Correlations

Data Analysis for Psychology in R 1

dapR1 Team

Department of Psychology
The University of Edinburgh

Weeks Learning Objectives

1. Understand how to calculate covariance and correlation.

2. Understand how to interpret the magnitude and direction of correlation coefficients.#

3. Understand which form of correlation to compute for different types of data.

Topics for today

- Recording 1: What is a correlation?
- Recording 2: Variance, covariance and correlation
- Recording 3: Pearson correlation
- Recording 4: Other forms of correlation

Purpose

- Correlations measure the degree of association between two variables.
  - If one goes up does the other go up (positive association)?
  - If one variable changes (varies) does the other change (vary) too.
  - If one goes up does the other go down (negative association)?
- The value ranges from -1 to 1.
  - Values close to $|1|$ indicate stronger associations.
  - Values close to 0 indicate no association.

Data Requirements

Variable 1	Variable 2	Correlation Type
Continuous	Continuous	Pearson
Continuous	Categorical	Polyserial
Continuous	Binary	Biserial
Categorical	Categorical	Polychoric
Binary	Binary	Tetrachoric
Rank	Rank	Spearman
Nominal	Nominal	Chi-square

- There is a form of correlation for almost all data types.

Scatterplots

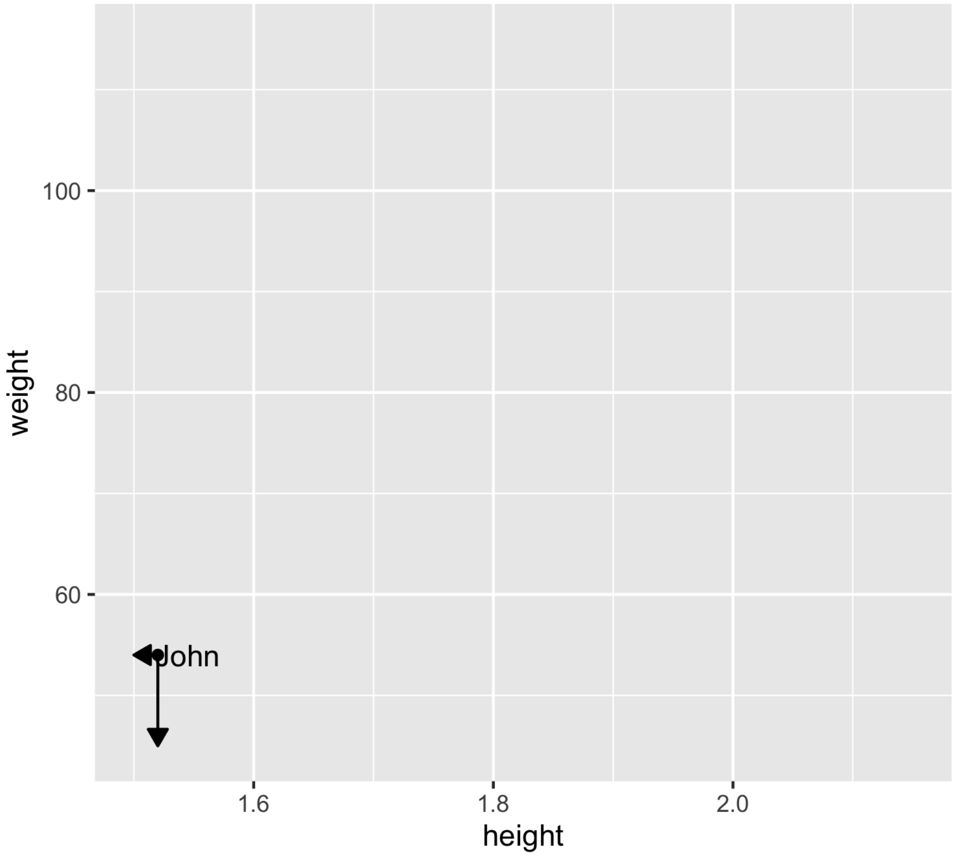
- Typical visualization of correlations is through scatterplots.
- Scatterplots plot points at the (x,y) co-ordinates for two measured variables.
- We plot these points for each individual in our data set.
  - This produces the clouds of points.

