Lancaster University

Quantitative Research Methods

Matthew Ivory

matthew.ivory@lancaster.ac.uk

1. Introduction to the course

- Session 1: Introduction to quantitative research methods using R
- Session 2: Data management and data wrangling
- Session 3: Exploratory data analysis
- Session 4: Data visualization
- Session 5: Live Coding Walkthrough
- Session 6: Probability and distributions
- Session 7: Tests for discrete variables: Analysing contingency tables
- **Session 8: Correlations and t-tests**
- Session 9: ANOVA and linear regression
- Session 10: Multiple regression, introduction to generalised linear regression

In this session

- Hypothesis Refresher
  - probabilities
- Correlations
- t-tests
- Confidence Intervals

Hypothesis Testing

- The null hypothesis states that there is no difference between the two means (or groups) of interest
- Hypotheses can be directional or non-directional
- Directional means that we are predicting a difference between groups as well as the direction of the effect
- Non-directional means we predict a difference but don't know which way it will go
- Hypothesis should be worded so that it can be tested and must include both independent and dependent variables
  - **Exam scores (DV)** will be higher for **students who studied more than four hours a week (IV)** than those who did not

Hypothesis Testing and *p*-values

- The ***p*-value** is the probability of finding a difference equal to or greater than the one found if no difference exists in the population. Let's say our *p*-value = .010
- This indicates a very small probability of finding a difference equal or greater in the population if there was no difference. The obtained *p*-value is also smaller than the standard cut-off that we use in Psychology of .05
- As such we would reject our null hypothesis and suggest that there is a significant difference between the two groups.
- *If* the p-value was *greater* than .05 (e.g. .1),we would *fail to reject* the null hypothesis, **not** *accept the null*

Effect sizes and *p*-values

- When we construct an experiment, we construct a research hypothesis:
  - Two variables are related
  - An experimental manipulation will affect another variable
- We also construct a null hypothesis:
  - Two variables are not related
  - The experimental manipulation has no effect
- *p*-value: how confident can we be in rejecting the null hypothesis
  - A p-value of 0.05 tells us that 5% of the time when we find an effect like this, the null hypothesis is true (a **false positive**)
  - If the p-value is less than 0.001, we write by convention: $p < .001$
    - If there's a chance less than one in a thousand that we find this effect when the null hypothesis is true then we're quite confident in our result

Correlations

- Correlation is a test that measures the relationship between two variable.
- You measure two variables and the correlation analysis tells you whether they are related in some manner, either positively or negatively, and how strong that relationship is
- We can conduct correlations on data when variables are continuous or ordinal

