## 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 4 problems, each worth 12 points. This exam is worth **48 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!
- Tables 1 and 2 are attached at the end if you need to do any calculations for a normal distribution or a $t$ distribution.
- You have 55 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. The presence of high arsenic (As) levels in water wells in southeastern New Hampshire has been recognized by the United States Environmental Protection Agency since the late 1970's. Geologists recently observed a random sample of $n = 32$ water wells in southeastern New Hampshire and recorded the arsenic concentration (in parts per billion, ppb) for each well. The data are shown below:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Well | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| As (ppb) | 45.3 | 12.6 | 25.5 | 97.3 | 56.8 | 31.1 | 29.6 | 43.3 | 180.2 | 33.6 | 51.1 | 82.4 |
| Well | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| As (ppb) | 1.8 | 55.2 | 6.3 | 5.7 | 40.9 | 11.2 | 10.1 | 96.2 | 43.6 | 4.8 | 44.6 | 11.9 |
| Well | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | | | | |
| As (ppb) | 8.5 | 112.2 | 137.8 | 118.9 | 44.0 | 34.9 | 59.1 | 4.6 | | | | |

(a) What is the population in this example? There is no "right" answer. Give a reasonable one.

(b) I used R to calculate the 95% confidence interval

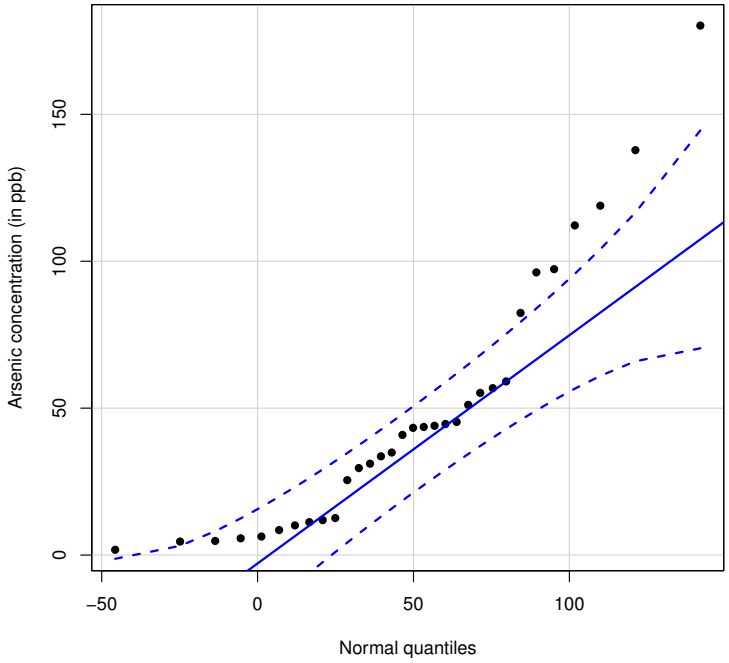
$$\overline{y} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}.$$

```
> t.test(arsenic,conf.level=0.95)$conf.int
[1] 32.5 63.9
```

Interpret what this interval means.

**Part (c) is on the next page.**

(c) I used R to construct a normal quantile-quantile (qq) plot of the data on the previous page. This plot is shown below:




Do you think these data are well represented by a normal population distribution? Explain what effect, if any, your answer here has on your conclusions in part (b).

2. The manager of a large hotel chain with locations all throughout the southeastern US would like to estimate the population mean amount spent by adult residents per day (in dollars) "after accounting for the cost of their room." This includes things like mini-bar snacks and drinks, room service, hotel restaurant meals, late charges, and other extra expenses.

He is willing to assume the population distribution of daily extra expenses per adult resident is $\mathcal{N}(\mu, \sigma^2)$. In addition,

- He wants to write a 99% confidence interval using
$$\overline{Y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$
- He wants the margin of error of the confidence interval above to equal 2.5 dollars.
- He doesn't know the population variance $\sigma^2$, but he is willing to elicit a **range** of daily extra expenses per adult resident. This range is from 0 dollars to 120 dollars.

(a) Determine the sample size (the number of adult residents that should be sampled) to accomplish his goals. Show all of your work.

(b) What are two things he could do to **reduce** the sample size required?

3. The NFL's Scouting Combine provides an opportunity for participants to display their professional football potential. A special teams coach at last year's event wanted to compare the population mean punting distance (in yards) between two types of footballs:

Type 1: Air-filled footballs

Type 2: Helium-filled footballs.

He recruited $n = 15$ punters and had each of them punt each type of ball. The order in which each punter punted an air-filled football and a helium-filled football was randomized. Here were the observed punt distances (measured in yards):

| Punter | Air | Helium |
|---|---|---|
| 1 | 66.4 | 65.2 |
| 2 | 54.8 | 57.9 |
| 3 | 53.7 | 51.8 |
| 4 | 47.1 | 48.3 |
| 5 | 43.8 | 43.5 |
| 6 | 47.1 | 50.2 |
| 7 | 49.9 | 53.2 |
| 8 | 43.2 | 45.5 |
| 9 | 52.8 | 56.1 |
| 10 | 58.7 | 60.6 |
| 11 | 42.7 | 47.3 |
| 12 | 54.6 | 58.1 |
| 13 | 50.4 | 50.3 |
| 14 | 54.9 | 56.1 |
| 15 | 56.1 | 57.5 |

(a) Explain why this is a matched-pairs study.

