Unit III

Measure of Variation

Measure of variation is the way to extract meaningful information from a set of provided data. Variability provides a lot of information about the data and some of the information it provides is mentioned below:

- It shows how far data items lie from each other.
- It shows the distance from the centre of the distribution.
- It measures the central tendency of the data.
- It also provides a descriptive analysis of the picture.

Variance and Standard Deviation

Variance and Standard Deviation are the two important measurements in statistics. Variance is a measure of how data points vary from the mean, whereas standard deviation is the measure of the distribution of statistical data. The basic difference between both is standard deviation is represented in the same units as the mean of data, while the variance is represented in squared units.

Variance

According to layman's words, the variance is a measure of how far a set of data are dispersed out from their mean or average value. It is denoted as 'σ^2'.

Properties of Variance

- It is always non-negative since each term in the variance sum is squared and therefore the result is either positive or zero.
- Variance always has squared units. For example, the variance of a set of weights estimated in kilograms will be given in kg squared. Since the population variance is squared, we cannot compare it directly with the mean or the data themselves.

Standard Deviation

The spread of statistical data is measured by the standard deviation. Distribution measures the deviation of data from its mean or average position. The degree of dispersion is computed by the method of estimating the deviation of data points. It is denoted by the symbol, 'σ'.

Properties of Standard Deviation

- It describes the square root of the mean of the squares of all values in a data set and is also called the root-mean-square deviation.
- The smallest value of the standard deviation is 0 since it cannot be negative.
- When the data values of a group are similar, then the standard deviation will be very low or close to zero. But when the data values vary with each other, then the standard variation is high or far from zero.

Variance and Standard deviation Relationship

Variance is equal to the average squared deviations from the mean, while standard deviation is the number's square root. Also, the standard deviation is a square root of variance. Both measures exhibit

variability in distribution, but their units vary: Standard deviation is expressed in the same units as the original values, whereas the variance is expressed in squared units.

Quartiles

- The data is divided into four equal parts, or quarters, by quartiles. This data is divided in ascending order, with two lower quartiles and two upper quartiles.
- The first quartile, second quartile, third quartile, and fourth quartile are used by statisticians to divide their data into percentages: the lowest and second-lowest 25%, and the highest and second-highest 25%, which are referred to as the first, second, third, and fourth quartiles, respectively.
- The quartiles are represented by the symbols Q1, Q2, Q3 & Q4.

Interquartiles Range

The midpoint of your data distribution, or the middle of your four quartiles, is referred to as the interquartile range (IQR), which is in the middle of the lower and upper quartiles. The IQR is a measurement of how evenly the data is distributed around the average.

The formula for Interquartile Range is given below:

