Homework 2 Answer Key

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-14

Objectives

- 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

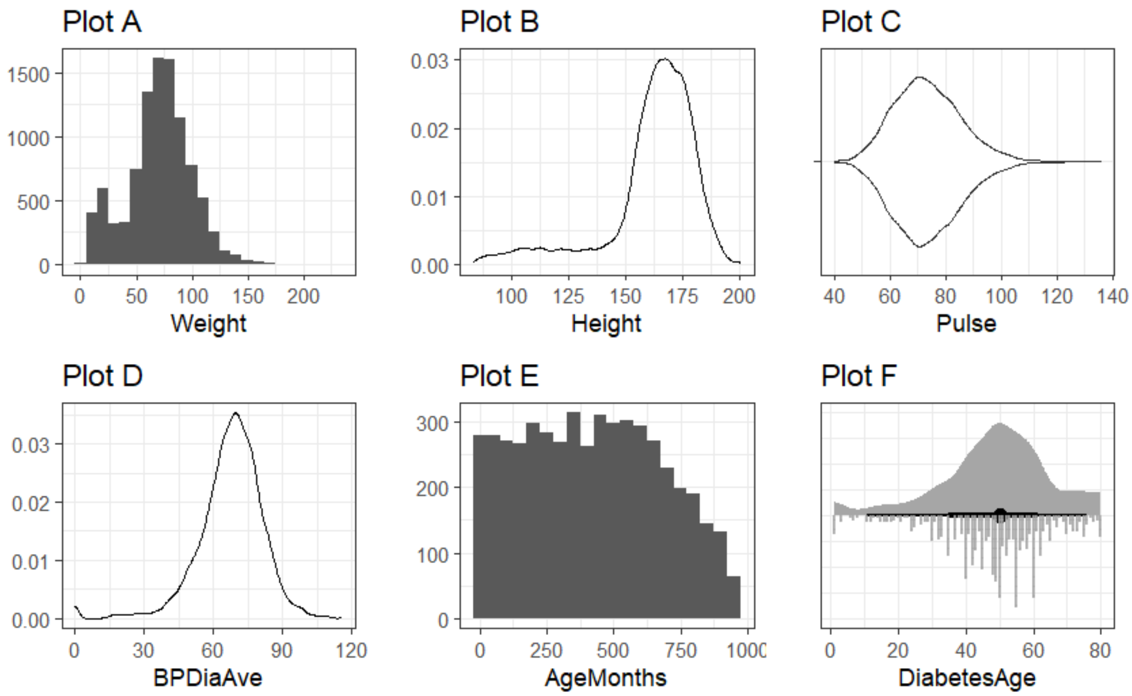
Problem 1 - Describing Numerical Distributions

For this exercise, I was looking for you to identify the plot type and describe the modality of the distribution. I was also looking for you to identify if the distribution was skewed or symmetrical and if there were any outliers. Because this is real data, there wasn't necessarily one "correct" description of each distribution, which is one of the challenges in statistics.

For each plot, students earned 1 point for correctly identifying the plot and 2 points for properly describing the distribution. Points were deducted for incorrect answers or when an item from the distribution checklist was missing (i.e. modality, skew/asymmetry, outliers). Points were deducted down to half credit, as long as the question was attempted. Students received 0 points for not attempting the question(s).

For bonus point 1, I was looking for you to recognize that when there is skew or asymmetry, the mean is expected to be different than the median. Because these differences can be large or small depending on the degree of skew and/or number of outliers, I accepted multiple answers where it was reasonable to do so.

For bonus point 2, I was looking for you to recognize that some values are MUCH more common than others, which indicates that the underlying data might have a pattern that's worth investigating. This pattern is not visible in the density curve, which is why combining strategies can be useful. I also accepted any answer that described in general why it is useful to combine visualization techniques in general.



Plot A

This plot is a histogram. It is bimodal and asymmetrical. Although this distribution does not actually have skew or outliers, I did accept answers these as answers. This distribution is actually made up of 2 separate normal distributions, neither of which are skewed. This distribution only looks like it has a long right-hand tail, but it's actually because the other mode/distribution is covering up the left-hand tail. You know the peak on the left-hand side of the plot is not made up of outliers because of how many observations there are. Outliers are both unusual and RARE.

Plot B

This plot is a density curve. It is unimodal and has outliers. I accepted answers that said left-skew, but your clue that these are outliers and not skew is that the overall shape of the distribution does not "lean" to the left. I accepted both symmetrical and asymmetrical for this distribution, because when you don't have too many outliers, it is often acceptable to use methods that require symmetry in distributions. There is not a hard cutoff for deciding when outliers become skew/asymmetry.

