Sampling and Representativeness

Department of Government
London School of Economics and Political Science

1 Simple Statistics

2 Research Design Proposal

3 Representativeness

4 Sampling

Relationship

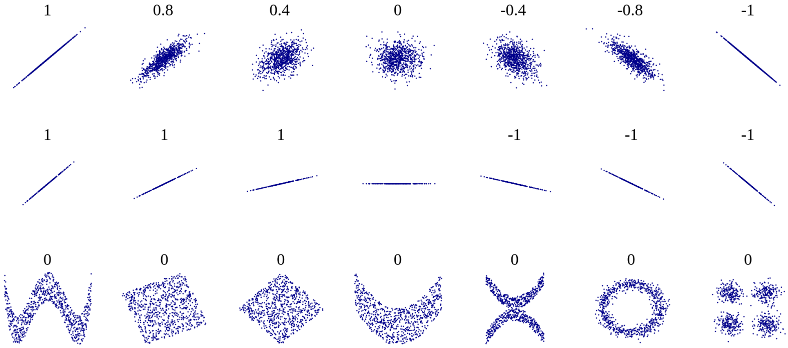
- Covariance:

$$Cov(X,Y) = \Sigma_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Relationship

- Covariance:
  $$Cov(X,Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
- Correlation:
  $$Corr(X,Y) = r_{x,y} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_x s_y}$$

Correlation is linear!



Source: Wikimedia

Guess the Correlation!

1. Go to:

   `http://guessthecorrelation.com/`

2. Play a few rounds

1 Simple Statistics

2 Research Design Proposal

3 Representativeness

4 Sampling

Rubric for Research Design Proposal

- Politically interesting and important topic and research question
- Research question, theory, and methods situated in relevant political science literature
- Thorough discussion of theory and explicitly stated, empirically testable hypotheses
- Appropriate, well-designed research methodology
    - Strengths and weaknesses
    - Trade-offs discussed against alternative designs
    - Able to generate data that will allow a test of hypotheses
- Well-written, coherent, well-structured

1 Simple Statistics

2 Research Design Proposal

3 Representativeness

4 Sampling

Case selection

- Earlier in the course we talked about case selection
- What did we discuss?

Discuss in Pairs!

What does it mean for a sample to be representative of a population?

Different conceptualizations of representativeness

- **Design-based**: A sample is representative because of how it was drawn (e.g., randomly)
- **Demographic-based**: A sample is representative because it resembles in the population some way (e.g., same proportion of women in sample and population, etc.)
- **Expert judgement**: A sample is representative as judged by an expert who deems it "fit for purpose"

Obtaining Representativeness

Obtaining Representativeness

- Census

Obtaining Representativeness

- Census
- Convenience/Purposive samples

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)
- Simple random sampling

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)
- Simple random sampling
- Complex survey designs

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (common before 1940s)
- **Simple random sampling**
- Complex survey designs

1 Simple Statistics

2 Research Design Proposal

3 Representativeness

4 Sampling

Inference from Sample to Population

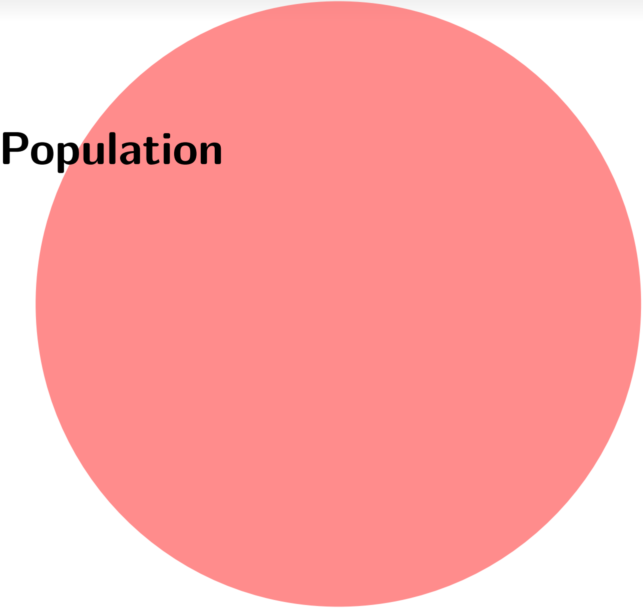
- We want to know population parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

