

Summarizing and Graphing Data

Butron 2017

Characteristics of Data

1. **Center** – a representative or average that indicates where the middle of the data set is located
2. **Variation** – a measure of the amount that the data values vary
3. **Distribution** – the nature or shape of the spread of the data over the range of values (such as bell shaped, uniform or skewed)
4. **Outliers** – sample values that lie very far away from the vast majority of the other sample values
5. **Time** – changing characteristics of the data over time

Frequency Distributions

- Organize and summarize data using a table
- Shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number of data values in each of the categories

Frequency Distribution

- Lower class limit – are the smallest numbers that can belong to the different classes
- Upper class limit – are the largest numbers that can belong to the different classes
- Class boundaries – are the numbers used to separate the classes, but without the gaps created by class limits.
- Class midpoints – are the values in the middle of the classes
- Class width – is the difference between two consecutive lower class limits or two consecutive lower class boundaries in a frequency distribution

Why?

- Large data set can be summarized
- We can analyze the nature of data
- We have a basis for constructing graph

Procedure for Constructing a Frequency Distribution

1. Determine the number of classes. It should be between 5 to 20.
2. Calculate the class width. $\square (\text{max.data value} - \text{min.data value})/\text{\#classes}$
3. Choose either the minimum data value or a convenient value below the minimum data value as the first lower class limit.
4. Using the first lower class limit and the class width, list the other lower class limits.
5. List the lower class limits in a vertical column and then enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.

Distribution

1. Relative Frequency Distribution

relative frequency = class frequency / sum of all frequency

percentage frequency = (class frequency / sum of all frequency) x 100

The sum of the relative frequencies in a relative frequency distribution must be close to 1 (or 100%)

Cumulative Frequency Distribution – for a class is the sum of the frequencies for that class and all previous classes.

Visual Presentation

- Histogram is a graph consisting of bars of equal width drawn adjacent to each other (without gaps). The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights the bar correspond to the frequency values
- Frequency Polygons uses line segments connected to points located directly above class midpoint values.
- Relative frequency polygon is a variation of the basic frequency polygon that uses relative frequencies for the vertical scale

Ogive

- Involves cumulative frequencies
- Useful for determining the number of values below some particular value
- Line graph that depicts cumulative frequencies
- An ogive use class boundaries along the horizontal scale and cumulative frequencies along the vertical scale

Dotplot

- A dotplot consists of a graph in which each data value is plotted as a point (or dot) along a scale of values. Dots representing equal values are stacked.

Stemplots

- A stemplot (or stem-and-leaf plot) represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit)

Bar Graphs

- Bar graph uses bars of equal width to show frequencies of categories of qualitative data. The vertical scale represents frequencies or relative frequencies. The horizontal scale identifies the different categories of qualitative data.
- Multiple bar graph has two or more sets of bars, and is used to compare two or more data sets.

Pareto Charts

- A pareto chart is bar graph for qualitative data, with added stipulation that the bar are arranged in descending order according to frequencies. The vertical scale in a Pareto chart represents frequencies. The horizontal scale identifies the different categories of qualitative data.

Pie Charts

- Pie chart is a graph that depicts qualitative data as slices of a circle, in which the size of each slice is proportional to the frequency count for the category

Scatterplots

- A scatterplot (or scatter diagram) is a plot of paired (x,y) quantitative data with a horizontal x-axis and a vertical y-axis. The horizontal axis is used for the first (x) variable, and the vertical axis is used for the second variable.

Time-Series Graph

- A time-series graph is a graph of time-series data, which are quantitative data that have been collected at different points in time.

Statistics for Describing, Exploring and Comparing Data

Review

- Descriptive statistics – summarize or describe relevant characteristics of data
- Inferential statistics – used to make inferences or generalizations, about a population

Basic Concepts of Measures of Center

- Measure of center – is a value at the center or middle of a data set
 1. Mean – is generally the most important of all numerical measurements used to describe data, and it is what most people call an average
 2. Median – of a data set is the measure of center that is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude.
 3. Mode – of a data set is the value that occurs with the greatest frequency
 4. Midrange – of a data set is the measure of center that is the value midway between the maximum and minimum values in the original data set.

Mean from a Frequency Distribution

- Mean from frequency distribution

Skewness

- A distribution of data is skewed if it is not symmetric and extends more to one side than to the other
 1. Skewed to the left (negatively skewed) have a longer left tail, and the mean and median are to the left of the mode.
 2. Skewed to the right (positively skewed) have a longer right tail, and the mean and median are to the right of the mode

The mean and median cannot always be used to identify the shape of the distribution.

Measures of Variation

1. Range – of a set of data values is the different between the maximum data value and the minimum data value
$$\text{range} = (\text{maximum data value}) - (\text{minimum data value})$$
2. Standard Deviation of a Sample – of a set of sample values, denote by s , is a measure of variation of values about the mean. It is a type of average deviation of values from the mean that is calculated by using the previous formula.

Standard Deviation

1. The standard deviation is a measure of variation of all values from the mean.
2. The value of the SD is usually positive. It is zero only when all of the data values are the same number. (It is never negative). Also, larger values of s indicate greater amounts of variation.
3. The value of the SD can increase dramatically with the inclusion of one or more outliers (data values that are very far away from all of the others)
4. The units of the SD (such as minutes, feet, pounds, and so on) are the same as the units of the original data values.