Prelim 2 for BTRY6010/ILRST6100

November 2, 2017: 7:30pm-9:30pm

Name:________________________________

NetID:_______________________________

Lab: (circle one)

Lab 402: Tues 1:25PM - 2:40PM

Lab 403: Tues 2:55PM - 4:10PM

Lab 404: Wed 2:55PM - 4:10PM

Lab 405: Tues 7:30PM - 8:45PM

Score: ______ / 60

Instructions

1. Please **do not turn to next page** until instructed to do so.
2. You have 120 minutes to complete this exam.
3. The last page of this exam has some useful formulas.
4. No textbook, calculators, phone, computer, notes, etc. allowed (please keep your phones off or do not bring them to the exam).
5. Please answer questions in the spaces provided.
6. When asked to calculate a number, it is sufficient to write out the full expression in numbers or code without actually calculating the value. E.g., $\frac{1+3\times\frac{4}{7}}{3+0.7}$ or `1 - pnorm(3)^2` are valid answers.
7. Please read the following statement and sign before beginning the exam.

Academic Integrity

I, ____________________, certify that this work is entirely my own. I will not look at any of my peers' answers or communicate in any way with my peers. I will not use any resource other than a pen/pencil. I will behave honorably in all ways and in accordance with Cornell's Code of Academic Integrity.

Signature: _______________ **Date:**_____________

For each problem in this section, please circle one of the following answers. No justification is required.

1. [2 points] Your waiting time to catch a bus to campus on a weekday is distributed uniformly between 5 minutes and 10 minutes. The average waiting time over 5 days is a random variable with mean
   (a) 7.5 ***minutes.***
   (b) more than 7.5 minutes.
   (c) less than 7.5 minutes.
   (d) Not enough information to conclude any of the above.

2. [2 points] Your waiting times to catch a bus to campus on a weekday has a standard deviation of 1.4 minutes. The average waiting time over 5 days is a random variable with a standard deviation of
   (a) ***less than*** 1.4 ***minutes.***
   (b) more than 1.4 minutes.
   (c) 1.4 minutes.
   (d) Not enough information to conclude any of the above.

3. [2 points] Rainfall in Ithaca during the months of September-November is historically known to be distributed normally with a mean of 2 inches and a standard deviation of 1 inch. Which of these `R` commands is **incorrect** for calculating the chance of 5 inches or more rain during September-November 2018?
   (a) `pnorm(5, mean = 2, sd = 1, lower.tail=FALSE)`
   (b) `1 - pnorm(5, mean = 2, sd = 1)`
   (c) ~~***pnorm(5, mean = 2, sd = 1)***~~
   (d) `pnorm(-1, mean = 2, sd = 1)`

4. [2 points] You want to compare average rainfall in Ithaca during September-November to the average rainfall during January-March. Which of the following scenarios is the most relevant?
   (a) one population mean (μ)
   (b) mean difference of two unrelated populations ($\mu_1 - \mu_2$)
   (c) ***mean difference of two related populations (μ_d)***
   (d) none of the above

5. [2 points] Suppose the average rainfall in Ithaca during September-November is not different from the average rainfall in January-March. You conducted a hypothesis test at 5% significance level to compare the two. What is the chance that your test will reject H_0?
   (a) ***5%***
   (b) 2.5%
   (c) none of the above
   (d) Not enough information to conclude any of the above.

6. [2 points] Alex and Bob want to test if the average rainfall in Ithaca during September-November is 3 inches or if it is more than that. Alex decided to reject H_0 if the average rainfall in September-November of 2018 is more than 4 inches, while Bob decided to reject H_0 only if the average rainfall exceeds 5 inches. Who has a more powerful test?

   (a) ***Alex***

   (b) Bob

   (c) Their tests have equal power.

   (d) Not enough information to conclude any of the above.

True or False?

For each question in this section, please answer either *True* or *False.* While justification is not required, it is encouraged and may allow in some cases for partial credit to be awarded.

7. [2 points] Scientists tested a new drug and did not find evidence that it increases sleep hours for patients with insomnia. This conclusion may have been due to a type-I error.

