Homework 2

Due: 2024-09-20 @ 6:00PM

Sarah E. Grabinski

2024-09-15

Objectives

- Effectively describe numerical distributions
- Select the appropriate summary statistics based on distribution shape
- Match numerical distributions to their summary statistics
- Calculate proportions from a contingency table

Instructions

1. Go to File > New Project to create a new project in new folder. Make sure to put that folder on your computer somewhere you can find it (see this Campuswire post for 2 tutorials on how to better organize your files).
2. Download the `homework2_template.qmd` file under Chapter 2 on the Modules page in Canvas. Move this file from your Downloads folder into the new project folder you just created. If you've done this correctly, it should now show up under the "Files" tab in the bottom right-hand corner of RStudio.
3. Open `homework2_template.qmd` in your project in RStudio. Go to File > Save As, and save a copy as `homework2_yourlastname.qmd` to your project folder. This file should also now appear in the "Files" tab.
4. If RStudio puts up a light yellow alert just underneath the tool bar that prompts you to install packages, please install those packages.
5. Go to the toolbar at the top and click "Render" next to the blue arrow.

   a. If your file won't render because you don't have required packages installed, please install those packages using the "Install" button in the "Packages" tab in the bottom right-hand corner. Now try rendering again.

   b. If your file won't render for another reason, try Googling the error message to see if you can figure out what is going on. If you still don't understand the problem, copy-paste the error message or upload a screenshot to Campuswire for help.

6. Once you have rendered your file successfully, it should show up under the "Files" tab as `homework_yourlastname.html`. If it does not, and you cannot find the file, come see me for help.

7. Return to your `homework_yourlastname.qmd` file. Read the rest of this document through before proceeding.

8. Run the very 1st code chunk where all the packages are loaded with the `library()` function. Each time you open a new project, you need to reload your packages first. You can't access the functions in these packages until this is completed.

9. Run the 2nd code chunk where the data is saved as `nhanes_df` to your global environment using the `<-` symbol.

10. Once you have run this two code blocks, you will insert your answers into the rest of the document and modify the code if indicated in the template in order to answer the problems. As you work, render the document relatively often to check that everything is working correctly.

11. When you are finished, download the `homework2_example.html` file under Chapter 2 on the Modules page in Canvas. Compare your completed and rendered `homework2_yourlastname.html` file to `homework2_example.html`.

    a. If they look similar, great. Please go to Canvas and upload ***BOTH*** your `homework2_yourlastname.qmd` file and your `homework2_yourlastname.html` file to Homework 2 on the Assignments page in Canvas.

    b. If they don't look similar, please ask a question on Campuswire or reach out to me so we can troubleshoot the problem. Once it has been corrected, please go to Canvas and upload ***BOTH*** your `homework2_yourlastname.qmd` file and your `homework2_yourlastname.html` file to Homework 2 on the Assignments page in Canvas.

Packages

You need the following packages installed before you can render this document.

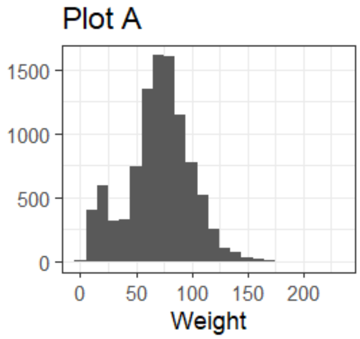
- `NHANES` - this is the package the data is sourced from

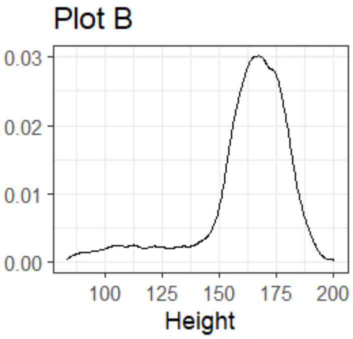
- `Hmisc` - this package contains the `describe()` function for generating numerical summaries
- `janitor` - this package contains the `tabyl` function for construction contingency tables, along with helper functions like `adorn_totals()` for analyzing them
- `kableExtra` - this package contains the `kbl()` function for cleaning up data tables for publication, as well as built-in themes like `kable_classic()`
- `tidyverse` - bundles the packages `ggplot2` with other data management packages like `dplyr`

