Problem Set #2: t-tests, correlations, and regressions

(1) Kai, an undergraduate in a CSUN Ecology class, noticed that in the intertidal area of southern California, there did not seem to be many sea urchins at a site that had relatively flat rocks, compared to a nearby site that had more complex topography. She wanted to know if there was statistical support for this casual observation. So she sampled densities of sea urchins in randomly placed 1-m^2 quadrats at the two sites. She sampled 10 replicate quadrats at each of the two sites and collected these data:

Site 1 (flat rocks): 3, 3, 4, 5, 2, 3, 2, 3, 4, 5
Site 2 (complex rocks): 3, 5, 2, 1, 7, 8, 7, 4, 11, 9

(a) (2pts) First calculate the following statistics for each of the two sites:

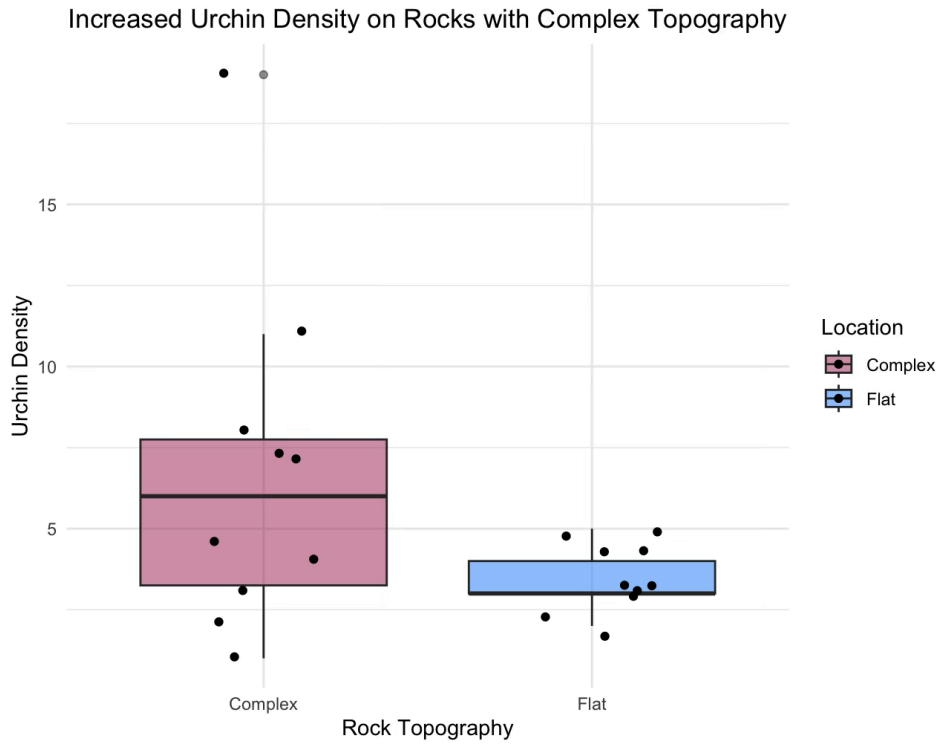
	Site 1	Site 2
Mean	3.4	5.7
Standard deviation	1.1	5.3

(b) (2 pts) Next use a two-sample *t*-test to test whether the mean density of urchins differed statistically between the two sites. Try both the more traditional "pooled variances" test and the "separate variances" test.

Pooled variances *t*-test	Separate variances *t*-test
t = 2.1345	*t* =2.1345
p =0.04681	*p* =0.05622
Are the means significantly different?	Are the means sig different?
Yes	No

Which *t*-test do you think is most appropriate for these data? Why?
I believe the most appropriate is the pooled test because it allows for more degrees of freedom. The welch's test is very conservative thus reducing the allowed variance of the data, this makes it less representative of the real population.

(c) (2 pts) Make a publication-quality graph of the data provided in question 1. Show means and standard error of the mean (SEM). Provide a figure legend. Write one sentence to be used in the results section of this paper that describes the conclusion based on your hypothesis testing.
Graph Below



(2) The El Segundo blue butterfly is endangered and depends on coast buckwheat, which has declined in abundance due to coastal development. To facilitate recovery of the butterfly, native coastal plants are being restored in certain areas. The literature suggests that a density of 4 coast buckwheat plants per 25-m^2 is necessary to support the butterfly. The data in the file "CoastBuckwheat.csv" contains the density of buckwheat in replicate 25 m^2 quadrats in an area that is intended to be a restored habitat for the El Segundo blue butterfly. Use an appropriate *t*-test to test the hypothesis that buckwheat plant density has reached the 4-plants-per-25-m^2 standard in this restored area and thus the plant restoration is a success.

