Correlation

Data Analysis for Psychology in R 1

Semester 2 Week 10

Dr Emma Waterston

Department of Psychology
The University of Edinburgh

Course Overview

Exploratory Data Analysis	Research design and data
	Describing categorical data
	Describing continuous data
	Describing relationships
	Functions
Probability	Probability theory
	Probability rules
	Random variables (discrete)
	Random variables (continuous)
	Sampling

Foundations of inference	Confidence intervals
	Hypothesis testing (p-values)
	Hypothesis testing (critical values)
	Hypothesis testing and confidence intervals
	Errors, power, effect size, assumptions
Common hypothesis tests	One sample t-test
	Independent samples t-test
	Paired samples t-test
	Chi-square tests
	Correlation

Learning Objectives

- Understand the difference between variance, covariance, and correlation
- Know how to calculate both covariance and correlation
- Understand how to interpret the correlation coefficient
- Know how perform and interpret the results of a significance test of your correlation
- Understand which form of correlation is most appropriate to use with your data

Example

Running Program

- Suppose a running coach wants to know whether there is an association between a participants' time spent enrolled in their running program (recorded in days) and the maximum distance they can run (recorded in miles)
- We are provided with information on the last 100 participants to enroll on the program

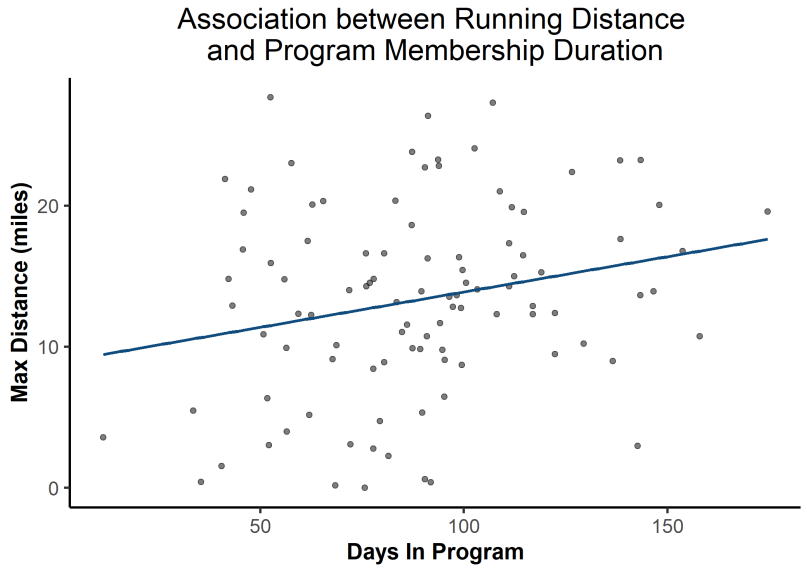
Data

```
head(dat)
```

```
##   daysInProgram maxDistance
## 1      95.32143    9.076514
## 2     174.62360   19.581428
## 3      11.38827    3.575377
## 4      91.17004   26.352013
## 5      56.37932    9.923332
## 6     122.36747   12.384450
```

Visualisation

```
ggplot(dat, aes(daysInProgram, maxDistance)) +
  geom_point() +
  geom_smooth(method = 'lm', se=FALSE) +
  labs(x='Days In Program',
       y='Max Distance (miles)',
       title = "Association between Running
       Distance \nand Program
       Membership Duration")
```



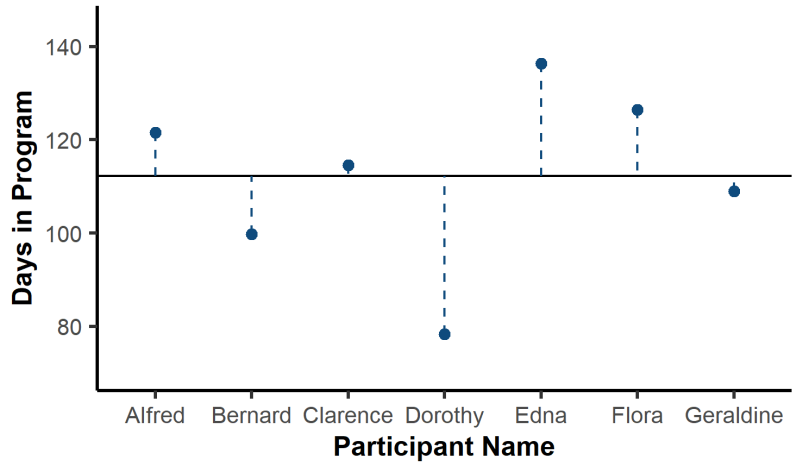
Questions?

Variance & Covariance

Variance Recap

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

- **Variance:** Deviance around the mean of a single variable
- Raw deviation is the distance between each person's days in the program and the mean number of days in the program
- To get the variance, we:
    1. Square the values to get rid of the negative
    2. Sum them up and divide by $n - 1$ to get the average deviation of the group from its mean



Variance

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = s_x^2 = \frac{2192.57}{7-1} = s_x^2 = 365.43$$

- where $\bar{x} = 112.3$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
121.57	9.27	86.01
99.73	-12.57	157.9
114.63	2.33	5.45
78.37	-33.93	1150.95
136.41	24.11	581.5
126.42	14.12	199.5
108.94	-3.36	11.26
		2192.57

Covariance Recap

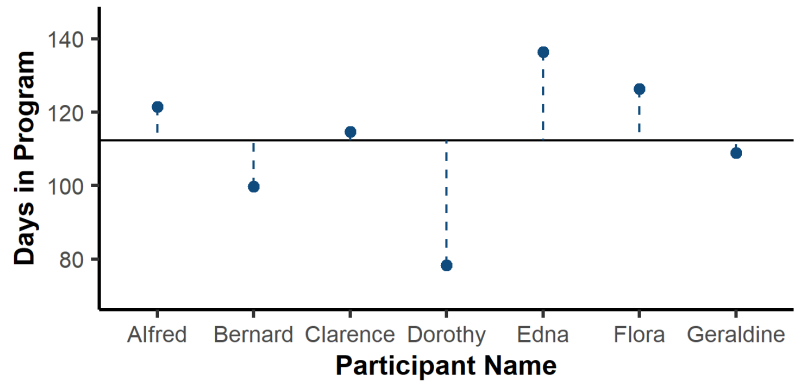
- **Covariance:** A value that represents how two variables change together
- Does y differ from its mean in a similar way to x?
- Mathematically similar to variance:

