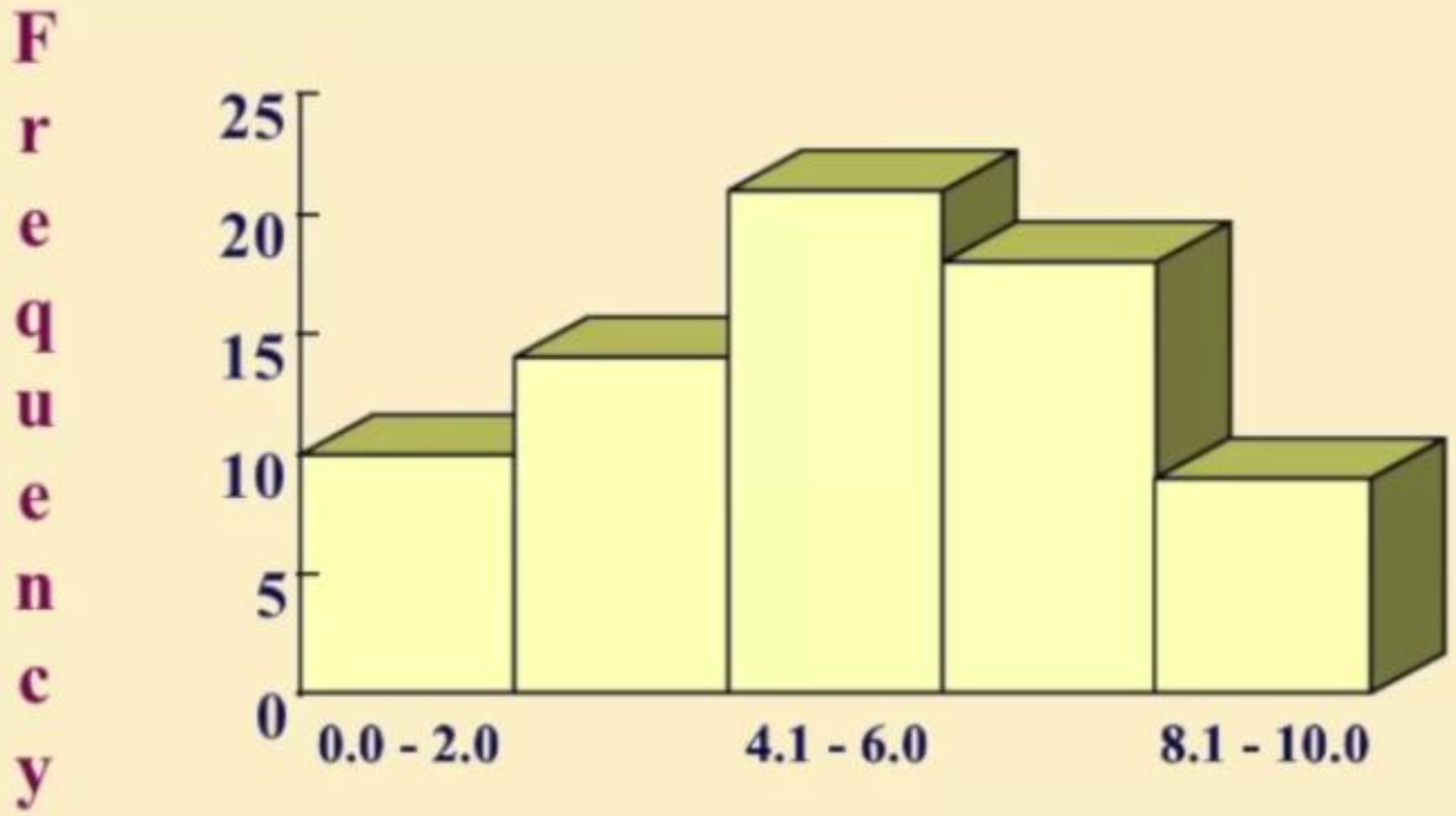
Histograms

Histograms

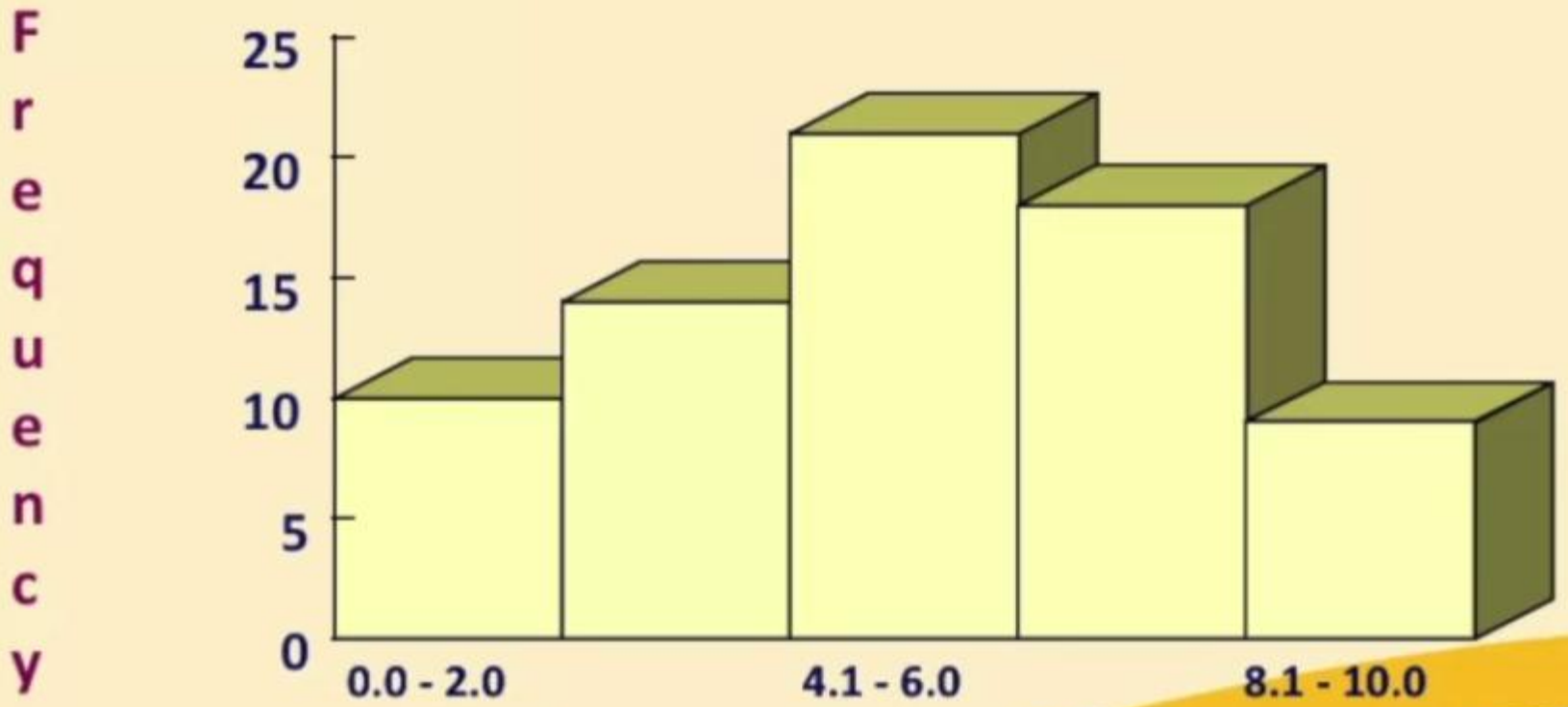


- A useful tool to help us to understand our data is a histogram.
- Remember with check sheets we counted how many times different things occurred.
- That number is also called a frequency.

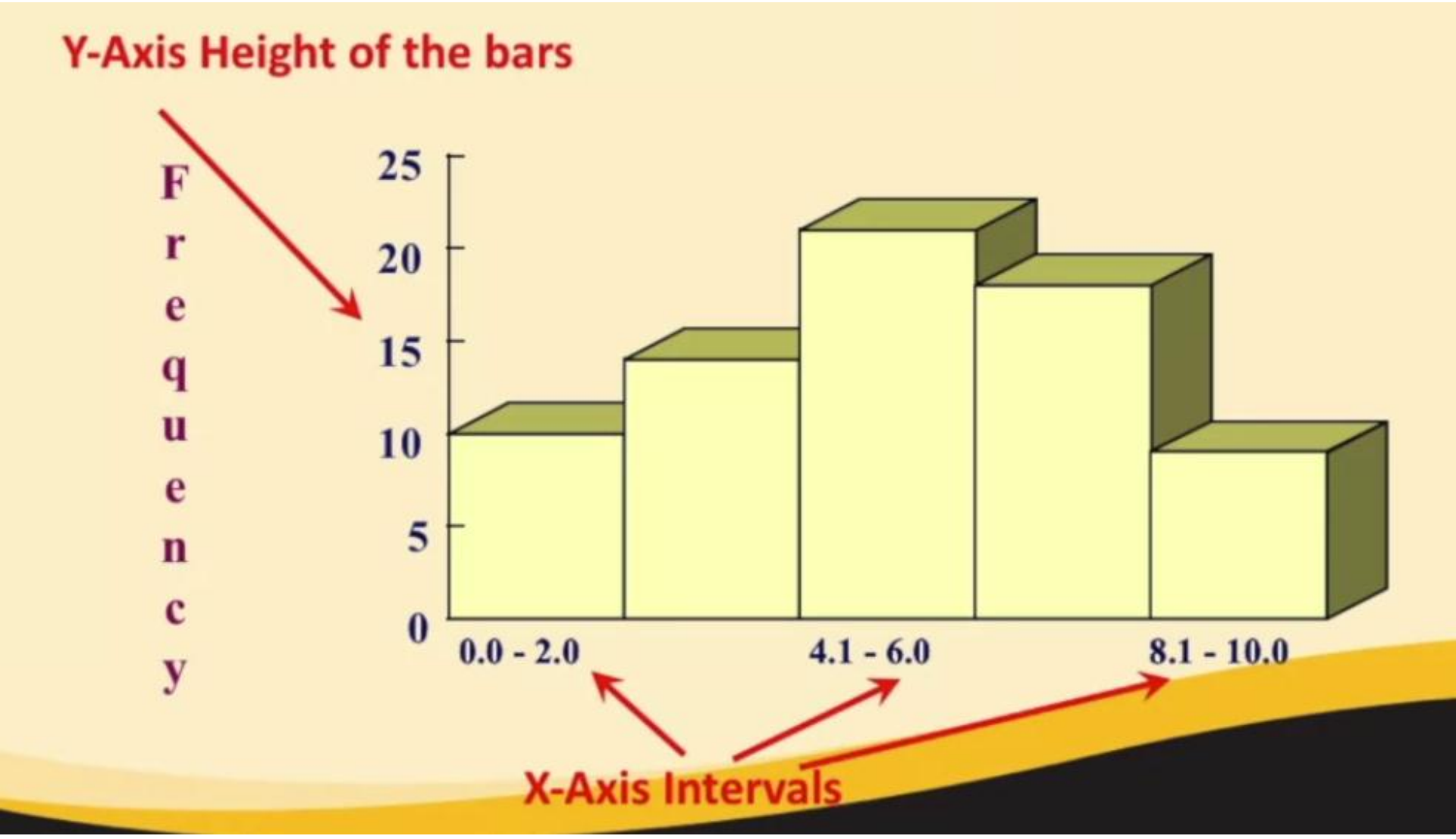
- The histogram is a graphic representation or a picture of what that frequency distribution looks like.
- Typically, it's a special type of bar graph.
- The vertical axis represents the frequency or how many times something happened.
- And the horizontal axis represents the various events or the things that happen.

Histograms

Customer Call Response Time in Seconds for October



- We can use our frequency table to construct a histogram.
- This might be simple counts of things or events, this is also called the discrete data.
- Or, if we have measurements, which may be continuous data, we will need to group those measurements and count how many times our measures fell in those groups.


