# Prelim 2 for BTRY6010/ILRST6100

*November 1, 2018: 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:** ______ / 65

## 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 two pages of this exam have 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.

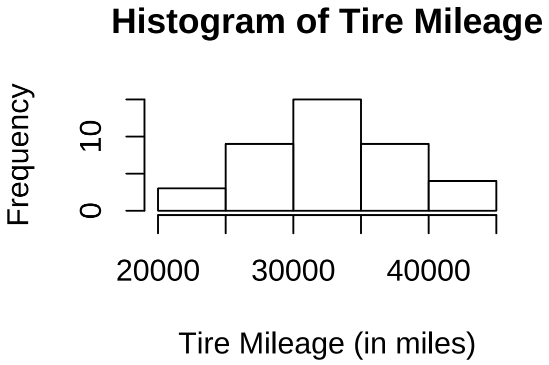
## Academic Integrity

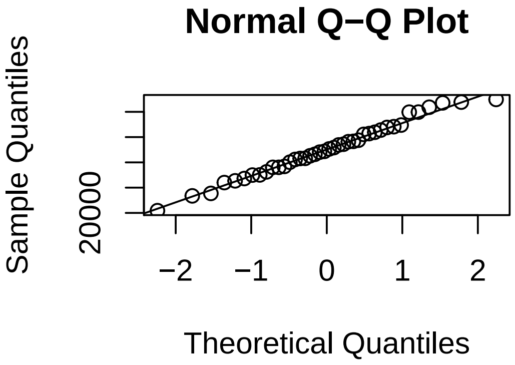
I, _____________________, certify that this work is entirely my own. I will not look at any of my peers' answers or communicate in any way with my peers. I will not use any resource other than a pen/pencil. I will behave honorably in all ways and in accordance with Cornell's Code of Academic Integrity.

**Signature:** ________________ **Date:**______________

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

1. A factory makes $2,000$ tires a day, and mileage of tires (denoted by a random variable $X$) are independent of each other. One day, a simple random sample of 40 tires was drawn and their mileages were tested. It was found that the sample standard deviation of their mileage was 5347.58 miles. Figures and summary statistics of the mileages of the 40 sample tires are shown below.





| minimum | 1st quartile | median | mean | 3rd quartile | maximum |
|---|---|---|---|---|---|
| 20444.65 | 29020.58 | 32420.57 | 32590.95 | 36069.02 | 42454.10 |

We are interested in the mean mileage of tires made on that day in the factory.

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

(b) [3 points] Give a point estimator of $\mu$ and specify its distribution. What is the realized value of your point estimator in this sample? [Hint: point estimator is a random variable, its realization is a numeric value).

(c) [3 points] Clearly state and/or check any assumption you made to arrive at the conclusion in (b).

(d) [4 points] Based on this sample of 40 tires, construct a 95% confidence interval for $\mu$, and interpret it in the context of the problem.

Now, suppose we want to test whether or not the mean mileage of tires made on that day is $30,000$ miles, with a significance level of 5%.

(e) [3 points] What are the null and alternative hypotheses? Interpret the significance level $\alpha = 5\%$ in the context of the problem.

(f) [3 points] What is the test statistic (expressed as a random variable) and what is its distribution under the null hypothesis?

(g) [2 points] Calculate the realized value of the test statistic in this example.

(h) [3 points] What is the p-value of this test? (You should draw a picture as you answer this, clearly mark the area used to calculate p-value, and provide an R command you would use to calculate this.)

(i) [2 points] Clearly write down the assumptions you would check to see if the distribution of test statistic you stated before is acceptable.

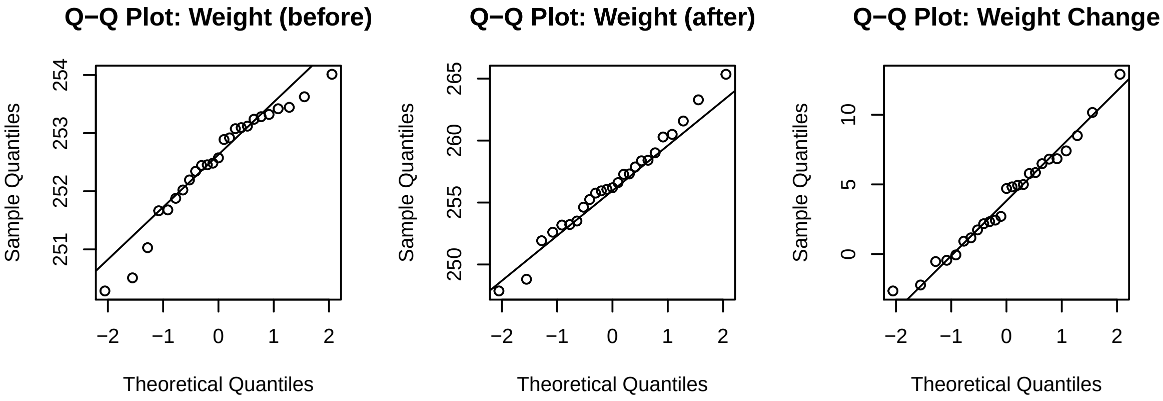
(j) [2 points] Let us assume that the p-value of this test is smaller than 0.03. Based on this, is there sufficient evidence in the data to say that the mean mileage of tires made on that day is different from $30,000$ miles?

