Lecture 2: Describing Bivariate Relationships

Dr. Rachel Blum

October 29, 2019

Table of contents

What are descriptive statistics?

Bivariate relationships

Today we're only looking at ways to describe relationships between two variables,

a.k.a. bivariate relationships.

How would we describe this variable?

How many emails you've sent me this semester:

1, 3, 5, 2, 0, 5, 1, 1, 8, 6, 2

Options

Mean The typical or average value of the variable.

Median The halfway point. A random observation will be above this value half of the time, and below it half of the time.

Mode If we picked a random observation, this would be the most likely result.

Mean

$$\frac{\sum_1^N X_i}{N} =$$

$$\frac{1+3+5+2+0+5+1+1+8+6+2}{11} =$$

$$\frac{34}{11} =$$

$$3.09$$

Where x_i is each data point from 1 to N, and N is the number of observations.

Median

Order your data from lowest to highest and fine the middle number.
If no one middle exists, take the average of the two middle numbers.

```
sort(0, 1, 1, 1, 2, 2, 3, 5, 5, 6, 8)
```

What is the median?

Mode

Find the most common value.

```
sort(0, 1, 1, 1, 2, 2, 3, 5, 5, 6, 8)
```

What is the mode?

Mean, median, and mode

- The mean number of emails per student so far has been 3.
- The median number of emails per student has been 2.
- The modal number of emails per student has been 1.

Mean, median, and mode

- Mode is useful when you have variables with few attributes.
- Mean and median will often be similar. If the two diverge substantially, use the median.
- What about categorical variables?

Review: Variable Types

1. **Categorical** (aka *nominal*): can be put in categories. Order is arbitrary.
    - **Ordinal** (aka *ranked*): categorical, but with a clear order.
    - **Binary** (aka *dummy* or *indicator*): two categories, reference and comparison. Order is arbitrary.
2. **Continuous** (aka *interval*): potentially infinite number of values. Order matters.
    - **Ratio**: similar to continuous/interval variables, but with meaningful zero. Order matters.

What to do with categorical data? Make tables!

- Main title should have descriptive variables names
- Each row should be a different attribute

Party	2016 vote
Democrat	Clinton
Democrat	Clinton
Republican	Trump
Democrat	Clinton
Republican	Trump
Democrat	Clinton
Republican	Trump
Independent	Stein
Republican	Trump
Libertarian	Johnson

What to do with categorical data? Make tables!

- It is useful to list the total number of observations at the bottom.
- You can show frequencies.

Party	Frequency
Democrat	4
Republican	4
Independent	1
Libertarian	1
N	10

What to do with categorical data? Make tables!

- You can also show proportions.

Party	Proportion
Democrat	40%
Republican	40%
Independent	10%
Libertarian	10%
N	10

Forming and testing hypotheses

Why do hypothesis testing?

Our goal = making statements about a *population* of interest.

Our challenge = only having access to information about particular observations from a *random sample* from that population.

Complications: error

Randomness, or *error*, comes in many forms.

Complications: error

Randomness, or *error*, comes in many forms.

1. **Sampling error:** Samples are only a single snapshot of the population and may not accurately represent the population.

Complications: error

Randomness, or *error*, comes in many forms.

1. **Sampling error:** Samples are only a single snapshot of the population and may not accurately represent the population.
2. **Measurement error:** We cannot perfectly measure most concepts/phenomena.

Complications: error

Randomness, or *error*, comes in many forms.

1. **Sampling error:** Samples are only a single snapshot of the population and may not accurately represent the population.
2. **Measurement error:** We cannot perfectly measure most concepts/phenomena.
3. **Complexity:** There will often be sources of error that we can't account for because we don't even know about them.

An answer, sort of

The framework of HYPOTHESIS TESTING allows us to evaluate whether there is evidence that our findings are real and not merely **the result of random chance.**

What can hypothesis tests answer?

