GROUND RULES:

- **Print** your full name clearly at the top of this page. Use the name that appears on university records.
- This is a closed-book and closed-notes exam. You can not use external notes of any kind.
- You may use a calculator. You may not use your phone as a calculator.
- This exam contains three parts:
  - Part 1. Multiple Choice. 25 questions, 2 points each (**50 points** total)
  - Part 2. Short Answer. 1 question, 10 points each **(10 points** total)
  - Part 3. Extra Credit. 1 question (5 points total).

  This exam is worth **60 points** (but it is possible to get up to 65 points).
- Any discussion or inappropriate communication between you and another examinee, as well as the appearance of any unnecessary material, is not allowed. All violations will be reported to the Student Conduct and Academic Integrity Office immediately.
- You have **50 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.*

PART 1: MULTIPLE CHOICE. Circle the best answer. Make sure your answer is clearly marked. Ambiguous responses will be marked wrong.

1. For a sample of 100 healthy dogs, a veterinarian measured the glucose concentration in the right eye and also in the blood serum. The data below are the eye measurements as a percentage of the blood measurements.

```
> stem(glucose)

   6 | 69
   7 | 04
   7 | 77789
   8 | 00000122222333444
   8 | 55555556777788889999
   9 | 000011111111222233333444
   9 | 55555666677888889
  10 | 001111123
  10 | 59
  11 | 0
```

The stems above denote the tens digit and the leaves denote the units digit.

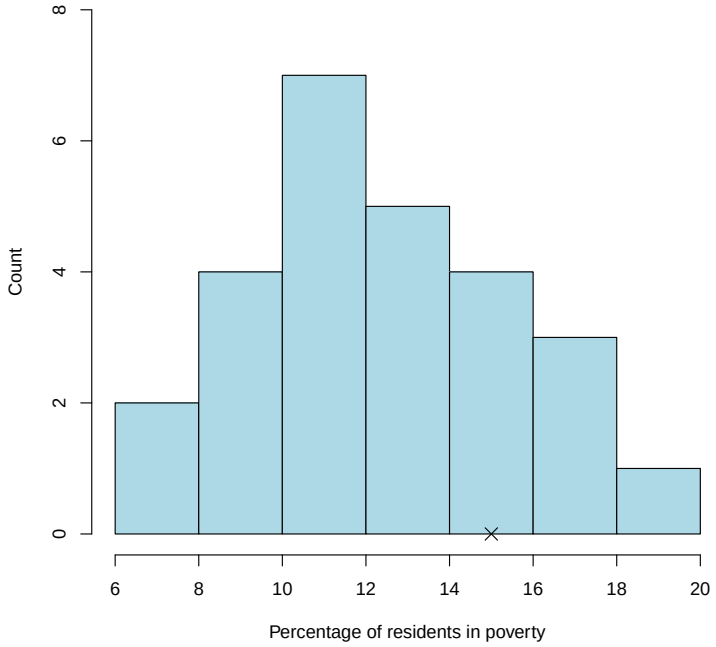
Glucose percentages at 100 or more identify dogs who are likely to become blind in the future. What percentage of dogs in the sample have glucose percentages that are larger than or equal to 100?

(a) 5 percent
(b) 3 percent
(c) 30 percent
(d) 12 percent

2. As a follow-up to Question 1, the veterinarian now thinks about how the results in the sample of 100 dogs might generalize to the entire population of healthy dogs. This is a question about

(a) statistical inference
(b) correlation
(c) least squares
(d) randomization

3. The histogram below shows the percentage of residents living below the poverty level for the 26 states east of the Mississippi River. For example, 15% of South Carolina's residents live below the poverty level (shown with an "×" below).