Y-Axis Height of the bars
Frequency
25
20
15
10
5
0
0.0 - 2.0
4.1 - 6.0
8.1 - 10.0
X-Axis Intervals

Grouping Data

- How many groups?
  - 5 – 20 depending on the data
  - Number of groups = $\sqrt{n}$
- How wide
  - Subtract the lowest value from the highest value
  - Divide by the number of groups
  - Round up
  - Equal size

What is it?

- A histogram is constructed from a frequency table.

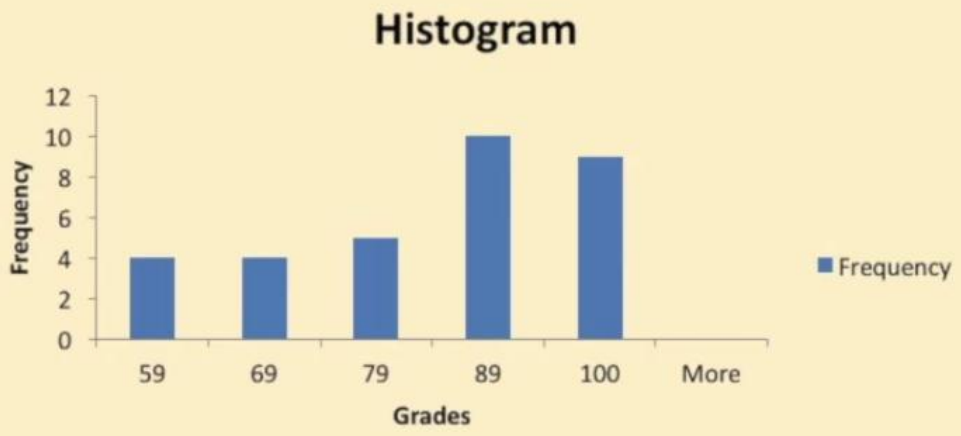
Grades									
<60	x	x	x						
60-69	x	x	x						
70-79	x	x	x	x					
80-89	x	x	x	x	x	x	x	x	x
90-100	x	x	x	x	x	x	x	x	

- A histogram is constructed from a frequency table.

Grades													Totals
<60	x	x	x										3
60-69	x	x	x										3
70-79	x	x	x	x									4
80-89	x	x	x	x	x	x	x	x	x				9
90-100	x	x	x	x	x	x	x	x					8
													27

Histograms

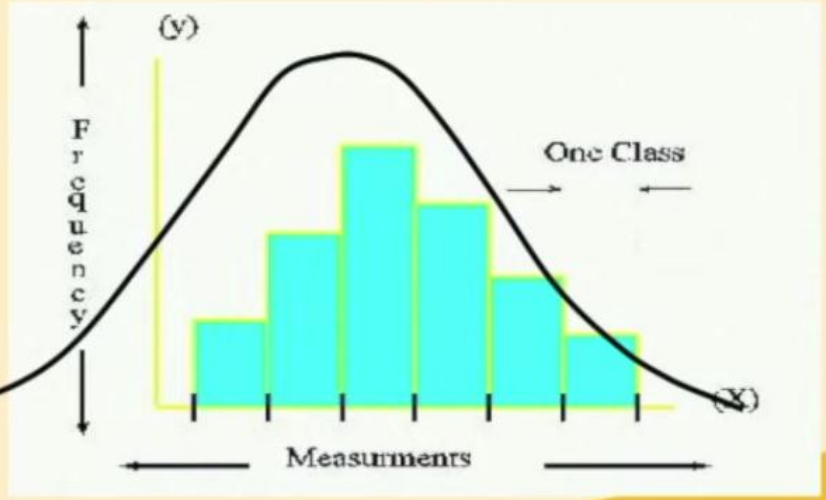
<60	60-69	70-79	80-89	90-100
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X
		X	X	X
			X	X
			X	X
			X	X
			X	X



What does this tell us?

Histograms

- Look for data to be “approximately” normal



- The first thing that we look for is whether or not the data is **normally distributed**.
- You've probably seen representations of the normal curve or the bell curve.
- Here we've placed a rough sketch of the normal curve over our histogram.

- We're not looking for an exact fit.
- Often, we may not have a lot of data.
- Small samples will not look exactly normal.
- For small amounts of data, like fewer than 50 data points, is the distribution higher near the middle and smaller as you move right to left?
- There are statistical tests for normality, but those are beyond the scope.