Simple Data

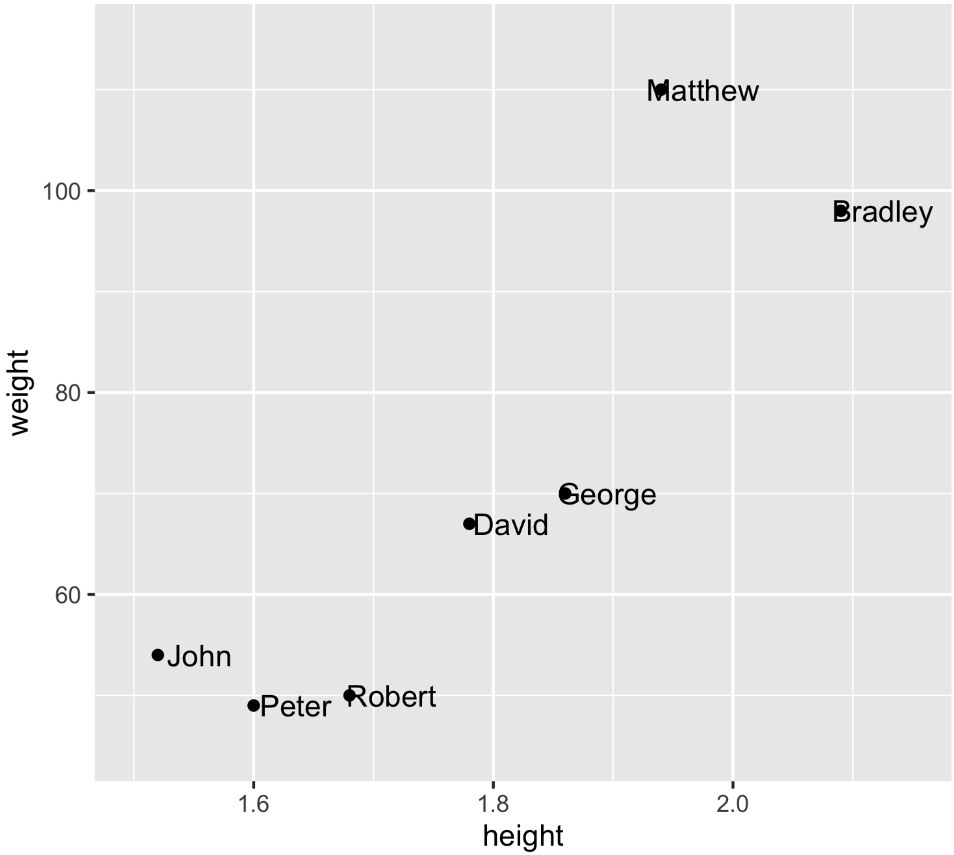
```r
data <- tibble(
  name = as_factor(c("John", "Peter","Robert","David","George","Matthew", "Bradley")),
  height = c(1.52,1.60,1.68,1.78,1.86,1.94,2.09),
  weight = c(54,49,50,67,70,110,98)
)
```

```
## # A tibble: 6 × 3
##   name     height weight
##   <fct>     <dbl>  <dbl>
## 1 John       1.52     54
## 2 Peter      1.6      49
## 3 Robert     1.68     50
## 4 David      1.78     67
## 5 George     1.86     70
## 6 Matthew    1.94    110
```

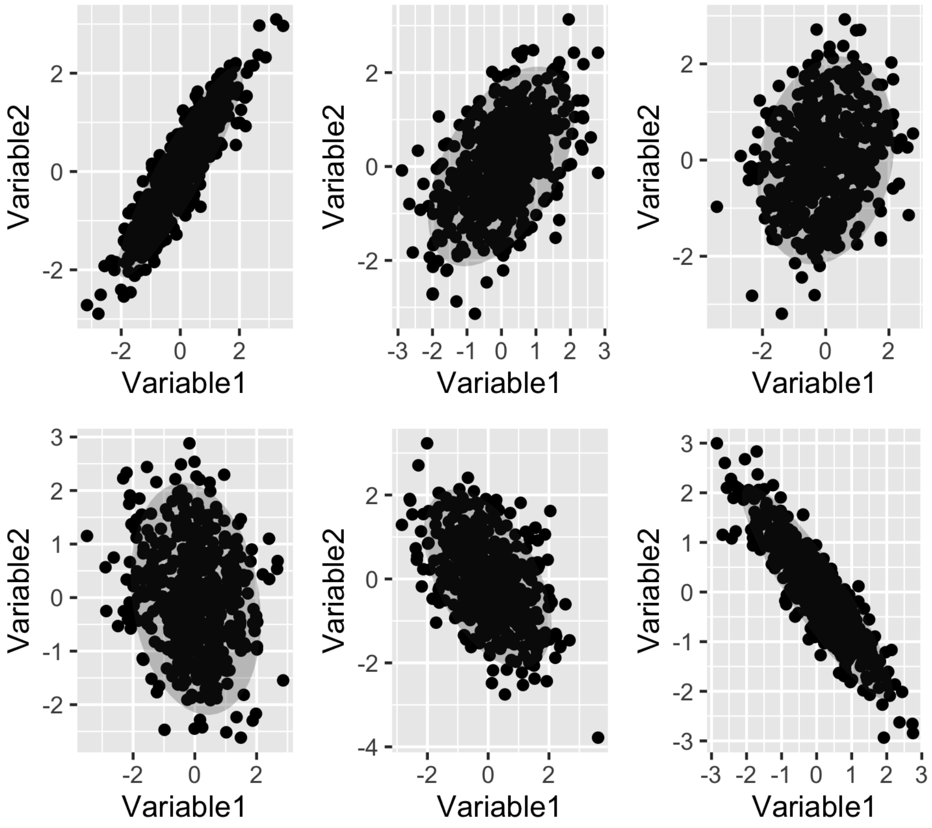
Scatterplot



Scatterplot



Strength of correlation



Time for a break

Welcome Back!

