STAT 503 – Statistical Methods for Biology

Homework 7

30 Points (30 available). Due at 11:59 PM on Tuesday, August 4, 2020.

Optional practice: Practice questions in Whitlock and Schluter Chapters 7-8 and 14-17

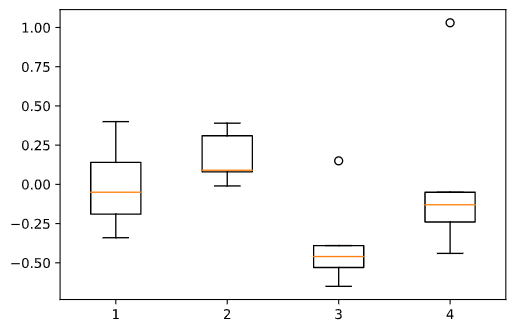
**Part 1:** Caffeine is an alkaloid chemical that occurs in many plant tissues and helps to deter herbivory. It also occurs at low concentrations in flower nectar, where it appears to improve the abilities of honey bees to remember floral rewards and to associate those rewards with flower scents, ultimately improving pollination success (Wright et al. 2013, *Science* 339:1202-1204). To see whether or not bees preferred caffeinated nectar, Singaravelan et al. (2005, *Journal of Chemical Ecology* 31:2791-2804) presented bees with feeding stations that were each randomly assigned to one of 4 treatment groups. At each station, bees could choose between a 20% sucrose solution without added caffeine or a 20% sucrose solution to which caffeine was added in one of four concentrations: 20, 100, 150, or 200 ppm.

For each station, the researchers calculated the difference in grams between the amount of the caffeinated sucrose solution that bees consumed and the amount of the pure sucrose solution that they consumed. Below, I refer to this difference as the *preference index*. If bees consumed more of the caffeinated solution at a station, then the preference index was positive. If they consumed less of the caffeinated solution, it was negative. There were stations per treatment. **The data as recorded by the researchers appear below (in wide format to save space on the page):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Caffeine concentration | | | | | | |
| 50 ppm |  | 100 ppm |  | 150 ppm |  | 200 ppm |
| 0.4 |  | -0.01 |  | -0.65 |  | -0.24 |
| -0.34 |  | 0.39 |  | -0.53 |  | -0.44 |
| -0.19 |  | 0.08 |  | -0.39 |  | -0.13 |
| -0.05 |  | 0.09 |  | 0.15 |  | -1.03 |
| 0.14 |  | 0.31 |  | -0.46 |  | -0.05 |

The data are also on the website in a long-format data file, singaravelan.csv. Please use these data for Questions 1 – 10.

* 1. [1 point] Given that caffeine concentration is a numeric variable, why is it appropriate to use one-way ANOVA to analyze these data, instead of correlation or regression?
     1. While it is numeric, it is in discrete steps of 50 ppm. If you had a continuous set of values for caffeine, you could use a regression. ANOVA is very similar but used to predict a continuous outcome from categorical predictors. While I said that the caffeine concentrations are a numeric and discrete variable type, there is an inherent order to them making them very similar to categorical and ordinal variables
  2. [1 point] State the null and alternative hypotheses for the ANOVA in this problem.
     1. NULL: The mean outcome is the same across all caffeine levels
     2. ALTERNATIVE: At least one pair of means are different from each other
  3. Load singaravelan.csv into R and **make sure that ppmCaffeine is a factor**. Then fit the model and obtain the residuals (I recommend that you use mutate() to add the residuals to your data frame as a new column). Once you have obtained the residuals, complete the following preliminary tasks:
     1. [1 point] **Check for possible outliers**. Present any graphs you are using (with a caption, as usual), and briefly explain what you found. If the plot suggests that an outlier exists, identify the row number for each outlier, describe the nature of the potential problem, and explain how you responded. If you find a mistake, correct it. If you leave an outlier unchanged, say so, and explain your reasoning.



* + - 1. Figure 1. Box Plot showing the data. Category 1 is 50 ppm, and Category 4 is 200 ppm of caffeine. 2 Outliers visible, 1 in either of the last 2 categories.
      2. Station # 16, and Station # 15. Station #16 should be negative, and so should be changed. Station #15 is an outlier, but it has a value that is theoretically possible as compared to 100 ppm caffeine, so it shall remain unchanged.
  1. [2 points] Present the ANOVA table for your analysis, and report your interpretation of this table: decide whether or not to reject the null hypothesis at , and explain what your conclusion means. Your answer should cite specific information from the table.

