INFO370 Problem Set: CLT and Hypothesis testing

January 29, 2023

Instructions

This problem set revolves around Central Limit Theorem, statistical hypotheses and confidence intervals.

Note that there is also an extra credit task, that one is quite different.

- Please write clearly! Answer questions in a way, that if the code chunks are removed from your document, the result is still readable!
- A recommendation–this problem set contains a number of repetitive tasks. Consider writing a function that does most of the job, and then just feed different data to this function, and comment what you see there.

Good luck!

1 Explore Central Limit Theorem (60pt)

In this section you will see how does Central Limit Theorem (CLT) work. CLT states two things:

a) The means of a sample of random numbers tend to be normally distributed if the sample gets large.

b) Variance of the mean tends to be $\frac{1}{S}\operatorname{Var} X$ where S is the sample size and X is the random variable we are analyzing.

(This is actually a property of expectation and independence, not really CLT. But CLT is closely related to this result.)

CLT, and how variance and mean value change when sample size increases, plays a very important role in computing confidence intervals later.

The task is structured in a way that you may want to create a function that takes in sample size S and outputs all needed results, including the histogram. There will be quite a bit of repetitive coding otherwise.

We start with a distribution that does not look at all normal. We create a RV

$$X = \begin{cases} -1 & \text{with probability } 0.5 \\ 1 & \text{with probability } 0.5. \end{cases}$$

(You can imagine we flip a fair coin and label heads as 1 and tails as -1.) One way to sample from such RV is something like this

```
import numpy as np
np.random.randint(0,2, size=10)*2 - 1

## array([-1, -1,  1,  1, -1, -1, -1,  1,  1,  1])
```

Detailed tasks:

1. (5pt) What is Random Variable (RV)? What makes X a RV?

   Hint: check out Lecture Notes 1.3.4 *Random Variable*.

2. (6pt) Calculate the expected value and variance of this random variable. Explain what is the difference between expected value and the sample mean.

   Note: these are theoretical values and not related to any samples. If you use functions like `mean` or `var` here then you have misunderstood the concepts!

   Hint: read lecture notes 1.3.4 (Expected Value and Variance), and Openintro Statistics 3.4 (Random variables), in particular 3.4.2 (Variability). I recommend to use the shortcut formula $\operatorname{Var} \mathsf{X} = \mathbb{E}\, \mathsf{X}^2 - (\mathbb{E}\, \mathsf{X})^2$.

3. (1pt) Choose your number of repetitions R. 1000 is a good number but you can also take 10,000 or 100,000 to get smoother histograms.

   Note: number of repetitions R *is not* the same as sample size S here. You will create samples of size S for R times below. For instance, you will create $\mathsf{R} = 1000$ times a sample of size $\mathsf{S} = 5$. Please understand the difference, it is a fequent source of confusion!

4. (5pt) Create a vector of R random realizations of X. Make a histogram of those. Comment the shape of the histogram.

   Note: in this case we have $\mathsf{R} = 1000$ repetitions and samples are of size $\mathsf{S} = 1$ as we look at individual realizations.

   Hint: it takes some tweaking to get nice histograms of discrete distributions. The simplest way is just to make many bars (most of which will be 0) by adding argument `bins=100` to `plt.hist`.

5. (3pt) Compute and report mean and variance of the sample you created (just use `np.mean` and `np.var`). NB! Here we talk about *sample mean* and *sample variance*. Compare these numbers with the theoretical values computed in question 2 above.

   Hint: they should be fairly close.

6. (6pt) Now create R *pairs* of random realizations of X (i.e. sample size $\mathsf{S} = 2$). For each pair, compute its mean. You should have R mean values. Make the histogram. How does this look like?

7. (6pt) Compute and report mean of the R pair means, and variance of the means. NB! we talk about *sample mean* and *sample variance* again, where sample is your sample of R pair means.

8. (5pt) Compute the expected value and variance of the pair means, i.e. the theoretical concepts. This mirrors what you did in 2.

   Compare the theoretical values with the sample values above. Are those fairly similar?

   Note that according to CLT, the variance of a pair mean should be just 1/2 of what you got above as for pairs $S = 2$.

9. (4pt) Now instead of pairs of random numbers, repeat this with 5-tuples of random numbers (i.e. $S = 5$ random numbers per one repetition, and still the same $R = 1000$ or whatever you chose repetitions in total). Compare the theoretical and sample version of mean and variance of 5-tuples. Are they similar? Do you spot any noticeable differences in the histogram compared to your previous histogram?

10. (3pt) Repeat with 25-tuples...

    (Also compute the expectation and theoretical variance, and compare those with sample mean, sample variance)

11. (3pt) ... and with 1000-tuples. Do not forget to compare with theoretical results.

12. (2pt) Comment on the tuple size, and how the shape of the histogram changes when the tuple size increases.

13. (7pt) Explain why do the histograms resemble normal distribution as S grows.

    In particular, explain what happens when we move from single values $S = 1$ to pairs $S = 2$. Why did two equal peaks turn into a "ılı"-shaped histogram?

14. (4pt) Explain what is the difference between R and S. How do changing these values affect the histograms?

---

2 Is povery in Azraq refugee camp falling? (60pt)

A lot of refugees from Syrian civil war are housed in camps in Jordan. A report https://reliefweb.int/attachments/32604cb5-1e86-49eb-8497-8f43a4721135/WFP-0000142211.pdf (also available on canvas/readings/WFP-0000142211.pdf) shows results of a survey of living conditions in Zaatri and Azraq camps.