Plot C

This plot is a violin plot. The distribution is asymmetric and unimodal with both right skew AND outliers. Many people described this plot as bimodal, and I'm not entirely sure why. Many people also described this as a normal distribution, but it is not.

Plot D

This plot is also a density curve. This distribution is unimodal and symmetric with outliers. This is one of the few scenarios where its acceptable to say this is a normal distribution.

Plot E

This plot is also a histogram. This distribution is uniform but asymmetrical with right skew. It has no outliers.

Plot F

This plot is a raincloud plot. The distribution is close to symmetric, although you could argue it has some left skew. Although the density plot appears unimodal, you can see from the dot plot portion that this distribution might also be considered multimodal, so I accepted this answer as well. This distribution has both positive and negative outliers, which you can tell because they occur outside of the boundary of the whiskers. Again, real data is not always going to follow an easy-to-describe distribution, which is one of the challenges of data analysis.

Bonus Point 1

Although you should have been able to get these answers just by looking at the plots, you also had a numerical summary of the data available to you just above this question. This numerical summary included both the mean and median, so this question was really easy if you paid attention.

A. The mean would be *less than* the median.

B. The mean would be *less than* the median.

C. The mean would be *greater than* the median.

D. I accepted both that the mean would be *less than* or *approximately equal* to the mean.

E. I accepted both that the mean would be *greater than* or *approximately equal* to the mean.

F. I accepted both that the mean would be *less than* or *approximately equal* to the mean.

Bonus Point 2

I was hoping you would recognize here that from the dot plot, you can see that there are actually many modes in the data, that it is not the smooth, continuous distribution that the density plot implies. If you look closely, you can see that these modes occur at ages that are multiples of 5 (e.g. 40, 45, 50, 55, 60…). It is important to investigate patterns like these in your data, as it could inform how you need to analyze it.

This sort of pattern is likely due to the fact that this measure is not based on medical records but the subject's ability to recall the age at which they were diagnosed. When people have to recall things like ages, weights, etc. from the past or guess an arbitrary number (e.g. how many jellybeans are in this jar?), they tend to round their answers to whole numbers and numbers that are divisible by 5 or 10. When you see this type of pattern in your numerical data, it might be beneficial not to analyze it as a continuous number, but to group the ages into "bins" as you do in a histogram and analyze them as categories.

Problem 2 - Selecting Summary Statistics

For this question, I wanted you to recognize that when there is a meaningful amount of skew and/or outliers in a distribution, you should use *robust statistics* like the median and interquartile range, which are less sensitive to those characteristics. Only when a distribution is mostly symmetrical with few outliers and little-to-no skew, specifically when it follows the classic "bell curve" shape, is it appropriate to summarize a distribution with the mean and standard deviation.

Points were only deducted for incorrect answers down to half credit, as long as the question was attempted. Students received 0 points for not attempting the question(s).

Many students lost points here because they did not read the instructions and repeated the activity in problem 1. Its unfortunate to take off so many points for such a silly mistake, so please make sure you're following directions carefully.

