Visualizing Frequency Distributions

Ernesto Diaz
Masters of Computer Science
Florida International University
ediaz188@fiu.edu

Abstract—
A frequency distribution is a list, table or graph that organizes all distinct values of some variable within a given interval. They are mostly used to summarize categorical variables and organize data into a meaningful form so that a trend, if any can easily be spotted. In practice, frequency distributions grant researchers and stakeholders alike the opportunity to glance at an entire dataset conveniently. They can highlight whether observations are high, low, concentrated in one area or spread out across an entire scale. Understanding how to display frequency distributions and some of the options available is a critical first step in fully comprehending a dataset.

***Keywords*—Data visualization, Frequency Distribution, Range, Standard Deviation, Dispersion, Histogram, Frequency Polygon, Box and Whisker Plot, Bubble Chart, Multi-set Bar Chart.**

I. Introduction

In today's digital landscape, data has become complex and bulky as it continues to grow from independent sources. In fact, there is so much data available that the term "Big Data" has become regularly used across industries with data analytics as a driving force behind innovation. For companies hoping to leverage datasets, fully understanding them is key to effectively create strategic advantages in their respective industries. In a typical workflow, the data preprocessing step is the ideal phase to organize the data into a meaningful form so that a trend, if any, emerging out of the dataset can be recognized and further analyzed [10].

One of the most common methods used for organizing data are frequency distributions. A frequency distribution which is an overview of all the distinct values in some variable and the number of times they occur is a standard visualization technique. In practice, frequency distributions are most commonly used to summarize categorical variables in datasets. If constructed well a frequency distribution is sometimes enough to make a detailed analysis of the structure of a population with respect to a given characteristic. Furthermore, one can easily spot whether observations are high or low and concentrated in one area or spread out across the entire scale. In this paper, we will briefly discuss some of the basic characteristics of frequency distributions and the visualization techniques available to illustrate them along with their strengths and weaknesses.

II. Properties Of Frequency Distributions

There are three important characteristics of frequency distributions to comprehend that are consistent regardless of the visualization used.

A. Measures of Central Location

Oftentimes when a frequency distribution is graphed it is common for a significant amount of data points to cluster around a central value. This clustering is referred to as the central location or central tendency of a frequency distribution. Once the value that a distribution centers around is known, it can be used to further characterize the rest of the data in the distribution. To calculate a central value several methods exist with each method producing somewhat of a different result. Collectively these methods can be referred to as "Measures of Central Location" and the three most commonly used are:

1) Mean: the sum of all values divided by the total number of values.
2) Median: the middle number in an ordered data set.
3) Mode: the most frequent value.

These three measures are best used in combination with one another. This is because they have complementary strengths and weaknesses. The mode can be used for any level of measurement, but it is most meaningful for nominal and ordinal values. The median can only be used on data that exhibits some type of order and the mean can only be used on interval and ratio values of measurement. This because it requires equal spacing between adjacent values or scores in the scale [1]. Most of the time depending on the dataset, only one or two of these measures are applicable at any given time.

B. Measures of Dispersion

A second property of frequency distributions is dispersion also know as variation, which is defined as the spread of a distribution out from its central value. The dispersion of a frequency distribution is independent of its central location. Figure 1 illustrates this fact, by showing the graph of three theoretical frequency distributions that have the same central location but different amounts of dispersion. Some of the more common measures of dispersion that are used include the following:

1) Range: the difference between the largest and the smallest observation in the dataset.
2) Interquartile Range (IQR): the difference between the 25th and 75th percentile (also called the first and third quartile).
3) Standard Deviation: Measures the spread of data about the mean.

Much like measures of central location, measures of dispersion have their own set of strengths and weaknesses. For

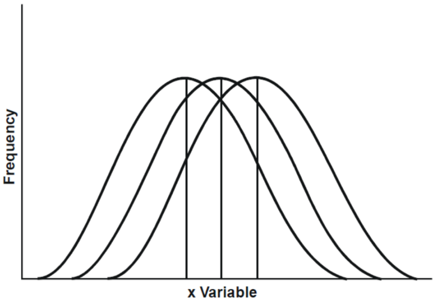

Fig. 1. Three curves identical in shape with different central locations [7].

instance, the biggest advantage of the range is that it is very easy to calculate. But the main disadvantage to be aware of is its sensitivity to outliers. Moreover, the range does not take into account all the observations in a data set. Likewise, in some situations it might be more informative to actually provide the minimum and maximum values rather than the range a singular value [11].

