Module 3: Assignment 2

Ellen Bledsoe

2022-11-03

Assignment Details

Purpose

The goal of this assignment is to assess your ability to compare *multiple* means numerically, visually, and statistically and interpret the results.

Task

Write R code which produces the correct answers and correctly interpret the results of visualizations and statistical tests.

Criteria for Success

- Code is within the provided code chunks
- Code is commented with brief descriptions of what the code does
- Code chunks run without errors
- Code produces the correct result
  - Code that produces the correct answer will receive full credit
  - Code attempts with logical direction will receive partial credit
- Written answers address the questions in sufficient detail

Due Date

November 8 at midnight MST

Assignment Questions

In this assignment, we're going to explore data from 4 bays down in Antarctica: Emporer Bay, Hope Bay, Iceberg Bay, and Sulzberger Bay.

We've assessed our fleet and feel like we have enough fishing boats to fish from *two of the four* nearby bays. How do we decide which two bays to fish?

First, we need to assess which how many fish are in each bay. Then we will assess how many leopard seals hang around each bay. Based on these two aspects, we will make a determination for which two bays we can most safely fish and still get enough fish for the station.

Set-Up

1. As always, our first order of business is to load the `tidyverse` and our datasets. For this assignment, we have two different datasets: one for our survey of fish amounts and one for seal observations. Read both of them into the workspace and name them `fish` and `seals`, respectively.

```r
library(tidyverse)
fish <- read_csv("../data/fish.csv")
seals <- read_csv("../data/seals.csv")
```

Fish

The first question we've been asked to tackle is if the four bays have different amounts of fish. Let's find out.

Explore the data set by using your favorite function to look at the data frame.

```r
head(fish)
```

```
## # A tibble: 6 x 5
##   date         net time  bay     fish_kg
##   <date>     <dbl> <chr> <chr>     <dbl>
## 1 2021-03-01     1 AM    Emporer    33.7
## 2 2021-03-01     2 AM    Emporer    25.4
## 3 2021-03-01     3 AM    Emporer    21.0
## 4 2021-03-01     4 AM    Emporer    38.8
## 5 2021-03-01     5 AM    Emporer    15.4
## 6 2021-03-01     1 PM    Emporer    27.0
```

Early last year, we had our fishers head out into each of the four bays everyday for a month to sample fish. They sampled using 5 nets in both the morning and the afternoon. They recorded how many kilograms of fish they caught in each net. The two variables that we will be using to help us make our evaluation are `bay` and `fish_kg`.

Conceptual

2. Write out our two hypotheses for these data.

   Null Hypothesis H_0: no difference in amount of fish per bay

   Alternative Hypothesis H_0: difference in amount of fish per bay

Numeric

3. Let's first find out what the mean (average) and standard deviation (spread) of `fish_kg` is in each bay. Save this data in a new data frame called `fish_summary`.

```r
fish_summary <- fish %>%
  group_by(bay) %>%
  summarise(mean_fish = mean(fish_kg),
            sd_fish = sd(fish_kg))
```

