

Name:

FINAL EXAM SOLUTIONS

Instructions. You are permitted to use pencils, pens, erasers, a handwritten sheet of notes, a clean new session of RStudio, and your mind. Using anything else (internet, textbook, etc) is a violation of the Honor Code. If you use RStudio to calculate something that isn't just a basic arithmetic operation, you must write down exactly what command you used (eg, `pnorm`).

You have 2.5 hours to work on the exam (from 9:15am until 11:45am). After you finish:

1. Make sure you've signed the Honor Code at the very end.
2. Remove (or cut off) the staple.
3. Use a scanner on campus (either in the math department, or in Tutt Library) to scan your exam as a PDF. Make sure to select the following settings:
 - At least 200 DPI
 - Black and White
 - 2-Sided (to make sure both sides of each page get scanned)
4. Wait a few minutes. You should receive an email with your scanned exam attached as a PDF soon.
5. Use your computer to upload the PDF of your exam to Gradescope under "Final Exam."

In recognition of the fact that it might take you some time to find a scanner, get a turn using the scanner, scan, and then figure out how to upload, the deadline to upload the exam to Gradescope is set to 1pm.

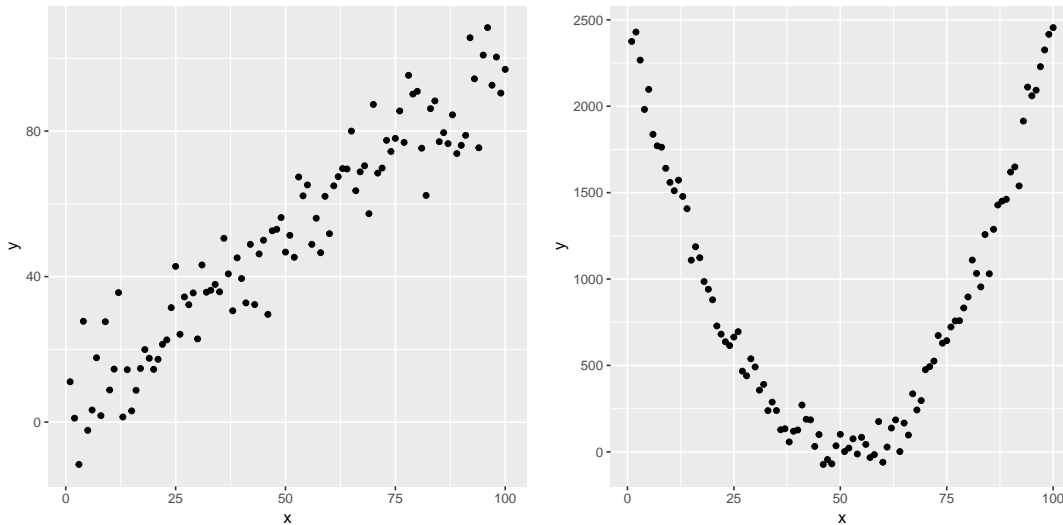
If you've been allocated extra time on exams by Accessibility Resources, you are permitted to work for 3.75 hours (from 9:15am until 1pm). So, if you do not finish in time for the 1pm upload deadline, email me your scanned PDF by 2pm.

Good luck! ☺

Problem 1 (1 point each). Circle T for true or F for false. No explanations are required.

Note: There will be no revisions for the final exam!

- (1) A randomized experiment can provide evidence for causation. T F
- (2) The correlation between x and y in the scatterplot on the left is larger than the correlation on the right. T F



- (3) Suppose the mass of a banana is normally distributed with standard deviation 1 g. Then the standard deviation of the total mass of 25 bananas is 5 g. T F
- (4) The closer that R^2 is to 1, the better that the relationship between two variables is explained by a least squares regression. T F
- (5) The sampling method described in the following scenario is multistage sampling. T F

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. To ensure that your sample contains respondents from each state, you choose a simple random sample within each state.

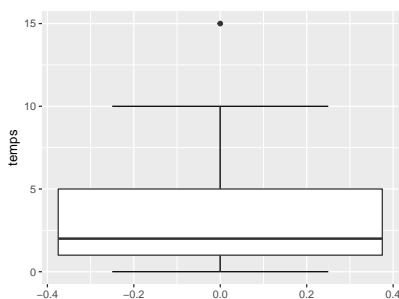
Problem 2 (3 points). You've collected data about the daily high temperature in C during February 2020 from several towns.

