SO5041 Unit 10:

Brendan Halpin, Sociology

Autumn 2023/4

SO5041 Unit 10

Outline

This unit

- Understanding imprecision due to sampling
- Estimating imprecision: confidence intervals
  - For means
  - For proportions
- Reading: Agresti Ch 5, sections 1 to 3.

SO5041 Unit 10

Estimating imprecision: confidence intervals

Samples are uncertain

- The characteristics of representative samples approximate those of the reference population
- But with uncertainty
- How do we characterise this uncertainty?
- With margins of error such as "confidence intervals"

Point estimates

- Agresti: A point estimator of a parameter is a sample statistic that predicts the value of that parameter
- For instance, our sample mean $\bar{X}$ is a point estimator of the population mean μ
- Good point estimators require two things:
  - To be centred around the true value (unbiased)
  - To have as small a sampling error as possible (efficient)

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, “on average”
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, “on average”
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,
- Sample mean:

$$\bar{X} = \frac{\sum X_i}{n}$$

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, "on average"
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,
- Sample mean:

$$\bar{X} = \frac{\sum X_i}{n}$$

- Sample standard deviation:

$$\hat{\sigma} = s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, “on average”
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,
- Sample mean:

$$\bar{X} = \frac{\sum X_i}{n}$$

- Sample standard deviation:

$$\hat{\sigma} = s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

- Sample proportion:

$$\hat{\pi}_1 = \frac{n_1}{n_1 + n_2}$$

sociology
UNIVERSITY OF LIMERICK

Sampling distributions and sampling error

- We have seen that sample estimates have sampling error, and now understand something of its characteristics, by exploring sampling distributions
- Sample estimates can be considered as being drawn from an imaginary random distribution, which (if the estimator is unbiased) will centre on the true population parameter, with a level of imprecision measured by the standard error (which will be as low as possible if the estimator is efficient)

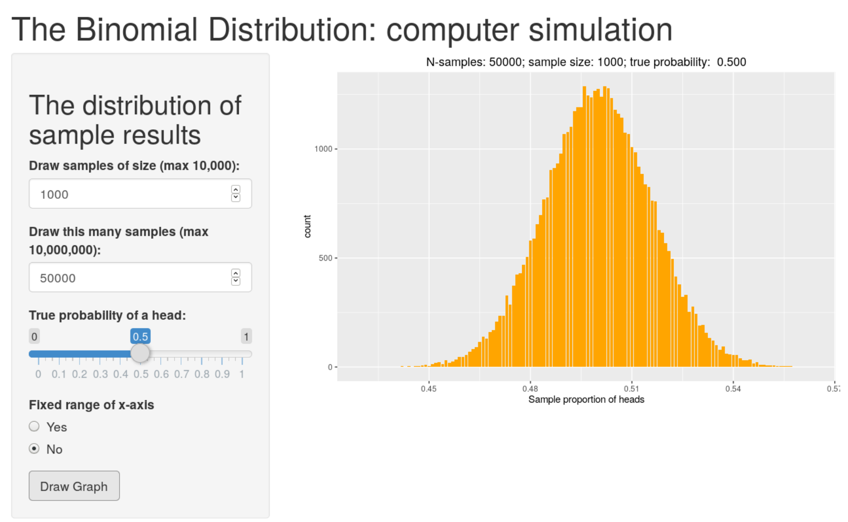
Sampling distributions: Central Limit Theorem

- Where the sample is sufficiently large, the Central Limit Theorem tells us that the sampling distribution is normal, with mean μ and standard deviation $\sigma_{\bar{X}}$
- The standard deviation of the sampling distribution is called the standard error

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}}$$

Simulation: Yes/No, 1000 cases, 50000 repeats

https://teaching.sociology.ul.ie/apps/binsim

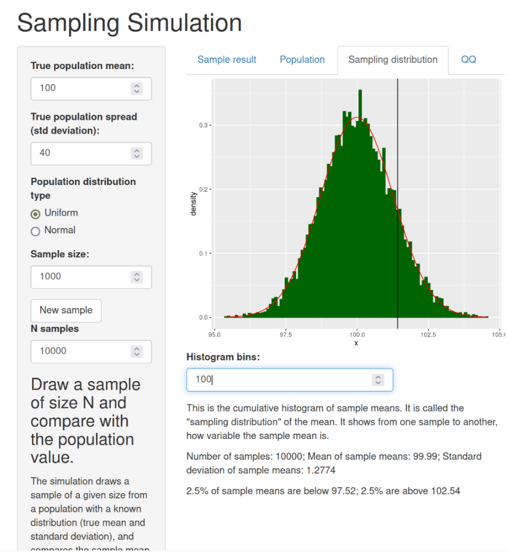


Simulation: Uniform population distribution

- Population: mean 100, Std Dev 40
- Sampling distribution
  - mean 99.99
  - Std Dev = Standard Error 1.2774
- Theoretical SE:

$$\frac{40}{\sqrt{1000}} = 1.2650$$

https://teaching.sociology.ul.ie/apps/so4046/sampling



sociology
UNIVERSITY OF LIMERICK

Changing sample size and population SD

- With the same simulation, we can see what happens to sampling variability
  - for different sample sizes
  - for different population standard deviations
  - for uniform and normal population distributions
- The previous simulation shows the same for binomial variables

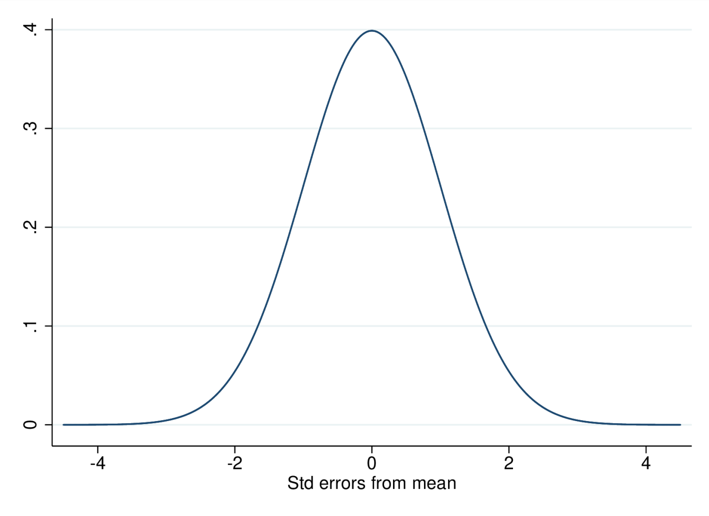
Confidence intervals

- We can use the CLT to add to our point estimate a measure of its precision: the *Confidence Interval*
- Agresti: “A confidence interval for a parameter is a range of numbers within which the parameter is believed to fall”
- “The probability that the confidence interval contains the parameter is called the confidence coefficient” which is a number close to 1, like 95% or 99%

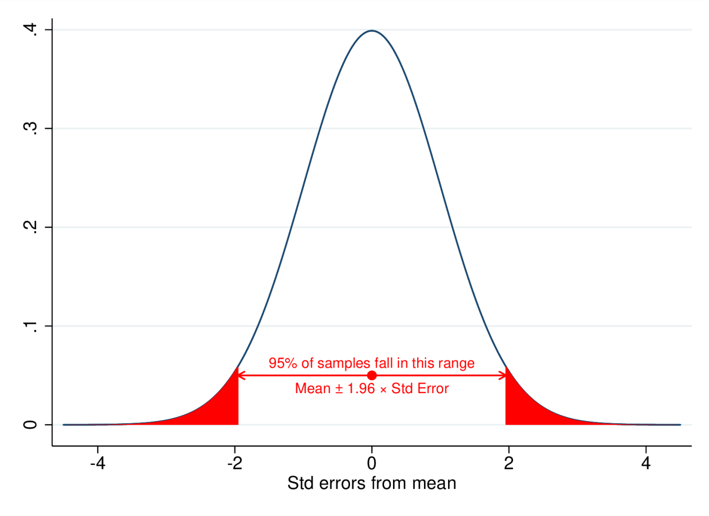
Estimating intervals

- A confidence interval is a band around our point estimate within which we can claim that there is a, for instance, 95% chance that the true population value lies – how do we calculate this?
- We work from the sampling distribution – if we are estimating a mean value, we know that 95% of estimates will fall within $\pm$ 1.96 standard errors of the mean
- The 1.96 comes from the normal distribution: 95% of the distribution is between 1.96 standard deviations above and below the mean

