Midterm 1 Suggested Review Problems

Here are problems that are similar to the ones you might see on the exam. Be sure to also review old quiz and workshop questions too.

Data Basics

Know all the terminology: Populations, Samples, Individuals (i.e., Observational Units), Variables, Statistics, Parameters, etc.

1. (Exercise 1.17 from OpenIntro Statistics) The General Social Survey asked the question, "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

   (a) An American in the sample.

   (b) Number of hours spent relaxing after an average work day.

   (c) 1.65.

   (d) Average number of hours all Americans spend relaxing after an average work day.

2. (Exercise 1.44 from OpenIntro Statistics) The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).

	1957-1999		2000-2018	
	Failure	Success	Failure	Success
Private	13	295	10	562
State	281	3751	33	711
Startup	-	-	5	65

   (a) What variables were collected on each launch in order to create to the summary table above?

   (b) State whether each variable is categorical or quantitative.

   (c) Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

Sampling

Make sure you know the difference between random error and bias. What is the best way to avoid bias? What is the best way to minimize random error?

3. (Exercise 1.26 from OpenIntro Statistics) Suppose we want to estimate household size, where a "household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

4. (Exercise 1.39 from OpenIntro Statistics) Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

    (a) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

    (b) A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later. However, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

    (c) An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

Examining Numerical Data

Make sure you know all of the different ways to plot a quantitative variable: Boxplots, Stemplots, Histograms. There will also be questions about how we measure the shape, center, and spread of a quantitative variable.

5. (Exercise 2.28(b) from OpenIntro Statistics) The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates for 224 countries for which such data were available in 2014. Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.

Fraction of Countries vs. Infant Mortality (per 1000 Live Births)
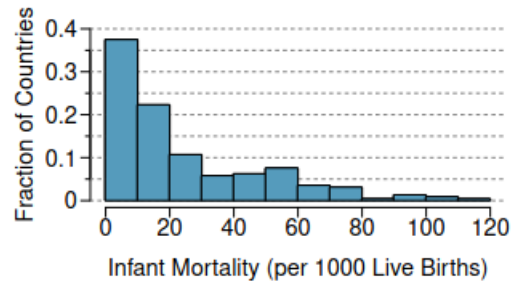
6. (Exercise 2.33 from OpenIntro Statistics) Below are the final exam scores of twenty introductory statistics students.

$$57, \ 66, \ 69, \ 71, \ 72, \ 73, \ 74, \ 77, \ 78, \ 78,$$

$$79, \ 79, \ 81, \ 81, \ 82, \ 83, \ 83, \ 88, \ 89, \ 94$$

Create a box plot of the distribution of these scores.

Normal Distributions

Make sure you comfortable with standardized values (z-values) and finding normal distribution proportions. Be sure to memorize the 68-95-99.7 rule.

7. (Exercise 4.4 from OpenIntro Statistics) In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the Men, Ages 30 - 34 group while Mary competed in the Women, Ages 25 - 29 group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups:

   • The finishing times of the Men, Ages 30 - 34 group has a mean of 4313 seconds with a standard deviation of 583 seconds.

- The finishing times of the Women, Ages 25 - 29 group has a mean of 5261 seconds with a standard deviation of 807 seconds.
- The distributions of finishing times for both groups are approximately Normal.

Remember: a better performance corresponds to a faster finish.

(a) Write down the short-hand for these two normal distributions.

(b) What are the Z-scores for Leo's and Mary's finishing times? What do these Z-scores tell you?

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

(d) What percent of the triathletes did Leo finish faster than in his group?

(e) What percent of the triathletes did Mary finish faster than in her group?

(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

8. (Exercise 4.8 from OpenIntro Statistics) The Capital Asset Pricing Model (CAPM) is a financial model that assumes returns on a portfolio are normally distributed. Suppose a portfolio has an average annual return of 14.7% (i.e. an average gain of 14.7%) with a standard deviation of 33%. A return of 0% means the value of the portfolio doesn't change, a negative return means that the portfolio loses money, and a positive return means that the portfolio gains money.

(a) What percent of years does this portfolio lose money, i.e. have a return less than 0%?

(b) What is the cutoff for the highest 15% of annual returns with this portfolio?

9. (Exercise 4.44 from OpenIntro Statistics) Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.

(a) What is the probability that a randomly chosen 10 year old is shorter than 48 inches?

