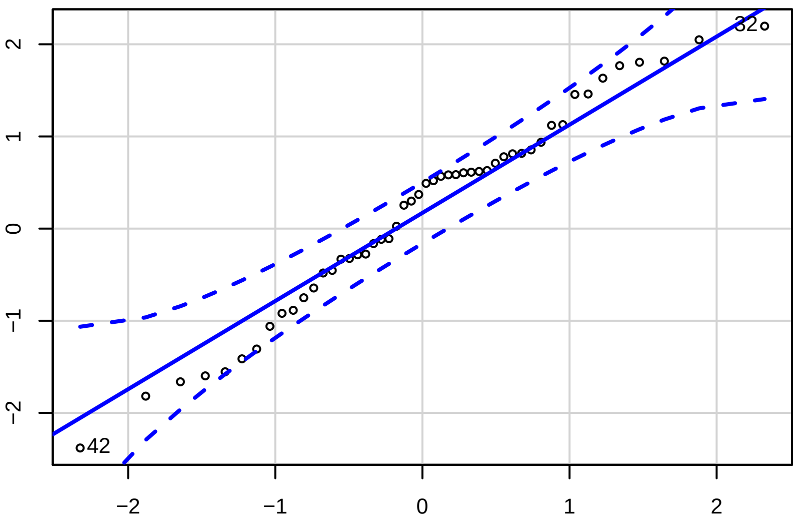
There are two ways to do a QQ-plot in R. The first works as follows:

```
> y = rnorm(50) # sample Normal values
> qqnorm(y) # plot the QQ-plot
> qqline(y) # add the line to the plot showing Normality
```

An alternative is to use the car add-on library:

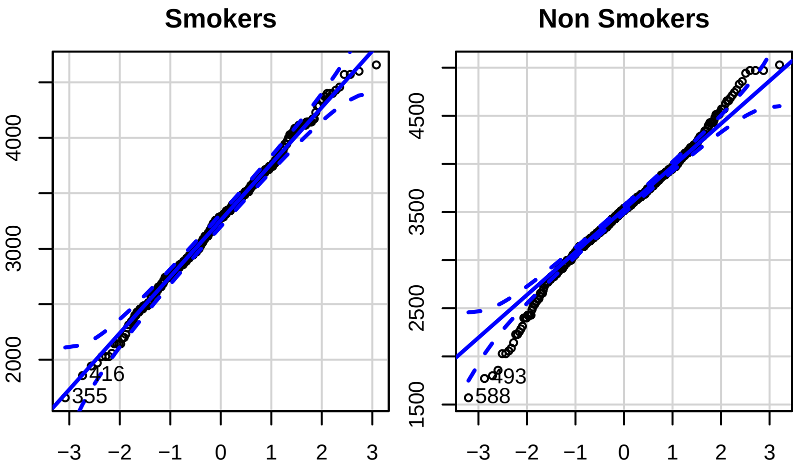
```
> library(car)
> qqPlot(y) # plot a QQ-plot with a line
```

In both cases, a guide line is added through the **quartiles** of the data. In the second type a **confidence envelope** is also added, while looks like:



Is the birth weight data normal?

Equipped now with our new technique, we can now examine if the birth weight/smoking data are normally distributed. We don't expect the smoking+non-smoking data combined to be normal because they have different means: it's two distributions on top of one another (unless the null hypothesis that the means had of been true of course!). It thus makes sense to look at the smoking and non-smoking groupings separately:



Non-smokers look a bit non-normal. However, the Central Limit Theorem tells us the mean distribution will be approximately normal.

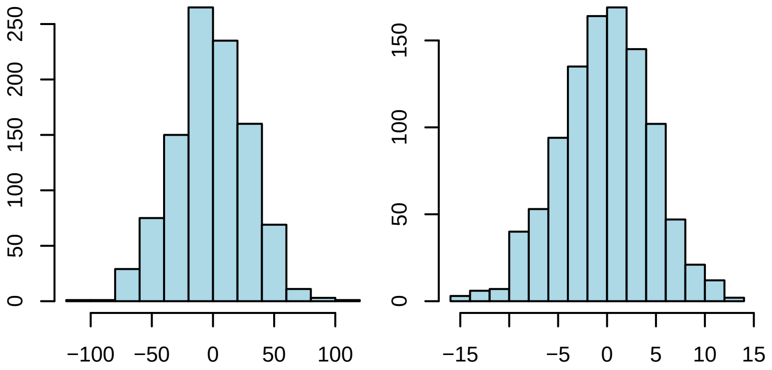
10.3 The t-test

The *t*-test was developed by William Gosset in 1908. Gosset worked for Guinness at the time, and needed to publish under the pen name Student for various political reasons. Hence the full name Student's *t*-test. The *t*-test is one of the most misunderstood/feared/abused methods in statistics. Actually a t-test is any statistical hypothesis test whose test statistic follows a Student's t-distribution under the null hypothesis. We will get to what that means shortly.

In this lecture, we will consider one-sided, two-sided and paired t-tests. In order to perform the t-test we first need to obtain the t-statistic.

Standard error

Below are the histograms of the differences in means for the randomised data sets (birth weight and store location).



Standardisation

Let $(x_1, \ldots, x_n)$ be data with mean $\bar{x}$ and standard deviation s.

1. Shift data by $\bar{x}$ so that its mean is 0.
2. Then scale by s to get a standard deviation of 1.

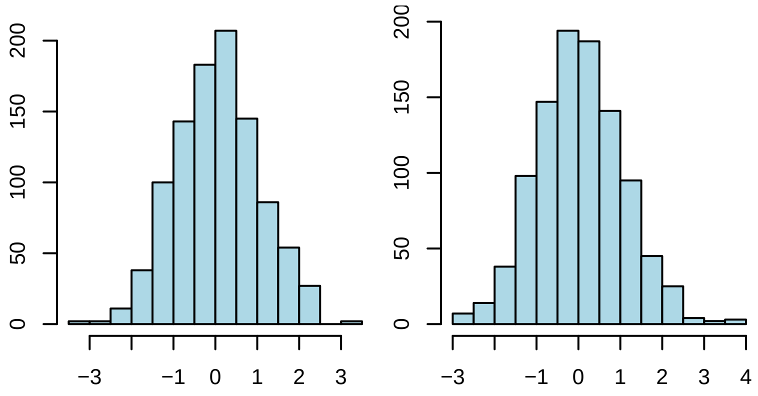
This gives standardised values:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Exercise

Prove that the set of all z_i have zero mean and standard deviation of 1.

Below are the same birth weight and sales histograms after standardisation.



After standardisation, the scales are very similar.

Now we define the notion of standard error.

Standard Error

1. The standard error (SE) of a statistic is the standard deviation of its sampling distribution.
2. When the statistic is the sample mean, it is called the standard error of the sample mean (SEM).

It's worth thinking how the SEM is different to the (population) variance. For a single sample of (say) infant birth weights, we can compute the mean as a measure of location. If we had a sample from different infants, we will in general get a different mean. How variable is that mean going to be. In the end, it depends on the variability (spread) of the data and the sample size.

We can estimate the SEM using a simulation such as calculating 1000 means and computing their standard deviation. In practise, the population mean is usually unknown, so we use an approximation.

SEM approximation

$$\text{SEM} = \frac{s}{\sqrt{n}},$$

where s is the data standard deviation and n the sample size.

One-sample t-test

Suppose have some data whose sample mean is $\bar{x}$, and our hypothesis is that $\mu = \bar{x}$. We wish to evaluate if it's true.

So we use the t-statistic:

$$t = \frac{\bar{x} - \mu}{\text{SEM}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where s is the sample standard deviation and n is the sample size.

The hypotheses that arise are:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0 \text{ etc.}$$

The t-statistic:

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

can be shown to follow what is known as a t-distribution with $n - 1$ degrees of freedom.

From that, we get a 95% confidence interval for a single mean

$$\bar{x} \pm t_{\alpha/2} s/\sqrt{n}$$

Where $t_{\alpha/2}$ is derived from a t-distribution with $n - 1$ degrees of freedom.

Two-sample t-test

In our maternal smoking problem, we are interested in seeing if there is a statistically significant difference in means. This is a two-sample problem. We are going to assume the underlying variances in the populations are the same[1]

Let n_1 and n_2 be the number of elements in each group, and s_1 and s_2 be the (sample) standard deviations of each group.

The t-statistic in this case is:

$$\frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

s_p is called the pooled standard deviation.

To compare two groups (e.g., smoking versus non-smoking mothers) we can compute this t-statistic, and could proceed as before: simulate using permutations, obtain a p-value, and decide if there is enough evidence to accept the alternative hypothesis.

There is another, standard way. The t-statistic has the property that if the null hypothesis is true, the t-statistic will follow a t distribution, for large enough sample size (at least 30).

This brings us to the **t-test**. Before we show that let's illustrate the t-statistic using our standard examples.

[1] If they are not, that requires Welch's t-test, not discussed in this unit.

