Module 3: ANOVA

Mila Pruiett

Your name: Mila :)

Date:

Main Question: We want to know if the population of leopard seals differs between the bays.

1. To answer this question, which variable in the data set is the independent variable? Which is the dependent? Which variable is categorical? Which variable is numeric? (4pts)

```
# independent is the bay

# dependent is the number of seals

# categorical is bay

# numerical is number of seals
```

2. What are our null and alternative hypotheses? (2pts)

```
# Null: The mean number of leapord seals is the same at each of the bays

# Alternative: The mean number of leapord seals is not the same at all of the bays. At least one is different.
```

3. Load tidyverse and read in the data, called “antarctic_seals_many_bays.csv” (2pts)

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

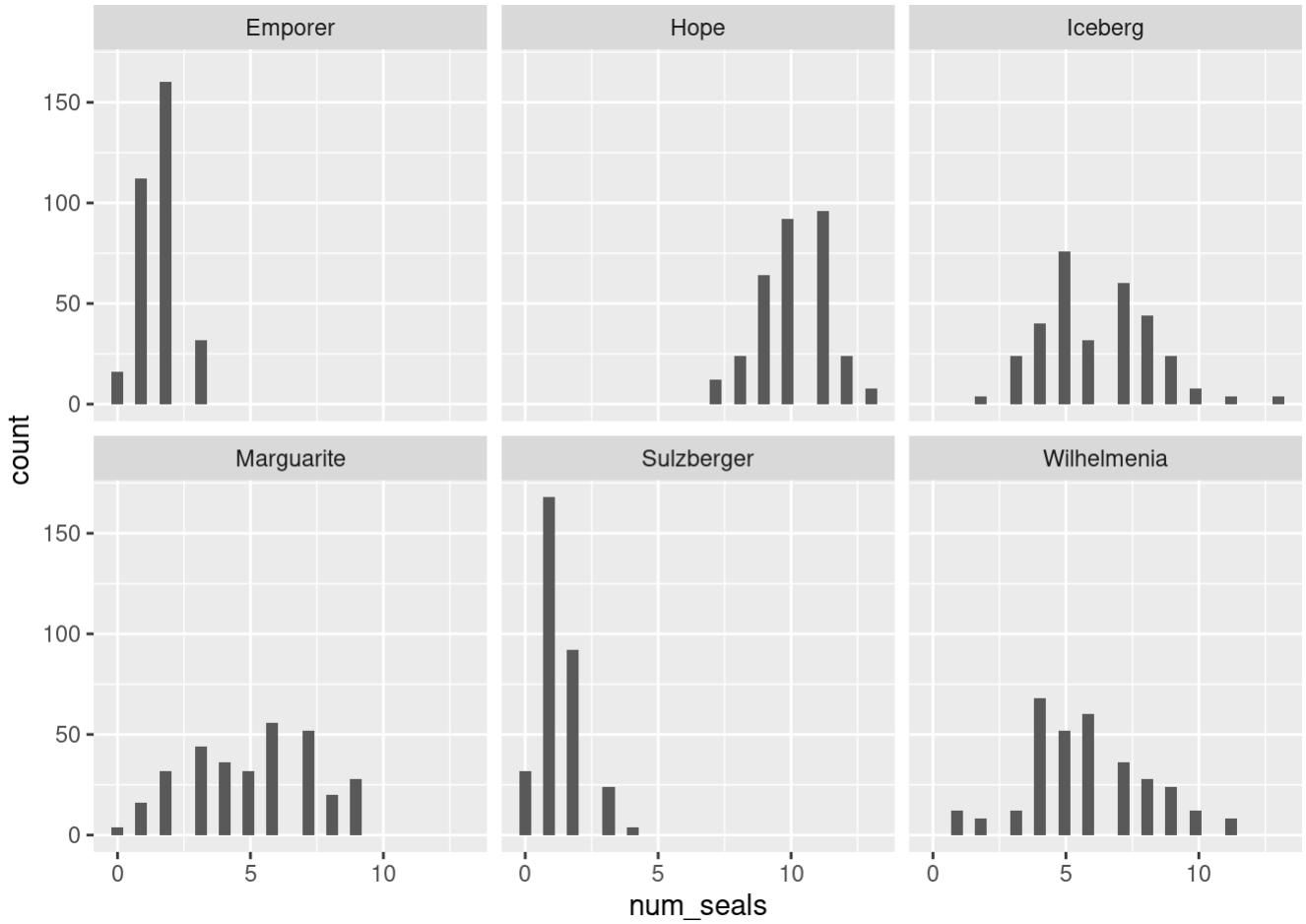
```
sealData <- read_csv("antarctic_seals_many_bays.csv")
```

```
## Rows: 1920 Columns: 5
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (2): time, bay
## dbl  (2): area, num_seals
## date (1): date
##
## 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.
```

4. Create a histogram of the number of seals to see the distribution. Use facet_wrap(~ bay) to create 6 histograms, one for each bay. (2pts)

```
ggplot(sealData, aes(num_seals)) +
  geom_histogram() +
  facet_wrap(~bay)
```