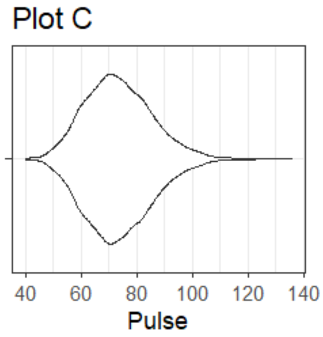
Problems

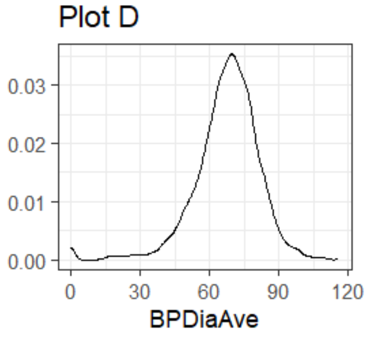
Problem 1 - Describing Numerical Distributions

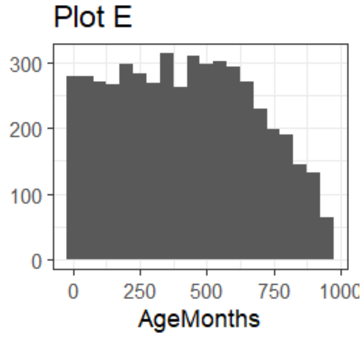
In 1-3 sentences, for plots A-F below, name the type of plot used to show the distribution and describe its shape. Follow the distribution checklist if you need help.

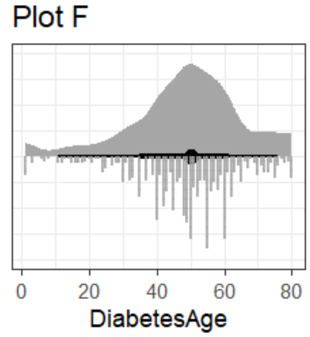

Plot A


Plot B


Plot C


Plot D


Plot E

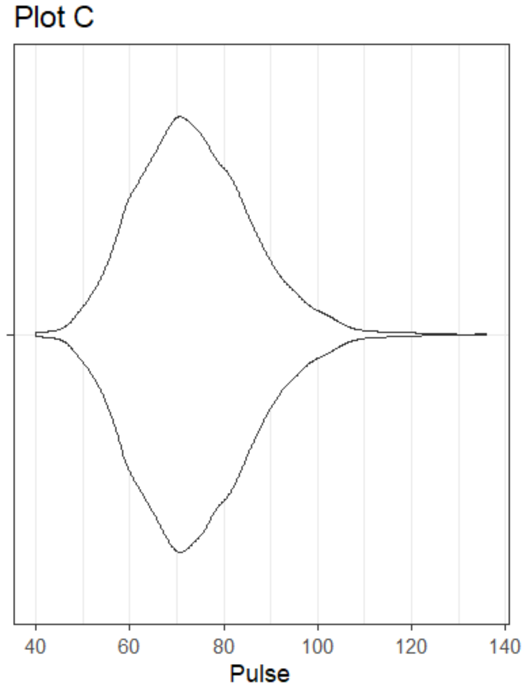

Plot F

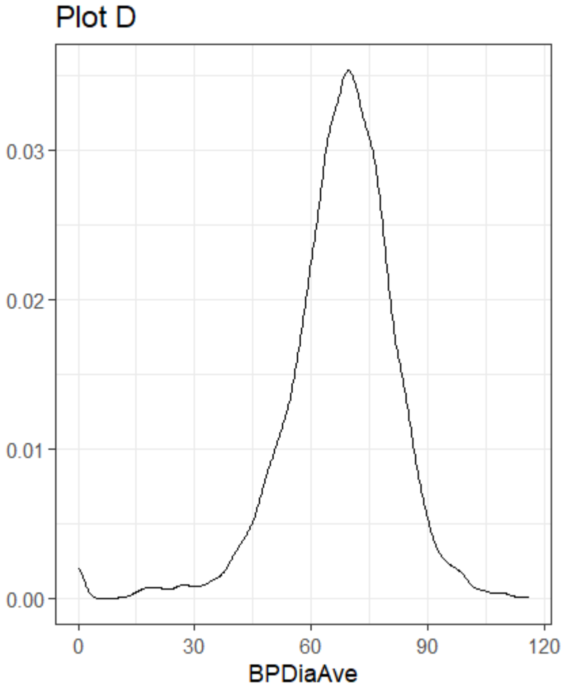
BONUS POINT 1: State whether the mean would be greater than, lesser than, or about equal to the median for each of the 6 distributions.

BONUS POINT 2: In 1-2 sentences, describe what the bottom half of Plot F shows that the top half does not, and why it is useful to combine summary techniques when visualizing distributions.

Problem 2 - Selecting Appropriate Summary Statistics

In 1-3 sentences, for plots C and D below, state whether the mean and standard deviation or the median and interquartile range are the more appropriate summary statistic for that distribution and briefly explain your decision.