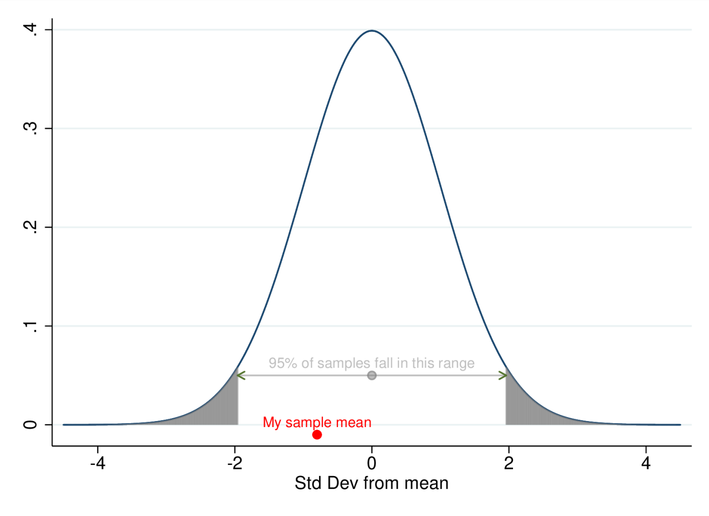
Central limit theorem: sample statistics normal



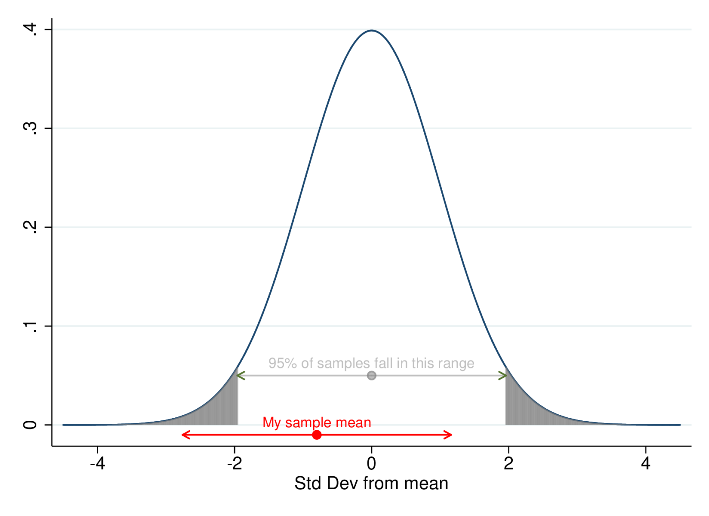
95% within ± 1.96 SE of mean



A sample: one of 95% within ± 1.96

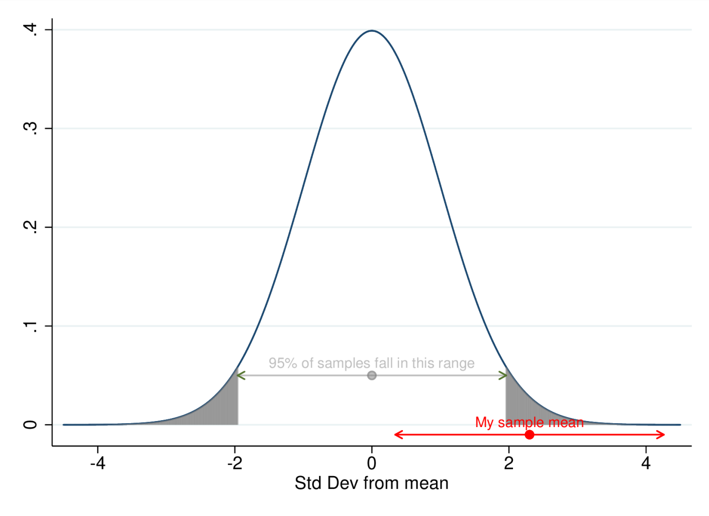


Sample mean ± 1.96 SEs contains true mean



sociology
UNIVERSITY OF LIMERICK

But 5% of sample means do not



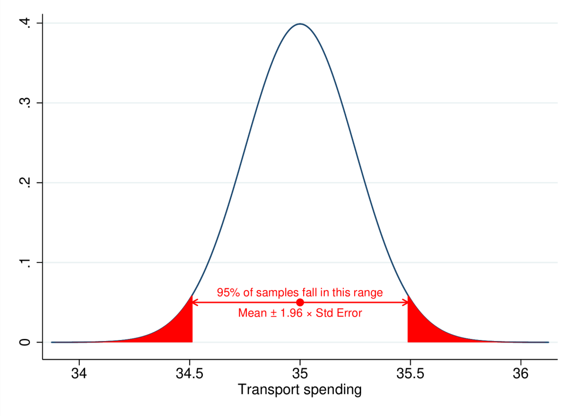
Example: transport spending

- Let's say spending on transport has a true mean of €35 per week, standard deviation €10. With a sample of 1600, 95% of all possible sample estimates will fall between:

$$35 \pm 1.96 \times \frac{10}{\sqrt{1600}}$$

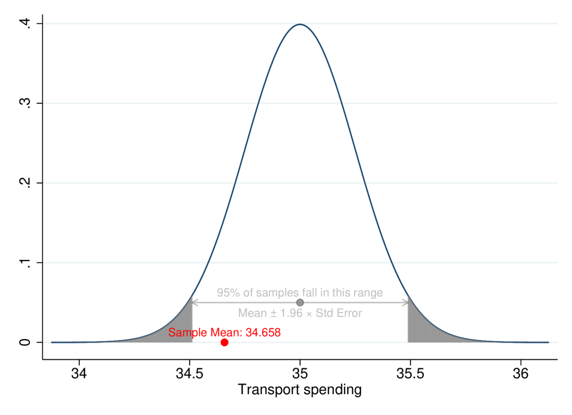
which is

$$\mu \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$



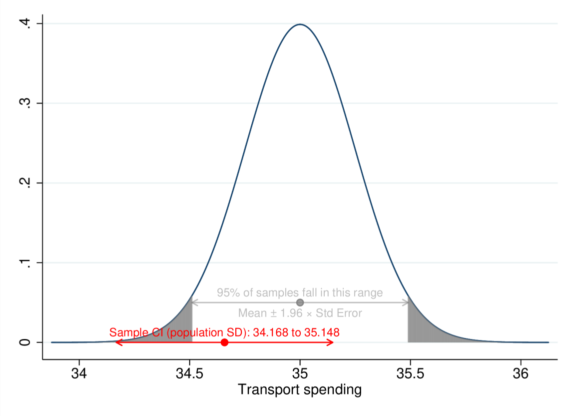
Sample results

- Let's say our sample gives a mean of €34.658, with a standard deviation of €10.123



Reverse the reasoning

- We don't know $\mu = 35$, only $\bar{X} = 34.658$
- We can reverse the reasoning and say that the true value has a 95% chance of falling in the range $\bar{X} \pm 1.96 \times \sigma_{\bar{X}}$



sociology
UNIVERSITY OF LIMERICK

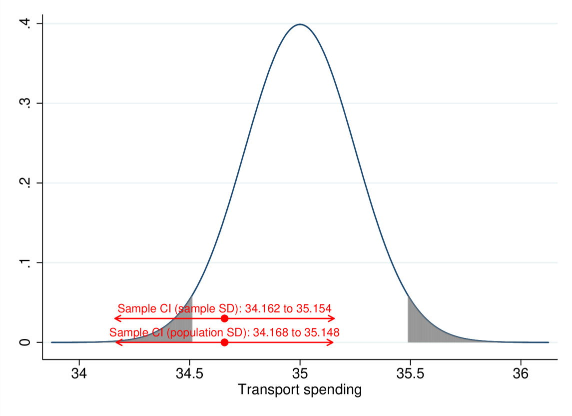
Sample SD

- But we don't know the true standard error
- We can use the sample estimate instead:

$$34.658 \pm 1.96 \times \frac{10.123}{\sqrt{1600}}$$

- In this case, a very slightly wider interval

$$\bar{X} \pm z_{0.95} \times \hat{\sigma}_{\bar{X}}$$



sociology
UNIVERSITY OF LIMERICK

Calculating the interval

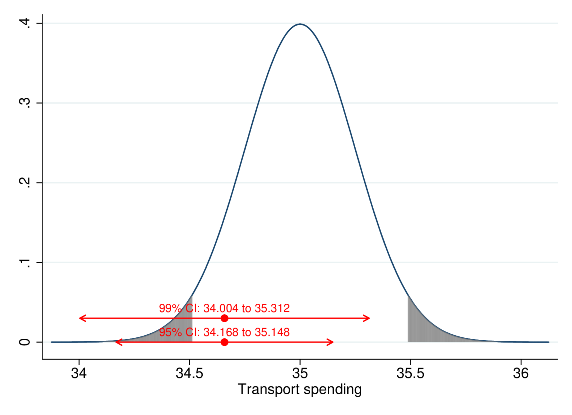
- Thus with sample mean is €34.658 and sample standard deviation €10.123 we calculate:
  - Standard Error is $\frac{10.123}{\sqrt{1600}} = \frac{10.123}{40} = 0.2531$
  - The lower bound is $34.658 - 1.96 \times 0.2531 = 34.162$
  - The upper bound is $34.658 + 1.96 \times 0.2531 = 35.154$
- We can interpret this as saying we are 95% confident that the true value is in this range
- This is because with 95% of all possible samples, such a confidence interval will include the true value