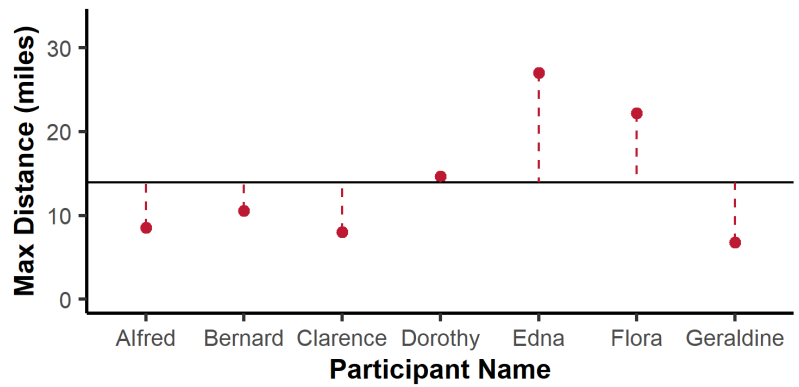
Variance

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

Covariance

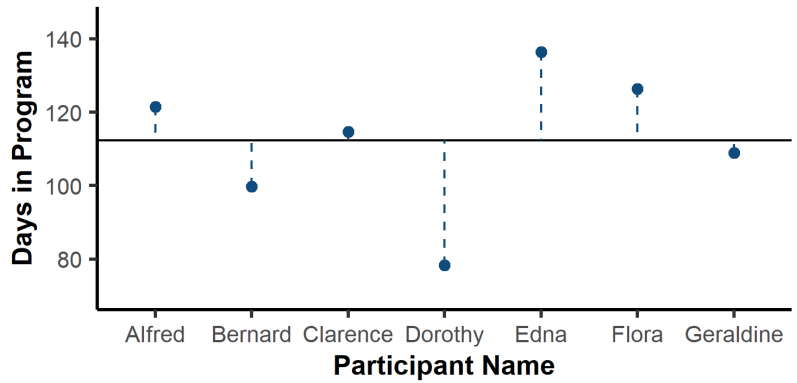
$$Cov_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

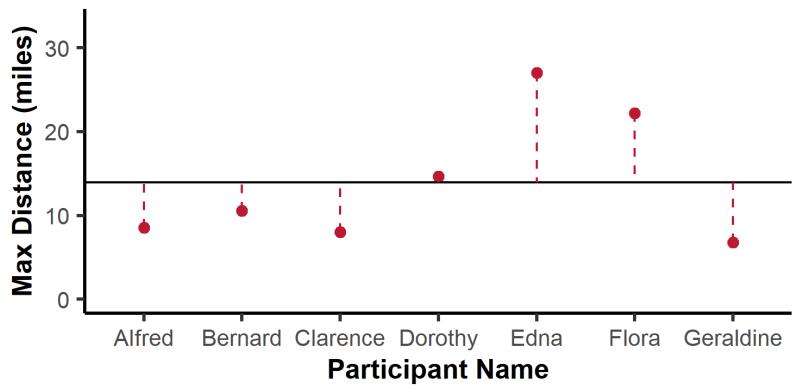




Covariance Recap

- It's possible two variables are related if their observations differ proportionally from their means in a consistent way
- Covariance gives us a sense of this...
  - High covariance suggests a stronger association than a lower covariance
  - Why can't we stop here?
  - Why is correlation necessary?

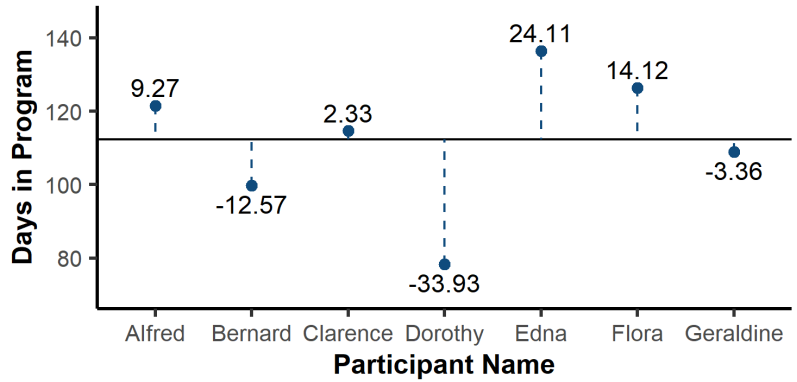


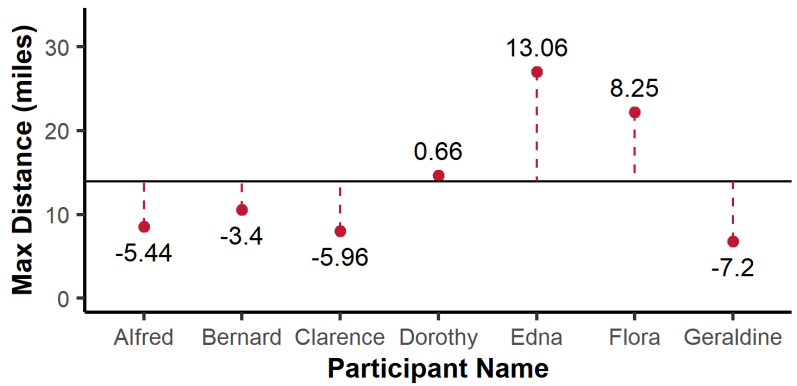


The Trouble with Covariance

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

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
9.27	-5.44	-50.41
-12.57	-3.4	42.67
2.33	-5.96	-13.9
-33.93	0.66	-22.54
24.11	13.06	315.04
14.12	8.25	116.59
-3.36	-7.2	24.15
		411.59

