Comments for MEDB 5501, Week 13

What this talk will cover

- Calculation of the covariance and correlation.
- Interpretation of the correlation
- Missing values
- SPSS calculations of correlations
- Spearman correlation
- Large correlation matrices
- Confidence intervals and hypothesis tests
- Partial correlations

Covariance

- $Cov(X,Y) = \frac{1}{n-1}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$
  - $(X_i - \bar{X})(Y_i - \bar{Y})$ is positive if
    - X_i and Y_i both above average
    - X_i and Y_i both below average
  - $(X_i - \bar{X})(Y_i - \bar{Y})$ is negative if
    - X_i above average and Y_i below average
    - X_i below average and Y_i above average

Speaker notes

I want to start this section with a discussion of covariance. Covariance is a term that the more mathematically oriented statisticians love to use. It is an interesting statistic from a theoretical perspective and it forms the foundation for a large number of statistical tests.

It doesn’t, however, have as much practical application, compared to the correlation, which you will see in just a bit. I am introducing it here to get you familiar with the terminology.

It is sort of analogous to the term variance versus standard deviation. The variance is interesting from a theoretical perspective, but the standard deviation has far more practical implications.

To compute the covariance, you add up a bunch of terms, each of which is a product. The product is positive if both X and Y are above average or if both X and Y are below average. A positive times a positive is positive and a negative times a negative is also positive.

The product is negative if one value is above average and the other value is below average. A positive times a negative or a negative times a positive produces a negative product.

Think of the covariance as measuring the tendency for two variables to co-vary in a positive sense (large pairing with large and small pairing with small) or in a negative sense (large pairing with small and small pairing with large).