Hypothesis testing is all about answering *questions about probability*:

How likely is it that I would have drawn a sample that looks like the one that I got, if in fact the null hypothesis were true?

Does **NOT** answer: "is the null hypothesis true?" or: "is the null hypothesis false?"

Steps in hypothesis testing

1. Formulate hypothesis
2. Draw a random sample from population of interest
3. State a null hypothesis
4. Choose a level of statistical significance (*decision rule for areas under sampling distribution that include unlikely sample outcomes)*
5. Compute test statistic (*Z score, t-statistic,* χ^2 *statistic, F-statistic, etc.)* and compare with critical value

Example

Hypothesis: Congressional campaigns in 2018 spent more than congressional campaigns in 2016.

Example

Hypothesis: Congressional campaigns in 2018 spent more than congressional campaigns in 2016.

Where could error come from?

Null and alternative hypotheses

$$H_{null} = Spending_{2018} \leq Spending_{2016}$$

There was no difference between 2016 and 2018, or
Candidates in 2016 spent *more* than candidates in 2018.

$$H_{alt} = Spending_{2018} > Spending_{2016}$$

2018 candidates spent more than 2016 candidates.

More on the null

Two types of null hypotheses:

1. **One-tailed**: If your alternative hypothesis is that something is greater or lesser than another thing, your null will always be $\leq$ or $\geq$ (it will be in one direction only).
2. **Two-tailed**: If your alternative hypothesis is that something is equal to another thing, your null will always be $=$.

Can you:

Think of examples of an one-tailed hypothesis?

A two tailed?

P-values

What is a good p-value?

- **p-value** $< 0.1 =$ weak evidence
- **p-value** $< 0.05 =$ good evidence (generally accepted in social science)
- **p-value** $< 0.01 =$ strong evidence

Confidence intervals

Quite simply, a 95% confidence interval includes the true value 95% of the time.

REMEMBER

1. You cannot find evidence *for* or *prove* the alternative hypothesis.
2. You can only find evidence that allows you to **reject** the null hypothesis.

Difference of means

AKA t-tests

- Difference of means tests involve comparing the mean of Y for one group in our sample against the mean of Y for a different group in our sample.
- We need a *dichotomous* variable and a *continuous* variable.

