Applied Probability II

Section 1: Module admin

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 1: Module admin

Section 1.1: Background information

About me

Professor Caroline Brophy

Associate Professor of Statistics

Research interests: developing and applying statistical models for Ecology, Agronomy and Climate Change studies.

About you!

Students taking this module come from a wide range of programmes:

- Management Science and Information Systems
- Mathematics major and Economics minor
- Mathematics Single Pathway
- Mathematics major and Irish minor
- Mathematics major and Statistics minor
- English Literature and Mathematics
- Mathematics and Philosophy
- Mathematics and Psychology
- Computer Science Single Pathway
- Visiting students

You have all taken Applied Probability I (pre-requisite).

There 180+ students registered to this module.

Section 1.2: Module details

Lectures

There will be two lectures per week.

- These will be held 'live' online at 9am on Mondays and 3pm on Thursdays.
- Lectures will be recorded and available 'on-demand' after each session; however, please attend the live lectures.

Assessment

- Final two-hour exam worth 85%.
- Continuous assessment worth 15%.
  - There will be four continuous assessments sheets during the semester that will each be worth an equal amount.
  - The four continuous assessment sheets will be handed out in weeks: 2, 4, 6 and 8 respectively.
  - There will be one week to complete each sheet, so deadlines in weeks: 3, 5, 7 and 9.

Laboratory sessions

There will be one laboratory session per week, starting in week 2.

- In these sessions, you will be split into smaller groups and will attend at your allocated 1-hour slot each week.
- Some weeks, a laboratory activity sheet will be provided to work through during the laboratory. Laboratory activity sheets do not count towards your continuous assessment grade.
- Some weeks, the material from a continuous assessment sheet that has been already submitted will be covered during the laboratory session.

Learning outcomes

- LO1: Derive confidence intervals and hypothesis tests for means and variances
- LO2: Derive prediction intervals for simple statistical models and explain how they differ from confidence intervals
- LO3: Conduct and explain the outputs of hypothesis testing in regression analysis
- LO4: Define maximum likelihood estimates and how to compute them
- LO5: Implement a bootstrap to construct confidence intervals
- LO6: Construct a q-q plot and use simple transformations of data that can make it more Normally distributed
- LO7: Construct a probability plot for any given distribution where its distribution function is known
- LO8: Calculate the properties of multivariate distributions
- LO9: Derive marginal and conditional probabilities of the bivariate Normal distribution

Topics covered

- Derivation of the confidence interval and tests of hypothesis for normal data; the difference between a confidence interval and a prediction interval
- The Central Limit Theorem and what it says about confidence intervals and tests of hypothesis
- Hypothesis testing for regression analysis
- The bootstrap approach to confidence intervals and tests of hypothesis
- Introduction to maximum likelihood estimation and computation
- The q-q plot and transforming data to make it more Gaussian
- Introduction to multivariate distributions

Let's do an experiment!

We are going to do an experiment in class today. I ask that you will each participate by carrying out a simple task. I will be recording the length of time it takes you to complete the task.

Where will I find this experiment?

- Under 'MODULE MATERIAL' there is a link called 'In-class experiment'. If you click on that page, you will just see an announcement saying that we will do an experiment during our first class.
- During our class, a new link will appear on that page called 'Experiment 2021'.

How will I participate?

- When it is time, I will tell you to refresh your Blackboard page and click into the experiment.
- You will have to answer just one question to complete the experiment.
- You will be asked to put a list of five words in alphabetical order. Put a number 1 beside the first alphabetical word, a 2 beside the second one, and so on.
- Please try to complete the question as quickly as you can BUT please make sure you have the correct order before you submit your answer.

Let's do an experiment!

Important notes:

- 'Experiment 2021' It will appear as a test on Blackboard. BUT IT WILL NOT COUNT TOWARDS YOUR GRADE FOR THIS MODULE!
- 'Experiment 2021' will appear as a test that is overdue. PLEASE IGNORE THIS OVERDUE MESSAGE!
- Doing this experiment is for fun and so that we can work on some data that you have been involved in collecting. When I download the dataset from Blackboard, I will not record your name. The data will be anonymised before we use it in class.
- I will stay in the live lecture. When you have completed your task, please come back and put a message in the chat to say 'finished'.
- The link to 'Experiment 2021' will disappear at the end of the first lecture. If you are watching the recorded version of the lecture, or just reading the lecture notes back, the link will not be available.
- Thank you for participating in this in-class experiment!

Applied Probability II

Section 2: Recap on previous material

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 2: Recap on previous material

Section 2.1: Applied Probability I

Topics covered

- Random variables: discrete and continuous
- Discrete distributions
- Continuous distributions

Random variables: discrete and continuous

- What is a random variable?
  - the 'outcome' from an 'experiment'
- You won't know the outcome in advance, but you can think about:
  - what are the possible values or range of values?
  - with what probability are certain values, or ranges of values, likely to occur?
- Random variables can be discrete or continuous.
- We talk about the 'distribution' of a random variable.

Discrete distributions

- To characterise a discrete distribution, we can think about:
  - possible values
  - probability mass function (pmf), $P(X = x)$
  - cumulative distribution function (cdf), $P(X \leq x)$
  - mean or expected value, $\mathrm{E}[X]$
  - variance, $\mathrm{Var}(X)$

- Side note: when do we use capital X and when do we use little x?

- Some commonly known discrete distributions
  - Bernoulli
  - Binomial
  - Poisson
  - Geometric
  - Negative binomial
  - Hypergeometric
  - Uniform

To help think about discrete distributions here is a link to some music from Dr Rafael de Andrade Moral, a talented collaborator of mine! https://youtu.be/ZINXFoQMZVs

Discrete distributions contd.

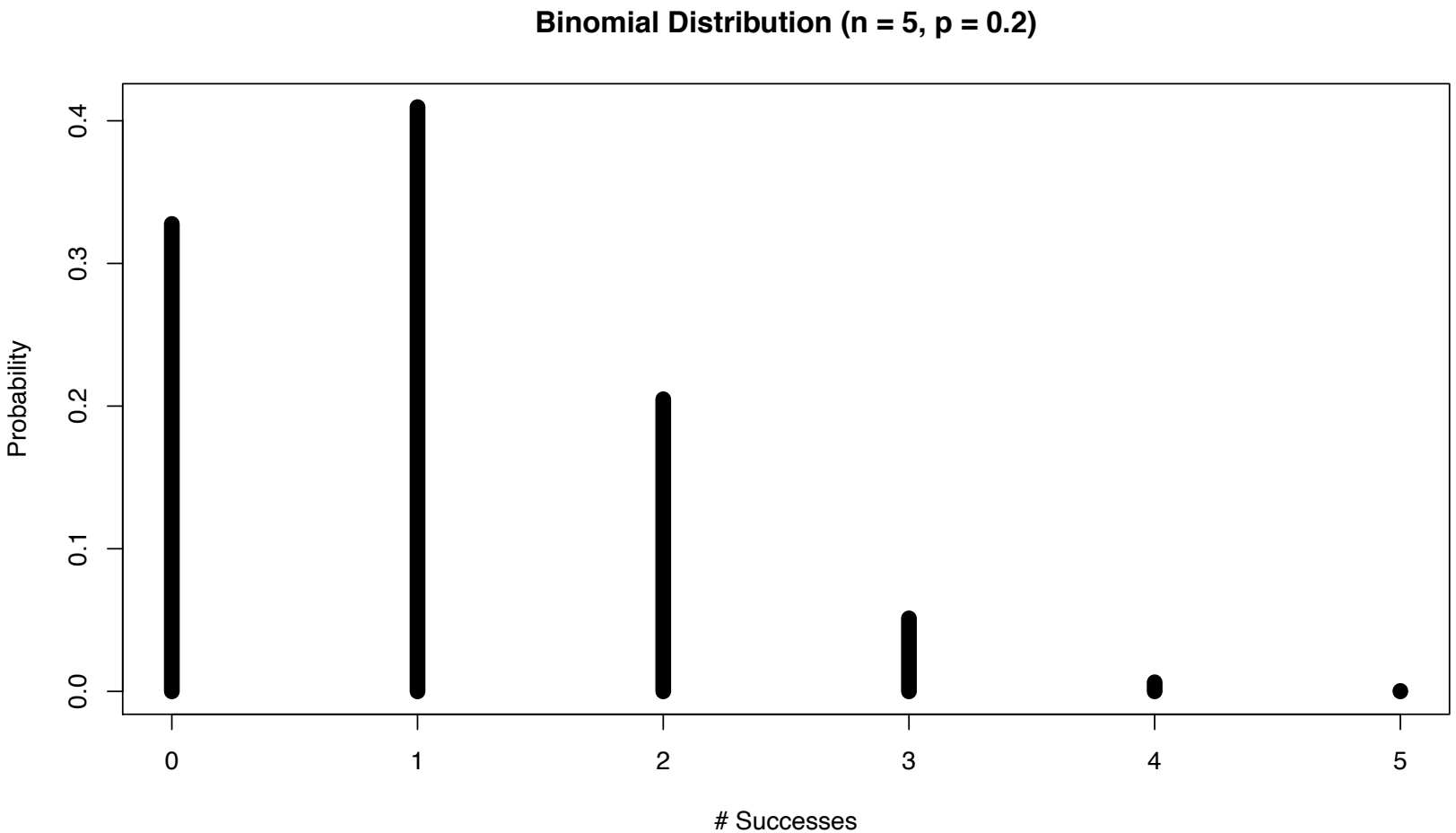
For the binomial distribution, consider $X \sim \text{Bin}(n = 5, p = 0.2)$. The random variable X is the number of 'successes' in $n = 5$ trials.

How might we characterise this distribution?

- possible values: $\{0, 1, 2, 3, 4, 5\}$
- probability mass function (pmf) $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{5}{k} 0.2^k (0.8)^{5-k}$
- cumulative distribution function (cdf), $P(X \leq x)$
- mean or expected value, $\text{E}[X] = np = 5 \times 0.2 = 1$
- variance, $\text{Var}(X) = npq = 5 \times 0.2 \times 0.8 = 0.8$

Discrete distributions contd.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{5}{k} 0.2^k (0.8)^{5-k}$$



Continuous distributions

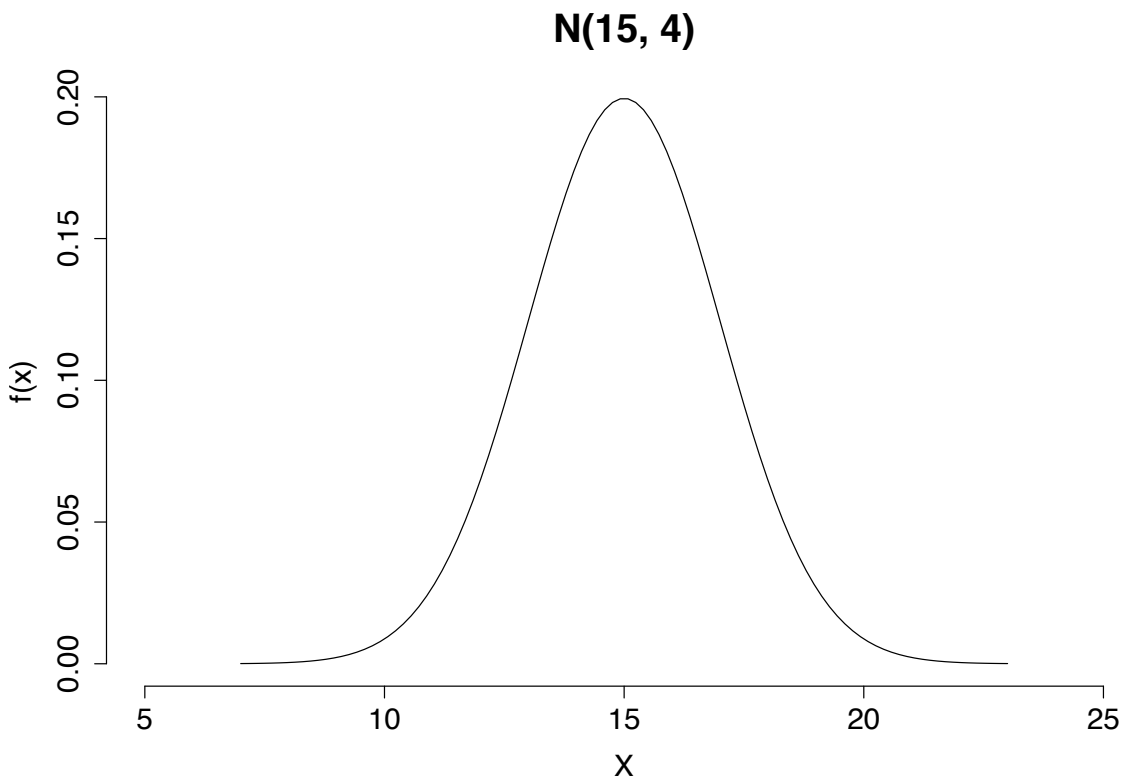
- To characterise a continuous distribution, we can think about:
  - the range of values
  - probability density function (pdf), $f(x) \geq 0$, and $\int_{-\infty}^{\infty} f(x)dx = 1$
  - cumulative distribution function (cdf), $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$.
- mean or expected value, $\mathrm{E}[X]$
- variance, $\mathrm{Var}(X)$

- Some commonly known continuous distributions
  - Normal
  - Uniform
  - Exponential

Continuous distributions contd.

The normal (or Gaussian) distribution, consider $X \sim \mathrm{N}(\mu = 15, \sigma^2 = 4)$.

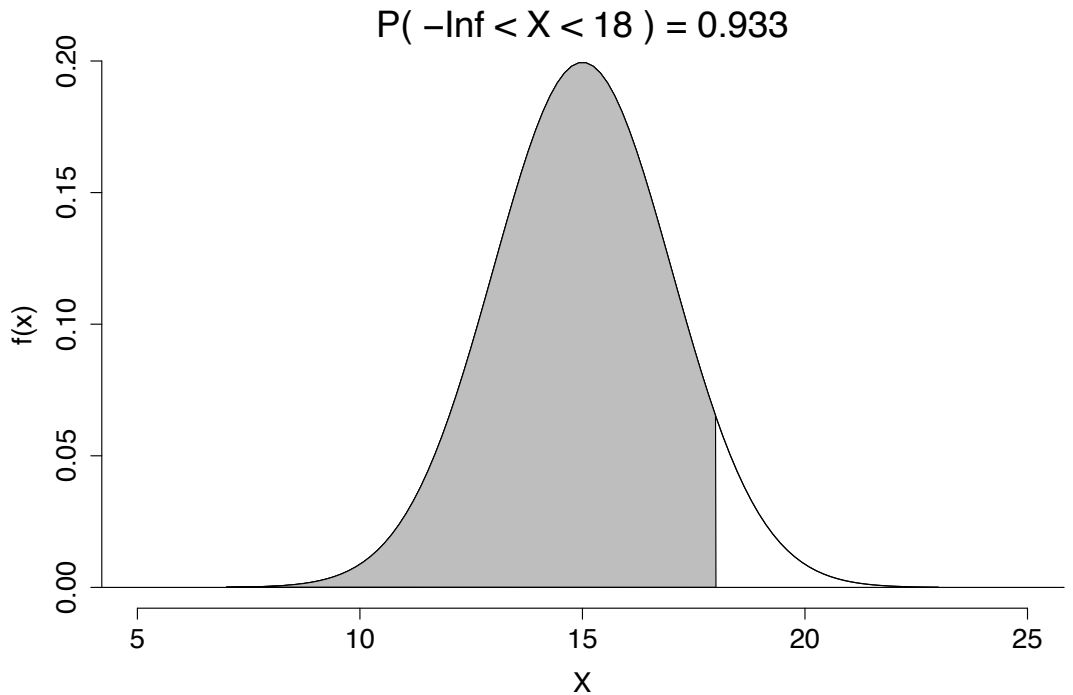
Probability density function (pdf): $\int_{-\infty}^{\infty} f(x)dx = 1$



Continuous distributions contd.

The normal (or Gaussian) distribution, consider $X \sim \mathrm{N}(\mu = 15, \sigma^2 = 4)$.

Cumulative distribution function (cdf), $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$.

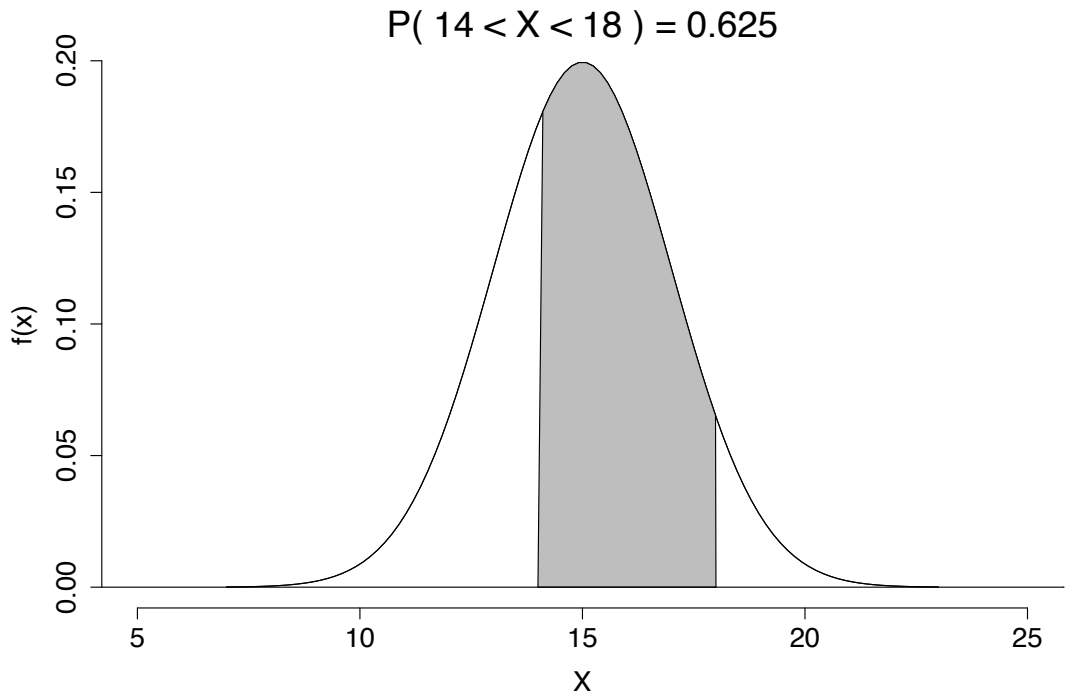


$$F(18) = P(X \leq 18) = \int_{-\infty}^{18} f(x)dx$$

Continuous distributions contd.

The normal (or Gaussian) distribution, consider $X \sim \mathrm{N}(\mu = 15, \sigma^2 = 4)$.

Cumulative distribution function (cdf), $F(a) = P(a \leq X \leq b) = \int_a^b f(x)dx$.



$$F(18) = P(14 \leq X \leq 18) = \int_{14}^{18} f(x)dx$$

Other topics

- Joint distributions
  - discrete
  - continuous
  - marginal distributions
  - independence
  - expectation
  - moment generating functions
  - covariance and correlation
  - conditional distributions
  - Bayes' rule
  - Bivariate normal distribution
- Law of Large Numbers
- Monte-Carlo Simulation
- Introduction to regression analysis

Section 2.2: Populations versus samples

Populations

Examples include

- All adults in Ireland
- People at risk of a particular disease
- All trees of a particular species

What population 'parameters' might we be interested in?

Samples

Consider the population:

- adults in Ireland

and population parameter of interest:

- the mean height.

How might we take random samples from this population?

- wrong answers ("bad" samples!) first!
- and good samples?

Terminology

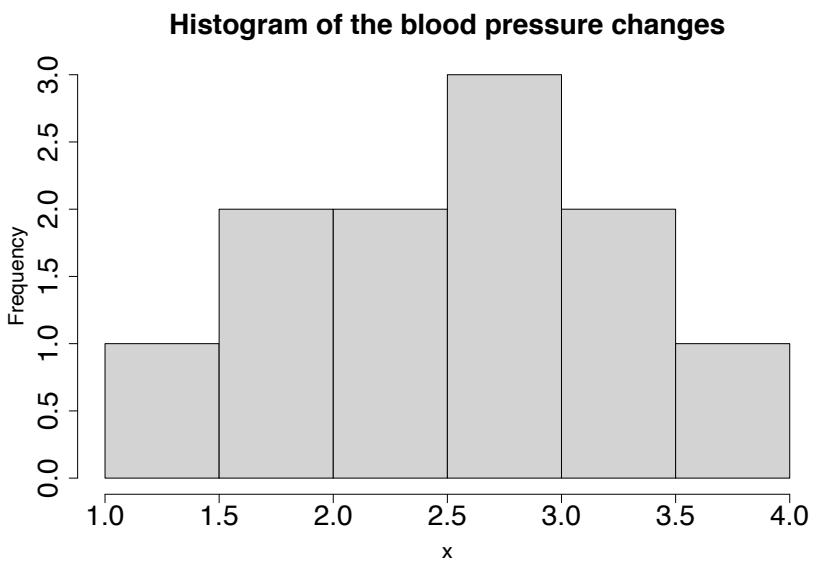
- **Parameter**: Population characteristic.
  - For example, μ, σ, σ^2, π.
- **Sample statistic**: Any quantity computed from values in a sample.
  - For example $\bar{x}$, s, s^2, p.
- But what about $\bar{X}$, S, S^2, P? What do these represent?
- **Statistical inference**?

Section 2.3: Hypothesis testing and confidence intervals

Pharmaceutical example

A pharmaceutical company has developed a drug they believe will help cure a particular disease and has acquired legal permission to test the drug on humans. However, there are safety concerns that the drug will have an adverse effect on blood pressure by increasing it. They run an experiment to estimate the average change in blood pressure for patients who take the new drug. They only have permission to test the drug on eleven patients.

The eleven recordings are: 1.1, 1.8, 2, 2.4, 2.5, 2.8, 2.9, 3, 3.4, 3.4, 4. The mean, $\bar{x} = 2.66$, standard deviation, $s = 0.824$ and the number of observations, $n = 11$.



Pharmaceutical example: confidence interval

Population parameter of interest: μ, the true mean change in blood pressure for the population of people who take the drug.

Suppose the data represent a random sample from the population of all people who take the drug. Then we would estimate the population mean μ by the sample mean $\bar{x}$, i.e. we **estimate** the mean change in blood pressure to be $\bar{x} = 2.66$.

95% confidence interval for μ: (2.11, 3.22)

This is computed as

$$\bar{x} \pm t_{\nu, \frac{0.05}{2}} \frac{s}{\sqrt{n}}$$

Let's do a Poll. How should we interpret the confidence interval?

A. We are 95% sure the sample mean lies in the interval.

B. We know that 95% of the observations studied lie in the interval.

C. If the experiment is repeated, we are 95% sure the next sample mean will lie in the interval.

Pharmaceutical example: confidence interval

Here is our confidence interval again:

95% CI for μ: (2.11, 3.22)

How is this confidence interval interpreted?

Let's answer these questions first:

- Would we get the same CI if we repeated the experiment?
- What do we know about $\bar{x}$ and our 95% confidence interval?
- What do we know about μ and our 95% confidence interval?

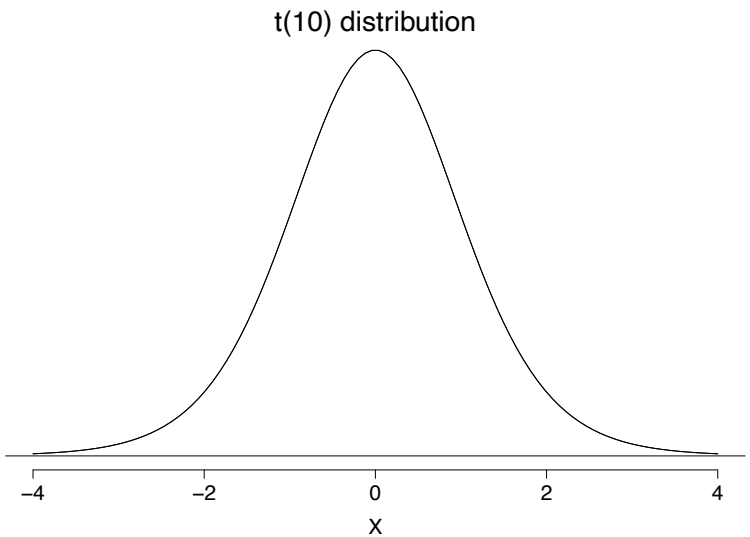
We are 95% confident that our population parameter of interest, μ, lies in this interval.

What would we expect to happen if we repeated the experiment 100 times?

Pharmaceutical example hypothesis test

- $H_0 : \mu = 0$ vs $H_A : \mu > 0$.
- The observed test statistic is $T_{obs} = \frac{\bar{x}-0}{s/\sqrt{n}} = 10.72$
- P-value $= P(t_{10} \geq 10.72) < 0.001$.
- Conclusion: We reject H_0 that $\mu = 0$ at $\alpha = 0.05$. We have evidence that the population average blood pressure change is greater than 0, i.e., evidence that the drug increases blood pressure.

Note about p-values. The p-value is the probability of observing a test statistic as extreme or more extreme than what was observed, given that the null hypothesis is true. Here, we assume that the distribution of the test statistic under the null hypothesis is:



How do confidence intervals and hypothesis tests work?

In this module, we will look at the underlying theory behind CIs and hypothesis tests.

Some closing thoughts:

- What assumptions did we make for the confidence interval for the pharmaceutical example?
- Are those assumptions reasonable?
- What assumptions did we make for the hypothesis test?

Applied Probability II

Section 3: Getting to grips with uncertainty

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 3: Getting to grips with uncertainty

Section 3.1: Probability versus Statistics

Uncertainty

If there is one certainty in life - it is uncertainty!

Think about a situation in your own life recently, where an outcome was uncertain. How did you try to make sense of it?

Some examples:

- car breaks down and waiting on road-side assistance.
- waiting on a bus or DART to arrive
- planning a cycle, will it be windy?
- choosing a treatment for a disease that has side-effects

How do we make sense of uncertainty?

We can try to quantify it, but how do we do that formally?

Probability

In Applied Probability I, you learned about distributions and probability rules.

If you make an assumption about the distribution of some outcome, you can then assign probabilities to the possible outcomes.

What else can you do?

Collect data!

Collecting data

Let's go back to one of the examples we looked at earlier:

Car has broken down and you call roadside assistance to come and need to know if they will arrive in the timeframe they said they would. This happened to me recently!

- Observation 1. How long did they take to arrive and was it in the timeframe they said?
- Observation 2. Again, how long did they take to arrive and was it in the timeframe they said?
- What will happen the third time? How high are the stakes?

Probability versus Statistics

- PROBABILITY: We start with a probability model, i.e., we consider outcomes that could occur and assign likelihood of those outcomes occurring. Outcomes are random, therefore we don't know what will happen so we 'quantify our degree of surprise'.
- STATISTICS: We assume that data at hand were generated by some probability model and that our task is to approximate what that model was.

Probability	**Statistics**
Predicting probabilities of events	Modelling data to understand events.
Rules -> data	Data -> rules

"All models are wrong, but some are useful", George Box

What are the stakes?

How high stakes are the various outcomes?

There may be times when quantifying uncertainty is more important than others.

For many of the uncertain phenomena that surround us, quantifying uncertainty is a necessity. It can, for example:

- Protect us (healcare, public safety, pandemic)
- Make money from us: business (decision making)

Decision making.

"If it was an easily solvable problem, or even a modestly difficult but solvable problem, it would not reach me, because, by definition, somebody else would have solved it." - Barack Obama

Some people see the world as black and white, however, better decision making can arise from moving away from black and white thinking, and considering the probabilities associated with the possible outcomes.

Section 3.2: Introduction to estimation

Random variables and realisations

Suppose we wish to describe a random process and have collected data

$$y_1, y_2, ..., y_n$$

- These values $y_1, y_2, ..., y_n$ are realisations of random variables $Y_1, Y_2, ..., Y_n$.
- In a random sample, these random variables are independent and follow the same probability distribution.

For example, commute to college on the DART.

Random sample of days' journey times: $Y_1, Y_2, ..., Y_n$ with realisations (in minutes):

$$y_1 = 20, y_2 = 19.6, ..., y_n = 21.5$$

A model for the mean

We may want to investigate the "centre" of the process.

- The true (population) mean is μ.
- As a first step, we compute the sample mean and say this is a fair estimate.

$$\bar{y} = \frac{\Sigma_{i=1}^{n} y_i}{n} = \hat{\mu}$$

- But how close or how far away will $\bar{y}$ be from μ?

Before we can go any further, we need to make some assumptions about the distribution of $Y_1, Y_2, ..., Y_n$.

For example, we might assume that $Y_i \sim N(\mu, \sigma^2)$, IID (independent and identically distributed).

Important note on terminology:

- What does the $\hat{}$ on $\hat{\mu}$ mean?
- What is the difference between μ and $\hat{\mu}$?

Assumptions in Statistics

Has anyone ever fit a statistical model to data before?

When we fit any statistical model to data, or perform any statistical analysis, we must:

1. Know what (if any) assumptions are being made by the model or analysis.
2. Verify that those assumptions are reasonable.

This is crucial!

Section 3.3: Sampling distributions

What is a sampling distribution?

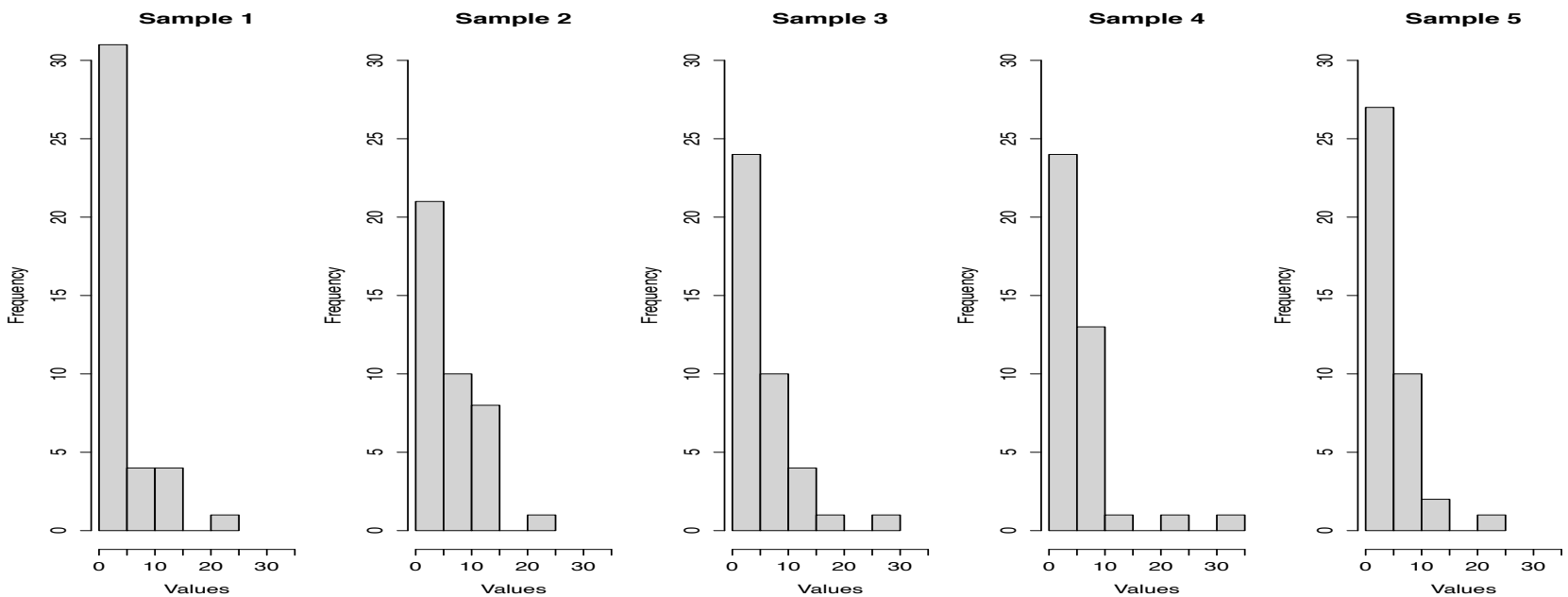
- **Parameter**: Population characteristic.
  - For example, μ, σ, σ^2, π.
- **Sample statistic**: Any quantity computed from values in a sample.
  - For example $\bar{x}$, s, s^2, p.
- The value of a population characteristic is fixed, but if you take, for example, ten samples from a population and compute $\bar{x}$ for each sample, would you expect each $\bar{x}$ to be the same?
- A sample statistic (considered in the context of any possible sample from the population) is a random variable and it has a probability distribution called the 'sampling distribution'. For example, we may talk about the sampling distribution of $\bar{X}$.
- **NB: The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.**

Illustration of the sampling distribution of the mean

From previous slide: The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.

Let's simulate from the exponential distribution with $\lambda = 0.2$. (Aside: what is the mean of this distribution? And the variance?)

We simulate 1000 samples from this distribution, each containing 40 values. Here are histograms of the first five samples:



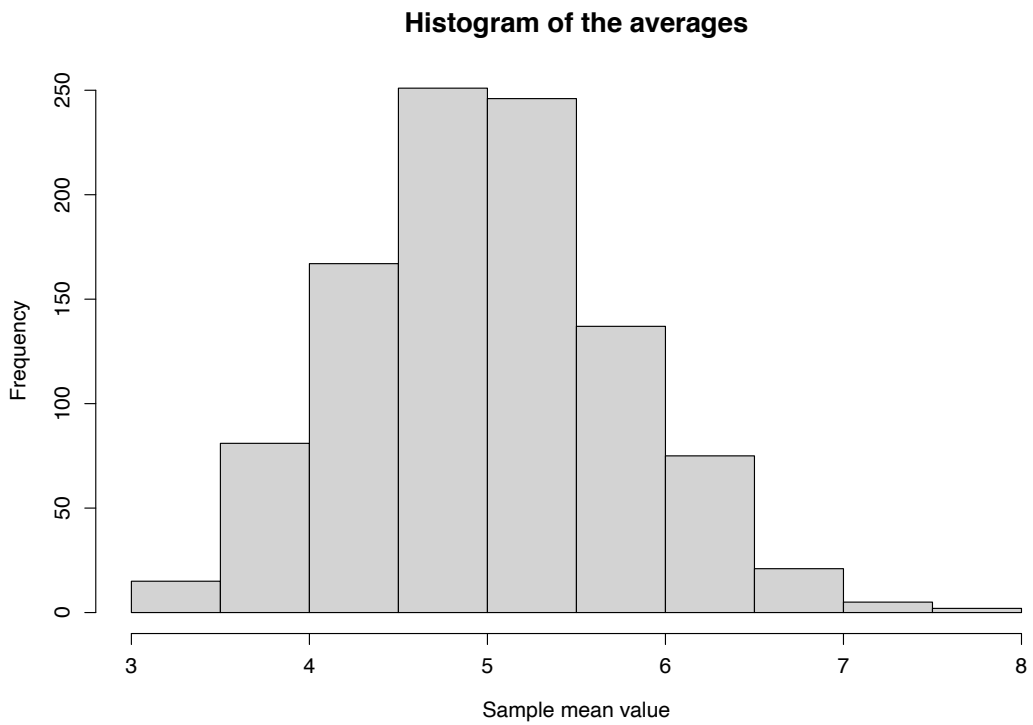
We could look at the histogram for any of the 1000 simulated samples.