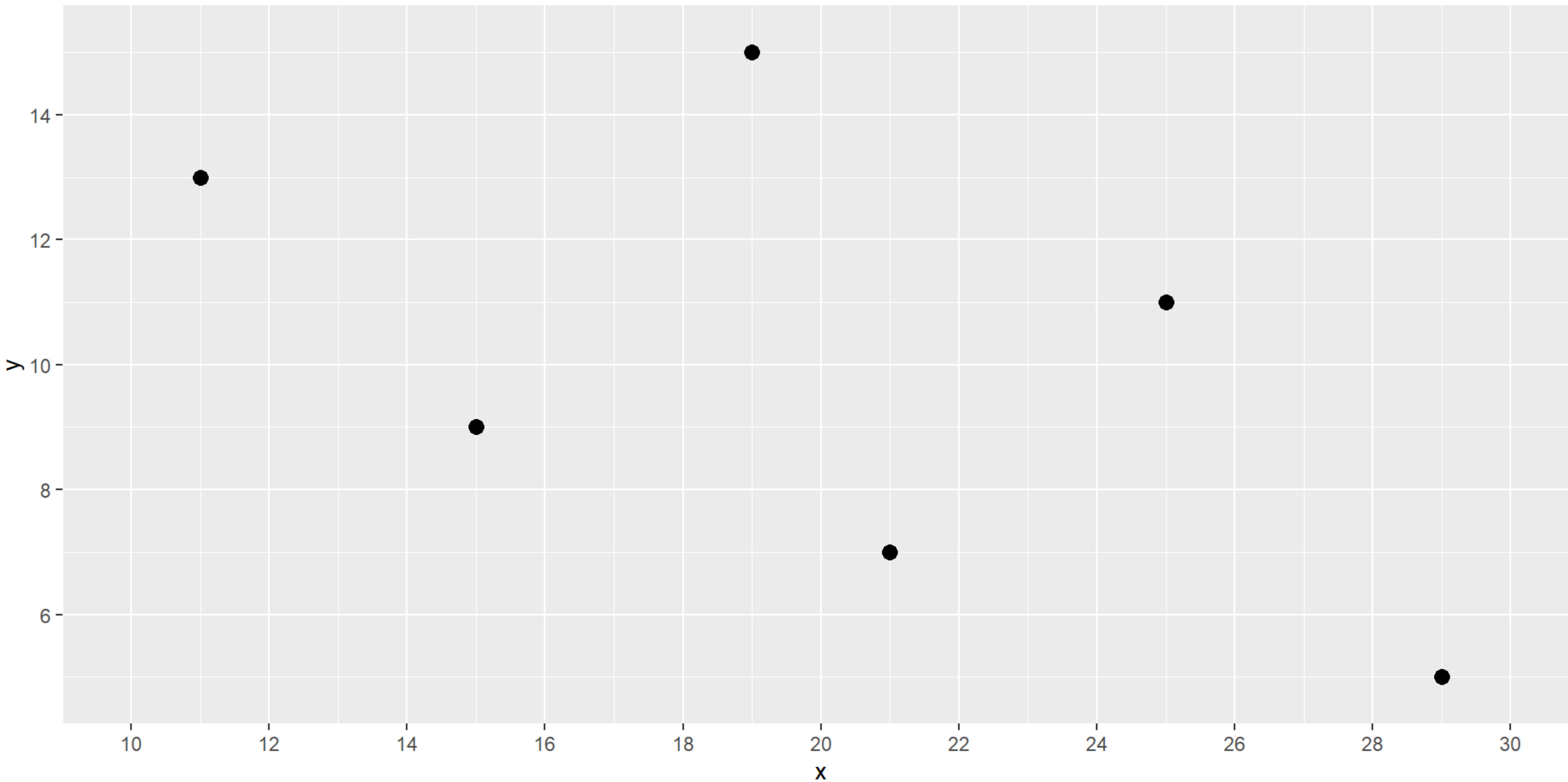
x	y
11	13
15	9
19	15
21	7
25	11
29	5

$$\bar{X} = 20;$$

$$\bar{Y} = 10;$$

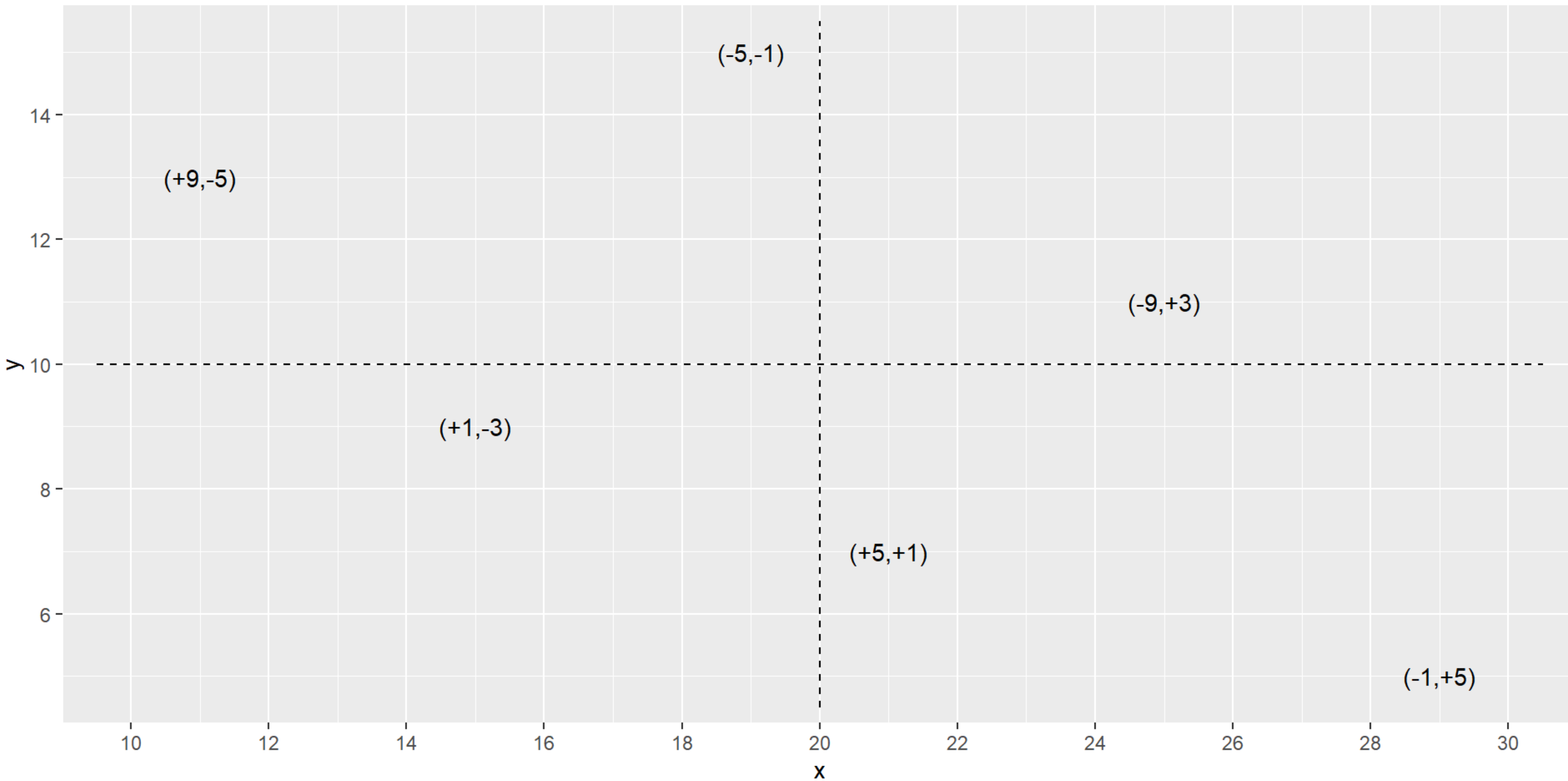
$$S_X = 6.5;$$

$$S_Y = 3.7$$


14
12
10
8
6
y
10
12
14
16
18
20
22
24
26
28
30
x

Speaker notes

Here is an artificial data set that I chose to make some of the calculations simpler. There are two variables, X and Y, and you want to measure how much they co-vary.



(-5,-1)
(+9,-5)
(-9,+3)
(+1,-3)
(+5,+1)
(-1,+5)
14
12
10
8
6
y
10
12
14
16
18
20
22
24
26
28
30
x

Speaker notes

The covariance measures how much each value deviates from the mean. Notice for two of the data points in the upper left corner of the graph, the X value is below average and the Y value is above average.

There is only one data point in the lower left corner, representing a data value where both X and Y are below average.

There are two points in the lower right corner, representing a data value where X is above average and Y is below average.

Finally, there is a single point in the upper right corner, representing a data value where both X and Y are above average.

Calculation of covariance

```
 x_centered y_centered product
          9         -5     -45
          1         -3      -3
         -5         -1       5
          5          1       5
         -9          3     -27
         -1          5      -5
```

$$

- $Cov(X,Y) = \frac{1}{5}(-70) = -14$

Speaker notes

Take the products of the terms in the previous graph, add them up and divide by n-1. In this example, n-1=5.

Correlation

- $Corr(X, Y) = \frac{Cov(X,Y)}{S_X S_Y}$
  - Also use r_{XY}
  - Population correlation is ρ_{XY}
- Other names
  - Pearson correlation
  - Product moment correlation

Speaker notes

The correlation is just the covariance divided by the two standard deviations. This calculation makes the quantity unitless, which is both an advantage and a disadvantage (but mostly an advantage).

While I will normally just use the word “correlation” you will sometimes see reference to the Pearson correlation or the product moment correlation or even the Pearson product moment correlation.

Calculation of correlation

- $r_{XY} = \frac{-14}{6.5 \times 3.7} = -0.571929$
  - Always round!
    - $r_{XY} = -0.57$ or -0.6

Speaker notes

Here is an example of how to compute a correlation. It’s easy once you have the covariance. Be sure to round your correlations to two decimal places. I’m in a minority here, but I often think that rounding to a single decimal place is appropriate. It may be a bit extreme, but you’ll see some examples where this amount of rounding makes it much easier to see patterns.

Break #1

- What have you learned
  - Calculation of the covariance and correlation.
- What is coming next
  - Interpretation of the correlation

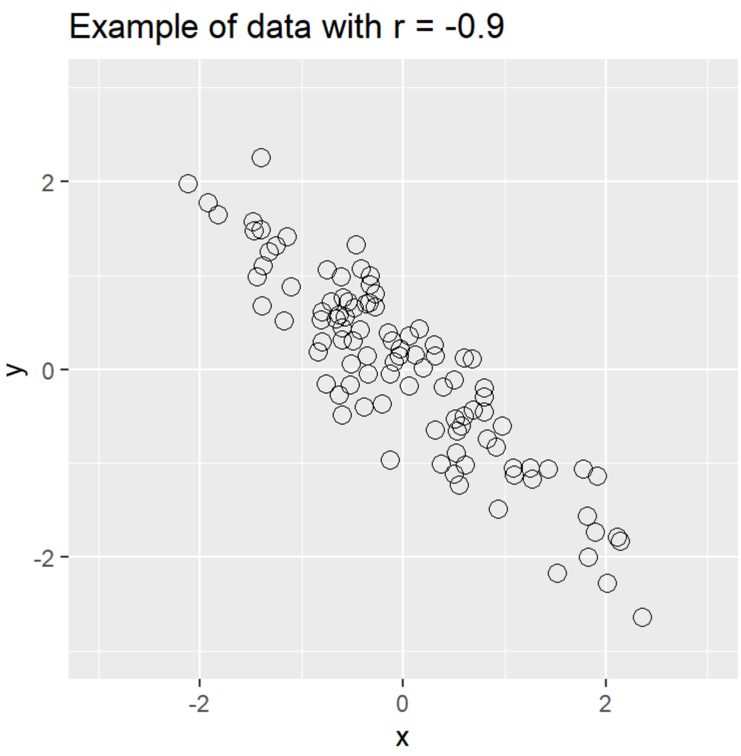
Interpretation of correlation

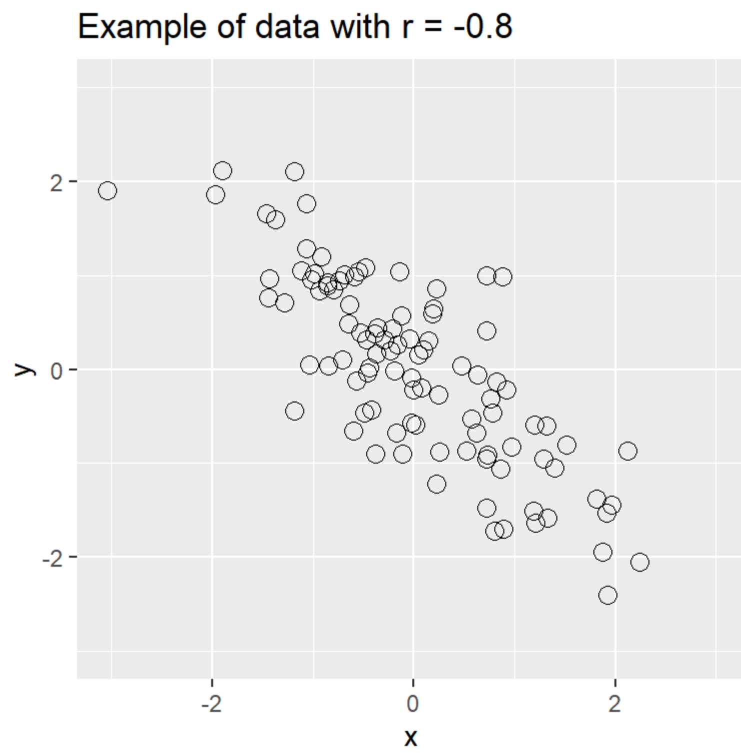
- r is always between -1 and +1
  - Positive values imply positive association
  - Negative values imply negative association
  - Strongest associations closest to -1 or +1

Speaker notes

The correlation is always between -1 and +1. That’s because two measurements cannot co-vary more than the variation that the individual measurements have. It’s pretty easy to show this, actually, but I won’t do it here.

