## GROUND RULES:

- **Print** your name clearly at the top of this page.
- This is a closed-book and closed-notes exam. You can not use external notes of any kind. You may use a calculator.
- This exam contains **two parts**:
  - Part 1. Multiple Choice. 22 questions, 1 point each (22 points total)
  - Part 2. Short Answer. 4 questions, 7 points each (28 points total).

  This exam is worth 50 points.
- Any discussion or inappropriate communication between you and another examinee, as well as the appearance of any unnecessary material, will result in a very bad outcome for you (it will be very bad).
- 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.*

**MULTIPLE CHOICE.** Circle the best answer. Make sure your answer is clearly marked. Ambiguous responses will be marked wrong.

1. The distribution of household incomes in South Carolina is highly skewed right, and the median income is $42,367. Which statement is **true**?

(a) The histogram of all South Carolina household incomes has a "balance point" at $42,367.
(b) Fifty percent of all South Carolina household incomes are greater than $42,367.
(c) The mean South Carolina household income is less than $42,367.
(d) If we added up all of the South Carolina household incomes and divided by the number of households, we would get $42,367.

2. In class, we talked about Carl Friedrich Gauss, who is credited with having invented the normal distribution in the early 1800s. In **what application** did Gauss discover this population density curve?

(a) He was characterizing the errors made by astronomers in measuring distances to celestial bodies.
(b) He was describing the large population growth in Europe in the late 1700s.
(c) He was describing the recidivism rates of criminals in Australia.
(d) He was using probability to model the sales of German tea and coffee.

3. In class, we discussed a study which found a strong negative correlation between

$$\begin{aligned} x &= \text{number of lemons imported from Mexico} \\ y &= \text{highway fatality rate in the United States.} \end{aligned}$$

Which statement is **true**?

(a) Importing more lemons from Mexico causes drivers in the United States to have fewer fatal accidents.
(b) The least squares regression line has $r^2$ close to 0.
(c) As the number of lemons imported from Mexico increases, the highway fatality rate in the United States tends to decrease.
(d) All of the above.

4. True or False. The area under any population density curve is equal to 1.

(a) True
(b) False

5. Which numerical value is **not** part of the 5-number summary?

(a) mean
(b) median
(c) minimum
(d) maximum

6. The length of human pregnancies from conception to birth follows a normal distribution with mean $\mu = 266$ days and standard deviation $\sigma = 16$ days. If the next baby born at Richland Palmetto Hospital has a **standard score** of $-2.5$, what does this mean?

(a) The length of this pregnancy is greater than 2.5% of the human pregnancies in the population.
(b) The length of this pregnancy is less than 2.5% of the human pregnancies in the population.
(c) The length of this pregnancy is 2.5 standard deviations above the mean.
(d) The length of this pregnancy is 2.5 standard deviations below the mean.

7. Regarding the correlation $r$, which of the following statements does **not** contain an error?

(a) "There was a weak correlation ($r = -0.10$) between the tire lifetime (in miles) and the average speed traveled (in miles/hour)."
(b) "We found a very high correlation ($r = 1.09$) between the horsepower of a car and the gas mileage of the car."
(c) "The correlation between the weight of the car and the gas mileage of the car was found to be $r = 0.53$ miles per gallon."
(d) "There is a weak to moderate correlation between the manufacturer of a car and the gas mileage of the car."

8. We used the **least-squares criterion** to calculate the regression line that "best fits" the data. What does this criterion say?

(a) The slope of the regression line will be chosen so that the line goes through as many of the points as possible.
(b) The line will be chosen to make $r^2$ as large as possible.
(c) The horizontal deviations that result from comparing the points to the line will all be equal to 0.
(d) None of the above.

9. We talked in class about how the statistics $\overline{x}$ and $s$ could be calculated from a sample of data but that the parameters $\mu$ and $\sigma$ could **not** be calculated. Why is this?

(a) The values of $\mu$ and $\sigma$ do not describe center and spread when we talk about populations.
(b) The parameters $\mu$ and $\sigma$ are biased for samples. We should use the median $M$ and IQR instead.
(c) We cannot calculate the parameters $\mu$ and $\sigma$ unless we know the sampling design.
(d) The parameters $\mu$ and $\sigma$ describe population-level characteristics. We cannot calculate them unless we see every individual in the population.

