# Data Analysis 7: Comparing Multiple Means

In this data analysis assignment, you'll apply inference procedures we've learned in class in two different data scenarios and questions of interest.

If you haven't already done so, work through the tutorial provided on the Data Analysis 7 Canvas page. Once you've worked through the tutorial, write up your responses to the questions listed throughout the tutorial. The same questions are included below to help you format your submissions.

Submit a PDF copy of your responses to Gradescope by the deadline stated on Canvas.

## Part 1: Comparing Lizard Speeds

### Question 1 (1 point)

In this portion of the assignment, you'll estimate and test the difference in average speeds of the two lizard species when completing the maze. Write a question of interest for which a two-sided t test for the difference in populations means can be used to answer.

What are the average speeds of Sagebrush lizards and Western fence lizards when completing the maze are equal?

### Question 2 (2 points)

Compute summary statistics for each sample.

A. (1 point) Report the sample mean, sample standard deviation, and sample size for the sampled maze speeds of Sagebrush lizards. Hint: use the mean(), sd(), and length() functions in R to do this. Note that the data are each stored in individual vectors (not a shared data frame), so you can simply input each vector name in the summary functions to compute the corresponding values.

   mean (Sagebrush) = 1.612692
   sd(Sagebrush) = 0.3241056
   length (Sagebrush) = 26

B. (1 point) Report the sample mean, sample standard deviation, and sample size for the sampled maze speeds of Western fence lizards.

   mean(Western_fence) = 2.314545
   sd(Western_fence) = 0.5554983
   length(Western_fence) = 22

Question 3 (1 point)

The parameter of interest is the difference in the population means of maze speeds for Sagebrush and Western fence lizards, $\mu_S - \mu_W$. Using the sampled data, calculate the point estimate for this parameter of interest.

Point estimate = $\mu_S - \mu_W$ = 1.612692- 2.314545 = -0.701853

Question 4 (1 point)

The point estimate in the previous problem has an associated estimate for the standard error. Calculate this standard error estimate and report the value below.

Standard Error = sqrt ((0.3241056^2 / 26) + (0.5554983^2 / 22)) = 0.1344115

Thus, standard error is 0.1344115.

Question 5 (0.5 points)

The point estimate reported in question 3 also has associate degrees of freedom. Using one of the two tools provided in the Week 9 module on Canvas for calculating Satterthwaite degrees of freedom, calculate the degrees of freedom for this point estimate.

Satterthwaite degrees of freedom = 32.57

Question 6 (1 point)

Use the boxplot to assess the sample size condition required to construct a two-sample t confidence interval and to perform a two-sample t test for the difference in population means.

The width of the Western fence lizard boxplot is slightly wider than the width of the Sagebrush lizard boxplot. Additionally, both boxplots appear to be approximately symmetric. These observations suggest that the sample size condition required for constructing a two-sample t-confidence interval and performing a two-sample t-test for the difference in population means is met.

Question 7 (2 points)

Construct a 90% confidence interval for the difference in the average maze speeds of the two lizard species. Report the numerical values for the lower and upper bounds of the interval. Show your work.

Based on previous question we already know that: n1 = 26, n2= 22, s1 = 0.3241056,

s2 = 0.5554983, Critical value for 90% confidence interval is: qt (0.95, 32.57) = 1.693006, and point estimate: -0.701853, SE= 0.1344115, Satterthwaite degrees of freedom = 32.57.

Lower bound = (1.612692 -2.314545) - 1.693006* 0.1344115 = -0.9294125

Upper bound = (1.612692 -2.314545) + 1.693006* 0.1344115 = -0.4742935

Thus the 90% confidence interval for the difference in the average maze speeds pf the two lizard species.

Question 8 (1 point)

Set up the null and alternative hypotheses needed to answer the two-sided question of interest from question 1.

$H_0$: The average speeds of Sagebrush lizards and Western fence lizard when completing the maze are equal.

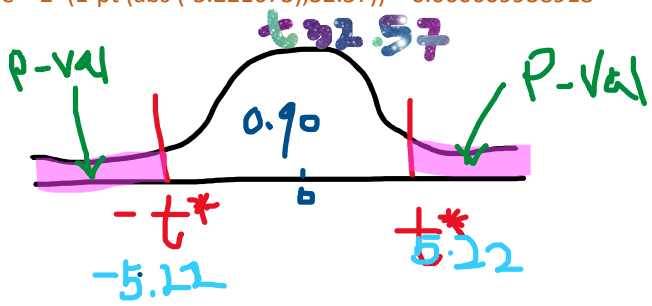
$H_A$: The average speeds of Sagebrush lizards and Western fence lizard when completing the maze are not equal.

