Name:

Final Exam Solutions

Instructions. You may use *one handwritten sheet of notes* and a *regular calculator*. Using anything else (the textbook, the internet, a friend, RStudio, ...) is a violation of the Honor Code. You may print out this document and work directly on it, or you can work on your own separate sheet of paper. You have 2 hours (enforced by the Honor Code, but a 5-hour time limit will be enforced by Gradescope). Please make sure to sign the Honor Code statement at the end. Good luck! ☺

Problem 1 (1 point)**.** You are interested in studying per capita cookie consumption in the United States. You suspect that per capita cookie consumption might vary from state to state. What sampling strategy would you use for your study?

Solution. Stratified sampling (ie, take a simple random sample within each state)

Problem 2 (2 points)**.** Masak Hijau is a cultivar of bananas that originates in Malaysia. The masses of individual Masak Hijau banans are normally distributed with mean 110 g and standard deviation 5 g. What is the standard deviation of the total mass of 100 randomly selected Masak Hijau bananas?

Solution. Let X_i denote the mass of the ith banana. The variance $\text{Var}(X_i)$ is 25. Since the 100 bananas are randomly selected, the total mass $X = \sum X_i$ is $\text{Var}(X) = 100 \cdot \text{Var}(X_i) = 2500$, so the standard deviation of X is 50 g.

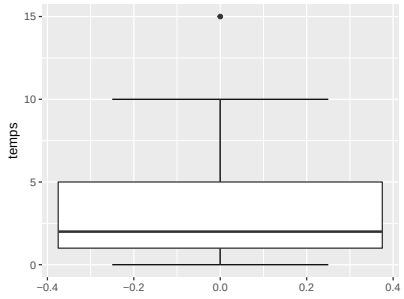
Problem 3 (3 points)**.** You've collected data about the daily high temperature in degrees Celsius during February 2021 from several towns.

(a) After sorting the data you've collected from town A, you get the following list: `0,0,1,...,5,10,15`. The quartiles of this data are given below.

Min	Q1	Median	Q3	Max
0	1	2	5	15

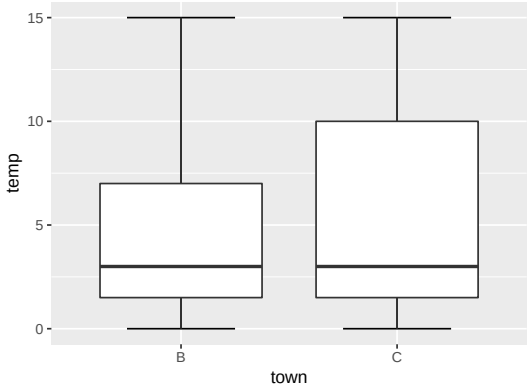
Use this information to sketch a box plot of temperatures in town A.

Solution. The interquartile range is 4, so $1.5 \cdot \text{IQR} = 6$. The minimum observation 0 is within 6 units of Q1, and the observation 10 is within 6 units of Q3, but the maximum observation 15 is not. Thus the whiskers extend from 0 to 10, and there is a point representing the outlier at 15.



Note: Just because the upper whisker *can* extend to 11 doesn't mean it does! It stops at the highest data point that's within range, which in this case is 10.

(b) The following figure shows side-by-side box plots of the temperatures in towns B and C.



Let μ_B be the mean temperature in town B and μ_C the mean temperature in town C. Circle one of the following, and justify briefly.

$$\mu_B < \mu_C \qquad \mu_B = \mu_C \qquad \mu_B > \mu_C$$

Solution. $\mu_B < \mu_C$. The distribution of temperatures in town C is more skewed right, which pulls the mean temperature in town C higher.

Problem 4 (1 point each)**.** You are a rainforest botanist. In a recent exploration of a remote corner of the Amazon, you encountered a previously undiscovered species of a flowering plant. Some of these plants had caterpillars on them. For each of the 1000 specimens that you found, you recorded the following data: leaf length (in mm), flower color (orange or pink), petal length (in mm), and whether or not you saw caterpillars on it (yes or no). You collect the data into a data matrix as follows.

	leafLength	flowerColor	petalLength	caterpillars
1	52	orange	61	yes
2	48	pink	60	yes
⋮	⋮	⋮	⋮	⋮

For each of the following pairs of variables, state *one* type of plot that you might use to visualize the relationship between the two variables. (If there is more than one valid option, just pick one!)