r between -1 and -0.7, strong negative association

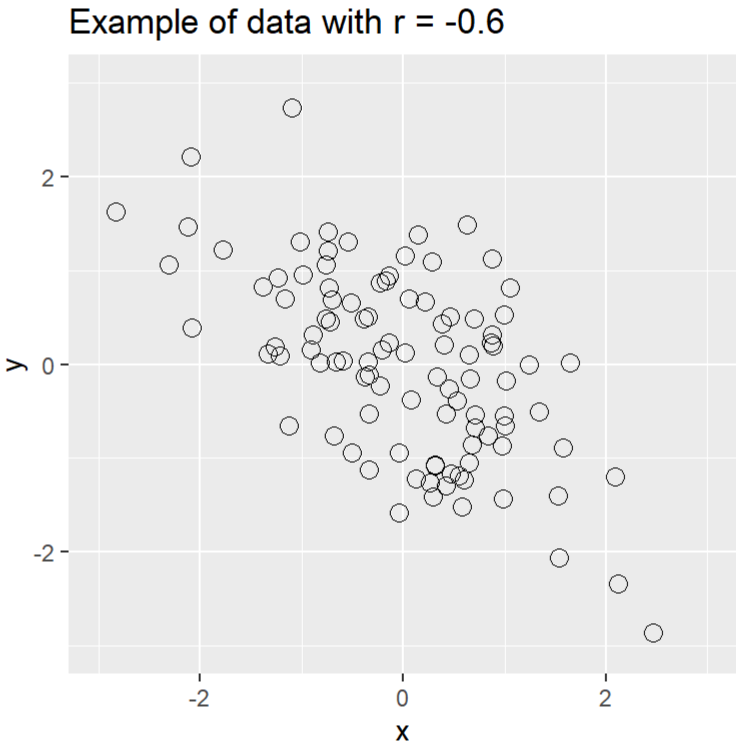


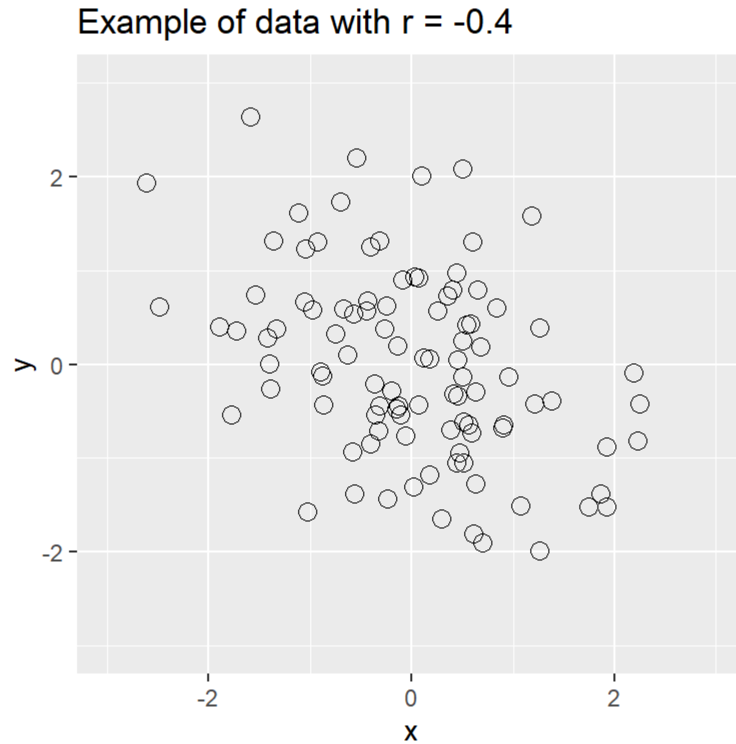


Speaker notes

A correlation close to -1 indicates a strong negative relationship or association. Here are two artificial examples illustrating what a correlation of -0.9 and -0.8 look like.

r between -0.7 and -0.3, weak negative association

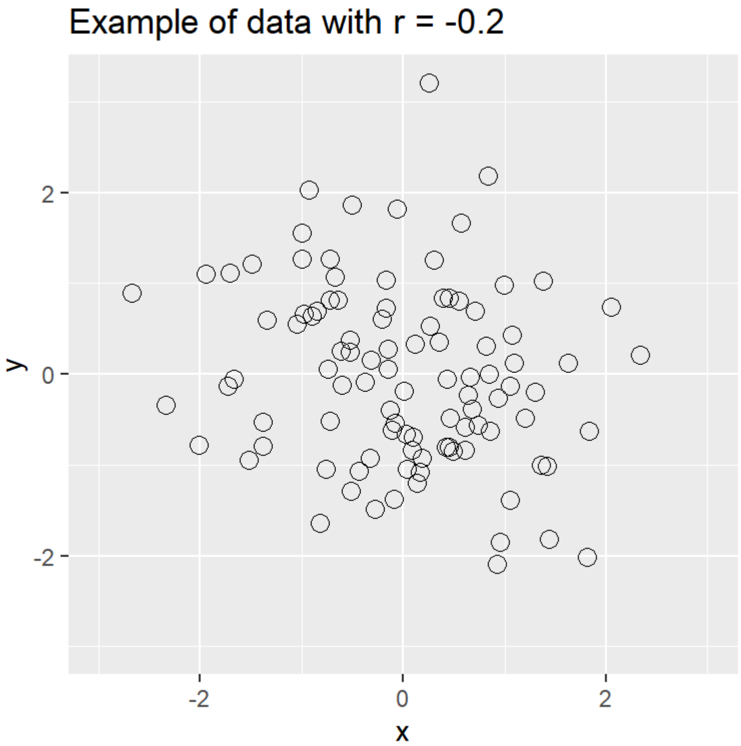


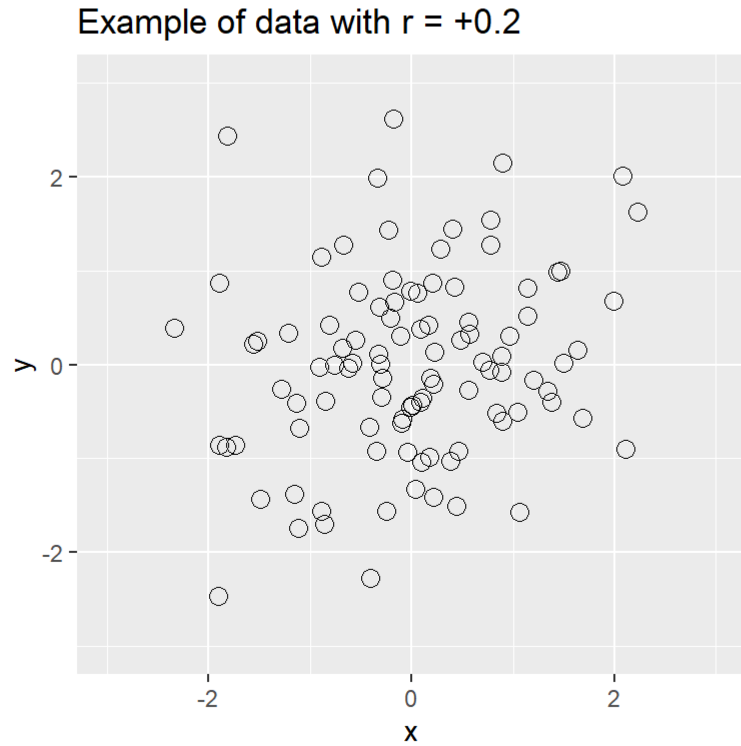


Speaker notes

A correlation between -0.3 and -0.7 indicates a weak negative relationship or association. Here are two examples with correlations of -0.6 and -0.4.