Birth weight data

Let's look at how to computer the t-statistic for the difference in means of the birth weight data. Here's a summary again:

Summary	Smoke=No	Smoke=Yes
n	742	484
Mean	3515.64	3260.29
Standard Deviation	497.1	517.11

Then

$$s_p^2 = \frac{(742-1)497.1^2 + (484-1)517.11^2}{742+484-2} = 255114.55$$

and

$$s_p = 505.09.$$

Thus the t-statistic is:

$$t = \frac{3515.64 - 3260.29}{505.09\sqrt{\frac{1}{742}+\frac{1}{484}}} = 8.653$$

Sales data

Here's the sales data again:

Summary	Office=East	Office=West
n	52	48
Mean	162.7	154.04
Standard Deviation	22.81	21.93

Thus

$$s_p^2 = \frac{(52-1)\times 22.81^2 + (48-1)\times 21.93^2}{52+48-2} = 501.42$$

and

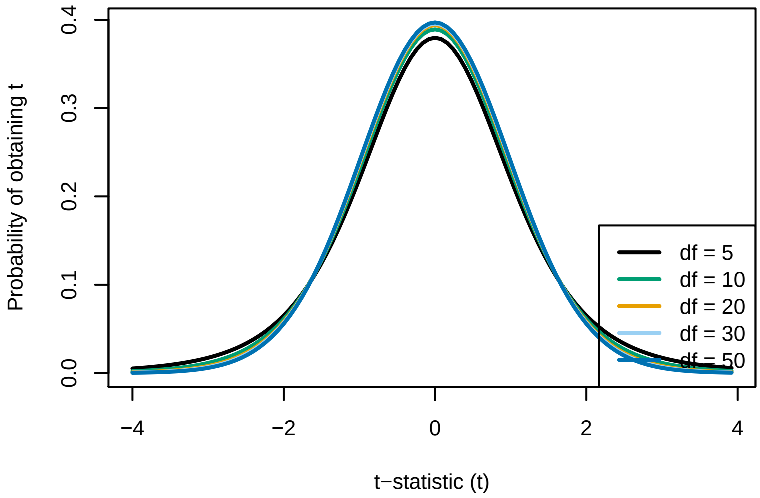
$$s_p = 22.39.$$

Thus the t-statistic is:

$$t = \frac{162.7 - 154.04}{22.39\sqrt{\frac{1}{52}+\frac{1}{48}}} = 1.931$$

The t-test

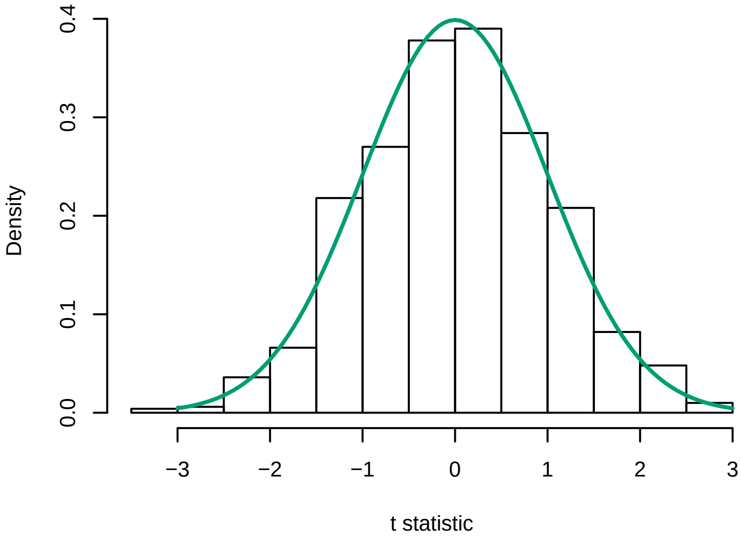
Here is the t-distribution. It has one parameter called degrees of freedom (df). In the two-sample test this is $n_1 + n_2 - 2$.



It represents the sample distribution the statistic will follow provided the null hypothesis holds and the population distribution of the statistic follows a normal distribution. Fortunately, the central limit theorem helps establish this as a reasonable assumption.

So to estimate a p-value for a given hypothesis we can use this distribution rather than simulating as before.

Let's compare using the permutations and t-distribution methods of the birth weight data:



This difference in means t-test is common, and has a function is provided in R to perform the complete set of calculations.

```
> t.test(bwt~smoke, data=birthwt, var.equal=TRUE)

    Two Sample t-test

data:  bwt by smoke
t = 8.6527, df = 1224, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 197.4554 313.2520
sample estimates:
 mean in group no mean in group yes
         3515.639          3260.285
```

For the office sales data, we get:

```
    Two Sample t-test

data:  sales by office
t = 1.9314, df = 98, p-value = 0.05632
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2379317 17.5511689
sample estimates:
mean in group east mean in group west
          162.6991           154.0425
```

10.4 Confidence intervals from a t-distribution

Previously we estimated confidence intervals using bootstrapping.

Just as with hypothesis testing, we can also approximate confidence intervals using the t-distribution.

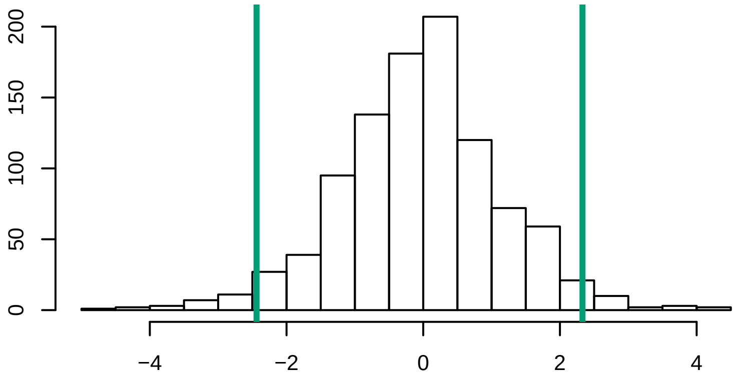
From the t-test above, provided the true difference in means is zero, then

$$\frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, provided the sample sizes are large enough (thanks to the guarantees of the central limit theorem).

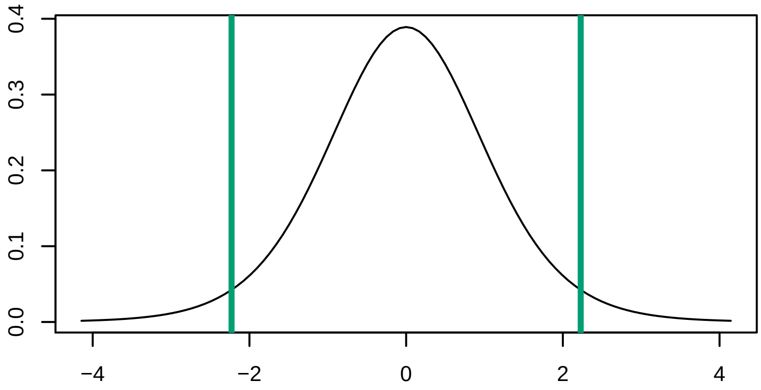
So instead of bootstrapping, we could simulate from the t-distribution, and find those points that have 2.5% less and 2.5% greater.

For example, simulating from a t-distribution for 1000 simulations yields something like:



In fact, this is overkill. We know already the formula for the t-distribution so we can use it directly.

Here is the 95% interval of t distribution with df $= 10$:



The middle interval of a t distribution will always be plus and minus the same number because of symmetry of the t distribution.

The dark lines haves 2.5% of the t-distribution to left and 2.5% to right. For 10 degrees of freedom, these values are -2.228, 2.228. We often write this number as $t_{0.025,10}$, which is in this case 2.228.

It means that 95% of the time a t-statistic with 10 degrees of freedom, is between $-t_{0.025,10}$ and $+t_{0.025,10}$, or more generally 95% of the time:

$$-t_{0.025,(n_1+n_2-2)} < \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{0.025,(n_1+n_2-2)}.$$

A bit of manipulation gives that 95% of the time:

$$(\bar{x}_1 - \bar{x}_2) - t_{0.025,(n_1+n_2-2)}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{0.025,(n_1+n_2-2)}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Birth weight and sales data confidence intervals

For the birth weight data we get:

```
    Two Sample t-test

data:  bwt by smoke
t = 8.6527, df = 1224, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 197.4554 313.2520
sample estimates:
 mean in group no mean in group yes
         3515.639          3260.285
```

while for the sales data we have:

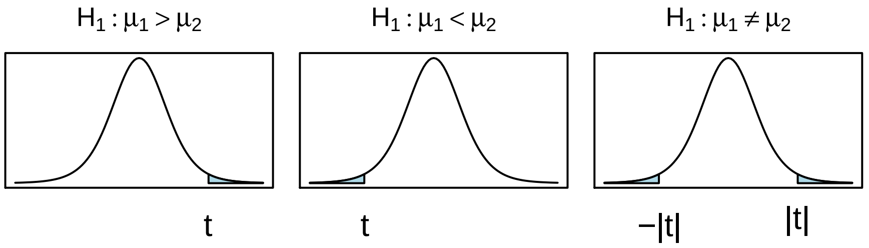
```
    Two Sample t-test

data:  sales by office
t = 1.9314, df = 98, p-value = 0.05632
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2379317 17.5511689
sample estimates:
mean in group east mean in group west
          162.6991           154.0425
```

Two-sample t-test summarised

- $H_1 : \mu_1 > \mu_2$ `1-pt(t, df)`,
- $H_1 : \mu_1 < \mu_2$ `pt(t, df)`,
- $H_1 : \mu_1 \neq \mu_2$ `2*(1-pt(abs(t), df))`

`t.test(x~grp, var.equal=TRUE)` or `t.test(x1,x2, var.equal=TRUE)`

The area representing the p value for each hypothesis test is then:



Confidence Interval for difference in means

A 95% confidence interval for the actual difference in means is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{1/n_1 + 1/n_2}$$

where $t_{\alpha/2}$ is derived from a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

For 95% $\alpha = 0.05$, `t = qt(1-0.05/2, df)`

10.5 Paired data

Paired data is an example of blocking, where two measurements are taken on each block. The analysis is simplified because there are only two measurements. This data comes from http://www.statsci.org/data/oz/nzhelmet.html.

After purchasing a batch of flight helmets that did not fit the heads of many pilots, the NZ Airforce decided to measure the head sizes of all recruits.

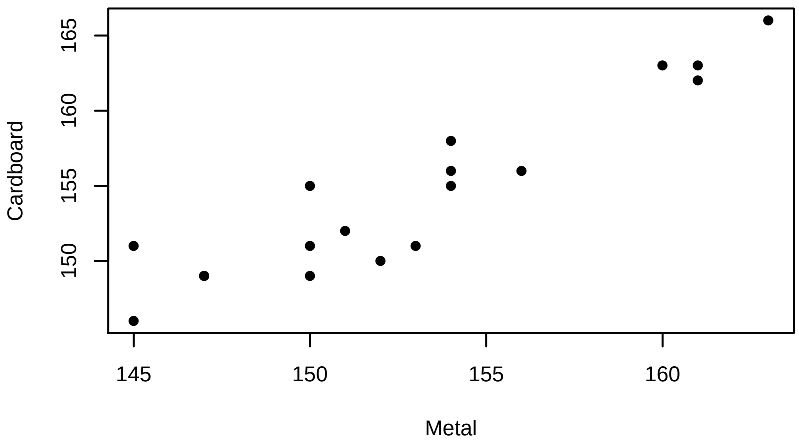
Before carrying this out, information was collected to determine the feasibility of using cheap cardboard callipers to make the measurements, instead of metal ones which were expensive and uncomfortable. The data lists the head diameters of 18 recruits measured once using cardboard callipers and again using metal callipers. The question was whether there is any systematic difference between the two sets of callipers.

