Math 154 – Fall 2021
Jo Hardin
sample problems, exam 1

1. Consider any of these dozen rules / plots `http://www.stat.columbia.edu/~gelman/communication/Wainer1984.pdf` [just the plots are provided here: `http://www.stat.berkeley.edu/~nolan/stat133/Fall05/lectures/DirtyDozen.pdf`]

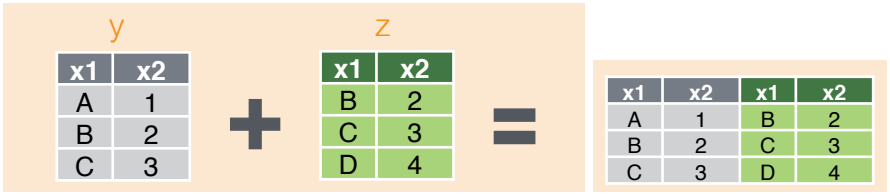
   (a) explain what is fundamentally wrong with it

   (b) explain how to improve the graphic with the goal of better communicating the information

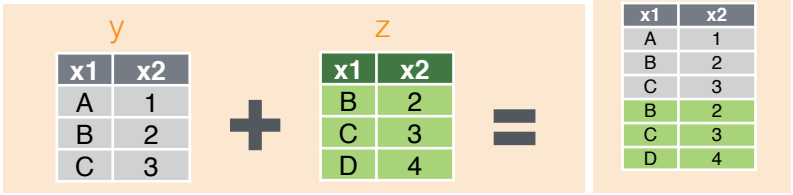
2. For each of the following diagrams, identify the data wrangling operation that is illustrated by the arrow or the equal sign in the diagram. There is only one correct answer for each arrow (or equal sign) in the diagram. Choose from the following operations:

   `arrange` `cbind` `filter` `mutate` `rbind` `pivot_wider` `select` `summarize` `group_by` `pivot_longer`

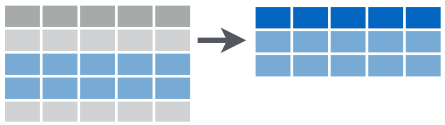
   (a)



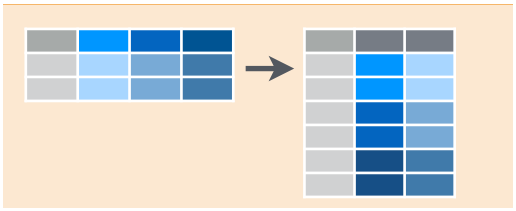
   (b)



   (c)



   (d)



(e)

(f)

3. This question concerns the `airlines` database which is very big. The database consists of four different types of information (in four different data tables): data about 148 million *flights* (occupying 15 gigabytes on disk), data about 1,491 *carriers* (airlines) ($\leq$ 1 megabyte), data about 3,376 *airports* ($\leq$ 1 megabyte), and data about 5,029 *planes* ($\leq$ 1 megabyte).

   (a) Suppose that a friend told you she was going to analyze this data by opening it in a spreadsheet application on her laptop. Why is this not a good idea?

   (b) Now suppose that the friend tells you that she wants to reproduce the route map for Delta Airlines during the month of her birth, which occurred in June of 1995. What data would be needed to create such a map? Describe how you would retrieve that data from the `airlines` data set. Be as specific as you can, but you do not need to write any code (answer should be in words).

4. The lonely recording device: This problem demonstrates the ways that empirical simulations can complement analytic (closed-form) solutions. Consider an example where a recording device that measures remote activity is placed in a remote location. The time, T, to failure of the remote device has an exponential distribution with mean of 3 years. Since the location is so remote, the device will not be monitored during its first two years of service. As a result, the time to discovery of its failure is X = max(T, 2). The problem here is to determine the average of the time to discovery (in probability parlance, the expected value of the observed variable X, E[X]).

   The analytic solution is fairly straightforward, but requires calculus. We need to evaluate:

$$E[X] = \int_0^2 2f(u)du + \int_2^\infty uf(u)du$$

   where $f(u) = 3exp(-3u)$ for $u > 0$. But is calculus strictly necessary here? Lay out the steps to estimate (/check) the value for the average time to discovery.

5. Nurses in an inner-city hospital were unknowingly observed on their use of latex gloves during procedures for which glove use is recommended. The nurses then attended a presentation on the importance of glove use. One month after the presentation, the same nurses were observed again. Here are the proportions of procedures for which each nurse wore gloves:

Nurse	Before	After	Nurse	Before	After
1	0.500	0.857	8	0.000	1.000
2	0.500	0.833	9	0.000	0.667
3	1.000	1.000	10	0.167	1.000
4	0.000	1.000	11	0.000	0.750
5	0.000	1.000	12	0.000	1.000
6	0.000	1.000	13	0.000	1.000
7	1.000	1.000	14	1.000	1.000

(a) Describe the null hypothesis and test statistic of interest.