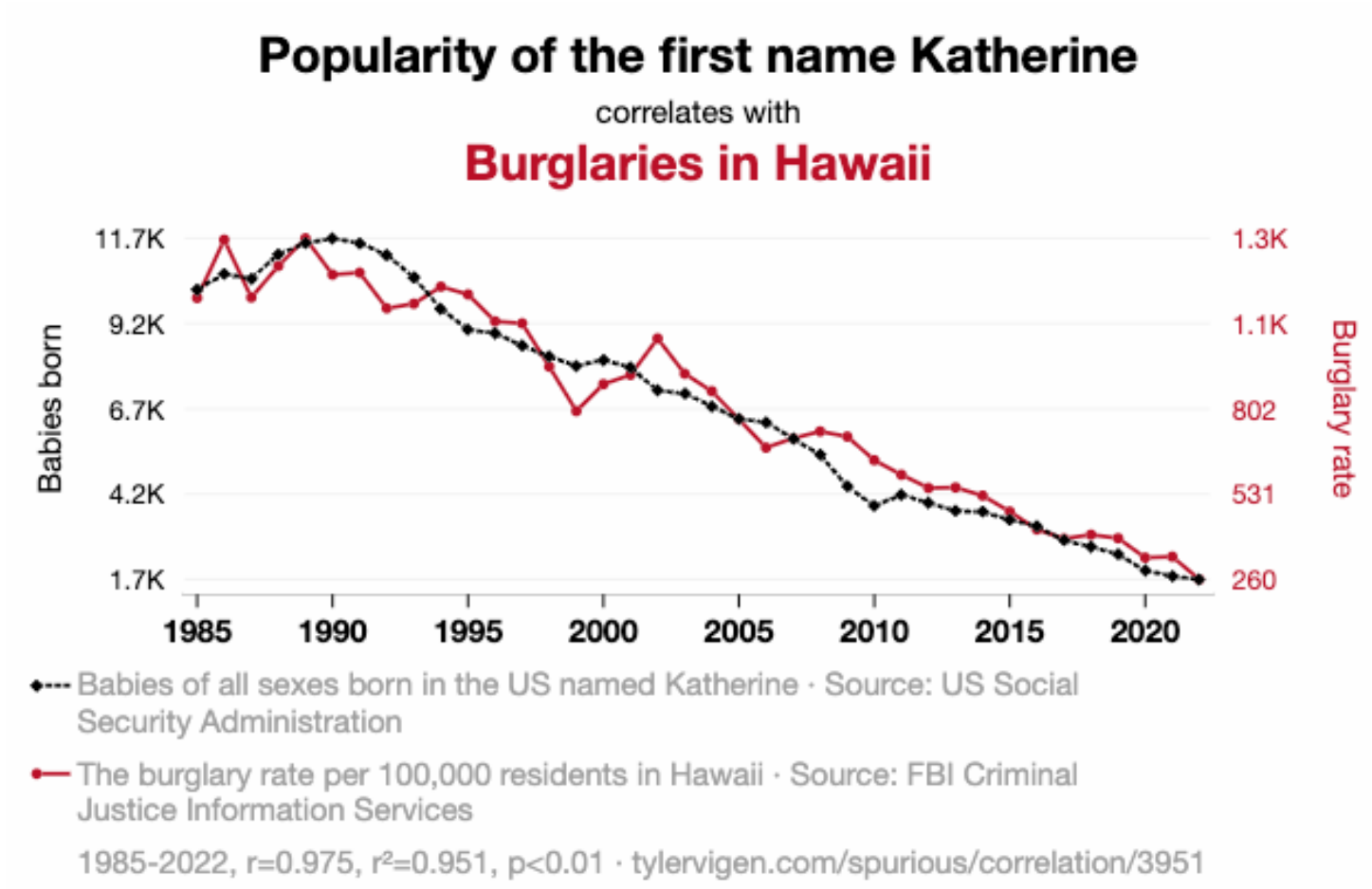
Correlations

- When dealing with correlations you should always refer to relationships and not predictions. In a correlation, X does not predict Y, nor does X cause an effect in Y
- **Correlation != Causation**
- For correlation, all we can say is that X and Y are related/associated/correlated

