Notes:

Scatter Diagrams

Key Learning Points

1. Describe the importance of Scatter Diagrams.
2. Explain how to develop and interpret Scatter Diagrams.
3. Utilize Scatter Diagrams in improvement projects.

What is a Scatter Diagram?

A scatter diagram (also known as a scattergram or scatter plot) is a numerical relationship. When looking at the relationship between variables in a scatter diagram, one can only say that the variables are correlated, you can't say that X causes Y. The scatter diagram is an ideal way to display data when a team is trying to evaluate a cause-effect relationship.

Relationship

A scatter diagram is a graphic presentation of the relationship between two variables.

Correlation Coefficient

It is possible to summarize the relationship between two measures quantitatively using a correlation coefficient.

Verify Root Cause

Scatter diagrams are used to help verify potential root causes because the premise is that a change in the cause (the X) will produce a change in the effect (the Y).

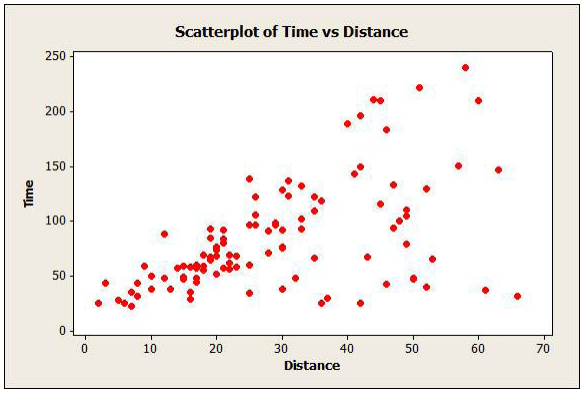
If the potential root cause is correct, the Y variable will change as the X changes. Scatter diagrams are used with continuous data, and when the team has ruled out special causes and is looking at common cause variation.

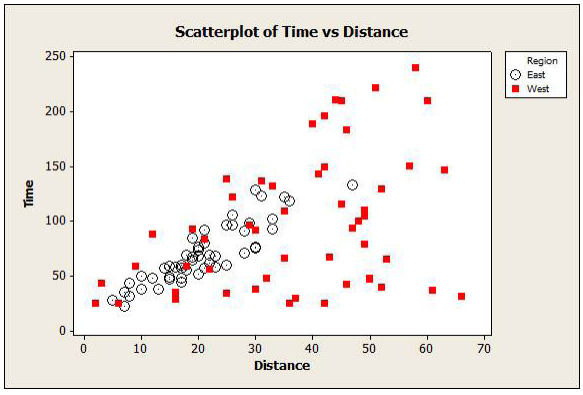
Example Scatter Diagram

A team was trying to improve the response speed of their ambulance services, which involved traveling to the patient's location of call. One of their theories was that the delays were primarily the result of the distance that they had to travel to deliver their service.

Earlier data had shown that the East region had faster response times than the West, so some tension had developed on the team over the reason for the difference. On the one hand, team members from the West claimed that they had greater distance to travel, which led to their longer response times. Team members from the East countered that traffic was more congested in their region, so they had to travel more slowly.

- Is the scatter diagram consistent with either of these theories?
- What might you do to further test the theories?





Characteristics

Remember: Correlation does not imply causation

Notes:

To be successful when creating Scatter Diagrams, you need:

- A good theory
- Correctly paired data
- Accuracy
- Complete information
- Representative data

Steps in Constructing Scatter Diagrams

1. Obtain the table of raw data, and determine the high and low values for each variable.
2. Decide which variable will be plotted on the horizontal axis.
3. Draw and label the horizontal and vertical axes.
4. Plot the paired data.
5. Title the chart, and provide other appropriate notations.
6. Identify and classify the pattern of correlation.
7. Check for potential pitfalls in your analysis. Consider confounding factors and other possible explanations for the correlation pattern, and decide on the team's next steps.

Additional Tips:

- As you gather the data, make sure that only the two variables of interest are changing—all other factors should remain constant and should be representative of normal conditions.
- Use the largest sample size possible, consistent with budget and time constraints—40 or more points are recommended.
- Keep in mind that your objective is to learn something about the relationship between two variables by examining their joint variability.

Interpretation

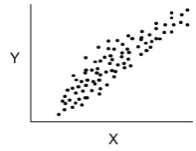
Analyzing scatter diagrams is a four-step process:

1. Develop a plausible and relevant theory about the suspected relationship between two variables of interest.
2. Collect appropriate paired data, and construct a scatter diagram.
3. Identify and classify the pattern of correlation.
4. Question your original theory and consider other explanations for the observed pattern of correlation.

Notes:

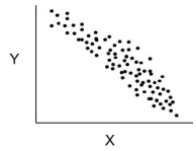
Interpreting Patterns of Correlation

Strong, Positive



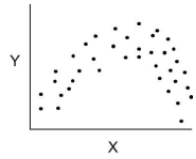
If one variable increases at the same time the other variable increases, they are said to be positively correlated.

