# Motivating Scenario

Maria is a chemist interested in calibration of fluorescence spectrophotometry information from an experiment that they have recently run. They have collected a dataset where the concentration (pg/mL) was controlled over four unique values to create 3 replicates of the intensity at each value. The data from the experiment are provided below.

|  |  |
| --- | --- |
| **Concentration (pg/mL)** | **Intensity** |
| 2 | 1.476413 |
| 2 | 0.874620 |
| 2 | 2.096216 |
| 5.3 | 1.222438 |
| 5.3 | 3.814496 |
| 5.3 | 2.198255 |
| 8.7 | 3.327383 |
| 8.7 | 4.396896 |
| 8.7 | 5.358306 |
| 12 | 6.573142 |
| 12 | 7.847182 |
| 12 | 6.111933 |

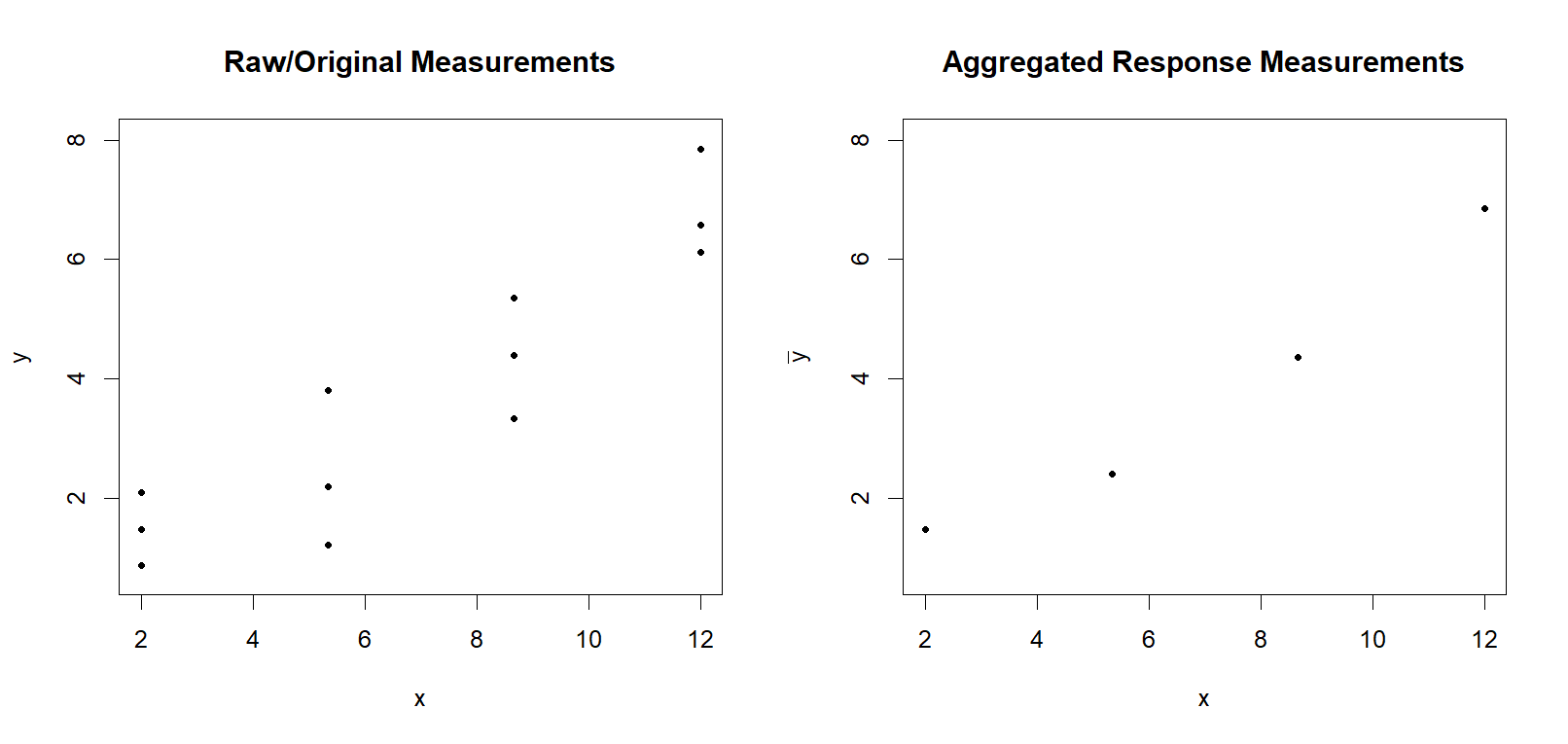
Maria’s colleague, Adam, suggests that Maria should average the replicated values of intensity for each concentration before creating a calibration line:

|  |  |
| --- | --- |
| **Concentration (pg/mL)** | **Intensity** |
| 2 | 1.482417 |
| 5.3 | 2.411729 |
| 8.7 | 4.360862 |
| 12 | 6.844086 |

What should Maria do -- Use the original set of 12 measurements or take Adam’s advice to reduce to four averaged response values to fit the calibration model?