   FALSE.

   H_0 is not rejected. If H_0 was true then there is no error. If H_0 is false then then this is a type-II error.

8. [2 points] Consider testing a population mean $H_0 : \mu = 10$ vs. $H_A : \mu \neq 10$. Based on a sample of size $n = 25$, a test provided a p-value of 0.02. Then a 95% confidence interval for μ constructed using the same sample will contain the value 10.

   FALSE.
   $\left|\frac{\bar{X}_n-10}{s/5}\right| = t_{0.01,24} \implies \bar{X}_n = 10 + t_{0.01,24} \cdot \frac{s}{5}$ *or* $\bar{X}_n = 10 - t_{0.01,24} \cdot \frac{s}{5}$.
   In the first case, when $\bar{X}_n = 10 + t_{0.01,24} \cdot \frac{s}{5}$, the lower bound of the 95% CI is $\bar{X}_n - t_{0.025,24} \cdot \frac{s}{5} = 10 + (t_{0.01,24} - t_{0.025,24}) \cdot \frac{s}{5} > 10$.
   In the second case, when $\bar{X}_n = 10 - t_{0.01,24} \cdot \frac{s}{5}$, the upper bound of the 95% CI is $\bar{X}_n + t_{0.025,24} \cdot \frac{s}{5} = 10 + (t_{0.025,24} - t_{0.01,24}) \cdot \frac{s}{5} < 10$.
   In either case 10 is outside the CI.

9. [2 points] You are interested in testing if the average number of customers in Purity Ice Cream on a Tuesday night exceeds 150. Sales records over the last 12 Tuesdays show that the store had 98 customers on average. We know that the number of customers ariving on a Tuesday follows an approximate normal distribution, and a t-test is applicable. Without calculating standard error or test statistic, we can say that this test will fail to reject H_0 at $\alpha = 0.05$.

   TRUE.

   $H_0 : \mu = 150, H_A : \mu > 150$. $\bar{X}_n = 98$ *means the test statistic* $(98 - 150)/SE(\bar{X}_n)$ *will be negative. Since alternative hypothesis is* $\mu > 135$*, the p-value will be more than* 0.5.

10. [2 points] A one-sample mean test $H_0 : \mu = 5$ vs. $H_A : \mu > 5$ rejected the null hypothesis at $\alpha = 0.05$, since the p-value for the given dataset was 0.03. If we tested $H_0 : \mu = 5$ vs. $H_A : \mu \neq 5$ with the same dataset, then we would have rejected H_0 at $\alpha = 0.05$.

    FALSE.

    $\frac{\bar{X}_n-5}{s/\sqrt{n}} = t_{0.03,n-1} < t_{0.025,n-1}$

11. [2 points] A one-sample mean test $H_0 : \mu = 5$ vs. $H_A : \mu \neq 5$ rejected the null hypothesis at $\alpha = 0.05$, since the p-value for the given dataset was 0.03. Based only on this information, we can conclude if we tested $H_0 : \mu = 5$ vs. $H_A : \mu > 5$ with the same dataset, then we would reject H_0 at $\alpha = 0.05$.

FALSE.

$\left|\frac{\bar{X}_n - 5}{s/\sqrt{n}}\right| = t_{0.015,n-1} \implies \frac{\bar{X}_n - 5}{s/\sqrt{n}} = \pm\, t_{0.015,n-1}$. *Now* $t_{0.015,n-1} > t_{0.05,n-1}$ *but* $-t_{0.015,n-1} < t_{0.05,n-1}$ *and either could happen.*

12. [2 points] Suppose $[2, 5]$ is a 90% confidence interval for μ. Then there is a 90% probability that μ is in the interval $[2, 5]$.

FALSE.

μ is unknown, but not random.

13. [2 points] Suppose $[2, 5]$ is a 95% confidence interval for μ. Then if we could repeat the experiment over and over, we would expect that 95% of the time μ would be in the interval $[2, 5]$.

FALSE.

CI changes if we repeat the experiment.