Examine the means from the 1000 samples

We saw the exponential shape in the histograms of the first five of 1000 samples simulated.

The averages from the 40 values from each of the 1000 samples was computed. The first five are: 4.38, 5.87, 5.44, 5.54, 4.15.

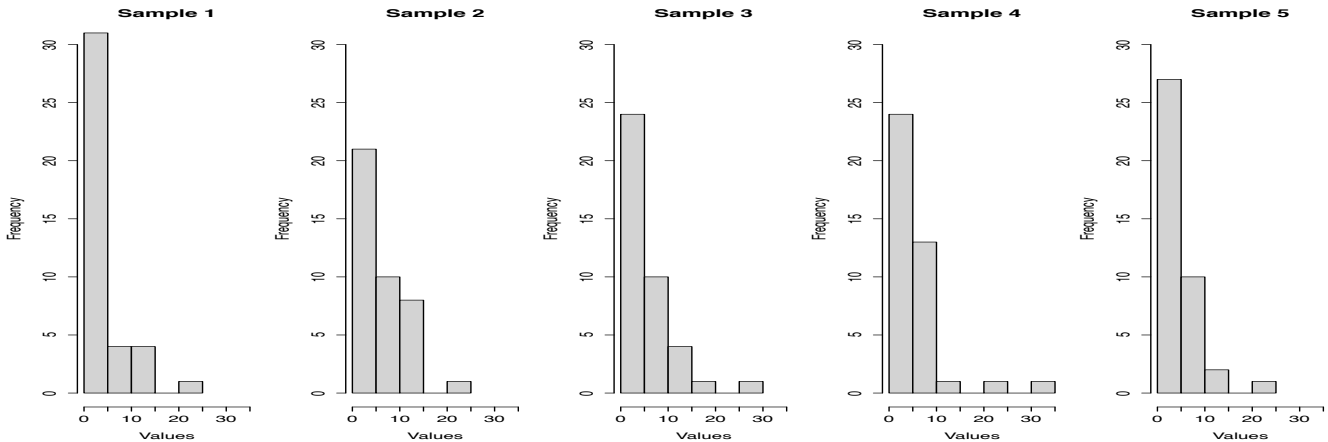
Let's plot the averages from all 1000 samples generated.



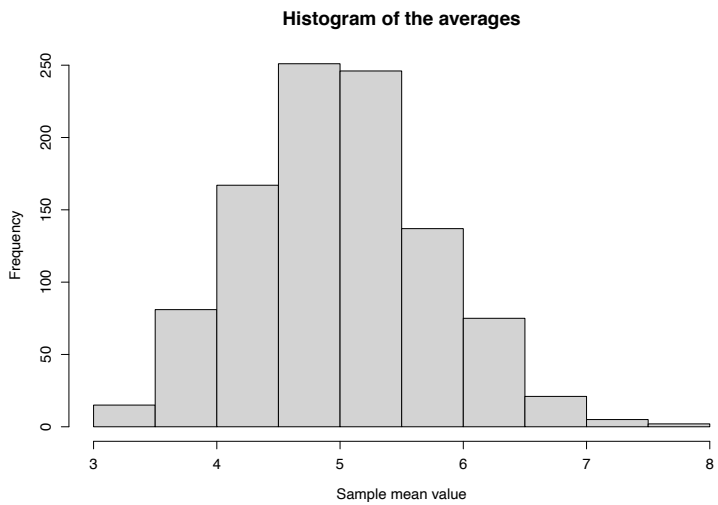
Does this look to be exponential? How would you describe this distribution?

Sample distribution versus sampling distribution

Sample distributions:



Sampling distribution:



Theory underlying sampling distributions

Recap:

The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample.

In later sections in this module, we will examine the theory behind the observation that we have just seen.

Section 3.4: In-class experiment

What were the questions?

- During the first lecture, many of you participated in an in-class expeirment. Thank you again for this!
- What did we test during this experiment?
- Each person was asked just one question, but there were six different questions.
- Each question required you to put a list of words in alphabetical order; they were:
  - bouncing handle kitchen tracksuit university
  - washing weird which wisdom wonderful
  - mobile model moment mountain movie
  - stack stake standard stapler statistics
  - starboard stardom starfish starry startle
  - crossbar crossfire crossing crossroad crossword

What was I trying to prove?

What hypothesis do you think I was trying to prove?

Here are the questions again.

- Each question required you to put a list of words in alphabetical order; they were:
  - bouncing handle kitchen tracksuit university
  - **w**ashing **w**eird **w**hich **w**isdom **w**onderful
  - **mo**bile **mo**del **mo**ment **mo**untain **mo**vie
  - **sta**ck **sta**ke **sta**ndard **sta**pler **sta**tistics
  - **star**board **star**dom **star**fish **star**ry **star**tle
  - **cross**bar **cross**fire **cross**ing **cross**road **cross**word

How did things work out?

Here is a quick look at the first few rows of data:

```
##   Question Seconds Correct
## 1        1      16       1
## 2        1      18       1
## 3        1      23       1
## 4        1      24       1
## 5        1      25       1
## 6        1      26       1
```

What do we need to think about to continue making sense of this data?

Group level information?

Here are some summary statistics for each 'group'.

```
##   Question Count Av_Seconds Sd_Seconds
## 1        1    13      31.85     20.663
## 2        2    25      34.48     11.034
## 3        3    25      46.12     27.213
## 4        4    27      47.00     19.213
## 5        5    18      39.11     10.272
## 6        6    31      42.97     14.775
```

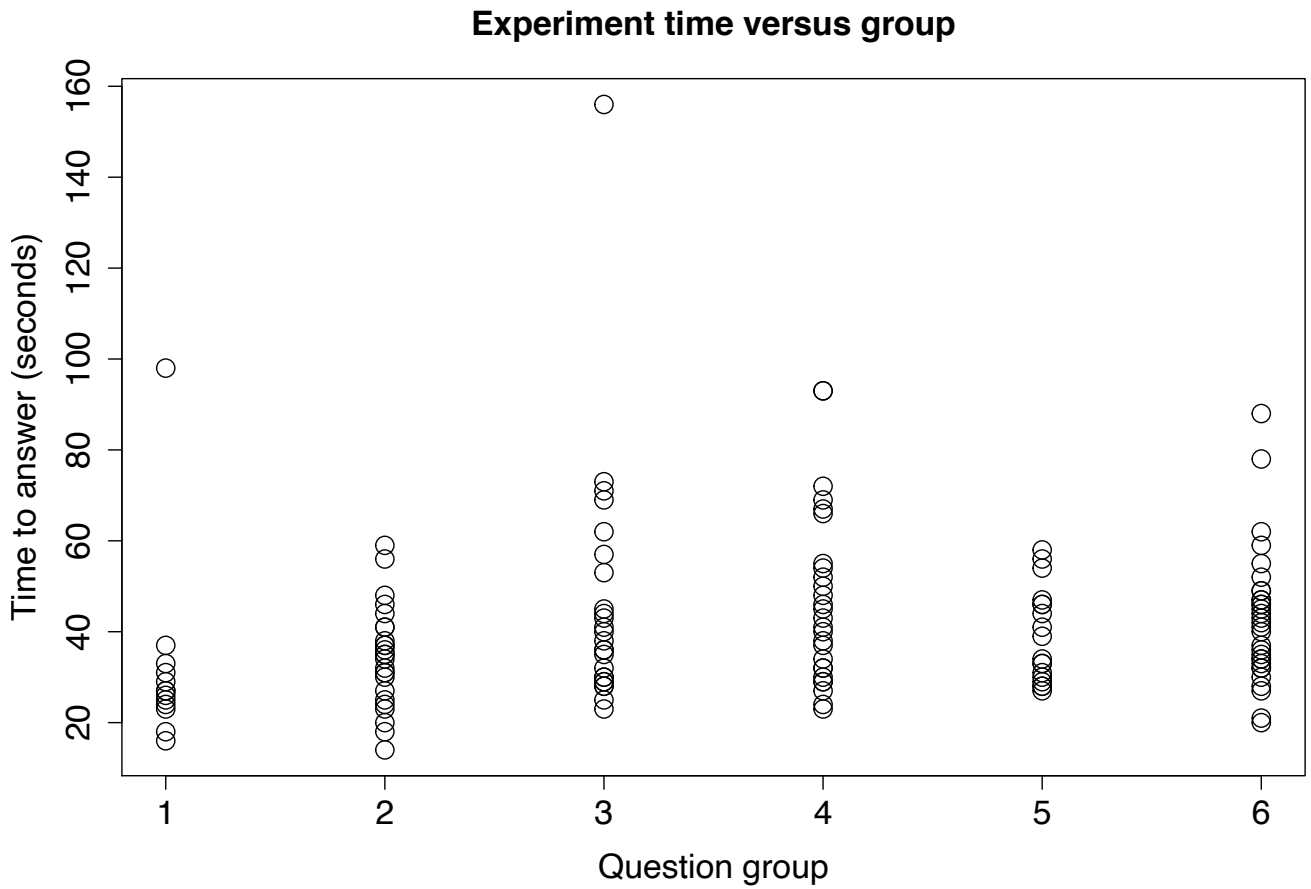
What do we need to think about to continue making sense of this data?

How many got the question wrong?

- One person in each group, except the first one.
- Does that matter?

Visualise the data

Here is a quick look at the first few rows of data and a scatter plot:



What might be the next steps to analyse this data?

Applied Probability II

Section 4: Confidence intervals and hypothesis tests

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 4: Confidence intervals and hypothesis tests

Section 4.1: The sampling distribution of $\bar{Y}$

Using the sample mean to estimate the population mean

In Section 3.2 (Introduction to estimation), we introduced the idea of using the sample mean ($\bar{y}$) to estimate the population mean (μ).

That is, suppose we have collected $y_1, y_2, ..., y_n$, where these values are realisations of random variables $Y_1, Y_2, ..., Y_n$, and come from a random sample and so are independent and follow the same probability distribution.

Then

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \hat{\mu}$$

How good is our estimate $\bar{y}$? Or how close to μ is it?

To answer this, we need to make an assumption about the distribution of the Y_i's.

How precise is our estimate?

We will assume that $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$.

The expected value of any Y_i is $E[Y_i] = \mu$ and the variance is $Var(Y_i) = \sigma^2$.

Consider the random variable

$$\bar{Y} = \frac{\Sigma_{i=1}^{n} Y_i}{n}$$

If we can describe the distribution of $\bar{Y}$, then we can quantify how precise our estimates of μ are (and where we are using $\bar{y}$ to estimate μ).

Aside: properties of expectation and variance

Before we examine the distribution of $\bar{Y}$, here are two useful results, where a and b are constants:

1. The linearity of expectation

$$E[aY_1 + bY_2] = aE[Y_1] + bE[Y_2]$$

2. Properties of the variances:

$$Var(aY_1 + bY_2) = a^2 Var(Y_1) + b^2 Var(Y_2) + 2abCov(Y_1, Y_2)$$
$$Var(aY_1 - bY_2) = a^2 Var(Y_1) + b^2 Var(Y_2) - 2abCov(Y_1, Y_2)$$

The expected value of $\bar{Y}$

Remember, we have assumed that $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$.

Or we can say: $E[Y_i] = \mu$, $Var(Y_i) = \sigma^2$, and the Y_i are normally distributed.

Remember too: $E[aY_1 + bY_2] = aE[Y_1] + bE[Y_2]$.

Then:

$$\begin{aligned} E[\bar{Y}] &= E[\frac{\Sigma_{i=1}^n Y_i}{n}] = \frac{1}{n} E[\Sigma_{i=1}^n Y_i] \\ &= \frac{1}{n} E[Y_1 + Y_2 + ... + Y_n] \\ &= \frac{1}{n} (E[Y_1] + E[Y_2] + ... + E[Y_n]) \\ &= \frac{1}{n} \Sigma_{i=1}^n E[Y_i] = \frac{1}{n} \Sigma_{i=1}^n \mu = \frac{1}{n} n\mu \\ &= \mu \end{aligned}$$

So, our expected value of our estimator of μ is μ. This is good news!

What about the variance of $\bar{Y}$?

The variance of $\bar{Y}$

What is the $Var(\bar{Y})$?

Remember, we have assumed that $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$.

Remember too: $Var(aY_1 - bY_2) = a^2 Var(Y_1) + b^2 Var(Y_2) - 2abCov(Y_1, Y_2)$

$$
\begin{aligned}
Var(\bar{Y}) &= Var(\frac{\Sigma_{i=1}^n Y_i}{n}) = \frac{1}{n^2} Var(\Sigma_{i=1}^n Y_i) \\
&= \frac{1}{n^2} Var(Y_1 + Y_2 + ... + Y_n) \\
&= \frac{1}{n^2} [Var(Y_1) + Var(Y_2) + ... + Var(Y_n)] \\
&= \frac{1}{n^2} \Sigma_{i=1}^n Var(Y_i) = \frac{1}{n^2} \Sigma_{i=1}^n \sigma^2 = \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

What about the covariances in line 3 of the equation?

If two random variables are independent, then their covariance is equal to 0.

The sampling distribution of $\bar{Y}$

Recap, we want to know the distribution of $\bar{Y}$.

We have assumed

- $Y_i \sim N(\mu, \sigma^2)$, and
- the Y_i's are independent.

We have already proven

- $E[\bar{Y}] = \mu$, and
- $Var(\bar{Y}) = \frac{\sigma^2}{n}$

Final step

$\bar{Y}$ can be written as: $\bar{Y} = \frac{1}{n}Y_1 + \frac{1}{n}Y_2 + ... + \frac{1}{n}Y_n$

and so is a linear combination of normal random variables.

We can assume: the linear combination of independent normal random variables is also normal.

Therefore $\bar{Y}$ is itself a normal random variable. Hence, we've shown that

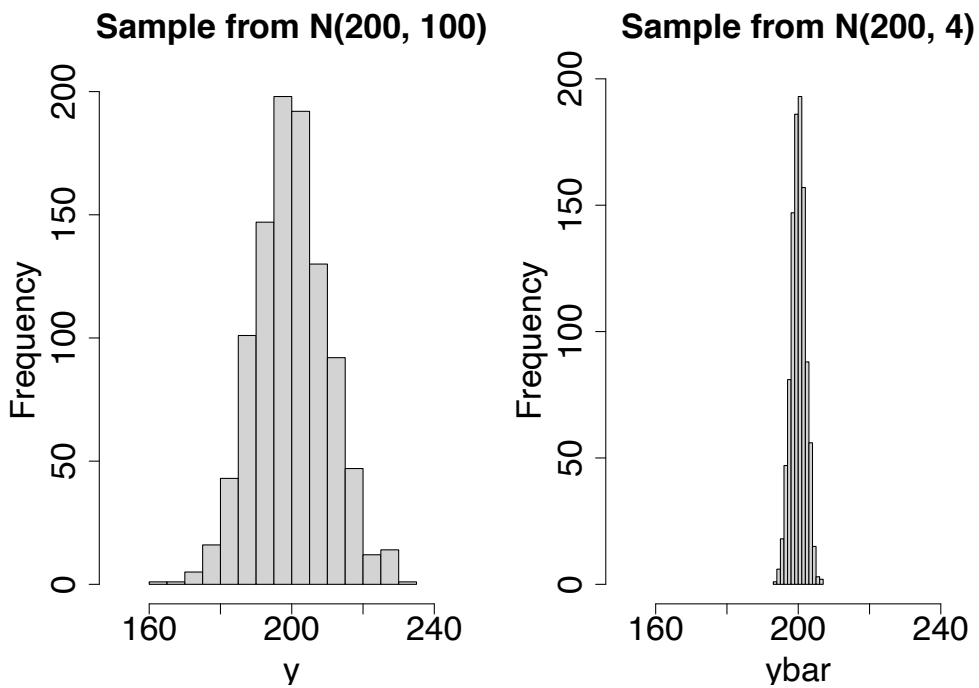
$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$$

An example

Assume we have $Y_i \sim N(200, 100)$.

Then what is the sampling distribution of $\bar{Y}$ for samples of size 25?

$\bar{Y} \sim N(200, \frac{100}{25}) = N(200, 4)$.



Final note

Since

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$$

we can see that the variance of $\bar{Y}$ decreases as n increases. Or, we can say that our estimate of the mean is more precise the larger that n is (which makes sense intuitively!).

Recall from previous modules, that if $X \sim N(\mu_X, \sigma_X^2)$, then

$$\frac{X - \mu_X}{\sigma_X} \sim N(0, 1)$$

This is referred to as standardising X.

Applying this to $\bar{Y}$, we get:

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

We will use this in the next Section to construct a confidence interval for μ.

Section 4.2: Confidence interval for μ (variance known)

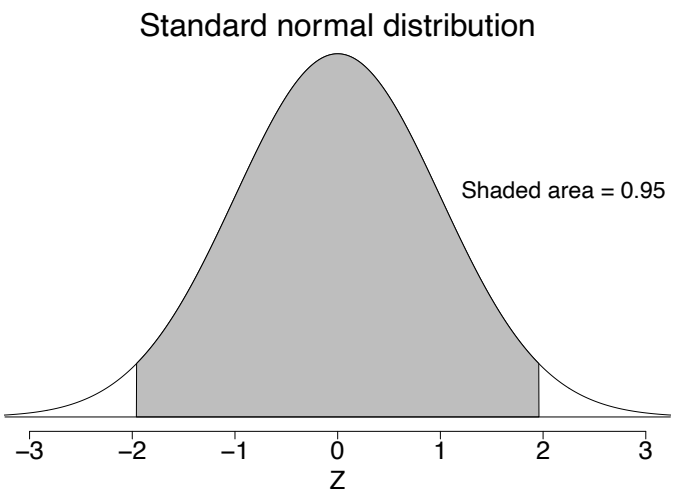
Assume we know σ

We will construct a CI for a population mean.

- To make things easier, assume we know the true value of σ.
- We are still assuming that $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$.
- Therefore we know from the previous section that $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$.

For the standard normal distribution, $Z \sim N(0, 1)$, we know that

$$Pr(-1.96 \leq Z \leq 1.96) = 0.95.$$



Constructing the CI

Since $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, this tells us that:

$$Pr(-1.96 \leq \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

$$Pr(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{Y}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

$$Pr(-1.96\frac{\sigma}{\sqrt{n}} \leq \mu-\bar{Y} \leq 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

$$Pr(\bar{Y}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y}+1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

This tells us that the probability the true value of μ lies in the interval

$$(\bar{Y}-1.96\frac{\sigma}{\sqrt{n}}, \bar{Y}+1.96\frac{\sigma}{\sqrt{n}})$$

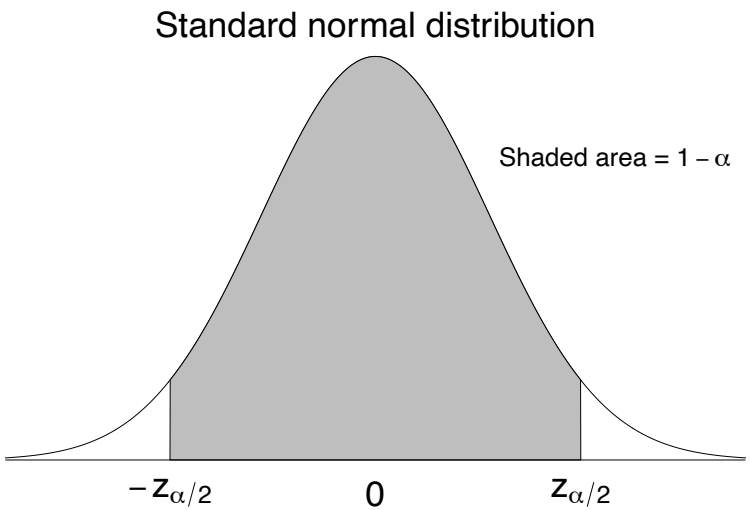
is 0.95, or we have 95% confidence that μ will be in any randomly selected interval.

95% of all realised intervals $\bar{y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ will contain μ.

Confidence interval for $100(1-\alpha)\%$

We don't need to restrict ourselves to 95% confidence. We can find the relevant $z_{\frac{\alpha}{2}}$ values (called critical values) from tables to construct a $100(1-\alpha)\%$ confidence interval for μ.

Where:



The 100(1-α)% confidence interval for μ is then:

$$\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Section 4.3: Confidence interval for μ (variance unknown)

What about when we don't know σ?

In practice, we generally don't know σ, but in the previous section, we assumed that σ was known.

What to we do when σ is unknown?

We use the sample standard deviation, $S = \sqrt{\frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$, and replace σ by this estimate.

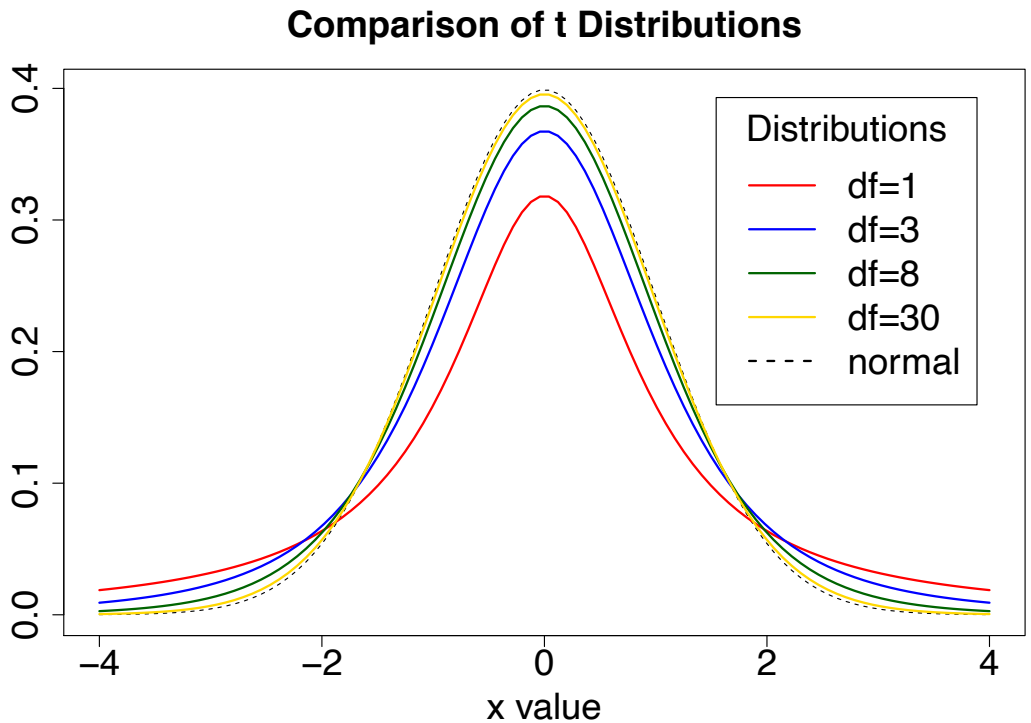
Comparing $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ with $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$, we know one less thing in the second one. Thus we have more uncertainty.

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

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

The t distribution

The t distribution has a heavier tail than the normal distribution.



As the degrees of freedom increase, the t-density curve approaches the N(0,1) density curve and the $t \approx N(0,1)$ curve for degrees of freedom ≥ 30.

Construct the CI

We are still assuming that $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$, but σ^2 is unknown.

We use the t-distribution to construct CIs.

We can write $t_{\nu, \frac{\alpha}{2}}$ to represent the critical values from the t distribution with $\nu = n - 1$ the degrees of freedom.

$$P(-t_{\nu, \frac{\alpha}{2}} < \frac{\bar{Y} - \mu}{S/\sqrt{n}} < t_{\nu, \frac{\alpha}{2}}) = 1 - \alpha$$

$$P(\bar{Y} - t_{\nu, \frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \bar{Y} + t_{\nu, \frac{\alpha}{2}} \frac{S}{\sqrt{n}}) = 1 - \alpha$$

Therefore, a 100(1 - α)% confidence interval when σ isn't known is

$$\bar{y} \pm t_{\nu, \alpha/2} \frac{s}{\sqrt{n}}$$

where we read $t_{\nu, \alpha/2}$ from the t-tables.

Example

Let's go back to the example we did in Section 2 of the notes.

A pharmaceutical company has developed a drug they believe will help cure a particular disease and has acquired legal permission to test the drug on humans. However, there are safety concerns that the drug will have an adverse effect on blood pressure by increasing it. They run an experiment to estimate the average change in blood pressure for patients who take the new drug. They only have permission to test the drug on eleven patients.

Here are the data values:

```
x <- c(1.1, 1.8, 2, 2.4, 2.5, 2.8, 2.9, 3, 3.4, 3.4, 4)
x
```

```
##  [1] 1.1 1.8 2.0 2.4 2.5 2.8 2.9 3.0 3.4 3.4 4.0
```

Population parameter of interest: μ, the true mean change in blood pressure for the population of people who take the drug.

Assume that the data represent a random sample from the population of all people who take the drug.

Example contd.

95% confidence interval for μ: (2.11, 3.22)

This is computed as

$$\bar{x} \pm t_{\nu, \frac{0.05}{2}} \frac{s}{\sqrt{n}}$$

where $\bar{x} = 2.664$, $s = 0.8237$, $n = 11$, $t_{10,0.025} = 2.228$

Before we interpret the confidence, what assumptions have we made for the inference (interpretation) to be valid?

Recall: When we fit any statistical model to data, or perform any statistical analysis, we must:

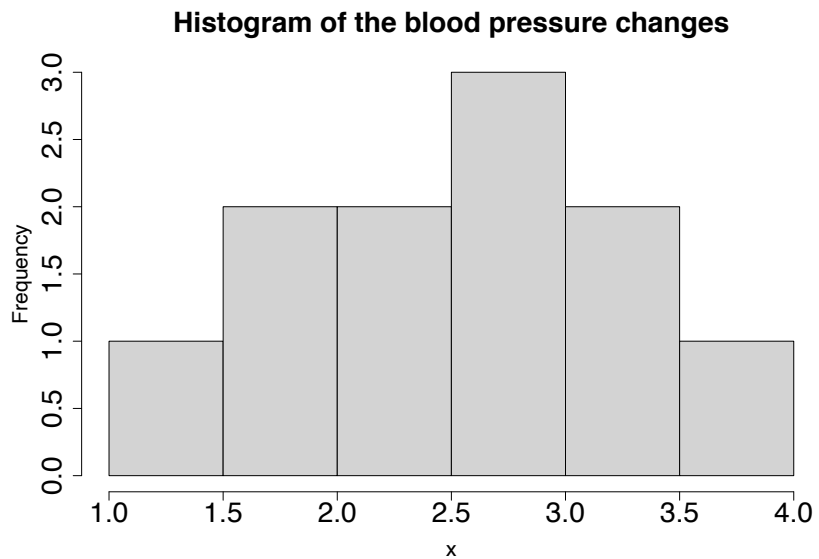
1. Know what (if any) assumptions are being made by the model or analysis.
2. Verify that those assumptions are reasonable.

Example contd.

We have assumed that the observations are independent of each other.

We have assumed that the population from which the observations were drawn is normally distributed.

- Is this reasonable?
- How might we check this?



We are 95% confident that μ, the true population mean change in blood pressure lies in the interval (2.11, 3.22).

Output from R

```
x
```

```
##  [1] 1.1 1.8 2.0 2.4 2.5 2.8 2.9 3.0 3.4 3.4 4.0
```

```
t.test(x, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  x
## t = 10.725, df = 10, p-value = 8.342e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  2.110241 3.217032
## sample estimates:
## mean of x
##  2.663636
```

Section 4.4: Hypothesis tests for μ (variance unknown)

Construction of a hypothesis test

We can carry out hypothesis tests for the population mean (μ) in the following way.

- Start by specifying the hypotheses. For example:

  H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$.

  This is a two-sided alternative hypothesis, one-sided would be:

  H_A: $\mu < \mu_0$ or H_A: $\mu > \mu_0$.
- Construct a test statistic.
- Evaluate the test statistic against the "null" distribution.
- Make a conclusion.

For now, we will assume $Y_1, ..., Y_n \sim N(\mu, \sigma^2)$ independent and that σ is unknown.

What is the logic behind a hypothesis test?

We assume that the null hypothesis, $\mu = \mu_0$, is true. We then examine the observed data to see if there is evidence to the contrary.

Computing the test statistic

If the H_0 is in fact true, then:

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

This is used as a reference to compare what we see in the data.

Calculate the observed test statistic (a realisation from the above distribution)

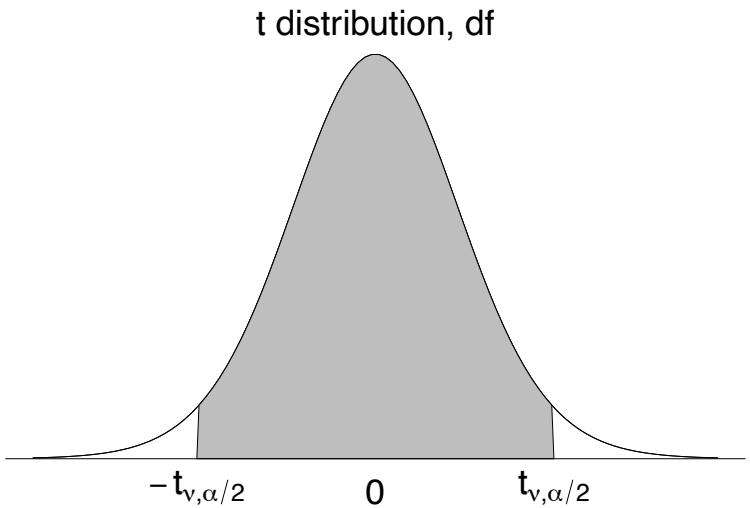
$$T_{obs} = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

A test of level α will compare T_{obs} with the corresponding critical value(s) and based on the comparison we either:

- Reject the H_0 and accept the H_A, or
- Fail to reject the H_0 (which does not mean that we have proven it is true)

Evaluating the hypothesis test - two sided test

The sampling distribution of the test statistic T, with possible values along the x axis:

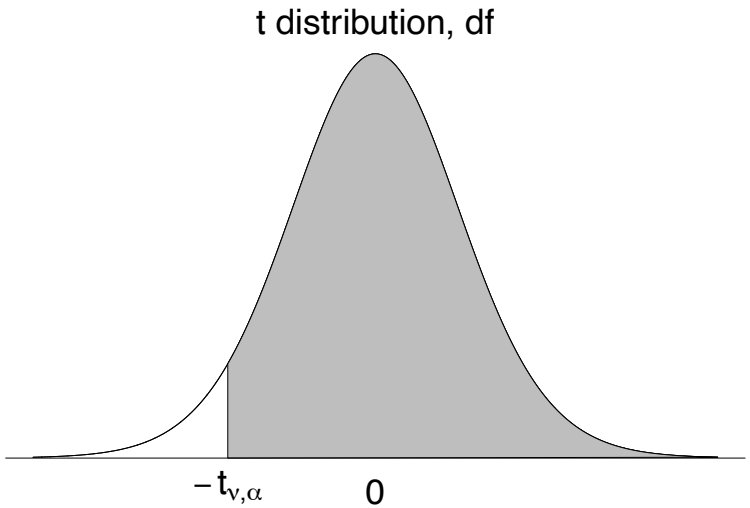


If an observed test statistic value T_{obs} lies along the region shaded in grey, this is 'typical' for this distribution and consistent with the null hypothesis. If the observed test statistic lies outside this, it is evidence for the alternative hypothesis, at the specified α level. Formally, for a two-sided test:

- Reject the H_0 and accept the H_A, when $T_{obs} \leq -t_{\nu,\alpha/2}$ or $T_{obs} \geq t_{\nu,\alpha/2}$.
- Fail to reject the H_0 when $-t_{\nu,\alpha/2} < T_{obs} < t_{\nu,\alpha/2}$.

Evaluating the hypothesis test for one-sided alternatives

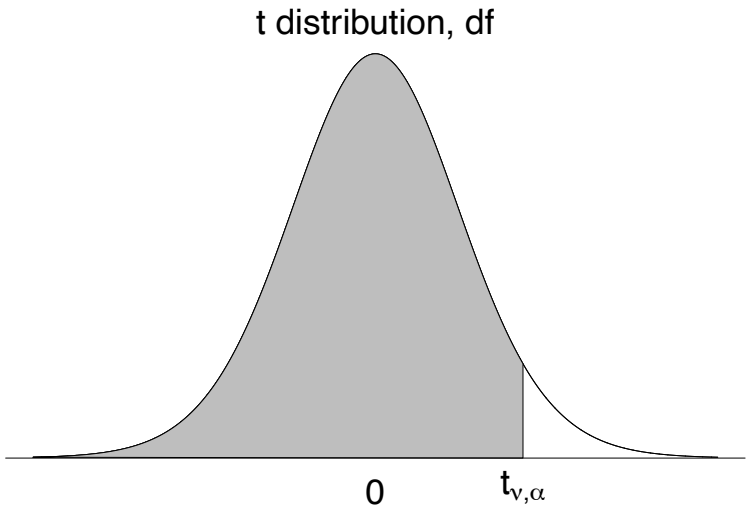
For a lower tailed test for a specified α level:



- Reject the H_0 and accept the H_A, when $T_{obs} \leq -t_{\nu,\alpha}$.
- Fail to reject the H_0 when $T_{obs} > -t_{\nu,\alpha}$.

Evaluating the hypothesis test for one-sided alternatives

For an upper tailed test for a specified α level:



- Reject the H_0 and accept the H_A, when $T_{obs} \geq t_{\nu,\alpha}$.
- Fail to reject the H_0 when $T_{obs} < t_{\nu,\alpha}$.

Additional notes

- It is possible that a null hypothesis is rejected, when in fact it is true. The probability of this happening here is α. We call this a Type I Error.

$$P(\text{Reject } H_0 \text{ when true}) = \alpha = P(\text{Type I Error})$$

- When specifying the hypotheses, the decision about what value μ_0 takes must be made **before** collecting and examining the data. We don't ever observe our sample mean, and then decide on a hypothesis to test.

- When it comes to specifying the alternative hypothesis, the default is to pick a two-sided test. If there is a reason before collecting the data for a belief that the true mean is in one direction, then a one-sided test can be done.

- NB: Don't ever decide on the value of μ_0, or decide to do a one-sided test based on the observed sample mean.