The interquartile range has a rather interesting advantage given that it is not affected by extreme values. Due to this the interquartile range can be used as a measure of variability if extreme outliers do exist in the dataset. However, its main disadvantage is that it is not amenable to mathematical manipulation.

Standard Deviation (SD) is perhaps the most famous and widely used measure in dispersion calculations. The reason why is because if the observations are from a normal distribution, then we can assume that 68% of observations lie between the mean and ± 1 SD, 95% of observations lie between the mean and ± 2 SD and 99.7% of observations lie between the mean and ± 3 SD [11]. The other advantage of standard deviation is that along with the mean it can be used to detect skewness. However, its biggest disadvantage is its inability to be used as an appropriate measure of dispersion for data this is already skewed.

C. Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point. The skewness for a normal distribution is zero, and any symmetric data should typically have a skewness near zero. Negative values for the skewness indicate that a dataset has a majority of its data points skewed left while positive values indicate a majority of data points are skewed right. For context when we say ”skewed left”, we mean that the left tail is long relative to the right tail. Similarly, ”skewed right” means that the right tail is long relative to the left [7]. Luckily, we can define the skewness of a distribution with the following formula:

$$Skewness = \sum \frac{(X_i - \bar{X})^3}{ns^3} \tag{1}$$

where n is the sample size, X_i is the i^{th} X value, $\bar{X}$ is the average and s is the sample standard deviation [9]. However, most software tools such as Microsoft Excel take into account the sample size as well. Therefore, we can slightly modify the formula to the following:

$$\begin{aligned} Skewness &= \frac{n}{(n-1)(n-2)} \sum \frac{(X_i - \bar{X})^3}{s^3} \\ &= \frac{n}{s^3(n-1)(n-2)}(S_{above} - S_{below}) \end{aligned} \tag{2}$$

In practice, as the sample size increases the difference in the results that these two formulas potentially produce is relatively small so either one can be used with confidence.

III. Displaying Frequency Distributions

Frequency distributions can be displayed in a table, or pictorial graphs to fully highlight a dataset.

A. Frequency tables

A frequency distribution is a table that shows “classes“ or “intervals“ of data entries with a count of the number of entries in each class. The frequency f of a class is the number of data entries in the class. Each class will have a “lower limit“’ and an “upper limit“ which can be interpreted as the lowest and highest numbers in each class. The class width is defined as the distance between the lower limits of consecutive classes. Before constructing a frequency table, some consideration should be given about the range of values in the dataset. In situations where there are to many class intervals, the likelihood of reducing the bulkiness of the data is highly unlikely. On the other hand, if the total number of classes is minimal, then the shape of the distribution itself cannot be successfully determined. Generally, for most datasets 6-14 intervals is considered an ideal benchmark [3]. However, this should not be interpreted as the defacto standard as a lot depends on the dataset itself. With that being said, the following are a few general guidelines one can follow when constructing a frequency table.

1) The ideal number of classes can be determined or approximated by the formulas:

$$C = 1 + 3.3 \log n \tag{3}$$

$$C = \sqrt{n} \tag{4}$$

where n is the total number of observations in the dataset. Formally, equation 3 is known as Sturge’s Rule while equation 4 is known as the Square Root Choice Rule.

2) Calculate the range of the data by finding the minimum and maximum data values.
3) Using the range, find the width of the classes which can be determined using the formula:

$$\text{Class Width} = \frac{\text{range}}{\text{number of classes}} \tag{5}$$

4) To find the class limits use the minimum data entry as the lower limit of the first class. Then to get the lower limit of the next class, add the class width. Continue

TABLE I
AGE FREQUENCY'S FOR MALES

Age	Frequency
17	1
20	2
21	6
22	5
23	6
24	5
25	9
26	15
27	18
28	17

*Sample of a simple frequency table.

until you reach the last class. Then find the upper limits of each class.