(a) `leafLength` and `petalLength`

Solution. Scatterplot

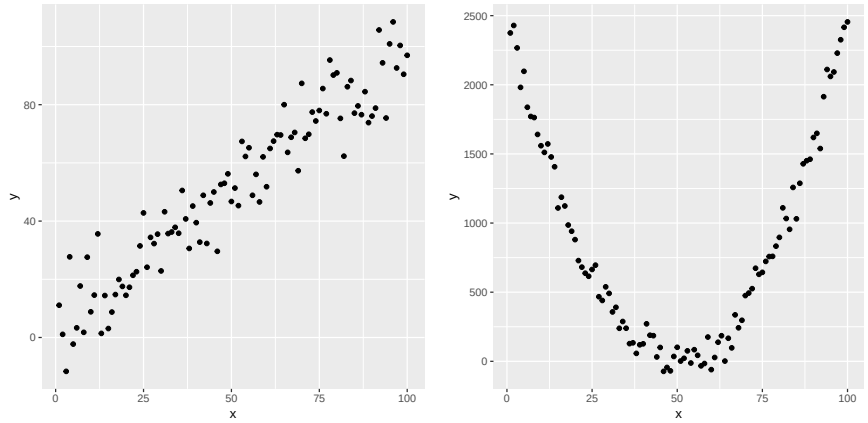
(b) `flowerColor` and `caterpillars`

Solution. Side-by-side bar plot, stacked bar plot, or (two variable) mosaic plot.

(c) `flowerColor` and `leafLength`

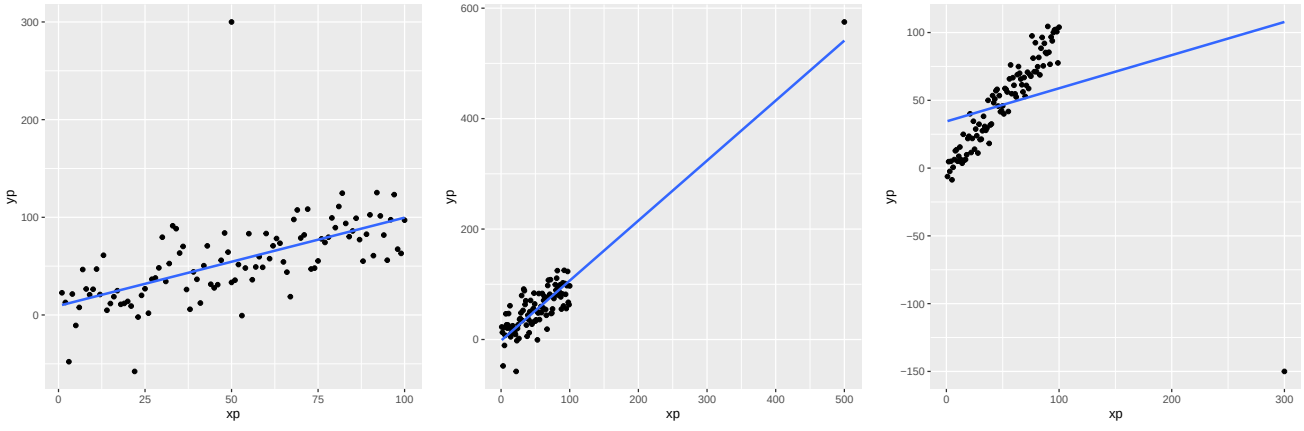
Solution. Side-by-side box plot, or hollow histogram.

Problem 5 (1 point)**.** Two scatterplots are shown below. Which data has a greater correlation between x and y? Explain briefly.



