NOTE: practice problems are not exhaustive! Please go back through your notes and problem sets!

1. Carbon monoxide (CO) for a certain kind of car vary with mean 2.9 gm/mi and standard deviation 0.6 gm/mi. A company has 100 of these cars, acquired from various (i.e., random) sources.

   (a) What is the probability that a randomly selected car from the fleet has CO emissions in excess of 3.1 gm/mi? State any assumptions you must make.

   (b) What is the probability that the average CO emissions for all 100 cars is in excess of 3.0gm/mi?

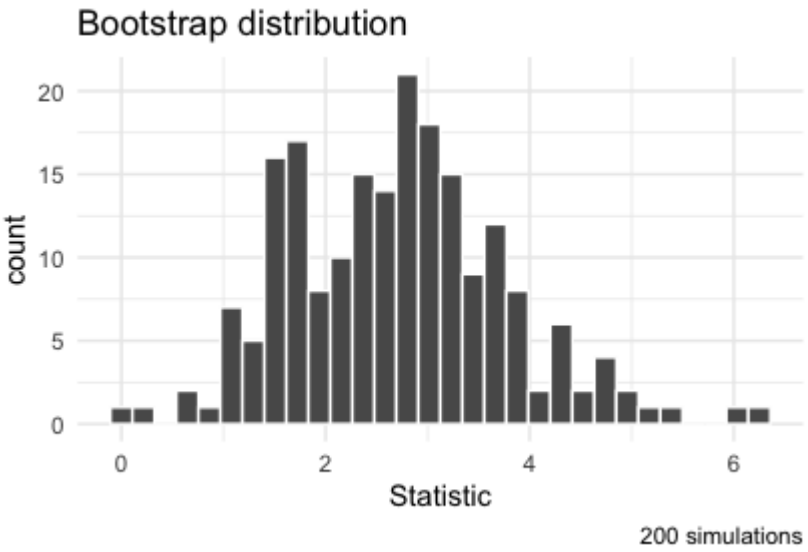
   (c) There is only a 1% chance that the company's car mean CO level is greater than what value?

2. In a double-blind experiment, a sample of male college students were asked to tap their fingers at a rapid rate. The sample was then divided at random into two groups of 10 students each. Each student drank the equivalent of about two cups of coffee, which included about 200 mg of caffeine for the students in one group but was decaffeinated coffee for the second group. After a two hour period, each student was tested to measure finger tapping rate (taps per minute). The average number of taps in the caffeine group was 246.53 and in the no caffeine group was 243.85, and both distributions were reasonably symmetric.

   (a) The goal of the experiment was to determine whether caffeine produces an increase in the average tap rate. Which of the following method(s) may be used to answer this research question? Circle **all** that apply.

   - Test for a single proportion
   - Test for a difference in proportions
   - z-test for a single mean
   - t-test for a single mean
   - t-test for a difference in means
   - Simple linear regression

   (b) We would like to calculate a 95% confidence interval for the average difference in the number of taps in the caffeine and no caffeine groups via bootstrapping. The bootstrap distribution below is created using 200 simulations. Using this distribution, estimate the 95% confidence interval, clearly state the bounds of the interval as well as marking them on the plot, and interpret your interval in context of the data.



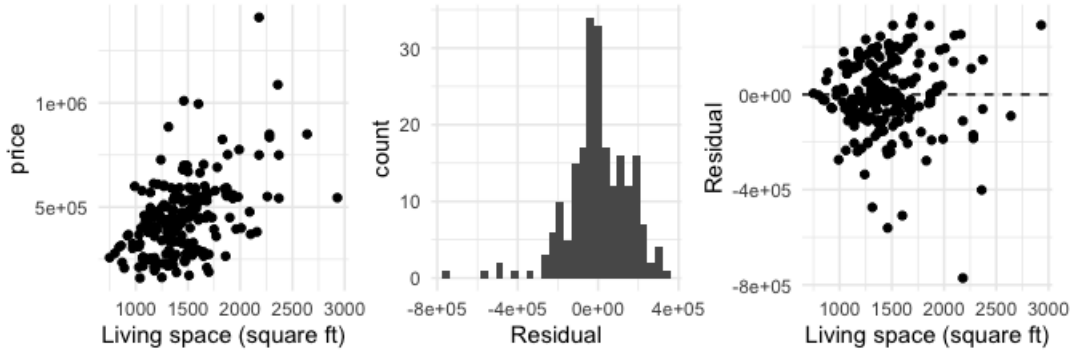
(c) For the test of whether caffeine produces an increase in the average tap rate, the p-value is 0.0212. Based on all of the information you have so far, which of the following intervals are plausible at a 98% confidence level for the average difference in the number of taps in the caffeine and no caffeine groups. Circle all that apply.

