**Notes – Ch 2 Descriptive Statistics: Tabular and Graphical Representation**

**Summarizing Qualitative Data**

**Frequency distribution -** A tabular summary of data showing the number (frequency) of data values in each of several nonoverlapping classes

**Relative Frequency distribution** - A tabular summary of data showing the fraction or proportion of data values in each of several nonoverlapping classes.

**Percent frequency distribution** – A tabular summary of data showing the percentage of data values in each of several nonoverlapping classes.

**Bar Graph or Bar chart** - A graphical device for depicting qualitative data that have been summarized in a frequency, relative frequency, or percent frequency distribution. On one axis of the graph (usually the horizontal axis), we specify the labels that are used for the classes (categories). A frequency, relative frequency, or percent frequency scale can be used for the other axis of the chart (usually the vertical axis). Then, using a bar of fixed width drawn above each class label, we extend the length of the bar until we reach the frequency, relative frequency, or percent frequency of the class. For qualitative data, the bars should be separated to emphasize the fact that each class is separate.

In quality control applications, bar charts are used to identify the most important causes of problems. When the bars are arranged in descending order of height from left to right with the most frequently occurring cause appearing first, the bar chart is called a **pareto diagram**. This diagram is named for its founder, Vilfredo Pareto, an Italian economist.

**Pie chart** - A graphical device for presenting data summaries based on subdivision of a circle into sectors that correspond to the relative frequency for each class. To construct a pie chart, we first draw a circle to represent all the data. Then we use the relative frequencies to subdivide the circle into sectors, or parts, that correspond to the relative frequency for each class.

Pie Chart angle is given by = relative frequency \* 360o

**Summarizing Quantitative Data**

**Frequency Distribution -** A frequency distribution is a tabular summary of data showing the number (frequency) of items in each of several nonoverlapping classes. The three steps necessary to define the classes for a frequency distribution with quantitative data are:

1. Determine the number of nonoverlapping classes.

2. Determine the width of each class.

3. Determine the class limits.

**Number of classes -** Classes are formed by specifying ranges that will be used to group the data. As a general guideline, between 5 and 20 classes are used. For a small number of data items, as few as five or six classes may be used to summarize the data. Fora larger number of data items, a larger number of classes is usually required. The goal is to use enough classes to show the variation in the data, but not so many classes that some contain only a few data items.

**Width of the classes -** The second step in constructing a frequency distribution for quantitative data is to choose a width for the classes. The width should be the same for each class. If the classes were unequal and the width of the intervals differed among the classes, then we would have a distribution that is much more difficult to interpret than one with equal intervals. Thus, the choices of the number of classes and the width of classes are not independent decisions. A larger number of classes means a smaller class width, and vice versa. To determine an approximate class width, we begin by identifying the largest and smallest data values. Then, with the desired number of classes specified, we can use the following expression to determine the approximate class width

**Class limits** - Class limits must be chosen so that each data item belongs to one and only one class. The **lower class limit** identifies the smallest possible data value assigned to the class.

The **upper class limit** identifies the largest possible data value assigned to the class.

**Class midpoint**- In some applications, we want to know the midpoints of the classes in a frequency distribution for quantitative data. The class midpoint is the value halfway between the lower and upper class limits.

**Relative Frequency** is the proportion of the observations belonging to a class. With n observations,

The **percent frequency of a class** is the relative frequency multiplied by 100.

**Dot Plot -** One of the simplest graphical summaries of data is a dot plot. A horizontal axis shows the range **f**or the data. Each data value is represented by a dot placed above the axis. Dot plots show the details of thedata and are useful for comparing the distribution of the data for two or more variables.

**Histogram** - A common graphical presentation of quantitative data is a histogram. This graphical summary can be prepared for data previously summarized in either a frequency, relative frequency, or percent frequency distribution. A histogram is constructed by placing the variable of interest on the horizontal axis and the frequency, relative frequency, or percent frequency on the vertical axis. The frequency, relative frequency, or percent frequency of each class is shown by drawing a rectangle whose base is determined by the class limits on the horizontal axis and whose height is the corresponding frequency, relative frequency, or percent frequency. One of the most important uses of a histogram is to provide information about the shape, or form, of a distribution. A histogram is said to be **skewed to the left** if its tail extends farther to the left. A histogram is said to be **skewed to the right** if its tail extends farther to the right. An example of this type of histogram would be for data such as housing prices; a few expensive houses create the skewness in the right tail. In a **symmetric histogram**, the left tail mirrors the shape of the right tail. Histograms for data found in applications are never perfectly symmetric, but the histogram for many applications may be roughly symmetric. Data for SAT scores, heights and weights of people, and so on lead to histograms that are roughly symmetric.

**Cumulative Frequency Distribution** - A variation of the frequency distribution that provides another tabular summary of quantitative data is the cumulative frequency distribution. The cumulative frequency distribution uses the number of classes, class widths, and class limits developed for the frequency distribution. However, rather than showing the frequency of each class, the cumulative frequency distribution shows the number of data items with values less than or equal to the upper class limit of each class.

**Cumulative relative frequency distribution** shows the proportion of data items, and a cumulative percent frequency distribution shows the percentage of data items with values less than or equal to the upper limit of each class. The cumulative relative frequency distribution can be computed either by summing the relative frequencies in the relative frequency distribution or by dividing the cumulative frequencies by the total number of items. The cumulative percent frequencies were again computed by multiplying the relative frequencies by 100. The last entry in a cumulative frequency distribution always equals the total number of observations.

