DISCOVERING STATISTICS USING R

Correlation

1

DISCOVERING STATISTICS USING R

Aims and Objectives

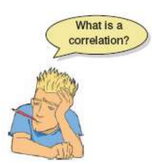
- Measuring relationships
  - Scatterplots
  - Covariance
  - Pearson's correlation coefficient
- Nonparametric measures
  - Spearman's rho
  - Kendall's tau
- Interpreting correlations
  - Causality
- Partial correlations

2

DISCOVERING STATISTICS USING R

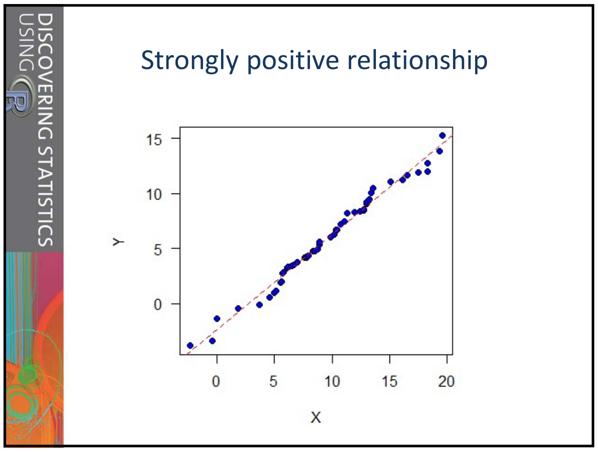
What is a correlation?

- A way of measuring the extent to which two variables are related
- It measures the pattern of responses across variables



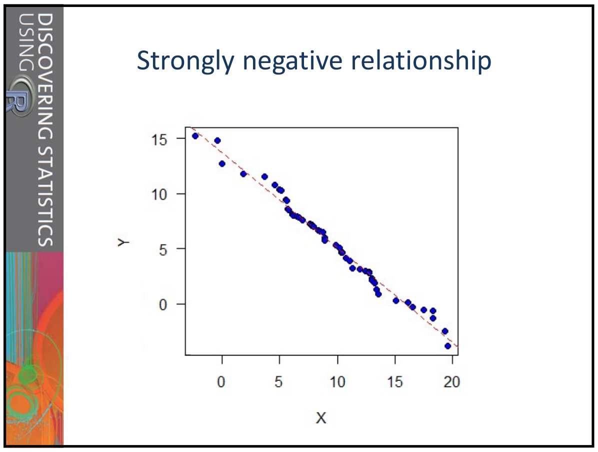
3

Strongly positive relationship



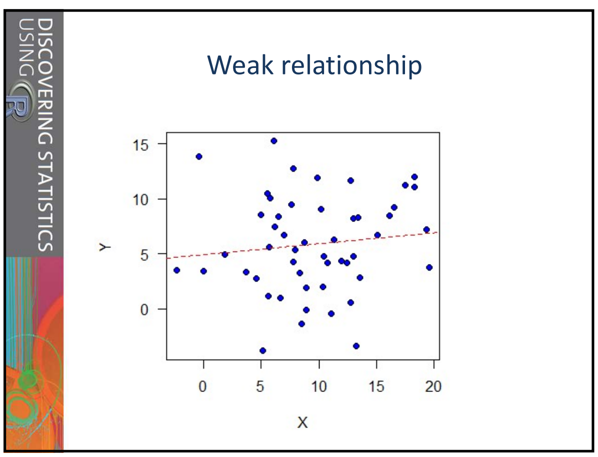
4

Strongly negative relationship



5

Weak relationship



6

DISCOVERING STATISTICS USING R

Measuring relationships

- We want to see whether as one variable increases, the other increases, decreases or stays the same
- This can be done by calculating the covariance
  - We look at how much each score deviates from the mean
  - If both variables deviate from the mean by the same amount, they are likely to be related

7

DISCOVERING STATISTICS USING R

Revision of variance

- The variance tells us by how much scores deviate from the mean for a single variable
- It is closely linked to the Sum of Squares (*SS*)
- Covariance is similar:
  - Tells us by how much scores on two variables differ from their respective means

