Sampling and Representativeness

Department of Government
London School of Economics and Political Science

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive
- Comparative

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive
- Comparative
- Representative

Case selection

Our ambitions about what kind of inferences we want to derive from our descriptions influence how we select cases.

- Purposive
- Comparative
- Representative
  - Unrepresentative

Population

"The complete population of units (observations) we want to understand."

Population

"The complete population of units (observations) we want to understand."

- We rarely observe all population units

Population

"The complete population of units (observations) we want to understand."

- We rarely observe all population units
- A "sample" is a set of units we actually observe

Population

"The complete population of units (observations) we want to understand."

- We rarely observe all population units
- A "sample" is a set of units we actually observe
- Sometimes we aim to *generalize* from the sample to the population

Discuss in Pairs!

What does it mean for a "sample" (set of cases) to be representative of a population?

Different conceptualizations

- **Design-based**: A sample is representative because of how it was drawn (e.g., randomly)
- **Model-based**: A sample is representative because it resembles in the population with respect to certain variables (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 (pre-1940s, post-2000s)

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (pre-1940s, post-2000s)
- Simple random sampling

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (pre-1940s, post-2000s)
- Simple random sampling
- Complex survey designs

Obtaining Representativeness

- Census
- Convenience/Purposive samples
- Quota sampling (pre-1940s, post-2000s)
- **Simple random sampling**
- Complex survey designs

