Statistical Literacy — MINT

Lecture 7: A first step in inference - the confidence intervals for proportions

Rémi Viné

The Graduate Institute | Geneva

November 6th, 2023

Outline

Recap' - Main concepts covered last week

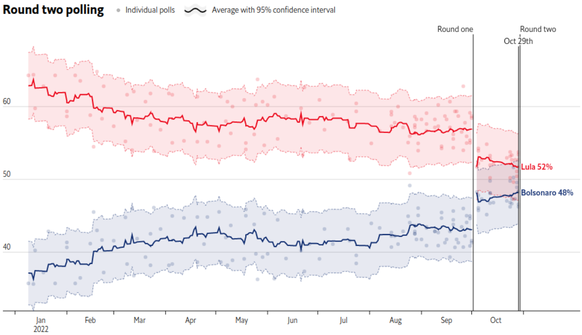
- Assume high speed train top speed is 250 km/h with a standard deviation of 32 km/h.
- What is the probability to be on a train whose top speed is between 258 km/h and 282 km/h?
  - $P(258 < X < 282) = P(\frac{258-250}{32} < X < \frac{282-250}{32}) = P(0.25 < Z < 1) = P(Z > 0.25) - P(Z > 1) = 0.4013 - 0.1587 = 0.2426$
- Now, constructing a sample - from this Normally distributed population - of size 16 different trains, what is the probability that the sample average has an average top speed beyond 282 km/h?
  - $P(\bar{X} > 282) = P(Z > \frac{282-250}{\frac{32}{\sqrt{16}}}) = P(Z > 4) = 0.000317$

Housekeeping

- Problem set 7 is now available
- Last tutorial will be about correcting PS11
  - $\rightarrow$ Very important because the tools covered will be the key aspects to address if you run regressions including several variables (which is most likely)
  - $\rightarrow$ The mock exam is revision material, solutions will be given so that you'll know what you did right

What is inference and why does it matter?

- From a sample, predict indicators of the whole population
- For example, sample randomly a few thousand individuals among Brazilian voters and make predictions for population
    - It would be cumbersome, time consuming and very costly (if possible at all) to ask all the 140+ million voters



Source: The Economist

What do we start with?

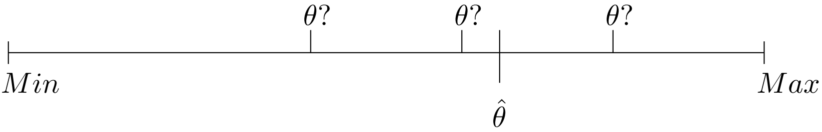
- From the data, we know
  - The **sample size**, n
  - The **point estimate**, $\hat{\theta}$

Inference - general recipe

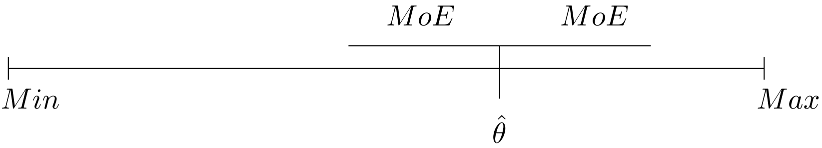
- Statistical inference applies inductive reasoning but uses probabilities to proceed
- From the sample, extract an indicator (a statistic) to use as the **point estimate** $\rightarrow \hat{\theta}$
  - It is important that the point estimate ($\hat{\theta}$) is not **biased** and is **consistent**; (no need to remember the formulas):
    - *Unbiased estimator*: the expected value of the point estimate is equal to the true value ($E[\hat{\theta}] = \theta$, with θ the population parameter of interest)
    - *Consistent estimator*: as the sample gets (very) large, the point estimate converges to the true value ($plim_{n \rightarrow +\infty} \hat{\theta} = \theta$)
- Because the point estimate will vary across random samples, take into account this variability to approximate the indicator in the population

Generalities on confidence intervals

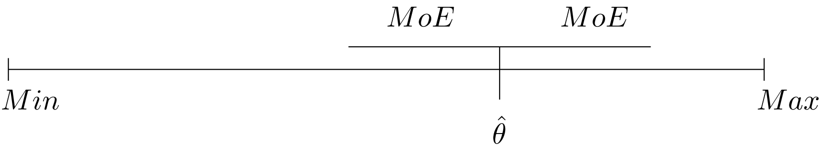
- The true parameter is unknown



- Hence, one builds an interval, composed of **Margin of Error** (MoE) on each side, in which the true parameter is likely to be
  - The likeliness can be measured
    - This likeliness is the **confidence level**



Generalities on confidence intervals



- A confidence interval is therefore $PointEstimate \pm MoE$
- The $Margin\ of\ Error$ is constructed using two elements
  1. The **standard error** (standard deviation of the sampling distribution, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)
  2. The **critical value** (cv) taken from a distribution

$$\Rightarrow MoE = \sigma_{\bar{x}} \times cv$$

Confidence intervals in this class

1. Confidence intervals for proportions
   - The critical value is always taken from a Normal distribution
   - The standard error is even easier to compute (need only of the sample size and the proportion)
2. Confidence intervals for means
   - The critical value is taken from a Normal distribution **only** if the population variance (σ) is available; otherwise, need to use another distribution (*Student distribution - to be seen in subsequent lectures*)

