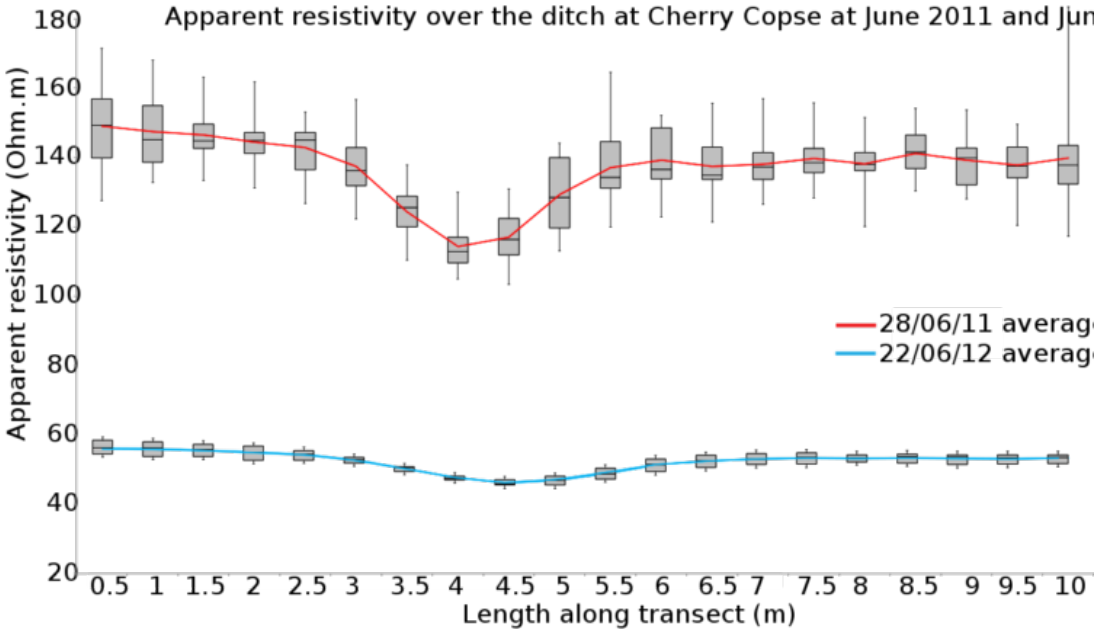
Statistics: Summarizing Quantitative Variables

See online here

This article explains the graphical displays, including quantitative variables, histogram, stem and leaf plot and shape of a distribution. The modes discussed in this article are one mode, two modes and more than two modes. The article also highlights the difference between a histogram and stem and leaf plot for the presentation of distribution in a graph. The spread of distribution includes a range i.e. interquartile range.



Quantitative Variables

Variables are values which can be changed situation to situation in data. Quantitative variables provide a measurable value which is countable, e.g. we have taken a herd as data which means we have considered the number of sheep. The number of sheep in the herd is a measurable quantity, hence the variable is quantitative. Normally, three kinds of graphs are used for statistical representation i.e:

- Histogram
- Stem and leaf plot
- Box plot

Others graphs include: frequency polygons, bar charts, line graphs, scatter plots and dot plots.

Some graph types such as stem and leaf displays are best-suited for small to moderate amounts of data, whereas others such as histograms are best-suited for large amounts of

data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

Histogram

A histogram is a graphical display of statistical data by using **rectangular bars** for the presentation of the frequency of data items. A histogram slices up the quantitative variables with an equal width interval in a single rectangular bar or bin. The number of values has to be counted which fall in a single bar or bin.

There are two types of variables in a histogram; one is dependent and the other is independent. Dependent variables are plotted on a vertical axis, and independent variable on a horizontal axis of the graph.

Example: To make a histogram from continuous variables, there is a need to first divide the whole data into intervals. Each interval is known as a bin. The data taken for the example comprises of the following age numbers:

- The first step is to arrange data in a chronological sequence.
- The second step involves a selection of bin width. In this example, the bin width is 10 years
- A histogram tells us the variable quantities which fall in a single bin.
- After the age numbers are arranged in chronological form, and the bin width is set, the frequency of each bin can be measured as given in the table below.

