Module 3: Comparing (Multiple) Means

Ellen Bledsoe

2024-03-21

Comparing (Multiple) Means

So far in this module, we have discussed comparing the means of two groups in multiple ways: numerically, visually, and statistically. Specifically, we learned about running t-tests to statistically compare the means of two groups.

What if we have more than 2 groups, though? Often, we have 3, 4, 12…any number of groups that we might want to compare against each other. How do we do that?

ANOVA

What is an ANOVA? It stands for **AN**alysis **O**f **VA**riance.

Why is it called that? Well, ANOVAs compare the amount of variability in the data *between* and *within* groups. In plainer terms, it is looking at how different the data in each group are from each other (within) and also how different the groups are from the other groups (between).

The group of statistical tests called ANOVAs are extensions of the t-test, using the same concepts but allowing us to compare the means of more than two groups at a time. There are lots of different types of ANOVAs, but we are going to stick with the most simple version.

When to use an ANOVA

To run an ANOVA, the following things need to be true:

- The variable that we use to create our groups (independent variable) is qualitative (categorical).
- The variable that we want to compare between groups (dependent or response variable) is qunatitative (numerical).

These two aspects above might sound familiar, because they also need to be true for t-tests!

- When we only have 2 groups to compare, we can use a t-test.
- When we have *more* than 2 groups to compare, we use an ANOVA.

Hypothesis Testing

Like t-tests, ANOVAs are another example of statistical hypothesis testing. We always want to formulate a null hypothesis and an alternative hypothesis. The generic version of these for any ANOVA are:

Null Hypothesis (H_0): The means of the populations we sampled from are all equal.
Alternative Hypothesis (H_A): The means of the populations we sampled from differ from each other.

Let's Run an Example

Last week, we ran t-tests on the battery life and signal distance variables in order to determine if there was a statistically significant difference between two groups (the collar makers).

What if our large mammal team had deployed collars from *three* different companies rather than just two? How would we compare the means numerically, visually, and statistically? Let's find out.

First up, in small groups, work through the set-up, numeric, and visual sections. We'll spend 15-20 minutes on this.

Set-Up

First, as always, we need to load the `tidyverse` and the read in our collars data. This time, use the file `more_collars.csv`.

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
collars <- read_csv("../data/more_collars.csv")
```

```
## Rows: 150 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): maker
## dbl (4): collar_id, battery_life, signal_distance, fail
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Conceptual

What questions are we trying to answer about these collars?

- Which variable is the independent variable? What about the dependent variable?

Let's write out our hypotheses.

First, for battery life:

Null hypothesis: There is no difference in the average battery life between the three collar makers.
Alternative hypothesis: There is a difference in the average battery life the three collar makers.

Next, for signal distance:

Null hypothesis: There is no difference in the average signal distance between the three collar makers.
Alternative hypothesis: There is a difference in the average signal distance between the three collar makers.

Numeric

Let's calculate some summary values for our data to get some preliminary information about our collar companies.

Calculate the means and the standard deviations (`sd()`) for the battery life and the signal distance for *each* collar maker.

```
summary_stats <- collars %>%
  group_by(maker) %>%
  summarise(mean_battery = mean(battery_life),
            sd_battery = sd(battery_life),
            mean_signal = mean(signal_distance),
            sd_signal = sd(signal_distance))

summary_stats
```

```
## # A tibble: 3 x 5
##   maker                     mean_battery sd_battery mean_signal sd_signal
##   <chr>                            <dbl>      <dbl>       <dbl>     <dbl>
## 1 Budget Collars LLC                86.8       9.64       4302.      31.4
## 2 Collarium Inc.                   121.       11.8        4195.      35.8
## 3 Darn Tuff Collars Company        118.       10.7        4172.      41.4
```

Let's examine the data. Do you think the collar companies are all different from each other? Are there pairs of companies that look more similar?

