homework_1

2025-09-17

You may write your answers directly in this file or you may make a new file. Either is acceptable. HINT: Use earlier labs to guide you

Question 1

1. What does it mean for two variables to be associated?
2. What does it mean for two variables to be independent?
3. Give an example of a pair of variables that you'd expect to be associated. They may be from any field (agriculture, physics, medicine, cooking, etc. . .).
4. Give an example of a pair of variables that you'd expect to be independent. They may be from any field (agriculture, physics, medicine, cooking, etc. . .).

Question 2

(Stolen from the textbook) Florence Nightingale was the founder of modern nursing, served as a nurse in the Crimean War, and was an early statistician. In her notes, she opined, "In comparing the deaths of one hospital with those of another, any statistics are justly considered absolutely valueless which do not give the ages, the sexes, and the diseases of all the cases."

1. Name the three variables that Nightingale thought was associated to the number of deaths in a hospital and mention why type of variables they are
2. Name another two variables that might be associated with the number of deaths at different hospitals.
3. Why does Nightingale say that the statistics are "valueless" if given without being broken down by age, sex, and disease? Explain.

Question 3

Run the following code:

```
data("OrchardSprays")
head(OrchardSprays)
```

```
##   decrease rowpos colpos treatment
## 1       57      1      1         D
## 2       95      2      1         E
## 3        8      3      1         B
## 4       69      4      1         H
## 5       92      5      1         G
## 6       90      6      1         F
```

1. Go to the help page for the data set OrchardSprays and briefly name and describe the variables "decrease" and "treatment". Please be sure to give the type of variable; if the answer is ambiguous please explain why. NOTE: you do not need to talk about the variables "rowpos" nor "colpos" (which deals with the spatial layout for the experiment)

2. Find the mean number of the variable 'decrease' observed in the data

3. Create a grid of boxplots, one boxplot for each treatment group. You should end up with 8 different boxplots. Make sure you run the code library(ggplot2) before you use the ggplot() function. NOTE: you shouldn't need faceting to do this; instead think of which variable you want as the x-axis and which as the y-axis

4. Comment on whether you believe your answer to Question 3 part 2 is a "useful' ' statistic given the graph you made in Question 3 Part 3

5. Comment on which treatment groups have outliers

6. Comment on whether any treatment group has a skewed distribution and how you came to that conclusion (ie what did you look at to decide?). Be sure to mention which treatment(s) might be skewed please.

7. Does this graph suggest the response variable "decrease" is independent of the explantory variable "treatment"

8. Add a (useful) main title and a subtitle which includes your name. Also give axis labels that look more professional (eg capitalize the first letter)

9. As your graph stands, is the graph color-blind friendly? We are missing alt-text for the moment but that is not related to color blind friendliness; alt-text is more for people with poor vision or the blind

10. Add alt-text. Be sure to mention the type of of graph, what your axis are, and major trends/takeaways you'd like your audience to know. You may assume the audience you are writing for has already read the help page for OrchardSprays in it's entirety.

11. Comment on the general relationship between the median of the boxplots and the general spread of the boxplots (ie IQR). In otherwords, as the medians increase what generally happens to the boxplot widths.

12. Focus just on treatment H. Compared to the median (solid black line in the middle of the boxplot), where do you think the mean will be?

BONUS POINT: Feel free to make the graphic aesthetically appealing. Whoever makes the best (as rated by the grader) will receive 2 (yes, 2!) bonus points. This is optional; I want you to work on making high quality graphics.

Question 4

This question deals with a somewhat famous data set. William "Student" Gosset was a super important statistician and he used data collected on scallywags, ruffians, and hooligans that made up the British prisons at the turn of the twentieth century. The data included the length of the middle finger and the prisoner's height. Let's see if they are related.

Run the following code:

```
criminals <- read.csv('https://vinnys-classes.github.io/data/criminals.csv')
head(criminals)
```

```
##   finger height
## 1   10.0 142.24
## 2   10.3 144.78
## 3    9.9 147.32
## 4   10.2 147.32
## 5   10.2 147.32
## 6   10.3 147.32
```

1. Calculate the mean for the fingers of the criminals
2. Create a histogram for the finger variable and describe the distribution. Be sure to mention the number of modes, skews/symmetry, and if there are outliers
3. Calculate the mean for the height of the criminals
4. Create a histogram for the height variable and describe the distribution. Be sure to mention the number of modes, skews/symmetry, and if there are outliers

HINT: The spacing in the data can be odd. Think about playing around with the "bins" parameter in geom_histogram(). Go to the Arguments section of the help page for geom_histogram() to read about it.

5. Create a scatterplot of the two variables and describe the form, strength, direction, and if there are outliers. NOTE: It's not obvious what should be the x-axis and what should be the y-axis; as such use "finger" for the x-axis
6. Find and report the number of observations in the data set.
7. Does it look like there is that many data points on your scatterplot?
8. Why might there not be as many data points as there are rows? HINT: look at the first six rows of the data that I printed out above.
9. Copy and paste your scatterplot code. Change geom_point() to geom_jitter() and run the code. Comment on what, if anything, changed. HINT: if needed look at the Description section of the help page for geom_jitter
10. Add a main title, subtitle, professional looking axis labels, and alt text to the graph.