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. 48 questions, 1 point each (48 points total)
  - Part 2. Short Answer. 5 questions, 8 points each (40 points total).

  This exam is worth 88 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 **2.5 hours** 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.*

HELPFUL FORMULAS

$$\textbf{margin of error} = \frac{1}{\sqrt{n}}$$

Measured value = True value + Bias + Random error.

$$\text{percentage change} = \frac{\text{amount of change}}{\text{starting value}} \times 100\%.$$

$$\overline{x} = \frac{1}{n}\sum x \qquad s = \sqrt{\frac{1}{n-1}\sum (x-\overline{x})^2}$$

$$z = \frac{x-\mu}{\sigma}$$

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

$\widehat{p}$ is (approximately) normal with mean p and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$

$$\widehat{p} \pm z^* \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

C	80%	90%	95%	99%
z^*	1.28	1.64	1.96	2.58

$\overline{x}$ is (approximately) normal with mean μ and standard deviation $\dfrac{\sigma}{\sqrt{n}}$

$$\overline{x} \pm z^* \left(\frac{s}{\sqrt{n}}\right)$$

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

1. The tuition and technology fees for in-state students at USC were \$11,454 during 2015-2016. For 2016-2017, they were \$11,854. The **percentage increase** between these two academic years is closest to

(a) 1.5%
(b) 2.5%
(c) 3.5%
(d) 4.5%

2. Long-term high systolic blood pressure (SBP) is an important factor in predicting heart disease in American males. What does this mean?

(a) Long-term high SBP has predictive validity.
(b) The correlation between long-term high SBP and heart disease is equal to 1.
(c) The probability an individual male with long-term high SBP gets heart disease is equal to 1.
(d) All of the above.

3. SAT mathematics exam scores are **normally distributed** with mean $\mu = 500$ and standard deviation $\sigma = 100$. What percentage of the scores will be less than 400?

(a) 2.5%
(b) 16%
(c) 32%
(d) 68%

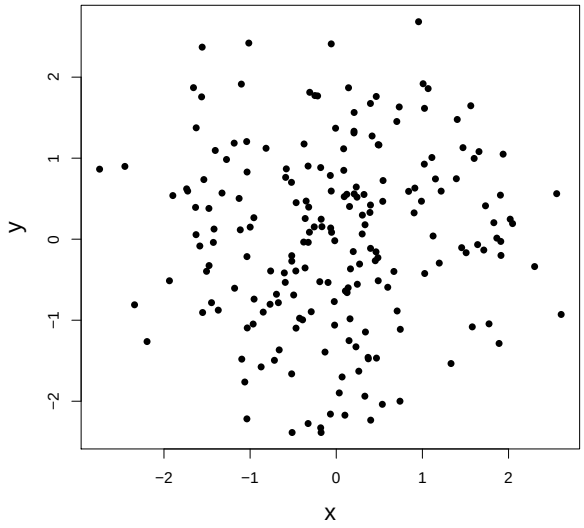
4. In class, we talked about a 1994 PhD dissertation whose author claimed that

> *"Every year since 1950, the number of American children gunned down has doubled."*

What was the **main point** of our discussion on this?

(a) Statistics show that gun violence is steadily getting worse in the United States.
(b) Some researchers make claims that are complete nonsense.
(c) Using quantitative numerical summaries can distort reality when you are measuring a categorical variable.
(d) We cannot understand the root cause of gun violence unless we use an experiment to control for the effects of lurking variables.

5. In the scatterplot below, the **correlation** r is closest to which value?



(a) -0.88
(b) -0.55
(c) 0.02
(d) 0.62

6. My colleague and his wife have 4 children and they are all boys. If they become pregnant again (and have a single birth), the **probability** they will have a 5th boy is closest to which value?

(a) 0.03
(b) 0.2
(c) 0.5
(d) 0.97

7. A university has 20,000 undergraduate students and 10,000 graduate students. A sample survey of student opinion on health care services first selects 200 undergraduate students at random and then 100 graduate students at random. This is a

(a) stratified sample
(b) simple random sample
(c) cluster sample
(d) convenience sample

8. True or False. The **margin of error** in a confidence statement only accounts for random sampling errors. It does not include non-sampling errors.