(k) [2 points] Comment on whether we could have known the decision of this hypothesis test (reject or fail to reject null) based on the 95% confidence interval constructed before.

(l) [2 points] Based on this p-value and the summary statistics, is it possible to know the decision of the hypothesis test $H_0 : \mu = 30,000$ vs. $H_A : \mu < 30,000$ at 5% level of significance?

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

2. We administered a drug believed to affect weight to $n = 25$ rats. The rats were weighed immediately before administration of the drug, and after one month, the rats were weighed again. Before, the mean rat weight was 252.5 ounces. One month later, the mean rat weight was 256.5 ounces. The sample standard deviation of the weight change was 5 ounces. The following quantile-quantile plots suggest approximate normality of rat weights after 1 month and their weight changes.


Q−Q Plot: Weight (before)
Q−Q Plot: Weight (after)
Q−Q Plot: Weight Change

We are interested in seeing if, on average, there was an *increase* in rat weight one month after drug administration.

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

(b) [4 points] Specify the distribution of your sample statistic, and calculate its realized value in this data. Clearly state and/or check any assumption you made to arrive at this conclusion.

(c) [4 points] Construct a 90% confidence interval for $\hat{\mu}_d$, and interpret it in the context of the problem.

(d) [2 points] Looking at your confidence interval, your colleague had the impression that there is a 90% probability that the true increase in average weight is somewhere in your confidence interval. Please provide him with an explanation (in words) of what "90% confidence level" means in the context of your problem.

(e) [2 points] Comment on how you would expect the length of the 90% confidence interval to change if we had only $n = 4$ rats in our sample. Assume the sample standard deviation is still 5 ounces.

(f) [1 point] Specify all the Q-Q plots that you used to justify normality assumption (if any) for your answer in (b).

Now, suppose that we want to test if there was an *increase* in the weight after drug administration, with a significance level of 0.5%.

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

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

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

(j) [2 points] What is the p-value of this test? (You should draw a picture as you answer this, clearly mark the area used to calculate p-value, and provide an R command you would use to calculate this.)

(k) [2 points] Assume the p-value is less than 0.01. Based on this, report the conclusion of your test in words.

(l) [2 points] Comment on what information in your data (other than the p-value) can help a scientist decide if the results are *practically* significant.

(m) [2 points] Explain in words what "power of a test" means in the context of this problem. What can be done to increase power of this test when significance level is set at 1%?

(n) [2 points] Suppose we would like to test if, on average, there was a *decrease* in rat weight between drug administration and one month later. A new set of data is collected, where the sample means of weights before and after drug administration are 252 and 255 ounces, respectively. What can be said about the p-value of this test?

3. [6 points] The following code attempts to generate a sampling distribution of $\bar{X}_4 = (X_1+X_2+X_3+X_4)/4$, where $X_1, X_2, X_3, X_4$ are independent normal random variables with mean 50 and *variance* 4. Then it attemps to overlay the true density function of $\bar{X}_4$ on top of the probability histogram obtained from the simulated data.

```r
#set n
n = 4


#initialize realizations
num.simulations = 10000


xbar.realizations = rep(NA, num.simulations)



#simulate realizations
for (i in 1:num.simulations) {


  xbar.realizations[i] = mean(4*rnorm(n)+50)


}


#plot histogram
hist(xbar.realizations,
main = expression(paste('Histogram of 10,000 Draws of ',bar(X)[n])))



#overlay true density
xvalues = seq(47, 53, 0.01)


yvalues = dnorm(xvalues,50, sqrt(2/n))


lines(xvalues,yvalues)
```

Now this code has some error/omission in it. Your job is to fix that. At the end of your corrections we should have a code chunk such that running it in `R` would give us the correct graphs.

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

# 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 | $\mu$ | $p$ | $\mu_D$ |
| Statistic / Point Estimate | $\bar{x}$ | $\hat{p} = \frac{X}{n}$ | $\bar{x}_D$ |
| Standard Error (of point estimate) | $SE(\bar{x}) = \sigma/\sqrt{n}$ ($\sigma$ known)<br>$SE(\bar{x}) = S/\sqrt{n}$ ($\sigma$ unknown) | $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ | $SE(\bar{x}_D) = \sigma_D/\sqrt{n}$ ($\sigma_D$ known)<br>$SE(\bar{x}_D) = S_D/\sqrt{n}$ ($\sigma$ unknown) |
| Confidence Interval | $(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ ($\sigma$ known)<br>$(\bar{x} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}})$ ($\sigma$ unknown) | $\hat{p} \pm z_{\alpha/2}SE(\hat{p})$ | $(\bar{x}_D \pm z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}})$ ($\sigma_D$ known)<br>$(\bar{x}_D \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}})$ ($\sigma_D$ unknown) |
| Hypothesis Testing | $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ ($\sigma$ known)<br>$T = \frac{\bar{x}-\mu_0}{S/\sqrt{n}}$ ($\sigma$ unknown) | $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ | $Z = \frac{\bar{x}_D-0}{\sigma_D/\sqrt{n}}$ ($\sigma_D$ known)<br>$T = \frac{\bar{x}_D-0}{S_D/\sqrt{n}}$ ($\sigma_D$ unknown) |

- **Sampling Distribution of** $\bar{X}$: If $X$ has expectation $\mu$ and standard deviation $\sigma$. 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 $N(0,1)$.