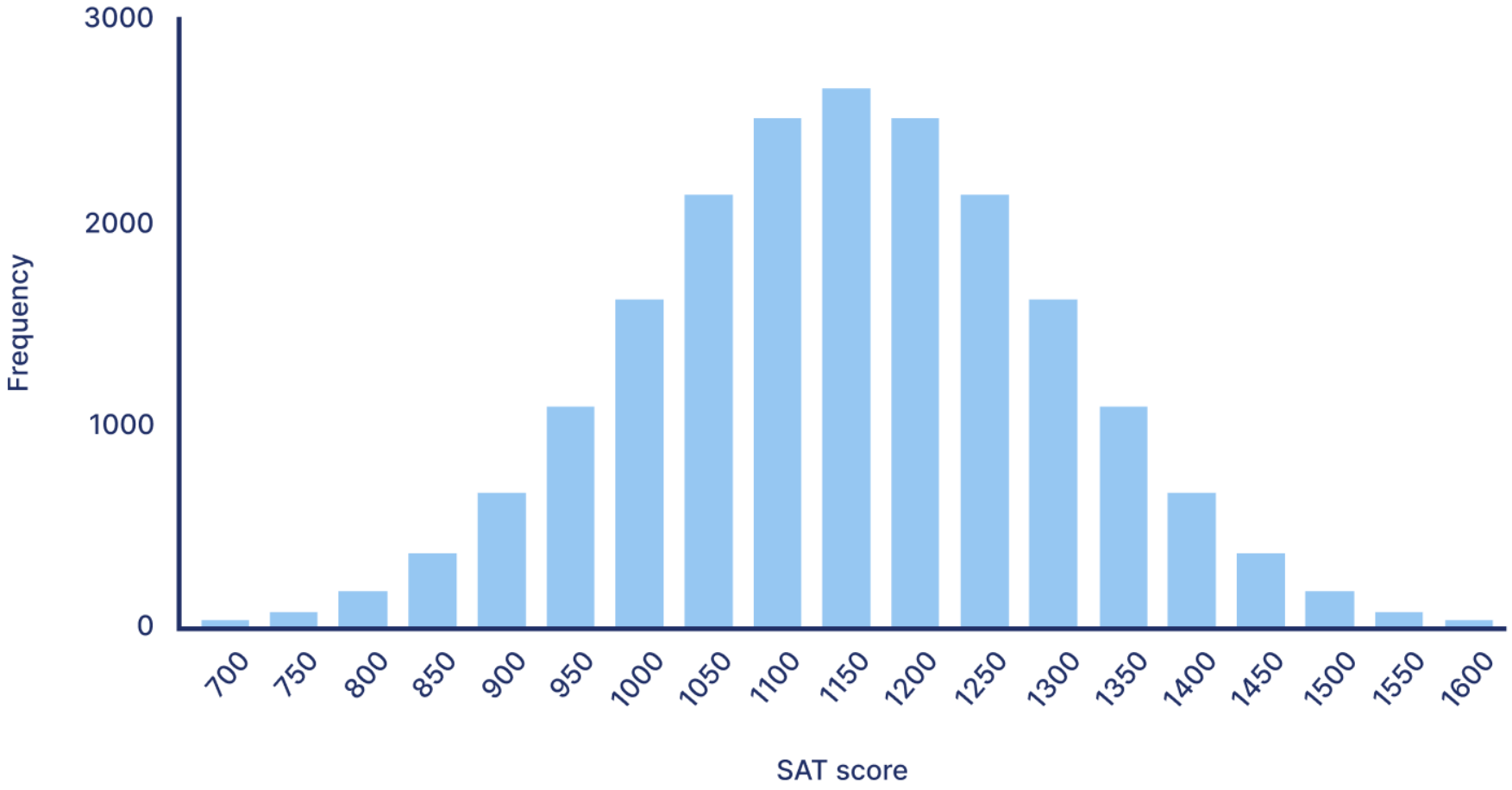
Purpose of Histogram

- Is the data normally distributed?
- Where's the approximate center of the data?
- What's the spread of the data?
- That is, what are the differences between the lowest and highest values?
- Are there any outliers or unusual values?
- The answers to these questions can help you to determine what type of analysis you can do with the data.
- And you can get an idea of those values just by glancing at the histogram.

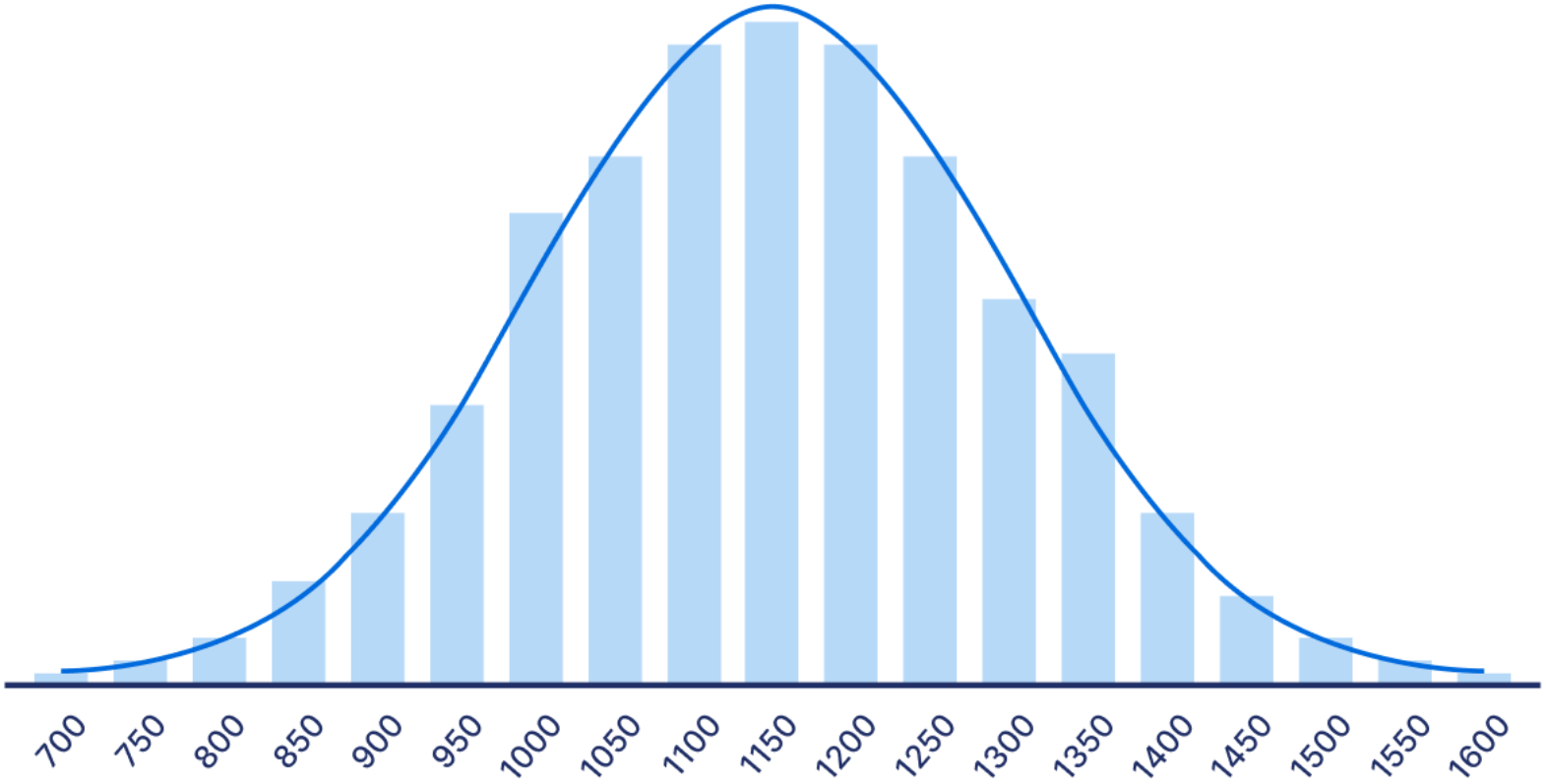
Normal Distribution

- A normal distribution, also known as a Gaussian distribution or a bell curve, is a specific type of distribution that is symmetric and characterized by a mound-shaped curve. In a normal distribution:
  - The data is symmetrically distributed around the mean.
  - The majority of data points cluster around the mean, with fewer data points in the tails.
  - The mean, median, and mode are all equal and located at the center of the distribution.
  - The distribution is fully described by its mean and standard deviation.