This data is different to the birth weight data in that every head is measured twice: once with cardboard and once with metal callipers.

Here's an extract of the data

```
  Cardboard Metal
1       146   145
2       151   153
3       163   161
4       152   151
5       151   145
6       151   150
```

Let's visualise that:



The data are paired. As we shall see, ignoring calliper type across all measurements can lead to the inability to deduce a hypothesis. We benefit by respecting the pairing.

Paired t-statistic

To use a standardised t-statistic with paired data we first take differences between the pairs.

$$d_i = x_i - y_i$$

If there is no difference the d_i would have mean zero. So a t-statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $\bar{d}$ is the mean and s_d is the standard deviation of the d_i.

We could obtain an p-value by simulation as follows.

```
> d = helmets$Cardboard - helmets$Metal
> n = length(d)
> t.stat0 = mean(d)/(sd(d)/sqrt(n))
> x = replicate(1000, {
+   s = sample(c(-1,1), replace=TRUE, size=n)
+   mean(s*d)/(sd(s*d)/sqrt(n))
+ })
```

In one case we got p = 0.009.

That's not really necessary as we now know, and just using the t-distribution can give us an estimated p-value.

```
> t.test(helmets$Cardboard, helmets$Metal, paired=TRUE)


    Paired t-test

data:  helmets$Cardboard and helmets$Metal
t = 3.1854, df = 17, p-value = 0.005415
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5440163 2.6782060
sample estimates:
mean of the differences
               1.611111
```

Paired data versus unpaired data

What would have happened if we “unpair” the data? We could have measured a set of heads with cardboard and a different set of heads with metal callipers. Then the random variation in head size would be part of the variation in measurements, as well as the difference in measurement methods.

size	calliper
153	Metal
163	Metal
155	Cardboard
154	Metal
160	Metal
151	Cardboard
147	Metal
163	Cardboard
150	Metal
154	Metal

Let’s try it out:

```
> t.test(size~calliper, helmetLong, var.equal=TRUE)
```

with output:

```
    Two Sample t-test

data:  size by calliper
t = 0.85076, df = 34, p-value = 0.4009
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.237425  5.459647
sample estimates:
```

```
mean in group Cardboard     mean in group Metal
               154.5556                152.9444
```

That's a p-value of 0.40 – we aren't able to reject the null hypothesis by this method. The difference between calliper measurements is washed out amongst the variability in head sizes.

The unpaired version of the test, while not technically invalid, cannot detect a difference. Using the *blocks* (individual heads) and tailoring the analysis gives much more ability to detect the difference in callipers. This ability is called **power**.

Type I+II errors

When testing a *null* H_0 and *alternative* H_1 hypothesis we can make two types of error.

- Type I error — finding evidence **against** the null although it is actually **true**
- Type II error — failing to find evidence **against** the null when it is actually **false**

Type I error is controlled by the threshold at which the p-value is said to be small enough to find evidence. If we choose 5% and say there is evidence against the null (because the p-value is smaller than this) then the probability of type I error is 5%.

The probability of Type II error is harder to control, but larger sample sizes usually give smaller probabilities of type II error.

Power

$$\text{power} = 1 - \text{Prob(type II error)}$$

It is the probability of finding evidence **against** the null when it is actually **false**.

Usually we choose the way of testing that gives most power. If a test makes more assumptions, and they are correct, that test usually has more power. We often determine the sample size so that we have adequate power (say 80%).

10.6 The χ^2 distribution

Recall our deployment of the χ^2 statistic. Given a set of counts in K categories, $x_1, \ldots, x_K$ we want to see whether these could have come from a given set of probabilities $p_1, \ldots, p_K$.

If q_k are the probabilities that generated the counts; and $N = \sum_{i=1}^{K} x_k$ is the total number of values, then the hypothesis become:

$$H_0 : q_k = p_k \quad k = 1, \ldots, K$$

$$H_1 : q_k \neq p_k \text{ for some } k$$

with the test statistic given by:

$$\chi^2 = \sum_{k=1}^{K} \frac{(x_k - Np_k)^2}{Np_k}$$

To compute the distribution of the χ^2 statistic given that H_0 is true, we previously used permutations of the data.

If the expected frequencies are at least 5, this distribution can be approximated using a χ^2 distribution with $K - 1$ degrees of freedom.

So, if we compute the χ^2 statistic, the p-value can be computed using:

$$p = 1 - \text{pchisq(X2, K-1)}.$$

In R, we can compute the χ^2 statistic for the data, then compare it to the χ^2 distribution using `chisq.test(x, p=p)`.

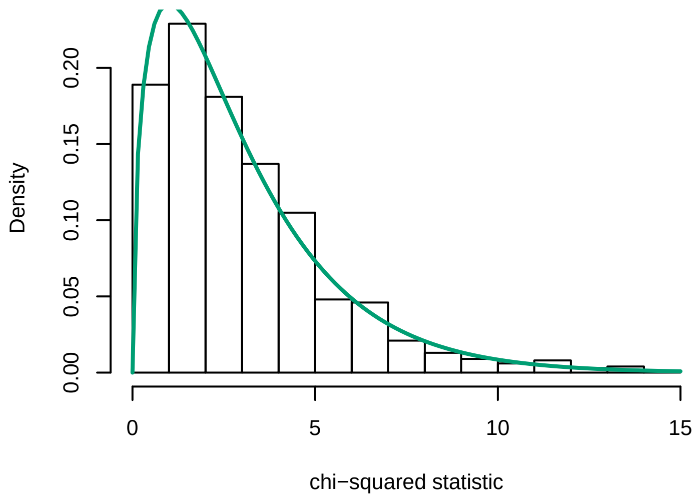
Example: χ^2 goodness of fit — Iraqi Refugees

Compare the counts of Iraqi refugees in various K10 categories to the Australian population

low	moderate	high	very high
123	70	93	157

low	moderate	high	very high
70.65	18.5	7.41	3.43

Let's compare a permutation simulation histogram and using the χ^2 distribution:



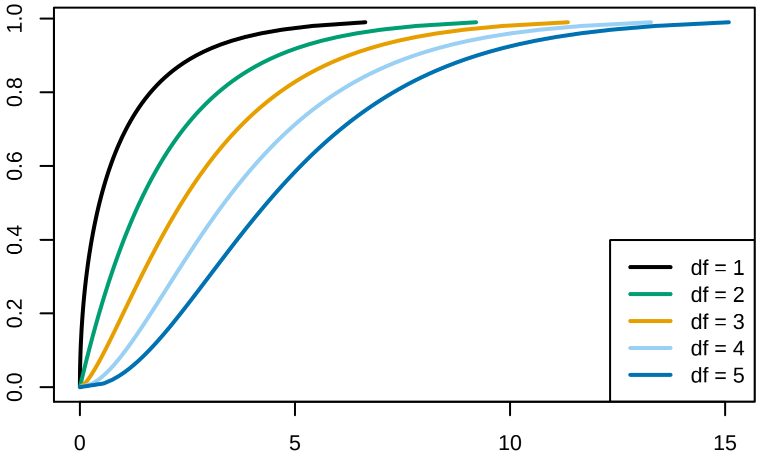
For the Iraqi refugee dataset, we can calculate a p-value estimate using the χ^2 distribution by:

```
> chisq.test(iraqi, p=aihw)
```

The output is then:

```
    Chi-squared test for given probabilities

data:  iraqi
X-squared = 1550.1, df = 3, p-value < 2.2e-16
```

It lets us compute the proportion of times a value is exceeded if the null hypothesis is true. It is actually a family of distributions that depend on a parameter related to the number of counts being considered, and requires the extra assumption that the expected cell counts are not too small (at least 5). The χ^2 distributions with different degrees of freedom (df) looks like:



For the first example — comparing to a known distribution — the chi-square distribution df to use is *one less than the number of cells in the count table* — 3 for iraqi refugees.

The actual chi-square distance was 1550.08. Using the graph, the proportion less than this would be virtually 1. So the proportion greater (the p-value) would be virtually zero.

For the second example — comparing two sets of counts — the chi-square distribution df to use is *one less than the number of rows in the count table multiplied by one less than the number of columns in the count table* $(r-1)\times(c-1) = 2$ for the eels example.

The actual χ^2 distance was 6.26. Using the graph (or tables or a command in R/excel), the proportion less than this would be over 0.9 (90%). So the proportion greater would be under 10%.

In fact, the proportion greater (p-value) is 0.0437. In one simulation run we got $46/1000 = 0.046$, so pretty close.

Chi-square test

Refers to the independence of rows and columns in a table of counts.

$$H_0 : \text{Row and Column features are independent}$$

$$H_1 : \text{Row and Column features are NOT independent}$$

We compute the proportions in each column, p_i $(i = 1, \ldots, C)$ and multiply by the row totals r_j $(j = 1, \ldots, R)$ to get expected values, under assumed independence

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - r_j p_i)^2}{r_j p_i}$$

To compute the distribution of the χ^2 statistic given that H_0 is true, we previously used permutations of the data.

If the expected frequencies are at least 5, this distribution can be approximated using a χ^2 distribution with $(C-1)(R-1)$ degrees of freedom.

Example: χ^2 test of independence — Eel habitat

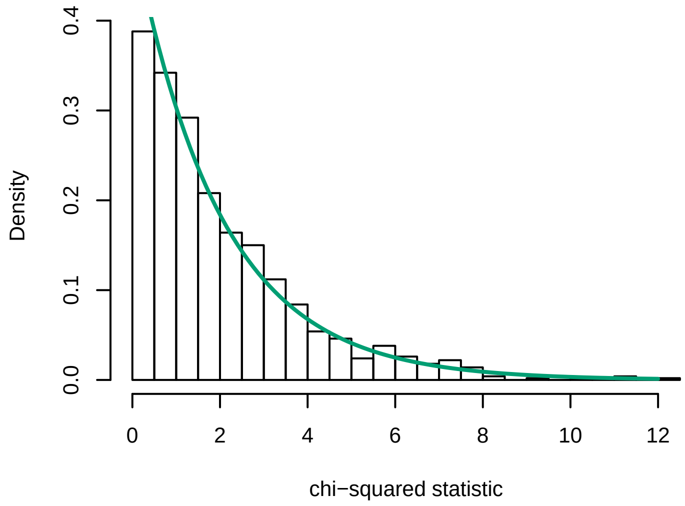
Is eel habitat independent of species?

