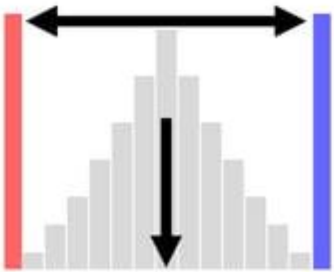
Continuous Improvement Toolkit

Descriptive Statistics



The Continuous Improvement Map

- Descriptive Statistics

- **Statistics** is concerned with the describing, interpretation and analyzing of data.
- It is, therefore, an essential element in any improvement process.
- Statistics is often categorized into **descriptive** and **inferential** statistics.
- It uses **analytical methods** which provide the math to model and predict variation.
- It uses **graphical methods** to help making numbers visible for communication purposes.

- Descriptive Statistics

Why do we Need Statistics?

- To find why a process behaves the way it does.
- To find why it produces defective goods or services.
- To center our processes on ‘Target’ or ‘Nominal’.
- To check the accuracy and precision of the process.
- To prevent problems caused by assignable causes of variation.
- To reduce variability and improve process capability.
- To know the truth about the real world.

- Descriptive Statistics

Descriptive Statistics:

- Methods of describing the characteristics of a data set.
- Useful because they allow you to make sense of the data.
- Helps exploring and making conclusions about the data in order to make rational decisions.
- Includes calculating things such as the average of the data, its spread and the shape it produces.

- Descriptive Statistics

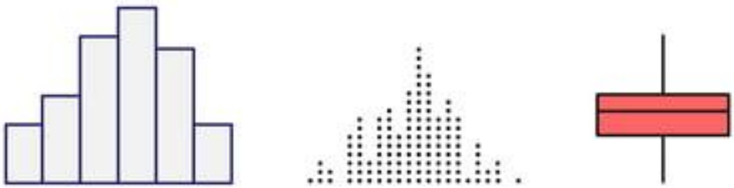
- For example, we may be concerned about **describing**:
  - The weight of a product in a production line.
  - The time taken to process an application.

- Descriptive Statistics

- Descriptive statistics involves describing, summarizing and organizing the data so it can be easily understood.
- **Graphical displays** are often used along with the quantitative measures to enable clarity of communication.

- Descriptive Statistics

- When analyzing a graphical display, you can draw conclusions based on several characteristics of the graph.
- **You may ask questions such ask:**
  - Where is the approximate middle, or center, of the graph?
  - How spread out are the data values on the graph?
  - What is the overall shape of the graph?
  - Does it have any interesting patterns?



- Descriptive Statistics

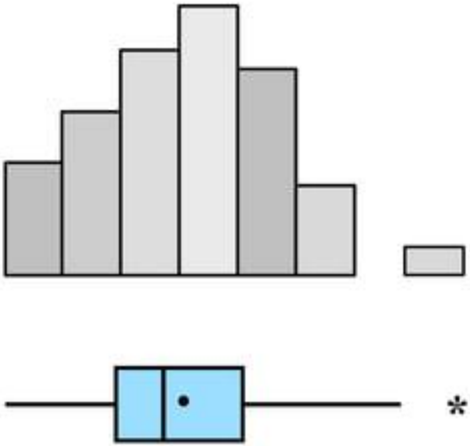
Outlier:

- A data point that is significantly greater or smaller than other data points in a data set.
- It is useful when analyzing data to identify outliers
- They may affect the calculation of descriptive statistics.
- Outliers can occur in any given data set and in any distribution.

- Descriptive Statistics

Outlier:

- The easiest way to detect them is by **graphing the data** or using graphical methods such as:
  - Histograms.
  - Boxplots.
  - Normal probability plots.



- Descriptive Statistics

Outlier:

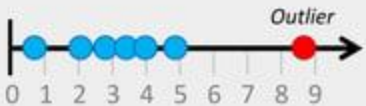
- Outliers may indicate an experimental error or incorrect recording of data.
- They may also occur **by chance**.
  - It may be normal to have high or low data points.
- You need to decide whether to exclude them before carrying out your analysis.
  - An outlier should be excluded if it is due to measurement or human error.

- Descriptive Statistics

Outlier:

- This example is about the time taken to process a sample of applications.