- $(-0.265, 4.384)$
- $(0.225, 5.269)$
- $(1.583, 6.249)$

(d) Describe in words how you would obtain the p-value in (c) using simulation-based methods.

3. We have data on house sale prices for King County, USA. The homes were sold between May 2014 and 2015, and we focus on 216 houses with 2 bedrooms and 2 bathrooms only. We will examine the relationship between housing prices and the square footage of the living space of the houses.

(a) A linear model for predicting price using square footage of the living space (square feet) has been fit, and the following diagnostic plots have been produced based on this model. Which of the following is true based on these plots? Circle all that apply.



- The observations are not independent.
- The relationship between price and square footage of the house is strongly linear and positive.
- The linear model is appropriate for predicting price of a house using these data.
- The residuals are not nearly normally distributed.
- The residuals do not have equal variance.

(b) After checking the model diagnostics you realize that a linear model is not appropriate here. You decide to instead model the **log(price)** (i.e. the log price) of these houses. The model output for the log housing price using square footage as the predictor is as follows:

term	estimate	std.error	statistic	p.value
(Intercept)	12.1418376	0.1022114	118.791437	< 0.0001
sqft_living	0.0005472	0.0000683	8.008842	< 0.0001

Assuming the linear model is appropriate, obtain a 90% confidence interval for β_0.

(c) Assuming the model is appropriate, what is the interpretation of the slope parameter?

(d) Someone tells you that because the estimated slope coefficient is so close to 0, there is no evidence of a relationship between the square footage of the house and its price. Is their reasoning correct? If not, explain why not.

(e) True or False: the predicted price of a house with 2000 square feet of living space is \$13.24.

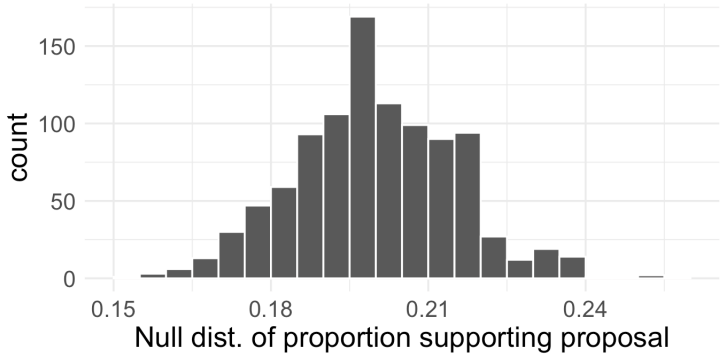
4. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries includes a sample mean hours of 7.73 and standard deviation of 0.77. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. Is the result statistically significant at the 0.05 level? Make a conclusion based on the your decision.

5. A Survey USA poll conducted in Seattle, WA in May 2021 reports that of the 650 respondents (adults living in this area), 159 support proposals to defund police departments.

   (a) A journals writing a news story on the poll results wants to use the headline: "More than 1 in 5 adults living in Seattle support proposals to defund police departments". You caution the journalist that they should first conduct a hypothesis test to see if the poll data provide convincing evidence for this claim. Write the hypotheses for this test using proper notation, defining any necessary quantities.

   (b) Describe in words a simulation scheme that would be appropriate for testing these hypotheses. Also describe how the p-value can be calculated using the simulation results.

