GROUND RULES:

- Print your name at the top of this page. Do not put your name on any other page.
- This is a closed-book and closed-notes exam.
- This exam contains 5 questions, each worth 12 points. This exam is worth **60 points** total.
- You may use a calculator, but this calculator cannot have internet access. You cannot use your phone as a calculator. You cannot share calculators with another student. Show all of your work; use a calculator only to do final calculations or to check your work.
- Each problem contains parts. On each part, there is opportunity for partial credit, so show all of your work and explain all of your reasoning. **Translation:** No work/no explanation means no credit.
- On any problem, you may use the back of the page if you need more space. I also have extra paper if you need it.
- Any discussion or inappropriate communication between you and another examinee, as well as the appearance of any unnecessary material, will result in a declaration of academic dishonesty. Don't risk it!
- You have 75 minutes to complete this exam.

HONOR PLEDGE FOR THIS EXAM:

After you have finished the exam, please read the following statement and sign your name below it.

> *I promise that I did not discuss any aspect of this exam with anyone other than the instructor, that I neither gave nor received any unauthorized assistance on this exam, and that the work presented herein is entirely my own.*

1. Johnson Controls claims the activation temperature of its new sprinkler system for residential use (X, in deg F) is normally distributed with population mean $\mu = 130$ and population variance $\sigma^2 = 2.25$. A random sample of 20 sprinkler systems is observed and the activation temperatures $X_1, X_2, ..., X_{20}$ are recorded.

(a) What is the population in this example? Give a reasonable answer.

(b) Let $\overline{X}$ denote the sample mean of the 20 activation temperatures. What is the sampling distribution of the sample mean $\overline{X}$?

(c) Calculate the standard error of the sample mean $\overline{X}$. What does this measure?

(d) The sample mean and sample standard deviation of the 20 activation temperatures (`temp`) are shown below:

```
> mean(temp)
[1] 135.3
> sd(temp)
[1] 1.15
```

I calculated a 99% confidence interval for the population mean μ using R's `t.test` function:

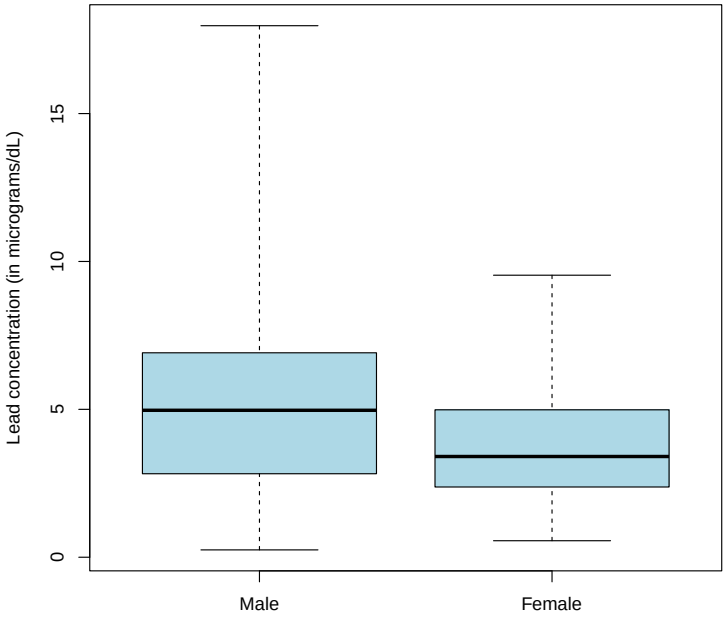
```
> t.test(temp,conf.level=0.99)$conf.int
[1] 134.5 136.0
```

Is this confidence interval consistent with Johnson Controls' claim that the population mean activation temperature is 130 deg F? Explain.

2. Public health officials in New Jersey wanted to compare lead levels in the blood of male and female hazardous waste workers employed in the state. Two independent random samples were observed:

- $n_1 = 152$ male workers
- $n_2 = 86$ female workers.

Lead concentrations (in micrograms per deciliter) were measured on each worker. Here are side-by-side boxplots of the data:



(a) If you wanted to write a confidence interval for $\Delta = \mu_1 - \mu_2$, the difference of the two population mean lead concentrations (1 = male; 2 = female), select the confidence interval you would use:

- the one that assumed equal population variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- the one that did not assume the population variances were equal.

Explain why you chose the answer you did. In addition, what statistical inference procedure could be used to determine which assumption is more reasonable?

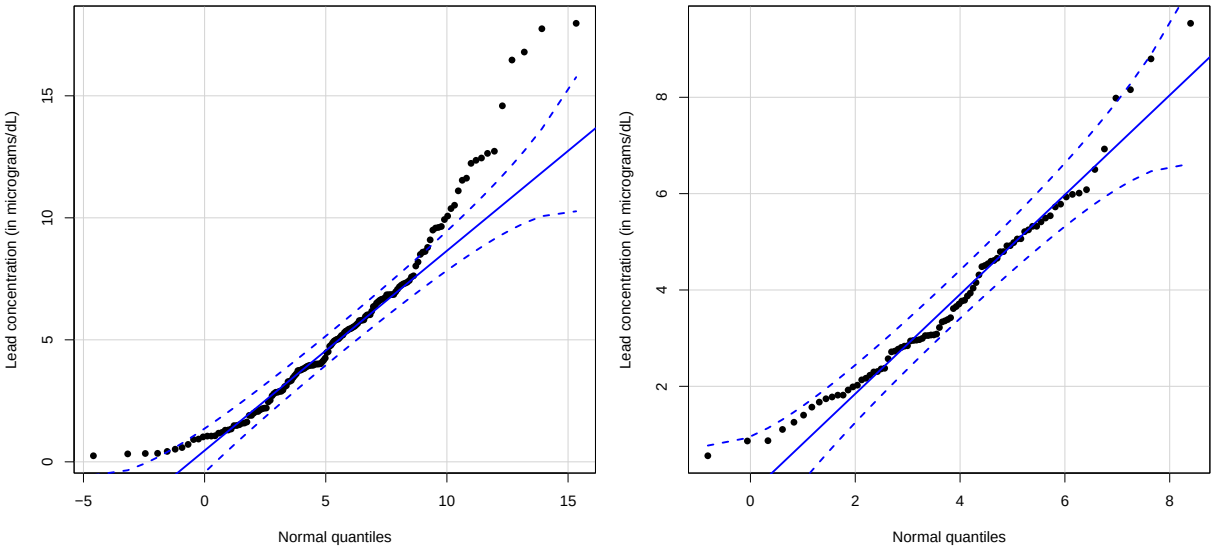
Two more questions are on the next page.

(b) I asked R to get both confidence intervals in part (a), each at the 95% confidence level. Here are the intervals:

```
> t.test(male,female,conf.level=0.95,var.equal=TRUE)$conf.int
[1] 0.75 2.41
> t.test(male,female,conf.level=0.95,var.equal=FALSE)$conf.int
[1] 0.88 2.28
```

For the interval you picked in part (a), interpret what it means. What does your interval suggest about the mean lead concentrations for the two populations?

(c) Here are the normal quantile-quantile plots for the male (left) and female (right) lead concentrations:



Interpret the plots. Are you concerned about your conclusions in part (b)? Explain.

3. In Connecticut, a random sample of 200 legally registered automobiles was selected from DOT records. Among the 200 automobiles selected, only 124 passed the state's emission test for pollution.

I calculated a 90% confidence interval for the population proportion of automobiles that met the state's emissions standards to be $(0.56, 0.68)$. I used the formula

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

and recorded all calculations using 2 digits.

(a) What is the population here? What is the sample?

(b) Draw a detailed picture showing me how $z_{\alpha/2}$ is determined. I am looking for a clear answer.

(c) An environmental engineer would like to design a larger study to estimate the population proportion. She would like to write a 99% confidence interval ($z_{0.01/2} \approx 2.58$) that will have margin of error equal to 0.01. How many cars will she need to sample? Comment on whether you think this is feasible and then list two ways she could reduce the number of automobiles needed.

4. New York City's Yellow Taxi has about 15,000 taxis in its fleet. A manager is trying to decide whether using radial tires or belted tires improves his fleet's fuel economy on average.

- He randomly samples $n = 12$ cars equipped with radial tires and has them driven over a test course.
- Using the same drivers, the same cars are then equipped with belted tires and are driven through the same test course.

The gasoline consumption (in kilometers per liter) was recorded for each car and tire type:

Car	1	2	3	4	5	6	7	8	9	10	11	12
Radial	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2
Belted	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.7	6.0	4.9

(a) Explain why this is a matched pairs study. Are the radial and belted samples independent or dependent?

(b) I used R to write 95% and 99% confidence intervals for $\Delta = \mu_1 - \mu_2$, the difference of the population mean gasoline consumption for the two types of tires (1 = radial; 2 = belted):

```
> t.test(diff,conf.level=0.95)$conf.int
[1] 0.02 0.27
> t.test(diff,conf.level=0.99)$conf.int
[1] -0.04 0.32
```

(i) What does `diff` mean in the code above? *Hint:* How are data from a matched pairs study analyzed?

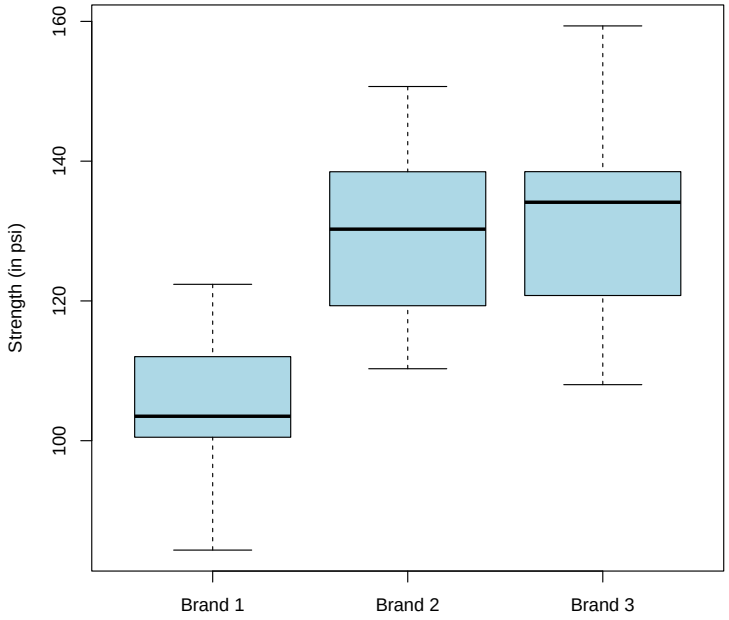
(ii) Pick one confidence interval above (tell me which one) and interpret it for the manager.

(c) The analysis above shows that different conclusions would be made about how the population means compare depending on which confidence level is used. Explain to the manager why this happens and why the two different conclusions would both be supported by the data above.

The next page is blank if you would like to use it for your answers to Problem 4.

This is a blank page for your answers to Problem 4. Use it if you wish.

5. Investigators wanted to perform a one-way classification analysis with three brands of "double wall" boxes. Twelve boxes of each brand were subjected to a compression test and the strength of each box was measured in pounds per square inch (psi). There were 36 boxes in all; 12 of each brand. Here are side-by-side boxplots of the data:



Here is the analysis of variance (ANOVA) table for these data:

```
> fit = lm(Strength ~ Brand)
> anova(fit)
Analysis of Variance Table

Response: Strength
          Df Sum Sq Mean Sq F value  Pr(>F)
Brand      2   5387  2693.6   16.29 1.2e-05
Residuals 33   5458   165.4
```

(a) The F statistic ($F = 16.29$) is used to test two hypotheses: H_0 and H_1. Write out what these hypotheses are. You can do this using notation (that you must clearly define) or you can write this out in words. Which one of your hypotheses is more supported by the data? Why?

Two more questions are on the next page.

(b) One of the statistical assumptions in a one-way classification analysis is that the variances of the populations being compared are the same. Using the information on the preceding page, report an estimate of what this common population variance σ^2 is. What are the units attached to your estimate?

(c) Here is the R output to do a follow-up Tukey analysis:

```
> TukeyHSD(aov(fit),conf.level=0.95)

Tukey multiple comparisons of means
95% family-wise confidence level

                  diff    lwr   upr  p adj
Brand.2-Brand.1 25.25  12.37 38.13 0.0001
Brand.3-Brand.1 26.60  13.72 39.48 0.0001
Brand.3-Brand.2  1.35 -11.53 14.23 0.9643
```

(i) Explain what is meant by "95% family-wise confidence level."

(ii) If you were advising the investigators on which box type to use (to maximize population mean strength), what would you tell them? Defend your conclusions with statistical evidence.