NAME:

STAT 503 – Statistical Methods for Biology

Homework 2

24 Points (26 available). Due at 11:59 PM on Friday, February 12, 2021

**Please use complete sentences unless the question is marked with an asterisk (\*).**

Please download the Dalgleish *et al*. tree data from the class website and save it to your hard drive (the file is titled dalgleish\_et\_al\_trees.csv). This dataset describes the species and diameter-at-breast-height (dbh, about 1.4 m above the ground, measured in cm) of 5681 individual trees at a site in Maine. It also contains age, reproductive data, and disease status for American chestnut (*Castanea dentata*) trees at this site, as well as map coordinates for each tree. We will use these data to practice working with large datasets and calculating summary statistics.

1. Follow the steps below to build and present a frequency distribution for dbh in yellow birch.
   1. \*[0.5 point] After you read in the data, use the filter() and arrange() functions from the ***dplyr*** package to create a new data frame that contains only data for yellow birch trees and has the rows arranged in order of increasing dbh (i.e., the first row has the smallest tree, the last has the largest). There should be a total of 45 dbh values, ranging from 11.430 – 49.022 cm). Please write the first 5 dbh values here.

**Answer**

**11.430, 11.938, 13.462, 14.224, 14.224**

* 1. \*[4 points] Use the 45 dbh values that you obtained in question 1a to build a frequency **table** for dbh in yellow birch at this site. Please begin your table at dbh 14 cm, use 4 cm wide bins, and include columns for both absolute and relative frequency. Do not use R for this question, except to get the data. See the PDF file *Guidelines for reporting statistical results* on the website under Cheat Sheets for guidelines on table formatting, and round decimals to 3 decimal places.

|  |  |  |  |
| --- | --- | --- | --- |
| **Species** | **cut\_width (dbh, width = 4)** | **abosulate frequency** | **relative frequency** |
| yellow birch | 14 -18 | 16 | 0.381 |
| yellow birch | 18-22 | 8 | 0.190 |
| yellow birch | 22-26 | 7 | 0.167 |
| yellow birch | 26-30 | 2 | 0.048 |
| yellow birch | 30-34 | 4 | 0.095 |
| yellow birch | 34-38 | 3 | 0.071 |
| yellow birch | 38-42 | 1 | 0.024 |
| yellow birch | 46-50 | 1 | 0.024 |

* 1. \*[2.5 points] In **Tutorial 1**, you learned how to control the number of bins in a histogram with the bins argument to qplot(). As an alternative, you can specify the width of the bins instead of the number of bins, using the argument binwidth. For example, if I have assigned my yellow birch data to a data frame named yb, then the following statement will draw a histogram of the table from question 1b (don’t forget to load ***ggplot2***!):

qplot(x = dbh, data = yb, binwidth = 4)

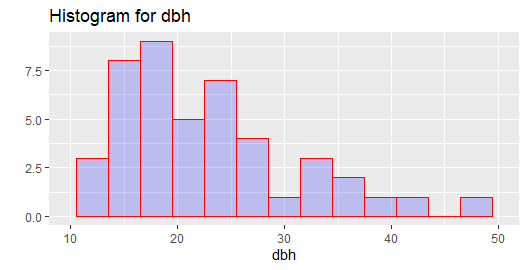
Finally, if you would like to, you can define the specific break points that divide the bins by passing a vector of break points to the breaks argument. For example:

qplot(x = dbh, data = yb, breaks = seq(from = 10, to = 50, by = 4))

This last option gives you complete control over the histogram (for example, try changing the start and end points to 11 and 51 cm, respectively).

Whether you choose to specify the number of bins, the binwidth, or the exact breaks, the goal of building a histogram is to approximate the distribution of the variable in the population. To do this, we must "focus" or "tune" the number of bins so that we can see the overall shape of the distribution while minimizing the effects of random noise due to sampling error. If we use too few bins, then we “oversmooth” the distribution and make it difficult to see any patterns. On the other hand, using too many bins “undersmooths” the plot and visualizes random noise in our particular dataset rather than the underlying distribution of the population that the data came from.

With this in mind, please use qplot() to generate a series of 10 histograms with 3-12 bins, and then select **one** of those graphs and present it here with a caption that includes the number of bins used in the graph (refer to the *Guidelines for reporting statistical results* for information on presenting graphics and captions; the number of bins is not normally reported – I am asking for it for the convenience of the graders). In addition, briefly explain why you chose the plot that you selected.

binwidth = 3

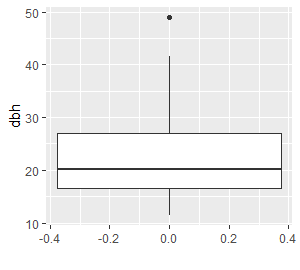
To properly show the pattern in the distribution of the data, a smaller bin was utilized thus, changing the bins affect the appearance of the histogram.

* 1. [2 points] You can use qplot() to generate a boxplot instead of a histogram using syntax similar to this:

qplot(y = myVariable, data = myData, geom = “boxplot”)

(see **Tutorial 4** for additional information on plotting). Please present a boxplot of the yellow birch dbh data, and briefly explain how to interpret this plot. What does each part of the plot represent? Do you see any evidence of potential outliers? If not, what would they look like if they were present?

