Chapter 6 Statistics

Statistics - numerical data, or the collection, organization, and analysis of numerical data.

First, write a **hypothesis (a theory or statement)** that clearly states what you want to prove or disprove. For example, you could start with the hypothesis that the accident rate for red cars is higher than that for other cars. You could then use data from accident reports or insurance claims to see if your hypothesis is correct.

You need data to test a hypothesis. Researchers must decide whether to collect new data or use data that other people have already collected. **Primary data** come from experiments and surveys done by the researchers. Researchers can find **secondary data** in sources such as publications, the Internet, and surveys done by Statistics Canada.

Examples for primary data: a) Daniel telephoned 100 families in his town to ask them how many pets they have. b) Tomas checked the Web sites of 24 stores for the price of the latest Harry Potter DVD.

Examples for secondary data: a) Cathy used data from Statistics Canada to determine the proportion of households in Canada that have at least one car. b) Anja found a Web site with the results from a survey on the spending habits of teenagers across Canada.

Survey - A question or questions asked to a sample a population.

Sample - any group of people or items selected from a population.

Population - the whole group of people or items being studied.

Census - a survey of all members of a population

Example: Identify the population in each situation. Then, indicate whether each researcher should survey a sample of the population or do a census.

a) A teacher wishes to know how early his students wake up in the morning.

Solution: The population *is the students in the teacher's class*. He should do a *census* since the population is small and easy to survey.

b) The principal of a school with 2100 students wants to find out how much homework her students have each day.

Solution: The population is *the students in the school*. The principal should use a *sample*, since the school population is quite large and all students in any particular class may have the same amount of homework for that subject.

c) A clothing store needs to find out whether its customers are happy with its service.

Solution: The population is *the store's customers*. A random *sample* is probably best because it could be difficult and time-consuming to reach all of the store's customers.

d) A newspaper wants to know the public's opinion of a federal political party.

Solution: The population is *everyone in Canada*. The newspaper will have to use a *sample* since it is next to impossible to get the opinion of every person in Canada.

e) A polling firm wants to know how people will vote in the next federal election.

Solution: The population *is every person who can vote in the next federal election.* Again, a census is not practical. It will take far less time and expense to interview a *sample* of voters from across the country.

You can never be completely certain that a sample is representative of the population. However, a **random sample** usually gives reasonably accurate results. You can use several different methods to select a random sample.

Random sample - a sample in which all members of a population have an equal chance of being chosen.

Types of Samples

Example: A principal of a school with 1600 students wants to know whether they favour introducing school uniforms. Describe three methods he could use to select a random sample of 200 students.

Simple random sample - choosing a specific number of members randomly from the entire population.

The principal takes an alphabetical list of all the students at the school and numbers the names in sequence. He then uses a graphing calculator or a spreadsheet to generate 200 random numbers between 1 and 1600. He selects the names on the list that correspond to these numbers.

Systematic random sample - choosing members of a population at fixed intervals from a list.

The principal finds a starting point on the list of students by picking a single random number between 1 and 1600. To get a random sample with 200 students, he then selects every eighth name before and after the starting point.

Stratified random sample - dividing a population into distinct groups and then choosing the same fraction of members from each group.

The principal uses lists of the students in each grade. He then randomly selects the same fraction of students from the list for each grade. Since he wants a sample of 200 students out of a total of 1600, he needs to choose 200/1600 = 1/8 of the students in each grade. Thus, if there are 480

students in grade 9, he would randomly select 480 x 1/8 = 60 of these grade 9 students to be part of the sample.

Non-random sample - using a method that is not random to choose a sample from a population. It can be cheaper or more convenient than random sampling, but the results are less likely to be accurate. Samples that are not random may tend to choose a certain type of member from the population. As a result of this **bias**, the sample does not properly represent the whole population.

Bias - error resulting from choosing a sample that does not represent the whole population.

Scatter plots

Scatter plot is a graph to display the relationships between two quantitative variables.

Scatter-plot

- A scatter-plot show the relationship between two quantitative variables measured on the same individuals;
- The values of the independent variable appear on the horizontal axis;
- The values of the dependent variable appear on the vertical axis;
- Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Independent variable – a variable that affects the value of another variable.
Dependent variable – a variable that is affected by some other variable.