Page 5 of the report discusses poverty in the refugee camps, in particular "abject poverty", inability to afford the "survival minimum expenditure basket" (SMEB) of food and basic hygiene. The survey finds that while the abject poverty has stayed fairly similar in Zaatari camp, it has fallen substantially (from 66% to 51%) in Azraq camp.

But is it actually improving? Maybe the survey wave of 2022-Q2 just got "more lucky" by randomly sampling fewer poor households? Or the way around–2022-Q1 survey was just "unlucky"? Let's find it out. We are going to compute the confidence interval for the 2022-Q2 survey, and check if the 2022-Q1 result falls into this confidence interval.

2.1 Background (2pt)

1. (1pt) What was the abject poverty in Azraq camp in Q1 and Q2 2022 (when including all assistance)? Lets call these variables p_1 and p_2.
2. (1pt) How many households were surveyed in the camp? Call this sample size S.

2.2 Simulations (30pt)

Now let's create a large number of such random surveys, and see how much variability do we see in the results. Note though that we create these assuming they were conducted using independent uniform sampling, if the actual surveys were done differently, then our analysis below may be incorrect.

Let's describe all households in the Azraq camp as a RV X that can be either "1" if the household is poor (abject poor) or "0" is it is not poor. This is a Bernoulli RV

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases} \tag{1}$$

You can create a sample from this RV along the lines

```
S = 8
p = 0.6
np.random.binomial(1, p, size=S)

## array([0, 0, 1, 1, 1, 1, 0, 1])
```

1. (3pt) Create a random sample using the correct values of S and p_2 you found in 2.1 above.
2. (2pt) Compute the sample mean and compare it with p_1 and p_2 above. How close is it to these figures?
3. (1pt) Pick your number of replications R (something like 1000 or 10,000 are good numbers).
4. (6pt) Repeat the points 1 and 2 for R times: create the sample, compute the average, but also store the average in an array. You should have R averages now.
5. (6pt) What is the average of the averages? Which probability from 1 does it resemble? Why?
6. (5pt) Plot a histogram of the averages. Which distribution does it resemble? What do you say, by just eyeballing the plot, what are the largest and smallest values that are "reasonably" common?
7. (6pt) Finally, compute 2.5th and 97.5th percentile and the 95% confidence intervals. Does the Q1 poverty value fall into this interval?

2.3 Theoretical CI (28pt)

But simulations are often hard to do. Fortunately, we can compute the confidence intervals by theoretical considerations.

1. (4pt) Compute variance of your sample of X. You can use the Bernoulli variance formula $\operatorname{Var} X = p(1-p)$. You can also use the sample you created in 2.2.1, or create a new sample, and find the sample variance.

2. (8pt) But this was variance of X (or *sample variance* if that was what you computed). What we need is variance of sample mean. What does CLT tell about relationship b/w sample variance and variance of the sample mean?

3. (7pt) Compute the standard deviation of sample mean using CLT.

4. (3pt) Compare the standard deviation you got here with the standard deviation of the sample of averages you computed in 2.2.4.

5. (6pt) Use this standard deviation to compute the confidence interval. Compare it to what you got in 2.2.7. Does p_1 falls inside or outside of this interval?

   You may want to check Lecture Notes 1.5.2 "Doing Statistical Inference".

   Hint: they should be fairly similar, and your conclusion regarding p_1 should be the same.

Finally tell us how many hours did you spend on this PS.

2.4 Challenge (extra credit, 1 EC point)

How long time do you need to simulate the Q2 poverty in Azraq camp in order to get an outlier that is as big as the Q1 poweverty value? If you did the Question 2 correctly, then you noticed that the simulated Q2 poverty values are all much smaller than what was reported in Q1. But if you simulate long enough, then you can get an arbitrarily large value. But how long do you need to run the simulations? Will a few minutes suffice? Or do you need to keep computer running for year, or even millenia?

1. First, time your simulations. Run Q2.2.4 many times (do not store the values this time as we may run out of memory) and measure how many seconds does it take. Your computer should run a least five seconds before proceeding (this will help with accuracy). Based on that figure, calculate how long it would take to run 10^{12} or however many simulations you need.

   Hint: check out `%timeit` and `%time` magic macros, alternatively you may check out the "time" module and "time.time()" method.

2. Next, we have to find the t-value. Let's do it in a little wrong way by assuming the Q1 value is 0.66 without any errors. You computed the standard deviation of the mean in Q2.2.. What is `t` value for this difference?

3. Second, what is the probability to receive such t-values? You may need to calculate your t-values yourself, they will not be on any tables. Assume we are dealing with normal distribution. (Not quite but we are close.) You have to compute the probability you get a value larger than the `t` value you just computed. This can be done along the lines:

   ```
   from scipy import stats
   norm = stats.norm()
   norm.cdf(-1.96)  # close to 0.025

   ## 0.024997895148220435
   ```

   Except that you replace 1.96 with your actual t-value.

   Explain: why does the example use `norm.cdf(-1.96)` instead of `norm.cdf(1.96)`?

4. How many iterations do you need? Let's do a shortcut—if probability `p` is small, you need roughly `3/p` iterations. So if $p = 0.001$, you need 3000 iterations.

5. Based on the timings you did above, how many years do you have to run the simulations? Could you get it done by the assignment deadline? Or if one had started the computer the year your grandfather was born, would it be there now? If the first Seattle inhabitants had started it when they moved here following the melting ice, 10,000 or so years ago?