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 $1.96 \approx 2$.)

Solution: One standard deviation is 3cm, so this is just the probability that we're outside 2 standard deviations, which is 5%.

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.

Solution: We can either pass the function our mean and standard deviation, or compute the location in terms of standard deviations away from the mean. `scipy.stats.norm.cdf(-0.5) = scipy.stats.norm.cdf(-1.5, 0, 3) = 0.3085`

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

Solution:

`scipy.stats.norm.cdf(0.5) - scipy.stats.norm.cdf(-0.5) = 0.383`. Or, alternately, `scipy.stats.norm.cdf(1.5,0,3) - scipy.stats.norm.cdf(-1.5,0,3) = 0.383`.

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.

Solution: The sampling distribution of the mean is itself a Gaussian with mean $\mu = 0$ and standard deviation equal to the standard error, $\sigma/\sqrt{N} = 3/3 = 1$. So it turns out that this is the same CDF calculation that we just did, checking whether we're within half a standard deviation to either side of the mean, and the probability is again 0.383.

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 $(\text{min/day})^2$.

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?

Solution: A Gaussian distribution.

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

Solution: These both are just summed, so the expectation is 20 billion min/day, and the variance is 10 billion $(\text{min/day})^2$.

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.

Solution: The mean is $36/6 = 6$ minutes. The biased estimate of the variance is $1/6((-5)^2 + (-2)^2 + 2^2 + (-4)^2 + 9^2 + 0^2) = 130/6 = 65/3 = 212/3$. The unbiased estimate divides by the degrees of freedom $N - 1$ instead of N, $130/5 = 26$.

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.)

Solution: The standard error is $\sigma_{unbiased}/\sqrt{N} = \sqrt{26}/\sqrt{6} = 2.08$. Our t-value is $2/2.08 = 0.96$. The critical value for $p < 0.05$ is 2.57 for 5 degrees of freedom. So this difference is not significant.

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.)

Solution: N is pretty small here, and the difference between the means was actually rather large compared to the size of our values. So the most reasonable conclusion is that we didn't collect enough data, and that we might discover a significant difference if we do collect more. (6 samples is quite small for most experiments; nevertheless, that also means it's possible this difference is an anomaly.)

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.)

Solution: Averaging the variance estimates to 0.34, the standard error is $\sqrt{2 \times 0.34/30} = 0.15$. With a difference in means of 0.7, the t-value is then $0.7/0.15 = 4.67$. The critical value for $29 \times 2 = 58$ degrees of freedom is roughly 2, which we've exceeded. So, yes, the Mac users are significantly more satisfied.

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?

Solution: Averaging the variances to 0.43, we have a standard error of $\sqrt{2 \times 0.43/30} = 0.17$. The t-value for the difference is then $0.4/0.17 = 2.35$, which is still larger than the value of 2 that we established for 95% confidence with 58 degrees of freedom. So this effect is significant, too.

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?

Solution: The confidence interval for the difference is $0.4 \pm 2 \times 0.17 = [0.06, 0.74]$. This includes smaller effect sizes than 0.3, so we must sadly conclude that we're not yet certain the change is worth pushing to production.

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?

Solution: In a world where we believe every difference that is just an artifact of our random sample, we'll make a fair number of decisions that are based on "facts" that just aren't true. We might also waste a lot of time implementing things one way, only to have a study contradict our original conclusions, and then we'll have to do things another way. (And then if that gets contradicted ...) If we cared about the truth of our experiment at all, we should probably care whether our results are real instead of fictitious.

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?

Solution: The main way of doing this is to increase the sample size for the experiment, N. It's also possible to get a better p value by having less variance.

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?)

Solution: When it comes to the values at the edges of the confidence interval, it's already very unlikely that those values are chosen, since they're in the tails of the distribution. To have a difference of zero would require both distributions to choose values in this unlikely range. That's not impossible, but it's more unlikely than either distribution choosing something in this range alone. So if the confidence intervals overlap but the means are still significantly different, it means that even though each distribution has a chance of a value in that overlapping range, the chance that both are in that range is relatively small, as it's the product of the two probabilities.

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?

Solution: The average hours measured are 50 (over 50K) versus 38.95 (under 50K), while the average ages are 42.5 (over 50K) versus 38.8 (under 50K). The hours difference is significant ($p = 0.045$), but the age difference is not ($p = 0.365$).

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?

Solution: Average hours are 45.4 for over 50K, versus 38.8 for under 50K. Average ages are 44.3 for over 50K, versus 36.9 for under 50K. Both have significance with the reported p value of 0. We shouldn't take that literally; there is *some* chance that these results are all a fluke, but this tells us that the probability is so small that Python can't represent it. So yes, the findings are 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?

Solution: The clearest way in which this is true is that there are many more men than women in the sample. We're probably biased in other ways, too, but this is the one where we have a clear idea of what the rate really should be.