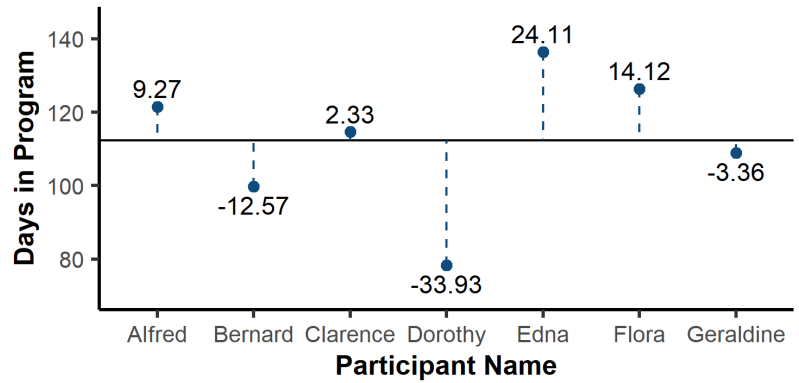


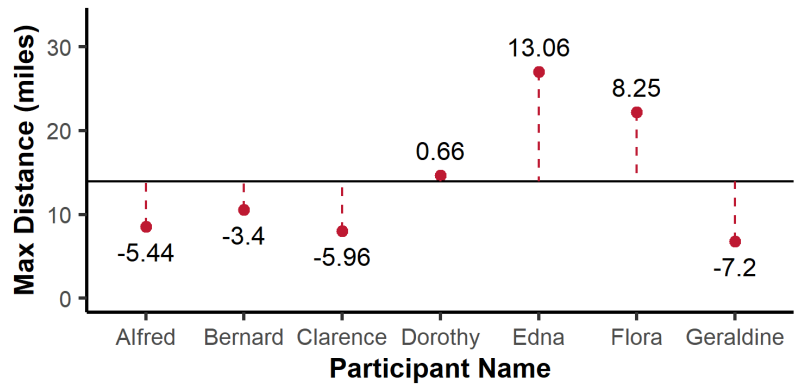


The Trouble with Covariance

$$Cov_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{411.59}{7-1} = 68.6$$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
9.27	-5.44	-50.41
-12.57	-3.4	42.67
2.33	-5.96	-13.9
-33.93	0.66	-22.54
24.11	13.06	315.04
14.12	8.25	116.59
-3.36	-7.2	24.15
		411.59



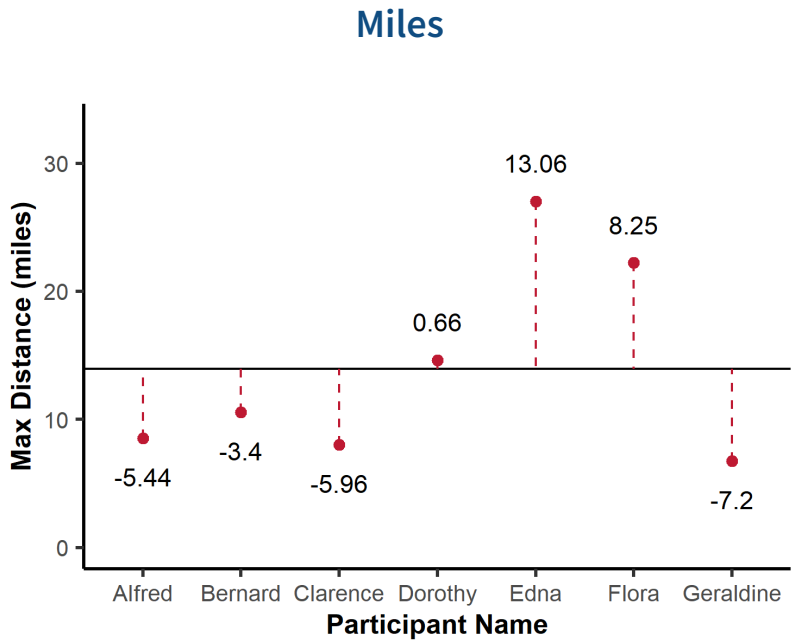


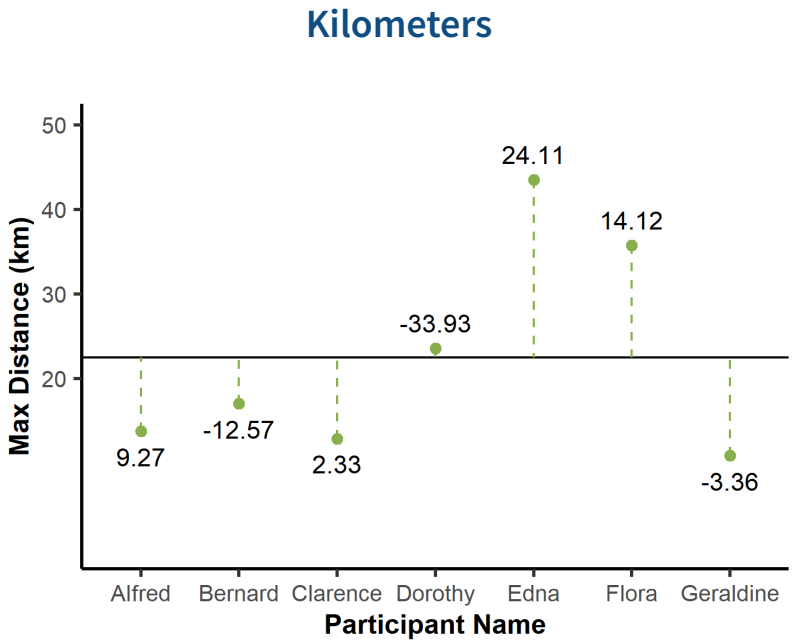
The Trouble with Covariance

- A value of 68.6 seems high...I think. Is it?
  - Maybe? But maybe not
  - Covariance is related specifically to the scales of the variables we are analysing
  - Variables with larger scales will naturally have larger covariance values

The Trouble with Covariance

- Consider what would happen if we converted our distance data to kilometers instead of miles





The Trouble with Covariance

Miles

$$Cov_{xy} = \frac{411.59}{7-1} = 68.6$$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
9.27	-5.44	-50.41
-12.57	-3.4	42.67
2.33	-5.96	-13.9
-33.93	0.66	-22.54
24.11	13.06	315.04
14.12	8.25	116.59
-3.36	-7.2	24.15
		411.59

Kilometers

$$Cov_{xy} = \frac{662.66}{7-1} = 110.44$$

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
9.27	-8.75	-81.16
-12.57	-5.47	68.7
2.33	-9.59	-22.38
-33.93	1.07	-36.28
24.11	21.03	507.21
14.12	13.29	187.7
-3.36	-11.59	38.88
		662.66

Questions?

