Understanding Sick Fish

MP & ED

2022-10-10

In class practice to understand why the fish are sick

We know that there are tanks whose temperature are below the critical threshold for the immune systems of the fish species we are farming. However, there could be other factors contributing to the numbers of sick fish. After our class brainstormed more factors, the ichthyologists (fish scientists) measured: oxygen concentration and ammonia concentration (a proxy for waste buildup). We are going to look at these factors as well, to ensure we can address all of the factors affecting the fish health.

```
# load the tidyverse
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# read in the data, sick-fish.csv
sick <- read_csv("sick-fish.csv")
```

```
## Rows: 1000 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): species
## dbl (11): tank_id, avg_daily_temp, num_fish, day_length, tank_volume, size_d...
## lgl  (1): below
##
## 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.
```

```
# look at the data
glimpse(sick)
```

```
## Rows: 1,000
## Columns: 13
## $ tank_id          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16~
## $ species          <chr> "tilapia", "tilapia", "tilapia", "tilapia", "tilapia"~
## $ avg_daily_temp   <dbl> 22.95922, 23.98088, 23.97097, 24.26474, 24.29623, 23.~
## $ num_fish         <dbl> 95, 96, 101, 98, 93, 101, 98, 109, 97, 102, 99, 99, 9~
## $ day_length       <dbl> 9, 11, 11, 10, 10, 11, 12, 10, 10, 10, 9, 11, 11, 10,~
## $ tank_volume      <dbl> 399.6975, 399.8071, 398.8427, 399.8410, 399.7561, 398~
## $ size_day_30      <dbl> 2784.895, 2781.003, 2785.807, 2785.253, 2786.946, 278~
```

```
## $ ammonia          <dbl> 0.10561057, 0.09073854, 0.10867733, 0.09421766, 0.093~
## $ density          <dbl> 0.2376798, 0.2401158, 0.2532327, 0.2450974, 0.2326418~
## $ avg_daily_temp_F <dbl> 73.32660, 75.16558, 75.14774, 75.67654, 75.73322, 75.~
## $ below            <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE~
## $ num_sick         <dbl> 63, 18, 12, 0, 5, 11, 11, 0, 55, 23, 7, 65, 61, 62, 6~
## $ oxygen           <dbl> 9.480023, 9.288952, 9.467007, 9.322897, 9.327849, 9.4~
```

Our ichthyologist friends told us that density often contributes to the spread of any disease present in a system. We want to look at how density relates to the number of sick fish. Because we are in Antarctica, and obtaining supplies is quite difficult, not all of our tanks are from the same manufacturer and shipment. We have tanks of many different sizes. We know the size of each tank and the number of fish, so we can calculate the density. (Density = number / volume).