(c) The histogram below shows the distribution of 1000 simulated proportions under H_0. Estimate the p-value using the plot and use it to evaluate your hypotheses (i.e. make a conclusion). Assume a significance level of 0.05.



6. In order to assess whether habitat conditions are related to the sunlight choices a lizard makes for resting, researchers observed, at random, 332 Western fence lizards across three different microhabitats (Adolph 1990; Asbury and Adolph 2007). The distribution of habitats and sunlight choices are shown below:

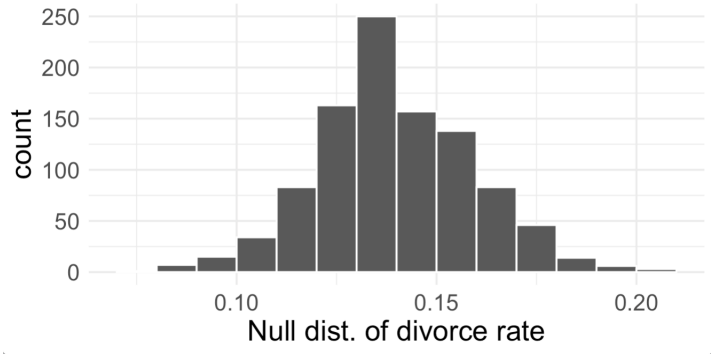
	sunlight			
site	sun	partial	shade	Total
desert	16	32	71	119
mountain	56	36	15	107
valley	42	40	24	106
Total	114	108	110	332

Help shed light on the researcher's question by stating appropriate null and alternative hypotheses and conducting a hypothesis test for these hypotheses. Be sure to conclude the test in the context of the problem.

7. A study conducted in 2020 found that the U.S. adjusted divorce rate was 14 per 100 married women. Joe is suspicious and disagrees with the stated divorce rate. Joe somehow collected data from 323 married or previously-married women, and asked them if they had a divorce in 2020. 55 of the women responded that they indeed had a divorce in 2020.

(a) Write out the hypotheses corresponding to this scenario.

(b) The histogram below shows the distribution of 100 simulated proportions under H_0. Estimate the p-value using the plot and use it to evaluate Joe's hypotheses (i.e. make a conclusion). Assume a significance level of 0.05.



(c) Joe has some free time and also created a 90% bootstrap confidence interval for the divorce rate.

He obtained the following interval: (0.136, 0.207). Interpret this interval in context.

(d) Based on this interval, would it be appropriate for Joe to conclude that the study's reported rate was wrong? Explain your reasoning.

(e) How do your conclusions from (c) and (e) compare?

8. For each of the statements (a) - (d), indicate if they are true or false interpretation of the following confidence interval. If false, provide or a reason or correction to the misinterpretation.

"You collect a large sample and calculate a 95% confidence interval for the average number of cans of soda consumed annually per adult to be (440, 520), i.e. on average, adults in the US consume just under two cans of soda per day".

(a) 95% of adults in the US consume between 440 and 520 cans of soda per year.

(b) There is a 95% chance that the true population average per adult yearly soda consumption is between 440 and 520 cans.

(c) The true population average per adult soda consumption is between 440 and 520 cans, with 95% confidence.

(d) The average soda consumption of the people who were is sampled is between 440 and 520 cans of soda per year, with 95% confidence.

9. A recent poll found that 11% of US adults say they have smoked cigarettes in the past week, a historical low. In a random sample of 730 randomly selected students at four-year colleges, it was found that 66 students have smoked cigarettes in the past week. Test the claim that the smoking rate of students at four-year colleges is the same the national US adult average at the 0.05 significance level.

10. A certain graduate department hopes to receive applicants with a verbal GRE scores over 210. From a sample of 29 applicants, the mean verbal GRE score is 213.8 with a standard deviation of 8.5.

(a) Using a significance level of 0.05, is the mean of this year's applicants' verbal GRE scores greater than the desired lower bound of 210? Be sure to state any assumptions you make.

(b) Obtain a 90% confidence interval for the mean verbal GRE score of all of this year's applicants to this graduate program.