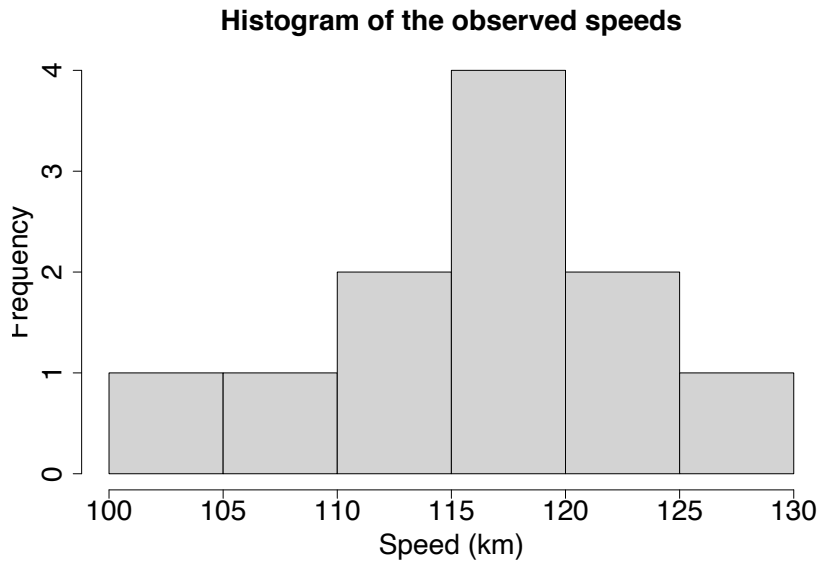
Example

There is a speed camera on a motorway where the speed limit is 120km. The camera observes the speed of 11 cars at random and records:

The 11 recordings are: 104, 109, 112, 114, 117, 118, 118, 120, 121, 124, 130.

Test the hypothesis that the true mean speed of vehicles on the motorway is under the speed limit at the $\alpha = 0.05$ level.

First, let's examine the data.



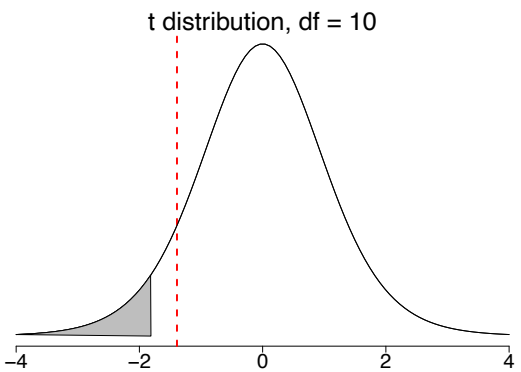
Is it reasonable to assume a normal population?

Example worked out

- Let μ be the mean population speed of all cars on the motorway.
- $H_0 : \mu = 120$ vs $H_A : \mu < 120$.
- Under the H_0, the test statistic $T \sim t(10)$. Find the observed test statistic:

$$T_{obs} = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{117 - 120}{7.15542/\sqrt{11}} = -1.39$$

- The critical value is $t_{\nu=10,\alpha=0.05} = -1.812$.



- We fail to reject the H_0 ($\alpha = 0.05$, $t_{10,0.05} = -1.812 < -1.39$, lower-tailed test) and therefore have no evidence that the true mean speed of cars on the motorway is lower than 120, the speed limit.

Output from R

```
speeds <- c(104, 109, 112, 114, 117, 118, 118, 120, 121, 124, 130)
speeds
```

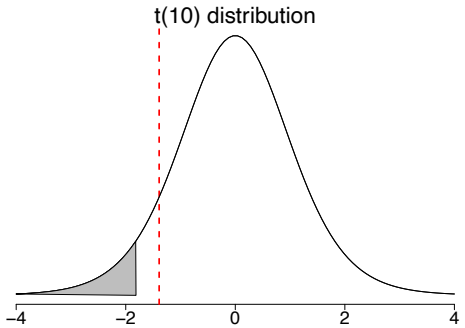
```
##  [1] 104 109 112 114 117 118 118 120 121 124 130
```

```
t.test(speeds, mu = 120,  alternative = "less")
```

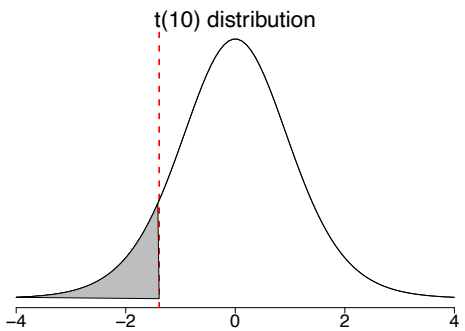
```
##
##  One Sample t-test
##
## data:  speeds
## t = -1.3905, df = 10, p-value = 0.09727
## alternative hypothesis: true mean is less than 120
## 95 percent confidence interval:
##      -Inf 120.9103
## sample estimates:
## mean of x
##       117
```

Using p-values or critical values to evaluate a hypothesis test

The critical value = -1.812; this means the P($t(10) <$ -1.812) = 0.05.



P-value = probability of observing a test statistic as extreme, or more extreme, under the null hypothesis. $T_{obs} = -1.39$. P-value = P($t(10) <$ -1.39) = 0.09727.



If the p-value $< \alpha$ we reject the H_0.

Section 4.5: Confidence intervals and hypothesis tests for σ^2

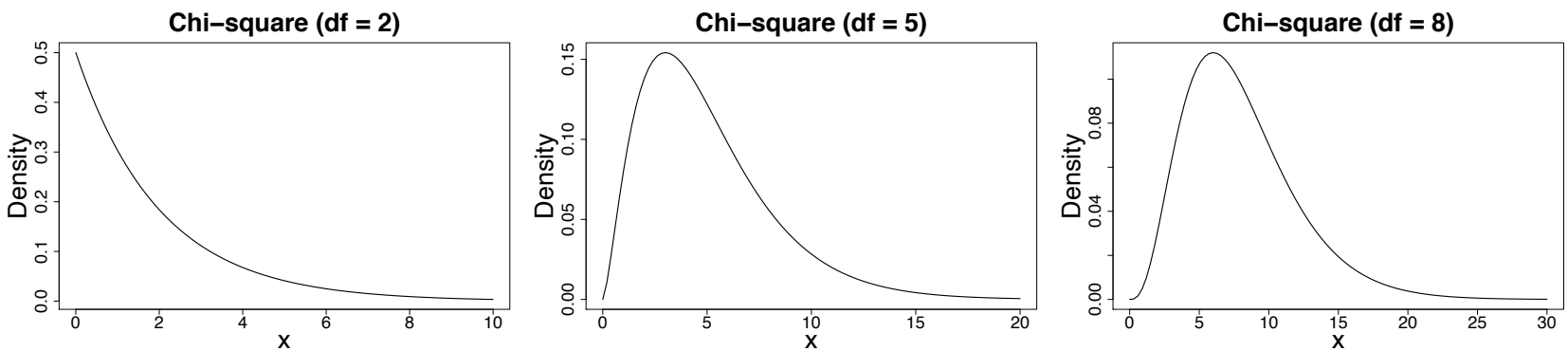
The chi-square distribution

If $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$, then we can use the following information to construct confidence intervals for σ^2:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

(We will assume this without proof.)

The shape of the chi-square distribution depends on the degrees of freedom:



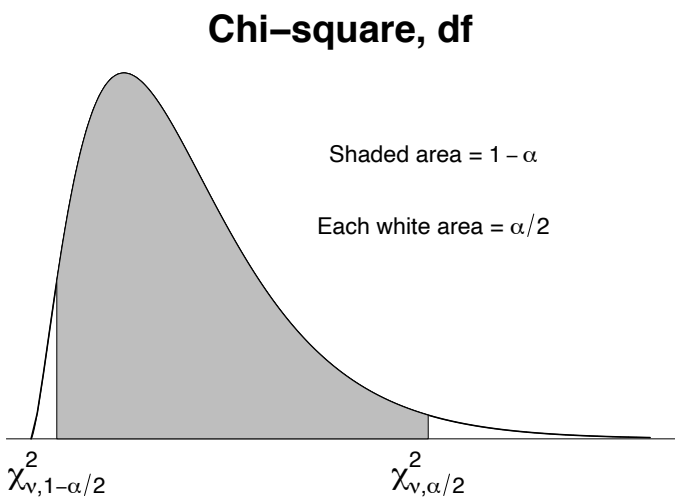
Construct a CI for σ^2

For degrees of freedom $\nu = n - 1$

$$P(\chi^2_{\nu,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\nu,\alpha/2}) = 1 - \alpha$$

and a $100(1-\alpha)\%$ confidence interval for σ^2 can be shown to be

$$(\frac{(n-1)s^2}{\chi^2_{\nu,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\nu,1-\alpha/2}})$$



Hypothesis test for σ^2

We will follow the same general steps to do a hypothesis test as before. That is:

- Specify the hypotheses, for example

  H_0: $\sigma^2 = \sigma_0^2$ versus H_A: $\sigma^2 > \sigma_0^2$.

- Construct a test statistic.
- Evaluate the test statistic against the "null" distribution.
- Make a conclusion.

If the H_0 is true, then,

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

and for a given sample, the test statistic

$$\chi^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

is a realisation of a $\chi^2(n-1)$ random variable.

Evaluating the test statistic

If H_0: $\sigma^2 = \sigma_0^2$ versus H_A: $\sigma^2 > \sigma_0^2$, and

the test statistic has been computed:

$$\chi^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

We evaluate the test statistic against the "null" distribution. Under the H_0, the test statistic comes from a $\chi^2(n-1)$.

Here are the possible outcomes (with degrees of freedom $\nu = n-1$):

- We reject the H_0 and accept the H_A if $\chi^2_{obs} \geq \chi^2_{\nu,\alpha}$.
- We fail to reject the H_0 if $\chi^2_{obs} < \chi^2_{\nu,\alpha}$.

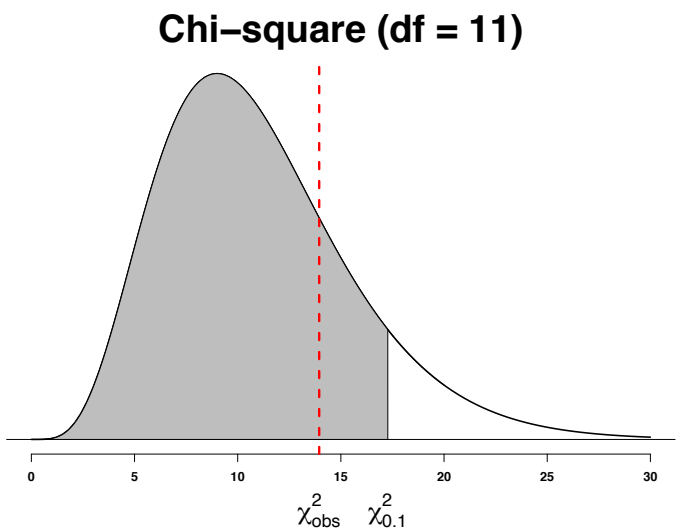
Example

A sample of data of size 12 was collected from a normally distributed population. The standard deviation was computed as 4.362.

We wish to test the hypothesis that H_0: $\sigma^2 = 15$ versus H_A: $\sigma^2 > 15$, using $\alpha = 0.1$.

$$\chi^2_{obs} = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(12-1)4.362^2}{15} = 13.95$$

The test statistic $\chi^2_{obs} = 13.95$ which is $< \chi^2_{0.1}(11) = 17.275$. At $\alpha = 0.1$, we fail to reject the H_0 and have no evidence that $\sigma^2 > 15$.



Final word

- Up to now, when you have done hypothesis testing or used confidence intervals, you have taken the methods at face value.
- Now, you know about the underlying theory.

- In this section, we have focused on data that is from populations that are normally distributed.
- In the next section, we will explore beyond that.

Applied Probability II

Section 5: The Central Limit Theorem

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 5: The Central Limit Theorem

Section 5.1: Statement of the Central Limit Theorem (CLT)

Introduction

In Section 4, we saw the following results.

When $X_1, ...X_n \sim N(\mu, \sigma^2)$,

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

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T(n-1)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

We used these to construct confidence intervals and hypothesis tests for μ and σ^2.

However, if the X_is depart from the normal distribution, how can we carry out hypothesis tests and construct confidence intervals for the population mean μ?

The Central Limit Theorem is a very general (and important) results describing the properties of $\bar{X}$ in general.

Statement of the Central Limit Theorem (CLT)

Theorem (without proof)

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with mean μ and variance σ^2, i.e., $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$. Then, any linear combination of the X_i random variables, follows a normal distribution.

Let that sink in for a moment...

In particular, we have not assumed anything about the shape of the distribution of the X_i, only that they are IID with mean μ and variance σ^2.

We will now look at some particular cases of the CLT.

Case 1

Let $X_1, ..., X_n$ be independent random variables.

Let $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$.

Consider $\sum_{i=1}^{n} X_i$. This is a linear combination of the X_is. Therefore, by the CLT,

$$\sum_{i=1}^{n} X_i \underset{approx}{\sim} N(n\mu, n\sigma^2)$$

Where do the mean and variance come from? We can see that

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu = n\mu$$

and

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

(with covariances equal to 0).

Case 2

Let $X_1, ..., X_n$ be independent random variables.

Let $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$.

Consider $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. This is a linear combination of the X_is. Therefore, by the CLT,

$$\bar{X} \underset{approx}{\sim} N(\mu, \frac{\sigma^2}{n})$$

Where do the mean and variance come from? We can see that

$$E[\frac{\sum_{i=1}^{n} X_i}{n}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

and

$$Var(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

(with covariances equal to 0).

The value of the CLT

The Central Limit Theorem is a powerful result.

Case 2 on the previous slide is one of the most common uses of the CLT.

It also tells us that, if $X_1, ..., X_n$ are independent random variables with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$, then the random variable

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

The larger the n, the closer the approximation (asymptotic result).

How large is large? Typically, as a rule of thumb, we say that the approximation works well for samples ≥ 30.

This opens the door for us to do tests of hypothesis and construct approximate confidence intervals for population parameters for data that comes from distributions other than the normal distribution.

Common misunderstanding of the CLT

Please do not make this mistake!

If $X_1, ..., X_n$ are independent random variables, and $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$. Then the CLT tells us that, as n increases, $\bar{X}$ tends to a normal distribution. This does not mean that the $X_1, ..., X_n$ values are changing distribution as n increases!

For example, if $X_1, ..., X_n$ are independent and exponentially distributed, as n increases, $\bar{X}$ tends to a normal distribution. But the $X_1, ..., X_n$'s remain exponentially distributed, no matter the n.

Section 5.2: Application of the CLT - finding probabilities

Example

Suppose scores from a national test have a distribution with mean 50 and standard deviation 10.

Suppose 35 results are selected at random from the test scores. What is the probability that their average test result is greater than 55?

- Let the random variable X_i denote an individual test score result.
- We know the $E[X_i] = \mu = 50$ and $Var(X_i) = \sigma^2 = 100$, but we don't know the shape of the distribution (and cannot make the assumption that the X_i are normally distributed).
- Let $\bar{X} = \frac{\sum_{i=1}^{35} X_i}{35}$. The CLT tells us that $\bar{X}$ is approximately normally distributed, or that

$$\bar{X} \underset{approx}{\sim} N(\mu, \frac{\sigma^2}{n}) = N(50, \frac{100}{35})$$

We can use this to work out an approximate probability.

$$\begin{aligned} P(\bar{X} > 55) &= P(\frac{\bar{X} - 50}{\sqrt{100/35}} > \frac{55 - 50}{\sqrt{100/35}}) \\ &= P(Z > 2.96) = 1 - 0.9985 \\ &= 0.0015 \end{aligned}$$

Section 5.3: Application of the CLT - confidence intervals

Example

A study was conducted where a random sample of 500 people were surveyed and asked if they supported a particular measure being imposed by government policy.

- Let p be the true proportion of people in the (large) population that support the policy.
- Let X be a random variable representing the number in a randomly selected sample of 500 people that will answer yes. What distribution does X have?
- $X \sim \text{Binomial}(n, p)$.

Suppose that 315 people of those surveyed support the policy. Then we can estimate the true population proportion p, using

$$\hat{p} = \frac{315}{500} = 0.63$$

But how reliable is this estimate?

Example contd.

We will use the CLT to construct a confidence interval for p.

- We have that $X \sim \text{Binomial}(n = 500, p)$. We have an estimate for p, but do not know the true population value.
- Remember that for the Binomial distribution, $\text{E}[X]{=}np$ and $\text{Var}(X){=}np(1-p)$. Aside challenge (recap from Applied Probability I): prove these in your own time!
- We can write $X = X_1 + ... + X_{500}$, where $X_i = 1$ if the person answers yes, and 0 if the person answer no. Each $X_i \sim \text{Binomial}(n = 1, p)$, or Bernoulli(p).
- This means that X is a sum of independent random variables $X_1, ..., X_{500}$ and each X_i has $\text{E}[X_i]{=}p$ and $\text{Var}(X_i){=}p(1-p)$.
- Then, by the CLT, since n is sufficiently large,

$$X \underset{approx}{\sim} N(np, np(1-p))$$

$$\frac{X - np}{\sqrt{np(1-p)}} \underset{approx}{\sim} N(0, 1)$$

And, we can then get:

$$\frac{X/n - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{approx}{\sim} N(0, 1)$$

Example contd.

To construct a confidence interval for p (the population parameter), use

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This gives us the (approximate) confidence interval

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Back to our example, the 95% confidence interval for p, the true proportion of people in the population that support the government policy, is:

$$0.63 \pm 1.96 \times \sqrt{\frac{0.63(1-0.63)}{500}}$$

$$0.63 \pm 1.96 \times 0.021592$$

$$= (0.588, 0.672)$$

We are 95% confident that the true population proportion of people that support the measure lies in this interval.

Section 5.4: Application of the CLT - hypothesis testing

Example dataset: Lateral bias

- There is a well observed phenomenon that human mothers tend to nurse their babies on the left hand side.
- One belief is that the left field of vision causes flow of information in the right hemisphere of the brain (which controls many aspects of social behaviour).

- Is there a preference for the left for other mammals?
- This has been studied in two phylogenetically distant species of mammal: Pacific Walrus and Indian Flying Fox. (Source: Giljov, Karenina and Malashichev, 2018, Biology Letters, 14, 20170707.)

- Want to test whether there is a preference for the left side.
- Think of this like tossing a (possibly biased) coin in each instance, where p (probability of heads) is the probability of the baby being on the left.
- We wish to carry out the hypothesis test

$$\mathrm{H}_0 : p = 0.5, \text{ versus } \mathrm{H}_A : p > 0.5$$

Pacific Walrus and Indian Flying Fox

What was observed?

Whether or not there was a preference for the left was observed for the following scenarios

1. floating on side of mother for Pacific Walrus

and

2. hanging at side of mother for Indian Flying Fox.

Is there evidence of a preference for the left?

Setting up the hypothesis test

In each experiment we are assuming to carry out n independent Bernoulli trials $X_1, ..., X_n$ each with outcome 0 (no lateral bias) or 1 (lateral bias).

We have: $P(X_i = 1) = p, P(X_i = 0) = 1 - p$.

Aside challenge: Prove that the $E[X_i] = p$ and that $Var(X_i) = p(1 - p)$.

Let $X = \sum_{i=1}^{n} X_i$.

We have, $E[X/n] = p$ and $\text{Var}(X/n) = \frac{p(1-p)}{n}$.

With collected data, $\hat{p} = \frac{x}{n}$.

From the CLT,

$$\frac{X/n - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{approx}{\sim} N(0, 1)$$

If we assume a null hypothesis of $H_0 : p = p_0$, then this becomes:

$$\frac{X/n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \underset{approx}{\sim} N(0, 1)$$

Performing the hypothesis test - Pacific Walrus

For the Pacific Walrus data, $n = 44$.

There were 36 on the left, giving $\sum x_i = 36$.

H_0: $p = p_0$ versus $H_A : p > p_0$.

(One-sided because there was *a priori* belief that there is left lateral bias.)

$\hat{p} = 36/44 = 0.8182$.

Assume that the H_0 is true. If H_0 really is true, then

$$z_{obs} = \frac{0.8182 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{44}}} = 4.221$$

is a realisation from a N(0, 1).

Using $\alpha = 0.05$, we reject the test if $z_{obs} > 1.645$ (where $P(Z > 1.645) = 0.05$). We reject the H_0 and conclude that there is evidence of left preference in Pacifc Walrus.

Performing the hypothesis test - Indian Flying Fox

For the Indian Flying Fox data, $n = 59$.

There were 43 on the left, giving $\sum x_i = 43$.

H_0: $p = p_0$ versus $H_A : p > p_0$.

(One-sided because there was *a priori* belief that there is left lateral bias.)

$\hat{p} = 43/59 = 0.7288$.

Assume that the H_0 is true. If H_0 really is true, then

$$z_{obs} = \frac{0.7288 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{59}}} = 3.515$$

is a realisation from a N(0, 1).

Using $\alpha = 0.05$, we reject the test if $z_{obs} > 1.645$ (where $P(Z > 1.645) = 0.05$). We reject the H_0 and conclude that there is evidence of left preference in Indian Flying Fox.

Applied Probability II

Section 6: Linear Regression

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 6: Linear Regression

Section 6.1: Simple linear regression model

A model for the mean

In Section 4.1, we assumed that $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$.

If we collect a sample of data from this distribution, we can estimate μ using $\bar{y}$, the sample mean. We also described the sampling distribution of $\bar{Y}$, and used it to construct CIs and test hypotheses about μ.

We could also have referred to this as 'fitting a model for the mean' and written it as:

$$Y_i = \mu + \epsilon_i$$

where ϵ_i is a random variable which is normal with mean 0 and variance σ^2.

I.e., $E[\epsilon_i] = 0$, and $Var(\epsilon_i) = \sigma^2$, and $\epsilon_i \sim N(0, \sigma^2)$.

We can see that

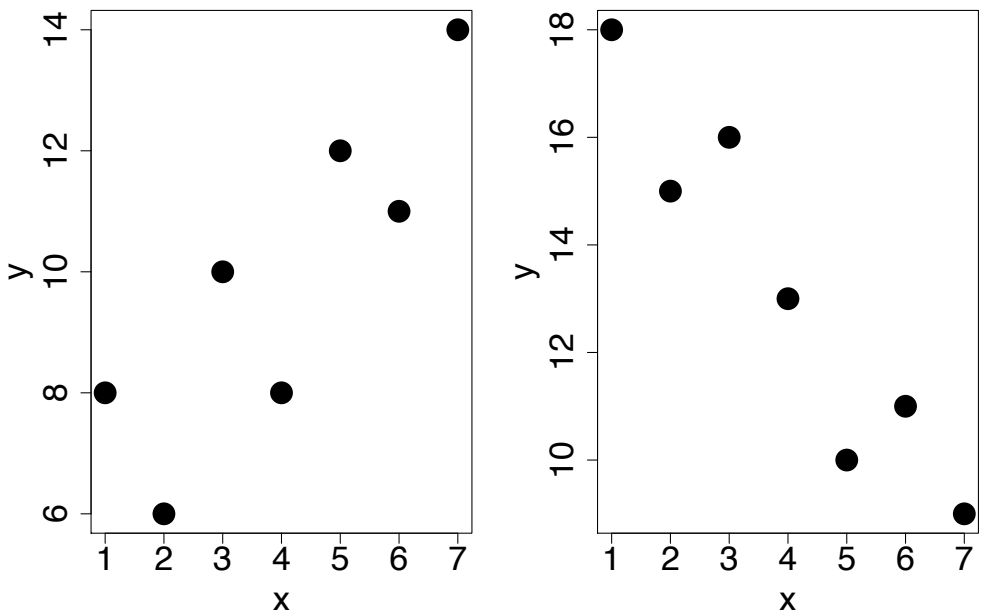
$$E[Y_i] = E[\mu + \epsilon_i] = \mu + E[\epsilon_i] = \mu$$
$$Var(Y_i) = Var(\mu + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

The ϵ_i terms are called the 'errors', i.e., how far away Y_i is from μ.

Simple linear regression motivation

Suppose now, in addition to observations $y_1, y_2, ..., y_n$, we have further data information $x_1, x_2, ...x_n$. We believe that knowing x can help us to predict y.

For example, a scatter plot could look like:



Simple linear regression equation

The simple linear regression model equation is

$$Y_i = \mu_{Y|x_i} + \epsilon_i$$

$$\mu_{Y|x_i} = E[Y|x_i] = \beta_0 + \beta_1 x_i$$

Or the more common way to express the simple linear regression model is:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where

β_0: is the intercept parameter, the expected mean of Y when $x = 0$.

and

β_1: is the slope parameter, the change in the expected mean of Y for a one unit increase in x.

The simple linear regression model assumptions

For the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

We assume that

1. For a fixed x, Y is a normally distributed random variable with mean $\beta_0 + \beta_1 x$.
2. The variance of Y does not depend on x. I.e. $Var(Y|x) = Var(Y) = \sigma^2$.
3. The values of Y are independent (uncorrelated).
4. The model is linear: $E[Y|x] = \beta_0 + \beta_1 x$.

We can also express the assumptions in terms of the errors.

1. $E[\epsilon_i] = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ (and does not depend on i).
3. ϵ_i are independent.
4. $\epsilon_i \sim N(0, \sigma^2)$.

Predicting and residuals

To predict from a fitted simple linear regression model for any x value, we use:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where $\hat{\beta}_0$ is the estimated intercept, and $\hat{\beta}_1$ is the estimated slope.

Side note on terminology: $\hat{\beta}_0$, β_0, $\hat{\beta}_1$, β_1. What's the difference?

To assess the model fit and assumptions, we can use residuals, where residuals are defined as

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

If we have fitted a simple linear regression model to n paired (x_i, y_i) data values, then we will have n observed values (y_i), n corresponding fitted or predicted values $(\hat{y}_i)$ and can find n corresponding residuals (observed minus predicted).

Simple linear regression example

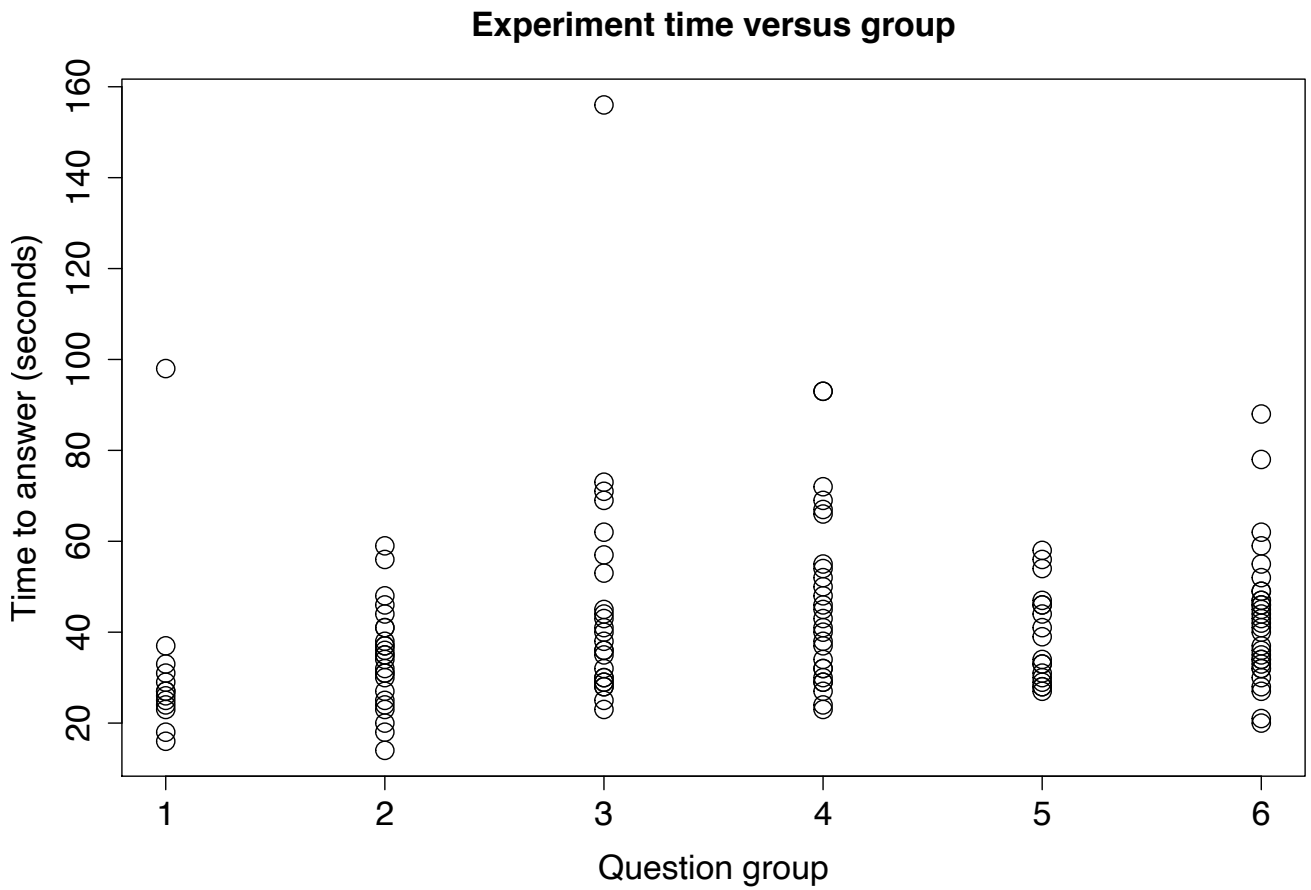
Remember the class experiment we did on the first day of term?

- Each person was asked just one question, but there were six different questions.
- Each question required you to put a list of words in alphabetical order; they were:
  - bouncing handle kitchen tracksuit university
  - **w**ashing **w**eird **w**hich **w**isdom **w**onderful
  - **mo**bile **mo**del **mo**ment **mo**untain **mo**vie
  - **sta**ck **sta**ke **sta**ndard **sta**pler **sta**tistics
  - **star**board **star**dom **star**fish **star**ry **star**tle
  - **cross**bar **cross**fire **cross**ing **cross**road **cross**word

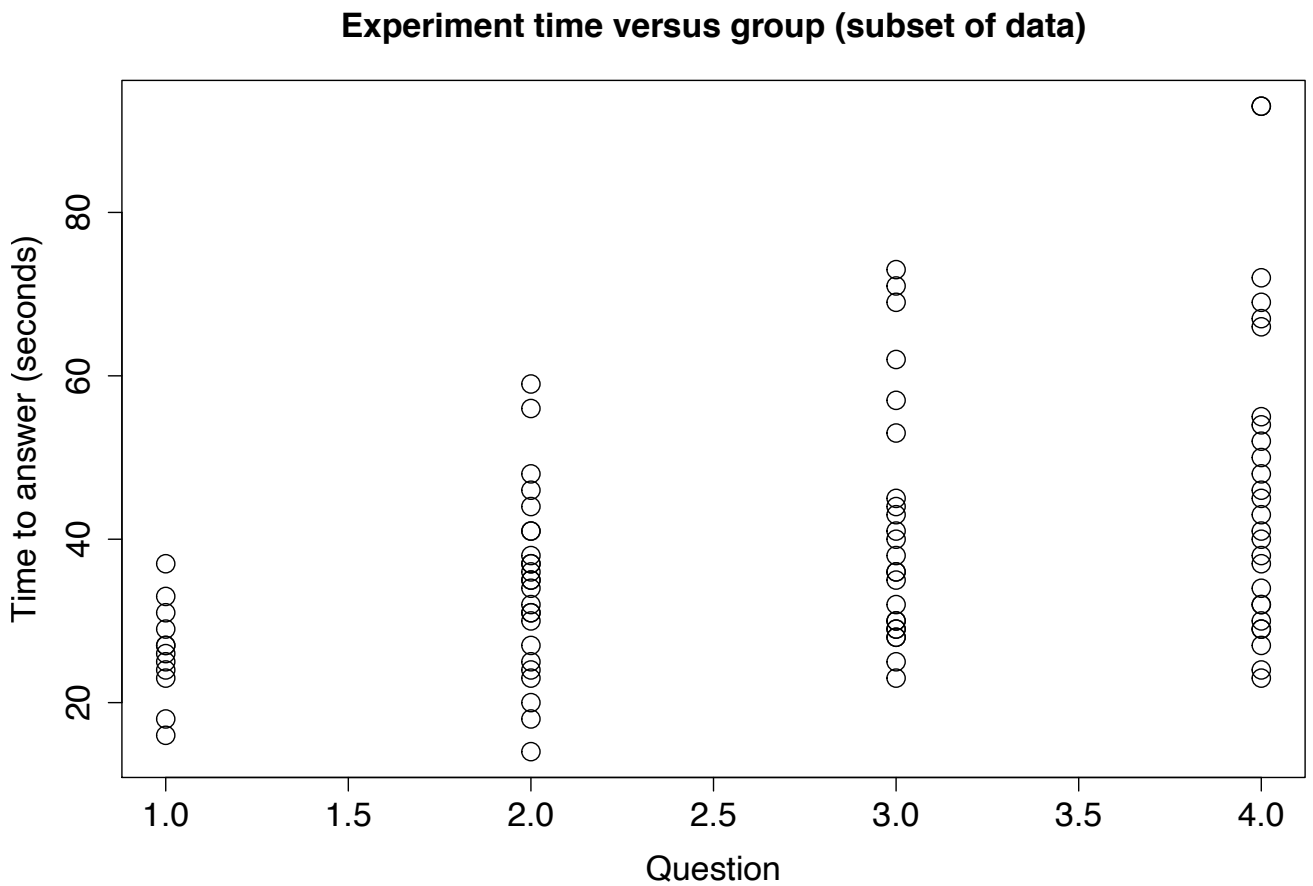
The first few rows of data:

```
##   Question Seconds Correct
## 1        1      16       1
## 2        1      18       1
## 3        1      23       1
## 4        1      24       1
## 5        1      25       1
## 6        1      26       1
```

A graph of the raw data



Let's focus in on the first four groups and omit outliers



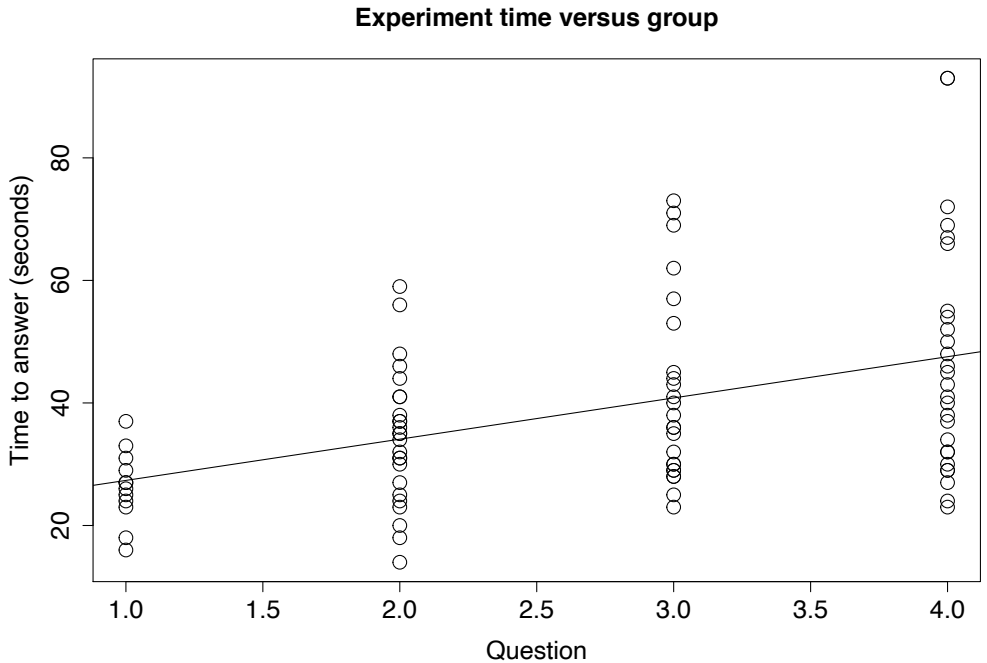
(We will come back to the full dataset during the lab session next week!)