**Note:** qplot() will place arbitrary numeric values on the x-axis. You may ignore them.



Explanation:

There is evidence of potential outlier

* 1. [1 point] Based on your graphs, please describe the distribution of dbh for yellow birch in this dataset (imagine trying to explain the distribution to your advisor on the phone). Your description does not need to be numerically exact, but remember that we are interested in shape, location, and spread (in this context, it is acceptable to use range to describe spread).

Spread is not normally distributed. The dataset is slightly skewed to the right.

1. Repeat the procedure that you used in question 1a to get the dbh data for eastern white pine. This should give you 14 values ranging from 10.160 – 70.866 cm. Please use these data to calculate the following summary statistics. Then, for each statistic, briefly explain its interpretation.

You should do these calculations by hand (calculators are okay), and you should **always** show your work on any calculation. For guidance see *Answering calculation problems* on the website under Cheat Sheets > Student Advice. Explanations should be brief and to the point (1-2 sentences), but they do need to give more detail than just “it represents location” or “spread.” What, specifically, does the statistic represent?

**Please round answers to 3 decimal places.** To avoid rounding errors, carry at least 2 more decimal places than you need in any intermediate calculations (e.g., use the mean to 5 decimals when calculating variance).

**OPTIONAL (not to be turned in):** use summarize() to double-check your calculations in R (note that your manual calculation for IQR will be differ from the answer in R).

* 1. [2 points] Mean

26.198

* 1. [2 points] Median

13.335

* 1. [3 points] Variance and standard deviation

555.878 and 23.577

* 1. [3 points] Approximate interquartile range (the hand calculation will differ from R)

20.260

1. In this exercise, you will use the Dalgleish dataset to generate a table that summarizes the root collar diameter (rcd) data for American chestnut seedlings of different ages, and then use that summary to generate a graph. Root collar diameter is the diameter, in cm, of the stem where it emerges from the ground. In this dataset, age is recorded in years (age 0 seedlings germinated in the year when the data were collected). Once you have summarized the data, we will use it to learn about the trees at this site.
   1. \*[2 points] Begin by generating a summary table for rcd in American chestnut seedlings that are 5 years old. Your table should include 6 rows of results (one for each age group that is included in the analysis). **Tutorial 3** discusses summary statistics and how to generate a summary table. You may export the final table using the procedures in **Tutorial 3 Section 4**, or you may copy the values into your homework manually.

**Follow the steps below to build a table that include the following statistics, in order:** sample size, mean, standard deviation, coefficient of variation, the 25% quantile, median, the 75% quantile, and IQR. See the select() function in **Tutorial 2** for instructions on how to reorder columns.

* + 1. Use filter() to extract the rows for American chestnut trees with age 5 years. You will also need to remove any trees that have missing values for rcd. To do this, you can use the search criterion !is.na(rcd) in filter(). This process is often necessary when we want to work with a subset of a larger dataset.
    2. Use summarize() and group\_by() to generate your table.
    3. The ***dplyr*** package uses a special type of data frame called a “tibble,” that can behave oddly when printed. To ensure that your summary table prints with the correct number of decimal places, you can modify the following code (note the use of the pipe – see **Tutorial 2, Section 3.4**):

myTibble %>% # a summary table or other tibble

mutate\_if(is.numeric, round, 3) %>% # see below

as.data.frame() # convert the tibble to a data frame

The mutate\_if() function applies the function named in its first argument to see if each column in a dataset meets a criterion. If it does, then the function named in the second argument is applied to the column. All subsequent arguments to mutate\_if are passed to the function in the second argument. Here, we check to see if each column is.numeric(). If it is, we round() the value to three decimal places. See the help files for mutate\_if(), is.numeric(), and round() for details.

* + 1. Present your table here. Remember to follow the table guidelines in *Guidelines for reporting statistical results*.
  1. [1 point] Based on your table, would the mean or median of rcd provide a better representation of "typical" trees in this dataset? Explain how you reached your answer.
  2. Because some individuals grow more quickly than others, we expect distributions of variables related to body size to become more spread out as organisms get older (see the SD and IQR columns in the table for question 3a). However, the *shape* of the distribution may also change as a function of age, and this can tell us something about the underlying biology of the system.

To make a complete inference about the relationship between age and CV in this population, we would need to find a confidence interval for each of the estimated CVs in the table. However, since we have not learned how to do this yet, we will proceed with a more exploratory analysis.

* + 1. [1 point] Use your summary from 3a to plot the relationship between age and CV in rcd, and present your plot here (remember to follow the guidelines).
    2. [1 point] Based on the plot, how does the shape of the rcd distribution change as seedlings age? What might account for the relationship? I am inviting you to speculate here; to make a complete inference, we would need to find confidence intervals for each of the CVs in the plot (and we do not know how to do this yet).

**OPTIONAL (not to be turned in):** Use ggplot() to graph the quantiles (25%, 50%, and 75%) of rcd against age. To do this, you will want to use three geom\_line() statements (one for each quantile) or use geom\_area() for the 25% and 75% quantiles and geom\_line() for the median. **Tutorial 4** discusses ggplot().

1. [1 point] Please paste your script here. Remember that your script should be an edited, clean document, and should contain comments that mark the code for each question in the assignment.