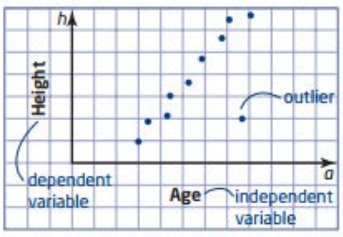
Example: Identify the independent and dependent variable in each situation.

a) Does the outdoor temperature affect the amount of fuel needed to heat a house?
Since you want to know whether the outdoor temperature affects the amount of fuel required for heating, the independent variable is the outdoor temperature and the dependent variable is the amount of fuel required.

b) Is there a relationship between people's ages and their heights?
The independent variable is age and the dependent variable is height.

c) Does the amount of rain in a region depend on its latitude?
The dependent variable is the amount of rain and the independent variable is the latitude. Note that the latitude might not actually affect the amount of rain. However, to analyze the data, you treat latitude as the independent variable and the amount of rain as the dependent variable.

An **outlier** is a point separated from the main body of data on a graph. Sometimes, an outlier results from a measurement error or from some factor that affects only a few of the observed values for a variable. If you can show that an outlier is inaccurate or unrepresentative, you can leave it out of your calculations. Otherwise, you should include the outlier in the data set.

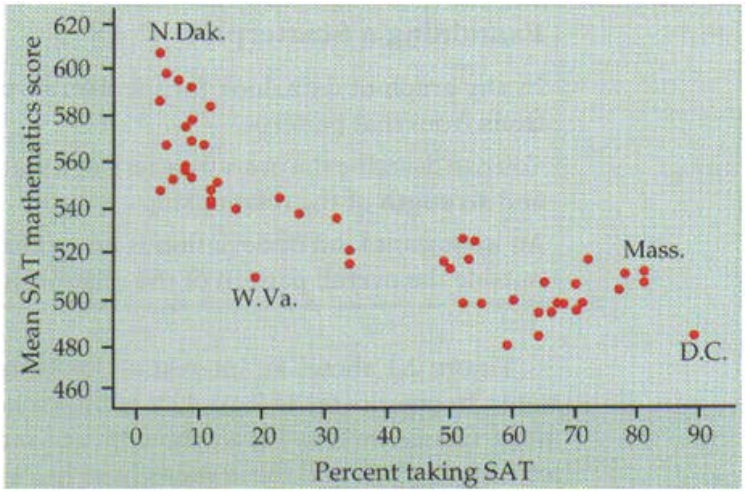


How to Plot a Scatter-plot?

1) Always plot the independent variable, if there is one, on the horizontal axis (the x axis) of a scatter-plot;

2) The explanatory variable (independent) is usually called x and the response variable (dependent) y:

3) If there is no explanatory-response distinction, either variable can go on the horizontal axis.



Each point on the plot represents a single individual-a single state.

The percent taking the exam influences mean score, hence, the percent taking is the explanatory variable, x; and we plot it horizontally. Mean SAT mathematics score is the response variable, y; and we plot it vertically. Therefore, every point of the scatter-plot can be describe by two variable (x; y) = (percent taking SAT; mean SAT mathematics score).
For example, West Virginia appears as the point (19; 511) in the scatter-plot, above 19 on the x axis and to the right of 511 on the y axis.

Analysis of the Scatter-plot is above Figure. State average SAT score is closely related to the percent of students who take the SAT. North Dakota has a high mean score, but only 4 presents of that state's seniors take the SAT. In Massachusetts, 78 percents of seniors take the exam.

Two states that may be outliers have been labeled individually above Figure. The District Columbia is a city rather than a state: 89 percents of its seniors take the SAT. West Virginia appears to have a lower score than we would expect in a state where only 19 percents take the exam.

Interpreting Scatter-plots

Examining a Scatter-plot
1) Look for the overall pattern and for striking deviations from that pattern;

2) Describe the overall pattern of a scatter plot by the form, direction, and strength of the relationship.

3) Look for the outlier, an individual value that falls outside the overall pattern of the relationship.

Interpreting Scatter-plot in above Figure.

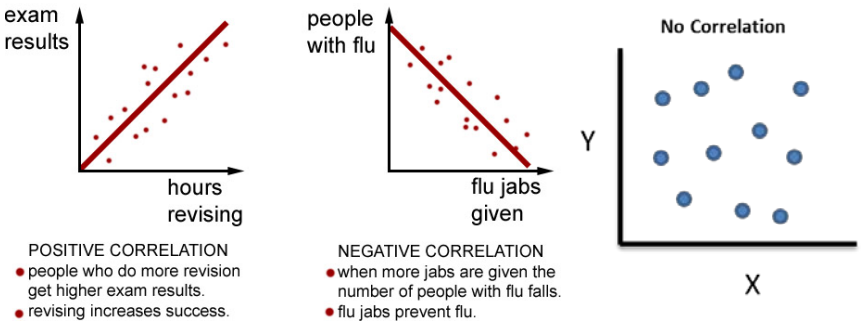
Form: above Figure shows an interesting form-there are two distinct clusters of states. In one cluster, at least 49 percents of high school seniors take the SAT, and the mean score are low. Fewer than 35 percents of seniors in states in the other cluster take the SAT, and these states have higher mean scores.