Fit a simple linear regression model

```
lm1 <- lm(Seconds ~ Question, data = exp_subset)
summary(lm1)
```

```
##
## Call:
## lm(formula = Seconds ~ Question, data = exp_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.549 -10.617  -1.819   5.844  45.451
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   20.628      4.390   4.699 0.00000985 ***
## Question       6.730      1.494   4.505 0.00002075 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.52 on 86 degrees of freedom
## Multiple R-squared:  0.1909, Adjusted R-squared:  0.1815
## F-statistic:  20.3 on 1 and 86 DF,  p-value: 0.00002075
```

Fit a simple linear regression model

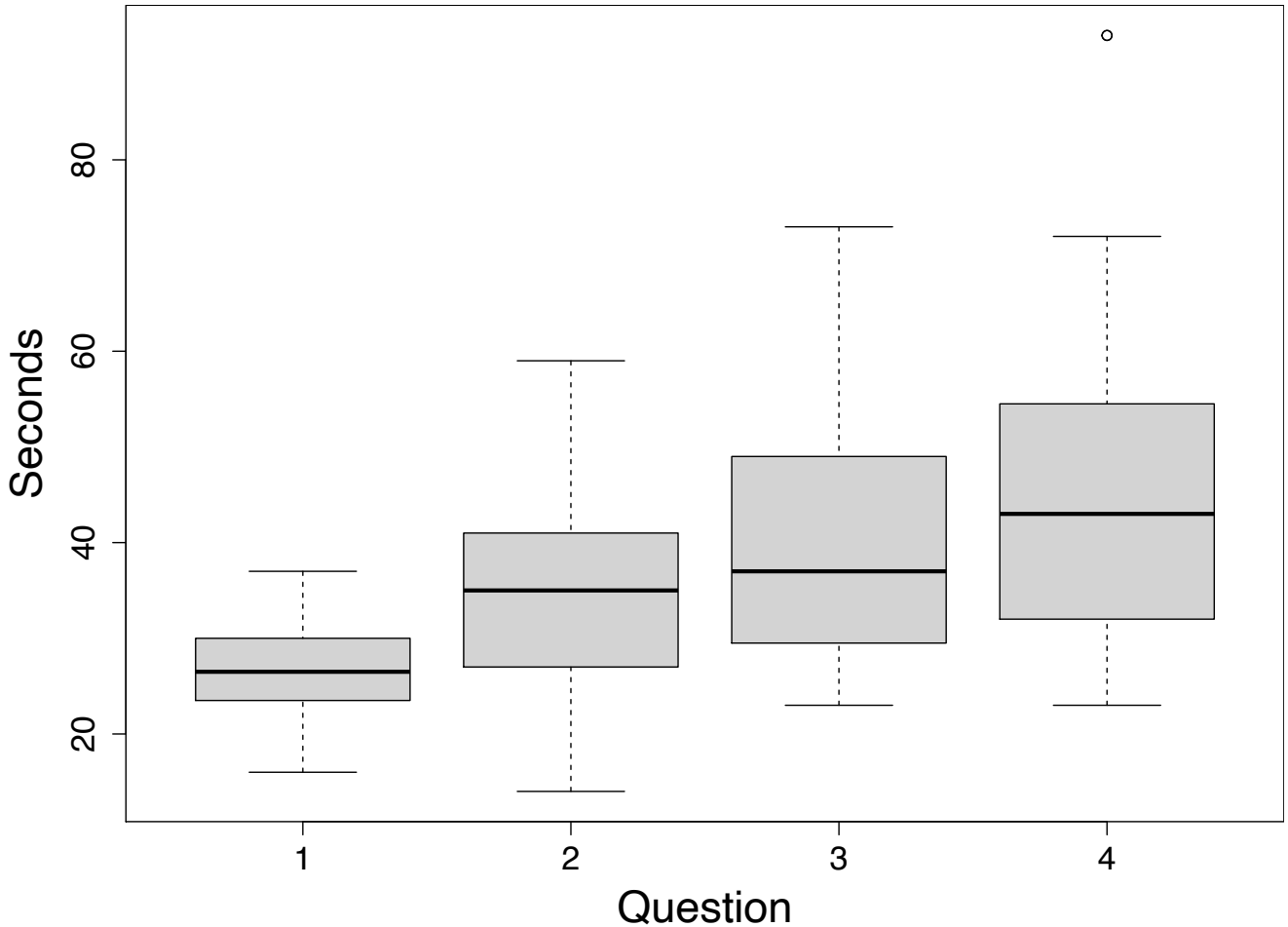


The equation of the line is: $\hat{y} = 20.63 + 6.73x$.

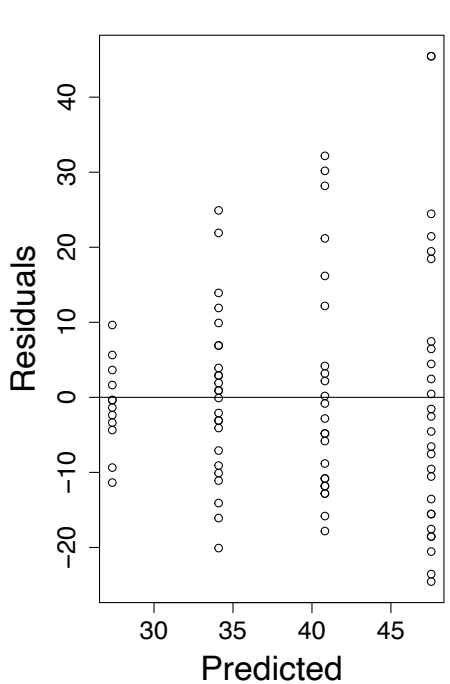
Intercept: the estimated average length of time to put the words in alphabetical order when question $= 0$ is 20.63 seconds.

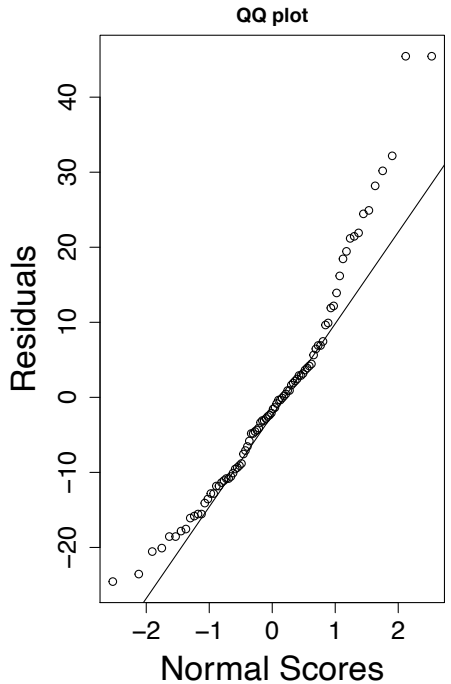
Slope: the estimated average length of time to put the words in alphabetical order increases by 6.73 seconds for each unit increase in question.

Assess the model



Assess the model (residuals)

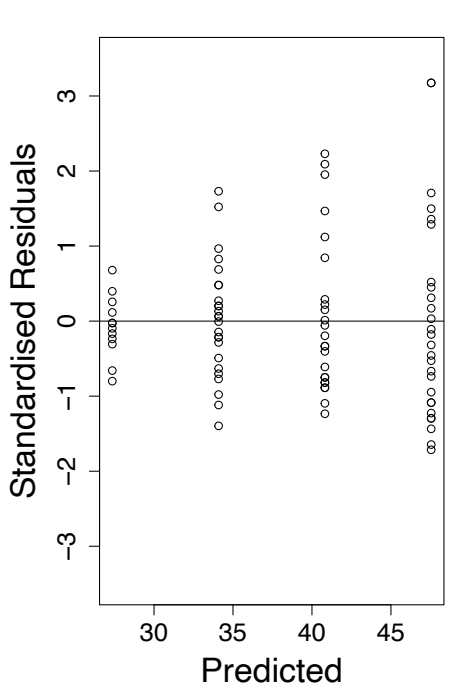


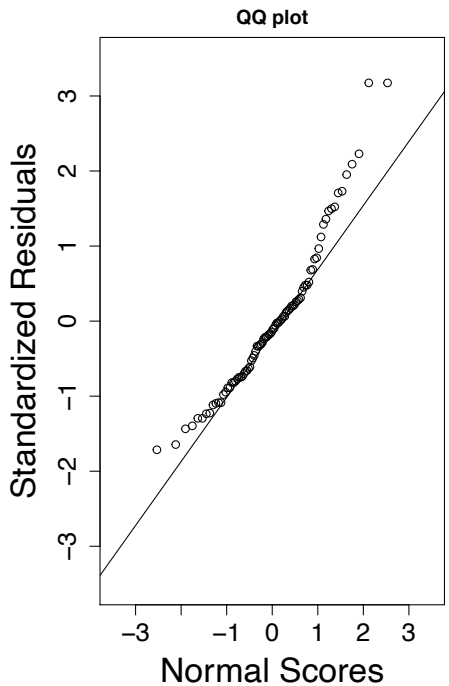


Assumptions:

1. $E[\epsilon_i] = 0$.
2. $\text{Var}(\epsilon_i) = \sigma^2$ (and does not depend on i).
3. ϵ_i are independent.
4. $\epsilon_i \sim N(0, \sigma^2)$.

Assess the model (standardised residuals)





Comparing models

It is often useful to think about the connections between different models.

The model for the mean is:

$$Y_i = \mu + \epsilon_i \qquad \text{with } \epsilon_i \sim N(0, \sigma^2) \text{ and IID.}$$

The simple linear regression model (SLR) is:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{with } \epsilon_i \sim N(0, \sigma^2) \text{ and IID.}$$

We can see that the mean model is a special case of the SLR model.

In fact, if we set $\beta_1 = 0$ in the SLR model, we get the mean model.

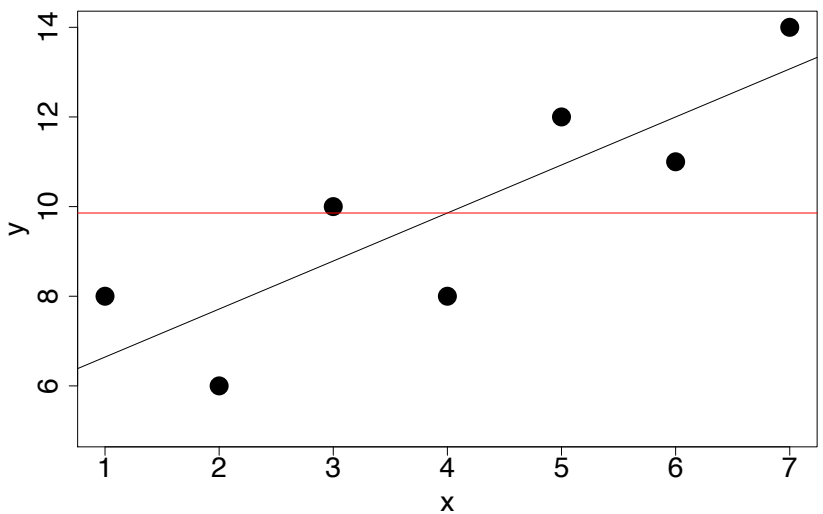
We will often be interested to test whether or not $\beta_1 = 0$. (Why?)

Comparing models visually

$$Y_i = \mu + \epsilon_i$$
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

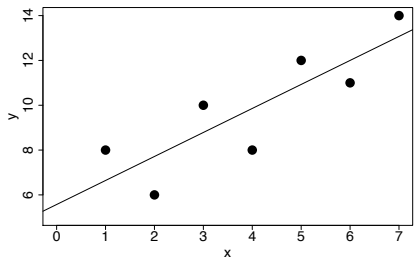
```
x <- c(1, 2, 3, 4, 5, 6, 7)
y <- c(8, 6, 10, 8, 12, 11, 14)
mean(y)
```

```
## [1] 9.857143
```



Note: $Y_i = \mu + \epsilon_i$ and $Y_i = \beta_0 + \epsilon_i$ are the same models, just different notation!

Simple linear regression in R




```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       1       2       3       4       5       6       7
##  1.3571 -1.7143  1.2143 -1.8571  1.0714 -1.0000  0.9286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5714     1.3477   4.134  0.00905 **
## x             1.0714     0.3014   3.555  0.01629 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.595 on 5 degrees of freedom
## Multiple R-squared:  0.7166, Adjusted R-squared:  0.6599
## F-statistic: 12.64 on 1 and 5 DF,  p-value: 0.01629
```

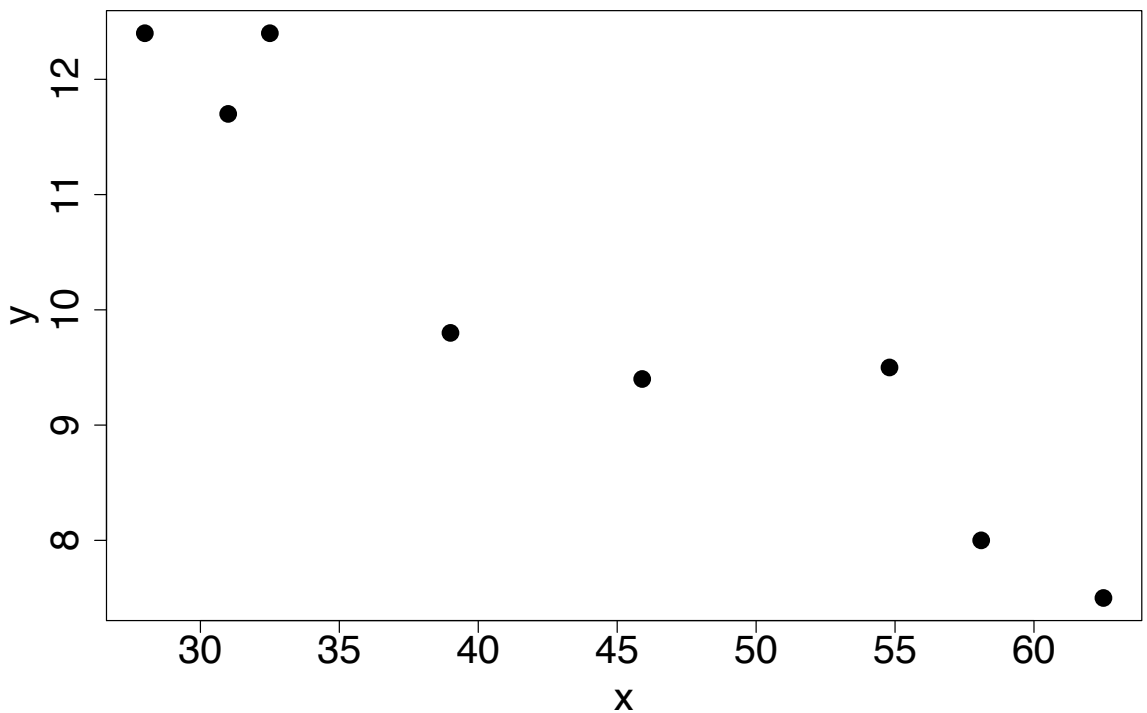
In the next sections, we will look at how to estimate the SLR parameters, and the underlying theory for conducting tests of hypothesis and condidence intervals for them.

Section 6.2: Least squares estimation

The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

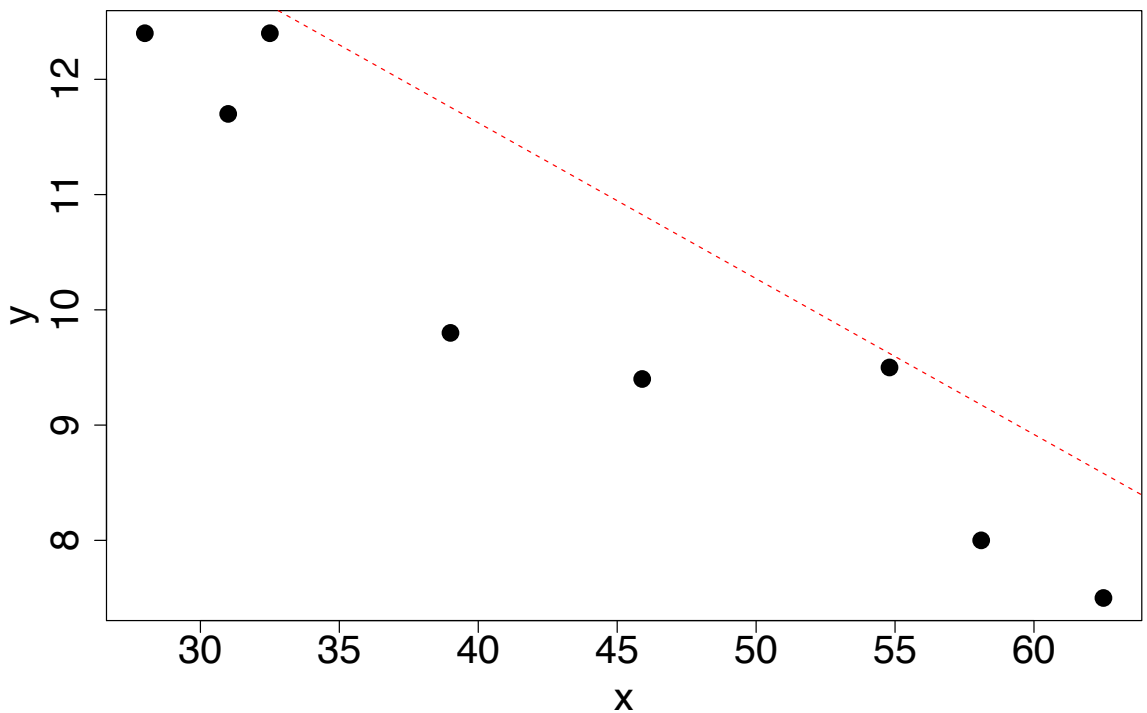
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

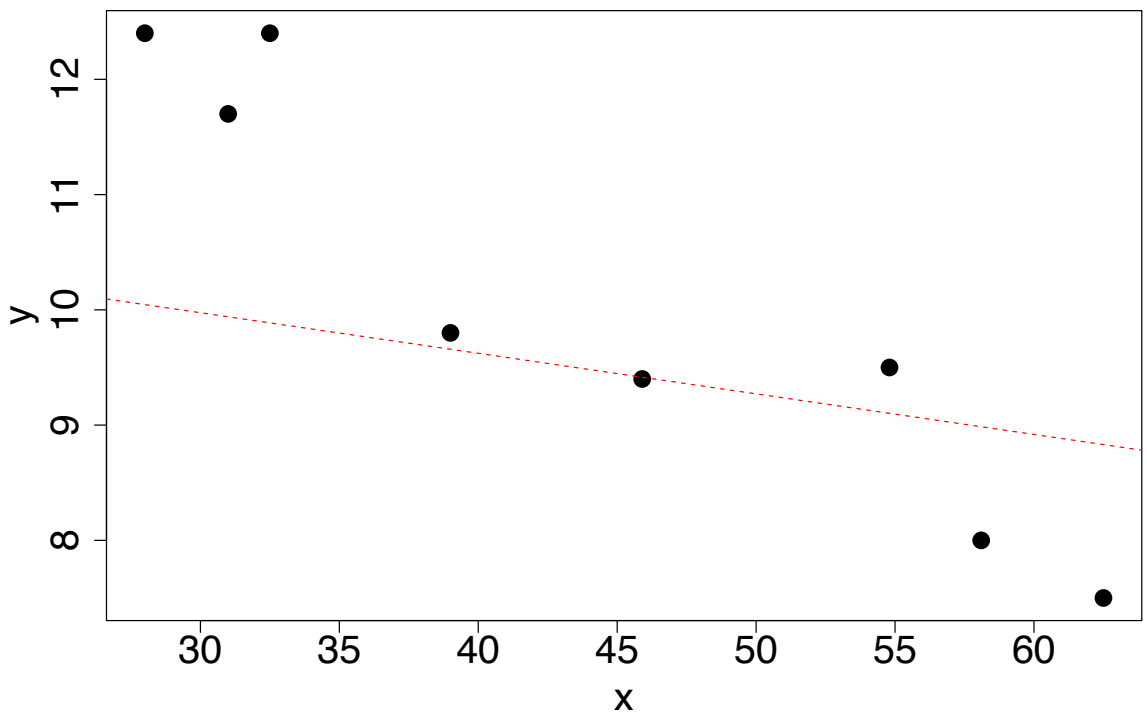
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

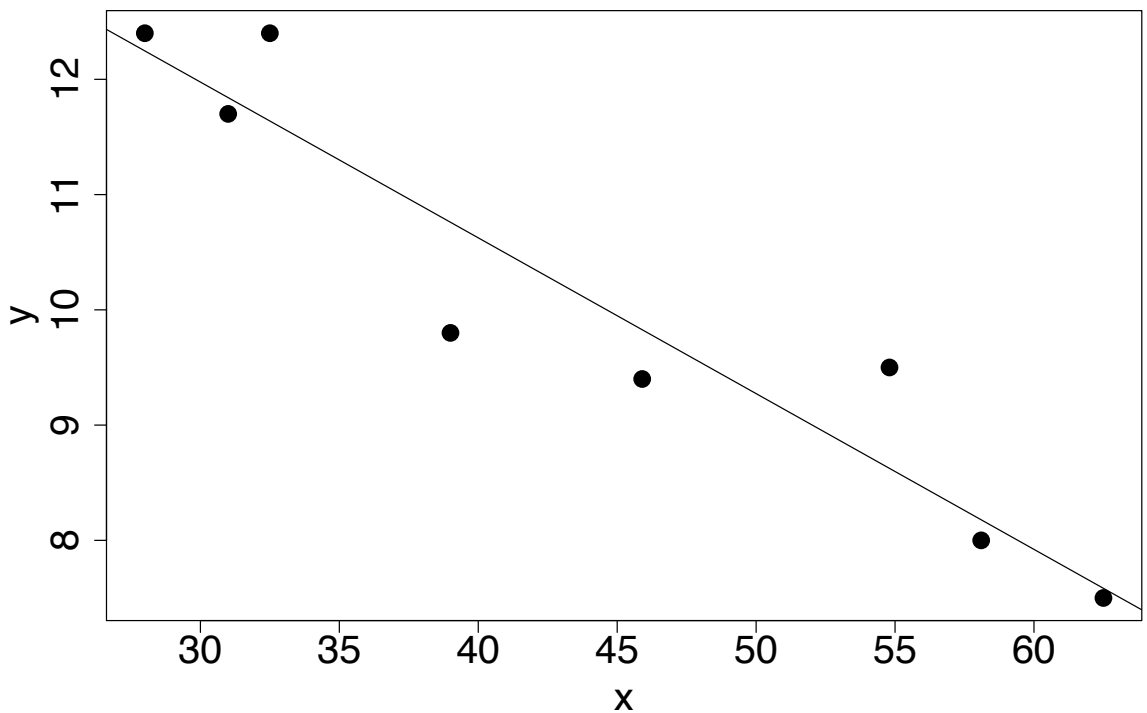
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

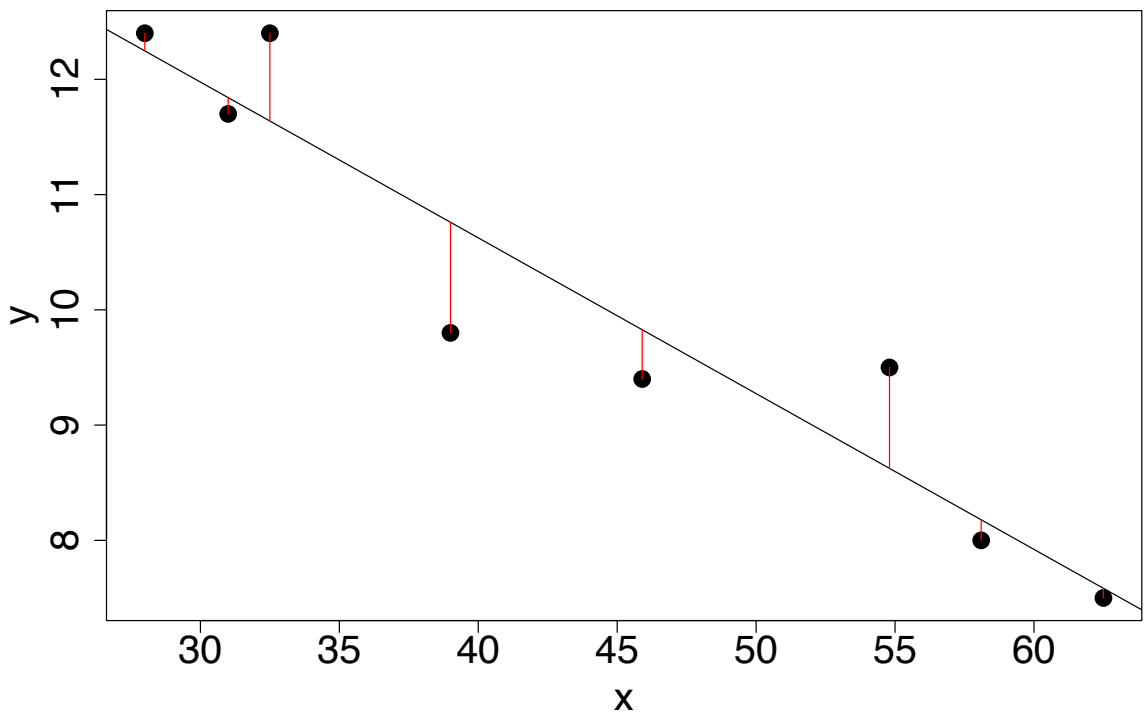
The method of ordinary least squares minimises the squared deviations from the line.



The concept behind least squares estimation

Least squares estimation can be used to fit a simple linear regression model.

The method of ordinary least squares minimises the squared deviations from the line.



Deriving the least squares estimates

Recall: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Define S, the sum of squared errors:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ must satisfy: $\frac{\delta S}{\delta \beta_0} = 0$ and $\frac{\delta S}{\delta \beta_1} = 0$.

$$\frac{\delta S}{\delta \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\delta S}{\delta \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

Deriving the least squares estimates (contd)

Taking the derivative with respect to β_0:

$$\frac{\delta S}{\delta \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

and setting it equal to 0 at $\hat{\beta}_0$ gives:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

And we can get:

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Deriving the least squares estimates (contd)

Taking the derivative with respect to β_1

$$\frac{\delta S}{\delta \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

and setting it equal to 0 at $\hat{\beta}_1$ gives:

$$\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Substituting $\hat{\beta}_0$ in (where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$):

$$\sum_{i=1}^{n} x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x})$$

Rearranging gives:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Some notation:

$$S_{xx} = \sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n} y_i(y_i - \bar{y}) = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$S_{xy} = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

So, the equation of the ordinary least squares (OLS) fitted line is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

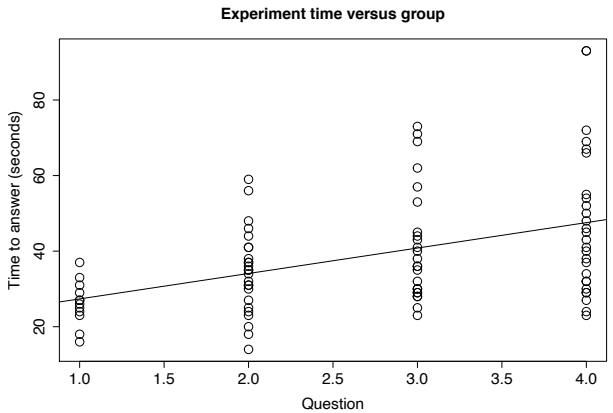
where

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Back to our class experiment data example



The equation of the line is: $\hat{y} = 20.63 + 6.73x$. Let's confirm these values.

Some summary values: $n = 88$, $\sum x_i = 242$, $\sum y_i = 3444$, $\bar{x} = 2.75$, $\bar{y} = 39.1363636$, $\sum x_i^2 = 760$, $\sum y_i^2 = 157202$, $\sum x_i y_i = 10107$.

The slope

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{636}{94.5} = 6.73$$

The intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 39.1364 - 6.73 * 2.75 = 20.63$$

Section 6.3: Parameter inference

Sampling distributions of OLS estimators

When the simple linear regression model assumptions hold, then

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

are the sampling distributions of the least squares estimators.

We can think of $\hat{\beta}_1$ and $\hat{\beta}_0$ as being random variables in the context of any data collected from the population of interest.

In this section, we will prove these sampling distributions. This will allow us to perform tests of hypotheses and construct confidence intervals for the population parameters β_0 and β_1.

Sampling distribution of $\hat{\beta}_1$: Expected value

Recall that $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, and $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ IID $N(0, \sigma^2)$. Want: $E[\hat{\beta}_1] = \beta_1$.

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n a_i Y_i$$

where the a_i values depend only on x and are NOT random. By linearity of expectation:

$$E[\hat{\beta}_1] = \sum_{i=1}^n a_i E[Y_i] = \sum_{i=1}^n a_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n a_i + \beta_1 \sum_{i=1}^n a_i x_i = \beta_1$$

Since: $$\sum_{i=1}^n a_i = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0, \text{ and}$$

$$\sum_{i=1}^n a_i x_i = \frac{\sum_{i=1}^n x_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{S_{xx}} = \frac{S_{xx}}{S_{xx}} = 1, \text{ giving } E[\hat{\beta}_1] = \beta_1 \text{ as required.}$$

Sampling distribution of $\hat{\beta}_1$: Variance

We now want to show that $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$.

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 Var(Y_i) \text{ (since } Y_i\text{s are independent)} \\
&= \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 \\
&= \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \\
&= \sigma^2 \frac{S_{xx}}{(S_{xx})^2} = \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

Finally, the normality assumption follows as $\hat{\beta}_1$ is a linear combination of normal random variables (Y_is). So we have proven that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Sampling distribution of $\hat{\beta}_0$: Expected value

Recall that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, and $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ IID $N(0, \sigma^2)$, and $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$.

Want to show that $E[\hat{\beta}_0] = \beta_0$.

$$
\begin{aligned}
E[\hat{\beta}_0] &= E[\bar{Y} - \hat{\beta}_1 \bar{x}] = E[\bar{Y}] - \beta_1 \bar{x} \\
&= \frac{1}{n} \sum_{i=1}^{n} E[Y_i] - \beta_1 \bar{x} = \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) - \beta_1 \bar{x} \\
&= \frac{1}{n} (n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i) - \beta_1 \bar{x} \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0
\end{aligned}
$$

Sampling distribution of $\hat{\beta}_0$: Variance

We now want to show that $\text{Var}(\hat{\beta}_1) = = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$.

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{Y} - \hat{\beta}_1 \bar{x}) \\ &= \text{Var}(\bar{Y}) + \bar{x}^2 \text{Var}(\hat{\beta}_1) - 2\bar{x}\text{Cov}(\bar{Y}, \hat{\beta}_1)\end{aligned}$$

Aside: $(\text{Cov}(aU + bV, cY + dZ) = ac\text{Cov}(U, Y) + bc\text{Cov}(V, Y) + ad\text{Cov}(U, Z) + bd\text{Cov}(V, Z))$.

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = \text{Cov}\left(\frac{1}{n} \sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} a_i Y_i \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{n} a_i \text{Cov}(Y_i, Y_j) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \text{Cov}(Y_i, Y_j)$$

$$= \frac{1}{n} \sum_{i=1}^{n} a_i \text{Cov}(Y_i, Y_i) \text{ (since } Y_i \text{ are indep.)}$$

$$= \frac{\sigma^2}{n} \sum_{i=1}^{n} a_i = \frac{\sigma^2}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 0$$

Sampling distribution of $\hat{\beta}_0$: Variance (contd)

Back to:

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{Y} - \hat{\beta}_1\bar{x}) \\ &= \text{Var}(\bar{Y}) + \bar{x}^2\text{Var}(\hat{\beta}_1) - 2\bar{x}\text{Cov}(\bar{Y}, \hat{\beta}_1)\end{aligned}$$

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{Y}) + \bar{x}^2\text{Var}(\hat{\beta}_1) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(Y_i) + \bar{x}^2\frac{\sigma^2}{S_{xx}} \\ &= \frac{1}{n^2}n\sigma^2 + \bar{x}^2\frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\end{aligned}$$

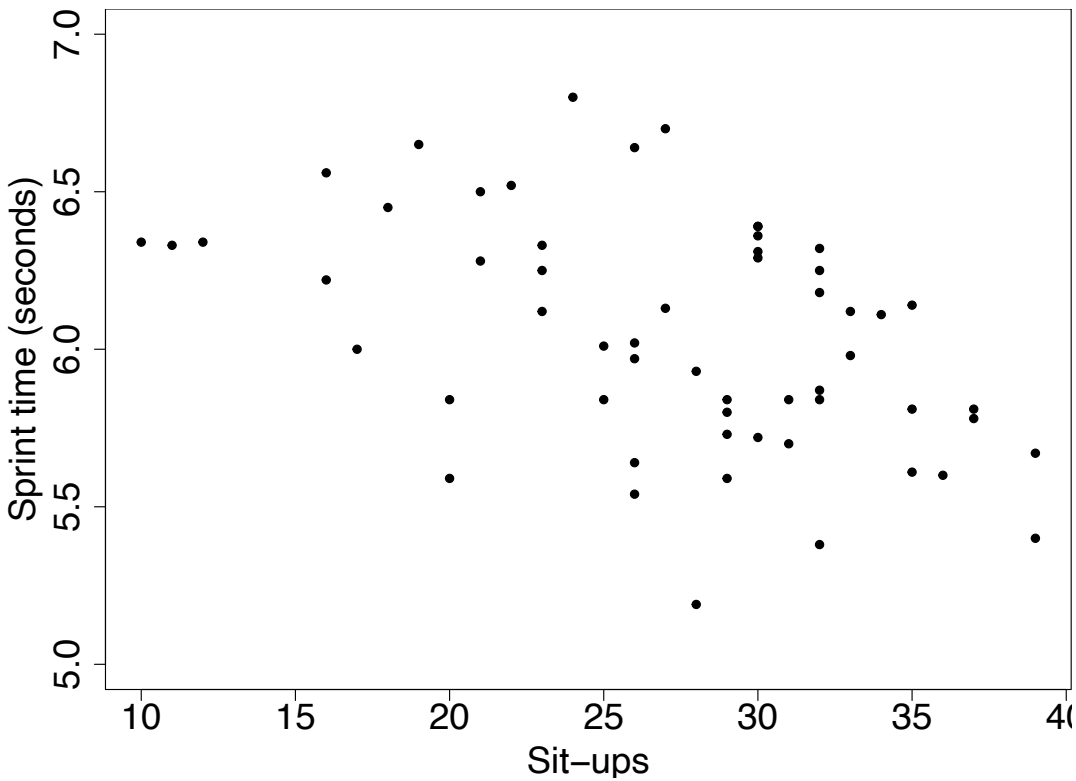
Finally, the normality assumption follows as $\hat{\beta}_0$ is a linear combination of normal random variables (Y_i values and $\hat{\beta}_1$). So we have proven that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

Athlete Example

Is there a relationship between how many sit-ups you can do and how fast you can sprint 40 yards? A study set up to examine this recorded details on 57 randomly selected female athletes.