8

DISCOVERING STATISTICS USING R

Revision of variance

- The variance tells us by how much scores deviate from the mean for a single variable
- It is closely linked to the Sum of Squares (SS)

$$\text{variance} = \frac{\Sigma(x_i - \bar{x})^2}{N-1}$$
$$= \frac{\Sigma(x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

9

DISCOVERING STATISTICS USING R

Covariance

- Calculate the error between the mean and each subject's score for the first variable (*x*)
- Calculate the error between the mean and each subject's score for the second variable (*y*)
- Multiply these error values
- Add to get the cross product deviation
- Covariance is the average cross-product deviation

$$\mathrm{cov}(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

10

DISCOVERING STATISTICS USING R

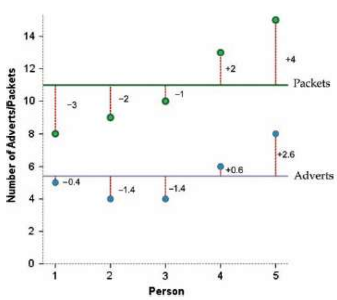
Measuring relationships

Participant:	1	2	3	4	5	Mean	S
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

11

DISCOVERING STATISTICS USING R

Covariance



12

Covariance

$$\begin{aligned}\mathrm{cov}(x,y) &= \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N-1}\\ &= \frac{(-0.4)(-3)+(-1.4)(-2)+(-1.4)(-1)+(0.6)(2)+(2.6)(4)}{4}\\ &= \frac{1.2+2.8+1.4+1.2+10.4}{4}\\ &= \frac{17}{4}\\ &= 4.25\end{aligned}$$

13

Problems with covariance

- It depends upon the units of measurement
  - *E.g.* the covariance of two variables measured in miles might be 4.25, but if the variable was expressed in kilometres, the covariance would be 11
- Need to standardize it
  - Divide by the standard deviations of both variables
- The standardized version of covariance is known as the correlation coefficient
  - It is (relatively) unaffected by units of measurement

14

The correlation coefficient

$$\begin{aligned}r &= \frac{\mathrm{cov}_{xy}}{s_x s_y}\\ &= \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{(N-1)s_x s_y}\\ &= \frac{4.25}{1.67\times 2.92}\\ &= .87\end{aligned}$$

15

DISCOVERING STATISTICS USING R

The correlation coefficient

- Varies between -1 and +1
  - 0= no relationship
- Is an effect size
  - ±.1= small effect
  - ±.3= medium effect
  - ±.5= large effect
- Coefficient of determination, r^2
  - By squaring the value of *r* you get the proportion of variance in one variable shared by the other

16

DISCOVERING STATISTICS USING R

Example

- Anxiety and exam performance
- Participants:
  - 103 students
- Measures
  - Time spent revising (hours)
  - Exam performance (%)
  - Exam Anxiety (the EAQ, score out of 100)
  - Gender

17

DISCOVERING STATISTICS USING R

Example

- To compute basic correlation coefficients there are three main functions that can be used

  'cor()', *'cor.test()'* and *'rcorr()'*

Function	Pearson	Spearman	Kendall	*p*-values	CI	Multiple Correlations?	Comments
cor()	✓	✓	✓			✓	
cor.test()	✓	✓	✓	✓	✓		
rcorr()	✓	✓		✓		✓	2 d.p. only

18

DISCOVERING STATISTICS USING R

Example

- Pearson correlations:

```
> cor(examData, use= "complete.obs", method=
      "pearson")
> cor.test(examData$Exam, examData$Anxiety,
           method= "pearson")
> rcorr(as.matrix(examData2), type= "pearson")
```

- If we predicted a negative correlation:

```
> cor.test(examData$Exam, examData$Anxiety,
           alternative= "less"), method=
           "pearson")
```

19

DISCOVERING STATISTICS USING R

Example

- Output '*cor()*' function

	Exam	Anxiety	Revise
Exam	1.0000000	-0.4409934	0.3967207
Anxiety	-0.4409934	1.0000000	-0.7092493
Revise	0.3967207	-0.7092493	1.0000000

- Reporting results
  - Exam performance was significantly correlated with exam anxiety, $r = -.44$, and time spent revising, $r = .40$; the time spent revising was also correlated with exam anxiety, $r = -.71$ (all $ps < .001$).

