Kaitlyn Valentinetti

November 18, 2022

ECo 634: Michael France Nelson

Lab 10: ANOVA

1. Submit the code you used to build your ANOVA by hand. Make sure you use the code
   template so that you use the same variable names as those which we'll use for the
   grading.

   rm(list = ls())

   rope <- read.csv(here("data", "rope.csv"))

   rope$rope.type = factor(rope$rope.type)

   levels(rope$rope.type)

   rope_type <- rope$rope.type
   rope_type

   n_obs = nrow(rope)
   n_obs
   n_groups = length(unique(rope$rope.type))
   n_groups

   grandmean = mean(rope$p.cut)
   obs = rope$p.cut
   resids = obs - grandmean
   ss_tot = sum(resids^2)

```r
df_tot = n_obs - 1


agg_mean = aggregate(
  x = rope$p.cut,
  by = list(rope$rope.type),
  FUN = mean)

aggregate(
  x = rope$p.cut,
  by = list(rope$rope.type),
  FUN = function(x) mean (x)
)

agg_resids = aggregate(
  x = rope$p.cut,
  by = list(rope$rope.type),
  FUN = function(y) y-mean(y)
)
str(agg_resids)

agg_sum_sq_resids = aggregate(
  x = rope$p.cut,
  by = list(rope$rope.type),
  FUN = function(y) sum(y-mean(y))^2)
str(agg_sum_sq_resids)

ss_within = sum(agg_sum_sq_resids$x)
ss_within

df_within = n_obs - n_groups
```

df_within

ss_among = ss_tot - ss_within
ss_among

df_among = n_groups - 1
df_among

ms_within = ss_within / (n_obs - n_groups)
ms_within

ms_among = ss_among / (n_groups - 1)
ms_among

f_ratio = ms_among / ms_within
f_ratio

f_pval = 1 - pf(f_ratio, df_among, df_within)
f_pval

2. Examine the conditional boxplot in the Partitioning Variance: Within-Group section of the walkthrough. Based on the figure, do you think there are equal variances among the groups?
   Based on the conditional boxplot for partitioning variance (within-group), there is not equal variances among the groups.

3. Conduct a Bartlett test to assess the homogeneity of variances of the percent cut among the rope type groups.
   bartlett.test(p.cut ~ rope.type, data = rope)

4. Given your graphical assessment (question 2) and the Bartlett test, do you think an ANOVA-type analysis is appropriate on the raw data? Explain why or why not.

Given the graphical assessment from question 2 and the Bartlett test, the ANOVA-type analysis is not appropriate on the raw data because the homogeneity assumption is not met.

5. Which rope type is the base case?

The rope type "BLAZE" is the base case.

6. What is the mean percent cut of the base case rope? Show your calculation using value(s) from the model coefficient table.

The mean percent cut of the base case rope is 0.36714. This calculation is shown in the first row of the model coefficient table:

```
Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.36714    0.04493   8.172 4.45e-13 ***
```

7. What is the mean percent cut rope type XTC? Show your calculation using value(s) from the model coefficient table.

The mean percent cut rope type XTC is 0.2655. The calculation is: mean_p_XTC = 0.36714 - 0.10164.

8. Use the residuals() function to retrieve the residuals from your model and perform an overall normality test. Report the p-value.

p-value = 7.238e-07

9. Do your model residuals meet the normality assumption, and how do you know?

The model residuals do meet the normality assumption because the p-value is less than 0.05.

10. Perform normality tests on the residuals within each group. How many groups meet the normality assumption?

Three out of six groups meet the normality assumption.

11. Given the results of your tests for residual normality, do you think that a one-way Analysis of Variance is appropriate for this dataset?

Since only 50% of the groups meet the normality assumption, I don't think that the one-way ANOVA is appropriate for this dataset.