I used R to calculate the five-number summary for these data:

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

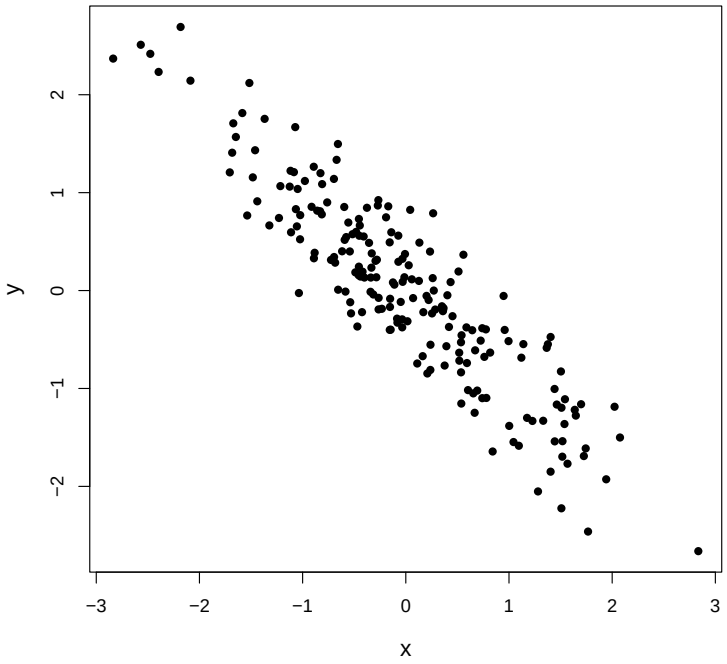
What is the **interquartile range**?

(a) 10.1
(b) 4.2
(c) 14.3
(d) 11.0

4. Refer to Question 3. What is the **percentile** associated with South Carolina in this distribution?

(a) 70th percentile
(b) 80th percentile
(c) 20th percentile
(d) 99th percentile

5. A scatterplot of $n = 200$ observations is below:



The correlation r is closest to

(a) 0.99
(b) 0.73
(c) −0.91
(d) −0.06

6. The number of home runs Barry Bonds hit in his 22 major league seasons are

16	25	24	19	33	25	34	46	37	33	42
40	37	34	49	73	46	45	45	5	26	28

I used R to find the mean and standard deviation of these 22 observations:

```
> mean(homeruns) # mean
[1] 34.6
> sd(homeruns) # standard deviation
[1] 14.0
```

How do we interpret the **standard deviation**?

(a) It is an average of the 22 observation distances from the mean.
(b) It counts the number of outliers as determined by our 1.5(IQR) rule.
(c) It is the largest of all standard scores.
(d) It is how wide the middle 50% range is under a population density curve.

7. According to the World Health Organization, the systolic blood pressure (SBP, mm Hg) of all American women is normally distributed with mean $\mu = 120$ and standard deviation $\sigma = 15$. What percentage of all American women will have a SBP within **two** standard deviations of the mean SBP?

(a) 81.5%
(b) 32%
(c) 95%
(d) 68%

8. Which value of the correlation below indicates the **strongest** linear relationship between two variables?

(a) $r = 0.26$
(b) $r = 0.90$
(c) $r = -0.99$
(d) $r = -0.03$

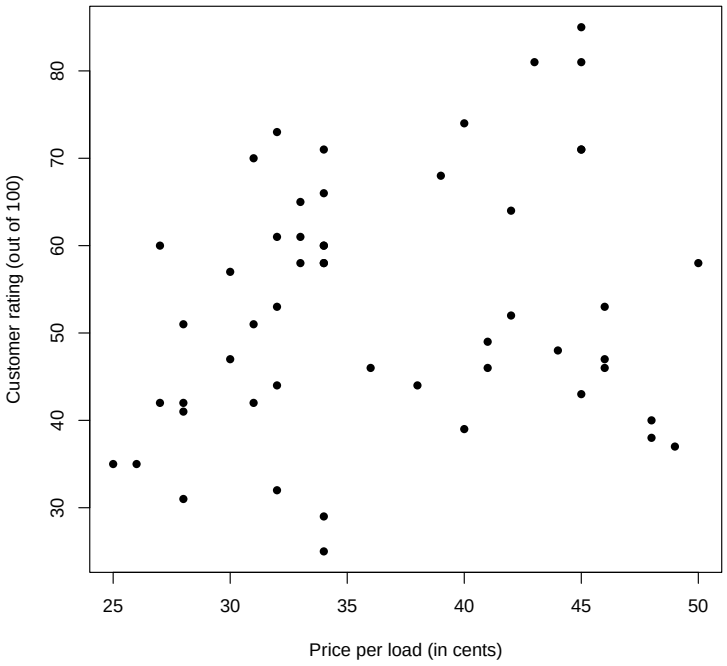
9. The HDL cholesterol level (mg/dl) of all American males is described by a normal distribution with mean $\mu = 48$ and standard deviation $\sigma = 12$. An individual in this population has a HDL cholesterol level of 30 mg/dl. What is his **standard score**?