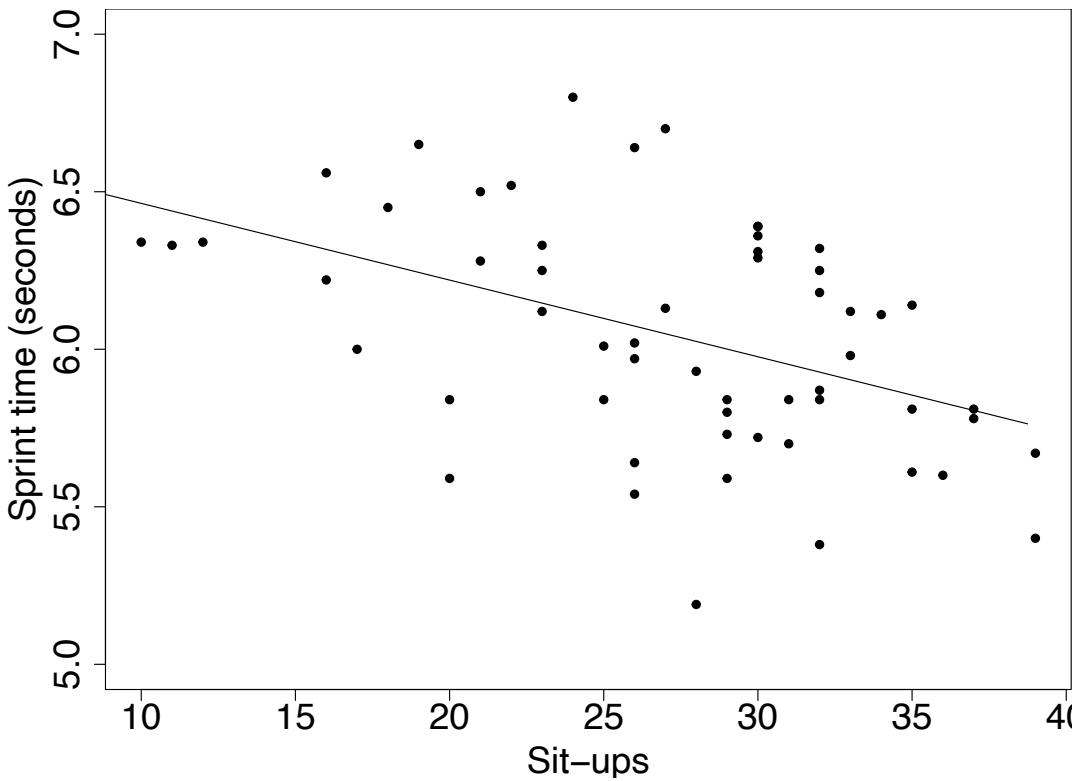
Here is a scatter plot of the data:



Athlete Example

Is there a relationship between how many sit-ups you can do and how fast you can sprint 40 yards? A study set up to examine this recorded details on 57 randomly selected female athletes.

Here is a scatter plot of the data with the linear regression line fitted:



Athlete Example

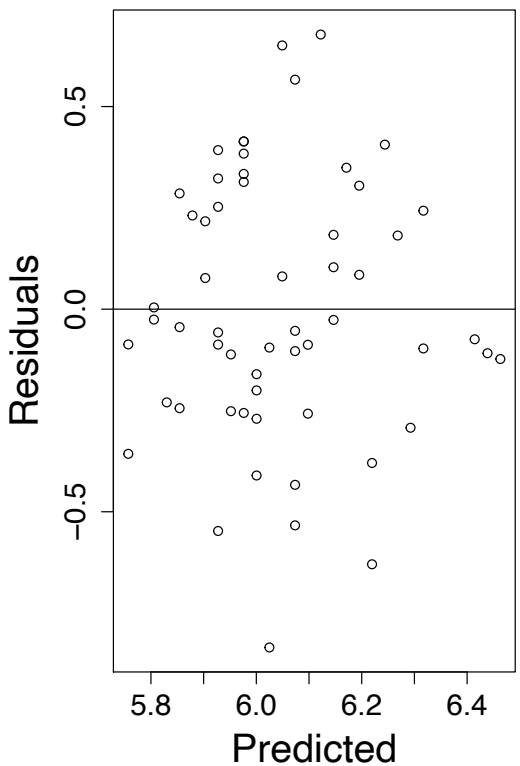
Let's fit the simple linear regression model and find confidence intervals for the parameters.

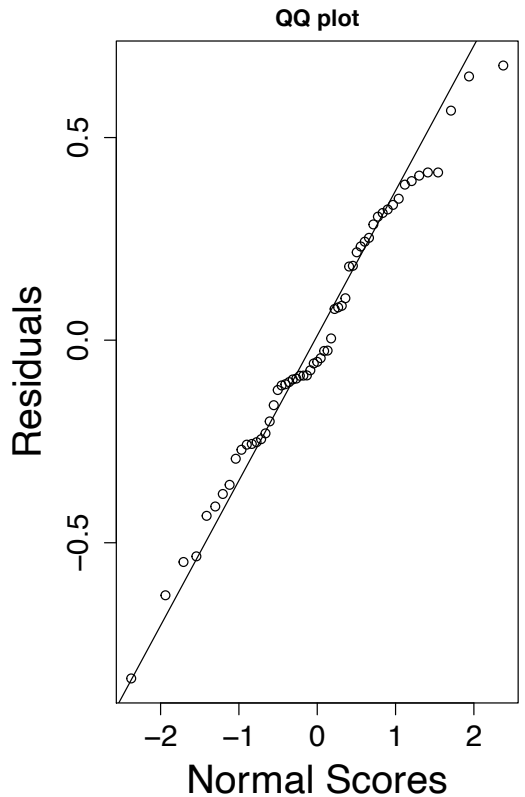
```
##
## Call:
## lm(formula = Sprint ~ Situp, data = SS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83484 -0.23007 -0.05353  0.25255  0.67778
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  6.706527   0.177891  37.700 < 0.0000000000000002 ***
## Situp       -0.024346   0.006349  -3.835             0.000326 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3272 on 55 degrees of freedom
## Multiple R-squared:  0.211,  Adjusted R-squared:  0.1966
## F-statistic: 14.71 on 1 and 55 DF,  p-value: 0.0003257
```

```
##               2.5 %  97.5 %
## (Intercept)  6.3500  7.0630
## Situp       -0.0371 -0.0116
```

Athlete Example

Testing the model assumptions:





Confidence interval for β_1

What theory allows us to construct a confidence interval for β_1?

We know that $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$, or equivalently $\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$.

And, when we replace σ by $\hat{\sigma}$ we have:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}.$$

The degrees of freedom (df) are $n-2$ because we have estimated two parameters (β_0 and β_1).

We estimate σ^2 by the mean squared error (MSE) by calculating:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

A $(1-\alpha)\times 100\%$ confidence interval for β_1 is:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \times \sqrt{\frac{MSE}{S_{xx}}}$$

Hypothesis tests for β_1

We may wish to test

$H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$.

The null hypothesis here is that $E[Y] = \beta_0$, i.e., $E[Y]$ is not linearly related to x.

Under H_0 (i.e., if the null hypothesis is true):

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}.$$

Using a specified α level, we reject H_0 for extreme values of t_{obs}.

We could also test $H_0 : \beta_1 = b$ by computing:

$$\frac{\hat{\beta}_1 - b}{\text{S.E.}(\hat{\beta}_1)}.$$

where $b \neq 0$.

Back to the athlete example

Here are the model estimates and the confidence intervals:

```
##                Estimate  Std. Error
## (Intercept)  6.70652722 0.177891008
## Situp       -0.02434606 0.006348777

##                2.5 %  97.5 %
## (Intercept)  6.3500  7.0630
## Situp       -0.0371 -0.0116
```

The 95% confidence interval for β_1 is

$$\begin{aligned}\hat{\beta}_1 \pm t_{n-2,\alpha/2} \times \sqrt{\frac{MSE}{S_{xx}}} \\ = -0.02434606 \pm 2.0040448 \times 0.006348777 \\ = (-0.0371, -0.0116)\end{aligned}$$

The athlete example

Here are the model estimates and hypothesis tests:

```
##                Estimate  Std. Error   t value
## (Intercept)  6.70652722 0.177891008 37.700204
## Situp       -0.02434606 0.006348777 -3.834764
##                                                      Pr(>|t|)
## (Intercept) 0.00000000000000000000000000000000000000005563697
## Situp       0.00032571993194159000982457197181929586804471910
```

Now we'll look at the hypothesis test for β_1, using $\alpha = 0.05$.

$H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$.

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{-0.02434606}{0.006348777} = -3.835$$

If the H_0 is true, then, t_{obs} is a random draw from a $t(55)$ distribution.

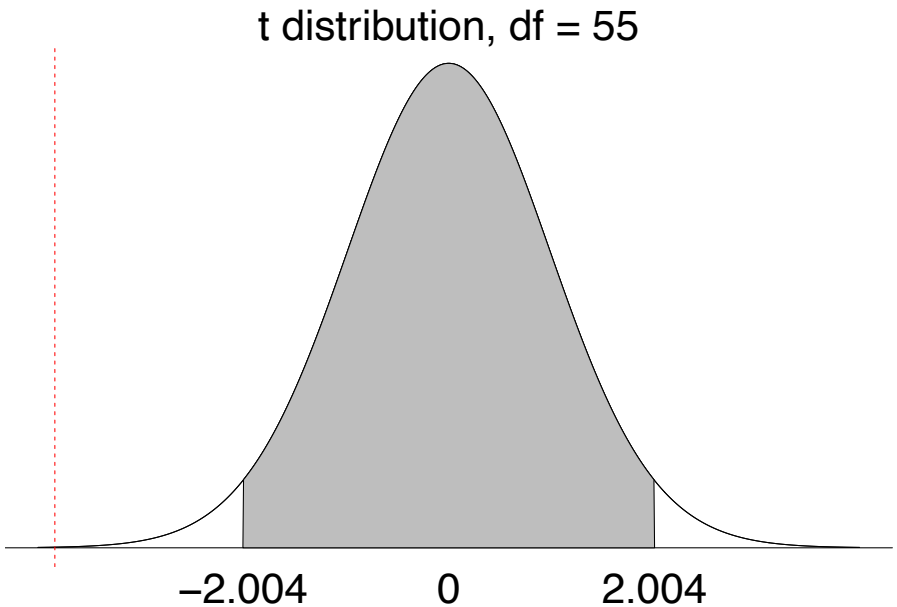
The critical values are -2.004 and 2.004. We reject H_0 if our observed test statistic is lower than -2.004, or greater than 2.004.

Let's take a look at this graphically...

The athlete example

Evaluate the hypothesis test using critical values.

The test statistic = -3.835, and is shown by the red line:

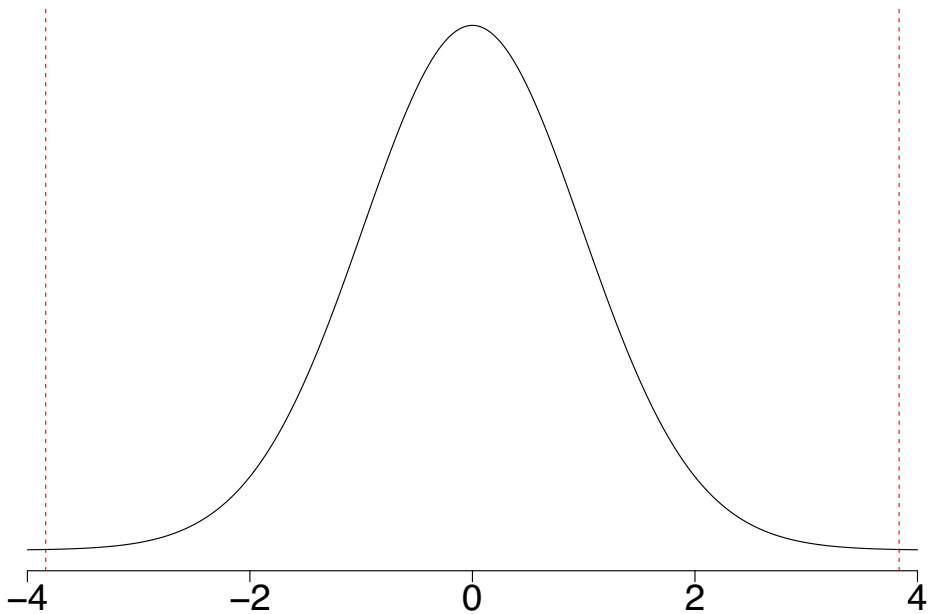


We reject the H_0, using $\alpha = 0.05$, and conclude that $\beta_1 \neq 0$. We have evidence that the true mean sprinting time for 40 yards is linearly related to the number of sit-ups that female athletes can do.

The athlete example

We can also evaluate the hypothesis test using the p-value. We find the probability of observing a test statistic as extreme, or more extreme than what we observed.

Here, p-value $= P(T(55) \geq |t_{obs}|) = 0.0003$.



As before, we reject the H_0, using $\alpha = 0.05$.

Confidence interval for β_0

What theory allows us to construct a confidence interval for β_0?

We know that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

We use this to construct a 95% C.I. for β_0 as:

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} \times \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Hypothesis for β_0

If we wish to test for a particular value of β_0, we can perform a hypothesis test:

$H_0 : \beta_0 = 0$ vs. $H_A : \beta_0 \neq 0$

The null hypothesis here is that $E[y] = \beta_1 x$, i.e., the line passes through the origin.

Under the H_0, the test statistic

$$t_{obs} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t(n-2)$$

where $\hat{\sigma}^2$ is estimated by the MSE, as before.

Using a specified α level, we reject H_0 for extreme values of t_{obs}.

Back to the athlete example

Here are the model estimates, hypothesis tests and confidence intervals:

```
##                Estimate  Std. Error   t value
## (Intercept)  6.70652722 0.177891008 37.700204
## Situp       -0.02434606 0.006348777 -3.834764
##                                                      Pr(>|t|)
## (Intercept) 0.00000000000000000000000000000000000000005563697
## Situp       0.00032571993194159000982457197181929586804471910

##                2.5 %  97.5 %
## (Intercept)  6.3500  7.0630
## Situp       -0.0371 -0.0116
```

The 95% confidence interval for $\beta_0 = (6.350, 7.063)$.

For the hypothesis test, the test statistic $= 37.7$, with p-value < 0.0001. Using $\alpha = 0.05$, we reject the null hypothesis and conclude that $\beta_0 \neq 0$.

In practice, are we interested in the confidence interval and hypothesis test for the intercept in this example?

Section 6.4: Confidence intervals and prediction intervals

Using slr models: Confidence intervals and prediction intervals

Once we have estimated a simple linear regression model, there are two different ways to think about using it:

- estimating the mean of Y for a given x value,
- predicting Y for a new observation with a given x value.

These two scenarios are conceptually different; they yield the same point estimate or prediction, however, their standard errors are different. We can construct a confidence interval or a prediction interval using the respective standard errors.

For the athlete data with $x_0 = 30$, we get $\hat{y} = 5.976$, with:

Confidence interval

```
##        fit      lwr      upr
## 1 5.976145 5.882149 6.070141
```

Prediction interval

```
##        fit      lwr      upr
## 1 5.976145 5.313703 6.638588
```

Let's take a look at how to construct these.

Confidence intervals

Suppose we want to estimate $\mu = E[y]$ at a particular value of x.

At x_0 let:

$$\mu_0 = E[y_0] = \beta_0 + \beta_1 x_0$$

We can estimate μ_0 by:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

For the athlete example, let $x_0 = 30$, then

$$\hat{y}_0 = 6.7065 - 0.02435 \times 30 = 5.976$$

A 95% confidence interval for this estimate is:

$$\hat{y}_0 \pm t_{n-2,\alpha/2}\ \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 5.976 \pm 2.004 \times 0.3272 \times \sqrt{\frac{1}{57} + \frac{(30 - 27.17544)^2}{2656.246}}$$

$$= (5.88, 6.07)$$

Prediction intervals

Now, let's assume that we have a new female athlete that can do 30 sit-ups. What time would we predict for their sprint and how certain would we be of the prediction?

The point estimate prediction is:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 6.7065 - 0.02435 \times 30 = 5.976$$

This is the same as before.

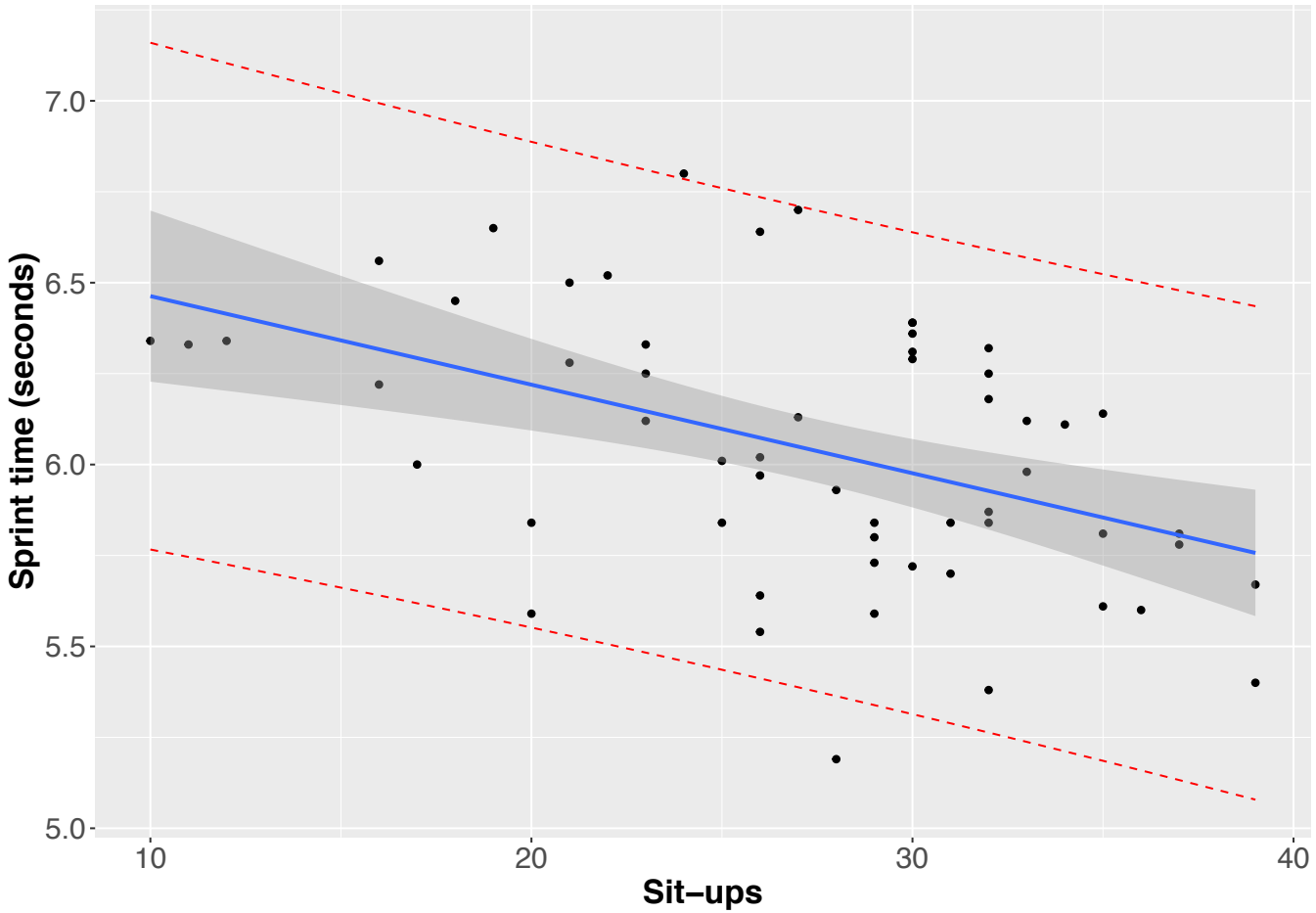
A 95% prediction interval is:

$$\hat{y}_0 \pm t_{n-2,\alpha/2}\ \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$= 5.976 \pm 2.004 \times 0.3272 \times \sqrt{1 + \frac{1}{57} + \frac{(30 - 27.17544)^2}{2656.246}}$$

$$= (5.31, 6.64)$$

Confidence intervals and prediction intervals graphically



The blue line shows the fitted regression line, the shaded grey region shows the confidence bounds and the dotted red lines show the prediction bounds.

Section 6.5: Matrix notation

Using matrix notation

We can express the simple linear regression model in matrix notation.

The regular form of the SLR model is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim$ IID $N(0, \sigma^2)$.

This can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

where $\mathbf{Y}$ is the $n \times 1$ response vector,

$\mathbf{X}$ is an $n \times 2$ design matrix,

β is a 2×1 parameter vector,

and ϵ is the $n \times 1$ error vector

Using matrix notation contd.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

The variance of $\mathbf{Y}$ can be expressed as

$$\text{Var}(\mathbf{Y}) = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn}^2 \end{bmatrix}$$

Where the diagonal values are the variances, and off-diagonals are the covariances. But remember, for simple linear regression model we assume that the Y_i values are independent and have constant variance, giving

$$\text{Var}(\mathbf{Y}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \text{Var}(\epsilon)$$

Least squares estimation in matrix notation

The least squares estimate of the unknown parameter vector β is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

assuming that $\mathbf{X}^T\mathbf{X}$ is invertible.

The least squares estimate of β is unbiased, i.e., $E[\hat{\beta}] = \beta$, and the variance is given by $\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

The predicted $\mathbf{Y}$ values are given by

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as the hat matrix.

The corresponding vector of residuals is

$$\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$$

The simple linear regression model in matrix notation

The $\mathbf{X}$ matrix is:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$\begin{aligned}(\mathbf{X}^T\mathbf{X})^{-1} &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \\ &= \frac{1}{n(\sum x_i^2 - n\bar{x}^2)}\begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \\ &= \frac{1}{S_{xx}}\begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\end{aligned}$$

The simple linear regression model in matrix notation contd.

Then to find the least squares estimates:

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum x_i y_i \end{bmatrix}$$

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \frac{1}{S_{xx}} \begin{bmatrix} \bar{y}\sum x_i^2 - \bar{x}\sum x_i y_i \\ -n\bar{x}\bar{y} + \sum x_i y_i \end{bmatrix} \end{aligned}$$

With some algebra, this gives:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

as before.

The simple linear regression model in matrix notation contd.

The variance of $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

$$= \frac{\sigma^2}{S_{xx}} \begin{bmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

which gives

$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{S_{xx}} \sum x_i^2/n = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$

$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$, as before, and

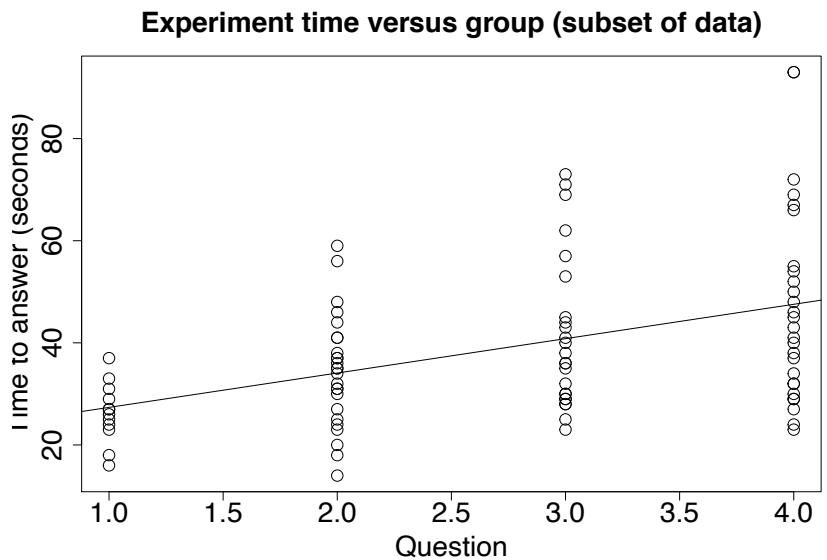
$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\frac{\sigma^2}{S_{xx}}$.

Section 6.6: Multiple regression

Multiple regression

Up until now, we have considered the simple linear regression model, with only one x predictor. We will now extend that so that the Y_i can depend on a number of possible independent variables $x_{i1}, x_{i2}, ..., x_{ik}$.

Let's think back to our in-class experiment. We looked at a simple linear regression model using question group as a predictor.



What other variables could we have recorded or considered?

The multiple regression model

A multiple regression model takes the form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \qquad \text{for } i = 1 \text{ to } n$$

and we can express this using matrix notation as

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

with dimensions $(n \times 1), (n \times p), (p \times 1)$ and $(n \times 1)$, where $p = k + 1$ is the number of model parameters.

β is a vector of unknown parameters to be estimated from observed data. β_j is the change in the mean value of Y per unit change in x_j, assuming all other independent variables are held constant. Consequently, the β_j depend on which x's are included in the model.

Model assumptions

For the model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

we have

$$E[\mathbf{Y}] = \mathbf{X}\beta, \quad E[\epsilon] = 0, \quad \text{Var}(\mathbf{Y}) = \text{Var}(\epsilon) = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

The assumptions are:

1. Linearity: $E[\epsilon] = \mathbf{0}$, hence $E[\mathbf{Y}] = \mathbf{X}\beta$.
2. Constant variance and 0 covariances: $\text{Var}(\epsilon) = \sigma^2 I_n$ and $\text{Var}(\mathbf{Y}) = \sigma^2 I_n$.
3. Multivariate normal (MVN) distribution: $\epsilon \sim N_n(\mathbf{0}, \sigma^2 I_n)$ and $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 I_n)$

Sampling distribution of $\hat{\beta}$

Let $\hat{\beta}$ be the OLS estimator of β.

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

As before,

$$\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is known as the hat matrix.

Also:

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

When the model assumptions hold:

$$\hat{\beta} \sim N_p(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

and

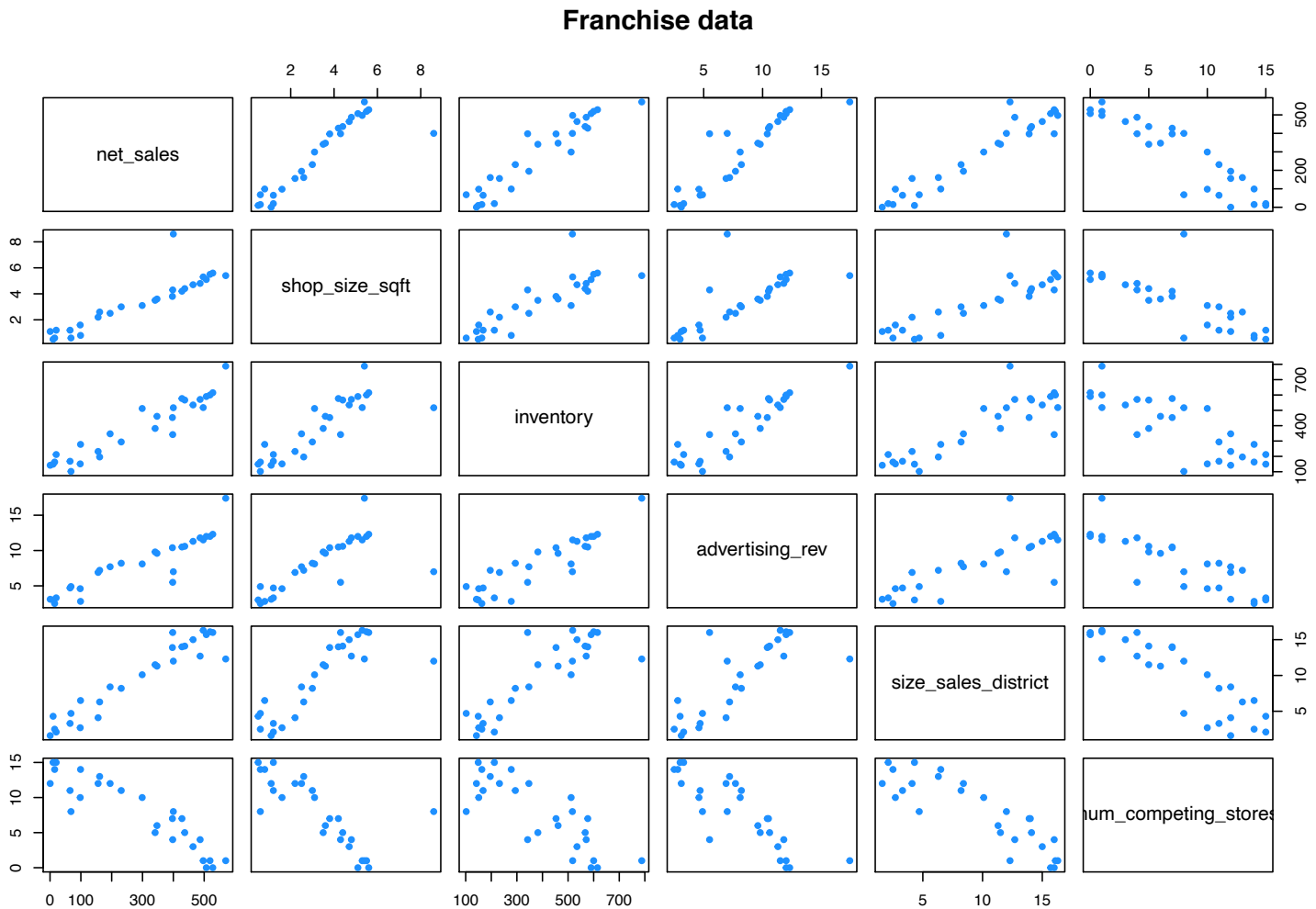
$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$$

where c_{jj} is the jj entry of $(\mathbf{X}^T\mathbf{X})^{-1}$ for $j = 0, ..., k$.

Franchise data example

- Data on 27 "All Greens" franchise branches
- There are multiple variables recorded on each branch
  - net sales (in thousands)
  - shop size, inventory (in thousands)
  - advertising revenue (in thousands)
  - size of sales district (thousands of families)
  - the number of competing stores in the sales district.
- Try to predict net sales
  - We could look at how net sales varies with each of the possible predictors.
  - Predictors could have an effect in tandem! For example advertising revenue and size of sales district could have different predictive power if considered together.
  - Which of the predictors are most important?

Franchise data matrix plot



Fitted model in R

```
##
## Call:
## lm(formula = net_sales ~ shop_size_sqft + inventory + advertising_rev +
##     size_sales_district + num_competing_stores, data = greens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.338  -9.699  -4.496   4.040  41.139
##
## Coefficients:
##                       Estimate Std. Error t value    Pr(>|t|)
## (Intercept)          -18.85941   30.15023  -0.626    0.538372
## shop_size_sqft        16.20157    3.54444   4.571    0.000166 ***
## inventory              0.17464    0.05761   3.032    0.006347 **
## advertising_rev       11.52627    2.53210   4.552    0.000174 ***
## size_sales_district   13.58031    1.77046   7.671 0.000000161 ***
## num_competing_stores  -5.31097    1.70543  -3.114    0.005249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.65 on 21 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9916
## F-statistic: 611.6 on 5 and 21 DF,  p-value: < 0.00000000000000022
```

Model interpretation

- Shop site: $\hat{\beta}_1 = 16.202$
- Inventory: $\hat{\beta}_2 = 0.175$
- Advertising revenue: $\hat{\beta}_3 = 11.526$
- Size sales district: $\hat{\beta}_4 = 13.580$
- Number competing stores: $\hat{\beta}_5 = $ -5.311

How should we interpret these estimates? The intercept?

Advertising revenue:

For every extra \$1,000 spent on advertising, it is estimated that average sales increase by \$11,526, holding all other predictors constant.

Competing shop numbers:

For every extra competing shop in the district, it is estimated that the average sales decrease by \$5,311, keeping all other predictors constant.

Confidence interval for model parameters

To construct a 100(1 - α)% confidence interval for β_j, we use:

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= \hat{\beta}_j \pm t_{n-p,\alpha/2}\sqrt{\mathrm{MSE}\ c_{jj}}$$

where

$$\mathrm{MSE} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \sum_{j=i}^{k}\hat{\beta}_j x_{ij})^2$$

$$c_{jj} = \text{the } jj^{th} \text{ diagonal element of } (\mathbf{X}^T\mathbf{X})^{-1}$$

The degrees of freedom are equal to $n - p$ because we have estimated $p = k + 1$ parameters, where k is the number of predictors in the multiple regression model.

Confidence interval example for the franchise data

For the advertising revenue parameter for franchise data, the 95% confidence interval is

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= \hat{\beta}_3 \pm t_{27-6,0.05/2}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= \hat{\beta}_3 \pm t_{21,0.025}\ \mathrm{SE}(\hat{\beta}_j)$$

$$= 11.52627 \pm 2.080 \times 2.53210 = (6.260, 16.793)$$

We are 95% confident that the true mean increase in average sales for a \$1000 increase in advertising expenditure lies between \$6,260 and \$16,793.

Hypothesis tests for model parameters

We can test the hypothesis:

H_0: $\beta_j = b$ vs H_0: $\beta_j \neq b$.

The test statistic is:

$$T_{obs} = \frac{\hat{\beta}_j - b}{\text{SE}(\hat{\beta}_j)}$$

Under the null hypothesis (that is, assuming that the H_0 is true), the test statistic is a random draw from a $t(n-p)$ distribution.

We evaluate the test by deciding if the observed test statistic is extreme, relative to the null hypothesis distribution.

Hypothesis test example for the franchise data

Advertising revenue. β_3 is the expected change in the mean sales (in thousands) for a \$1,000 increase in advertising revenue, holder other predictors constant.

H_0: $\beta_3 = 0$ vs H_0: $\beta_3 \neq 0$.