2.8	8.7	0.7	4.9	3.4	2.1	4.0

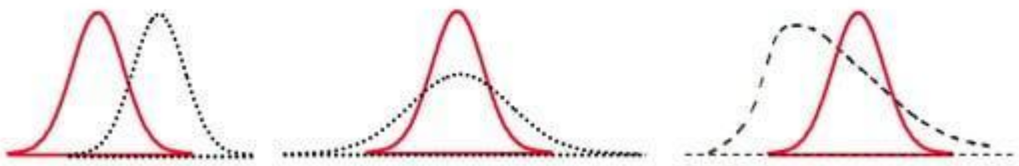


It is clear that one data point is far distant from the rest of the values.
This point is an 'outlier'

- Descriptive Statistics

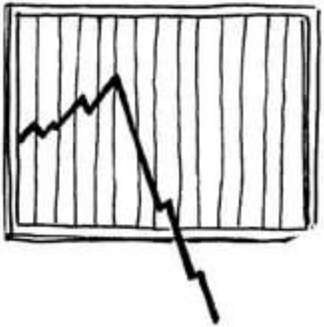
The following measures are used to describe a data set:

- Measures of position (also referred to as central tendency or location measures).
- Measures of spread (also referred to as variability or dispersion measures).
- Measures of shape.



- Descriptive Statistics

- If assignable causes of variation are affecting the process, we will see changes in:
  - Position.
  - Spread.
  - Shape.
  - Any combination of the three.



- Descriptive Statistics

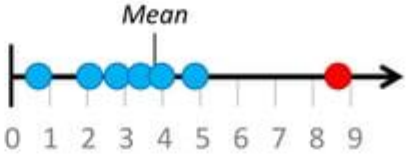
Measures of Position:

- Position Statistics measure the data central tendency.
- Central tendency refers to where the data is centered.
- You may have calculated an average of some kind.
- Despite the common use of average, there are different statistics by which we can describe the average of a data set:
  - Mean.
  - Median.
  - Mode.

- Descriptive Statistics

Mean:

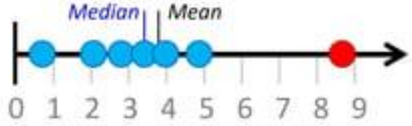
- The total of all the values divided by the size of the data set.
- It is the most commonly used statistic of position.
- It is easy to understand and calculate.
- It works well when the distribution is symmetric and there are no outliers.
- The mean of a sample is denoted by '**x-bar**'.
- The mean of a population is denoted by '**μ**'.



- Descriptive Statistics

Median:

- The middle value where exactly half of the data values are above it and half are below it.
- Less widely used.
- A useful statistic due to its robustness.
- It can reduce the effect of outliers.
- Often used when the data is nonsymmetrical.
- Ensure that the values are ordered before calculation.
- With an even number of values, the median is the mean of the two middle values.



- Descriptive Statistics

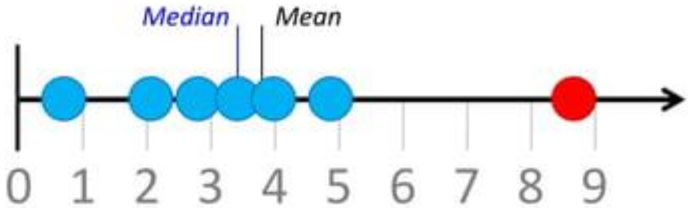
Median Calculation:

23	12
33	30
34	31
36	37
38	38
40	40
41	41
41	41
44	44
	45

Median = 38 + 40 / 2 = 39

- Descriptive Statistics

- Why can the mean and median be different?



- Descriptive Statistics

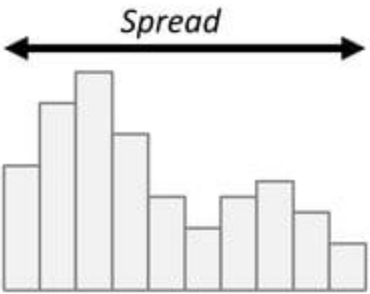
Mode:

- The value that occurs the most often in a data set.
- It is rarely used as a central tendency measure
- It is more useful to distinguish between unimodal and multimodal distributions
  - When data has more than one peak.

- Descriptive Statistics

Measures of Spread:

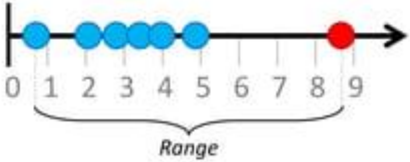
- The **Spread** refers to how the data deviates from the position measure.
- It gives an indication of the amount of variation in the process.
  - An important indicator of quality.
  - Used to control process variability and improve quality.
- All manufacturing and transactional processes are variable to some degree.
- There are different statistics by which we can describe the spread of a data set:
  - Range.
  - Standard deviation.