Example of Normal Distribution



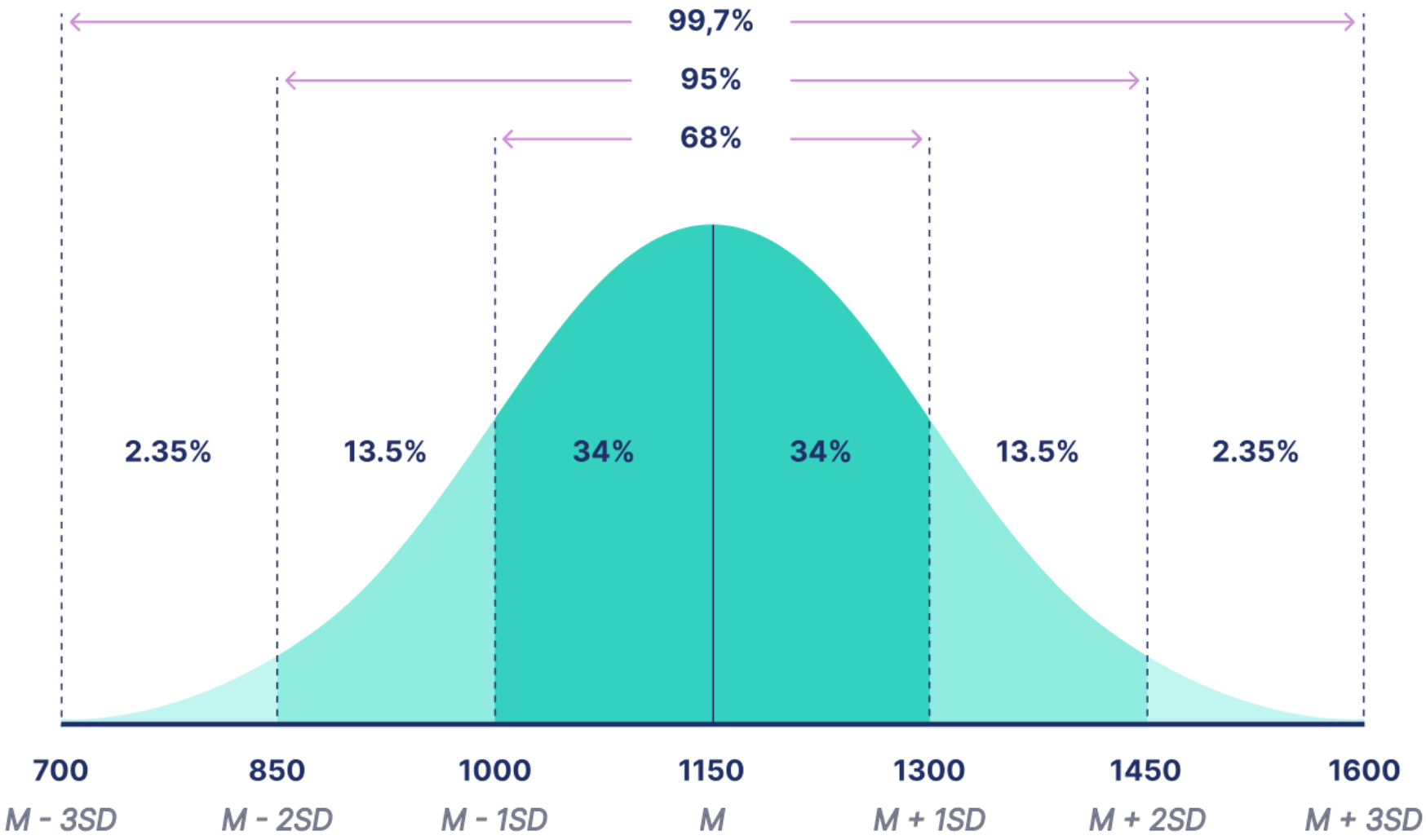
The Bell Curve



Example: Using the Empirical Rule in a Normal Distribution

- You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (*M*) of 1150 and a standard deviation (*SD*) of 150.
- Following the empirical rule:
  - Around 68% of scores are between 1,000 and 1,300, 1 standard deviation above and below the mean.
  - Around 95% of scores are between 850 and 1,450, 2 standard deviations above and below the mean.
  - Around 99.7% of scores are between 700 and 1,600, 3 standard deviations above and below the mean.