(a) -1.5
(b) 93.32%
(c) 6.68%
(d) 1.5

10. The distribution of all household incomes in South Carolina is skewed right, and the median household income is \$67,804. Which statement is true?

(a) The mean South Carolina household income is less than \$67,804.
(b) The histogram of all South Carolina household incomes has a "balance point" at \$67,804.
(c) Fifty percent of all South Carolina household incomes are greater than \$67,804.
(d) The distribution of all South Carolina household incomes follows the 68-95-99.7% rule.

11. A marketing researcher wants to determine if a laundry detergent's customer satisfaction rating is related to the price of the detergent. He records these variables for 52 brands of laundry detergent. Here is a scatterplot of the 52 observations:



The correlation is $r = 0.21$. What **units** are attached to this number?

(a) rating score (out of 100)
(b) cents
(c) cents per rating score (out of 100)
(d) none of the above

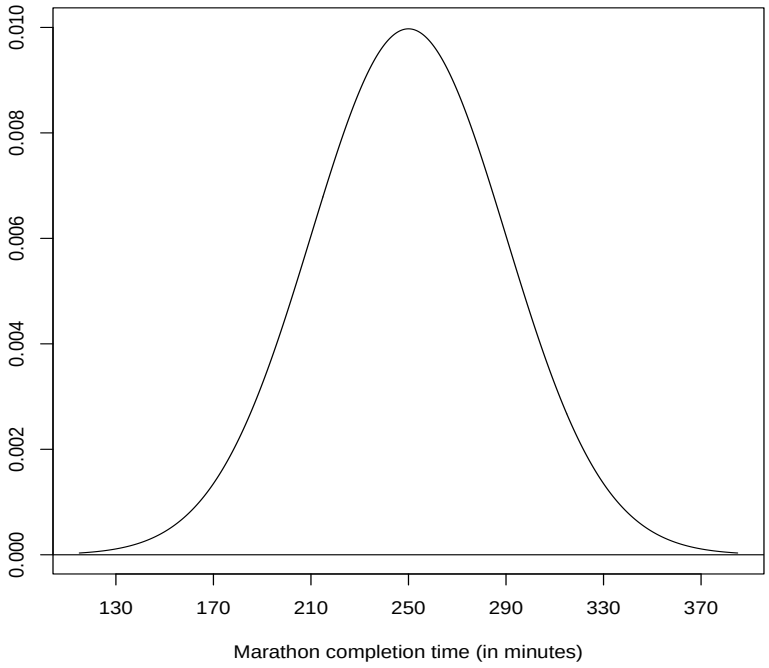
12. Refer to Question 11. The square of the correlation is

$$r^2 = (0.21)^2 \approx 0.04.$$

How do we interpret this?

(a) Four percent of the observations will be within the margin of error of the least-squares regression line.
(b) Four percent of the least-squares predictions will be unbiased.
(c) Four percent of the relationship between the two variables is linear and the remaining 96% is curved.
(d) Four percent of the variation in the customer rating data is explained by the straight-line relationship with the price per load.

