SO5041 Unit 10:

Brendan Halpin, Sociology

Autumn 2019/0

Point estimates

- Agresti/Finlay: 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
- Estimators such as the sample mean, standard deviation, or sample proportion are reasonably efficient and unbiased
  - 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:

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

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)
- Where the sample is sufficiently large, the Central Limit Theorem tells us that the sampling distribution is normal, with mean μ and standard deviation $\sigma_{\bar{X}}$
- We can use this information to add to our point estimate a measure of its precision: the *Confidence Interval*

Confidence intervals

- Agresti/Finlay: "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%
- 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

Example: transport spending

- Assume we have a national sample, and are estimating spending on transport – if the true value is €35 per month, with a standard deviation of €5, and the sample size is 1600, then 95% of all possible sample estimates will fall between

$$35 - 1.96 \times \frac{5}{\sqrt{1600}}$$

and

$$35 + 1.96 \times \frac{5}{\sqrt{1600}}$$

which is

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

- The 1.96 comes from the normal distribution: 95% of the distribution is between 1.96 standard deviations above and below the mean

Reverse the reasoning

- We don't know μ, but we can reverse the reasoning and say that the true value has a 95% chance of falling in the range $\bar{X} \pm 1.96 \times \sigma_{\bar{X}}$
- We don't know \$ $\sigma_{\bar{X}}$ \$, the standard error either:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- However, we do know the standard deviation of the sample, s, and this is an unbiased estimator of the population standard deviation, σ, so we can estimate the standard error:

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Calculating the interval

- Thus in our example, if our sample mean is €34.75 and our sample standard deviation is €7.7 we can calculate:
  - The standard error is $\frac{7.7}{\sqrt{1600}} = \frac{7.7}{40} = 0.1925$
  - The lower bound is $34.75 - 1.96 \times 0.1925 = 34.37$
  - The upper bound is $34.75 + 1.96 \times 0.1925 = 35.13$
- Thus our point estimate is 34.75 and our 95% confidence interval runs from 34.37 to 35.13
- We can interpret this as saying there is a 95% chance 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 ± 2.58 standard deviations
- Again in our example (sample mean is 34.75%, standard deviation 7.7)
  - The lower bound is $34.75 - 2.58 \times 0.1925 = 34.25$
  - The upper bound is $34.75 + 2.58 \times 0.1925 = 35.25$
- To be more sure, we are necessarily less precise

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

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

CI for proportions

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