Inference from Sample to Population

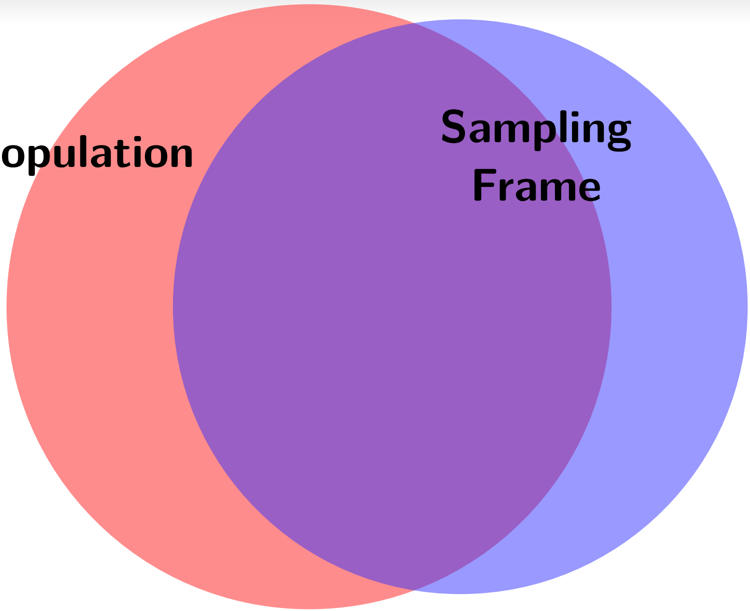
- We want to know population parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain
- What range of values for θ does our $\hat{\theta}$ imply?