- Descriptive Statistics

Range:

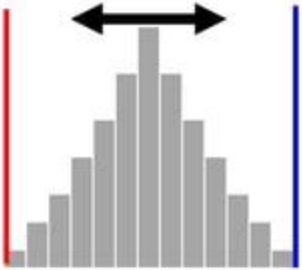
- The difference between the highest and the lowest values.
- The simplest measure of variability.
- Often denoted by '**R**'.
- It is good enough in many practical cases.
- It does not make full use of the available data.
- It can be misleading when the data is skewed or in the presence of outliers.
  - Just one outlier will increase the range dramatically.



- Descriptive Statistics

Standard Deviation:

- The average distance of the data points from their own mean.
- A low standard deviation indicates that the data points are clustered around the mean.
- A large standard deviation indicates that they are widely scattered around the mean.
- The standard deviation of a sample is denoted by ‘**s**’.
- The standard deviation of a population is denoted by “**μ**”.



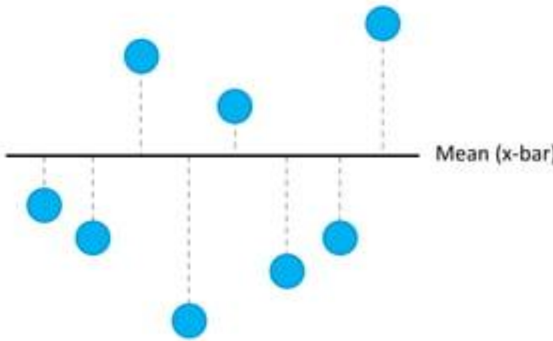
- Descriptive Statistics

Standard Deviation:

- Perceived as difficult to understand because it is not easy to picture what it is.
- It is however a more robust measure of variability.
- Standard deviation is computed as follows:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

s = standard deviation
$\bar{x}$ = mean
x = values of the data set
n = size of the data set



- Descriptive Statistics

Exercise:

- This example is about the time taken to process a sample of applications.
- Find the mean, median, range and standard deviation for the following set of data: 2.8, 8.7, 0.7, 4.9, 3.4, 2.1 & 4.0.

Time allowed: 10 minutes

- Descriptive Statistics

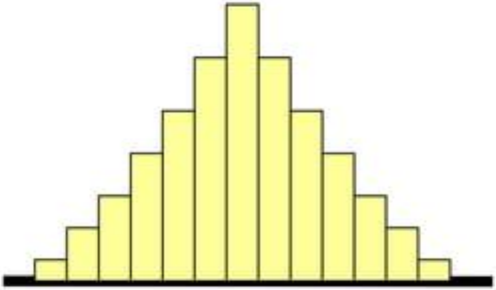
- If someone hands you a sheet of data and asks you to find the mean, median, range and standard deviation, what do you do?

21	19	20	24	23	21	26	23
25	24	19	19	21	19	25	19
23	23	15	22	23	20	14	20
15	19	20	21	17	15	16	19
13	17	19	17	22	20	18	16
17	18	21	21	17	20	21	21
21	17	17	19	21	22	25	20
19	20	24	28	26	26	25	24

- Descriptive Statistics

Measures of Shape:

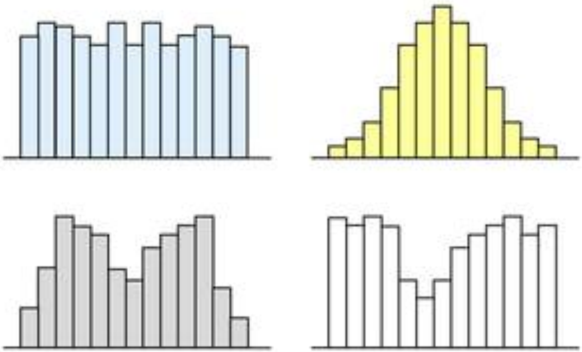
- Data can be plotted into a histogram to have a general idea of its shape, or distribution.
- The shape can reveal a lot of information about the data.
- Data will always follow some know distribution.



- Descriptive Statistics

Measures of Shape:

- It may be symmetrical or nonsymmetrical.
- In a symmetrical distribution, the two sides of the distribution are a mirror image of each other.
- Examples of **symmetrical** distributions include:
  - Uniform.
  - Normal.
  - Camel-back.
  - Bow-tie shaped.