Levels of confidence

- The 95% confidence level is often used but sometimes the chance of being wrong one time in twenty is not acceptable
- We can use the normal distribution to choose other levels of confidence: for instance, 99% of the normal distribution lies within ± 2.575 standard deviations
- Check for yourself on `https://teaching.sociology.ul.ie/apps/snd`

99% Confidence

- 99% CI:
  - Lower: $34.658 - 2.575 \times 0.2531 = 34.006$
  - Upper: $34.658 + 2.575 \times 0.2531 = 35.310$
- To be more sure, we are necessarily less precise



SO5041 Unit 10

CIs for proportions

Standard deviation/error for proportions

- The example above is for sample means: the CI for sample proportions is similar
- The sampling distribution for a proportion (percent unemployed, percent voting Republican, percent satisfied etc.) is also normal for large samples
- The standard deviation of a 0/1 or yes/no variable depends on the proportions in the population however:

$$\sigma_\pi = \sqrt{\pi \times (1 - \pi)}$$

- We can estimate the standard error then as

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

See https://teaching.sociology.ul.ie/apps/binsim

CI for proportions

- The 95% confidence interval for a proportion uses this, the same way as for the mean:

$$\hat{\pi} \pm z_{0.95} \times \hat{\sigma}_{\hat{\pi}}$$

- If we have a sample 1000 and find 45% of voters expressing an intention to vote yes we calculate as follows:

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{0.45(1-0.45)}{1000}} = 0.0157$$

and thus the CI is

$$0.45 \pm 1.96 \times 0.0157$$

- That is $\hat{\pi} \pm 0.031$, i.e., plus or minus 3%

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union

      union |
     worker |      Freq.     Percent        Cum.
------------+-----------------------------------
   nonunion |      1,417       75.45       75.45
      union |        461       24.55      100.00
------------+-----------------------------------
      Total |      1,878      100.00
```

- Proportion = 461/1878 = 0.2455

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union
```

union worker	Freq.	Percent	Cum.
nonunion	1,417	75.45	75.45
union	461	24.55	100.00
Total	1,878	100.00	

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union
```

union worker	Freq.	Percent	Cum.
nonunion	1,417	75.45	75.45
union	461	24.55	100.00
Total	1,878	100.00	

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$
- Std Err = $\frac{0.4304}{\sqrt{1878}} = 0.0099$

sociology
UNIVERSITY OF LIMERICK

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union
```

union worker	Freq.	Percent	Cum.
nonunion	1,417	75.45	75.45
union	461	24.55	100.00
Total	1,878	100.00	

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$
- Std Err = $\frac{0.4304}{\sqrt{1878}} = 0.0099$
- Half width = 1.96*0.0099 = 0.0195

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union
```

union worker	Freq.	Percent	Cum.
nonunion	1,417	75.45	75.45
union	461	24.55	100.00
Total	1,878	100.00	

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$
- Std Err = $\frac{0.4304}{\sqrt{1878}} = 0.0099$
- Half width = 1.96*0.0099 = 0.0195
- Low: 0.2455 - 0.0195 = 0.2260
- High: 0.2455 + 0.0195 =0.2650

sociology
UNIVERSITY OF LIMERICK

Summary

- CLT: Sampling distribution of a sample statistic is normally distributed
  - Centred on true value, standard deviation is SE = $\frac{\sigma}{\sqrt{n}}$
- 95% of samples fall in range $\mu \pm SE \times 1.96$
- 95% Confidence interval: $\bar{X} \pm \hat{SE} \times 1.96$
  - "95% confident the true value lies in this interval"
- 99% Confidence interval: $\bar{X} \pm \hat{SE} \times 2.575$
- Similar process from proportions where SE is $\sigma_\pi = \sqrt{\pi \times (1 - \pi)}$

$$\hat{\pi} \pm z_{0.95} \times \hat{SE}_\pi$$

- A way of presenting a sample statistic with a measure of its precision

sociology
UNIVERSITY OF LIMERICK