Analysis of Variance Across Groups

Written by Rafael Romero

2023-12-29

Contents

Lab 6 – Hypothesis Testing & ANOVA/Post-hoc Tests

This lab I will apply data visualization skills, hypothesis testing, knowledge about residuals & how to interpret ANOVAs/Post-hoc Tests.

Lab Questions

Question 1

Please answer **True or False** for 1a and b.

a) For ANOVAs, we are testing the normality of individual variables (i.e. not residuals) **FALSE**
b) If your ANOVA is significant you know which level is different from the other levels **FALSE**

Question 2

What are the four parameters needed to run a power analysis?

- β probabibility of making a type II error
- (1 - β) the power of the test
- Effect size
- Variability

Question 3

In Exercise 6 file, we stated that the H_0 of an ANOVA could be written as $H_0 : \alpha_1 = \alpha_2 ... = \alpha_k$. What does k refer to in this situation?

ANSWER: K refers to the number of groups or levels that are being tested

Question 4

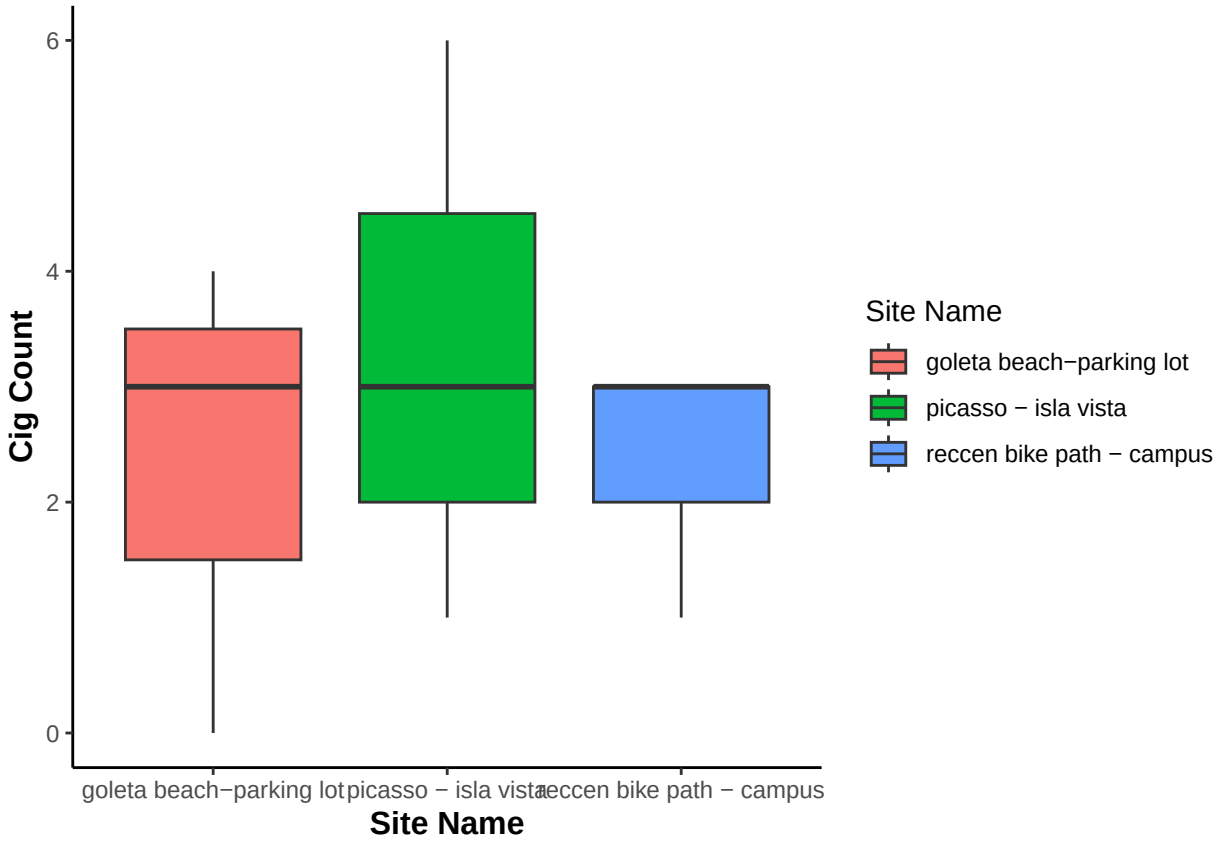
For Question 4 you will be using the data set `litter`. These data come from the UCSB Litter Survey conducted around Isla Vista, California on 23 & 25 May 2023. The litter was systematically collected by

k-pass removal surveys. Keep IV beautiful - pick up a piece of litter!