In this example, the frequency is a dependent variable, whereas age is an independent variable which does not change in any case.

Histograms can be based on relative frequencies instead of actual frequencies.

Relative frequency histogram

This type of histogram gives a graphical representation of the occurrence of data in percentage form. A relative frequency histogram adds a further column of frequency next to the frequency column. The percentage of each frequency is calculated by dividing the frequency of each bin by the total frequency.

Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions).

You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

Bin	Frequency	Percentage of each frequency
20-30	2	10%
30-40	4	20%
40-50	4	20%
50-60	5	25%
60-70	3	15%
70-80	1	5%
80-90	0	0%

90-100	1	5%

The percentage can be calculated as follows:

Percentage of each frequency = frequency number/sum of frequency ×100

F1= 2/20 ×100

= 10%

F2=4/20 ×100

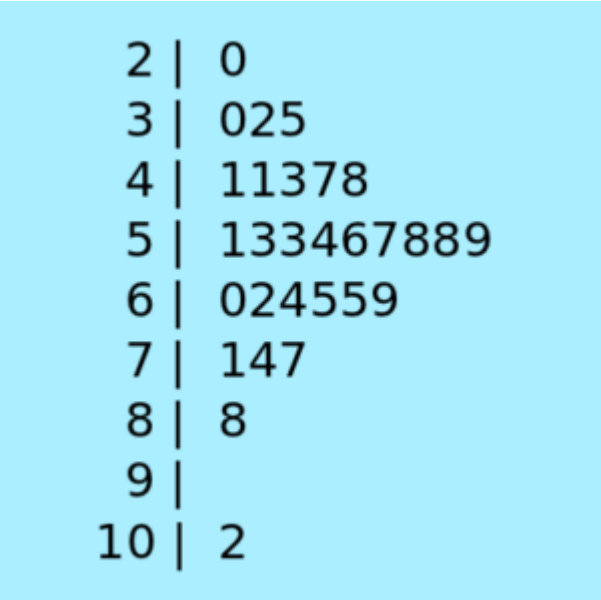
= 20% and so on

Stem and Leaf Plot

Stem and leaf plot is the simplest method of statistical distribution. In this model of presentation of data in a graphical form, the stem is known for a bunch or bin having so many more numbers as subset. Here, a range is considered as stem e.g. 1-10 and subset numbers i.e. 2,3,4,5,6,7,8 and 9 are taken as a leaf.

The difference between histogram and stem and leaf plot

Stem and leaf plots show individual values, whereas a histogram shows the frequency of summarized data falling a single bin. Stem and leaf plot can be turned into a histogram by arranging data in different frequencies. In another way, data in raw form, as originally assembled, can be displayed without any graphical presentation. The raw presentation of stem and leaf plot is shown as follows:



Advantages

- It is an easy approach to display data in a graphical form.

- It provides an easy to understand the distribution of lengthy quantitative data.

Disadvantages

- In the case of lengthy quantitative data, the stem and plot model of statistical distribution becomes burdensome to handle properly.

Frequency polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

Characteristics of a Distribution

There are several factors which are required to be observed for the distribution of quantitative variables of data. These factors are as follows:

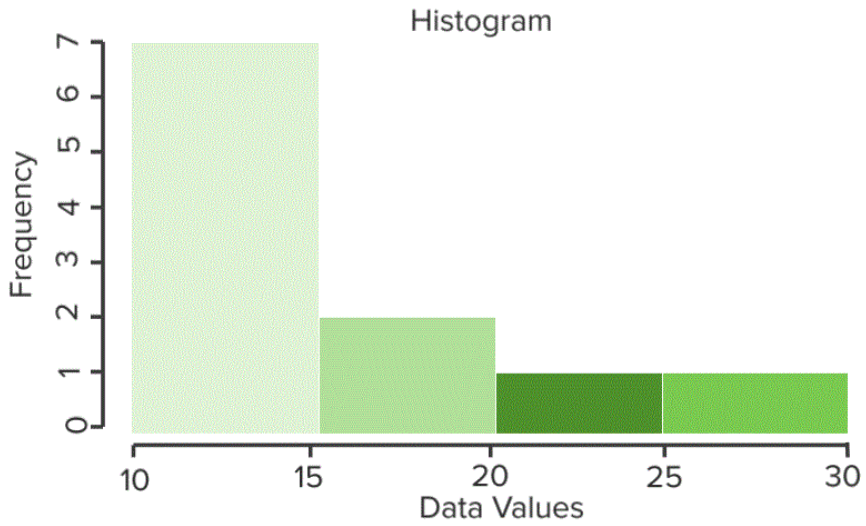
Shape of the distribution

It depicts the tendency of data to skew, its number of peaks, its uniformity and the possession of symmetry involved.

Symmetric (U shaped): This shape of distribution forms a self-mirror image when it reflects in its vertical center line like below graph.

Skewed right (positively skewed): This shape of distribution refers to a large cluster of data which is distributed towards the right side. The tail of this kind of graph is skewed towards the right side of the graph showing high positive numbers. In this kind of data distribution, the mean is greater in the median.

Skewed left (negatively skewed): In this shape, the “tail” of the graph is directed towards small or negative numbers or towards the left side. When the mean of the data gets pulled towards the left side, it becomes less than the median value.



“Right-Skewed Distribution” Image created by Lecturio

Modes

It is the peak value of quantitative variable distribution through a histogram. The histogram has three kinds of peaks or modes.

- **One mode**: When there is a single peak in a histogram, it is known as a unimodal histogram.
- **Two modes**: In case where there is more than one peak in a histogram, it is known as a bimodal histogram.
- **More than two modes**: There is a possibility that more than two peaks are there is a histogram; it is termed as a multimodal histogram.

The Center of a Distribution (Finding Middle Ground)

The second well-known characteristic required for a distribution is to find out the center of a distribution.

Center

Suppose in a data 1,2,3,4,5,6,7, 4 is the middle or center of a distribution. In the case of odd numbers or small data, the isolated number left in the middle after the pairing of data is considered as the middle of the distribution. In the case of large population with even numbers of distribution of quantitative variables, mean or average of data set can be used to find out the middle of data. Another way to find of the center of distribution is through finding out median.

Median

The **median of data** can be calculated through the following steps:

1. Put a complete set of given data in a sequence from lowest to highest. Once arranged, the middle number can be termed as median or the data.
2. If data is comprised of “n” number of variables with odd number, median can be calculated by using formula (n+1)/2.
3. In case the number of values in a data “n” is even, then the median can be

calculated by taking the average of two values which are calculated through n/2 and (n+1)/2.

Example: Suppose the data comprise of 12, 16,17,17,18,19,10,21,21,22,22,22,22,22,23,23,24,24,25,25,25,26,27,28,29.

There are 25 observations. The **correct sequence** of data is as follows:

10,16,17,17,18,19, 21,21,22,22,22,22,22,23,23,24,24,25,25,25,26,27,28,29.

After arranging data in sequence, the value in median position by using formula comes out.

Median = (25+1)/2

= 13

Now, let's check which data comes **in the 13th position** of arranged data i.e. 22.

So the median is "22". Median is a point where 50% of the data on the left side is higher and 50% of the data on the other side (the right side) is lower than the median value.

Spread of a Distribution

The third characteristic which has to be considered while observing the distribution of a quantitative variable data set is spread. It shows how spread out a distribution of data is, or how it varies in its occurrence. The spread of distribution comprises of a range of the data set.

1. Range

The difference between maximum and minimum values of a data set is known as a range. i.e.