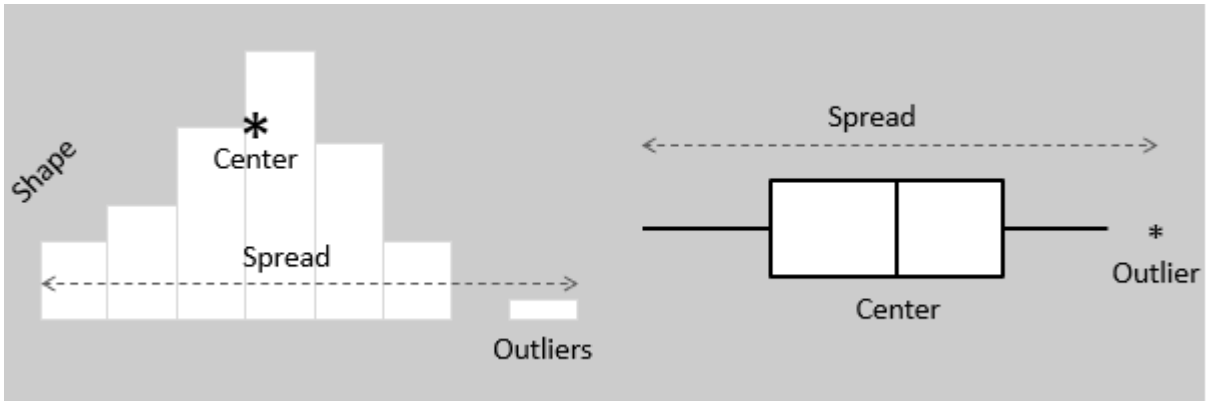
Interquartile Range $Q3 - Q1$

Robustness

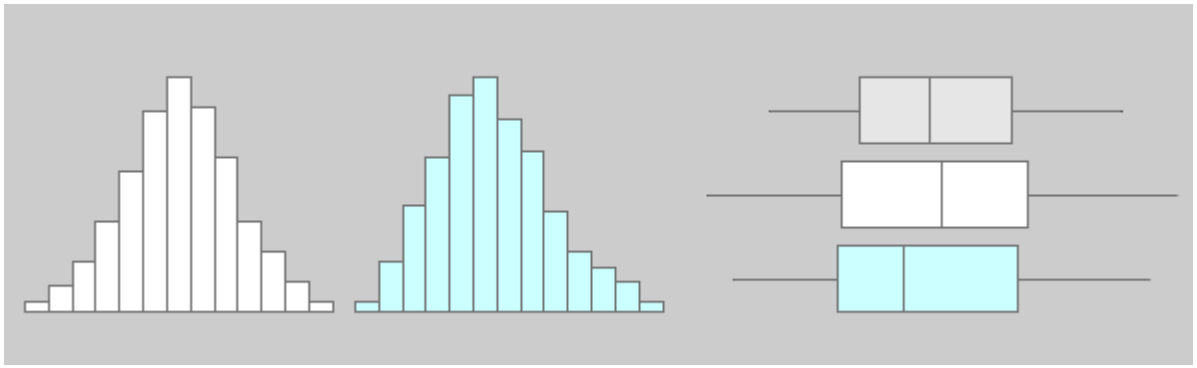
- In statistics, the term robust or robustness refers to the strength of a statistical model, tests, and procedures according to the specific conditions of the statistical analysis a study hopes to achieve. Given that these conditions of a study are met, the models can be verified to be true through the use of mathematical proofs.
- Many models are based upon ideal situations that do not exist when working with real-world data, and, as a result, the model may provide correct results even if the conditions are not met exactly.
- Robust statistics, therefore, are any statistics that yield good performance when data is drawn from a wide range of probability distributions that are largely unaffected by outliers or small departures from model assumptions in a given dataset. In other words, a robust statistic is resistant to errors in the results.
- One way to observe a commonly held robust statistical procedure, one needs to look no further than t-procedures, which use hypothesis tests to determine the most accurate statistical predictions.

Histograms and Box Plots

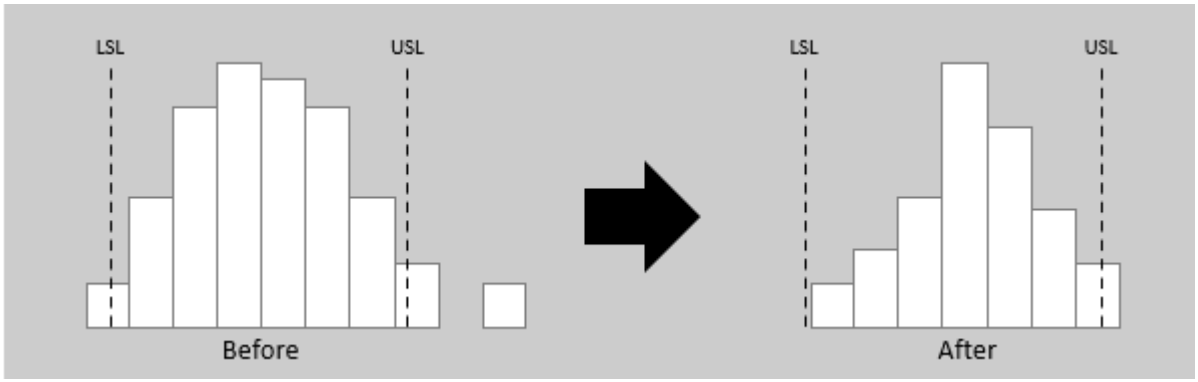
- Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.
- Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



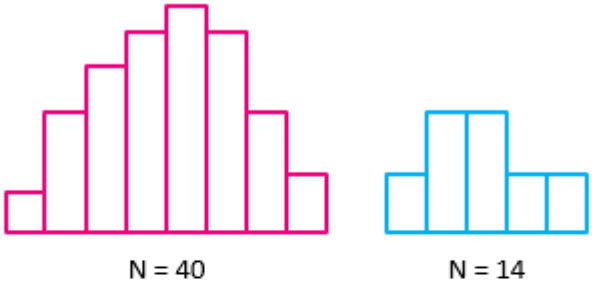
Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.