(a) (2 pts) Do the data meet the assumptions of a t-test?
Yes

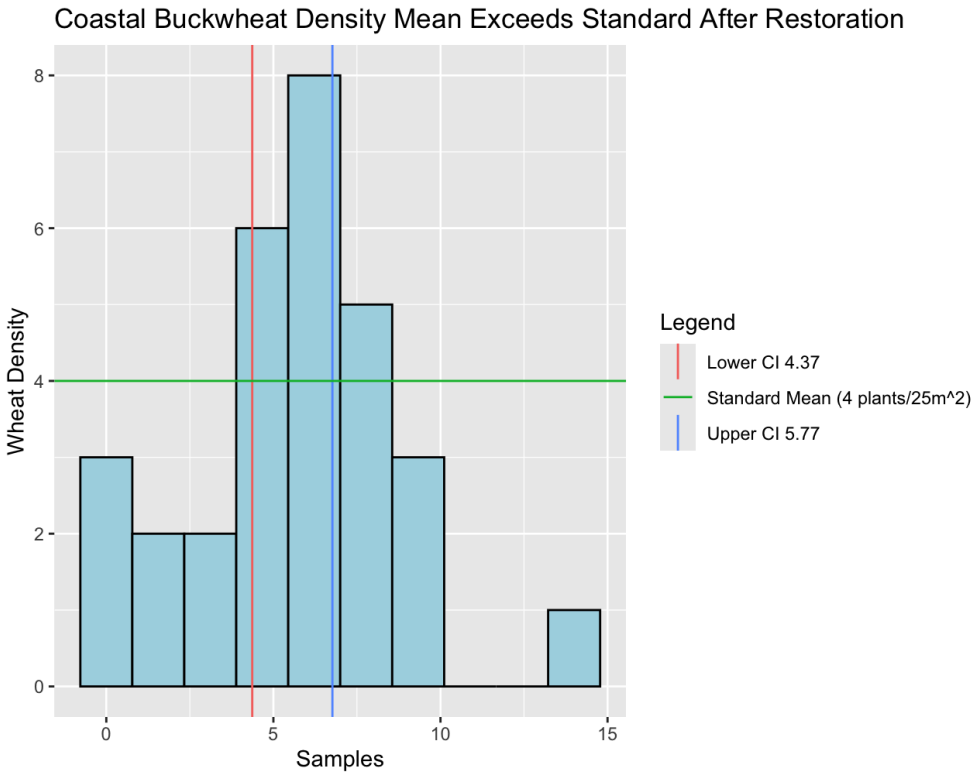
The data is normally distributed p=0.1408 > 0.05. Variances are equal. Observations are independent.