- Descriptive Statistics

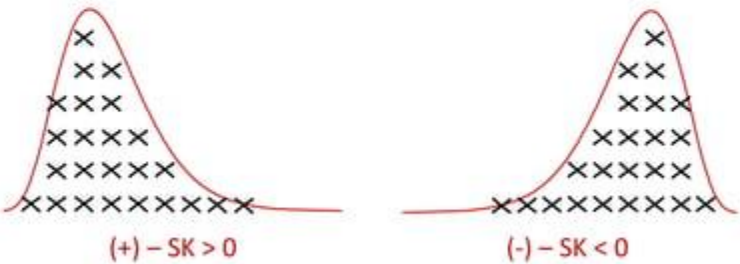
Measures of Shape:

- The shape helps identifying which descriptive statistic is more appropriate to use in a given situation.
- If the data is symmetrical, then we may use the mean or median to measure the central tendency as they are almost equal.
- If the data is skewed, then the median will be a more appropriate to measure the central tendency.
- Two common statistics that measure the shape of the data:
  - Skewness.
  - Kurtosis.

- Descriptive Statistics

Skewness:

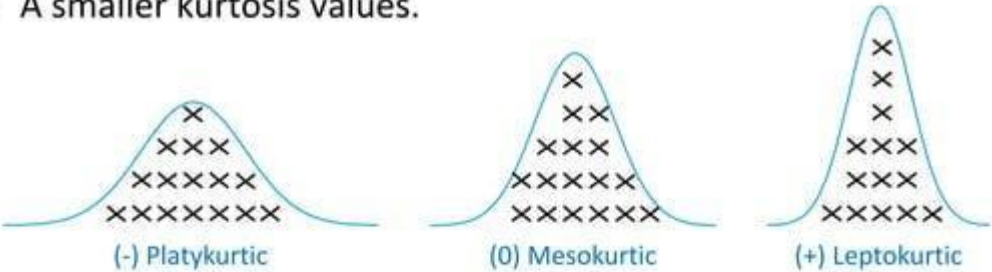
- Describes whether the data is distributed symmetrically around the mean.
- A skewness value of zero indicates perfect symmetry.
- A negative value implies left-skewed data.
- A positive value implies right-skewed data.



- Descriptive Statistics

Kurtosis:

- Measures the degree of **flatness** (or **peakness**) of the shape.
- When the data values are clustered around the middle, then the distribution is more peaked.
  - A greater kurtosis value.
- When the data values are spread around more evenly, then the distribution is more flatted.
  - A smaller kurtosis values.



- Descriptive Statistics

- Skewness and kurtosis statistics can be evaluated visually via a histogram.
- They can also be calculated by hand.
- This is generally unnecessary with modern statistical software (such as Minitab).

- Descriptive Statistics

Further Information:

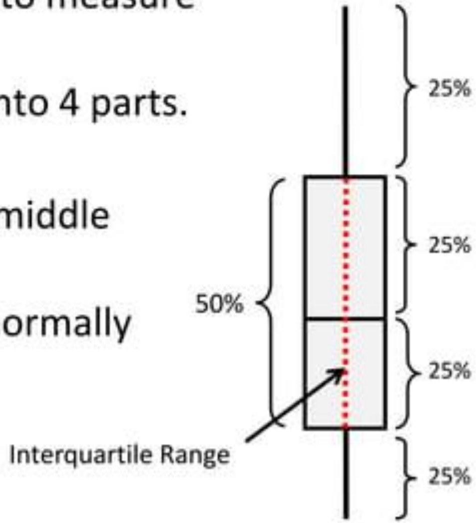
- **Variance** is a measure of the variation around the mean.
- It measures how far a set of data points are spread out from their mean.
- The units are the square of the units used for the original data.
  - For example, a variable measured in meters will have a variance measured in meters squared.
- It is the square of the standard deviation.

Variance = s^2

- Descriptive Statistics

Further Information:

- The **Inter Quartile Range** is also used to measure variability.
- Quartiles divide an ordered data set into 4 parts.
- Each contains 25% of the data.
- The inter quartile range contains the middle 50% of the data (i.e. Q3-Q1).
- It is often used when the data is not normally distributed.



- Descriptive Statistics in Minitab

- **Minitab** is a statistical software that allows you to enter your data to perform a wide range of statistical analyses.
- It can be used to calculate many types of descriptive statistics.
- It tells you a lot about your data in order to make more rational decisions.
- Descriptive statistics summaries in Minitab can be either quantitative or visual.

Descriptive Statistics

- Descriptive Statistics in Minitab

Example:

- A hospital is seeking to detect the presence of high glucose levels in patients at admission.
- You may use the glucose_level_fasting worksheet or use data that you have collected yourself.
- Remember to copy the data from the excel sheet and paste it into Minitab worksheet.