# Exploration toward a solution

Maria began the investigation by creating the scatterplots shown below (scaled to have equivalent horizontal and vertical axis scales)



1. In your opinion and just based on the scatterplots, which approach (original measurements, aggregated responses) seems to have the strongest linear relationship? Why?
2. Maria fit a least squares regression model using each of the datasets and found that the estimated model was exactly the same! However, the R2 values differed.

A comparison of measurements and measurements

AI-generated content may be incorrect.  
If you were only using this information, which approach seems to be associated with a “better” model? Why?

1. For each of the dataset approaches, Maria created the traditional linear regression model summary. She was surprised to see that, even though the fitted models were equivalent, the hypothesis tests of a linear relationship produced different results.

**Model using the 12 original measurements:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.01240 0.60500 -0.020 0.984

predictor 0.54102 0.07629 7.092 3.33e-05

Residual standard error: 0.9849 on 10 degrees of freedom

Multiple R-squared: 0.8341, Adjusted R-squared: 0.8176

F-statistic: 50.29 on 1 and 10 DF, p-value: 3.328e-05

**Model using the four aggregated response measurements:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.01240 0.59021 -0.021 0.9852

groupPredictor 0.54102 0.07443 7.269 0.0184

Residual standard error: 0.5547 on 2 degrees of freedom

Multiple R-squared: 0.9635, Adjusted R-squared: 0.9453

F-statistic: 52.84 on 1 and 2 DF, p-value: 0.0184

Examine Maria’s results closely. What aspect(s), beyond p-values, do you notice differ?

1. If the research goal were to find evidence of a linear relationship between Concentration and Intensity, which dataset approach is associated with finding a statistically discernable relationship at α=0.01? Briefly justify your answer.
2. What is one item from the data and/or analysis that you found most surprising? Why?

## Review: Errors in Hypothesis Testing

When implanting a hypothesis test, you set the **significance level (α)** before performing any calculations.

**α = P(Type I error) = P( rejecting H0 | H0 true)**

* The significance level is typically set to a value between 0.01 and 0.1.

**Remember, the significance level (α) could be interpreted as a long run error rate.** For example,

Suppose we had the ability to gather 100 independent samples, each sample of size n, from the population of interest. When the null hypothesis is true, we expect approximately (100α) of the hypothesis tests to result in a decision where the null hypothesis is incorrectly rejected.

1. As the researcher, you set α=0.01 and you (somehow) can gather 10,000 independent samples, each sample of size 6, from the population of interest. If the null hypothesis were true, approximately how many of your hypothesis test results should result in an incorrect decision?
2. 1
3. 6
4. 10
5. 60
6. 100
7. 600
8. Cannot be determined
9. In a simulation study with α=0.05, 1,000 independent samples, each of size 10, were generated from a population distribution where the null hypothesis is true. Which of the following numbers might be associated with the number of times the null hypothesis was rejected, if the statistical methodology used to implement the test works as expected?
10. 48
11. 52
12. 47
13. 54
14. Any of the above is highly plausible

## Power: A Long-Run Rate of Correct Decisions

The significance level describes the probability of rejecting the null hypothesis, H0, when the null hypothesis is true. This is one kind of mistake that could occur when implementing a hypothesis test. There is a second kind of mistake that could occur.

**Type I Error:** Reject H0 when H0 is true

**Type II Error:** Do not reject H0 when H0 is false

Example: You are testing hypotheses about a population slope, β1:

H0: β1 ≤ 8  
 H1: β1 > 8

You gather a sample of size 11, conduct a hypothesis test at significance level α=0.05 with a resulting p-value of 0.07. The null hypothesis was not rejected. However, the true population slope is β1 = 9.1. Your hypothesis test resulted in a Type II error.

**β** is common notation used to denote the probability that H0 is not rejected when H0 is false. This is not the same as β1 or β0 parameters in a simple linear theoretical model.

