Code-a-long 3.2 Key - Statistical inference and ANOVAs

Mila Pruiett

Student learning outcomes

- Students will understand basic concepts underlying inference
- Students will be able to formulate statistical hypotheses using data
- Students will be able to perform and interpret the results of an ANOVA

Wrapping up t-tests

Last week we used t-tests to think about which bay we should go fishing to minimize our impact on leopard seals.

1. Are the bays equal in their fish populations?

```r
library(tidyverse)
```

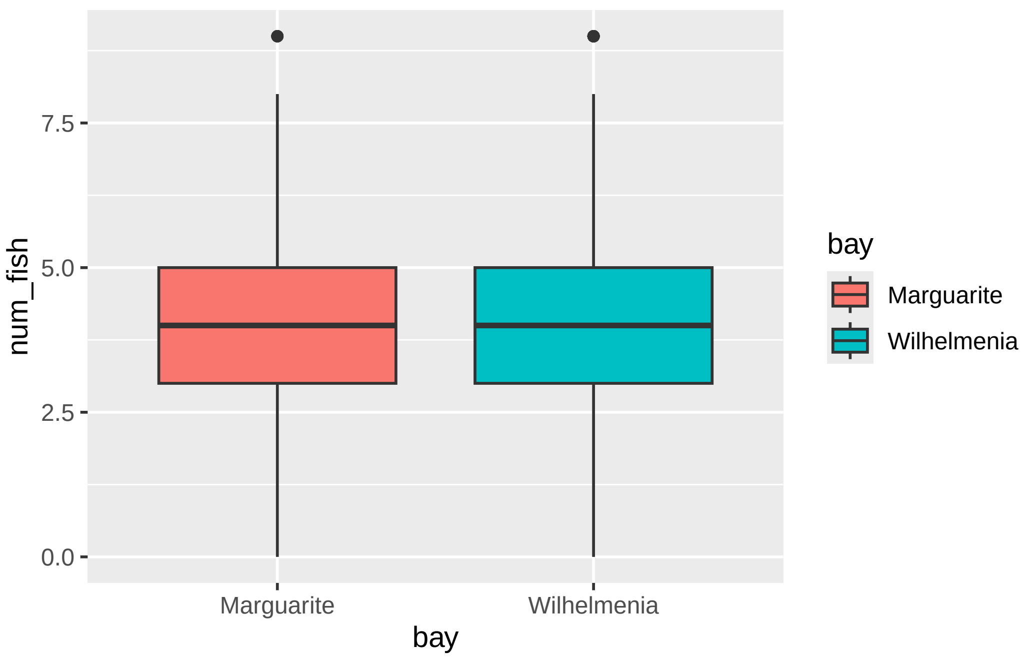
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ 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 becom
```

```r
fish <- read_csv("arctic-fish.csv")
```

```
Rows: 640 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): time, bay
dbl  (2): net, num_fish
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.
```