Strong, Negative



If one variable decreases at the same time the other variable increases, or vice versa, they are said to be negatively correlated.

Complex



The data points are scattered in a curved pattern. The shape may look like a rainbow or an arch. The two variables are correlated, though not linearly. As X increases, Y first increases, then it decreases (or vice versa).

Weak Relationships

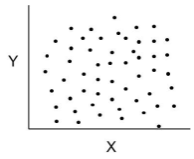
A weak correlation does not necessarily mean that the factor being studied is not a cause. It may simply be a weak cause or a cause that requires the presence of another contributing factor to bring about the effect. In this latter case, both the factor under study and the contributing factor are perfectly good causes; you just need them both to
be active simultaneously to get the effect.

Weak, Positive

Notes:

Weak, Negative

None



The data points are scattered in a shapeless pattern. You can conclude that the two variables are not correlated over the ranges for which the data was collected.

When Should a Scatter Diagram Be Used?

Scatter diagrams are an ideal way to display data when trying to evaluate a cause-effect relationship.

They are also a valuable tool for analyzing relationships between two variables to test:

- if both variables are from a common cause that is unknown or difficult to measure
- if one effect can be used as a surrogate for the other

They also are used to plot the relationship between two possible causes

Scatter Diagrams are Easy to Misinterpret

Stratification

Stratification can make a correlation pattern appear, or it can completely change the interpretation of an apparent correlation. It is a good idea to plot stratifying factors on scatter diagrams to test their effect on the correlation patterns.

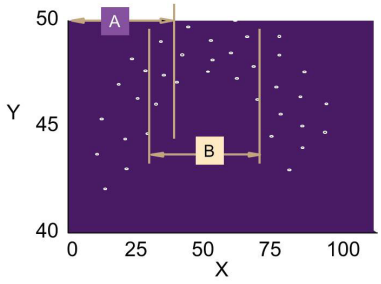
Range of the Data

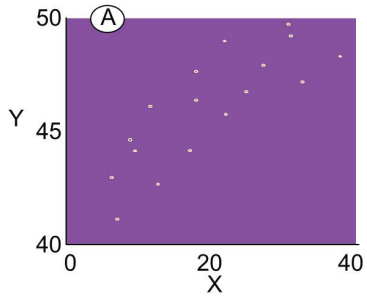
It is critically important that a team restrict its interpretation of the scatter diagram to the range of the observations. The comparison below illustrates this potential pitfall. The actual relationship between variables X and Y is shown in the top left of the diagram below.

Suppose that a team had only gathered data on X for the range from 0 to 40. Their scatter diagram would resemble the area labeled “A” (enlarged as follows). On the

Notes:

basis of the strong positive correlation displayed in diagram A, the team might be tempted to assume that a large value of X, say X = 80, would yield a correspondingly large value of Y. Clearly, this conclusion about the relationship between X and Y, which goes outside the range of the data shown in diagram A, would be incorrect.



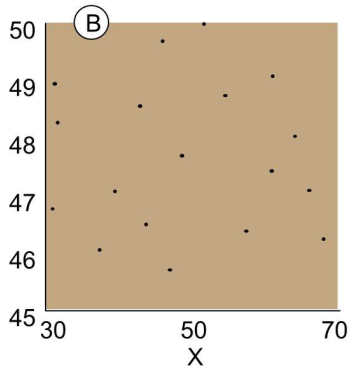


Range of Operation

While it is incorrect to state conclusions beyond the range of the data, it is important to be equally wary of the value of gathering and analyzing data beyond the range of the operation of the process. To clarify and illustrate this point, recall on the Range of Data slide that the diagram shows the actual relationship between X and Y for the range X = 0 to 100 and Y = 40 to 50. There is a definite, albeit complex, correlation. If a team had all the data shown in the diagram, they might be tempted to conclude that control of one variable would result in control of the other. Suppose that the process only operates within the range X = 30 to 70. The scatter diagram in this "region of actual operation" is shown on as diagram B. In the "region of actual operation," there is no correlation; control of one variable will not help to control the other. While it may be interesting to know the relationship between two variables in the region outside the normal range of operation, it must be understood that such knowledge is useful only if changes can be made to allow the process to operate in this new range.

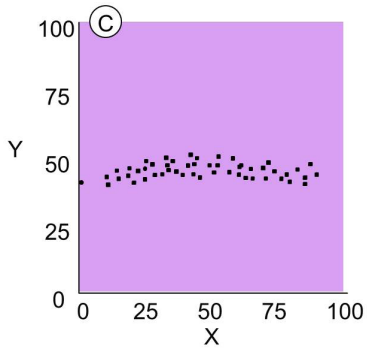
Notes:

Notes:



Effect of Scale

Another interpretation problem is illustrated in diagram C below. Diagram C shows the same data, but here the team has chosen to make the X and Y axes the same. This simple change of scale makes it appear that the value of Y is more or less constant, regardless of X. The point is that proper interpretation of a scatter diagram depends on careful attention to construction details. Generally, the axes of scatter diagrams should be drawn so that (1) the maximum and minimum values on each axis correspond closely to the maximum and minimum values of each variable, and (2) the axes are roughly the same length.



