STA 9708, Baruch College, **11-2-22** Prof. L. Tatum

Second Draft: due date corrected, and a clarification

**Final Project Assignment: Sections A, B, and C**

Due Date: Monday, December 19, 2022, 11:59 PM

*General Instructions*

This an individual assignment, not a group assignment. Please remember that I am chiefly interested in documented evidence of engagement with the material presented; do not fear error.

Submit a *single* Microsoft Word file as the project. Accompany that with a single Excel file containing back-office work. I will only look to the Excel file to double-check issues arising in the Word document.

Write in the first person.

Share some of your (a) thought processes, (b) miscues, (c) workarounds, and (d) insights.

Demonstrate engagement with the *concepts* and *approaches* taken in the Lecture Notes.

Do not share your project work with anyone else; that is cheating.

Do not plagiarize; do not plagiarize me.

Number the questions in the same way as shown: A1, A2,...

Answer each question at its numbered position. I will not look elsewhere.

Number your pages.

Do not show the text of my questions. Instead, *fold the question into the answer*. For example, instead of repeating my first question, you could say, "A.1. I am now going to develop a research hypothesis. To get started, I am going to think about ways in which the effects of the Irish Potato Famine might be reflected in..."

Put a title page on your project; on it give your name, course number and date.

For your filename, put your *last name* first.

You are permitted to consult with published sources from outside the course. If you elect to consult such sources, list them in an appendix. If you quote or paraphrase, cite the source with a footnote. If you quote, use quotation marks to avoid plagiarism.

Never share your material with another student. First, that is cheating. Second, because you cannot trust them; they may take your work, unchanged, and present it as their own.

**Part A.**  **One-sided, Two-Sample z-Test of Population Proportions** (See LN8.)

In Part A, get your data from the **U.S. Census** using the methods discussed in LN8. I recommend using FamilySearch.org.

40 points

A1. *Develop* your own research hypothesis to use in a one-sided two-sample z-test for population proportions. One way to start is to imagine the potential social impacts of a major historical event, such as the Civil War, Industrialization, or the Great Migration. Then, think how such an impact might be reflected in a quantitative form in the Census.

Full Census data is available at 10-year intervals from 1790 to 1940. The full records of the 1950 Census have just recently been released and are not yet in a form usable to us. What can be measured depends on what questions were asked in Census used at that time. You can easily search for "index of census questions" to find a breakdown of which questions were asked and when. I fear that sometimes two versions of a census were used in the same year, as appears to be the case with 1820.

Finding a good topic requires working back-and-forth between history and data availability. For example, consider the Pike's Peak Gold Rush, occurring 1858-1861 in parts of what became Kansas and Nebraska. In principle, comparing data from the 1850 and 1860 Census could be used to study the impact of that event. However, Kansas and Nebraska would not appear in either of the 1850 or 1860 Census because they had not yet achieved statehood. Be flexible!

A2. State your null and alternate hypothesis in symbols. Then explain in words what those hypotheses say, being specific to your setting. Explainto a novice how your particular population proportions would in principle be computed.

A3. Gather your data from U.S. Census using original handwritten records. Those are readily accessed using FamilySearch. For each of the two samples, I recommend working from all persons shown on a single sheet of the census, which is usually between 30 and 40 people per sheet. If you are restricting yourself to a smaller group, say, school-age children, then your working sample sizes will of course be smaller.

Describe what you are doing as you present and process the data to get your two sample proportions. Show the reader where and how you went from the census sheets to the proportions. Those images will not be self-explanatory; they must be accompanied with text. Write as if you are interested in the subject and the people, and are addressing someone else who is interested, too.

A4. Walk the reader through the steps of the hypothesis test in the context of your data. As you go, explain how the test progressively answers the question, “How far is far?”

A5. Graph the test, labeling all relevant portions.

A6. As we have stressed, taking a single page from the U.S. Census does not give a truly *random* sample. What specific problems could arise your sample as a result? How would using a random sample, of the same sample size, have helped overcome those problems?

**Part B. One-sided, Two-Sample t-Test of Population Averages**

(see LN7 and Section 4 of LN8)

40 pts

In Part B, get your data from **ship manifests**. I recommend using either FamilySearch.org or the "Ellis Island." I use that second site in Section 4 of LN8. If you plan to use a different site, please come to me office hours, and clear it with me, first. No surprises, please!

B1. *Develop* a research hypothesis to use in a one-sided two-sample test of population averages concerning age, height, or family size. To do so, consider how an event such as a World War, famine, or the shifts in immigration policy *could* produce a change in who immigrates and when. As always, write as if you are interested and even curious as to the nature of things that you encounter, and are addressing someone else who is interested, too.