10. An agronomist records

$$\begin{aligned} x &= \text{amount of precipitation (in inches)} \\ y &= \text{yield (in bushels/acre)} \end{aligned}$$

for a sample of plots of land. Despite the two variables showing a strong curved relationship in a scatterplot, the correlation is only $r = 0.03$. He's confused that the correlation is so small. What would you tell him?

(a) "Switch the variables. You will get a stronger correlation."
(b) "The correlation does not describe curved relationships."
(c) "If there is a strong relationship, then you probably calculated $r$ incorrectly."
(d) "You need to square the correlation to see how strong the relationship is."

11. On the use of statistics and establishing causation, which statement is **correct**?

(a) A strong relationship between two variables does not mean that there is a causal link between them.
(b) The relationship between two variables can be influenced by other variables lurking in the background.
(c) The best evidence for causation comes from randomized comparative experiments−not from observational studies.
(d) Each statement above is correct.

12. Jeannie's score of 650 on the SAT mathematics exam translated to her being in the **93rd percentile** among all students who took the SAT. What does this mean?

(a) Taking into account variability, Jeannie's true SAT mathematics score is about 605.
(b) She scored better than 93% of all students who took the SAT mathematics exam.
(c) She is 93 standard deviations above the mean SAT mathematics score.
(d) Jeannie scored below the median SAT mathematics score.

13. In an observational study, researchers find a correlation of $r = -0.96$ between two of the variables studied. Which statement is **true**?

(a) There is likely a cause-and-effect relationship between the variables.
(b) The square of the correlation will also be negative.
(c) These two variables have a very strong linear relationship.
(d) The correlation does not have any meaning here; correlation is only useful in randomized comparative experiments.

14. The mean score for an exam in a large undergraduate statistics class is 74 and the standard deviation is 8. Suppose you make a 70 on the exam. What is your **standard score**?

(a) 1.5
(b) $-1.5$
(c) $-0.5$
(d) None of the above

15. Weather scientists in Mongolia collected annual rainfall amounts (mm) and maximum daily temperatures (deg C) from regions throughout the country. They then calculated the least-squares regression equation relating rainfall amount ($y$) to temperature ($x$) to be

$$\text{Rainfall} = 295 - 16 \times \text{Temperature}.$$

Using the regression equation, what would you **predict** the annual rainfall amount to be for a region whose maximum daily temperature was 10 deg C?

(a) 27.9 mm
(b) 135.0 mm
(c) 174.4 mm
(d) None of the above.

16. Because the correlation

$$r = \frac{1}{n-1} \sum \left( \frac{x - \overline{x}}{s_x} \right) \left( \frac{y - \overline{y}}{s_y} \right)$$

uses the standard scores for the observations (the quantities within the parentheses), which statement is **true**?

(a) The value of $r$ does not change when we change the units of measurement.
(b) The correlation can exceed 1 if there are extreme outliers in the data set.
(c) The square of the correlation can exceed 100%.
(d) All of the above.

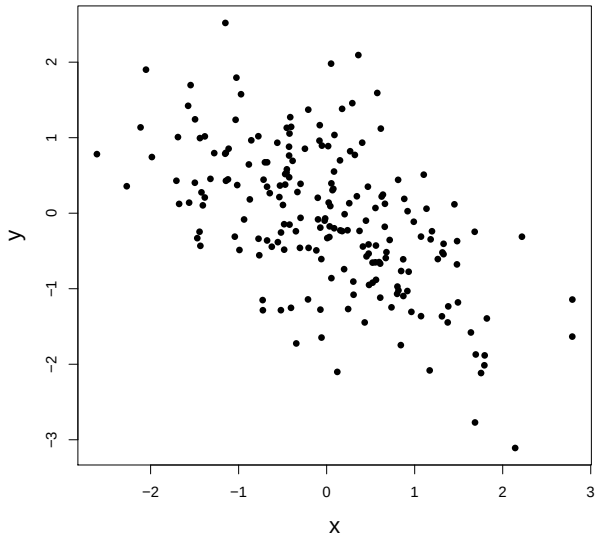
17. Consumer Reports recorded the number of calories in 20 brands of beef hot dogs, 17 brands of meat hot dogs, and 17 brands of poultry hot dogs. Here are the summaries:

| Type | Number of brands | Mean | Standard deviation |
|---|---|---|---|
| Beef | $n = 20$ | $\overline{x} = 156$ | $s = 22$ |
| Meat | $n = 17$ | $\overline{x} = 159$ | $s = 25$ |
| Poultry | $n = 17$ | $\overline{x} = 122$ | $s = 28$ |