The test statistic is:

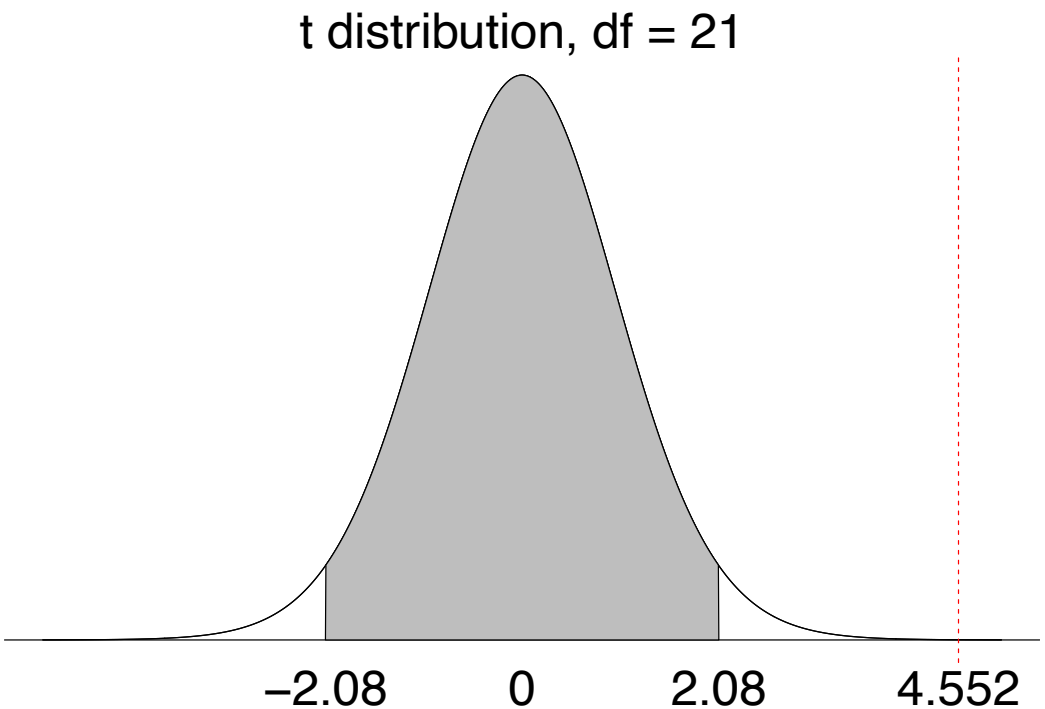
$$T_{obs} = \frac{\hat{\beta}_j - b}{\text{SE}(\hat{\beta}_j)} = \frac{11.52627 - 0}{2.53210} = 4.552$$

Using $\alpha = 0.05$, the critical values are $\pm t_{21,0.025} = \pm 2.080$. We reject H_0 if the observed test statistic is less than -2.080 or greater than 2.080.

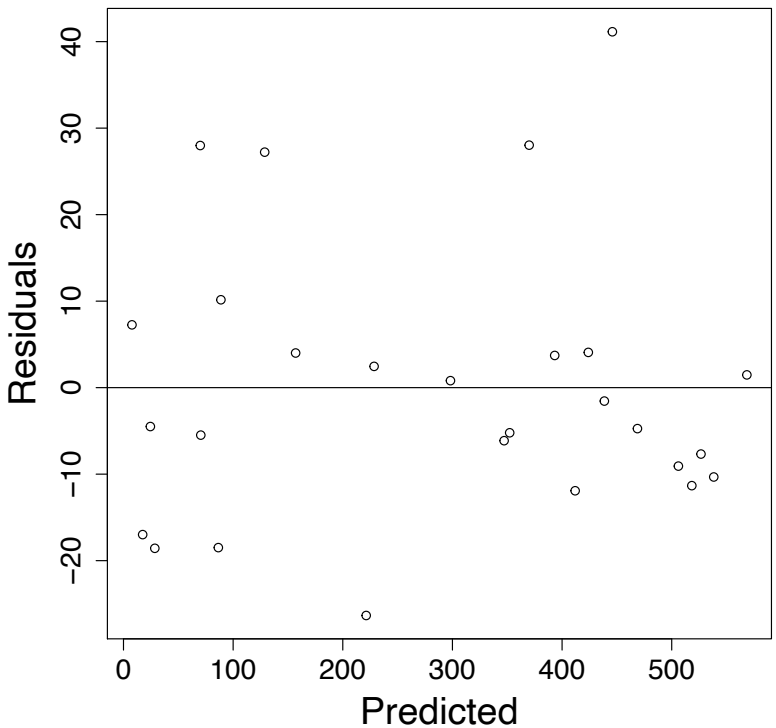
In this case, we reject the H_0 and conclude that $\beta_3 \neq 0$, since $4.552 > 2.080$.

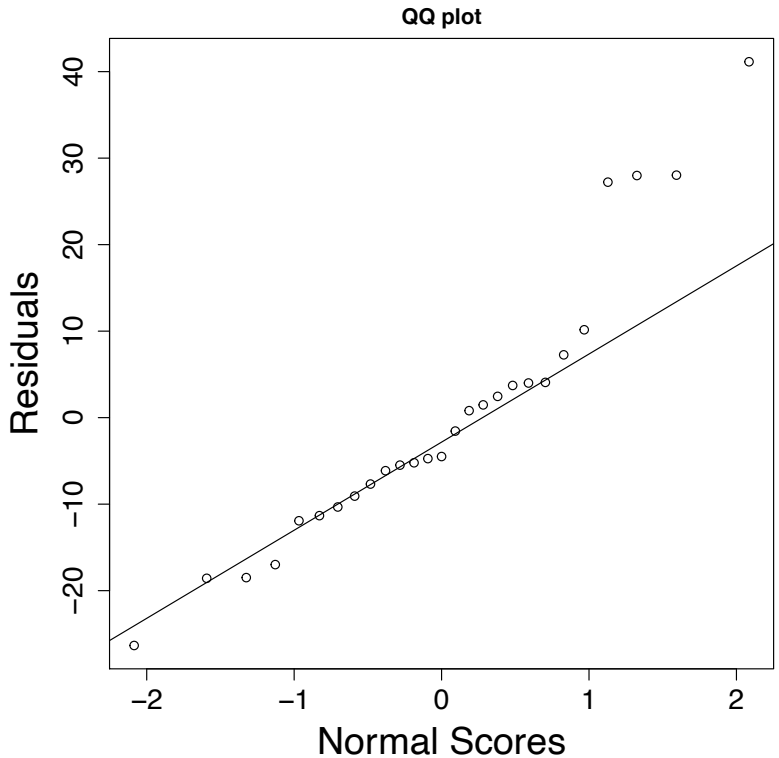
We could also evaluate the test from the p-value in the R output, the p-value = 0.00174, since the p-value < 0.05, we reject the H_0 and conclude that $\beta_3 \neq 0$.

Assessing the hypothesis test visually



Model assumptions - residual plots for the franchise data





Further considerations

Next steps in the analysis of the franchise data

- Further diagnostic tests
- Multicollinearity (test using variance inflation factors - VIF)
- Interactions among the predictors?

Multiple regression in general

- Powerful and widely used tool.
- For valid inference:
  - must have an awareness of how the model should be correctly interpreted.
  - assumptions must be validated.
  - data source must be reliable and fit for purpose.

Applied Probability II

Section 7: Estimation

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 7: Estimation

Section 7.1: Recap of standard probability laws

The distribution game!

What is the distribution of the random variable in each of the following?

- Suppose you are playing a game of Ludo and so you need to roll a 6 to get your first counter out of home and onto the board. What is the probability that you will get onto the board in between 4 and 6 goes?

- A recent study has shown that 30% of women in a certain large population suffer from anemia (iron deficiency). A random sample of 8 women is taken from the population and tested. What is the probability that three or more of the women in the sample are anemic?

- A basketball player scores a basket from the free throw line with probability 0.45. You start observing a training session on free throw attempts at a random point. What is the probability that her third basket occurs on the sixth shot?

- Suppose that items in a vending machine get stuck on the way out at random with probability 0.03. You arrive at the vending machine, put you money in and select your item. What is the probability you end up not getting your item (and are really annoyed!)?

- Consider a lotto draw with 45 balls where 6 balls are chosen at random. What is the probability of matching 4 balls?

Bernoulli

Consider an experiment with two outcomes, success and failure. Such an experiment is known as a Bernoulli trial.

Let X be a Bernoulli random variable.

$$X \sim \text{Bernoulli}(p)$$

X can take values: 0, 1, where 1 is defined as a 'success'.

Parameter: p, where $0 < p < 1$ and is the probability of success in a trial.

Probability mass function (pmf): $P(X = x) = p^x(1-p)^{1-x}$

Mean: $\text{E}[X] = p$

Variance: $\text{Var}[X] = p(1-p)$

Example

Suppose that items in a vending machine get stuck on the way out at random with probability 0.03. You arrive at the vending machine, put you money in and select your item. What is the probability you end up not getting your item (and are really annoyed!)?

Binomial

Bernoulli random variables provide the building blocks for defining other discrete random variables. An experiment which consists of n repeated independent Bernoulli trials, each with probability of success p, is called a **binomial** experiment.

Let X be the total number of successes in a binomial experiment with n trials. Then X is a binomial random variable with parameters n and p.

$$X \sim \text{Binomial}(n, p)$$

X can take values: 0, 1, 2,. . . ,n.

Parameter: $n = 1, 2, 3...$, and p, where $0 < p < 1$ and is the probability of success in an individual trial.

Probability mass function (pmf): $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Mean: $\text{E}[X] = np$

Variance: $\text{Var}[X] = np(1-p)$.

Example

A recent study has shown that 30% of women in a certain large population suffer from anemia (iron deficiency). A random sample of 8 women is taken from the population and tested. What is the probability that three or more of the women in the sample are anemic?

Geometric

Suppose independent Bernoulli trials with success probability p are performed until a success occurs. Let X be the number of trials required. Then

$$X \sim \text{Geometric}(p)$$

X can take values: $1, 2, 3, \ldots$.

Parameter: p, where $0 < p < 1$ and is the probability of success in an individual trial.

Probability mass function (pmf): $P(X = x) = (1 - p)^{x-1} p$

Mean: $\mathrm{E}[X] = \frac{1}{p}$.

Variance: $\mathrm{Var}[X] = \frac{(1-p)}{p^2}$.

Example

Suppose you are playing a game of Ludo and so you need to roll a 6 to get your first counter out of home and onto the board. What is the probability that you will get onto the board in between 4 and 6 goes?

Negative binomial

The negative binomial distribution is a generalisation of the geometric distribution.

Suppose independent Bernoulli trials with success probability p are performed until r successes occurs. Let X be the number of trials required.

Then $X \sim$ Negative binomial (r, p).

The probability mass function for X is:

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

for $k = r, r+1, r+2, \ldots$

Example

A basketball player scores a basket from the free throw line with probability 0.45. You start observing a training session on free throw attempts at a random point. What is the probability that her third basket occurs on the sixth shot?

Hypergeometric

Suppose a box contains N items, A of which are of type A and $N - A$ of which are of type B. A sample of size n is taken without replacement.

Let X denote the number of type A items drawn.

Then X is a hypergeometric random variable with parameters N, A and n.

The probability mass function is

$$p(k) = P(X = k) = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}}$$

for $k = 0, 1, 2, \ldots, \min(A, n)$.

Example

Consider a lotto draw with 45 balls where 6 balls are chosen at random. What is the probability of matching 4 balls?

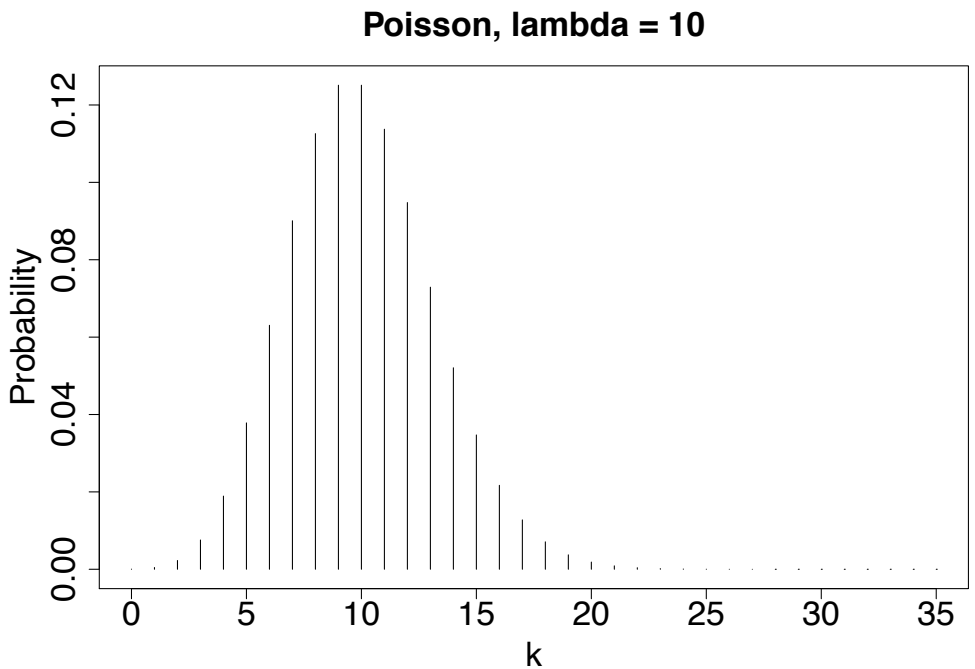
Poisson

The Poisson is a 'counting' distribution. A random variable X has a Poisson distribution with parameter λ if

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

for $k = 0, 1, 2, \ldots$ and $\lambda > 0$.

We say that $X \sim \text{Poisson}(\lambda)$. $P(X = k)$ is the probability mass function.

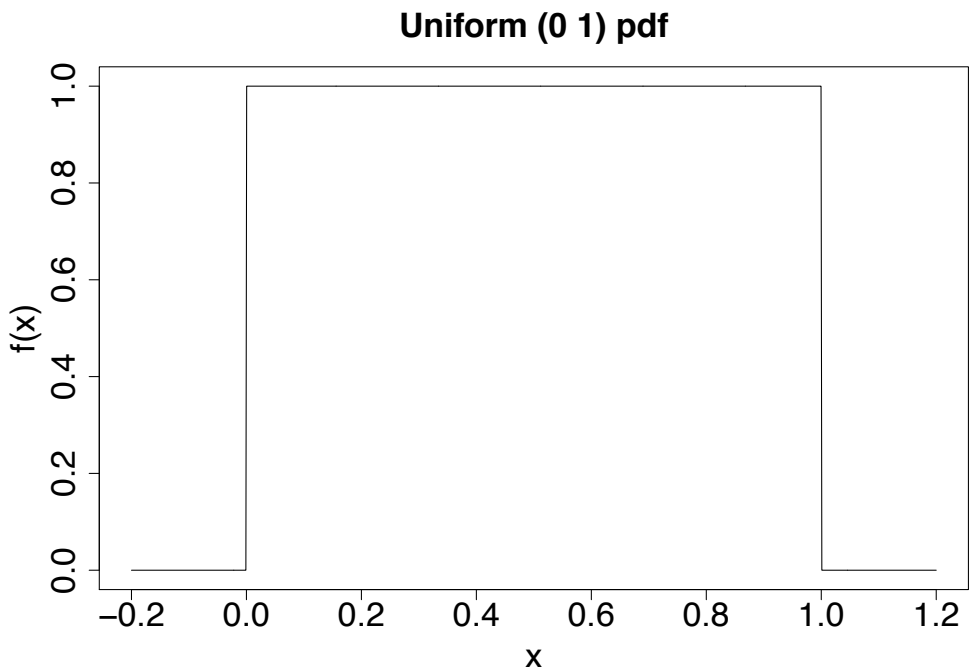


Uniform

X is said to have a uniform distribution on the interval (a, b) if

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$$

$$F(x) = P(X \leq x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$$



Exponential

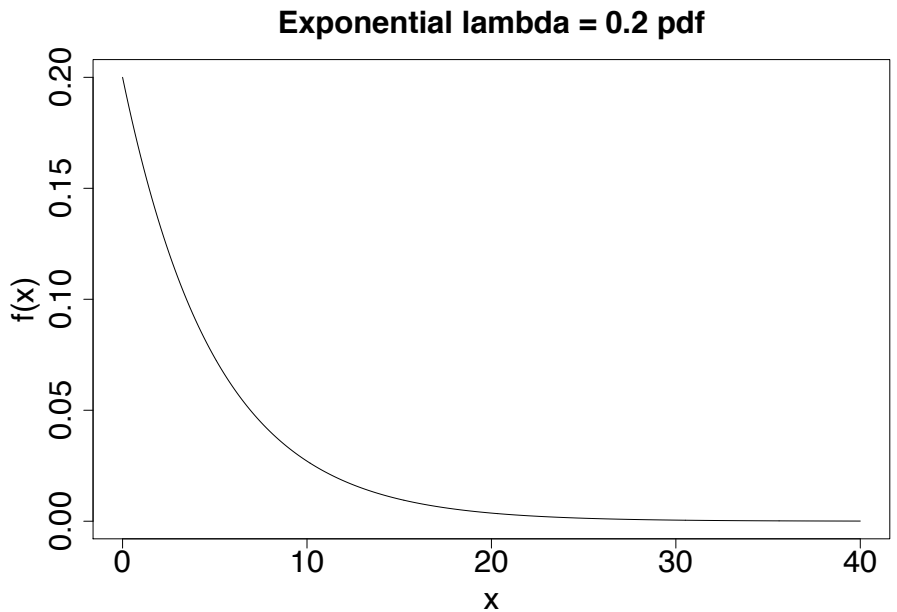
A continuous random variable whose probability density function (pdf) is given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

with $\lambda > 0$, is said to have an exponential distribution.

The exponential distribution is often used to model the time until a specific event occurs. The cumulative distribution function (cdf) for an exponential random variable is:

$$F(t) = P(X \leq t) = \int_0^t \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^t = 1 - e^{-\lambda t}, t \geq 0.$$

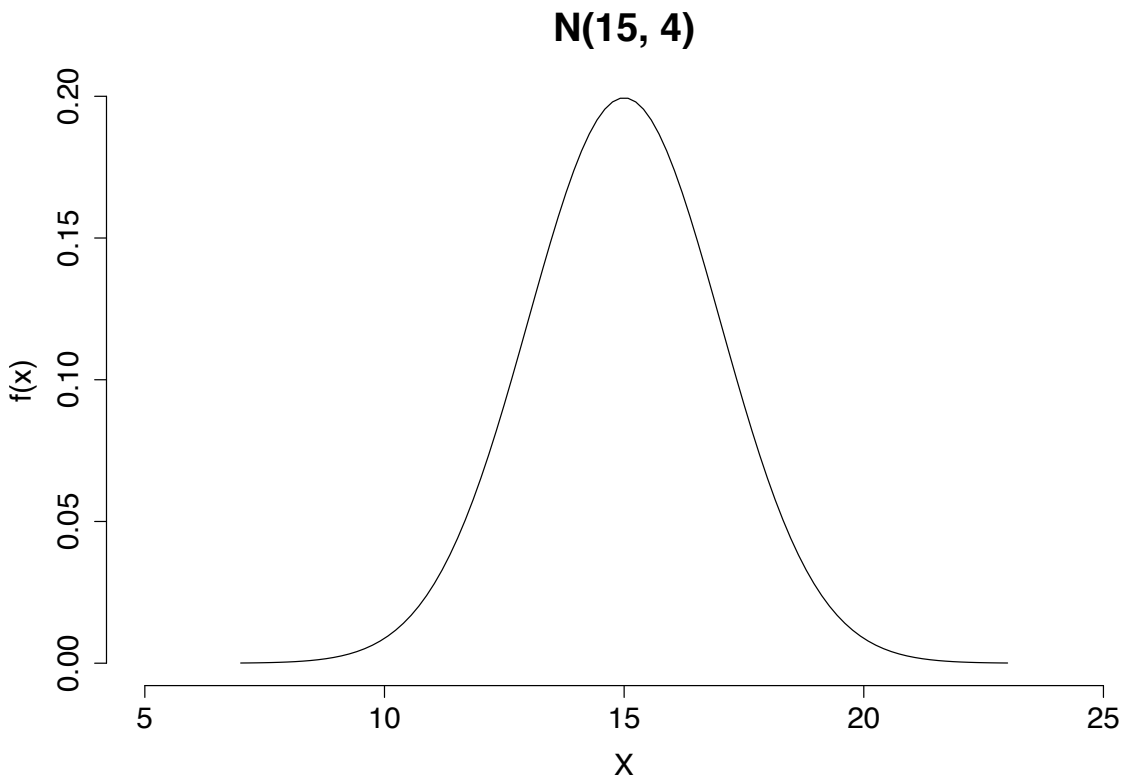


Normal

A random variable X is said to have a normal (or Gaussian) distribution with parameters μ and σ^2 if it has the probability density function (pdf):

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

For example, consider $X \sim \mathrm{N}(\mu = 15, \sigma^2 = 4)$.



Section 7.2: Likelihood

Binomial example

Think about tossing a coin 10 times, but where we know nothing about the fairness of the coin. The outcome from each toss is a Bernoulli distributed random variable with parameter p.

The number of heads we see is: $X \sim \text{Binomial}(n = 10, p)$.

Suppose that we observe $x = 8$ heads.

We don't know what the value of p is, but what can we say about p from our own experiment?

- Information about p is not complete, so there is uncertainty.
- p cannot be zero and is unlikely to be small as then $\text{P}(X = 8)$ would be tiny.
- Likely values for p are 0.7, 0.8 or 0.9, because:
  - if $p = 0.7$, then $\text{P}(X = 8) = 0.2335$, and
  - if $p = 0.8$, then $\text{P}(X = 8) = 0.3020$, and
  - if $p = 0.9$, then $\text{P}(X = 8) = 0.1937$.

A way to compare candidate values of p is to compare the observed probabilities across different values for p. We can do this formally via a likelihood function.

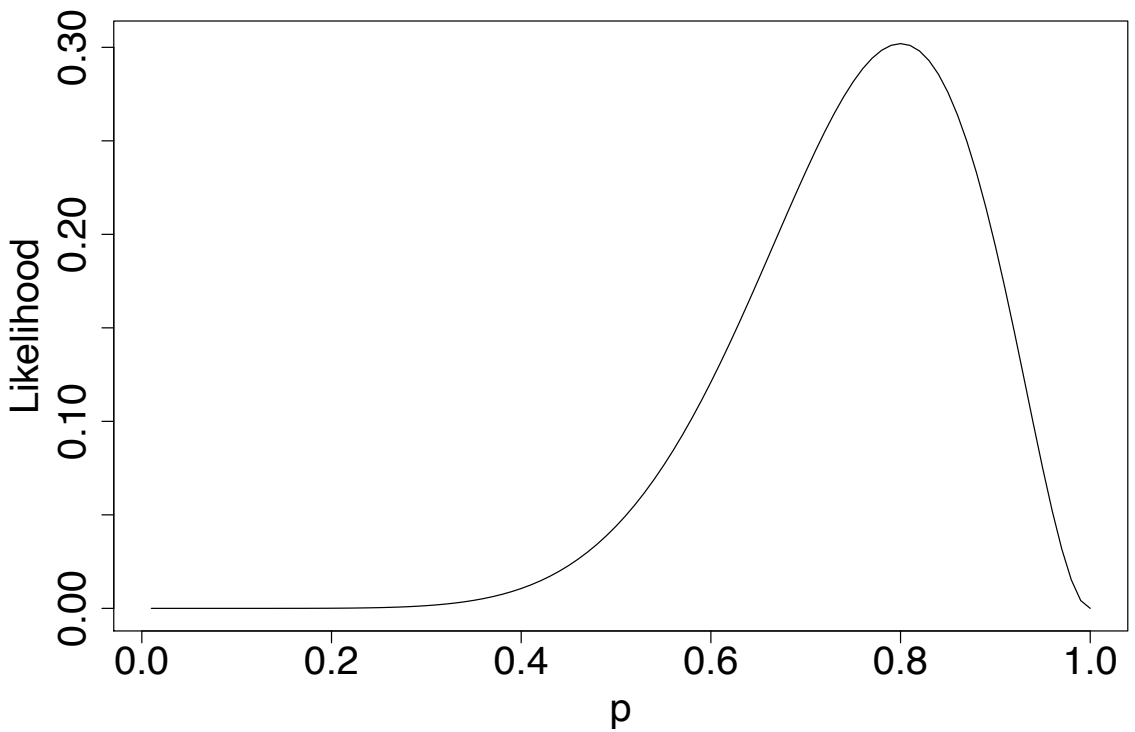
Binomial example contd.

In our experiment, we observe $x = 8$ heads in 10 trials.

The likelihood function for p is

$$L(p) = P(X = 8 \mid p) = \binom{10}{8} p^8 (1-p)^2$$

A plot of $L(p)$ versus p is:



Definition of likelihood

Definition: Assuming a statistical model is parametrised by a fixed and unknown parameter θ, the likelihood is the probability of the observed data x considered as a function of θ.

Back to the coin experiment

Suppose we repeat the coin experiment three times:

First run, observe $x_1 = 8$, second run: $x_2 = 6$, third run: $x_3 = 7$.

Now the likelihood is the probability of observing $\mathbf{x} = (x_1, x_2, x_3)$.

Assuming the experiments are independent, the probability of observing all three is

$$P(X = 8)P(X = 6)P(X = 7)$$

where, $X \sim \text{Binomial}(n = 10, p)$.

The likelihood function is

$$L(p) = P(X = x_1)P(X = x_2)P(X = x_3) = \prod_{i=1}^{3} P(X = x_i)$$

We know that $P(X = x) = \binom{10}{x} p^x (1-p)^{10-x}$. So,

$$L(p) = \prod_{i=1}^{3} P(X = x_i) = \prod_{i=1}^{3} \binom{10}{x_i} p^{x_i} (1-p)^{10-x_i}$$

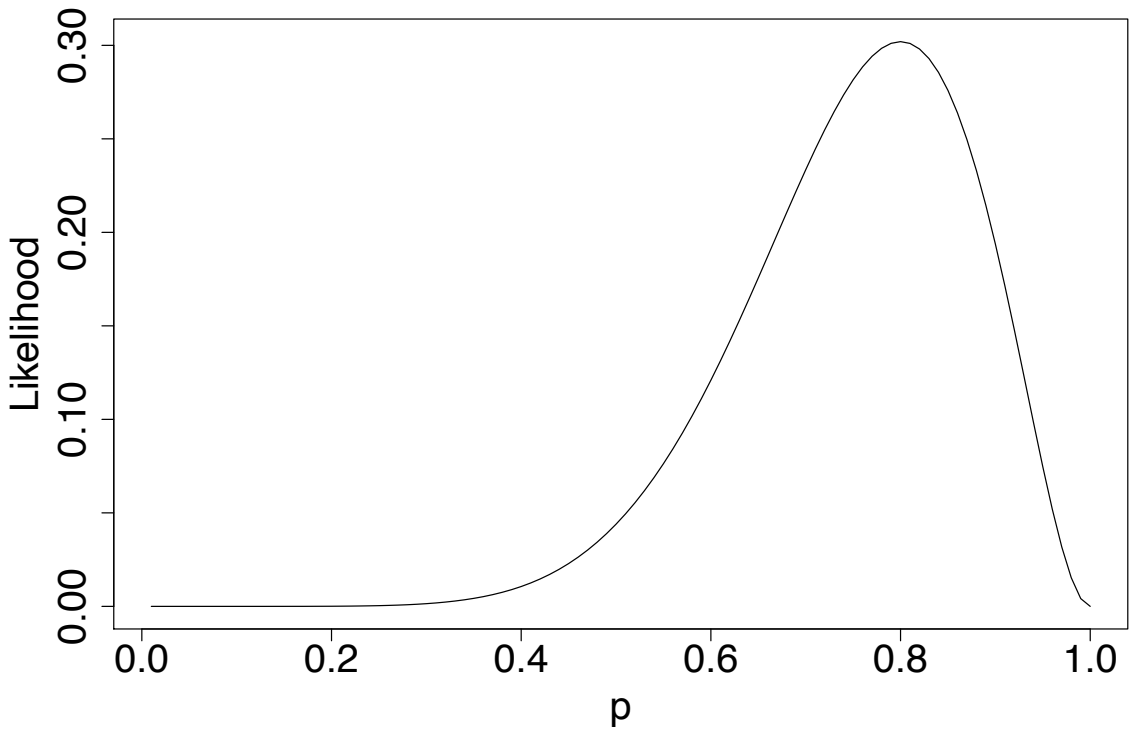
$$= \left[p^{\sum_{i=1}^{3} x_i} \right] \left[(1-p)^{30-\sum_{i=1}^{3} x_i} \right] \left[\prod_{i=1}^{3} \binom{10}{x_i} \right]$$

Section 7.3: Maximum likelihood estimation

How does maximum likelihood (ML) estimation work?

If we have an unknown parameter θ in a statistical model, the maximimum likelihood estimate (MLE) $\hat{\theta}$ is the value of θ which maximises $L(\theta)$.

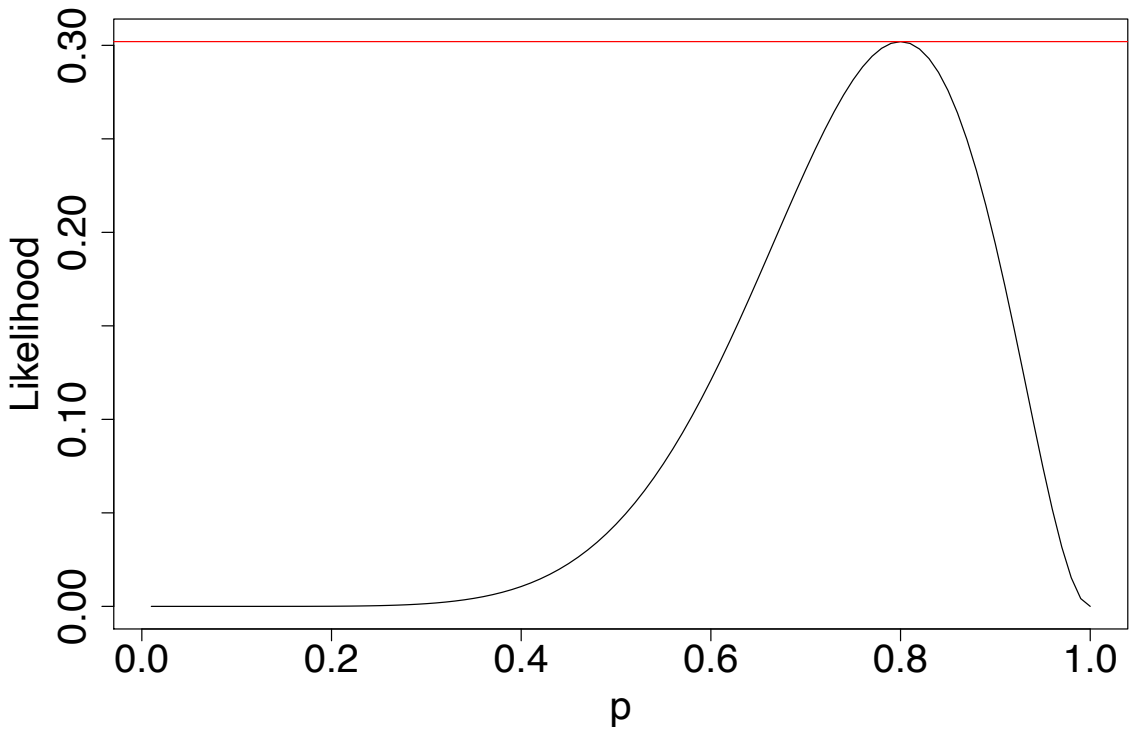
Think back to the coin example earlier, when we observed $x = 8$ heads from 10 independent trials.



How does maximum likelihood (ML) estimation work?

If we have an unknown parameter θ in a statistical model, the maximimum likelihood estimate (MLE) $\hat{\theta}$ is the value of θ which maximises $L(\theta)$.

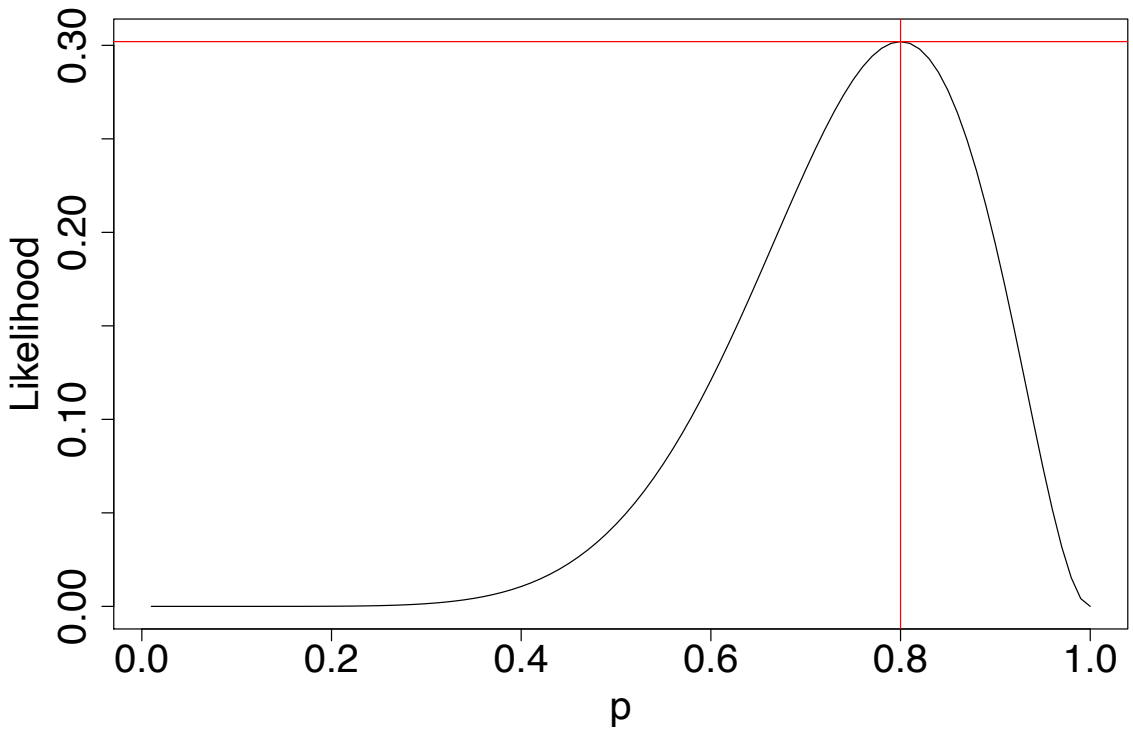
Think back to the coin example earlier, when we observed $x = 8$ heads from 10 independent trials.



How does maximum likelihood (ML) estimation work?

If we have an unknown parameter θ in a statistical model, the maximimum likelihood estimate (MLE) $\hat{\theta}$ is the value of θ which maximises $L(\theta)$.

Think back to the coin example earlier, when we observed $x = 8$ heads from 10 independent trials.



How does maximum likelihood (ML) estimation work?