Normal Distribution

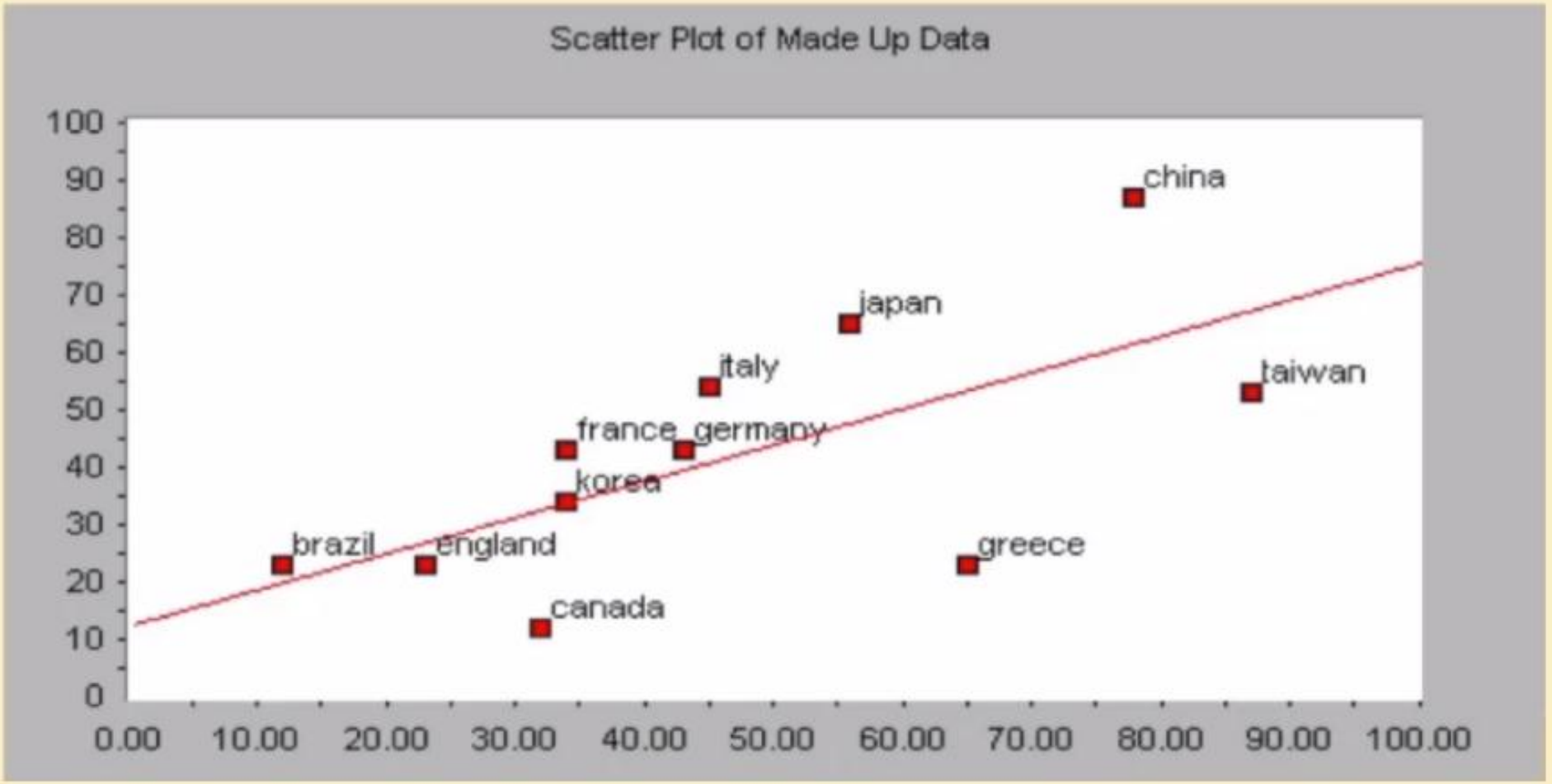


Scatter Diagram

Scatter Diagram

- Scatter plots, or scatter diagrams, are **used to investigate the possible relationship between two variables** that **both relate to the same event**.
- A scatter diagram gives you a visual representation of relationships that can be confirmed through **correlation regression analysis**.

Scatter Diagram



Scatter Diagram

How to Construct

- Not cause and effect
- Software
- Plot by hand

How to construct

- Scatter diagrams are useful for exploring root cause.
- Although they do not specifically indicate cause and effect, used with other evidence cause and effect might be implied.
- Scatter diagrams are usually created using statistical software, such as Many Tab, Excel, or some other spreadsheet application, but they're easily plotted by hand.

How to construct

- The independent variable, the one you think might be the cause, is the y axis or vertical axis.
- The dependent variable, the one you're trying to fix, is the x, or horizontal axis.
- For every event or occurrence, two pieces of data are reported.
- One for the independent variable and a related one for the dependent variable.
- Data is recorded in x, y pairs. Each point on the plot represents the pairs of values for x and y.

Example

How to Construct

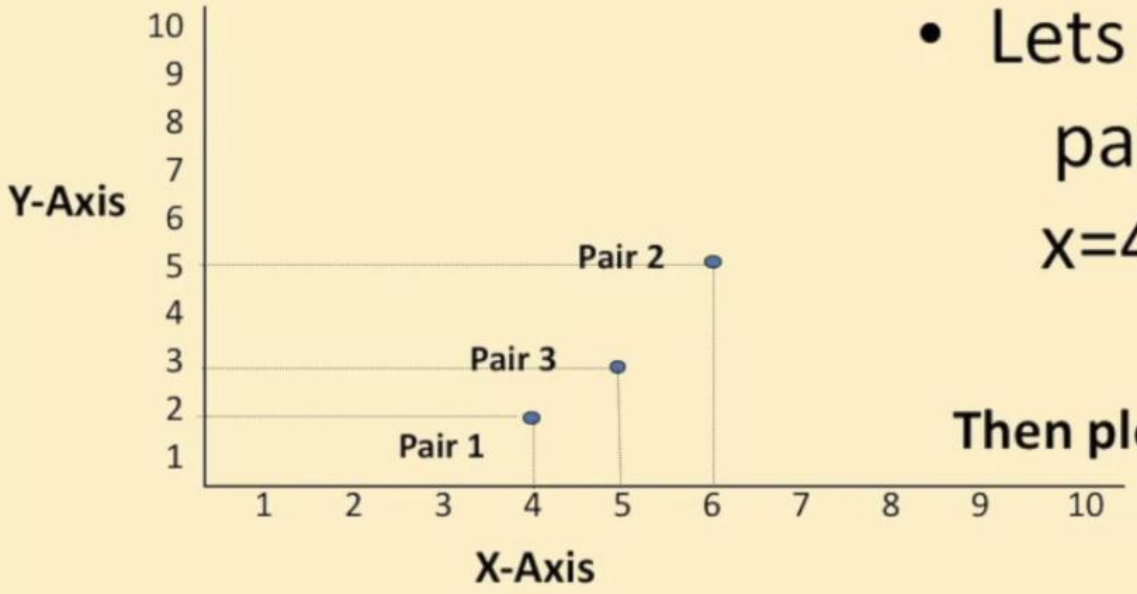
- Let's start with three pairs of x-y data:

  x=4 y=2 x=6 y=5 x=5 y=3.

x	y
4	2
6	5
5	3

Example

How to Construct



- Lets start with three pairs of x-y data:

x=4 y=2, x=6 y=5

x=5 y=3.