(a) True
(b) False

9. In an observational study or experiment involving human subjects, what is the primary responsibility of an **institutional review board**?

(a) To ensure that sampling is representative of the population of individuals
(b) To ensure that the statistical results can eventually be published
(c) To ensure that the results are statistically significant
(d) To ensure that the participants are protected from harm

10. I tracked the number of unique hits to my professional web site during a five-day period:

```
20   36   16   20   28
```

The **standard deviation** is closest to

(a) 2
(b) 8
(c) 32
(d) 128

11. What is one way to remove the effects of **lurking variables** when designing a comparative experiment?

(a) Use enough subjects to eliminate confounding.
(b) Weight the influence of each subject differently by transforming the response variable.
(c) Use blocking to separate out the effects of these variables beforehand.
(d) Calculate statistical significance measures by averaging over these variables.

12. A recent article I saw online contained this statement:

> *"A sample survey in the United States revealed that 50% of obese people earn less than the national median income."*

What **informed conclusion** can be made from this statement?

(a) Obese people probably have a harder time finding jobs, so the jobs they do get pay less.
(b) Our state and federal governments can do more to end employment discrimination on the basis of obesity status.
(c) This sample's results suggest that obese Americans are underpaid when compared to the rest of the population.
(d) None of the above.

13. In class, we talked about this main result:

> **Result:** Take a SRS of size n from a large population of individuals, where p denotes the population proportion. Let $\widehat{p}$ denote the sample proportion. For large samples (i.e., for large n),
>
> - the sampling distribution of $\widehat{p}$ is represented by a normal density curve
> - the **mean** of the sampling distribution is p
> - the **standard deviation** of the sampling distribution is
>
> $$\sqrt{\frac{p(1-p)}{n}}.$$

What is meant by the term **sampling distribution**?

(a) This distribution gives the probability that $\widehat{p}$ will be equal to p.
(b) This distribution summarizes how we assign subjective probabilities to samples.
(c) This distribution describes how $\widehat{p}$ varies in repeated sampling from a population.
(d) This distribution quantifies the non-sampling errors associated with the population proportion.

14. Refer to Question 13. What does the **second bullet** (about the mean) say about the sample proportion $\widehat{p}$ as an estimate of the population proportion?

(a) It is equal to the population proportion.
(b) It is unbiased.
(c) It is reliable.
(d) It is within the margin of error.

15. True or False. An event with probability 0 can never occur.

(a) True
(b) False

16. In one experiment dealing with contracting poison ivy, 13 out of 13 healthy subjects receiving a liquid solution developed a rash−despite the fact that the solution was completely harmless. This is an example of what?

(a) blinding
(b) non-adherence
(c) the Hawthorne principle
(d) the placebo effect

17. In class recently, we talked about the role that **actuaries** play in insurance companies. What is their role?

(a) To market insurance products to individuals who may experience innumeracy.
(b) To calculate the risk associated with insuring individuals.
(c) To interact with law enforcement following an accident to (hopefully) minimize the company's payout.
(d) To use statistical and mathematical models to explain weather patterns.

18. Samples taken from shopping malls are fast and cheap, but people contacted at shopping malls are not representative of the entire US population. Such a sampling design is likely to suffer from

(a) blocking
(b) blinding
(c) undercoverage
(d) non-replication

19. I read the following survey question recently online:

> *"Do you agree with the following statement: The federal government is too large and has gotten totally out of control and threatens our basic liberties. Should we or should we not clear house and commit to drastic change?"*

Using this question in a survey will likely lead to bias caused by a serious

(a) reliability error
(b) random sampling error
(c) non-sampling error
(d) margin of error

20. What is a **clinical trial**?

(a) It is a study that matches an individual's response with a response from another individual with the same explanatory variables.
(b) It is a random process that when performed over time explains where probability comes from.
(c) It is a controlled observational study where scientists observe the effects of measurement error in a laboratory setting.
(d) It is an experiment that examines the effect of new medical treatments in human subjects.