	Border	Grass	Sand
G.moringa	264	127	99
G.vicinus	161	116	67

```
> chisq.test(eels)
```

```
	Pearson's Chi-squared test

data:  eels
X-squared = 6.2621, df = 2, p-value = 0.04367
```

We can see how the permutation simulation method and the χ^2 distribution are very similar:



10.7 Summary

When we have large sample sizes, we can use known distributions for the tests:

- χ^2 test for goodness of fit: χ^2 distribution.
- χ^2 test for independence: χ^2 distribution.
- Two sample t test and confidence interval for difference in means: t distribution.
- Paired t test and confidence interval: t distribution.
- Equality of a single mean: t distribution.
- Test and confidence interval for regression slope: t distribution.

- If n is large and p not small, then we can approximate a Binomial distribution with a Normal distribution.
- We can compute probabilities from a Normal distribution by measuring the area under the density curve.
- The Central Limit Theorem tells us that the sum of a set of random variables is Normal, where each variable follows the same distribution and are independent from each other.

Lecture 11

Do redheads have a lower pain threshold?

Last week we looked at using the t-test for comparing two means from two population groups. What happens when we have more? This lecture is about that generalisation.

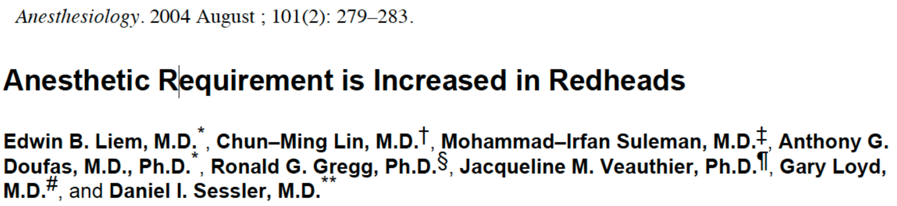
Previously we looked at the following scenarios:

- one qualitative variable — e.g., Iraqi Refugees.
- two qualitative variables — e.g., eel species and habitat.
- one quantitative and one qualitative variable — eg. Birth weight and smoking.
- two quantitative variables — e.g., height and income.

The "one quantitative and one qualitative variable" case focused on a binary qualitative variable. There were only 2 possibilities: smoking or non-smoking). This lecture looks at extending this to multiple levels in the qualitative variable.

The question and data driving us through is that of hair colour and if it plays a role in the ability to withstand pain. At first glance that might seem like an odd, perhaps absurd question. Of course (natural) hair colour, in particular redheadedness, on face value would seem to be a strange influencer of pain. However, one's hair colour is an outcome from multiple gene expressions, and thus may be a proxy for an underlying reason. More importantly for a data-driven unit like ours, there is evidence.

Redheads, on average, need more anaesthetic:

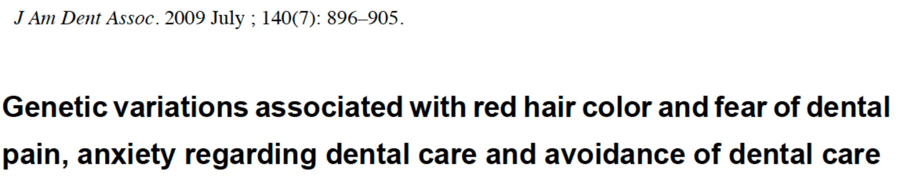


Anesthesiology. 2004 August ; 101(2): 279–283.

Anesthetic Requirement is Increased in Redheads

Edwin B. Liem, M.D.[*], **Chun–Ming Lin, M.D.**[†], **Mohammad–Irfan Suleman, M.D.**[‡], **Anthony G. Doufas, M.D., Ph.D.**[*], **Ronald G. Gregg, Ph.D.**[§], **Jacqueline M. Veauthier, Ph.D.**[¶], **Gary Loyd, M.D.**[#], and **Daniel I. Sessler, M.D.**[**]

fear dentists more:



J Am Dent Assoc. 2009 July ; 140(7): 896–905.

Genetic variations associated with red hair color and fear of dental pain, anxiety regarding dental care and avoidance of dental care

are more sensitive to thermal pain:

Anesthesiology. 2005 March ; 102(3): 509–514.

Increased Sensitivity to Thermal Pain and Reduced Subcutaneous Lidocaine Efficacy in Redheads

Edwin B. Liem, M.D.[*], **Teresa V. Joiner, B.S.N.**[†], **Kentaro Tsueda, M.D.**[‡], and **Daniel I. Sessler, M.D.**[§]

report more bodily pain:

Pigment Cell Melanoma Res. 2016 March ; 29(2): 239–242. doi:10.1111/pcmr.12445.

Natural hair color and questionnaire-reported pain among women in the United States

Wen-Qing Li[1,2], **Xiang Gao**[3,4], **Shelley S. Tworoger**[4,5], **Abrar A. Qureshi**[1,2,4], and **Jiali Han**[4,6,7]

and it is has been shown that redhead genetics affects pain sensitivity in mice[1] and humans:

Melanocortin-1 receptor gene variants affect pain and μ-opioid analgesia in mice and humans

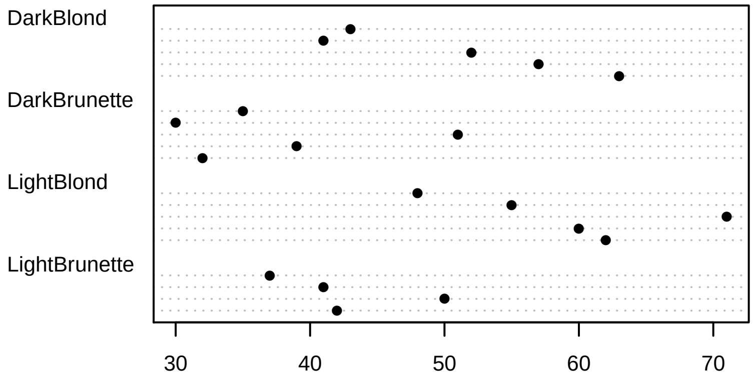
J S Mogil, J Ritchie, S B Smith, K Strasburg, L Kaplan, M R Wallace, R R Romberg, H Bijl, E Y Sarton, R B Fillingim, A Dahan

J Med Genet 2005;**42**:583–587. doi: 10.1136/jmg.2004.027698

Pain tolerance and hair colour

The chart above shows pain tolerance scores of 19 individuals as they vary with hair colour.

[1] You must have seen all those redheaded mice scurrying around.



This is called a dot chart, which is an alternative to box plots for small data sets. The question of interest here is “does pain tolerance vary by hair colour?”. So does this plot provide evidence for that?

We need to be more formal than this. More precise is to ask if *average* pain tolerance varies by hair colour.

Let’s put it in hypothesis terms. Let there be K groups each with population mean pain tolerance μ_k for $k = 1, \ldots K$. The question we then ask is if there is evidence that the μ_k are not all equal.

Mathematically this becomes

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : &\quad \mu_i \neq \mu_j \text{ for at least one pair } i, j. \end{aligned}$$

So we just need to have evidence of one pair who is different and the house of cards falls.

11.1 F-statistic

When we compared two groups (smoker vs. non-smokers), we computed a standardised difference between the group means, called a t-statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where the pooled variance is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Let’s see how to extend this difference from equal means for more than two groups.

First some notation. Let the data consist of K groups with each group k having sample size n_k.

Then let each group has a sample mean $\bar{x}_k$ and sample standard deviation s_k.

Now consider all data merged into one global group containing $n = n_1 + \cdots + n_K$ elements. Its mean, the global mean, is then

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k \bar{x}_k.$$

That's the average across groups. For a given group k, its square difference from the average $\bar{x}$ is $(\bar{x}_k - \bar{x})^2$. When we sum all those differences across the K groups, we get the variance between them, called the *sum of squares between groups*:

$$SS_B = \sum_{i=1}^{K} n_k (\bar{x}_k - \bar{x})^2.$$

This captures the variability in the group sample means when compared to the overall mean. If the population group means were equal, namely $\bar{x} = \bar{x}_k$ for all k, then $SS_B = 0$, which makes sense, as there would be no variation between groups.

Sum of squares within groups: To get a measure of the overall variation in groups, we need a function of all the standard deviations $s_1, s_2, \ldots, s_k$. This should take into account the size of each group[2] This results in the sum:

$$SS_W = \sum_{i=1}^{K} (n_k - 1) s_k^2$$

The sum of squares within groups captures the variability within each group around its own mean. In fact, $SS_W/(n-K)$ is a pooled estimate of variance, analogous to s_p^2 (assuming each group has the same variance).

A standardised statistic for measuring the variation in group means is then given by the F-statistic.

F-statistic

$$F = \frac{SS_B/(K-1)}{SS_W/(n-K)}$$

is called the F-statistic.

Note that if $K = 2$ (only two groups) then we have $F = t^2$, or the square of the t-statistic. This shows that the F-statistic is a generalisation of the t-statistic. In theory, we could have presented everything in terms of the F-statistic, and just then considered the special case of $K = 2$.

The denominator of the F-statistic is also called the mean-square error, and written $MSE = SS_W/(n-K)$.

So here's a table summary of the hair data:

[2] We will take $n_k - 1$ rather than n_k in order to get an unbiased estimate.

	DarkBlond	DarkBrunette	LightBlond	LightBrunette
ns	5.0	5.0	5.0	4.00
means	51.2	37.4	59.2	42.50
vars	86.2	69.3	72.7	29.67

We have $K = 4$ groups and $n = 19$ total measurements. From the table we can calculate that the global mean $\bar{x} = 47.84$, and $SS_B = 1360.73$ and $SS_W = 1001.8$.

We now have enough calculate the F-statistic:

$$F = \frac{SS_B/(K-1)}{SS_W/(n-K)} = \frac{1360.73/(4-1)}{1001.8/(19-4)} = 6.791.$$

All well and good, but is this F-statistic large enough to reject that all category means are equal? For that we need what other F-statistic values arise when the null hypothesis is true, namely that all means are equal.