The distribution of calories for each type of hot dog is approximately symmetric. Which type of hot dog has the largest amount of **variability** in its distribution?

(a) beef
(b) meat
(c) poultry

18. In the scatterplot below, the correlation $r$ is closest to which value?



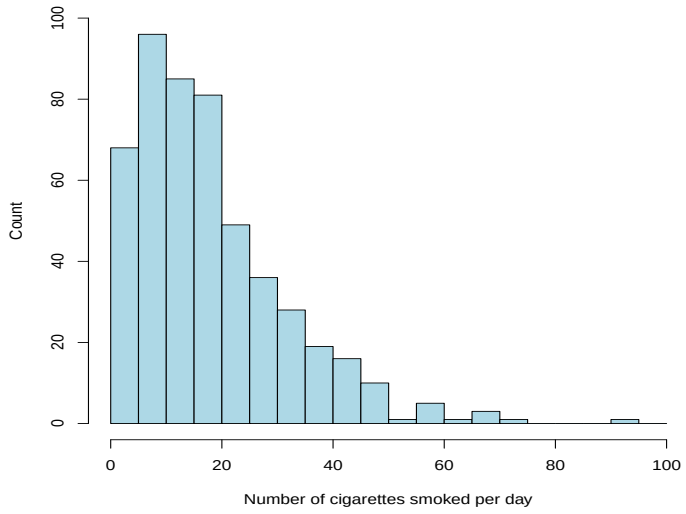

(a) $-0.99$
(b) 0.43
(c) 0.04
(d) $-0.62$

19. In a regression analysis, what is meant by the term "extrapolation?"

(a) It refers to the extra standard deviation used to estimate the margin of error.
(b) It refers to the numerical process R uses to calculate the least-squares regression line.
(c) It refers to using the correlation to assess causal effects between variables.
(d) It refers to the process of predicting a response variable outside of the range of the data.

Use the information provided below to answer **Questions 20-22**.

As part of a public health study, researchers took a simple random sample (SRS) of $n = 500$ adult smokers in South Carolina in 2012. For each smoker in the sample, researchers recorded the number of cigarettes smoked per day. A histogram of the 500 observations is shown below:




20. The **shape** of the histogram above is best described as

(a) skewed right
(b) bimodal
(c) skewed left
(d) symmetric

21. Which statistics should we use to numerically summarize the **center** and **spread** of this distribution?

(a) mode (center) and range (spread)
(b) mean (center) and variance (spread)
(c) mean (center) and standard deviation (spread)
(d) median (center) and interquartile range (spread)

22. The **standard deviation** of the data above is $s = 15$. How is this interpreted?

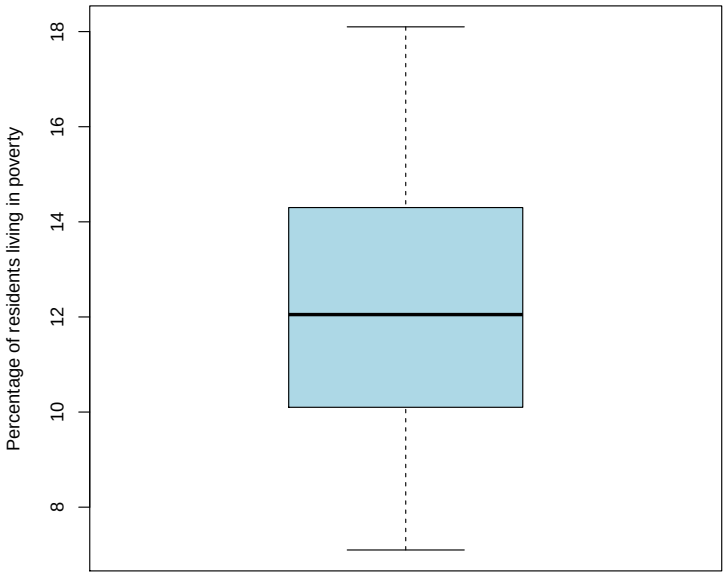
(a) Most SC smokers smoke more than 15 cigarettes per day.
(b) About 68 percent of the SC smokers are within 15 cigarettes from the median.
(c) This represents the average smoker's distance from the mean (in terms of the number of cigarettes smoked).
(d) The median is greater than the mean by about 15 cigarettes.