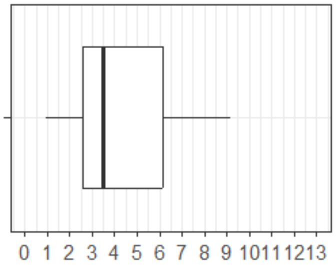




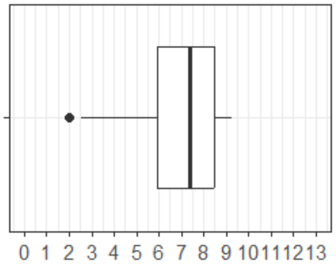
Problem 3 - Matching Distributions to Summary Statistics

You are given a set of 6 boxplots G-L below that describe different numerical distributions.

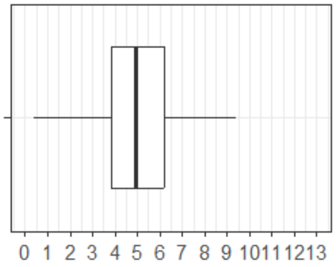
Plot G



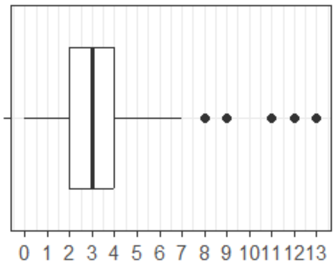
Plot H



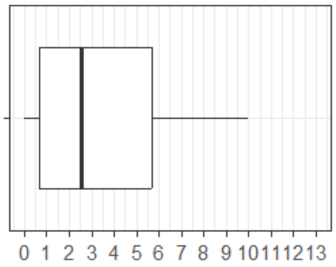
Plot I



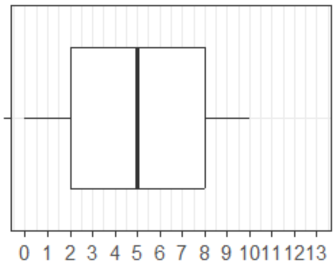
Plot J



Plot K



Plot L



You are also given a table with 6 different 5-number summaries, each belonging to one of the boxplots above.

Minimum	Q1	Median	Q3	Maximum	Boxplot	Mean vs Median
0.95	**2.58**	**3.48**	**6.15**	**9.20**		
0.001	**2**	**5**	**8**	**10**		
0.001	**0.66**	**2.55**	**5.66**	**10**		
0	**2**	**3**	**4**	**13**		
0.38	**3.86**	**4.91**	**6.20**	**9.37**		
2	**5.91**	**7.39**	**8.44**	**9.28**		

In the column "Boxplot", you will enter which boxplot you think the 5-number summary belongs to. In the column "Mean vs Median", you will enter whether you think the median is greater than (higher), lesser than (lower), or about equal (same) to the median given.

Problem 4 - Calculating Contingency Table Probabilities

You are given a 2 x 2 contingency table of counts for the variables `Gender` (male, female) and `SleepTrouble` (yes, no) from the NHANES data set. It looks like the table below.

	SleepTrouble		
Gender	**No**	**Yes**	**Total**
Female	2789	1164	3953
Male	3010	809	3819
Total	5799	1973	7772

You are to calculate the proportion of observations which fall into each combination of categories: Female/No, Female/Yes, Male/No, Male/Yes.

1. Calculate proportions by row using the row totals as the denominator. Enter them into the empty table.
2. Calculate proportions by row using the column totals as the denominator. Enter them into the empty table.

BONUS POINT 3: What is the *range* of a proportion?

BONUS POINT 4: In 1-2 sentences, explain why we use proportions in addition to counts to describe the distributions of categorical data.

Additional Help

- ***Watch this short video*** in the Tutorials section under Modules on Canvas that walks through the instructions for this homework. It will help if you also watched this video for homework 1 first.
- ***Read the textbook.*** Many of you are asking for additional examples. Luckily, there are tons we didn't go over in the textbook.
- ***Ask a question on our Campuswire class feed.*** I'm only one person, and I may not be able to give you a prompt answer. However, the 20+ other people in the class might be able to.
- ***Come to office hours.*** I will be available after class Monday 9/23/2024 and Wednesday 9/25/2024 from 2:30pm - 4:00pm. If you cannot make it, reach out to me to try and schedule an appointment.

Late Policy

- This homework is due by 6:00pm on Friday, 9/20/24.
- No credit will be lost for assignments received by 7:00pm on Friday, 9/20/24 to account for issues with uploading.
- Assignments received by 9:00am on Saturday 9/21/24 can get a maximum grade of 90%.
- Assignments turned in after 9:00am on Saturday 9/21/24 have a maximum grade of 50%, so it benefits you to turn in whatever you have completed by the due date.

Updates

This document was last updated on 2024-09-15.