```r
ggplot(fish, aes(bay, num_fish, fill = bay)) +
  geom_boxplot()
```



```r
t.test(data = fish, num_fish ~ bay)
```

```

    Welch Two Sample t-test

data:  num_fish by bay
t = -1.7366, df = 630.63, p-value = 0.08295
alternative hypothesis: true difference in means between group Marguarite and group Wilhelmen
95 percent confidence interval:
 -0.54602183  0.03352183
sample estimates:
 mean in group Marguarite mean in group Wilhelmenia 
                  3.90625                   4.16250 
```

Response:

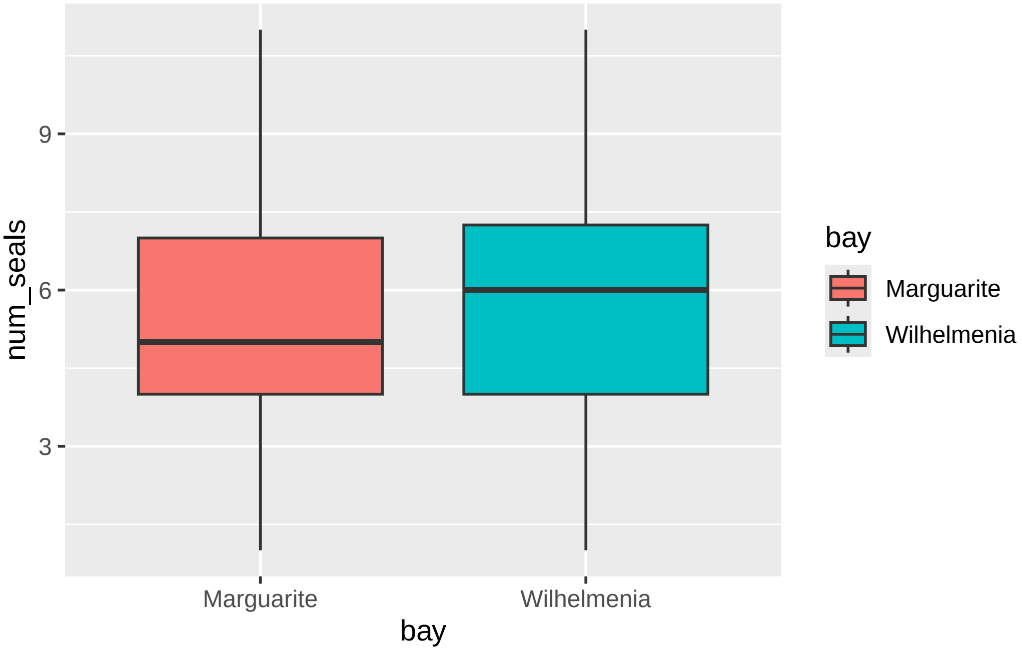
2. Are the bays equal in their leopard seal populations?

```
seals <- read_csv("arctic-seals.csv")
```

```
Rows: 640 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.
```

```
ggplot(seals, aes(bay, num_seals, fill = bay)) +
  geom_boxplot()
```



```
t.test(data = seals, num_seals ~ bay)
```

```

	Welch Two Sample t-test

data:  num_seals by bay
t = -4.2182, df = 638, p-value = 2.82e-05
alternative hypothesis: true difference in means between group Marguarite and group Wilhelme
95 percent confidence interval:
 -1.0258729 -0.3741271
sample estimates:
 mean in group Marguarite mean in group Wilhelmenia 
                     5.25                      5.95 
```

Response:

3. What should we do?

```
# How to report a t test: We can reject null and conclude x bc (DF< alpha, p)
```

Figure 1: Cute Seal

ANOVA: ANalysis Of VAriance

When can I use an ANOVA? Why would I?

- Independent variable is categorical and the response is numerical
- Goal: to compare means among groups

Assumptions of ANOVA

- Data are “normally distributed” => look at the histogram
- Data are “equally varied” => standard deviations reasonably similar
- Samples are independent of one another

The null and alternative hypotheses

H_0 (null hypothesis) - The means of the populations we sampled from **are all equal:** $\mu 1 = \mu 2 = ... = \mu i$

H_a (alternative hypothesis) - The means of the populations we sampled from **are not all equal**

Let's jump in with an example

We have figured out the best option for minimizing our impact on leopard seals while keeping ourselves fed between two bays: Wilhelmina and Marguerite. But there are more bays! And ideally we would use two or more bays to spread out our fishing efforts among multiple humped rock cod populations.

Our team has collected similar data, as we had for Wilhelmina and Marguerite, on four more bays: Emperor, Hope, Sulzberger, and Iceberg.

We are going to examine the fish populations in class, and you will work with the leapard seals for your homework.

1. What is it that we want to know about these six bays? Which variable is the independent variable? Which is the dependent?

2. What are our null and alternative hypotheses?

```
# H0:
```

3. Load tidyverse and read in the data

```
library("tidyverse")
fishManyBays <- read_csv("antarctic_fish_many_bays.csv")
```

