STA 35C – Homework 0 (Self-Assessment), due: Never

Instructor: Dogyoon Song

Instructions: This assignment is for your self-assessment and practice only. It will *not* be collected or graded, nor will solutions be provided. It reviews key topics from STA 35A and STA 35B, along with a brief check on your familiarity with R and RStudio. The symbol (♠) indicates topics that may have been skipped in your iteration of STA,35A/35B; don't worry too much if you find them difficult.

If you find any part especially challenging or need help with R or RStudio (e.g., installation), please make time to review your STA 35A/35B notes, textbooks, or online resources before STA 35C begins, and *attend discussion sessions in the first week (Tue, April 1, 2025).* If you need additional help, please feel free to attend office hours and consult with the instructor or TA during the first week of class.

Problem 1. Probability

(a) Suppose you roll a fair six-sided die twice.

(i) What is the probability that the sum of the two rolls is exactly 7?

(ii) What is the probability that at least one roll is a 6?

(b) A coin is flipped three times. Let A be the event "exactly two heads occur," and B be the event "the second flip is a head."

(i) Compute $\Pr(A)$ and $\Pr(B)$.

(ii) Compute $\Pr(A \cap B)$.

(iii) Use your results to find $\Pr(A \mid B)$.

Problem 2. Distributions

(a) Let X be a Binomial random variable $\text{Binomial}(n = 10, p = 0.3)$.

(i) What does X represent in words?

(ii) How would you calculate $\Pr(X = 3)$? (No need for an exact decimal; just give the formula or expression.)

(iii) How do you compute $\mathbb{E}[X]$ and $\text{Var}(X)$?

(b) A random variable Y is normally distributed with mean $\mu = 50$ and standard deviation $\sigma = 10$.

(i) Write the formula for the probability density function of $Y \sim \mathcal{N}(50, 10^2)$.

(ii) Describe how you would find $\Pr(45 \leq Y \leq 60)$ approximately (e.g., using the standard normal distribution).

(c) Suppose $Z_1, Z_2, \ldots, Z_n$ are i.i.d. random variables from a distribution with mean μ and variance σ^2. Let $\overline{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$ be the sample mean.

(i) What are the mean and variance of $\overline{Z}$?

(ii) If the Z_i are normally distributed, what is the distribution of $\overline{Z}$?

(iii) If the Z_i are *not* necessarily normal but n is large, how would you approximate the distribution of $\overline{Z}$?

Problem 3. Statistical Inference

(a) You collect an i.i.d. random sample $(X_1, X_2, \ldots, X_n)$ of size $n = 36$ from a population with unknown mean μ and known standard deviation $\sigma = 4$.

(i) Write down a 95% confidence interval for μ.

(ii) If you wanted to test $H_0 : \mu = 10$ versus $H_1 : \mu \neq 10$, which test statistic would you use, and why?

(iii) If you wanted a 99% confidence interval instead, how would it differ from the 95% interval? Explain briefly why one is wider or narrower than the other.

(b) Suppose you have data $X_1, \ldots, X_n$ from a population with unknown mean μ and unknown variance σ^2.

(i) How would you construct a confidence interval for μ if n is large and the data appear approximately normal? Could a normal-based (Wald-type) approximation work in this case, and if so, why?

(ii) (♠) How might the approach change (or not) if n is relatively small and the data are still approximately normal?

Problem 4. Linear Regression

(a) Suppose that we have data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and posits a simple linear regression model of the form $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

(i) How do we conceptually find $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing the sum of squared residuals?

(ii) Write down the closed-form formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$.

(iii) Briefly explain the interpretation of $\hat{\beta}_1$.

(b) Continuing with the same setup as in **(a)**:

(i) What is R^2, and what does it measure?

(ii) How do we compute R^2 using the total sum of squares (TSS) and the residual sum of squares (RSS)?

(iii) Why is R^2 sometimes called the "coefficient of determination"?

(iv) (♠) What is the adjusted R^2, and why might it be preferred over R^2? Provide a simple example.

(c) Consider testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ in the simple linear model. Which test statistic would you use, and how would you interpret the result?

(d) (♠) List the main assumptions of the simple linear regression model (e.g., linearity, independence, homoskedasticity, normal errors). Why are these assumptions important in practice?

Problem 5. R and RStudio

(a) Access: Confirm you have installed or can access both R and RStudio (on your own machine, a lab computer, or in the cloud). If not, please do so before classes begin.

(b) Basic Familiarity: Ensure you can comfortably answer the following questions.

- What are some frequently used data types in R (e.g., vectors, matrices, data frames, factors)?
- How would you read a CSV file into R (e.g., `read.csv()`)?
- Which command would you use to fit a simple linear regression model (e.g., `lm()`)?

What to Do If You Struggle

- **Review:** Revisit your STA 35A/35B notes, textbooks, or online resources.
- **Attend discussion sessions:** In the first week (Tue, April 1, 2025), the TA will help you with reviewing necessary concepts and possibly with R and RStudio.
- **Ask for Help:** If multiple areas are unclear, contact the instructor or TA, or form a study group with your peer students.
- **Practice:** Solve additional example problems or run small test scripts in R to strengthen your understanding.