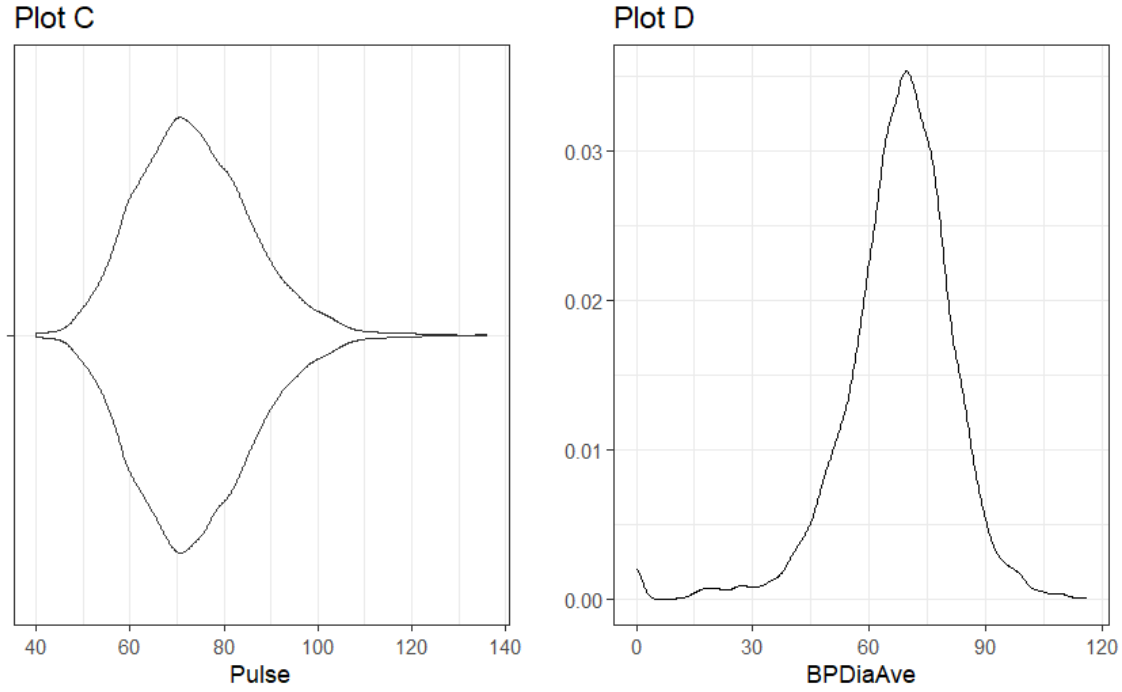


Figure 1: en the instructions

Plot C

For this plot, the only correct answer was the median and interquartile region. This distribution has both right skew AND outliers, so the median and standard deviation would not accurately represent the “center” of the data. The median and interquartile region are *robust* to skew and outliers, so they will be better descriptors of the “center” of the data.

Plot D

This is a beautiful example of a normal distribution in real-world data, because although it has a couple outliers at the low end of its range, they are few and not very extreme. Furthermore, even with the outliers, the distribution is still very symmetrical, and it has that classic “bell curve” shape. This distribution is one of the few scenarios where it would be acceptable to use the mean and standard deviation as a numerical summary.

Many of you chose the mean and standard deviation and justified it because the distribution is symmetrical. While a symmetrical distribution is a start, the distribution requires that “bell curve” shape for the mean and standard deviation to accurately describe the “center” of the data. For example, a uniform distribution is symmetrical, but the mean and standard deviation are not appropriate because it does not have the right shape.

However, it is basically never wrong to use the median and interquartile region to summarize a distribution numerically. When the distribution is close to normal, the median/IQR and mean/SD are going to describe extremely similar ranges of the data. The mean + SD describes the center 68% of the data, whereas the median + IQR describes the center 50% of the data. When the distribution is symmetrical with few-to-no outliers, the mean and median will be approximately equal. Therefore, it was not incorrect to say median + IQR, and I gave partial credit for this. However, because the mean+SD describes MORE of the distribution (68% of the data) than the median+IQR (50% of the data), it is better to use the mean + SD when the distribution looks normal.

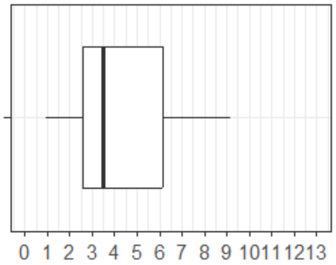
Problem 3 - Matching Distributions to Statistics

When looking at a boxplot, the box is always defined on the lower end by the value for quartile 1 (Q1, 25th percentile) and on the upper end by the value for quartile 3 (Q3, 75th percentile). The bold line in the middle of the box indicates the value of the median (Q2, 50th percentile). You should have been able match the approximate values of these features to the actual values in the table.

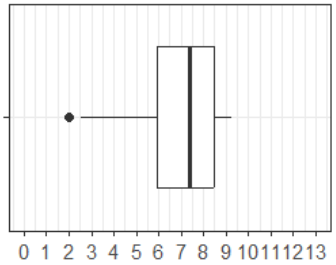
When a distribution is asymmetrical or there are outliers, the ranges for data less than the median will differ from the ranges for data greater than the median. For example, in plots G and K, the area of box on the right-hand side of the boxplot is larger than the area on the left-hand side. The whisker on the right-hand side is also longer than the whisker on the left-hand side. For plots H and J, extreme values (outliers) are indicated by the points. Although the median is robust to asymmetry and outliers, the mean is not. The mean will trend towards the side with more data and/or more extreme values.

Points were only deducted for incorrect answers down to half credit, as long as the question was attempted. Students received 0 points for not attempting the question(s).

