STAT 110: Week 6

University of Otago

Outline

- Previous lectures:
  - Explored statistical models for normally distributed data
  - Data are modelled as normal with mean μ and variance σ^2
  - Found confidence interval for μ
  - Hypothesis test for μ
- Today: begin to look at relationships between variables
  - Relationship between a continuous variable and a categorical variable
  - Continuous variable: can take any value
    - e.g. height, weight, time to run 100 m
    - It could be limited a range (e.g. height must be positive)
  - Categorical variable: represents categories or groups
    - e.g. sex, country of birth, blood type, etc.

Motivation

- What is the effect of sensory deprivation?[1]
  - Study designed to explore this question, where all participants were prisoners
- Twenty participants were selected
  - 82 inmates initially volunteered
    - Removed: medically unfit, low IQ, history of behaviour or psychiatric problems in prison
- The 20 participants were randomly allocated into two groups
  - Solitary confinement
  - Control (ordinary prison life)
- EEG[2] frequencies were obtained on day 7
  - Is there a difference in arousal levels? (as measured by EEG frequency)

[1] From Journal of Abnormal Psychology, 1972, **79**, 54–59

[2] EEG (Electroencephalogram) measures the frequency of brain waves

Data: EEG frequencies

- Import the data

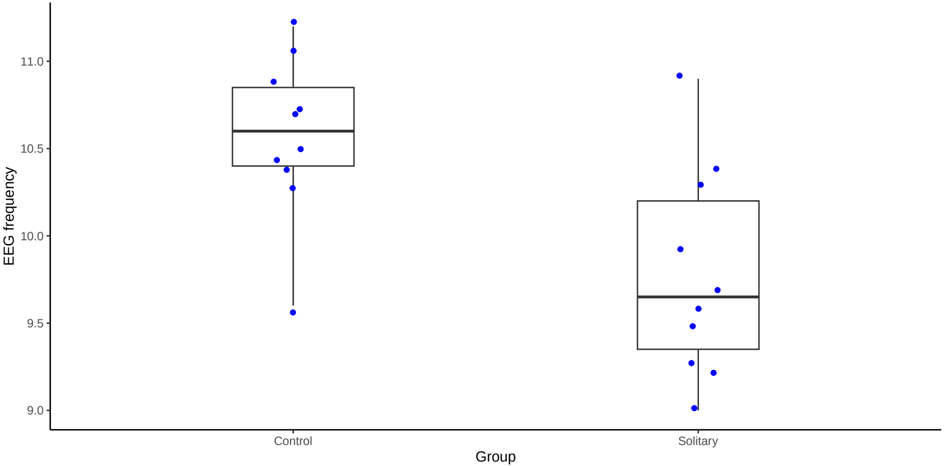
```r
EEG = read.csv('EEG.csv')
```

- Have a look at the data:

```r
head(EEG)
```

```
##     Group Freq
## 1 Control 10.7
## 2 Control 10.7
## 3 Control 10.4
## 4 Control 10.9
## 5 Control 10.5
## 6 Control 10.3
```

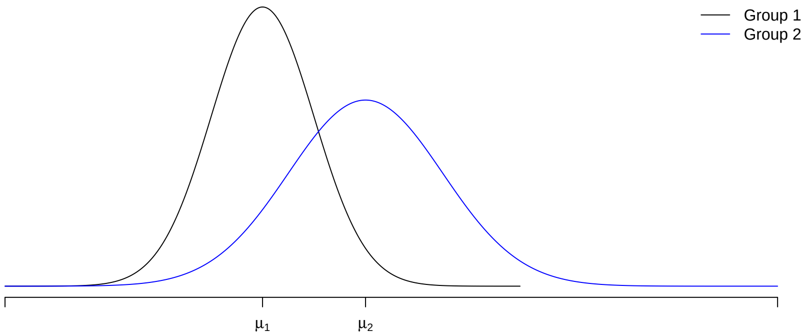
Visualise the data



Problem

- We have looked at models:
  - Data are normally distributed with mean μ and variance σ^2
  - Focus has been on the estimation of a (single) mean μ
- We need to extend our model to allow for two groups of data
  - Group 1 (experimental): normally distributed with mean μ_1 and variance σ_1^2
  - Group 2 (control): normally distributed with mean μ_2 and variance σ_2^2
- Interest is in the difference in means between the two groups
  - $\mu_1 - \mu_2$ (or $\mu_2 - \mu_1$)
- Difference in the mean arousal level between the deprived and the controls

Model (graphical representation)



Other examples

- There are other applications we could have used to motivate:
  - Cuckoos are avian brood parasites: they lay their eggs in the nest of other birds
    - Compare the length of cuckoo eggs in wren and robin nests
  - Explore differences in chemical composition of wine or olives
    - Different cultivars (wine)
    - Different regions (olives)
  - Comparing athletic performance
    - Comparing resistance training and traditional training for athletes in some sport
  - Survival time for breast cancer patients
    - Comparing candidate drug and placebo
  - Gene expression in a section of the brain
    - Comparing diseased, with healthy controls
  - You will see some of these in the Assignment

How to find a confidence interval

- Much of what we have learned previously 'carries over'
- Use statistics (from sample) to estimate parameters (from population)
  - Parameter: $\mu_1 - \mu_2$
  - Statistic: $\bar{y}_1 - \bar{y}_2$
- Standard error for $\bar{y}_1 - \bar{y}_2$
  - Tells us about the variation in $\bar{y}_1 - \bar{y}_2$ in repeated samples
  - Estimated standard error: $s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- The confidence interval is given as

$$\underbrace{\bar{y}_1 - \bar{y}_2}_{\text{statistic}} \pm \underbrace{t_{\nu,1-\alpha/2}}_{\text{multiplier}} \underbrace{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}_{\text{standard error}}$$