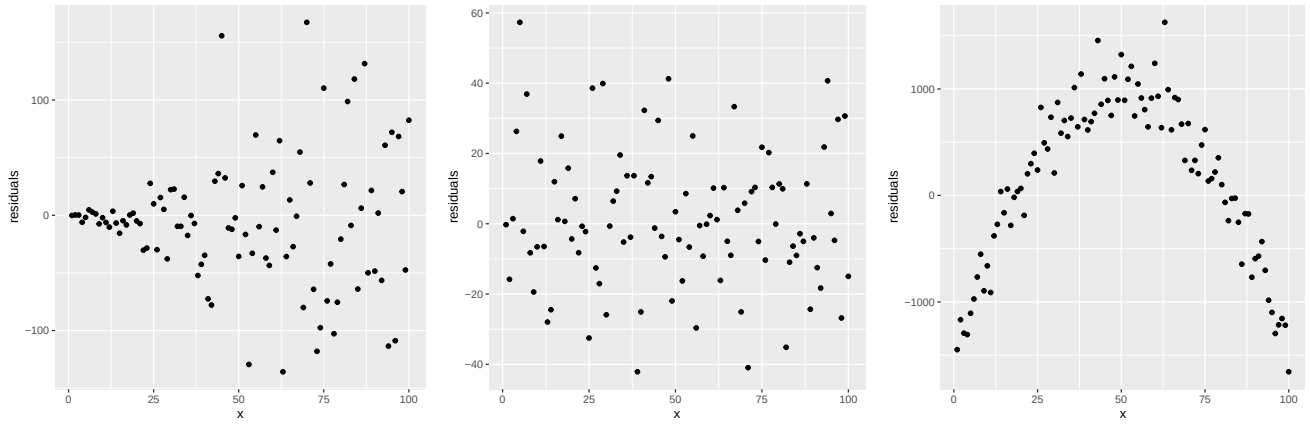
Solution. The data on the left has greater correlation, because the relationship there is linear.

Problem 6 (1.5 points)**.** Three scatterplots are depicted below, and each one has an outlier. Underneath each one, state whether or not the outlier has high leverage and/or is influential.



Solution. (a) not high leverage, (b) high leverage but not influential, (c) high leverage and influential

Problem 7 (1.5 points)**.** The plots below show residuals after fitting least squares regression lines to three different data sets. In each case, do you have concerns about using applying least squares regression? Write "concerned" or "not concerned" below each plot, and justify briefly (less than 5 words).



Solution. (a) concerned: non-constant variability, (b) not concerned: residuals look random, (a) concerned: pattern in residual plot

Problem 8 (3 points)**.** On a distant alien planet, 10% of the alien population has *hypertentacloma*, a genetic condition which causes them to sprout fifteen additional tentacles when they reach the age of 42 (normally, they only have three tentacles throughout their lives). A test for hypertentacloma given at birth has a false negative rate of 1% and a false positive rate of 44%.

Suppose a newborn alien named Zwq has just tested positive for hypertentacloma. What is the probability that Zwq actually has hypertentacloma? (If you don't know how to do this, write clearly that you don't know and then sketch a picture of baby Zwq for 1 point! ☺)

Solution. Let A be the event that someone has this genetic condition, and B be the event that someone tests positive. We know that $P(A) = 0.10$. A false negative rate of 1% tells us that $P(B^c|A) = 0.01$, which also means $P(B|A) = 0.99$. A false positive rate of 44% tells us that $P(B|A^c) = 0.44$. Then, by Bayes's theorem, we have

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{0.99 \cdot 0.10}{0.99 \cdot 0.10 + 0.44 \cdot 0.90} \\ &= 0.20. \end{aligned}$$

Thus there is a 20% chance that Zwq has hypertentacloma.

Problem 9 (1 point each)**.** For each of the following, sketch a picture and write down how you would compute the indicated quantity in R using the functions in the table below. You don't need to actually compute the values!

Distribution	Observation to Percentile	Percentile to Observation
Normal distribution	`pnorm`	`qnorm`
t-distribution	`pt`	`qt`
chi-square distribution	`pchisq`	`qchisq`

(a) The percentage of observations in a t-distribution with 5 degrees of freedom that are greater than 2.

Solution. `1-pt(2,5)`

(b) The percentage of observations in a standard normal distribution that are within 1.5 standard deviations of the mean.

Solution. `pnorm(1.5)-pnorm(-1.5)`

(c) The observation in a standard normal distribution that is smaller than 80% of observations.

Solution. `qnorm(0.2)`

(d) The percentage of observations in a chi-square distribution with 12 degrees of freedom that are between 10 and 14.

Solution. `pchisq(14,12)-pchisq(10,12)`

(e) The number t^* such that 50% of observations in a t-distribution with 4 degrees of freedom are betweeen $-t^*$ and t^*.

Solution. `qt(0.75, 4)`

Problem 10 (3 points)**.** Johann is a sociologist studying psychedelic use in the United States. He takes simple random samples of 100 men and 100 women, all over the age of 21. He finds that 22 of the men and 12 of the women in his sample have experience using psychadelics.

(a) Describe the data matrix that Johann would have used to generate these estimates. How many observations are in this data matrix? How many variables? What types of variables are they?