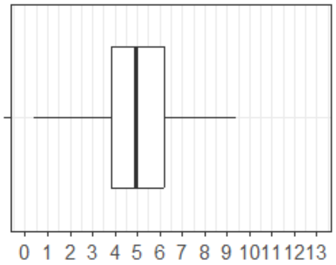
Plot G



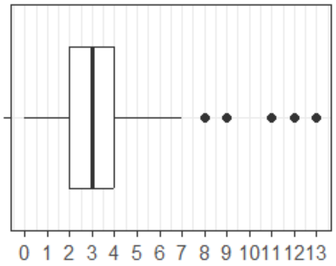
Plot H



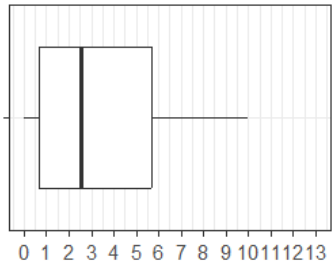
Plot I



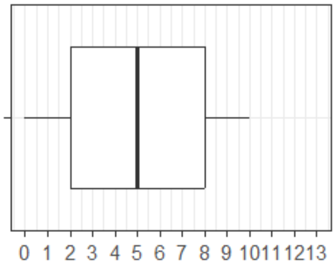
Plot J



Plot K



Plot L



Minimum	Q1	Median	Q3	Maximum	Boxplot	Mean vs Median
0.95	**2.58**	**3.48**	**6.15**	**9.20**	Plot G	Higher
0.001	**2**	**5**	**8**	**10**	Plot L	Same
0.001	**0.66**	**2.55**	**5.66**	**10**	Plot K	Higher
0	**2**	**3**	**4**	**13**	Plot J	Same/ Higher
0.38	**3.86**	**4.91**	**6.20**	**9.37**	Plot I	Same
2	**5.91**	**7.39**	**8.44**	**9.28**	Plot H	Same/ Lower

Problem 4 - Calculating Proportions for Contingency Tables

Here, I wanted you to calculate the proportions by row and by column for the provided contingency table. When the proportions were done by row, your row totals should all have equaled 1 or 100%. When the proportions were done by column, your column totals should all have equaled 1 or 100%.

Table 2: Self-Reported Trouble Sleeping by Gender in the NHANES study for years 2009-2012 (n = 7,772)

Gender	No	Yes	Total
female	2789	1164	3953
male	3010	809	3819
Total	5799	1973	7772

Points were deducted for incorrect answers down to half credit, as long as the question was attempted. Students received 75% credit if they had a correct answer but mixed up row and column proportions. Students received 0 points for not attempting the question(s).

```r
nhanes_df |>
  filter(!is.na(Gender), !(is.na(SleepTrouble))) |>
  janitor::tabyl(Gender, SleepTrouble) |>
  janitor::adorn_totals(where = c('row', 'col')) |>
  kableExtra::kbl(caption = 'Self-Reported Trouble Sleeping by Gender in the NHANES study fo
  kableExtra::kable_classic(full_width = F)
```

Row Proportions

Table 3: Proportions of those who do or do not have trouble sleeping by gender in NHANES subjects from the years 2009-2012 (n = 10,000)

	No	**Yes**	**Total**
Female	0.706 or 70.6%	0.294 or 29.4%	1 or 100%
Male	0.788 or 78.8%	0.212 or 21.2%	1 or 100%
Total	0.746 or 74.6%	0.254 or 25.4%	1 or 100%

Column Proportions

Table 4: Proportions of men and women by whether they have trouble sleeping or not in NHANES subjects from the years 2009-2012 (n = 10,000)

	No	**Yes**	**Total**
Female	0.481 or 48.1%	0.590 or 59.0%	0.509 or 50.9%
Male	0.520 or 52.0%	0.410 or 41.0%	0.491 or 49.1%
Total	1 or 100%	1 or 100%	1 or 100%

Bonus Point 3

The range of a proportion is ALWAYS 0 to 1. A proportion is the *count* of items in a subset divided by the total *count* of all items. A proportion CANNOT be less than 0 because you cannot have fewer than 0 items in a subset ($p = \frac{0}{n} = 0$). A proportion CANNOT be greater than 1, because you cannot have greater than n items in a subset when your total number of items is n ($p = \frac{n}{n} = 1$).

Bonus Point 4

We use proportions in addition to counts when describing categorical data because it is hard to compare across categories with different total amounts of subjects. For example, it is not immediately obvious from the counts that females are more likely to report trouble sleeping than males ($p = \frac{1164}{1973} = 0.590$) but females and males are about equally likely to have no trouble sleeping ($p = \frac{2789}{5799} = 0.481$). Proportions put counts onto the same scale, so we can more easily identify when there are differences between subsets.