

Visualizing Distributions & Missing Data

PSC7475: Week 4

Prof. Weldzius

Villanova University

Slides Updated: 2025-02-12

Concepts and measurement

Concepts and measurement

- Social science is about understanding **causal relationships**

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?
- Relationships are between **concepts**:

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?
- Relationships are between **concepts**:
 - Minimum wage, unemployment, outgroup contact, views on immigration

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?
- Relationships are between **concepts**:
 - Minimum wage, unemployment, outgroup contact, views on immigration
 - We took these for granted when talking about causality

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?
- Relationships are between **concepts**:
 - Minimum wage, unemployment, outgroup contact, views on immigration
 - We took these for granted when talking about causality
- Important to consider how we **measure** these concepts

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?
- Relationships are between **concepts**:
 - Minimum wage, unemployment, outgroup contact, views on immigration
 - We took these for granted when talking about causality
- Important to consider how we **measure** these concepts
 - Some straightforward: what is your age?

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?
- Relationships are between **concepts**:
 - Minimum wage, unemployment, outgroup contact, views on immigration
 - We took these for granted when talking about causality
- Important to consider how we **measure** these concepts
 - Some straightforward: what is your age?
 - Others more complicated: what does it mean to “be liberal”?

Concepts and measurement

- Social science is about understanding **causal relationships**
 - Does minimum wage change levels of employment
 - Does outgroup contact influence views on immigration?
- Relationships are between **concepts**:
 - Minimum wage, unemployment, outgroup contact, views on immigration
 - We took these for granted when talking about causality
- Important to consider how we **measure** these concepts
 - Some straightforward: what is your age?
 - Others more complicated: what does it mean to “be liberal”?
 - **Operational definition**: mapping of concept to numbers in our data

Example

Example

- Concept: presidential approval

Example

- Concept: presidential approval
- Conceptual definition:

Example

- Concept: presidential approval
- Conceptual definition:
 - Extent to which US adults support the actions and policies of the current US president

Example

- Concept: presidential approval
- Conceptual definition:
 - Extent to which US adults support the actions and policies of the current US president
- Operational definition:

Example

- Concept: presidential approval
- Conceptual definition:
 - Extent to which US adults support the actions and policies of the current US president
- Operational definition:
 - “On a scale from 1 to 5, where 1 is least supportive and 5 is most supportive, how much would you say you support the job that Donald Trump is doing as president?”

Measurement error

Measurement error

- **Measurement error:** chance variation in our measurements

Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error

Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error
 - chance errors tend to cancel out when we take averages

Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error
 - chance errors tend to cancel out when we take averages
- No matter how careful we are, chance error can always affect a measurement.

Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error
 - chance errors tend to cancel out when we take averages
- No matter how careful we are, chance error can always affect a measurement.
 - Panel study of 19,000 respondents: 20 reported being a citizen in 2010 and then a non-citizen in 2012

Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error
 - chance errors tend to cancel out when we take averages
- No matter how careful we are, chance error can always affect a measurement.
 - Panel study of 19,000 respondents: 20 reported being a citizen in 2010 and then a non-citizen in 2012
 - Data entry errors

Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error
 - chance errors tend to cancel out when we take averages
- No matter how careful we are, chance error can always affect a measurement.
 - Panel study of 19,000 respondents: 20 reported being a citizen in 2010 and then a non-citizen in 2012
 - Data entry errors
- **Bias:** systematic errors for all units in the same direction.

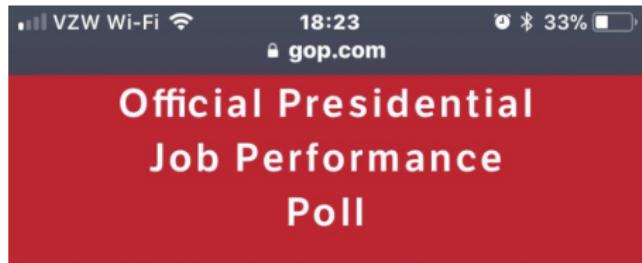
Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error
 - chance errors tend to cancel out when we take averages
- No matter how careful we are, chance error can always affect a measurement.
 - Panel study of 19,000 respondents: 20 reported being a citizen in 2010 and then a non-citizen in 2012
 - Data entry errors
- **Bias:** systematic errors for all units in the same direction.
 - individual measurement = exact value + bias + chance error

Measurement error

- **Measurement error:** chance variation in our measurements
 - individual measurement = exact value + chance error
 - chance errors tend to cancel out when we take averages
- No matter how careful we are, chance error can always affect a measurement.
 - Panel study of 19,000 respondents: 20 reported being a citizen in 2010 and then a non-citizen in 2012
 - Data entry errors
- **Bias:** systematic errors for all units in the same direction.
 - individual measurement = exact value + bias + chance error
 - “What did you eat yesterday?” \rightsquigarrow underreporting

A biased poll?



1. How would you rate President Trump's job performance so far?
 - Great
 - Good
 - Okay
 - Other
2. (Optional) Please explain why you selected your response.

1936 Literary Digest Poll

The Literary Digest
NEW YORK OCTOBER 21, 1936

Topics of the day

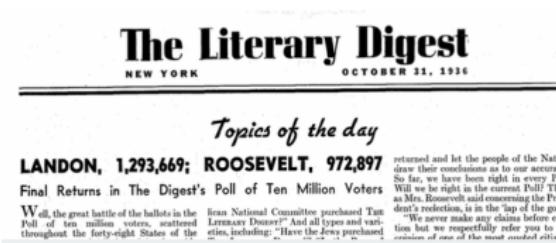
LANDON, 1,293,669; ROOSEVELT, 972,897

Final Return in The Digest's Poll of Ten Million Voters

With the great battle of the ballot in the Poll of ten million voters, scattered throughout the forty-eight States of the

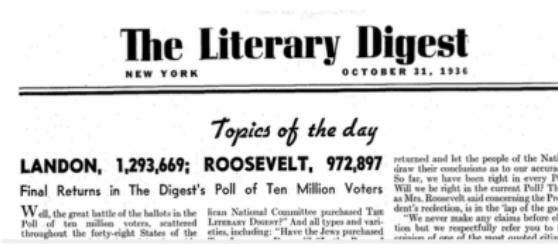
Item National Committee purchased That Literary Digest? And all types and varieties, including: "Have the Jews purchased returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mr. Roosevelt said, "is a question, perhaps, for reflection, is in the 'lap of the gods.' We never make any claims before election, but we respectfully refer you to the opinion of one of the most *respected* officers

1936 Literary Digest Poll



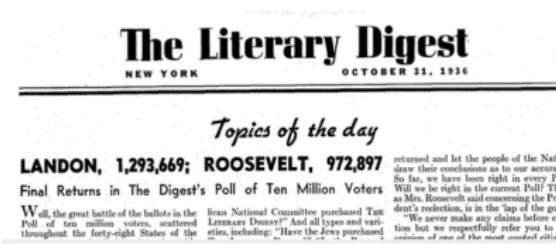
- Literary Digest predicted elections using mail-in polls

1936 Literary Digest Poll



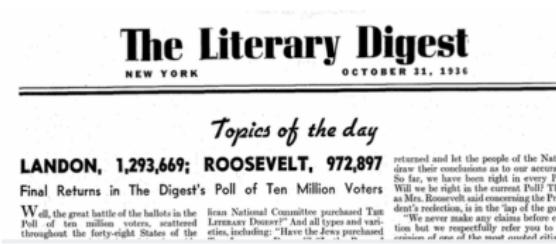
- Literary Digest predicted elections using mail-in polls
- Source of addresses: automobile registrations, phone books, etc.

