module_3_2

Keaton Wilson

2/12/2020

Plotting and Means Comparisons

Last class, we examined the collars data set, read a bit about the grammar of graphics and analyzed some code for a ggplot of some of the data we've been working with.

Let's take a stab at making our ggplot from scratch.

In your groups, work together to develop code that utilizes the collars data https://tinyurl.com/sp7b25x and plots:
1. Two histograms one for battery life and one for signal length
2. Color coded for each manufacturer
3. Correct and appropriate labels
4. A vertical line that plots the median of each group (this is trickier - remember to use the google)

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyv
```

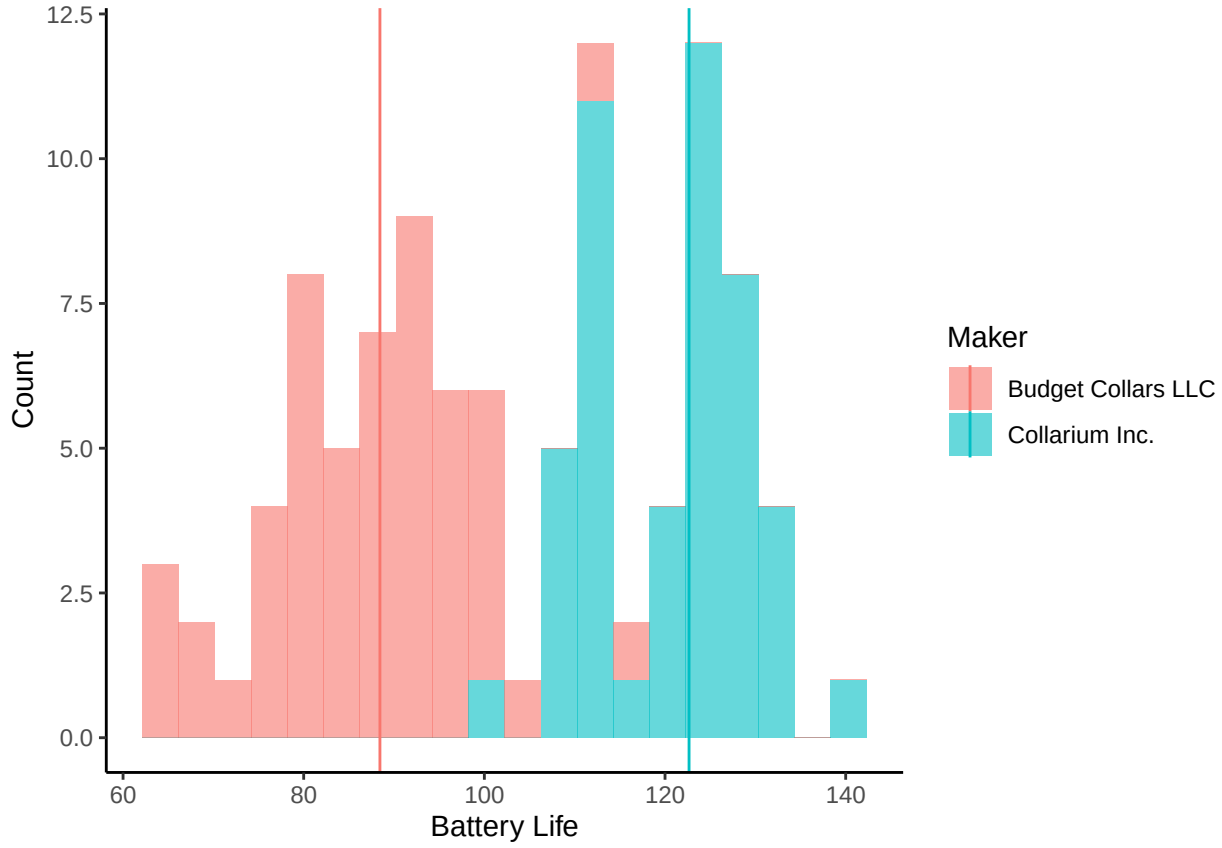
```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_c
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