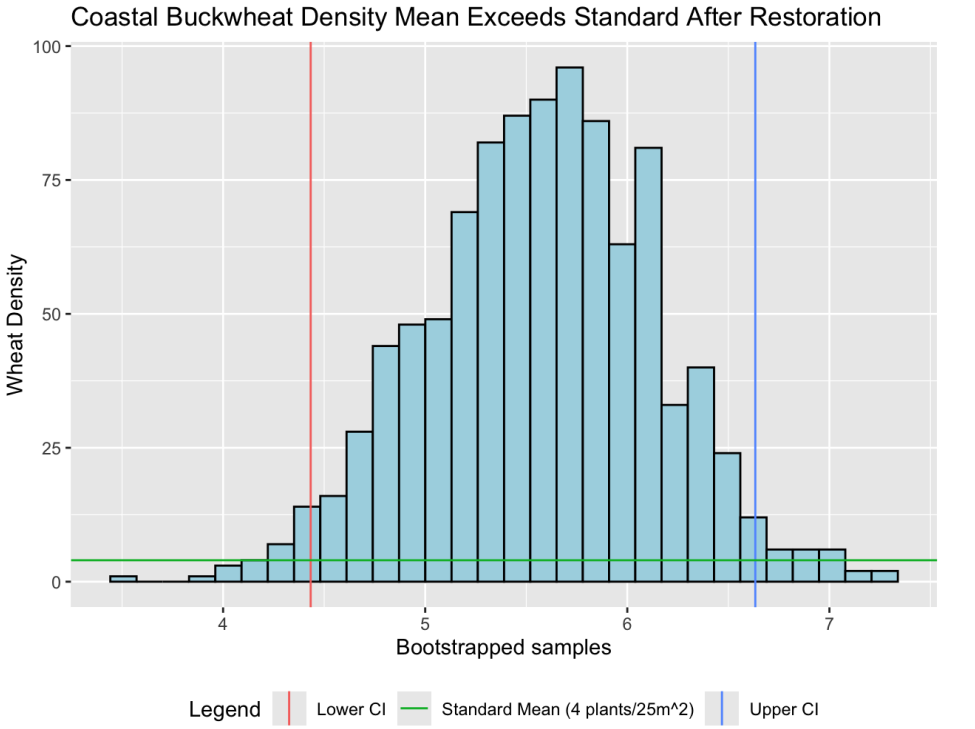
(b) (2 pts) Write a sentence that states whether this standard (4-plants-per-25-m^2) been met (support your answer with t, df, and P).

My one-tailed test (with mu=4) shows that the standard has been met because the p-value < 0.05 shows the significance and the t-value is above the necessary threshold (1.699) in t-distribution when using 29 degrees of freedom.

t = 2.6708, df = 29, p-value = 0.01228

(c) (3 pts) Make a bar graph, showing the mean +/- 95% CI and indicate the null hypothesis with a horizontal line on your plot. Nullhypothesis=standard expected mean





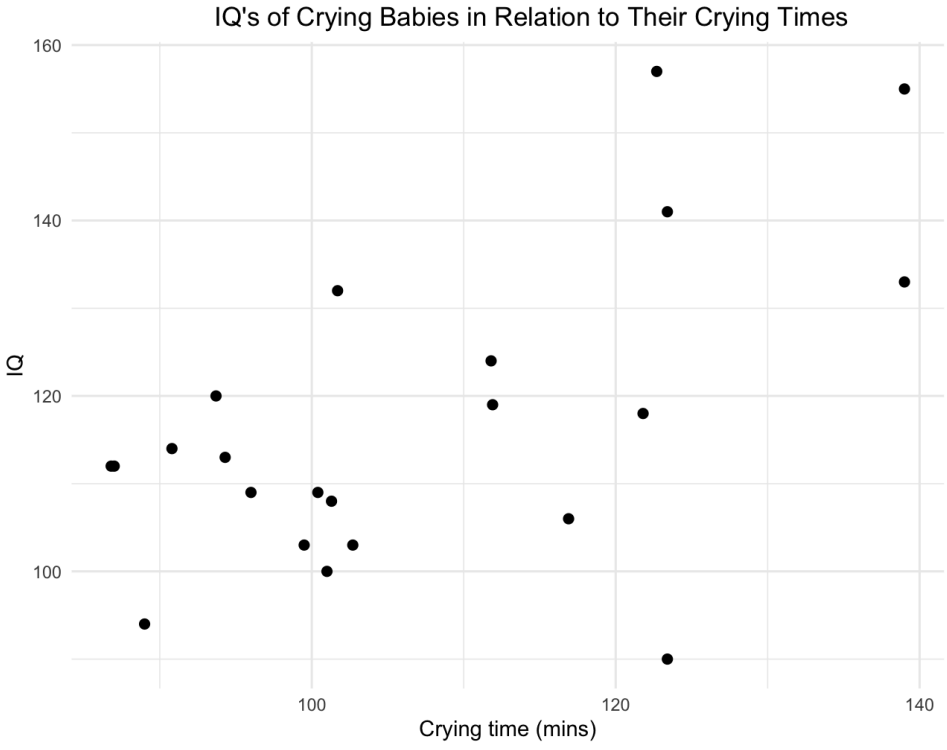
(3) (5 pts) You are testing the effects of a newly developed sports drink on athletic performance. You recruit 20 student athletes at CSUN for the experiment. Each student runs 5 km as fast as they can, twice, once after drinking a liter of the new sports drink and once after drinking a liter of water. The two runs are separated by a week, and the order is randomized among students (i.e., some get sports drink the first time and water the second, and others get the reverse order). The data for each treatment (in seconds) are in the file "RunTimes.csv". Use an appropriate *t*-test to test the hypothesis that the sports drink improves athletic performance. Is there compelling evidence that the drink altered running speed? Support your answer with appropriate statistics.