```
Rows: 1920 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): time, bay
dbl  (2): net, num_fish
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.
```

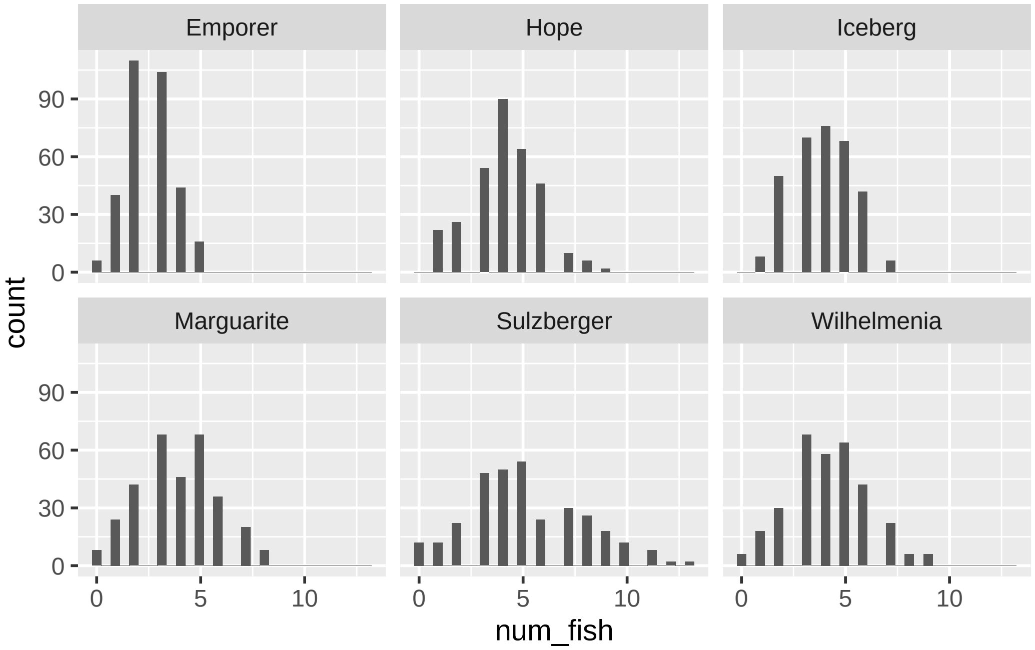
It's always a good idea to visualize your data first. This gives you some perspective on the distribution of the data. What type of data viz is best for viewing the distribution of one variable?

4.

```
#Histogram

ggplot(data = fishManyBays, aes(x = num_fish)) +
  geom_histogram()  +
  facet_wrap(~ bay)
```

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



5. Now let's calculate some summary statistics. What do you notice?

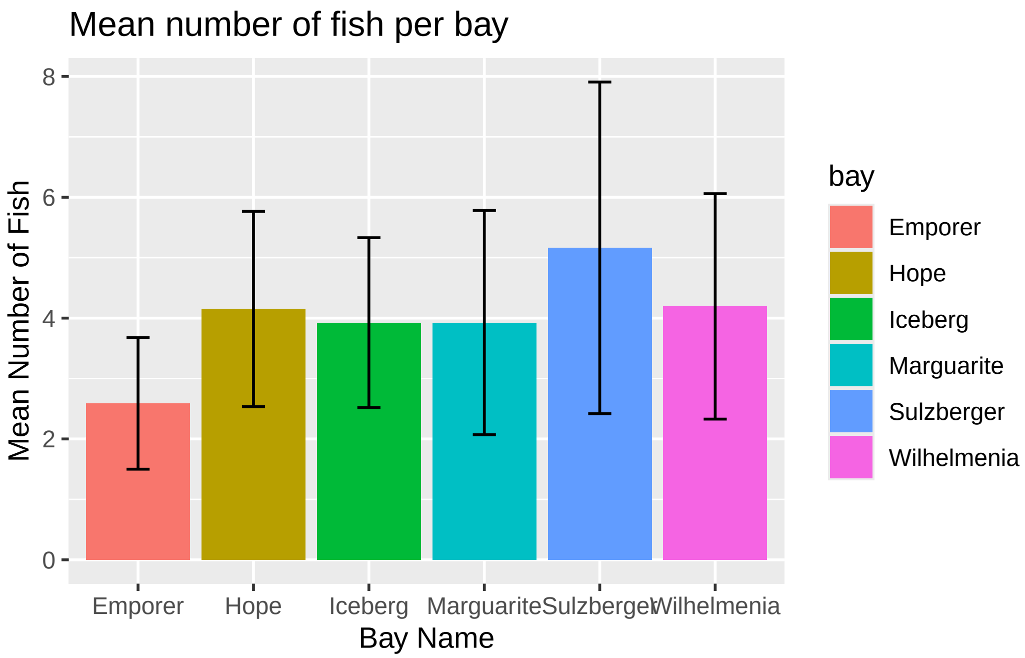
```
fishSummary <- fishManyBays %>%
  group_by(bay) %>%
  summarize(meanFish = mean(num_fish),
            standDevFish = sd(num_fish),
            sampleSize = n())