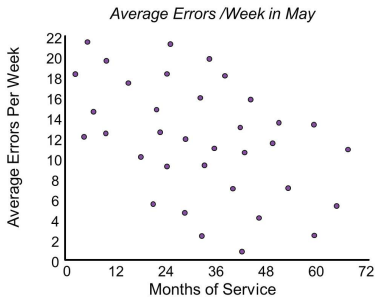
Pitfalls to Avoid

Confounding Factors

With a scatter diagram, you are studying one cause-effect pair among many. It is therefore possible that the correlation you observe in your scatter diagram is due to a cause other than the one you are plotting. This other unseen cause is sometimes called a confounding factor.

The best time to deal with confounding factors is before the data is gathered. Identify the potential confounding factors and arrange to hold them reasonably constant. After constructing the scatter diagram, consider confounding factors again. Look for other explanations for the correlation pattern, and if possible, stratify the data by confounding factors.

Example: Errors vs. Experience



A scatter diagram indicated some correlation between experience level and number of errors.

"So," one team member concluded, "we'll just have to wait until the new clerks get up to speed."

Another team member, not content to settle for this wait-and-see conclusion, noted that the clerks were in two different supervisory groups. One group, which admittedly had the more experienced clerks, had spent more time refining the process, developing job aids, conducting 30 minute refresher training sessions, distributing new-product offering announcements, etc. "What would the scatter diagram look like if the data were stratified by a supervisory group?" he asked. The stratified scatter diagram is shown here.

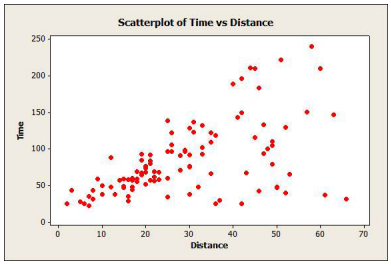
The scatter diagram showed a correlation between errors and experience, but the interpretation was "confounded" by the fact that there was another variable at work beside experience—namely, different support systems in the two groups. The presence of these different support systems was a "confounding factor" in the analysis. In this case, the confounding factor turned out to be the most important factor.

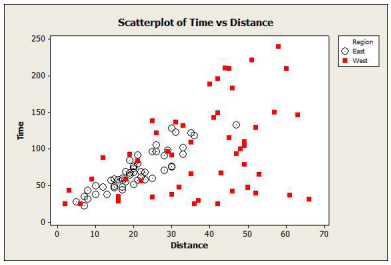
Data Problems

The data must be accurate and representative of normal conditions. Otherwise, the scatter diagram interpretation might be eloquent but totally useless.

Notes:

Example: Service Response Time





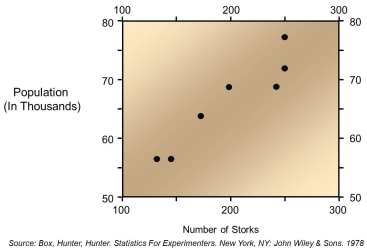
A team was trying to improve the response speed of their ambulance services, which involved traveling to the patient's location of call. One of their theories was that the delays were primarily the result of the distance that they had to travel to deliver their service.

Earlier data had shown that the East region had faster response times than the West, so some tension had developed on the team over the reason for the difference. On the one hand, team members from the West claimed that they had greater distance to travel, which led to their longer response times. Team members from the East countered that traffic was more congested in their region, so they had to travel more slowly.
Is the scatter diagram consistent with either of these theories?
What might you do to further test the theories?

Assuming Correlation Equals Causation

One of the main pitfalls of scatter diagrams is the assumption that correlation equals causation.

Correlation and Causation Examples



Source: Box, Hunter, Hunter. Statistics For Experimenters. New York, NY: John Wiley & Sons. 1978

- Rum prices and ministers' salaries are known to be correlated. Does an increase in salary cause a rum price increase? Or the other way around?

Notes:

- This is a diagram showing the population of Oldenburg at the end of each year versus the number of storks observed in that year, 1930-1936.

Even strong correlations do not imply causation.

Mixing Up X and Y Variables

In analyzing variables for potential cause and effect, the process variable as the predictor is on the "X" axis and the output or process performance variable as the response is on the "Y" axis.

Incorrectly Paired Data

Scatter diagrams require that the "X" and "Y" variables be paired. It means that there has to be a logical connection between the data to appropriately study the correlation.

Improper Scaling

To ensure proper interpretation, it is necessary to use the fullest possible extent of both the X and Y axes to cover the range of the data collected. Improper scaling can result in the pattern being skewed.

Incorrect Increment Spacing Between Tick Marks

The increment, or the (space) width between tick marks on a scatter diagram, should be a value that makes it easy to plot the data and is easy to read. A general guideline is to have between 5 to 15 increments of equal width along the full length of each axis.

Notes: