CSCI 3022 EXAM 3 REVIEW PROBLEMS

These are additional problems to use as practice when studying for Exam 3 (in addition to lecture examples, in-class notebooks, HW and quiz questions). These questions are not comprehensive , so be sure to also review lecture examples, in-class notebooks, HW and quiz questions as part of your review.

Potentially Useful Python Output

Standard Normal Distribution: Here $\Phi(z)$ is the cumulative distribution function for the standard normal distribution evaluated at z. Its equivalent form in Python is $\Phi(z) =$ stats.norm.cdf(z),

$\Phi(4.75) \approx 1.000$	$\Phi(3.00) = 0.999$	$\Phi(2.58) = 0.995$	$\Phi(2.32) = 0.990$	$\Phi(2.00) = 0.977$
$\Phi(1.96) = 0.975$	$\Phi(1.90) = 0.971$	$\Phi(1.75) = 0.960$	$\Phi(1.64) = 0.950$	$\Phi(1.50) = 0.933$
$\Phi(1.44) = 0.925$	$\Phi(1.2) = 0.885$	$\Phi(1.15) = 0.875$	$\Phi(1.04) = 0.850$	$\Phi(1.00) = 0.841$
$\Phi(0.93) = 0.825$	$\Phi(0.84) = 0.800$	$\Phi(0.76) = 0.775$	$\Phi(0.67) = 0.750$	$\Phi(0.60) = 0.725$
$\Phi(0.52) = 0.700$	$\Phi(0.5) = 0.691$	$\Phi(0.45) = 0.675$	$\Phi(0.44) = 0.67$	$\Phi(0.39) = 0.650$
$\Phi(0.32) = 0.625$	$\Phi(0.25) = 0.600$	$\Phi(0.19) = 0.575$	$\Phi(0.13) = 0.550$	$\Phi(0.06) = 0.525$
$\Phi(0.00) = 0.500$				

```
stats.norm.pdf(1) = 0.24
stats.norm.pdf(0.5) = 0.35
stats.norm.ppf(0.5) = 0
```

1. Definitions/Key Theorems.

   Explain in words what the following mean and WHAT THEY ARE USED FOR:

   a). $\rho(X, Y)$

   b). Power of a hypothesis test

   c). What does the Central Limit Theorem state? What is it used for? What conditions must hold for it to apply?

   d). What is a p-value? Give your definition (a): As a conditional probability and (b): In a sentence that explains what it is used for.

   e). Standard Error of a Statistic

   f). Conditional Independence

   g). A 95% confidence interval for a statistic.

2. Let n be the size of a random sample drawn from a population. Which of the following are **NOT** well-modeled by a normal distribution? (Select all that apply).

   A. the distribution of the sample mean from a normal population when $n = 100$

   B. the distribution of the sample from a normal distribution when $n = 100$.

   C. the distribution of the sample from an exponential population when $n = 100$

   D. the distribution of the sample mean from an exponential population when $n = 100$

   E. the distribution of the sample from a uniform population when $n = 100$

   F. the distribution of the sample mean from a uniform population when $n = 100$

   G. the distribution of the sample median from an exponential population when $n = 100$

   H. the distribution of the sample proportion $\hat{p}$ when $n\hat{p} = 20$ and $n(1 - \hat{p}) = 30$

   Solution: C, E and G are NOT well modeled by normal distributions.

   A is because the sampling distribution of the sample mean of a normal distribution will also be normal (for all values of n)

   B is because the sample is random and so it will have the same distribution as its population.

   D, F and H are all normal thanks to the Central Limit Theorem.

3. Suppose a procedure generates confidence intervals with fixed significance level α which **FAIL** to cover the true mean 2 times out of 20 *on average*. What is the significance level α?

A. 0.01

B. 0.05

C. 0.1

D. 0.20

Solution: C (since we expect $2/20 = \frac{1}{10}$ of the intervals to NOT contain the population mean).

4. Suppose you're only given the following information about two joint random variables X and Y:

$$\mu_X = 2, \;\; \mu_Y = 5, \;\; \sigma_X^2 = 4, \;\; \sigma_Y^2 = 9 \text{ and } \rho(X,Y) = \frac{3}{8}$$

.

