INFO370 Lab: Compare proportions

January 30, 2022

Instructions

This lab asks you to compute confidence intervals (CI) for an election poll. Imagine a country that is heading to polls sometime in the hear future, for instance on April 3rd. Let's call this country "Hungary", and assume there are two main political parties, authoritarian *Fidesz* (F), and a more democratic conglomerate *Opposition electoral alliance* (O). Let's call the leader of F by Viktor Orbán and the leader of O as Péter Márki-Zay. Imagine they look like this:

(a) Viktor Orbán, European People's Party, CC BY 2.0, via Wikimedia Commons

(b) Peter Marki-Zay, Photo by Draskovics Ádám, CC BY 4.0, via Wikimedia Commons

Further imagine a polling firm (call it *IDEA Intézet*) conducted a poll between Januay 4th and January 14th where they sampled 1840 respondents who would vote for either F or O (they also sampled respondents who will not vote for either party but we can ignore those here). Out of these 1840, 960 (i.e. 52.174%) intend to vote for F and 880 (i.e. 47.826%) would vote for O. Let's look away from all the questions about representativeness, non-response, don't knows and such, and just ask "how precise are these numbers"? Can we say confidently that the authoritarian F is leading in the polls or does the opposition a chance to break the creeping authoritarianism?

We suggest you read Openintro Statistics 5.2 "Confidence intervals for a sample proportion".

We might use some data here but the data would be a little bit stupid: it would contain 960 "F"-s and 880 "O"-s. So we haven't uploaded any data but you can easily create it yourself if you want. Instead, we denote "O" by "1" and "F" by "0" and note that average response, percentage of voters who prefer O, is $\bar{X} = 0.47826$.

You have two tasks that closely mirror each other. In the first one, you create data according your H_0 and see if the actual result is close enough. In the second one you create data according to the actual result and see if H_0 is close enough.

The first approach corresponds closely to the idea of the "hypothesis world" (H_0 world). The second approach corresponds to what is normally done in linear regression.

1 Test data under H_0

First, let's simulate the polls according to H_0 and test if the actual number falls into the CI.

1. What is your H_0–the claim about the world you want to test? Explain!

   Hint: what value you must reject to be confident in this claim?

2. Choose the number of polls (repetitions R) you conduct. 1000 is a good choice. More is better but slower.

   Note: this is not the same thing as the number of respondents N! What is your number of respondents N?

3. Now conduct R polls of N respondents. Each respondent should be represented as 0 or 1, with 1 occuring with the probability 0.5 (your H_0).

   Hint: there are several ways to do it. You can loop R times over single polling. You can make an $\mathsf{R} \times \mathsf{N}$ matrix of Bernoullis. You can also make an length-R vector of binomials. Do whatever you like, most important that you understand what you are doing.

   Suggestion: as random numbers are, well, random, your results and conclusions my change from run to run. Use `np.random.seed` to fix the random sequence.

4. For each poll, compute the average value (the proportion of respondents who prefer O). Now you should have R averages.

5. Find the 95% CI of your means by computing 2.5-th and 97.5-th percentiles.

6. Does your actual poll result (0.47826) fall into the CI?

7. Based on the CI you computed answer the question: can you confidently (at 5% confidence level) say that F is ahead and O has no chance?

---

2 Test H_0 under data

Now we repeat the same procedure but swap around how we generate data and what we test.

1. Now conduct R polls of N respondents. Each respondent should be represented as 0 or 1, with 1 occuring with the probability 0.47826 (the probability in data).

2. For each poll, compute the average value (the proportion of respondents who prefer G). Now you should have R averages.

3. Find the 95% CI of your means by computing 2.5-th and 97.5-th percentiles.

4. Is 0.5—your H_0 inside of the CI?

5. Based on the CI you computed answer the question: can you confidently (at 5% confidence level) say that Greens are leading in the polls?