21. Any experiment or observational study that uses human subjects requires **informed consent**. What does this mean?

(a) The authors of the study must inform the institutional review board about the study and obtain the board's permission to go ahead.
(b) The institutional review board must agree that the study will benefit science and that it will not harm the subjects.
(c) The nature of the study must be explained in advance to the subjects who then must voluntarily agree to take part.
(d) The authors of the study must agree to inform the public of the study results.

22. In class, we discussed how the **randomized response** survey method could be used to ask the sensitive question:

Have you ever drink-spiked someone to take advantage of them sexually?

Does this survey method guarantee the confidentiality of its participants or anonymity of its participants?

(a) confidentiality
(b) anonymity

23. We want to write a **100%** confidence statement for the population of American adults. What must we do?

(a) Take the range in a 95% confidence statement and increase it by 5%.
(b) Take two random samples: one to make a 95% confidence statement and a second one to verify it.
(c) Eliminate all non-sampling errors.
(d) None of the above.

24. Dr. Joel Best, author of *Damned Lies and Statistics*, uses the term **innumeracy** to describe members of the media and people who consume information from the media. What is the best description of this term?

(a) Not knowing how to tell the difference between statistical significance and practical significance
(b) People generally get their information from the media outlets with whom they agree
(c) A lack of understanding of basic mathematical ideas and numbers
(d) Reporting observational study findings without discussing studies that contradict the findings

25. What are the **three basic principles** of experimental design?

(a) margin of error, confidence level, statistical inference
(b) realism, equipoise, and blinding
(c) randomization, replication, and control
(d) correlation, regression, and probability

26. When I weighed myself one morning during my recent trip to Kansas, the scale read 226.6. When I stepped on the scale again, it read 226.5. When I stepped on the scale one final time, it read 226.7. What do we know?

(a) It is not possible to determine if these measurements are biased.
(b) These three measurements do not have random error.
(c) The standard deviation of these three measurements is less than zero.
(d) None of the above.

27. A very rare trait is observed in newborn children. Geneticists estimate the probability of this trait to be 0.0003; i.e., 3 out of 10000. What is the **law of averages** interpretation of this probability?

(a) If we observed 9,997 consecutive children without the trait, the next 3 children should have the trait.
(b) After observing many births over the long run, the proportion of children with this trait will be close to 0.0003.
(c) We are 95% confident that 3 out of the first 10,000 children born will have the trait.
(d) The margin of error associated with the sample proportion of newborns with the trait is 0.0003.

28. In class, we discussed this excerpt taken from your textbook:

> *"There's a delicate balance between when to do or not do a randomized trial. On the one hand, there must be sufficient belief in the agent's potential to justify exposing half the subjects to it. On the other hand, there must be sufficient doubt about its efficacy to justify withholding it from the other half of subjects who might be assigned to placebos."*

What **concept** is being expressed in this quote?

(a) the Hawthorne principle
(b) the placebo effect
(c) control
(d) None of the above

29. Here is a **probability model** for blood types in the population of all Caucasians in the United States:

Outcome	A	B	AB	O
Proportion	0.40	0.10	0.05	0.45

A person is selected at random from this population. What is the probability that his/her blood type is A or B (but not AB)?

(a) 0.04
(b) 0.45
(c) 0.50
(d) This cannot be determined with the information provided.

30. During the 2015-2016 Republican primary, Donald Trump claimed that he received more votes than any other candidate in previous Republican primary contests. This claim is **true**. However, I was sharply critical of its intended meaning in class. Why?

(a) His statement dealt with only one quantitative variable. He should have used correlation with a relevant explanatory variable.
(b) He should have cited the margin of error in the number of votes he received.
(c) The size of the Republican electorate has changed substantially because of population growth over the years.
(d) The statement has no predictive validity; he could not have possibly known at the time he would win the general election.