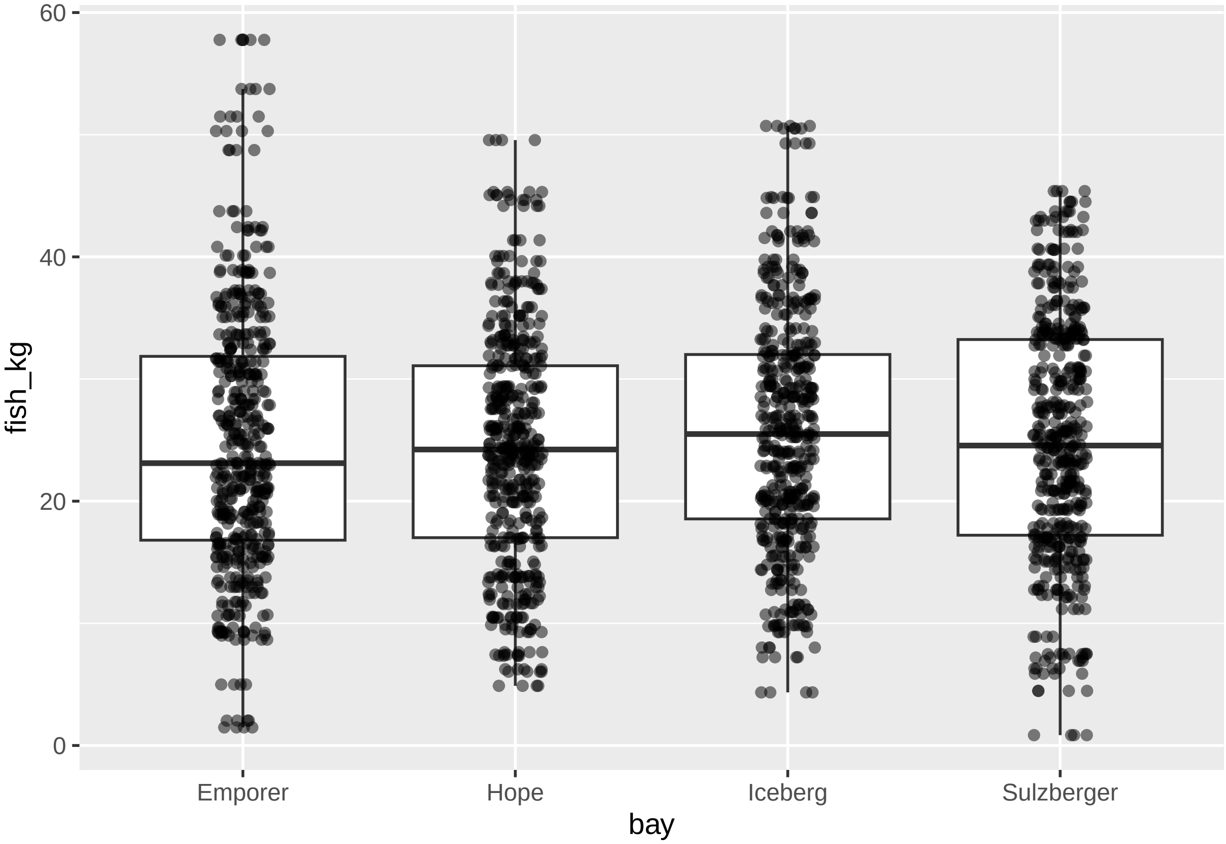
Take a look at the values. Do you think the bays have different amounts of fish based on these values? Let's keep exploring.

Visual

4. Plot the fish data using a box-and-whisker plot. It doesn't need to be fancy (no need to fix axis labels and such) but be sure it has the following: Make sure your plot has the following:

    - data points overlaying the boxes
    - the points should be partially transparent and jittered

```
ggplot(fish, aes(bay, fish_kg)) +
  geom_boxplot() +
  geom_jitter(alpha = 0.5, width = 0.1)
```



Statistics

5. In our data set, which variable is the independent variable? Is it categorical or continuous? Which is the dependent variable and is it categorical or continuous?

    Independent: bay, categorical

    Dependent: fish_kg, continuous

6. Run an ANOVA to determine if there is an overall difference in means in the four bays. Save the model as an object called `fish_model`, then display the results of the model using the `summary()` function. Remember, the syntax for the model is: `dependent ~ independent`

```
fish_model <- aov(data = fish, fish_kg ~ bay)
summary(fish_model)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## bay            3    542   180.6   1.751  0.155
## Residuals   1916 197625   103.1
```

7. Interpret the results of the model and answer the following questions (2 points):

   - What is the value of the test statistic? 1.751
   - What is our p-value? 0.155
   - Is our p-value higher or lower than our cut-off value of 0.05? higher
   - Write 2-3 sentences explaining that this result means. How do we interpret this value? What does it mean for our hypotheses?

