Final Exam Practical Exam

Support all of your results statements with statistics. State which statistical test you used. Be sure to interpret the biological meaning of all results. It is your responsibility to confirm that all test assumptions have been met, in order to do the analysis correctly, but you do not need to show evidence of the assumptions having been met. Use R for all of your statistical analyses and graphs. You only need to provide graphs where indicated.

1. (15 pts) Maria from the Robertson Lab wants to know the weight of the frogs she finds in the field, but it is not convenient to carry a balance into remote field sites. It is much easier to measure snout-vent length (SVL). She wants to know whether this is a good predictor of the weight of the frog. She catches 100 frogs and brings them back to the lab in Costa Rica to measure their SVL and weight. Her data are in the file *TreeFrog.csv*.

Are weight and SVL significantly related?

A regression test on a linear model shows that SVL is significantly related to the weight of treefrogs (F-stat= 1016, p<0.001).

How strongly?

This is a strong relationship r^2 = 0.91 very close to 1.

```
Call:
lm(formula = weight ~ SVL, data = froggy)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4998 -0.3738 -0.1239  0.2716  2.1960

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.713754   0.437651  -19.91   <2e-16 ***
SVL          0.255992   0.008032   31.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7148 on 98 degrees of freedom
Multiple R-squared:  0.912,     Adjusted R-squared:  0.9111
F-statistic:  1016 on 1 and 98 DF,  p-value: < 2.2e-16
```

Given a particular SVL, what equation would she use to predict the weight?

Use the following formula: y = a + b(x) → Weight = -8.714 + 0.256(SVL)

In the formula, a refers to the y-intercept when x=0 and b refers to the slope of SVL.

2. (15 pts) As you drive the remote back roads to your field study site, you notice that the number of dead animals you see along the road seems different from what you see closer to home. You start counting. You divide the roads you travel evenly into rural or urban. After a year, you have **counted** 214 road-killed animals. The data are: 119 animals on rural roads and 95 animals on urban roads. Was there a difference in the number of animals killed along rural roads compared to urban roads?

After running a G test and a X^2 test on observed and expected values I found no significant difference between the amount of roadkill on rural roads and urban roads.

G test: G-value= 2.697, p>0.05 X^2 test: X^2= 2.691, p>0.05

3. Sophia is a psychologist studying the effects of cognitive behavioral therapy (CBT) on anorexia nervosa. She found 55 people to participate in the study. Each person was randomly assigned to the control treatment (no therapy) or assigned to receive cognitive behavioral therapy once a week for four months. She measured the weight of each person at the beginning and end of the study and determined their change in weight over four months. The data are in the file *anorexia.csv*.

(a) (10 pts) Use R to determine the following descriptive statistics for weight change. Are the data normally distributed?

