Bonnie Turek

Eco 634 – Lab 9

11/10/21

Chi-Square Tests

Q1. State the null hypothesis of the Chi-square test.

Make sure you state the null hypothesis in terms of Brown Creeper presence/absence and edge/interior habitats.

According to the help in R, the null hypothesis of the Chi-square test is that the joint distribution of the cell counts in our 2-dimensional contingency table is the product of the row and column marginals.

In other words, the null hypothesis is that the observed counts of Brown Creeper presence/absence in edge and interior habitats do not differ from expected values if presence/absence was independent of habitat type. I.e. Brown creeper presence or absence does not depend on habitat type. The expected values can be calculated from the values in the rows and columns of the Brown Creeper presence/absence data.

Q2. Consider the results of your test and explain whether you think that Brown Creepers show a significant habitat preference.

Make sure your use the output of your statistical test to support your answer.

OUTPUT:

```
> br_creeper_table

   TRUE FALSE

 E   29   144

 I  314   559

> chisq.test(br_creeper_table)

Pearson's Chi-squared test with Yates' continuity correction

data:  br_creeper_table

X-squared = 23.3, df = 1, p-value = 1.386e-06
```

Yes, I think the results of the Chi Squared test show a very low p-value which could signify that we have strong evidence that the Brown Creepers **DO** show a habitat preference between exterior and interior habitats. In other words, the low p-value gives us strong evidence to reject the Chi-square null hypothesis stated above. The observed values in the contingency table are different from the expected

values for Presence/Absence in each habitat. IN just looking at the contingency table, I can't tell which way the preference would go though.

Penguins: Building Models for ANOVA

Q3. Show the R-code you can use to create a model fit (call it fit_species) of penguin body mass as predicted by penguin species.

```
fit_species =
  lm(
  formula = body_mass_g ~ species,
  data = penguins)
```

Q4. Show the R-code you can use to create a model fit (call it fit_sex) of penguin body mass as predicted by sex.

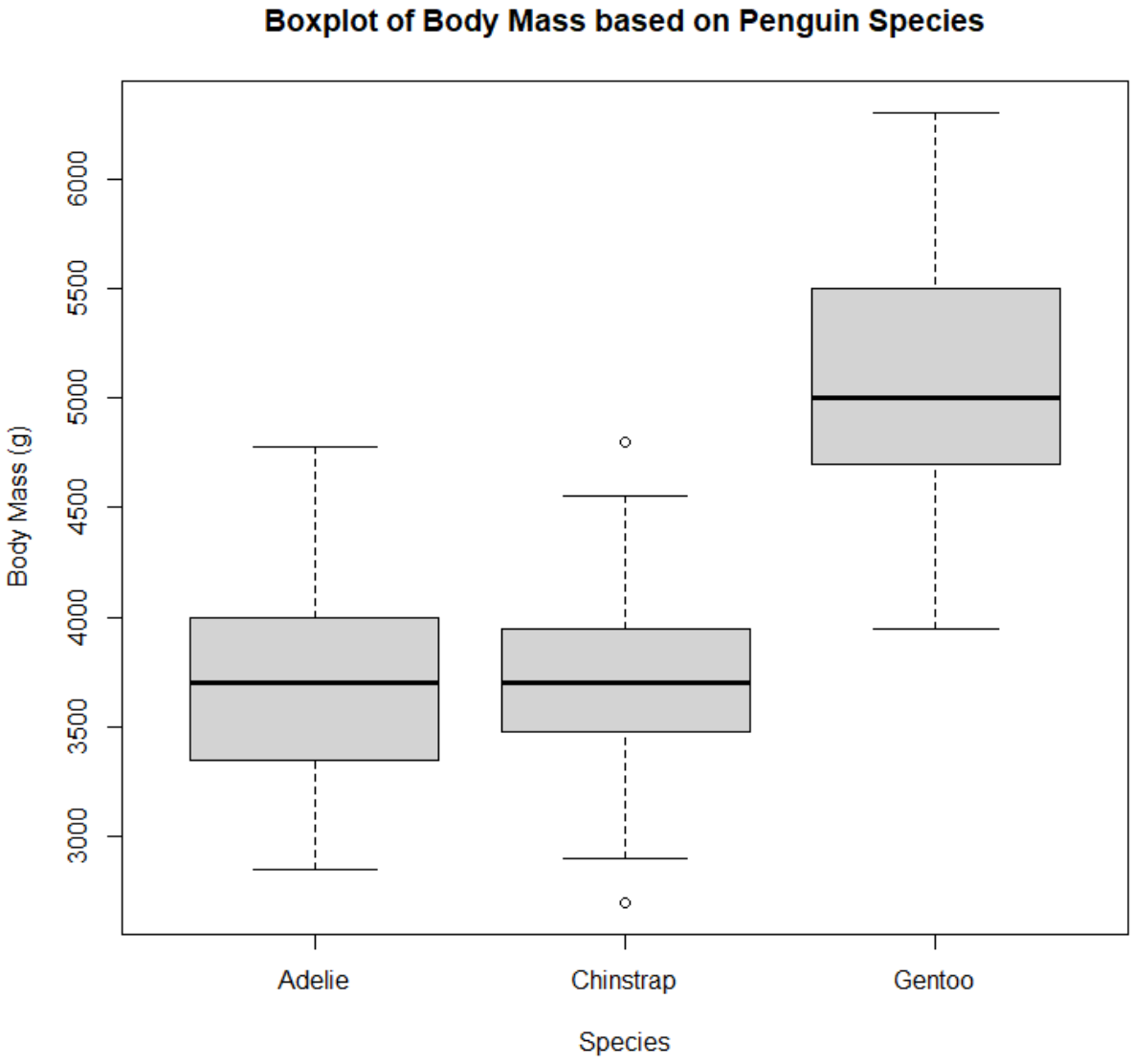
```
fit_sex =
  lm(
    formula = body_mass_g ~ sex,
    data = penguins)
```