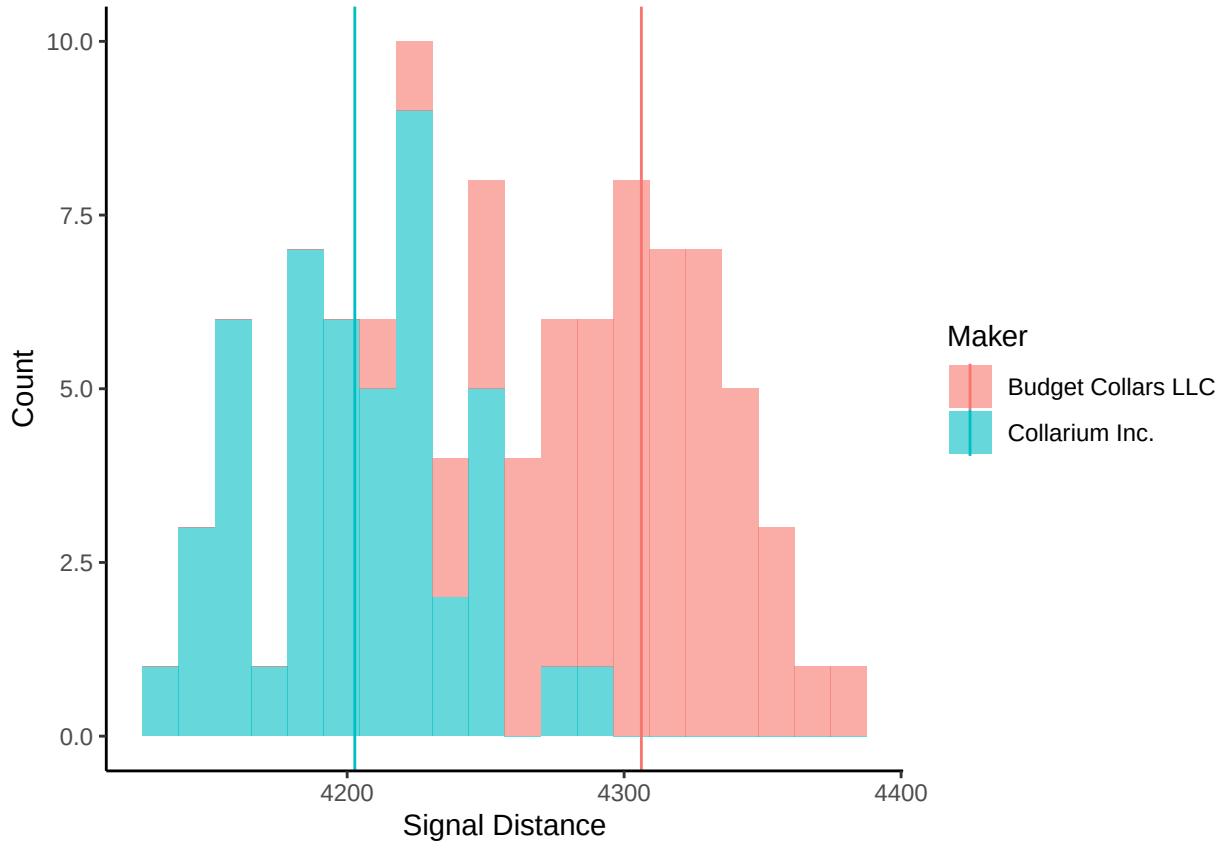
```
collars = read_csv("https://tinyurl.com/sp7b25x")
```

```
## Parsed with column specification:
## cols(
##   collar_id = col_double(),
##   maker = col_character(),
##   battery_life = col_double(),
##   signal_distance = col_double(),
##   fail = col_double()
## )
```

```
collars %>%
  ggplot(aes(x=battery_life, fill = maker)) +
  geom_histogram(alpha = 0.6, bins = 20) +
  geom_vline(data = collars %>%
               group_by(maker) %>%
               summarize(median_battery = median(battery_life)),
             aes(xintercept = median_battery, col = maker)) +
  ylab("Count") +
  xlab("Battery Life") +
  scale_color_discrete(name = "Maker") +
  scale_fill_discrete(name = "Maker") +
  theme_classic()
```




```
collars %>%
  ggplot(aes(x=signal_distance, fill = maker)) +
  geom_histogram(alpha = 0.6, bins = 20) +
  geom_vline(data = collars %>%
               group_by(maker) %>%
               summarize(median_signal = median(signal_distance)),
             aes(xintercept = median_signal, col = maker)) +
  ylab("Count") +
  xlab("Signal Distance") +
  scale_color_discrete(name = "Maker") +
  scale_fill_discrete(name = "Maker") +
  theme_classic()
```



What is another way to plot these data?