$$\bar{Y}_1 - \bar{Y}_0 \sim t_{df}$$

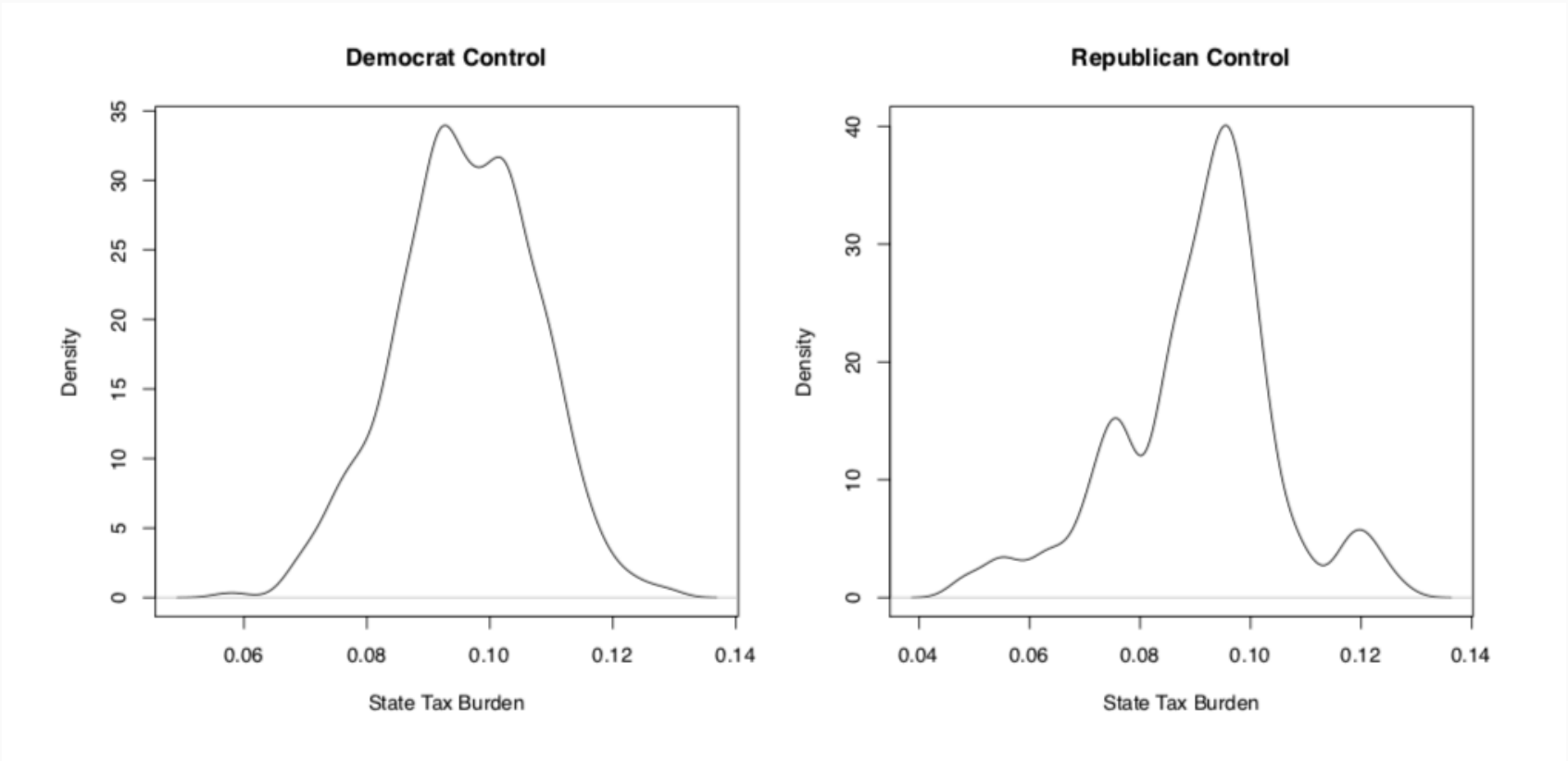
Example: tax burdens

Hypothesis: States where Democrats control the lower house will have higher tax burdens than states that control where Republicans control the lower house.