By now you know the drill. `R` does this in one operation:

```
> oneway.test(Pain~HairColour, data=hair, var.equal=TRUE)
```

The output for our data is then:

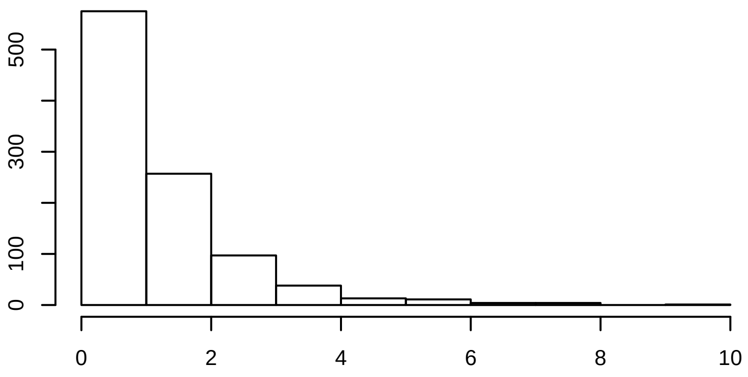
```

	One-way analysis of means

data:  Pain and HairColour
F = 6.7914, num df = 3, denom df = 15, p-value = 0.004114
```

It's even gone and given us a p-value (a small one too!). Similar to how we used the t-distribution to obtain a p-value, this one uses an F-distribution.

It's also possible to estimate a p-value using permutation simulation. The procedure is similar to the permutation version of the two group case. We just need to randomly permute all the group labels.

```
> ## compute the data F statistic
> Fstat = oneway.test(Pain~HairColour, data=hair, var.equal=TRUE)$statistic
>
> ## compute the F statistic when the category means are equal
> x = replicate(1000,{
+   ## shuffle the categories to force equal means
+   hair.perm = sample(hair$HairColour)
+   ## compute the F statistic using the shuffled data
+   oneway.test(Pain~hair.perm, data=hair, var.equal=TRUE)$statistic
+ })
```

We can now compare the data $F = 6.7914$ to the distribution of F-statistics from the simulation. Here is one output:



In R, the p-value can be computed as

```
> sum(x > Fstat)/1000
```

It turns out, in this instance, to be 0.007.

Note that F-statistics are always positive, and F-tests are always in effect two-sided.

`R` has a function (that uses unstandardised SS_B) and allows us to do permutations. It uses the `coin` library (Note the underscore).

```
> oneway_test(Pain~HairColour, data=hair, distribution=approximate(B=1000))
```

The output is then:

```
    Approximative K-Sample Fisher-Pitman Permutation
    Test

data:  Pain by
     HairColour (DarkBlond, DarkBrunette, LightBlond, LightBrunette)
chi-squared = 10.367, p-value = 0.005
```

The test is equivalent to the permutation version we did with our own code.

Back to the use of the F-distribution to obtain a p-value. It's validity requires that the data be approximately normally distributed, or that the data set is large enough (so that the central limit theorem works).

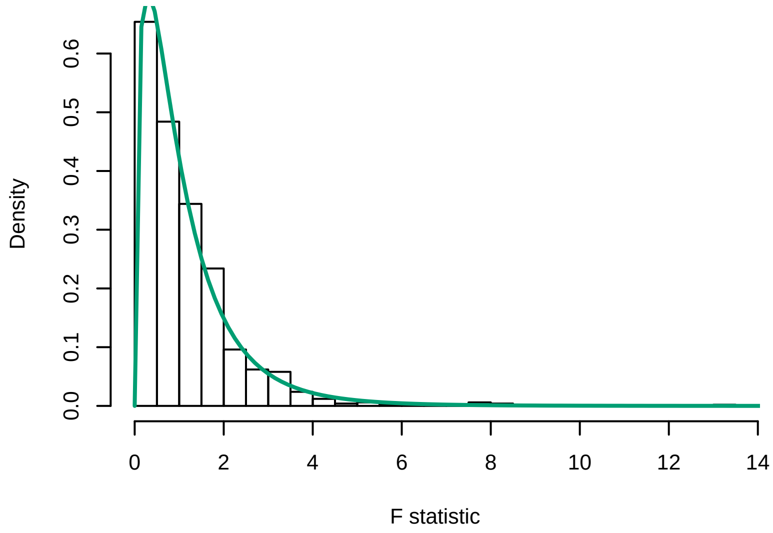
If either of those conditions are true, then under the null hypothesis, the F-statistic has the pre-determined distribution called the F-distribution. The F-distribution has two parameters called the numerator $K-1$ and denominator $n-K$ degrees of freedom.

To obtain the F-distribution directly use 'pf":

```
> pf(Fstat, 4-1, 19-4, lower=FALSE)
```

```
          F
0.004114227
```

We can see how the permutation and F-distribution methods compare in this histogram comparison:



11.2 ANOVA table

The procedure above is called a *one-way Analysis of Variance* (one-way ANOVA). It is traditional to present the results in an ANOVA table.

	df	SSQ	Mean Sq	F stat	p-value
Between	$K-1$	SS_B	$SS_B/(K-1)$	F	
Within	$N-K$	SS_W	$SS_W/(N-K)$		

Funnily enough, `oneway.test` does not produce this table.

```

    One-way analysis of means

data:  Pain and HairColour
F = 6.7914, num df = 3, denom df = 15, p-value =
0.004114
```

For that we need (more general) function called `aov`:

```
> fit = aov(Pain~HairColour, data=hair)
> summary(fit)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)   
HairColour   3   1361   453.6   6.791 0.00411 **
Residuals   15   1002    66.8                   
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11.3 Post-hoc tests

The F-test in ANOVA just tells us that at least one pair of means differs. It doesn't however tell us which one(s).

Obviously we would like to get that information. The first idea might simple be to generate a t-statistic for each pair (indeed we might question why we didn't skip the F-test altogether and go for that approach).

It's not quite so straightforward however. For K groups there are $K(K-1)/2$ possible pairs, which gets very large very quickly as K increases. That's a lot of computation, but even more importantly it massively increases our chances of rejecting the null hypothesis in error (a Type I error). Remember that keeping the Type I error rate to 0.05 (by only determining statistical significance $p < 0.05$) gives an expectation of rejecting the null hypothesis incorrectly 1 out of 20 times.

Let's take an example: $K = 10$ groups and $n_k = 20$ for each group k. It's our simulation, so we build in that the null hypothesis in the population is true. Then we need to calculate each t-statistic. The code looks as follows:

```
> n = 20 # sample size per category
> K= 10 # number of categories
> grp = rep(1:K, each=n)
> x = rnorm(length(grp)) # generate sample, all with same mean
> mns = tapply(x, grp, mean) # compute category mean
> # compute pooled variance
> sp = sqrt(sum((table(grp)-1)*tapply(x, grp, var))/
+                            (length(grp)-length(mns)))
> # compute the t-statistics of the largest difference in means
> maxT = diff(range(mns))/(sp*sqrt(2/n))  ## This code is for equal groups sizes
> print(maxT)
```

The output for one run is then

```
[1] 2.461777
```

This results in a p-value of 0.0147. So at the very least we have falsely detected one-pair (it may be more!).

One fix would be to require statistical significance only for very small p values (say $p < .01$), but this might be too conservative or maybe even not be small enough for a given number of groups K.

We need a more robust way, which leads to the **Tukey's range test** (also known as Tukey's Honest Significant Difference test or Tukey's HSD).

John Tukey proposed this technique that allows for the multiple testing by considering the distribution of the *maximum* t-statistic across all categories. It replaces that intolerable error rate by just lots of t-tests by controlling the *family-wise error rate.*

Here again is the t-test for the ith vs jth group comparison:

$$t_{i,j} = \frac{\bar{x}_i - \bar{x}_j}{s_p^{ij}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

We replace the sample variance s_p^{ij} term (which depends on i and j) in the denominator by the full *pooled sample variance* across all groups $1, \ldots, K$:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{n_1 + n_2 + \cdots + n_k - k}.$$

As you can see it is just a generalisation of the original pooled variance for two groups. Notice that this is the denominator of the F-statistic, and that

$$s_p^2 = \frac{SS_W}{N - k} = MSE.$$

It is also given in the ANOVA table by ”Mean Sq” (see example above).

Then we end up with a statistic for groups i and j:

$$q_{i,j} = \frac{|\bar{x}_i - \bar{x}_j|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which we use for the resulting statistical test.

Consider a significance α (e.g., $\alpha = 0.05$), so that $(p < \alpha)$. If $q_{i,j} > q_\alpha$, where q_α is the appropriate value from the *Studentized range distribution*, then the null hypothesis that the population means for groups i and j are the same can be rejected.

This test works provided either the data is normally distributed or the n_j are large enough. In summary:

Tukey’s range test

1. For each group pair (i, j), calculate the statistic:

$$q_{ij} = \frac{|\bar{x}_i - \bar{x}_j|}{s_p\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}.$$

2. If $q_{i,j} > q_\alpha$, then the null hypothesis $H_0 : \mu_i = \mu_j$, for population means μ_i, μ_j can be rejected with significance α.

By using the pooled sample variance rather than each paired sample variances, we have brought the family-wise error rate to be less than α.

Tukey’s range test using permutations

Rather than use the Studentized range distribution, it is also possible to use a permutation-based approach to test for statistical significance.

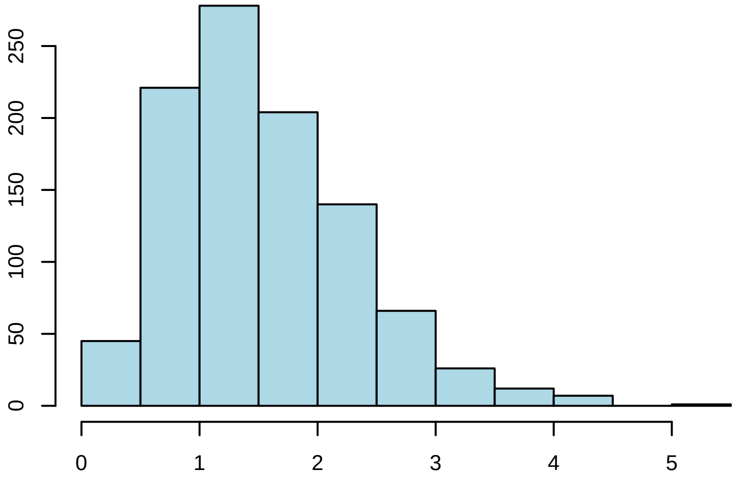
```
> ns = table(hair$HairColour)
> x = replicate(1000,{
+   # permute the categories (to satisfy H0)
+   hair.perm = sample(hair$HairColour)
+   fit0 = aov(Pain ~ hair.perm, data=hair)
+   # compute MSE  and means on permuted categories
+   MSE = summary(fit0)[[1]][2,3]  ## Extract the residual MSE
+   means =  aggregate(Pain ~ hair.perm, data=hair, mean)[,2]
+   # compute t statistic for all pairs
+   Ts = outer(means, means, "-")/sqrt(outer(1/ns,1/ns, "+"))
+   Ts = Ts/sqrt(MSE)
+   # keep only the largest t statistic
+   max(abs(Ts))
+ })
```

We needed to generate the maximum t-statistic for each permuted set of data. The above permutation approach uses the function `outer` that takes every combination of items in x and y and applies the operation.