r between -0.3 and +0.3, little or no association

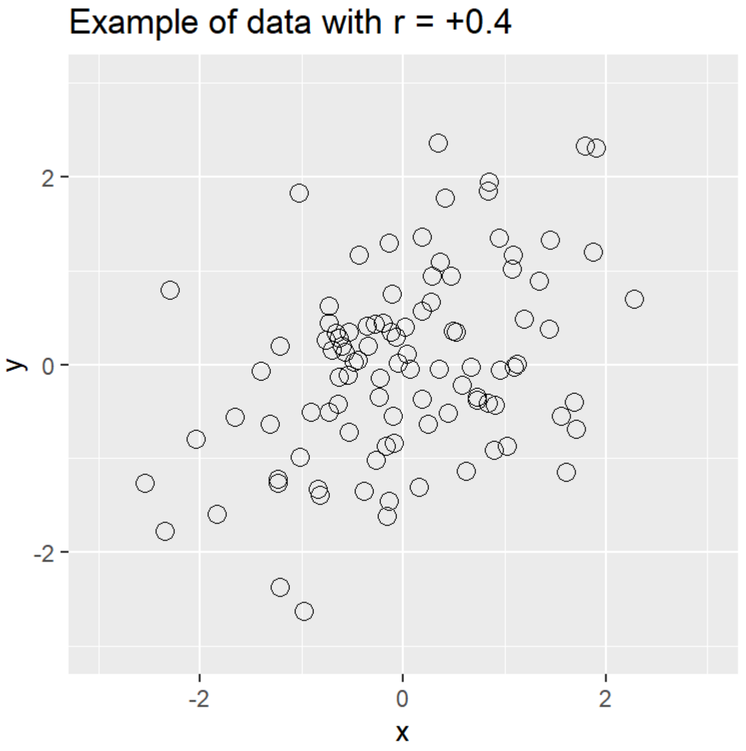


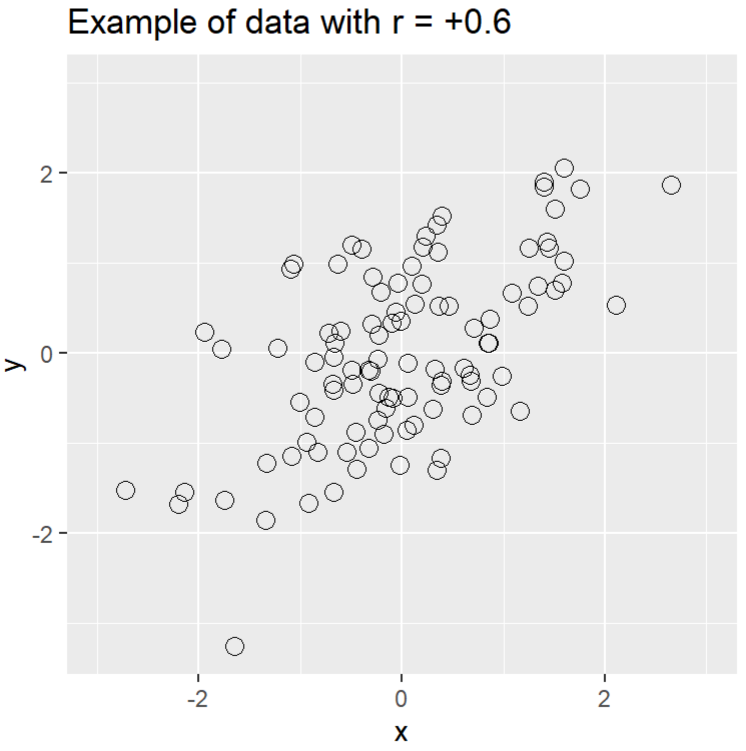


Speaker notes

A correlation between -0.3 and 0.3 indicates little or no association. Here are two examples with correlations of -0.2 and +0.2.

r between +0.3 and +0.7, weak positive association

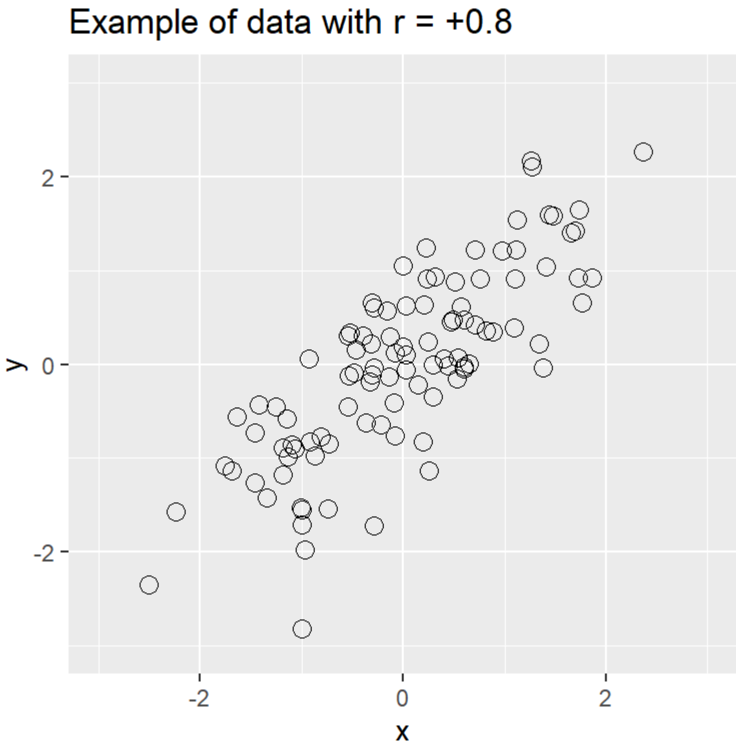


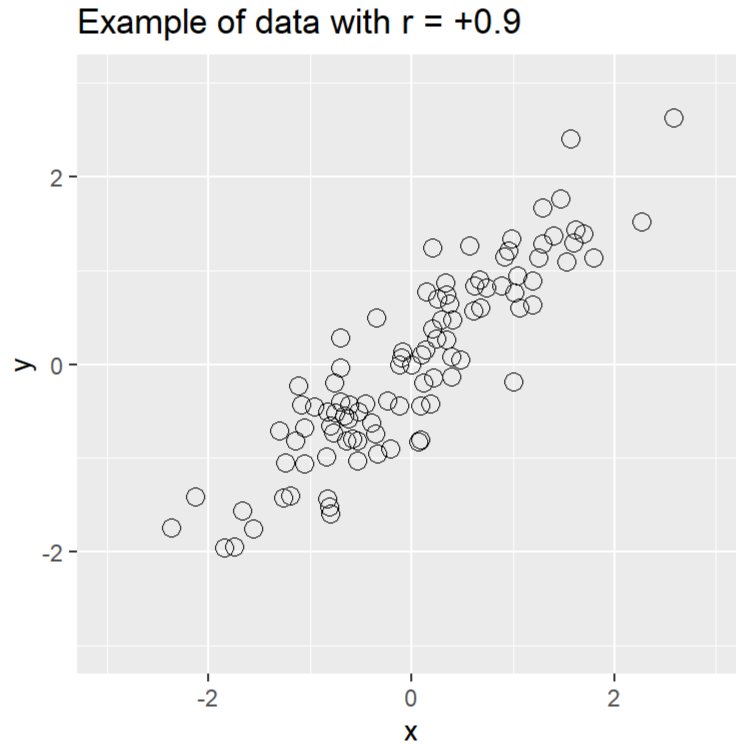


Speaker notes

A correlation between +0.3 and +0.7 indicates a weak positive association. Here are two examples with correlations of +0.4 and +0.6.