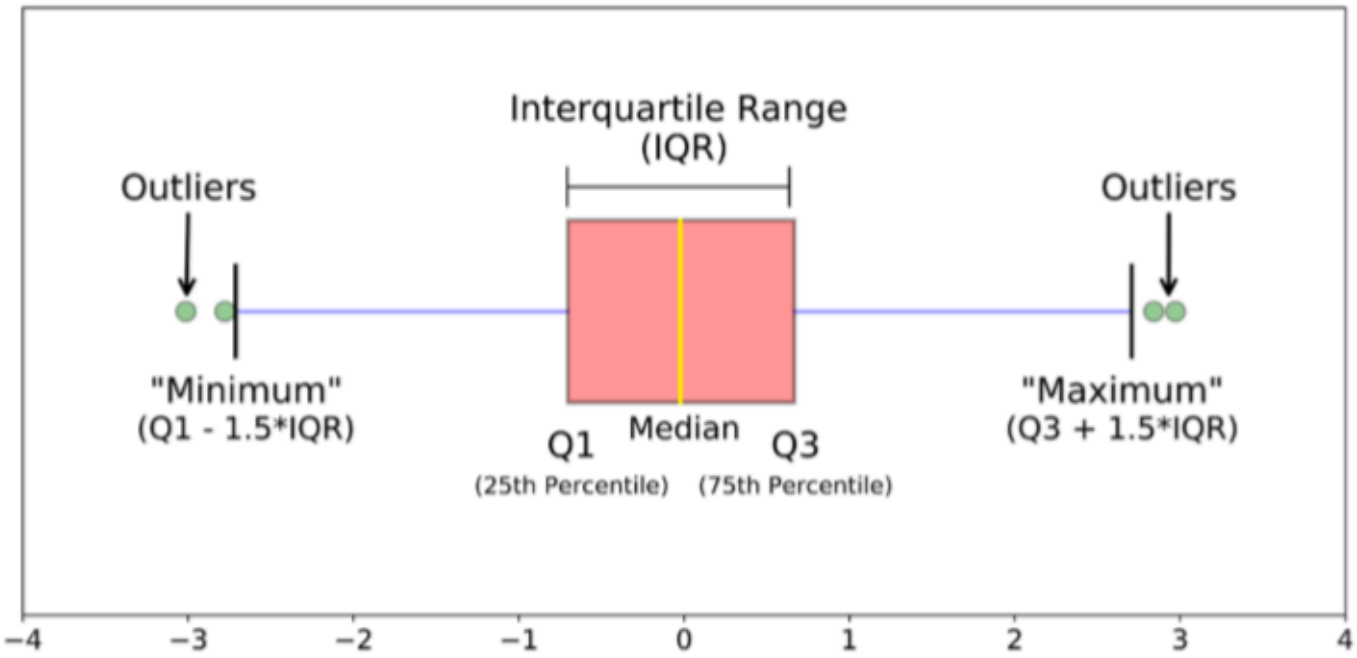
Dichotomous (binary) variable: Democratic control of the lower chamber (Klarner 2013).

Continuous variable: State tax burden (Caughey and Warshaw 2015).

Comparing densities

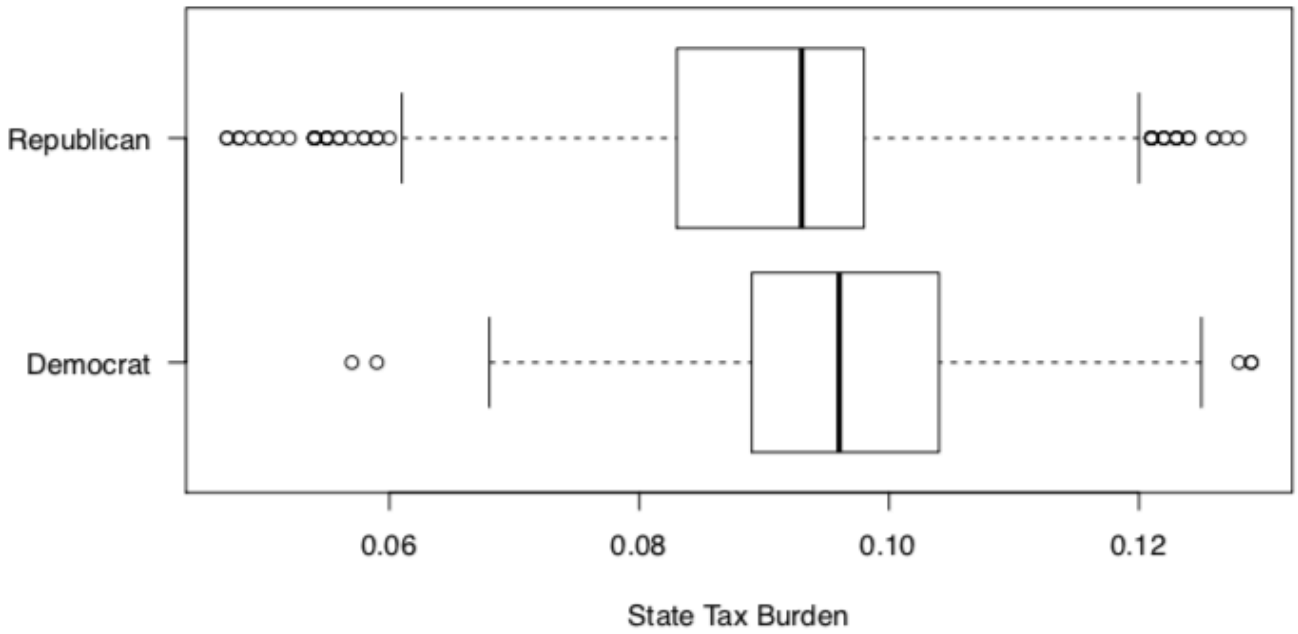


Comparing means using boxplots



From: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Comparing means using boxplots