31. Which is the best example of **statistical inference**?

(a) Weighing myself 10 times (back to back) in the morning.
(b) Writing a confidence interval for the population mean SAT mathematics exam score.
(c) Randomizing subjects to treatments within blocks in a randomized block experiment.
(d) Calculating the sample mean length and the sample standard deviation of the lengths from a simple random sample of turtle shells.

32. In class, we watched a short video featuring **Hans Rosling**, who was an internationally known expert in health and medical statistics. The highlight of the video was his use of visual displays to show data in multiple dimensions. What was he was talking about?

(a) the Tuskegee syphilis experiment involving sharecroppers
(b) increased radon and arsenic levels in third-world manufacturing settings
(c) measurements recorded in new experiments for breast cancer
(d) life expectancy and wealth over time for every country

33. The Bureau of Labor Statistics announced that last month it interviewed a sample of individuals in 60,000 households; **9.7%** of the people interviewed were unemployed or underemployed. The number "9.7%" is an example of a

(a) statistic
(b) parameter

34. In class recently, we discussed the **birthday problem**. What important lesson did we learn?

(a) Probabilities of things happening can disagree with our intuition.
(b) Probabilities can be 0 or 1 but only in extreme situations.
(c) There is a spurious correlation between the month that people are born and the day that people are born.
(d) Probabilities follow a normal sampling distribution.

35. Researchers want to compare the effectiveness of two ways to treat prostate cancer. The two treatments are traditional surgery (where the prostate is removed) and a newer method that does not require surgery. The researchers have 300 prostate cancer patients who have agreed to be subjects in the experiment. Could this experiment be carried out as a **matched pairs experiment**?

(a) No, a matched pairs experiment requires many more patients.
(b) Yes, each of the 150 subjects in the surgery group can be paired with someone from the non-surgery group.
(c) No, it is not possible to give both treatments to the same person.
(d) Yes, randomize each patient to one of the two treatment groups.

36. Here are some statements about the correlation:

A. The correlation is a number between 0 and 1.

B. The correlation between x and y is the same as the correlation between y and x.

C. The correlation does not have any units attached to it.

Which statements are **true**?

(a) A and B
(b) A and C
(c) B and C
(d) Each statement above (A, B, and C) is true.

37. At the 2017 Masters golf tournament in Augusta, GA, statisticians recorded (a) the time it took each golfer to complete his round and (b) the distance of the putts each golfer made. These are measurements on

(a) two categorical variables
(b) two quantitative variables
(c) one categorical variable and one quantitative variable

38. Researchers at the University of Tennessee sampled 565 adults from the state and recorded many variables on them, including the number of years of education, smoking status, and whether the individual suffered from any type of sleep disorder. What group of individuals best describes the **population** in this example?

(a) the 565 adults sampled
(b) all adults in Tennessee
(c) all adults in the United States
(d) all adults in the North America

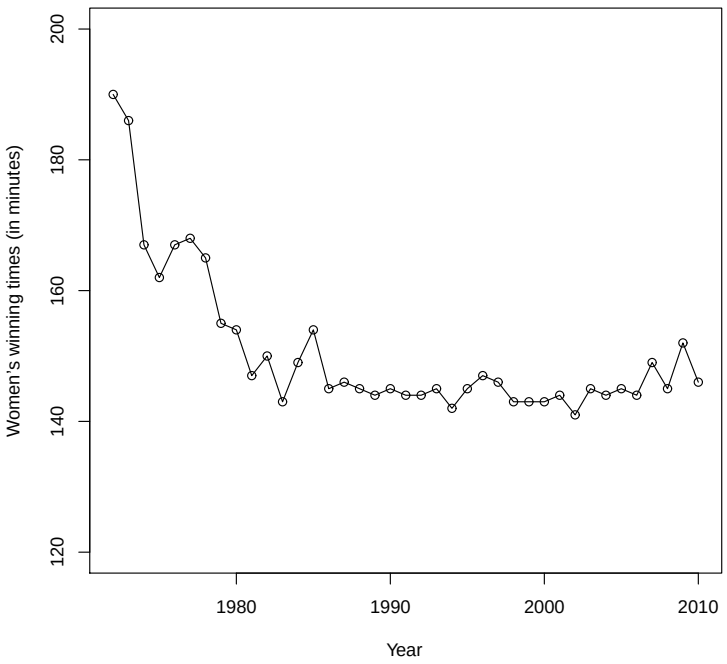
39. Does regular exercise reduce bone loss in post-menopausal women? A researcher finds 100 post-menopausal women (80 white; 20 non-white). She assigns white women to a regular exercise program. She assigns non-white women to a program that does not involve exercise. At the end of the study, she finds a statistically significant difference between the two groups. Her paper summarizing the experiment gets rejected from every journal. Why?