The Correlation Coefficient

Correlation

- Correlation allows you to compare continuous variables across different scales without the magnitude of the variables skewing your results
- **Pearson's product moment correlation**, r, is the standardised version of covariance:

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

- By dividing covariance by the product of the standard deviations of x and y, we remove issues with scale differences in the original variables
- Because of this, you can use r to investigate the association(s) between continuous variables with completely different ranges

Correlation

Miles

$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
9.27	86.01	-5.44	29.55
-12.57	157.9	-3.4	11.53
2.33	5.45	-5.96	35.47
-33.93	1150.95	0.66	0.44
24.11	581.5	13.06	170.68
14.12	199.5	8.25	68.13
-3.36	11.26	-7.2	51.78
	2192.57		**367.58**

$$s_x = \sqrt{\frac{2192.57}{7-1}} = 19.12 \quad s_y = \sqrt{\frac{367.58}{7-1}} = 7.83$$

Kilometers

$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
9.27	86.01	-8.75	76.59
-12.57	157.9	-5.47	29.89
2.33	5.45	-9.59	91.94
-33.93	1150.95	1.07	1.14
24.11	581.5	21.03	442.41
14.12	199.5	13.29	176.61
-3.36	11.26	-11.59	134.21
	2192.57		**952.8**

$$s_x = \sqrt{\frac{2192.57}{7-1}} = 19.12 \quad s_y = \sqrt{\frac{952.8}{7-1}} = 12.6$$

Correlation

Miles

$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
9.27	86.01	-5.44	29.55
-12.57	157.9	-3.4	11.53
2.33	5.45	-5.96	35.47
-33.93	1150.95	0.66	0.44
24.11	581.5	13.06	170.68
14.12	199.5	8.25	68.13
-3.36	11.26	-7.2	51.78
	2192.57		**367.58**

$$r = \frac{Cov_{xy}}{s_x s_y} = \frac{68.6}{19.12 \cdot 7.83} = 0.46$$

Kilometers

$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
9.27	86.01	-8.75	76.59
-12.57	157.9	-5.47	29.89
2.33	5.45	-9.59	91.94
-33.93	1150.95	1.07	1.14
24.11	581.5	21.03	442.41
14.12	199.5	13.29	176.61
-3.36	11.26	-11.59	134.21
	2192.57		**952.8**

$$r = \frac{Cov_{xy}}{s_x s_y} = \frac{110.44}{19.12 \cdot 12.6} = 0.46$$

Correlation

- Correlations measure the degree of association between two variables
- If one variable changes, does the other variable also change?
- If so, do they rise and fall together, or does one rise as the other falls?

Correlation in R

- To run a simple correlation, you can use `cor()`
- Let's compute the correlation between the number of days in the program and the max running distance for our entire sample of 100:

```
cor(dat$daysInProgram, dat$maxDistance)
```

```
## [1] 0.2332039
```

- Now we have a correlation value. But what does it mean?

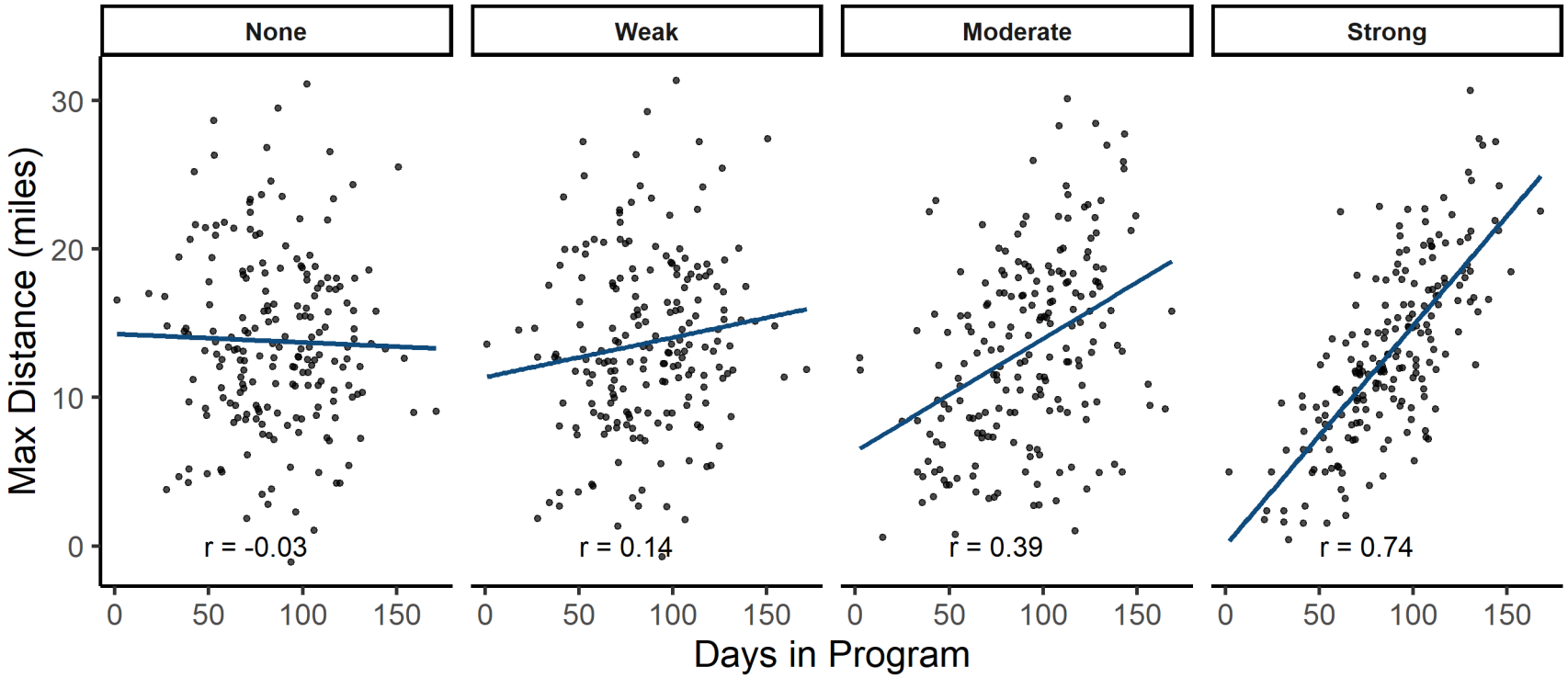
Interpreting r

- Values of r fall between -1 and 1.
    - Values closer to 0 indicate a weaker association
    - More extreme values indicate a stronger association
- Interpretation:

Strength	Value
Weak	$.1 < \|r\| < .3$
Moderate	$.3 < \|r\| < .5$
Strong	$\|r\| > .5$