Hypothesis testing with tax burdens

Hypothesis: States where Democrats control the lower house will have higher tax burdens than states that control where Republicans control the lower house.

$$H_{null} = TaxBurden_{Dem} \geq TaxBurden_{Rep}$$
$$H_{alt} = TaxBurden_{Dem} < TaxBurden_{Rep}$$

T-test with tax burdens

```
> t.test(x_tax_burden~party, alternative="less",
+        data=df, var.equal=T)

Two Sample t-test

data:  x_tax_burden by party
t = 8.4598, df = 1467, p-value = 1
alternative hypothesis: true difference in means
        is less than 0
95 percent confidence interval:
-Inf 0.006832579
sample estimates:
mean in group Democrat mean in group Republican
0.09604682                    0.09032704
```

Correlations

Correlations: basic association between variables

Questions we might ask:

- Does an association exist?
- If an association exists, how strong is it?
- What is the pattern and/or direction of association?

Crosstabs

Also known as **contingency tables**

- These show the relationship between two *ordinal* or *interval* variables.
- Each cell shows the frequency (or proportion) of observations that have both attributes.
- Column percentages make it easier to compare differences.

Crosstabs

Candidate support by party

	Dem	Ind	Not sure	Lib	Rep
Clinton	84.58	25.85	40	4.35	8.38
Johnson	1.4	9.76	10	17.39	3.35
Stein	1.4	2.44	0	4.35	0
Trump	11.68	54.63	30	60.87	84.36
Other	0.93	7.32	20	13.04	3.91
N	214	205	10	23	170

Correlation coefficient

- **Numerator:** sum of cross-products
  - How a particular ordered pair (X_i, Y_i) are related to each other
  - Positive correlation will exist when a high score for one variable (X) occurs with high score for second variable (Y).
  - High score assessed in relationship to avariable's mean value
- **Denominator:** standardizes the numerators across scales to remove the effects of unit size

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\sum_i (X_i - \bar{X})^2\right]\left[\sum_i (Y_i - \bar{Y})^2\right]}}$$