(a) The significant exercise results are completely confounded with the white/non-white race of the subjects.
(b) She did not use nearly enough subjects; 100 is too small.
(c) A different number of white and non-white subjects was used.
(d) This is an observational study, and it does not contain a control group.

40. The distribution of BMI measurements for fourth-grade children is strongly **skewed to the right** (high) side. How does the mean BMI compare with the median BMI?

(a) the mean BMI and the median BMI are about the same
(b) the mean BMI is larger than the median BMI
(c) the mean BMI is smaller than the median BMI

41. In a homework assignment, you examined the winning times for women in the Boston Marathon each year from 1972 to 2010:



What can we say from this graph?

(a) There is a seasonal pattern in the winning times.
(b) The correlation between winning times and year would be close to 1.
(c) The least squares regression line (using year as an explanatory variable) would have a negative y-intercept.
(d) None of the above.

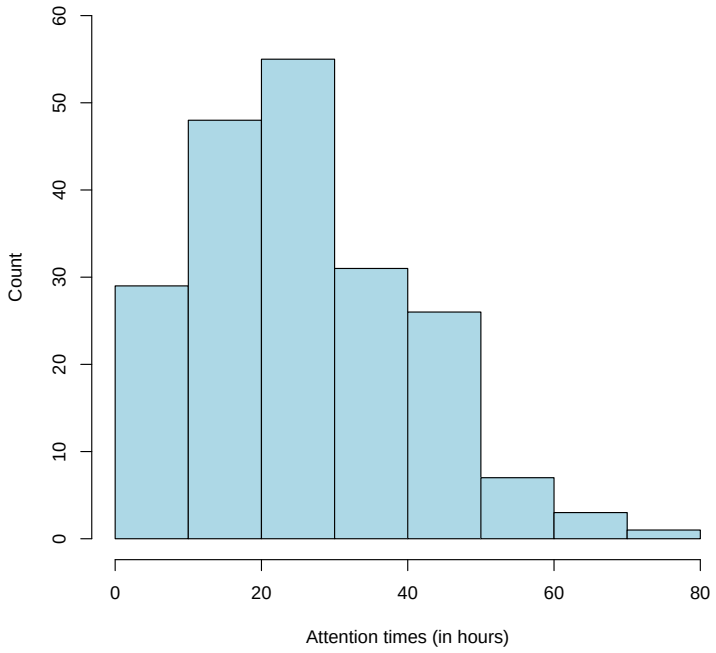
42. The graph in Question 41 shows the winning times for women each year. This year, there were 12,661 women who ran the Boston Marathon. If we wanted to represent the percentage of women who ran the race in these different age groups:

$$<25 \text{ years}, \quad 25\text{-}34 \text{ years}, \quad 35\text{-}44 \text{ years}, \quad 45\text{-}54 \text{ years}, \quad \geq 55 \text{ years}$$

and do so visually in a graph for this year only, what graph would you use?

(a) boxplot
(b) stemplot
(c) line graph
(d) pie chart

43. Twins tend to have lower IQ scores in their early years than non-twins. Psychologists believe this may be explained by the fact that parents do not spend enough time with twins. A large study involving $n = 200$ pairs of twins took place in Wisconsin. Researchers recorded the number of hours parents spent giving attention to their twins during a one-week period. Here is a histogram of these times:



For this sample of twins, the **third quartile** Q_3 is closest to

(a) 10
(b) 20
(c) 40
(d) 60

44. Refer to Question 43. A **95% confidence interval** for the population mean attention time given to twins is 23.9 to 28.1 hours per week. What does this mean?