Direction: above Figure shows a clear direction: states in which a higher percent of students take the SAT tend to have lower mean scores. This is true both between the clusters and within each cluster. This is a negative association between the two variables.

Given the scatter plot, you can **interpolate** an estimate a value between two measurements in a set of data if certain data wasn't surveyed. You can also **extrapolate** – estimate a value beyond the range of a set of data.

A **scatter plot** can be used for data in the form of ordered pairs of numbers. The result will be a bunch of points "scattered" around the plane.

If the general tendency is for the points to rise to the right of the graph, then we say there is a **positive correlation** between the two variables measured. If the points fall to the left of the graph, we say there is **negative correlation**. If there is no general tendency, then there is **no correlation**.



If the tendency is not very pronounced – that is, the points are scattered widely – then we say the variables are **weakly correlated**. If the correlation is more pronounced, we say the variables are **strongly correlated**.

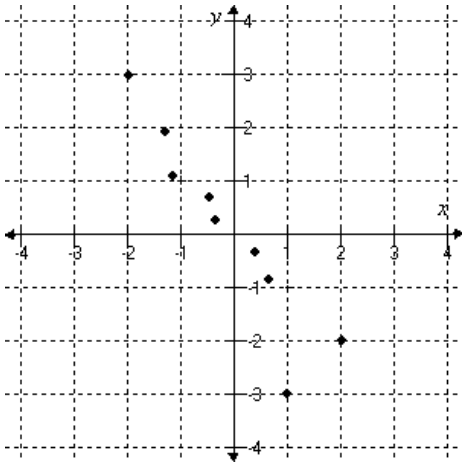
Examples:

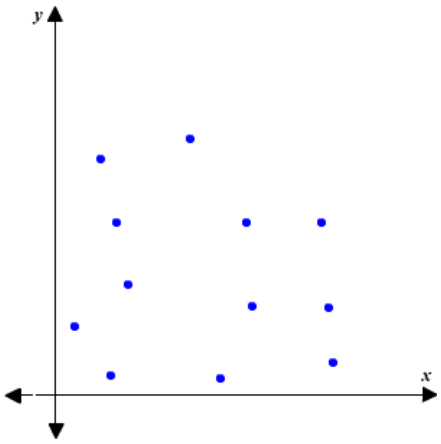
If you graphed a person's height on one axis and their weight on the other, you would probably get a strong positive correlation (because taller people generally weigh more).

If you graphed a man's age and the number of hairs on his head, you would probably get a weak negative correlation (because some men have a tendency for baldness as they get older).

If you graphed a woman's shoe size and the length of her hair, you would probably get no correlation. (These variables are unrelated.)

What correlation can be observed between *x* and *y* in the figure?

1) 

2) 

Find the coordinates of two points that lie on the line best fitting the scatter plot.

It is not possible to tell the exact points that lie on the line of best fit. However, we can estimate the possible points which could be (-2, 3) and (0, 0) for 1). Since there is no scale for 2), possible points can't be found.

How can the relation between the variables in the given scatter plot be best described?

1) Strong negative correlation.
2) Correlation approximately 0. The points shown are scattered randomly. Therefore, the graph has correlation approximately 0.

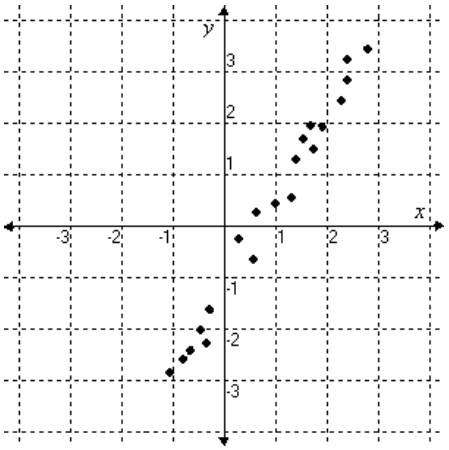
Line of Best Fit

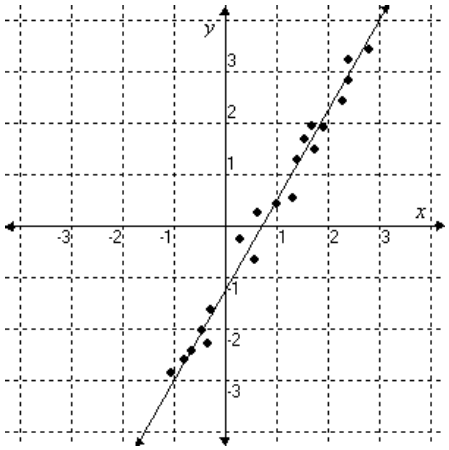
When data is displayed with a **scatter plot**, it is often useful to attempt to represent that data with the equation of a straight line for purposes of predicting values that may not be displayed on the plot.
Such a straight line is called the "**line of best fit**."
It may also be called a "trend" line.
A **line of best fit** is a straight line that best represents the data on a scatter plot.