Standard error

- The standard error is different from before, but similar
  - Follows from variance rules (week 3; ice cream)
  - Observations in the two groups are independent

$$Var(\bar{y}_1 - \bar{y}_2) = Var(\bar{y}_1) + Var(\bar{y}_2)$$
$$= \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

Multiplier

- The multiplier is again given by the t-distribution
  - The use of the t-distribution relies on an approximation
    - Approximation is accurate provided we have more than a handful of observations $(n_1 > 5, n_2 > 5)$
- The degrees of freedom, ν, we use is given by a complicated formula
  - You have no need to know or learn this

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

- If software isn't available, simpler approximations for ν are sometimes used
  - e.g. using smaller of $n_1 - 1$ and $n_2 - 1$
  - Conservative

Calculating the confidence interval

- We could calculate the confidence interval by hand:
  - Find the sample mean in each group: $\bar{y}_1, \bar{y}_2$
  - Find the sample variance in each group: s_1^2, s_2^2
  - Find the standard error
  - Calculate the degrees of freedom
  - Find the t-multiplier
  - Construct the confidence interval
- Tedious task
  - Important to know how the interval is constructed
    - You may be asked to do various aspects of it for assignment/test/exam
  - Easier to use R to calculate the interval

In R

- We use the same function as before: `t.test`
  - This requires us to have the data for each group separately
  - Currently our data are in a single data frame

```
head(EEG)

##     Group Freq
## 1 Control 10.7
## 2 Control 10.7
## 3 Control 10.4
## 4 Control 10.9
## 5 Control 10.5
## 6 Control 10.3
```

- The variable `Group` distinguishes which group the observation is from
  - Either `Control` or `Solitary`

In R

- There are several ways in R we could separate into two groups
  - We will use `subset`
    - Subsets the data based on a specified criteria
  - Only cover 'basic' data handling in STAT 110
    - See STAT 260

```
control = subset(EEG, Group == "Control")
solitary = subset(EEG, Group == "Solitary")
```

- We use two equal signs (==) to *check* equality
  - `Group == "Solitary"` is checking which observations are `Solitary`

In R

- Check each of these objects

```
control
```

```
##      Group Freq
## 1  Control 10.7
## 2  Control 10.7
## 3  Control 10.4
## 4  Control 10.9
## 5  Control 10.5
## 6  Control 10.3
## 7  Control  9.6
## 8  Control 11.1
## 9  Control 11.2
## 10 Control 10.4
```

```
solitary
```

```
##        Group Freq
## 11 Solitary  9.6
## 12 Solitary 10.4
## 13 Solitary  9.7
## 14 Solitary 10.3
## 15 Solitary  9.2
## 16 Solitary  9.3
## 17 Solitary  9.9
## 18 Solitary  9.5
## 19 Solitary  9.0
## 20 Solitary 10.9
```

In R

- Each of the groups is a separate argument in `t.test`

```
out = t.test(control$Freq, solitary$Freq)
out

##
##  Welch Two Sample t-test
##
## data:  control$Freq and solitary$Freq
## t = 3, df = 17, p-value = 0.004
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.297 1.303
## sample estimates:
## mean of x mean of y
##     10.58      9.78
```

R output

- R calculates the degrees of freedom for us: $\nu = 16.875$
- R gives us the means

```r
out$estimate # gives the samples means of the two groups

## mean of x mean of y
##     10.58      9.78

out$estimate[1] - out$estimate[2] # find the diff in sample means

## mean of x
##       0.8
```

- When interpreting, we must be careful to not confuse the order
  - Mean of x corresponds to the first argument: controls
  - Mean of y corresponds to the second argument: solitary
  - Confidence interval is for $\mu_x - \mu_y$, or $\mu_{\text{control}} - \mu_{\text{solitary}}$

Confidence interval

- The confidence interval is

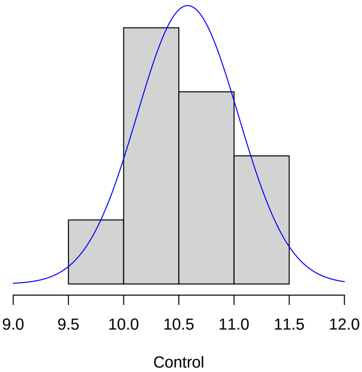
```
out$conf.int
```

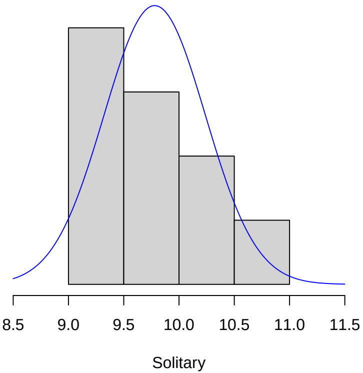
```
## [1] 0.297 1.303
## attr(,"conf.level")
## [1] 0.95
```

- We are 95% confident that the mean EEG frequency for the control group is between (0.297, 1.303) higher than those in solitary confinement
- The confidence interval has the same properties as before
  - In the long run, we would expect 95% of the confidence intervals we calculate to include the true difference $\mu_1 - \mu_2$
    - If we were to repeatedly sample from the population and repeat this analysis

Checking assumptions

- We are assuming a normal model for each group
- Check fitted model





Checking assumptions

- No major departures from normality
- Enough to make us cautious
  - Small sample size: normality assumption very important
    - It is hardest to assess normality assumptions, when it matters the most
- Want to be cautious in our conclusions

Hypothesis test

- This study was set up to look into a specific hypothesis
  - Confirmatory
- Theory was that sensory deprivation changes EEG frequency
- Null hypothesis: status quo / assumption of no difference
  - The two groups have the same mean: $\mu_1 = \mu_2$
  - $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$
- The alternative hypothesis
  - The two groups differ: $\mu_1 \neq \mu_2$
  - $\mathrm{H}_A : \mu_1 - \mu_2 \neq 0$