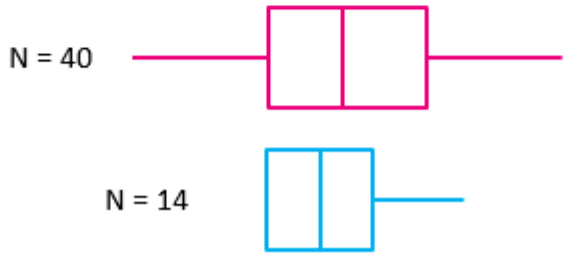


You can use histograms and box plots to verify whether an improvement has been achieved by exploring the data before and after the improvement initiative. Both tools can be helpful to identify whether variability is within specification limits, whether the process is capable, and whether there is a shift in the process over time.



Both histograms and box plots are ideal to represent moderate to large amount of data. They may not accurately display the distribution shape if the data size is too small. In practice, a sample size of at least 30 data values would be sufficient for both tools.





What is a Contingency Table?

- A contingency table displays frequencies for combinations of two categorical variables. Analysts also refer to contingency tables as cross tabulation and two-way tables.
- Contingency tables classify outcomes for one variable in rows and the other in columns. The values at the row and column intersections are frequencies for each unique combination of the two variables.
- Use contingency tables to understand the relationship between categorical variables. For example, is there a relationship between gender (male/female) and type of computer (Mac/PC)?
- I love these tables because they organize your data and allow you to answer diverse questions. In this post, learn about contingency tables, including how to interpret, graph, and analyse them.

Example Contingency Table

The contingency table example below displays computer sales at our fictional store. Specifically, it describes sales frequencies by the customer's gender and the type of computer purchased. It is a two-way table (2 X 2). I cover the naming conventions at the end.

	PC	Mac	Row Totals
Male	66	40	106
Female	30	87	117
Column Totals	96	127	223

- In this contingency table, columns represent computer types and rows represent genders. Cell values are frequencies for each combination of gender and computer type. Totals are in the margins. Notice the grand total in the bottom-right margin.
- At a glance, it's easy to see how two-way tables both organize your data and paint a picture of the results. You can easily see the frequencies for all possible subset combinations along with totals for males, females, PCs, and Macs.

- For example, 66 males bought PCs while females bought 87 Macs. Furthermore, there are 117 females, 106 males, 96 PC sales, 127 Mac sales, and a grand total of 223 observations in the study.
- Marginal and Conditional Distributions in Contingency Tables
- Contingency tables are a fantastic way of finding marginal and conditional distributions. These two distributions are types of frequency distributions. Learn more about Frequency Tables: How to Make and Interpret.

Marginal Distribution

- These distributions represent the frequency distribution of one categorical variable without regard for other variables. Unsurprisingly, you can find these distributions in the *margins* of a contingency table.
- The following marginal distribution examples correspond to the blue highlights.

	PC	Mac	Row Totals
Male	66	40	106
Female	30	87	117
Column Totals	96	127	223

For example, the marginal distribution of gender without considering computer type is the following:

- **Males**: 106
- **Females**: 117

Alternatively, the marginal distribution of computer types is the following:

- **PC**: 96
- **Mac**: 127

Conditional Distribution

For these distributions, you specify the value for one of the variables in the contingency table and then assess the distribution of frequencies for the other variable. In other words, you *condition* the frequency distribution for one variable by setting a value of the other variable. That might sound complicated, but it's easy using a contingency table. Just look across one row or down one column.

The following conditional distribution examples correspond to the green highlights.

	PC	Mac	Row Totals
Male	66	40	106
Female	30	87	117
Column Totals	96	127	223

For example, the conditional distribution of computer type for females is the following:

- **PC**: 30
- **Mac**: 87

Alternatively, the conditional distribution of gender for Macs is the following:

- **Males**: 40
- **Females**: 87

Finding Relationships in a Contingency Table

In the contingency table below, the two categorical variables are gender and ice cream flavor preference. This is a two-way table (2 X 3) where each cell represents the number of times males and females prefer a particular ice cream flavor. The CSV datasheet shows one format you can use to enter the data into your software: Flavor Preference.

