**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-m2 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

Standard deviation

# (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* = | *t* = |
| *p* = | *p* = |
| Are the means significantly different? | Are the means sig different? |
|  |  |

Which *t*-test do you think is most appropriate for these data? Why?

(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.

**(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-m2 is necessary to support the butterfly. The data in the file “CoastBuckwheat.csv” contains the density of buckwheat in replicate 25 m2 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-m2 standard in this restored area and thus the plant restoration is a success.

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

(b) (2 pts) Write a sentence that states whether this standard (4-plants-per-25-m2) been met (support your answer with t, df, and P).

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

**(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.

**(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?

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

(c) Based on the above, do you think the butterfly ballot affected the outcome of the election?

**(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.

**(7)** Jim Hogue intensively sampled riffles in 49 streams for invertebrate species. The mass of all invertebrates per unit area (mg/m2) 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?

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

(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.

**(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 (cm2), 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.

(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?

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

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

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

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