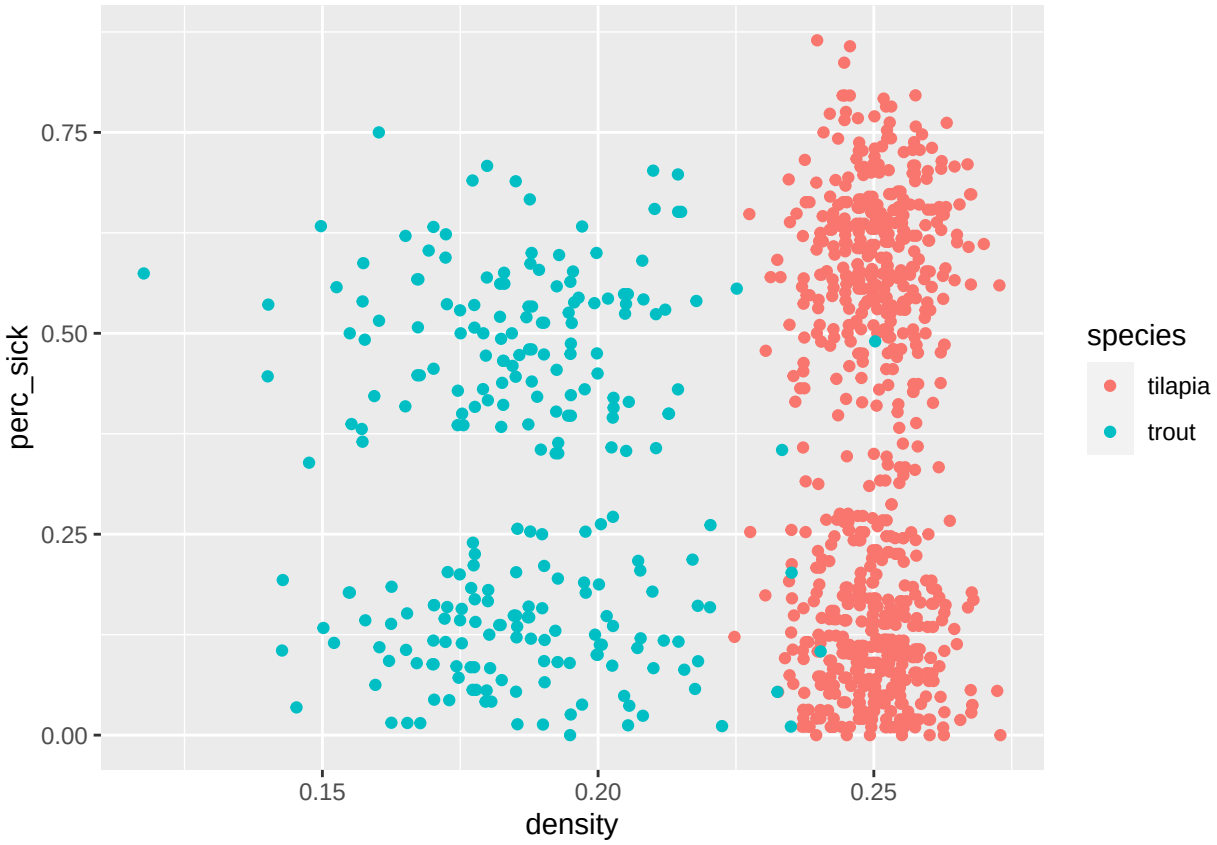
1.Create a variable in the data set for the density of fish per tank. Create a variable in the data set for the percentage of sick fish per tank.

```
sick <- sick %>% mutate(density = num_fish / tank_volume)

sick <- sick %>% mutate(perc_sick = num_sick / num_fish)
```

2. Create a scatter plot to examine the relationship between density and the percentage of sick fish. In comments, explain why we are looking at the relationship between the density and the number of sick fish in a tank instead of the total number of fish in a given tank and the percentage of sick fish.

```
ggplot(data = sick, mapping = aes(density, perc_sick, color = species)) +
  geom_point()
```




```
# percentage sick is standardized across all the tank volumes and fish populations, so it puts everythi
```

In your group, discuss which of the following variables you'd like to examine in more detail.

- Temperature
- Oxygen concentration
- Ammonia (NH3) concentration

PAUSE. CLASS DISCUSSION OF CHOICES.

3. What is your variable and species?

```
# comment it out
# tilapia and temperature
```

Your task is now to learn all about this variable and how it may contribute to the problem. Each person in the group will turn in this .Rmd file as a homework assignment, so be sure to fill out your copy completely. For every visualization, be sure to label the axes clearly (with units) and provide a title. Feel free to customize the appearance as you like.