Mean 7.069091

Median 6.6

S.D. 8.904199

S.E.M. 1.200642

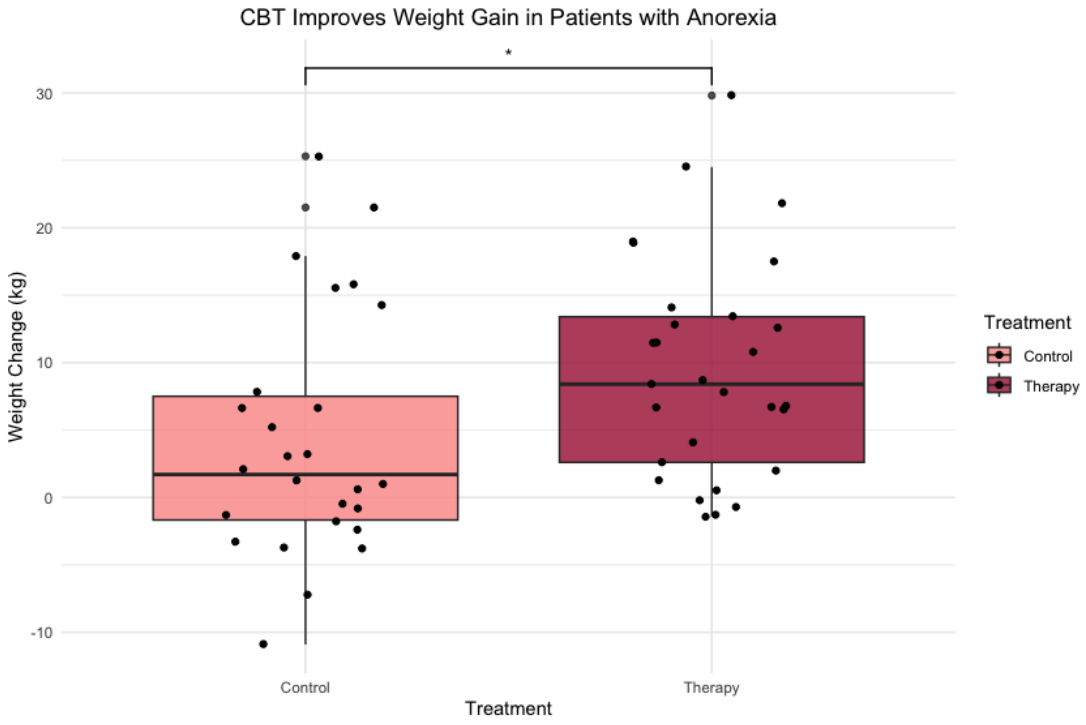
C.V. 1.259596 or 125.96%

Even though we have a high CV the data is normally distributed.

(b) (10 pts) Use an appropriate test to determine whether CBT affects weight change. Include a graph that illustrates the result.

I used Welch's test because the variances between treatment groups are not equal. There is a significant effect of treatment on Weight change (t=-2.2418, p<0.05).

Significance is annotated with one asterisk (*) to show a p value<0.05 but still bigger than 0.01.



4. Erin in the Steele Lab studied kelp bass at several different sites along the coast of California. She wanted to know if density of this fish was related to the habitat attributes of the site. At each site she

sampled the site using four independent transects. In addition to recording the number of kelp bass on each transect at each of the 7 sites, she recorded six attributes of the habitat: the approximate volume of space occupied by kelp, water temperature, salinity, and the percent cover of sand, seagrass, and rock. The data are in the file *habitats.csv*.

(a) (5 pts) First, convert all habitat variables to z-scores and then use Principal Components Analysis (PCA) to derive components that summarize the variation in the **six** habitat variables (do not include kelp bass density).

How many components would you need to include to capture 70% of the variance in the original data?

Three (3) components must be included to capture at least 70% of the variance. The summary shows that the three components capture 71.2%.

```
> PCAmodel <- princomp(bass.scale, cor=FALSE)
> summary(PCAmodel)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6
Standard deviation     1.3583524 1.1448698 0.9811799 0.9102352 0.8011466 0.44360805
Proportion of Variance 0.3189098 0.2265454 0.1663950 0.1432024 0.1109346 0.03401276
Cumulative Proportion  0.3189098 0.5454552 0.7118503 0.8550526 0.9659872 1.00000000
```

How much variance in the original data is explained by the first two principal components?

The first two components explain 54.5% of the variance.

Which variables are most strongly related to PC1 and PC2?

By extracting the loadings, I found that Kelp Volume (0.611) and Temperature (-0.667) are the most related to PC1 while Sand (-0.608) and Seagrass (0.578) are strongly related to PC2.

(b) (5 pts) Use perMANOVA with the original (i.e. use the raw data, not z-scores) six habitat variables to determine whether the seven sites differ in habitat variables. Does habitat differ among sites?

perMANOVA on raw data shows that habitat differs significantly among sites ($p<0.001$).

(c) (5 pts) Finally, let's return to Erin's original question about kelp bass densities. Use multiple regression to determine whether the original (raw data) six habitat variables have any effect on kelp bass density. Describe any significant effects.

Using multiple regression on the raw data shows that habitat, specifically with Seagrass has a significant effect on Kelp Bass density ($F=5.8816$ $p<0.05$).

However, the residual plot for the raw data showed heterostadicity. After log transformation, there were no significant effects of habitat on Kelp Bass density.