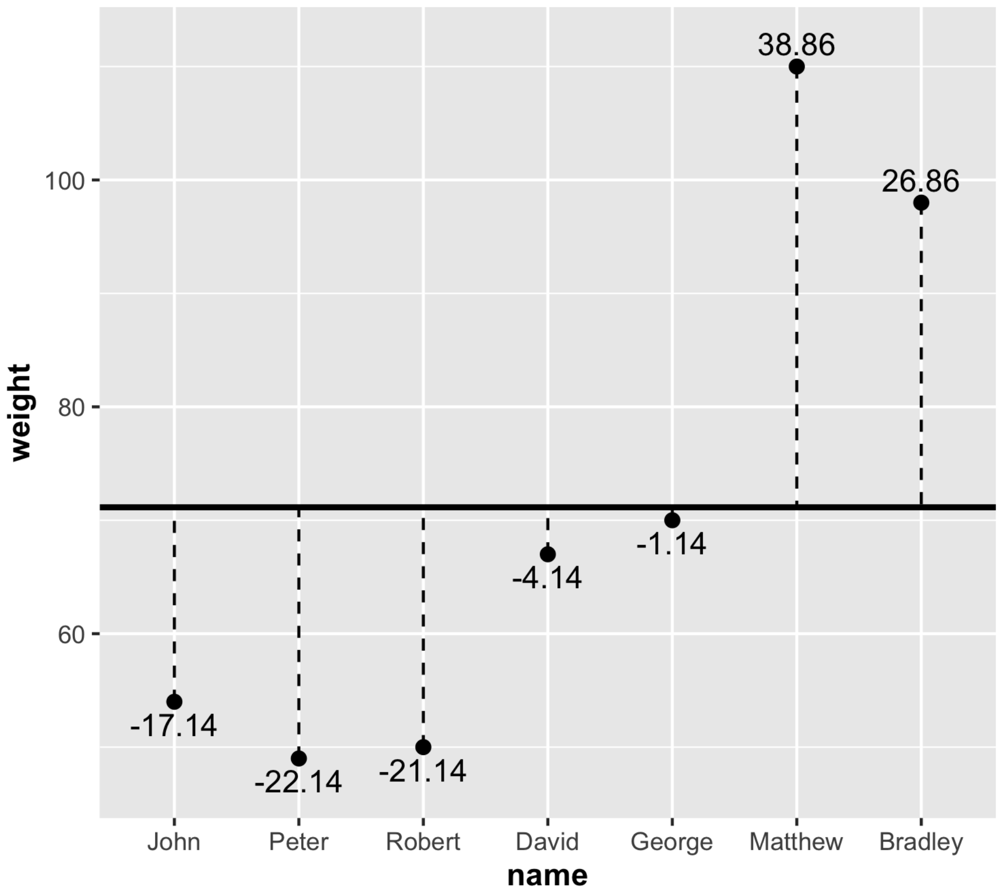
We have discussed what a correlation is and how to visualize it. Now let's move on to consider the relation to variance and covariance

Variance

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

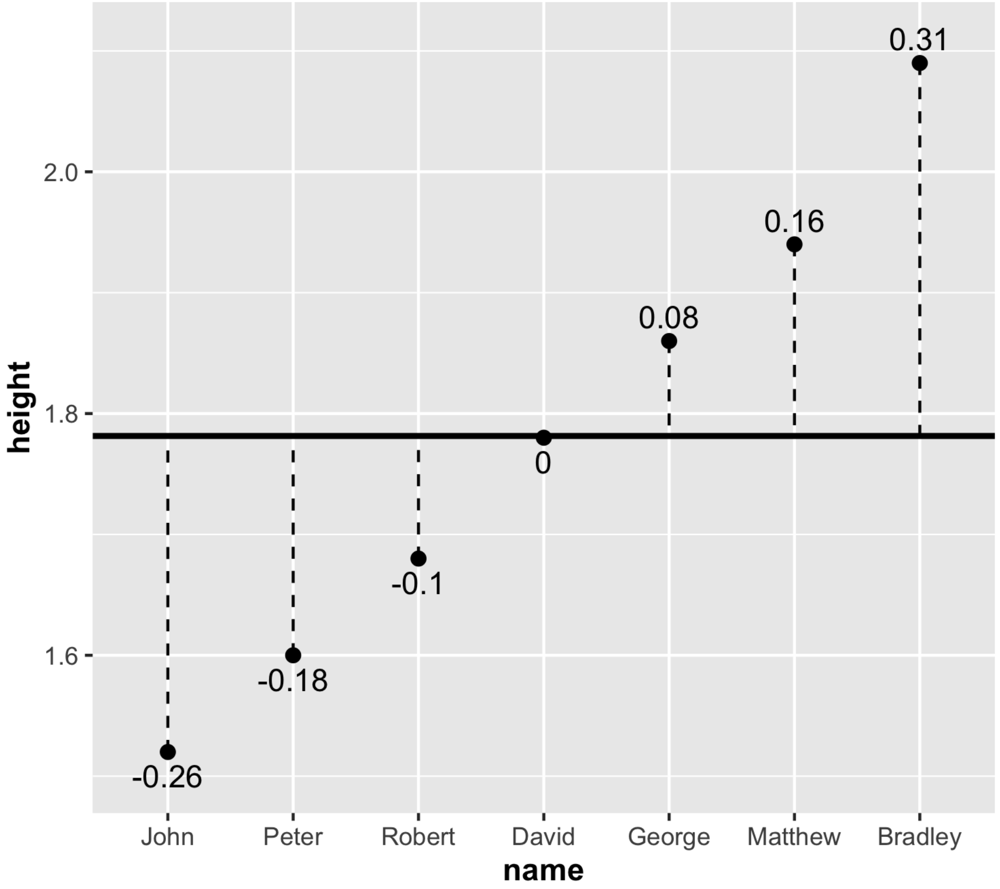
- Variance is the mean squared deviation from the mean.