Solution. The data matrix would have 200 observations or rows, each corresponding to each person in the sample. Two variables or columns are recorded about each person: `sex` (male/female) and `psychadelics` (yes/no). Both are categorical variables. In other words, the data matrix is something that looks like this:

	`sex`	`psychadelics`
1	M	yes
2	F	yes
3	M	no
4	M	no
5	F	yes
6	F	no
⋮	⋮	⋮
200	M	yes

Note: The following thing is a contingency table, not a data matrix:

	M	F	total
yes	22	12	34
no	78	88	166
total	100	100	200

(b) Johann wants to test the hypothesis H_0 that equal proportions of men and women have used psychedelics. Help Johann calculate a p-value for this hypothesis test. You can leave your answer in terms of the R functions listed in the table for problem 9.

Solution. Let p_1 and p_2 be the proportion of men and women who use psychadelics, respectively. Under the hypothesis H_0 that $p_1 = p_2$, our best estimate for the overall percentage of Americans who use psychadelics is $\hat{p} = (22 + 12)/200 = 0.17$, and the sampling distribution of $p_1 - p_2$ is normal with mean 0 and standard deviation

$$\sqrt{\frac{0.17 \cdot 0.83}{100} + \frac{0.17 \cdot 0.83}{100}} \approx 0.0531.$$

We observed $\hat{p}_1 - \hat{p}_2 = 0.10$, which has z-score 1.882. To calculate the p-value, we use R to calculate `2*pnorm(-1.882)` (which is 0.06).

Note 1: You need to calculate the "pooled" percentage $\hat{p} = 0.17$ for standard error. Otherwise, you're not using your H_0. (If you don't compute the pooled percentage, you end up with a standard error of 0.0526 and a z-score of 1.899.)

Note 2: You could also do this using a chi-square test. If you do this, you would find $\chi^2 = 3.54$ with 1 degree of freedom (and a p-value of 0.06).

Problem 11 (3 points)**.** Kwame wants to know if the average number days per month that Coloradans exercise is different than that for Kansans. Suppose that the number of days that people exercise in each state is normally distributed. Kwame collects the following data.

State	Mean	SD	Sample Size
Colorado	22	2	225
Kansas	18	4	100

Help Kwame compute a p-value for testing the hypothesis H_0 that the average number of exercise days is the same between Coloradans and Kansans. You can leave your answer in terms of the R functions listed in the table for problem 9.

Solution. Let μ_1 and μ_2 be the averages in Colorado and Kansas, respectively. The null hypothesis H_0 is that $\mu_1 = \mu_2$ and the alternative hypothesis H_A is that $\mu_1 \neq \mu_2$. Assuming H_0, the sampling distribution for the test statistic

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{225} + \frac{\hat{\sigma}_2^2}{100}}}$$

follows a t-distribution with at least 99 degrees of freedom. We calculate that, with our data, the standard error (denominator in the above expression) is about 0.42, and that $T \approx 9.5$. Thus the p-value is `2*pt(-9.5, 99)`.

Problem 12 (3 points)**.** Anastasia is interested in studying reading habits of students. This summer, she is planning to conduct a survey on a large random sample of students. For each student, she records how many books they read during the 2019–2020 academic year, and what level of school they were in (elementary school, middle school, high school, college, or graduate school).

(a) Suppose that, after Anastasia collects this data, she'll want to test the hypothesis that the average number of books read is the same for students in all five of these levels. What kind of a hypothesis test would you recommend to Anastasia?

Solution. We're comparing a numerical variable against a categorical one, so ANOVA.

(b) Anastasia has taken great pains to ensure that she has a truly random sample, so she's pretty sure the data points she's collected are independent. What other condition(s) should hold before Anastasia proceeds with her hypothesis test?

Solution. Nearly normal distribution within each category, and constant variability across groups.

(c) Suppose all conditions needed to apply the appropriate hypothesis test are met. Anastasia finds some instructions online to run this test in R, and it returns a p-value of 0.003. She's unsure what to make of this number. Explain to Anastasia how to interpret this p-value in context, and what conclusions she should draw from this p-value.

Solution. Interpretation: If there was no difference in the average number of books read by students of various levels, there would be a 0.3% chance of seeing data at least as extreme as the data that Anastasia actually saw.

Conclusion: Since this number is quite small, this study gives evidence to reject the hypothesis that there is no difference in the average number of books read by students of the five levels.