13. For a population of runners, the time it takes to complete the New York Marathon (in minutes) is normally distributed with mean $\mu = 250$ and standard deviation $\sigma = 40$. This population density curve is shown below:



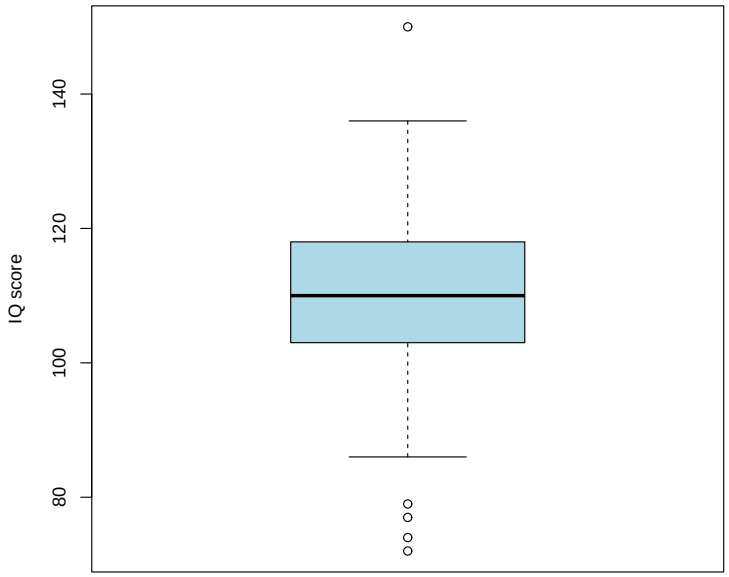
What percentage of runners in this population will finish the marathon in shorter than (**less than**) 230 minutes? Round your answer to the nearest whole percent.

(a) 31%
(b) 16%
(c) 8%
(d) 69%

14. What does the **correlation** between two variables measure?

(a) the degree to which two variables have a causal relationship
(b) the proportion of standard scores that are positive
(c) the percentage of observations that are within one standard deviation of the mean in either direction
(d) the strength and direction of the straight-line relationship between the variables

15. A graduate student in the College of Education observes a sample of $n = 78$ seventh grade students from Columbia area schools and measures the IQ of each student. Below is a boxplot of the IQ scores for the sample. Five outliers have been identified.



The **median** IQ score is closest to

(a) 85
(b) 135
(c) 110
(d) 15

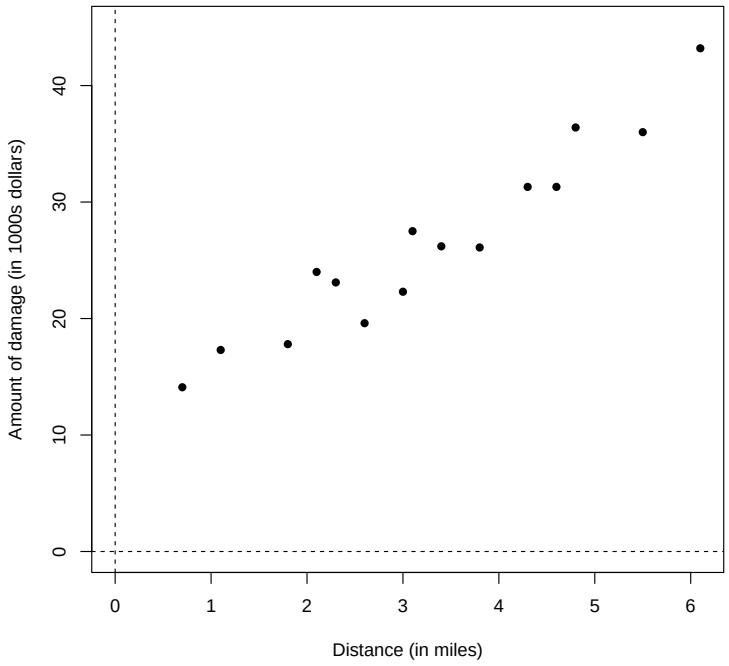
16. Scores are released for a major exam and you scored a 73.5% with a standard score of $z = 2.5$. Which statement below is true?

