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.

   Without knowing the underlying distribution of the data, we cannot answer this question. If we assume normality, we can find the z-score and use the z-table: $z = \frac{3.1-2.9}{0.6} = \frac{1}{3}$. *We want* $Pr(Z > z) \approx 1 - 0.63 = 0.37$

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

   Since the cars were randomly sampled (independent) and $n = 100$ *is quite large, CLT will kick in assuming no particularly extreme outliers. CLT says* $\bar{X} \sim N(2.9, 0.6/10)$*, approximately. The z-score here would be* $z = \frac{3.1-2.9}{0.06} = 3.33$. *So* $Pr(Z > z) \approx 1 - 0.9996 = 0.0004$.

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

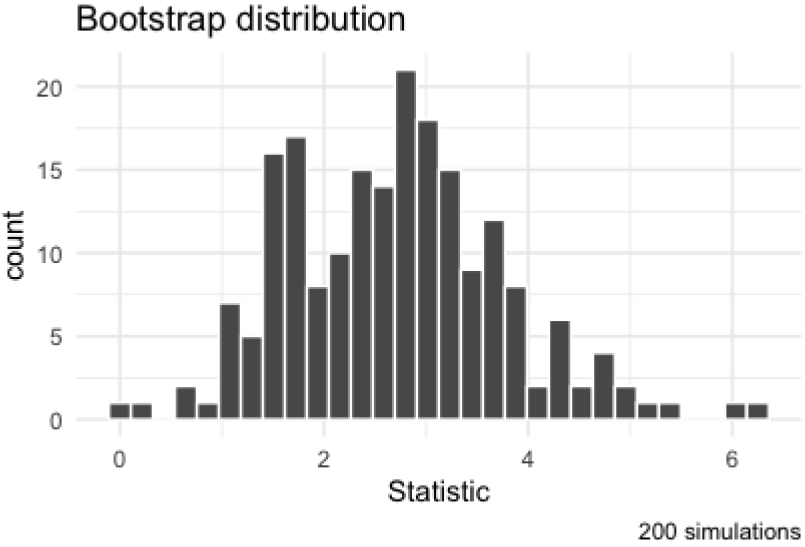
   Using the z-table, the 99-th percentile of the standard normal occurs at a z-score of about 2.33. So we want to find the x *such that* $2.33 = \frac{x-2.9}{0.06}$. *This is* $x = 3.03958$.

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 *This one!*
   - Simple linear regression *This one!*

(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.



We are looking for bounds that will give us roughly 5 simulations below and above. This is roughly $(1, 5)$. We are 95% confident that the true difference in mean tapping rates per minutes between the two groups (caffeine - no caffeine) is between 1 and 5.

(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.

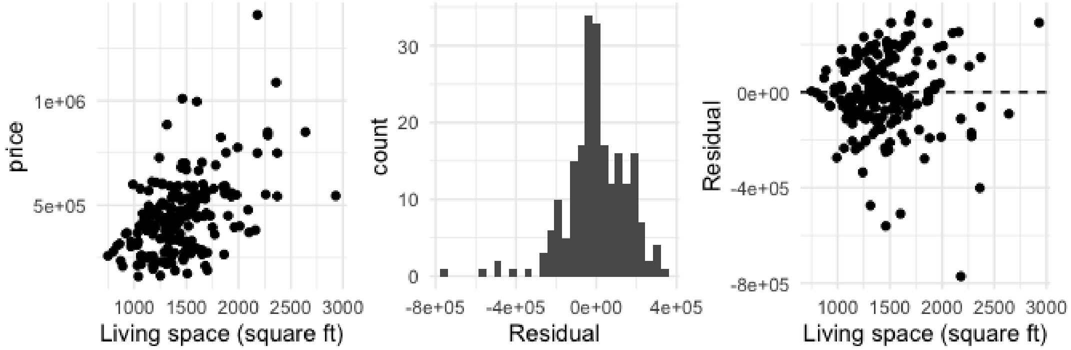
- $(-0.265, 4.384)$ *Not reasonable given that the p-value was 0.0212, so the interval shouldn't include 0.*
- $(0.225, 5.269)$ *This one seems reasonable!*
- $(1.583, 6.249)$ *Based on the bootstrap distribution above, this interval would be way too narrow!*

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

Write down the observed tapping rates across both groups on 20 cards. Shuffle them, then deal ten into a pile representing the caffeine group and the remaining ten into a pile representing the no caffeine group. Take the means of each group, and then record the difference (caffeine - no caffeine). Repeat this lots and lots of times. To get the p-value, we want to find the proportion of the simulated difference in sample means was greater than or equal to the observed difference of 2.68 taps per minute.

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. *Could argue this one!*
   - The residuals do not have equal variance. *Definitely this one!*

   (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.