The main advantage of a frequency table is their ability to quickly reveal outliers and even significant trends within a dataset with not much more than a quick inspection. However, the milage might vary on this as extremely large datasets might make it difficult for a simple inspection to be enough. This leads to a major disadvantage of a frequency table where they are not always readily apparent depending on the dataset size. For instance, there may be situations where unless other visualization graphs are used, characteristics such as skewness are not inherently obvious.

B. Frequency Distribution Graphs

A frequency distribution graph is a diagrammatic illustration of the information in the frequency table.

1) Histogram: A histogram is a graphical representation of the variable of interest in the X axis versus the number of observations (frequency) in the Y axis. Percentages can also be used if the goal is to compare two histograms with a different number of categories. Typically, a histogram is used to depict the frequency when data is measured against an interval or ratio scale. They are great to highlight skewness and work well with large ranges in a dataset. Coincidentally, there is a striking resemblance between a bar chart and a histogram. However, they are nothing alike with three important distinctions between them to be aware of. First off, in a histogram there are no gaps between bars as the variable is continuous. In a bar chart, oftentimes there is a noticeable amount of space between the bars depending on the aspect ratio. Secondly, in histograms the width of the bars have a meaning and do not need to be of equal length as this depends on the class interval. Whereas in a bar chart all the bars widths are equal in length. Finally, in histograms the area of each bar corresponds to the frequency whereas in a bar chart, it is the height [10]. Figure 2 is an example of a histogram created in *R* using the *ggplot2* library. Unfortunately, this figure also exposes a major weakness when interpreting histograms. It is extremely difficult and practically impossible to extract the exact amount of "input" unless the frequencies are listed above the bars. Luckily, there is a version of a histogram that does just this called a frequency histogram should that feature become a necessity.

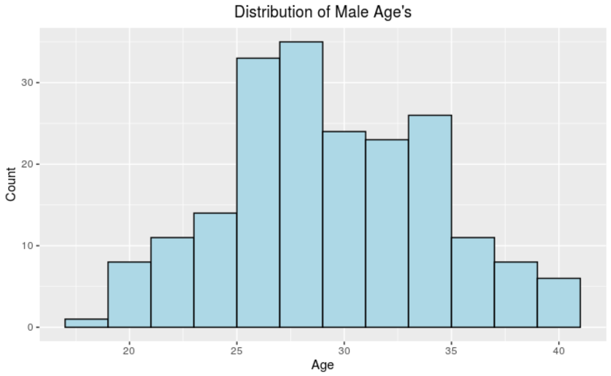

Fig. 2. Histogram built in R using the ggplot2 library.

2) Frequency Polygon: A frequency polygon is very similar to a histogram. In fact, they are almost identical but differ in that frequency polygons are used to compare sets of data or to display a cumulative frequency distribution. A cumulative distribution is a form of a frequency distribution that represents the sum of a class and all the classes below it. They are extremely useful when you need to determine the frequency up to a specific threshold or to easily compare two frequency distributions quickly. Visually, there is also a slight difference where histograms use rectangles while a frequency polygon resembles a line graph. Constructing a frequency polygon is done by connecting all midpoints of the top of the bars in a histogram by a straight line and then removing the bars. Furthermore, when the total frequency is large and the class intervals are narrow, the frequency polygon has the tendency to become a smooth curve which is formally known as the "frequency curve". Figure 3 is an example of a frequency polygon built using the same tools as the histogram. The main advantage of a frequency polygon involves their ability to show central tendencies and dispersion clearly on large datasets. However, given their close relationship to histograms, unsurprisingly frequency polygons share many of their downsides too such as the loss of details on relative numbers and proportions.