r between +0.7 and +1, strong positive association

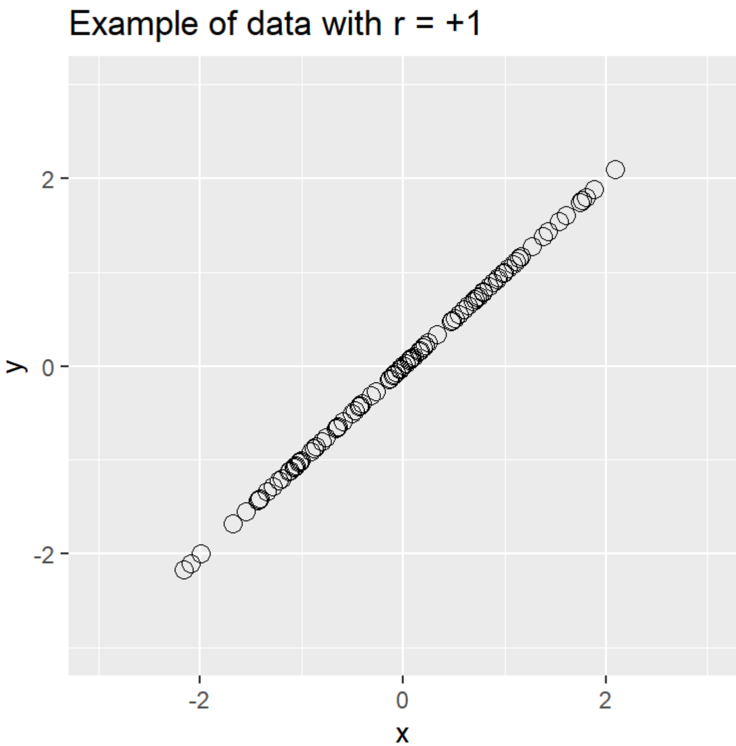


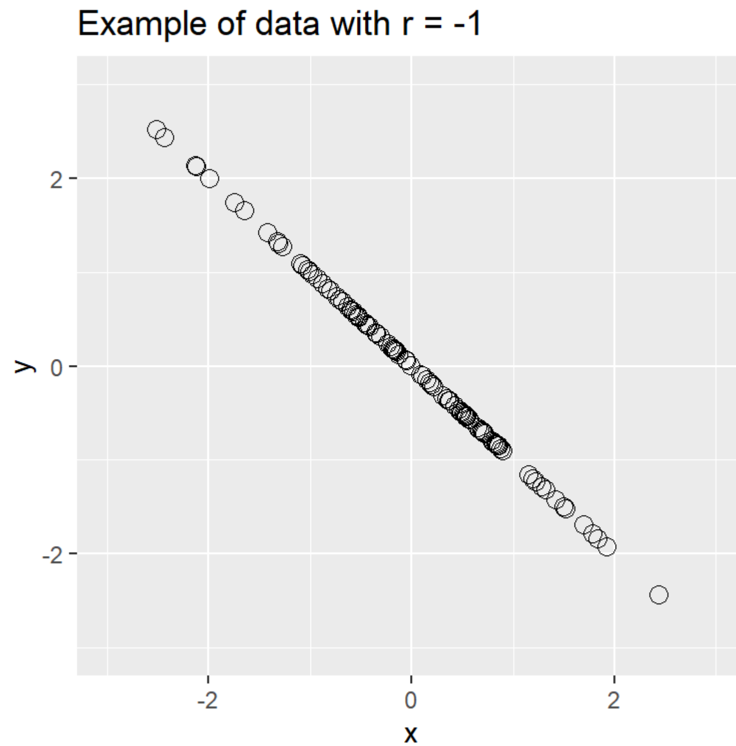


Speaker notes

A correlation close to +1 indicates a strong positive association. Here are two examples with correlations of 0.8 and 0.9.

Extreme case, perfect association





Speaker notes

A correlation equaling +1 or -1 exactly implies a perfect association.

Break #2

- What have you learned
  - Interpretation of the correlation
- What is coming next
  - Missing values

Sleep data dictionary, 1 of 6

```
---
data_dictionary:
  sleep.txt

source:
  This dataset is part of the Austrasian Data and
  Story Library (OZDASL). Please cite this data as
  Smyth, GK (2011). Australasian Data and Story
  Library (OzDASL). http://www.statsci.org/data.
  The data comes originally from Allison, T., and
  Cicchetti, D. V. (1976). Sleep in mammals.
  ecological and constitutional correlates.
  Science 194 (November 12), 732-734.
```

Speaker notes

You will see some practical applications of correlation using a dataset from OzDASL looking at sleep patterns in mammals.

Sleep data dictionary, 2 of 6

```
description:
  This dataset has information about sleep patterns
  in 62 common mammals, along with other information
  that might help you understand what influences
  variations in sleep.

download:
  text-format: http://www.statsci.org/data/general/sleep.txt
  additional-information: http://www.statsci.org/data/general/sleep.html

copyright:
  There is no information about the copyright for this
  dataset. You should, however, be able to use this
  data for individual educational purposes under the
  Fair Use guidelines of U.S. copyright law.
```

Speaker notes

Note that there is no information about how you might use or share this data. This is a common problem with data sources on the Internet.

Sleep data dictionary, 3 of 6

```
format:
  delimiter: tab
  varnames: included in the first row of data
  missing-value-code: NA
  rows: 62
  columns: 11
```

Speaker notes

Here’s the interesting thing about this data. It has missing value codes. I want to talk about missing values in more detail in just a bit.

Sleep data dictionary, 4 of 6


```
vars:
  Species:
    label: Species of mammal

  BodyWt:
    label: Body weight
    unit: kg

  BrainWt:
    label: Brain weight
    unit: g
```

Speaker notes

Here’s the information on the first few variables…

Sleep data dictionary, 5 of 6

```
NonDreaming:
  label: Time spent in non-dreaming sleep
  unit: hours

Dreaming:
  label: Time spent in dreaming sleep
  unit: hours

TotalSleep:
  label: Total time spent in sleep
  unit: hours

LifeSpan:
  unit: years
```

Speaker notes

…and the next set of variables…

Sleep data dictionary, 6 of 6

```
  Gestation:
    unit: days

  Predation:
    scale: likert
    range: 1-5

  Exposure:
    scale: likert
    range: 1-5

  Danger:
    scale: likert
    range: 1-5
---
```

Speaker notes

…and the last few variables.

What does a missing value represent

- Dropout
- Refuse to answer survey question
- Survey question is not applicable
- Lab result is lost
- Concentration below detectable limit
- Many other reasons

Speaker notes

There are many reasons why a data value might be designated as missing. If you are involved with data analysis, you need to understand WHY a data value is missing and adjust the statistical analysis plan appropriately. How you adjust your plan is difficult to say. It does depend a lot on the context.

Common missing value codes

- A single dot (.)
  - SPSS and SAS
- NA
  - R
- Asterisk (*) and other symbols
- Unusual number codes (-1, 9, 99, 999)

Speaker notes

There are a variety of codes for missing values. You will see all of these if you work with data long enough.

A single dot is common, and is the default option in SPSS and SAS. The letters “NA” are also common. This is the default for the R programming language. I’ve seen an asterisk used frequently for missing values.

Also common is the use of unusual number codes. These are numbers outside the range of reasonable values. A negative value is common for many variables that can only take on positive values. A birthweight of -1, for example, either means missing or a baby that floats to the ceiling after it is born.

For other variables a field with one, two, or three nines is common.

Importing missing values

- No problems for default value
- NA and * convert numeric to string
  - Fix during import, or
  - Convert back after import
- Unusual number codes
  - Designate after import
  - **Don’t forget!**

Speaker notes

When you are importing a dataset with missing data into SPSS, use of the default code, a single dot, will usually work just fine.

The problem occurs when the data you get uses a different code, like NA or asterisk. SPSS will often take a column of numbers with one or more missing value codes, and convert it into a string. This makes it impossible for you to run most of the analyses that you would want to run.

You can tell SPSS during the import to designate the column of data as numeric, and SPSS will automatically convert any NA or asterisk to missing. Or you can convert from string to numeric after import. The conversion process will convert your NA and asterisk to missing.

If your dataset uses number codes for missing, you should have no trouble during import, but you do need to designate which numeric code or codes represents a missing value.

Don’t forget this, or any statistics that you compute will be wrong, sometimes very wrong.

Imputing missing values, 1 of 2

- Several simple (simplistic?) imputation choices
  - No news is bad news
  - No news is good news
  - No news is average news (MCAR)
  - No news is last week’s news (LOCF)

Speaker notes

Sometimes you can use a bit of knowledge about the context of your research to help infer what the missing value might be.

