Math 141

Midterm Exam Review Sheet

The midterm exam will take place on Thursday March 12. **You are allowed one 3x5 index card with handwritten notes on both sides, but otherwise, you are not allowed to use *any* notes, textbooks, or any other kind of resources on the exam.**

The midterm exam will involve several, but not all, of the concepts covered in this review sheet. The questions listed are meant to help you self-assess your understanding of these topics - they are not all necessarily representative of how the midterm questions will be asked. Feeling comfortable answering the conceptual questions and being able to complete the selected practice problems, while timing yourself and without viewing your notes, are very good indicators that you are ready for the exam.

The midterm will not involve evaluating or writing any R code beyond possibly showing you code with the functions that are mentioned below. However, the final is much more likely to involve R code. It may be beneficial to take this time to create a "cheat sheet" of the data wrangling, visualization, and modeling functions that we have learned so far. The act of creating a sheet that lists the functions taught in lecture and explains their uses in a way that is memorable to you will help your long-term memory of them.

Remember that office hours (both with me and the course assistants), the Slack channel, your peers, and the Reed tutoring program are available to you as resources while you prepare for the exam, in addition to the online lecture materials and your completed assignments.

Linear Regression

1. What do we mean by the "line of best fit"?
2. Draw a picture distinguishing data points, fitted values, residuals, and the regression line.
3. What do the `lm()` and `get_regression_table()` functions do?
4. Practice writing the model equation, the fitted line equation, and interpreting coefficients in these settings (examples in parentheses):
    a. One quantitative explanatory variable (End of Lecture 2/16)
    b. One categorical explanatory variable (HW 4 #3, end of Lecture 2/20)
    c. One categorical and one quantitative variable (Slides 17-21 in Lecture 2/23 - equal **and** varying slopes cases)

d. Multiple quantitative variables (Slides 7-8 in Lecture 2/25)

5. How can we use R^2 and adjusted R^2 to select a model? What is the difference between these two metrics?
6. What are the LINE assumptions and how do we assess whether they are met?

Probability

1. Define probability, and explain how it relates to the Law of Large Numbers.
2. Conditional Probability, Multiplication Rule, Bayes Rule, and the Law of Total Probability

    a. What are the definitions of these probability rules?
    b. When are each of them useful?
    c. Solve OI 3.22 (exit poll). You should get 0.487.
    d. For more practice later, without looking at your notes and timing yourself, revisit:
        - 2/9 lecture slide 10
        - HW 3 Exercise 1

3. Independence

    a. How do we know, mathematically, when two events are independent? Give an example of two independent events and an example of two dependent events.
    b. How does the intuition of the definition of independence relate to LINE assumptions?
    c. Without looking at your notes, write pseudo-code for how you would assess independence between gender and survival in the Titanic data set (lab 3)

Data Types and Visualization

1. When does it make sense to consider "year" as a quantitative variable vs a categorical variable? Does this depend on the task (ex. visualizing vs linear model)?
2. What types of visualizations/summaries are best for the following variable type(s):

    a. One quantitative
    b. One categorical
    c. One quantitative and one categorical
    d. Two categorical
    e. Two quantitative

(*Hint: Review slides from 1/30 and 2/2 if you are unsure*)

3. What summary statistics are relevant or helpful for interpreting boxplots? Histograms?
4. What are the elements of the grammar of graphics, and what purpose does each serve?

Sampling Schemes and Study Design

1. Revisit slide 25 from lecture on 2/11. Explain where and how simple random sampling, cluster random sampling, and stratified random sampling strategies each appear in the different stages, and what advantages they offer at each stage.
2. Without looking at your response, practice answering and explaining your solution to Exercise 4 from Lab 3.
3. Recall potential sources of sampling bias. Without looking at your response, practice answering and explaining HW 3 #3.
4. How can you distinguish an observational study from an experiment?

Sampling Distributions and Bootstrap Distributions

1. What is a sampling distribution?
2. **Example:** In a recent study, 23 rats showed compassion that surprised scientists. Twenty-three of the 30 rats in the study freed another trapped rat in their cage, even when chocolate served as a distraction and even when the rats would then have to share the chocolate with their freed companion. (Rats, it turns out, love chocolate.) Rats did not open the cage when it was empty or when there was a stuffed animal inside, only when a fellow rat was trapped. We wish to use the sample to estimate the proportion of rats that show empathy in this way.

    a. **Population**:
    b. **Sample**:
    c. **Parameter**:
    d. **Statistic**:
    e. Explain how we would create the true sampling distribution in this case, if we could.
    f. Explain how we could create a bootstrap distribution using our sample.