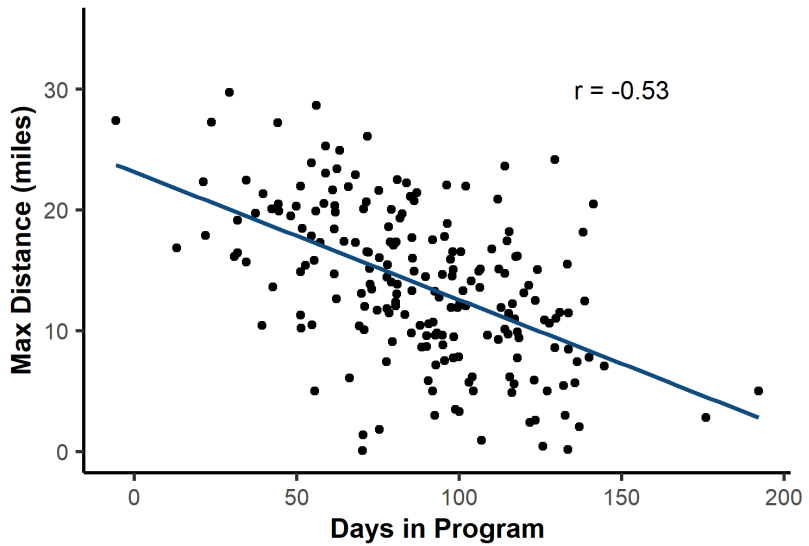
Interpreting r

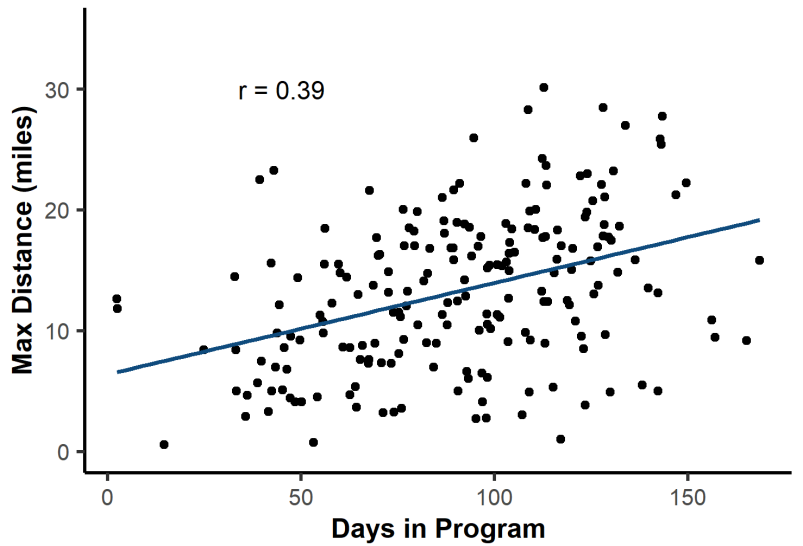
- Values of r fall between -1 and 1
  - Values closer to 0 indicate a weaker association
  - More extreme values indicate a stronger association



Interpreting r

- The sign of r says nothing about the strength of the association, but its direction
  - Positive values indicate that the two variables rise together or fall together.
  - Negative values indicate that as one variable increases, the other decreases, and vice versa





Interpreting r

- In some cases, r is considered a descriptive statistic
  - r is actually a direct measure of effect size:
    - It provides information about the strength of the association between two variables
    - It is a standardized measure

Questions?

Hypothesis Testing with r

Hypotheses

- There may be times that a correlation is the test of interest, and we can formulate associated hypothesis tests
- There is no association between two random variables, so the null hypothesis should reflect this:
  - $H_0 : r = 0$
  - $H_{1\ two-tailed} : r \neq 0$
  - $H_{1\ one-tailed} : r > 0 \ \vee \ r < 0$

Significance Testing

Remember the key steps of hypothesis testing:

1. Compute a test statistic
2. Locate the test statistic on a distribution that reflects the probability of each test statistic value, given that H0 is true.
3. Determine whether the probability associated with your test statistic is lower than α

Calculation

Compute a test statistic

- The sampling distribution for r is approximately normal with a large n, and is t distributed when n is small.
  - Thus, significance is assessed using a t-distribution
- The t-statistic for a correlation is calculated as:

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

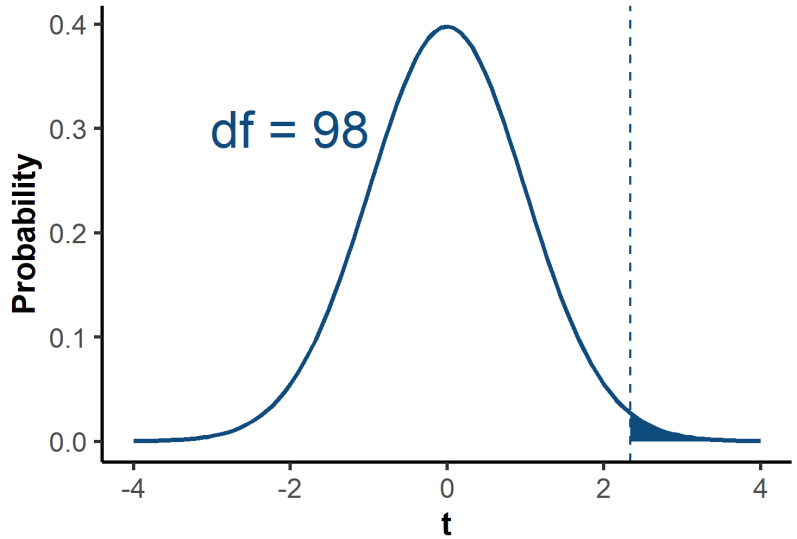
- In our example:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.23 \cdot \sqrt{\frac{100-2}{1-0.23^2}} = 0.23 \cdot \sqrt{\frac{98}{0.95}} = 0.23 \cdot \sqrt{103.16} = 2.34$$

Is our Test Significant?

Locate the test statistic on a distribution

- We have all of the pieces we need:
  - Our t-statistic = 2.34
  - We use a t distribution with $n - 2$ degrees of freedom, so degrees of freedom = $100 - 2 = 98$
    - $n - 2$: we had to calculate the means of *two* variables (e.g., in our example `daysInProgram` and `maxDistance`)
  - We will use two-tailed α = .05
- Now all we need is to calculate the p-value in order to make our decision



Is our Test Significant?