Hypothesis test

- The same function (`t.test`) is used to calculate a hypothesis test

```
out = t.test(control$Freq, solitary$Freq)
out
##
##  Welch Two Sample t-test
##
## data:  control$Freq and solitary$Freq
## t = 3, df = 17, p-value = 0.004
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.297 1.303
## sample estimates:
## mean of x mean of y
##     10.58      9.78
```

Interpretation

- The p-value is 0.004
  - Evidence of incompatibility between data and null hypothesis
  - Data provide support for the alternative hypothesis
    - Difference in EEG frequency between the control and solitary groups
- Given the small sample and cautiousness in checking assumptions
  - We have provided evidence in support of EEG differing
  - Larger studies desirable to provide further confirmation

Confidence intervals vs hypothesis testing

- In this example we look at both confidence intervals and hypothesis test
- The p-value does not tell us how strong an effect is
  - We could have p-value of 0.05 with $\bar{y}_1 - \bar{y}_2 = 10$
    - Small sample size
  - We could have p-value of 0.001 with $\bar{y}_1 - \bar{y}_2 = 0.002$
    - Large sample size
- Confidence interval gives an interval estimate of effect

Independent groups

- We have assumed the two groups are independent
  - Important assumption
- What does that mean?
  - The outcome from one group does not affect the outcome from the other group
- This will not always be the case:
  - Students take a test before undertaking a course
  - Same students undertake the same test after the course
    - Same participants in each 'group'
    - It is likely that someone who scored well in first test will also score well in the second test
- Look into this more tomorrow

Summary

- First look at relationship between variables
  - How EEG frequency varies by sensory deprivation
- Relationship between a continuous variable and a categorical variable
  - EEG frequency (continuous); sensory deprivation yes/no (categorical)

Outline

- Previous:
  - Started to look at relationships between variables
    - Frequency of brain waves (EEG) and sensory deprivation
  - Examples of relationship between one continuous and one categorical variable
    - Two groups are independent
- Today:
  - Look at paired data (two groups are not independent)
  - Start looking at relationships between two continuous variables

Motivating example

- Reaction time (ms) for 23 participants (press a button after stimulus)
  - University students
- There are two stimuli:
  - Auditory (a burst of white noise)
  - Visual (a circle flashing on a computer screen)
- Each participant exposed to both stimuli
  - Shouldn't use the approach from previous lecture
  - The two groups are not independent
    - We might expect someone with fast reaction time (auditory) to have a fast reaction (visual)
- Example of paired data
  - Each observation in group one has correspondence to an observation in group two
- This is an exploratory study

Data

```r
AV = read.csv('AV.csv')
head(AV)
```

```
##   auditory visual
## 1      226    256
## 2      188    309
## 3      280    364
## 4      234    379
## 5      181    268
## 6      178    288
```

Paired: find the differenceback to the future

- Look at the difference in the outcomes for each pair

```
AV$differ = AV$visual - AV$auditory
# this adds another variable (called differ) to the data frame AV
head(AV)
```

```
##   auditory visual differ
## 1      226    256   29.3
## 2      188    309  121.9
## 3      280    364   83.7
## 4      234    379  144.8
## 5      181    268   87.1
## 6      178    288  109.9
```

Paired: back to the future

- Model the differences as if they were a single sample
  - The data are the differences and are given by y_d
  - The differences y_d are assumed to be normal with mean μ_d and variance σ_d^2
  - μ_d is a parameter representing the mean difference in the population
- For our example:
  - y_d is the difference in reaction time (visual - auditory)
  - μ_d is the population mean difference in reaction time (visual - auditory)

In R

- For paired data: two ways to find confidence intervals and hypothesis tests in R
- Option 1: use `t.test` on the differenced values

```
t.test(AV$differ)

##
##  One Sample t-test
##
## data:  AV$differ
## t = 4, df = 22, p-value = 2e-04
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  32.3 87.9
## sample estimates:
## mean of x
##      60.1
```

In R

- For paired data: two ways to find confidence intervals and hypothesis tests in R
- Option 2: specify the two groups and include option `paired = TRUE`

```
t.test(AV$visual, AV$auditory, paired = TRUE)

##
##  Paired t-test
##
## data:  AV$visual and AV$auditory
## t = 4, df = 22, p-value = 2e-04
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  32.3 87.9
## sample estimates:
## mean difference
##            60.1
```

Output and interpretation

- Both approaches give identical confidence intervals
- Minor differences
  - Input differs: (1) input the differences; (2) input each group
  - Wording differences in output
    - 'One sample t-test' vs 'Paired t-test'
    - 'true mean' vs 'true mean difference'
    - 'mean of x' vs 'mean difference'
- Interpretation:
  - We are 95% confident that mean difference in the reaction times between visual and auditory stimuli is between (32.3, 87.9) ms

Hypothesis test

- Often with an exploratory study: use confidence interval
  - Calculate hypothesis test here as an example
- The hypothesis test is in terms of μ_d
- Null hypothesis: assumption of no difference ($\mu_d = 0$)
  - $\text{H}_0 : \mu_d = 0$
  - $\text{H}_A : \mu_d \neq 0$
- The p-value is 1.85×10^{-4}
  - Evidence that data are incompatible with the null hypothesis
  - There is evidence (at the $\alpha = 0.05$ level) that the data are incompatible with assumption of no difference

Extension

- Many applications may have more than two groups
  - Data from multiple independent groups
  - Multiple observations of each subject (repeated measures)
- There are statistical models for both cases
  - Independence: ANOVA (analysis of variance)
    - We will see this later in the course
  - Repeated measures: complex model
    - Outside the scope of this course

Relationship between continuous variables

- Previous examples: relationship between a continuous variable and a categorical variable
  - Continuous: reaction time; categorical: stimuli
  - Continuous: EEG frequency; categorical: sensory status (solitary/control)
- We are now going to consider relationships between two continuous variables