(b) For a permutation test, describe the process for permuting the observations above. Hint: an important factor is recognizing the dependency between before and after for each nurse.

(c) When calculating a p-value, is interest in whether the observed difference is unusually large, unusually small, or either unusually small or large?

(d) If you are mainly interested in whether or not the effect of the intervention is significant at the 5% level, an alternative approach is to give a BS CI for the difference in means. After constructing a CI, how would you evaluate the interval to determine whether or not the null hypothesis is true?

6. When constructing a 97% confidence interval, which percentiles of the bootstrap distribution (of your test statistics) give the end points of the interval [for a percentile BS interval]?

7. Explain what is wrong with each of the following statements:

   (a) The bootstrap distribution is created by resampling with replacement from the population.

   (b) The bootstrap distribution is created by resampling without replacement from the original sample.

   (c) When generating the resamples, it is best to use a sample size larger than the size of the original sample.

   (d) The bootstrap sampling distribution will be similar to the true sampling distribution in shape, center, and spread.

8. We often talk about the standard deviation of the data and the standard error of the statistic. Explain the difference between the two (not the difference in the words, but the difference in the ideas/concepts) as if to someone who has not taken a statistics class before.

9. In trying to decide which type of bootstrap CI to use, what would you want to know about the intervals? That is, given infinite information (e.g., about the population), and infinite ability to compute (e.g., create intervals), the best interval would be the one that: _____ . Explain (you should be able to come up with at least two things about the interval).

10. What is the primary reason to bootstrap data? Or said differently, what comes from bootstrapping?

11. What is the primary reason to permute data (in the context of a permutation test)? Or said differently, what comes from permuting data?

12. Consider the following scenario: researchers are interested in testing the **ratio** (instead of difference) of average number of hours spent watching TV for kids who don't have a stay-at-home parent versus kids who do have a stay-at-home parent. The researchers collect a random sample of size 50 from each group and calculate the mean number of hours spent watching TV for each group.

    (a) Why can't traditional methods (e.g., those learned in intro stats) be used to approach this problem? (just 2-3 sentences.)

    (b) Suppose you have used a permutation approach to the problem.

i. One of the steps in the permutation test is "permute the data." Explain (using words or an example but not R code) what it means to "permute the data." (Explain only this one step in the algorithm.)

ii. Explain how a permutation approach to the research problem works. That is, give the **rationale** for a permutation test (not the algorithm for a permutation test). Suppose I don't know anything about permutation tests, but I do know statistics (i.e., demonstrate that you understand permutation tests.)

(c) Suppose you have used a bootstrap approach to this problem (same problem of trying to discern whether the average number of hours spent watching TV for kids who don't have a stay-at-home parent as compared to those that do is the same - as measured by the ratio).

i. One of the steps in the bootstrap algorithm is "bootstrap the data." Explain (using words or an algorithm, but not R code) what it means to "bootstrap the data." (Explain only this one step in the algorithm.)

ii. Describe what each of: θ, $\hat{\theta}$, and $\hat{\theta}^*$ are in this context.

iii. You and a friend are discussing bootstrapping. You both agree that the logic of building standard confidence intervals is quite clear in your heads:

$$P\left(z_{.05} < \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} < z_{.95}\right) = 0.9$$
$$P(\hat{\theta} - z_{.95}SE(\hat{\theta}) < \theta < \hat{\theta} - z_{.05}SE(\hat{\theta})) = 0.9$$

where the endpoints of the CI are random.

- Report why it is often not appropriate to use $z_{\alpha/2}$ (e.g., $z_{0.05}$, the quantiles from the standard normal distribution) quantiles to create confidence intervals for arbitrary parameters.
- How can a bootstrap-t interval be made when the value $z_{\alpha/2}$ is not appropriate? Give specific details (but no R code) for how to create the interval above without using the t multiplier.

13. The reverse percentile interval is a bootstrap confidence interval we did not discuss in class. It is based on the key assumption that:

$$\hat{\theta} - \theta \overset{D}{\approx} \hat{\theta}^* - \hat{\theta}.$$

That is, the *shifted* distribution of $\hat{\theta}$ is approximately equal to the *shifted* distribution of $\hat{\theta}^*$.

Let's say that through the bootstrap procedure, you find 1000 values for $\hat{\theta}^* - \hat{\theta}$. [Feel free to visualize / sketch those values in a histogram.]

The confidence procedure starts by considering the following probability statement:

$$P(c_1 \leq \hat{\theta} - \theta \leq c_2) = 1 - \alpha.$$

(a) (+8 pts) Using your bootstrap information, what values do you use to approximate c_1 and c_2? Be specific.