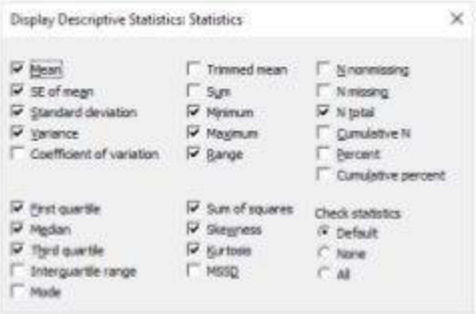
79	72	77	85	76	120	78	94
93	70	79	75	68	73	79	85
98	77	77	88	79	79	70	113
75	80	74	83	85	79	87	82
104	106	81	76	68	72	61	95
78	106	84	70	96	70	90	98
69	60	74	67	71	75	105	79
71	75	131	80	75	52	152	106
81	96						

- Descriptive Statistics in Minitab

Example:

- To create a **quantitative summary** of your data:
  - Select **Stat > Basic Statistics > Display Descriptive Statistics**.
  - Select the variable to be analyzed, in this case ‘glucose level’.
  - Click OK.

- Here is a screenshot of the various descriptive statistics you may choose when doing your analysis.



- Descriptive Statistics in Minitab

Example:

- Here is a screenshot of the example result:

Descriptive Statistics: Glucose level

Variable	Total Count	Mean	SE Mean	StDev	Variance	Sum of Squares	Minimum	Q1	Median
Glucose level	66	83.38	2.06	16.71	279.25	476985.00	52.00	73.75	79.00

Variable	Q3	Maximum	Range	Skewness	Kurtosis
Glucose level	90.75	152.00	100.00	1.63	4.06

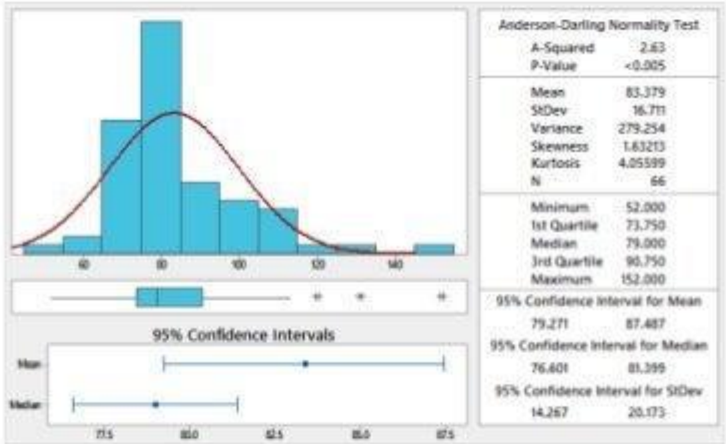
Quantitative Summary

- Descriptive Statistics in Minitab

Example:

- To create a **visual summary** of your data:
  - Select Stat > Basic Statistics > Graphical Summary.
  - Select the variable to be analyzed, in this case ‘glucose level’.
  - Click OK.

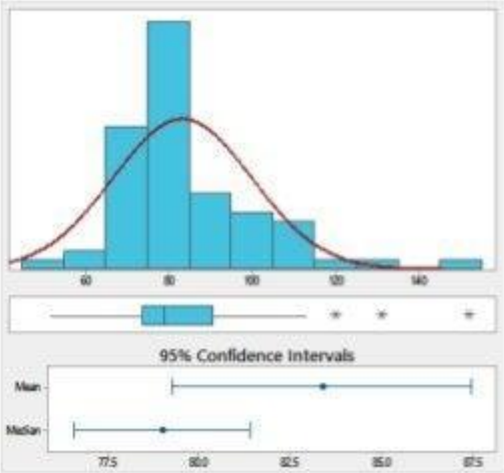
- Here is a screenshot of the example result:



- Descriptive Statistics in Minitab

Example:

- By default, Minitab fits a normal distribution curve to the histogram.
- A boxplot will also be shown to display the four quartiles of the data.
- The 95% confidence intervals are also shown to illustrate where the mean and median of the population lie.



- Descriptive Statistics in Minitab

Example:

- Mean, standard deviation, sample size, and other descriptive statistic values are shown in the adjacent data table.
- The skewed distribution shows the differences that can occur between the mean and median.
- The mean is pulled to the right by the high value outliers.
- The positive value for skewness indicates a positive skew of the data set.

Anderson-Darling Normality Test	
A-Squared	2.63
P-Value	<0.005
Mean	83.379
StDev	16.711
Variance	279.254
Skewness	1.63213
Kurtosis	4.05599
N	66
Minimum	52.000
1st Quartile	73.750
Median	79.000
3rd Quartile	90.750
Maximum	152.000
95% Confidence Interval for Mean	
79.271	87.487
95% Confidence Interval for Median	
76.601	81.399
95% Confidence Interval for StDev	
14.267	20.173