Doing a paired t-test because we have data from same individuals with or without treatments and assumptions are met. H0= does not improve, Ha= improves, Hb=altered running speed.

The p-value = 0.01296 showing a significant difference in the running speeds. The mean difference = 15.25 showing that runs with water were 15.25s longer that with the sportsdrink. The sports drink altered and improved run times.

(4) In a psychology experiment, investigators seek to determine if there is an association between how much babies cry and their IQ. The data are minutes spent crying during a day (taken at 3 months old) and IQ at age 3. The data are in the worksheet "*cryingbabies*".

(a) (2 pts) First, graph these data using a scatterplot.



(b) (5 pts) Analyze the data with three different tests of correlation: Pearson's *r*, Spearman's rho, and Kendall's tau. Is there an association between crying babies and their IQ? Does the answer depend on which test of correlation is used? Do you think one of these correlation tests is more appropriate than the others?

Looking at the scatterplot along with results from the tests, I believe that Spearman's rho is the most trustworthy since we observe a monotonic increase in IQs as babies cry longer. Since there are some outliers, this is not linear (constant rate of increase or decrease).

Pearsons r = 0.5655392. Closer to 1 than it is to 0, so we know that there is a positive relationship + p-value is 0.006087 showing significance. This says there is correlation

Spearmans rho = 0.3935539 (issues with exact p-value even after ranking) also a positive relationship + p value = 0.006087. There is correlation.

Kendalls tau = 0.2450772 (issues with exact p-value even after ranking) positive relationship + p-value = 0.1136 suggest the correlation is not statistically significant.

(5) (5 pts) In 2000, using statistics could have changed history. The US Presidential election was a contest between George W. Bush (Republican) and Al Gore (Democrat). It came down to a recount in Florida to determine the outcome of the election. One concern was whether the "butterfly ballot" used in Palm Beach County caused voters intending to vote for Gore to accidentally cast their vote for Pat Buchanan (a conservative candidate for the Reform Party). We could have used statistics to analyze the relationship between votes cast for Bush and those for Buchanan by county (n=67 counties in Florida). We would expect their votes to be correlated, as both candidates had similar political views. We could then determine whether the vote totals for Palm Beach County for Buchanan were similar to other counties, or an outlier with respect to voting patterns for all other counties in Florida. The data are in the worksheet *butterflyballot*. They represent vote totals (in thousands) by county for the state of Florida. The last observation in the data set is for Palm Beach County.

(a) If you leave this data point out of the analysis, are vote totals between Bush and Buchanan correlated?
Yes they are correlated because pearsons, spearmans and kendalls tests show a high positive correlation. Their scores are all close to +1 and they have p values showing statistical significance.

(b) If you included Palm Beach County, does it appear to be outlier with respect to the other counties?

Palm County is the outlier because the correlation scores change dramatically when it is added back to the dataset.

(c) Based on the above, do you think the butterfly ballot affected the outcome of the election?
Yes because Palm County was the only group to use the butterfly ballot.