Motivating examples

- We are going to introduce three motivating examples
  1. The size of brushtail possums
     - Compare total length (mm) to head length (cm)
     - $n = 104$ observations
  2. Height of STAT 110 students
     - Compare father's height (cm) to son's height (cm)
     - $n = 279$ observations
  3. Squat weight of international power lifters
     - Comparing body weight (kg) to max squat weight (kg)
     - Photo from powerliftingtechnique.com
     - The athlete pictured (Kelly Branton) is in the dataset
     - $n = 9045$ observations (athletes)
- All of these involve two continuous variables

Brushtail possums

- Import the data

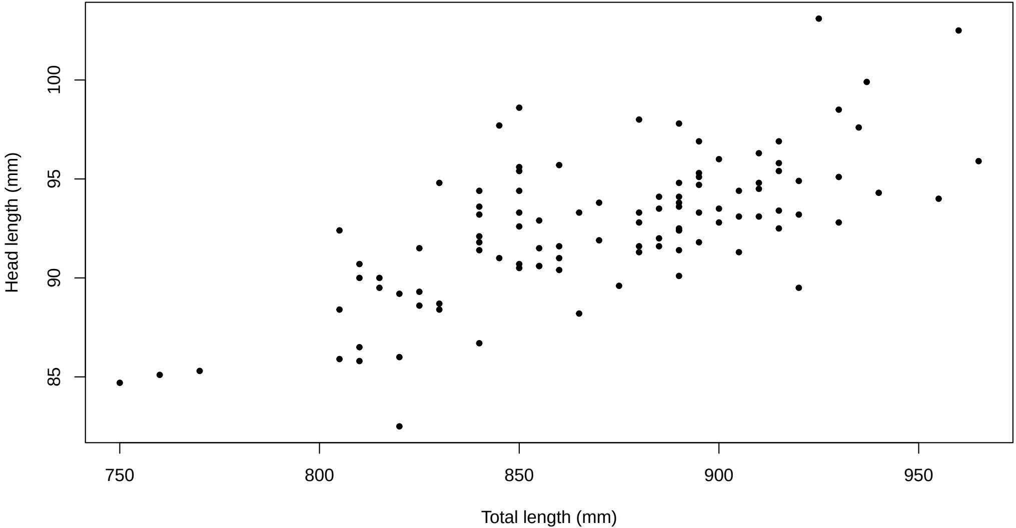
```
possum = read.csv('possum.csv')
```

- Have a look at the data:

```
head(possum)

##   total_l head_l
## 1     890   94.1
## 2     915   92.5
## 3     955   94.0
## 4     920   93.2
## 5     855   91.5
## 6     905   93.1
```

Brushtail possums: scatterplot



Father & son height

- Import the data

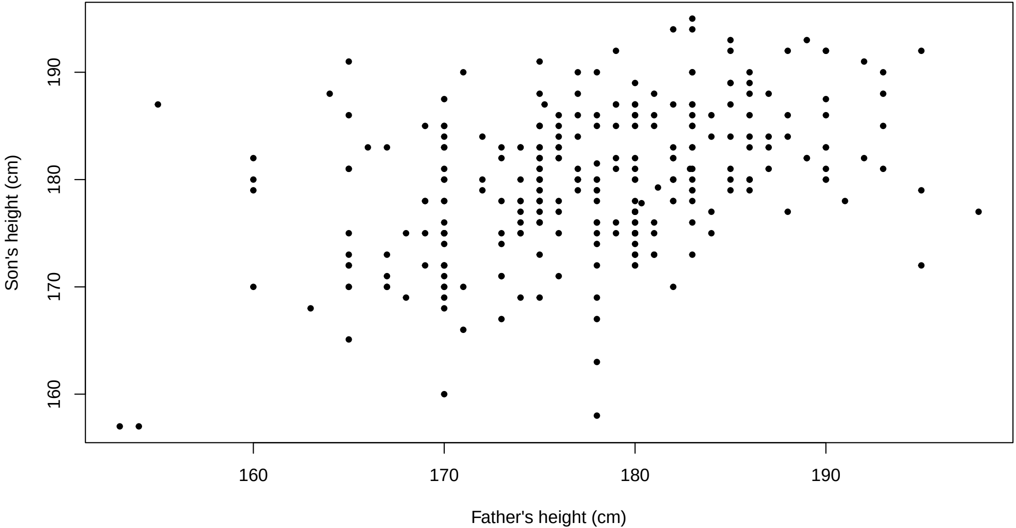
```r
height = read.csv('height.csv')
```

- Have a look at the data:

```r
head(height)
```

```
##   son father
## 1 176    178
## 2 180    190
## 3 180    174
## 4 181    179
## 5 184    187
## 6 180    182
```

Father & son height: scatterplot



Powerlifting

- Import the data

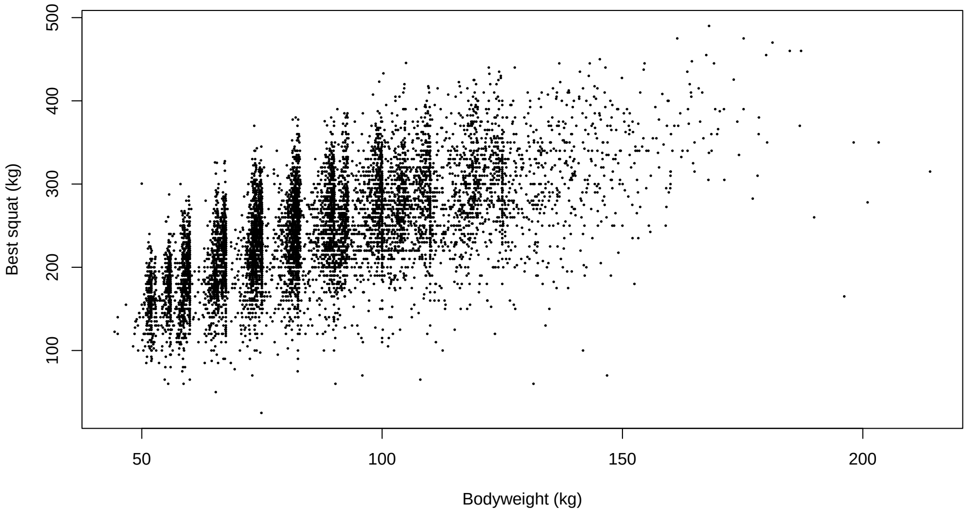
```
powerlift = read.csv('powerlift.csv')
```