(a) Your score is 2.5 standard deviations above the mean.
(b) Your true score (after accounting for measurement error) is somewhere between 71 and 76.
(c) The highest score on the exam was 81%.
(d) You did better than 2.5% of all students who took the exam.

17. An insurance company wants to relate the amount of fire damage in residential fires to the distance between the burning house and the nearest fire station. Actuaries record

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

for a sample of $n = 15$ houses. Here is a scatterplot for the data collected in the sample. A vertical line at $x = 0$ has been added.



Which statement is true about the equation of the least-squares regression line?

(a) The slope b is negative and the intercept a is positive.
(b) The slope b and the intercept a are both negative.
(c) The slope b is positive and the intercept a is negative.
(d) The slope b and the intercept a are both positive.

18. Refer to Question 17. Which variable is the **response variable**?

(a) amount of damage
(b) distance between house and nearest fire station

19. We know the total area under any population density curve is 1 (or 100% as a percentage). What is a conceptual interpretation of this?

(a) A population density curve describes the distribution for all individuals in the population.
(b) A population density curve has total area equal to 1 only when the mean equals the median.
(c) A population density curve tells us if two variables have a perfect straight-line relationship.
(d) A population density curve only includes those individuals within the margin of random sampling error.

20. 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	$n = 17$	$\overline{x} = 122$	$s = 28$

Which of the three calorie distributions has the largest variability?

(a) Poultry
(b) Beef
(c) Meat

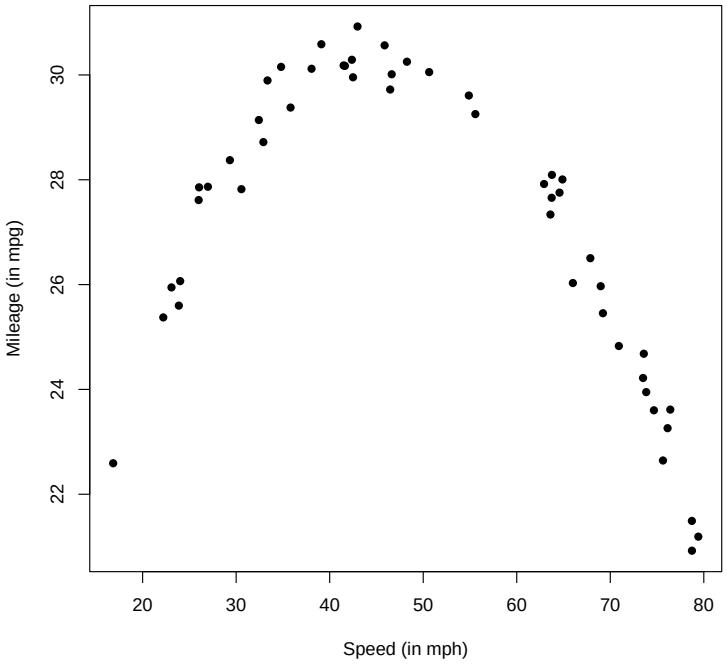
21. In a scatterplot of observations, we defined the "best-fit line" for regression as the one that makes the

(a) sum of the vertical distances of the data points from the line as large as possible.
(b) sum of the squared vertical distances of the data points from the line as large as possible.
(c) sum of the vertical distances of the data points from the line as small as possible.
(d) sum of the squared vertical distances of the data points from the line as small as possible.

22. An observational study was done to examine the relationship between gas mileage and vehicle speed. A sample of $n = 50$ sports utility vehicles was used with the following variables recorded on each one:

$$\begin{aligned} x &= \text{speed (in miles per hour)} \\ y &= \text{gas mileage (in miles per gallon).} \end{aligned}$$

A scatterplot of the observations is below:



Which statement is true?

(a) The correlation r has no meaning here because this was an observational study−not a randomized comparative experiment.
(b) The correlation r is not useful in this study because the variables have a curved relationship.
(c) The correlation r does not make sense in this study because both variables are categorical.
(d) The correlation r is close to 1 because the two variables (mileage and speed) are strongly related.

