STA 35C – Homework 0 (Self-Assessment)

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; do not 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 review your STA 35A/35B notes, textbooks, or online resources before the first graded homework is released, and bring any remaining questions to the first discussion section (Thu, September 24, 2026). If you need additional help, please attend office hours or consult 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 $\Pr(A \mid B)$.

(c) (♠) Let S denote the event that an email is spam and F the event that it is flagged by a spam filter. Suppose

$$\Pr(S) = 0.20, \qquad \Pr(F \mid S) = 0.90, \qquad \Pr(F \mid S^c) = 0.10.$$

(i) Compute $\Pr(F)$ using the law of total probability.

(ii) Compute $\Pr(S \mid F)$ using Bayes' rule.

Problem 2. Distributions

(a) Let X be a Binomial random variable 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 $\mathrm{Var}(X)$?

(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) Standardize Y and express $\Pr(45 \le Y \le 60)$ in terms of the standard normal cumulative distribution function. No decimal approximation is required.

(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 $H_1 : \mu \ne 10$, 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? Explain why a normal-based (Wald-type) approximation using the sample standard deviation can work, when n is large, without assuming that the population itself is normally distributed.

(ii) (♠) If n is relatively small and the population is approximately normal, which reference distribution should replace the standard normal distribution, and why?

Problem 4. Linear Regression

(a) Suppose that we observe data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and posit 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 to R^2 when comparing regression models with different numbers of predictors?

(c) Consider testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \ne 0$ 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, zero conditional mean, independence, and homoskedasticity). Briefly explain which conclusions rely on these assumptions and what additional role normality plays in exact small-sample inference.

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 R data structures and object types (e.g., vectors, matrices, data frames, and factors)?
- How would you read a CSV file into R (e.g., `read.csv()`)?
- How would you install and load an R package (e.g., using `install.packages()` and `library()`)?
- Which command would you use to fit a simple linear regression model (e.g., `lm()`)?

(c) End-to-End Check: Create a short R Markdown or Quarto document containing a brief sentence of written explanation and a code chunk that runs the following commands:

```
head(mtcars)
fit <- lm(mpg ~ wt, data = mtcars)
summary(fit)
plot(mpg ~ wt, data = mtcars)
abline(fit)
```

Render the document to PDF with both the code and its output visible. Identify the estimated slope and briefly interpret its sign. If the document does not render successfully, resolve the software or LaTeX installation issue before the first graded homework.

What to Do If You Struggle

- **Review:** Revisit your STA 35A/35B notes, textbooks, or online resources.
- **Attend the first discussion section:** Bring questions about these prerequisite concepts or R/RStudio setup to the first discussion section (Thu, September 24, 2026) so that the TA can help you with reviewing necessary concepts and possibly with software skills.
- **Ask for help:** If multiple areas are unclear, contact the instructor or TA, or form a study group with your classmates.
- **Practice:** Solve additional example problems or run small test scripts in R to strengthen your understanding.