STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 18: Paired Data

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 auditory reaction time to have a fast visual reaction
- 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    226.3  255.5
## 2    187.5  309.4
## 3    279.8  363.5
## 4    233.8  378.7
## 5    180.8  268.0
## 6    178.2  288.1
```

Paired: find the difference

- 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    226.3  255.5  29.26
## 2    187.5  309.4 121.91
## 3    279.8  363.5  83.73
## 4    233.8  378.7 144.83
## 5    180.8  268.0  87.14
## 6    178.2  288.1 109.87
```

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 = 4.5, df = 22, p-value = 2e-04
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  32.29 87.86
## sample estimates:
## mean of x
##     60.08
```

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 = 4.5, df = 22, p-value = 2e-04
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  32.29 87.86
## sample estimates:
## mean difference
##           60.08
```

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 1.8498×10^{-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

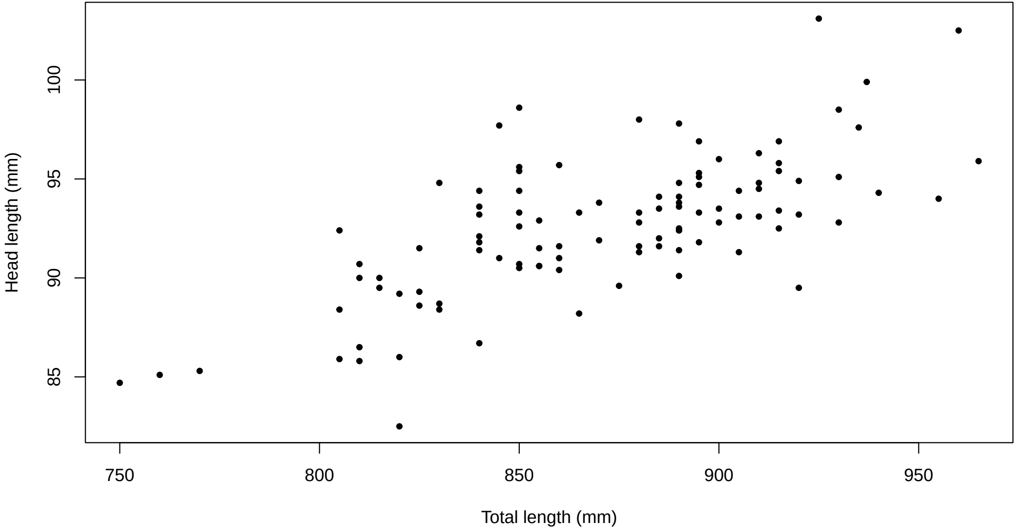
```r
possum = read.csv('possum.csv')
```

- Have a look at the data:

```r
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

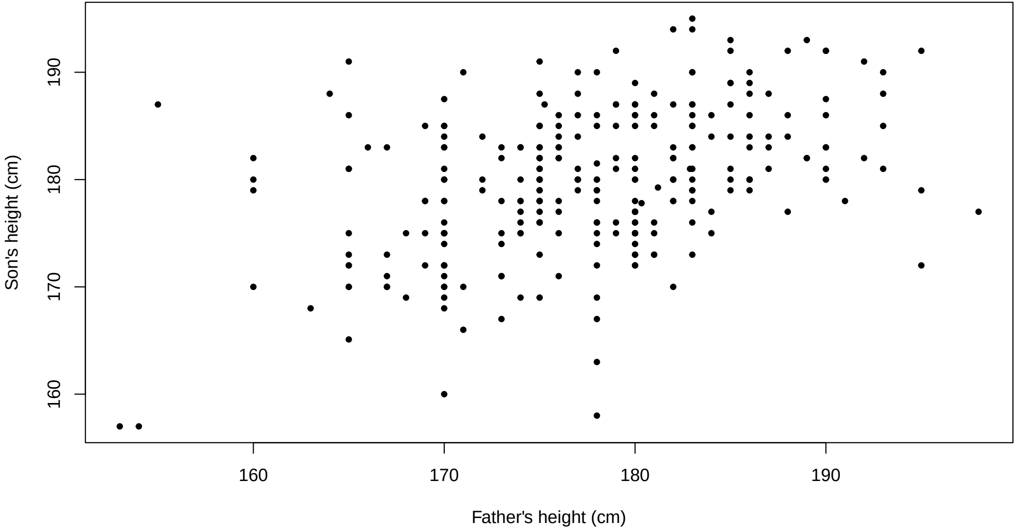
```
height = read.csv('height.csv')
```

- Have a look at the data:

```
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

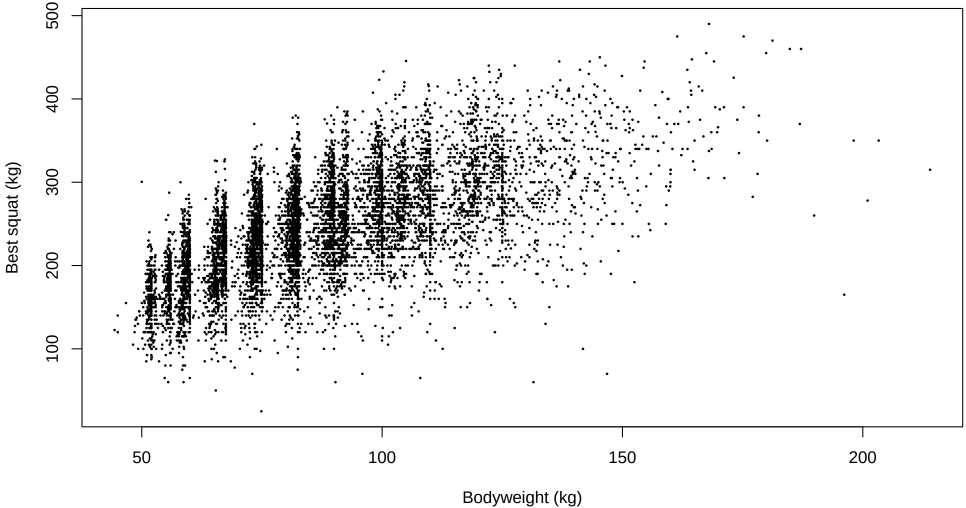
```
powerlift = read.csv('powerlift.csv')
```

- Have a look at the data:

```
head(powerlift)

##   bodyweight bestsquat
## 1       59.6     227.5
## 2       67.2     255.0
## 3       67.4     270.0
## 4       59.9     260.0
## 5       59.9     250.0
## 6       56.0     210.0
```

Powerlift: scatterplot



Back to the beginning

- What was the first thing we did when we first encountered data in STAT115?
  - 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`

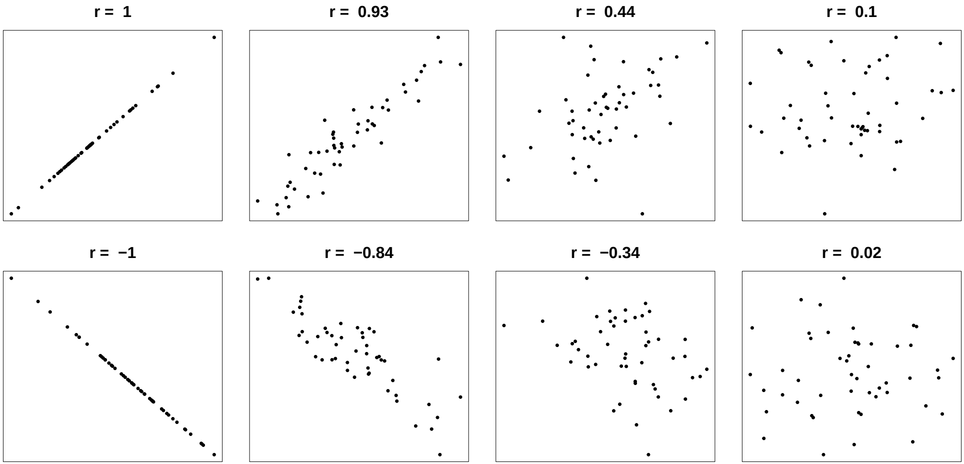
```
cor(possum$total_l, possum$head_l)

## [1] 0.6911
```

Understanding correlation

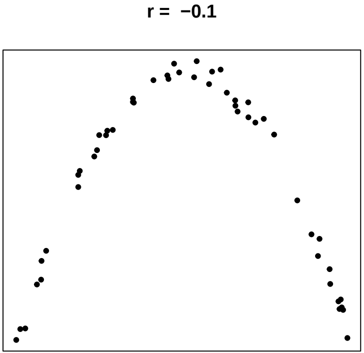
- Positive correlation:
  - If y is above its mean, then x is likely to be above it's mean (and vice versa)
- Negative correlation
  - If y is above its mean, then x is likely to be below it's mean (and vice versa)
- If the relationship is strong and positive
  - r will be close to 1
- If the relationship 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

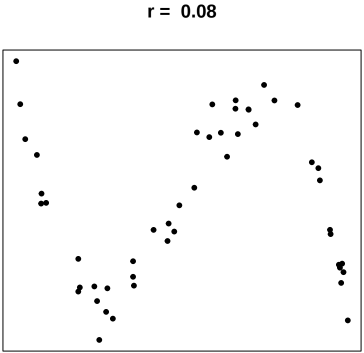
Understanding correlation: graphically I



Understanding correlation: graphically II

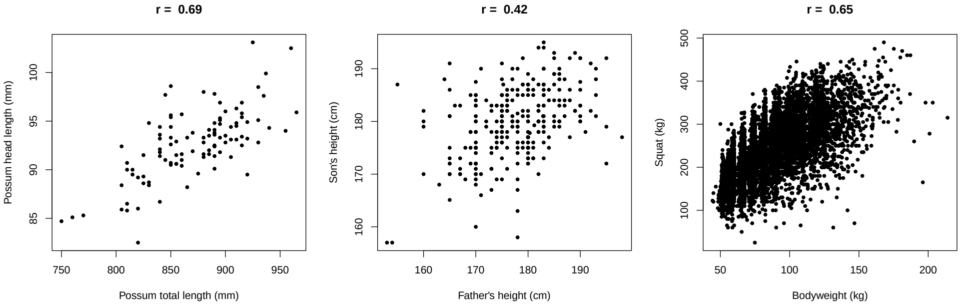
- r measures the strength of the linear relationship
  - Strong non-linear relationships can produce r values that do not reflect the strength of the relationship





Data

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



Test yourself at `https://www.guessthecorrelation.com/`

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 STAT115
- 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