Histograms and Boxplots

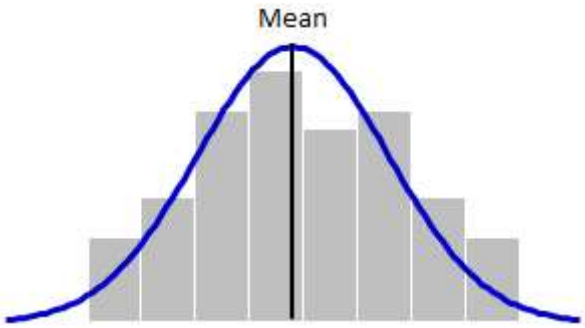
Histograms:

A **histogram** is a graphical representation of a frequency distribution for numeric data. It is a **bar chart** that is often used as the first step to determine the probability distribution of a data set or a sample. It allows to visually and quickly assess the shape of the distribution, the central tendency, the amount of variation in the data, and the presence of gaps, outliers or unusual data points.

A histogram can tell you about the underlying distribution of the data and whether you can apply certain statistical tests to perform potential improvement opportunities. It shows if the variability in the data is within specification limits and if the process is capable or not. It is also used to identify shifts in the process and to verify that the changes you made were a real improvement.

Histograms are normally used to represent moderate to large amount of continuous data and we need at least 25 data points to determine if a histogram follows a particular distribution. If the data size is too small or the measurement system has a low resolution, the histogram may show very few columns and may not accurately display the shape of the distribution. **Dotplots** are preferred over histograms when representing small amount of data and when comparing between multiple distributions.

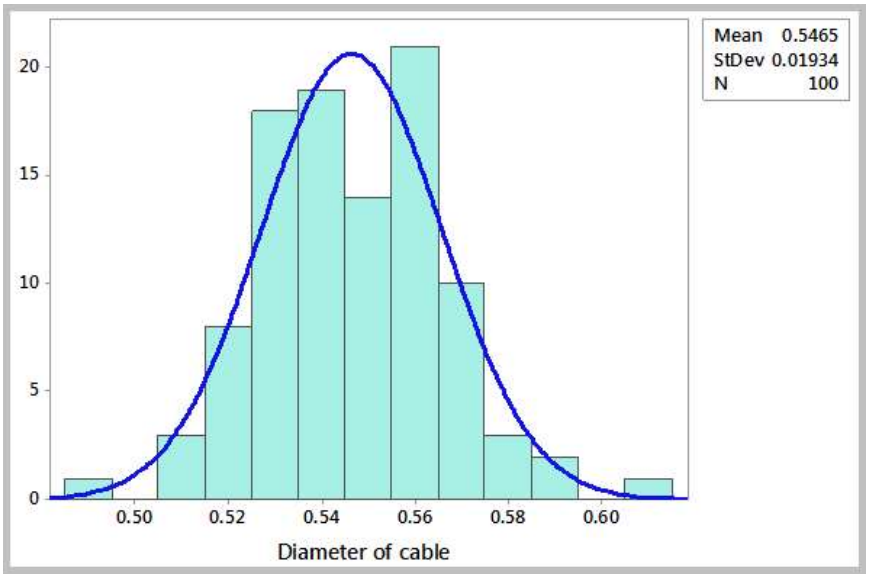
It is always a good practice to plot your data in a histogram after collecting the data. This will give you an insight about the nature of the data, the minimum and maximum values, the shape of the distribution and whether it is normal, exponential, chi-squared, etc. It also tells if the distribution is symmetric or non-symmetric, and whether it is unimodal, bimodal, or multimodal. If the data is symmetrically distributed and centered around the mean, we can say that the data is normally distributed.



To construct a histogram, you first need to split the data into intervals called bins. A bar should be constructed above each bin to represent the frequency of the data values within each interval. The bars should be adjacent with no gaps between them to indicate the continuity of the data. There should be a small gap before the first bar of a histogram. The mean of the data and the specification limits are often indicated on the histogram.

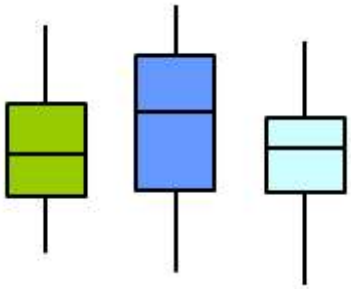
The following is a histogram that represents the distribution of cable diameters in a manufacturing process. The result

should be summarized using day to day language such as: *"The distribution looks symmetric around the cable diameter mean (0.546 cm) and appears to fit the Normal Distribution fairly well"*.



The above chart shows the results of a data set that belongs to Minitab Inc.

