Exam 2

## Exam 2 - Good Fish Gone Bad

### Question 1

You’re interested in exploring a data set on pH levels (a measure of acidity) in some of the aquaculture tanks we discussed during this module. You have data from ~750 tanks that includes, fish density, temperature and disease rate. You’re interested in examining the distribution of pH values across all tanks - what type of plot do you build to visualize these data?  
a. a scatterplot  
b. a histogram  
c. a box and whisker plot  
d. a pie chart  
e. a 3-dimensional scatterplot that includes all variables

### Question 2

You want to create a scatterplot of pH and disease rate from the dataframe described in question 1 (750 tanks that includes data on fish density, temperature, disease rate and pH). Select the code chunk below that best produces a scatter plot from these data.  
a. hist(x=tilapia\_data$pH, y=tilapia\_data$disease\_rate, type = "scatter")  
b. scatter.plot(x=tilapia\_data$pH, y=tilapia\_data$disease\_rate, pch = 16)  
c. plot(x = tilapia\_data$pH, y = tilapia\_data$disease\_rate, pch = 19, main = "Disease Rate as a Function of pH" , xlab = "pH", ylab = "Disease Rate")  
d. plot(x = disease\_rate, y = pH, data = tilapia\_data, main = "Disease Rate as a Function of pH" , xlab = "pH", ylab = "Disease Rate")

### Question 3

Your team has been tasked with creating a function that automates the calculation of the standard error of a set of data (something that isn’t built into Base R). A bit of context, standard error is another measure of spread, compared to standard deviation (which we’ve talked about a bit). Standard deviation measures the amount of variability or dispersion of a data set, while standard error of the mean measures how far the mean of the data is likely to be from the true population mean. This distinction isn’t important for this question, but it is something that we’ll talk about later in the course. One of your team members submitted some code for you to review that is a first draft of this function. Pick a choice below that highlights problems in the function.

standard\_error = function(data = NULL){  
 se = sd(data)/sqrt(length(data))  
}

1. The data argument should be set to a real number (instead of NULL), to avoid breaking the function.
2. The functions sd and sqrt aren’t defined explicitly.
3. Naming an object se within a function is confusing, as it overwrites an existing named standard error function in R.
4. The function doesn’t return anything.

### Question 4

Pick one of the topics/skills we worked on in this module (from the list below), and write a brief summary of what it is, how we used it and why you think it may a foundational tool for more complicated tasks and questions.  
a. Iteration (loops)  
b. Simulations  
c. Plotting and exploratory data analysis  
d. Functions

### Question 5

Explain in your own words (1-5 sentences) what the code block below is doing. Feel free to reference line numbers if that’s helpful.

range\_diff = function(vector\_1, vector\_2){  
 range\_1 = max(vector\_1) - min(vector\_1)  
 range\_2 = max(vector\_2) - min(vector\_2)  
 return(range\_1-range\_2)  
}  
  
diffs = c()  
for(i in 1:ncol(df\_test)){  
 diffs[i] = range\_diff(i, i+1)  
}

### Question 6

Imagine you’re drawing simulated data from two normal distributions using the rnorm() function we’ve discussed in class. The two distributions (and functions) both have a mean of 4.6, but the first has the sd argument set to 1 and the second has it set to 5. What statement below is TRUE regarding the two datasets generated from the rnorm() function?  
a. The second set will have a much wider range of values that deviate more from the mean compared to the first set. b. The second set of data will be much tighter around the mean of 4.6.  
c. The data sets will be similarly shaped, except the sd=5 argument will start to turn the distribution into something more closely resembling the Poisson distribution.  
d. The first set will demonstrate higher variance because of a smaller standard deviation. e. It depends on the sample size.

### Question 7

Write some psuedo-code (it doesn’t have to be syntaxically perfect, but it should generally convey your thought process) to solve the following challenge. Write a function that iterates through a vector and generates an output vector where each value is the value in the original vector divided by the mean of the entire vector.

### Question 8

You are tasked with creating a simulation that for the average number of trout in a recirculating aquaculture tank that are infected with a parasite that causes whirling disease. Rates of this disease are typically low but there is high variance, and your simulation randomly draws values from a normal distribution with a mean of 1.2 (low) and a standard deviation of 4 (high). What is one problem with this simulation (hint: think about the data that would be generated and whether it makes sense in the real world). In 1-5 sentences, describe the problem.

### Question 9

Pick the option below that best describes a major flaw in the code block presented.

input\_vec = c(1,2,3,4,5,6,7,8,9,10)  
output\_vec = c()  
for(k in 1:10){  
 value = sqrt(k)/(k^2)  
 return(k)  
}

1. There is nothing wrong with this code block, it returns an appropriate output vector
2. The length of the output vector isn’t defined on initialization
3. k is an innappropriate choice of variable, it should be i
4. The output vector isn’t populated within the for loop
5. The loop isn’t generalizeable beyond a vector 10 units long

### Question 10

Examine the figure below that explores the data we used for review (size of fish at day 30 of development as a function of different amounts of soy protein in their diet. What conclusions can you draw from this data exploration? How might the figure be improved? What other variables would you want to examine for further analysis? (5-10 sentences).

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────────────────── tidyverse 1.3.0 ──

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

## ── Conflicts ────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

df = read\_csv("https://tinyurl.com/yg886rqz")

## Parsed with column specification:  
## cols(  
## tank\_id = col\_double(),  
## fish\_id = col\_double(),  
## perc\_soy\_protein = col\_double(),  
## day\_30\_weight = col\_double(),  
## avg\_tank\_temp = col\_double()  
## )

glimpse(df)

## Observations: 320  
## Variables: 5  
## $ tank\_id <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
## $ fish\_id <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…  
## $ perc\_soy\_protein <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.…  
## $ day\_30\_weight <dbl> 148.140833, 496.072560, 140.918247, 150.143106, 376.…  
## $ avg\_tank\_temp <dbl> 74.97763, 77.59575, 76.50525, 77.38325, 71.85171, 74…

df %>%  
 ggplot(aes(x = perc\_soy\_protein, y = day\_30\_weight, color = factor(tank\_id))) +  
 geom\_jitter(width = 0.025)