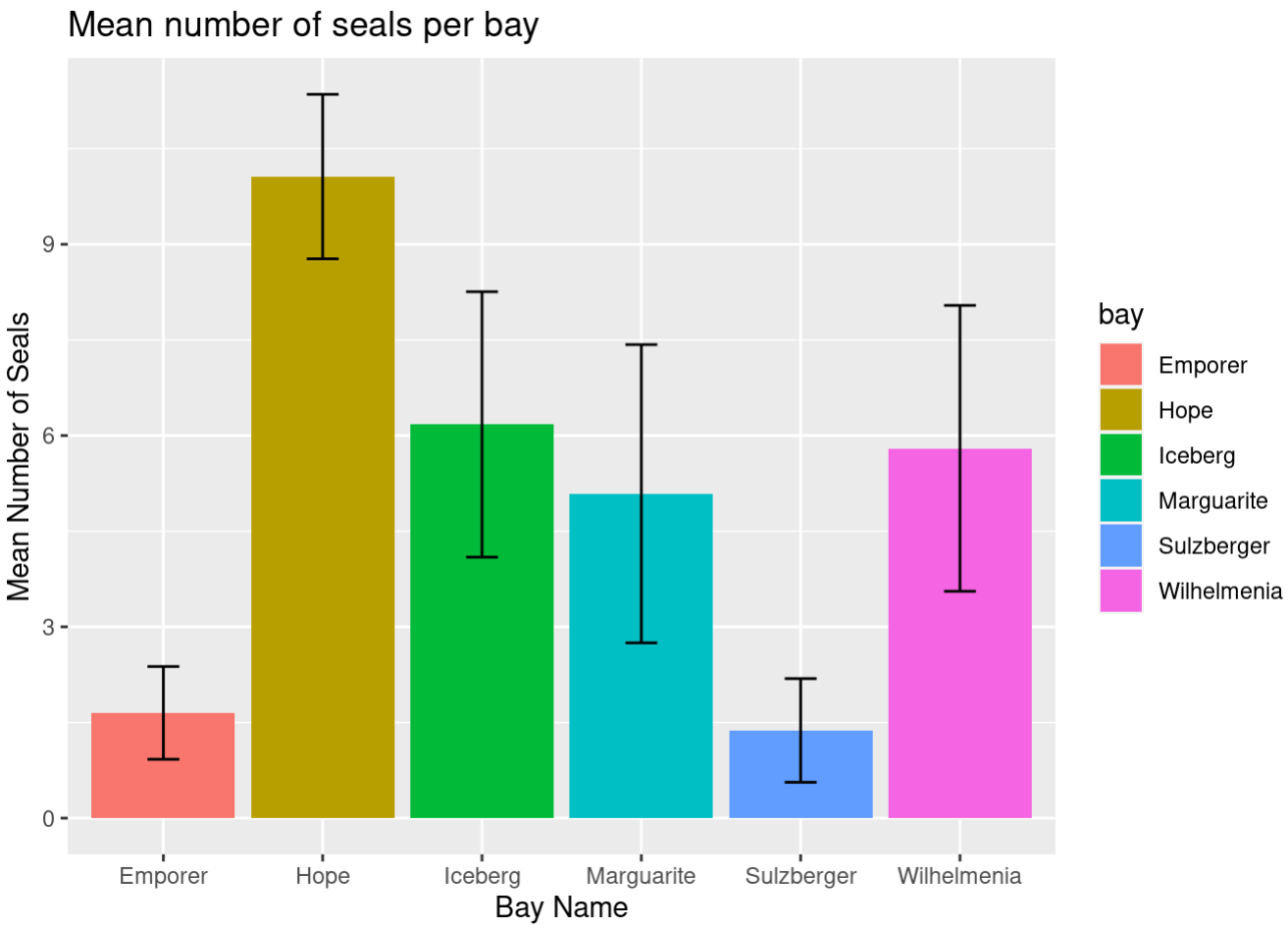
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



5. Now let's calculate some summary statistics of the number of leopard seals per bay (mean and standard deviation). Which bay has the most leopard seals? Which has the least? (4pts)

```
sealSummary <- sealData %>% group_by(bay) %>% summarize(meanSea
ls = mean(num_seals), sdSeals = sd(num_seals))
# bay with lowest mean number of seals: Sulzberger
# bay with the greatest mean number of seals: Hope

ggplot(data = sealSummary, aes(bay, meanSeals, fill = bay)) +
  geom_bar(stat = "identity") +
    geom_errorbar( aes(ymin = meanSeals-sdSeals, ymax = meanSea
ls + sdSeals),
                width = 0.2)  +
  xlab("Bay Name") +
  ylab("Mean Number of Seals") +
  ggtitle("Mean number of seals per bay")
```



6. Run an ANOVA to test if the mean number of seals varies between the bays. (1pts)

```
sealModel <- aov(data = sealData, num_seals ~ bay)
summary(sealModel)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## bay            5  16645    3329    1129 <2e-16 ***
## Residuals   1914   5641       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Interpret the ANOVA table you just created: What is the p-value? What is the F-statistic? What is our alpha level (cutoff value for p)? Do you accept or reject the null hypothesis? What does that mean? (5pts)

```
# the p-value is smaller than 2 x 10 ^ - 16
# the f-stat is 1129
# I reject the null hypothesis because the p-value of 2e-16 is smaller than my alpha of 0.05
# this means that at least one of the mean number of leopard seals is different in one of the bays
```