- Have a look at the data:

```
head(powerlift)
```

```
##   bodyweight bestsquat
## 1       59.6       228
## 2       67.2       255
## 3       67.4       270
## 4       59.9       260
## 5       59.9       250
## 6       56.0       210
```

Powerlift: scatterplot



Back to the beginning

- What was the first thing we did when we first encountered data in STAT 110?
  - Found data summaries: sample mean and sample variance
- What summary describes the relationship between two continuous variables?

Correlation

- Correlation describes the strength of a linear relationship between two variables (let's call them x and y)
  - Always takes a value between -1 and 1
  - Population correlation represented by ρ (greek letter rho)
  - Sample correlation represented by r
- With data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the correlation is given by

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

- We will calculate the correlation using the R function `cor`

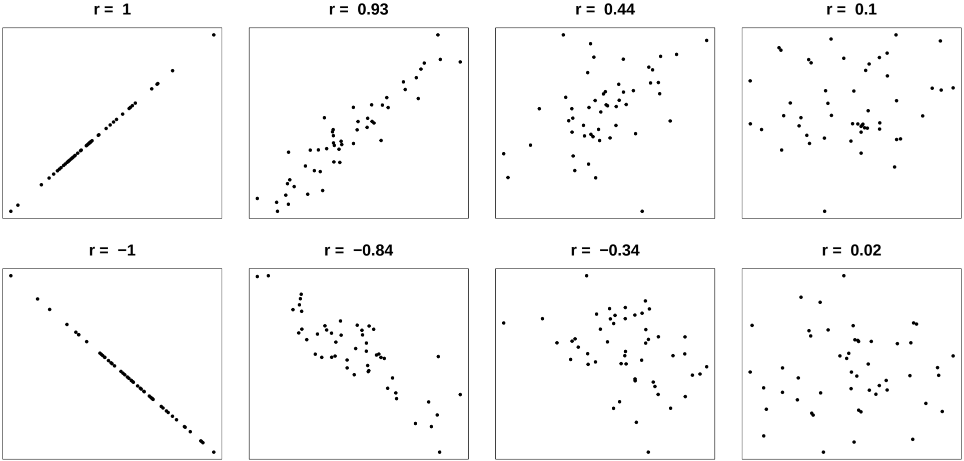
```
cor(possum$total_l, possum$head_l)

## [1] 0.691
```

Understanding correlation

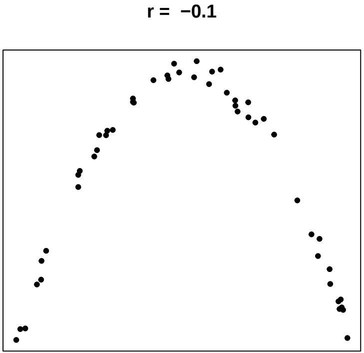
- Positive correlation:
  - If y is above its mean, then x is likely to be above it's mean (and vice versa)
- Negative correlation
  - If y is above its mean, then x is likely to be below it's mean (and vice versa)
- If the relationship is strong and positive
  - r will be close to 1
- If the relationship is strong and negative
  - r will be close to -1
- If there is no apparent (linear) relationship between x and y
  - r will be close to 0

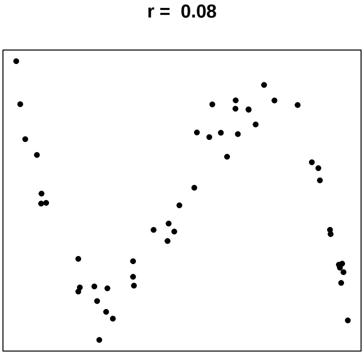
Understanding correlation: graphically I



Understanding correlation: graphically II

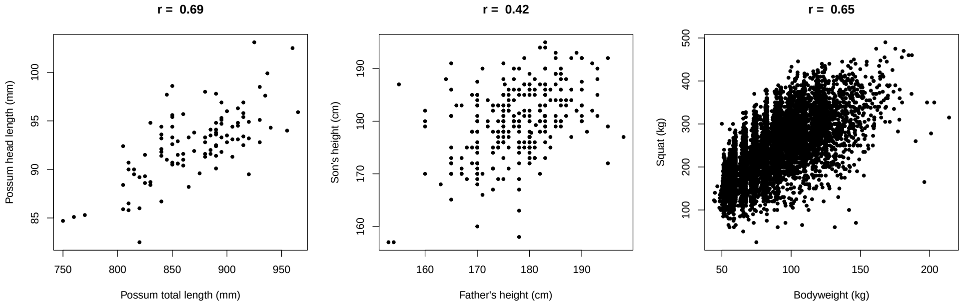
- r measures the strength of the linear relationship
  - Strong non-linear relationships can produce r values that do not reflect the strength of the relationship





Data

```
rposs = cor(possum$total_l, possum$head_l)
rheight = cor(height$son, height$father)
rpower = cor(powerlift$bodyweight, powerlift$bestsquat)
```



Practice

- Guess the correlation

Limitations

- The correlation r is a useful summary
  - We may want to learn how precise it is: confidence interval
  - Such intervals can be found: `cor.test` in R
    - We will not consider them in STAT 110
- The correlation as a summary is limited
- What might we want to know?