You may use R to get the ANOVA table, or you may calculate it by hand (use the pf() function in R to obtain a -value). If you want to export the ANOVA table from R, first assign the output from anova() to an object, and then treat it like a regular table:

anovatable <- anova(myModel)

# follow procedure in Tutorial 3 to output anovatable to Word.

* 1. Suppose that we are interested in a planned contrast between the 50 ppm and 200 ppm treatments.
     1. [1 point] Using the output from summary() and confint(), report a point estimate, standard error, and 95% confidence interval for contrast.

* + 1. [1.5 points] The standard error for this contrast is calculated as,

Please explain why this structure makes sense. Your answer should explain what each term in the equation represents (the terms are , , and ) and what a contrast represents. It may be helpful to compare this equation with the equations for the SE of and the SE for the exact two-sample -test.

* 1. Look at the results from the summary() method.
     1. [0.5 points] At , which treatments have mean preference indices that differ significantly from the mean for the 50 ppm treatment (if none of them are different, say so)?
        1. 150 and 200 ppm differs from 50
     2. [0.5 points] Why does the result in 6a **not** contradict the -test result in Question 4?

* 1. There are a total of 5 possible pairwise comparisons (or unplanned contrasts) that we can make among the four treatment levels in this study. This exercise will illustrate how our choice of multiple testing correction can affect the results of these comparisons. The cheat sheet titled *Methods to control error inflation in multiple comparisons* (available on the website) may be helpful.
     1. [1 point] Each of the five contrasts is listed in the table below. Please fill in the -values that you obtain for each comparison using no correction, the Bonferroni correction, the Bonferroni-Holm correction, and the Tukey-Kramer method. The first three of these may be found using the pairwise.t.test() function. For Tukey-Kramer, use aov(fit) %>% TukeyHSD(), where fit is the output from lm(). I have filled in the first row so that you can confirm that your code is correct.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Contrast | None | Bonferroni | Bonferroni-Holm | Tukey-Kramer |  |
| 100 – 50 | 0.358 | 1.000 | 0.716 | 0.781 |  |
| 150 – 50 |  |  |  |  |  |
| 200 – 50 |  |  |  |  |  |
| 150 – 100 |  |  |  |  |  |
| 200 – 100 |  |  |  |  |  |
| 200 – 150 |  |  |  |  |  |

* + 1. [1 point] Which method is the best choice for the current analysis, and why? Your answer should explain the -values that you got in 7a, but should not be based on those -values.
    2. [1 point] Suppose that instead of an ANOVA, we had decided to run a non-parametric Kruskal-Wallis test. Which multiple-testing method would be preferable in this scenario, and why?

* 1. [2 points] Use the predict() method with interval = "conf" to estimate the mean preference index for the each of the four treatment groups (see Tutorial 9, Section 2.5.6). A quick way to obtain a dataset for the newdata argument is to use the first four rows of the singaravelan.csv data, which contain one row for each treatment group.

Use the point estimates and confidence intervals from predict() to generate a graph that shows (i) the data, (ii) the mean for each group, and (iii) a 95% confidence interval for the mean of each group, as illustrated in Tutorial 9, Section 2.6. Your caption should clearly identify each type of symbol in the plot. Finally, (iv) label your graph with letter codes to indicate which group means are significantly different from each other (see Lecture 5.10 for example code). The cheat sheet on the website titled *Identifying differences in pairwise comparisons* explains how to figure out the set labels.

* 1. [1 point] Briefly explain your biological interpretation of the results. For evidence, your discussion can cite the ANOVA table results, hypothesis testing results for pairwise comparisons, and/or the graph in question 8. You do not need to repeat any statistics (simply explain what they mean, biologically).

* 1. [0.5 points] In general terms, how confident are you in the repeatability of these results? Briefly explain your reasoning.