In general, based on a sample of size n independent observations $x_1, x_2, ..., x_n$, the likelihood function can be written as

$$L(\theta) = \prod_{i=1}^{n} P(X = x_i \mid \theta) \qquad \text{for } X \text{ for discrete}$$

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta) \qquad \text{for } X \text{ for continuous}$$

Effectively we want to maximise $L(\theta)$ with respect to θ.

It is often much easier to work with the log (natural log) of the likelihood function $l(\theta)$.

$$l(\theta) = logL(\theta)$$

Maximising $l(\theta)$ with respect to θ and taking $\hat{\theta}$ as the maximiser, gives the MLE.

Example 1 Poisson - likelihood function

The number of cars arriving at a car park per hour from 9 to 10am is assumed to be Poisson with rate λ. The number of cars from 9-10am on six randomly selected days were: 50, 47, 82, 91, 46, 64.

We have a random sample of $X_1, X_2, ..., X_6$ from a Poisson(λ) distribution.

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

The likelihood function for λ is:

$$L(\lambda) = P(X_1 = 50)P(X_2 = 47)P(X_3 = 82)P(X_4 = 91)P(X_5 = 46)P(X_6 = 64)$$
$$= \left(\frac{\lambda^{50} e^{-\lambda}}{50!}\right)\left(\frac{\lambda^{46} e^{-\lambda}}{46!}\right)...\left(\frac{\lambda^{64} e^{-\lambda}}{64!}\right)$$

Or more generically:

$$L(\lambda) = \prod_{i=1}^{6} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{6} x_i} e^{-6\lambda}}{\prod_{i=1}^{6} x_i!}$$

Example 1 Poisson - MLE

What is the maximimum likelihood estimate for λ?

Starting with the likelihood function:

$$L(\lambda) = \frac{\lambda^{\sum_{i=1}^{6} x_i} e^{-6\lambda}}{\prod_{i=1}^{6} x_i!}$$

We take the log of the likelihood function:

$$log(L(\lambda)) = l(\lambda) = log\left(\frac{\lambda^{\sum_{i=1}^{6} x_i} e^{-6\lambda}}{\prod_{i=1}^{6} x_i!}\right)$$

$$= log(\lambda^{\sum_{i=1}^{6} x_i}) + log(e^{-6\lambda}) - log(\prod_{i=1}^{6} x_i!)$$

$$= \left(\sum_{i=1}^{6} x_i\right) log(\lambda) - 6\lambda - \sum_{i=1}^{6} log(x_i!)$$

Example 1 Poisson - MLE contd.

$$l(\lambda) = \left(\sum_{i=1}^{6} x_i\right) log(\lambda) - 6\lambda - \sum_{i=1}^{6} log(x_i!)$$

We maximise $l(\lambda)$ with respect to λ, by first differentiating:

$$\frac{dl}{d\lambda} = \frac{\sum_{i=1}^{6} x_i}{\lambda} - 6$$

And then setting equal to 0 and evaluating at $\hat{\lambda}$:

$$\frac{\sum_{i=1}^{6} x_i}{\hat{\lambda}} - 6 = 0$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{6} x_i}{6} = \bar{x} = 63.33$$

Our MLE is $\hat{\lambda} = 63.33$.

Example 2 exponential - likelihood function

Consider the time to failure for a piece of equipment. The observed times to failure are: 30.4, 7.8, 1.4, 13.1 and 67.3 hours.

Assume the lifetime follows an exponential distribution with parameter λ.

Let X be the lifetime, then

$$f_X(x) = \lambda e^{-\lambda x}$$

The likelihood function is

$$L(\lambda) = f_X(x_1)f_X(x_2)f_X(x_3)f_X(x_4)f_X(x_5)$$

$$= (\lambda e^{-\lambda 30.4})(\lambda e^{-\lambda 7.8})(\lambda e^{-\lambda 1.4})(\lambda e^{-\lambda 13.1})(\lambda e^{-\lambda 67.3})$$

Or more generically:

$$L(\lambda) \quad = \quad \prod_{i=1}^{5} \lambda e^{-\lambda x_i} \quad = \quad \lambda^5 e^{-\lambda \sum_{i=1}^{5} x_i}$$

Example 2 exponential - MLE

Take the log of the likelihood function:

$$log(L(\lambda)) = l(\lambda) = log(\lambda^5 e^{-\lambda \sum_{i=1}^{5} x_i})$$

$$= log(\lambda^5) + log(e^{-\lambda \sum_{i=1}^{5} x_i}) = 5log\lambda - \lambda \sum_{i=1}^{5} x_i$$

Maximise $l(\lambda)$ with respect to λ, by first differentiating:

$$\frac{dl}{d\lambda} = \frac{5}{\lambda} - \sum_{i=1}^{5} x_i$$

And then setting equal to 0 and evaluating at $\hat{\lambda}$:

$$\frac{5}{\hat{\lambda}} - \sum_{i=1}^{5} x_i = 0$$

$$\frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{5} x_i}{5} = \bar{x}$$

$$\hat{\lambda} = \frac{1}{\bar{x}} = 0.042 \qquad \text{this is our MLE}$$

Final word on maximum likelihood estimation

- Widely used and powerful tool.
- Here we have focused on some simple examples, but ML estimation is also very useful in estimating complex models.

Applied Probability II

Section 8: The Bootstrap

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 8: The Bootstrap

Section 8.1: The concept of bootstrapping

Introduction

So far, we have generally assumed a sample $X_1, ..., X_n$ to arise from some known distribution. That means that outside our data observations, we make additional assumptions about the shape of the underlying distribution, for example:

$$X_1, ..., X_n \sim \text{IID Normal}(\mu, \sigma^2)$$

or

$$X_1, ..., X_n \sim \text{IID Poisson}(\lambda)$$

The bootstrap is based on the idea that, without further information about the underlying distribution, the observed sample $x_1, ..., x_n$ contains all available information about $F_X(t)$, and hence the actual distribution of the data.

Essentially with bootstrapping, we resample the sample and consider it to be a sample from the population of interest.

Sampling distributions

Suppose we use a sample $X_1, ..., X_n$ to estimate some unknown θ from the true distribution of X. This could be the mean or the variance.

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

In both of these cases, we can write the estimate as a function of the sample:

$$\hat{\theta} = h(X_1, ..., X_n)$$

I.e., the definition of a statistic.

In earlier sections of the module, we used $\bar{X}$ as an estimator of μ the population mean and derived its sampling distribution:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Using the sampling distribution, we constructed CIs and hypothesis tests for μ.

With bootstrapping, we construct the sampling distribution by repeatedly resampling the observed data and evaluating $\hat{\theta}$ for each of these samples.

Overview of the bootstrap approach

When the underlying population of $X_1, ..., X_n$, call it F_X, is not known, we can find the sampling distribution of $\hat{\theta}$ using bootstrapping.

The steps are:

1. Consider $X_1, ..., X_n$ from distribution F_X.
2. θ is the parameter of interest.
3. Sample n values from $X_1, ..., X_n$ **with replacement**, giving $X_1^*, ..., X_n^*$, a bootstrap sample.
4. Compute the bootstrap estimator
$$\hat{\theta}^* = h(X_1^*, ..., X_n^*)$$
5. Repeat the previous two steps B times. Index the samples by $b = 1, ..., B$: $X_1^{*(b)}, ..., X_n^{*(b)}$ and similarly index the bootstrap statistic estimate: $\hat{\theta}^{*(b)}$.
6. We now have B bootstrapped samples and B $\hat{\theta}^*$ values.

The distribution of the $\hat{\theta}^{*(b)}$'s approximates the sampling of $\hat{\theta}$ under F_X.

The bootstrap is a very general method and relies on no assumptions about F_X, the actual distribution of X.

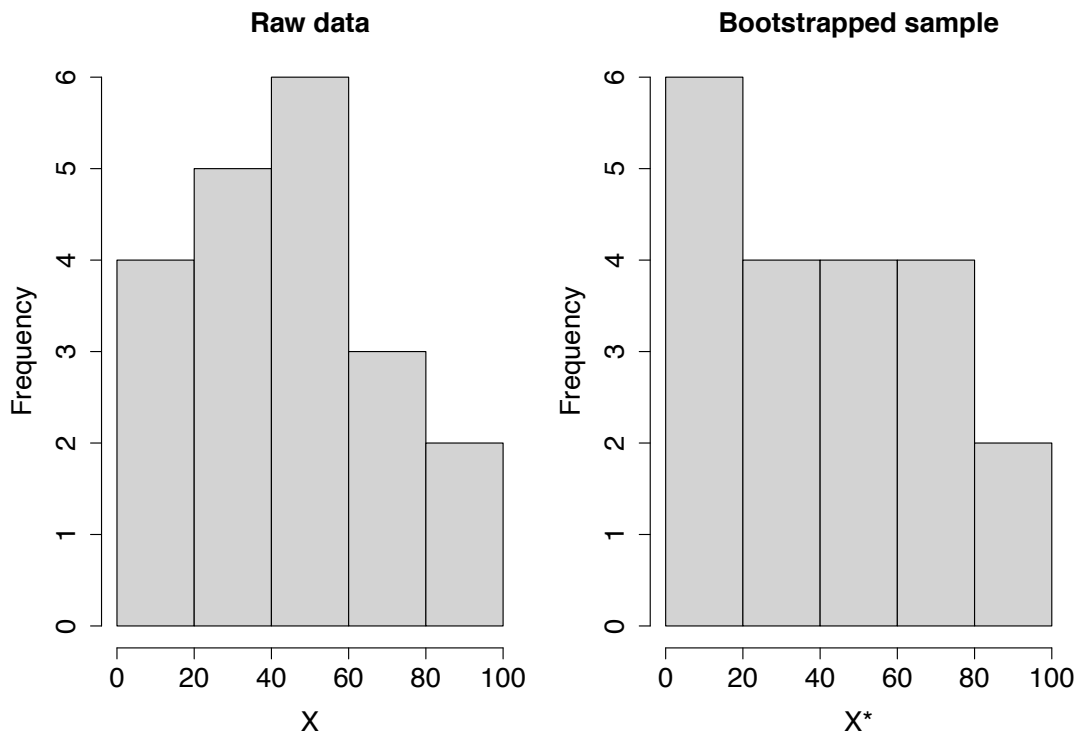
Example of a bootstrap sample

```
Raw_data
```

```
##  [1]  8 10 15 19 28 35 36 38 39 41 43 44 49 54 57 62 69 75 85 95
```

```
Bootstrapped_sample
```

```
##  [1]  8 10 10 15 15 15 28 35 38 39 41 41 41 49 62 62 69 75 95 95
```

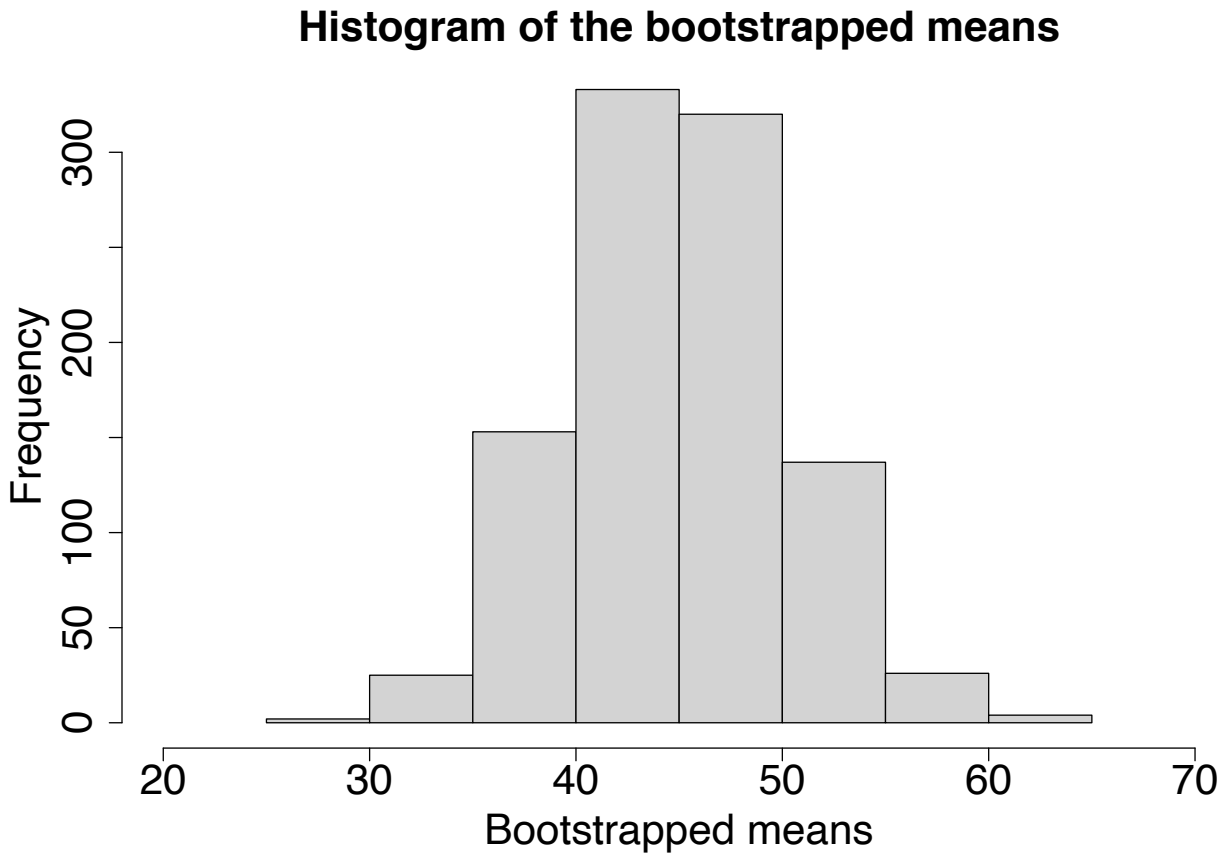


The mean of the raw data is 45.1, and the median of the bootstrapped sample is 42.15.

Sampling distribution

Continuing with the previous example, the mean from the raw data is 45.1 and the mean of the bootstrapped sample is 42.15.

Let's take 1000 bootstrap samples and compute the mean each time.



Fun fact about bootstrapping

The method is called the 'bootstrap', to suggest pulling oneself up by the bootstraps (as an example of an impossible task!).

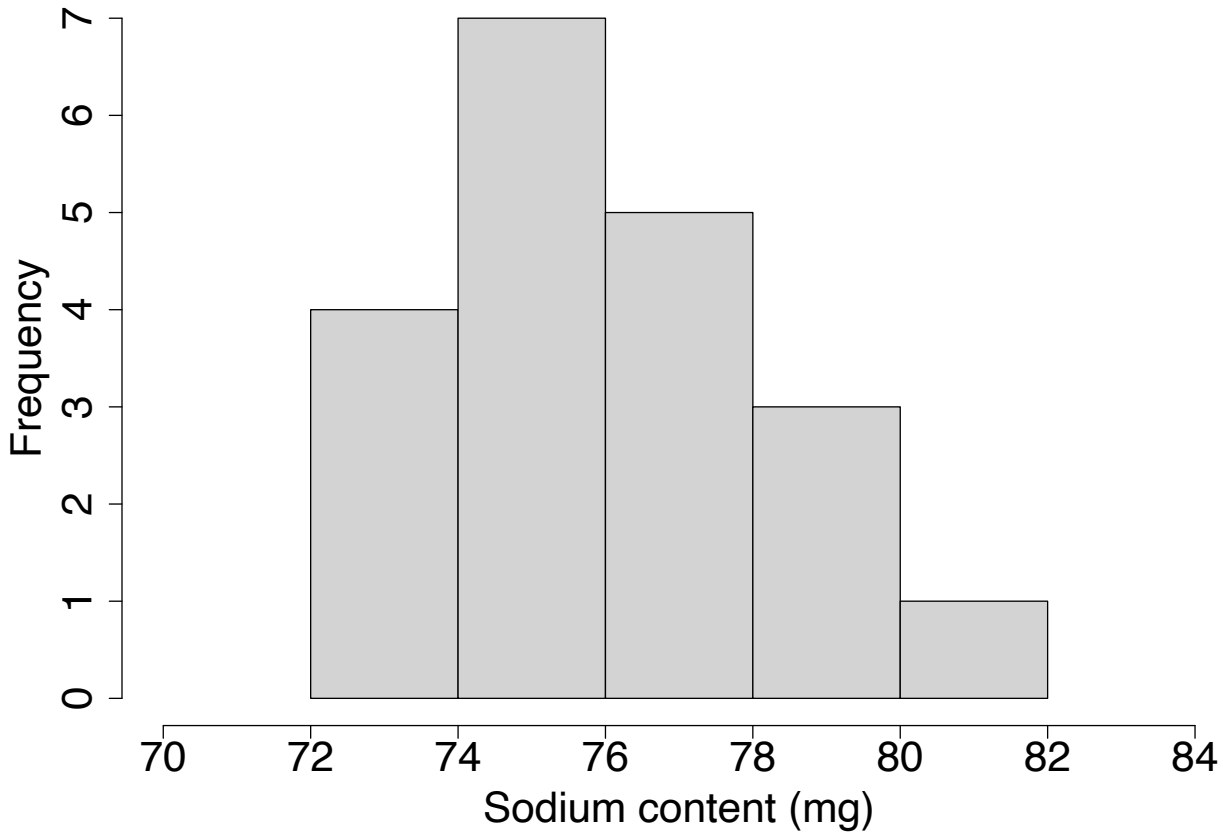
Section 8.2: Bootstrap standard error for the median

Bootstrap application

Let's have a look at a practical way that bootstrapping can be used.

Example

The sodium content (mg) of 20 fast food products was recorded. The dataset:



Sodium example contd.

We can easily find the median of the Sodium data: 75.35.

We will now use bootstrapping to find the standard error of the median.

Here are the steps:

- Take the original sample of data. Note the sample size n.
- Sample n values from it **with replacement** B times. These are called the bootstrapped samples.
- Compute the median for each bootstrapped sample and denote $M^{*(i)}$ for $i = 1, ..., B$.
- Create a histogram of the B $M^{*(i)}$ values. (Useful but not required.)
- Compute the standard deviation of $M^{*(1)}, \ldots, M^{*(B)}$. This gives us our standard error.

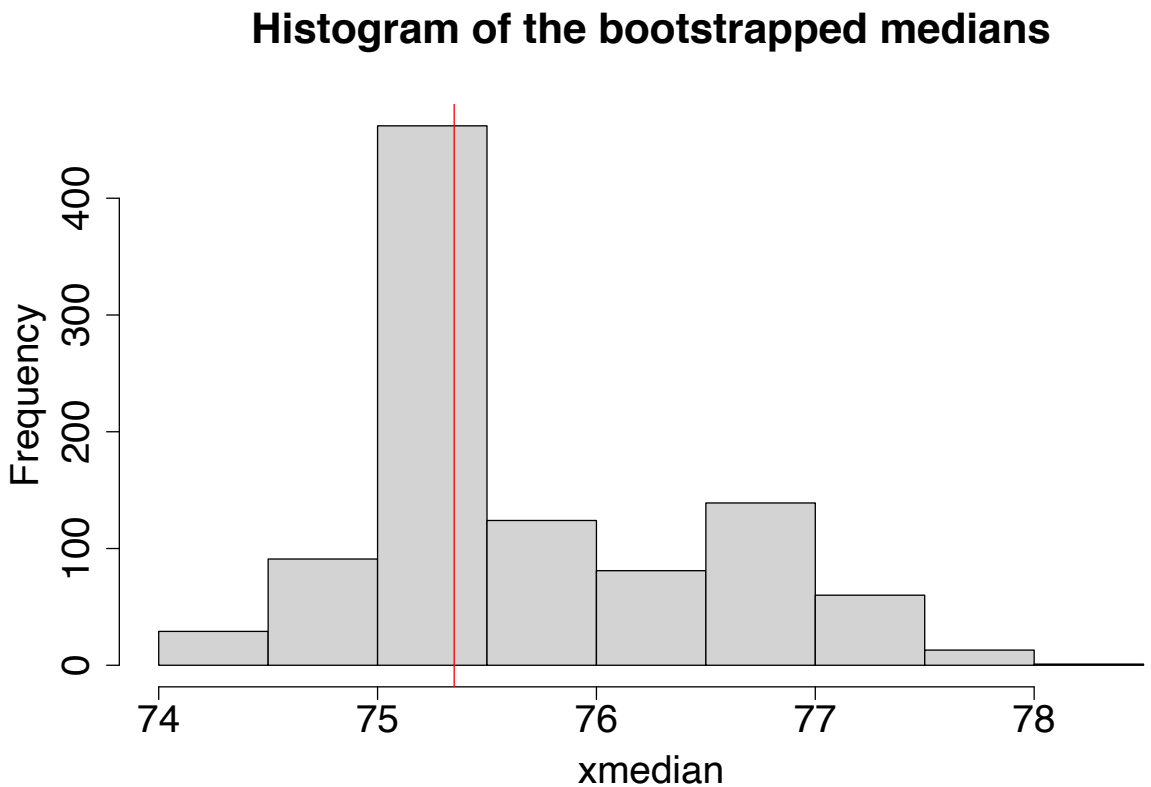
This has been implemented using R.

Here are the first 20 medians:

```
##  [1] 75.95 75.20 74.95 75.15 75.15 75.30 75.80 75.10 75.30 75.35 77.00 75.35
## [13] 74.20 77.10 75.20 75.95 75.10 77.25 75.95 75.15
```

Sodium example contd.

Here are all the bootstrapped medians in a histogram, with the median of the Sodium dataset (= 75.35) highlighted in red:



The standard deviation across the bootstrapped medians = 0.7922.

Sodium example contd.

We have found the standard error for the median for the Sodium dataset using the bootstrap.

Summary of our process:

- The original Sodium dataset had sample size $n = 20$.
- We sampled 20 values with replacement from the Sodium dataset 1000 times to create 1000 bootstrapped datasets.
- We computed the median for each of the 1000 bootstrapped datasets.
- We examined a histogram of the 1000 bootstrapped medians.
- We computed the standard deviation of the 1000 bootstrapped medians.

Our median for the Sodium dataset was 75.35.

Our bootstrapped standard error for the median was 0.7922.

Section 8.3: Bootstrap confidence intervals

Introduction

We can use bootstrapping to construct confidence intervals.

Here we will illustrate for the population median, but other statistics of interest could also be used.

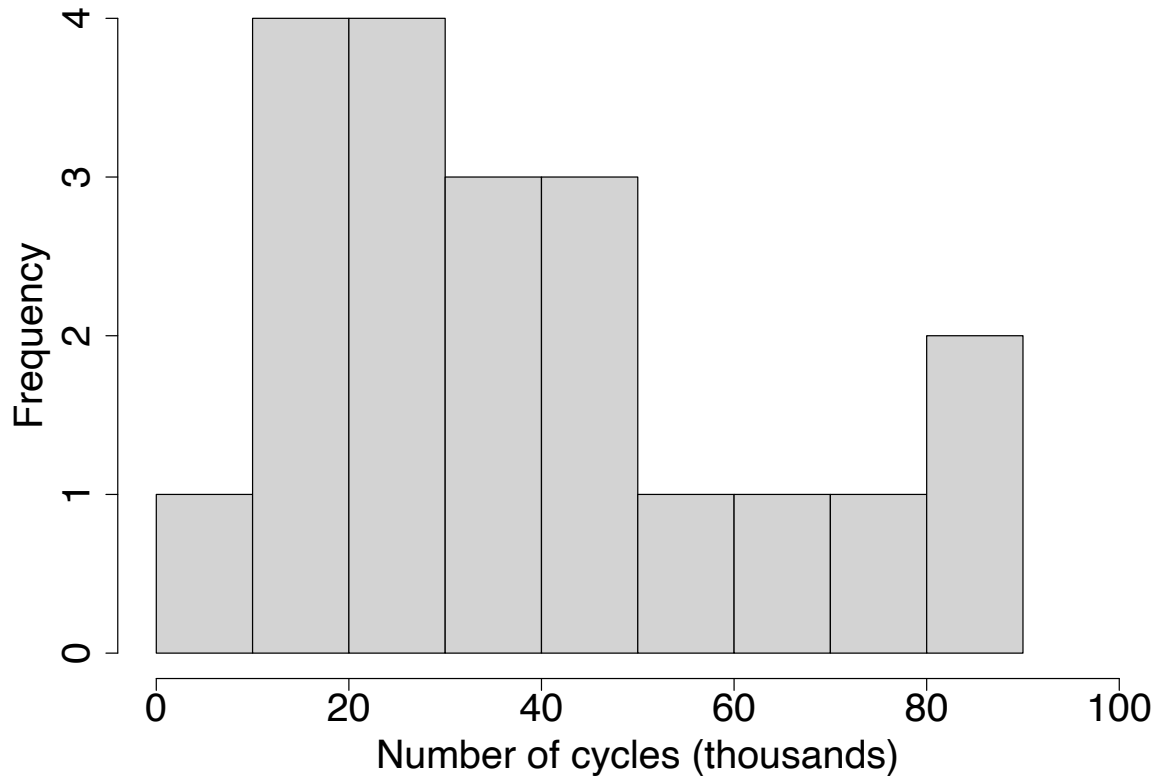
Let θ be the population median. Suppose we have a sample $x_1, ..., x_n$ from the population. We can find the estimate of the median $\hat{\theta}$ from this sample. We can then use bootstrap methods to find a confidence interval for the median.

Example

Testing of electrical and mechanical devices often involves an action such as turning a device on and off or opening and closing a device many times. The number of open-close cycles (in thousands) that it took 20 door latches to fail was recorded.

Here is the dataset and a histogram of it.

```
##   [1]   7  11  15  16  20  22  24  25  29  33  34  37  41  42  49  57  66  71  84  90
```



Steps to construct a bootstrap confidence interval

Let θ be the population median. Let $\hat{\theta}$ be the sample median for a sample of size n.

Here are the steps to construct a bootstrap confidence interval for θ:

- From the original sample of size n, sample values with replacement B times. These are called the bootstrap samples.
- Compute the median for each bootstrap sample and denote M^{*i}.
- Create a histogram of the M^{*i} values. (Useful but not required.)
- Order the M^{*i} values and denote each ordered value by $M^{*(i)}$. So for example, $M^{*(1)}$ is the lowest and $M^{*(B)}$ the highest of the bootstrap sample median estimates.
- Let $a = (\alpha/2) * B$. The ordered bootstrap median estimates are $M^{*(1)}, M^{*(2)}, ..., M^{*(B)}$. Then $(M^{*(a)}, M^{*(B-a)})$ is an approximate $(1 - \alpha) * 100\%$ confidence interval for θ.

Example Step 1 and 2

Here is the raw data again:

7 11 15 16 20 22 24 25 29 33 34 37 41 42 49 57 66 71 84 90

The estimate of the median is: 33.5. Let's now construct a confidence interval.

Step 1: From the original sample of size $n = 20$, sample values with replacement B = 1000 times.

Here are the first five bootstrap samples down to the 1000th:

37 66 16 33 42 7 41 29 20 41 16 66 29 71 24 37 16 42 29 15
20 42 90 25 71 15 90 20 84 20 15 7 90 37 25 37 15 33 22 66
16 49 71 22 66 16 49 16 22 16 25 16 20 41 25 57 15 84 22 22
29 20 37 22 15 37 66 42 42 11 42 7 29 15 25 20 25 66 34 42
84 42 71 33 66 20 15 34 66 41 90 29 15 42 37 24 49 84 71 22

⋮

33 33 71 25 33 25 57 29 20 29 37 71 29 16 42 15 66 37 42 71

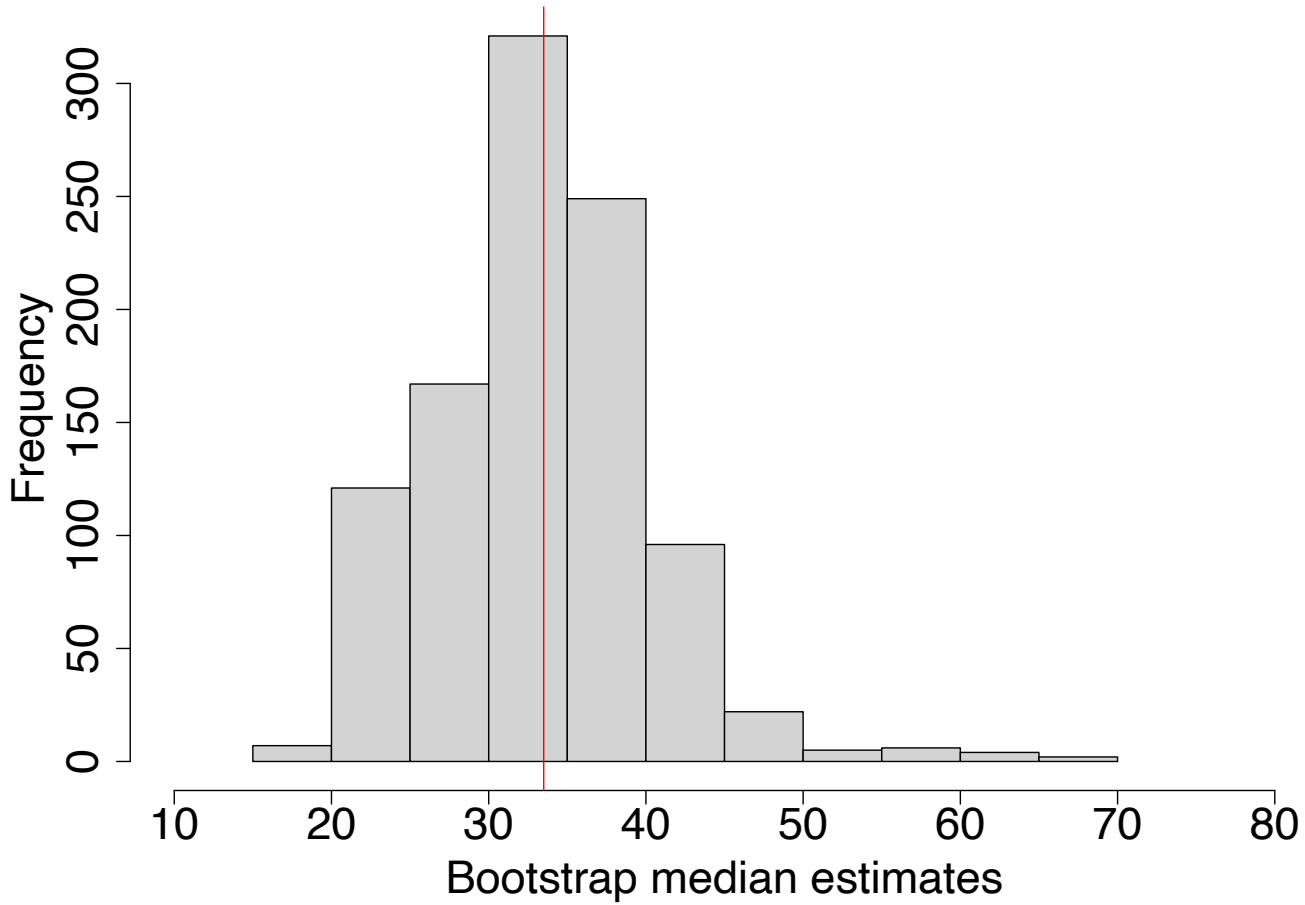
Step 2: Compute the median for each bootstrap sample and denote M^{*i}.

The medians $M^{*1}, M^{*2}, M^{*3}, M^{*4}, M^{*5}, ..., M^{*1000}$ are:

31, 29, 22, 29, 41.5, . . . , 33.

Example Step 3

Step 3: Create a histogram of the 1000 M^{*i} bootstrap median estimates. The red lines shows the sample median.



Example Step 4

Step 4: Order the M^{*i} values and denote each ordered value by $M^{*(i)}$. So for example, $M^{*(1)}$ is the lowest and $M^{*(B)}$ the highest of the bootstrap sample median estimates.

Here are the ordered values (1000 in total, showing first 50 and last 50):

```
 [1] 19.0 20.0 20.0 20.0 20.0 20.0 20.0 21.0 21.0 22.0 22.0 22.0 22.0 22.0 22.0
[16] 22.0 22.0 22.0 22.0 22.0 22.5 22.5 22.5 23.0 23.0 23.0 23.0 23.0 23.0 23.0
[31] 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.5 23.5 23.5
[46] 23.5 23.5 23.5 23.5 24.0
```

⋮

```
 [1] 42.0 42.0 42.0 42.0 42.0 42.0 42.0 42.0 43.0 45.0 45.0 45.0 45.5 45.5 45.5
[16] 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 49.0 49.0 49.0 49.0 49.0
[31] 49.0 49.5 49.5 49.5 53.0 53.0 53.0 53.0 53.0 57.0 57.0 57.0 57.0 57.0 57.5
[46] 61.5 61.5 61.5 61.5 66.0 66.0
```

Step 5

Step 5: Let $a = (\alpha/2) * B$. The ordered bootstrap median estimates are $M^{*(1)}, M^{*(2)}, ..., M^{*(B)}$. Then $(M^{*(a)}, M^{*(B-a)})$ is an approximate $(1 - \alpha) * 100\%$ confidence interval for θ.

From the ordered medians, we want to pull out the values $(M^{*(a)}, M^{*(B-a)})$, where $a = (\alpha/2) * B$.

Let's construct a 95% confidence interval.

Then, $a = (0.05/2) * 1000 = 25$ and $B - a = 1000 - 25 = 975$, so we want to pull out the 25th and the 975th medians from the ordered vector of bootstrap medians.

Step 5

Let's go back to the ordered medians.