Q5. Show the R-code you can use to create a model fit (call it fit_both) of penguin body mass as predicted by species and sex.

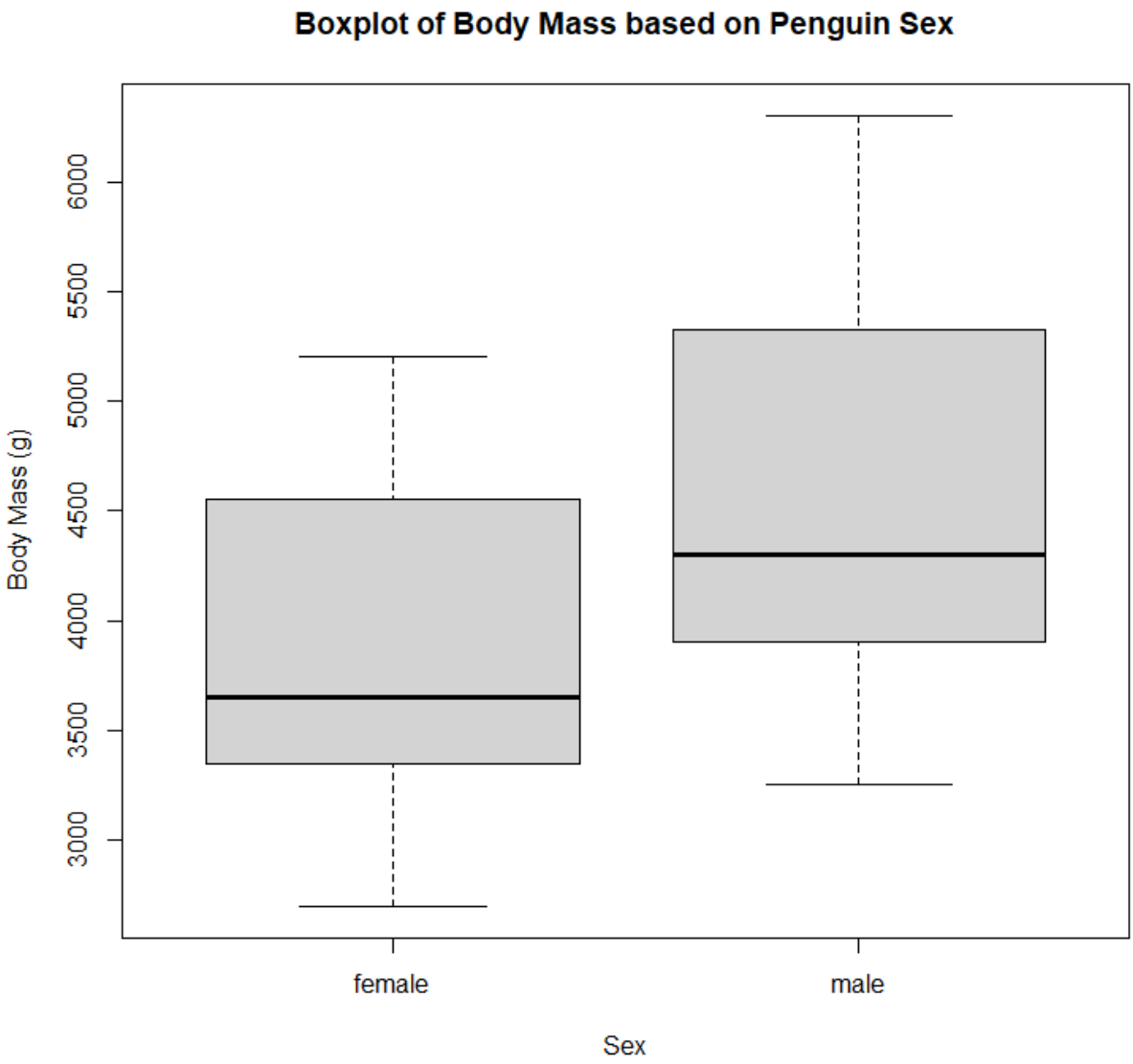
```
fit_both =
  lm(
    formula = body_mass_g ~ sex + species,
    data = penguins)
```