Confidence intervals: step by step

1. Verify the conditions of validity (*i.e.*, can the sampling distribution be considered as Normally distributed?)
2. Select a level of confidence
3. Find the point estimate
4. Compute the standard error
5. Find the critical value from the appropriate distribution
6. Compute the Margin of Error and build the confidence interval
7. Conclude

Confidence intervals for proportions

- A proportion is a share, between 0 and 1
  - It is about binary variables: vote for candidate A or B, being vegetarian or not, being a MINT student in IHEID or not, *etc.*
  - p is the proportion of observations (in population) with the characteristic of interest, q is its complement (in population)

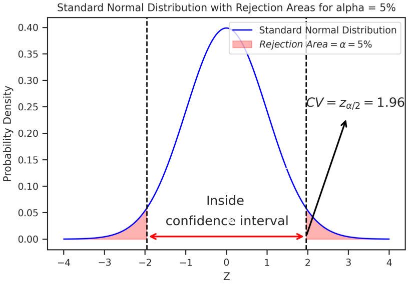
1. **Condition of validity**
   - $npq >> 1$ (much larger) or $np > 5$ together with $nq > 5$
2. **Choose the level of confidence**
   - For example, a level of confidence of 95% implies that we are 95% confident that the population parameter is in the confidence interval
     - The reminder, here 5%, is called the **level of significance** and denoted α

Confidence intervals for proportions

3. **Find the point estimate**
   - $\hat{p} = \frac{Number\ of\ people\ with\ characteristic(s)\ of\ interest\ (sample)}{All\ people\ in\ the\ sample}$
4. **Compute the standard error**: $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

5. **Find the critical value**

   - In confidence intervals for proportions, we always use the Normal distribution
   - The critical value is the *z-score*
     - $z_{\alpha/2}$ is the *z-score* corresponding to a given probability to be outside the interval (on either side!), called the **rejection area**



Confidence intervals for proportions

MoE and Interpretation

- $MoE = z_{\alpha/2} \times s_{\hat{p}}$

6. **Confidence interval** (different equivalent notations):
   - $\hat{p} \pm MoE$
   - $\hat{p} - MoE < p < \hat{p} + MoE$
   - $[\hat{p} - MoE; \hat{p} + MoE]$

7. **Interpret**
   - We can be $(1 - \alpha)\%$ confident that the proportion in the population is located in the interval $[\hat{p} - MoE; \hat{p} + MoE]$
   - Constructing 100 intervals, the population proportion would belong to the constructed interval $[\hat{p} - MoE; \hat{p} + MoE]$ about $(1 - \alpha)$ times

A simple example

Brexit : compare polls with referendum results

- Commentators were adamant before the referendum
- Among the last survey, YouGov made a poll on the 23rd of June (actual day of the referendum)
  - Results of the survey (weighted): 2307 for "Remain", 2162 for "Leave"

    → *Statistics* from the sample
  - Results of the referendum: 48.1% for "Remain"

    → *Parameter* from the population
- From the poll let's see whether the actual proportion would have been part of a 95% confidence interval

A simple example

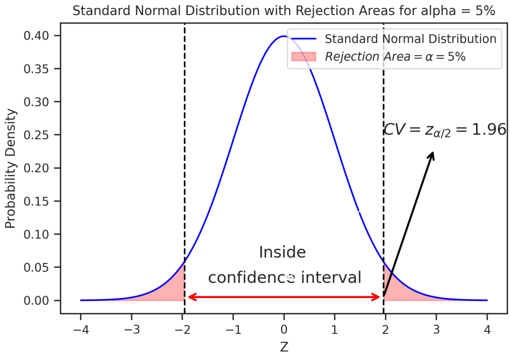
Proceed step by step

- Can we do inference based on proportions being Normally distributed?
  - Yes, as $npq >> 1$, at least $n\hat{p}\hat{q} >> 1$
  - Sample size: $n = 2307 + 2162 = 4469$
  - "Remain" proportion: $\hat{p} = \frac{2307}{4469} = 0.5162$
    - $\rightarrow$ This is the point estimate, $\hat{p}$
  - "Leave" proportion: $\hat{q} = 1 - 0.5162 = 0.4838$
  - Hence, $n\hat{p}\hat{q} = 4469 * 0.5162 * 0.4838 = 1116.0772$
- So the sampling distribution of the proportion is (approximately) Normally distributed

A simple example

Find the critical value

- A 95% confidence interval implies that 5% of the area of the Normal distribution belongs to the rejection area



- $z_{0.025} = 1.96$ (from a standard Normal distribution table, or a software's command)

A simple example

Compute the standard error

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.5162 * 0.4838}{4469}}$$

$$s_{\hat{p}} \approx \sqrt{\frac{0.2497}{4469}} \approx 0.0075$$

- The standard deviation of the sampling distribution (*the standard error*) of proportions is 0.0075.
  - One expects polls to differ from each other - about the average proportion - at about this amount

A simple example

Put things together

- The margin of error is $z_{0.025} \times s_{0.5162} \approx 1.96 \times 0.0075 \approx 0.0147$
- The 95% CI in the example is

$$\hat{p} \pm MoE \Rightarrow 0.5162 \pm 0.0147$$

$$\Rightarrow (0.5015, 5309)$$

A simple example

Correct interpretations

- It means that if we repeatedly took samples of size 4469 from the same population, and constructed a 95% confidence interval around each sample proportion, the actual population proportion is expected to be found in 95% of them.
- We can be 95% confident that the population proportion of British voters in favor of "Remain" is located between 50.15% and 53.09%.