   We get the critical value using the t with 214 degrees of freedom. Looking at the t-table, we can get the critical value by looking at the row with 200 degrees of freedom (rounding to closest df) and the column 0.95. This yields critical value of 1.653. $12.142 \pm 1.653(0.102)$. *This yields a 90% CI of* $(11.97, 12.31)$.

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

   For every additional square foot of living space, the log of the sale price of the home is expected to increase by about $0.0005472.

(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.

No! The magnitude of the coefficient doesn't matter so much as it depends on the scale of the variable it is associated with. We should instead look to the test-statistic or p-value. Since the p-value for the test of $H_0 : \beta_1 = 0$ *versus* $H_A : \beta_1 \neq 0$ *is so small, we would conclude that there is actually strong evidence of a linear relationship between square foot of living space and the log price of the home.*

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

False! This would be predicted price on the log scale!

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.

   $H_0 : \mu = 8$ *versus* $H_A : \mu < 8$ *where* μ *is the average hours of sleep of a New Yorker. We have independence via random sample. We don't have a histogram of the data and the sample size* $n = 25$ *is small. We will cautiously proceed, but note that we are assuming there are no clear outliers in the data. Conducting a t-test, we have* $t = \frac{7.73-8}{0.77/\sqrt{25}} = -1.75$. *Our p-value is* `pt(-1.75, df = 24)` $= 0.046$. *(To get an approximate p-value from t-table, see that* $P(T_{24} \geq 1.711) = 0.95 \Rightarrow P(T_{24} \geq 1.75) > 0.95 \Rightarrow P(T_{24} \leq -1.75) < 0.05$*). Since this is less than 0.05, we reject* H_0*. The data provide some convincing evidence that the true average hours of sleep New Yorkers receive is less than 8 hours,*

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.

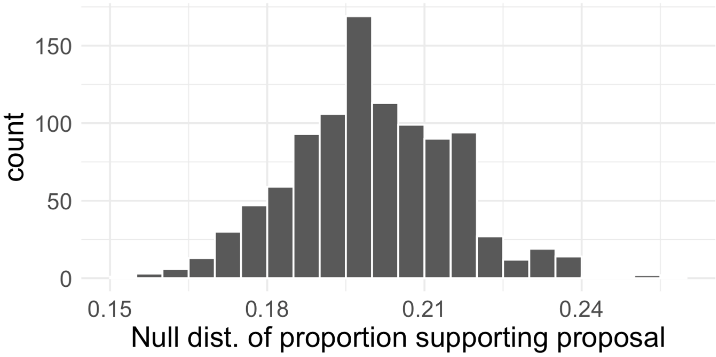
   (a) A journals writing a news story on the poll results wants to use the headline: "More than 1 in 5 adults living in Seattle support proposals to defund police departments". You caution the journalist that they should first conduct a hypothesis test to see if the poll data provide convincing evidence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

   $H_0 : p = 0.20$ *versus* $H_A : p > 0.20$ *where* p *is the true proportion of Seattle adults who support proposals to defund.*

   (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.

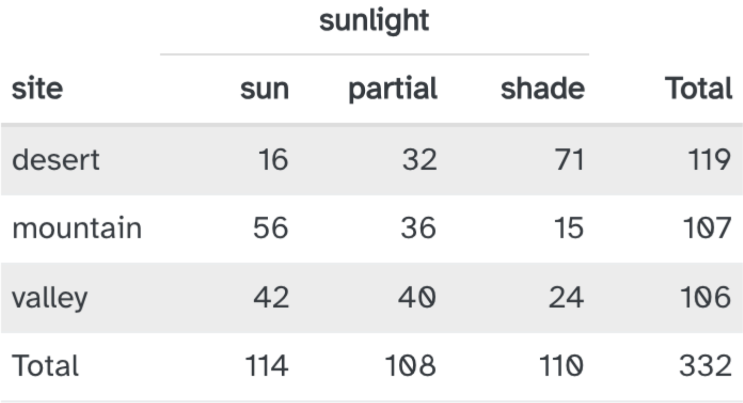
Example solution (there are others): Take 10 cards, 2 black cards representing those who support proposals to defund police departments and 8 red cards representing those who do not. Shuffle the cards and draw with replacement (shuffling each time in between draws to get our "infinite population") 650 cards representing the 650 respondents to the poll. After each iteration, calculation $\hat{p}_{sim}$, the proportion of black cards which represents the simulated proportion of adults in favor. The p-value will be the proportion of simulations where $\hat{p}_{sim} \geq 0.245$.

(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.



There is only one simulated proportion that is at least 0.245, therefore the approximate p-value is 0.001. Since $0.001 < 0.05$, reject H_0. The data provide convincing evidence that the proportion of Seattle adults who support proposals to defund police departments is greater than 0.20.

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.