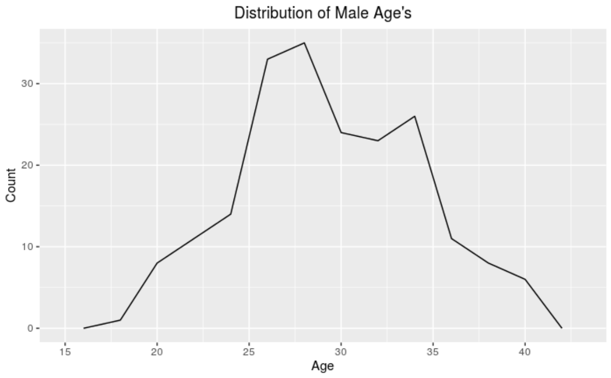

Fig. 3. Frequency Polygon built in R using the ggplot2 library.

3) Box and Whisker Plot: First described by John Tukey in 1977, the box and whisker plot is a histogram type visualization that can be used to illustrate frequency distributions [10]. Constructing a box and whisker plot requires a vertical or horizontal rectangle (box) where the ends are used to represent the upper and lower quartiles. The middle 50% of observations is represented by the box itself and the length of the box indicates the variability of the data while the line inside represents the median. Interestingly enough, the position of the median visually indicates whether or not the dataset is skewed. In fact, for instances where the median is closer to the upper quartile we say that the dataset is positively skewed and for instances where the median is closer to the lower quartile we say it is negatively skewed. The lines outside the box on either side are known as whiskers and each whisker is roughly 1.5 times the length of the box. The ends of whiskers are called the "inner fence"' and any data point beyond its boundary is considered an outlier. Furthermore, the dataset distribution plays a role in the length of the whiskers where if symmetrical the whiskers will be of equal length. But if the dataset is sparse on one side, the corresponding side whisker will be short. The "outer fence" which is roughly defined as the section farthest away from the whiskers is at a distance of three times the IQR on either side of the box. The reasoning behind having the inner and outer fence at 1.5 and 3 times the IQR, is due to the assumption that 95% of observations fall within 1.5 times the IQR, and another 99% for 3 times the IQR [10]. The biggest advantage of the box and whisker plot is its ability to handle large amounts of data and clearly highlight outliers as illustrated in Figure 4. However, the biggest downside is its inability to to retain the exact values and details of a distribution. Unfortunately, this type of visualization is meant to only show a simple summary of the distribution and should be used in combination with other visualizations when analyzing a dataset.

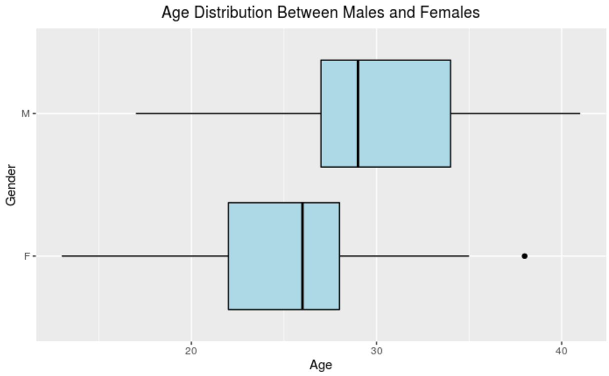

Fig. 4. Box and Whisker Plot built in R using the ggplot2 library.

4) Bubble Chart: A Bubble chart is a type of multi-variable graph that can be viewed as a variation of the Scatterplot. In Bubble charts, the additional dimension of the data is represented in the size of the bubble. Much like a Scatterplot, Bubble charts are plotted on a cartesian coordinate system where the X and Y axis represent separate variables. However, unlike a Scatterplot the Bubble chart assigns a label or category for each point. Colors can also be used to distinguish between categories or used to represent an additional data variable. It is also possible to convey time in the Bubble chart by either having it as a variable on one of the axis or by animating the data variables changing over time through the use of software tools. In practice, Bubble charts are used to compare and show the relationships between categorized circles, through the use of positioning and proportions. The overall picture of Bubble charts can be used to analyze for patterns and correlations. But if to many bubbles are used it can make the chart difficult to read and comprehend as illustrated in Figure 5. Therefore, Bubble charts should be considered as having a limited size capacity when constructed. However, this is mostly true for static versions of Bubble charts as this limitation is somewhat nonexistent for their interactive counterparts.