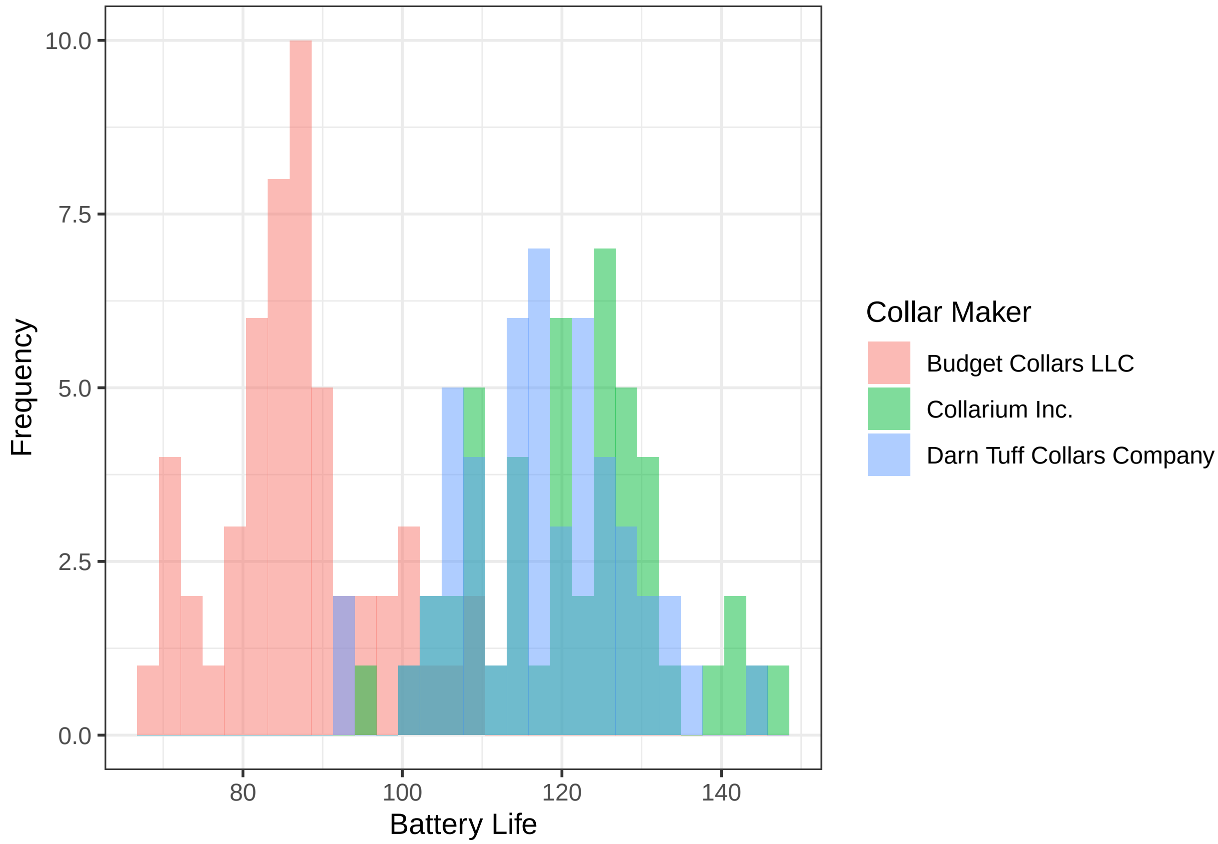
Visual

As always, we love (<3) visualizing our data, especially as a form of data exploration and comparing means.

First, explore the distribution of battery life by maker using a multiple histogram (no need to add a line for the mean, but you can if you really want to!).