Example

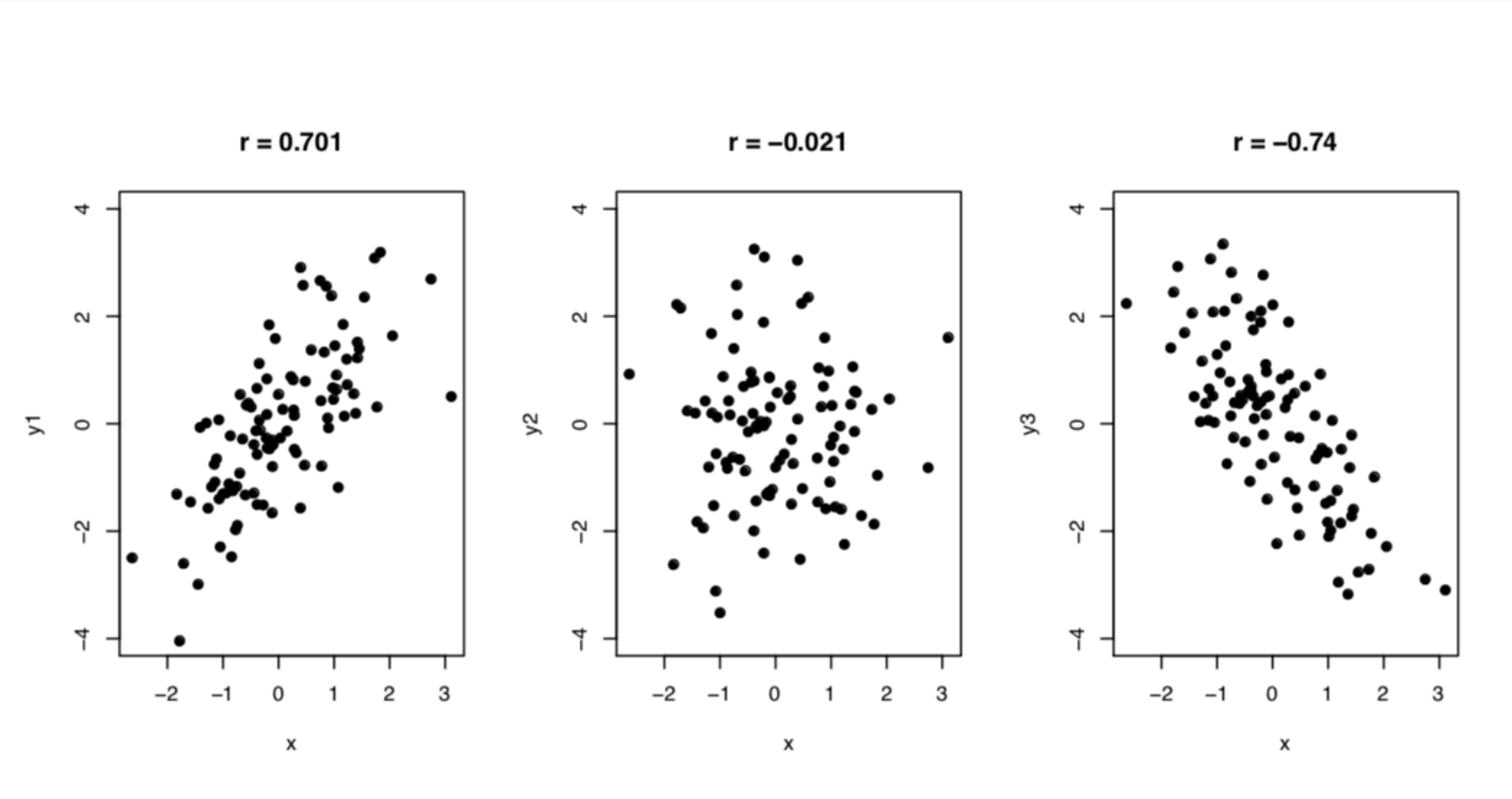
Research question: What is the relationship between the economy and the incumbent's chance of victory in presidential elections?