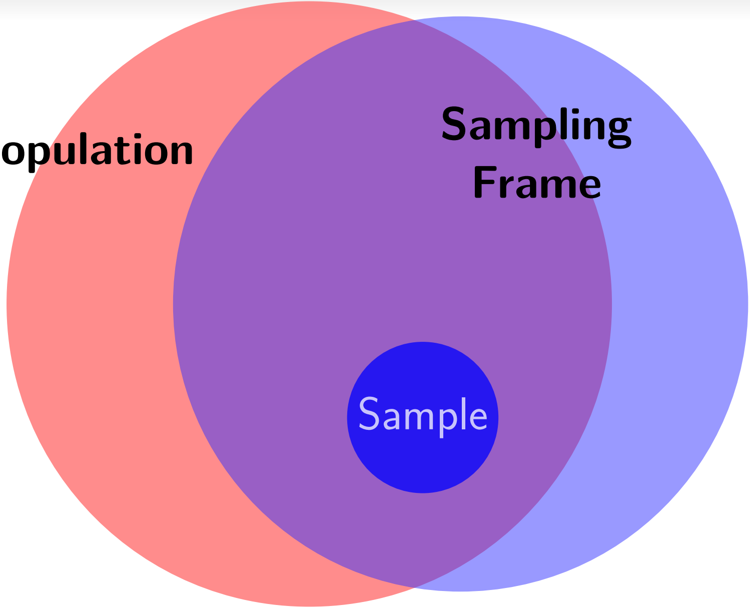
Simple Random Sampling

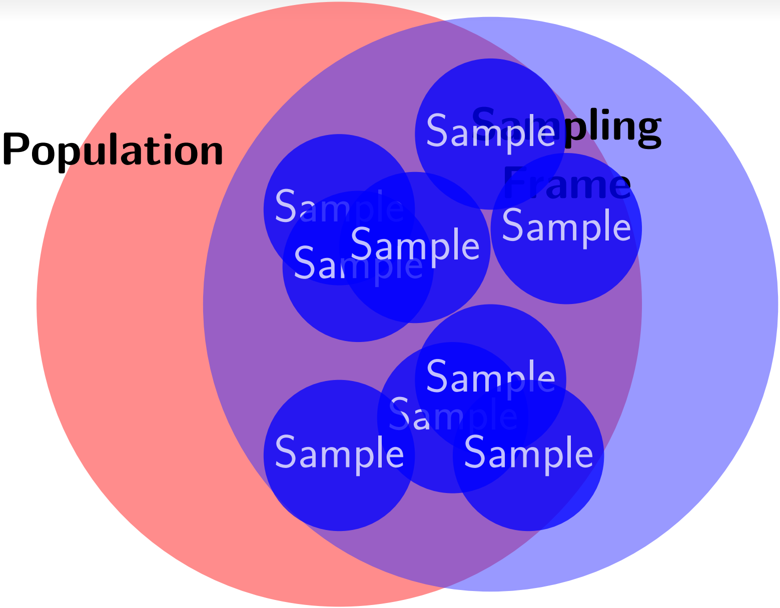
1. Define target population
2. Create “sampling frame”
3. Each unit in frame has equal probability of selection
4. Collect data on each unit
5. Calculate sample *statistic*
6. Draw an inference to the population



Population


Population
Sampling Frame


Population
Sampling
Frame
Sample


Population
Sampling Frame
Sample
Sample
Sample
Sample
Sample
Sample
Sample
Sample
Sample

Simple Random Sampling

1. Define target population
2. Create "sampling frame"
3. Each unit in frame has equal probability of selection
4. Collect data on each unit
5. Calculate sample *statistic*
6. Draw an inference to the population

Statistical Inference I

To calculate a sample mean (or proportion):

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1}$$

where $y_i =$ value for a unit, and

$n =$ sample size

Statistical Inference II

- If we calculate $\bar{y}$ in our *sample*, what does this tell us about the $\bar{Y}$ in the *population*?

Statistical Inference II

- If we calculate $\bar{y}$ in our *sample*, what does this tell us about the $\bar{Y}$ in the *population*?
- The sample *estimate* is our guess at the value of the population *parameter* within some degree of uncertainty

Law of Large Numbers

- Definition: The *mean* of the $\hat{\theta}$ from each of a number of samples will converge on the population θ, as the number of samples increases

Sampling Variance

- The $\hat{\theta}$ in any particular sample can differ from the population value θ
- This variation is calling “sampling variance” or “sampling error”
- The standard error describes the average amount of variation of the $\hat{\theta}$'s around θ

How Uncertain Are We?

- Our uncertainty depends on sampling procedures
- Most importantly, *sample size*
    - As $n \to \infty$, uncertainty $\to 0$
- We typically summarize our uncertainty as the *standard error*

Standard Errors (SEs)

- Definition: "The standard error of a sample estimate is the average distance that a sample estimate ($\hat{\theta}$) would be from the population parameter (θ) if we drew many separate random samples and applied our estimator to each."

Standard Errors (SEs)

- Definition: “The standard error of a sample estimate is the average distance that a sample estimate ($\hat{\theta}$) would be from the population parameter (θ) if we drew many separate random samples and applied our estimator to each.”
- Square root of the sampling variance

Sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2}$$

where y_i = value for a unit, and
n = sample size

$$SE_{\bar{y}} = \sqrt{(1-f)\frac{s^2}{n}} \tag{3}$$

where f = proportion of population sampled,
s^2 = sample (element) variance, and
n = sample size

Sample proportion

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4}$$

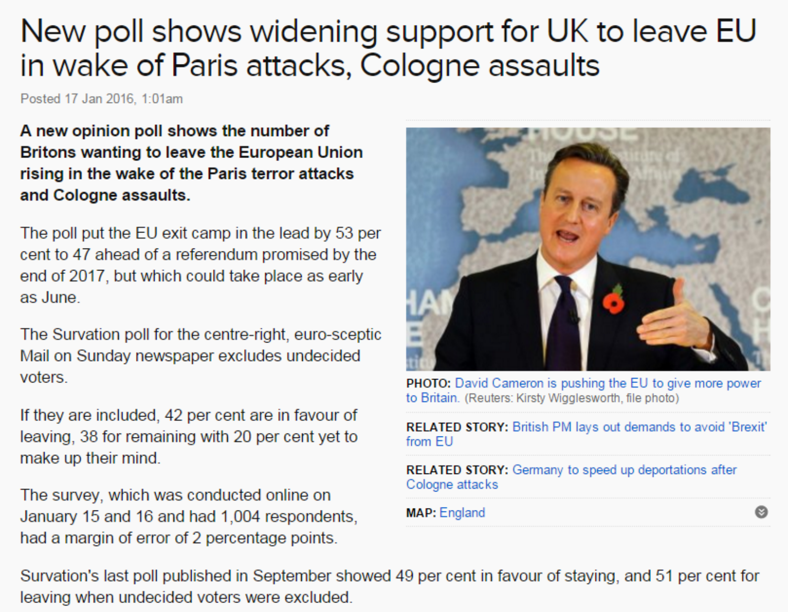
where y_i = value for a unit, and
n = sample size