Inference from Sample to Population

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

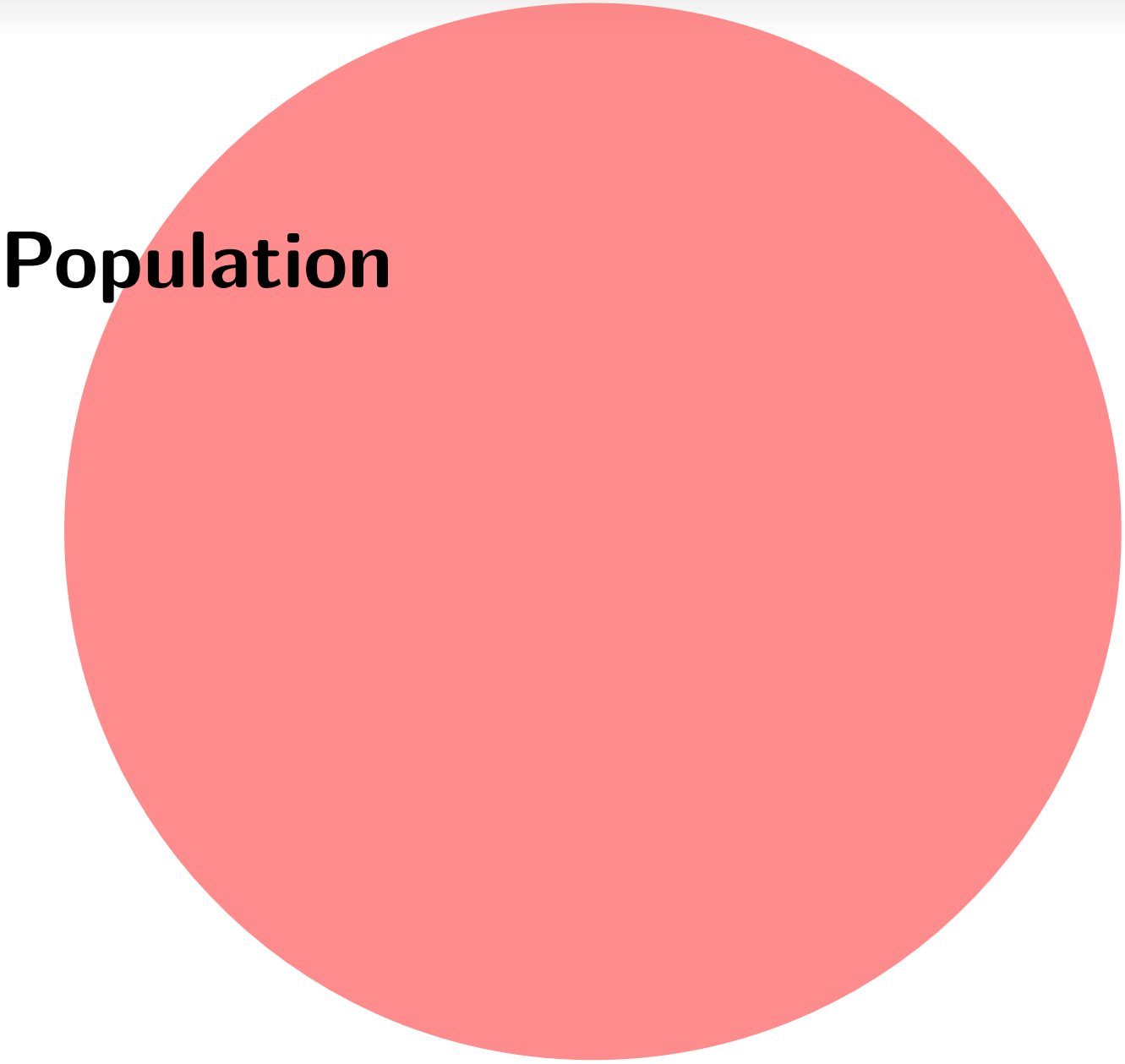
Inference from Sample to Population

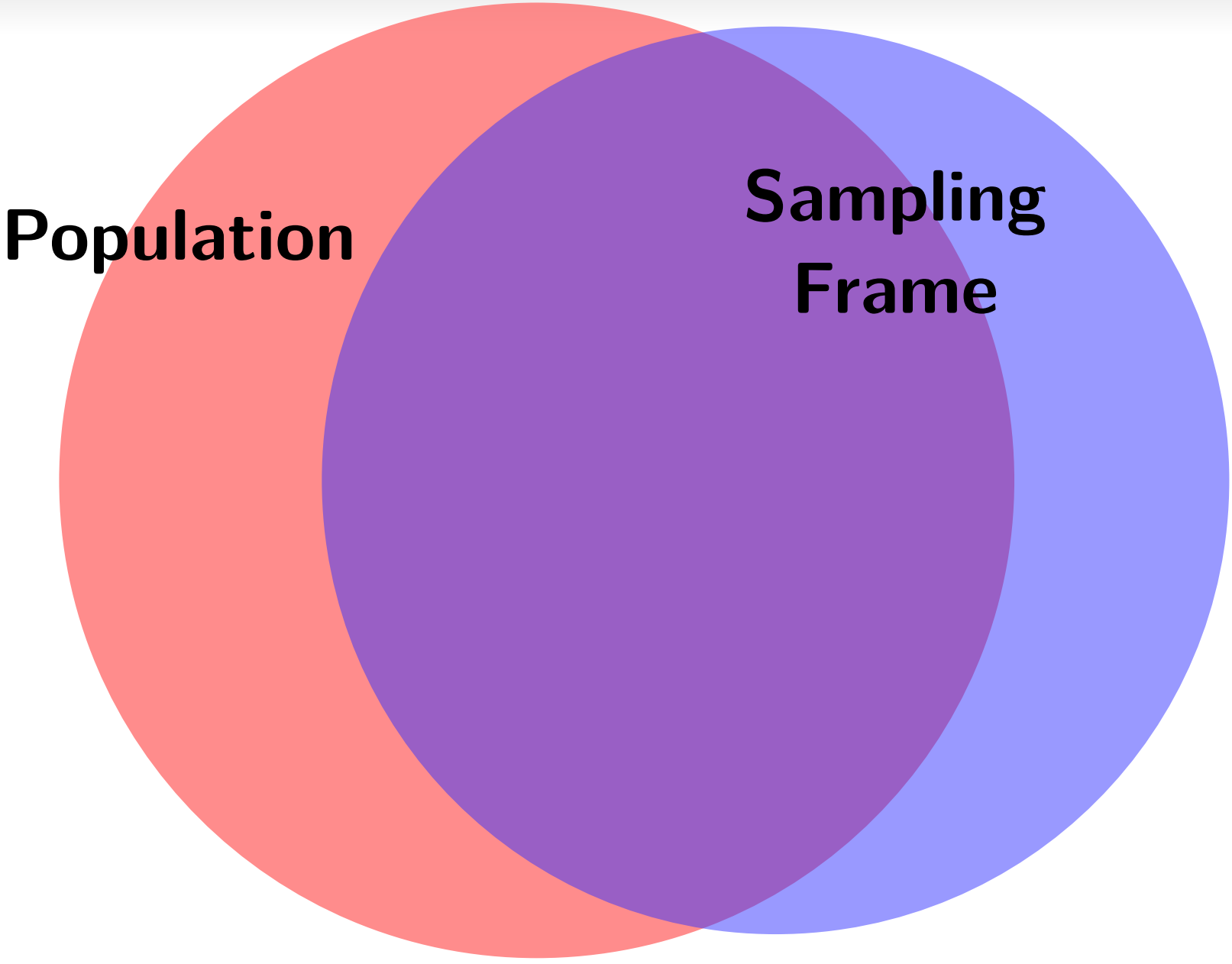
- We want to know pop. parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