1936 Literary Digest Poll



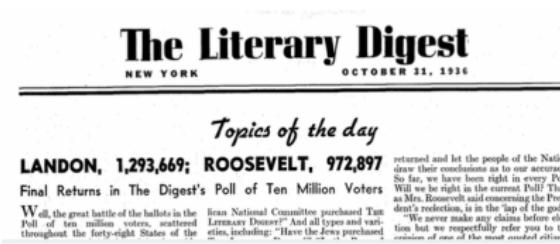
- Literary Digest predicted elections using mail-in polls
- Source of addresses: automobile registrations, phone books, etc.
- In 1936, sent out 10 million ballots, over 2.3 million returned

1936 Literary Digest Poll



- Literary Digest predicted elections using mail-in polls
- Source of addresses: automobile registrations, phone books, etc.
- In 1936, sent out 10 million ballots, over 2.3 million returned
- George Gallup used only 50,000 respondents

1936 Literary Digest Poll



- Literary Digest predicted elections using mail-in polls
- Source of addresses: automobile registrations, phone books, etc.
- In 1936, sent out 10 million ballots, over 2.3 million returned
- George Gallup used only 50,000 respondents

	FDR's vote share
Literary Digest	43
George Gallup	56

Poll fail



	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

Poll fail



	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

- **Selection bias:** ballots skewed toward the wealthy (with cars, phones)

Poll fail



	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

- **Selection bias:** ballots skewed toward the wealthy (with cars, phones)
 - Only 1 in 4 households had a phone in 1936

Poll fail



	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

- **Selection bias:** ballots skewed toward the wealthy (with cars, phones)
 - Only 1 in 4 households had a phone in 1936
- **Nonresponse bias:** respondents differ from nonrespondents

Poll fail



	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

- **Selection bias:** ballots skewed toward the wealthy (with cars, phones)
 - Only 1 in 4 households had a phone in 1936
- **Nonresponse bias:** respondents differ from nonrespondents
 - ↵ when selection procedure is biased, adding more units won't help!

Poll fail



	FDR %
Literary Digest	43
George Gallup	56
Actual Outcome	62

- **Selection bias:** ballots skewed toward the wealthy (with cars, phones)
 - Only 1 in 4 households had a phone in 1936
- **Nonresponse bias:** respondents differ from nonrespondents
 - ↵ when selection procedure is biased, adding more units won't help!

1948 Election



The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

- **Quota sampling:** fixed quota of certain respondents for each interviewer

The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

- **Quota sampling:** fixed quota of certain respondents for each interviewer
 - If Black women make up 5% of the population, stop interviewing them once they make up 5% of your sample

The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

- **Quota sampling:** fixed quota of certain respondents for each interviewer
 - If Black women make up 5% of the population, stop interviewing them once they make up 5% of your sample
- Sample resembles the population on these characteristics

The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

- **Quota sampling:** fixed quota of certain respondents for each interviewer
 - If Black women make up 5% of the population, stop interviewing them once they make up 5% of your sample
- Sample resembles the population on these characteristics
- Potential unobserved confounding ↽ **selection bias**

The Polling Disaster

	Truman	Dewey	Thurmond	Wallace
Crossley	45	50	2	3
Gallup	44	50	2	4
Roper	38	53	5	4
Actual Outcome	50	45	3	2

- **Quota sampling:** fixed quota of certain respondents for each interviewer
 - If Black women make up 5% of the population, stop interviewing them once they make up 5% of your sample
- Sample resembles the population on these characteristics
- Potential unobserved confounding ↪ **selection bias**
- Republicans easier to find within quotas (phones, listed addresses)

Sample surveys

Sample surveys

- **Probability sampling** to ensure representativeness

Sample surveys

- **Probability sampling** to ensure representativeness
 - Definition: every unit in the population has a known, non-zero probability of being selected into sample

Sample surveys

- **Probability sampling** to ensure representativeness
 - Definition: every unit in the population has a known, non-zero probability of being selected into sample
- **Simple random sampling**: every unit has an equal selection probability.

Sample surveys

- **Probability sampling** to ensure representativeness
 - Definition: every unit in the population has a known, non-zero probability of being selected into sample
- **Simple random sampling**: every unit has an equal selection probability.
- Random digit dialing:

Sample surveys

- **Probability sampling** to ensure representativeness
 - Definition: every unit in the population has a known, non-zero probability of being selected into sample
- **Simple random sampling**: every unit has an equal selection probability.
- Random digit dialing:
 - Take a particular area code + exchange: 310-495-XXXX.

Sample surveys

- **Probability sampling** to ensure representativeness
 - Definition: every unit in the population has a known, non-zero probability of being selected into sample
- **Simple random sampling**: every unit has an equal selection probability.
- Random digit dialing:
 - Take a particular area code + exchange: 310-495-XXXX.
 - Randomly choose each digit in XXXX to call a particular phone

Sample surveys

- **Probability sampling** to ensure representativeness
 - Definition: every unit in the population has a known, non-zero probability of being selected into sample
- **Simple random sampling**: every unit has an equal selection probability.
- Random digit dialing:
 - Take a particular area code + exchange: 310-495-XXXX.
 - Randomly choose each digit in XXXX to call a particular phone
 - Every phone in the US has an equal chance of being included in sample

Sampling lingo

Sampling lingo

- **Target population:** set of people we want to learn about

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample
 - Frame bias: list of registered voters (frame) might include nonvoters!

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample
 - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample:** set of people contacted

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample
 - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample:** set of people contacted
- **Respondents:** subset of sample that actually responds to the survey

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample
 - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample:** set of people contacted
- **Respondents:** subset of sample that actually responds to the survey
 - Unit non-response: sample \neq respondents

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample
 - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample:** set of people contacted
- **Respondents:** subset of sample that actually responds to the survey
 - Unit non-response: sample \neq respondents
 - Not everyone picks up their phone

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample
 - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample:** set of people contacted
- **Respondents:** subset of sample that actually responds to the survey
 - Unit non-response: sample \neq respondents
 - Not everyone picks up their phone
- **Completed items:** subset of questions that respondents answer

Sampling lingo

- **Target population:** set of people we want to learn about
 - Example: people who will vote in the next election
- **Sampling frame:** list of people from which we will actually sample
 - Frame bias: list of registered voters (frame) might include nonvoters!
- **Sample:** set of people contacted
- **Respondents:** subset of sample that actually responds to the survey
 - Unit non-response: sample \neq respondents
 - Not everyone picks up their phone
- **Completed items:** subset of questions that respondents answer
 - Item non-response: refusing to disclose their vote preference

Difficulties of sampling

Difficulties of sampling

- Problems of telephone survey

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)
 - Caller ID screening (unit non-response)

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)
 - Caller ID screening (unit non-response)
 - Response rates down to 9%

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)
 - Caller ID screening (unit non-response)
 - Response rates down to 9%
- An alternative: internet surveys

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)
 - Caller ID screening (unit non-response)
 - Response rates down to 9%
- An alternative: internet surveys
 - Opt-in panels, respondent-driven sampling \rightsquigarrow **non-probability sampling**

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)
 - Caller ID screening (unit non-response)
 - Response rates down to 9%
- An alternative: internet surveys
 - Opt-in panels, respondent-driven sampling \rightsquigarrow **non-probability sampling**
 - Cheaper, but non-representative

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)
 - Caller ID screening (unit non-response)
 - Response rates down to 9%
- An alternative: internet surveys
 - Opt-in panels, respondent-driven sampling \rightsquigarrow **non-probability sampling**
 - Cheaper, but non-representative
 - Digital divide: rich vs. poor, young vs. old

Difficulties of sampling

- Problems of telephone survey
 - Cell phones (double counting for the wealthy)
 - Caller ID screening (unit non-response)
 - Response rates down to 9%
- An alternative: internet surveys
 - Opt-in panels, respondent-driven sampling \rightsquigarrow **non-probability sampling**
 - Cheaper, but non-representative
 - Digital divide: rich vs. poor, young vs. old
 - Correct for potential sampling bias via statistical methods