Range: maximum value – minimum value
Suppose a data set comprises of the following values:

23 56 45 65 59 55 62 54 85 25

In the data, the highest value is "85", whereas the lowest or minimum value is "23". By using the formula, the range of the data set is:

Range = 85 – 23

= 62

Problem: The spread of distribution can only be measured by using the range method when the data value has limitations or boundaries. If a variable has a critically low or high data limit, in that case spread can be measured by using range. Range can detect errors only while data is entered. If there are extreme values in data, it makes range very large and critical to be calculated. In case "29" value is mistakenly entered as "299", the range can critically deviate i.e.

Range = 29 – 17 = 17

Range = 299 – 17 = 287 (**mistaken range**)

2. Interquartile range (focusing on the middle of the data)

This range comprises of 3 quartiles or division of data set in order to measure the variability of the dataset. This kind of range has the ability to ignore extreme values and describe the spread of distribution in a clearer way. The step to measure interquartile range comprises of:

1. Firstly, give sequence to a data properly from lowest to highest values.
2. Now divide properly ordered data into half to find out the median value.
3. Now further calculate the median of divided data i.e. median of lower half termed as “lower quartile” and then calculate the median of the data set in the upper half known as the **“upper quartile.”**
4. We can find out the interquartile range by using the following formula:

Interquartile range = upper quartile – lower quartile
Example: If data set comprises of:
12,16,17,17,18,19,20,21,21,22,22,22, 22,23,23,24,24,25,25,25,26,27,28,29.

The lower half of data comprises of 12,16,17,17,18,19, 20,21,21,22,22,22.

As 12 observations are given, in this case, median comes out an average of 6^{th} and 7^{th} value. The lower quartile can be calculated by using the above-given instructions as follows:

Lower quartile = 1/2 (19+20) = 19.5

The upper half of data set comprises of 22,23,23,24,24,25,25,25,26,27,28,29 i.e. 12 observations. Again the median will be the average of 6^{th} and 7^{th} value.

Upper quartile = ½ (25+25) = 25

Hence, using the values of upper and lower quartile, the interquartile range can be measured as follows:

Interquartile range = 25 – 19.5

= 5.5

3. The five number summary

The five number summary of center and spread distribution comprises of:

- Minimum
- First Quartile
- Median
- Third Quartile
- Maximum

4. Box plots

A box plot is a type of interquartile range. A data set can be presented graphically using box plots. A five number summary can be displayed with a box plot model by using the following steps:

1. Firstly, the entire range of data should be drawn on a vertical axis of a graph.

2. Secondly, horizontal lines will be drawn at the lower quartile, the upper quartile and the median position of data. The upper and lower quartiles will be connected with vertical lines to draw a box. The line connecting the median will be drawn inside the box.
3. Now, in the third step, draw fences around the data. There will be two fences:
    - Upper Fence = Q3 + 1.5× IQR (Interquartile range)
    - Lower Fence = Q1 + 1.5× IQR (Interquartile range)

    Note: Fences will be excluded from boxplot.
4. Now whiskers will be grown out of the box plot. One whisker will connect the upper edge and the other one with the lower edge of the box connecting the maximum and minimum values in the range.
5. A whisker will be connected to the horizontal line marking the lower fence in case the data point falls below the lower fence.
6. In case data point falls above the upper fence, the upper whisker will be connected to the horizontal line which marks the upper fence.

If we calculate the numerical value IQR, keeping in view the five number summary for the box plot, we have no extreme values.

Suppose minimum value = 12

Q1 = 19.5

Median = 22

Q3 = 25

Maximum value = 29

Interquartile range = 25 – 19.5

= 5.5

Fences in this example are:

Lower Fence = 19.5 – 1.5(5.5)

= 11.25

Upper fence = 25 + 1.5(5.5)

References

Histograms viastatistics.laerd.com

Visualizing Numerical Data via researchhubs.com

Shapes of Distributions via mathbitsnotebook.com

Statistics How To via statisticshowto.com

Notes