Introduction to Statistics (SOC-GA 2332)

Assignment II

Due Date: October 16th, 11:59pm

Instructions

Submit a pdf-file that answers the following problems together with the `R` script that is used in your analysis. Make sure that the script runs without error from start to end before submitting it.

The files you submit should be named as follows: `[Last Name]_Assignment2.pdf` and `[Last Name]_Assignment2.R`. If you are Bob Smith, your document should be named `Smith_Assignment2.pdf` and your `R` script `Smith_Assignment2.R`. The pdf-file should be created with LaTeX or Rmarkdown.

You are encouraged to work with others on the problem sets. However, the documents and `R` scripts that you submit have to be created on your own. Both the pdf-file and the `R` script must be submitted to me via e-mail **before October 16th, 11:59pm**.

Important Notes:

1. These problems are designed to help you figure out what to do when you are confronted with a statistical problem that is not obvious at first sight, rather than to test your knowledge. You've learned all the relevant concepts to answer these questions, either in the lecture or in the math camp. The difficult part is putting these things together.

2. Some of these problems are easy, while others might be rather challenging. In either case, you probably won't be able to solve all of the problems in one day. So, please **plan ahead**. Read through the problems and take your time thinking about them.

3. To reiterate, you are encouraged to **discuss the problems with your classmates**. Some of you will be better in dealing with probability, others will be more advanced in coding; so I hope that you will learn from each other. Notice as well that we have students in this class who are not in your cohort. It would be great if you could reach out to them and work together.

4. When answering the questions, try to be as detailed as possible. Otherwise, it is difficult to give you feedback.

5. If there are one or more problems for which you remain completely clueless by **October 11rd**, send me an e-mail (`bp1094@nyu.edu`). In this e-mail, **please tell me what you have tried so far**, so I can understand exactly why you are struggling and help you efficiently. If you forget to include your prior trials, my first response will be most likely "can you tell me what you have tried so far?"

6. On October 17th, I will share a solution manual that explains the answers to the problems. In the lab following that date, we will have some time to discuss the problems. So, **please read the solution manual before you come to class**.

Less Important:

1. In all numerical answers, round the numbers to three decimal digits. That is, if you get 2.2014542, you should report 2.201.
2. When you include figures in your document, include them within the text rather than pushing them to the end of the document.
3. When creating your `R` script, please keep in mind the **Style Suggestions** that I've provided throughout the Math Refresher and the labs.
4. Make sure that your `R` script is well-documented with comments. In addition, make clear which parts of your code are used for what problems in the problem set. I have numbered all problems, so this should not be difficult.

Problems

Problem 1 (Definitions)

Explain the following concepts:

1. What is the relationship/difference between a parameter, an estimator for that parameter, and an estimate?
2. What is the difference between unbiasedness and consistency? Are these properties of population parameters, estimators, or estimates?
3. In hypothesis testing, what is an α-level?
4. What is a Type I error and what is a Type II error?
5. What is statistical power?
6. What is a 95% confidence interval?
7. What is a p-value?

Problem 2 (Hypothesis Testing)

Siwei and me are discussing whether letting students simulate sampling distributions helps them understand how statistical inference works. We both believe that it does, but to be rigorous we sample $n = 2,000$ students from all sociology departments in the U.S., where half of them ($n_1 = 1,000$) used simulations in their Introduction to Statistic course and the other half did not ($n_0 = 1,000$). We ask students questions regarding their statistical knowledge to see whether there is any difference between those who simulated

and those who did not. Consider the following scenarios. For all the statistical tests, we want to control the probability of falsely rejecting the null hypothesis, when it is in fact true, to be lower than 0.05.

1. Suppose that we measure the statistical knowledge of the students with a categorical variable that has two categories: "high" and "low."
   (a) What statistical test would you conduct in order to examine whether using simulations and statistical knowledge were related?
   (b) What would be the null and alternative hypothesis of this test?
   (c) What would be the test-statistic you would use for this test?
   (d) Assume that the null hypothesis is true. What would be approximate distribution of your test statistic?
   (e) For how large or small of an observed test statistic would you reject the null hypothesis? Explain why.
   (f) Suppose that Siwei and me have gathered the data. We calculate the test statistic on our observed sample and found that the test statistic is 1.98. Calculate the p-value of the test (using R or any other method; you don't have to include the code). Should we reject or not reject the null hypothesis?
2. Suppose we have collected the statistical knowledge with a continuous variable.
   (a) What statistical test would you use to examine whether there is a difference in the mean statistical knowledge between the students who used simulation and those who did not in the population?
   (b) What would be the null and alternative hypothesis of this test?
   (c) What would be the test-statistic you would use for this test?
   (d) Assume that the null hypothesis is true. What would be an approximate distribution your test statistic? (Hint: note that the sample size is "large.")
   (e) For how large or small of an observed test statistic would you reject the null hypothesis? Explain why.
   (f) Suppose that Siwei and me have gathered the data. We calculate the test statistic on our observed sample and found that the test statistic is 1.98. Calculate the p-value of the test (using R or any other method; you don't have to include the code). Should we reject or not reject the null hypothesis?
3. Assume the same scenario as in Problem 2.2. Let the group that used simulations in their Intro to Stats class as group 1 and the group that has not as group 0. We do a literature review of prior studies and figure out that simulations have been shown to be an effective pedagogical tool in related fields. Indeed, it has been shown that it increases the test scores on a very similar statistical knowledge tests as ours by 20 units. So, our best *prior* guess for the average difference of the test scores (in the population) is $\mu_1 - \mu_0 = 20$. Suppose further that we know that the standard deviations of the test score distributions within the two groups is $\sigma_0 = 100$ and $\sigma_1 = 150$, where the subscripts denote the groups. Together, this would indicate that the group that has used simulation has, on average, higher test scores but that their scores varies more than students who didn't use simulations.