Gender	Chocolate	Strawberry	Vanilla	Total
Female	37	17	12	66
Male	21	18	32	71
Total	58	35	44	137

How do we go about identifying a relationship between gender and flavor preference?

If there is a relationship between ice cream preference and gender, we'd expect the conditional distribution of flavors in the two gender rows to differ. From the contingency table, females are more likely to prefer chocolate (37 vs. 21), while males prefer vanilla (32 vs. 12).

Both genders have an equal preference for strawberry. Overall, the two-way table suggests that males and females have different ice cream preferences.

The Total column indicates the researchers surveyed 66 females and 71 males. Because we have roughly equal numbers, we can compare the raw counts directly. However, when you have unequal groups, use percentages to compare them.

Row and Column Percentages in Contingency Tables

Row and column percentages help you draw conclusions when you have unequal numbers in the margins. In the contingency table example above, more women than men prefer chocolate, but how do we know that's not due to the sample having more women? Use percentages to adjust for unequal group sizes. Percentages are relative frequencies. Learn more about Relative Frequencies and their Distributions.

Here's how to calculate row and column percentages in a two-way table.

- **Row Percentage**: Take a cell value and divide by the cell's row total.
- **Column Percentage**: Take a cell value and divide by the cell's column total.

For example, the row percentage of females who prefer chocolate is simply the number of observations in the Female/Chocolate cell divided by the row total for women: 37 / 66 = 56%.

The column percentage for the same cell is the frequency of the Female/Chocolate cell divided by the column total for chocolate: 37 / 58 = 63.8%.

Interpreting Percentages in a Contingency Table

The contingency table below uses the same raw data as the previous table and displays both row and column percentages. Note how the row percentages sum to 100% in the right margin while the column percentages sum to 100% at the bottom.

Gender	Chocolate	Strawberry	Vanilla	Row Total
Female	Raw: 37 Row%: 56% Col%: 63.8%	Raw: 17 Row%: 25.8% Col%: 48.6%	Raw: 12 Row%: 18.2% Col%: 28.8%	Raw: 66 Row%: 100%
Male	Raw: 21 Row%:29.6% Col%: 36.2%	Raw: 18 Row%: 25.4% Col%: 51.4%	Raw: 32 Row%: 45.0% Col%: 71.2%	Raw: 71 Row%: 100%
Total	Raw: 58 Col%: 100%	Raw: 35 Col%: 100%	Raw: 44 Col%: 100%	137

- Whether you focus on row percentages or column percentages in a contingency table depends on the question you're answering. In our case, we want to know whether flavor preference depends on gender. Because the two genders display in separate rows, we'll look for differences in the row percentages.
- 56% of females prefer chocolate versus only 29.6% of males. Conversely, 45% of males prefer vanilla, while only 18.2% of females prefer it. These results reconfirm our previous findings using the raw counts.

What is Relative Risk?

- Relative risk is the ratio of the probability of an adverse outcome in an exposure group divided by its likelihood in an unexposed group.
- This statistic indicates whether exposure corresponds to increases, decreases, or no change in the probability of the adverse outcome. Use relative risk to measure the strength of the association between exposure and the outcome.
- Analysts also refer to this statistic as the risk ratio.

Exposure can be a protective factor that reduces the chances of an adverse outcome (such as a vaccine or education program) or a risk factor that increases its likelihood (such as a toxin or a harmful environment).

- In this manner, the risk ratio determines whether exposure is a protective or risk factor.
- Given its ability to identify protective and risk factors, analysts frequently use relative risk in medical, intervention, and ecological studies. Additionally, it can intuitively convey the treatment effect in randomized controlled trials with binary outcomes, such as infected or not infected.

How to Calculate Relative Risk

- In statistics, risk is a probability, usually relating to an adverse event (i.e., something bad). To calculate relative risk (RR), you must know all subjects' exposure statuses and outcomes. Before learning how to calculate it, you first need to know about *absolute* risk.
- Absolute risk (AR) is simply the number of events divided by the number of people in the group. In the context of RR, we're working with two groups, those who were and were not exposed to something.
- For example, if 1 in 10 people exposed to a substance gets sick, the exposed AR is 0.1. If 1 in 100 people who are *not* exposed get sick, the unexposed AR is 0.01.
- In its simplest form, the relative risk formula is the ratio of AR for the two exposure groups, as shown below:

$$RelativeRisk = \frac{\text{AR exposed group}}{\text{AR unexposed group}}$$

- Using the example values above, let's plug the exposed and unexposed ARs into the formula:
- The relative risk result indicates that people exposed to the substance are ten times more likely to get sick! That's the relative increased probability associated with exposure.

