Problem set 2

Ec626|USF
Spring 2020|Kyle Carlson

Intro: Simulating uncertainty

1. I know this problem set looks long, but it is primarily repeating simulations with various parameters and interpretation of results.

2. The problem set builds on this notebook: "Random variables and Python review".

3. Our goal is to develop a portfolio of simulation, analytical, and graphical patterns for understanding random sampling and hypothesis tests. These will be building blocks for the next problem set and future ones!

1 Notebook from class (50 points)

1.1 Simulating basic distributions

1. In the "Normal distribution" section we simulated from `np.random.normal` and then plotted a histogram of the data. Let's repeat the same exercise but try two other distributions: Bernoulli$(p = 0.25)$ and Exponential$(\text{scale} = 3)$. Confirm that the sample mean matches the theoretical mean.

2. In the notebook section "Sampling distribution of $\bar{X}_n$" we generated normally-distributed samples and plotted the sample means. Repeat this exercise using your Bernoulli and Exponential distributions. The code is fairly modular, so only minimal changes are needed. **Do not worry about plotting the pdf of $\bar{X}_n$.** Reference the slides following "Sampling distribution of the mean" for this question. Plot the underlying mean on the graph.

1.2 Simulating experiment statistics

In the section "Simulating other sample statistics for experiments" we simulated an experiment with two groups and estimated the ATE in each simulated sample.

1. In the histogram we plotted the distribution of these estimates. What is the standard deviation of these estimates (we also call this the standard error)? Use the `.std()` method.

2. The experiment in the notebook creates equal-sized treatment and control groups every time. Change the simulation so that the split is always 20% treatment, 80% control. (i) Plot the histogram again and calculate the standard deviation of the estimates. (ii) Plot histograms of the sample means of each group individually. (iii) How does the percentage split affect the "noise" in our estimate of the ATE? (iv) What seems to be the best way to make our treatment and control groups?

3. Returning to the original simulation in the notebook, what is the probability (across repeated samples) that we would estimate the ATE to be larger than 0.7?

4. What is the approximately probability that we would estimate the ATE to be more than two standard errors away from the true ATE (above or below)? Confirm with your simulation data.

2 Simulating t-tests under the null (30 points)

In the section "Simulating a two-sample t-test" we simulated a t-test with normally distributed data in the underlying sample. Please also review the new section "Comparing simulations with different parameters" at the end of the notebook before your work on this problem. Most of this problem will be small variations on that code.

For the final results of your simulations use $B = 5000$ to keep the simulation variability reasonably low. But when you are prototyping and working out bugs you should use a lower number (like 500 or 1000) to increase speed.

1. Define "false positive rate" in your own words (in the context of a statistical test).

2. Let's simulate experiments under the null hypothesis of no treatment effect. Adapt the code from "Comparing simulations with different parameters". Use $E[Y^0] = 3$ and leave the `sigma` (standard deviation, `scale` parameter in `np.random.normal`) of the underlying data at 1. Simulate with a total sample size of 1000 and use varying proportions of units in the treatment group: 10%, 25%, 50%, 25%, 90%. Consider (i) the distribution of 'diff', the estimated difference between the treatment and control group means and (ii) the distribution of t, and (iii) the rejection rate for the hypothesis test ($p < 0.05$). Which of these are related to the treatment-control proportions and what does the pattern seem to be? Show your results.

3. Repeat the previous question but this time vary the total sample size using 100, 1000, 10000. Use a 50/50 split between treatment and control.

4. Simulate with a 50/50 and split total sample size of 1000 (benchmark parameters we have used before). Does the rejection rate change if you change $E[Y^0] = 10$ and sigma to 20?

5. Repeat question 2.3 ($N = 100; 1000; 10000$) but set the false positive rate to 0.10. Show that you get the expected results. (Hint: What variable name in our results is the false positive rate? What code is fundamentally "setting" the false positive rate?)

6. In the slides about the "Hypothesis test machine" it says that the false positive rate is directly set by the analyst. Interpret that with respect to these simulation results.

3 Simulating t-tests when there is an ATE (20 points)

This problem will use much of the same code from the last one, but now we will simulate with an ATE $>$ 0 rather than the null hypothesis.

1. Define "power" in your own words (in the context of a statistical test).

2. Now we will continue the pattern of simulations but we'll add a treatment effect. Let's use the benchmark parameters $N = 1000$, $E[Y^0] = 3$, $ATE = 0.15$, $\sigma = 1$, and a 50/50 split. Run simulations for these six α (alpha) values = 0.01, 0.05, 0.10, 0.15, 0.25, 0.35. Your α settings define six different hypothesis tests. (i) Show a table of power vs. α. (ii) Explain the relationship in the results.

3. Run the same simulations with benchmark parameters from Problem 2.2 (and $\alpha = 0.05$) but vary the treatment % like you did in problem 1.2. Plot power vs. the treatment %. What seems to be the best treatment % in this case?

4. Suppose you are working with a product manager to launch a new product feature based on an A/B test. The feature may have a positive effect on the key metric (Y) but it could also have zero effect. We are confident it will not hurt the metric. We are committed to doing the A/B test and are now planning it based on our simulation results in this problem. (i) Scenario 1: The product manager says that it would be worth it to launch if the ATE is at least 0.15 (same as our simulations), but a full launch of the product would have a very large up-front cost (more engineering time, marketing). What false positive rate do you recommend? Why? (ii) Scenario 2: The product manager says we can launch without any additional costs (beyond the sunk cost of the A/B test). What false positive rate do you recommend? Why? *There isn't necessarily an exactly right answer for these, but you should be able to explain the reasoning and trade-offs involved.*

5. Suppose we are in scenario 1 above and recommended a false positive rate of 0.05. What is the power associated with that test when $N = 1000$? Find 'power_function' in the notebook and adapt it to answer the question. (i) Show that this matches your simulations from before. (ii) Suppose that the product manager says this is not enough power. How else can we increase power without changing the false positive rate? What would be another set of hypothesis test parameters that could achieve at least 80% power? Experiment with 'tt_ind_solve_power' until you find something.

6. **(*Bonus 1: +5 points)** The product manager says she is not comfortable allocating more than 10% of users to the treatment during the A/B test. How large should our overall sample be to reach at least 80% power? Assume we choose a false positive rate of 0.05 and the data-generating process is otherwise the same. Again play with the power function to get an approximate answer.

7. **(*Bonus 2: +5 points)** Repeat Bonus 1 but use a false positive rate of 0.01. What sample size do we need now to get at least 80% power? What are the key trade-offs you are making with the product manager?