CS2810 Mathematics of Data Models	Gold
Spring 2019	March 28, 2019

Homework 08

Due: April 3, 9PM
Point total: 60

Instructions:

- Submit your PDF and/or .py file to Blackboard by the due date and time. Please do not zip your files together, as this interferes with Blackboard's preview functionality. Always show all your work, and for full credit, you must use the method that the problem instructs you to use (unless none is mentioned). Handwritten or typeset solutions are both acceptable, but unreadable submissions will be penalized. You may discuss problems with other students, but you may not write up solutions together, copy solutions from a common whiteboard, or otherwise share your written work or code. Do not use code or language that is copied from the Internet or other students; attribute the ideas *and* rephrase in your own words.
- *Note for this homework:* You should at least set up the initial equation for a distribution correctly for full credit, but you're welcome to use an online solver to get the value after that. For a Gaussian, you don't need to give the PDF, but do give the mean and standard deviation.

Problem 1 (12 pts [2,3,3,4])

Sensor error is often normally distributed. Suppose I have a depth camera where the reading on an object has normally distributed error with mean 0cm, variance 9cm^2. (The error is continuous and can be negative.)

i. What is the probability the sensor is off by more than 6cm in either direction? (You can use the approximation that $1.96 \approx 2$.)

ii. For the same distribution, use the Python function `scipy.stats.norm.cdf()` to find the probability that the error is -1.5cm, or more negative than that.

iii. Again using the CDF function, find the probability that the absolute value of the error is no greater than 1.5cm.

iv. Suppose we were to average the error across 9 measurements from this sensor. Use the sampling distribution of the mean and the CDF function to find the likelihood that our mean error measurement is within 0.5cm of the true mean of 0.

Problem 2 (15 points [2,2,5,4,2])

Suppose the amount of time somebody spends watching videos on YouTube in one day is 4 minutes/day. The distribution is not very smooth at all, since its affected by video times that arent uniformly distributed, but the variance for an individual user is 2 (min/day)^2.

i. According to the Central Limit Theorem, what distribution will model the *total* time spent by all users on YouTube in one day, if there are 5 billion (5,000,000,000) daily users with identical time distributions?

ii. Using the information given so far, calculate the expectation and variance of the distribution you named above.

iii. Suppose we wanted to estimate the average watch time for our own websites videos. It hasnt been up very long, so we only have 6 samples: 1 minute, 4 minutes, 8 minutes, 2 minutes, 15 minutes, 6 minutes. Calculate our mean, then a biased estimate of the variance, then an unbiased estimate of the variance.

iv. Calculate a t-value for our mean watch time's difference from 4 minutes, and use the table at `https://www.ruf.rice.edu/ bioslabs/tools/stats/ttable.html` (or a similar one) to determine whether our site's average watch time is significantly different from YouTube's. (Assume the value of 4 for YouTube is just known as a fact; thus we are comparing to a fixed value.)

v. Given our experimental results, which of the following is the most reasonable conclusion: people watch our videos for longer than YouTube's; people watch YouTube's videos for longer than ours; people actually watch both sites' videos the same amount; or our data was insufficient to come to a conclusion. (Justify your answer.)

Problem 3 (12 points, 4 each)

i. In a small user study of satisfaction with our software, 30 Mac users reported an average satisfaction of 7.5 on a 10 point scale, with unbiased variance estimate .36, while 30 Windows users reported an average satisfaction of 6.8, unbiased variance estimate .32. Can we say that these populations are significantly different with $p < 0.05$? (Again, you can use the critical value table mentioned earlier in the assignment; notice that the degrees of freedom don't really need to be exact if the number is large.)

ii. We improved the Mac features further and found average satisfaction rise to 7.9 for 30 different Mac users, unbiased variance 0.5. Is this significantly better than our previous Mac score?

iii. Suppose we actually want to be very confident the difference between the means for the Mac improvement is at least 0.3, or else this effect size may be too trivial to warrant forcing users to update. Do smaller effect sizes than this lie within a 95% confidence interval of the difference between means that we discovered?

Problem 4 (9 points, 3 each)

i. In your own words, what exactly is the point of testing for statistical significance? Why not always just act as if the differences between means can be trusted?

ii. A large effect size is one way to achieve a good (small) p-value threshold when testing whether populations are significantly different. What's another way?

iii. It's possible for the 95% confidence intervals for two means to overlap, and yet, still have the means be significantly different with $p < 0.05$. Explain how this might happen. (Hint: Suppose the intervals just barely overlap in one place. What independent events have to happen for the difference to be zero?)

Problem 5 (12 points [5,5,2])

For this problem, you'll work with the "adult" census dataset, taken from the UCI Machine Learning repository (`https://archive.ics.uci.edu/ml/datasets/adult`) although the data is provided for you in a cleaner form). You can use the provided `censusreader.csv` to skip the steps of loading and cleaning the data from a CSV. It will return 3 lists, corresponding to three of the columns of the data: "ages" for the ages of census participants, "hours" for the hours-per-week worked, and "over50K" which is 1 if the person in question made more than $50K in a year, and 0 if the person did not.

i. Create a sublist using just the first 30 entries in this file, which you can easily access with the syntax `mylist[:30]`. Use this smaller dataset and the Python function

```
scipy.stats.ttest_ind()
```

to test the following hypotheses: people who make over $50K are older than those who don't; and, people who make over $50K work more hours than those who don't. (This t-test function will work correctly even if N is different between groups.) Report the population means and p-values in both cases. Which differences are significant?

ii. Repeat the experiment using the whole dataset, again reporting means and the p-values calculated by

```
scipy.stats.ttest_ind()
```

. Should we interpret the p-values here literally? Are the findings significant?

iii. Though we talked about bias in estimators in class, you can also have bias if you just don't sample your population in a truly uniform way; selecting *arbitrarily* as we did may not be the same thing as selecting without bias (especially if we don't know how the list was created). Look at our first 30 entries. What's one way in which they are *definitely* not representative of the population at large?