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

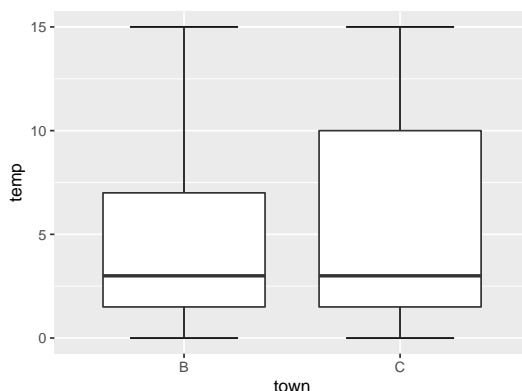
Min	Q1	Median	Q3	Max
0	1	2	5	15

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.



- (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.

$$\mu_B < \mu_C$$

$$\mu_B = \mu_C$$

$$\mu_B > \mu_C$$

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

Problem 3 (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
\vdots	\vdots	\vdots	\vdots	\vdots

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

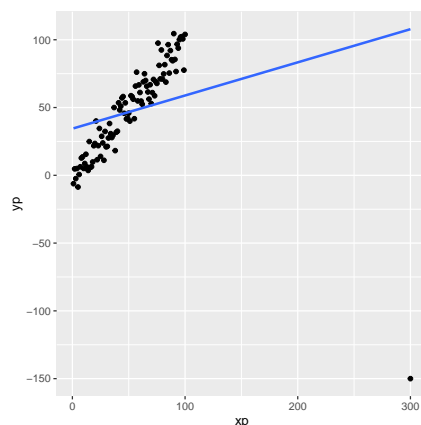
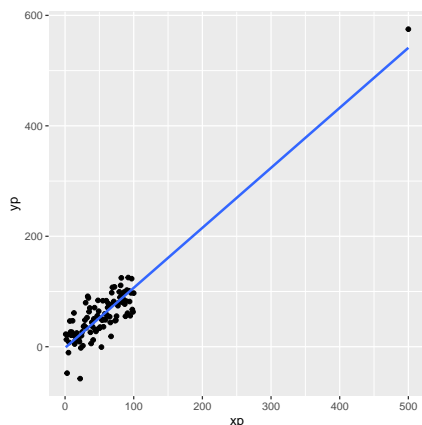
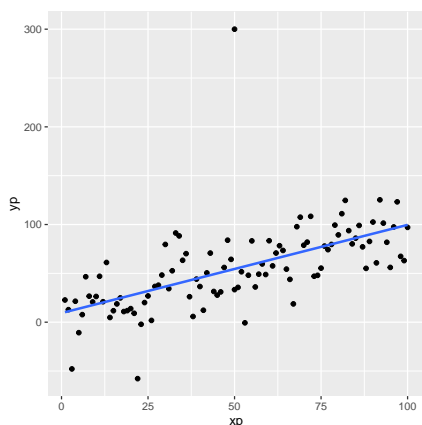
(b) flowerColor and caterpillars

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

(c) flowerColor and leafLength

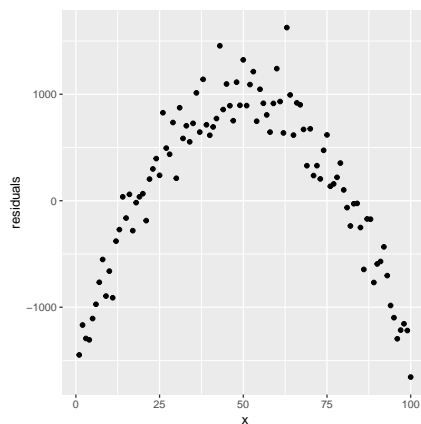
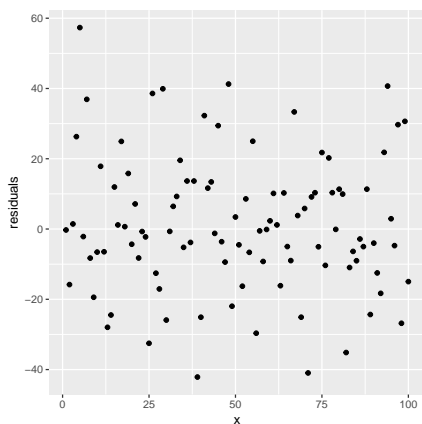
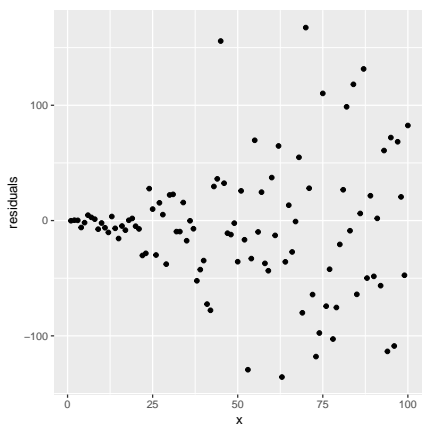
Solution. Side-by-side box plot, or hollow histogram.