Effect of assassination attempts

Effect of assassination attempts

```
library(tidyverse)
data(leaders, package = "qss")
head(leaders[,1:7])
```

```
##   year      country      leadername age politybefore
## 1 1929 Afghanistan Habibullah Ghazi  39      -6
## 2 1933 Afghanistan          Nadir Shah  53      -6
## 3 1934 Afghanistan          Hashim Khan 50      -6
## 4 1924      Albania           Zogu    29       0
## 5 1931      Albania           Zogu    36      -9
## 6 1968      Algeria        Boumedienne 41      -9
##   polityafter interwarbefore
## 1 -6.000000
## 2 -7.333333
## 3 -8.000000
## 4 -9.000000
## 5 -9.000000
## 6 -9.000000
```

Contingency tables

Contingency tables

- With two categorical variables, we can create **contingency tables**

Contingency tables

- With two categorical variables, we can create **contingency tables**
 - Also known as **cross-tabs**

Contingency tables

- With two categorical variables, we can create **contingency tables**
 - Also known as **cross-tabs**
 - Rows are the values of one variable, columns the other

Contingency tables

- With two categorical variables, we can create **contingency tables**
 - Also known as **cross-tabs**
 - Rows are the values of one variable, columns the other

```
leaders %>%
  group_by(civilwarbefore, civilwarafter) %>%
  count() %>%
  spread(civilwarafter, n)
```

```
## # A tibble: 2 x 3
## # Groups:   civilwarbefore [2]
##   civilwarbefore '0'   '1'
##             <int> <int> <int>
## 1                 0    177     19
## 2                 1     27     27
```

- Quick summary how the two variables “go together”

Cross-tabs with proportions

```
leaders %>%
  group_by(civilwarbefore, civilwarafter) %>%
  count() %>%
  ungroup() %>%
  mutate(prop = n/ sum(n)) %>%
  select(-n) %>%
  spread(civilwarafter, prop, drop = T)
```

```
## # A tibble: 2 x 3
##   civilwarbefore   '0'   '1'
##             <int> <dbl> <dbl>
## 1                 0  0.708  0.076
## 2                 1  0.108  0.108
```

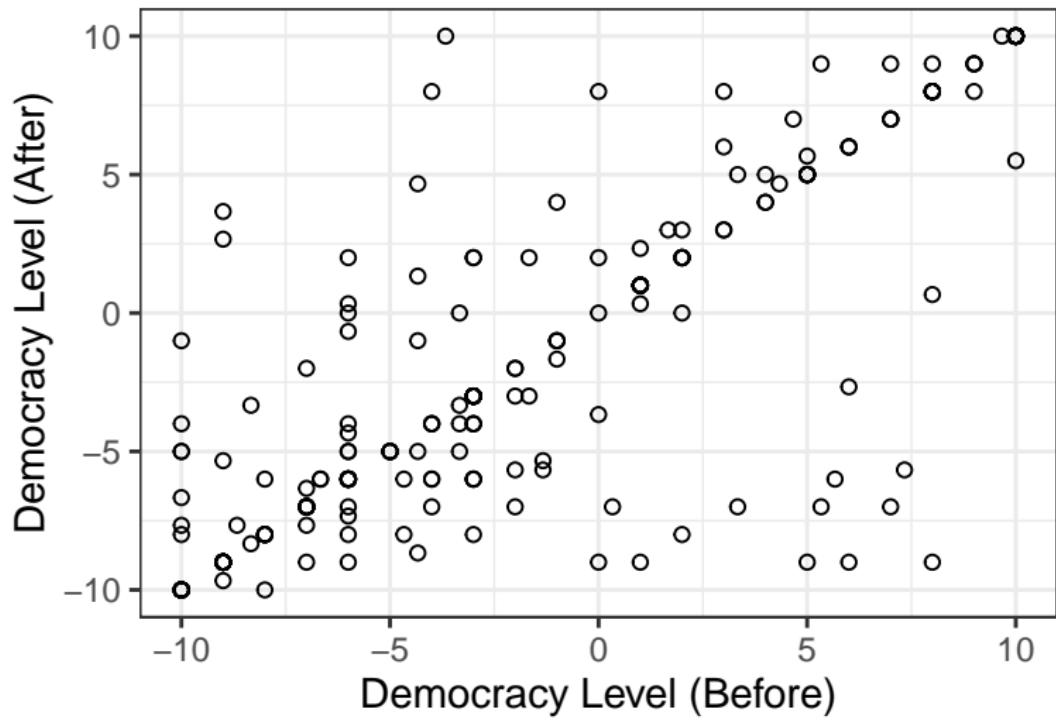
Cross-tabs with proportions (by row)

```
leaders %>%
  group_by(civilwarbefore,civilwarafter) %>%
  count() %>%
  ungroup() %>%
  group_by(civilwarbefore) %>%
  mutate(prop = n/ sum(n)) %>%
  select(-n) %>%
  spread(civilwarafter, prop, drop = T)
```

```
## # A tibble: 2 x 3
## # Groups:   civilwarbefore [2]
##   civilwarbefore '0'    '1'
##             <int> <dbl>  <dbl>
## 1                 0  0.903  0.0969
## 2                 1  0.5    0.5
```

Scatterplot

Democracy before/after assasination attempts



Scatterplot

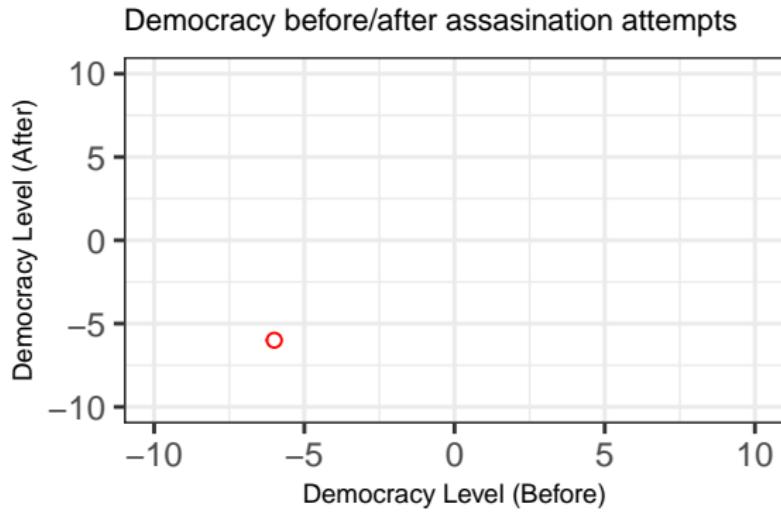
- Each point on the scatterplot (x_i, y_i)
- Use `geom_point()` function in `ggplot`

```
leaders %>%
  ggplot(aes(x = politybefore, y = polityafter)) +
  geom_point(shape = 21) +
  labs(title = "Democracy before/after assassination attempts",
       x = "Democracy Level (Before)",
       y = "Democracy Level (After)") +
  theme_bw() +
  theme(plot.title = element_text(size=12))
```

Scatterplot

```
leaders[1, c("politybefore","polityafter")]
```

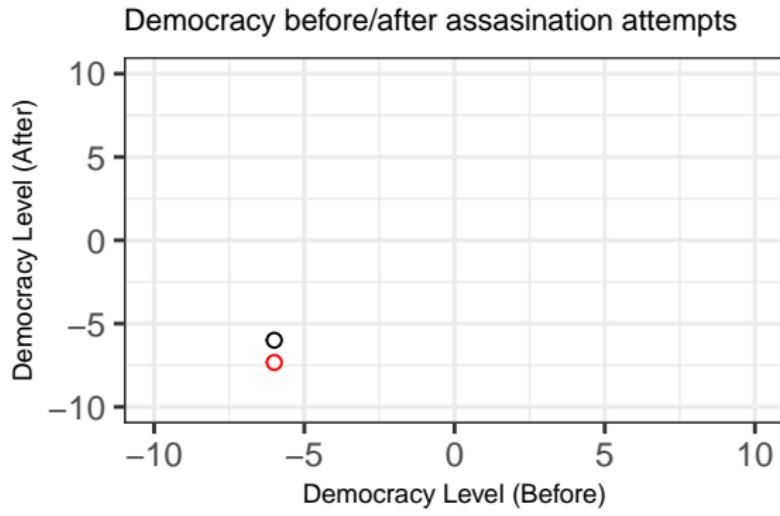
```
##   politybefore polityafter  
## 1           -6           -6
```