1. For the hypotheses below, what value(s) of the true population slope could result in a hypothesis test with a Type II error?  
    H0: β1 ≥ -5  
    H1: β1 < -5
   1. (-5, 5)
   2. [-5, ∞)
   3. (-∞, -5]
2. For the hypotheses below, what value(s) of the true population mean could result in a hypothesis test with a Type II error?  
    H0: β1 = 2  
    H1: β1 ≠ 2
3. (2, ∞)
4. (-∞, 2)
5. Both choices above
6. None of the choices above

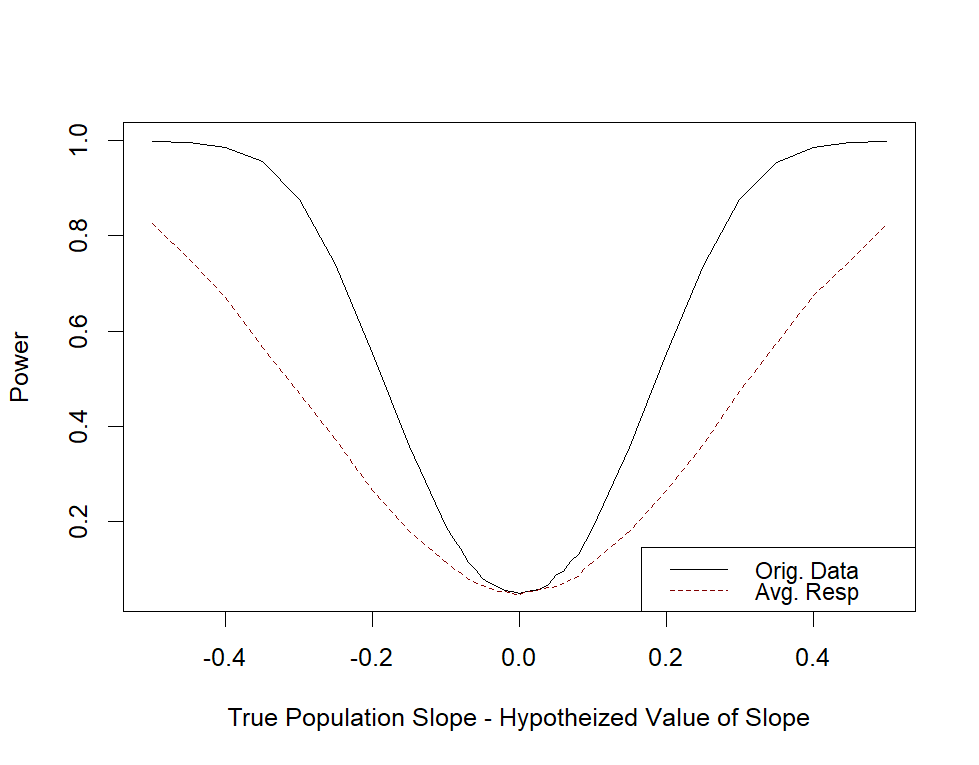
Instead of discussing β = Pr(Type II error), most researchers are interested in the probability of rejecting H0 when H0 is false – that is, what is the probability of making a correct decision that we have support for H1 when, in reality, H1 is true?

**Power (1-β)**: Probability of rejecting H0 when H0 is false.

Similar to the significance level, power can be interpreted as a **long-run rate of making a correct decision** when H0 is false. For example,

Suppose we had the ability to gather 100 independent samples, each sample of size n, from the population of interest. When the null hypothesis is false and we conduct a hypothesis test with significance level, α, we expect approximately (100(1-β)) of the hypothesis tests to result in a decision where the null hypothesis is correctly rejected.

1. As the researcher, you set α=0.05 and you (somehow) can gather 10,000 independent samples, each sample of the same size which allows your specified test to have power of .90, from the population of interest. If the null hypothesis were false, approximately how many of your hypothesis test results should result in a correct decision (e.g. H0 is rejected)?
2. 10
3. 90
4. 100
5. 900
6. 1000
7. 9000
8. Cannot be determined
9. In a simulation study with α=0.05, 1,000 independent samples, each of the same size which allows your specified test to have power of .78, were generated from a population distribution where the null hypothesis is not true. Which of the following numbers might be associated with the number of times the null hypothesis was rejected, if the statistical methodology used to implement the test works as expected?
10. 33
11. 52
12. 77
13. 325
14. 779
15. Any of the above is highly plausible
16. Shown below are power curves (e.g. long-run error rates) associated with testing the slope in linear regression. In this case, we specifically examine a scenario similar to Maria’s data: Four unique values of the predictor, each replicated three times. (There are other underlying conditions that we do not explicitly address as well.)

  
What trend(s) do you notice from this graph?

1. Tests with larger/higher power are more preferable (essentially, in the long-run, fewer mistakes occur). Based on the power curve graph, which approach for the data format would you recommend? Briefly justify.

# Reflection

1. Based on your observations from this activity, what are the two most interesting items or “take-aways” that you observed? How would you relate these to future approaches you might seek when collecting data where a predictor has repeated values?
2. Do you have any remaining questions or concerns that arose from this activity? If so, please include those below so that the instructor can reach out to you with more information.