For each of the quantities below, calculate if you have enough information, showing all steps. If not, explain what additional info you'd need.

i). $E[X+Y]$

Solution: $E[X+Y] = E[X] + E[Y] = 2 + 5 = 7$

ii). $Cov(X,Y)$

Solution:

Since

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

we know $\frac{3}{8} = \frac{Cov(X,Y)}{(2)(3)} \implies Cov(X,Y) = \frac{3}{8}(6) = \frac{9}{4}$

iii). $Var[X+Y]$

Solution:

$Var[X+Y] = Var[X] + Var[Y] + 2Cov(X,Y) = 4 + 9 + 2(\frac{9}{4}) = \frac{35}{2}$

iii). $E[XY]$

Solution:

$Cov(X,Y) = E[XY] - E[X]E[Y] \implies \frac{9}{4} = E[XY] - (2)(5) \implies E[XY] = \frac{9}{4} + 10 = \frac{49}{4}$

5. Suppose in the general population 5% of people are genetically predisposed to alcoholism. You take a random sample of 4900 people from the general population. Use the Central Limit Theorem to approximate the probability that fewer than 216 of them are genetically predisposed to alcoholism.

Solution

Let $\hat{p}$ be the sample proportion.

We want to calculate $P(\hat{p} < \frac{216}{4900})$.

To calculate any probability we need to know the distribution of the random variable.

By the Central Limit Theorem, we can approximate $\hat{p}$ with a normal distribution (we can use this approximation because $n\hat{p} = 245 > 15$ and $n(1-\hat{p}) = 4655 > 15$)

$$\hat{p} \sim N(p, \frac{p(1-p)}{n})$$

where $p = 0.05$ be the probability of being predisposed to alcoholism and $n = 4900$

i.e.

$$\hat{p} \sim N(0.05, \frac{0.05(0.95)}{4900})$$

(SIDEBAR: wait, why does this work again? The central limit theorem tells us $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ (approximately) where μ is the population mean and σ^2 is the population variance.

Notice that we can write $\hat{p} = \bar{X} = \frac{X_1 + X_2 + X_3 + \ldots X_{4900}}{4900}$, where

$X_i \sim Ber(0.05)$. Thus, the central limit theorem tells us $\hat{p} = \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ (approximately) , where $\mu = E[Ber(0.05)] = 0.05$ and $\sigma^2 = Var[Ber(0.05)] = (0.05)(0.95))$

Now, back to the question.

$P(\hat{p} < \frac{216}{4900}) = P(Z < \frac{\frac{216}{4900} - 0.05}{\sqrt{\frac{0.05(0.95)}{4900}}}) = P(Z < -1.90) = 1 - P(Z < 1.90) = 1 - \Phi(1.9) = 1 - 0.971 = \boxed{.029}$

We can check our answer by calculating the exact probability. Notice that this is the same as asking $P(Y < 216)$ where $Y \sim Bin(4900, 0.05)$.

Using Python we find:

```
stats.binom.cdf(216, 4900, 0.05)
```

```
0.029102295468562835
```

6. Suppose the height of 1st graders at a given elementary school are normally distributed with a mean 45 inches and std deviation of 3 inches. And suppose the height of 2nd graders at this same school are normally distributed with a mean of 50 inches and a std deviation of 4 inches. Suppose there are 110 first graders and 90 second graders at this school. A student is randomly sampled out of all 1st and 2nd graders and all you are told is that their height is 48 inches. What is the probability that the student is a first grader?

   Solution:

   Let F be a *discrete* Bernoulli random variable that denotes whether or not a student is a first grader ($F = 1$ implies first grader, $F = 0$ implies 2nd grader).

   Let H be a *continuous* random variable that denotes the height.

   We are given that random variable $(H|F = 1) \sim N(45, 9)$ and $(H|F = 0) \sim N(50, 16)$

   We are interested in calculating:

   $P(F = 1|H = 48)$.

   We can use Bayes' Theorem (the mixed discrete and continuous version for this):

   $P(F = 1|H = 48) = \frac{f(H=48|F=1)P(F=1)}{f(X=48)}$

   Using the law of total probability in the denominator we can rewrite this as:

   $P(F = 1|H = 48) = \frac{f(H=48|F=1)P(F=1)}{f(X=48|F=1)P(F=1)+f(X=48|F=0)P(F=0)}$

   $P(F = 1) = \frac{110}{200}$

   $P(F = 0) = \frac{90}{200}$

   $f(H = 48|F = 1) = stats.norm.pdf(\frac{48-45}{3}) = stats.norm.pdf(1) = 0.24$ (this is given to us on the top of the first page of this review).