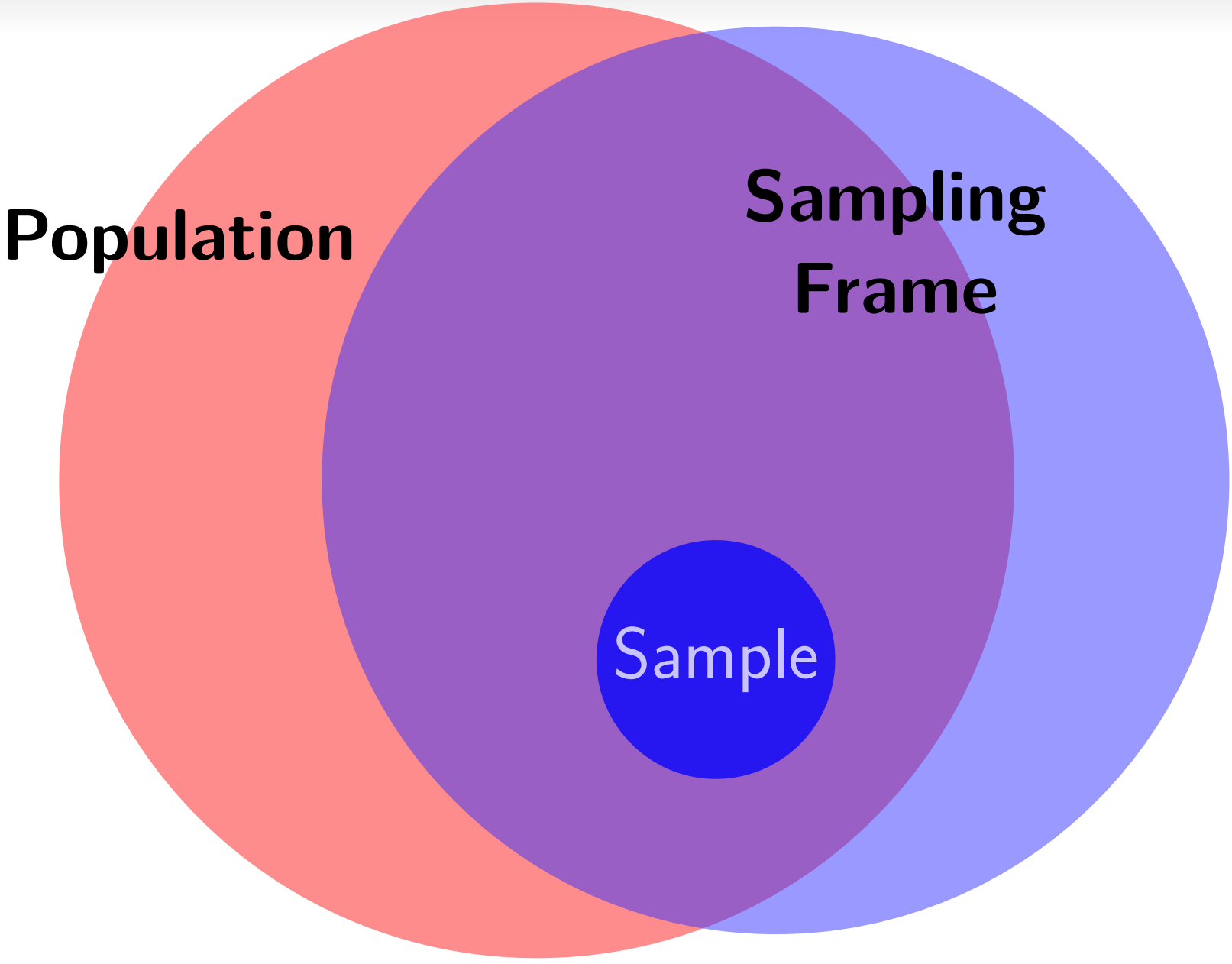
- What range of values for θ does our $\hat{\theta}$ imply?