```
> outer(2:6, 2:6, "*")
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    4    6    8   10   12
[2,]    6    9   12   15   18
[3,]    8   12   16   20   24
[4,]   10   15   20   25   30
[5,]   12   18   24   30   36
```

The resulting output from a simulation is then:



Actual t-statistics

If the t-statistics for our data difference in means are greater than the maximum difference in means when H_0 is true (all population means are equal), then we have evidence that the given pair have different population means.

	DarkBlond	DarkBrunette	LightBlond	LightBrunette
DarkBlond	0.000	2.670	-1.548	1.587
DarkBrunette	-2.670	0.000	-4.218	-0.930
LightBlond	1.548	4.218	0.000	3.046
LightBrunette	-1.587	0.930	-3.046	0.000

To compute the p-value, we compare the actual t-statistics to those from the permuted data. So, for example, p-value for DarkBrunette versus DarkBlond

```
> sum(x > 2.670)/length(x)
```

```
[1] 0.083
```

Using the Studentized range distribution

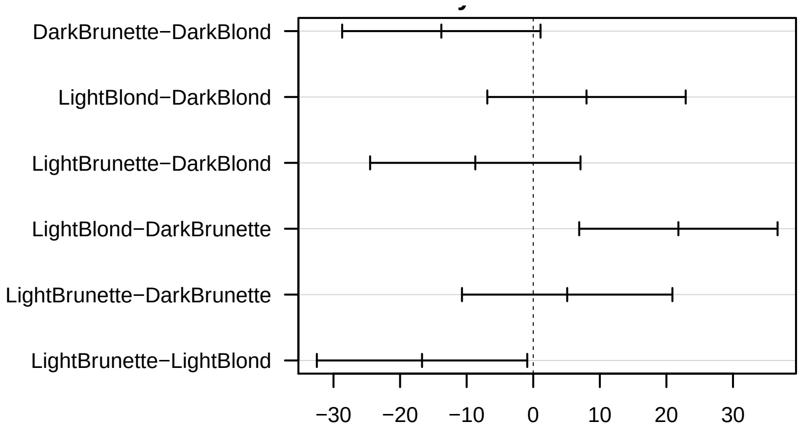
To use the studentized range distribution method, R provides various functions. The simplest is TukeyHSD.

```
> fit = aov(Pain~HairColour, data=hair)
> TukeyHSD(fit)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Pain ~ HairColour, data = hair)

$HairColour
                            diff        lwr        upr
DarkBrunette-DarkBlond     -13.8 -28.696741  1.0967407
LightBlond-DarkBlond         8.0  -6.896741 22.8967407
LightBrunette-DarkBlond     -8.7 -24.500380  7.1003795
LightBlond-DarkBrunette     21.8   6.903259 36.6967407
LightBrunette-DarkBrunette   5.1 -10.700380 20.9003795
LightBrunette-LightBlond   -16.7 -32.500380 -0.8996205
                               p adj
DarkBrunette-DarkBlond     0.0740679
LightBlond-DarkBlond       0.4355768
LightBrunette-DarkBlond    0.4147283
LightBlond-DarkBrunette    0.0037079
LightBrunette-DarkBrunette 0.7893211
LightBrunette-LightBlond   0.0366467
```