Variance



- On the plot on the left we see the raw deviations for weight (y-axis) for each person (x-axis).
    - Each point is a person's weight.
    - The solid black line is the average weight.
    - The dashed lines highlight the distance from the mean of the individual weights.
    - The raw deviations are show by each point.
- Raw deviations are the distance of each person's weight from the average weight.
- To get the variance, we square each value (to get rid of the negative values) and sum them up.

Variance



- On the left is the same figure but for height.

Covariance

- So variance = deviation around the mean of a single variable.
- **Co**variance concerns variation in two variables.
- To think about the equation for covariance, suppose we re-write variance as follows. Instead of:

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

- we use

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

Covariance

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

- So our covariance is identical to our variance, with the exception that our summed termed is the combined deviance from the respective means of both x and y.

Covariance

- For our data:

```
round(cov(data$height, data$weight),4)
```

```
## [1] 4.1681
```

Scale & Covariance

- So what does a covariance of 4.1681 between height and weight mean?
    - I have no idea!
- Covariance is related to the scale of the variables we are analysing.
    - Makes sense right? variance was just the same.
- What about if we had measured height in centimetres not metres?

```
round(cov(data$height*100, data$weight),2)
```

```
## [1] 416.81
```

Correlation

- How do we deal with problems of scale?
    - We standardize.
- And how do we standardize?
    - We divide by an estimate of the variability.
    - Here, the product of standard deviations of x and y.
- The resulting statistic is the Pearson Product Moment Correlation (r)

Correlation

$$r = \frac{Cov_{xy}}{SD_x SD_y}$$

- Or in full

$$r = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}}$$

Correlation

- In our data:

```
cov(data$height, data$weight)/ (sd(data$height)*sd(data$weight))
```

```
## [1] 0.8687186
```

- or we can use built in functions:

```
cor(data$height, data$weight)
```

```
## [1] 0.8687186
```

Correlation = ES

- For some other tests we have discussed associated measures of effect size.
- Remember, an effect size is a standardized measures of the type relationship of interest.
    - So Cohen's D is a standardize raw mean difference.
- Well our correlation **is** standardized
    - It is a standardized covariance.
    - Or a standardize measure of association

Time for a break

Welcome Back!

In the last recording we considered the relationships between variance, covariance and correlation. Now we will consider inferential tests for the Pearson's correlation.

Hypotheses

- For many people, correlations are descriptive statistics.
    - As such, they do not require significance tests.
- But in other circumstances a correlation may be a test of interest, and we can formulate associated hypothesis tests.

Hypotheses

- The association between two random variables = 0.
- This leads to the null for a correlation being:

$$H_0 : r = 0$$

- And the two-tailed alternative:

$$H_1 : r \neq 0$$

- The sampling distribution of r is approximately normal with large N, and is t distributed when N is small.
    - Thus we assess the significance using the t-distribution with n-2 degrees of freedom.
    - The minus 2 is because we have had to calculate the means of both variables from our data.

Hypothesis testing & significance

- The t-statistic for a given correlation is calculated as:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

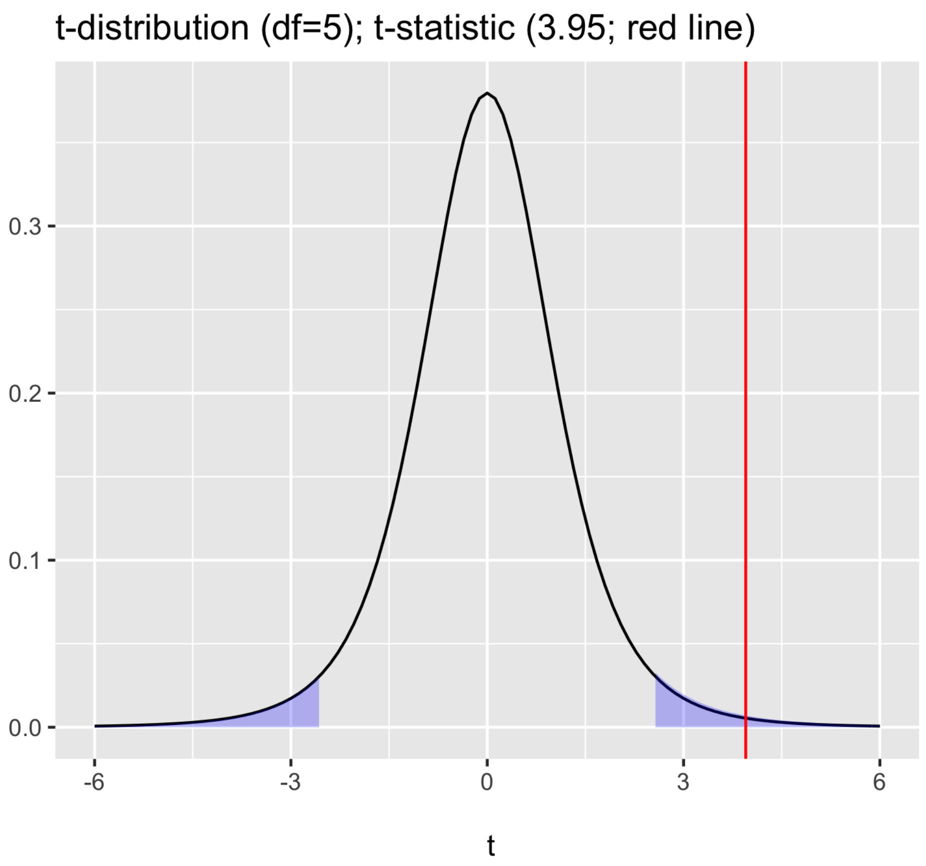
- So for our data:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.87\sqrt{\frac{5}{1-0.87^2}} = 0.87\sqrt{\frac{5}{0.2431}} = 0.87 * 4.535 = 3.95$$