4. Create a dataframe with only your fish species.

```
tilapiaOnly <- sick %>% filter(species == "tilapia")
```

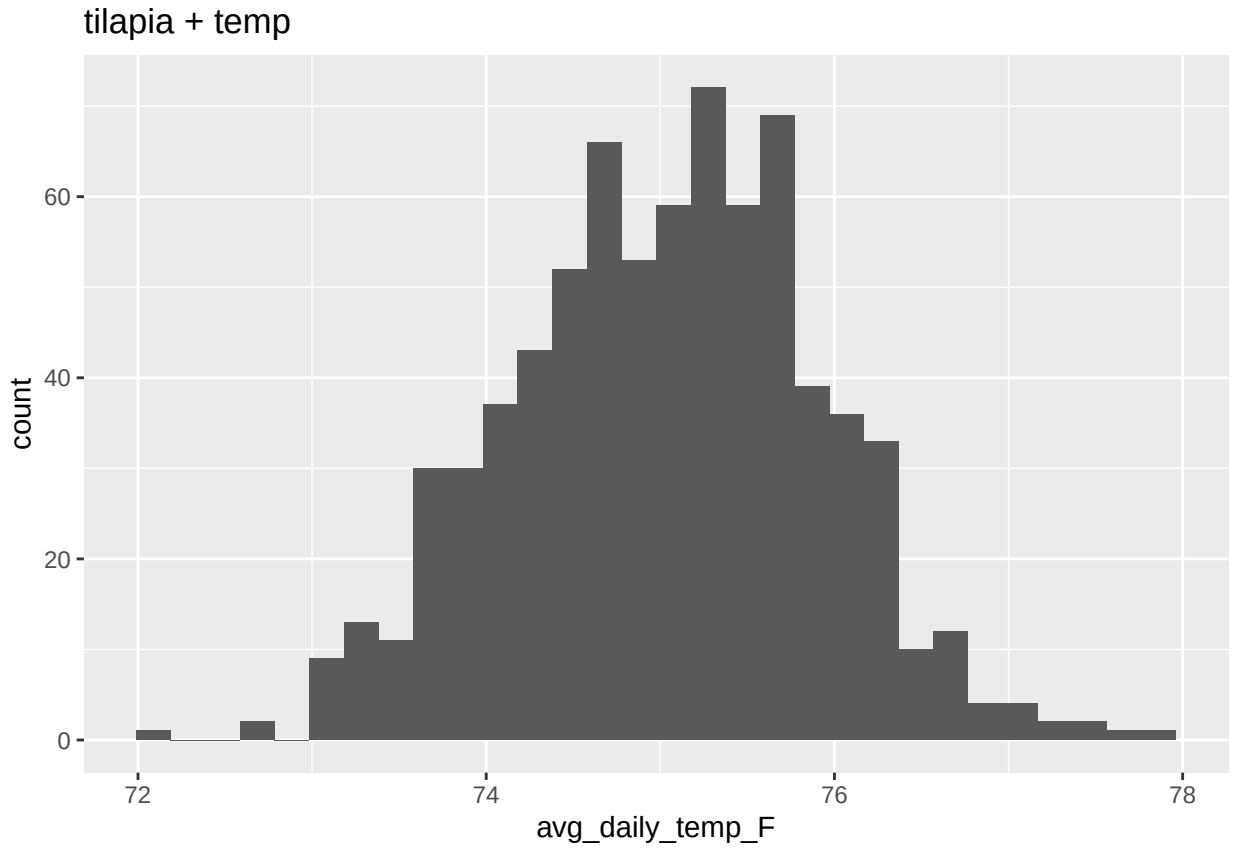
5. Create a histogram of your variable. What is (approximately) the most common value of your variable? One person from each group: put a copy of this histogram to this shared google slide deck to show the class on Wednesday.

11:30 Class Section: https://docs.google.com/presentation/d/1AYbRlU9NB36EfF4R8C5_nTr0dQyTWpcAPTeSiW7XKhY/edit?usp=sharing

1:50 Class Section: https://docs.google.com/presentation/d/1CnL45KGLifZypKY9J5uAC8qoXEjEToFmxZSdiy6oKqQ/edit?usp=sharing

```
ggplot(tilapiaOnly, aes(x = avg_daily_temp_F)) +
  geom_histogram() +
  labs(title = "tilapia + temp")
```

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




```
# most common is around 59 but it could be different depending on their bin number. The correct answer
```

6. What is the mean and standard deviation of your variable? Add to google slide deck.

```
tilapiaOnly %>% summarize(mean(avg_daily_temp_F), sd(avg_daily_temp_F))
```

```
## # A tibble: 1 x 2
##   `mean(avg_daily_temp_F)` `sd(avg_daily_temp_F)`
##                      <dbl>                  <dbl>
## 1                     75.0                  0.876
```