1. Possum data: predict head length from a measurement of total length
2. Height data: understanding and quantifying heritability of height as a trait
3. Powerlifting: compare the squat weight of an athlete to their peers of a similar weight

- Correlation does not help us for 1 and 3
  - Limited for 2: quantifies the linear relationship, but does not describe it
    - What is the expected difference in height between a son with father who is 170 cm tall, and a son with father who is 180 cm tall?

Summary

- Looked at paired data
  - Model the difference between the two groups
  - Confidence intervals
  - Hypothesis test
- Looked at relationships between two continuous variables
- Explored a data summary: correlation
  - Gives the strength of a linear relationship between two variables
  - Always between -1 and $+1$
  - Easy to calculate in R

Outline

- Continue to explore relationships between two variables
- Go beyond summary statistics
  - Look into a statistical model for the relationship
    - What the model looks like
    - Fitted model
    - Residuals

Recall: motivating examples

- The size of brushtail possums
  - Compare total length (mm) to head length (cm)
- Height of STAT 110 students
  - Compare father's height (cm) to son's height (cm)
- Squat weight of international power lifters
  - Comparing body weight (kg) to max squat weight (kg)

Recall: correlation

- The correlation r measures the strength of linear relationship between two variables x and y
- The correlation is limited
- What might we want to know?
  1. Possum data: predict head length from a measurement of total length
  2. Height data: understanding and quantifying heritability of height as a trait
  3. Powerlifting: compare the squat weight of an athlete to their peers of a similar weight
- Correlation does not help us for 1 and 3
  - Limited for 2: quantifies the linear relationship, but does not describe it
    - What is the expected difference in height between a son with father who is 170 cm tall, and a son with father who is 180 cm tall?

Statistical model

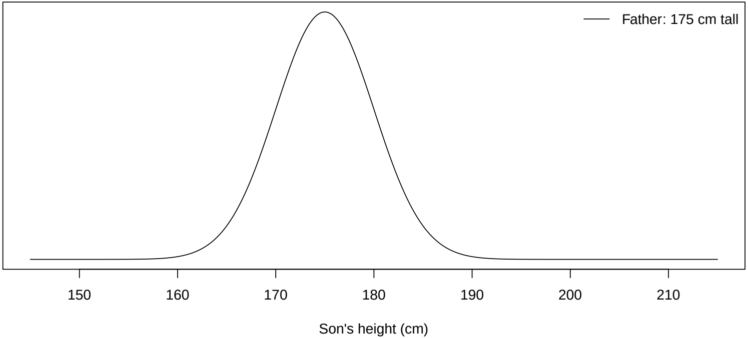
- To overcome these problems we will look to a statistical model
  - Extension of our previous models
- Explore relationship between continuous variables x and y
  - e.g. x is father's height, y is son's height
- The variable y is referred to as the outcome variable
  - Can also be called the response variable, or dependent variable
- The variable x is referred to as the predictor variable
  - Can also be called the explanatory variable, or independent variable
- The idea: the predictor variable helps us 'predict' the outcome variable

Statistical model

- Our description will make use of the father/son height example
  - Interest is in understanding the relationship the height of NZ male university students and their fathers
  - Sample is from (former) students in STAT 110
- Using probability to describe data
- Recall concept of conditional probability: $Pr(A|B)$
  - Here we are looking at a probability density for $y|x$
    - We have the height of a father (x) and son (y)
    - Given a father's height (x), we specify a model for son's height (y)
    - We will specify a normal model
- Look at it graphically

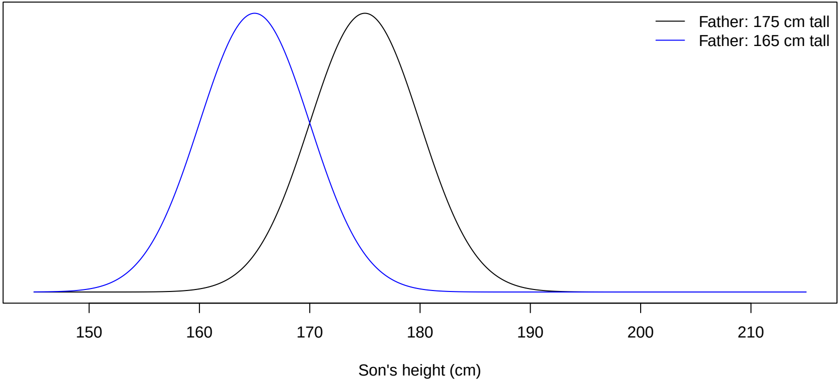
Statistical model

- Consider the subpopulation at particular value of x
  - e.g. sons with fathers who are 175 cm tall ($x = 175$)
  - Assume that son's height is normally distribution
    - For the sake of explanation: sons are expected to be the same height as their fathers



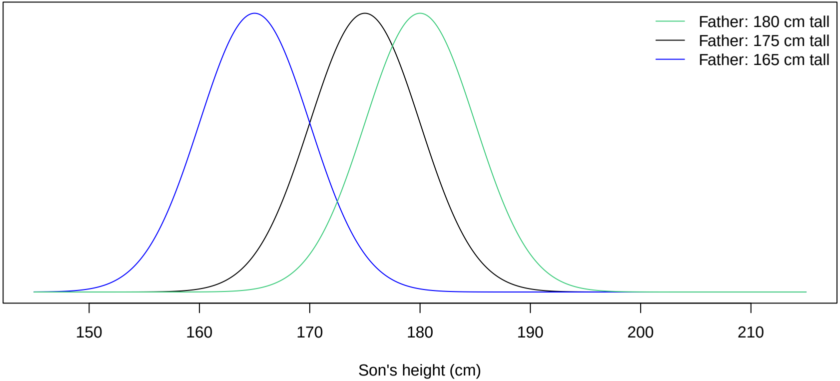
Statistical model

- Subpopulation at a given value of x: outcome variable is normally distributed
- For fathers who are 165 cm tall (blue)