Then plot the pairs on the chart

Correlation!

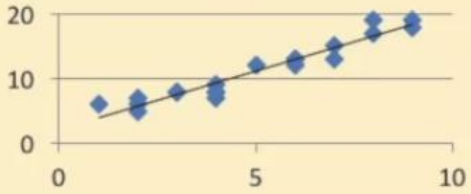
- Correlation and regression analysis is a topic for future courses but for now, it's enough to know that correlation tells you the direction and the strength of a relationship.
- A correlation will be some value between -1 and +1.
- A correlation of +1 indicates a perfect positive relationship.
  - That is, all of the points will be right on the line.
- A correlation of -1 indicates a perfect negative relationship, and;
- A correlation of zero means no significant relationship between x and y.

Regression!

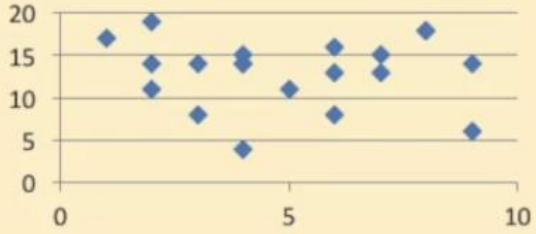
- Regression analysis will tell you the nature of the relationship between an independent variable and a dependent variable and regression allows you to create and test a predictive equation.

