

## Q1

0 Points

### Instructions

This is a 75-minute exam, though you have up to 2 hours to complete it. There are 100 points in total, broken down as follows:

Q2: 21 points

Q3: 20 points

Q4: 24 points

Q5: 15 points

Q6: 20 points

You may use R, as well as any notes, books, or *existing* internet resources to answer the questions. However, you may not collaborate or communicate to anyone except the instructor regarding the exam (e.g., you may not communicate with other students, the TAs, or post/solicit help on the internet or via any other communication means). Note that you may need to install some packages if you are using RStudio Cloud.

By taking this exam, you pledge to uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

Please sign your name electronically if you agree to abide by the Duke Community Standard in completing this exam (note that failure to do so will result in consequences outlined in the course syllabus).

## Q2

21 Points

Researchers are interested in the relationship between atmospheric pollution levels and temperature. They examined daily levels of various atmospheric pollutants at a monitoring station and summarized daily pollution levels as being acceptable, elevated, dangerous, and hazardous.

### Q2.1

3 Points

What type of variable would best describe this daily pollution summary?

- ☐ Categorical nominal
- ☒ Categorical ordinal
- ☐ Numeric discrete
- ☐ Numeric continuous

### Q2.2

4 Points

Were these data from an observational study or from an experiment? Explain.

**Q2.3**

3 Points

What distribution might best be used to model the total number of days in a given month where air quality was dangerous or hazardous?

**Q2.4**

6 Points

Are the assumptions needed for your distribution in part 1.3 satisfied? Explain.

**Q2.5**

5 Points

One component variable of the pollution summary was atmospheric PM2.5 level (very fine particulate matter), measured in  $\mu g/m^3$ . Suppose the researchers wanted to visualize the distribution of PM2.5 levels and were considering either a boxplot or histogram. Discuss the pros/cons of each, being sure to mention something apparent from a histogram that is not easily captured in a boxplot *and vice-versa*.

### Q3

20 Points

Continuing from Question 1, the researchers' data may be found at the following GitHub repository (this dataset is based on real data from a monitoring station in Beijing):

<https://classroom.github.com/a/rOkhiEDs>

The researchers are interested in potential associations between atmospheric pollution and weather conditions, and asked you to create a data visualization. The variables in the dataset are below. Missing values are coded as -99.

PM2.5: PM2.5 levels in  $\mu g/m^3$

PM10: PM10 levels in  $\mu g/m^3$

S02: Sulfur dioxide levels in  $\mu g/m^3$

N02: Nitrogen dioxide levels in  $\mu g/m^3$

TEMP: Mean temperature during the day in degrees Celsius

PRES: Mean barometric pressure during the day in hPa

month: Month (1 = January, 2 = February, etc.)

#### Q3.1

7 Points

Create and upload a visualization that effectively summarizes the relationship between nitrogen dioxide levels, mean daily temperature, and season, being sure to use visualization best practices. For the purposes of your visualization, you may treat the months of Dec./Jan./Feb. as Winter, Mar./Apr./May as Spring, Jun./Jul./Aug. as Summer, and Sep./Oct./Nov. as Fall.


 No files uploaded

### Q3.2

6 Points

In the space provided, provide any code you used to create your visualization, including data import and any potential processing/manipulation.

You may upload any supporting documentation if needed.

 No files uploaded

### Q3.3

7 Points

Describe any relationship(s) you see.

## Q4

24 Points


The global coronavirus pandemic illustrates the need for accurate testing of COVID-19 as its extreme infectivity poses a significant public health threat. Due to the time-sensitive nature of the situation, the FDA has enacted emergency authorization of a number of serological tests for COVID-19. The Abbott Alinity CMIA test for COVID-19 has an estimated sensitivity of 100% and specificity of 99%.

### Q4.1

7 Points

Assume the prevalence of COVID-19 in a population is 6%. What is the probability that the Alinity test administered to a randomly selected individual in this population will be positive? **Round to 4 decimal places, and show your work/code for full credit.**

You may upload any supporting documentation if needed.