B2. Gather your two samples from immigration data. Use sample sizes between 5 and 70. Show details of the sourcing and context. Show and use images of the manifests of the ships. Explain what you are doing as you go. Annotate the images so the reader can follow your explanation,

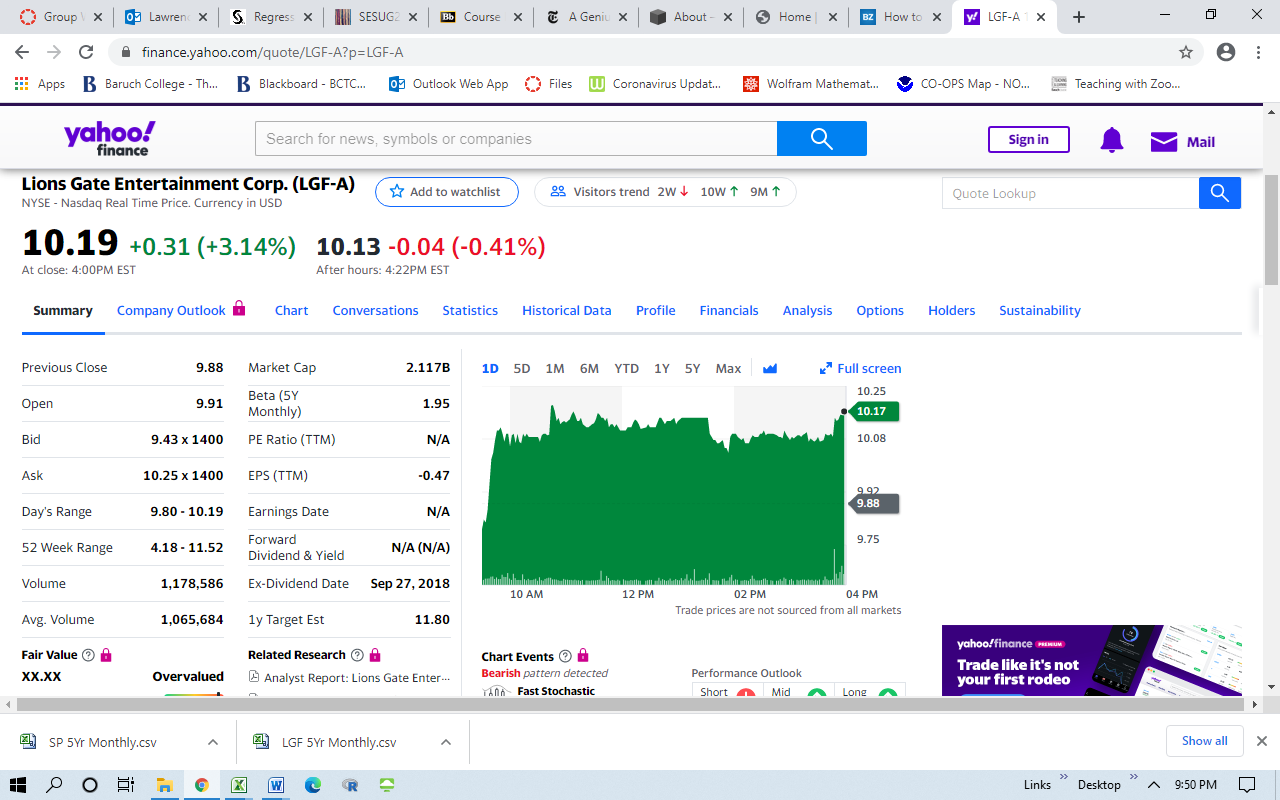
B3. State the null hypotheses in words that apply to the particular topic you are addressing. Define the populations you are referencing, their approximate sizes, and how the population averages would (in principle) be computed.

B4. Perform a one-sided, two-sample t-test. Explain what the software is doing on each line of the table of output. In particular, name the sample statistics that are computed and explain how those are connected with population parameters.

B5. State the conclusion of the test and the grounds. Explain the *reasoning* behind the conclusion.

B6. All serious studies are preceded by a small trial run to look for problems, including lack of clarity in definitions, problems in acquiring data, or a mismatch between the available measurements and what was desired to be measured. Think of your study as a trial run for a larger study, in which many ships will be selected at random. In terms of preparing for that larger study, what problems did you encounter and what changes would you make?