(a) Ninety-five percent of the twins in the sample will have an attention time between 23.9 to 28.1 hours per week.
(b) Ninety-five percent of the twins in the population will have an attention time between 23.9 to 28.1 hours per week.
(c) Ninety-five percent of the twins will have an average attention time between 23.9 to 28.1 hours per week.
(d) None of the above.

45. Researchers recruited 100 amateur boxers to participate in a study that compared rest and massage before boxing. After a 10-minute workout in which each boxer threw 400 punches, the boxers were randomized to one of two groups:

- **Group 1:** Massage (boxer receives a 20 minute massage)
- **Group 2:** Rest (boxer rests for 20 minutes).

Before they returned for a second workout, the heart rate (beats per minute) and the blood lactate level (micromoles) were measured.

Is this an experiment or an observational study?

(a) experiment
(b) observational study

46. Refer to Question 45. The heart rate results were **not statistically significant** between the two groups of boxers. What does this mean?

(a) Measuring a second variable (blood lactate level) that is correlated with the heart rate prevented researchers from detecting a difference.
(b) Every boxer had a different heart rate.
(c) There may have been differences between the groups of boxers, but the differences were small and could have arisen by chance.
(d) The heart rates could not be measured without bias and random error.

47. What **numerical summary** quantifies the variability associated with the middle 50 percent of a distribution?

(a) mean
(b) median
(c) standard deviation
(d) interquartile range

48. In class, we used R to select random numbers from a list. The authors of your textbook used the Table of Random Digits to do this instead. This is helpful in what situation below?

(a) To decide who to sample from a population of individuals
(b) To calculate the margin of error for a confidence statement
(c) To gauge the impact of using blocking in a randomized experiment
(d) To calculate areas under a normal population density curve

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

1. During March 24-25, 2017, Rasmussen Reports conducted a national telephone and online survey using a simple random sample (SRS) of $n = 1000$ American adults. Each participant was asked:

Are there too many lawyers in America?

The survey found that 540 of the 1000 adults in the sample answered "Yes" to this question.

(a) Calculate a **99 percent** confidence interval for the population proportion and interpret what it means. When you interpret your interval, make sure you identify what the population is.

(b) How could you decrease the margin of error associated with your confidence interval in part (a)? List two ways you could do this.

2. A paper published in February in *PLOS One* by Estelle Dumas-Mallet and colleagues tracked 156 observational studies that had been the subject of stories in major English-language newspapers. The studies dealt with a wide range of issues, including the biology of attention deficit hyperactivity disorder (ADHD), new breast-cancer treatments, and a reported link between pesticide exposure and Parkinson's disease. Follow-up studies, they showed, overturned half of those initially statistically significant results. Of course, these follow-up studies rarely got news coverage!

(a) Using what you have learned in this class, comment on possible reasons why 1/2 of the statistically significant findings were overturned in follow-up studies.