**SHORT ANSWER.** Give detailed responses showing all of your calculations. Please write clearly and legibly.

1. The table below gives the percentages of residents living below the poverty level in 26 states (all east of the Mississippi River):

| State | Percent | State | Percent | State | Percent |
|---|---|---|---|---|---|
| Alabama | 16.9 | Connecticut | 7.9 | Delaware | 10.5 |
| Florida | 12.1 | Georgia | 14.3 | Illinois | 11.9 |
| Indiana | 12.3 | Kentucky | 17.3 | Maine | 12.0 |
| Maryland | 8.3 | Massachusetts | 9.9 | Michigan | 14.0 |
| Mississippi | 18.1 | New Hampshire | 7.1 | New Jersey | 8.6 |
| New York | 13.7 | North Carolina | 14.3 | Ohio | 13.1 |
| Pennsylvania | 11.6 | Rhode Island | 12.0 | South Carolina | 15.0 |
| Tennessee | 15.9 | Vermont | 10.1 | Virginia | 9.9 |
| West Virginia | 16.9 | Wisconsin | 10.8 | | |

Here is a boxplot for the percentages as well as the 5-number summary:




```
> quantile(poverty,type=2)
   0%   25%   50%   75%  100%
 7.10 10.10 12.05 14.30 18.10
```

**Three questions are on the next page.**

(a) Would you characterize the distribution of percentages as being approximately symmetric? Or would you say the distribution is strongly skewed to one side? Explain.

(b) **Based on your answer in part (a)**, how would the mean compare to the median? Would the mean be larger than the median? Smaller than? Or about equal to? You don't have to do any calculations here, but explain why you are correct.

(c) You will note that the boxplot on the last page did not show any outliers in it. Use the 1.5(IQR) rule to explain why. Show all of your calculations.

2. In a population of the herring *Pomolobus aestivalis*, the lengths of individual fish follow a normal distribution with mean $\mu = 55$ mm and standard deviation $\sigma = 5$ mm.

(a) Draw the normal population density curve in this example. Identify on the horizontal axis where the mean falls. **Neatness counts! Be precise!!**

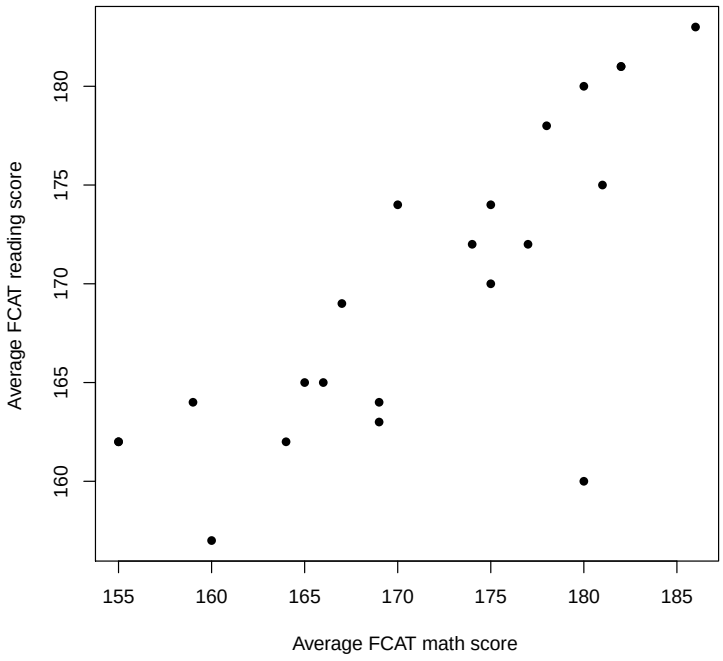
(b) Form intervals 1, 2, and 3 standard deviations from the mean. Interpret each interval by writing a complete sentence for each one.