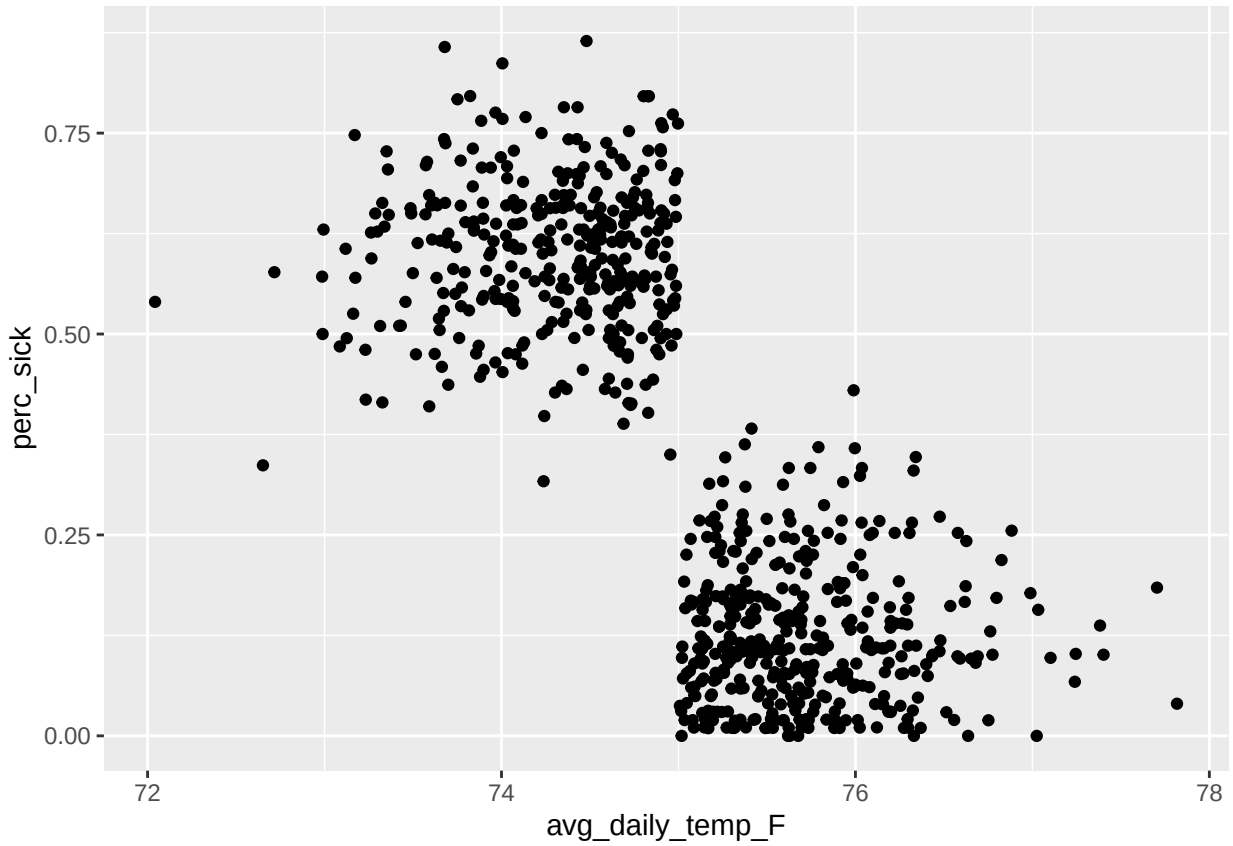
7. What is the motivating question? That is, what can your variable tell us about the sick fish? What data visualization will you use to answer this question?

```
# we want to know if temperature is related with the percentage of sick fish. I will use a scatterplot.
```

PAUSE. SHARE YOUR QUESTION AND PROPOSED ANALYSIS WITH THE GROUP THAT IS STUDYING THE SAME VARIABLE ON THE OTHER FISH SPECIES.