Correlation Examples

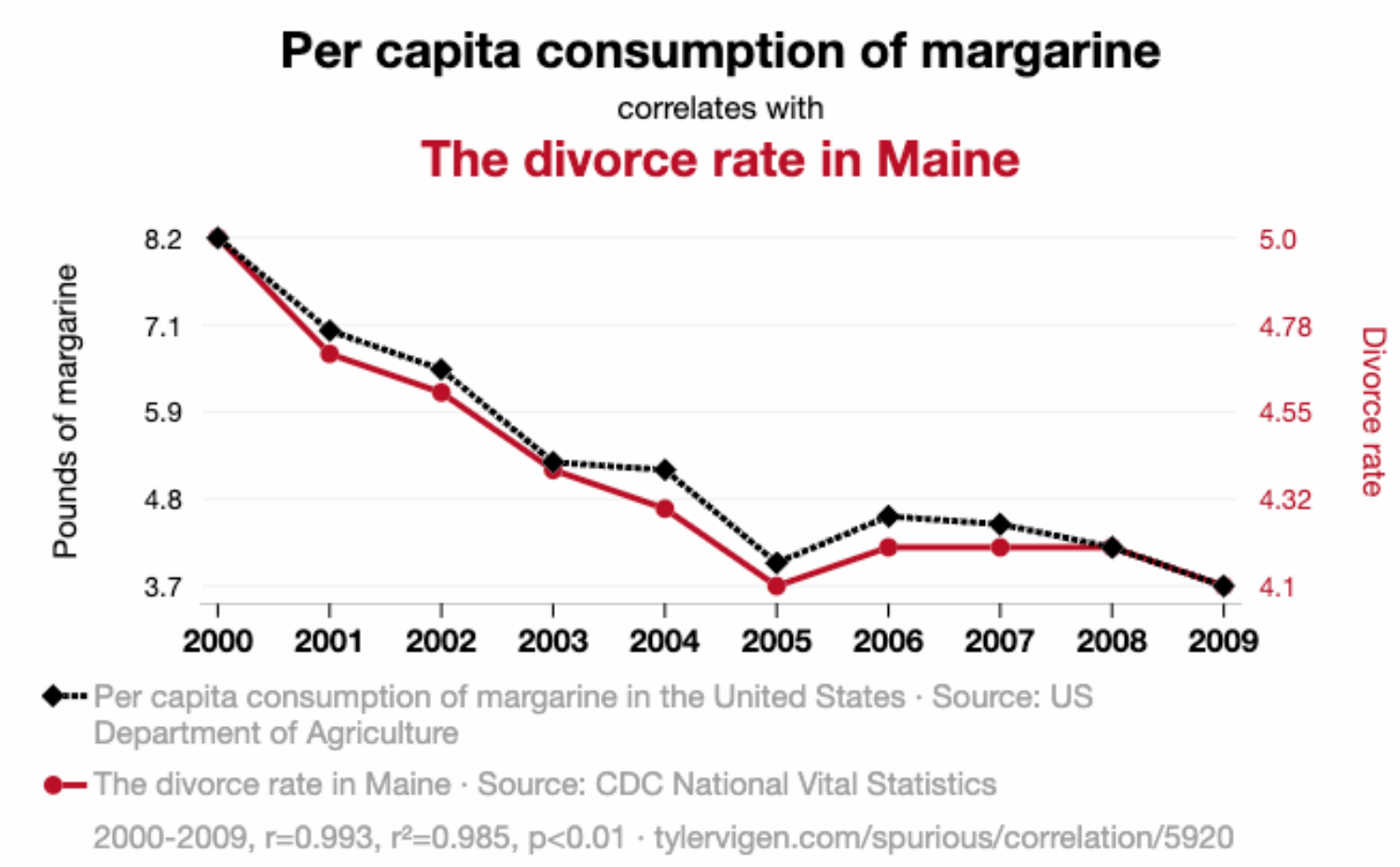
https://www.tylervigen.com/spurious-correlations

Bad Correlation Examples



https://www.tylervigen.com/spurious-correlations

Bad Correlation Examples



https://www.tylervigen.com/spurious-correlations

Bad Correlation Examples

- Why do these correlations exist?
- In the two examples above: a combination of scientific dishonesty, ‘creative’ data wrangling, and **confounds**
- A *confounding variable* is an unmeasured variable that influences both the dependent variable and the independent variable