For this question **I am only interested in sites "goleta beach-parking lot", "picasso - isla vista", "reccen bike path - campus"** so please filter `litter` to those sites only.

a) Visualize `cig_count` by `site_description` with a boxplot. Must use informative labels to receive full credit.

```
filter_litter <- litter %>% filter(site_description %in% c("goleta beach-parking lot", "picasso - isla v
filter_litter %>% ggplot(aes(x=site_description, y=cig_count,fill=site_description)) +
  geom_boxplot() +
  theme_classic()+
  labs(x="Site Name",y="Cig Count", fill = "Site Name")+
  theme(axis.title = element_text(face = "bold", size = 12))
```

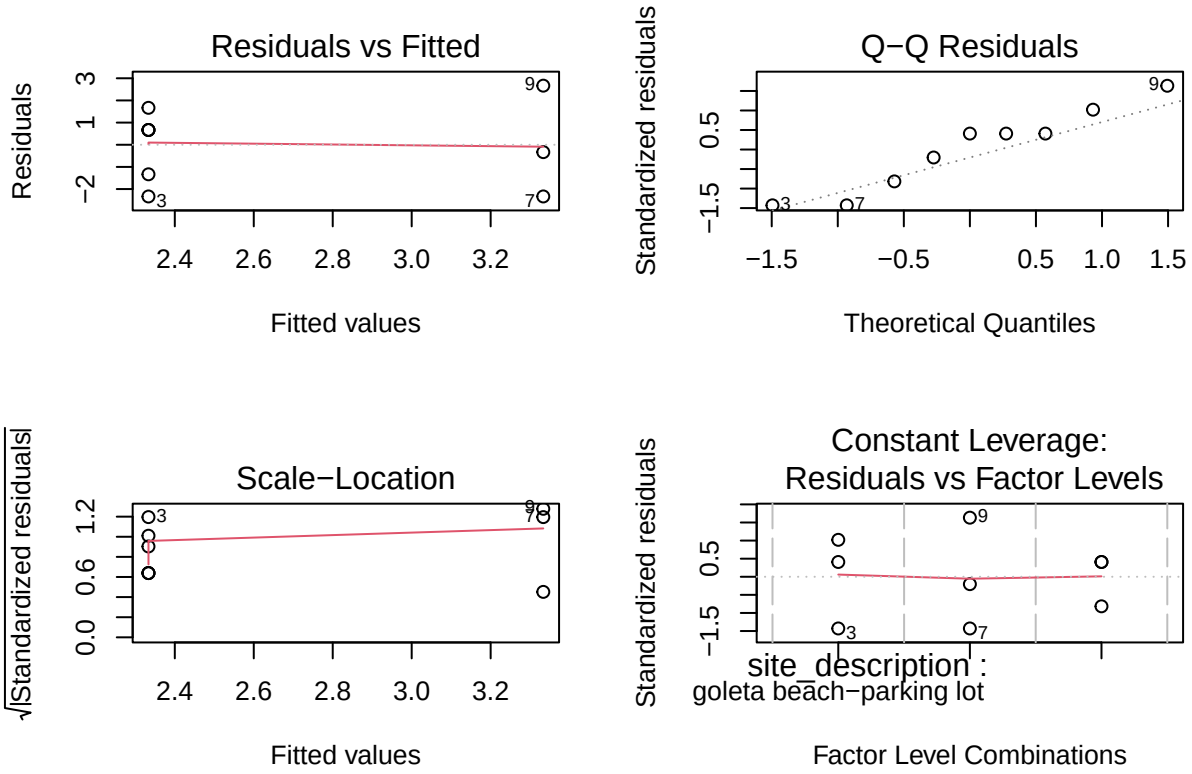


b) Fit ANOVA model with subset of data for the 3 sites

```
mod1 <- aov(cig_count ~ site_description, data =filter_litter)
```

c) Check if the residuals meet ANOVA assumptions using the plot(). Do your residuals meet the normality and variance assumptions (yes/no)? Must comment on the shape of the "Residual vs Fitted" and "Q-Q Residuals" graphs for full credit.

```
par(mfrow = c(2,2))
plot(mod1)
```



CONCLUSION: Our residuals do meet the assumption of the variances being equal across all groups becuase in the top left graph **Residuals vs Fitted** we notice that the points are relatively similar to each other which indicates that the variances among the group may be equal. On the other hand, in the plot of **Q-Q Residuals** we do not see a straight line by the residuals which indicates that the data may not be normal.

d) Regardless if you think the residuals meet the ANOVA assumptions, please interpret the ANOVA model. You must state the null and alternative hypotheses of the ANOVA model. And include the degrees of freedom, F-value, p-value, if you reject or fail the null hypothesis. After reporting these values, in 1-2 sentences discuss what this means about the 3 sites if they have similar or dissimilar amounts of cigarette counts.

```
summary(mod1)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## site_description  2      2       1    0.25  0.787
## Residuals         6     24       4
```

ANSWER: Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : Not\ all\ means\ are\ equal$$

The **DF** are 2 for sites and 6 for the residuals. The **F-value** is 0.25. The **P-value** is 0.787. Therefore, we fail to reject the null hypothesis which means that there is no signifcant difference between the cig count among the 3 sites.

e) Based on the result from the ANOVA model, should you run a post-hoc test (e.g. Tukey-Kramer)? Say yes or no and explain your answer.