Is our test significant?

- So the t associated with our correlation is 3.95
  - Our degrees of freedom are n-2 = 7-2 = 5
  - We will use two-tailed $\alpha = .05$

Is our test significant?




```
## # A tibble: 1 × 2
##   LowerCrit UpperCrit
```

In R

```
cor.test(data$height, data$weight)
```

```
## 
##      Pearson's product-moment correlation
## 
## data:  data$height and data$weight
## t = 3.9218, df = 5, p-value = 0.01116
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3344679 0.9804020
## sample estimates:
##       cor 
## 0.8687186
```

Write up

- Write up is very simple for small number of variables.

> There was a strong positive correlation between height and weight (r = .87, $t(5)$ = 3.92, p<.05) in the current sample. As height increased, so did weight.

- Often we report lots of correlations and do so in a correlation matrix.

Correlation matrices

- Off-diagonal values show the correlations between the variables.
  - Range from -1 to 1.
- Values in diagonal are correlations of each variable with itself.
  - Always 1.00
  - Not informative
  - Can omit or replace with e.g. reliability
- Symmetric.
  - Above and below diagonal = same values.
  - Do not need both.
  - Could switch with p-values or leave empty

Correlation matrices

```
pers_items <- bfi[,c(1:5)]
pers_cors <- hetcor(pers_items)
```

```
round(pers_cors$correlations, 2)
```

```
##       A1    A2    A3    A4    A5
## A1  1.00 -0.34 -0.27 -0.15 -0.18
## A2 -0.34  1.00  0.49  0.34  0.39
## A3 -0.27  0.49  1.00  0.36  0.51
## A4 -0.15  0.34  0.36  1.00  0.31
## A5 -0.18  0.39  0.51  0.31  1.00
```

Assumptions: Pearson correlation

1. Variables must be interval or ratio (continuous)
    - No test: about design.
2. Variables must be normally distributed.
    - Histograms, skew, QQ-Plots, Shapiro-Wilks.
3. Homoscedasticity (homogeneity of variance)
4. The relationship between the two variables must be linear.
    - Visualize: scatterplots.

Anscombe Quartet

- Anscombe quartet is a set of data designed to show the importance of visualizing data.
- There are four pairs of x and y variables.
  - Each x variable has the same mean and standard deviation.
  - Each y variable has the same mean and standard deviation.
  - Each pair has the same correlation.
- In other words, if you calculate descriptive statistics only, each pair is identical.
- BUT......

Anscombe Quartet

```r
round(cor(anscombe$x1, anscombe$y1),2)
```

```
## [1] 0.82
```

```r
round(cor(anscombe$x2, anscombe$y2),2)
```