(b) What is the probability that a randomly chosen 10 year old is between 60 and 65 inches?

(c) If the tallest 10% of the class is considered "very tall", what is the height cutoff for "very tall"?

Scatterplots & Regression

Make sure you can explain the meaning and the units of the slope of the least squares regression line.

10. (Exercise 1.40 from OpenIntro Statistics) The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.
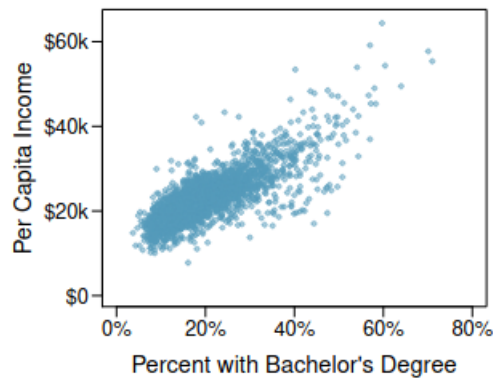

(a) What are the explanatory and response variables?

(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.

(c) Can we conclude that having a bachelor's degree increases one's income?

11. (Exercise 8.8 from OpenIntro Statistics) Match each correlation to the corresponding scatterplot.
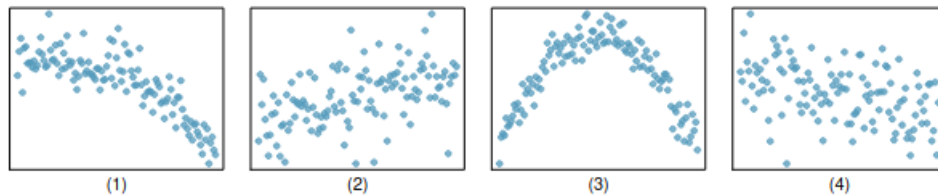
(a) $R = 0.49$

(b) $R = -0.48$

(c) $R = -0.03$

(d) $R = -0.85$

12. (Exercise 8.12 from OpenIntro Statistics) A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months. Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that's when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (°F) and age is measured in weeks.
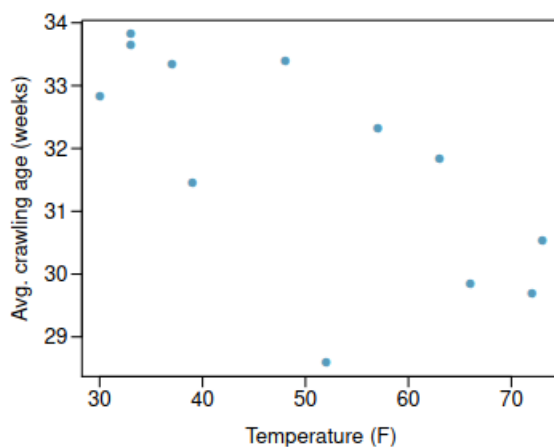
(a) Describe the relationship between temperature and crawling age.

(b) The correlation between temperature in °F and age in weeks was $r = -0.70$. If we converted the temperature to °C and age to months, what would the correlation be?

13. (Exercise 8.17 from OpenIntro Statistics) Consider a regression predicting weight (kg) from height (cm) for a sample of adult males. What are the units of the correlation coefficient, the intercept, and the slope?

14. (Exercise 8.24 from OpenIntro Statistics) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals. The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.
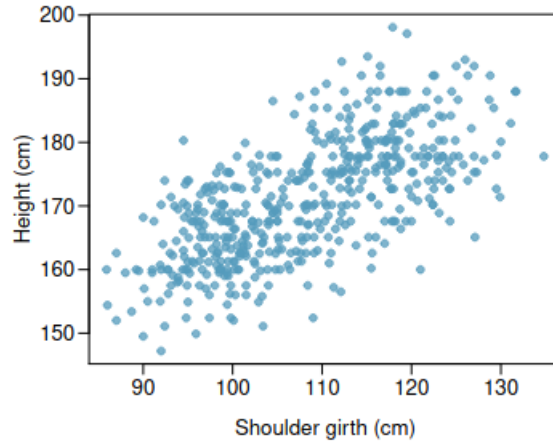
The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.

(a) Write the equation of the regression line for predicting height.

(b) Interpret the slope in this context explain what it means using units.

(c) Calculate R^2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application.

(d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

(e) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?