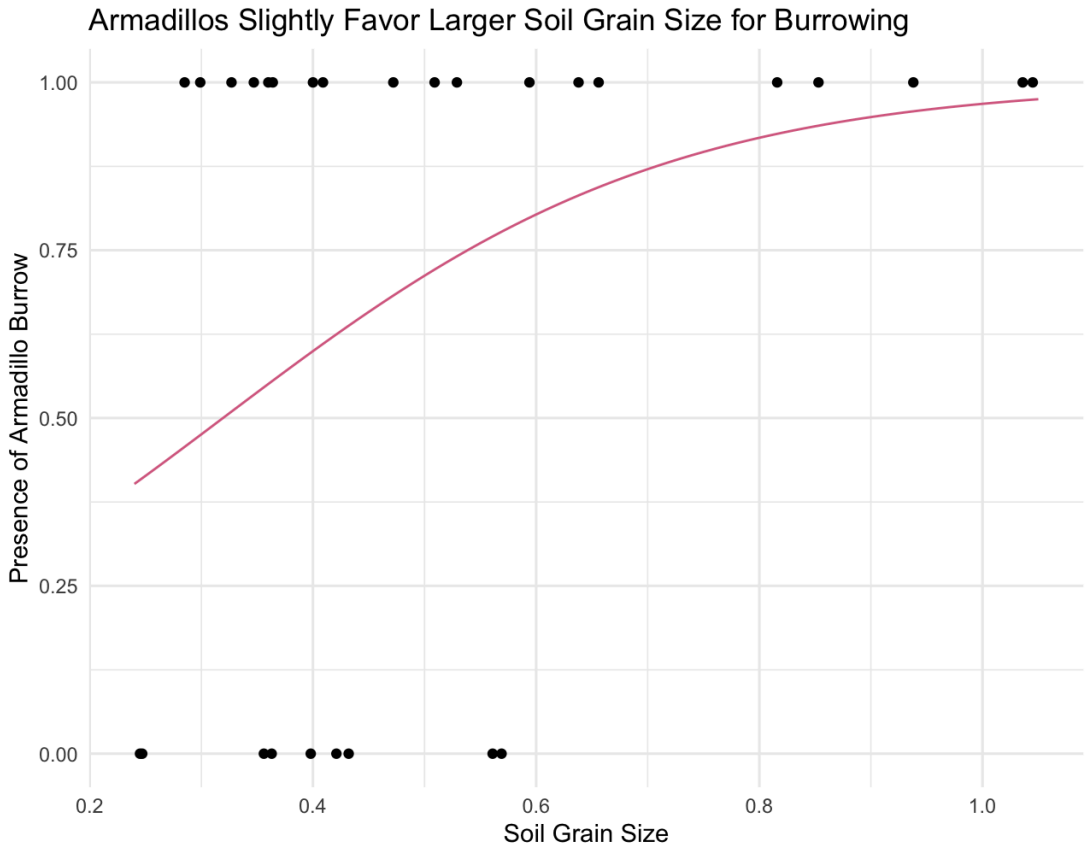
The log transformation model also showed a lower AIC score than the original which suggests a better fit for the data.

5. (10 pts) Jeffrey is studying armadillos and interested in whether the type of soil affects where they choose to make a burrow. He visits many sites and at each one, he measures the average particle size of the soil, and whether or not he finds an armadillo burrow there. The data are in the file *armadillos.csv*. In the file, 0 means no burrow was found, 1 means that at least one burrow was found.

Determine whether grain size affects the probability of finding armadillos at a site. Include a graph to illustrate your results.

Logistic regression on a generalized linear model shows that soil grain size **weakly affects** the probability of armadillo burrowing (about 12.54% McFadden R^2= 0.1253636, $p<0.05$).

This figure shows that while armadillos burrow in all soil types observed, they are more likely to be found in larger grain soils.



6. (15 pts) Luisa studies the growth of cancer cells in the human body. She developed two slightly different versions of a new drug that might reduce the rate of proliferation of cancer cells. She treated cancer cells in tissue culture with each of the new drugs (Drug A and Drug B). She also included a control in which only saline was added to the culture. She had 30 replicates of each treatment and measured the growth of cancer cells in each. The data are in the file *cancerdrug.csv*. Use these data to determine the effectiveness of these drugs.

With one-way ANOVA on the raw data, I found both treatments had a statistically significant and negative effect on the growth of cancer cells ($F= 36.38$, $p<0.001$ on 2 and 87 DF, and adjusted $R^2= 0.44$).

The original data for cell growth was not normally distributed. After log transformation, I found the same conclusion ($F=35.42$, $p<0.001$ on 2 and 87 DF, and adj$R^2=0.44$).

A Tukey posthoc test on the original data shows that Drug A vs the control and Drug B vs the control both have a p-value = 0 confirming the previous conclusion. Drug A vs Drug B did not have any significant difference (p= 0.77).

7. (10 pts) Kate is working on a recovery plan for the green sea turtle. As part of this plan, she needs to build a demographic model that will project population growth (or decline). To build a more accurate model, she wants to know if per-capita reproductive output of females varies among sites. She has access to a data set that gives reproductive output (eggs laid per nesting female) at three sites, pooled over a 30 year time span. Bear in mind that reproductive output is often a function of body size in animals, and this effect should be controlled for statistically. The file *SeaTurtle.csv* also includes the carapace length (i.e., shell length) as a measure of size.

Based on the data, does per-capita reproductive output differ among sites? Is it a function of body size? Does the effect of body size differ among sites? Include a graph illustrating your results.

Using ANCOVA (type III) to see the response of egg production to Site and Size shows that they individually have significant effects on reproductive output ($p<0.001$). This model proved to be the best fit for the data with a low AIC score but to answer the question of the effect of body size differing among sites, we must use a model including the interaction between site and size.

ANCOVA (type III) on a model with interaction shows no statistically significant effect of body size, site, or their interaction on egg production ($p>0.05$). This was the second best model for the data but it is the best model for answering the research question. The graph will reflect these results.