(a) Under the same null hypotheses as in Problem 2.2.b, what would be your test statistic to examine whether there is a mean difference between the two groups in the population? (Note: now I have given you much more information regarding the population; hence your answer should be different from the last one.)

(b) What would be the approximate distribution of

$$\frac{\bar{X}_1 - \bar{X}_0}{\sigma_{\text{diff}}}$$

under the null hypothesis? (σ_{diff} is the standard error of $\bar{X}_1 - \bar{X}_0$)

(c) What would be the approximate distribution of

$$\frac{\bar{X}_1 - \bar{X}_0}{\sigma_{\text{diff}}}$$

under the hypothesis that our best prior guess, namely $\mu_1 - \mu_0 = 20$, is the true mean-difference in the population? (Hint: Recall that if a random variable $Z \sim \text{Normal}(0, 1)$ and a and b are constants, then $a + bZ \sim \text{Normal}(a, b^2)$, i.e., $a + bZ$ will be Normally distributed with a mean of a and a variance of b^2.)

(d) Assume that the true population difference in the average scroes between the groups is $\mu_1 - \mu_0 = 20$. Hence, if we sample students, we would sample from the population in which $\mu_1 - \mu_0 = 20$ holds. Yet, when we test hypotheses, we consider the sampling distribution of our test statistic under the assumption that $\mu_1 - \mu_0 = 0$.

Keeping this in mind and using the results from Problem 2.3.c, answer the following questions:

i. What is the probability that we will fail to reject the null hypothesis?

ii. What is the probability that we will correctly reject the null hypothesis?

Problem 3 (With Small Power Comes Large Responsibility)

Suppose we live in a world full of scientists but in which federal funding has been cut so much that no one is able to conduct a study with more than $n = 20$ subjects. Further, the scientists are sincerely interested in whether having a cat at home will make you more liberal. They decide to test this claim rigorously. Yet, as they all want to get their ASR out first (we love "surprising" results!), they refuse to collaborate and conduct their studies independently.

Let μ_0 be the average liberalism in the population of individuals who don't have a cat and μ_1 be the average liberalism of cat owners. Suppose that cat owners and non-cat owners do not differ in their liberalism, i.e., $\mu_1 - \mu_0 = 0$. Suppose further that liberalism is a random variable with distribution Normal$(0, 1)$—i.e., liberalism in the population follows a standard Normal distribution. Lastly, assume that all scientists know that the variance of the population is equal to one, but don't know the value of $\mu_1 - \mu_0$. All of the scientists test the hypotheses

$$H_0 : \mu_0 = \mu_1 \qquad \text{vs.} \qquad H_1 : \mu_0 \neq \mu_1$$

and use an α-level of 0.05

1. Say a scientist randomly samples from the population $n = 20$ individuals, where half of the sample own a cat and the other half doesn't (This can be achieved if the scientists have a register of all individuals in the population who own a cat). Then, they use the statistic
$$T = \frac{\bar{X}_1 - \bar{X}_0}{2/\sqrt{n}}$$
to test their hypotheses. What is the *exact* distribution of this statistic? What would be the probability that s/he falsely rejects the null hypothesis?

2. After several scientists run the same test, one claims to have a significant result and publishes in ASR. Suppose the mean difference $\bar{D} = \bar{X}_1 - \bar{X}_0$ was positive. What is the smallest possible value of $\bar{D}$ that will get published? Why?

3. Compare this value to the distribution of the population. Does this teem to be a small or large difference?

4. With the publication of the ASR, scientists become disinterested in the topic of cat ownership and liberalism. So, they give all of their funding to a graduate student and recommend the topic for a dissertation. The graduate student uses all of the funds to gather a sample of size $n = 10,000$. S/he uses the same statistic as before, namely,
$$T = \frac{\bar{X}_1 - \bar{X}_0}{2/\sqrt{n}}$$
to test the hypotheses. What is the *exact* distribution of this statistic? What is the probability that s/he will falsely reject the null hypothesis?