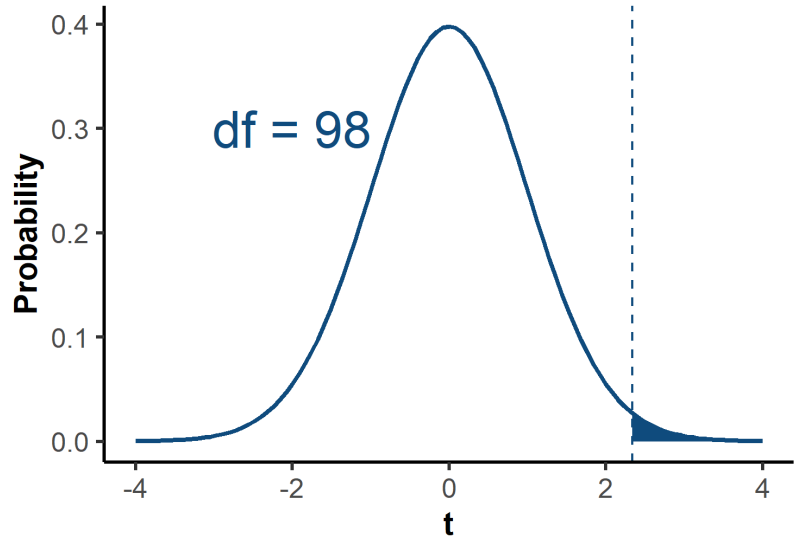
Determine whether the probability associated with your test statistic is lower than α

- What proportion of the plot falls in the shaded area?

```
tibble(
  Exactp = round(2*(1-pt(2.34, 98)),2)
)
```

```
## # A tibble: 1 × 1
##   Exactp
##    <dbl>
## 1   0.02
```

- The probability that we would have a t-statistic at least as extreme as 2.34 if H_0 were true is only 0.022
  - $0.022 < .05$, so we conclude our results are significant (i.e., since $p < \alpha$)



Correlation in R

- Use `cor.test()`

```
cor.test(dat$daysInProgram, dat$maxDistance,
         alternative = "two.sided")
```

```
##
##     Pearson's product-moment correlation
##
## data:  dat$daysInProgram and dat$maxDistance
## t = 2.3741, df = 98, p-value = 0.01954
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03855169 0.41080493
## sample estimates:
##       cor
## 0.2332039
```

*Note: There might be a small amount of rounding error when we compare to calculated in **R**.*

Write Up

There was a weak positive correlation between maximum distance and the number of days in program $(r = .23, t(98) = 2.37, p = .020)$. These results suggested that a greater number of days in the program was positively associated with a higher maximum running distance.

Questions?

Assumptions

Assumptions of Pearson Correlation

1. Variables must be interval or ratio (continuous)
2. Variables must be normally distributed
3. There must be no extreme outliers in your data
4. The association between the two variables must be linear
5. Homoscedasticity (homogeneity of variance)

Assumptions of Pearson Correlation

1. Variables must be interval or ratio (continuous)
    - Knowledge of your data
    - No Likert scales!

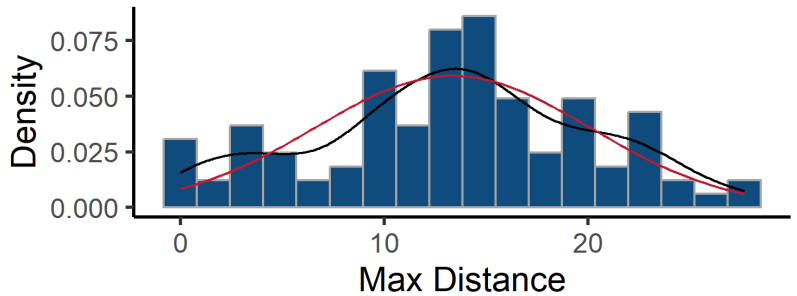
Assumptions of Pearson Correlation

2. Variables must be normally distributed

Max Distance

```
shapiro.test(dat$maxDistance)
```

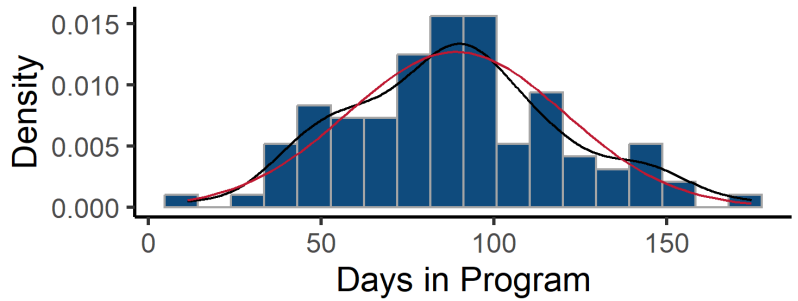
```
##
##     Shapiro-Wilk normality test
##
## data:  dat$maxDistance
## W = 0.98001, p-value = 0.1332
```



Days in Program

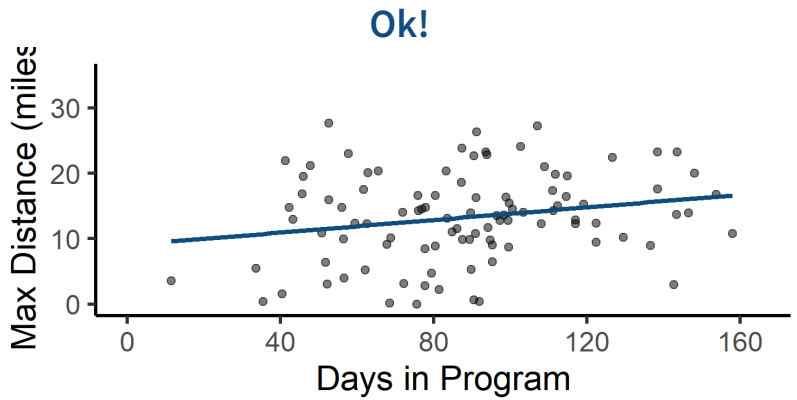
```
shapiro.test(dat$daysInProgram)
```

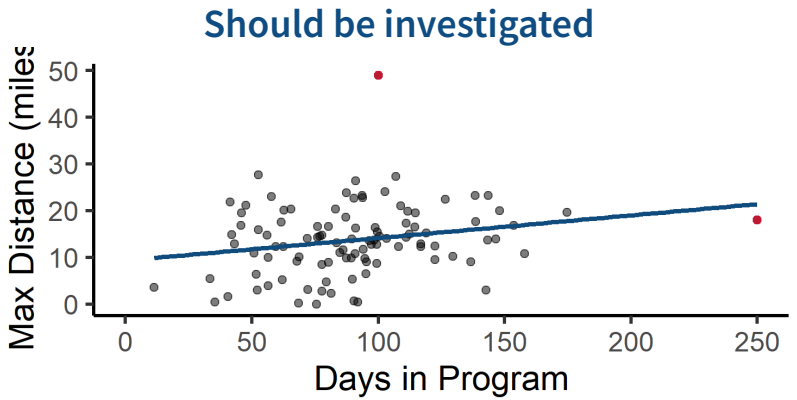
```
##
##     Shapiro-Wilk normality test
##
## data:  dat$daysInProgram
## W = 0.98833, p-value = 0.533
```



