R Lab #10b - ANOVA of indigobird feather reflectance

Contents

Indigobirds

Indigobirds (the subject of my PhD research) are birds in which different species are very difficult to tell apart because they look so similar to each other. Males have feathers that are generally black, although species also seem to have slightly different iridescent "sheen" coloration that is evident under optimal light conditions or when holding a bird. Are there really quantifiable differences among species in feather coloration? I used a spectrophotometer to measure feather hue, the wavelength of light with maximum reflectance, from samples of four species. Do all species have the same mean feather hue?

Import and explore data

First, we import the data (file: indigobirds_reflectance.csv), look at the structure of the data frame, and calculate some descriptive statistics.

```
indigoData <- read.csv("indigobirds_reflectance.csv", stringsAsFactors = T)
str(indigoData)
## 'data.frame':    94 obs. of  2 variables:
##  $ species: Factor w/ 4 levels "chalybeata","codringtoni",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ hue    : num  458 455 439 454 436 ...
tapply(indigoData$hue, indigoData$species, mean)
##   chalybeata  codringtoni      funerea purpurascens
##     459.2859     466.6666     407.4442     385.2294
tapply(indigoData$hue, indigoData$species, sd)
##   chalybeata  codringtoni      funerea purpurascens
##     15.82197     11.89263     11.83731     17.43100
tapply(indigoData$hue, indigoData$species, length)
##   chalybeata  codringtoni      funerea purpurascens
##           17           21           35           21
```

We see that there are two columns in the data frame: one factor (categorical) that designates species and one numerical that has the hue measurements.

From the descriptive statistics we see that the four species are somewhat similar in mean hue, standard deviation, and sample size.

Assumptions of ANOVA

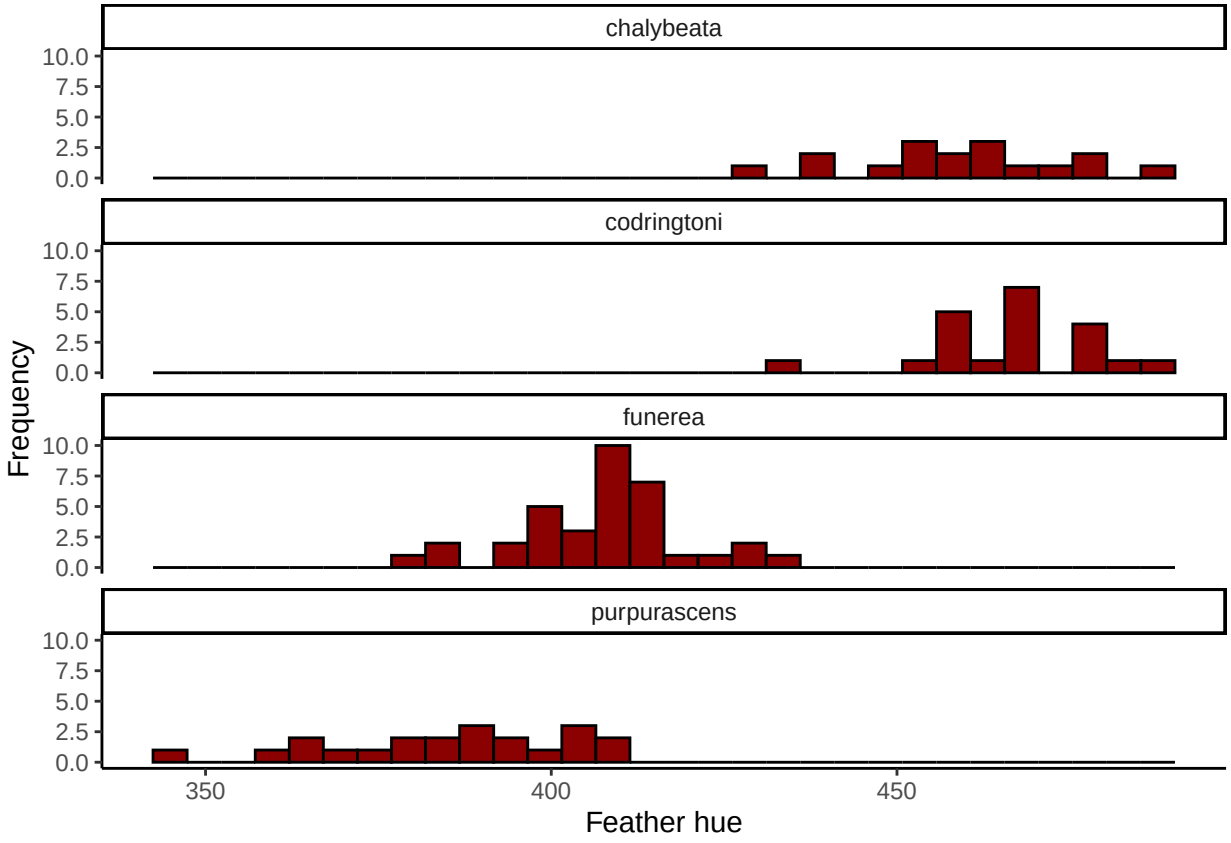
An analysis of variance has some underlying assumptions:

- Measurements from each group represent a random sample from its corresponding population
- Variable is normally distributed in each of the k populations
- Variance is the same in all k populations

Assumption of normality in each population

ANOVA is robust to deviations from normality, but let's have a look with frequency distributions:

```
library(ggplot2)
ggplot(indigoData, aes(x = hue)) + geom_histogram(color = "black",
    fill = "darkred") + facet_wrap(~species, ncol = 1) + xlab("Feather hue") +
    ylab("Frequency") + theme_classic()
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Hmm, not too bad but maybe an outlier in *codringtoni*. Let's check each group with a Shapiro-Wilk test:

```
tapply(indigoData$hue, indigoData$species, shapiro.test)
## $chalybeata
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98065, p-value = 0.9627
```

```
##
##
## $codringtoni
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.92087, p-value = 0.09033
##
##
## $funerea
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.96281, p-value = 0.2772
##
##
## $purpurascens
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.96391, p-value = 0.5981
```

Ok, for all four species the samples **fail to reject** the null hypothesis of normality.

Assumption of equal standard deviations

The textbook notes that ANOVA is robust to departures of this assumption, with up to a 10x difference allowed if sample sizes are large and about the same size.

Although only one species has a sample size above 30, the standard deviations are all within 2x of each other. Let's run the Levene's test to check:

```
library(car)
leveneTest(indigoData$hue ~ indigoData$species, center = mean)
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value Pr(>F)
## group  3  2.1412 0.1005
##       90
```

Since the P-value is above 0.05 we fail to reject the null hypothesis of equal variances among groups.

Running the ANOVA

Now we want to run the ANOVA to test the null hypothesis that the mean hue of each species is equal. This is done in R with the command lm(), which refers to linear model. Note that the syntax supplies the name of the data frame (indigo), and a "formula" of categorical variable ~ numerical variable (see R Lab #10). It's good practice to "store" the linear model to a variable and then run anova() on the stored model, which will generate an ANOVA table:

```
indigoAnova <- lm(hue ~ species, data = indigoData)
anova(indigoAnova)
## Analysis of Variance Table
##
```

```
## Response: hue
##           Df Sum Sq Mean Sq F value    Pr(>F)
## species    3 100444   33481  170.49 < 2.2e-16 ***
## Residuals 90  17675     196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-value of the test is <2.2 e-10, which is below 0.05 and leads us to **reject** the null hypothesis that all groups have equal means. So we know that we have significant variation among group means. That is, at least one group has a mean that is significantly different from at least one other group.

Calculating R-squared

With the summary() command you can see details of the ANOVA, including the R-squared statistic (0.829) that reports the proportion of total variance explained by group differences. If you only want to see the R-squared then output only that with $r.squared:

```
summary(indigoAnova)$r.squared
## [1] 0.8503631
```

This tells us that ~83% of the variation in the data is due to differences among groups, which is a high value.

Post-hoc comparisons

Based on the ANOVA result we can reject the null hypothesis, but we do not know which group means are different from each other. To do this we run a post-hoc analysis that does pairwise comparisons among groups but adjusts for testing multiple hypotheses (i.e., running multiple tests). The most commonly used test for this is the Tukey-Kramer method.

This is run with the command TukeyHSD(). Note that this is also called the Tukey Honest Significant Differences test. We also need to apply the command aov() to our model, so the syntax looks like this:

```
TukeyHSD(aov(indigoAnova))
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = indigoAnova)
##
## $species
##                                diff        lwr       upr     p adj
## codringtoni-chalybeata     7.380718  -4.587616  19.34905 0.3758429
## funerea-chalybeata       -51.841626 -62.686367 -40.99689 0.0000000
## purpurascens-chalybeata  -74.056451 -86.024785 -62.08812 0.0000000
## funerea-codringtoni      -59.222344 -69.348074 -49.09661 0.0000000
## purpurascens-codringtoni -81.437169 -92.758079 -70.11626 0.0000000
## purpurascens-funerea     -22.214825 -32.340555 -12.08910 0.0000007
```

The output includes the groups being compared, difference in means, the confidence interval for the difference in means, and the P-value (adjusted for multiple tests) of the comparison.