A common scenario is what I call “no news is bad news”. If someone drops out of a weight loss study, there’s a good chance it is because the intervention was not effective. You might make a similar assumption for a smoking cessation study. Now you might consider this a bit extreme. But it is a more realistic scenario than some of the other choices that you might make.

You might also consider the opposite scenario, “no news is good news”. If you have information about adverse events in a drug trial, perhaps it is because people forget to say “no problems” more often than they forget to specify problems.

A third scenario is “no news is average news”. This might apply when the reason an observation is missing stems from a cause that is totally unrelated to anything else in the study. You might not have results from a blood test because a technician lost one of your blood samples. That would not be related to any treatment received or any outcome that could have been measured. It was just dumb luck. In such a case, you might replace the missing value with the average of the non-missing values. This case is often called missing completely at random, known by the acronym MCAR.

A final scenario is “no news is last week’s news”. If you are measuring a patient longitudinally, and the patient misses the last visit, you might take the value from the next to last visit and assume that things have not changed too much from the previous week (or month). This is often referred to as last observation carried forward, known by the acronym LOCF.

Imputing missing values, 2 of 2

- Rigorous approaches (beyond the scope of this class)
  - Missing at random (MAR), Missing not at random (MNAR)
  - Ignorable, Non-ignorable
  - Single/Multiple imputation
  - Maximum likelihood/Bayesian approaches
- You cannot ignore missingness, you cannot avoid imputation

Speaker notes

Statisticians will sometimes classify missingness into two other categories besides missing completely at random. There is missing at random, where missingness might depend on some covariates like age and gender. This might allow you to predict missingness using some type of regression model.

If you use regression models to impute the missing values, you need to do it carefully. The regression model itself has some uncertainty associated with it. Failure to account for this uncertainty can lead to falsely precise results. Generally, you need to impute the missing values multiple times (multiple imputation), but practical considerations may force you to impute just once (single imputation).

In contrast, if the data is described as missing not at random, then missingness might be related to the missing value itself. The cases of "no news is bad news" and "no news is good news" are extreme examples of this. Generally, things get messy if a missing outcome is related to what the outcome might have been if you could have observed the value. You often have to make untestable assumptions about your data.

In addition to the regression approaches, there are methods based on maximum likelihood principles or Bayesian models. It turns out that Bayesian models are quite good at handling the messiest case, the missing not at random case.

While I will not expect you to apply these complex approaches, I did want to make you aware of some of the terminology used when discussing missing values.

One point that bears remembering is that you cannot ignore missing values. If you try, you are effectively adopting a "no news is average news" assumption.

It is somewhat analogous to the saying "not to decide is to decide." Ignoring missingness is effectively imputing a value for missing data that is averaging out the missing data. So you will end up imputing anyway.

SPSS investigation of missing data, 1 of 2

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
BodyWt	62	198.78998	899.158011	0	.0	0	10
BrainWt	62	283.1342	930.27894	0	.0	0	9
NonDreaming	48	8.673	3.6665	14	22.6	0	0
Dreaming	50	1.972	1.4427	12	19.4	0	3
TotalSleep	58	10.533	4.6068	4	6.5	0	0
LifeSpan	58	19.878	18.2063	4	6.5	0	2
Gestation	58	142.353	146.8050	4	6.5	0	2
Predation	62	2.87	1.476	0	.0	0	0
Exposure	62	2.42	1.605	0	.0	0	0
Danger	62	2.61	1.441	0	.0	0	0

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Speaker notes

SPSS has a nice set of procedures that help you investigate patterns in missing values that may help you understand the processes behind missingness. This might guide you towards an appropriate method of imputation. Here is one table from SPSS that shows how often values are missing and also tries to identify extreme values. These might be missing value codes that you forget to tell SPSS about, or they might represent values that you have to designate as missing because they are so extreme that they cannot be anything other than a coding error. Examples might be a body mass index of 3.2, which is incompatible with life.

Converting an extreme value to a missing value is something that you should not do with caution and only after discussing this choice with your research team. Sometimes the outlier itself may be the only interesting feature of your data and you don’t want to toss it aside without thinking in depth about it.

SPSS investiation of missing data, 2 of 2

Speaker notes

Another interesting table that SPSS produces shows patterns among missing values. If there is missing values for one variable, how often is a second value missing. In a longitudinal study, for example, some people drop out of the study and stay dropped out. Others may drop back in. They just missed an appointment, but didn’t really stop participating in your study.

Missing value approaches for correlations, 1 of 2

$$
\begin{array}{ccc}
A_1 & B_1 & C_1 \\
A_2 & B_2 & C_2 \\
A_3 & B_3 & C_3 \\
A_4 & B_4 & C_4 \\
A_5 & B_5 & C_5 \\
A_6 & B_6 & .
\end{array}
$$

Speaker notes

When you are computing more than one correlation, you have to face how to handle missing values right away.

Consider a simple scenario with three variables: A, B, and C. Assume that you have all the data (6 rows) for A and B, but only 5 rows for C because the last value is missing.

Missing value approaches for correlations, 2 of 2

- Listwise deletion (complete case analysis),
  - Use 5 pairs for r_{AB}, r_{AC}, and r_{BC}
- Pairwise deletion
  - Use 5 pairs for r_{AC}, and r_{BC}
  - Use all 6 pairs for r_{AB}
- My recommendation
  - Pairwise deletion for descriptive statistics
  - Multiple imputation for inferential statistics
  - Never use complete case analysis

Speaker notes

The two choices you are offered are pairwise deletion and listwise deletion (also known as complete case analysis).

For listwise deletion, you toss out the entire if any of the values in the list of variables is missing. In this example, it means tossing out the sixth row. All correlations, correlations between A and B, between A and C, and between B and C are all based on only five observations.

For pairwise deletion, you still use five observations for the correlation between A and C because C is unknown for one row. Likewise, you use five observations for the correlation between B and C. But you do have six pairs for A and B. Why not use all six pairs?

For simple settings, such as a descriptive analysis, pairwise deletion makes sense. Use the extra data when you have it for certain correlations.

When you use the correlations as part of a more complex inferential analysis, spend the extra time for multiple imputation or something similar like maximum likelihood or a Bayesian approach.

Never use complete case analysis. It provides a small amount of protection compared to pairwise deletion. It still makes some pretty strong assumptions about missing completely at random that are as difficult to justify for complete case analysis as it is for pairwise deletion.

You get a bit more precision with pairwise deletion, but it may cause problems if you do further analyses based on the correlations. In particular, factor analysis (a method beyond the scope of this class), can sometimes produce nonsensical results using correlations with pairwise deletion.

Break #3

- What have you learned
  - Missing values
- What is coming next
  - SPSS calculations of correlations

SPSS correlations with pairwise deletion

Correlations

		BodyWt	LifeSpan	Gestation
BodyWt	Pearson Correlation	1	.302	.651
	Sig. (2-tailed)		.021	<.001
	N	62	58	58
LifeSpan	Pearson Correlation	.302	1	.615
	Sig. (2-tailed)	.021		<.001
	N	58	58	55
Gestation	Pearson Correlation	.651	.615	1
	Sig. (2-tailed)	<.001	<.001	
	N	58	55	58

Speaker notes

Here is a set of correlations from SPSS using pairwise deletion. Notice the sample sizes. 58 for two of the correlations and 55 for the other.

