Visualizing Frequency Distributions

Ernesto Diaz
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 data.**

***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 time where organizing the data into a meaningful form so that a trend, if any, emerging out of the data can be recognized and further analyzed.

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 a few of the visualization techniques available to illustrate them.

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 frequency distribution data 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. 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 ?? 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: the difference between the 25^{th} and 75^{th} 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 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 that does not take into account all the observations in a data set. Likewise, in some situations it be more informative to actually provide the minimum and maximum values rather than the range a singular value.

The interquartile range has a rather interesting advantage given that it can be used as a measure of variability if there are extreme values in the dataset that are not recorded exactly. This leads to the other advantageous feature that the interquartile range is not affected by extreme values. However, the main disadvantage of the interquartile range 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 mean ± 1 SD, 95% of observations lie between mean ± 2 SD and 99.7% of observations lie between mean ± 3 SD. 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 tail. 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, X is the average and s is the sample standard deviation. 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 614 intervals is considered an ideal benchmark. 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:

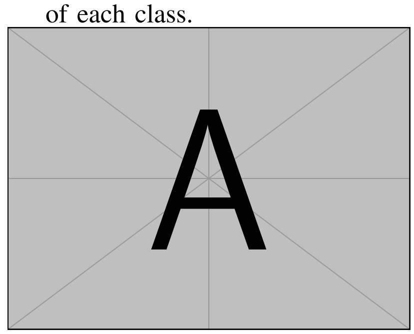
$$C = 1 + 3.3 \log n \tag{3}$$

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

where n is the total number of observations in the dataset.
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:

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

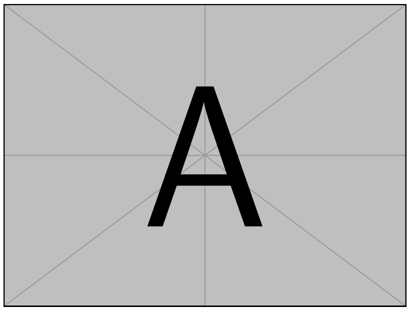
4) To find the class limits use the minimum data entry as the lower limit of the first class. Then to get the lower limit of the next class, add the class width. Continue until you reach the last class. Then find the upper limits of each class.



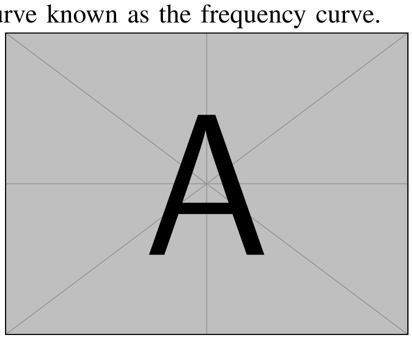
B. Frequency Distribution Graphs

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

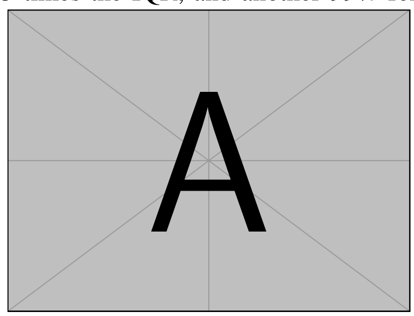
1) Histogram: A histogram is a graphical representation of the variable of interest in the X axis versus the number of observations (frequency) in the Y axis. Percentages can be used if the goal is to compare two histograms with a different number of subjects. Typically, a histogram is used to depict the frequency when data is measured against an interval or ratio scale. Coincidently, there is a striking resemblance between a bar diagram and a histogram. However, they are nothing alike with three important distinctions between them. First off, in a histogram, there is no gap between the bars as the variable is continuous. A bar diagram will oftentimes have a noticeable amount of space between the bars. Secondly, in histograms the width of the bars have meaning and do not need to be of equal length as this depends on the class interval. Whereas in a bar diagram all the bars widths are equal in length. Finally, the area of each bar corresponds to the frequency in a histogram whereas in a bar diagram, it is the height. Figure XX ...



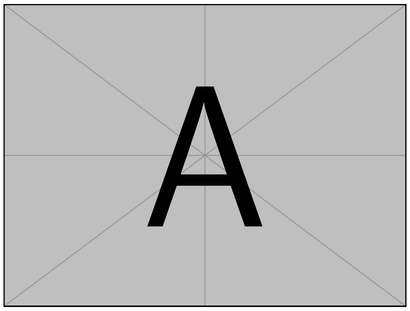
2) Frequency Polygon: A frequency polygon is very similar to a histogram. In fact, they are almost identical except that frequency polygons can be used to compare sets of data or to display a cumulative frequency distribution. A cumulative distribution is a form of a frequency distribution that represents 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 tend to have 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 without displaying the bars. Also, when the total frequency is large and the class intervals are narrow, the frequency polygon becomes a smooth curve known as the frequency curve.



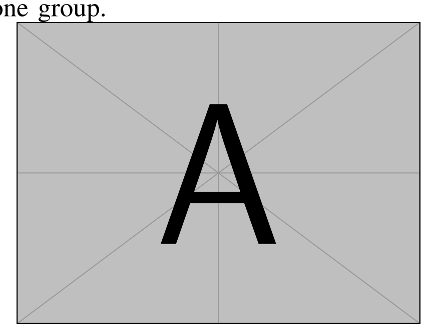
3) Box and whisker plot: First described by John Tukey in 1977, the box and whisker plot is a histogram type visualization that can be used to illustrate frequency distributions. Constructing a box and whisker plot requires a vertical or horizontal rectangle (box) where the ends are used to represent 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, then 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 fact that 95% of observations fall within 1.5 times the IQR, and another 99% for 3 times the IQR.



4) Bubble Chart: A Bubble chart is a type of multi-variable graph that can be views 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 analyse for patterns and correlations. If to many bubbles are used it can make the chart difficult to read and comprehend, therefore Bubble charts have a limited data size capacity. However, this is mostly true for static versions of Bubble charts as this limitation is somewhat remedied for their interactive counterparts. 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 in one group.



IV. Conclusion

References

[1] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.
[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] Bayo Lawal. *Frequency Distributions*, pages 11–31. 09 2014.
[8] Manikandan S. Frequency distribution. *Journal of pharmacology pharmacotherapeutics*, 2:54–6, 01 2011.
[9] Edward Schrock and Henry Lefevre. *Frequency Distribution Charts*, pages 167–175. 11 2020.
[10] Heiner Thiessen. *Frequency Distributions*, pages 216–237. 05 2013.
[11] 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.