Assumptions of Pearson Correlation

3. There must be no extreme outliers in your data

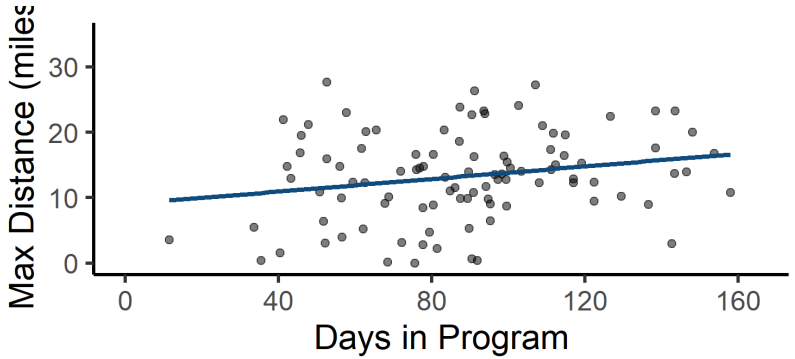




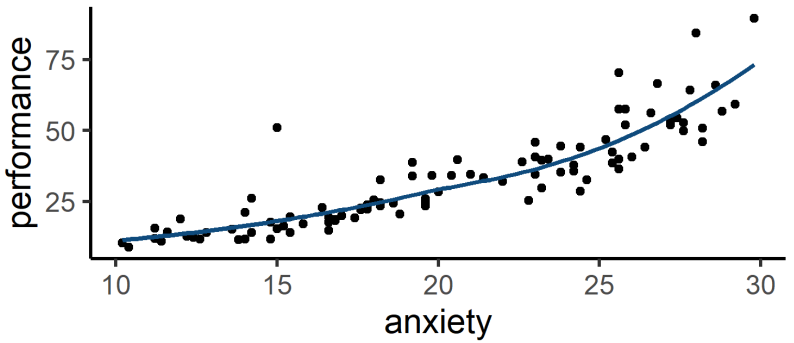
Assumptions of Pearson Correlation

4. The association between the two variables must be linear

Ok!

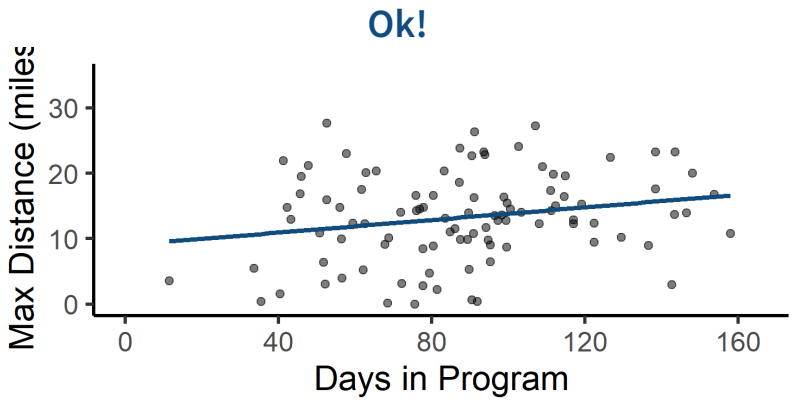


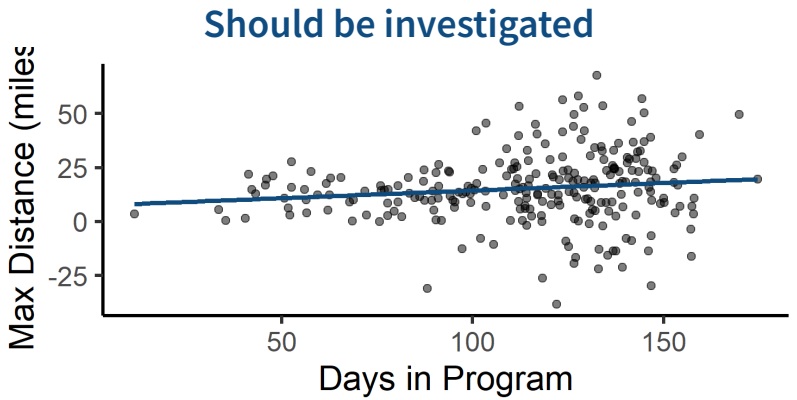
Should be investigated



Assumptions of Pearson Correlation

5. Homoscedasticity (homogeneity of variance)





Questions?

Other Types of Correlation

Types of Correlation

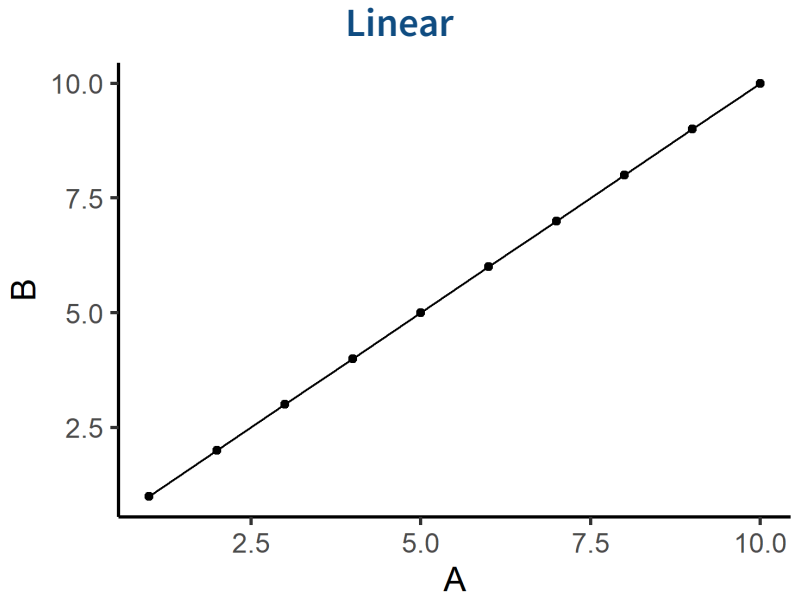
Variable 1	Variable 2	Correlation Type
Continuous	Continuous	Pearson
Continuous	Categorical	Polyserial
Continuous	Binary	Biserial
Categorical	Categorical	Polychoric
Binary	Binary	Tetrachoric
Rank	Rank	Spearman
Nominal	Nominal	Chi-square