(c) What percentage of fish in this population have lengths greater than 65 mm? Show your calculations; e.g., draw a picture. (Use the back of this page to show your work).

3. In class, we examined the relationship between poverty level and Florida Comprehensive Assessment Test (FCAT) reading scores for a simple random sample of $n = 22$ Florida elementary schools. Let's ignore poverty here and look at FCAT math scores instead. Specifically, let's examine

$$\begin{aligned} x &= \text{average FCAT math score} \\ y &= \text{average FCAT reading score} \end{aligned}$$

for this same simple random sample of $n = 22$ Florida elementary schools. Here is a scatterplot of the math and reading scores:




(a) Describe the relationship between math and reading scores for this sample of schools. Discuss the form of the relationship, the strength, the direction, and anything else that catches your eye.

(b) I used R to calculate the correlation between math and reading scores:

```
> cor(math,reading)
[1] 0.79
```

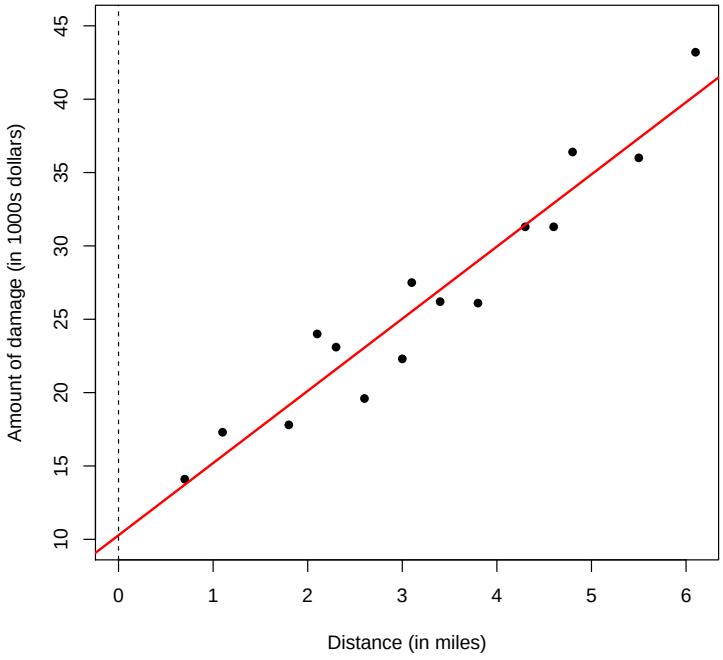
If we removed the school at $(180, 160)$, an obvious outlier, what would happen to the correlation? Would it increase, decrease, or stay about the same? Explain your reasoning.

(c) Citing the correlation above ($r = 0.79$), a new teacher claims that "if we improve our math scores, this will cause reading scores to increase too." How would you respond to this teacher?

4. An insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. Actuaries at the insurance company take a simple random sample of $n = 15$ residential houses with recent fires; on each house, they record

$$\begin{aligned} x &= \text{distance between house and nearest fire station (in miles)} \\ y &= \text{amount of damage (in 1000s of dollars).} \end{aligned}$$

Here is a scatterplot of the data with the least-squares regression line superimposed. A vertical line at $x = 0$ has been added.




I used R to calculate the least-squares regression line; here is the output:

```
> fit = lm(damage~distance)
> fit
Coefficients:
(Intercept)   distance
      10.28       4.92
```

(a) Which variable is the explanatory variable and which is the response variable? Explain why it is not the other way around.

(b) The slope of the least-squares regression line is $b = 4.92$. Explain what this means in words.

(c) The $y$-intercept of the least-squares regression line, $a = 10.28$, doesn't really have any practical meaning in this application. Why not?

(d) The square of the correlation is $r^2 = 0.92$. Explain what this means in words.