Relative Risk Formula

Now, let's expand the relative risk calculation to show the formula in more detail. The table below shows a standard format for RR in a two-way contingency table. Learn more about Contingency Tables.

	Events	**Non-Events**	**Absolute Risk**
Exposed	A	B	A / (A + B)
Unexposed	C	D	C / (C + D)

Each letter represents a count of events or non-events in the exposed and unexposed groups. A row represents a group defined by exposure status, exposed or unexposed. In an experiment, the exposed group receives a treatment, while the unexposed group is the control. Learn more about Control Groups.

How to Interpret Relative Risk

Because the relative risk formula is a ratio, that tells us how to interpret it. The value of 1 becomes an important benchmark because it indicates that the exposed and unexposed groups have equal absolute risks.
Consequently, analysts compare their risk ratio results to one during interpretation.
As the ratio moves away from one in either direction, the relationship between exposure and the outcome strengthens.

Relative Risk = 1: The risk ratio equals one when the numerator and denominator are equal. This equivalence occurs when the probability of the event occurring in the exposure group equals the likelihood of it happening in the unexposed group. There is no association between exposure and the outcome.

Relative Risk > 1: The numerator is greater than the denominator in the risk ratio. Therefore, the event's probability is greater in the exposed group than in the unexposed group. This result identifies a risk factor because exposure corresponds with a greater probability of an adverse outcome.

Relative Risk < 1: The numerator is less than the denominator in the risk ratio. Consequently, the probability of the event is lower for the exposed group than for the unexposed group. This exposure is a protective factor because it corresponds with a lower probability of an adverse outcome.

Difference Between Proportions

Statistics problems often involve comparisons between two independent sample proportions. This lesson explains how to compute probabilities associated with differences between proportions.

Difference Between Proportions: Theory

Suppose we have two populations with proportions equal to P_1 and P_2. Suppose further that we take all possible samples of size n_1 and n_2. And finally, suppose that the following assumptions are valid.

- The size of each population is large relative to the sample drawn from the population. That is, N_1 is large relative to n_1, and N_2 is large relative to n_2. (In this context, populations are considered to be large if they are at least 20 times bigger than their sample.)
- The samples from each population are big enough to justify using a normal distribution to model differences between proportions. The sample sizes will be big enough when the following conditions are met: $n_1P_1 \geq 10$, $n_1(1 - P_1) \geq 10$, $n_2P_2 \geq 10$, and $n_2(1 - P_2) \geq 10$. (This criterion requires that at least 40 observations be sampled from each population. When P_1 or P_2 is more extreme than 0.5, even more observations are required.)

- The samples are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.

Given these assumptions, we know the following.

- The set of differences between sample proportions will be normally distributed. We know this from the central limit theorem.
- The expected value of the difference between all possible sample proportions is equal to the difference between population proportions. Thus, $E(p_1 - p_2) = P_1 - P_2$.
- The standard deviation of the difference between sample proportions (σ_d) is approximately equal to:

$$\sigma_d = \text{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$$

It is straightforward to derive the last bullet point, based on material covered in previous lessons. The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

$$\sigma^2_d = \sigma^2_{P1 - P2} = \sigma^2_1 + \sigma^2_2$$

If the populations N_1 and N_2 are both large relative to n_1 and n_2, respectively, then

$$\sigma^2_1 = P_1(1 - P_1) / n_1 \qquad \text{And} \qquad \sigma^2_2 = P_2(1 - P_2) / n_2$$