14. [2 points] In order to reduce the margin of error in a 95% confidence interval for the difference of two related means by half, we need to quadruple the sample size.

TRUE.

$ME = \frac{constant}{\sqrt{n}} \implies \frac{ME}{2} = \frac{constant}{\sqrt{4n}}$.

15. [2 points] You want to find out the average weekly hours spent by a Cornell undergrad in the gymnasium. You constructed a 95% confidence interval (7, 12) for it using a sample of size $n = 12$. You later noticed that you used the t-distribution with 12 degrees of freedom to calculate your multipliers. Your friend pointed out that you should use $n - 1 = 11$ in your degree of freedom formula. After you correct your calculation, the resulting interval will be wider.

TRUE.

$t_{0.025,12} \times SE = \frac{12-7}{2} = 2.5$ *and* $t_{0.025,11} > t_{0.025,12}$.

16. [5 points] Suppose we want to estimate the proportion (p) of males in a population. We take a sample of 1000 people and record their gender in a vector `Gender`. If the i^{th} person is female then we set `Gender[i]` to be `TRUE`, otherwise we set it to be `FALSE`. Once we have the vector `Gender` we used the following `R` code to calculate a 99% confidence interval for p.

```
n = 1000    # <-- Correct line


# Calculate sample proportion
# phat = mean(Gender)    # <-- Wrong line

# We want proportion of males but we record males as FALSE, or in other words as 0.
phat = 1 - mean(Gender)

# Calculate standard error
# s = sd(Gender)    # <-- Wrong/unnecessary line

s = sqrt(phat*(1-phat))

# SE = s^2/sqrt(n)    # <-- Wrong line

SE = s/sqrt(n)
# There can be other correct combinations for the two above lines

# Calculate multipliers/quantile
# alpha = 1 - 0.95    # <-- Wrong line

# Confidence level is 99% not 95%
alpha = 1 - 0.99

# quant = qt(alpha/2, n-1)    # <-- Wrong line

# We use the normal distribution, not the t-distribution. Also note the sign.
quant = -qnorm(alpha/2)

# Calculate bounds
lower = phat - quant * SE    # <-- Correct line


upper = phat + quant * SE    # <-- Correct line


# Print confidence interval
c(lower, upper)    # <-- Correct line


# End
```

Now this code has multiple errors in it. Your job is to correct it. At the end of your corrections we should have a code chunk such that running it in `R` would give us the correct confidence interval.

If you think a line of code is correct then leave it as is. If you think a line of code has error(s) in it, then scratch it out and write the correct code in the space provided directly below that line.

17. [5 points] For each of the following questions, answer which of the five scenarios is most appropriate.

(a) Compare the average rent of studio apartments in downtown Ithaca and suburban Boston

(i) μ (ii) p (iii) μ_d (iv) $p_1 - p_2$ (v) $\underline{\boldsymbol{\mu_1 - \mu_2}}$

(b) Test if a drug enhances activity in mice by recording their movements before and after administering the drug

(i) μ (ii) p (iii) $\underline{\boldsymbol{\mu_\mathbf{d}}}$ (iv) $p_1 - p_2$ (v) $\mu_1 - \mu_2$

(c) Test if a majority of college drop-outs decided to do so because of high tuition rates

(i) μ (ii) $\underline{\mathbf{p}}$ (iii) μ_d (iv) $p_1 - p_2$ (v) $\mu_1 - \mu_2$

(d) Test if smokers have a higher incidence rate of lung cancer than non-smokers

(i) μ (ii) p (iii) μ_d (iv) $\underline{\mathbf{p_1 - p_2}}$ (v) $\mu_1 - \mu_2$

(e) Estimate the average increase in your monthly expense of gasoline after switching to a hybrid car

(i) μ (ii) p (iii) $\underline{\boldsymbol{\mu_\mathbf{d}}}$ (iv) $p_1 - p_2$ (v) $\mu_1 - \mu_2$

For the following question an answer without justification will not receive full credit.

18. 100 randomly selected residents of Tompkins county were asked if they believe that climate change is real. 56 of them said 'yes'. We are interested in estimating the proportion of Tompkins county residents who believe that climate change is real.