SPSS analysis with listwise deletion

Correlations[a]

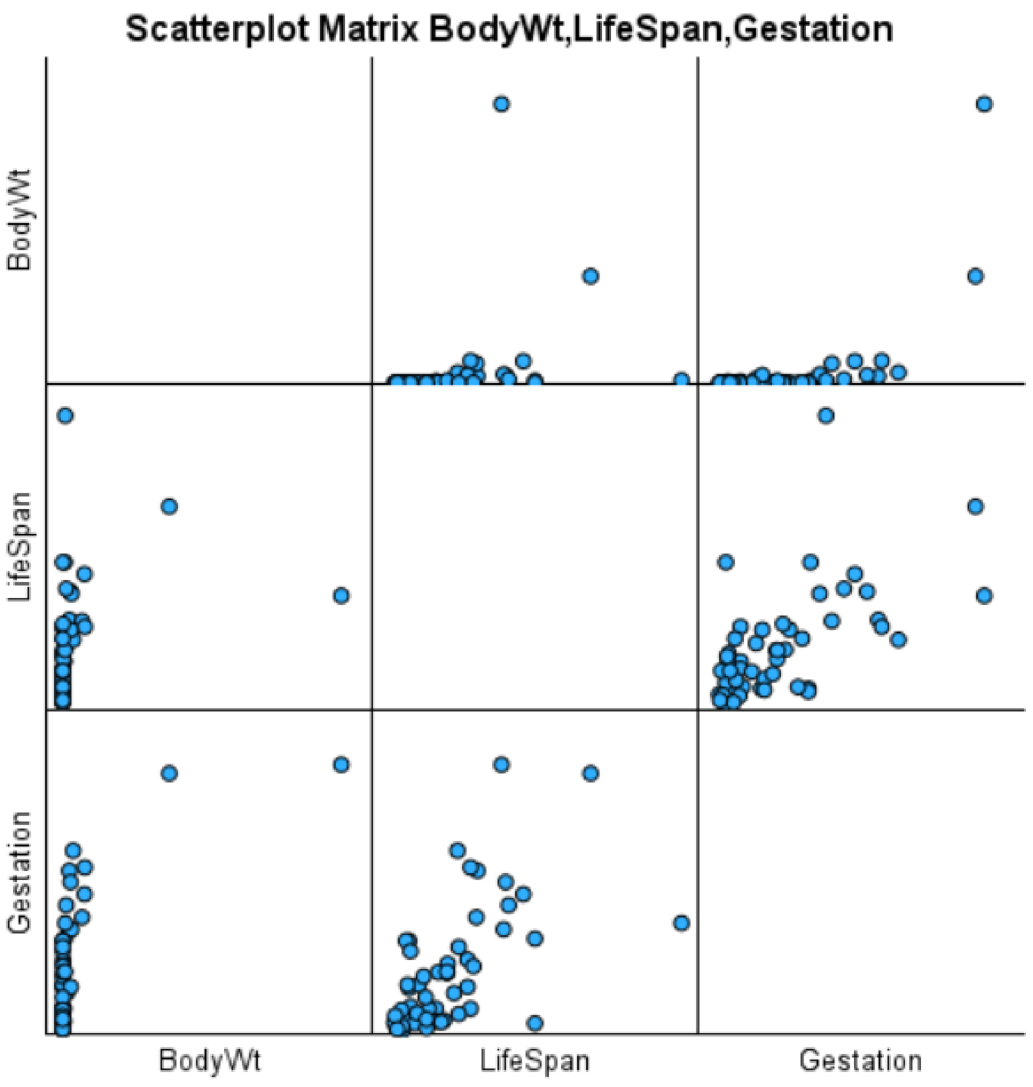
		BodyWt	LifeSpan	Gestation
BodyWt	Pearson Correlation	1	.305	.653
	Sig. (2-tailed)		.023	<.001
LifeSpan	Pearson Correlation	.305	1	.615
	Sig. (2-tailed)	.023		<.001
Gestation	Pearson Correlation	.653	.615	1
	Sig. (2-tailed)	<.001	<.001	

a. Listwise N=55

Speaker notes

With listwise deletion, every correlation is based on 55 observations.

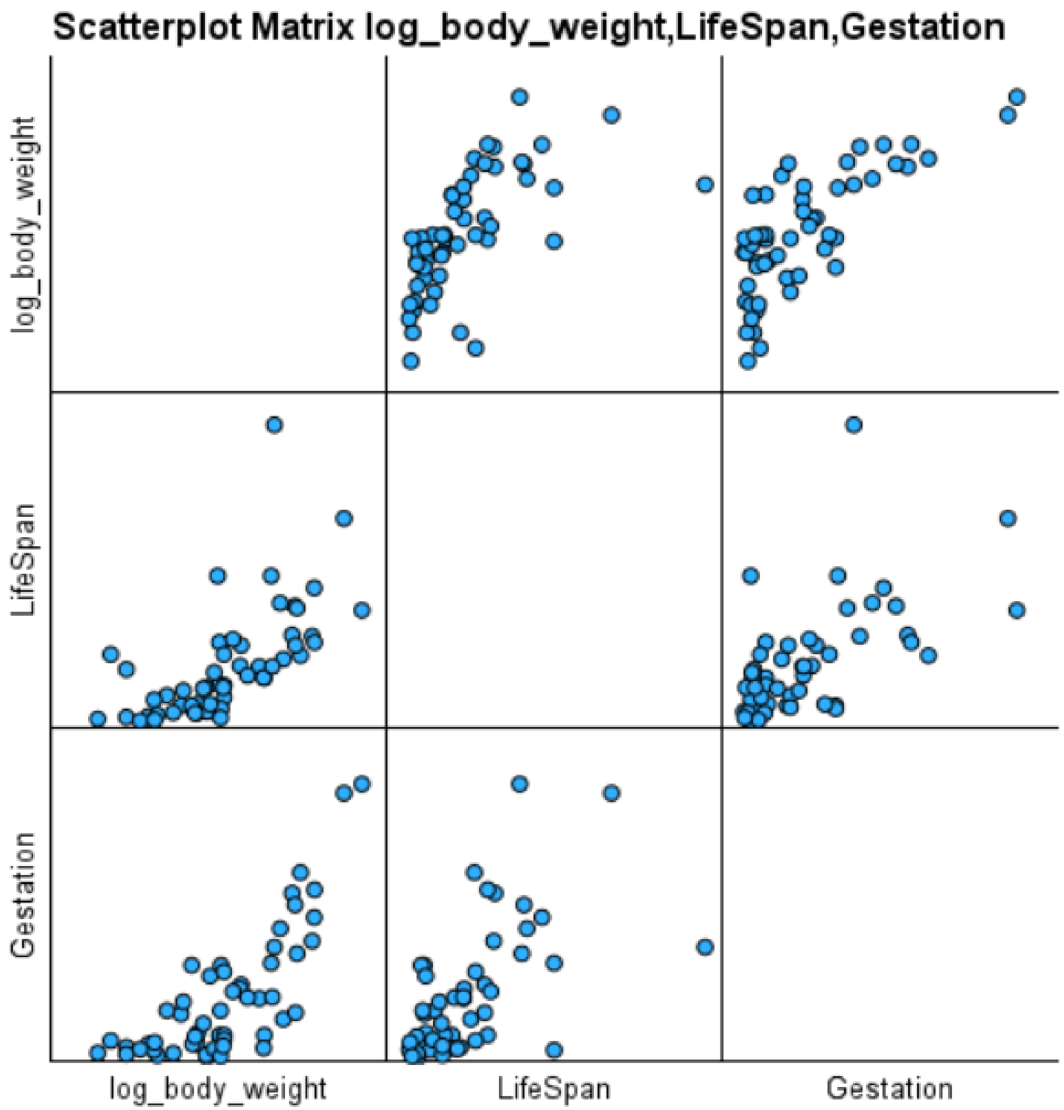
SPSS analysis, scatterplot matrix



Speaker notes

A scatterplot matrix is an interesting alternative to a matrix of correlations. Notice the clustering of body weight values near the low end.

SPSS analysis, 6 of 9



Speaker notes

A log transformation produces a better spread among body weights.

SPSS analysis with a log transformation

Correlations

		log_body_weight	LifeSpan	Gestation
log_body_weight	Pearson Correlation	1	.614	.767
	Sig. (2-tailed)		<.001	<.001
	N	62	58	58
LifeSpan	Pearson Correlation	.614	1	.615
	Sig. (2-tailed)	<.001		<.001
	N	58	58	55
Gestation	Pearson Correlation	.767	.615	1
	Sig. (2-tailed)	<.001	<.001	
	N	58	55	58

Speaker notes

The correlations with log body weight are also different. These correlations are not as strongly influenced by outliers on the high end.

