ANOVA as a linear model

KEY

Wed Mar 10 13:47:13 2021

Two species

We will work with data on the length of cuckoo eggs deposited in the nests of different host species. To begin import the "two_sp" worksheet from the cuckoo_all_species.xls spreadsheet you downloaded.

```
library(readxl)
twosp <- read_excel("cuckoo_all_species.xls","two_sp")
```

Add a "Wren" dummy variable to twosp:

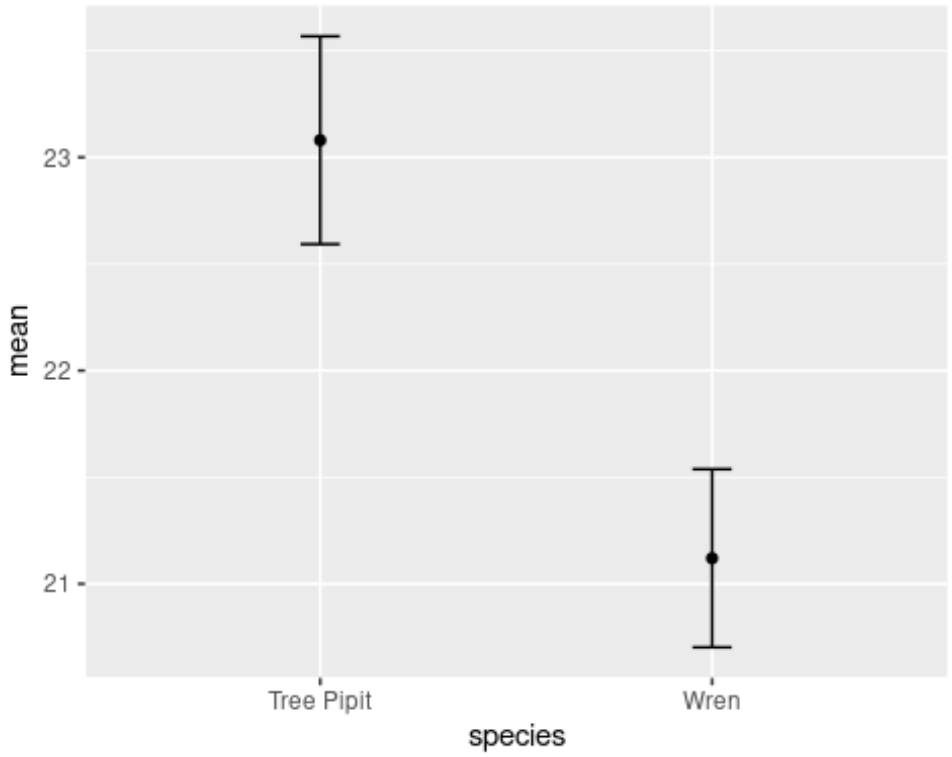
```
as.numeric(twosp$species == "Wren") -> twosp$Wren
```

Calculate summary statistics for the data, needed for plotting:

```
library(biol531)
summarySE(twosp, "length", "species") -> twosp.sumstats
```

Plot the means and confidence intervals:

```
library(ggplot2)
ggplot(twosp.sumstats, aes(x = species, y = mean)) + geom_point() +
geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.1)
```



Question: based on the graph does there appear to be a difference between species? How do you know?

Yes, Wrens have shorter cuckoo egg lengths than Tree Pipits do. Since the 95% CI's aren't overlapping this difference will be statistically significant.

Question: what does a 95% confidence interval tell you?

It predicts the range of values that has a 95% chance of containing the (unknown) population mean. Since a sample mean is an estimate of a population mean, we need a way of telling how precise the estimate is, and a confidence interval helps express how far from the true value we expect that we might be, and gives a range of plausible values for the population mean given the information we have about it in our sample of data.

Run the ANOVA, in the usual way (using length as the response variable, and species as the grouping variable):

```
lm(length ~ species, data = twosp) -> twosp.anova.lm
anova(twosp.anova.lm)

## Analysis of Variance Table
##
## Response: length
##           Df Sum Sq Mean Sq F value   Pr(>F)
## species    1 28.812 28.8120  42.893 4.23e-07 ***
## Residuals 28 18.808  0.6717
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now you can run the analysis using Wren as a predictor in a regression analysis:

```
lm(length ~ Wren, data = twosp) -> twosp.regression.lm
anova(twosp.regression.lm)

## Analysis of Variance Table
##
## Response: length
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Wren       1 28.812 28.8120  42.893 4.23e-07 ***
## Residuals 28 18.808  0.6717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(twosp.regression.lm)