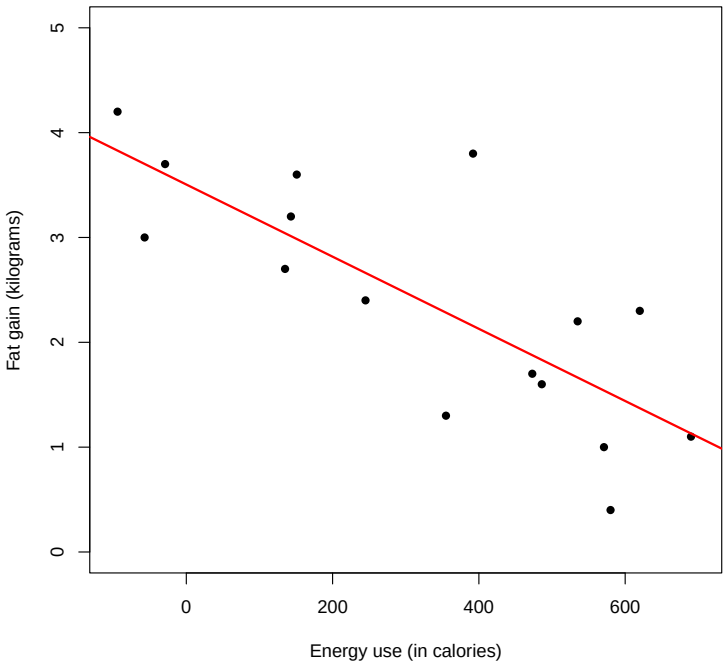
23. In an obesity study, researchers deliberately overfed 16 young healthy adults for 8 weeks. Afterwards, they measured the following variables on each adult:

x = amount of energy used in non-exercise activities (in calories)
y = amount of fat gain (in kilograms).

I used R to calculate the equation of the least-squares regression line:

$$y = 3.5 - 0.0035x.$$

Below is a scatterplot of the data with the least-squares regression line superimposed.



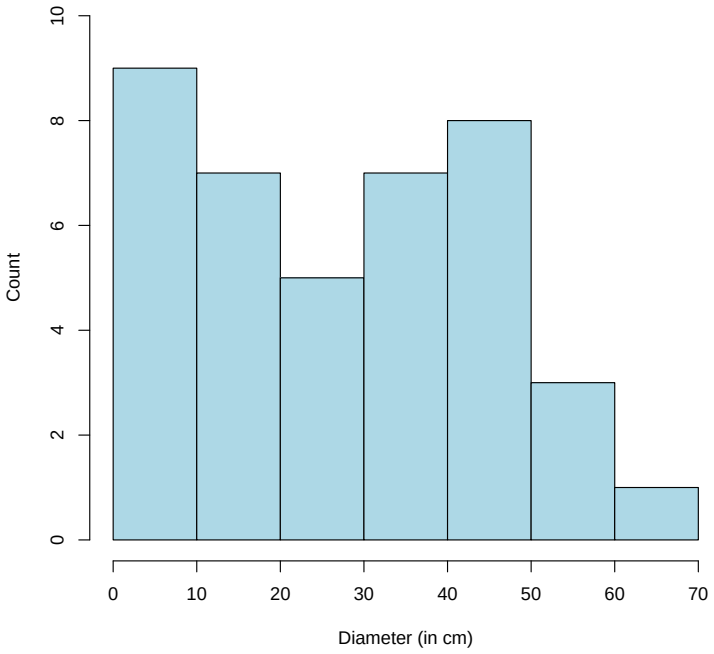
What would you predict the fat gain to be for a young healthy adult who burns 400 calories during non-exercise activities?

(a) 3.5 kilograms
(b) 2.1 kilograms
(c) 3.8 kilograms
(d) 1.4 kilograms

24. Which number is **not** part of the five-number summary?

(a) mean
(b) minimum
(c) median
(d) first quartile

25. The Wade Tract in Thomas County, Georgia, is an old growth forest which has 584 longleaf pine trees. An ecologist observes a sample of $n = 40$ longleaf pine trees from the Wade Tract forest. She measures the diameter of each tree (in cm) and constructs the histogram below with the 40 observations:



In this example, a **population density curve** would describe the diameter distribution for which group of individuals?