**Part 2:** Sloan and Keatinge (1973, *Journal of Applied Physiology* 35:371-375) investigated the insulating properties of subcutaneous body fat in human children and teenagers. To do this, they had subjects swim in 20.3C (68.54F) water for 18-40 minutes, and measured the change in body temperature that occurred during the swim. The data that we have were taken on 12 boys. For each child, we have an index of leanness in units of m/kg (higher values indicate lower amounts of body fat), and a measurement of the rate of heat loss, which was calculated as,

where and are temperature measurements in C, taken before and after swimming and is the length of time the child swam, in minutes. I have provided results, but if you would like to analyze these data yourself, they are available on the website in the file sloan\_keatinge\_1973\_cold\_swim.csv.

Figures 2.1 – 2.4 present diagnostic results for these data. Regression results from R are presented in Tables 2.1 and 2.2. Finally, Figure 2.5 presents the results from the regression analysis. Please use these all of these pieces of information to answer questions 11 – 16.

**Figure 2.1:** Scatterplot of the rate of heat loss in boys swimming in water at 20.3C, as a function of leanness (residual skinfold thickness, adjusted for body surface area, measured in m/kg).

**Figure 2.2:** Normal quantile plot of residuals from a linear regression of the rate of heat loss against leanness in boys swimming in water at 20.3C (Shapiro-Wilk:

).

(plots continued on the next page)



**Figure 2.3:** Residual plot for a linear regression of the rate of heat loss against leanness in boys swimming in water at 20.3C.

**Figure 2.5:** Boxplot of the residuals from a linear regression of the rate of heat loss against leanness in boys swimming in water at 20.3C.

> sloan <- file.choose() %>% read.csv()

> fit <- lm(heatLossRate ~ leanness, data = sloan)

> anova(fit)

Analysis of Variance Table

Response: heatLossRate

Df Sum Sq Mean Sq F value Pr(>F)

leanness 1 0.0148064 0.0148064 68.729 8.606e-06 \*\*\*

Residuals 10 0.0021543 0.0002154

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(results continued on next page)

> summary(fit)

Call:

lm(formula = heatLossRate ~ leanness, data = x)

Residuals:

Min 1Q Median 3Q Max

-0.027378 -0.007411 -0.002899 0.012901 0.023065

Coefficients:

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

(Intercept) -0.026913 0.010017 -2.687 0.0228 \*

leanness 0.018970 0.002288 8.290 8.61e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01468 on 10 degrees of freedom

Multiple R-squared: 0.873, Adjusted R-squared: 0.8603

F-statistic: 68.73 on 1 and 10 DF, p-value: 8.606e-06

# note: to export this, I would use

# summaryForWord <- summary(fit) %>% coef() %>% as.data.frame() %>%

# tibble::rownames\_to\_column(var = 'Coefficient')

> confint(m, level = 0.95)

2.5 % 97.5 %

(Intercept) -0.04923098 -0.004594039

leanness 0.01387120 0.024067881

A close up of a map

Description automatically generated

**Figure 2.5:** Relationship between leanness and the rate of heat loss in 12 boys swimming in water at 20.3C. The dark and light grey regions show 95% confidence and predictions bands, respectively.

* 1. [2 points] Please check the assumptions for the linear regression model. Your answer should clearly identify each assumption that you check, state whether or not it has been met, and cite specific evidence to support each conclusion. You may use the same graph for more than one assumption. You may also assume randomness and independence.
     1. All Assumptions met
  2. [1 point] Report the results of the regression analysis. Specifically, your answer should indicate whether or not a linear relationship exists and should estimate the slope of that relationship at . (Note: there are two different statistics that can be used to test the null hypothesis that the slope is equal to zero; you only need to cite one of them).
     1. R-squared of 0.873 using the linear model
     2. This implies that there is a fairly strong linear relationship between the two variables.
  3. [1 point] What is the expected rate of heat loss for a boy swimming in 20.3C water, given that he has a leanness index of 2.8 m/kg? Please show your work, including the symbolic equation that you are using to calculate your answer.
     1. About 0.03 degreesC/min (No equation, just following the line),
     2. Given a stand y = mx + b formatted linear equation, just plug 0.03 in for x and solve for y.