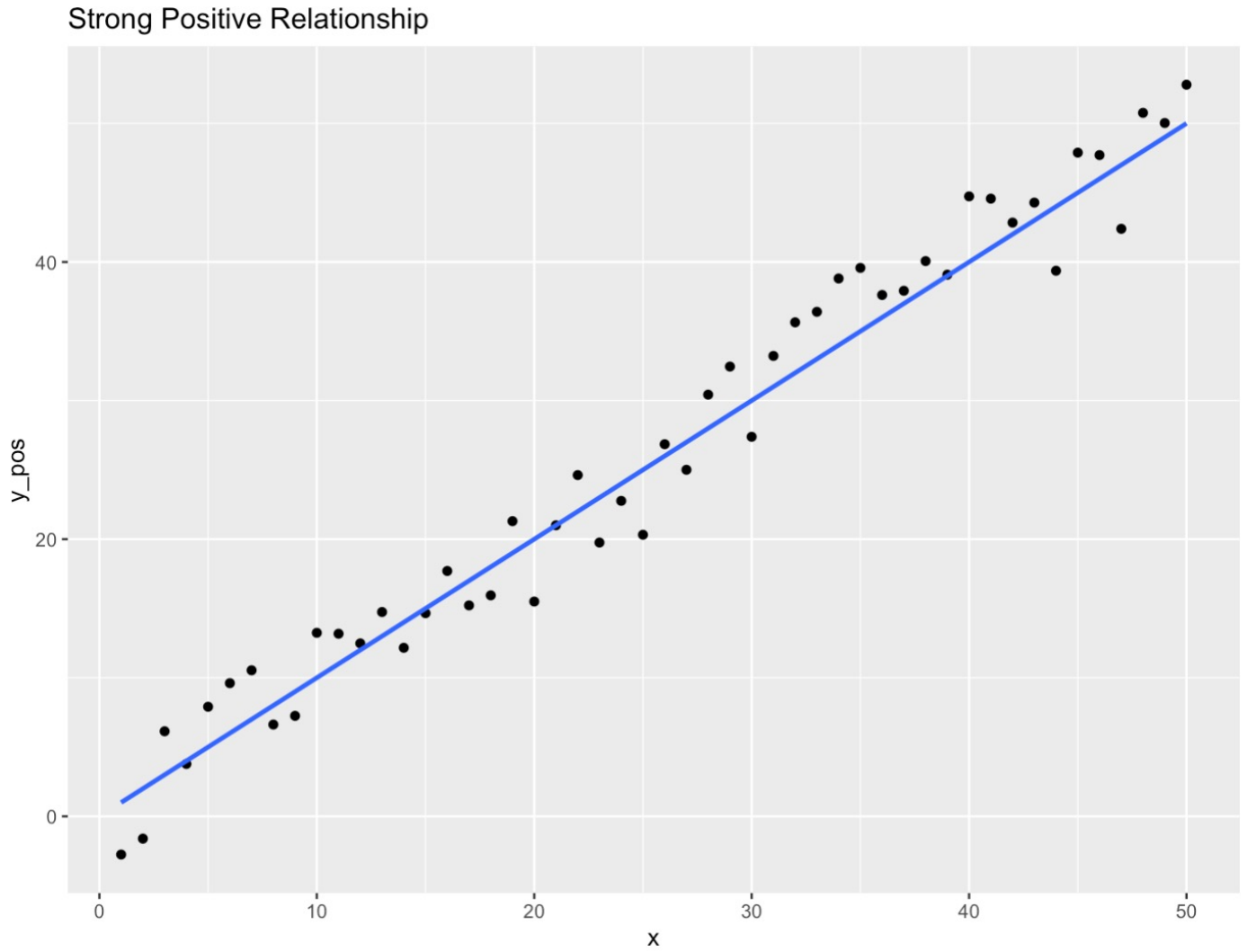
- Independent Variables (IV) – is independent to other variables in the study
- Dependent Variables (DV) – is dependent on changes in independent variables
  - The DV is what we are interested in explaining

Correlation

- So we must be careful about what claims or inferences that we make, and to consider possible confounds
- Now, let's look at some better and more accessible examples of correlations…