Homogeneity Assumption: Graphical

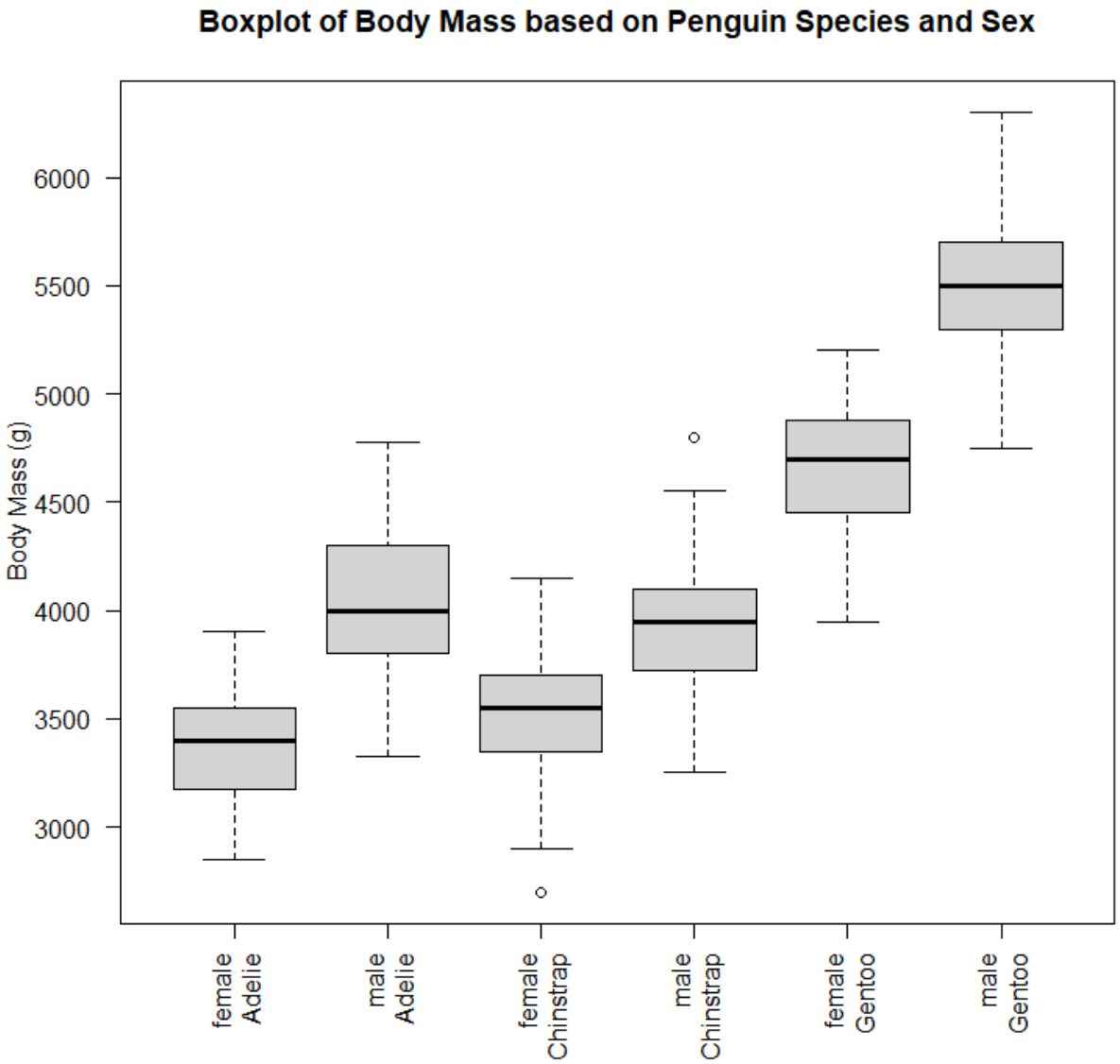
Q6. Include a conditional boxplot corresponding to your fit_species model.



Q7. Include a conditional boxplot corresponding to your fit_sex model.



Q8. Include a conditional boxplot corresponding to your fit_both model.