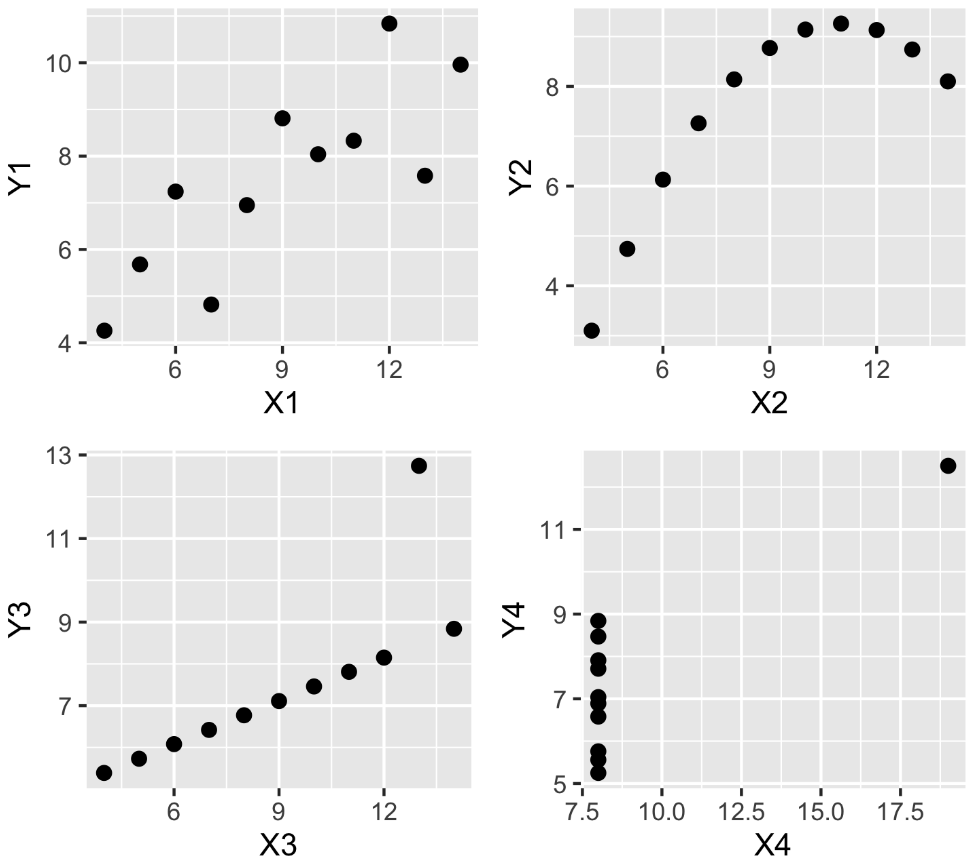
```
## [1] 0.82
```

```r
round(cor(anscombe$x3, anscombe$y3),2)
```

```
## [1] 0.82
```

```r
round(cor(anscombe$x4, anscombe$y4),2)
```

```
## [1] 0.82
```



Time for a break

Welcome Back!

We have now looked at the Pearson correlation, but what about different data types?

Types of correlation

Variable 1	Variable 2	Correlation Type
Continuous	Continuous	Pearson
Continuous	Categorical	Polyserial
Continuous	Binary	Biserial
Categorical	Categorical	Polychoric
Binary	Binary	Tetrachoric
Rank	Rank	Spearman
Nominal	Nominal	Chi-square

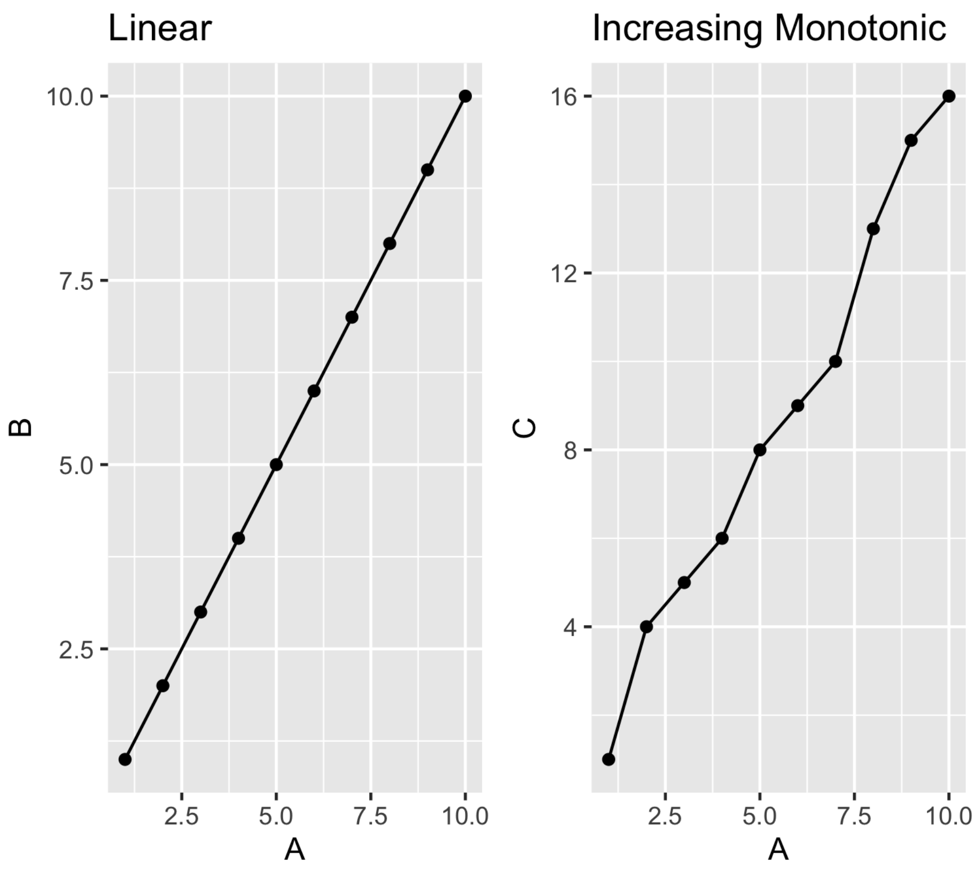
Spearman correlation

- Spearman's ρ (or rank-order correlation) uses data on the rank-ordering of x, y responses for each individual.
- When would we choose to use the Spearman correlation?
  - If our data are naturally ranked data (e.g. imagine a survey where the task is to rank foods and drinks in terms of preference).
  - If the data are non-normal or skewed.
  - If the data shows evidence of non-linearity.