(b) Your friend sends you a link to an article whose headline is

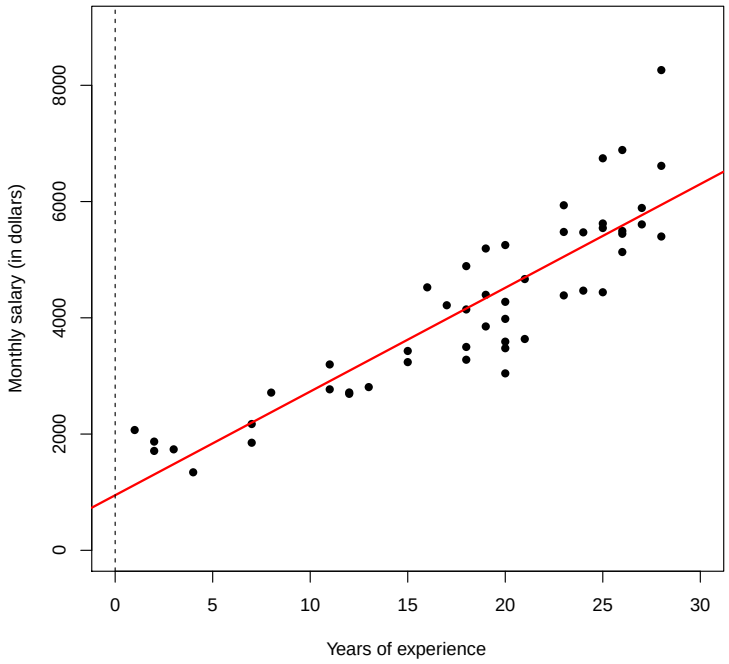
> *"Researchers find that increased sugar intake causes ADHD in young children."*

Using what you have learned in this class, explain to your friend why it might be good to be skeptical of this headline.

3. A researcher in the College of Social Work at USC reports on salaries for social workers in South Carolina who have a bachelor's degree. Using a simple random sample of $n = 50$ social workers in the state, she wants to describe the relationship between

$$\begin{aligned} x &= \text{years of experience the social worker has} \\ y &= \text{monthly salary (in dollars).} \end{aligned}$$

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(salary~years.of.experience)
> fit
Coefficients:
        (Intercept)  years.of.experience
              947.4                178.4
```

(a) From the output, the y-intercept of the least-squares regression line is $a = 947.4$ and the slope is $b = 178.4$. Explain what each of these numbers means in words.

(b) I calculated the correlation between the years of experience and monthly salary to be $r = 0.887$.

```
> cor(years.of.experience,salary)
[1] 0.887
```

Calculate the square of the correlation, express it as a percentage, and interpret precisely what it means in words.

(c) In an orientation seminar for high-school seniors, a professor in the College says,

> *"Ten years after you graduate from college, you can expect to make about 50,000 dollars per year by working as a social worker in the state."*

What do you think of this claim?

4. The Environmental Protection Agency (EPA) performs extensive tests on all new car models to determine their mileage ratings. A simple random sample (SRS) of $n = 100$ 2017 Ford Fiesta Titanium cars was tested for highway driving. The miles per gallon (mpg) was measured for each car.

Below is a stemplot of the mpg observations. The stem is the tens and units digit (e.g., 33). The leaf is the tenths digit (e.g., 0.1).

```
> stem(mpg,scale=2)
    The decimal point is at the |

  30 | 0
  31 | 8
  32 | 5799
  33 | 126899
  34 | 024588
  35 | 01235667899
  36 | 01233445566777888999
  37 | 000011122334456677899
  38 | 0122345678
  39 | 00345789
  40 | 0123557
  41 | 002
  42 | 1
  43 |
  44 | 9
```

(a) If you were going to use numerical summaries to describe the <u>center</u> and <u>spread</u> of this distribution, which ones would you use? Give justification for your answers; e.g., talk about the shape of the distribution.

(b) Here is the 5-number summary for the mpg data:

```
> quantile(mpg,type=2)
   0%   25%   50%   75%  100%
30.00 35.65 37.00 38.35 44.90
```

Construct a boxplot for the data using these values (don't worry about potential outliers). **Neatness counts! Be precise!!**

(c) I used R to calculate the sample mean and the sample standard deviation of the data:

```
> mean(mpg)
[1]  37.0
> sd(mpg)
[1]  2.4
```

Use this information to write a 95% confidence interval for the population mean. Interpret your interval.

5. The time it takes for an ambulance to arrive in response to a 911 call in Columbia, SC follows a normal distribution with mean $\mu = 10$ minutes and standard deviation $\sigma = 3$ minutes.

(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) Would you consider it to be unusual if it took an ambulance 20 minutes to arrive after receiving a 911 call? In defending your answer (yes/no), calculate this observation's standard score and interpret what it means.