**Part C. Regression Study** (See LN9-LN11)



20 points

**Introduction**

*LN11 Regression Part Two* contains an introduction to CAPM, the *Capital Assets Pricing Model*. The central interest of CAPM is the value of the sample slope coefficient found from regressing the percentage return on a stock against the percentage return on the S&P500. In informal settings in finance, that sample slope is called *Beta*.

I recently conducted an experiment to see if I could match the value of Beta reported by Yahoo Finance for Lions Gate Entertainment Corp (LGF-A). Yahoo reported the Beta as 1.95, as seen in the screenshot to the right.

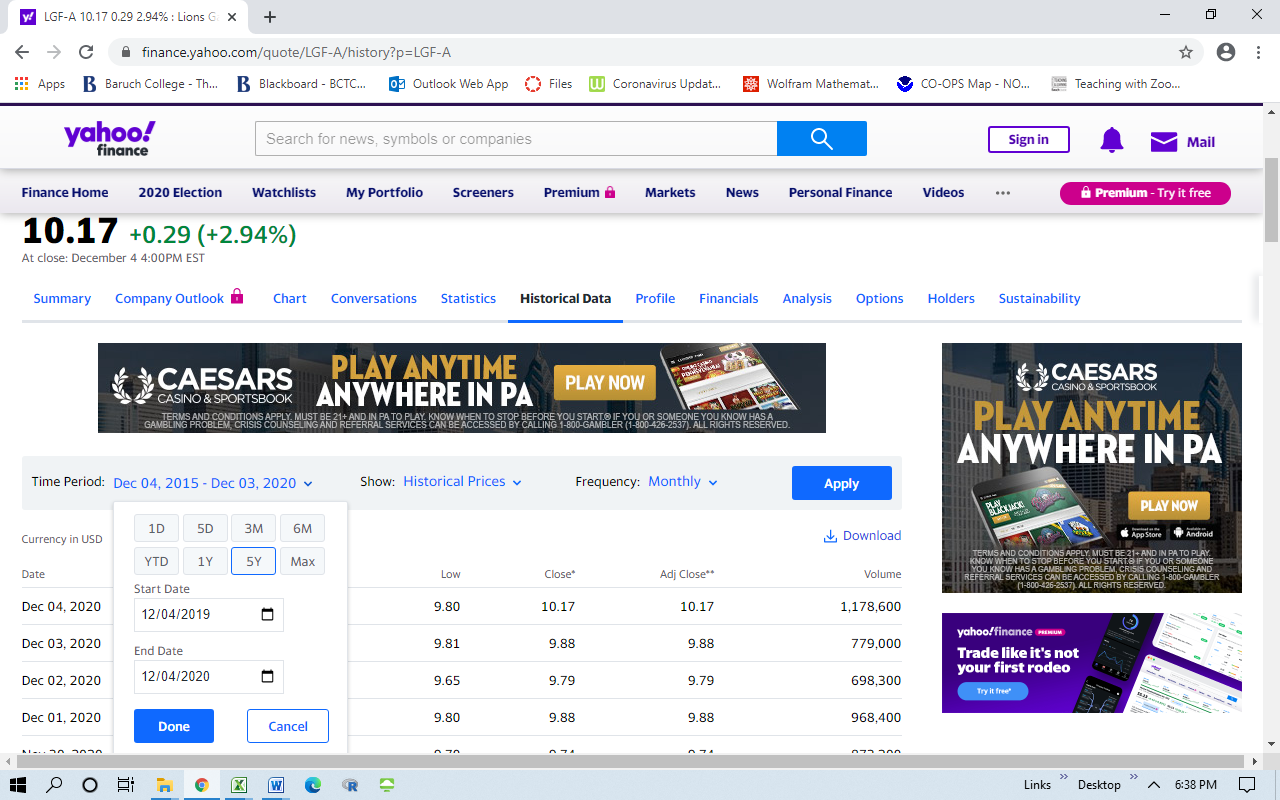
I was able to do reasonably well because I finally took the time to search carefully and read carefully about how Yahoo does the computation. Authoritative Yahoo sources state that Beta is computed using **monthly** returns for a period of **5** **years**. This idea is supported, perhaps, by the cryptic

*5Y Monthly*

modifier-notation used by Yahoo.