fishSummary
```

```
# A tibble: 6 x 4
  bay         meanFish standDevFish sampleSize
  <chr>          <dbl>        <dbl>      <int>
1 Emporer         2.59         1.09        320
2 Hope            4.15         1.62        320
3 Iceberg         3.92         1.41        320
4 Marguarite      3.92         1.86        320
5 Sulzberger      5.16         2.74        320
6 Wilhelmenia     4.19         1.87        320
```

6. Let's create a bar graph to compare the summary stats between the groups. Does it seem like the groups are different?

```
ggplot(data = fishSummary, aes(bay, meanFish, fill = bay)) +
  geom_bar(stat = "identity") +
    geom_errorbar( aes(ymin = meanFish-standDevFish, ymax = meanFish + standDevFish),
                   data = fishSummary, width = 0.2)  +
  labs(x="Bay Name", y="Mean Number of Fish", title="Mean number of fish per bay")
```



7. Finally, let's code for the ANOVA.

```
fishModel <- aov(data = fishManyBays, num_fish ~ bay)

summary(fishModel)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
bay            5   1094  218.71   64.88 <2e-16 ***
Residuals   1914   6452    3.37
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. How do we interpret this ANOVA?

```
# Reject the null hypothesis that all bays have the same mean number of fish.

# Based upon the p-value (<2e-16), at least one bay has a significantly different mean numbe
```

9. What is our recommendation for fishing based only off of this information?

```
# We need additional information.
```

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 do that, you need to use a class of tests called Post Hoc Tests. Post hoc tests take into account the problem of running multiple pairwise comparisons, which is the increasing chance of error rates. The most common is Tukey's HSD, but there are others depending on the specifics of your data set. You don't need to worry about understanding Tukey's test, but here I am going to show you how it works and an overview of the interpretation of it.

```
TukeyHSD(fishModel)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = num_fish ~ bay, data = fishManyBays)

$bay
                                diff        lwr       upr     p adj
Hope-Emporer             1.562500e+00  1.1484366 1.9765634 0.0000000
Iceberg-Emporer          1.337500e+00  0.9234366 1.7515634 0.0000000
Marguarite-Emporer       1.337500e+00  0.9234366 1.7515634 0.0000000
Sulzberger-Emporer       2.575000e+00  2.1609366 2.9890634 0.0000000
Wilhelmenia-Emporer      1.606250e+00  1.1921866 2.0203134 0.0000000
Iceberg-Hope            -2.250000e-01 -0.6390634 0.1890634 0.6316815
Marguarite-Hope         -2.250000e-01 -0.6390634 0.1890634 0.6316815
Sulzberger-Hope          1.012500e+00  0.5984366 1.4265634 0.0000000
Wilhelmenia-Hope         4.375000e-02 -0.3703134 0.4578134 0.9996682
```

```
Marguarite-Iceberg      4.440892e-16 -0.4140634  0.4140634 1.0000000
Sulzberger-Iceberg      1.237500e+00  0.8234366  1.6515634 0.0000000
Wilhelmenia-Iceberg     2.687500e-01 -0.1453134  0.6828134 0.4328039
Sulzberger-Marguarite   1.237500e+00  0.8234366  1.6515634 0.0000000
Wilhelmenia-Marguarite  2.687500e-01 -0.1453134  0.6828134 0.4328039
Wilhelmenia-Sulzberger -9.687500e-01 -1.3828134 -0.5546866 0.0000000
```

Want to nerd out about ANOVAs? I recommend Bio statistical Design and Analysis Using R: https://primo.lclark.edu/permalink/01ALLIANCE_LCC/pajj6s/alma99900585075901844 (p254) A Primer of Ecological Statistics: https://primo.lclark.edu/permalink/01ALLIANCE_LCC/pajj6s/alma991413743401018441