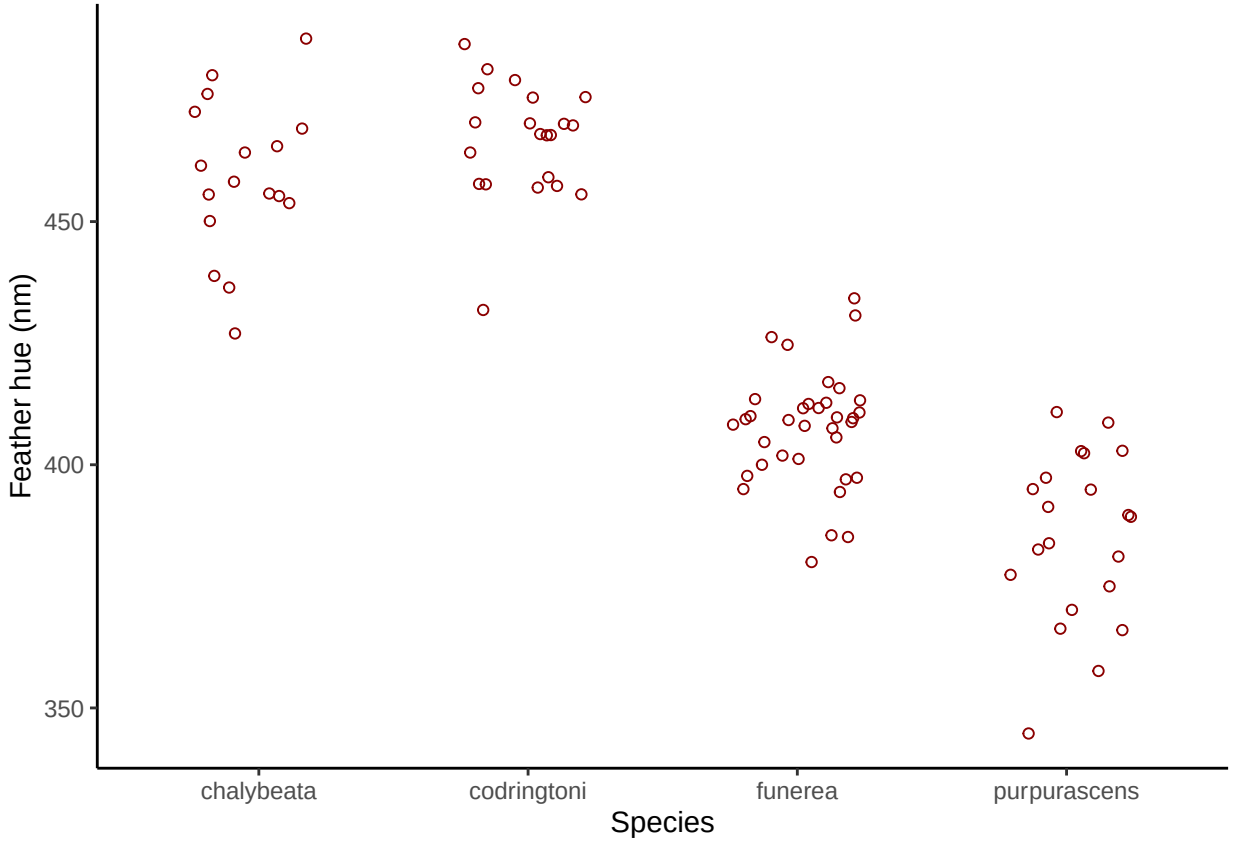
The output shows us that with the exception of chalybeata versus codringtoni the pairwise comparisons of means are all significant.

Illustrating results in a single plot

It would be useful to add summarize all of this work in a single plot.

First, let's create a strip plot (although a box plot or violin plot would be goood too).

```
ggplot(indigoData, aes(x = species, y = hue)) + geom_jitter(pch = 1,
    color = "darkred", width = 0.25) + xlab("Species") + ylab("Feather hue (nm)") +
    theme_classic()
```



Now let's add letters to the plot that show significantly different groups (see Figure 15.4-1).

It is typical to start with the group farthest to the left (chalybeata) and assign it "a." The next group is codringtoni, which is not significantly different from chalybeata, so it also get's an "a."

The next groups, funerea and purpuascens, are significantly different from all other groups (including each other). Therefore, each needs to have a unique letter. Keeping with a left-to-right pattern, we'll assign "b" to funerea and "c" to purpurascens.