8. Create a visualization to analyze the relationship between fish sickness and your variable. Add this to the google slide deck.

```
ggplot(data = tilapiaOnly, mapping = aes(x = avg_daily_temp_F, y= perc_sick)) +
  geom_point()
```



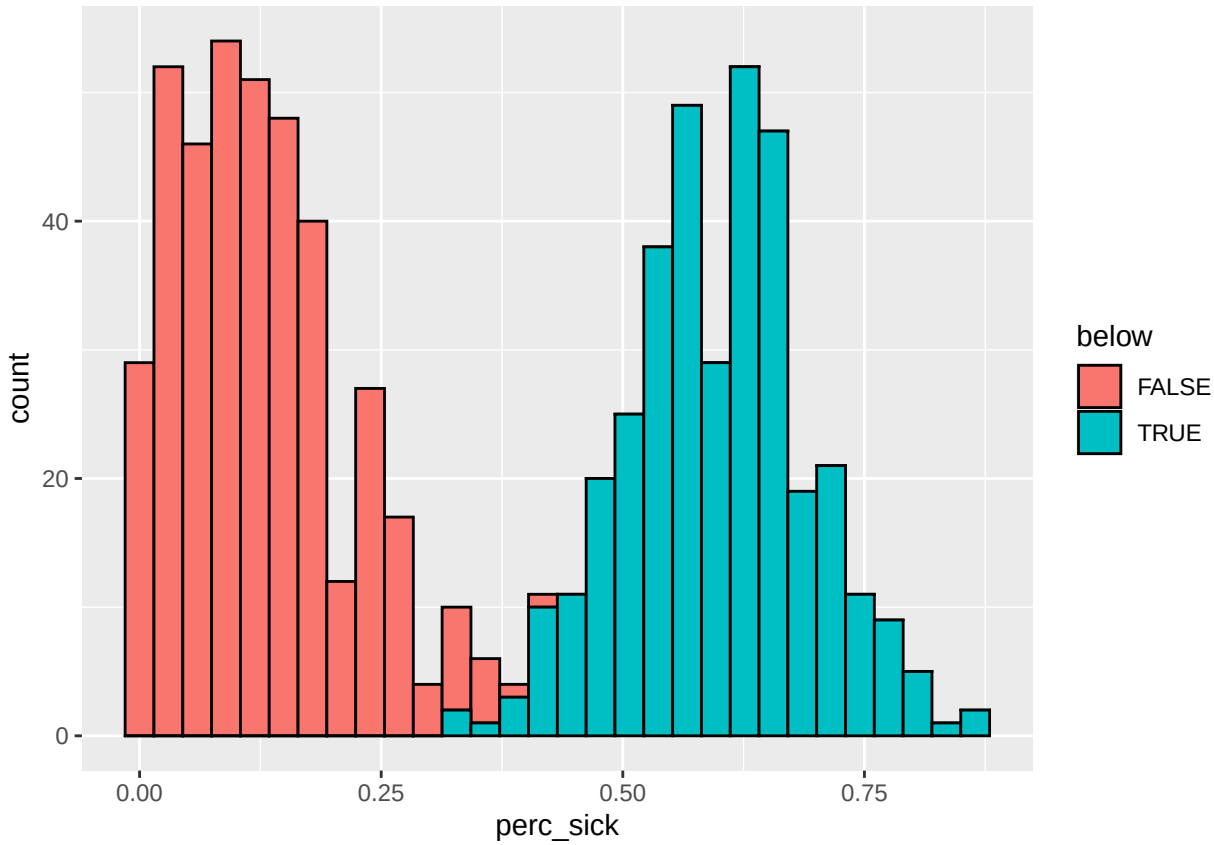
TEMPERATURE GROUPS ONLY

9. In the previous questions, you have examined fish sickness and temperature as a continuous variable. However, our data set also includes temperature as a categorical variable- is a given tank below the critical threshold for fish immune systems. If the column titled "below" is TRUE, then that tank is below the critical threshold. If the column titled "below" is FALSE, then that tank is above the critical threshold. Create a histogram that examines sick fish and the tank temperature as this categorical variable.