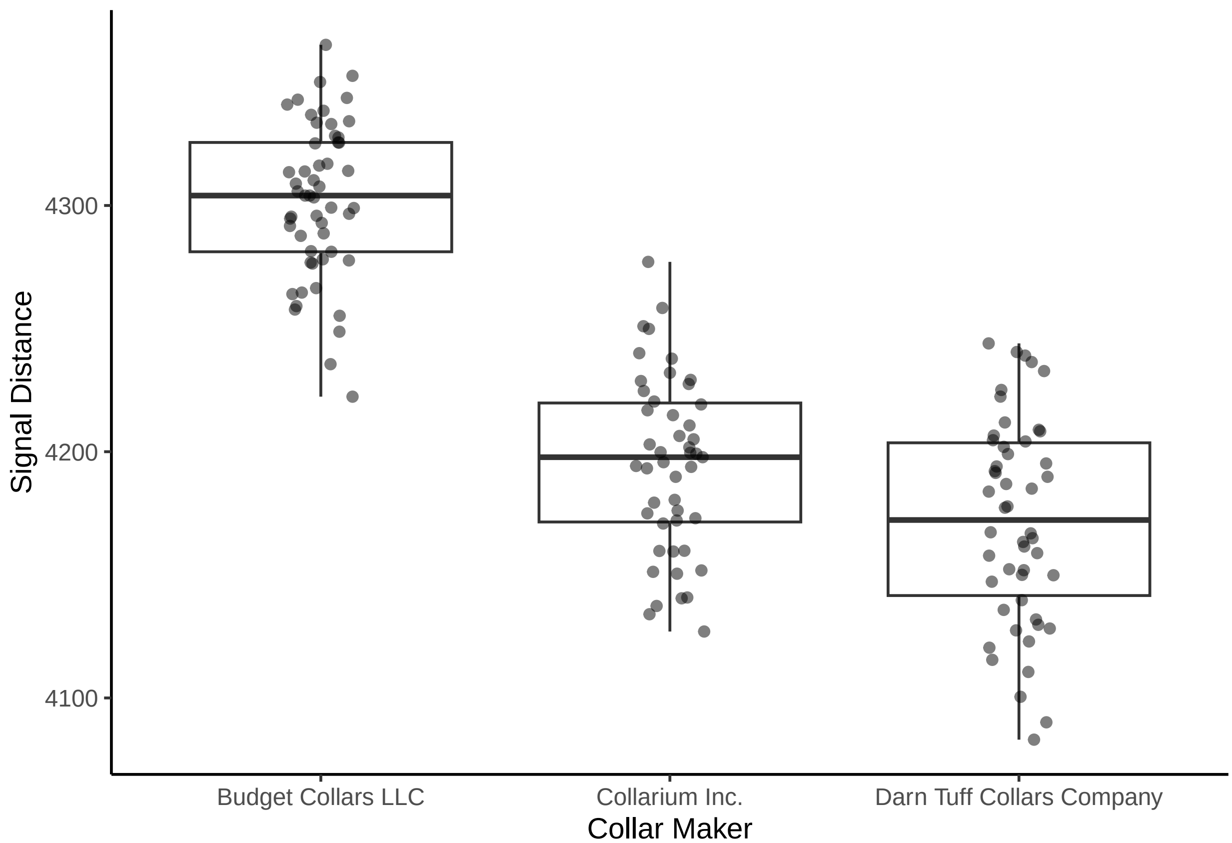
```
# battery life
ggplot(collars, aes(battery_life, fill = maker)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  labs(x = "Battery Life",
       y = "Frequency",
       fill = "Collar Maker") +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Let's do a different type of visualization for signal distance. What type of data visualization have we covered that does a good job of showing both central tendency (median, in this case) and variation (middle 50% and 95% confidence intervals) in one continuous variable?

```
# signal distance
ggplot(collars, aes(maker, signal_distance)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, alpha = 0.5) +
  labs(x = "Collar Maker",
       y = "Signal Distance") +
  theme_classic()
```



Based on the numbers and your visualizations, what is your prediction? Are the 3 groups different *enough*, whatever that means?

Statistic

Now that we have explored the data numerically and visually, let's interrogate the data statistically.

To run an ANOVA, we use the `aov()` function. Some helpful tips about the `aov()` function:

- unlike the `t.test()` function, the initial output for the `aov()` function isn't particularly helpful. We will want to save it to an object and then look at it using the `summary()` function
- like `t.test()`, the syntax for `aov()` is `dependent ~ independent`.

First, let's run an ANOVA on the battery life variable.