Therefore,

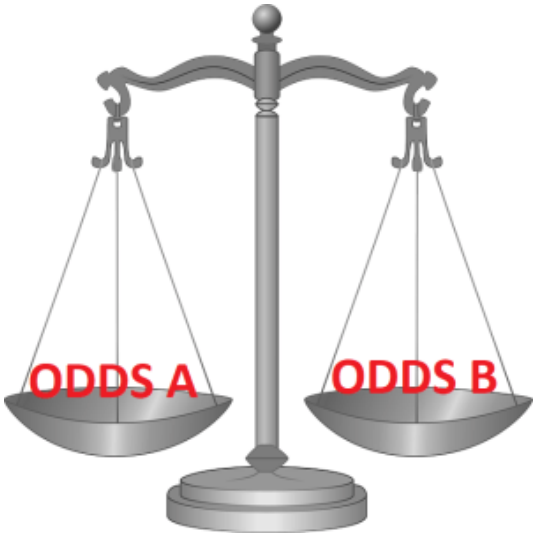
$$\sigma^2_d = [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2]$$

And

$$\sigma_d = \text{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$$

What is an Odds Ratio?

An odds ratio (OR) calculates the relationship between a variable and the likelihood of an event occurring. A common interpretation for odds ratios is identifying risk factors by assessing the relationship between exposure to a risk factor and a medical outcome. For example, is there an association between exposure to a chemical and a disease?



- To calculate an odds ratio, you must have a binary outcome. And you'll need either a grouping variable or a continuous variable that you want to relate to your event of interest.
- Then, use an OR to assess the relationship between your variable and the likelihood that an event occurs.
- When you have a grouping variable, an odds ratio interpretation answers the question, is an event more or less likely to occur in one condition or another?
- It calculates the odds of an outcome occurring in one context relative to a baseline or control condition. For example, your grouping variable can be a subject's exposure to a risk factor—yes or no—to see how that relates to disease status.
- With a continuous variable, calculating an odds ratio can determine whether the odds of an event occurring change as the continuous variable changes.

Odds definition: The probability of the event occurring divided by the probability of the event not occurring.

$$Odds = \frac{Probability\ Event\ Occurs\ (p)}{Probability\ Event\ Does\ Not\ Occur\ (1-p)}$$

Odds Ratios Interpretation for Two Conditions

Odds ratios with groups quantify the strength of the relationship between two conditions. They indicate how likely an outcome is to occur in one context relative to another.

The odds ratio formula below shows how to calculate it for conditions A and B.

$$Odds\ Ratio = \frac{Odds\ of\ an\ Event\ (Condition\ A)}{Odds\ of\ an\ Event\ (Condition\ B)}$$

How to Interpret Odds Ratios

Due to the odds ratio formula, the value of one becomes critical during interpretation because it indicates both conditions have equal odds. Consequently, analysts always compare their OR results to one when interpreting the results. As the OR moves away from one in either direction, the association between the condition and outcome becomes stronger.

Odds Ratio = 1: The ratio equals one when the numerator and denominator are equal. This equivalence occurs when the odds of the event occurring in one condition equal the odds of it happening in the other condition. There is no association between condition and event occurrence.

Odds Ratio > 1: The numerator is greater than the denominator. Hence, the event's odds are higher for the group/condition in the numerator. This is often a risk factor.

Odds Ratio < 1: The numerator is less than the denominator. Hence, the probability of the outcome occurring is lower for the group/condition in the numerator. This can be a protective factor.

How to Calculate an Odds Ratio

The equation below expands the earlier odds ratio formula for calculating an OR with two conditions (A and B). Again, it's the ratio of two odds. Hence, the numerator and denominator are also ratios.

$$Odds\ Ratio = \frac{Odds\ of\ an\ Event\ (Condition\ A)}{Odds\ of\ an\ Event\ (Condition\ B)} = \frac{\#\ Events\ (A) \Big/ \#\ Non\ Events\ (A)}{\#\ Events\ (B) \Big/ \#\ Non\ Events\ (B)}$$

In the infection example above, we assessed the relationship between treatment and the odds of being infected. Our two conditions were the treatment (condition A) and the control group (B). On the right-hand side, we'd enter the numbers of infections (events) and non-infections (non-events) from our sample for both groups.