   Answers: no significant difference in the amount of fish per bay, fail to reject null

8. Run some code to compute pairwise comparisons.

```
TukeyHSD(fish_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = fish_kg ~ bay, data = fish)
##
## $bay
##                           diff        lwr       upr     p adj
## Hope-Emporer       -0.53796408 -2.2236167 1.1476886 0.8447466
## Iceberg-Emporer     0.94195591 -0.7436968 2.6276086 0.4764075
## Sulzberger-Emporer  0.04106038 -1.6445923 1.7267130 0.9999118
## Iceberg-Hope        1.47991999 -0.2057327 3.1655727 0.1085321
## Sulzberger-Hope     0.57902446 -1.1066282 2.2646771 0.8135811
## Sulzberger-Iceberg -0.90089553 -2.5865482 0.7847571 0.5157262
```

9. Write 2-3 sentences interpreting the pairwise comparisons. Which pairs of bays are significantly different from each other, if any?

   Answer: none are different from each other (all above 0.05)

Seals

Let's do the same for our seal data. The `seals` data frame has data on how many seals were observed each day in each survey area of the bay. The columns we will be focusing on are `bay` and `num_seals`.

Conceptual

10. Write out our 2 hypotheses for the seal data.

    Answer: null: no difference in seals by bay

    alternative: difference in # of seals by bay

Numeric

11. Calculate the mean and standard deviation for the number of seals in each bay. Save this to a new data frame called `seal_summary`.

```
seal_summary <- seals %>%
  group_by(bay) %>%
  summarise(mean_fish = mean(num_seals),
            sd_fish = sd(num_seals))
seal_summary
```

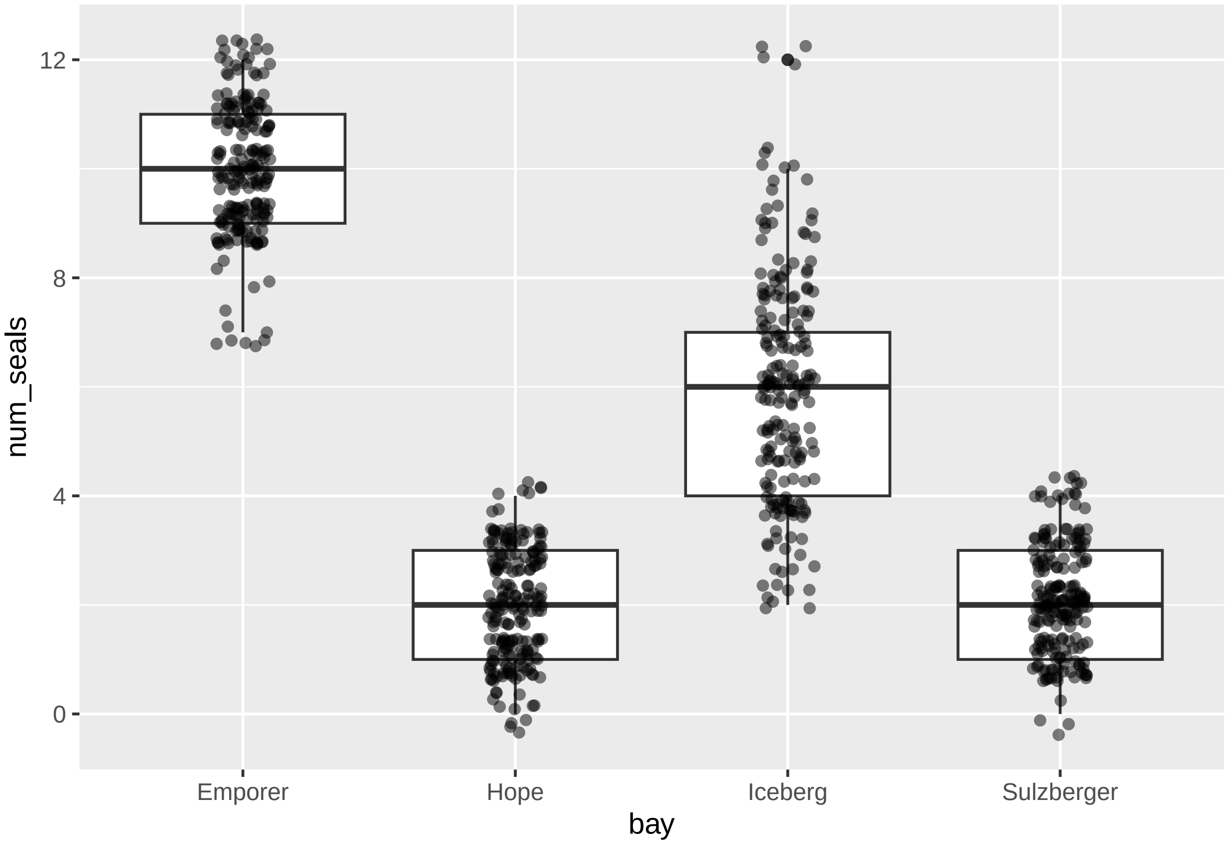
```
## # A tibble: 4 x 3
##   bay        mean_fish sd_fish
##   <chr>          <dbl>   <dbl>
## 1 Emporer         9.96   1.19
## 2 Hope            1.96   1.02
## 3 Iceberg         6.02   2.13
## 4 Sulzberger      2.08   0.958
```