5. S/he looks into the sample and finds a significant result. Suppose that $\bar{D}$ was positive. What is the smallest possible value of $\bar{D}$ that she found?

6. Compare this value to the distribution of the population. Does this teem to be a small or large difference?

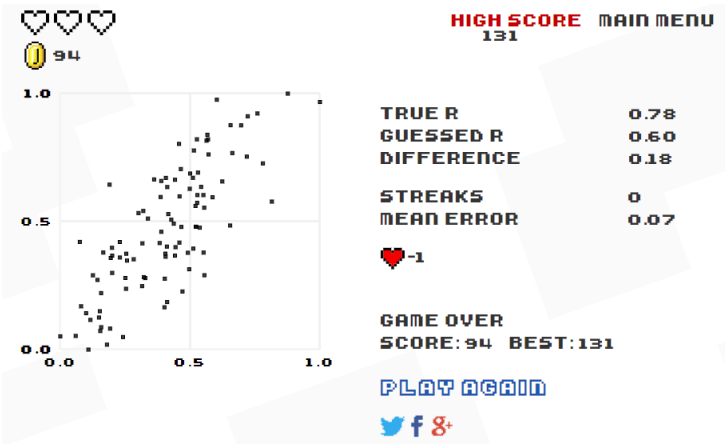
Problem 4 (Correlation & Simple Linear Regression)

1. Go to `http://guessthecorrelation.com/`. Play the game at least 5 times and report your best score as a screenshort, such as that shown in Figure 1

2. Simulate a iid sample of size $n = 5,000$ from the following data-generating process:
$$Y = \beta_0 + \beta_1 X + \epsilon$$
where $X \sim \text{Bernoulli}(\pi)$ and $\epsilon \sim \text{Normal}(0, \sigma^2)$. Let the parameters be $\beta_0 = -1.2$, $\beta_1 = 0.2$, $\pi = 0.4$ and $\sigma = 1.75$.

   (a) Recall that the regression model is a model of a conditional relationship and think carefully what to simulate first.

Figure 1: Example Screenshot



(b) Your code should return a `data.table` object named `regdat`, where the first column, named `y`, contains samples of the outcome variable and the second column, named `x`, contains samples of the predictor variable.

(c) Don't forget to set the seed!

3. Use the following code to download a dataset from my github repository:

```
# url to dataset
url = "https://raw.githubusercontent.com/baruuum/intro_to_stats_2019/master/data/regdat.csv"
# read dataset into R
regdat = data.table(read.csv(url))
```

4. Fit a regression of y_i on x_i using the `lm` function in `R`.

5. What are the hypotheses that you are testing in this model with your t-values in the `(Intercept)` and `x` row of the results?

6. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$.

7. What is the predicted mean response when $X = 1$?

8. Create a 95% confidence interval for the parameter β_1.

9. Create a vector `y1` containing only the y_i values for which $x_i = 1$. Create a vector `y0` similarly for $x_i = 0$. Conduct a two-sample t-test to compare the means of `y1` and `y0` using the `t.test` function in `R`. Make sure to set the `var.equal` option to `TRUE`.

10. State the hypotheses of this test.

11. Compare the results with the results you got about β_1. How do they compare? Why?

Problem 5 (Replication Project)

Download the article for the replication project: Hadas Mandel and Moshe Semyonov. 2016. "Going back in time? Gender differences in trends and sources of the racial pay gap, 1970 to 2010." *American Sociological Review* 81(5): 1039-1068.

1. Read the paper closely and respond the following:

   (a) What are the authors' research questions?

   (b) What is the gap in the literature that the authors aim to fill? How does their analysis advance the literature?

   (c) What is the population that they are making inferences about? Be specific and make sure to identify the geographical region that they are focusing on, the time period, demographic characteristics, and so on.

   (d) Do they have data on all individuals in the population? If they don't, how do they solve this?

   (e) Did the authors collect the data themselves? If they did, describe their sampling procedure. If they didn't, identify and describe the data source and discuss the sampling procedure that was used to collect these data.

2. As you read the text, make a list of all variables or characteristics of the population that the authors mention throughout the paper (e.g., gender, age, wages, occupation, ...). Make sure to read footnotes and table notes; they contain important information to understand who is in the sample. Submit a list with all variables and characteristics that you have identified in the text.

3. Register and login into your IPUMS account (https://usa.ipums.org). Based on your answers to prior questions, select the samples and variables that you think you will need to replicate the paper. Submit a screenshot of the page where you can see the samples and variables that you have selected (you can obtain this from your "Data Cart").