The last entry in a cumulative relative frequency distribution always equals 1.00 and the last entry in a cumulative percent frequency distribution always equals 100.

**Ogive** - A graph of a cumulative distribution, called an ogive, shows data values on the horizontal

axis and either the cumulative frequencies, the cumulative relative frequencies, or the cumulative percent frequencies on the vertical axis. It enables us to see how many observations lie above or below certain values, rather than merely recording the number of items within intervals.

**Stem and Leaf diagram -** The techniques of exploratory data analysis consist of simple arithmetic and easy-to-draw graphs that can be used to summarize data quickly. One technique—referred to as a stem-and leaf display —can be used to show both the rank order and shape of a data set simultaneously. To develop a stem-and-leaf display, we first arrange the leading digits of each data value to the left of a vertical line. To the right of the vertical line, we record the last digit for each

data value. With this organization of the data, sorting the digits on each line into rank order is simple. Although the stem-and-leaf display may appear to offer the same information as a histogram, it has two primary advantages.

1. The stem-and-leaf display is easier to construct by hand.

2. Within a class interval, the stem-and-leaf display provides more information than the histogram because the stem-and-leaf shows the actual data.

Just as a frequency distribution or histogram has no absolute number of classes, neither does

a stem-and-leaf display have an absolute number of rows or stems.

**Leaf unit** - Used when the data is very large and the last few digits are not very significant A single digit is used to define each leaf in a stem and-leaf display. The leaf unit indicates how to multiply the stem-and-leaf numbers in order to approximate the original data. Leaf units may be 100, 10, 1, 0.1, and so on. Although it is not possible to reconstruct the exact data value from this stem-and-leaf display, the convention of using a single digit for each leaf enables stem-and-leaf displays to be constructed for data having a large number of digits. For stem-and-leaf displays where the leaf unit is not shown, the leaf unit is assumed to equal 1.

**Stretched stem and leaf diagram** If we believe that our original stem-and-leaf display condensed the data too much, we can easily stretch the display by using two or more stems for each leading digit. In a stretched stem-and-leaf display, whenever a stem value is stated twice, the first value corresponds to leaf values of 0–4, and the second value corresponds to leaf values of 5–9.

Advantage of stem and leaf: Individual data points can be obtained from the plot

**Cross Tabulation** - A crosstabulation is a tabular summary of data for two variables. The frequency and relative frequency distributions constructed from the margins of a crosstabulation provide information about each of the variables individually, but they do not shed any light on the relationship between the variables. The primary value of a crosstabulation lies in the insight it offers about the relationship between the variables. Converting the entries in a crosstabulation into row percentages or column percentages can provide more insight into the relationship between the two variables.

Advantages: Shows Relationship between two variables that is not readily apparent

Useful in reviewing customer feedback, market research or survey responses Crosstabulation is widely used for examining the relationship between two variables. In practice, the final reports for many statistical studies include a large number of crosstabulation tables.

Crosstabulations can also be developed when both variables are qualitative and when both

variables are quantitative. When quantitative variables are used, however, we must first create

classes for the values of the variable.

**Simpsons Paradox** - The data in two or more crosstabulations are often combined or aggregated to produce a summary crosstabulation showing how two variables are related. In such cases, we must be careful in drawing a conclusion because a conclusion based upon aggregate data can be reversed

if we look at the unaggregated data. The reversal of conclusions based on aggregate and unaggregated data is called Simpson’s paradox.

**Uses:** Certain analysis with the individual value is different from the analysis of the aggregates Because of the possibility of Simpson’s paradox, realize that the conclusion or interpretation may be reversed depending upon whether you are viewing unaggregated or aggregate crosstabulation data. Before drawing a conclusion, you may want to investigate whether the aggregate or unaggregated form of the crosstabulation provides the better insight and conclusion. Especially when the crosstabulation involves aggregated data, you should investigate

whether a hidden variable could affect the results such that separate or unaggregated crosstabulations provide a different and possibly better insight and conclusion.

Example of Simpson’s Paradox: You and a friend each do problems, and your friend answers a higher proportion correctly than you on each of two days. Does that mean your friend has answered a higher proportion correctly than you when the two days are combined?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Day** | **You** | | | **Friend** | | |
| **Correct Ans** | **Total Ques** | **%** | **Correct Ans** | **Total Ques** | **%** |
| **Sat** | 7 | 8 | 87.5% | 2 | 2 | 100% |
| **Sun** | 1 | 2 | 50% | 5 | 8 | 62.5% |
| **Total** | 8 | 10 | 80% | 7 | 10 | 70% |

**Scatter diagram -** A Scatter diagram is a graphical presentation of the relationship between two quantitative variables. One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.

**Trendline -** A trendline is a line that provides an approximation of the relationship.

The positive relationship exists when y tends to increase as x increases. The negative relationship exists when y tends to decrease as x increases No relationship exists when y remains same as x increases

Disadvantage of scatter line

The correlation may be due to chance particularly when the data pertain to a small sample. A small sample bivariate series may show the relationship but such a relationship may not exist in the universe. It is possible that both the variables are influenced by one or more other variables. For example, expenditure on food and entertainment for a given number of households show a positive relationship because both have increased over time. But, this is due to rise in family incomes over the same period. In other words, the two variables have been influenced by another variable - increase in family incomes