Gov 50: 9. Survey Sampling

Matthew Blackwell

Harvard University

Roadmap

1. Proportion tables

2. Measurement

1/ Proportion tables

CCES Data

```r
library(gov50data)
cces_2020
```

```
## # A tibble: 51,551 x 6
##    gender race  educ                pid3  turnout_self pres_vote
##    <fct>  <fct> <fct>               <fct>        <dbl> <fct>
##  1 Male   White 2-year              Repu~            1 Donald J~
##  2 Female White Post-grad           Demo~           NA <NA>
##  3 Female White 4-year              Inde~            1 Joe Bide~
##  4 Female White 4-year              Demo~            1 Joe Bide~
##  5 Male   White 4-year              Inde~            1 Other
##  6 Male   White Some college        Repu~            1 Donald J~
##  7 Male   Black Some college        Not ~           NA <NA>
##  8 Female White Some college        Inde~            1 Donald J~
##  9 Female White High school gr~ Repu~            1 Donald J~
## 10 Female White 4-year              Demo~            1 Joe Bide~
## # i 51,541 more rows
```

Mutate after summarizing

```r
cces_2020 |>
  group_by(pres_vote) |>
  summarize(n = n()) |>
  mutate(prop = n / sum(n))
```

```
## # A tibble: 7 x 3
##   pres_vote                         n     prop
##   <fct>                         <int>    <dbl>
## 1 Joe Biden (Democrat)          26188 0.508
## 2 Donald J. Trump (Republican) 17702 0.343
## 3 Other                          1458 0.0283
## 4 I did not vote in this race     100 0.00194
## 5 I did not vote                   13 0.000252
## 6 Not sure                        190 0.00369
## 7 <NA>                           5900 0.114
```

Another approach

```
cces_2020 |>
  group_by(pres_vote) |>
  summarize(prop = n() / nrow(cces_2020))
```

```
## # A tibble: 7 x 2
##   pres_vote                        prop
##   <fct>                           <dbl>
## 1 Joe Biden (Democrat)         0.508
## 2 Donald J. Trump (Republican) 0.343
## 3 Other                        0.0283
## 4 I did not vote in this race  0.00194
## 5 I did not vote               0.000252
## 6 Not sure                     0.00369
## 7 <NA>                         0.114
```

Another approach

```
cces_2020 |>
  group_by(pres_vote) |>
  summarize(prop = n() / nrow(cces_2020))
```

```
## # A tibble: 7 x 2
##   pres_vote                         prop
##   <fct>                            <dbl>
## 1 Joe Biden (Democrat)          0.508
## 2 Donald J. Trump (Republican) 0.343
## 3 Other                         0.0283
## 4 I did not vote in this race  0.00194
## 5 I did not vote                0.000252
## 6 Not sure                      0.00369
## 7 <NA>                          0.114
```

Doesn't work if you have filtered the data in any way during the pipe

Multiple grouping variables

What happens with multiple grouping variables

```
vote_by_party <- cces_2020 |>
  filter(pres_vote %in% c("Joe Biden (Democrat)",
                          "Donald J. Trump (Republican)")) |>
  mutate(pres_vote = if_else(pres_vote == "Joe Biden (Democrat)",
                             "Biden", "Trump")) |>
  group_by(pid3, pres_vote) |>
  summarize(n = n()) |>
  mutate(prop = n / sum(n)) |>
  select(-n)

vote_by_party
```

```
## # A tibble: 10 x 3
## # Groups:   pid3 [5]
##    pid3        pres_vote   prop
##    <fct>       <chr>      <dbl>
##  1 Democrat    Biden     0.968
##  2 Democrat    Trump     0.0319
##  3 Republican  Biden     0.0712
##  4 Republican  Trump     0.929
##  5 Independent Biden     0.571
##  6 Independent Trump     0.429
##  7 Other       Biden     0.487
##  8 Other       Trump     0.513
##  9 Not sure    Biden     0.599
## 10 Not sure    Trump     0.401
```