Scatterplot

```
leaders[2, c("politybefore","polityafter")]
```

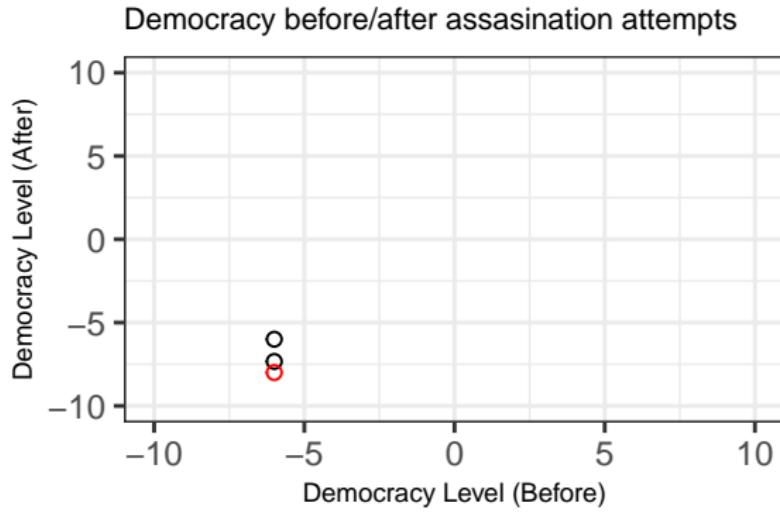
```
##   politybefore polityafter  
## 2           -6     -7.333333
```



Scatterplot

```
leaders[3, c("politybefore","polityafter")]
```

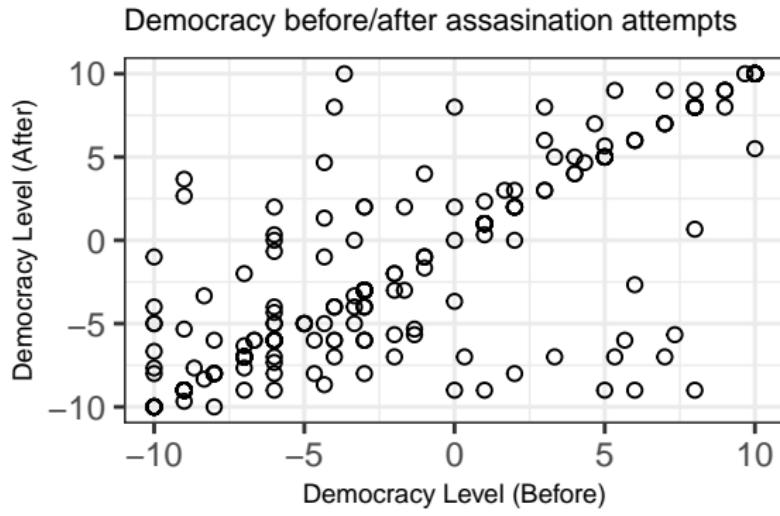
```
##   politybefore polityafter  
## 3           -6          -8
```



Scatterplot

```
leaders[3, c("politybefore","polityafter")]
```

```
##   politybefore polityafter  
## 3          -6           -8
```



How big is big?

How big is big?

- Would be nice to have a standard summary of how similar variables are

How big is big?

- Would be nice to have a standard summary of how similar variables are
 - Problem: variables on different scales!

How big is big?

- Would be nice to have a standard summary of how similar variables are
 - Problem: variables on different scales!
 - Needs a way to put any variable on common units

How big is big?

- Would be nice to have a standard summary of how similar variables are
 - Problem: variables on different scales!
 - Needs a way to put any variable on common units
 - **z-score** to the rescue!

How big is big?

- Would be nice to have a standard summary of how similar variables are
 - Problem: variables on different scales!
 - Needs a way to put any variable on common units
 - **z-score** to the rescue!

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

How big is big?

- Would be nice to have a standard summary of how similar variables are
 - Problem: variables on different scales!
 - Needs a way to put any variable on common units
 - **z-score** to the rescue!

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Crucial property: z-scores don't depend on units

How big is big?

- Would be nice to have a standard summary of how similar variables are
 - Problem: variables on different scales!
 - Needs a way to put any variable on common units
 - **z-score** to the rescue!

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- Crucial property: z-scores don't depend on units

$$\text{z-score of } (ax_i + b) = \text{z-score of } x_i$$

Correlation

Correlation

- How do variables move together on average?

Correlation

- How do variables move together on average?
- When x_i is big, what is y_i likely to be?

Correlation

- How do variables move together on average?
- When x_i is big, what is y_i likely to be?
 - Positive correlation: when x_i is big, y_i is also big

Correlation

- How do variables move together on average?
- When x_i is big, what is y_i likely to be?
 - Positive correlation: when x_i is big, y_i is also big
 - Negative correlation: when x_i is big, y_i is small

Correlation

- How do variables move together on average?
- When x_i is big, what is y_i likely to be?
 - Positive correlation: when x_i is big, y_i is also big
 - Negative correlation: when x_i is big, y_i is small
 - High magnitude of correlation: data cluster tightly around a line

Correlation

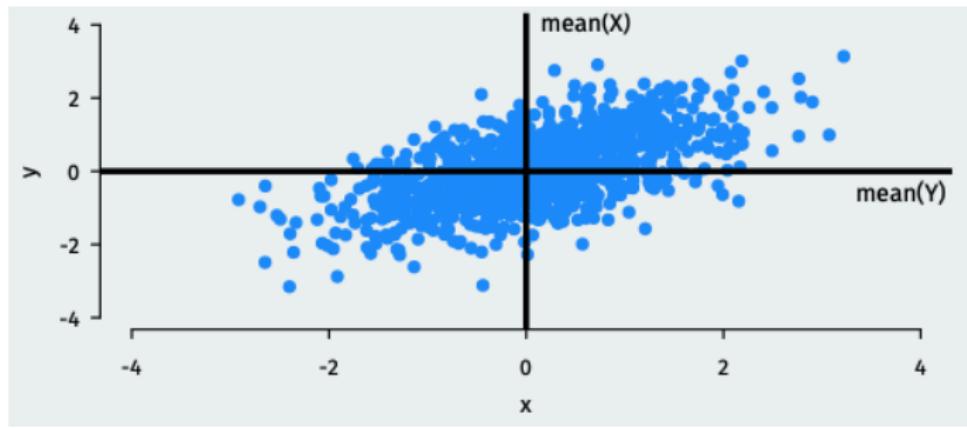
- How do variables move together on average?
- When x_i is big, what is y_i likely to be?
 - Positive correlation: when x_i is big, y_i is also big
 - Negative correlation: when x_i is big, y_i is small
 - High magnitude of correlation: data cluster tightly around a line
- The technical definition of the **correlation coefficient**:

Correlation

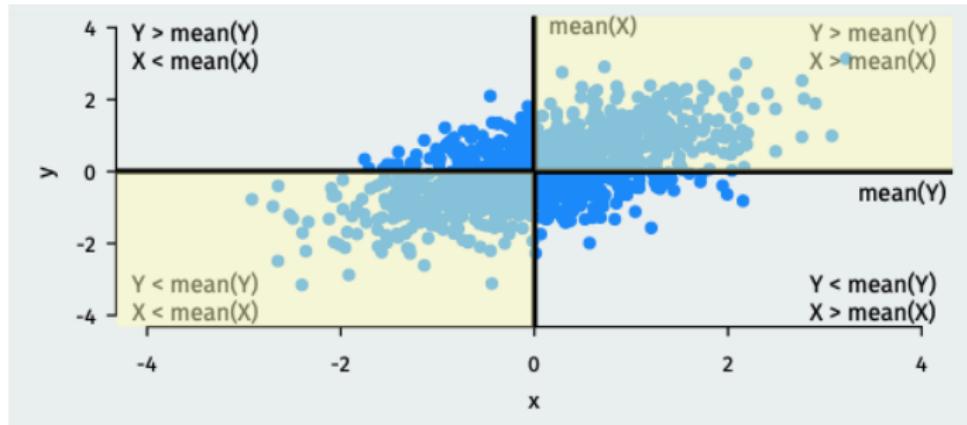
- How do variables move together on average?
- When x_i is big, what is y_i likely to be?
 - Positive correlation: when x_i is big, y_i is also big
 - Negative correlation: when x_i is big, y_i is small
 - High magnitude of correlation: data cluster tightly around a line
- The technical definition of the **correlation coefficient**:

$$\frac{1}{n-1} \sum_{i=1}^n [(\text{z-score for } x_i) \times (\text{z-score for } y_i)]$$

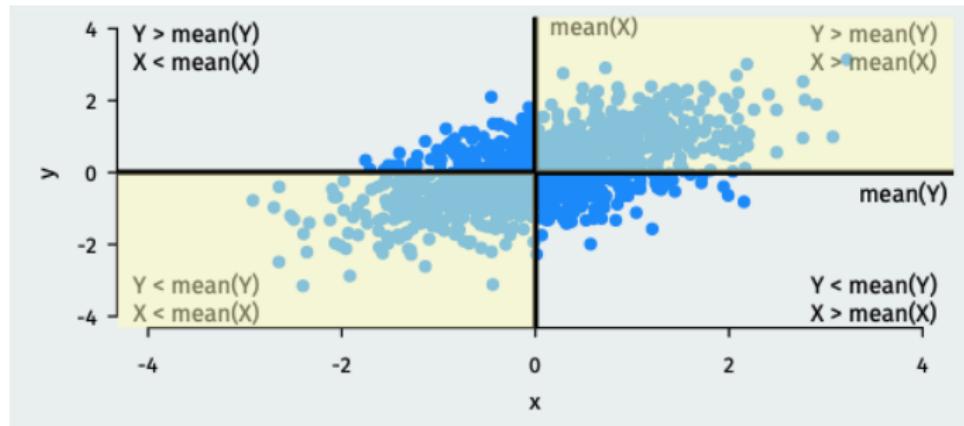
Correlation intuition:



Correlation intuition:

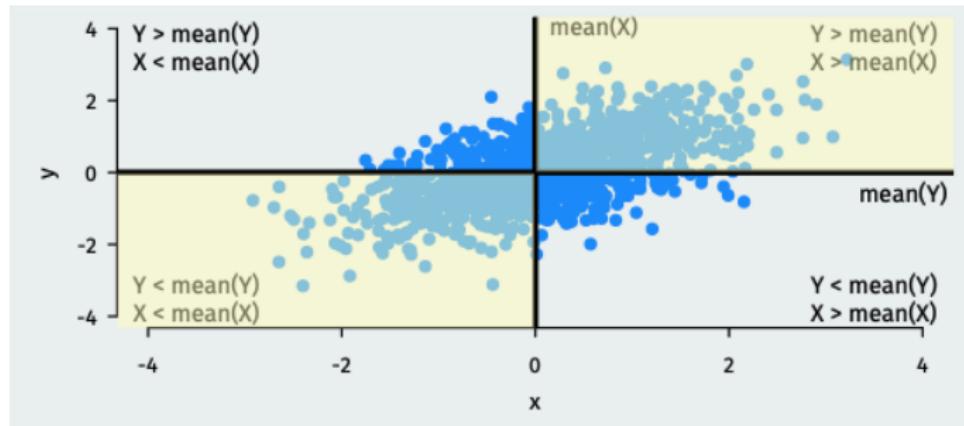


Correlation intuition:



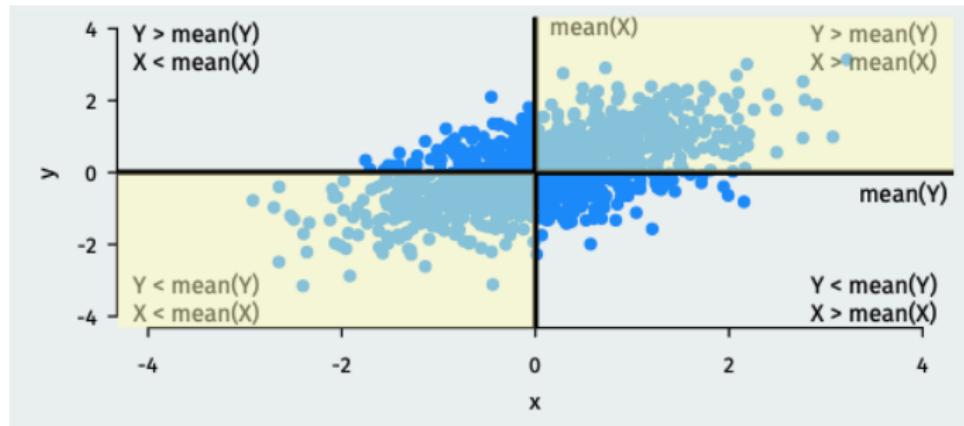
- Large values of X tend to occur with large values of Y

Correlation intuition:



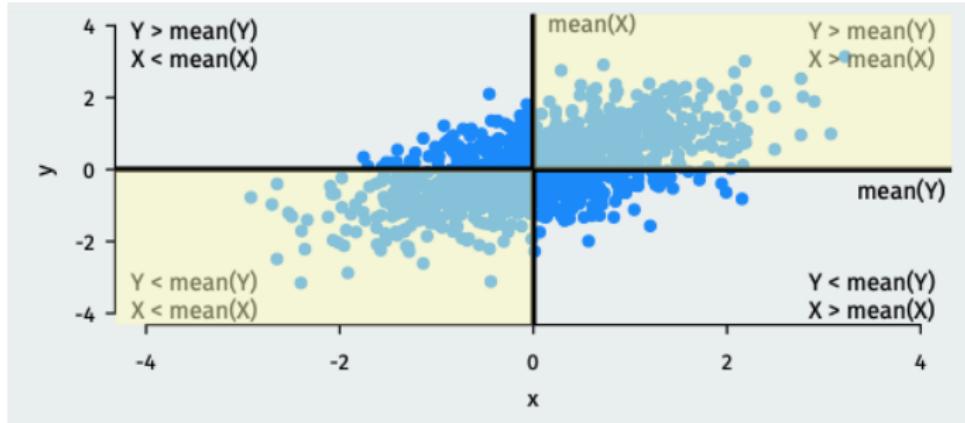
- Large values of X tend to occur with large values of Y
 - $(z\text{-score for } x_i) \times (z\text{-score for } y_1) = (\text{pos. num.}) \times (\text{pos. num.}) = +$

Correlation intuition:



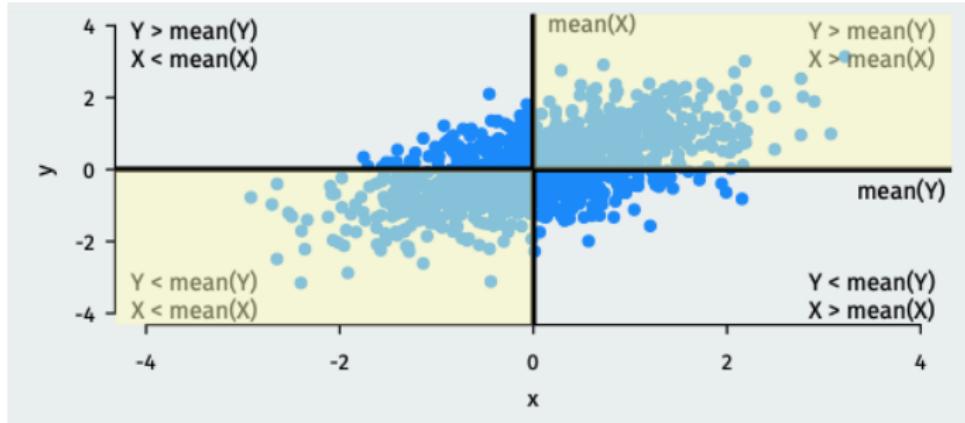
- Large values of X tend to occur with large values of Y
 - $(z\text{-score for } x_i) \times (z\text{-score for } y_1) = (\text{pos. num.}) \times (\text{pos. num.}) = +$
- Small values of X tend to occur with small values of Y

Correlation intuition:



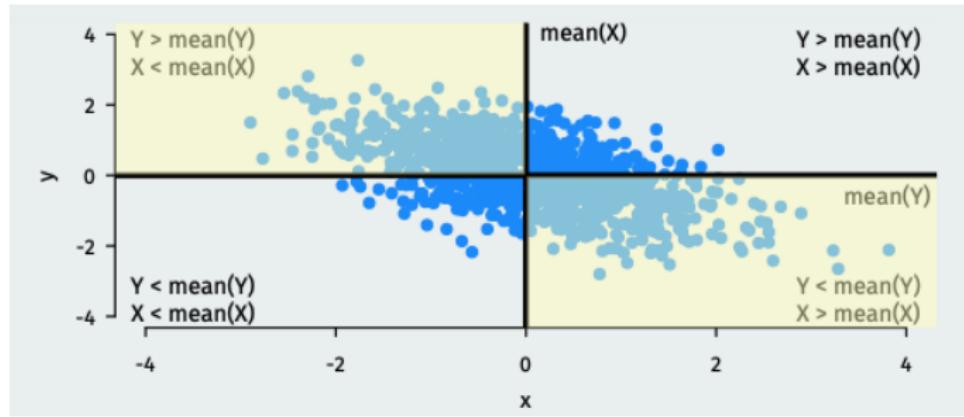
- Large values of X tend to occur with large values of Y
 - $(z\text{-score for } x_i) \times (z\text{-score for } y_1) = (\text{pos. num.}) \times (\text{pos. num.}) = +$
- Small values of X tend to occur with small values of Y
 - $(z\text{-score for } x_i) \times (z\text{-score for } y_1) = (\text{neg. num.}) \times (\text{neg. num.}) = +$

Correlation intuition:

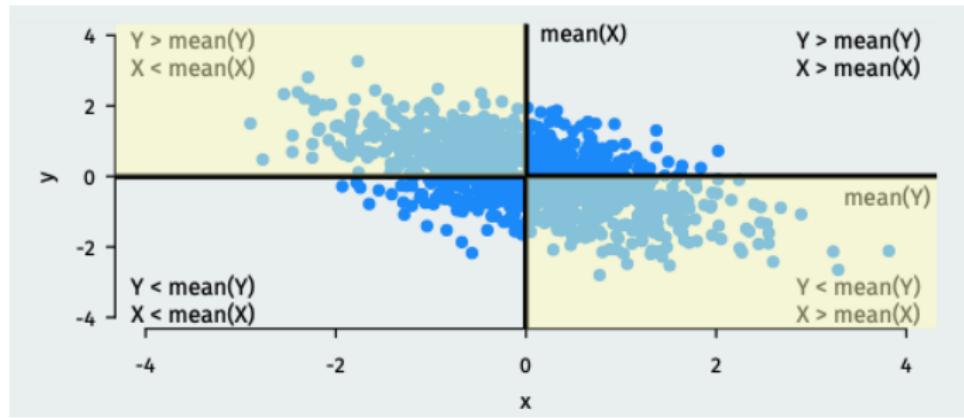


- Large values of X tend to occur with large values of Y
 - $(z\text{-score for } x_i) \times (z\text{-score for } y_1) = (\text{pos. num.}) \times (\text{pos. num.}) = +$
- Small values of X tend to occur with small values of Y
 - $(z\text{-score for } x_i) \times (z\text{-score for } y_1) = (\text{neg. num.}) \times (\text{neg. num.}) = +$
- If these dominate \rightsquigarrow positive correlation

Correlation intuition:

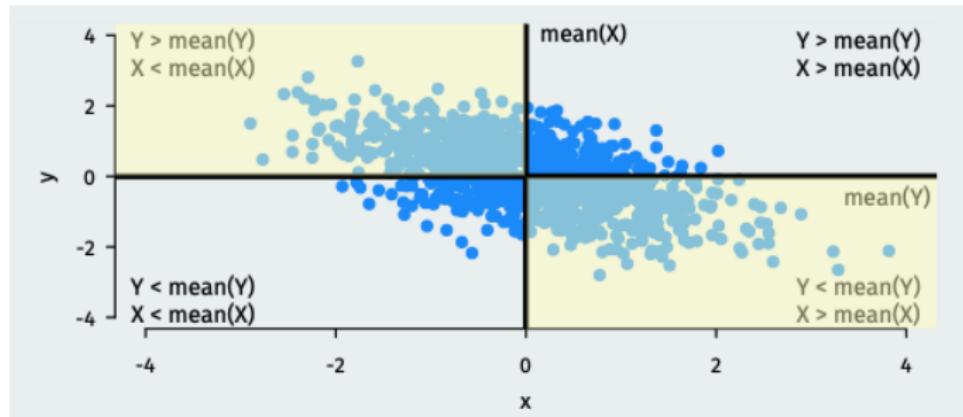


Correlation intuition:



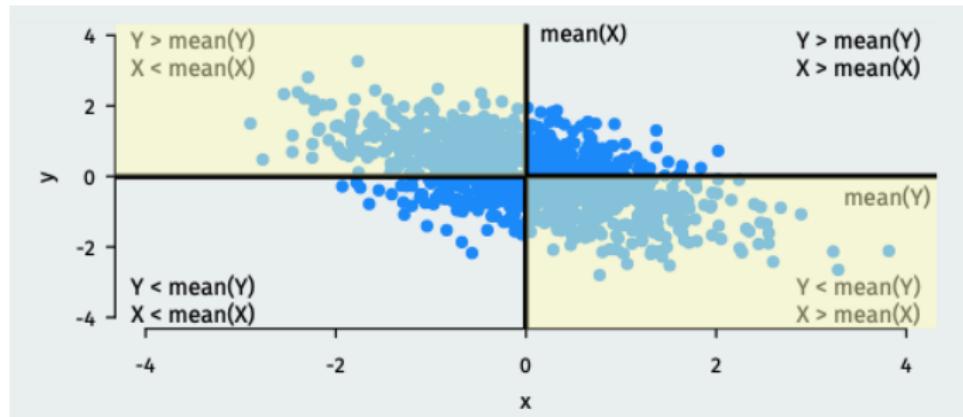
- Large values of X tend to occur with small values of Y

Correlation intuition:



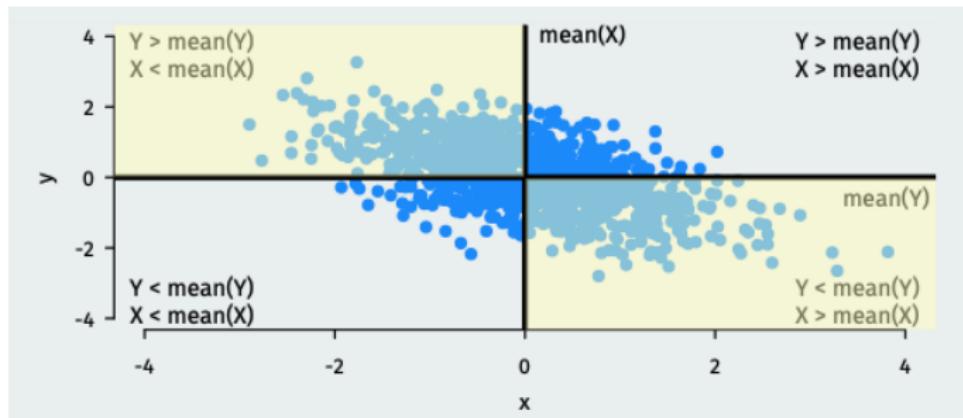
- Large values of X tend to occur with small values of Y
 - $(\text{z-score for } x_i) \times (\text{z-score for } y_1) = (\text{pos. num.}) \times (\text{neg. num.}) = -$

Correlation intuition:



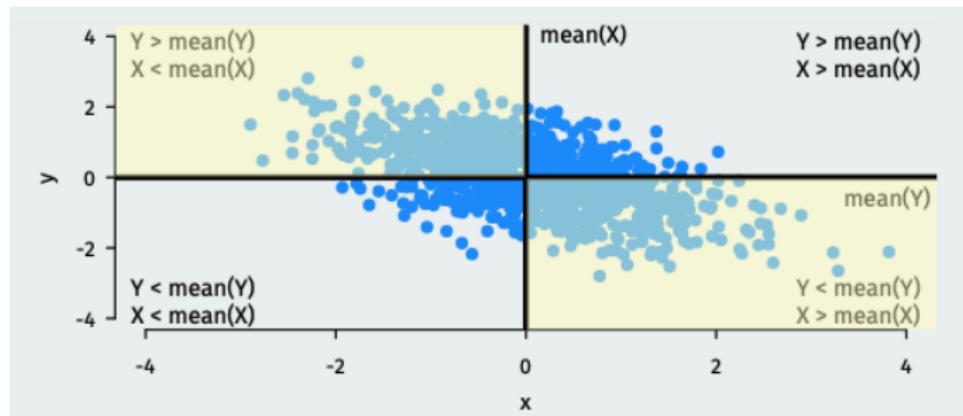
- Large values of X tend to occur with small values of Y
 - $(z\text{-score for } x_i) \times (z\text{-score for } y_1) = (\text{pos. num.}) \times (\text{neg. num.}) = -$
- Small values of X tend to occur with large values of Y

Correlation intuition:



- Large values of X tend to occur with small values of Y
 - $(\text{z-score for } x_i) \times (\text{z-score for } y_1) = (\text{pos. num.}) \times (\text{neg. num.}) = -$
- Small values of X tend to occur with large values of Y
 - $(\text{z-score for } x_i) \times (\text{z-score for } y_1) = (\text{neg. num.}) \times (\text{pos. num.}) = -$

Correlation intuition:



- Large values of X tend to occur with small values of Y
 - $(\text{z-score for } x_i) \times (\text{z-score for } y_1) = (\text{pos. num.}) \times (\text{neg. num.}) = -$
- Small values of X tend to occur with large values of Y
 - $(\text{z-score for } x_i) \times (\text{z-score for } y_1) = (\text{neg. num.}) \times (\text{pos. num.}) = -$
- If these dominate \rightsquigarrow negative correlation

Properties of correlation coefficient

Properties of correlation coefficient

- Correlation measures **linear** association.

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1
 - Correlation of 0 means no linear association

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1
 - Correlation of 0 means no linear association
 - Positive correlations \rightsquigarrow positive associations

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1
 - Correlation of 0 means no linear association
 - Positive correlations \rightsquigarrow positive associations
 - Negative correlations \rightsquigarrow negative associations

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1
 - Correlation of 0 means no linear association
 - Positive correlations \rightsquigarrow positive associations
 - Negative correlations \rightsquigarrow negative associations
 - Closer to -1 or 1 means stronger association

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1
 - Correlation of 0 means no linear association
 - Positive correlations \rightsquigarrow positive associations
 - Negative correlations \rightsquigarrow negative associations
 - Closer to -1 or 1 means stronger association
- Order doesn't matter: $\text{cor}(x,y) = \text{cor}(y,x)$

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1
 - Correlation of 0 means no linear association
 - Positive correlations \rightsquigarrow positive associations
 - Negative correlations \rightsquigarrow negative associations
 - Closer to -1 or 1 means stronger association
- Order doesn't matter: $\text{cor}(x,y) = \text{cor}(y,x)$
- Not affected by changes of scale:

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
 - Correlation is between -1 and 1
 - Correlation of 0 means no linear association
 - Positive correlations \rightsquigarrow positive associations
 - Negative correlations \rightsquigarrow negative associations
 - Closer to -1 or 1 means stronger association
- Order doesn't matter: $\text{cor}(x,y) = \text{cor}(y,x)$
- Not affected by changes of scale:
 - $\text{cor}(x,y) = \text{cor}(ax+b, cy+d)$
 - Celsius vs. Fahrenheit; dollars vs. pesos; cm vs. in.

Correlation in R

- Use the `cor()` function

Correlation in R

- Use the `cor()` function

```
leaders %>%  
  select(politybefore, polityafter) %>%  
  cor()
```

```
##           politybefore polityafter  
## politybefore     1.0000000   0.8283237  
## polityafter      0.8283237   1.0000000
```

- Very highly correlated!

Assassination attempts

- See the possible attempt results

```
unique(leaders$result)
```

```
## [1] "not wounded"  
## [2] "dies within a day after the attack"  
## [3] "survives, whether wounded unknown"  
## [4] "wounded lightly"  
## [5] "plot stopped"  
## [6] "hospitalization but no permanent disability"  
## [7] "dies between a day and a week"  
## [8] "dies, timing unknown"  
## [9] "survives but wounded severely"  
## [10] "dies between a week and a month"
```