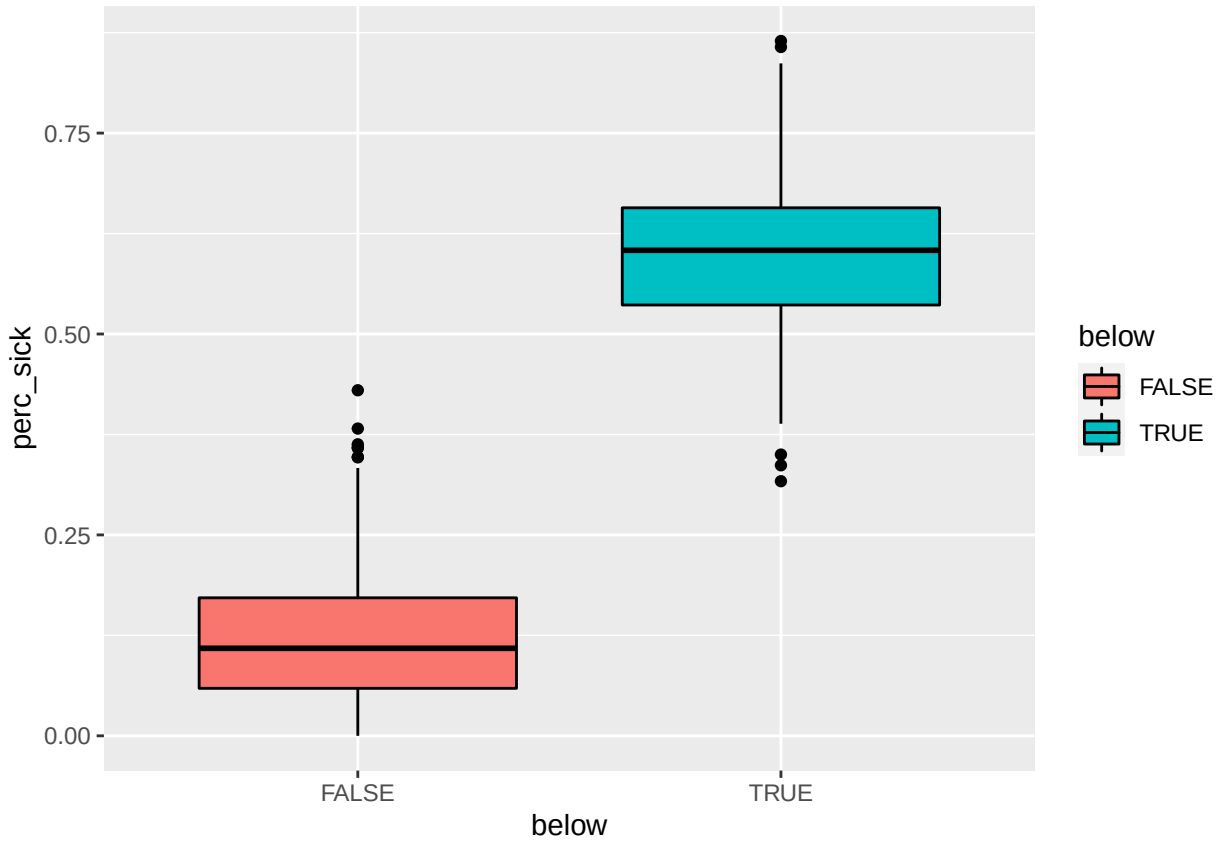
```
# okay so many ways to do this
# do not grade them down if they don't have anything here


# with a histogram
ggplot(data = tilapiaOnly, mapping = aes(x = perc_sick, fill = below)) +
  geom_histogram(color = "black")
```