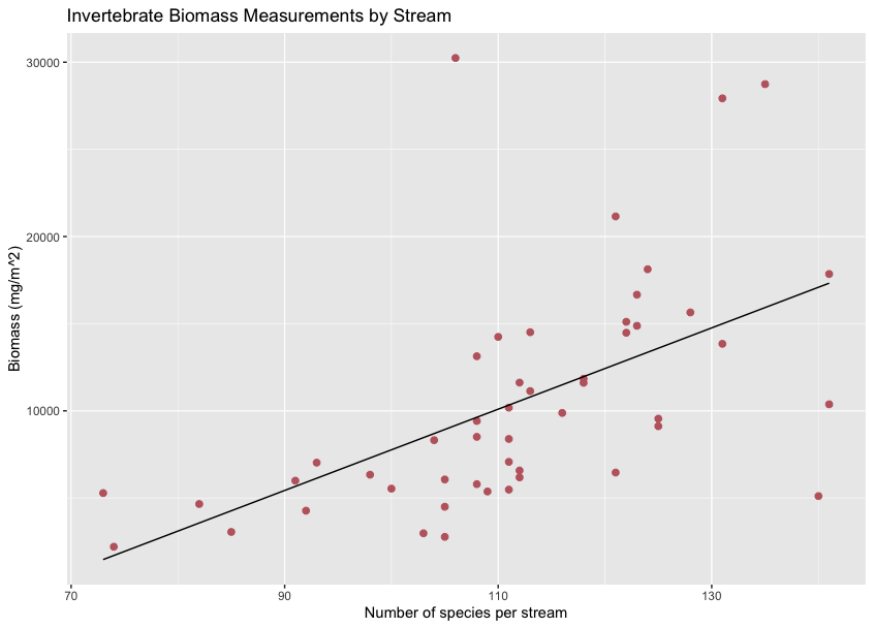
(6) (5 pts) We are interested in whether the age of a pregnant woman determines how much weight she gains during pregnancy. A few data are gathered:

Age (years)	Weight Gain (kg)
15	6.32
16	7.04
17	6.91
19	7.56
21	13.01
22	10.34
23	13.80
24	17.17

State whether or not there is a relationship between age and weight gain and support your answer with statistics. Here's the catch…do not use R. Instead, calculate F and df as we discussed in lecture, using a calculator of Excel/Numbers. Knowing F and df will allow you to use the "FDIST" function in Excel to obtain a P-value. You do not need to provide a graph.
Excel Correlation test:
r=0.913785915
p=0.007548453408 <0.05 Significant and positive correlation.

(7) Jim Hogue intensively sampled riffles in 49 streams for invertebrate species. The mass of all invertebrates per unit area (mg/m^2) was determined, as was the total number of species found in all the riffles of a stream. The data are in the worksheet "*streams*". He wants to know if species richness of invertebrates is a function of biomass.

(a) (2 pts) Provide an appropriate graph of these data. Include a best-fit line.



(b) (2 pts) Do the data appear to meet the assumptions of simple linear regression? Provide appropriate diagnostics. Is any transformation of either variable needed?

After shapiro tests, Biomass was not normally distributed but Number of Species is. We should log transform Biomass data.

(c) (1 pt) Are there any outliers, points with high leverage, or high influence? If so, which points?

Yes bonferonni test p-value=3.3009e-05 and p = 0.0016175 both <0.05. Showing strong evidence of a high influence/leverage on point 33.

Looking at the graph we may also inquire about points 14, 42. Their biomasses are way high.