 No files uploaded

### Q4.2

8 Points

Still assume the prevalence of COVID-19 in a population is 6%. What is the probability that, out of 20 patients who tested positive with the Alinity test, at least two are actually *negative* for COVID-19? **Round to 4 decimal places, and show your work/code for full credit.**

You may upload any supporting documentation if needed.


 No files uploaded

### Q4.3

9 Points

What COVID-19 prevalence levels would correspond to positive predictive values of 90% or greater for the Alinity test? **Round to 4 decimal places, and show your work/code for full credit.**

You may upload any supporting documentation if needed.

 No files uploaded

### Q5

15 Points

Lead poisoning in children is a serious medical issue that can lead to growth and developmental disabilities. It is known that the natural log of blood lead concentration is approximately normally distributed. Before major legislation such as the ban on lead paint and elimination of lead from gasoline, historical lead levels in children had a mean of  $2.7 \log(\mu\text{g}/\text{dL})$  and a standard deviation of  $0.39 \log(\mu\text{g}/\text{dL})$  (log refers to the natural log).

#### Q5.1

4 Points

Children are classified as having high blood lead levels and at risk of lead poisoning if their log-blood lead concentration is above the 95th percentile. At what log-blood lead levels would children have been considered "at risk" prior to the lead-reducing legislation? **Round to 4 decimal places, and show your work/code for full credit.**

You may upload any supporting documentation if needed.

 No files uploaded

### Q5.2

6 Points

What is the probability that a randomly selected child from this historical population would be between  $2.5 \log(\mu/dL)$  and  $3.0 \log(\mu/dL)$ ? **Round to 4 decimal places, and show your work/code for full credit.**

You may upload any supporting documentation if needed.


 No files uploaded

### Q5.3

5 Points

The 95th percentile of log-blood lead concentration based on the most recent NHANES study is now  $1.61 \log(\mu/dL)$ . What is the probability that a child from before the introduction of lead-reducing legislation would be classified as "at risk" using modern standards? **Round to 4 decimal places, and show your work/code for full credit.**

You may upload any supporting documentation if needed.

 No files uploaded



## Q6

20 Points

Mark each of the following statements as TRUE or FALSE.

### Q6.1

2 Points

Because we can make causal statements with experimental studies, they are always preferred to observational studies.

- ☐ TRUE
- ☐ FALSE

### Q6.2

2 Points

If a study is reproducible, then all raw data and analysis code should be available such that independent researchers can exactly reproduce the published results.

- ☐ TRUE
- ☐ FALSE

**Q6.3**

2 Points

Bar charts are an effective data visualization tool for examining continuous numerical data.

- ☐ TRUE
- ☐ FALSE

**Q6.4**

2 Points

The median is less sensitive to outliers than the mean.

- ☐ TRUE
- ☐ FALSE

**Q6.5**

2 Points

If a medical diagnostic test has a specificity of 1, then it is impossible to have a false positive using this test.

- ☐ TRUE
- ☐ FALSE

**Q6.6**

2 Points

If a medical diagnostic test has a specificity of 1, then the positive predictive value is guaranteed to be 1.

- ☐ TRUE
- ☐ FALSE

**Q6.7**

2 Points

For a fixed prevalence, a test that has higher sensitivity will always have higher positive predictive value than a test that has a lower sensitivity.

- ☐ TRUE
- ☐ FALSE

**Q6.8**

2 Points

If a numeric count variable has the same mean and variance, then it will always follow the Poisson distribution.

- ☐ TRUE
- ☐ FALSE

**Q6.9**

2 Points

A Bernoulli random variable is simply a binomial random variable where  $n = 1$ .

- ☐ TRUE
- ☐ FALSE

**Q6.10**

2 Points

It is impossible for the density of a random variable to be above 1.

- ☐ TRUE
- ☐ FALSE