12. Create a conditional boxplot of the female penguins: body mass conditioned on species.

Female Penguin Body Mass Conditional Boxplot
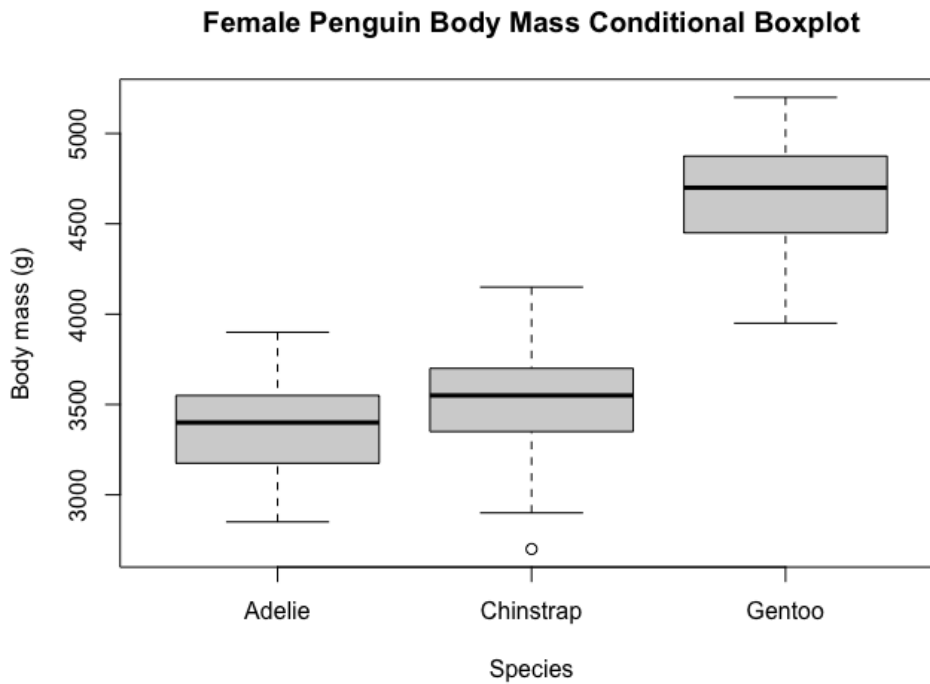
13. Based on the boxplot, do you anticipate any problems with residual normality, or homogeneity of variances? Why or why not?

Based on the boxplot, I don't anticipate any problems with residual normality or homogeneity of variances. These two assumptions are met because the boxes are all about equal width as well as the range of each species.

14. Conduct a Bartlett test for homogeneity of variances of body mass grouped by species. Hint: use the formula notation. Report the p-value. Is the homogeneity assumption met? Why or why not?

The p-value of the Bartlett test is 0.9056. Thus, the homogeneity assumption is met because the null hypothesis for the Bartlett test states the variance in all species is homogeneous, and we cannot reject it because the null is greater than 0.05.

15. Fit a linear model of body mass (the response) and species (the predictor) using the female penguin data. Conduct a test for normality of the residuals. Report the p-value. Is the residual normality assumption met? Why or why not?

The p-value is 0.3639. Therefore, the residual normality assumption is met because the null hypothesis for the Shapiro-Wilk test states the residuals are normally distributed, and we cannot reject it because the null is greater than 0.05.

16. Conduct a Tukey HSD post-hoc test on your model. Which pair or pairs of species have significantly different body masses?

    All pairs (Chinstrap-Adelie, Gentoo-Adelie, and Gentoo-Chinstrap) have significantly different body masses.

17. Describe how your HSD rest results match, or do not match, the graphical insight from the conditional boxplot.

    The results of the HSD overall matches well with the graphical insight of the conditional boxplot because both suggest there is a definite difference in body mass between the species. In the graph we can see that Adelie and Chinstrap are completely out of the minimum-maximum range of Gentoo penguins which supports the results of the HSD. However, the body masses of Adelie and Chinstrap penguins seem to overlap each other in the graphical model, but according to the HSD p-value (Chinstrap-Adelie, p-value = 0.0179471, they are significantly different.