(a) [1 point] Define the parameter of interest (p) in the context of this problem.

Parameter of interest (p) is the proportion of Tompkins county residents who believe climate change is real.

(b) [5 points] Calculate the value of the sample statistic ($\hat{p}$), and specify its distribution. Clearly state and/or check any assumption you made to arrive at this conclusion.

$$\hat{p} = \frac{\textbf{\# Residents in the sample who believe climate change is real}}{\textbf{\# Residents in the sample}} = \frac{56}{100} = 0.56$$

is the value for this specific sample. It's distribution is

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

The assumptions to be satisfied for this to hold true are:

i. **Each resident's opinion does not depend on anyone else, i.e., the datapoints are independent.**

ii. **There are enough success-failure cases *in the sample.***

$$n\hat{p} = 100 \times 0.56 = 56 \geq 10; \; n(1-\hat{p}) = 100 \times 0.44 = 44 \geq 10$$

(c) [4 points] Construct a 95% confidence interval for p, and interpret it in the context of the problem.

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ **is the standard error of $\hat{p}$. A 95% CI for p would be**

$$\begin{aligned}&\left(\hat{p} - z_{\alpha/2} \cdot SE(\hat{p}), \hat{p} + z_{\alpha/2} \cdot SE(\hat{p})\right)\\ =&\left(0.56 - z_{\alpha/2} \cdot \sqrt{\frac{0.56 \times 0.44}{100}}, 0.56 + z_{\alpha/2} \cdot \sqrt{\frac{0.56 \times 0.44}{100}}\right),\\ &\text{where } \alpha = 1 - 0.95 = 0.05,\ z_{\alpha/2} = -\texttt{qnorm(0.05/2)} = 1.96\end{aligned}$$

We are 95% confident that p, the true proportion of Tompkins county residents who believe climate change is real, is between the numbers $0.56 - 1.96 \cdot \sqrt{\frac{0.56 \times 0.44}{100}}$ and $0.56 + 1.96 \cdot \sqrt{\frac{0.56 \times 0.44}{100}}$.

Now suppose we want to test if a majority of residents in Tompkins county believe that climate change is real, with a significance level of 5%. Answer the following questions below, with justifications as needed.

(d) [2 points] What are the null and alternative hypotheses?

$$\begin{aligned}H_0 &: p = 0.5\\ H_A &: p > 0.5\end{aligned}$$

(e) [2 points] What is the test statistic (expressed as a random variable) and what is its distribution under the null hypothesis?

The test statistic is $Z = \frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}}$. Under H_0, $Z \sim N(0, 1)$.

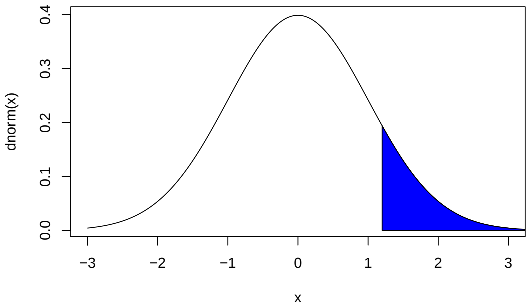
(f) [1 point] Calculate the realized value of the test statistic in this example.

Realised value of Z is

$$\hat{Z} = \frac{0.56 - 0.5}{\sqrt{\frac{0.5\times 0.5}{100}}} = \frac{0.06}{\frac{0.5}{10}} = \frac{0.06}{0.05} = \frac{6}{5} = 1.2$$

(g) [2 points] What is the p-value of this test? (You should draw a picture as you answer this.)

p-value for the test is `(1-pnorm(1.2))` = 0.1150697.



(h) [2 points] Clearly write down which assumptions you would check to see if the distribution you stated in (e) is acceptable. Are these different from the assumptions you used in (b)?

The assumptions to be satisfied for the p-value calculation to hold true are:

i. **Each resident's opinion does not depend on anyone else, i.e., the datapoints are independent.**

ii. **There are enough success-failure cases *under the null***

$$np_0 = 100 \times 0.5 = 50 \geq 10;\ n(1-p_0) = 100 \times 0.5 = 50 \geq 10$$

So the second assumption to be checked here is different compared to the previous case in (b).