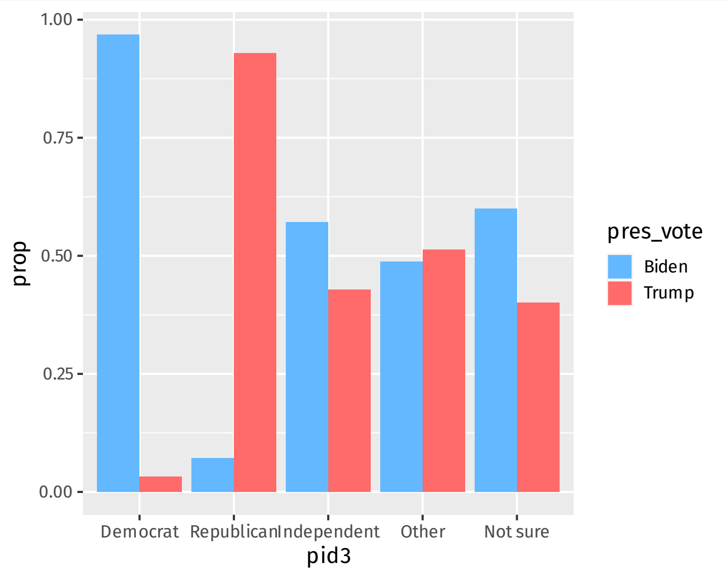
```
## # A tibble: 10 x 3
## # Groups:   pid3 [5]
##    pid3        pres_vote   prop
##    <fct>       <chr>      <dbl>
##  1 Democrat    Biden     0.968
##  2 Democrat    Trump     0.0319
##  3 Republican  Biden     0.0712
##  4 Republican  Trump     0.929
##  5 Independent Biden     0.571
##  6 Independent Trump     0.429
##  7 Other       Biden     0.487
##  8 Other       Trump     0.513
##  9 Not sure    Biden     0.599
## 10 Not sure    Trump     0.401
```

With multiple grouping variables, `summarize()` drops the last one.

Visualizing the cross-tab

We can visualize this using the `fill` aesthetic and `position="dodge"`:

```
ggplot(vote_by_party,
       aes(x = pid3, y = prop, fill = pres_vote)) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c(Biden = "steelblue1", Trump = "indianred1"))
```


1.00
0.75
0.50
0.25
0.00
prop
Democrat
Republican
Independent
Other
Not sure
pid3
pres_vote
Biden
Trump

Pivoting to create cross-tab

```r
cces_2020 |>
  filter(pres_vote %in% c("Joe Biden (Democrat)",
                          "Donald J. Trump (Republican)")) |>
  mutate(pres_vote = if_else(pres_vote == "Joe Biden (Democrat)",
                             "Biden", "Trump")) |>
  group_by(pid3, pres_vote) |>
  summarize(n = n()) |>
  mutate(prop = n / sum(n)) |>
  select(-n) |>
  pivot_wider(
    names_from = pid3,
    values_from = prop
  )
```

```
## # A tibble: 2 x 6
##   pres_vote Democrat Republican Independent Other `Not sure`
##   <chr>        <dbl>      <dbl>       <dbl> <dbl>      <dbl>
## 1 Biden       0.968      0.0712       0.571 0.487      0.599
## 2 Trump       0.0319     0.929        0.429 0.513      0.401
```

What if we want row proportions?

Switch the grouping variables to switch denominator:

```r
cces_2020 |>
  filter(pres_vote %in% c("Joe Biden (Democrat)",
                          "Donald J. Trump (Republican)")) |>
  mutate(pres_vote = if_else(pres_vote == "Joe Biden (Democrat)",
                             "Biden", "Trump")) |>
  group_by(pres_vote, pid3) |>
  summarize(n = n()) |>
  mutate(prop = n / sum(n)) |>
  select(-n) |>
  pivot_wider(
    names_from = pid3,
    values_from = prop
  )
```

```
## # A tibble: 2 x 6
## # Groups:   pres_vote [2]
##   pres_vote Democrat Republican Independent  Other
##   <chr>        <dbl>      <dbl>       <dbl>  <dbl>
## 1 Biden       0.674      0.0327       0.252 0.0281
## 2 Trump       0.0328     0.631        0.280 0.0437
## # i 1 more variable: `Not sure` <dbl>
```