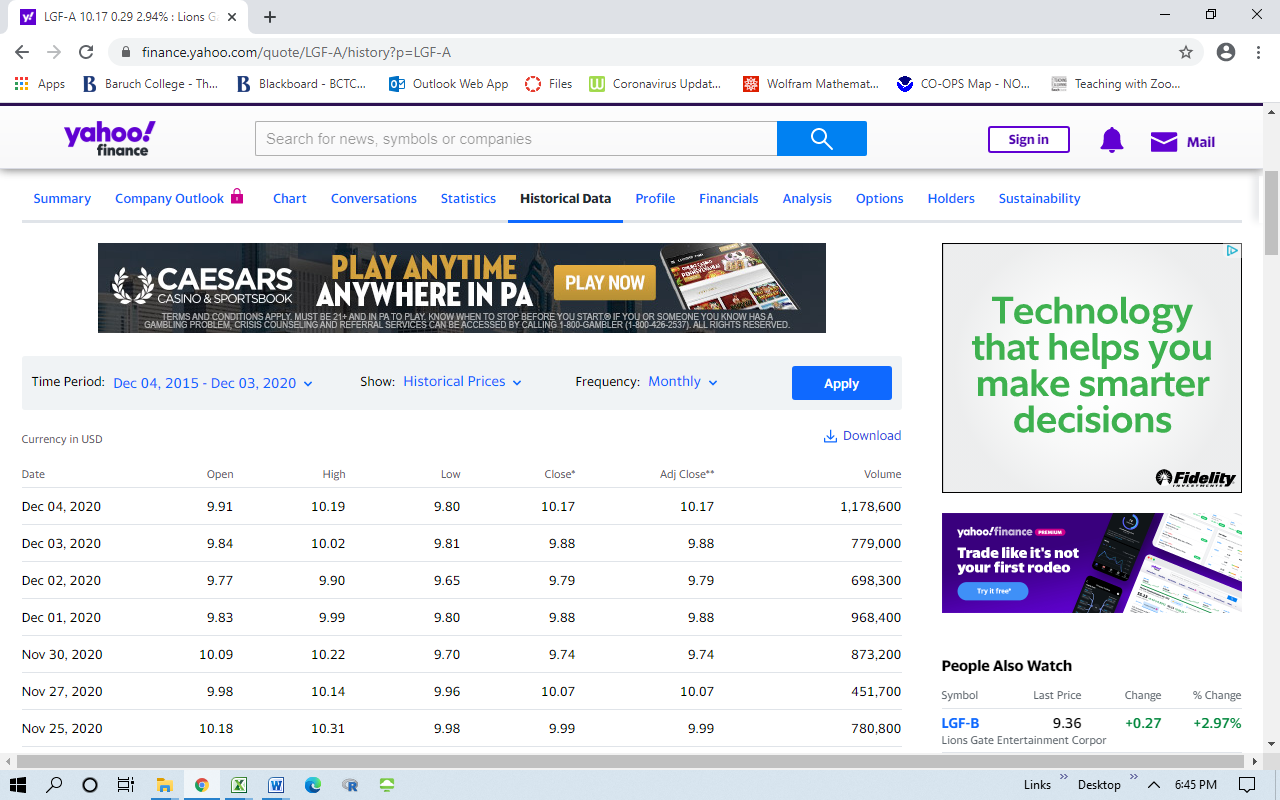
So, I downloaded five years of monthly data for both LGF and S&P500 (^GSPC), computed the monthly returns for both, and regressed the LGF returns on the S&P500 returns. Lo and behold, I got a sample slope of 1.903.

I was surprised and gratified to get a value so close to the one reported by Yahoo; I have not worried at this point about the difference between 1.95 and 1.903.



One time, I was able to employ the **5Y** (five year) button in Yahoo; other times I have typed in the date under "Start Date." In the screenshot I highlight the key features of using that button, and the selection of *monthly* returns.

Recently, the download button did not come on, perhaps because all the data fit on the screen. So, I copied and pasted the monthly returns into Excel: that worked fine for me.



**Assignment**

C1. For your assigned stock and for the S&P, download *monthly* prices for the past five years. Compute your *monthly returns* and those of the S&P. Using the Data Analysis option in Excel, regress your monthly returns on those of the S&P. Show a cropped screenshot that includes Yahoo's "Beta" for your stock and another for your regression. How closely does your sample slope agree with the Yahoo Beta? Explain in your own words what Beta is trying to measure.

C2. Plot your monthly stock returns against the S&P returns. Add the Excel's regression line, then add the approximate 95% prediction intervals, stating how those were computed. Add a vertical cut anywhere you choose and explain how the intersection of that line with the two prediction interval lines is informative.

C3. In the context of your data, explain what is meant by the standard error of the slope - see LN10 for details.

C4. Pick another stock. Download its monthly prices and compute monthly returns for the same time period as above. Add that as a second x-variable to your regression, so you are running a multiple regression! Run Excel to fit the model. You will find that you get a different sample slope for your stock, compared to the slope you found in C2. Explain as best you can why adding the second x-variable changed the slope of the original variable.