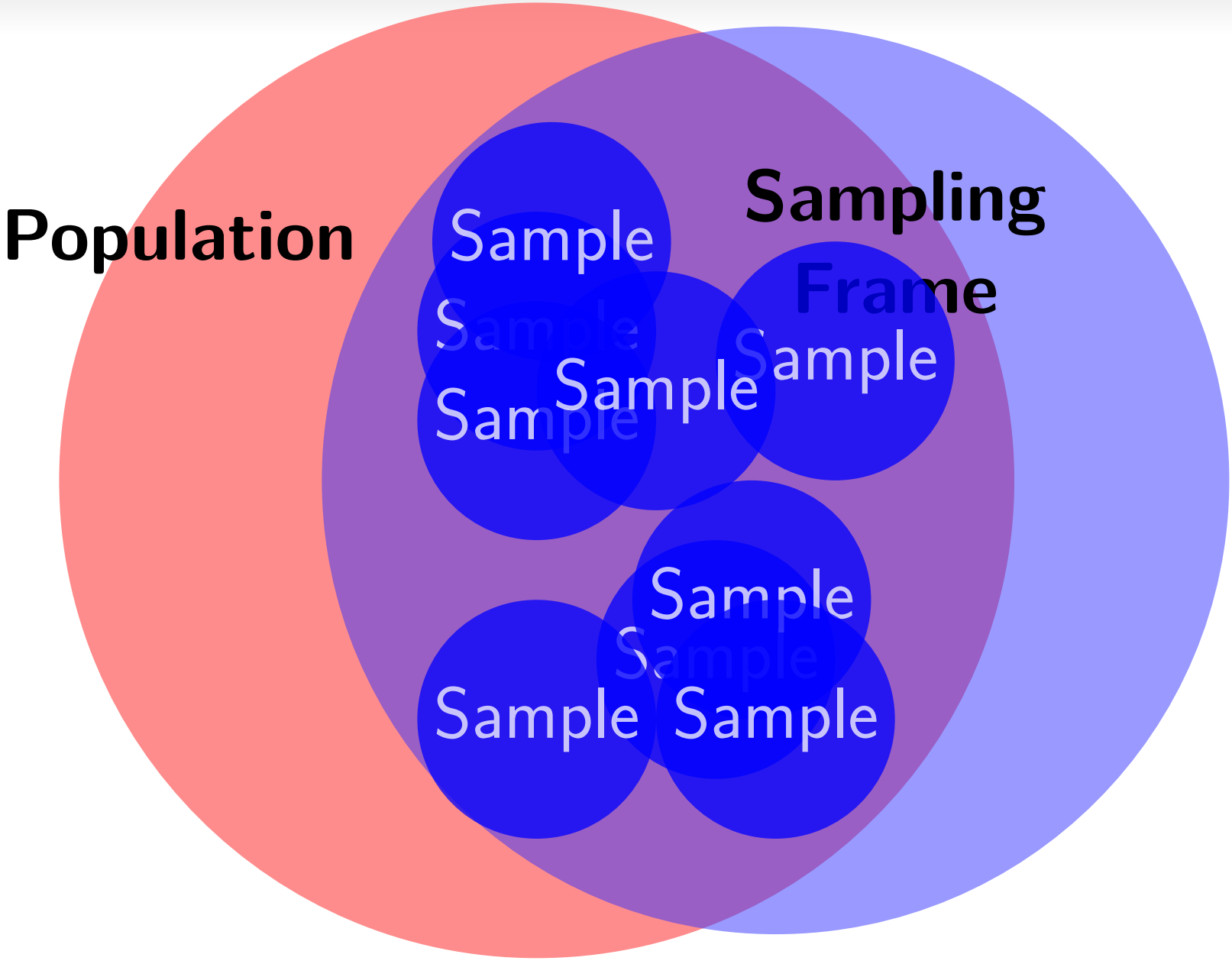
Simple Random Sampling

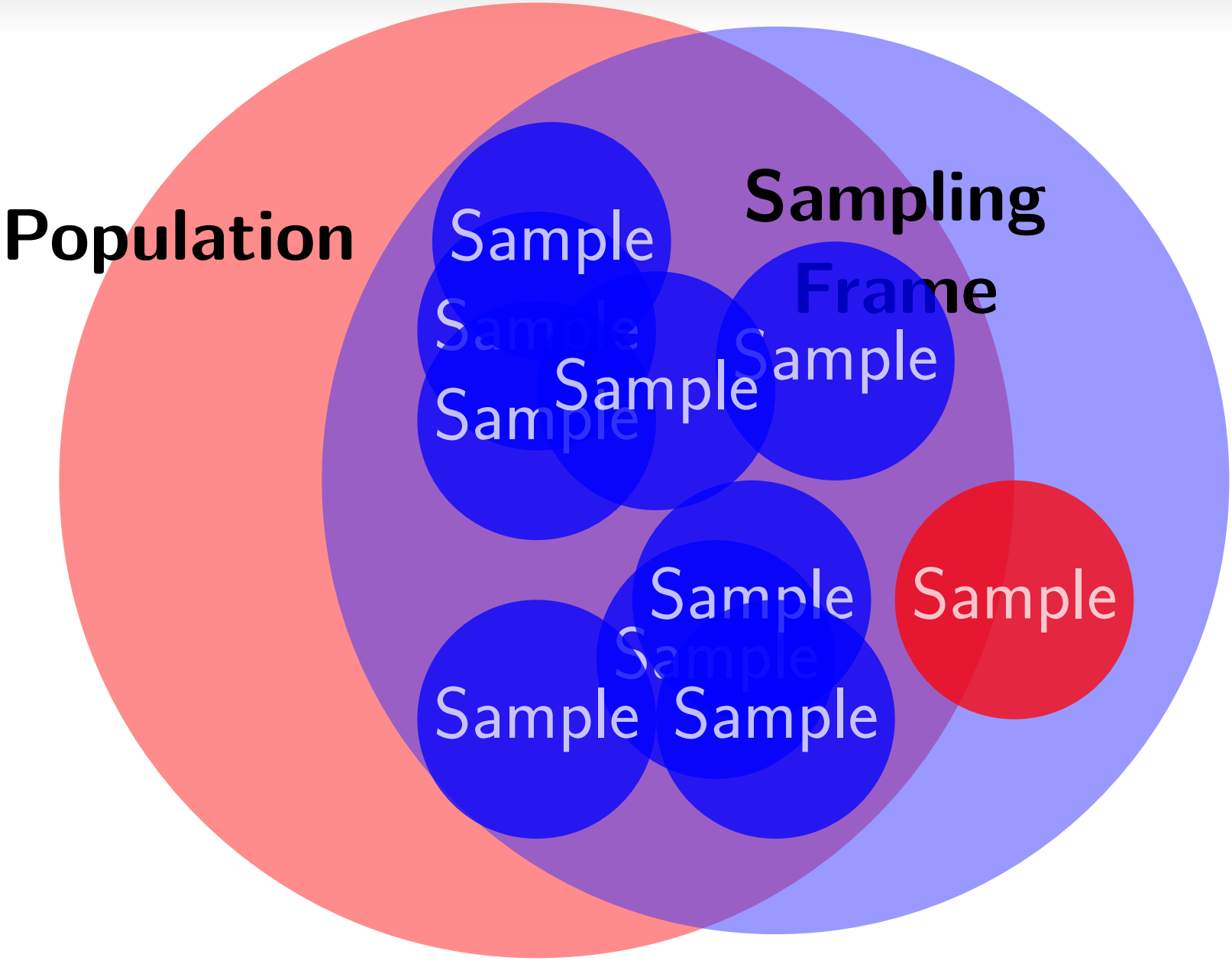
1. Define target population
2. Create "sampling frame"
3. Each unit in frame has equal probability of selection
4. Collect data on each unit
5. Calculate sample *statistic*
6. Draw an inference to the population