   Similarly: $f(H = 48|F = 1) = stats.norm.pdf(\frac{48-50}{4}) = stats.norm.pdf(-\frac{1}{2})$ (by symmetry of standard normal pdf) $= stats.norm.pdf(\frac{1}{2}) = 0.35$ (also given to us on the first page).

   Putting this altogether we have:

$$P(F = 1|H = 48) = \frac{f(H = 48|F = 1)P(F = 1)}{f(X = 48|F = 1)P(F = 1) + f(X = 48|F = 0)P(F = 0)} = \frac{(0.24)(.55)}{(0.24)(0.55) + (0.35)(0.45)} \approx \boxed{0.456}$$

7. Suppose that the rents paid by a random group of 900 people are i.i.d (independent and identically distributed) each with expectation \$1500 dollars and standard deviation \$810.

   There is about 90% chance that the average rent of the 900 people is in the range \$1500 plus or minus $\$x$. Find x.

   Solution:

   We want to solve $P(1500 - x < \bar{X} < 1500 + x) = 0.90$

   We need to know the distribution of $\bar{X}$ to calculate this.

   By the central limit theorem, $\bar{X} \sim N(E[X], \frac{Var[X]}{900})$ (approximately).

i.e. $\bar{X} \sim N(1500, \frac{810^2}{900})$

$P(1500 - x < \bar{X} < 1500 + x) = 0.90 \implies P(\bar{X} > 1500 + x) = 0.05$

Converting to Z-scores :

$P(\bar{X} > 1500 + x) = P(Z > \frac{1500+x-1500}{\frac{810}{30}}) = P(Z > \frac{30x}{810})$

Thus we want to solve

$P(Z > \frac{30x}{810}) = 0.05$ for x.

Recall:

$P(Z > z_{0.05}) = 0.05$

So we need to find the critical value $z_{0.05}$:

Using the chart on the first page we find:

$z_{0.05} = 1.64$

Thus, we solve:

$\frac{30x}{810} = 1.64 \implies x = 1.64\frac{810}{30} \approx \boxed{44.28}$

Notice where we've seen this before!

$(1.64)(\frac{810}{30}) = (z_{0.05})$(Standard error of sampling mean)

This is the same thing we add and subtract to the test statistic when calculating a 90% confidence interval around a sample mean.

It's also the same thing we would add and subtract to the null hypothesis when finding the acceptance region when testing the null hypothesis!

8. You are assigned to conduct a randomized control trial as part of an effort to encourage high school students from under-resourced communities to apply for college.

   Group A receives special coaching for the ACT. Group B receives no intervention.

   You are interested in determining whether there is a difference in **mean** ACT scores between Groups A and B that is significant at the $\alpha = 5\%$ level.

   a). State the null hypothesis and the alternative hypothesis.

   Solution:

   $H_0 : \mu_A = \mu_B$

   $H_A : \mu_A \neq \mu_B$

   Before conducting the study, you use a power analysis to determine the number of participants to include in your study.

   b). In a power analysis, the necessary sample size depends on what 3 other variables? Give their math notation AND explain what these variables mean in a sentence.

   Solution:

   i). The significance (α): The probability of a Type I error, which is incorrectly rejecting the null when it is in fact true.

   ii). The power ($1 - \beta$): The probability of correctly rejecting the null hypothesis when it is in fact false.

   iii). The expected effect size. The expected effect size is the minimum size effect you hope to be able to detect in a statistical test, such as a 20% improvement in click rates. (Note we haven't introduced a conventional variable for this, as there are different ways to calculate effect size depending on the evaluation design you use).

   c). What is an example of a Type I error in the context of this problem?

   Solution: You incorrectly conclude the coaching made a difference when in fact it didn't.

   i.e. There IS NOT in fact a real difference between mean ACT scores in both groups, but the test statistic you actually gather is very unlucky and (and thus in the tails of the null distribution) leading you to incorrectly reject the null and conclude that there is a difference when there's really not.

   d). What is an example of a Type II error in the context of this problem?