Things to Look for

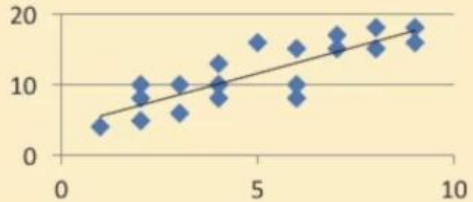
Strong Positive



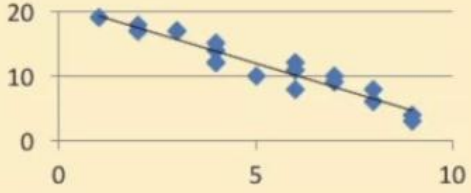
No Relationship



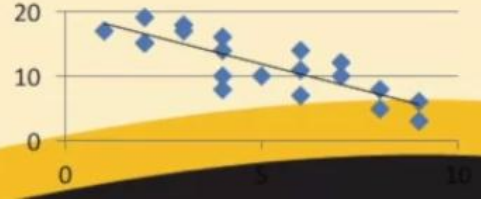
Weak Positive



Strong Negative

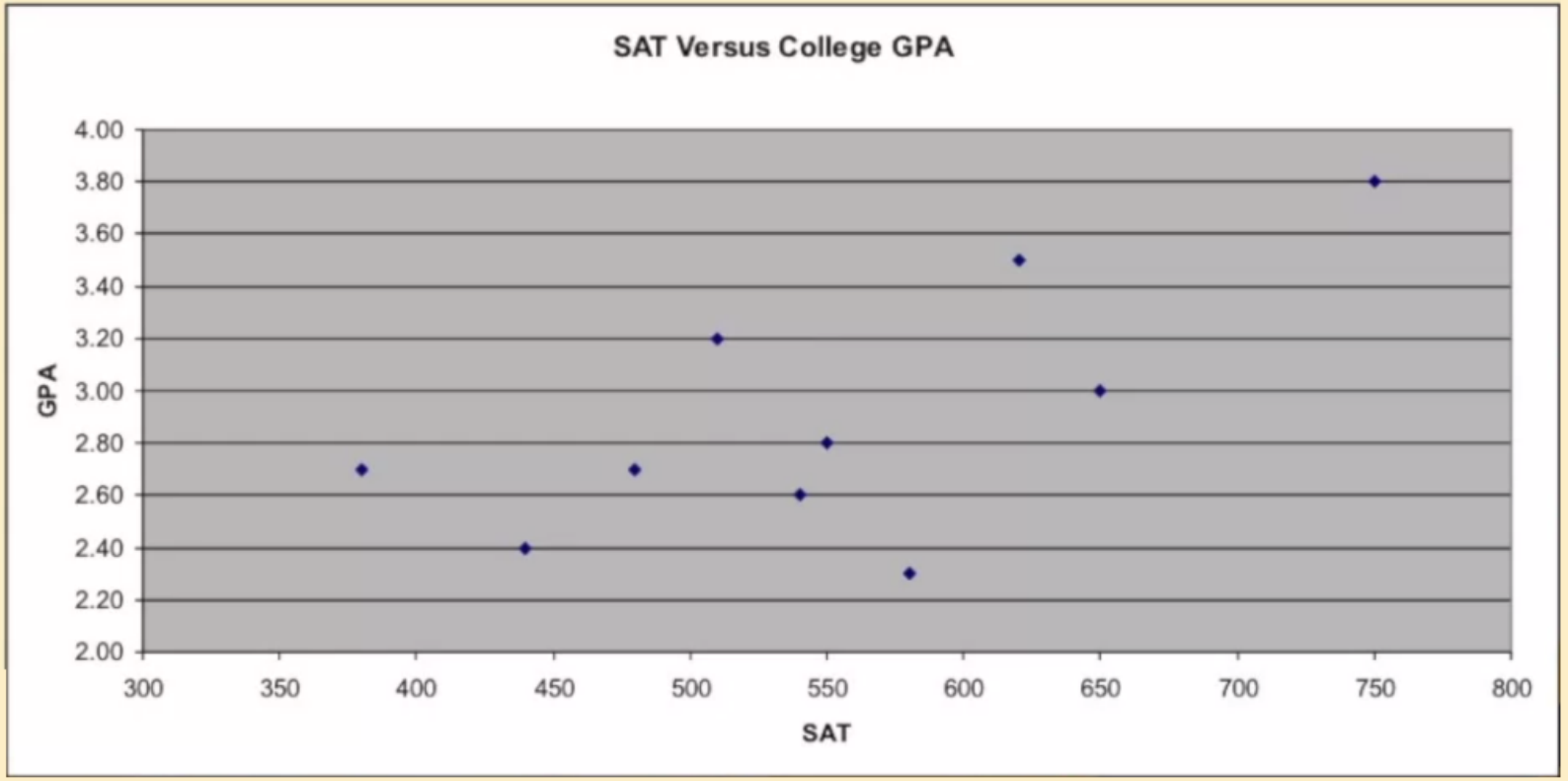


Weak Negative



Scatter Diagram

Positive Relationship

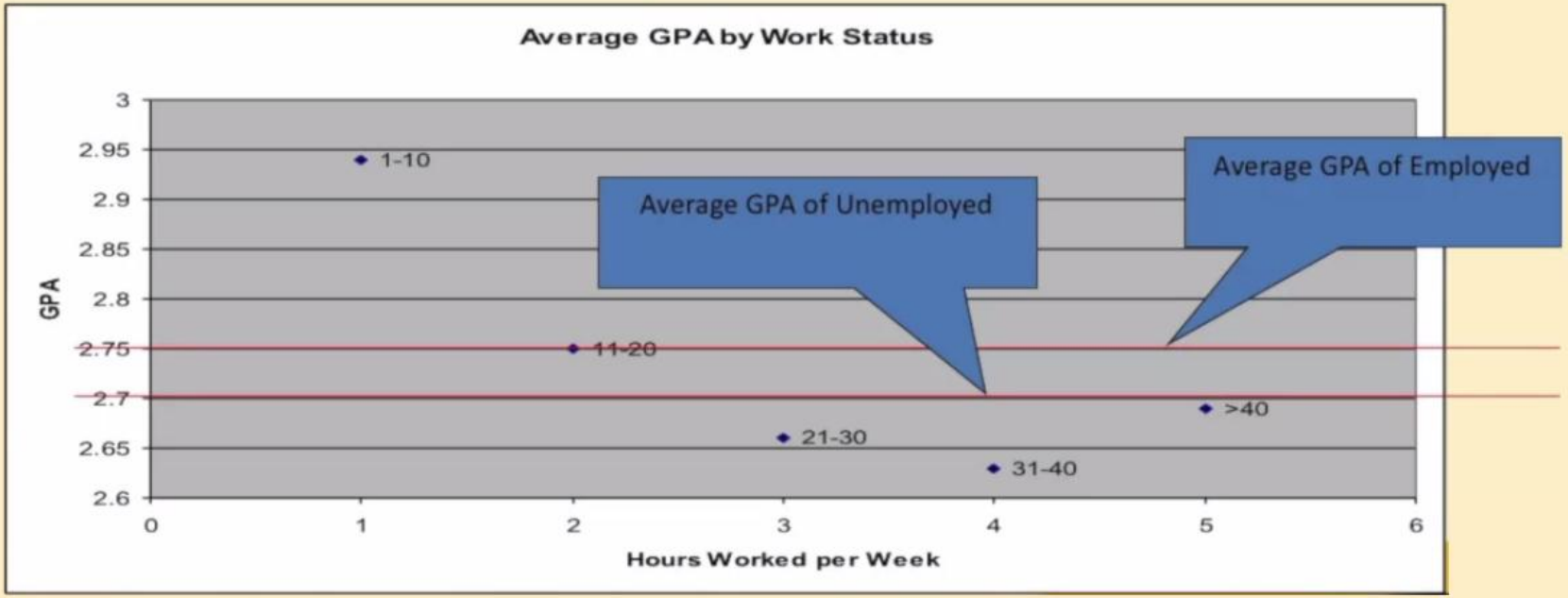


- In this scatter diagram, there appears to be a positive relationship between SAT scores and college grade point averages.
- It's positive but it's not very strong.

- We've so far talked about linear relationships.
- That is, they seem to follow a straight line but that's not always the case.
- Sometimes relationships follow a curved line.

Scatter Diagram

Negative Relationship



This scatter diagram shows the relationship between hours worked per week and grade point average for college students.

It is probably not surprising that as the number of hours increases, the GPA goes down but there's a surprising upturn at the end.

Students working over 40 hours a week seem to do better than those working 20 to 40 hours a week.

Correlation Example

Investigating the correlation between the frequency of exercise and heart health in adults.

Description: The study aims to explore the potential correlation between regular exercise and heart health outcomes in a sample of adults. Researchers collect data on the weekly exercise routines of participants and measure various heart health indicators such as blood pressure, cholesterol levels, and heart rate.

Findings: After analyzing the data, researchers observe a strong negative correlation between the frequency of exercise and elevated blood pressure. Individuals who engage in regular exercise tend to have lower blood pressure levels. However, it's crucial to note that correlation does not imply causation.

Causation vs. Correlation

- **Correlation:**
  - *Definition:* Correlation measures the statistical association between two variables. It describes how changes in one variable are associated with changes in another.
  - *Example:* There is a positive correlation between ice cream sales and drowning incidents. Both variables increase during the summer, but one does not cause the other.
- **Causation:**
  - *Definition:* Causation implies a cause-and-effect relationship between two variables. Changes in one variable directly lead to changes in another.
  - *Example:* A study finds a causal relationship between smoking and lung cancer. Smoking is identified as a direct cause of increased lung cancer risk.

Key Differences

- Correlation does not imply causation. Even if two variables are correlated, it does not mean that one causes the other.
- Causation requires evidence of a direct cause-and-effect relationship, often established through controlled experiments.
- Correlation is a statistical measure, while causation involves establishing a logical connection between variables.

Conclusion – Scatter Diagram

- Like histograms, scatter diagrams provide a useful snapshot. It's important to remember that this is only a snapshot.
- More conclusive analysis can be done with correlation and regression. Correlation tells you if there is a relationship between two variables.
- We can tell how strong the relationship is but we don't know if x causes y, if y causes changes in x, or if something else causes changes in both of them.
- If we want to know more, we need to do more analysis. Correlation and regression analysis is a topic for future courses but for now, it's enough to know that correlation tells you the direction and the strength of a relationship.