Scatter plots

Scatter plots **are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a** Cartesian system**. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. These plots are often called** scatter graphs **or** scatter diagrams**.**

Scatter plot Graph

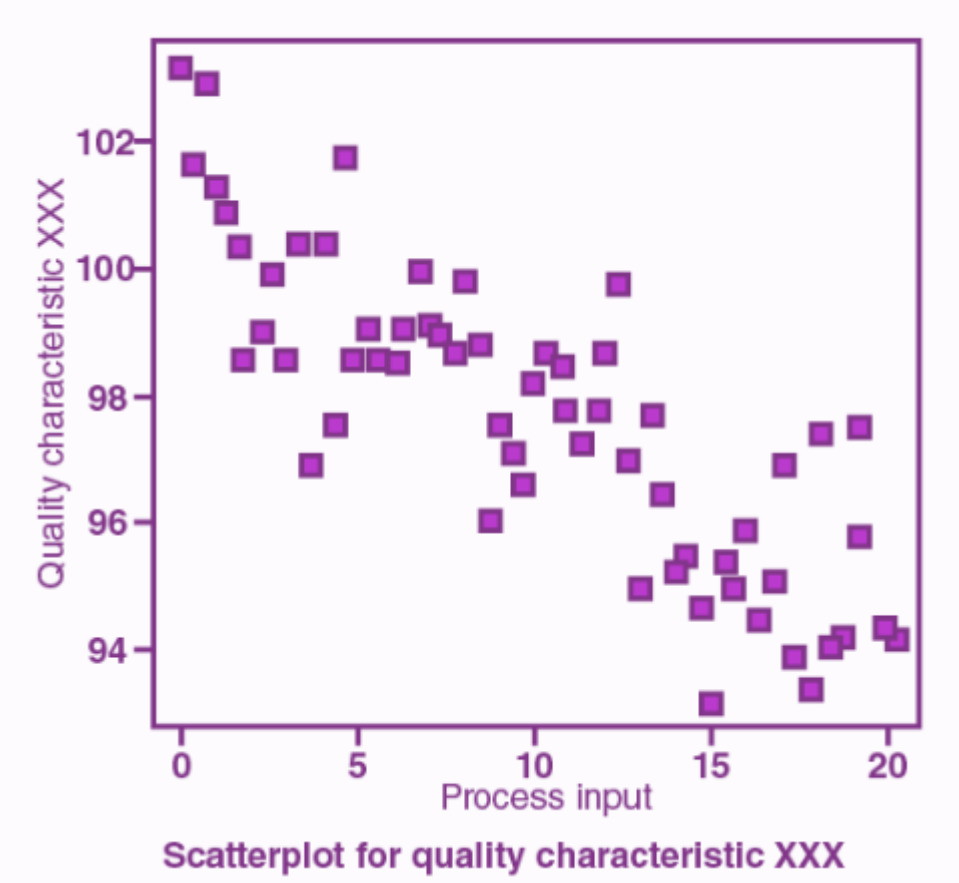
A scatter plot is also called a scatter chart, scattergram, or scatter plot, XY graph. The scatter diagram graphs numerical data pairs, with one variable on each axis, show their relationship. Now the question comes for everyone: **when to use a scatter plot**?

Scatter plots are used in either of the following situations.

- When we have paired numerical data
- When there are multiple values of the dependent variable for a unique value of an independent variable
- In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appear to be related both occur with the exact cause and so on.

Scatter Plot Uses and Examples

- For a large set of data points given
- Each set comprises a pair of values
- The given data is in numeric form



Scatter plot Correlation

We know that the correlation is a statistical measure of the relationship between the two variables' relative movements. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the closer the points will touch the line. This cause examination tool is considered as one of the seven essential quality tools.

Types of correlation

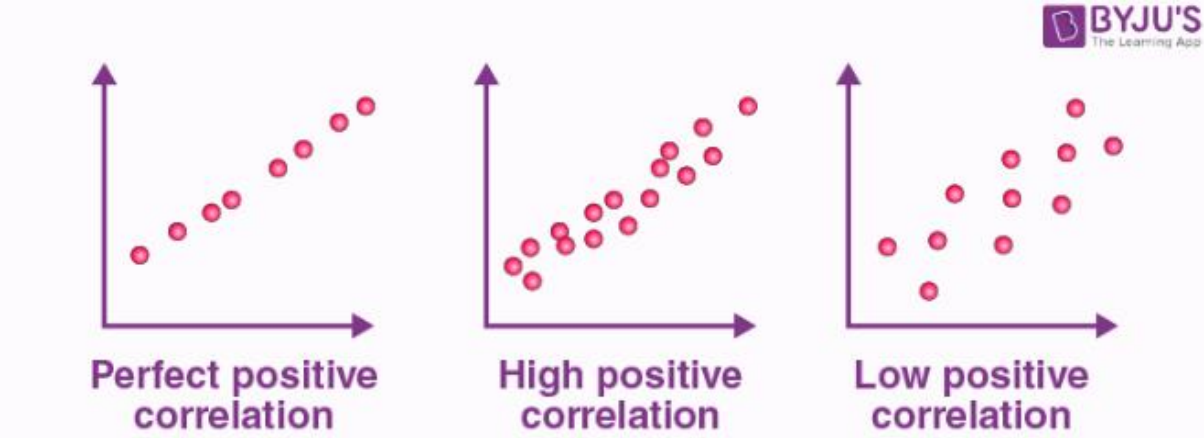
The scatter plot explains the correlation between two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