Q9. Based on the shapes of the boxes, which of the models (if any) do you think may have problems fulfilling the homogeneity assumption?

If the homogeneity assumption holds, we would expect that all of the groups of penguins (by sex, species, or both) would have similar variability – AKA the boxes should all be about the same width.

In the case of body mass by penguin species, the width of the boxes does NOT appear to be as uniform in width. I would say that the Adelie and Gentoo penguins have larger variance than the Chinstrap species. This is purely a graphical/visual interpretation. There may be less variance in body mass values for the Chinstrap species. We should confirm this with the Bartlett test.

In the case of body mass based on penguin sex, I would say that visually the boxes for male and female penguins are relatively similar in width, however the female box is slightly narrower in width than males.

This signifies the variance might be slightly smaller for females' body mass than males. We should confirm this with the Bartlett test.

In the case of body mass explained by the interaction of sex AND species, I would say that we cannot assume the homogeneity assumption is valid. The male Adelie box width is much larger than the rest of the categories. Perhaps this category has a larger variance than the rest of the sex and species categories.

Overall, I think all three of the fit_sex, fit_species and fit_both models would have problems fulfilling the homogeneity assumption. Actual statistical analysis Bartlett tests would help in verifying these graphical/visual interpretations.

Homogeneity Assumption: Bartlett Test 1

Q10. State the null hypothesis of the Bartlett test.

The null hypothesis of the Bartlett test is that the variances in each of the groups (samples) are the same. In this case, the Bartlett test null would be that the variances in the body masses of the groups of sex (male and female penguins) are the same, or the variances in body mass for the species of penguins (Adelie, Gentoo and Chinstrap) are the same. The same null hypothesis goes for the interaction of sex and species resulting in the same variances of body mass values.

Q11. What was the p-value from the Bartlett test of homogeneity for observations grouped by species? You can round your answer to 4 decimal digits.

```
bartlett.test(body_mass_g ~ species,
        data = penguins)
```

Bartlett test of homogeneity of variances

data: body_mass_g by species

Bartlett's K-squared = 5.9895, df = 2, **p-value = 0.05005**

Q12. What was the p-value from the Bartlett test of homogeneity for observations grouped by sex? You can round your answer to 4 decimal digits.

```
bartlett.test(body_mass_g ~ sex,
        data = penguins)
```

Bartlett test of homogeneity of variances

data: body_mass_g by sex

Bartlett's K-squared = 4.6017, df = 1**, p-value = 0.03194**

Homogeneity Assumption: Bartlett Test 2

Q13. What was the p-value from the Bartlett test of homogeneity for observations grouped by both factors? You can round your answer to 4 decimal digits.

CODE:

```
datpen_species = aggregate(
  body_mass_g ~ species,
  data = penguins,
  FUN = c)
str(datpen_species)
datpen_sex = aggregate(
  body_mass_g ~ sex * species,
  data = penguins,
  FUN = c)
str(datpen_sex)
#one way to do this
bartlett.test(body_mass_g ~ interaction(sex, species), data = penguins)
#the other way to do this (same output)
bartlett.test(datpen_sex$body_mass_g)
```

OUTPUT:

Bartlett test of homogeneity of variances

data: datpen_sex$body_mass_g

Bartlett's K-squared = 7.6908, df = 5, **p-value = 0.1741**

Q14. Based on the results of the Bartlett tests, do you anticipate any issues with heterogeneity in any of the models? Make sure you justify your response with the results of your tests.

Based on the Bartlett test for just species, our resulting p-value of 0.05 gives us some evidence that the body mass values between all male and female penguins have different variances. The 0.05 p-value is right at the threshold for rejection of the null hypothesis, so we don't have super strong evidence.

Therefore the fit_species model is likely to have issues with heterogeneity. We may not be able to assume the homoegeneity assumption of Group 1 models.

Based on the Bartlett test for just penguin sex, our resulting p-value of 0.03 gives us stronger evidence that the body mass values between Adelie, Chinstrap, and Gentoo penguins have different / heterogenous variances. The p-value gives us strong evidence to reject the null hypothesis and we can infer that the fit_sex model is likely to have issues with heterogeneity. We can likely not assume the homogeneity assumption of Group 1 models is met here either.

Based on the Bartlett test for both penguin species AND sex, and the resulting p-value of 0.1741, we DO NOT have strong evidence to reject the null hypothesis. Therefore, the body mass values based on the interaction of species AND sex likely do follow the homogeneity assumption of Group 1 models. However, we thought this resulting p-value was interesting since we did not make this homogeneity assumption for this model purely on visual/graphical interpretation.

Overall, we can expect issues of heterogeneity in the fit_species model and fit_sex model but not the fit_both model.