A simple example

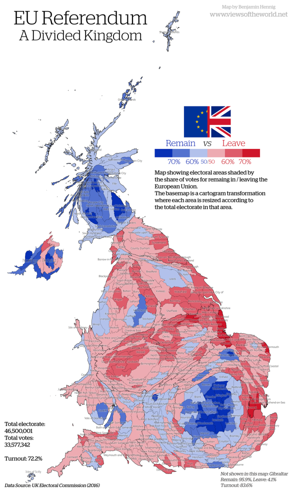
Extensions

- The referendum gave a result ($p = 0.48$) that was not part of the 95% confidence interval
- Not only the confidence interval failed to capture the real population proportion, but it also led to misinterpretations of the results!
  - A 99% confidence interval would have led to: (0.4969,0.5355)
    - $\rightarrow$ Can you compute this?
- Be cautious in the way of interpreting confidence intervals, and remember that the population parameter might not be captured
  - Can you raise a few concerns in the case of Brexit?

A simple example: what went wrong?

Some plausible avenues

- Sampling variability (OK), sampling design (not OK)
- Turnout: by age group, by education group
- Geographical divide $\Rightarrow$



Cartogram - Benjamin Hennig

Confidence intervals for proportions

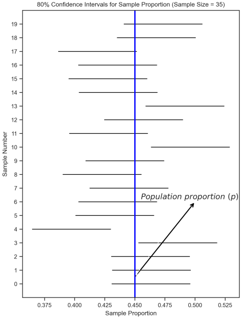
From MoE to the required sample size

- What if, working for a poll company, you cannot exceed a Margin of Error of 1%? $\rightarrow$ If MoE big, is a poll relevant?...
  - What is the required sample size?
  - $MoE = z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Longleftrightarrow n = (\frac{z_{\alpha/2}}{MoE})^2 \hat{p}(1-\hat{p})$
- *Example:* you assume the second round of an election has a proportion for candidate A to be 48.5%. For a 95% confidence level, and a Margin of Error set to be 1%, how many people should be surveyed?
  - $n = (\frac{1.96}{0.01})^2 0.485 \times 0.515 \approx 9595.4 \rightarrow n = 9596$
  - Note that if you do not assume proportions *a priori*, you should use $\hat{p} = 0.5$ (most conservative choice that maximizes the variance)

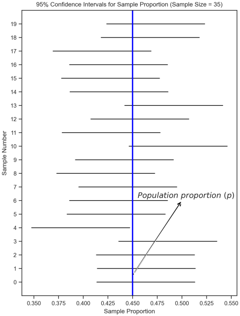
Confidence intervals for proportions

Link between α and MoE

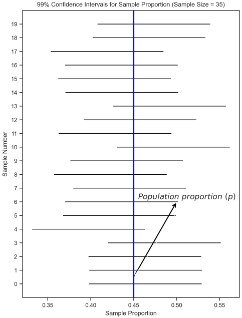
(a) 80% CI



(b) 95% CI



(c) 99%

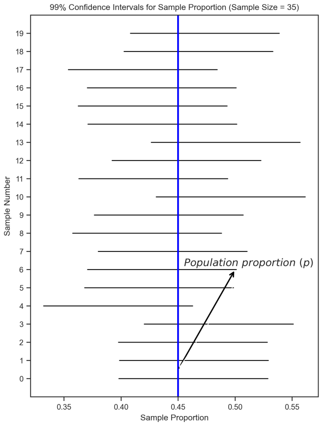


- $MoE \uparrow \Leftrightarrow \; \alpha \downarrow$

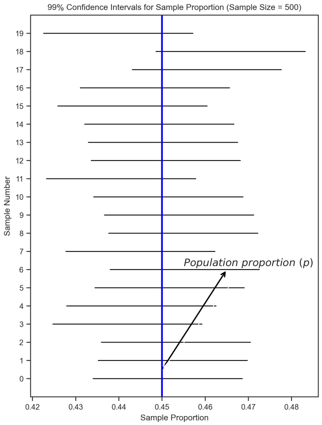
Confidence intervals for proportions

Link between n and MoE

(a) 99% CI, $n = 35$



(b) 99% CI, $n = 500$



- $MoE \downarrow \Leftrightarrow \; n \uparrow$ *(look at the x-axis)*

Next session

- Next session is on another type of confidence interval, focusing on means. This will lead us to introduce another distribution (the Student distribution).