Proportion of all observations

If we want the proportion of all rows, drop all groups

```
cces_2020 |>
  filter(pres_vote %in% c("Joe Biden (Democrat)",
                          "Donald J. Trump (Republican)")) |>
  mutate(pres_vote = if_else(pres_vote == "Joe Biden (Democrat)",
                             "Biden", "Trump")) |>
  group_by(pid3, pres_vote) |>
  summarize(n = n(), .groups = "drop") |>
  mutate(prop = n / sum(n)) |>
  select(-n) |>
  pivot_wider(
    names_from = pid3,
    values_from = prop
  )
```

```
## # A tibble: 2 x 6
##   pres_vote Democrat Republican Independent  Other
##   <chr>        <dbl>      <dbl>       <dbl>  <dbl>
## 1 Biden       0.402      0.0195       0.150 0.0167
## 2 Trump       0.0132     0.254        0.113 0.0176
## # i 1 more variable: `Not sure` <dbl>
```

2/ Measurement

Where does data come from?

- Social science is about developing and testing **causal theories**:

Where does data come from?

- Social science is about developing and testing **causal theories**:
    - Does minimum wage change levels of employment?

Where does data come from?

- Social science is about developing and testing **causal theories**:
    - Does minimum wage change levels of employment?
    - Does outgroup contact influence views on immigration?

Where does data come from?

- Social science is about developing and testing **causal theories**:
    - Does minimum wage change levels of employment?
    - Does outgroup contact influence views on immigration?
- Theories are made up of **concepts**:

Where does data come from?

- Social science is about developing and testing **causal theories**:
    - Does minimum wage change levels of employment?
    - Does outgroup contact influence views on immigration?
- Theories are made up of **concepts**:
    - Minimum wage, level of employment, outgroup contact, views on immigration.

Where does data come from?

- Social science is about developing and testing **causal theories**:
    - Does minimum wage change levels of employment?
    - Does outgroup contact influence views on immigration?
- Theories are made up of **concepts**:
    - Minimum wage, level of employment, outgroup contact, views on immigration.
    - We took these for granted when talking about causality.

Where does data come from?

- Social science is about developing and testing **causal theories**:
    - Does minimum wage change levels of employment?
    - Does outgroup contact influence views on immigration?
- Theories are made up of **concepts**:
    - Minimum wage, level of employment, outgroup contact, views on immigration.
    - We took these for granted when talking about causality.
- Need **operational definition** to concretely measure these concepts

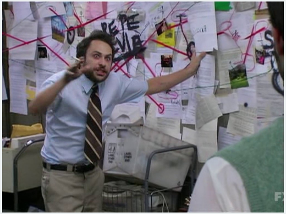
Concepts vary in how observable they are

Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws

Observable by survey



Not directly observable

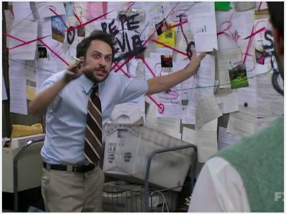
Concepts vary in how observable they are

Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws
- Sensor measurements

Observable by survey



Not directly observable

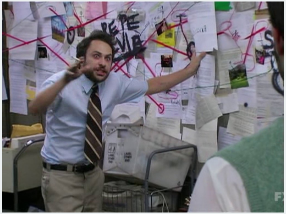
Concepts vary in how observable they are

Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws
- Sensor measurements
- Election results

Observable by survey



Not directly observable

Concepts vary in how observable they are

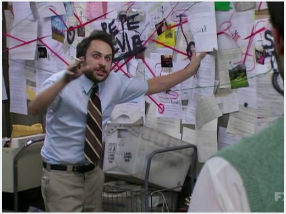
Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws
- Sensor measurements
- Election results

Observable by survey

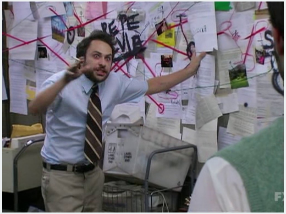
- Age of a person



Not directly observable

Concepts vary in how observable they are

Kinds of measurement arranged by how direct we can measure them:



Observable in the world

- Minimum wage laws
- Sensor measurements
- Election results

Observable by survey

- Age of a person
- Employment status

Not directly observable

Concepts vary in how observable they are

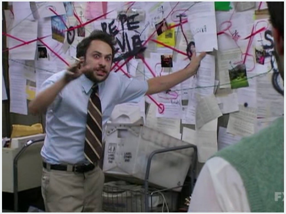
Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws
- Sensor measurements
- Election results