**Parts (b), (c), and (d) are on the next two pages.**

(b) I asked R to calculate a 95% confidence interval for the population mean difference $\mu_1 - \mu_2$ which acknowledges the matched-pairs design:

```
> t.test(air,helium,paired=TRUE,conf.level=0.95)$conf.int
[1] -2.7 -0.6
```

Interpret what this interval means.

(c) If you were to test

$$H_0 : \mu_1 = \mu_2$$
$$\text{versus}$$
$$H_a : \mu_1 \neq \mu_2$$

using a matched-pairs $t$ test with significance level $\alpha = 0.05$, would the probability value (p-value) be greater than 0.05 or less than 0.05? Explain. You shouldn't have to perform any calculations here.

**Part (d) is on the next page.**

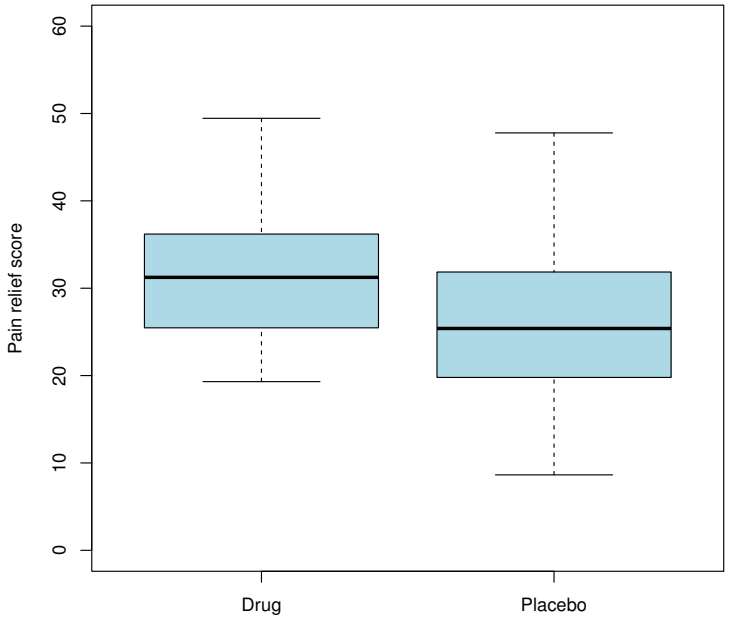
(d) If I incorrectly analyzed the data as data from two **independent** samples of punters, then a 95% confidence interval for $\mu_1 - \mu_2$ would be

```
> t.test(air,helium,conf.level=0.95,var.equal=TRUE)$conf.int
[1] -6.4  3.0
```

Why is this interval so much wider than the confidence interval in part (b)? Don't just say, for example, "You are analyzing the data incorrectly" or "These are not independent samples." Get to the heart of the issue.

4. A pain-killing drug was tested for efficacy in 52 women who were experiencing uterine cramping pain following childbirth. Twenty-six (26) women were randomly assigned to receive the drug, and the remaining 26 women received a placebo (an inert substance with no pain relief ingredients). Researchers then recorded $Y$, the pain relief score for each of the 52 women. *The higher the score, the more pain relief experienced.*

Side-by-side boxplots of the two samples of pain relief scores are shown below:




A table of summary statistics from the 52 women is shown below (SD = standard deviation):

| Treatment | Mean pain relief score | SD | $n$ |
|---|---|---|---|
| Drug (Group 1) | 31.9 | 7.6 | 26 |
| Placebo (Group 2) | 26.3 | 9.1 | 26 |

The researchers wanted to demonstrate taking the drug improved pain relief at the population level, so the hypothesis test

$$H_0 : \mu_1 = \mu_2$$
versus
$$H_a : \mu_1 > \mu_2$$

was performed using $\alpha = 0.05$. The researchers were willing to assume equal population variances ($\sigma_1^2 = \sigma_2^2$) for the two distributions of pain relief scores.

**Parts (a), (b), and (c) are on the next two pages.**

(a) Calculate the two-sample $t$ statistic which assumes equal population variances:

$$t = \frac{\overline{y}_{1+} - \overline{y}_{2+}}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Recall the pooled sample variance estimate $s_p^2$ is a weighted average of the two sample variances $s_1^2$ and $s_2^2$.

**Parts (b) and (c) are on the next page.**

(b) Identify the rejection region by drawing a picture of the appropriate reference distribution and shading in the region. Make sure you identify what the reference distribution is. Neatness and precision counts.

(c) State your decision. Write conclusions in terms of the problem.