Example

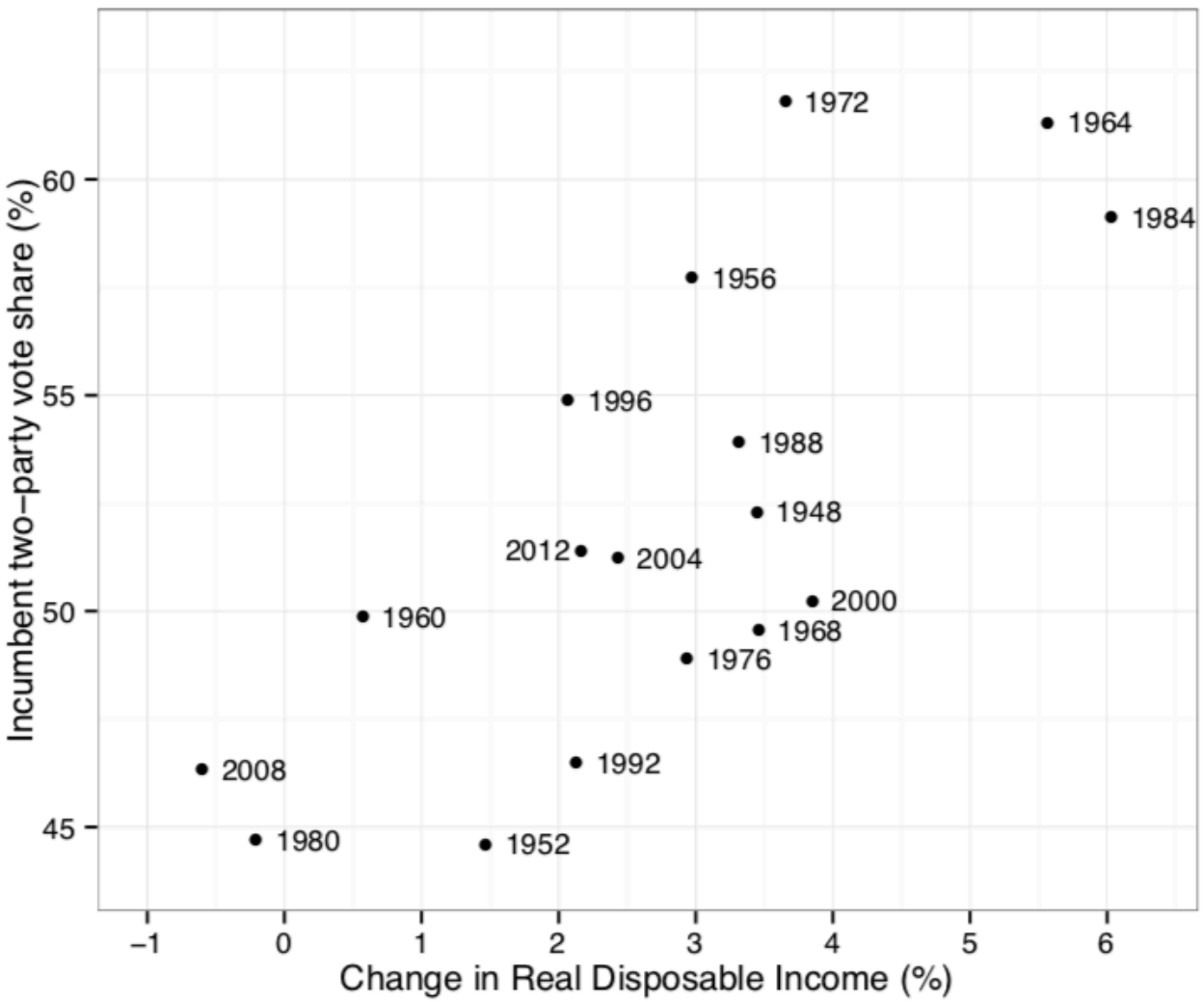
Research question: What is the relationship between the economy and the incumbent's chance of victory in presidential elections?

- What is our hypothesis?
- What's the direction of the relationship?

Visualizing correlations with scatterplots



Visualizing correlations with scatterplots



Use of the correlation coefficient

Correlation coefficients only tell you the *degree* and *direction* of a relationship.

- **Direction:** $+$, $-$, or 0.
- **Strength:** between -1.00 and 1.00, higher is stronger correlation.
- **Statistical significance:** Is the correlation we observe likely to have occurred by chance alone? Use test statistic.

Different types of correlation coefficients

- **Pearson's R:** most common. Based on covariance. Useful for continuous and interval variables.
- **Spearman's rank correlation:** Useful for variables that are ranked. Ranges from -1 to 1.
- **Kendall's Tau:** Useful for ordinal variables. Looks at "dischordant" and "concordant" pairs. Ranges from -1 to 1.

What is a (substantively) significant correlation?

- Correlation $< 0.3 =$ No correlation
- Correlation > 0.3 and $< 0.5 =$ Weak correlation
- Correlation > 0.5 and $< 0.7 =$ Moderate correlation
- Correlation > 0.7 and $< 1 =$ Strong correlation

Fun in R after the break!