Setup (for problems 13–16)**.** While rummaging through a dusty trunk in his attic, Daisuke discovers his great-great-great-grandmother Kaori's childhood diary. As he opens the diary, a gold coin falls out. Inside the diary, Kaori writes that this coin lands heads 95% of the time. The diary also records all of the mischief that Kaori caused using this coin.

Problem 13 (1 point)**.** Daisuke starts flipping Kaori's coin to test it out. How many times will he need to flip the coin until he expects to see the first tails?

Solution. Let X be the random variable which outputs the number of flips until the first tails. This is a geometric random variable, where "success" means getting a tails, which has probability 0.05. Thus the expected number of flips until the first tails is $1/0.05 = 20$.

Problem 14 (2 points)**.** Daisuke flips Kaori's coin 10 times. What is the probability that he sees at least one tails?

Solution. Let X be the random variable which outputs the number of tails out of 10. This is a binomial random variable, where "success" means getting a tails. We have

$$P(X \geq 1) = 1 - P(X = 0) = 1 - 0.95^{10} \approx 0.40.$$

There is about a 40% chance that he sees at least one tails.

Problem 15 (2 points)**.** Daisuke reads that Kaori was exceptionally fond of strawberries. One day, she decided to trick the neighborhood strawberry farmer's daughter Michiko into giving her strawberries through a coin flipping game. They each started with a basket of strawberries. Kaori would flip her gold coin; if it landed heads, Michiko gave Kaori a strawberry, and if the coin landed tails, Kaori gave Michiko a strawberry. Kaori writes that she played this game for 100 rounds in a row before Michiko went home crying. But Kaori did not write down how many strawberries she gained by playing this game. How many strawberries do you expect she gained?

Solution. Let X_i be the number of strawberries gained on round i. Then X_i takes the values 1 and -1, and $P(X_i = 1) = 0.95$ and $P(X_i = -1) = 0.05$. Thus

$$E(X_i) = 1 \cdot 0.95 - 1 \cdot 0.05 = 0.9$$

so the total number of strawberries that Kaori is expected to gain is

$$E(X_1 + \cdots + X_{100}) = 100 \cdot 0.9 = 90.$$

More intuitively, she can expect to win 95 of the 100 rounds, and lose 5 of the rounds. In other words, she's expected to win 95 strawberries and to lose 5 strawberries, so her net gain is expected to be 90 strawberries.

Problem 16 (3 points)**.** To show that her gold coin really lands heads 95% of the time, Kaori recorded the results of 1000 coin flips in her diary. Of these, 960 landed heads. Kaori seems to have been convinced by this data, but she did not indicate in writing that she conducted any statistical tests.

(a) Using the approximation $z^* \approx 2$, help Daisuke use Kaori's data to construct a 95% confidence interval for the true proportion of times that the coin lands heads.

Solution. Let p be the true proportion of times that the coin lands heads. We observed $\hat{p} = 0.96$. The standard error for the sampling distribution is

$$SE = \sqrt{\frac{p(1-p)}{1000}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} = \sqrt{\frac{0.96(1-0.96)}{1000}} \approx 0.006,$$

so the margin of error is $2 \cdot 0.006 \approx 0.012$. Thus the confidence interval is $(0.948, 0.972)$.

Note: When you're constructing confidence intervals, you use your point estimate in your standard error formula rather than the null value. In other words, you want to compute SE using 0.96, rather than 0.95.

(b) Based on this confidence interval, would you reject Kaori's hypothesis that her coin lands heads 95% of the time? Explain briefly.

Solution. No, because the confidence interval contains 0.95.

Problem 17 (2 points)**.** What is the smallest number C such that the C% confidence interval constructed using the data from Kaori's diary contains 0.95? Show your work, but you can leave your answer in terms of the R functions listed in the table in problem 9.

Solution. In order for the confidence interval to contain 0.95, we would need a margin of error of at least $0.96 - 0.95 = 0.01$. In other words, we need $z^* \cdot \text{SE} \geq 0.01$, so

$$z^* \geq 0.01/\text{SE} \approx 1.614,$$

using the formula for SE above. So, to calculate C, we have to calculate the percentage of data in a standard normal distribution that's contained between 1.614 and -1.614. To do this, we can calculate `pnorm(1.614)-pnorm(-1.614)` and get $C = 89.3\%$.

Honor code. If you have neither given nor received any unauthorized aid on this quiz, please write either "HCU" or "Honor Code Upheld" below, and sign your name next to it.