Spearman correlation

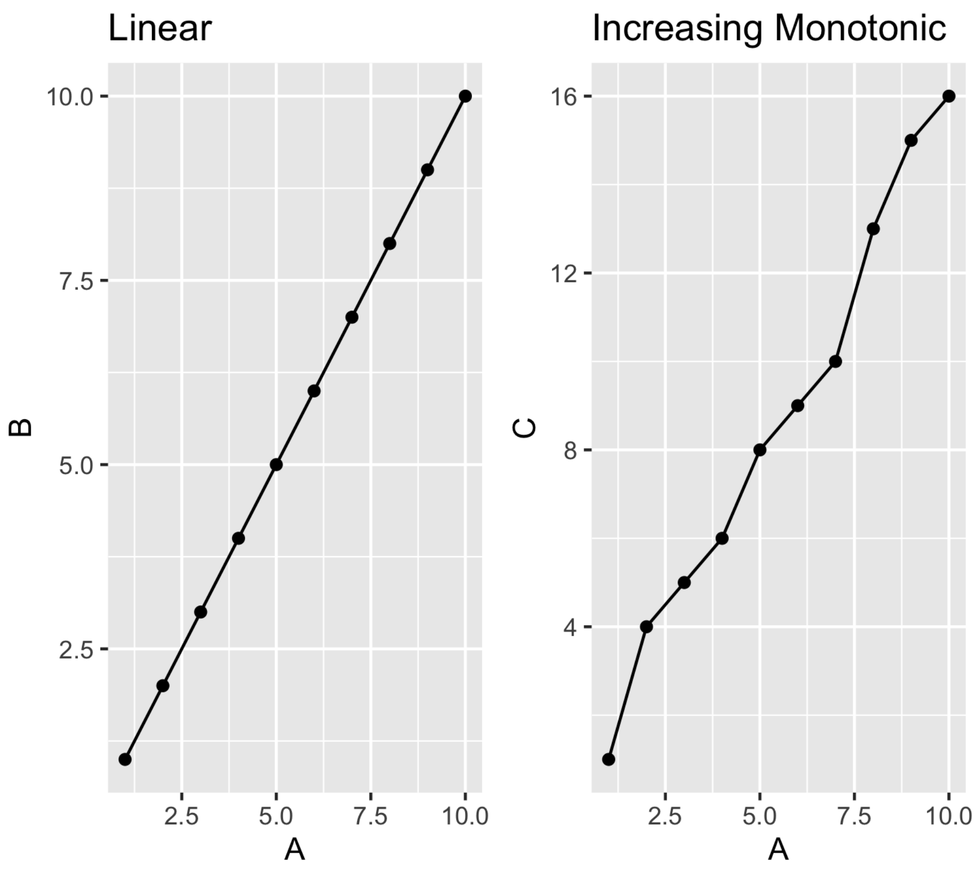
- Spearman's is not testing for linear relations, it is testing for increasing monotonic relationship.
  - Huh?

Linear vs. monotonic



- Left-hand plot shows a perfectly linear relationship between A and B.
- Right-hand plot shows a perfectly increasingly monotinic relationship between A and C.
  - The rank position of all observations on A, is the same a the rank position of all observations on C.

Linear vs. monotonic



ID	A	C	Rank_A	Rank_C
ID1	1	1	1	1
ID2	2	4	2	2
ID3	3	5	3	3
ID4	4	6	4	4
ID5	5	8	5	5
ID6	6	9	6	6
ID7	7	10	7	7
ID8	8	13	8	8
ID9	9	15	9	9
ID10	10	16	10	10

Steps in Spearman's

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

- Calculation steps:
  - Rank each variable from largest to smallest.
  - If there are ties in ranks, assign the average of the rankings to each case.
  - Calculate the difference in rank for each person on the two variables.
  - Square the difference.
  - Sum the squared values.

Quick example

```
rank <- tibble(
  ID = paste("ID", 1:6, sep = ""),
  RT =c(.264, .311, .265, .291, .350, .500),
  Caff = c(210,280,150,90,200,450)
)
rank
```

```
## # A tibble: 6 × 3
##   ID        RT  Caff
##   <chr> <dbl> <dbl>
## 1 ID1   0.264   210
## 2 ID2   0.311   280
## 3 ID3   0.265   150
## 4 ID4   0.291    90
## 5 ID5   0.35    200
## 6 ID6   0.5     450
```

Calculation

```
rank_calc <- rank %>%
  mutate(
    RT_rank = rank(RT),
    Caff_rank = rank(Caff),
    di = RT_rank - Caff_rank,
    di2 = di^2
  )
rank_calc
```

```
## # A tibble: 6 × 7
##   ID       RT  Caff RT_rank Caff_rank    di   di2
##   <chr> <dbl> <dbl>   <dbl>     <dbl> <dbl> <dbl>
## 1 ID1   0.264   210       1         4    -3     9
## 2 ID2   0.311   280       4         5    -1     1
## 3 ID3   0.265   150       2         2     0     0
## 4 ID4   0.291    90       3         1     2     4
## 5 ID5   0.35    200       5         3     2     4
## 6 ID6   0.5     450       6         6     0     0
```