### Question 9 (5 points)

Perform the hypothesis test using a significance level of $\alpha = 0.1$.

A. (1 point) Using the sample statistics determined in question 2, calculate the test statistic for the two-sample t test for the difference in population means.
Test statics = (1.612692 -2.314545) / sqrt ((0.3241056^2 / 26) + (0.55549829^2 / 22)) = -5.221673

B. (1 point) Using the degrees of freedom from question 4 and the test statistic from the previous question, calculate the p-value. Hint: sketch the curve and shade in the p-value! p-value = 2*(1-pt (abs (-5.221673),32.57)) = 0.000009938918




C. (1 point) Using the p-value from the previous problem, complete the following p-value interpretation by filling in the blanks.

*The probability of observing a difference in sample means of* ___-5.221673___ *or a difference even more extreme is* __0.000009938918_____, *when* _____*the population mean of two-sample difference is the same* _____.

D. (2 points) Write a 4-part conclusion to answer the question of interest.
There is convincing evidence to suggest the average speeds of Sagebrush lizards and Western fence lizard when completing the maze are equal. We reject null hypothesis since our p-value is less than 0.1. We are 90% confident that the average speeds of Sagebrush lizards and Western fence lizard when completing the maze are between -0.9294125 and -0.4742935 with a point estimate of -0.701853.

## Part 1: ANOVA

### Question 10 (2 points)

Write the null and alternative hypotheses for evaluating whether the average number of hours worked varies across the five groups.

$H_0$: The mean number of hours worked in the same(equal) across all five groups (Less than HS, HS, JR Coll, Bachelor's, Graduate)

$H_A$: At least one population means differs from at least one other, indicating that the average number of hours worked varies across the five groups

### Question 11 (1.5 points)

Using the information provided, assess whether the following conditions necessary to accurately perform an ANOVA F test are met:

A. Are the observations in the study independent?
yes, the observation in the study is independent.

B. Are the sample sizes sufficiently large? (Hint: the *n* row of the table above provides the sample sizes of each group.)
yes, sample size is sufficiently large, because all sample size in 5 groups is larger than 30

C. Is the variation in the groups about equal from one group to the next? (Hint: use the spread of the boxplots and standard deviation values from the table to assess this condition.)
Yes, the variation in the groups about equal from one group to the next. The mean is about 40.

Question 12 (1 point)

To assess whether there is a significant difference in the average number of hours worked between one or more of the groups, we need to determine the mean squares between groups (MSG) and the mean squares within groups (MSE). Each of these values has an associated degree of freedom.

A. Determine the degrees of freedom associated with the MSG.
Df of freedom associated with MSG: 5-1 = 4

B. Determine the degrees of freedom associated with the MSE.
Df of freedom associated with MSE:1172-5 = 1167

Question 13 (1 point)

An ANOVA was performed in R. The estimate for the mean squares between groups is MSG = 501.54 and the resulting *F* statistic is equal to 2.189. Determine the average variation within each group. That is, calculate the MSE.

We know that F= MSG/MSE therefore, MSE = MSG/F = 501.54/ 2.189 = 229.118319

Question 14 (1 point)

Using the *F* statistic from question 13 and the two values for the degrees of freedom in question 12, calculate the p-value for this test.

p_value = 1-pf (2.189,4,1167) = 0.06819242

Question 15 (2 points)

Using the p-value calculated in question 14, write a conclusion for this ANOVA F test using a significance level of $\alpha$=0.05. (Hint: your conclusion should include a statement of evidence in favor of the alternative and a statement as to whether the null hypothesis is rejected or not.)
p-value= 0.06819242 > $\alpha$ =0.05
There is no evidence to suggest that the average number of hours worked in the same(equal) across all five groups (Less than HS, HS, JR Coll, Bachelor's, Graduate). We fail to reject null hypothesis since our p-value is greater than 0.05.

**Gradescope Page Matching (2 points)**

When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page "Submitting Assignments in Gradescope". Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item "Gradescope Page Matching".