(i) [1 point] Let us suppose that the p-value of this test is more than 0.07. Based on this, is it justified to say that a majority of Tompkins county residents believe that climate change is real?

The p-value $>$ 0.07 $>$ 0.05, the significance level. Since the p-value for this sample is more than the significance level of this test, we *cannot* reject the null hypothesis ($p = 0.5$) in favour of the alternative ($p > 0.5$). Thus it is *not* justified to say that a majority of Tompkins county residents believe that climate change is real.

Formula Sheet

The law of total probability:

$$P(B) = P(A)P(B|A) + P(A^C)P(B|A^C)$$

Bayes' theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^C)P(B|A^C)}.$$

Binomial distribution: $X \sim \text{Binomial}(n, p)$

- probability mass function (pmf):

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n$$

- expected value $E(X) = np$
- variance $\text{Var}(X) = np(1-p)$

Poisson distribution: $X \sim \text{Poisson}(\lambda)$

- probability mass function (pmf):

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, \ldots$$

- expected value $E(X) = \lambda$
- variance $\text{Var}(X) = \lambda$

Uniform distribution: $X \sim Unif(a, b)$

- probability density function (pdf):

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

- expected value $E(X) = (a+b)/2$
- variance $\text{Var}(X) = (b-a)^2/12$

Normal distribution: $X \sim \text{Normal}(\mu, \sigma)$

- probability density function (pdf):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

- expected value $E(X) = \mu$
- variance $\text{Var}(X) = \sigma^2$
- If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma}$ has a $N(0, 1)$ distribution.

Scenario	**One population Mean**	**One population proportion**	**Paired Mean**
Parameter	μ	p	μ_D
Statistic / Point Estimate	$\bar{x}$	$\hat{p} = \frac{X}{n}$	$\bar{x}_D$
Standard Error (of point estimate)	$SE(\bar{x}) = \sigma/\sqrt{n}$ (σ known) $SE(\bar{x}) = S/\sqrt{n}$ (σ unknown)	$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE(\bar{x}_D) = \sigma_D/\sqrt{n}$ (σ_D known) $SE(\bar{x}_D) = S_D/\sqrt{n}$ (σ unknown)
Confidence Interval	$(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ (σ known) $(\bar{x} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}})$ (σ unknown)	$\hat{p} \pm z_{\alpha/2}SE(\hat{p})$	$(\bar{x}_D \pm z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}})$ (σ_D known) $(\bar{x}_D \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}})$ (σ_D unknown)
Hypothesis Testing	$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ (σ known) $T = \frac{\bar{x}-\mu_0}{S/\sqrt{n}}$ (σ unknown)	$Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$Z = \frac{\bar{x}_D-0}{\sigma_D/\sqrt{n}}$ (σ_D known) $T = \frac{\bar{x}_D-0}{S_D/\sqrt{n}}$ (σ_D unknown)

- **Sampling Distribution of $\bar{X}$**: If X has expectation μ and standard deviation σ. Then, $E(\bar{X}) = \mu$, $SD(\bar{X}) = \sigma/\sqrt{n}$.
  - If $X \sim N(\mu, \sigma)$, then $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
  - For large sample ($n \geq 30$), under suitable assumptions on X, $\bar{X}$ is *approximately* $N(\mu, \sigma/\sqrt{n})$

- **Sampling Distribution of $\hat{p}$**: $\hat{p} = X/n$, $E(\hat{p}) = p$, $SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$.
  - For large sample ($np \geq 10, \quad n(1-p) \geq 10$), $\hat{p}$ is *approximately* $N(p, \sqrt{\frac{p(1-p)}{n}})$.

- ***t*-distribution**: If $X \sim N(\mu, \sigma)$, then $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows a t-distribution with degree of freedom $n-1$.
  - With larger n, t-distribution with n degrees of freedom becomes more similar to $N(0,1)$.