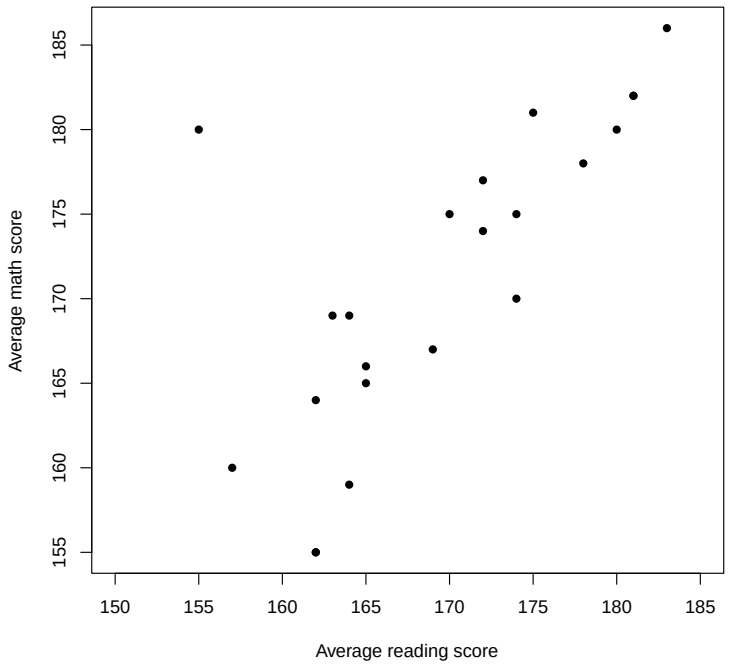
(a) all trees in the southeastern United States
(b) all trees in Thomas County, Georgia
(c) all 584 longleaf pine trees in the Wade Tract forest
(d) the 40 longleaf pine trees in the sample

PART 2: SHORT ANSWER. Give a detailed response. Please write clearly.

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

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

for this same sample of 22 schools. Here is a scatterplot of the math and reading scores:



(a) Describe the relationship between the average 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.

Answer: Our 4 characteristics are:

- Form: There is a linear (straight-line) relationship between average math score and the average reading score.
- Strength: The linear relationship is strong.
- Direction: The relationship is positive.
- Deviations: There is one outlier in the upper left corner of the scatterplot.

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

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

If we removed the school at $(155, 180)$, an obvious outlier, what would happen to the correlation? Would it increase, decrease, or stay about the same? Explain your reasoning.

Answer: The correlation would increase. If this observation was removed, the linear relationship between average math score and average reading score would strengthen considerably. The outlier does not adhere to the positive linear relationship between these two variables, so it is bringing the correlation down.

(c) A candidate in an upcoming Miami school board election claims,

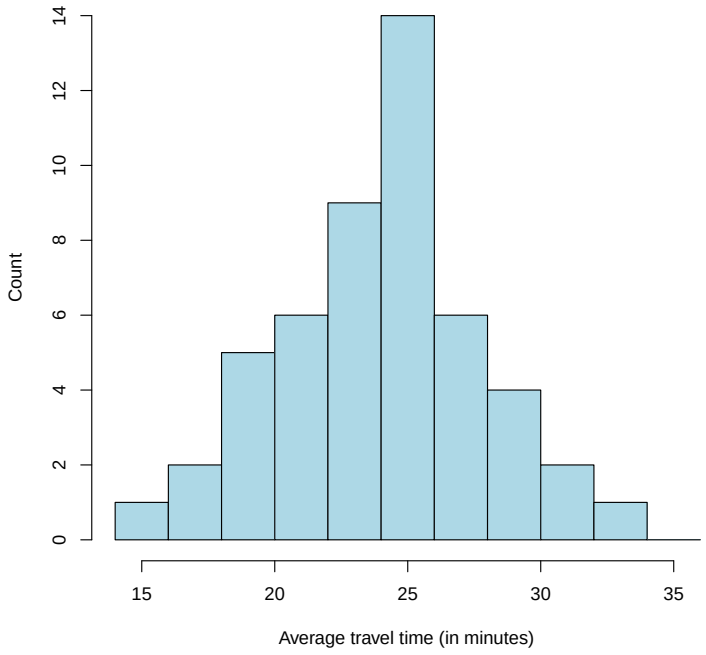
> *"We know that math and reading scores for our elementary schools in Florida are strongly correlated. Therefore, if we improve our math scores, this will cause reading scores to increase too."*