Problem 4 (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 5 (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, (c) concerned: pattern in residual plot

Problem 6 (3 points). On a distant alien planet, 10% of the alien population has *hypertentaclooma*, 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 hypertentaclooma 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 hypertentaclooma. What is the probability that Zwq actually has hypertentaclooma? (If you don't know how to do this, write clearly that you don't know and 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 hypertentaclooma.

Problem 7 (1 point each). For each of the following, 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	<code>pnorm</code>	<code>qnorm</code>
t-distribution	<code>pt</code>	<code>qt</code>
chi-square distribution	<code>pchisq</code>	<code>qchisq</code>

- (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 between $-t^*$ and t^* .

Solution. `qt(0.75, 4)`

Problem 8 (3 points). Johann is a sociologist studying psychedelic use in the United States. In a simple random sample of 100 men and 100 women, all over the age of 21, Johann found that 22% of men and 12% of women have experience using psychedelics.

- (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, corresponding to each person in the sample. Two variables are recorded about each person: **sex** (male/female) and **psychedelics** (yes/no). Both are categorical variables.

- (b) Johann wants to test the hypothesis H_0 that equal proportions of men and women use psychedelics against the hypothesis H_A that different proportions of men and women use psychedelics. Help Johann calculate a p -value for this hypothesis test.

Solution. Let p_1 and p_2 be the proportion of men and women who use psychedelics, respectively. Under the hypothesis H_0 that $p_1 = p_2$, our best estimate for the overall percentage of Americans who use psychedelics 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

$$\sqrt{\frac{0.17 \cdot 0.83}{100} + \frac{0.17 \cdot 0.83}{100}} \approx 0.053.$$

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 that you do 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, while if you do use the pooled percentage, you get 0.0531.

You could also do this using a chi-square test. If you do this, you would find $\chi^2 = 2.87$ with 1 degree of freedom and a p -value of 0.09.

Problem 9 (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 conduct a hypothesis test using p -values with a significance level of 0.05 to determine if the average number of exercise days is different between Coloradans and Kansans. Be sure to state your hypotheses and conclusions clearly and in context.

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,

$$T \approx 9.5,$$

and that $2*\text{pt}(-9.5, 99)$ is tiny. Thus we do find evidence to reject the hypothesis that average number of days that Coloradans and Kansans exercise is the same.

Problem 10 (2 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, middle, high, 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 finds some instructions online to run the appropriate hypothesis test in R. She finds a p -value of 0.003, but she's unsure how to interpret this number. Help Anastasia interpret this p -value in context.

Solution. This means that, 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. Since this number is quite small, this study gives evidence that there *is* a difference in the average number of books read by students of the five levels.

Setup (for problems 11–14). 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 11 (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 12 (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 13 (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 Kaori gains is

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

Problem 14 (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) Help Daisuke use Kaori's data to construct a 99% confidence interval for the true proportion of times that the coin lands heads.

Solution. We calculate z^* as `qnorm(0.995)`, which is 2.58. Let p be the true proportion of times that the coin lands heads. We observed $\hat{p} = 0.96$. Thus, the standard error 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 $2.58 \cdot 0.006 \approx 0.016$. Thus the confidence interval is (0.944, 0.976).

Note that, 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 15 (2 points). What is the smallest number R such that the $R\%$ confidence interval constructed using the data from Kaori's diary contains 0.95? Show your work.

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 R , 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 $R = 89.3\%$.

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.

This page has been left blank for scratchwork.