Calculation

```
## # A tibble: 6 × 7
##   ID       RT  Caff RT_rank Caff_rank    di   di2
##   <chr> <dbl> <dbl>   <dbl>     <dbl> <dbl> <dbl>
## 1 ID1   0.264   210       1         4    -3     9
## 2 ID2   0.311   280       4         5    -1     1
## 3 ID3   0.265   150       2         2     0     0
## 4 ID4   0.291    90       3         1     2     4
## 5 ID5   0.35    200       5         3     2     4
## 6 ID6   0.5     450       6         6     0     0
```

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6 * 18}{6(6^2-1)} = 1 - \frac{108}{210} = 1 - 0.514 = 0.486$$

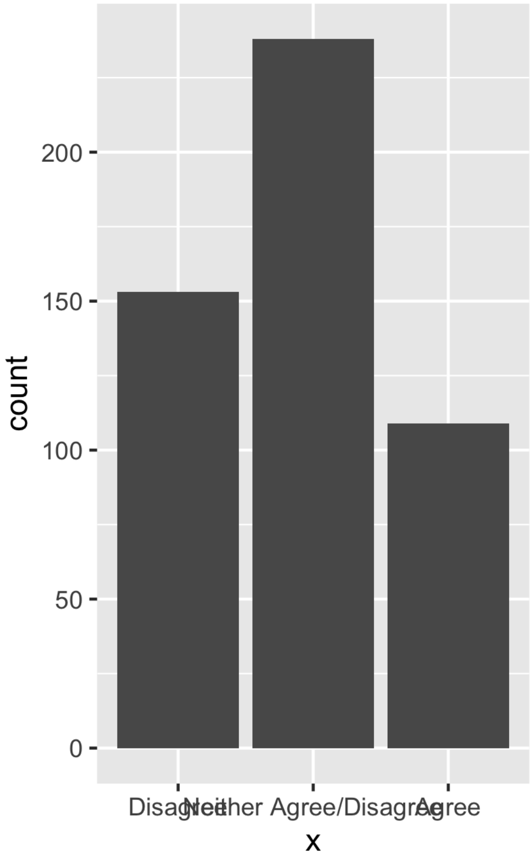
In R

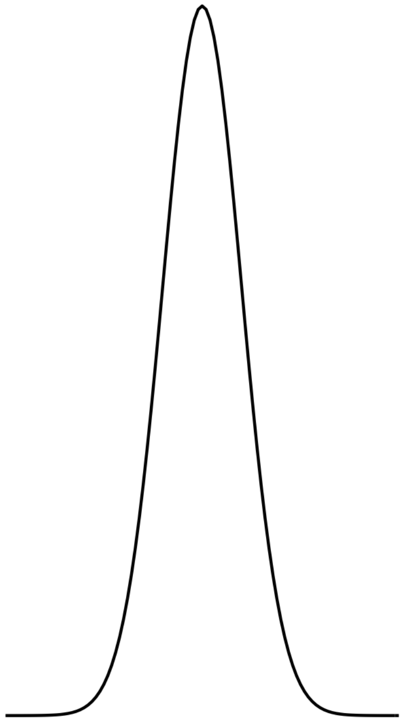
```
round(cor(rank$RT, rank$Caff, method = "spearman"),3)
```

```
## [1] 0.486
```

Other forms

- General principle (simplified a little) of the other forms of correlation is roughly the same.
- We assume that the categorical variable is a crude measurement of an underlying normal variable.
- Aiming to provide an estimate of the association between these underlying variables.





In R

- Estimating correlation is straight forward.
- All we need to do is make sure R knows the type of data we have, then use `hetcor`

```
pers_items <- bfi[,c(1:5)]
pers_items <- pers_items %>%
  mutate(
    A1 = as_factor(A1)
  )
pers_cors <- hetcor(pers_items)
```

In R

```
round(pers_cors$correlations,2)
```

```
##       A1    A2    A3    A4    A5
## A1  1.00 -0.37 -0.29 -0.16 -0.21
## A2 -0.37  1.00  0.49  0.34  0.39
## A3 -0.29  0.49  1.00  0.36  0.51
## A4 -0.16  0.34  0.36  1.00  0.31
## A5 -0.21  0.39  0.51  0.31  1.00
```

```
pers_cors$type
```

```
##      [,1]         [,2]         [,3]         [,4]         [,5]
## [1,] ""           "Polyserial" "Polyserial" "Polyserial" "Polyserial"
## [2,] "Polyserial" ""           "Pearson"    "Pearson"    "Pearson"
## [3,] "Polyserial" "Pearson"    ""           "Pearson"    "Pearson"
## [4,] "Polyserial" "Pearson"    "Pearson"    ""           "Pearson"
## [5,] "Polyserial" "Pearson"    "Pearson"    "Pearson"    ""
```

Correlation and causation

- You will talk more about this point in lab.
  - And forever more when discussing statistical results.
- Typically we hope to be able to explain *why* things happen.
- Though correlation is a fundamental metric in statistics, it actually does not help us (on it's own) with this.
- An association between two things does not mean it **causes** the other.
  - Much more on this to come in lab and next year.

Summary of today

- In these recordings we have discussed:
  - The basic principle and interpretation of correlations
  - The importance of visualization and how to "read" scatterplots.
  - Calculation of Pearson's and other forms of correlation
  - Inferential tests and effect sizes for correlations.