* 1. [2 points] Suppose that we collect data on a 13th boy under the same experimental conditions that were used in this study. He has a leanness index value of 6.0 m/kg. Give an approximate 95% prediction interval for his rate of heat loss and explain how we should interpret this interval (what does it tell us)?
     1. (0.075, 0.1) is the 95% CI. This means that if you had a group of people with 6.0 m/kg leanness index, 95% of them would fall in these bounds.
  2. [1point] Within this dataset, what percentage of the variation in the rate of heat loss is explained by variation in leanness (cite a specific statistic)? What does this value tell us about how well our model will predict new data values?
     1. 87% of the variation in the rate of heat loss is explained by variation in leanness. (R-squared value).
     2. This tells us nothing about how well the model will predict new data. In a model building cycle, you must train the model (as was done), but you must test the model with a testing set that has new data points in order to gauge it’s real-life accuracy
  3. [1 point] Based on these data, what can we conclude about the rate of heat loss as a function of leanness in girls? Please explain your answer.
     1. We can’t conclude anything about girls. I thought this was just a dataset about boys… I mean, If I had to stake a bet on it, I’d say they’re probably similar, but girls in general have a higher fat content, which would mean a lower rate of heat loss assuming leanness index takes biological sex into account.

**Part 3:** Wilson et al. (2011, *Journal of the National Cancer Institute*, 103:876–884) used a dataset that tracked the status of 10,383 male health professionals for 20 years, from 1984 to 2004, to look for associations between coffee consumption and the occurrence of prostate cancer. The data that we have available in wilson\_et\_al\_2011\_coffee\_and\_cancer.csv compares the men who drank no coffee to the heaviest coffee consumers, who drank 6 or more cups of coffee a day. The subjects' coffee-drinking habits are classified in the variable coffeeStatus. The second variable, cancerStatus, has the value "cancer" if the individual developed advanced prostate cancer during the study, and "no cancer" if he did not develop advanced cancer during the study.

* 1. [1 point] Present a 2 2 contingency table for this dataset, with the explanatory variable represented as columns. The simplest way to generate such a table in R is to use the xtabs() function, following the syntax:

ctable <- xtabs(~ response + explanatory, data = myData)

Exporting a contingency table generated by xtabs() is a little tricky because of the way it is stored. Use:

ctableForWord <- as.data.frame.matrix(ctable) %>%

tibble::rownames\_to\_column()

* 1. [1 point] Report the estimated probability of developing cancer, along with a 95% confidence interval, for both the non-coffee drinkers and the coffee drinkers, using the Agresti-Coull method. You may do this in R or by hand.
  2. [1 point] Briefly explain why it is possible to use relative risk instead of odds-ratios with this dataset (hint: how does the study design differ from a case-control study?).
     1. This is an observational study of a fairly common disease. Odds ratio can overestimate in these cases.
  3. [1 point] Estimate the relative risk for the coffee group versus the no coffee group, with a 95% confidence interval. By what factor does coffee consumption change the risk of prostate cancer, according to your analysis? Show your work.
  4. [1 point] Would it be legitimate to claim that this study shows that coffee drinking lowers the risk of developing advanced prostate cancer? Why or why not? (Note: your answer to this question should not depend on the results of your earlier analysis).
     1. No, just that there is a statistically significant difference between the distribution of cancer development rates. You don’t know for sure that the coffee is what is directly lowering the risk of developing advanced prostate cancer.
  5. [1 points] Please attach your code.

1. import pandas as pd
2. data\_frame = pd.read\_csv("singaravelan\_et\_al\_2005.csv")
3. data\_frame = data\_frame.set\_index("station")
4. data\_frame
5. data\_frame.values
6. import numpy as np
7. df\_50 = data\_frame.preferenceIndex[data\_frame.ppmCaffeine == 50]
8. df\_100 = data\_frame.preferenceIndex[data\_frame.ppmCaffeine == 100]
9. df\_150 = data\_frame.preferenceIndex[data\_frame.ppmCaffeine == 150]
10. df\_200 = data\_frame.preferenceIndex[data\_frame.ppmCaffeine == 200]
11. data\_to\_plot = [df\_50,df\_100,df\_150,df\_200]
12. from matplotlib import pyplot as plt
13. plt.boxplot(data\_to\_plot)
14. plt.show()
15. plt.savefig("fig1.png")