

HW 3

Hypothesis Tests and Interval Estimation

September 26, 2025

Assignment Goals

1. Conduct and interpret a t-test for the slope
2. Explore the relationship between hypothesis tests and confidence intervals
3. Determine which type of interval (confidence interval for the slope, confidence interval for the mean response, or prediction interval) is most appropriate for a particular question
4. Interpret those intervals correctly in context
5. Unpack the standard error formulas for predictions and subpopulation means

! Important

Please be sure to read and reference the [Homework Instructions and Grading Rubric](#) document on Moodle before beginning this assignment. Please also see [a guide](#) for typesetting fitted regression lines and other mathematical expressions in R using LaTeX!

Q1 (MAWR LEAF WIDTHS)

The `LeafWidth` data set from the `Stat2Data` package (which you also worked with in Homework 2) contains information on samples of leaves from the *Dodonaea viscosa* plant (better known as the broadleaf hopbush), collected in Southern Australia between the years 1882 and 2001.

Each row in the data set represents the average of all the hopbush leaves measured at a particular site during a particular year. All measurements were made in millimeters.

You can load these data into R using the following commands:

```
library(Stat2Data)
data("LeafWidth")
```

(a) Suppose we view leaf width as a function of (i.e., as the outcome of) leaf length. Conduct a t -test to answer the question: does a linear relationship exist between the width and length of a Hopbush plant's leaves? Take your significance level to be $\alpha = 0.05$.

(b) Compute a 99% confidence interval for the slope of the model you fit in part (a) in two ways. First, compute the interval “by hand”, showing all of your work along the way. Then, use an R

function that computes confidence intervals automatically to double-check your calculations. *Note that some rounding error is okay.*

(c) Interpret your confidence interval from (b) in context.

(d) Does the confidence interval that you constructed in (b) include the fitted slope? Does that tell us anything about the statistical significance of the slope?

(e) Does the confidence interval that you constructed in (b) include zero? Does that tell us anything about the statistical significance of the slope?

(f) Imagine you work in a broadleaf hopbush lab with two colleagues, Tia and Tamara, who are currently discussing the interpretation of a 95% confidence interval for a different regression model. You hear Tia tell Tamara, “We can be 95% confident that the true slope is between 15.7 and 22.1.” Tamara replies, “Interesting, so this means there is a 95% chance that the true population slope is between 15.7 and 22.1?” To which Tamara responds, “Maybe? I’m not exactly sure what ‘95% confident’ means in this situation.”

Help your colleagues out! Explain to them, in your own words, what the phrase “95% confident” represents, as well as whether Tamara’s rephrasing of “95% confident” (as a “95% chance”) is appropriate.

Q2 (Inference for Slopes vs. Inference for Population Means):

If the test $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ results in a small p -value, can we be confident that the regression model provides a good estimate of $\mu(x^*)$, the mean value of the response in the subpopulation with a specific value, x^* , of the explanatory variable? Explain your reasoning.

When evaluating whether you think the answer is “yes” or “no”, it may be helpful to see if you can come up with potential counter-examples: scenarios in which the slope is statistically significant but the regression model does not provide a good estimate of $\mu(x^*)$.

Q3 (Understanding the Standard Error Formulas)

For each of the following subparts, you may find it helpful to reference the formulas for the standard error of the estimated mean response, $\hat{\mu}(x^*)$, and of the predicted future response, $\hat{y}_{\text{new}}(x^*)$, below:

$$SE\{\hat{\mu}(x^*)\} = \sigma \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$
$$SE\{\hat{y}_{\text{new}}(x^*)\} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

(a) At what value x^* of the explanatory variable will there be the most precise estimate of the subpopulation mean response? In other words, what choice of x^* makes the standard error of $\hat{\mu}(x^*)$ the smallest (and the corresponding confidence interval the narrowest)?

- (b) At what value x^* of the explanatory variable will there be the most precise prediction of a future response? In other words, what choice of x^* makes the standard error of $\hat{y}_{\text{new}}(x^*)$ the smallest (and the corresponding prediction interval the narrowest)?
- (c) Can the standard error of $\hat{\mu}(x^*)$ be brought increasingly close to zero as the sample size, n , becomes large? If not, what value does it become close to?
- (d) Can the standard error of $\hat{y}_{\text{new}}(x^*)$ be brought increasingly close to zero as the sample size, n , becomes large? If not, what value does it become close to?
- (e) Write a few sentences explaining, in your own words, why your answers to (c) and (d) make sense.

Q4 (Old Faithful)

The Old Faithful Geyser in Yellowstone National Park in Wyoming, USA is world famous for the regularity of its eruptions. As one of the members of the expedition that discovered Old Faithful wrote:

“It spouted at regular intervals nine times during our stay, the columns of boiling water being thrown from ninety to one hundred and twenty-five feet at each discharge, which lasted from fifteen to twenty minutes. We gave it the name of ‘Old Faithful.’” – Nathaniel Langford, 1871.

The dataset `faithful` (which comes pre-loaded in R; you do not need to load any packages or use any `data()` statements to access it) contains information on 272 eruptions of Old Faithful. For each of these eruptions, it records both the duration of that eruption (in minutes) and the wait time until the next eruption occurs (also in minutes). The first six rows of the dataset are shown below:

```
faithful |> head()
```

	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55

- (a) To help park guests plan their visit, Yellowstone National Park rangers would like to post a sign indicating the average wait time after a four-minute-long eruption, the median duration of an Old Faithful eruption. Use simple linear regression to obtain a “best guess” for this average wait time.
- (b) Construct and interpret in context an appropriate interval (at 95% confidence) for your “best guess” in part (a).
- (c) Suppose you and a friend have been road-tripping across the United States and are currently at Yellowstone National Park. You just finished watching a four-minute-long eruption of Old Faithful. Your friend would like to run back to the car to grab some snacks for you both, but is worried about

missing the next Old Faithful eruption. Use simple linear regression to obtain a “best guess” for how much time your friend has (i.e., how long the wait until the next eruption will be).

(d) Construct and interpret in context an appropriate interval (at 95% confidence) for your “best guess” in part **(c)**.

(e) Are your answers to parts **(a)** and **(c)** the same? If so, explain why it makes sense for them to be the same. If not, explain why they are different.

(f) Are your answers to parts **(b)** and **(d)** the same? If so, explain why it makes sense for them to be the same. If not, explain why they are different.