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



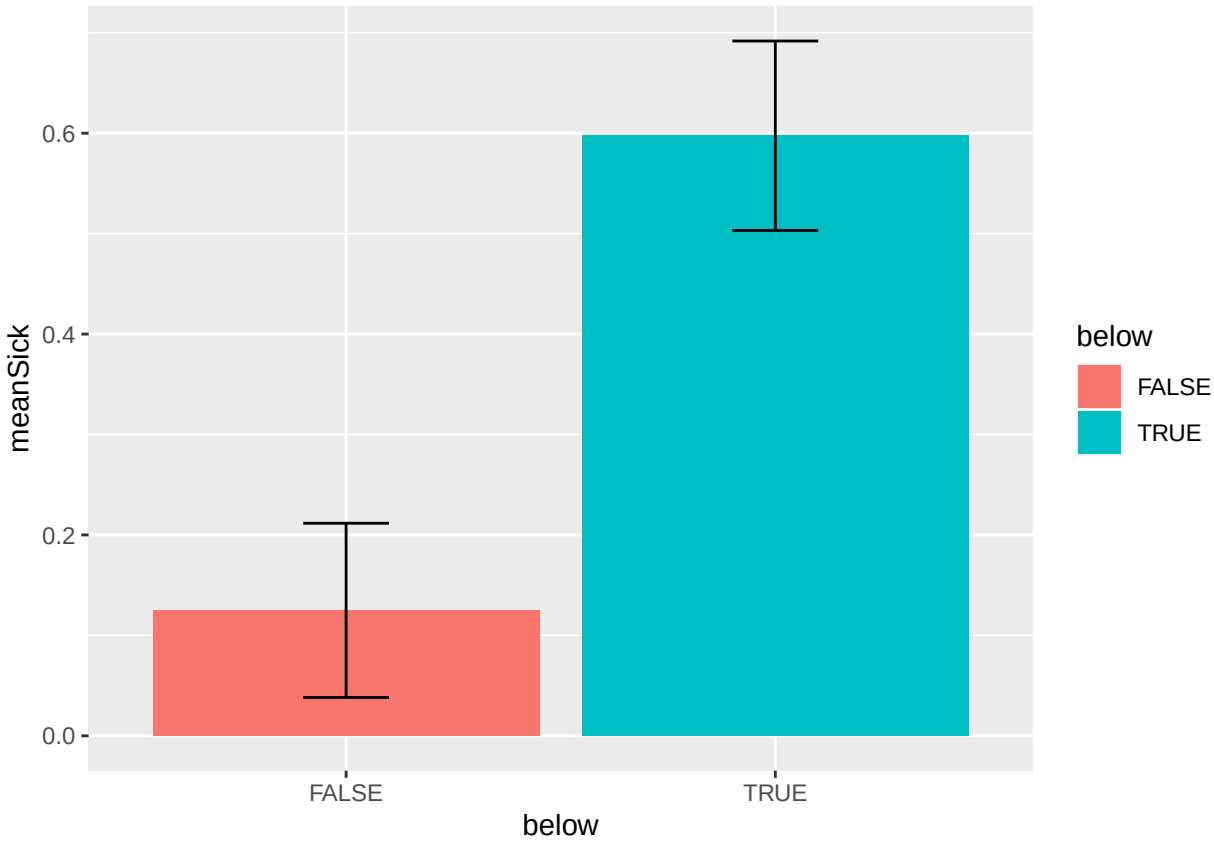

```
# with a box plot
ggplot(data = tilapiaOnly, mapping = aes(x = below, y = perc_sick, fill = below)) +
  geom_boxplot(color = "black")
```




```
# with a bar plot
summaryData <- tilapiaOnly %>% group_by(below) %>% summarize(meanSick = mean(perc_sick), sdSick = sd(pe
summaryData
```

```
## # A tibble: 2 x 3
##   below meanSick sdSick
##   <lgl>    <dbl>  <dbl>
## 1 FALSE    0.125 0.0867
## 2 TRUE     0.597 0.0943
```

```
ggplot(summaryData, aes(x = below, y = meanSick, fill = below)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = meanSick - sdSick, ymax = meanSick + sdSick, width = 0.2))
```


meanSick
0.6
0.4
0.2
0.0
FALSE
TRUE
below
below
FALSE
TRUE