STAT 512 Homework 7

Kathleen Wendt

04/08/2020

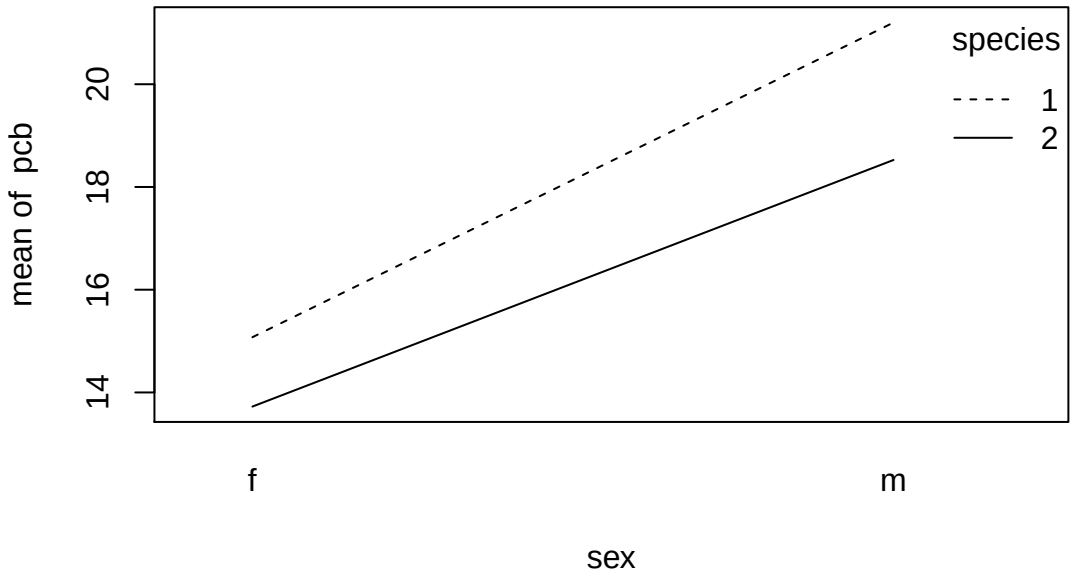
Part 1: PCB data

In this group of questions we use the "PCB.csv" data available from Canvas. A researcher is interested in comparing PCB concentrations by sex (m, f) and species (1, 2). This corresponds to four groups (sp1f, sp1m, sp2f, sp2m). Note that depending on the analysis you will use group OR sex and species in the analysis but not all three!

Question A: Table and plot

Create a table of summary statistics including sample size, mean and standard deviation for each sex, species combination. Then construct an interaction plot. For consistency, please put sex on the X axis. Include both the summary table and interaction plot in your assignment. (4 pts)

sex	species	n	mean	sd
f	1	4	15.075	1.236595
f	2	4	13.725	1.209339
m	1	4	21.200	1.329160
m	2	4	18.525	1.837344



Question B: One-way ANOVA

Fit a one-way ANOVA model to the data using group as the predictor. Construct the Type 3 ANOVA table.

```
## Anova Table (Type III tests)
##
## Response: pcb
##             Sum Sq Df  F value    Pr(>F)
## (Intercept) 4695.7  1 2309.112 4.317e-15 ***
## group        137.3  3   22.508 3.230e-05 ***
## Residuals     24.4 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question C: Estimated marginal means

With model from 2B, use `emmeans` to calculate (Tukey adjusted) pairwise comparisons for all four groups.

```
## $emmeans
##  group emmean    SE df lower.CL upper.CL
##  sp1f    15.1 0.713 12     13.5     16.6
##  sp1m    21.2 0.713 12     19.6     22.8
##  sp2f    13.7 0.713 12     12.2     15.3
##  sp2m    18.5 0.713 12     17.0     20.1
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast    estimate   SE df t.ratio p.value
##  sp1f - sp1m    -6.12 1.01 12 -6.074  0.0003
##  sp1f - sp2f     1.35 1.01 12  1.339  0.5576
##  sp1f - sp2m    -3.45 1.01 12 -3.421  0.0227
##  sp1m - sp2f     7.47 1.01 12  7.413  <.0001
##  sp1m - sp2m     2.67 1.01 12  2.653  0.0857
##  sp2f - sp2m    -4.80 1.01 12 -4.760  0.0023
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Question D: Two-way ANOVA

