

Probability Theory and Introductory Statistics

ALY 6010

Assignment 1

Title: Descriptive & Inferential statistics

Rahul Iyer

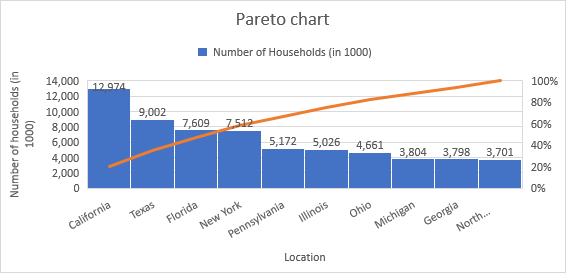
Professor Name: Amin Karimpour

**PART I**

**Introduction**

Pareto chart is one of the seven quality control tools to determine major factors that negatively influence the product/service, by displaying the results categorically. Pareto chart is used to relate majority of the frequency (generally 80%) to a small number of categories (generally 20%). Here, we are determining the 10 largest number of households based on the category-states to identify where most of the population is concentrated.

**Analysis**



The name of the states and its respective number of households have to be sorted in descending order to obtain the 10 largest states. Out of these states, total household must be calculated by adding the individual values of each state; other states must be represented as a percentage of this total calculated value. Cumulative percentage must be estimated with the sum of ‘percentage of households’ consecutively. The pareto chart must be represented with number of households on the vertical axis and categories (states) on the horizontal axis. The percentage axis is added at the right to display the household’s cumulative percentage (the orange curve).

**Conclusion**

It is clear from the Pareto chart the state of California, Texas, Florida and New York account for more than 58% of the total households in the 10 states. This directly means that the population is concentrated in these states, therefore, population density can be calculated to ensure whether or not there is an equal distribution of population area-wise.

**PART II**

**Introduction**

A histogram is different than the conventional bar chart as the former represents only one variable rather than establishing relationship between two variables, as in case of the latter. More specifically, the relative frequency histogram assists in understanding the frequency of variation within the data by comparing it to the total frequency. It can be done based on absolute values or range values. We shall use range values as the number of data points are widely spread and in large quantity.

**Analysis**

The horizontal axis represents the percentage of household with pets while the vertical axis represents the frequency of distribution among the various states. The span of each interval is 10. The first data point lies in the 20-30 interval, followed by more states only in the 50-60 interval. The distribution gradually decreases after this interval.

**Conclusion**

The distribution of the histogram is right-skewed since we have more higher values than lower value data points as it can be seen in the graph. Majority of the data points seem to exist at the interval of 50 to 60, followed by 60-70 interval. It can be inferred that in majority of the locations, 50-70% of the households have pets.

**PART III**

**Introduction**

Cumulative frequency line plot is used to indicate the running total of the frequencies across different data points with the help of a curve. Each data point in this graph is in the form of a range, bounded by upper and lower values. This helps in better understanding of the line plot.

**Analysis**

In case of the ‘percentage of household with pets’, it can be seen that the line plot shows no substantial increase initially until it approaches the range of 50-60, it shows a steep rise compared to the frequency in the 40-50 range which means that majority of the states lie in this region. The orange line indicates the running total while the blue line represents the actual frequency in that interval.

**Conclusion**

It can be concluded that out of the total 49 states, 47 of them have 50-70 percent pets in their household, more towards the lower half of this range (50-60%). While two states are exceptions with extremely low value in the range of 20-30 and higher value in the range of 70-80 each, respectively.

**PART IV**

**Introduction**

Numerical descriptive statistics is used to summarize the data of a sample in the most concise form by providing important quantitative information on the same. These details include mean, median, mode, maximum, minimum, range, sum, standard deviation that would help us obtain a better understanding and interpretation of the data as well as the entire sample.

**Analysis**

Based on the findings of the numerical descriptive statistics, outliers can be determined. Outliers are data points that behave differently from the overall sample. It can be identified based on the lower quartile (first quartile) and upper quartile (third quartile). Interquartile range (IQR) is calculated as the difference between these two values. Values below and above the lower quartile and upper quartile, respectively and/or within the IQR are minor outliers. Major outliers are estimated by multiplying the interquartile range by 1.5 to obtain new range values (upper and lower fence) in terms of upper and lower quartile.

**Conclusion**

The numerical descriptive statistics indicates that the distribution is rightly-skewed as the values of mean, median and the negative skewness value. The outlier among the various locations is determined to be ‘District of Columbia’ with 21.9% of households with pets, lesser than the lower fence of 42.1%.

**PART V**

**Introduction**

The steps for tasks 2 to 4 are iterated in the similar fashion for ‘percentage of dog owners’ and ‘percentage of cat owners’. The value of results is different as the data in each case is different.

**Analysis**

Data analysis is done in a similar manner for the relative frequency distribution histogram, cumulative frequency line plot and numerical descriptive statistics. The results add more meaning and significance to the analysis conducted on the ‘percentage of household with pets’. As dog and cats are peta animals. Therefore, combining the results obtained would give a more accurate understanding and better interpretation of the entire data.

**Conclusion**

The relative frequency distribution histogram in case of ‘percentage of cat owners’ and ‘percentage of dog owners’ have a comparatively normal distribution with respect to the one presented in ‘percentage of household with pets’. The distribution has a small skewness towards the right for ‘percentage of dog owners’ while minor left-skewness in ‘percentage of cat owners. Percentage of cat owners has a bigger range than percentage of dog owners.

In all the three scenario, District of Columbia had data points that were beyond the upper and lower fence values, basically representing an outlier.

The state of Maine and Vermont are identified as outliers while conducting numerical descriptive statistics on percentage of cat owners with the values exceeding the upper fence limit of 41in addition to the outlier of ‘District of Columbia’.

**PART VI**

**Introduction**

Scatterplots is a graphical representation of data to establish a relationship between two variables, namely, explanatory variable and response variable. The scatterplot gives us information about how a variable (response variable) is influenced with the change in another variable (explanatory variable).

**Analysis**

For this scatterplot, the two variables under consideration are the ‘mean number of dogs per household’ and the ‘mean number of cats’. The ‘mean number of dogs’ is assumed to be the explanatory variable, plotted along the horizontal axis while the ‘mean number of cats’ is plotted along the vertical axis, and considered to be the response variable. The values of each variable for the state are plotted in this graph to indicate the overall trend and relationship between the two variables.

**Conclusion**

From the scatterplot, it can be observed that most of the points follow the linear pattern with acceptable variation while few points are well beyond the limits. The graph shows a proportionally positive relationship between the ‘mean number of dogs per household’ and the ‘mean number of cats’. Basically, it means that an increase or decrease in mean number of dogs would lead to a proportional increase or decrease in mean number of cats. This holds true for most of the data points, i.e. majority of the states.

**Reference:**

Albright, D. (2017, December 15). How to calculate basic statistics in Excel. Retrieved November 2, 2018 from <https://www.makeuseof.com/tag/excel-basic-statistics/>

Taylor, C. (2018, March 2). Descriptive vs. Inferential Statistics. Retrieved November 4, 2018 from <https://www.thoughtco.com/differences-in-descriptive-and-inferential-statistics-3126224>