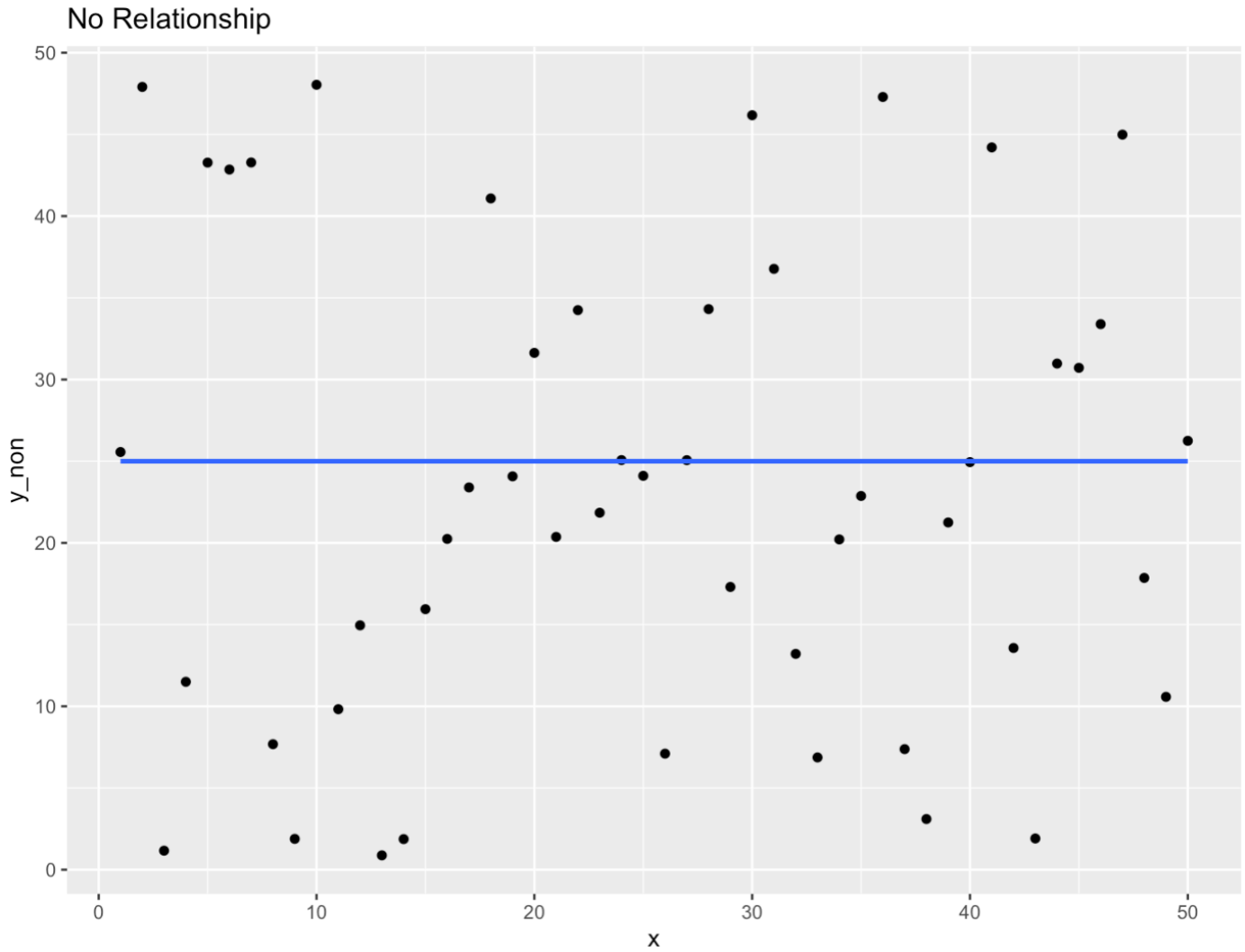
Types of Correlation

- As X increases, Y increases



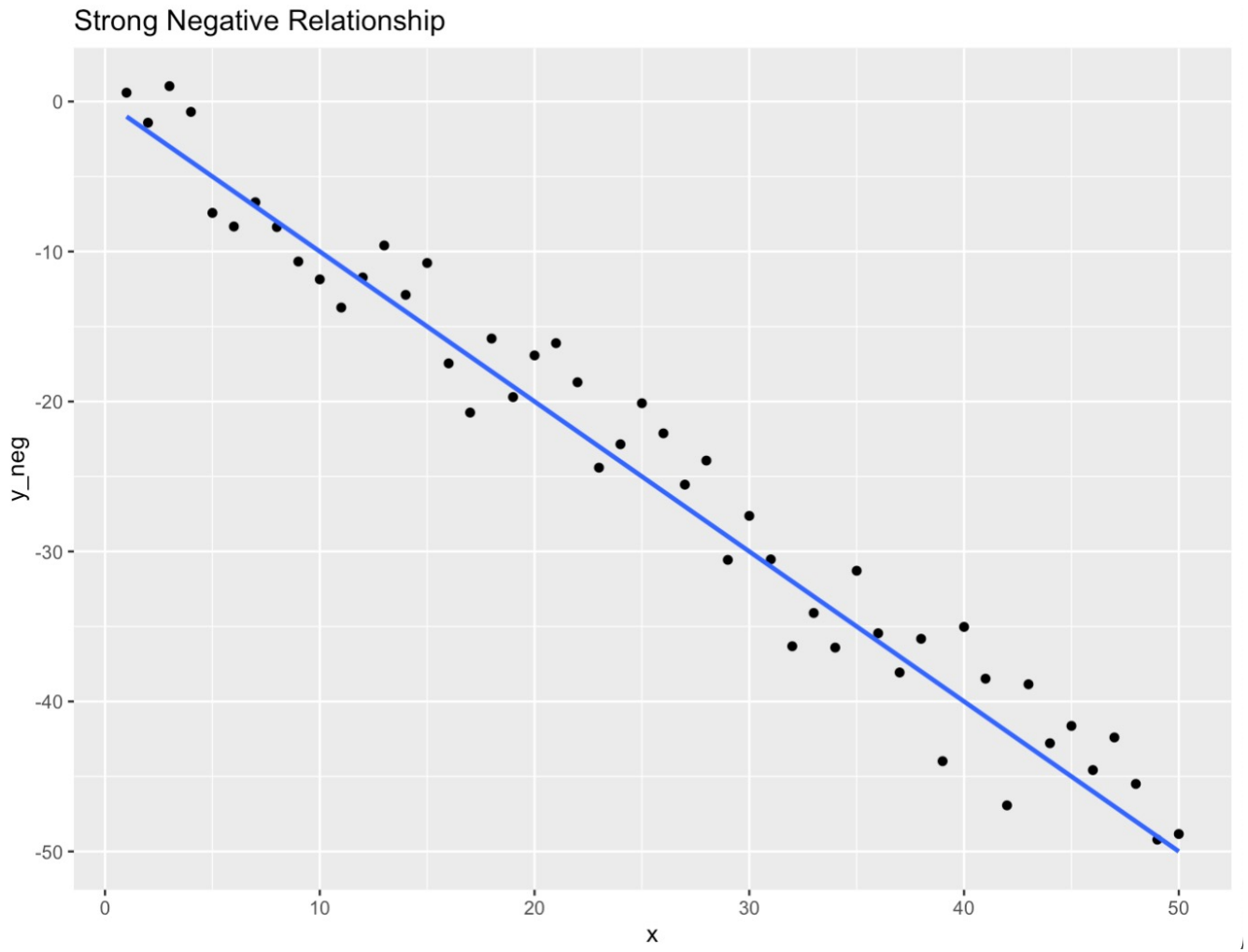
Types of Correlation

- As X increases, Y shows no pattern or correlation



Types of Correlation

- As X increases, Y decreases



Assumptions of the test

- before running an analysis, we should check the assumptions of the test
- Assumptions are the checks that the data must pass before we can use it
- The assumptions change depending on the test
- You should only use a given test based on how well the data meets the assumptions
- Every statistical test relies upon assumptions regarding the data
  - Some are more lenient than others
  - Others are very strict

Assumptions of the test

For correlations, the main assumptions we need to check are:

1. Is the data interval, ratio, or ordinal?
2. Is there a data point for each participant on both variables?
3. Is the data normally distributed in both variables?
4. Does the relationship between variables appear linear?
5. Does the spread have homoscedasticity?