Fit a two-way ANOVA model to the data using sex and species as predictors. Be sure to include the interaction. Construct the Type 3 ANOVA table.

```
## Anova Table (Type III tests)
##
## Response: pcb
##             Sum Sq Df   F value    Pr(>F)
## (Intercept) 4695.7  1 2309.1121 4.317e-15 ***
## sex          119.4  1   58.6935 5.839e-06 ***
## species       16.2  1    7.9667   0.01539 *
## sex:species    1.8  1    0.8633   0.37112
## Residuals     24.4 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question E: Sex/species emmeans

Use `emmeans( , pairwise ~ sex:species)` to calculate (Tukey adjusted) pairwise comparisons for all four groups.

```
## $emmeans
##  sex species emmean    SE df lower.CL upper.CL
##  f   1         15.1 0.713 12     13.5     16.6
##  m   1         21.2 0.713 12     19.6     22.8
##  f   2         13.7 0.713 12     12.2     15.3
##  m   2         18.5 0.713 12     17.0     20.1
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast  estimate   SE df t.ratio p.value
##  f,1 - m,1    -6.12 1.01 12 -6.074  0.0003
##  f,1 - f,2     1.35 1.01 12  1.339  0.5576
##  f,1 - m,2    -3.45 1.01 12 -3.421  0.0227
##  m,1 - f,2     7.47 1.01 12  7.413  <.0001
##  m,1 - m,2     2.67 1.01 12  2.653  0.0857
##  f,2 - m,2    -4.80 1.01 12 -4.760  0.0023
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Question F: Species emmeans

Use `emmeans( , pairwise ~ species)` to calculate the pairwise comparison corresponding to the main effect of species. *Note that the p-value from this comparison should match the F-test corresponding to species from the ANOVA table from question 4.*

```
## $emmeans
##  species emmean    SE df lower.CL upper.CL
##  1         18.1 0.504 12       17     19.2
##  2         16.1 0.504 12       15     17.2
##
## Results are averaged over the levels of: sex
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  1 - 2        2.01 0.713 12 2.823   0.0154
##
## Results are averaged over the levels of: sex
```

Question G: Reflection

Consider the output from the two previous questions. From question 5, for the "m,1 - m,2" comparison you should have found an estimate = 2.67, p-value = 0.0857. From question 6, for the "1 - 2" comparison you should have found an estimate = 2.01, p-value = 0.0154. Briefly explain why we find a smaller p-value for the "1 - 2" comparison even though the estimated difference is smaller.

When more tests are conducted (6 vs. 1), there is a Tukey correction ("penalty") to minimize the family-wise error rate.

Note: The two models from above are equivalent. This can be seen by comparing the ANOVA tables (questions 2 and 4) and the pairwise comparisons (questions 3, 5). Either analysis approach is acceptable. However, one benefit of the two-way analysis is that since the interaction is not significant, we can easily discuss main effects.