Observable by survey

- Age of a person
- Employment status
- Presidential approval



Not directly observable

Concepts vary in how observable they are

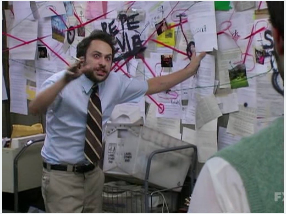
Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws
- Sensor measurements
- Election results

Observable by survey

- Age of a person
- Employment status
- Presidential approval



Not directly observable

- A person’s ideology

Concepts vary in how observable they are

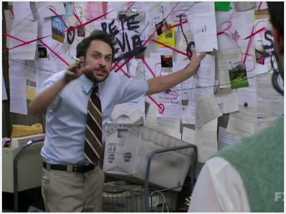
Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws
- Sensor measurements
- Election results

Observable by survey

- Age of a person
- Employment status
- Presidential approval



Not directly observable

- A person's ideology
- Levels of democracy

Concepts vary in how observable they are

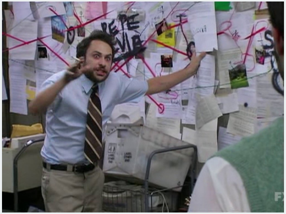
Kinds of measurement arranged by how direct we can measure them:

Observable in the world

- Minimum wage laws
- Sensor measurements
- Election results

Observable by survey

- Age of a person
- Employment status
- Presidential approval



Not directly observable

- A person's ideology
- Levels of democracy
- Extent of gerrymandering

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 more supportive, how much would you say you support the job that Joe Biden is doing as president?”

Measurement error

Table 1
Response to citizenship question across two-waves of CCES panel.

Response in 2010	Response in 2012	Number of respondents	Percentage
Citizen	Citizen	18,737	99.25
Citizen	Non-Citizen	20	0.11
Non-Citizen	Citizen	36	0.19
Non-Citizen	Non-Citizen	85	0.45

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

Measurement error

Table 1
Response to citizenship question across two-waves of CCES panel.

Response in 2010	Response in 2012	Number of respondents	Percentage
Citizen	Citizen	18,737	99.25
Citizen	Non-Citizen	20	0.11
Non-Citizen	Citizen	36	0.19
Non-Citizen	Non-Citizen	85	0.45

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

Measurement error

Table 1
Response to citizenship question across two-waves of CCES panel.

Response in 2010	Response in 2012	Number of respondents	Percentage
Citizen	Citizen	18,737	99.25
Citizen	Non-Citizen	20	0.11
Non-Citizen	Citizen	36	0.19
Non-Citizen	Non-Citizen	85	0.45

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

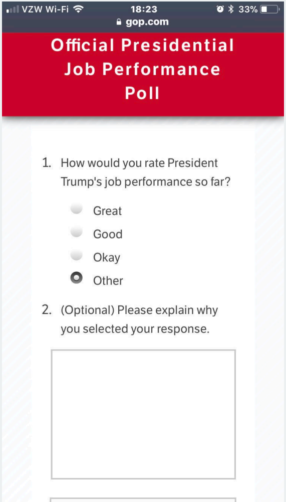
Measurement error

Table 1
Response to citizenship question across two-waves of CCES panel.

Response in 2010	Response in 2012	Number of respondents	Percentage
Citizen	Citizen	18,737	99.25
Citizen	Non-Citizen	20	0.11
Non-Citizen	Citizen	36	0.19
Non-Citizen	Non-Citizen	85	0.45

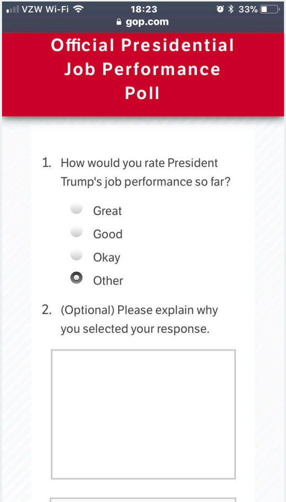
- **Measurement error**: chance variation in our measurements.
    - individual measurement = exact value + chance error
    - chance errors tend to cancel out when we take averages.
    - why? often data entry errors or faulty memories.