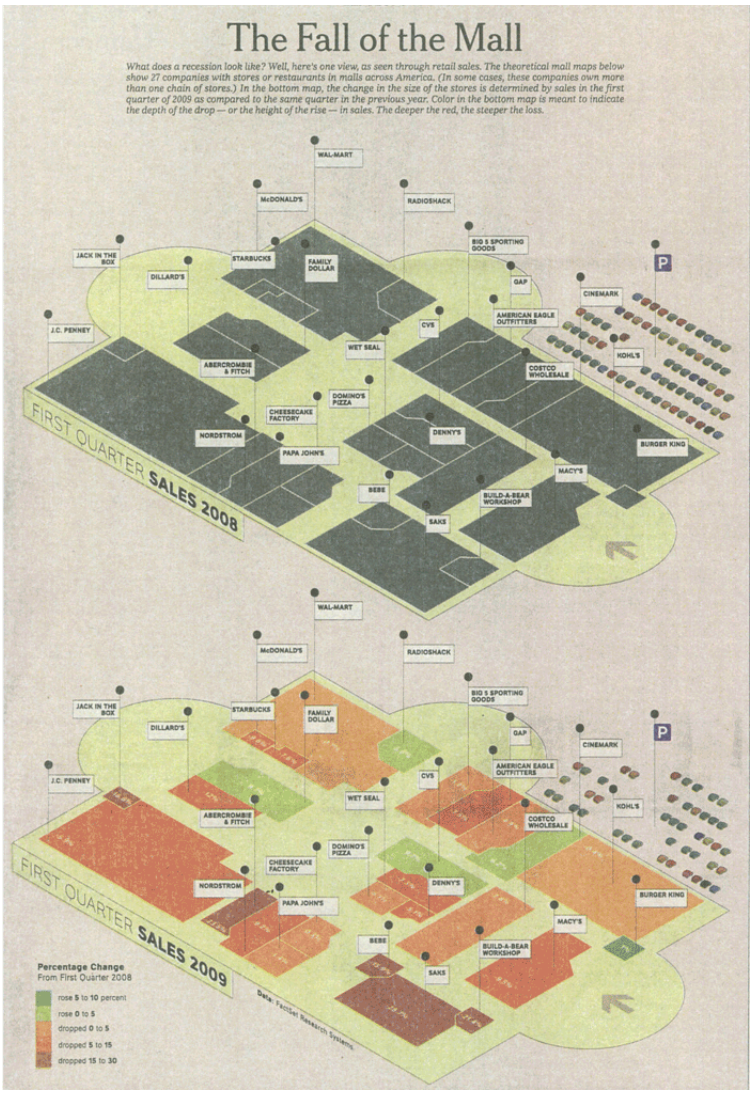
(b) (+8 pts) Given the values of c_1 and c_2 above, **derive** a $(1-\alpha)100\%$ confidence interval for θ. Show the steps you used to come up with the interval.

(c) (+8 pts) Comment on the following claim in two ways: (1) what is wrong with the claim? (2) what is a correct interpretation of the "95% of the time" part of the confidence interval conclusion?

Claim: The bootstrap confidence interval above captures 95% of the sample statistics, $\hat{\theta}$, across repeated samples.

14. Consider the NY Times mall graphic below.

What does a recession look like? Well, here's one view, as seen through retail sales. The theoretical mall maps below show 27 companies with stores or restaurants in malls across America. (In some cases, these companies own more than one chain of stores.) In the bottom map, the change in the size of the stores is determined by sales in the first quarter of 2009 as compared to the same quarter in the previous year. Color in the bottom map is meant to indicate the depth of the drop – or the height of the rise – in sales. The deeper the red, the steeper the loss. ("The Fall of the Mall", May 31, 2009, *NY Times*)



(a) List each of the variables displayed in the graphic, along with a few typical values for each.

(b) List the visual cues (at least 3) used in the data graphic and explain how each visual cue is linked to each variable.

(c) Assuming you can zoom in on the figure to read it clearly (i.e., don't say font size), name two (graphical) aspects of the figure that make it problematic.

(d) Examine the graphic carefully: Describe in words what *information* you think the data graphic conveys. Do not just summarize the data - interpret the data in the context of the problem and report what it means. [Note: *information* is meaningful to human beings - it is not the same thing as *data*.]

15. Suppose that we have two rectangular arrays of data, labeled *students* and *houses*. *students* contains information about individual Pomona students (e.g., student ID, name, date of birth, class year, dorm name, etc.). Each row in *students* contains data about one student. *houses* contains data about Pomona dorms (e.g., dorm name, capacity, street address, etc.). Each row in *houses* contains data about one Pomona dorm.

    Suppose further that we want to generate a student address book. The address book will consist of two columns of data: the first column will contain the student's name and the second will be the address of the dorm where he or she lives.

    (a) Describe, in words, a series of data wrangling operations that you could perform in order to achieve the address book. Be as specific as you can about what the operation will do and how it must be specified. Note: you do not have to write or reference any R code (and your answer must be in words).

    (b) It is important that every student appears in the address book, regardless of whether he or she lives on campus. But we'd like a concise address book, so empty entries should not occur. Would a full, left, or inner join be most appropriate here? Explain why.

    (c) Suppose now that only students from the Class of 2016 are to be included in the address book. What additional data wrangling step should you perform to achieve this? Again, be specific, but no need to write code.