Stepping back and thinking about the problem

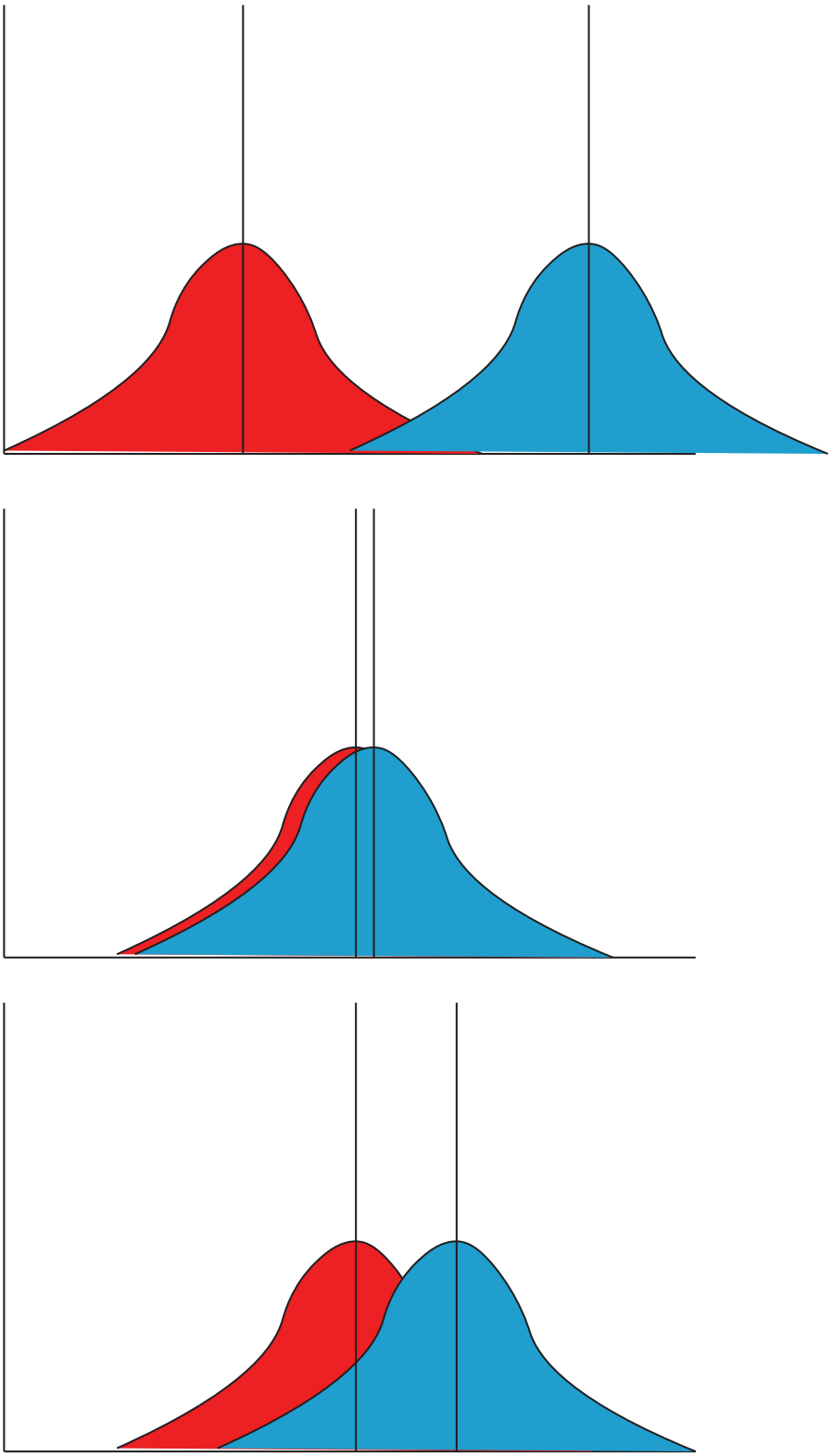
Ok, so stepping back - what is the fundamental question we're asking here?

Pose this to the class.

Fundamentally, we want to know if there a difference between the two manufacturers in these two variables - battery life and signal distance

They look different in the graphs you made, but is this enough? What do I mean?

Draw three different histograms on the board.



How do we tell whether there is an **actual** or **meaningful** difference in the means?

There is a formalized set of statistical tools to answer this question. Some of you may have heard of t-tests

before - which is what we use on this type of data.

Let's briefly talk through the logic here.
1. Our data are a sample of a larger population (think about all of the collars ever produced for both companies).
2. What we're really interested in is the the difference in the means between the two groups. If they're the same, this difference is 0. If they're different, the difference is something larger or smaller than 0.
3. the code is simple, but the output is a bit trickier:

```
t.test(battery_life ~ maker, data = collars)
```

```
##
##  Welch Two Sample t-test
##
## data:  battery_life by maker
## t = -16.733, df = 96.267, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -37.28164 -29.37458
## sample estimates:
## mean in group Budget Collars LLC     mean in group Collarium Inc.
##                         87.23787                         120.56598
```

4. Breaking down the information here:

a. *t*: is the test statistic - here, it's a metric of how different the difference of the two means is from 0. High (or low) values indicate a big difference, small values are closer to 0.

b. *df*: is degrees of freedom - it's a measure of your sample size and some other stuff - not something we're really going to focus on here.

c. *p-value*: a measure of certaintanty of the difference outlined in the *t-statistic* above, it is veryyyyyy small. This means that there is an extremely low probability that obtaining a difference in means of 0 is very, very unlikely. There are a lot of benchmarks in different fields for what this value should be below to consider the difference *significant*, typically < 0.05 is considered significant.
d. *95 percent condience interval*: this is the range that we can expect the test statistic (difference in means) to fall in 95% of the time given the data.