1. Positive Correlation
2. Negative Correlation
3. No Correlation

Positive Correlation

When the points in the graph are rising, moving from left to right, then the scatter plot shows a positive correlation. It means the values of one variable are increasing with respect to another. Now positive correlation can further be classified into three categories:

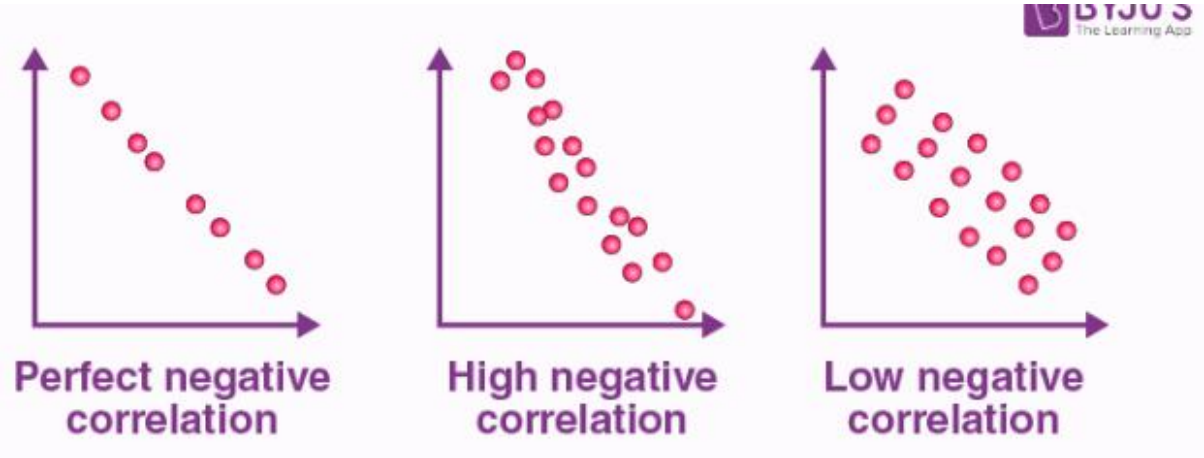
- **Perfect Positive** – Which represents a perfectly straight line
- **High Positive** – All points are nearby
- **Low Positive** – When all the points are scattered



Negative Correlation

When the points in the scatter graph fall while moving left to right, then it is called a negative correlation. It means the values of one variable are decreasing with respect to another. These are also of three types:

- **Perfect Negative** – Which form almost a straight line
- **High Negative** – When points are near to one another
- **Low Negative** – When points are in scattered form



No Correlation

When the points are scattered all over the graph and it is difficult to conclude whether the values are increasing or decreasing, then there is no correlation between the variables.

Measures of correlation

Correlation refers to a process for establishing the relationships between two variables. You learned a way to get a general idea about whether or not two variables are related, is to plot them on a "scatter plot". While there are many measures of association for variables which are measured at the ordinal or higher level of measurement, correlation is the most commonly used approach.

Correlation in Statistics

- This section shows how to calculate and interpret correlation coefficients for ordinal and interval level scales. Methods of correlation summarize the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1.
- A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables. A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.
- A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable. A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal.
- For ordinal scales, the correlation coefficient can be calculated by using Spearman's rho. For interval or ratio level scales, the most commonly used correlation coefficient is Pearson's r, ordinarily referred to as simply the correlation coefficient.

Simple linear regression

Simple linear regression is used to estimate the relationship between **two quantitative variables**. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

****