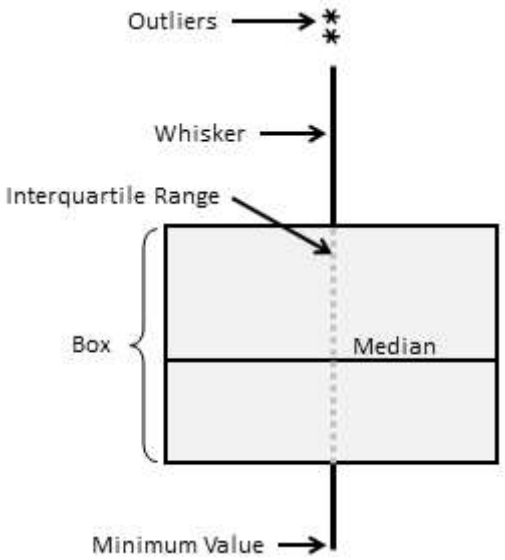
Box plots:



A **Boxplot** is a graphical way that summarizes the important aspects of the distribution of continuous data. It is particularly useful when comparing between several groups of data sets or samples. Like histograms, they should be used for moderate to large amount of data as the size of the boxplot can vary significantly if the data size is too small. However, they are less detailed and take up less space which allows easy comparison of multiple data sets.

Boxplots are primarily used when comparing several distributions against each other. They summarize key statistics from the data and display them in a **box-and-whiskers format**. They provide a quick way for examining the variation

present in the data. A wider range boxplot indicates more variability. Boxplots are also used to check if there is a significant difference in the process after implementing a process improvement initiative.

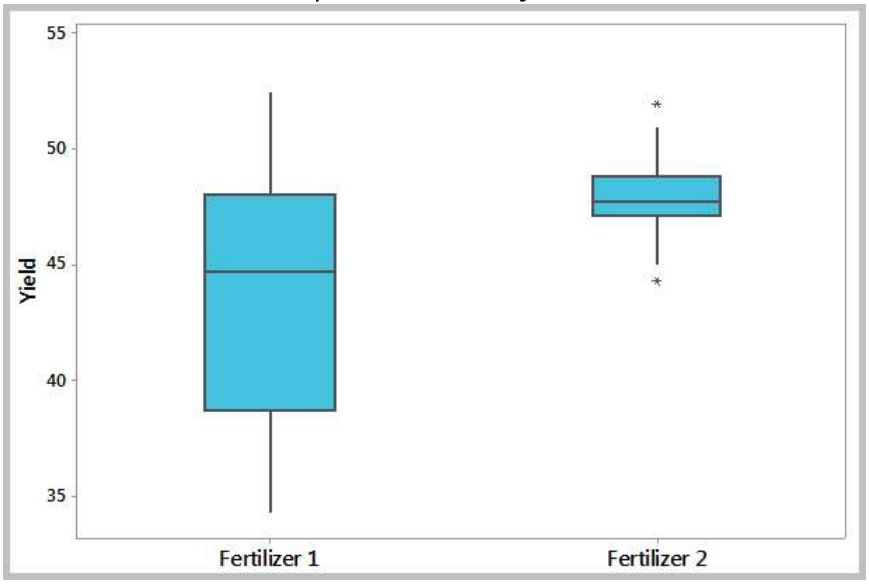


Boxplots can tell us whether the distribution is symmetrical or skewed and if there are outliers in the data. The spacings between the different parts of a boxplot indicate the spread and skewness present in the data. The data is plotted in a way that the middle 50% of the data points fits inside the box, the bottom 25% of the data points located below the box and the top 25% of the data points located above the box. Each whisker may extends up to 1.5 times the length of the box.

The middle line of the box is the median of the data points. Some boxplots also display the mean of the data points with an additional character. Any data beyond the whiskers are considered outliers and are plotted as asterisks (*). Outliers often reflect errors in data recording or data entry, and if the values are real you should investigate what was going on in the process at the time.

The following are boxplots that display the yield of a crop after applying two different fertilizers. Fertilizer 2 appears to have a higher yield than Fertilizer 1. What other comments would you make about the below boxplots? Think about the

variation as well as the presence of any unusual values.



The above chart shows the results of a data set that belongs to Minitab Inc.

Further Information:

- Histograms are sometimes called **Frequency Plots** while boxplots are referred to as **Box-and-Whisker Plots**.
- Histograms and boxplots can be drawn either vertically or horizontally. There are many graphical tools that can generate histograms and boxplots quickly and easily (such as Minitab).
- Histograms are often confused with bar charts. A histogram is normally used for continuous data while a bar chart is a plot of count data.
- Although histograms are efficient graphical methods for describing the distribution of the data, they can't see changes and trends over time.
- **Individual Value Plots** are preferred over boxplots when representing small amount of data.