ANSWER: No, we know that there is not enough evidence to show that there is a differnce in mean cig count among the three sites so we do not need to run a post hoc test to know which one is different.

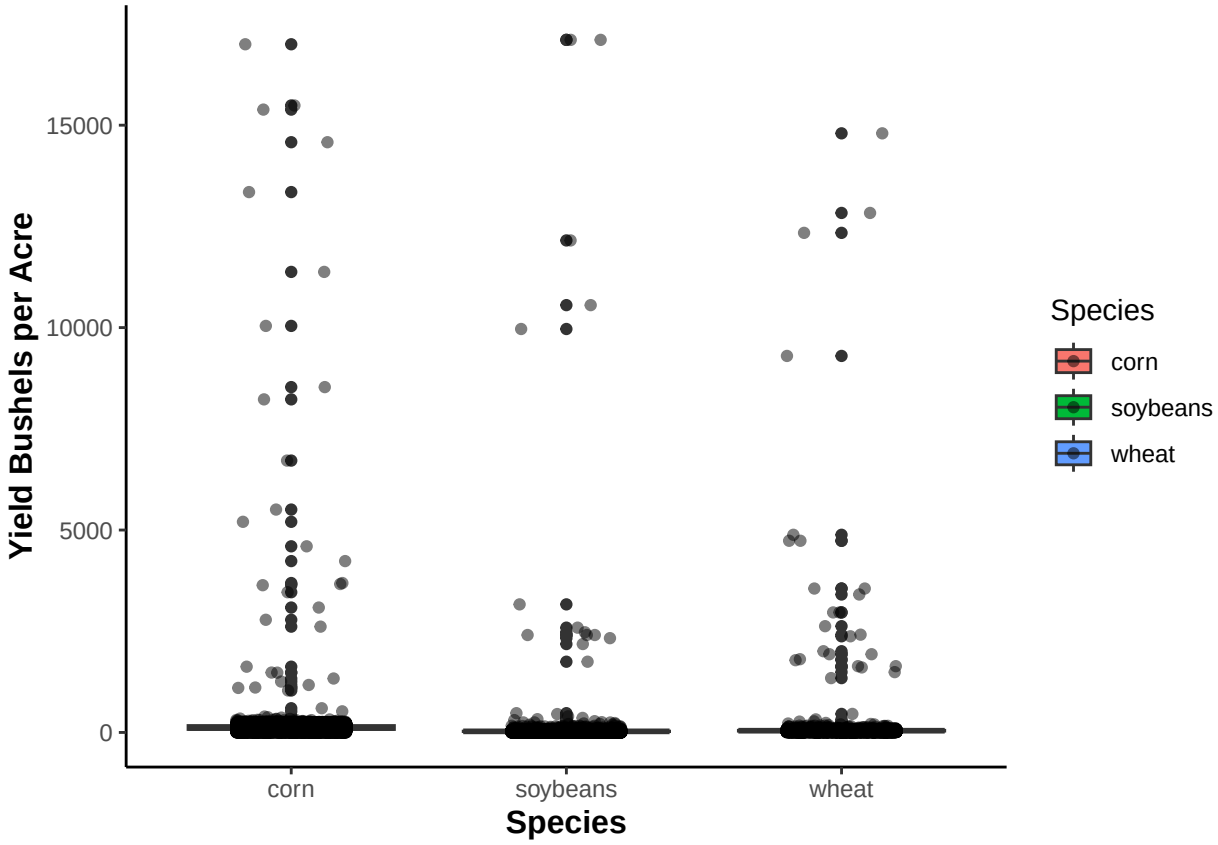
Question 5

For Question 5, you will be using the data set `crop`. These data came from the Konza Prairie Agroecosystem (KONA), which is part of another long term ecological program - NEON. The National Ecological Observatory Network NEON is another NSF supported program that has lots of job opportunities for recent grads - go check it out!

Say you are a post-doctoral researcher interested in a crop species that has a high biomass per acre (i.e. lots of data in a short amount of time). Let's run an ANOVA to test if there are certain crop `species` that have higher `yield_bushels_per_acre`.

a) Visualize `yield_bushels_per_acre` by `species` with a boxplot. Must use informative labels to receive full credit.

```
crop$species <- as.factor(crop$species)
crop %>% ggplot(aes(x=species, y=yield_bushels_per_acre, fill=species)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.5) +
  theme_classic() +
  labs(x="Species",y="Yield Bushels per Acre", fill = "Species") +
  theme(axis.title = element_text(face = "bold", size = 12))
```



b) Fit ANOVA model

```
mod2 <- aov(yield_bushels_per_acre~species, data=crop)
```

c) Check if the residuals meet ANOVA assumptions using the plot(). Do your residuals meet the normality and variance assumptions (yes/no)? Must comment on the shape of the "Residual vs Fitted" and "Q-Q Residuals" graphs for full credit.