We want the 25th and the 975th values from the ordered vector of bootstrap medians:

```
 [1] 19.0 20.0 20.0 20.0 20.0 20.0 20.0 21.0 21.0 22.0 22.0 22.0 22.0 22.0 22.0
[16] 22.0 22.0 22.0 22.0 22.0 22.5 22.5 22.5 23.0 23.0 23.0 23.0 23.0 23.0 23.0
[31] 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.0 23.5 23.5 23.5
[46] 23.5 23.5 23.5 23.5 24.0
```

⋮

```
 [1] 42.0 42.0 42.0 42.0 42.0 42.0 42.0 42.0 43.0 45.0 45.0 45.0 45.5 45.5 45.5
[16] 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 45.5 49.0 49.0 49.0 49.0 49.0
[31] 49.0 49.5 49.5 49.5 53.0 53.0 53.0 53.0 53.0 57.0 57.0 57.0 57.0 57.0 57.5
[46] 61.5 61.5 61.5 61.5 66.0 66.0
```

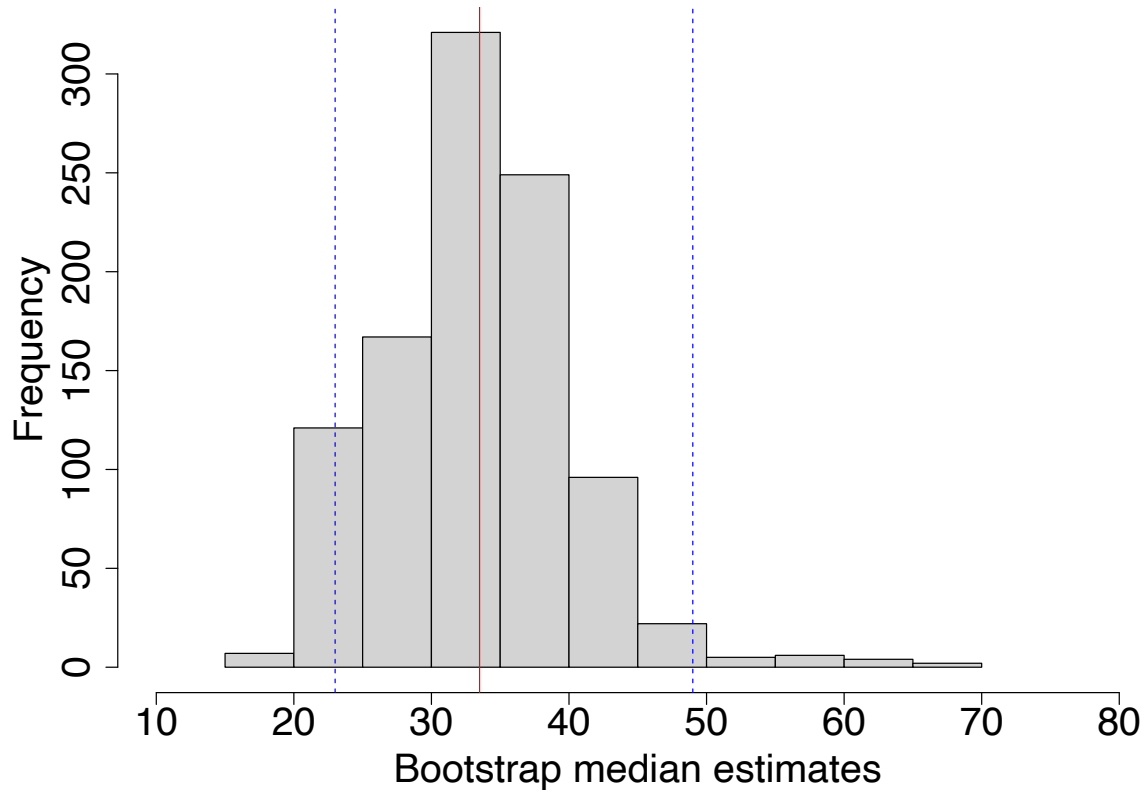
Our 95% CI is: (23, 49)

The confidence interval graphically

The estimated median (for the original sample of data) is 33.5.

The 95% confidence interval for the population median is (23, 49).

Graphically:



Example summary

The 95% confidence interval for the population median is (23, 49).

We are 95% confident that the true population median number of cycles (in thousands) that it takes for the door latch to fail lies between 23 and 49.

There are alternative bootstrap methods for constructing the CI that will adjust for potential biases.

Section 8.4: Bootstrap hypothesis tests

Hypothesis testing

We can use bootstrapping to test hypotheses of interest.

NB

In bootstrap hypothesis testing, the resampling is conducted under the conditions that ensure the null hypothesis, H_0, is true.

We will consider the 1-sample example, where we may wish to test the hypothesis:

$$H_0\text{: } \theta = \theta_0 \text{ versus } H_A\text{: } \theta \neq \theta_0$$

Where the parameter of interest may be the median or the mean as a measure of location.

Example

A set of data of size $n = 25$ was collected in a study. It was believed that the mean of the population from which it came from was > 1.

We will use bootstrapping to test the hypothesis:

$$\mathrm{H}_0\colon\ \mu = \mu_0 \text{ versus } \mathrm{H}_A\colon\ \mu > \mu_0$$

where $\mu_0 = 1$ in this case.

The steps in general to follow are:

- Let $x_1, \ldots, x_n$ denote the observed data. Let $\hat{\mu} = \bar{x}$ be an estimate of μ.
- Sample with replacement from $x_1 - \bar{x} + \mu_0, \ldots, x_n - \bar{x} + \mu_0$. This will ensure that the null hypothesis is true.
- Compute $\bar{x}^{*i}$, the mean for each bootstrap sample. It is useful to generate a histogram of these bootstrap estimates.
- Compute the p-value:

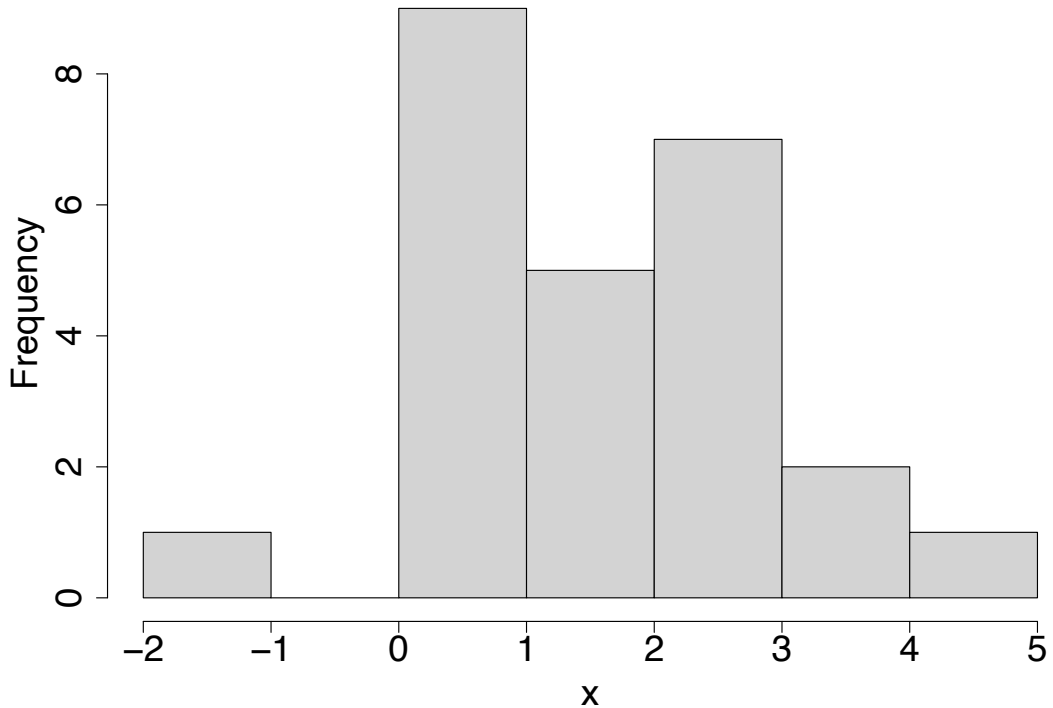
$$\text{p-value} = \frac{\#(\bar{x}^{*i} \geq \bar{x})}{B}$$

Example step 1

Step 1: Let $x_1, \ldots, x_n$ denote the observed data. Let $\hat{\mu} = \bar{x}$ be an estimate of μ.

Here is the raw data and a histogram of it:

```
 [1]  0.620  2.300  0.710  2.300  0.800  3.200  2.600  2.100  2.400  1.500
[11]  0.900  0.830  0.067  1.900  2.200  0.018  1.800  2.400  1.300 -1.100
[21]  4.600  1.600  0.380  3.800  0.500
```



The sample mean is: $\hat{\mu} = \bar{x} = 1.589$.

Example step 2

Step 2: Sample with replacement from $x_1 - \bar{x} + \mu_0, \ldots, x_n - \bar{x} + \mu_0$. This will ensure that the null hypothesis is true.

We are testing the hypothesis

$$\mathrm{H}_0\colon\ \mu = 1 \text{ versus } \mathrm{H}_A\colon\ \mu > 1$$

Therefore, we compute $x_i - \bar{x} + \mu_0 = x_i - 1.589 + 1$, which gives:

```
 [1]  0.031  1.711  0.121  1.711  0.211  2.611  2.011  1.511  1.811  0.911
[11]  0.311  0.241 -0.522  1.311  1.611 -0.571  1.211  1.811  0.711 -1.689
[21]  4.011  1.011 -0.209  3.211 -0.089
```

We sample from this with replacement B = 5000 times giving:

```
 [1] -1.689 -0.571 -0.209 -0.089  0.031  0.121  0.121  0.311  0.711  0.911
[11]  0.911  1.011  1.011  1.011  1.211  1.311  1.711  1.711  1.811  1.811
[21]  1.811  2.011  2.611  4.011  4.011
```

⋮

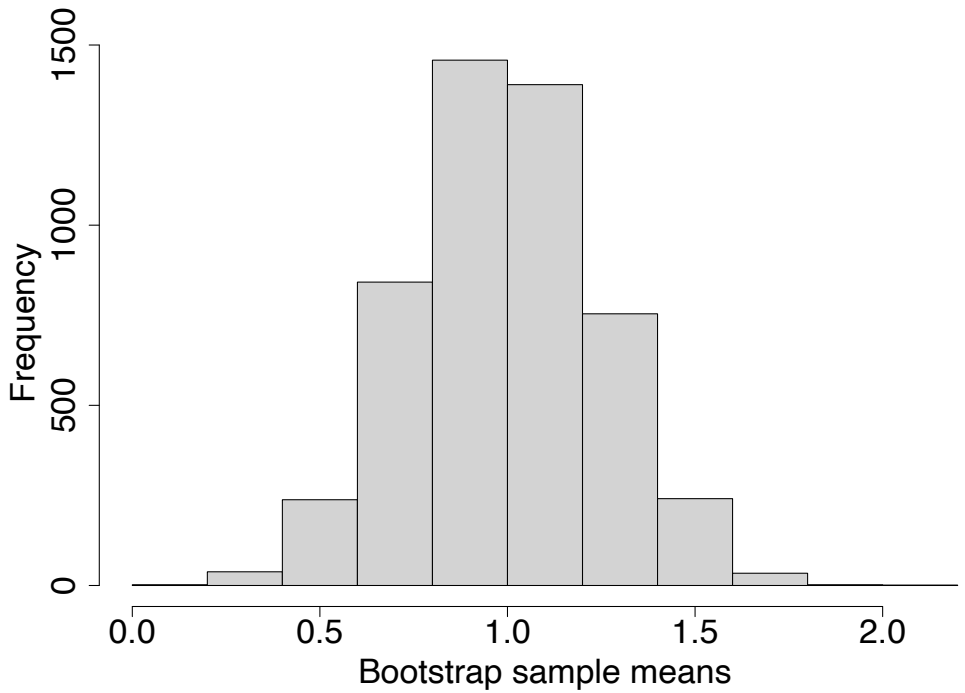
```
 [1] -1.689 -0.089  0.121  0.211  0.211  0.211  0.241  0.311  0.711  0.711
[11]  1.011  1.211  1.311  1.511  1.611  1.611  1.711  1.811  1.811  2.011
[21]  2.611  2.611  3.211  3.211  4.011
```

Example step 3

Step 3: Compute $\bar{x}^{*i}$, the mean for each bootstrap sample. It is useful to generate a histogram of these bootstrap estimates.

The means for the first 15 bootstrap samples and a histogram of all bootstrap means:

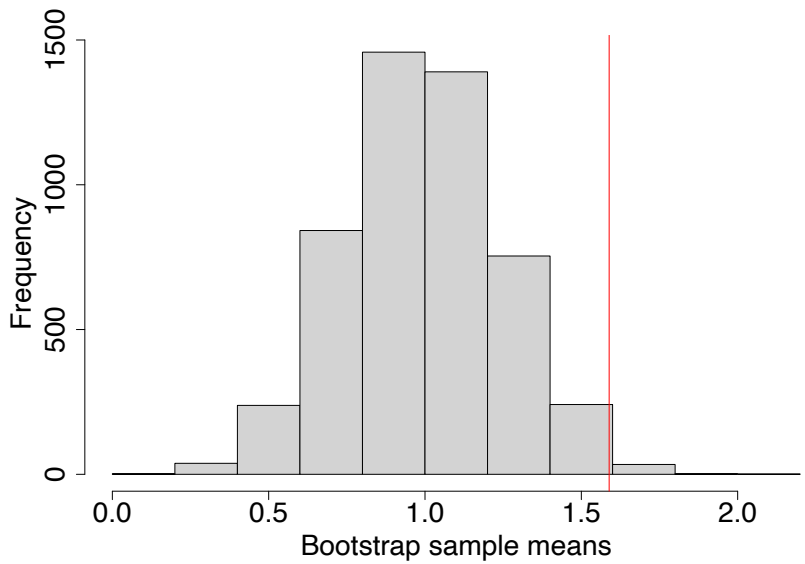
```
 [1] 1.10452 1.01788 1.27968 1.20768 1.31568 1.14280 1.04036 0.91676 0.86436
[10] 1.34540 1.20420 0.96940 0.80752 0.86052 1.03940 1.49180 1.33900 0.93840
[19] 1.06348 0.95316
```



Example step 4

Step 4: Compute the p-value: p-value $= \frac{\#(\bar{x}^{*i} \geq \bar{x})}{B}$

Graphically (with the observed mean highlighted by a red line):



The number of $\bar{x}^{*i} > 1.589 = 46$. Therefore the p-value $= 46 \ / \ 5000 = 0.0092$.

Notes on the p-value calculation:

- Remember a p-value is the probability of getting a value as extreme or more extreme than what was observed assuming that the null hypothesis is true.
- We have used bootstrapping to approximate the sampling distribution of the mean under the null hypothesis.

Example summary

H_0: $\mu = 1$ versus H_A: $\mu > 1$

$\hat{\mu} = 1.589$.

P-value $= 0.0092$.

We reject the H_0 and conclude that the true population mean is greater than 1.

Applied Probability II

Section 9: The Normal Distribution

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 9: The Normal Distribution

The normal or Gaussian distribution

The normal (or Gaussian) distribution has a central place in statistics, largely as a result of the central limit theorem.

In this Section we will examine various aspects of the normal distribution.

Section 9.1: Assessing normality

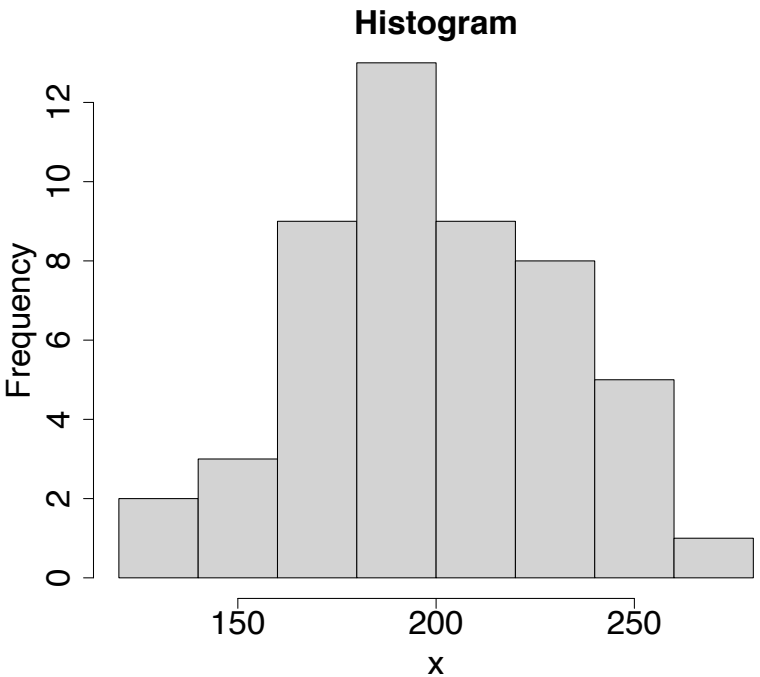
The univariate normal distribution

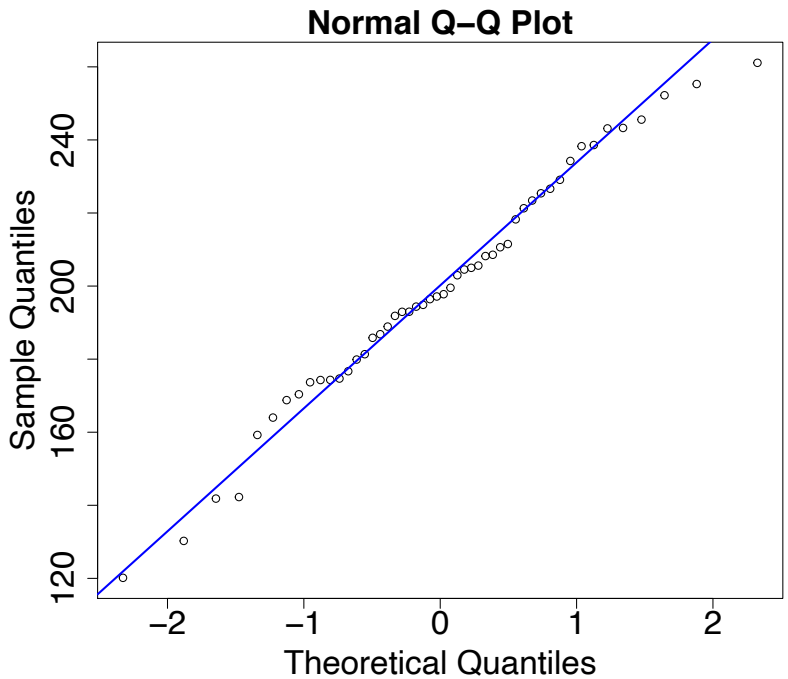
In some sections so far in this module, we have carried out statistical analyses or modelling where a normal distribution was assumed.

To validate such an assumption, we can assess the normality of the data for which the assumption is made. This can be done by observing a histogram of the data (which is an approximation of the probability density function), or we can use a quantile-quantile (QQ) plot, which is a little bit more formal, but also subjective.

We have already used QQ plots to assess normality in other Sections. In this sub-section, we will examine in more detail how to assess QQ plots.

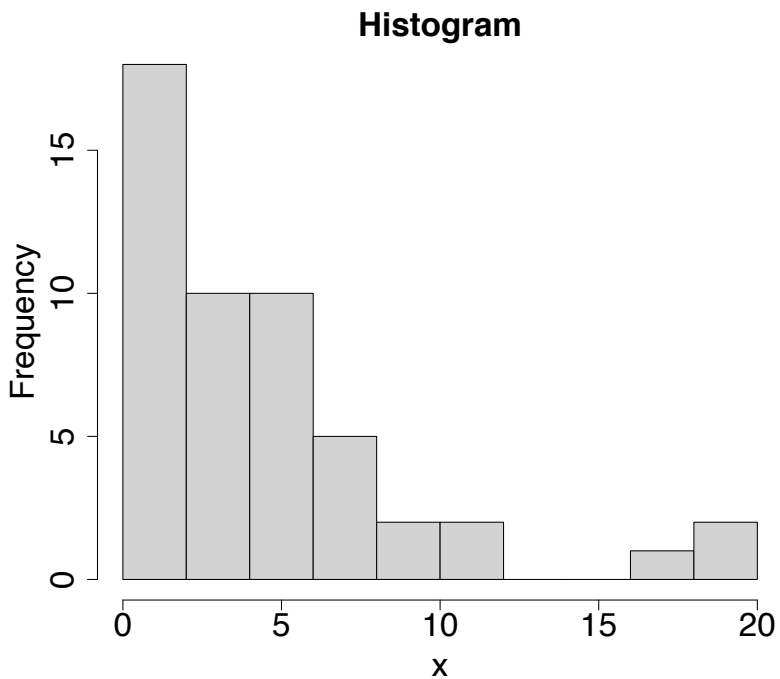
Example 1

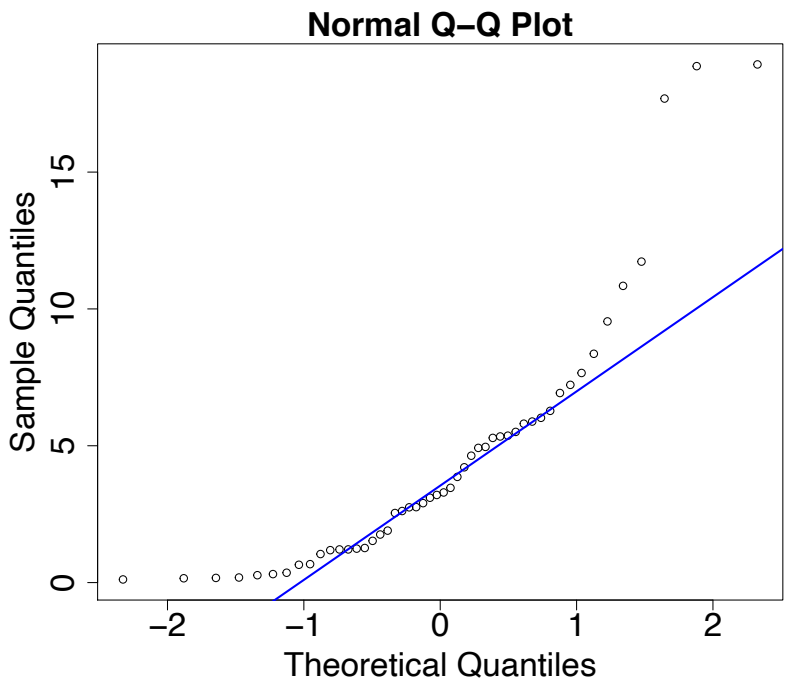




- Normality seems to be a reasonable assumption

Example 2

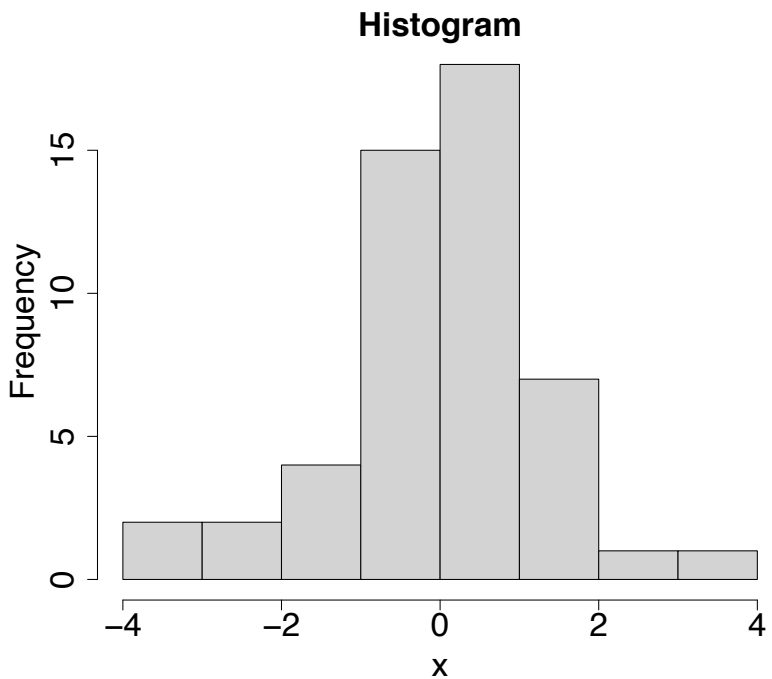


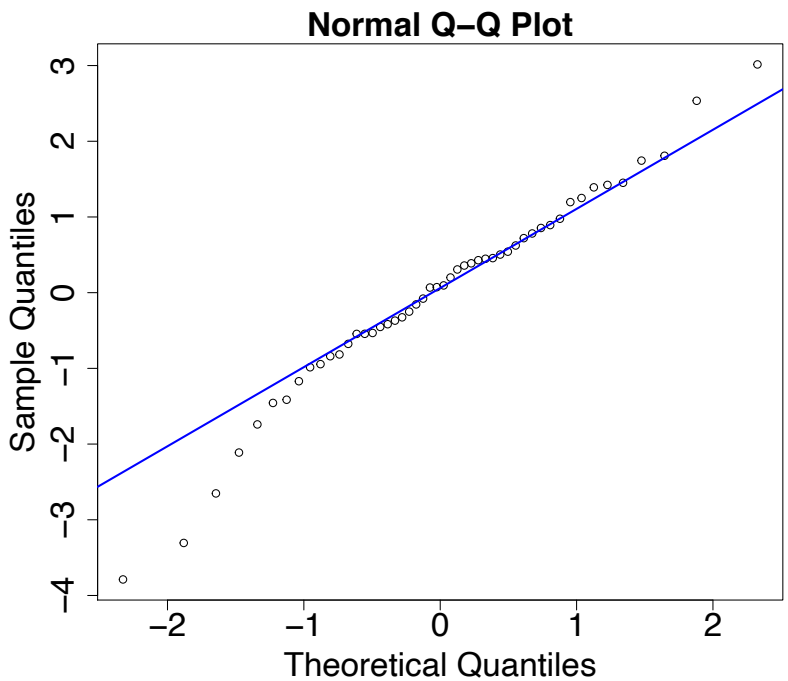


In the QQ plot:

- Short tail shown at lower end with points above the line (points would be below the line here if long tailed in this direction).
- Long tail shown at upper end with points above the line (points would be below the line here if short tailed in this direction).

Example 3





In the QQ plot:

- Long tail shown at lower end with points below the line (points would be above the line here if short tailed in this direction).
- Slight long tail shown at upper end with points above the line (points would be below the line here if short tailed in this direction).

Assessing normality summary

We have assessed the normality of a sample of data using the following methods

- Histogram: we observe if the histogram follows the bell shaped curve typical of the normal distribution.
- QQ plots: we plot the sample quantiles against the theoretical quantiles of the normal distribution and see if they follow a straight line.

There are also other plots that can be used (e.g., PP plots), and there are formal hypothesis tests that can assess normality (e.g., the Shapiro-Wilks test).

Section 9.2: The bivariate normal distribution

Joint discrete random variables

If X and Y are discrete random variables we can define their joint probability mass function (pmf) as:

$$p(x, y) = P(X = x, Y = y)$$

The marginal probability mass function of X is:

$$p_X(x) = P(X = x) = \sum_y P(X = x, Y = y)$$

Joint discrete random variables - Example

Toss a coin three times. Let X = the number of heads on the first toss. Let Y = the total number of heads. There are 8 equally likely outcomes:

S	X	Y
hhh	1	3
hht	1	2
hth	1	2
thh	0	2
htt	1	1
tht	0	1
tth	0	1
ttt	0	0

We can tabulate the joint pmf of X and Y:

		Y				
		0	1	2	3	
X	0	1/8	2/8	1/8	0	0.5
	1	0	1/8	2/8	1/8	0.5
		1/8	3/8	3/8	1/8	1

The entries in the final column and row give $p_x(x)$ and $p_x(y)$, the marginal probability functions of X and Y respectively.

Joint continous distributions

Let X and Y be continuous random variables with joint cumulative distribution function (cdf) $F(x, y)$. They are jointly continuous if there is a function $f(x, y) \geq 0$ such that

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

f is called the joint probability density function (pdf) of X and Y.

The marginal cumulative distribution function (cdf) F_x of X can be obtained as:

$$F_x(a) = P(X \leq a, Y \leq \infty) = \int_{-\infty}^{a} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx$$

Thus, the marginal density f_x of X is

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

Similarly, the marginal density f_y of Y is

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

The bivariate normal distribution

The univariate normal distribution probability density function is given by:

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

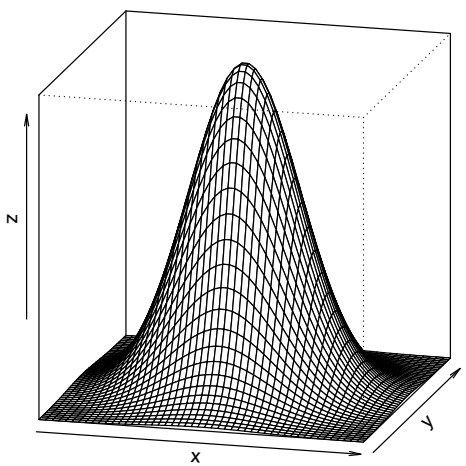
The bivariate normal density is given by:

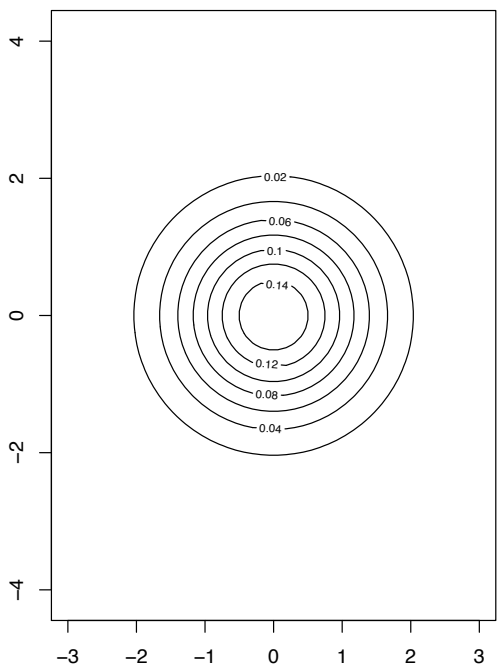
$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)}\right]$$

The bivariate normal distribution - Example

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)}\right]$$

Normal with $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ and $\rho = 0$:

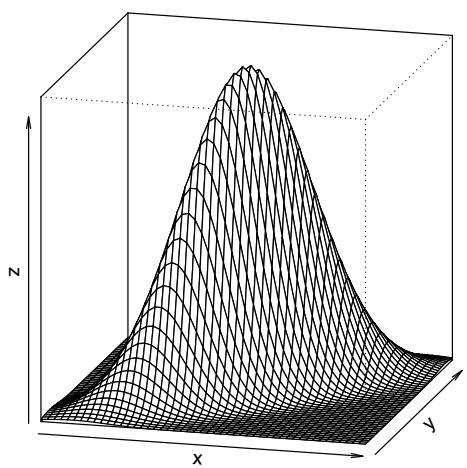


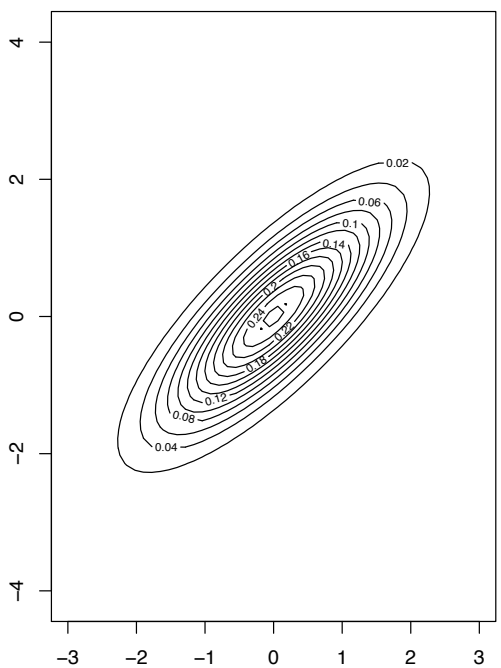


The bivariate normal distribution - Example

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)}\right]$$

Normal with $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ and $\rho = .8$:

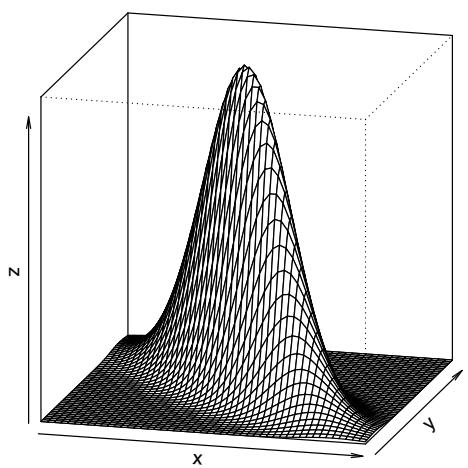


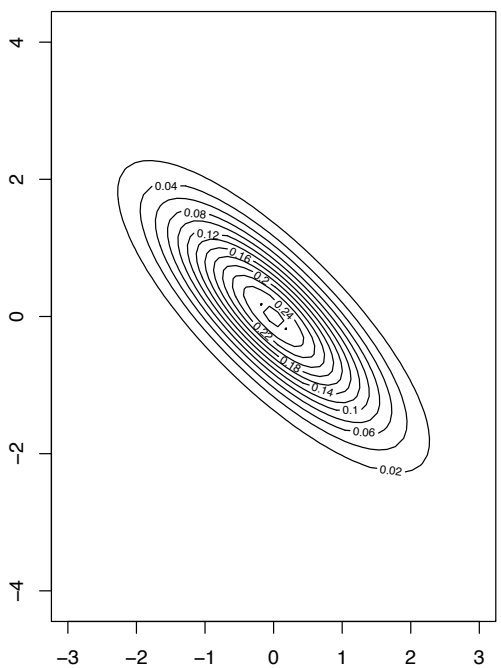


The bivariate normal distribution - Example

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)}\right]$$

Normal with $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ and $\rho = -.8$:





The marginal distributions of the bivariate normal

The joint pdf is:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}}{2(1-\rho^2)}\right]$$

The marginal distributions of X is:

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

Which we can show is equal to:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{(y-b_x)^2}{2\sigma_Y^2(1-\rho^2)}\right] dy$$

where

$$b_x = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x-\mu_X)$$

The marginal distributions of the bivariate normal

When we examine:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{(y-b_x)^2}{2\sigma_Y^2(1-\rho^2)}\right] dy$$

We can see that the right hand side is the integral of a normal probability density function with parameters: $N(b_x, \sigma_Y^2(1-\rho^2))$, and thus integrates to 1.

We arrive at:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{1}{2}\frac{(x-\mu_x)^2}{\sigma_x^2}\right)$$

Therefore $X \sim N(\mu_x, \sigma_x^2)$.

It can be shown similarly that the marginal distribution of Y is $N(\mu_y, \sigma_y^2)$.

Independence

In general, if two random variables are independent, we know that the correlation between them (ρ) equals 0. But a zero correlation does not imply independence.

However, if X and Y follow a bivariate normal distribution, then X and Y are independent *if and only if* ρ equals 0.

Section 9.3: The multivariate normal distribution

The multivariate normal distribution

Consider a set of n independent and identically distributed (IID) standard normal random variables

$$Z_i \underset{IID}{\sim} N(0,1)$$

The covariance matrix for $\mathbf{Z}$ is $\mathbf{I_n}$ and $\mathrm{E}[\mathbf{Z}] = 0$.

Let $\mathbf{B}$ be an $m \times n$ matrix of fixed coefficients and $\boldsymbol{\mu}$ be an m-vector of fixed coefficients.

Then, the m-vector $\mathbf{X} = \mathbf{BZ} + \boldsymbol{\mu}$ is said to have a multivariate normal distribution.

The mean of $\mathbf{X}$ is: $E[\mathbf{X}] = \boldsymbol{\mu}$.

The covariance matrix of $\mathbf{X}$ is: $Var[\mathbf{X}] = \mathbf{BB^T} = \boldsymbol{\Sigma}$.

Or, we can say that

$$\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The multivariate normal distribution pdf

The bivariate normal probability density function that we looked at in the last sub-section generalises to the multivariate case.

If

$$\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^m |\boldsymbol{\Sigma}|}} exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

for $\mathbf{x} \in \mathbb{R}^m$

If you study the module Multivariate Linear Analysis next year (and Data Analytics the following year), you will come across the MVN distribution again.