Take a look at the values. Do you think the bays have different numbers of seals based on these values? Let's keep exploring.

Visual

12. Plot the seal using a box-and-whisker plot. It doesn't need to be fancy (no need to fix axis labels and such) but be sure it has the following: Make sure your plot has the following:
    - data points overlaying the boxes
    - the points should be partially transparent and jittered

```
ggplot(seals, aes(bay, num_seals)) +
  geom_boxplot() +
  geom_jitter(alpha = 0.5, width = 0.1)
```



Statistics

13. Run the appropriate statistical model to determine if there is an overall difference in means in the four bays. Save the model as an object called `seal_model`, then display the results of the model using the `summary()` function.

    Remember, the syntax for the model is: dependent ~ independent

```
seal_model <- aov(data = seals, num_seals ~ bay)
summary(seal_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## bay           3   8682    2894    1467 <2e-16 ***
## Residuals   796   1570       2                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

14. Interpret the results of the model and answer the following questions (2 points):

    - What is the value of the test statistic? 1467
    - What is our p-value? <2e-16
    - Is our p-value higher or lower than our cut-off value of 0.05? lower
    - Write 2-3 sentences explaining that this result means. How do we interpret this value? What does it mean for our hypotheses?

    Answers: differences in num of seal per bay, reject null

15. Run some code to compute pairwise comparisons.

```
TukeyHSD(seal_model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = num_seals ~ bay, data = seals)
##
## $bay
##                      diff        lwr        upr     p adj
## Hope-Emporer        -8.00 -8.3615601 -7.6384399 0.0000000
## Iceberg-Emporer     -3.94 -4.3015601 -3.5784399 0.0000000
## Sulzberger-Emporer -7.88 -8.2415601 -7.5184399 0.0000000
## Iceberg-Hope         4.06  3.6984399  4.4215601 0.0000000
## Sulzberger-Hope      0.12 -0.2415601  0.4815601 0.8282111
## Sulzberger-Iceberg -3.94 -4.3015601 -3.5784399 0.0000000
```

16. Write 2-3 sentences interpreting the pairwise comparisons. Which pairs of bays are significantly different from each other, if any?

Answer: all pairs are sig. difference except Sulzberger and Hope

Decision-Making

Think through our goals of doing this analysis.

We want to choose two out of four bays to fish in. We want bays that have enough fish to feed our station but the least number of leopard seals to keep our fishing crews as safe as possible.

17. Based on *all* our results, which two bays do you think we should fish? Explain your rationale (2 points)

Because there is no difference in the number of fish per bay, we will want to make our decision based on seal numbers. We are looking for the lowest seal numbers, which Sulzberger and Hope have. The other bays have significantly higher seal numbers than either of these two, so we would choose them.