Break #4

- What have you learned
  - SPSS calculations of correlations
- What is coming next
  - Spearman correlation

Spearman correlation

x	y	rank_x	rank_y
11	13	1	5
15	9	2	3
19	15	3	6
21	7	4	2
25	11	5	4
29	5	6	1

Speaker notes

To compute the Spearman correlation, convert the data values into ranks and then compute the correlation of the ranks.

When to use the Spearman correlation

- Similar to considerations for other nonparametric tests
  - Non-normal data
  - Small sample size
  - Ordinal data
- Measures degree of monotonicity

Speaker notes

There is no hard and fast rule about when to use the Spearman correlation. Three important considerations are lack of normality, small sample size and/or ordinal data.

SPSS Spearman correlations, 1 of 2

Correlations

			TotalSleep	Predation	Exposure	Danger
Spearman's rho	TotalSleep	Correlation Coefficient	1.000	-.355**	-.606**	-.524**
		Sig. (2-tailed)	.	.006	<.001	<.001
		N	58	58	58	58
	Predation	Correlation Coefficient	-.355**	1.000	.567**	.918**
		Sig. (2-tailed)	.006	.	<.001	<.001
		N	58	62	62	62
	Exposure	Correlation Coefficient	-.606**	.567**	1.000	.718**
		Sig. (2-tailed)	<.001	<.001	.	<.001
		N	58	62	62	62
	Danger	Correlation Coefficient	-.524**	.918**	.718**	1.000
		Sig. (2-tailed)	<.001	<.001	<.001	.
		N	58	62	62	62

**. Correlation is significant at the 0.01 level (2-tailed).

Speaker notes

Add note.

SPSS Spearman correlations, 2 of 2

Correlations

			BodyWt	log_body_weight	LifeSpan	Gestation
Spearman's rho	BodyWt	Correlation Coefficient	1.000	1.000	.724	.728
		Sig. (2-tailed)	.	.	<.001	<.001
		N	62	62	58	58
	log_body_weight	Correlation Coefficient	1.000	1.000	.724	.728
		Sig. (2-tailed)	.	.	<.001	<.001
		N	62	62	58	58
	LifeSpan	Correlation Coefficient	.724	.724	1.000	.673
		Sig. (2-tailed)	<.001	<.001	.	<.001
		N	58	58	58	55
	Gestation	Correlation Coefficient	.728	.728	.673	1.000
		Sig. (2-tailed)	<.001	<.001	<.001	.
		N	58	58	55	58

Speaker notes

Add note.

Break #5

- What have you learned
  - Spearman correlation
- What is coming next
  - Large correlation matrices

SPSS large correlation matrix

Correlations

		BodyWt	log_body_weight	BrainWt	NonDreaming	Dreaming	TotalSleep	LifeSpan	Gestation
BodyWt	Pearson Correlation	1	.461	.934	-.376	-.109	-.307	.302	.651
	Sig. (2-tailed)		<.001	<.001	.008	.450	.019	.021	<.001
	N	62	62	62	48	50	58	58	58
log_body_weight	Pearson Correlation	.461	1	.540	-.584	-.230	-.533	.614	.767
	Sig. (2-tailed)	<.001		<.001	<.001	.109	<.001	<.001	<.001
	N	62	62	62	48	50	58	58	58
BrainWt	Pearson Correlation	.934	.540	1	-.369	-.105	-.358	.509	.747
	Sig. (2-tailed)	<.001	<.001		.010	.467	.006	<.001	<.001
	N	62	62	62	48	50	58	58	58
NonDreaming	Pearson Correlation	-.376	-.584	-.369	1	.514	.963	-.384	-.595
	Sig. (2-tailed)	.008	<.001	.010		<.001	<.001	.009	<.001
	N	48	48	48	48	48	48	45	44
Dreaming	Pearson Correlation	-.109	-.230	-.105	.514	1	.727	-.296	-.451
	Sig. (2-tailed)	.450	.109	.467	<.001		<.001	.044	.002
	N	50	50	50	48	50	48	47	46
TotalSleep	Pearson Correlation	-.307	-.533	-.358	.963	.727	1	-.410	-.631
	Sig. (2-tailed)	.019	<.001	.006	<.001	<.001		.002	<.001
	N	58	58	58	48	48	58	54	54
LifeSpan	Pearson Correlation	.302	.614	.509	-.384	-.296	-.410	1	.615
	Sig. (2-tailed)	.021	<.001	<.001	.009	.044	.002		<.001
	N	58	58	58	45	47	54	58	55
Gestation	Pearson Correlation	.651	.767	.747	-.595	-.451	-.631	.615	1
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	.002	<.001	<.001	
	N	58	58	58	44	46	54	55	58

Speaker notes

Add note.

Large correlation matrix after reduction, rounding

	A	B	C	D	E	F	G	H	I
1		BodyWt	BrainWt	NonDreaming	Dreaming	TotalSleep	LifeSpan	Gestation	
2	BodyWt		0.9	-0.4	-0.1	-0.3	0.3	0.7	
3	BrainWt	0.9		-0.4	-0.1	-0.4	0.5	0.7	
4	NonDreaming	-0.4	-0.4		0.5	1.0	-0.4	-0.6	
5	Dreaming	-0.1	-0.1	0.5		0.7	-0.3	-0.5	
6	TotalSleep	-0.3	-0.4	1.0	0.7		-0.4	-0.6	
7	LifeSpan	0.3	0.5	-0.4	-0.3	-0.4		0.6	
8	Gestation	0.7	0.7	-0.6	-0.5	-0.6	0.6		

Speaker notes

Add note.

Large correlation matrix after further reduction

	A	B	C	D
1		Gestation	LifeSpan	
2	BrainWt	0.7	0.5	
3	BodyWt	0.7	0.3	
4	Dreaming	-0.5	-0.3	
5	NonDreaming	-0.6	-0.4	
6	TotalSleep	-0.6	-0.4	
7				

Speaker notes

Often it helps to look at a rectangular subregion.

Break #6

- What have you learned
  - Large correlation matrices
- What is coming next
  - Confidence intervals and hypothesis tests

Confidence intervals and hypothesis tests

- r_{XY} is a statistic, ρ_{XY} is a parameter.
  - $H_0: \ \rho_X Y = 0$
  - Accept H_0 if r_{XY} is close to zero, or
  - Accept H_0 if confidence interval includes zero.

Speaker notes

Add note.

SPSS correlation confidence intervals

Correlations

Variable	Variable2	Statistic: Correlation	Count	Lower C.I.	Upper C.I.	Notes
Gestation	TotalSleep	-.631	54	-.769	-.438	
	LifeSpan	.615	55	.418	.757	
	Gestation	1.000	58	--	--	
LifeSpan	TotalSleep	-.410	54	-.611	-.160	
	LifeSpan	1.000	58	--	--	
	Gestation	.615	55	.418	.757	
TotalSleep	TotalSleep	1.000	58	--	--	
	LifeSpan	-.410	54	-.611	-.160	
	Gestation	-.631	54	-.769	-.438	

Missing value handling: PAIRWISE, EXCLUDE. C.I. Level: 95.0

Break #7

- What have you learned
  - Confidence intervals and hypothesis tests
- What is coming next
  - Partial correlations

Partial correlation

- $\rho_{XY \cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}}$

SPSS partial correlation

Correlations

Control Variables			TotalSleep	Gestation
LifeSpan	TotalSleep	Correlation	1.000	-.532
		Significance (2-tailed)	.	<.001
		df	0	48
	Gestation	Correlation	-.532	1.000
		Significance (2-tailed)	<.001	.
		df	48	0

Correlations

Control Variables			TotalSleep	LifeSpan
Gestation	TotalSleep	Correlation	1.000	.008
		Significance (2-tailed)	.	.956
		df	0	48
	LifeSpan	Correlation	.008	1.000
		Significance (2-tailed)	.956	.
		df	48	0

Summary

- Calculation of the covariance and correlation.
- Interpretation of the correlation
- Missing values
- SPSS calculations of correlations
- Spearman correlation
- Large correlation matrices
- Confidence intervals and hypothesis tests
- Partial correlations