```
> par(mar = c(2,11,0.5,0.5), cex=0.7)
> plot(TukeyHSD(fit), las = 1)
```

Plotting those mean difference intervals gives:



As you can see, there are two confidence intervals which don't overlap with 0, yielding statistical significance for the LightBlond-DarkBrunette, and LightBrunette-LightBlond data sets.

11.4 A major study of redheads

That was for a toy data set. Let's now look at data from a study that looked at data from two nurse health studies in the US, consisting of around 150,000 women of varying ages, questioned multiple times. Two questions were asked related to pain: the extent of the pain and the extent the pain interfered with normal work in the previous four weeks.

The study is from the reference: Li, Wen-Qing et al. "Natural hair color and questionnaire-reported pain among women in the United States." Pigment cell and melanoma research vol. 29,2 (2016): 239-42.
`https://doi.org/10.1111/pcmr.12445`

The data were age corrected, and multiple measurements per individual allowed for. That's much more complex analysis than our example above.

Here is an extract:

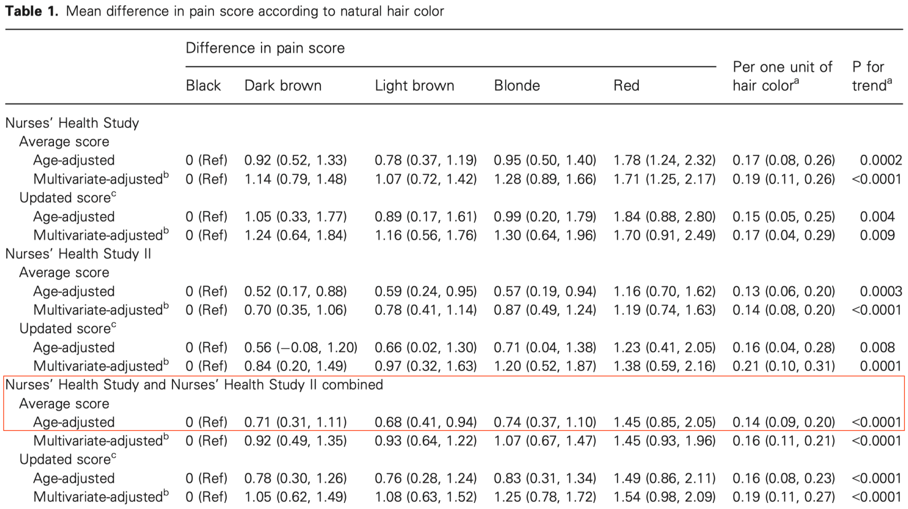


Table 1. Mean difference in pain score according to natural hair color

	Difference in pain score					Per one unit of hair color[a]	P for trend[a]
	Black	Dark brown	Light brown	Blonde	Red		
Nurses' Health Study							
Average score							
Age-adjusted	0 (Ref)	0.92 (0.52, 1.33)	0.78 (0.37, 1.19)	0.95 (0.50, 1.40)	1.78 (1.24, 2.32)	0.17 (0.08, 0.26)	0.0002
Multivariate-adjusted[b]	0 (Ref)	1.14 (0.79, 1.48)	1.07 (0.72, 1.42)	1.28 (0.89, 1.66)	1.71 (1.25, 2.17)	0.19 (0.11, 0.26)	<0.0001
Updated score[c]							
Age-adjusted	0 (Ref)	1.05 (0.33, 1.77)	0.89 (0.17, 1.61)	0.99 (0.20, 1.79)	1.84 (0.88, 2.80)	0.15 (0.05, 0.25)	0.004
Multivariate-adjusted[b]	0 (Ref)	1.24 (0.64, 1.84)	1.16 (0.56, 1.76)	1.30 (0.64, 1.96)	1.70 (0.91, 2.49)	0.17 (0.04, 0.29)	0.009
Nurses' Health Study II							
Average score							
Age-adjusted	0 (Ref)	0.52 (0.17, 0.88)	0.59 (0.24, 0.95)	0.57 (0.19, 0.94)	1.16 (0.70, 1.62)	0.13 (0.06, 0.20)	0.0003
Multivariate-adjusted[b]	0 (Ref)	0.70 (0.35, 1.06)	0.78 (0.41, 1.14)	0.87 (0.49, 1.24)	1.19 (0.74, 1.63)	0.14 (0.08, 0.20)	<0.0001
Updated score[c]							
Age-adjusted	0 (Ref)	0.56 (−0.08, 1.20)	0.66 (0.02, 1.30)	0.71 (0.04, 1.38)	1.23 (0.41, 2.05)	0.16 (0.04, 0.28)	0.008
Multivariate-adjusted[b]	0 (Ref)	0.84 (0.20, 1.49)	0.97 (0.32, 1.63)	1.20 (0.52, 1.87)	1.38 (0.59, 2.16)	0.21 (0.10, 0.31)	0.0001
Nurses' Health Study and Nurses' Health Study II combined							
Average score							
Age-adjusted	0 (Ref)	0.71 (0.31, 1.11)	0.68 (0.41, 0.94)	0.74 (0.37, 1.10)	1.45 (0.85, 2.05)	0.14 (0.09, 0.20)	<0.0001
Multivariate-adjusted[b]	0 (Ref)	0.92 (0.49, 1.35)	0.93 (0.64, 1.22)	1.07 (0.67, 1.47)	1.45 (0.93, 1.96)	0.16 (0.11, 0.21)	<0.0001
Updated score[c]							
Age-adjusted	0 (Ref)	0.78 (0.30, 1.26)	0.76 (0.28, 1.24)	0.83 (0.31, 1.34)	1.49 (0.86, 2.11)	0.16 (0.08, 0.23)	<0.0001
Multivariate-adjusted[b]	0 (Ref)	1.05 (0.62, 1.49)	1.08 (0.63, 1.52)	1.25 (0.78, 1.72)	1.54 (0.98, 2.09)	0.19 (0.11, 0.27)	<0.0001

These results are reported as differences between each pain score and a reference level. The reference hair colour was black hair. Confidence intervals are reported for the difference. If the interval does not contain zero there is evidence for a difference. The reported p-value is for a form of linear regression, adjusting for different variables.

It certainly seems that there is evidence that redheads report more pain than women with black hair colour.

Summary

To identify if one of more categories from a set of categories have a different mean, we compute the F statistic.

The results of an F test are presented as an ANOVA table.

To identify which pairs have different means, we use Tukey's range test.

Lecture 12

When it all goes wrong

Much of this lecture is taken from the book “Statistics Done Wrong” by Alex Reinhart.

12.1 Power and underpower

Suppose you testing a medication that is meant to shorten the length of colds. You find 20 patients with a cold, and give half the medication. Then you track the patients, comparing the average duration of colds in the two groups.

The null hypothesis is that the drug has no effect. We want to compute a p-value.

Suppose we actually observe a reduction in cold duration.

p-value: What is the chance of seeing a reduction as large or larger than that which we actually saw, if there really were no effect of the drug?

The book calls this a measure of surprise, in other words how surprising the result is.

As a thought experiment, suppose an oracle tells us the drug has no effect but we only use two patients: one treated, one not.

We would not be surprised if we saw a reduction in the treated group — unless that reduction were extremely large. The chance of seeing a reduction (even with no real effect) is still quite large.

If we had 100 patients though, and saw a reduction in days in the treated group, this would be surprising if there were no real effect.

But suppose the drug really does have some effect. Will we get a surprising result, of a small estimated p-value.

This obviously depends on the sample size, the size of the effect and the amount of variation in the measurements.

To illustrate this, consider the example of a friend who has a biased coin. You think it gives heads 60% of the time, yet she insists its fair and gives heads 50% of the time.

To see if she's right, you toss the coin 100 times and count the number of heads.

Although your friend might look at your strangely, you formulate the game as a null/alternative hypothesis:

$$H_0 : p = 0.5, \quad H_1 : p > 0.5.$$

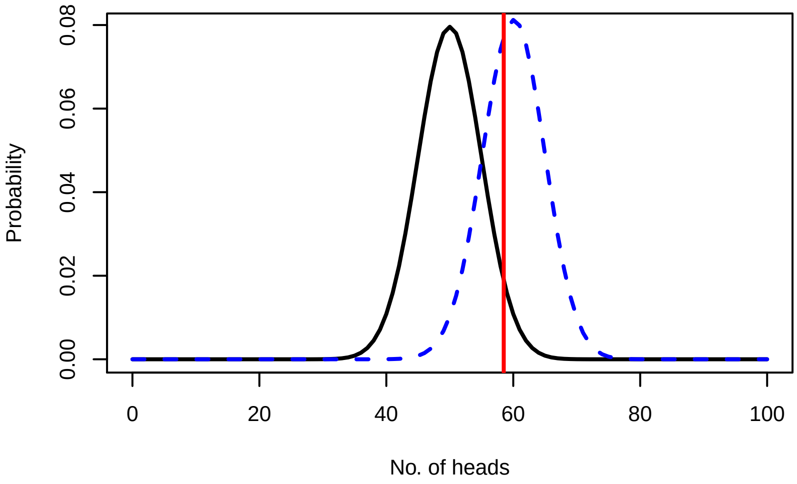
How many times would you need to see heads before the (exact) p-value for this test was less than 5%? Obviously 50 isn't enough as it would happen too often.

We use the binomial distribution to give us this answer. Let X be a binomial random variable with $p = 0.5$. Then[1] $P(X > 57) = 0.067$ and $P(X > 58) = 0.044$.

This means you would need to more than 58 heads before the (exact) p-value for this test was less than 5%.

If the true probability of a head was 0.6 (population parameter $p = 0.6$), then the chance of doing this would be only about 62%. Pretty close.

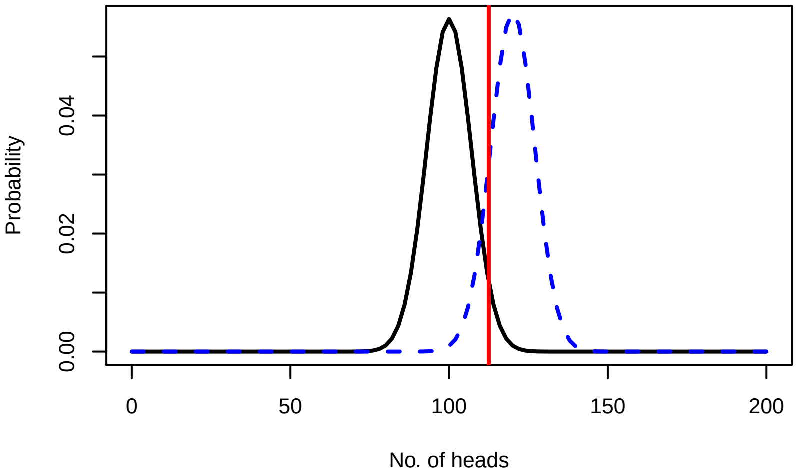
A diagram showing the distributions illustrate this well:



Here there are $n = 100$ samples (coin tosses) from each binomial distribution. The red line indicates the 5% threshold. Distinguishing which is the most likely scenario is tricky.

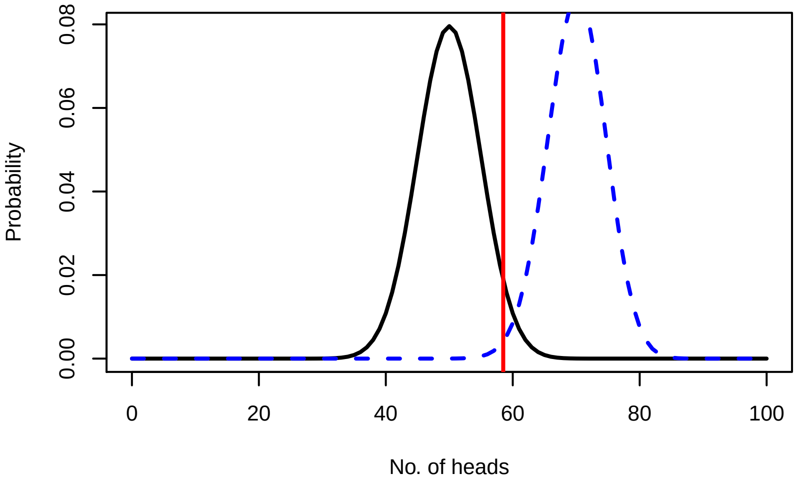
More evidence helps. You toss the coin another 100 times, getting 200 in total. Then the distributions become:

[1] In R you calculate this with `pbinom(58, 100, 0.5, lower=FALSE)` etc.



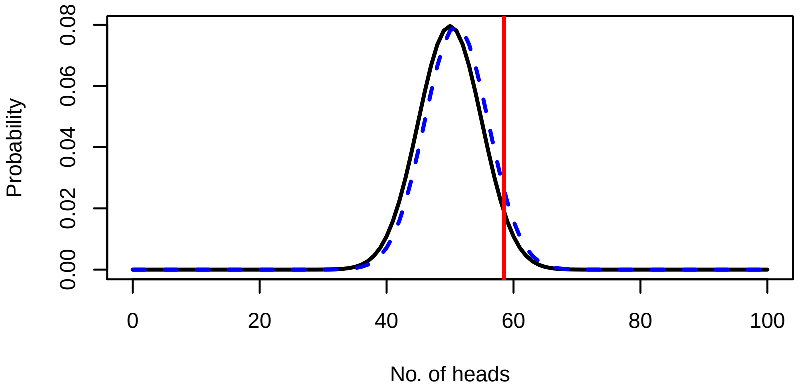
The distributions get separated apart. With 200 tosses, you would need to observe more than 112 heads, and the chance of doing so is 86%.

Let's go back to 100 tosses, this time with a true (population) probability of heads 0.7.



As expected, the distributions are pulled apart a lot more, as the difference between the two coins is more significant. Thus we can hope to distinguish them with the evidence of 100 coin tosses. We need to see 58 heads again, but the chance is more than 99%.

However, with a true probability of heads 0.51 the chance of seeing more than 58 heads is under 7%. Here are the distributions:



In fact, if the true chance of heads is 0.51 we need around 7,000 tosses to get a more than 50% chance of detecting this increase.

Under-powered tests

You might think it is rare for research to encounter this problem. But it occurs far more often than realised.

JAMA. 1994 “Statistical power, sample size, and their reporting in randomized controlled trials.” Moher D, Dulberg CS, Wells GA. Nearly two thirds of trials that reported no difference didn’t have the power to detect a 50% difference.

J Clin Oncol. 2007 “Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings.” Bedard PL, Krzyzanowska MK, Pintilie M, Tannock IF. Only about half of the trials had enough power to detect even large differences.

Right turns on red

In the 1970s several places in the USA introduced “right turn on red” provisions.

A study of 20 intersections before and after showed that 308 incidents occurred before, and 337 afterwards (over a similar time period). No statistical significance was reported by this or by many other studies.