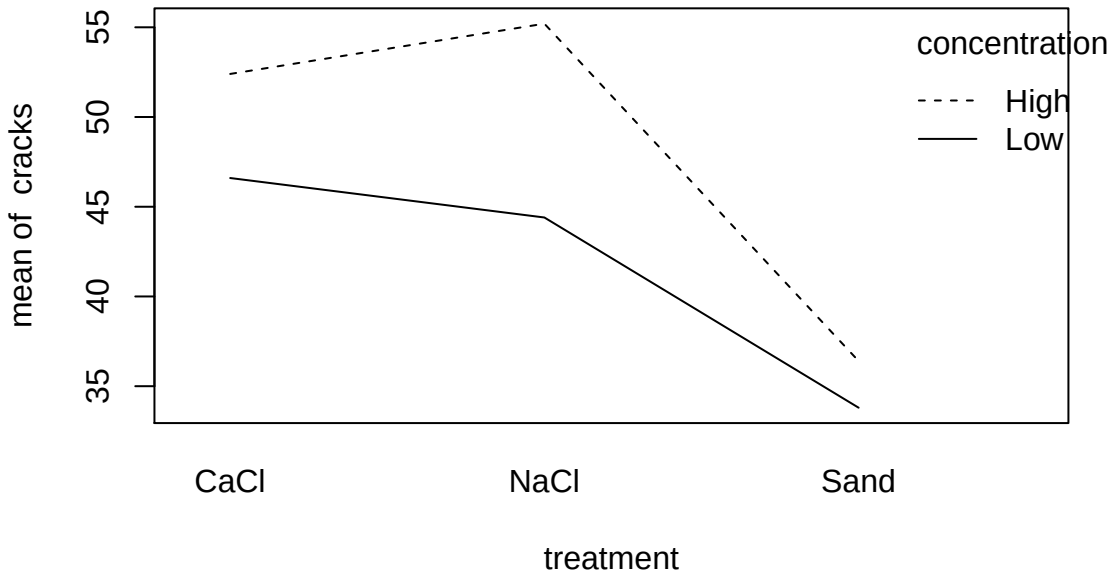
Part 2: Roadways

For this group of questions use the data described in Ott & Longnecker problem 15.14 (p 907 in the 7th Edition).

Question A: Table and plot

Create a table of summary statistics including sample size, mean and standard deviation for each Treatment*Concentration combination. Then construct an interaction plot. For consistency, please put Treatment on the X axis. Include both the summary table and interaction plot in your assignment. (4 pts)

treatment	concentration	n	mean	sd
CaCl	High	5	52.4	6.503845
CaCl	Low	5	46.6	5.594640
NaCl	High	5	55.2	7.791021
NaCl	Low	5	44.4	6.877499
Sand	High	5	36.4	4.979960
Sand	Low	5	33.8	7.463243



Question B: Describe

Describe the blocking and treatment structure.

The study examined the amount of damage (i.e., number of cracks per mile) associated with three methods for clearing snow and ice from roads (i.e., sodium chloride, calcium chloride, and sand). A *randomized block design* (i.e., traffic volume from previous winter) was used to try to minimize the effect of traffic on road damage. Each of the six *treatments* (i.e., high or low levels of each of the three aforementioned substances) was *randomly assigned* to five roads. Based on the textbook description, the treatment structure can be

considered one factor (treatment) with six levels (high and low levels of each of the three methods); according to data structure, this has a factorial design of treatment (3 levels) and concentration (2 levels) with blocking by roadway (5 roadways).

Question C: Fit model

Considering your answer to the previous question, fit an appropriate model and include the Type 3 ANOVA table in your assignment. (4 pts)

```
## Anova Table (Type III tests)
##
## Response: cracks
##                          Sum Sq Df   F value    Pr(>F)
## (Intercept)               60211  1 15999.433 < 2.2e-16 ***
## roadway                     973  4    64.646 3.740e-11 ***
## treatment                  1412  2   187.573 1.103e-13 ***
## concentration               307  1    81.630 1.694e-08 ***
## treatment:concentration      85  2    11.346 0.0005091 ***
## Residuals                    75 20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question D: Blocking

Discuss the value of the blocking for this data. Justify your response with appropriate test-statistic(s) and p-value(s).

Blocking by roadway is effective in this study, $F = 64.646$, $p = 0$. This reduces the variability in outcomes attributed to a nuisance variable (roadway).

Question E: Compare concentration by treatment

Compare mean response for High vs Low Concentration separately for each Treatment. This can be done using `emmeans`. Include the `emmeans contrasts` output in your assignment, but also (briefly) summarize your findings. (4 pts)

```
## treatment = CaCl:
##  contrast   estimate   SE df t.ratio p.value
##  High - Low      5.8 1.23 20 4.727   0.0001
##
## treatment = NaCl:
##  contrast   estimate   SE df t.ratio p.value
##  High - Low     10.8 1.23 20 8.803   <.0001
##
## treatment = Sand:
##  contrast   estimate   SE df t.ratio p.value
##  High - Low      2.6 1.23 20 2.119   0.0468
##
## Results are averaged over the levels of: roadway
```

For each treatment type, there was a higher number of cracks in the road for high concentrations, compared to low concentrations, using $\alpha = 0.05$.

Question F: Compare treatment by concentration

Compare mean responses between the 3 Treatments separately for each Concentration. This can be done using `emmeans` (default Tukey adjustment is fine). Include the `emmeans contrasts` output in your assignment, but also (briefly) summarize your findings for Concentration = Low. (4 pts)