##
## Call:
## lm(formula = length ~ Wren, data = twosp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.980 -0.365  0.100  0.670  1.180
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.0800     0.2116 109.066  < 2e-16 ***
## Wren         -1.9600     0.2993  -6.549 4.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8196 on 28 degrees of freedom
## Multiple R-squared:  0.605,  Adjusted R-squared:  0.5909
## F-statistic: 42.89 on 1 and 28 DF,  p-value: 4.23e-07
```

Question: Report the Wren coefficient; what does it represent?

> The wren coefficient is -1.96, and it is the difference between the Hedge Sparrow and Wren means.

Question: Is there a statistically significant difference between species? Do you need to run a Tukey HSD procedure to knwo which species are different for this analysis? Why or why not?

> Yes the difference is statistically significant, and since there are only two species being compared the p-value on the “species” term tells us those two species are different. We don’t need to do a Tukey procedure.

Six species

Now we will repeat this process using six species.

Import the data:

```
allsp <- read_excel("cuckoo_all_species.xls","all_sp")
```

Dummy code the columns:

```
as.numeric(allsp$species == "Meadow Pipit") -> allsp$Meadow.Pipit
as.numeric(allsp$species == "Pied Wagtail") -> allsp$Pied.Wagtail
as.numeric(allsp$species == "Robin") -> allsp$Robin
as.numeric(allsp$species == "Tree Pipit") -> allsp$Tree.Pipit
as.numeric(allsp$species == "Wren") -> allsp$Wren
```

Question: What should these dummy coded columns have in them for a row that contains a Hedge Sparrow? Why? Double check your allsp data frame and confirm that this is the case.

> All of the dummy coded variables should have a 0 when the row is from a Hedge Sparrow. This is because a Hedge Sparrow is not a Meadow Pipit, Pied Wagtail, Robin, Tree Pipit, or Wren, and these columns only contain a 1 when the species is the same as the column heading.

Question: Why isn't there a column for Hedge Sparrow? Does it have to be Hedge Sparrow, or could any of the species be omitted?

> The dummy coded columns collectively indicate the species. Each gets a 1 when the species is the same as the column heading, so five of the species are identified by having a 1 in the appropriate column. The final "baseline" species is indicated by 0's in every column. So, five columns are all that are needed to indicate six species.

Question: We are doing this dummy coding so that we can use regression analysis to conduct an ANOVA comoparing species means. Why not just use a single column with numbers instead of species names (1 for Hedge Sparrow, 2 for Meadow Pipit, etc.)? Would the analysis work correctly (why or why not)?

> This would force the species to be different by the same amount. If we had a single predictor ("species.number") with 1,2,3,4,5,6 assigned in place of the species name, then a slope coefficient for that column would impose a difference equal to the slope between every pair of species. For example, an intercept of 20 with a slope of 0.5 would result in predicted values of 20.5, 21.0, 21.5, 22.0, 22.5, and 23.0 for Hedge Sparrow, Meadow Pipit, Pied Wagtail, Robin, Tree Pipit, and Wren, respectively - the difference between species means is clearly not the same, so this would not work. Also, the assignment of numbers to species would be arbitrary, since species is not an ordered variable - whatever number assignment we used would force an ordering on the species that may not match reality. Using

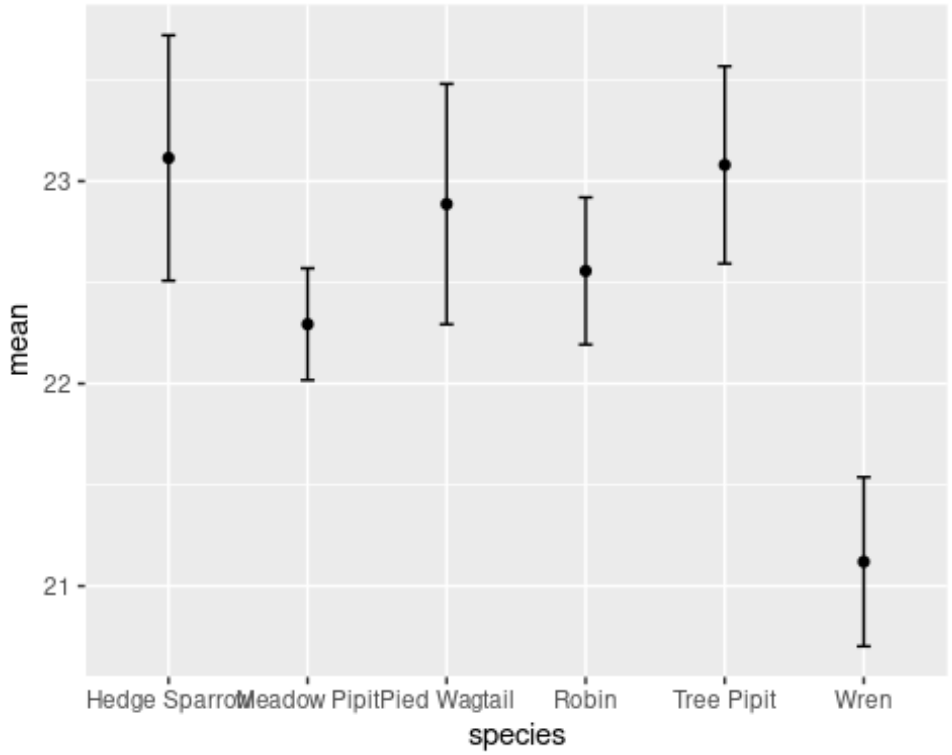
dummy variables allows us to reproduce the actual group means, with no distortion of the relationships between them.

Graph the six species data - first, summarize the data by species:

```
summarySE(allsp, "length","species") -> allsp.sumstats
```

Plot the means and 95% confidence intervals:

```
ggplot(allsp.sumstats, aes(x = species, y = mean)) + geom_point() +
geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.1)
```



Question: Based on the graph, which species is most different from the others?

Wrens are smaller than all the other species.

ANOVA of the six species data set:

```
lm(length ~ species, data = allsp) -> allsp.lm
anova(allsp.lm)
```

```
## Analysis of Variance Table
##
## Response: length
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     5  42.81  8.5620  10.449 2.852e-08 ***
## Residuals 114  93.41  0.8194
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: Are there significant differences between species? Can you tell at this point which species are different (why or why not)?

> Yes, the species term is statistically significant so there are differences between at least some of the species. We don't know which are different becasue there are more than two in the data set. We would need a Tukey HSD, or some other post-hoc procedure to find out which means are different.

Now analyze the same data using the dummy-coded variables. This is a multiple linear regression of the six species data set:

```
lm(length ~ Meadow.Pipit + Pied.Wagtail + Robin + Tree.Pipit + Wren, data =
allsp) -> allsp.regression
anova(allsp.regression)
```

```
## Analysis of Variance Table
##
## Response: length
##               Df Sum Sq Mean Sq F value    Pr(>F)
## Meadow.Pipit   1  1.767  1.7672  2.1567 0.1446995
## Pied.Wagtail   1  2.202  2.2016  2.6869 0.1039311
## Robin          1  0.209  0.2092  0.2553 0.6143459
## Tree.Pipit     1  9.832  9.8319 11.9991 0.0007507 ***
## Wren           1 28.800 28.8002 35.1486 3.329e-08 ***
## Residuals    114 93.410  0.8194
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Question: At this point only the residual line should be the same between this table and your ANOVA table, above. Is it the same?

> Yes it is.

To see if the rest of the table for the multiple regression version of the analysis matches the ANOVA version you need to sum the SS for each of the dummy coded variables. Confirm that the SS for the five dummy coded predictors is the same as the species SS in your ANOVA:

```
anova(allsp.regression) -> allsp.aovtable
sum(allsp.aovtable[c("Meadow.Pipit","Pied.Wagtail","Robin","Tree.Pipit","Wren
"),"Sum Sq"])
```

```
## [1] 42.81015
```

Question: Is the sum of the five dummy coded variable SS equal to the species SS from the ANOVA table above?

> Yes it is.

Confirm that the regression using five dummy coded variables is predicting the mean length for each species.

Coefficient estimates from the species lm:

```
coefficients(allsp.lm) -> allsp.coeff
allsp.coeff

##         (Intercept) speciesMeadow Pipit speciesPied Wagtail
speciesRobin
##         23.11428571         -0.82095238         -0.22761905         -
0.55803571
##   speciesTree Pipit         speciesWren
##         -0.03428571         -1.99428571
```

Question: Can you tell from these tests of coefficients if Meadow Pipit is different from Robin? If so, how? If not, what would you need to do to test for this difference?

No, these are all tests of the difference from Hedge Sparrow. To compare against any of the other species we would need to do a Tukey HSD, or some other post-hoc procedure.

```
allsp.sumstats$mean.glm <- c(allsp.coeff["(Intercept)"],
allsp.coeff["(Intercept)"] + allsp.coeff[c("speciesMeadow Pipit",
"speciesPied Wagtail", "speciesRobin", "speciesTree Pipit", "speciesWren")])
allsp.sumstats

##         species     mean        sd  n        se    lower    upper mean.glm
## 1 Hedge Sparrow 23.11429 1.0494373 14 0.2804739 22.50836 23.72021 23.11429
## 2  Meadow Pipit 22.29333 0.9195849 45 0.1370836 22.01706 22.56961 22.29333
## 3  Pied Wagtail 22.88667 1.0722917 15 0.2768645 22.29285 23.48048 22.88667
## 4         Robin 22.55625 0.6821229 16 0.1705307 22.19277 22.91973 22.55625
## 5    Tree Pipit 23.08000 0.8800974 15 0.2272402 22.59262 23.56738 23.08000
## 6          Wren 21.12000 0.7542262 15 0.1947404 20.70232 21.53768 21.12000
```

Question: Compare the "mean" column with this new "mean.glm" column in allsp.sumstats - are they the same?

Yes they are identical.

Question: Were the predicted values for each species, using the app on the web page, the same as the group means?

Yes they were the same.