Finally, years later, a rather larger study showed that collisions occurred 20% more frequently, and pedestrians were being hit 60% more frequently. The previous studies lack sufficient power to observe that.

12.2 Confidence intervals

Collapsing an entire data set to a yes/no call of statistical significance ($p<0.05$) is wasteful. Much more useful is to estimate the size of a difference using a confidence interval. However, this is rarely done...

12.3 Pseudo-replication

Randomisation prevents researchers unwittingly introducing bias into their study.

Consider the following example. 2000 patients are randomised to two blood pressure medications. After the medication takes effect, the average blood pressure of the two groups is measured.

Alternatively, suppose we have have only ten patients in each group, but measure each patients blood pressure 100 times. We still have 1000 measurements per group. Could we say that we still have the same sample size?

No, as we only have 10 unique patients per group and just know an awful lot about each patients blood pressure. This form of **pseudo-replication**.

Here's another example. A WSU researcher devised a study to compare two antibiotic treatments used on cows during the feed lot period just before slaughter.

They had data on 2000 cows split into the two treatments. Outcome was the amount of illness/infections. Analysed on face value, there were very significant differences.

However, the cows are put into the feed lots in batches. Other factors, like the weather and location of the lot potentially have an impact on the outcome.

The 2000 cows were in 8 batches, and treatment was allocated to batches. Ideally, treatment would be randomly allocated to cows in each batch, then batches would be blocks. But this was not practical.

To solved this the researchers we analysed with *linear mixed models.* Essentially this involves estimating within and between batch. variation as well as treatment difference. There was still evidence of treatment differences, but it was much weaker.

Unfortunately pseudo-replication is quite common in the published literature. see for example "Pseudoreplication and the Design of Ecological Field Experiments" Stuart H. Hurlbert, Ecological Monographs 1984.

Paired t-test and pseudo-replication

The paired t-test actually allows for multiple measurements of the same individual.

On the face of it, by analysing only the n differences and not the $2n$ observations it looks like we have less data. But the differences generally have much less random variation, because the individual effect is removed.

We demonstrated this already with the New Zealand helmet example.

12.4 Base rate fallacy and multiple testing

Think about testing multiple drugs. You have 100 cancer drugs and 10 of them are really effective. Using a threshold of 0.05 on the p-value, 13 drugs pass this test. An oracle tells us that 8 were really effective ones but there were 5 false positives.

Because a p-value threshold of 0.05 will occur around 1 time in 20 even for non-effective drug and 90/20 is approximately 5, only 8 out of 13 drugs (62%) called effective are truly effective.

Someone may incorrectly state that since the p-value is less than 5% there is a less than 5% chance that the drug was called effective by chance. But in fact we saw that 38% are of those called effective are in fact not, but were called effective by chance. This is called the **base rate fallacy**.

The base rate of truly effective drugs was low (10%) but this is forgotten.

12.5 p-value again

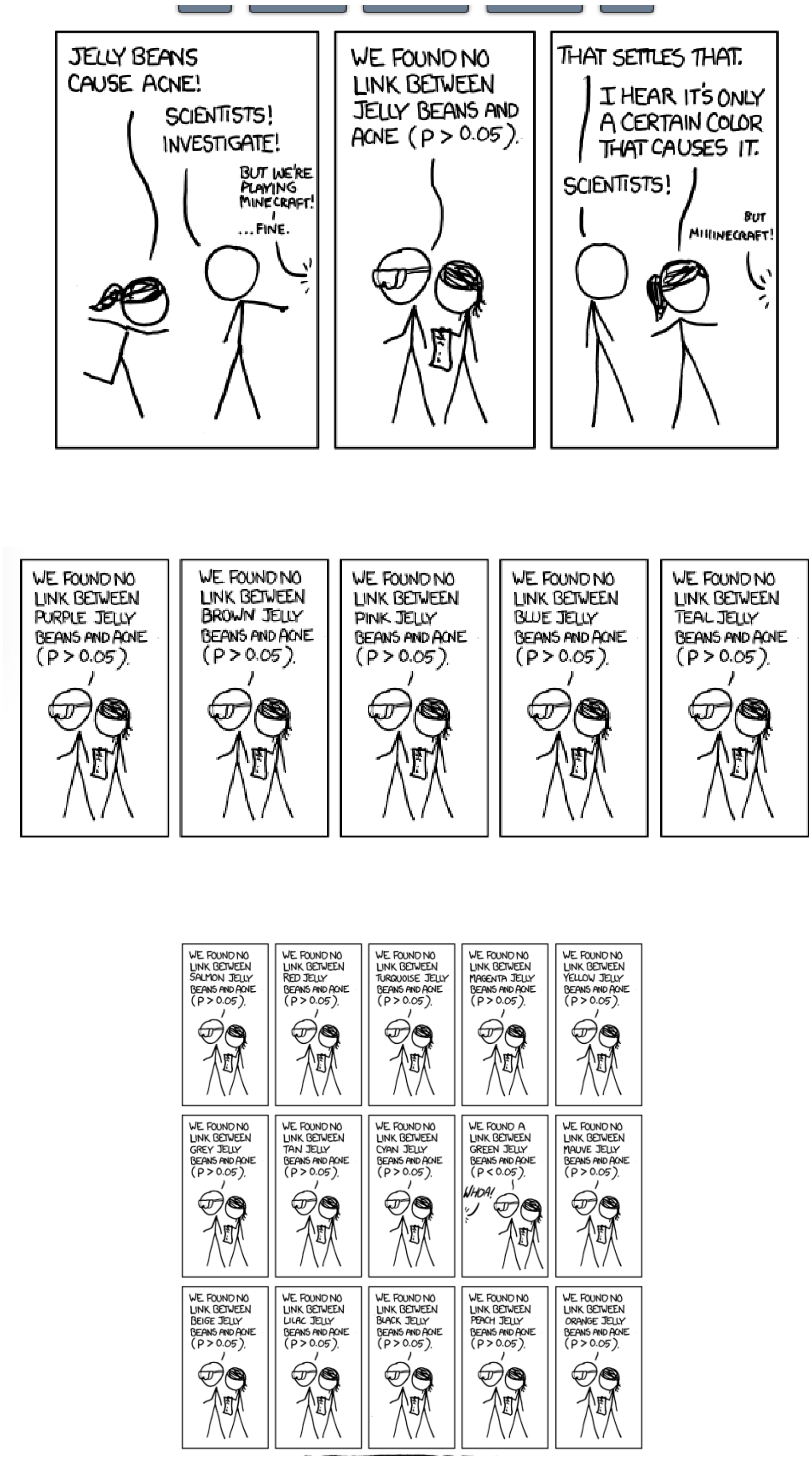
For a p-value of 0.01 which of the following is true?

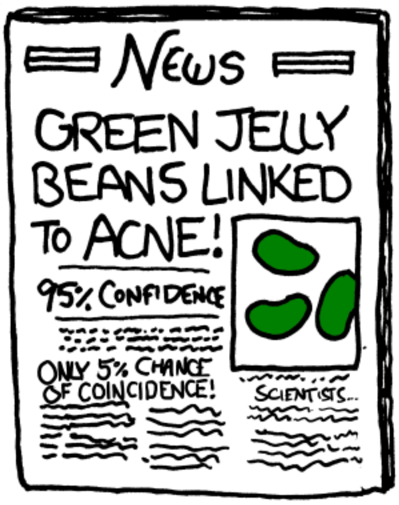
Problem

1. You have absolutely disproved the null hypothesis.
2. There is a 1% probability that the null hypothesis is true.
3. You have absolutely proved the alternative hypothesis.
4. You can deduce the probability that the alternative hypothesis is true.
5. You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.
6. You have a reliable experimental finding, in the sense that if your experiment were repeated many times, you would obtain a significant result in 99% of trials.

Multiple tests

This comic comes from `http://xkcd.com/882/`.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

12.6 False discovery rate

In genetics, its now routine to test the activity of tens of thousands of genes in one go. Often this involves huge sample sizes.

Suppose you are comparing the activity of 10,000 genes between cancer and normal tissue. Suppose all that 100 of these genes are really different in cancer and you have an 80% chance of detecting them (using a 5% p-value threshold).

Then you will detect 80 *true positive* genes, but also 5% of 9,900 genes as *false positives.* That is, you will detect around 575 genes, of which only 80 (14%) are real.

Fortunately, some of these 80 true positive genes will have a much smaller p-value than 5%. They will beat the threshold by a long way.

Of course, some of the false positives will have a very small p-value too, but this rate can be estimated. This leads to a number of methods to estimate the *False discovery rate.*

false discovery rate

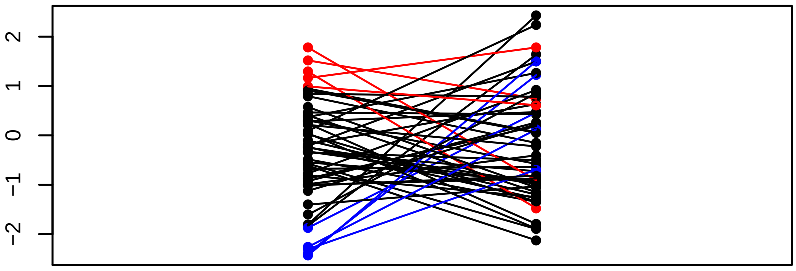
For a given, p-value threshold, the false discovery rate is the fraction of tests with a p-value less than this threshold that are there by random chance.

12.7 Double dipping and Simpson's paradox

Remember we talked about **regression to the mean**? Galton had noticed this effect as early as 1869.

When trialling a new blood pressure medication, suppose a number of people are tested, then those with the highest blood pressure selected. They a trialled on the drug for a period, and then re-tested. Generally they would have lower second measurements.

The plot below has 50 pairs of $N(0,1)$ points.



The highest 5 are coloured red, the lowest blue. Note how *most* of the reds go down, and most of the *blue* go up.

This is a common issue. In 1933, Horace Secrist published *The Triumph of Mediocrity in Business*. He suggested that unusually successful business tend to become less successful, and unsuccessful businesses tended to improve. The work contained heaps of data and graphs, and even cited Galton.

I'm guessing it sold a lot.[2]

Harold Hotelling reviewed the book in the *Journal of the American Statistical Association* and concluded that

> Secrist's arguments "really prove nothing more than that the ratios in question have a tendency to wander about"

Avoiding problems with regression to the mean

This can all be avoided by determining the groups to be treated in a way that is independent of any data that is used in the analysis.

For the blood pressure medication, we could determine a high blood pressure group either using historical data, or a separate screening test.

Then, after a few weeks, take a baseline measurement and use the drug, before finally taking the post treatment measurement.

Also a placebo/control should be considered.

[2] possibly not as much as "The Art of the Deal", which was a ghost-written book authored by a charlatan.

Simpson's paradox

In 1973 the University of California, Berkeley saw that of 12,763 applications for graduate study, 44% of males were accepted and 35% of females. Fearing a discrimination lawsuit the University investigated.

They found that of 101 departments only 4 showed a statistically significant bias against women **and** 6 showed a bias against men.

It turns out that men and women did not apply in equal ratio to all the departments. Two thirds of applicants in the English department were women, and only 2% of applicants in Mechanical Engineering were women. Also some departments had higher success proportions than others.

Overall, this was sufficient to explain the perceived bias.

To see it clearly consider the following constructed example. Two tennis players are to be compared across two seasons.

Player A had win stats:

	wins	games	percent
season 1	80	100	80%
season 2	20	40	50%

while Player B's stats were:

	wins	games	percent
season 1	78	100	78%
season 2	2	5	40%

When putting both seasons together however, it looks like:

	wins	games	percent
player A	100	140	71.4%
player B	80	105	76.2%

It is not a paradox that player A had better performance in both seasons taking individually, while player B had better performance overall.

12.8 Conclusions

We looked at a number of ways that analyses can go wrong. Care is needed when designing or analysing a data based study.