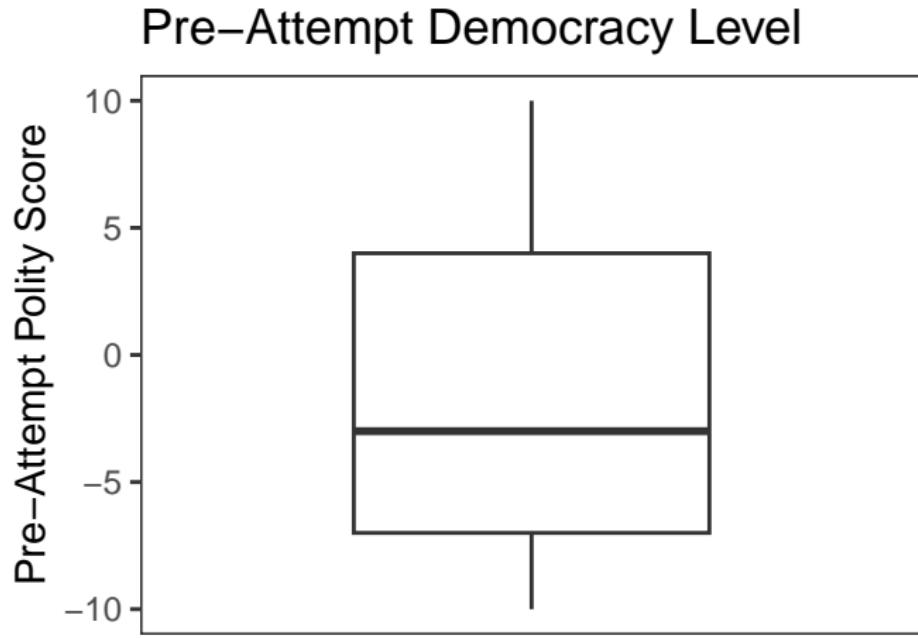
Creating an attempt fatal variable

- use ifelse to create a fatal variable

```
## create new vector of unique results of "result"
lev <- unique(leaders$result)
leaders <- leaders %>%
  mutate(fatal = ifelse(result %in% lev[c(2,7,8,10)], 1,0))
leaders %>%
  summarize(mean(fatal))

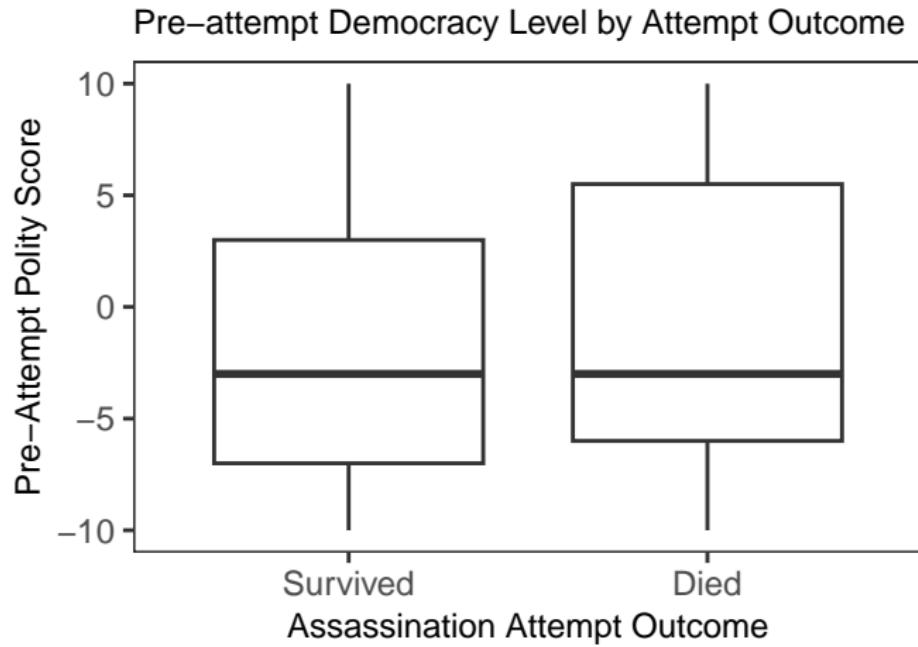
##     mean(fatal)
## 1      0.216
```

Remember boxplots?



Comparing distribution with the boxpot

- What if we want to know how the distribution varies by success?



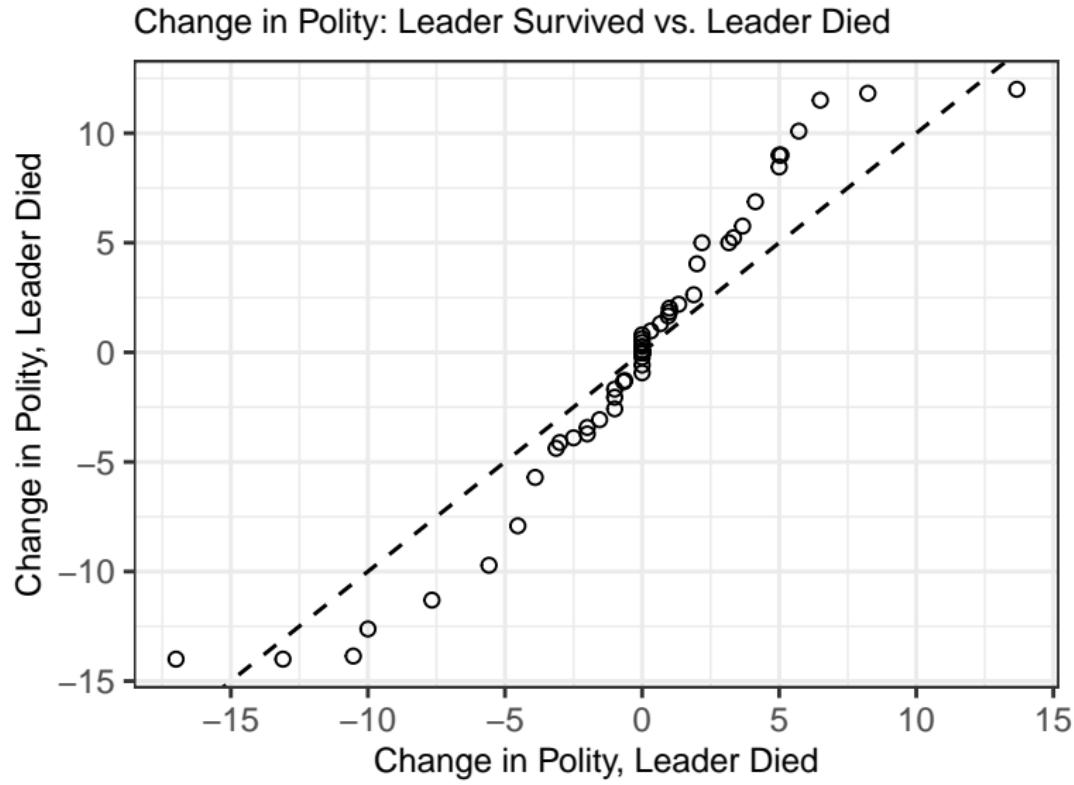
Boxplot comparisons in R

```
leaders %>%
  ggplot(aes(y = politybefore,
             x = factor(fatal, labels = c("Survived", "Died"))),
         geom_boxplot() +
         scale_y_continuous(breaks = seq(-10, 10, by = 5)) +
         labs(title = "Pre-attempt Democracy Level by Attempt Outcome",
              y = "Pre-Attempt Polity Score",
              x = "Assassination Attempt Outcome") +
         theme_bw() +
         theme(plot.title = element_text(size=9),
               axis.title.x = element_text(size = 9),
               axis.title.y = element_text(size = 9),
               panel.grid.major = element_blank(),
               panel.grid.minor = element_blank())
```

Quantile-Quantile Plot

- How do we compare distributions of two variables that are not in the same dataset?
 - Could use boxplots, but it's only a crude summary of the distributions.
- **Quantile-quantile plot (Q-Q plot)**: scatterplot of **quantiles**
 - (min of X , min of Y)
 - (median of X , median of Y)
 - (25th percentile of X , 25th percentile of Y)
- Intuitions:
 - If distributions are the same \rightsquigarrow all points on a 45-degree line
 - Points above 45° line \rightsquigarrow y-axis variable has larger value of the quantile
 - Point below 45° line \rightsquigarrow x-axis variable has larger value of the quantile
 - Steeper slope than 45° line \rightsquigarrow y-axis variable has more spread
 - Flatter slope than 45° line \rightsquigarrow x-axis variable has more spread

QQ-plot example



QQ-plot example (setup)

```
## calculate change in polity
leaders <- leaders %>%
  mutate(polity_change = polityafter - politybefore)

## set quantile vectors
quantile_probs <- seq(from = 0, to = 1, by = 0.01)
quantile_names <- as.character(quantile_probs)

## generate dataframe for plot
quantiles <- leaders %>%
  group_by(fatal) %>%
  summarize(politychng_quantile = quantile(polity_change, probs = q,
                                             quantile = quantile_names) %>%
    pivot_wider(names_from = fatal,
                values_from = politychng_quantile)
```

QQ-plot example (plot)

```
quantiles %>%
  ggplot(aes(x = `0`, y = `1`)) +
  geom_point(shape = 1) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") +
  scale_y_continuous(breaks = seq(-20, 15, by = 5)) +
  scale_x_continuous(breaks = seq(-20, 15, by = 5)) +
  labs(title = "Change in Polity: Leader Survived vs. Leader Died",
       y = "Change in Polity, Leader Died",
       x = "Change in Polity, Leader Died") +
  theme_bw() +
  theme(plot.title = element_text(size=9),
        axis.title.x = element_text(size = 9),
        axis.title.y = element_text(size = 9))
```