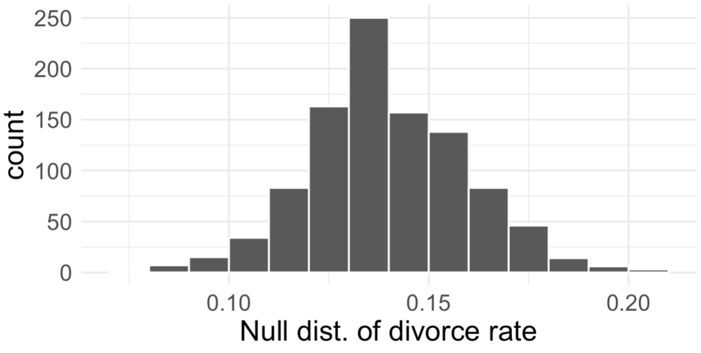
H_0: habitat conditions and sunlight preferences of this lizard species are independent versus H_A: habitat conditions and sunlight preferences of this lizard species associated. In order to perform inference, we should check conditions. The lizards were observed randomly, which hopefully guarantees independence. Then, we should check that the expected counts are indeed at least 5 for all the cells (they are!). Set $\alpha = 0.05$. Our test-statistic is 68.773. To find the p-value, we take `1 - pchisq(68.773, df = 4) = 4.118927e-14`. (Looking at the chi-square distribution table, we see that the 0.99-th percentile/quantile of a Chi-square distribution with 4 degrees of freedom is 13.280 (has 0.01 probability to the right*). Since our test-statistic 68.773, we can say that the p-value is less than 0.01.) Since this is lower than 0.05, reject H_0. There is sufficient evidence to suggest that habitat provides information about the likelihood of being in the different sunshine states (i.e. the two variables are associated).*

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.

   $H_0 : p = 0.14$ versus $H_A : p \neq 0.14$ where p is the true divorce rate among married women.

   (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.



   The observed proportion was $\hat{p}_{obs} = 55/323 = 0.17$. Since the alternative is two-sided, the p-value is approximately 0.13. Since this is larger than 0.05, fail to reject. The data do not provide convincing evidence that the divorce rate amount married women is different from 0.14.

   (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.

   Joe is 90% confidence that the true divorce rate among married women is between 0.136 and 0.207.

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

No! 0.14 is included in the interval, so it is a plausible value.

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

They agree!

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. *False.The interval is for the parameter (a number which describes the population), not for individual observational units.*

(b) There is a 95% chance that the true population average per adult yearly soda consumption is between 440 and 520 cans. *False. Although unknown, the parameter is either in the interval or it is not (so either with probability zero or probability one).*

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

(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. *The sample mean is always inside the interval (it is the center!). We are 100% confident that the sample mean is in the interval*

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.

$H_0 : p = 0.11$ *versus* $H_A : p \neq 0.11$ *where p is the smoking rate of students at four year colleges. To use CLT, we verify that we have independence via random sampling and the success-failure condition is met:* $np_0 = 730(0.11) = 80.3 \geq 10$ *and* $n(1-p_0) = 649.7 \geq 10$. *So the CLT tells us that our test statistic is* $z = \frac{66/730-0.11}{\sqrt{0.11(0.89)/730}} = -1.69$. *The p-value is* $2 * \texttt{pnorm(-1.69)} = 0.091$. *(We can get an approximation from the z-table, by noting that* $P(Z \leq 1.69) = 0.954$ *so* $P(Z \leq -1.69) = 1 - 0.954 = 0.046 \Rightarrow$ *p-value is* $2 * 0.046 = 0.092$.*) Since the p-value is greater than 0.05, we fail to reject* H_0. *The data do not suggest that the smoking rate of students at four-year colleges is different from the US adult average rate.*

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.

   $H_0 : \mu = 210$ and $H_A : \mu > 210$ where μ is the mean verbal GRE score of all of this year's applicants to this graduate program. We can assume that the samples are independent as one applicant's score won't tell us about another's. The sample size $n = 30$ is less than 30, so as long as there are no clear outliers, we can proceed with CLT. Since the population standard deviation is unknown, we conduct a t-test: $t = \frac{213.8-210}{8.5/\sqrt{29}} \approx 2.41$. Our p-value is `1 - pt(2.41, 28)` $= 0.011$ (To get from the t-table, note because $2.048 < 2.41 < 2.467$, $0.01 < P(T_{28} \geq 2.41) < 0.025$). Since this is less than 0.05, we reject H_0. The data provide convincing evidence that the mean verbal GRE of this year's applicants is above 210.

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

   Define μ as the stated. We checked conditions for CLT in (a). So to obtain our CI, we need our critical value. This is `qt(0.95, 28) = 1.701` (also obtained directly from the t-table). So the 90% CI is $213.8 \pm 1.701(8.5/\sqrt{29}) = (210.97, 216.63)$.