20

DISCOVERING STATISTICS USING R

Correlation and causality

- The third-variable problem
  - In any correlation, causality between two variables cannot be assumed because there may be other measured or unmeasured variables affecting the results
- Direction of causality
  - Correlation coefficients say nothing about which variable causes the other to change

21

DISCOVERING STATISTICS USING R

Non-parametric correlation

- Spearman's rho
  - Pearson's correlation on the ranked data
- Kendall's tau
  - Better than Spearman's for small samples
- World's Biggest Liar competition
  - 68 contestants
  - Measures
    - End-ranking in the competition (first, second, third, etc.)
    - Creativity questionnaire (maximum score= 60)

22

DISCOVERING STATISTICS USING R

Bootstrapping correlations

- Bootstrapping
  - Useful resampling 'trick', especially when:
    - The theoretical distribution of a statistic is unknown or complicated
    - Sample size is too small to allow parametric inference
  - *Use 'boot()'* function…

```
> object.B<- boot(data, function, replications)
```

  - … in combination with *'boot.ci()'* function

```
> boot.ci(object.B)
```

23

DISCOVERING STATISTICS USING R

Bootstrapping correlations

- Example for Kendall's tau
  - First, write a function

```
> bootTau<- function(liarData, i){
                    cor(liarData$Position[i],
                        liarData$Creativity[i],
                        use= "complete.obs",
                        method= "kendall")
                    }
```

  - Next, plug this new function into the *'boot()'* function

```
> boot_kendall<- boot(liarData, bootTau, 9999)
```

  - Finally, check whether CI contains 0 or not

```
> boot.ci(boot_kendall, 0.99)
```

24

Point-biserial and biserial correlations

- Point-biserial correlation, r_{pb}
  - Quantifies the relationship between a continuous variable and a variable that is a discrete dichotomy (*e.g.* dead or alive)

```
> cor.test(variable 1, variable 2)
```

- Biserial correlation, r_b
  - Quantifies the relationship between a continuous variable and a variable that is a continuous dichotomy (*e.g.* acidic or alkaline)

```
> polyserial(variable 1, variable 2)
```

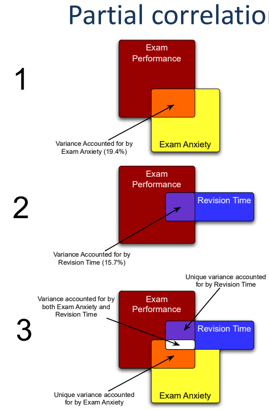
25

Partial and semi-partial correlations

- Partial correlation
  - Measures the relationship between two variables, controlling for the effect that a third variable has on them both
- Semi-partial correlation
  - Measures the relationship between two variables controlling for the effect that a third variable has on only one of the others

26

Partial correlation



27

Partial correlation

- Partial correlation '*pcor()*'

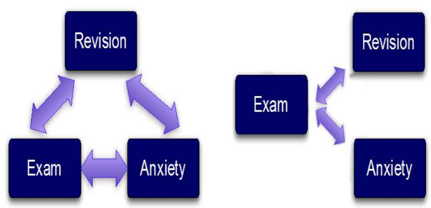
```
> pc.Exam<- pcor(c("Exam", "Anxiety", "Revise"),
              var(examData2))
```

- We can look at the partial correlation coefficient and it's *t* and *p*-value:

```
> pc.Exam
> pcor.test(pc.Exam, 1, 103)
```

28

Partial vs. semi-partial correlation



Partial Correlation

Look at unique relationship between two variables when other variables are ruled out

Semi-Partial Correlation

Explain the variance in one variable (outcome) from a set of predictor variables

29

Rest of morning & afternoon

- Practical Chapter 6
  - Read § 6.1, 6.2, 6.3 (skip 6.3.3 & 6.3.4), "Cramming Sam's Tips", and "What have I discovered about statistics?
  - Also read § 6.5.7 on bootstrapping and § 6.6.1 & 6.6.3 on part and partial correlation
  - Skip § 6.5.2 on R commander (Rcmdr) and § 6.7
  - Solve Smart Alex's tasks 1-3

30