Solution:

There IS in fact a real difference, but you conclude that the coaching made no difference.

e). You conduct your power analysis and find the minimum necessary sample size is 95 participants per group. You end up recruiting 200 participants. A simple random sample of 95 participants received special coaching for the ACT. The remaining participants received no intervention. You find that the 95% confidence interval for the difference in mean ACT scores between those who used the intervention and those who didn't is given by: $[-1, 3]$. What is the conclusion of your hypothesis test? Explain.

Solution:

Since our confidence interval contains 0, and we know that 95% of intervals in this manner contain the tru difference in scores, we do NOT reject the null hypothesis that the mean difference is 0 at the 5% level (i.e. this interval indicates it's possible that there is no difference in scores).

9. An artificial intelligence algorithm is going to be used to make a binary prediction (G for guess) for whether a person will repay a loan. The question has come up: is the algorithm "fair" with respect to a binary protected demographic (D for demographic)? To answer this question we are going to analyze predictions the algorithm made on historical data. We are then going to compare the predictions to the true outcome (T for truth). Consider the following joint probability table from the history of the algorithms predictions:

	$D=0$			$D=1$	
	$G=0$	$G=1$		$G=0$	$G=1$
$T=0$	0.21	0.32	$T=0$	0.01	0.01
$T=1$	0.07	0.28	$T=1$	0.02	

a). What is the value of the missing probability?

Solution: 0.08 (since all of the entries in both tables must add up to a total of 1.

b). What is $P(D=1)$?

Solution:

$$\begin{aligned} \mathrm{P}(D=1) &= \sum_{j\in\{0,1\}} \sum_{k\in\{0,1\}} \mathrm{P}(D=1, G=j, T=k) \\ &= 0.01 + 0.01 + 0.02 + 0.08 = 0.12 \end{aligned}$$

c). Are D and G independent? (Justify)

Solution: No. One way you can show this is to show $P(D=1|G=1) \neq P(D=1)$.

$P(D=1|G=1) = \frac{P(D=1,G=1}{P(G=1)} = \frac{0.01+0.08}{0.01+0.08+0.32+0.28} = 0.1304$

And from above $P(D=1) = 0.12$.

$0.1304 \neq 0.12$

Thus D and G are not independent.

(Sidenote: If these two HAD been equal, we still would have had to check this same thing with all the other possible values for D and G before concluding they were independent).

d). Are D and G conditionally independent given T? (Justify)

Solution: No. We just need to demonstrate $P(D=1|G=1,T=1) \neq P(D=1|T=1)$ with one counterexample.

Notice: $P(D=1|G=1,T=0) = \frac{P(D=1,G=1,T=0)}{P(G=1,T=0)} = \frac{0.01}{0.32+0.01} = \frac{1}{33}$

$P(D=1|T=0) = \frac{P(D=1,T=0}{P(T=0)} = \frac{0.01+0.01}{.21+.32+.01+.01} = \frac{2}{55}$

Since these aren't equal, we have found a counterexample and thus D and G are NOT conditionally independent given T.

e). To test whether this algorithm is fair, we will compare 2 different definitions of "fairness":

i). Definition 1: An algorithm satisfies "parity" if the probability that the algorithm makes a positive prediction ($G = 1$) is the same regardless of being conditioned on a demographic variable.

Does this algorithm satisfy parity? (i.e. does the following equality hold?)

$$P(G = 1|D = 1) \stackrel{?}{=} P(G = 1|D = 0)$$

Solution: No. See full solution here: Fairness in AI

ii). Definition 2: An algorithm satisfies "equality of odds" if the probability that the algorithm predicts a positive outcome ($G = 1$) is the same regardless of demographics *given* that the outcome will occur ($T = 1$). Does this algorithm satisfy equality of odds? (i.e. does the following equality hold?)

$$P(G = 1|D = 0, T = 1) \stackrel{?}{=} P(G = 1|D = 1, T = 1)$$

Solution: Yes. See full solution here: Fairness in AI

Fun Fact: It turns out, it can actually be proven that these cannot be jointly optimized, and this is called the Impossibility Theorem of Machine Fairness. In other words, any AI system we build will necessarily violate some notion of fairness. For a deeper treatment of the subject, here is a useful summary of the latest research: Pessach et al. Algorithmic Fairness.