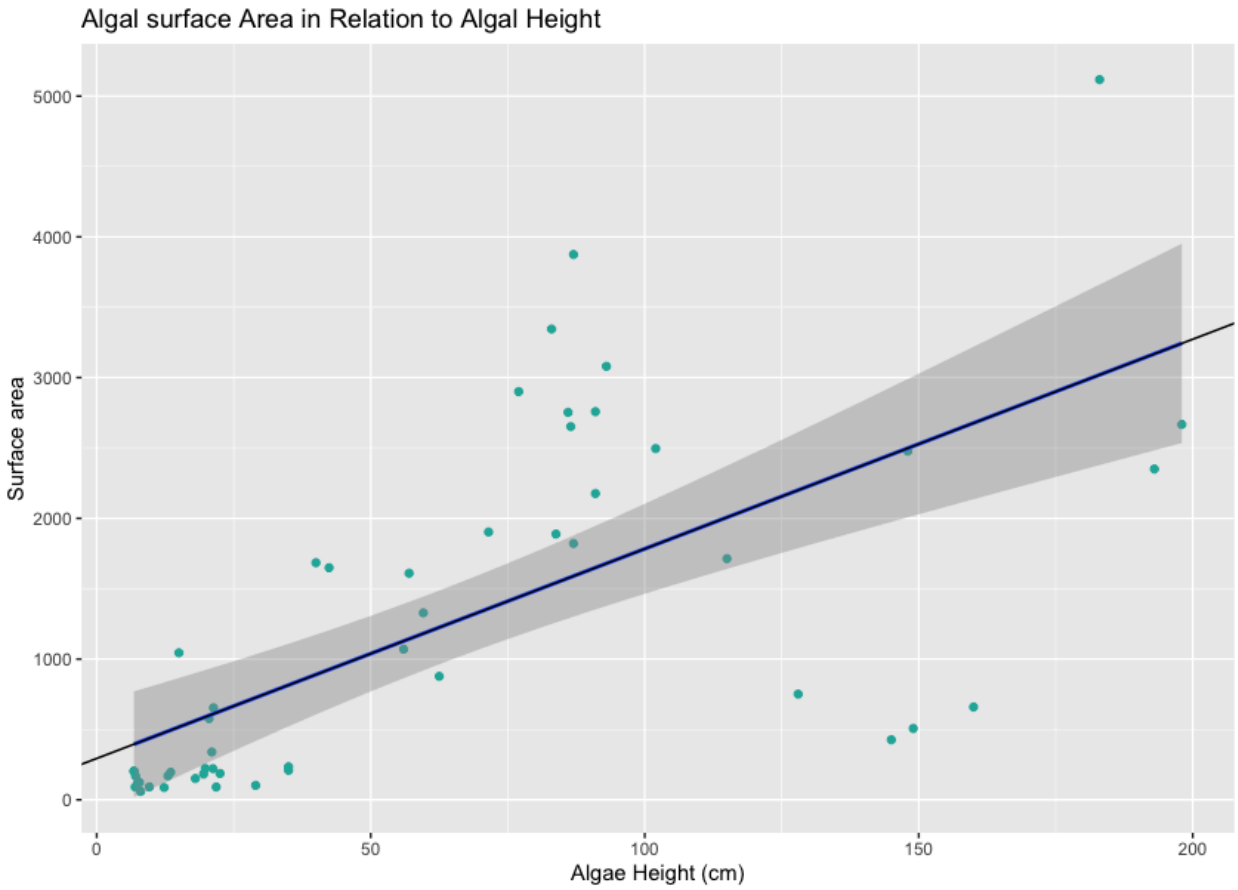
(d) (2 pts) Write a statement for the Results section of this paper that describes whether species richness is related to biomass of stream invertebrates.

Our analysis shows a significant positive relationship between Biomass and Species abundance in invertebrates recorded at 49 streams, suggesting that environments with higher species richness will yield larger groups of invertebrates.

(8) Griffin was a student in Mark Steele's lab and was interested in algal morphology and its effects on fish distribution. He wants to find a quick way to estimate algal surface area. He thinks that height (measured in cm), which is much quicker to measure in the field than surface area (cm^2), might be a good predictor of surface area. He collects fronds and measures their height and then carefully measures their surface

area. Data on height and surface area of the alga *Sargassum horneri* are in the worksheet "*algae*".

(a) (2 pts) Graph these data. Include a linear best-fit line and 95% confidence intervals.



The shaded area around the fitted line shows the 95% confidence interval

(b) (1 pt) Do the data appear to meet the assumptions of simple linear regression? Provide appropriate diagnostics. Is any transformation of either variable needed?

No, most of the data is outside of the CI and there are some visible outliers. It's also not linear, normally distributed. We can log transform the data or omit outliers to arrange the data.

(c) (1 pt) Are there any outliers, points with high leverage, or high influence? If so, which points?

Yes, point 12 according to the bonferoni test. (and 44 after removing that and running another test)

(d) (1 pt) Which is more appropriate for the research question, regression or correlation? Why?

Regression because correlation does not mean causation. We want to know if algae height causes/predicts surface area. In the lecture, it says we are looking at how efficient x(height) is at predicting y(surface).

(e) (1 pt) Is algal height a good predictor of algal surface area in this study species? Provide statistics to justify your answer.

I removed the outlier and got the same plot and a new outlier in my results.

When looking at residuals before and after removing my first outlier the results were often negative and far out of bounds. the p-values shown in the summaries were both <0.5 showing a significant influence of the outlier.

Our confidence intervals for the raw data were lower = -105.29 and upper = 693.27. If our outliers are within the confidence interval limits how can they be outliers? This is all shows that the data collected from this study cannot show a linear relationship between algae height and surface area

(f) (2 pts) How well does height predict surface area? If height could be measured in $1/10^{th}$ the time it takes to measure surface area, would you do that rather than measure surface area? Explain why or why not.

Unfortunately, height is not a good predictor here. If it could be measured faster I still would not use it as a part of dependent variables because of what I know after this statistics exercise.