Assumptions of the test

For correlations, the main assumptions we need to check are:

1. **Is the data interval, ratio, or ordinal?**

- If we want to run a Pearson correlation then we need interval or ratio data; Spearman correlations can run with ordinal, interval or ratio data.

Assumptions of the test

For correlations, the main assumptions we need to check are:

1. Is the data interval, ratio, or ordinal?
2. **Is there a data point for each participant on both variables?**

- all correlations must have a data point for each participant in the two variables being correlated. This should make sense as to why - you can't correlate against nothing

Assumptions of the test

For correlations, the main assumptions we need to check are:

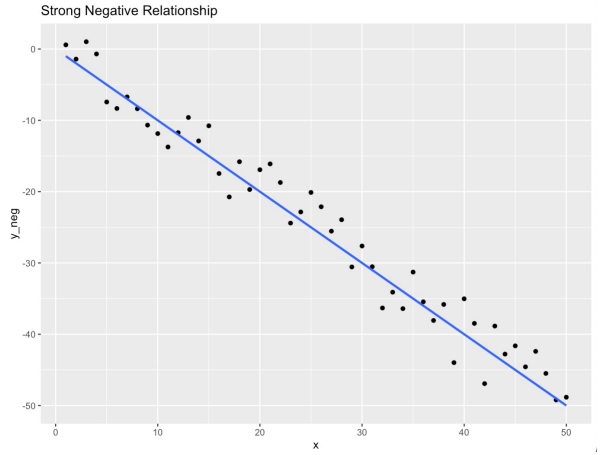
1. Is the data interval, ratio, or ordinal?
2. Is there a data point for each participant on both variables?
3. **Is the data normally distributed in both variables?**

- We went through distributions last week, if we plotted a histogram of the data, they should resemble a normal distribution (we will try this out in the worksheet)

Assumptions of the test

For correlations, the main assumptions we need to check are:

1. Is the data continuous, or ordinal?
2. Is there a data point for each participant on both variables?
3. Is the data normally distributed in both variables?
4. **Does the relationship between variables appear linear?**



- If we were to plot the variables against each other, would the line look straight?

Assumptions of the test

For correlations, the main assumptions we need to check are:

1. Is the data interval, ratio, or ordinal?
2. Is there a data point for each participant on both variables?
3. Is the data normally distributed in both variables?
4. Does the relationship between variables appear linear?
5. **Does the spread have homoscedasticity?**

- Homoscedasticity is an assumption of equal or similar variances in different groups being compared
- The error within the data is equal between the DV and IV

Homoscedasticity

- Homoscedasticity is an assumption of equal or similar variances in different groups being compared
- The error within the data is equal between the DV and IV
- When we take measurements, we get our observations
- Observations are made up of the ‘true’ score and additional error terms
- The error term can be made up of:
  - instrument accuracy e.g. scales that vary by ±100g
  - Researcher recording error e.g. too slow at stopping a timer
  - Participant error e.g. mistakes in completing surveys
  - Anything really that isn’t the true reality of what we are trying to measure

Two Correlation Tests

- Pearson's product-moment correlation (Pearson's R)
  - Used for Continuous data
- Spearman's rank correlation coefficient (Spearman's Rho)
  - Used for Ordinal or Continuous data

- You will have greater exposure to the correlations, how we interpret them and how to report them in the worksheet
- For now… t-tests!