Spearman's Correlation

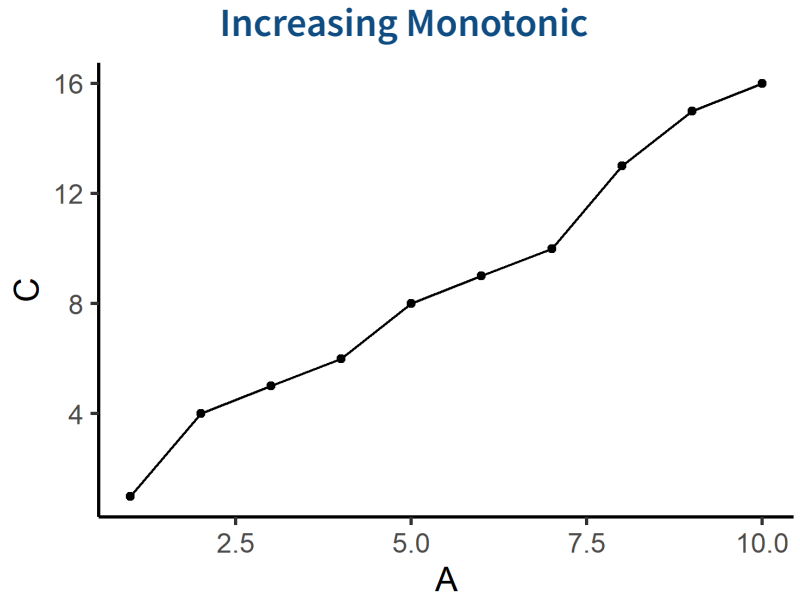
- Spearman's ρ (or rank-order correlation) uses data on the rank-ordering of x, y responses for each individual
- Spearman's ρ is a nonparametric version of Pearson's r, so it doesn't require the same constraints on your data
- When would we choose to use the Spearman correlation?
    - If our data are naturally ranked data (e.g. a survey where the task is to rank foods and drinks in terms of preference)
    - Our data are ordinal (e.g., Likert scales)
    - If the data are non-normal or skewed
    - If the data shows evidence of non-linearity

Spearman's Correlation

- Spearman's is not testing for linear associations, it is testing for increasing monotonic association
  - What?



A perfectly linear association between A & B

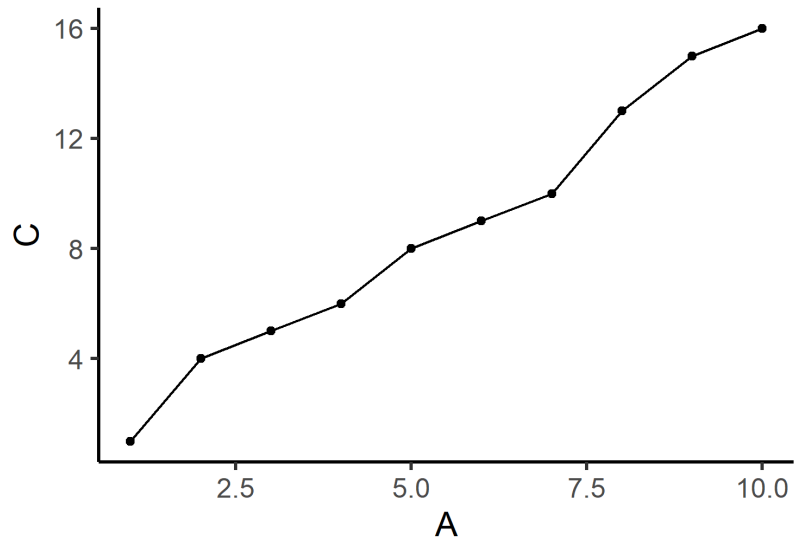


A perfectly increasing monotonic association between A & C

Monotonic Association

- **Perfect Monotonic Association:** The rank position of all observations on Variable A is the same as the rank position of all observations on Variable C

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



Calculating Spearman's ρ

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

- $d_i =$ rank(x_i) $-$ rank(y_i)
- Steps:

    1. Rank each variable from largest to smallest
    2. Calculate the difference in rank for each person on the two variables
    3. Square the difference
    4. Sum the squared values

Calculating Spearman's ρ by Hand

- Imagine we want to know whether the participants' ratings (on a 1-10 scale) of the program are associated with how difficult they found the program (on a 1-10 scale):

Names	ProgRating	Difficulty
Alfred	7	10
Bernard	8	9
Clarence	4	8
Dorothy	2	5
Edna	5	7
Flora	3	1
Geraldine	1	4

x_i	y_i	x_i ranked	y_i ranked	di	di^2
7	10	6	7	-1	1
8	9	7	6	1	1
4	8	4	5	-1	1
2	5	2	3	-1	1
5	7	5	4	1	1
3	1	3	1	2	4
1	4	1	2	-1	1
					10

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 10}{7 \cdot (7^2-1)} = 1 - \frac{60}{7 \cdot (49-1)} = 1 - \frac{60}{336} = 1 - 0.18 = 0.82$$

Spearman's ρ in R

- You can also use `cor()` and `cor.test()` to calculate Spearman's ρ in R

```
cor.test(ratings$ProgRating, ratings$Difficulty,
         method = "spearman",
         alternative = "two.sided")
```

```
##
##     Spearman's rank correlation rho
##
## data:  ratings$ProgRating and ratings$Difficulty
## S = 10, p-value = 0.03413
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8214286
```

- Necessary to use when you have ties in your ranks (i.e., your ranks are not unique)

Summary

- Today we have covered:
  - The differences between variance, covariance, and correlation
  - How to calculate both covariance and correlation
  - How to interpret both the correlation coefficient and the results of the associated significance test
  - Other methods for correlation and calculated Spearman's ρ

This Week

Tasks

- Attend both lectures
- Attend your lab and work on the assessed report with your group (due by 12 noon on Friday 27th of March 2026)
- Complete the weekly quiz
    - Opened Monday at 9am
    - Closes Sunday at 5pm

Support

- **Office Hours**: for one-to-one support on course materials or assessments
  (see LEARN > Course information > Course contacts)
- **Piazza**: help each other on this peer-to-peer discussion forum
- **Student Adviser**: for general support while you are at university
  (find your student adviser on MyEd/Euclid)