Population


Population
Sampling
Frame


Population
Sampling Frame
Sample


Population
Sampling
Frame
Sample
Sample
Sample
Sample
Sample
Sample
Sample
Sample


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

$$Pr(y = 1) = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{4}$$

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

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

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

Margin of Error

- Uncertainty often stated in terms of a "margin of error"
- Standard MoE is twice the SE (x1.96)
- For estimated proportions, expressed as: "p $\pm$ 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.

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

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.

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.

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

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.

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

Questions?

Questions?

(There is an R lab activity about this.)

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 $\hat{p} = \frac{\sum \heartsuit}{n}$

What proportion of all Haribo Starmix gummies are ♡s?

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

What proportion of all Haribo Starmix gummies are ♡s?

1. Everyone collect a random sample
2. Calculate $\hat{p} = \frac{\sum \heartsuit}{n}$
3. Calculate element variance:
   $Var(x) = p(1-p)$
4. Calculate MoE: $\hat{p} \pm \left(2 * \sqrt{\frac{Var(x)}{n}}\right)$

How large of a sample do we need?

[1] Population element variance is estimated by sample element variance.

How large of a sample do we need?

- Uncertainty is influenced by:
  - Sample size
  - *Element* variance[1]
  - Population size?

[1] Population element variance is estimated by sample element variance.

How large of a sample do we need?

- Uncertainty is influenced by:
  - Sample size
  - *Element* variance[1]
  - Population size?
- So what do we do?
  - Decide on desired uncertainty
  - Guess at element variance

[1] Population element variance is estimated by sample element variance.

How large of a sample do we need?

- Uncertainty is influenced by:
  - Sample size
  - *Element* variance[1]
  - Population size?
- So what do we do?
  - Decide on desired uncertainty
  - Guess at element variance
  - Adjust sample size based on feasibility

[1] Population element variance is estimated by sample element variance.

Estimating sample size

What precision (margin of error) do we want?

- ± 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{6}$$

Estimating sample size

What precision (margin of error) do we want?

- ± 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{6}$$

- ± 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \tag{7}$$

Estimating sample size

What precision (margin of error) do we want?

- ± 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{6}$$

- ± 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \tag{7}$$

- ± 0.5 percentage points: $SE = 0.0025$

$$n = \frac{0.25}{0.00000625} = 40,000 \tag{8}$$

Summary

- Various ways to select cases
- Various notions of representativeness
- Case selection is one way of addressing questions of "external validity" but there are others, too