Switching Gears

- We just looked at correlations – where we correlate continuous data against continuous data (or ordinal)
- But what if we want to compare continuous data between two groups or conditions?
- Then we can use **t-tests**
- Useful when we want to know if there is an effect of a given variable

One sample t-tests

- Used to compare a single group against a known norm
- “does the mean of the sample deviate significantly from the expected norm”
- E.g., comparing the mean IQ of a group against the expected population norm of 100

Independent-samples t-tests

- Otherwise known as between-subjects
- Different participants in different conditions
- Comparing the means of two groups against each other
- Test scores between two classes

Dependent-samples t-tests

- within-subjects, dependent-samples, paired-samples, or repeated-measures
- the same participants in all conditions
- Each participant experiences each condition
- E.g., each subject completes an easy, medium, and hard test

Matched-pairs t-tests

- Different people in different conditions but you have matched participants across the conditions so that they are “effectively the same person” (e.g. age, IQ)
- Used where you need to make sure that participants don’t demonstrate practice effects
- E.g., matching participants on age and IQ to test the effect of condition (maybe levels of glucose) on test scores

Assumptions of the test

- Two groups of data/participants (More than 2? Wait until next week)
- The data are continuous.
- The sample data have been randomly sampled from a population.
- There is homoscedasticity (i.e., the variability of the data in each group is similar).
- The distribution is approximately normal.

Confidence Intervals

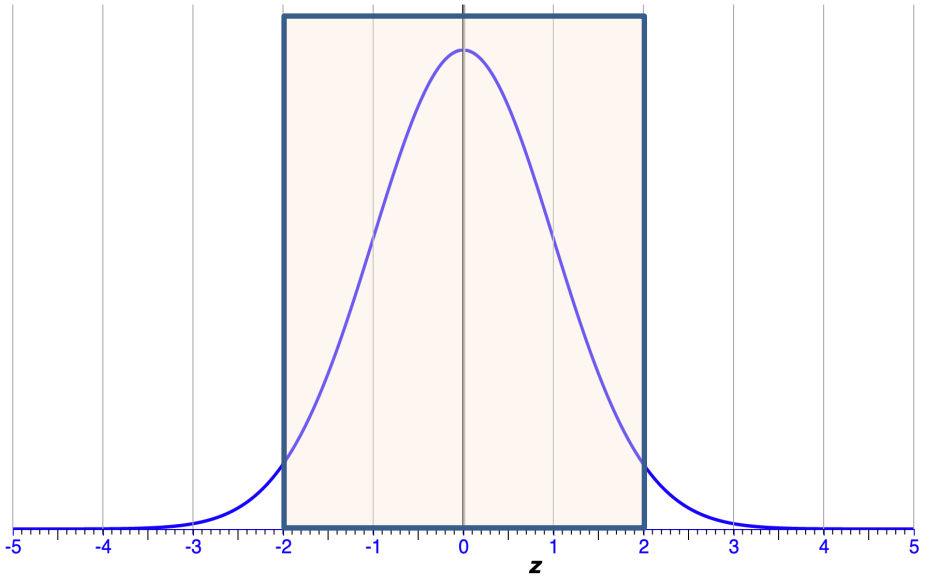
- Reported alongside our coefficients or estimates
- They give a 95% (typically) interval where we would believe the true population coefficient to lie.
- “I am 95% confident that the true value falls between point A and point B
- Uncertainty in an estimate of a population parameter based on a sample

- when a set of observations have a Normal distribution multiples of the standard deviation mark certain limits on the scatter of the observations. For instance, 1.96 (or approximately 2) standard deviations above and 1.96 standard deviations below the mean (±1.96SD mark the points within which 95% of the observations lie.

Confidence Intervals

- Assuming a normal distribution
- Recall that ±2 SDs contains 95% of our observations
- Offer an upper and lower bound that contains the *true* mean 95% of the time

- Why not 100% confidence? Because we are talking on probabilities



Worksheet this week

- Correlations
- t-tests
- Confidence intervals
- Writing up results for a report
- Interpreting the findings