How would you respond to this? Talk about how the candidate may be misinterpreting the correlation.

Answer: Correlation does not imply causation! Both average math score and average reading score are related to other variables, for example, the quality of the school/teachers, the percentage of the students below the poverty level, etc. It is the common relationship math scores and reading scores have to these other variables that are making math and reading scores correlated themselves.

PART 3: EXTRA CREDIT. Give a detailed response. Please write clearly.

How long does it take to commute to work? The histogram below shows the average travel time to work (in minutes) for each of the 50 states. That is, there are 50 average times depicted below−one for each state. For example, South Carolina's average travel time to work is 24.1 minutes.



(a) Describe the shape of this distribution. Is it symmetric or skewed?
(b) In the light of your answer in part (a), which numerical summary would you use to describe the center of this distribution? the variability in this distribution?
(c) Does it make sense to think about a population density curve in this example? Explain.

Answers:
(a) The distribution is symmetric.
(b) For symmetric distributions, it is most common to use the mean as a measure of "center." It is most common to use the standard deviation as a measure of "variability."
(c) No, it doesn't. These observations are state averages for the 50 states. These observations are not representative of a larger population of states.

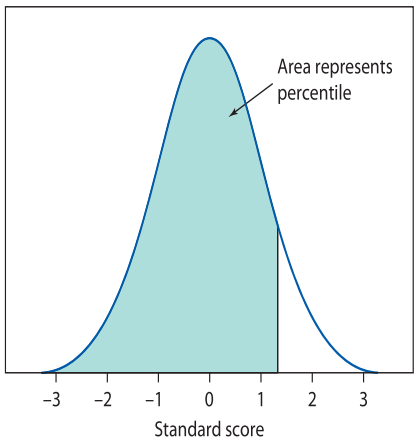


Table B Percentiles of the Normal distributions

Standard score → Percentile		Standard score → Percentile		Standard score → Percentile	
−3.4	0.03	−1.1	13.57	1.2	88.49
−3.3	0.05	−1.0	15.87	1.3	90.32
−3.2	0.07	−0.9	18.41	1.4	91.92
−3.1	0.10	−0.8	21.19	1.5	93.32
−3.0	0.13	−0.7	24.20	1.6	94.52
−2.9	0.19	−0.6	27.42	1.7	95.54
−2.8	0.26	−0.5	30.85	1.8	96.41
−2.7	0.35	−0.4	34.46	1.9	97.13
−2.6	0.47	−0.3	38.21	2.0	97.73
−2.5	0.62	−0.2	42.07	2.1	98.21
−2.4	0.82	−0.1	46.02	2.2	98.61
−2.3	1.07	0.0	50.00	2.3	98.93
−2.2	1.39	0.1	53.98	2.4	99.18
−2.1	1.79	0.2	57.93	2.5	99.38
−2.0	2.27	0.3	61.79	2.6	99.53
−1.9	2.87	0.4	65.54	2.7	99.65
−1.8	3.59	0.5	69.15	2.8	99.74
−1.7	4.46	0.6	72.58	2.9	99.81
−1.6	5.48	0.7	75.80	3.0	99.87
−1.5	6.68	0.8	78.81	3.1	99.90
−1.4	8.08	0.9	81.59	3.2	99.93
−1.3	9.68	1.0	84.13	3.3	99.95
−1.2	11.51	1.1	86.43	3.4	99.97