```
## concentration = High:
##  contrast    estimate   SE df t.ratio p.value
##  CaCl - NaCl     -2.8 1.23 20 -2.282  0.0817
##  CaCl - Sand     16.0 1.23 20 13.041  <.0001
##  NaCl - Sand     18.8 1.23 20 15.323  <.0001
##
## concentration = Low:
##  contrast    estimate   SE df t.ratio p.value
##  CaCl - NaCl      2.2 1.23 20  1.793  0.1974
##  CaCl - Sand     12.8 1.23 20 10.433  <.0001
##  NaCl - Sand     10.6 1.23 20  8.640  <.0001
##
## Results are averaged over the levels of: roadway
## P value adjustment: tukey method for comparing a family of 3 estimates
```

At a low concentration level, there were differences in damage for calcium chloride vs. sand and sodium chloride vs. sand, such that roads treated with calcium chloride or sodium chloride had more cracks than those treated with sand. There was no difference in damage between calcium chloride and sodium chloride.

Appendix

```
# load packages
library(tidyverse)
library(janitor)
library(kableExtra)
library(car)
library(emmeans)
library(broom)
# set global options
knitr::opts_chunk$set(fig.width = 6,
                      fig.height = 4,
                      fig.path = "figs/",
                      echo = FALSE,
                      warning = FALSE,
                      message = FALSE)
# 1. load and prepare pcb data
pcb_data <- readr::read_csv("data/PCB.csv") %>%
  dplyr::mutate(sex = as.factor(sex),
                species = as.factor(species),
                group = as.factor(group))
# 1a. create summary statistics table for pcb data
pcb_table <- pcb_data %>%
  dplyr::group_by(sex, species) %>%
  dplyr::summarize(
    n = n(),
    mean = mean(pcb),
    sd = sd(pcb))
# 1a. kable pcb sum stat table
kableExtra::kable(pcb_table)
# 1a. create interaction plot
with(interaction.plot(x.factor = sex,
                      trace.factor = species,
                      response = pcb),
     data = pcb_data)
# 1b. change contrast defaults
options(contrasts = c("contr.sum", "contra.poly"))
# 1b. construct one-way anova model with group as predictor
pcb_1anova <- lm(pcb ~ group, data = pcb_data)
car::Anova(pcb_1anova, type = 3)
# 1c. extract emmeans for group
emmeans::emmeans(pcb_1anova, pairwise ~ group)
# 1d. change contrast defaults
options(contrasts = c("contr.sum", "contra.poly"))
# 1d. build two-way anova model
pcb_2anova <- lm(pcb ~ sex*species, data = pcb_data)
car::Anova(pcb_2anova, type = 3)
# 1e. extract emmeans for group by sex*species
emmeans::emmeans(pcb_2anova, pairwise ~ sex:species)
# 1f. extract emmeans for group by species
emmeans::emmeans(pcb_2anova, pairwise ~ species)
# 2. load and prepare roadway data
road_data <- readxl::read_xlsx("data/ex15-14.xlsx") %>%
```

```
  janitor::clean_names() %>%
  dplyr::mutate(roadway = as.factor(roadway),
                treatment = as.factor(treatment),
                concentration = as.factor(concentration))
# 2a. create summary statistics table for road data
road_table <- road_data %>%
  dplyr::group_by(treatment, concentration) %>%
  dplyr::summarize(
    n = n(),
    mean = mean(cracks),
    sd = sd(cracks)
    )
# 2a. kable road stat table
kableExtra::kable(road_table)
# 2a. create interaction plot
with(interaction.plot(x.factor = treatment,
                      trace.factor = concentration,
                      response = cracks),
     data = road_data)
# 2c. adjust contrast default
options(contrasts = c("contr.sum", "contr.poly"))
# 2c. fit two-way anova
road_anova <- lm(cracks ~ roadway + treatment*concentration,
                 data = road_data)
# 2c. call anova type 3 table
car::Anova(road_anova, type = 3)
road_anova_tidy <- broom::tidy(car::Anova(road_anova, type = 3))
# 2e. extract emmeans for high vs low concentration by treatment
road_con_emmeans <- emmeans::emmeans(road_anova,
                                     pairwise ~ concentration|treatment)
# 2e. only show contrasts - diff between concen level for each trt
road_con_emmeans$contrasts
# 2f. extract emmeans for each trt by concentration level
road_trt_emmeans <- emmeans::emmeans(road_anova,
                                     pairwise ~ treatment|concentration)
# 2f. only show contrasts - diff between trt for each concen level
road_trt_emmeans$contrasts
```