We can add these letters to our plot by including annotate() to our ggplot line (see previous "making better graphs" lab). One option is to put each letter at the same y-value across the top of the graph:

```
ggplot(indigo, aes(x = species, y = hue)) + geom_jitter(pch = 1,
    color = "darkred", width = 0.25) + xlab("Species") + ylab("Feather hue (nm)") +
    theme_classic() + annotate(geom = "text", x = 1, y = 500,
    label = "a") + annotate(geom = "text", x = 2, y = 500, label = "a") +
    annotate(geom = "text", x = 3, y = 500, label = "b") + annotate(geom = "text",
    x = 4, y = 500, label = "c")
## Error in ggplot(indigo, aes(x = species, y = hue)): object 'indigo' not found
```

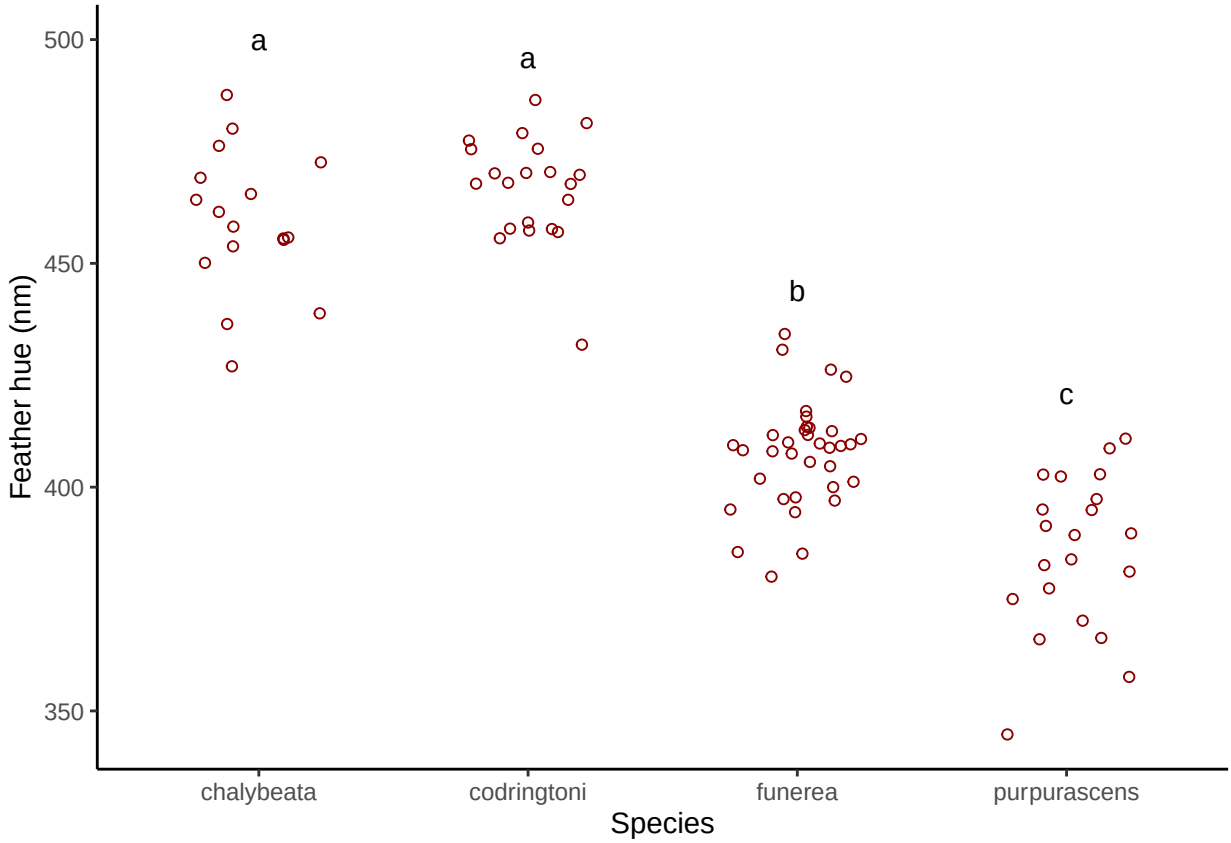
Another option is to put the letters just above the maximum value for each group. Let's get the max of each group and add 10 (how much you might add depends on the scale of the variable that you're working with):

```
tapply(indigoData$hue, indigoData$species, max) + 10
##   chalybeata  codringtoni      funerea purpurascens
```

```
##     497.6250     496.5000     444.2222     420.8333
```

. . . and use these approximate values for the y-coordinate of the corresponding letter:

```
ggplot(indigoData, aes(x = species, y = hue)) + geom_jitter(pch = 1,
    color = "darkred", width = 0.25) + xlab("Species") + ylab("Feather hue (nm)") +
    theme_classic() + annotate(geom = "text", x = 1, y = 500,
    label = "a") + annotate(geom = "text", x = 2, y = 496, label = "a") +
    annotate(geom = "text", x = 3, y = 444, label = "b") + annotate(geom = "text",
    x = 4, y = 421, label = "c")
```



Now we have a nice graph that shows all of the data and the results of the ANOVA post-hoc test!

R commands summary

- **One-way ANOVA**
  - lm(y~x,data=df)
- **ANOVA table**
  - anova(anovaModel)
- **R-squared statistic from anova**
  - summary(anovaModel)$r.squared
- **Tukey-Kramer test**
  - TukeyHSD(aov(anovaModel))