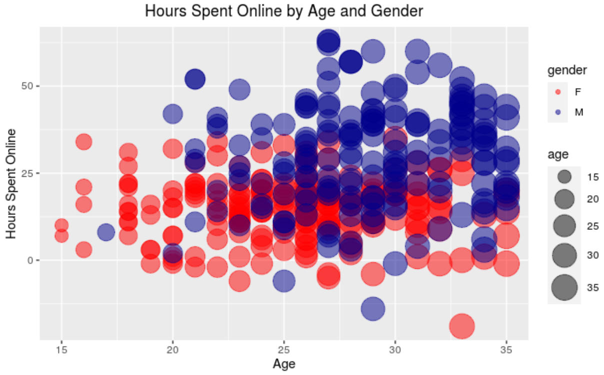

Fig. 5. Bubble chart built in R using the ggplot2 library.

When built with software tools, clicking or hovering over bubbles to display hidden information or having an option to reorganize and filter out grouped categories helps to increase the number of potential bubbles in the chart. Furthermore, to avoid misinterpretations of the bubbles their size should be based on the area of a circle, not the radius or diameter. This is to avoid the side effect of the circles changing exponentially and unintentionally tricking the human eye into interpreting drastic changes in the dataset that are potentially not true.

5) Multi-set Bar Chart: Multi-set Bar Charts which can also be refereed to as Grouped Bar Charts or Clustered Bar Charts are a variation of traditional Bar Charts. They are primarily used when two or more data series are plotted side-by-side and grouped together under categories, all on the same axis. Like a Bar Chart, the length of each bar is used to show discrete, numerical comparisons between the categories. Each data series is assigned an individual color, in order to distinguish amongst the categories. Furthermore, each group of bars is spaced apart to increase visual comprehension. Multi-set Bar charts are best suited to compare grouped variables or categories to other groups with those same variables or category types. The downside of Multi-set Bar charts is that they become harder to comprehend as more bars are added into a group.

IV. Conclusion

The main advantages of certain visualizations for frequency distributions is highly dependent on the dataset and use case. Frequency tables are simple and easy to use and should be constructed first over other more complex visualizations. However, given that their usefulness is highly dependent on data complexity and size caution should be taken to avoid potentially concealing valuable information. Graphs such as histograms are a nice upgrade from frequency tables but should not be the only visualization relied upon. In fact, a combination of multiple graphs like the ones mentioned is ideal to fully illustrate a datasets features and flaws. Ultimately, in practice frequency distribution graphs have the most to offer during the exploratory phases of a dataset and should be used as much as possible before complex algorithms are designed to further process data.

References

[1] Pritha Bhandari. Central tendency: Mean, median and mode.

[2] Isaac Gottlieb. *Frequency Distributions*, pages 90–99. 03 2012.

[3] Helmut Herrmann and Herbert Bucksch. *(frequency) distribution*. 01 2014.

[4] Robert Ho. *Frequency Distributions*, pages 35–60. 09 2017.

[5] Zealure Holcomb and Keith Cox. *Frequency Distribution with Percentages*, pages 8–10. 08 2017.

[6] Basil Jarvis. *Frequency distributions*, pages 13–45. 12 2016.

[7] Guang Jin. Properties of frequency distributions.

[8] Bayo Lawal. *Frequency Distributions*, pages 11–31. 09 2014.

[9] Dr. Bill McNeese. Are the skewness and kurtosis useful statistics?

[10] Manikandan S. Frequency distribution. *Journal of pharmacology & pharmacotherapeutics*, 2:54–6, 01 2011.

[11] Manikandan S. Measures of dispersion. *Journal of pharmacology & pharmacotherapeutics*, 2:315–316, 01 2011.

[12] Edward Schrock and Henry Lefevre. *Frequency Distribution Charts*, pages 167–175. 11 2020.

[13] Heiner Thiessen. *Frequency Distributions*, pages 216–237. 05 2013.

[14] J Wang and Z Ping. Visualization of frequency distribution. *Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine]*, 53:1188–1192, 11 2019.