```
battery_model <- aov(data = collars, battery_life ~ maker)
summary(battery_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## maker         2  36919   18460   161.2 <2e-16 ***
## Residuals   147  16828     114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For us, our main focus here in the ANOVA summary is on the `maker` row. We can ignore the `Residuals`.

We want to pay attention to the following:

- `F-value`: this is our test statistic. It determines our `p-value` (you don't need to know how)
- `Pr(>F)`: this is our `p-value`, which has roughly the same meaning as with a t-test.
  - We want to know if this value is larger or smaller than our cut-off value. We will stick with 0.05.
  - The notation here, `Pr(>F)`, is saying "What is the probability of us getting an F-value larger than the one we have by change alone?" If that probability is less that 0.05 (or 5%), then we reject the null hypothesis that the means are the same.
- `***`: the asterisks are reiterating what the p-value tells us, indicating just how small the p-value is.

What is our conclusion? Are these means different?

Practice How about for signal life? What is your interpretation of these results?

```
signal_model <- aov(data = collars, signal_distance ~ maker)
summary(signal_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## maker         2 499547  249773   189.1 <2e-16 ***
## Residuals   147 194137    1321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pairwise Comparisons

ANOVAs are incredibly useful to tell you if there is a difference in the means of *any* of the groups. However, they do not tell you *which* means differ from another.

To figure that out, you need to use a class of tests called "post hoc tests," meaning "after the event." Post hoc tests take into account the fact that we are running multiple pairwise comparisons, which unfortunately increases our chance of error. To account for this, we need to adjust our p-values based on the number of pairwise comparisons we run.

The most common post hoc test for these pairwise comparisons is `TukeyHSD()`, but there are others depending on the specifics of your data set. You don't need to worry about understanding the nuances of Tukey's test, but you should know how to run the code and interpret it.

```
TukeyHSD(battery_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = battery_life ~ maker, data = collars)
##
## $maker
##                                                   diff       lwr       upr
## Collarium Inc.-Budget Collars LLC            34.667433 29.591701 39.743164
## Darn Tuff Collars Company-Budget Collars LLC 30.765879 25.771504 35.760254
## Darn Tuff Collars Company-Collarium Inc.     -3.901553 -9.048355  1.245248
##                                                  p adj
## Collarium Inc.-Budget Collars LLC            0.0000000
## Darn Tuff Collars Company-Budget Collars LLC 0.0000000
## Darn Tuff Collars Company-Collarium Inc.     0.1748636
```

What do these different columns represent?

- `diff`: the difference in the means for that pair
- `lwr`: the lower end of the 95% confidence interval
- `upr`: the upper end of the 95% confidence interval
- `p adj`: the adjusted p-value for that pair. We interpret it as normal.

How should we interpret these results?

- Are Collarium and Bugdet significantly different in battery life? **Yes ($p < 0.05$)**
- Are Darn Tuff and Budget significantly different in battery life? **Yes ($p < 0.05$)**
- Are Darn Tuff and Collarium significantly different in battery life? **No ($p > 0.05$)**

Practice In your groups, run the Tukey's HSD test for signal distance and interpret the results.

```r
TukeyHSD(signal_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = signal_distance ~ maker, data = collars)
##
## $maker
##                                                    diff        lwr        upr
## Collarium Inc.-Budget Collars LLC            -107.25643 -124.49632  -90.01655
## Darn Tuff Collars Company-Budget Collars LLC -130.34487 -147.30843 -113.38132
## Darn Tuff Collars Company-Collarium Inc.      -23.08844  -40.56972   -5.60716
##                                                 p adj
## Collarium Inc.-Budget Collars LLC            0.000000
## Darn Tuff Collars Company-Budget Collars LLC 0.000000
## Darn Tuff Collars Company-Collarium Inc.     0.005997
```

How should we interpret these results?

- Are Collarium and Budget significantly different in signal distance? **Yes ($p < 0.05$)**
- Are Darn Tuff and Budget significantly different in signal distance? **Yes ($p < 0.05$)**
- Are Darn Tuff and Collarium significantly different in signal distance? **Yes ($p < 0.05$)**