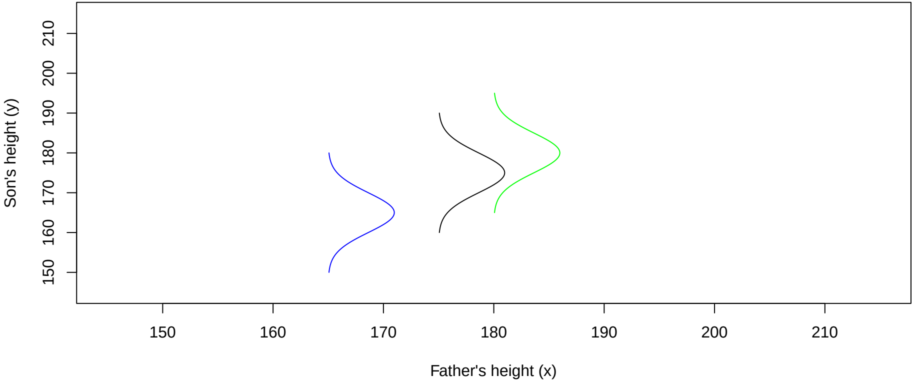
Statistical model

- Subpopulation at a given value of x: outcome variable is normally distributed
- For fathers who are 180 cm tall (green)



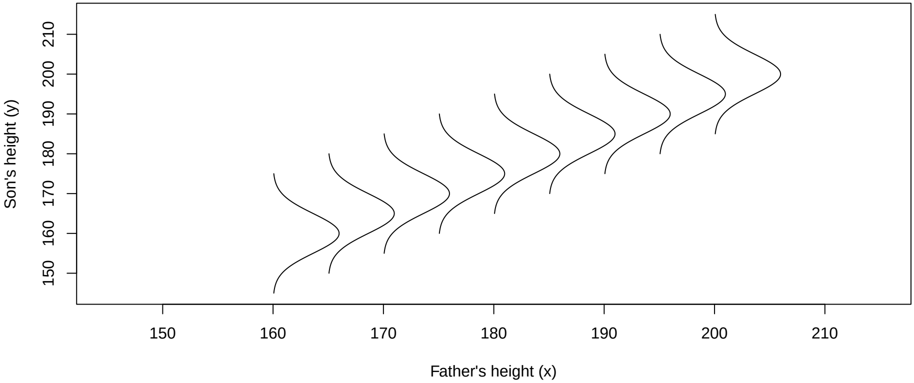
Turning it sideways

- Visualise it with outcome variable on y-axis, and predictor variable on x-axis
  - The same distributions are given below



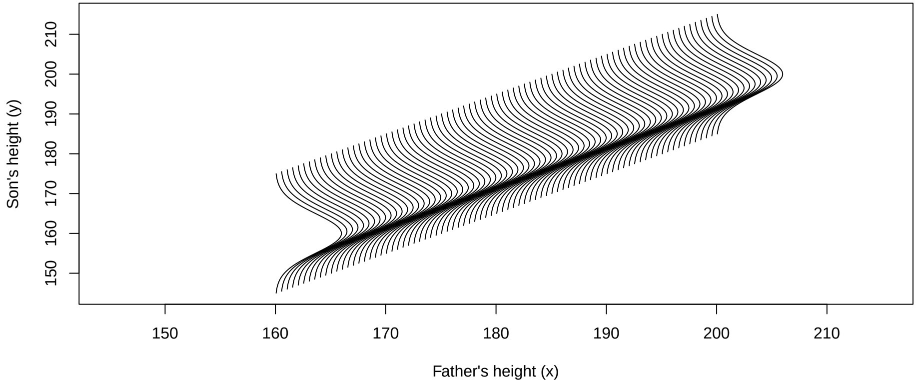
Turning it sideways

- Including some other values of x (father's height)



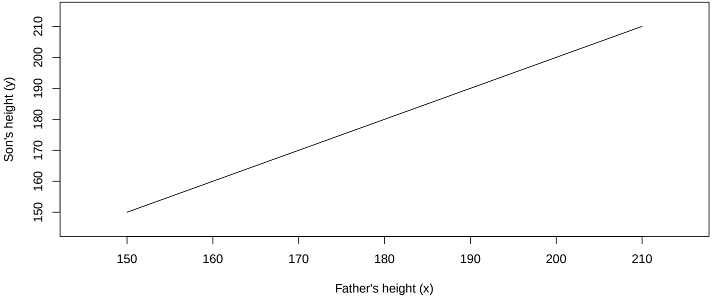
Turning it sideways

- Including even more values of x (father's height)



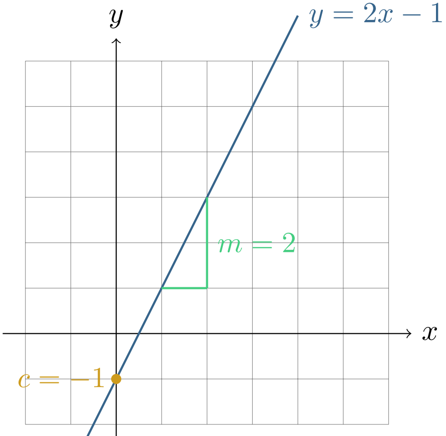
Linear regression

- The outcome variable, y, can be written in terms of two pieces:
  - outcome $=$ mean response $+$ error
- The mean response (what we expect) is assumed to vary with the predictor x
  - Expected height of a son is different if father is 165 cm vs father who is 180 cm
- We assume the mean response is a straight line
  - e.g. continuing the father and son height example, the mean response is