Bias



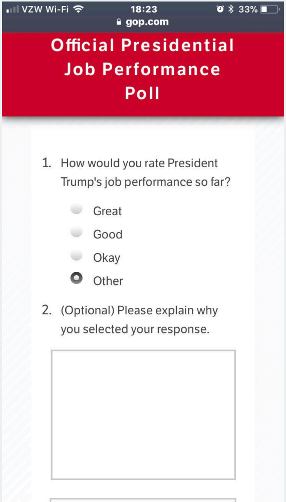
- **Bias**: systematic errors for all units in the same direction.

Bias



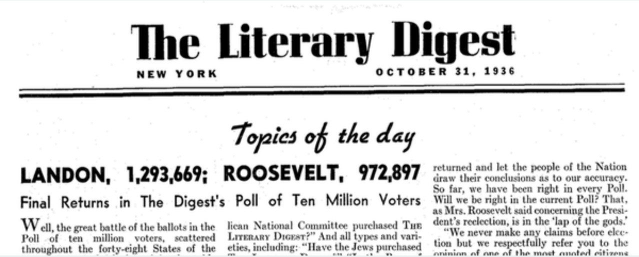
- **Bias**: systematic errors for all units in the same direction.
- individual measurement = exact value + bias + chance error.

Bias



- **Bias**: systematic errors for all units in the same direction.
- individual measurement = exact value + bias + chance error.
- "What did you eat yesterday?" $\rightsquigarrow$ underreporting

1936 Literary Digest Poll



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

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

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

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

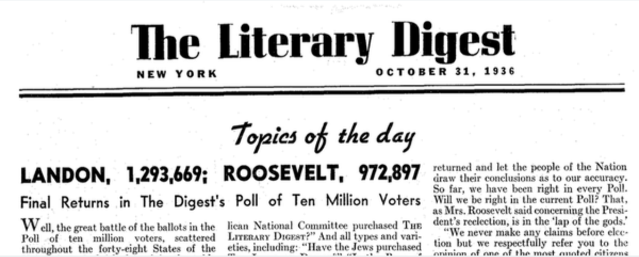
lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

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 Mrs. Roosevelt said concerning the President's reelection, 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 quoted citizens

- Literary Digest predicted elections using mail-in polls.

1936 Literary Digest Poll



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

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

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

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

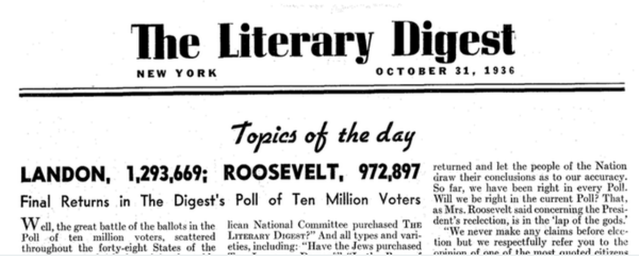
lican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased

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 Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

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

1936 Literary Digest Poll



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

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

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

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

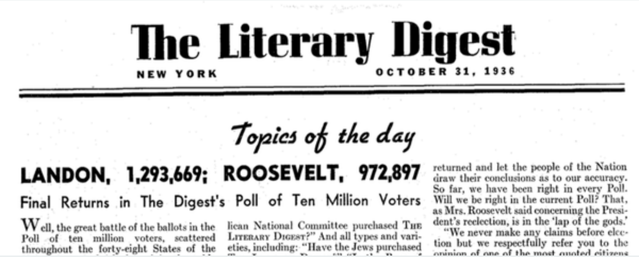
lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

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 Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

- 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



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

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

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

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

lican National Committee purchased THE 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 Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

- 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.

Poll fail

	FDR's Vote Share
Literary Digest	43%
George Gallup	56%
Actual Outcome	62%

Poll fail

	FDR's Vote Share
Literary Digest	43%
George Gallup	56%
Actual Outcome	62%

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

Poll fail

	FDR's Vote Share
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's Vote Share
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's Vote Share
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.
- $\rightsquigarrow$ 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	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	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	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	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	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 $\rightsquigarrow$ **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	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 $\rightsquigarrow$ **selection bias**
- Republicans easier to find within quotas (phones, listed addresses)

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: 617-495-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: 617-495-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: 617-495-XXXX.
    - Randomly choose each digit in XXXX to call a particular phone.
    - Every phone in America has an equal chance of being included in sample.

Sampling lingo

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

Sampling lingo

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

Sampling lingo

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

- 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.