$$SE_{\bar{y}} = \sqrt{\frac{(1-f)}{n}p(1-p)} \tag{5}$$

where f = proportion of population sampled,
p = sample proportion, and
n = sample size

Margin of Error

- Common to express SE as a “margin of error”
- MoE is twice the Standard Error
- For estimated proportions, expressed as: “+/- MoE percentage points”



New poll shows widening support for UK to leave EU in wake of Paris attacks, Cologne assaults

Posted 17 Jan 2016, 1:01am

A new opinion poll shows the number of Britons wanting to leave the European Union rising in the wake of the Paris terror attacks and Cologne assaults.

The poll put the EU exit camp in the lead by 53 per cent to 47 ahead of a referendum promised by the end of 2017, but which could take place as early as June.

The Survation poll for the centre-right, euro-sceptic Mail on Sunday newspaper excludes undecided voters.

If they are included, 42 per cent are in favour of leaving, 38 for remaining with 20 per cent yet to make up their mind.

The survey, which was conducted online on January 15 and 16 and had 1,004 respondents, had a margin of error of 2 percentage points.

Survation's last poll published in September showed 49 per cent in favour of staying, and 51 per cent for leaving when undecided voters were excluded.

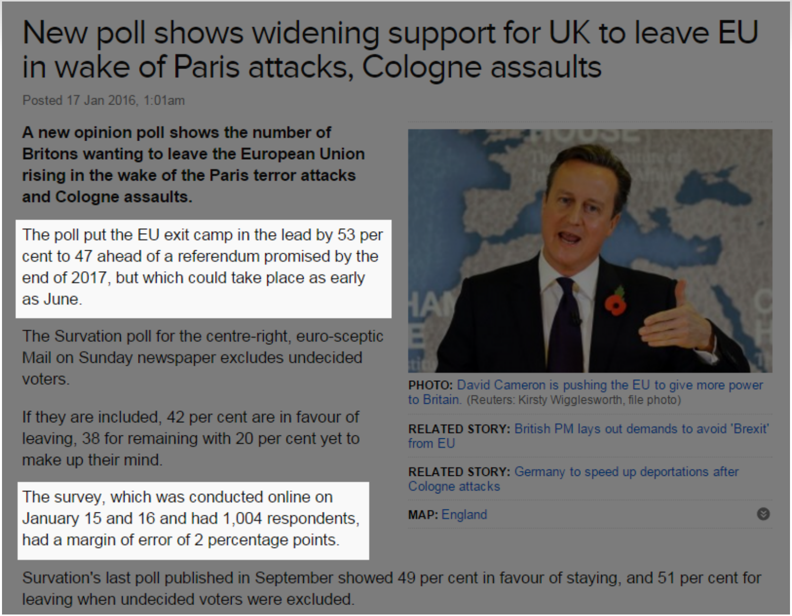
PHOTO: David Cameron is pushing the EU to give more power to Britain. (Reuters: Kirsty Wigglesworth, file photo)

RELATED STORY: British PM lays out demands to avoid 'Brexit' from EU

RELATED STORY: Germany to speed up deportations after Cologne attacks

MAP: England

Source: `http://www.abc.net.au/news/2016-01-17/new-poll-show-widening-support-for-uk-to-leave-eu/7093730`



New poll shows widening support for UK to leave EU in wake of Paris attacks, Cologne assaults

Posted 17 Jan 2016, 1:01am

A new opinion poll shows the number of Britons wanting to leave the European Union rising in the wake of the Paris terror attacks and Cologne assaults.