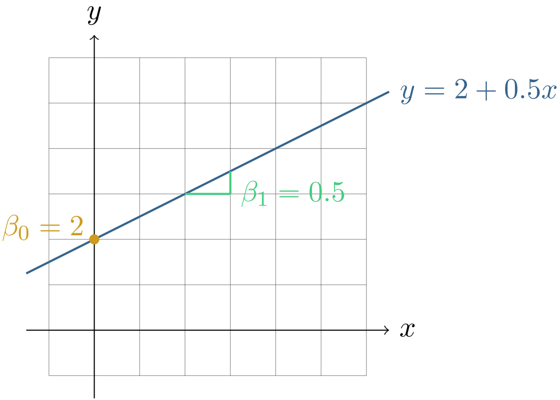
Revision: equation for a straight line

- Mathematical equation: $y = mx + c$
  - Intercept c: where it crosses the y-axis $(x = 0)$
  - Slope m



Revision: equation for a straight line

- We will use the equation: $\beta_0 + \beta_1 x$
  - Convention: use β_0 and β_1 in place of c and m
    - Intercept β_0: where it crosses the y-axis ($x = 0$)
    - Slope β_1



Understanding the model: population level

- Putting this together we have:

$$\underbrace{y}_{\text{outcome}} = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \underbrace{\varepsilon}_{\text{error}}$$

- The mean response is given by the straight line: $\mu_y = \beta_0 + \beta_1 x$
  - Gives us the expected value of y in the population for a given value of x
- The mean will be different for two different values of x
- For $x = 165$ cm:
  - Mean is: $\mu_y = \beta_0 + \beta_1 \times 165$
- For $x = 180$ cm:
  - Mean is: $\mu_y = \beta_0 + \beta_1 \times 180$

Interpretation

- What do β_0 and β_1 represent?
- The mean will be different for two different values of x
  - Mean is: $\mu_y = \beta_0 + \beta_1 x$
- For someone with a father one cm taller $(x + 1)$, the mean response is
  - Mean is: $\mu_y = \beta_0 + \beta_1(x + 1) = \beta_0 + \beta_1 x + \beta_1$
- β_1 is the difference between these
  - β_1 is the change in mean response when x increases by one unit
    - Change in the expected height of two male NZ university students whose fathers differ in height by 1 cm
- β_0 is the mean response when $x = 0$
  - May make no sense in many examples
    - Mean response for a son with a father of height 0 cm: physically impossible

From mean response to individual response

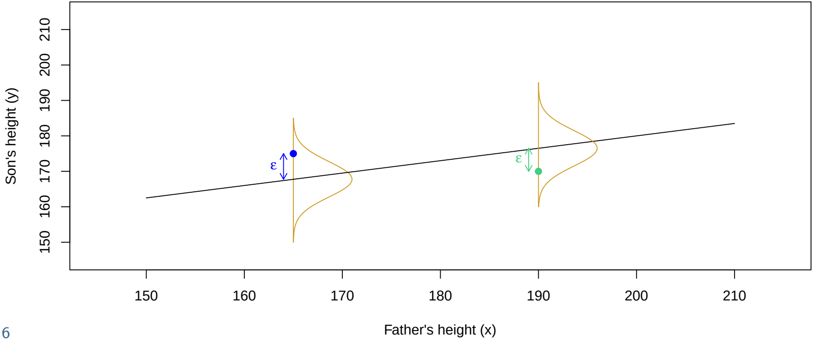
- The linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Error term ε (greek letter epsilon) describes how an individual response differs from the mean of their subpopulation
  - Subpopulation: all individuals in the population with the same value of x
- We assume that variation within a given subpopulation is normally distributed
  - ε is normally distributed with mean 0 and variance σ_ε^2
    - σ_ε tells us how variable individual observations are within their subpopulation

Visualising subpopulation

- Suppose that the true regression model for height is $y = 110 + 0.35x + \varepsilon$
  - Mean response (black line)
  - Normal model for the errors (gold)
  - Individual with $y = 175$ and $x = 165$ (blue point)
  - Individual with $y = 170$ and $x = 190$ (green point)

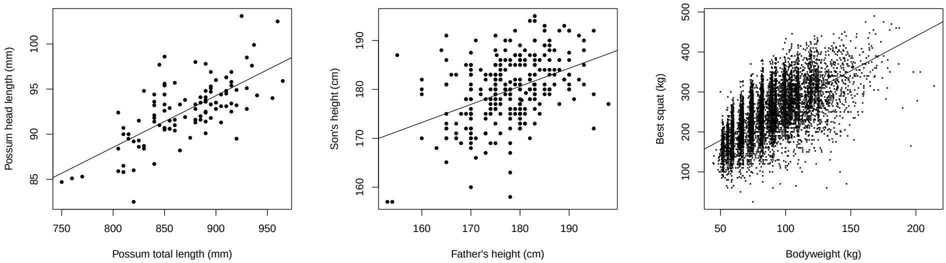


Statistical model: data

- The linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- The errors mean that data will not fall exactly on the line
  - Like the data we have!



It's quiz time!

- Suppose that the true regression model for height is

$$y = 110 + 0.35x + \varepsilon$$

- Decide whether the following statements are true or false:

1. Consider the subpopulation of all students with fathers of height $x = 200$ cm. The mean height of those students is 180 cm.
2. On average, students with fathers of height $x = 201$ cm are 0.35 cm taller than students with fathers of height $x = 200$cm.
3. All students with fathers of height $x = 190$ cm are taller than all students with fathers of height $x = 170$ cm.
4. Students with fathers of height $x = 0$ cm are 110 cm tall on average

Summary

- Introduced a statistical model for the relationship between x and y
  - Outcome variable, y
  - Predictor variable, x
  - For a given value of x, y is assumed to be normally distributed
- Understand the linear regression model
  - Mean response
  - Error
  - Interpretation
- Looking forward: how do we fit a linear regression to data?