The poll put the EU exit camp in the lead by 53 per cent to 47 ahead of a referendum promised by the end of 2017, but which could take place as early as June.

The Survation poll for the centre-right, euro-sceptic Mail on Sunday newspaper excludes undecided voters.

If they are included, 42 per cent are in favour of leaving, 38 for remaining with 20 per cent yet to make up their mind.

The survey, which was conducted online on January 15 and 16 and had 1,004 respondents, had a margin of error of 2 percentage points.

Survation's last poll published in September showed 49 per cent in favour of staying, and 51 per cent for leaving when undecided voters were excluded.

PHOTO: David Cameron is pushing the EU to give more power to Britain. (Reuters: Kirsty Wigglesworth, file photo)

RELATED STORY: British PM lays out demands to avoid 'Brexit' from EU

RELATED STORY: Germany to speed up deportations after Cologne attacks

MAP: England

Source: `http://www.abc.net.au/news/2016-01-17/new-poll-show-widening-support-for-uk-to-leave-eu/7093730`

Questions?

Activity!

What proportion of all Haribo Starmix gummies are ♡s?

What proportion of all Haribo Starmix gummies are ♡s?

1. Everyone collect a random sample

What proportion of all Haribo Starmix gummies are ♡s?

1. Everyone collect a random sample
2. Calculate $\hat{p} = \frac{\heartsuit}{n}$

What proportion of all Haribo Starmix gummies are ♡s?

1. Everyone collect a random sample
2. Calculate $\hat{p} = \frac{\heartsuit}{n}$
3. Report $\hat{p}$

What proportion of all Haribo Starmix gummies are ♡s?

1. Everyone collect a random sample
2. Calculate $\hat{p} = \frac{\heartsuit}{n}$
3. Report $\hat{p}$
4. Calculate element variance: $p(1-p)$

What proportion of all Haribo Starmix gummies are ♡s?

1. Everyone collect a random sample
2. Calculate $\hat{p} = \frac{\heartsuit}{n}$
3. Report $\hat{p}$
4. Calculate element variance: $p(1-p)$
5. What is your margin of error?

How large of a sample do we need?

How large of a sample do we need?

- Uncertainty is influenced by:
    - Sample size
    - *Element* variance
    - Population size?

How large of a sample do we need?

- Uncertainty is influenced by:
  - Sample size
  - *Element* variance
  - Population size?

- So what do we do?
  - Decide on desired uncertainty
  - Guess at element variance

How large of a sample do we need?

- Uncertainty is influenced by:
  - Sample size
  - *Element* variance
  - Population size?
- So what do we do?
  - Decide on desired uncertainty
  - Guess at element variance
  - Adjust sample size based on feasibility

Estimating sample size

Determining sample size requires:

- A possible value of p
- A desired precision (SE)

So:

$$n = \frac{p(1-p)}{SE^2} = \frac{0.5(1-0.5)}{SE^2} = \frac{0.25}{SE^2} \tag{6}$$

Note: Element variance is highest when $p = 0.5$.

Estimating sample size

What precision (margin of error) do we want?

- +/- 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{7}$$

Estimating sample size

What precision (margin of error) do we want?

- +/- 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \qquad (7)$$

- +/- 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \qquad (8)$$

Estimating sample size

What precision (margin of error) do we want?

- +/- 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{7}$$

- +/- 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \tag{8}$$

- +/- 0.5 percentage points: $SE = 0.0025$

$$n = \frac{0.25}{0.00000625} = 40,000 \tag{9}$$

How do we reduce uncertainty?

How do we reduce uncertainty?

- Increase sample size

How do we reduce uncertainty?

- Increase sample size
- Reduce element variance

How do we reduce uncertainty?

- Increase sample size
- Reduce element variance
- Change sampling design
    - Stratified sampling (tends to decrease SEs)
    - Cluster sampling (tends to increase SEs)

Big Caveat!

- Calculations from today assume SRS
- Lots of data comes from other kinds of samples
- Think of SRS as a baseline

Preview

- Feedback on PS5 this week
- PS6 due Feb. 23
- Next week we'll discuss hypothesis testing
- Remember Reading Week assignment!