This line may pass through some of the points, none of the points, or all of the points.

Example 1:

What would be the equation of a line that fits the given scatter plot the best?





Find the coordinates of two points that lie on the line best fitting the scatter plot.

From the scatter plot, the line passing through the points (3, 4) and (–1, –3) seems to best fit the scatter plot.

Use the coordinates to express the equation in the two–point form.
The two–point form of the equation of the line is given by:

$$\frac{y-y_1}{x-x_1}=\frac{y_2-y_1}{x_2-x_1}$$

Substituting the coordinates in the above equation:

$$\frac{y-4}{x-3}=\frac{-3-4}{-1-3} \qquad \frac{y-4}{x-3}=\frac{-7}{-4}$$

Simplify by cross multiplying: $-4(y-4)=-7(x-3)$

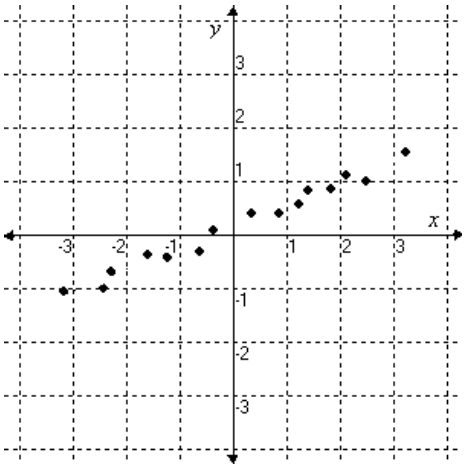
→ $-4y+16=-7x+21$

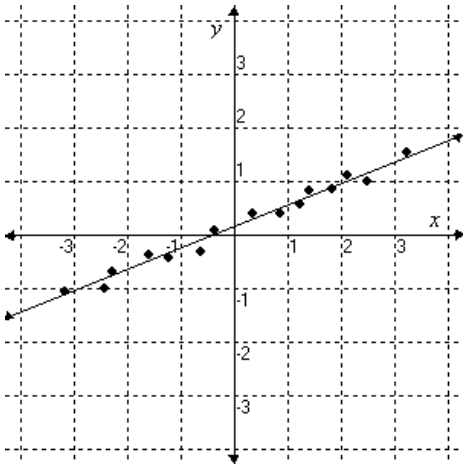
Subtracting 16 from both sides: $-4y=-7x+5$

Dividing throughout by –4: $y=\frac{7}{4}x-\frac{5}{4}$

This is the required equation.

Example 2:
What would be the equation of a line that fits the given scatter plot the best?





Find the coordinates of two points that lie on the line best fitting the scatter plot.

From the scatter plot, the line passing through the points (2, 1) and (–3, –1) seems to best fit the scatter plot.

Use the coordinates to express the equation in the two–point form.

The two–point form of the equation of the line is given by:

$$\frac{y-y_1}{x-x_1}=\frac{y_2-y_1}{x_2-x_1}$$

Substituting the coordinates in the above equation:

$$\frac{y-1}{x-2}=\frac{-1-1}{-3-2} \qquad \frac{y-1}{x-2}=\frac{2}{5}$$

Simplify by cross multiplying: $5(y-1)=2(x-2)$ → $5y-5=2x-4$

Adding 5 to both sides: $5y=2x+1$

Dividing throughout by 5: $y=\frac{2}{5}x+\frac{1}{5}$

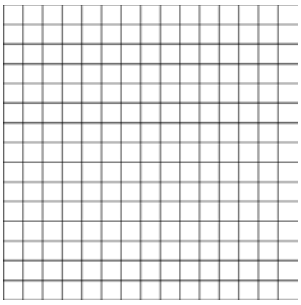
This is the required equation.

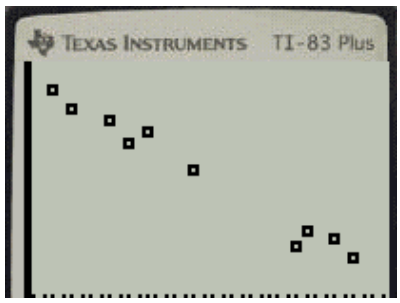
Practice: The following data was collected about the number of bathroom breaks and final mark.

Breaks	3	29	35	5	30	33	18	9	13	11
Mark	93	24	18	86	31	27	58	80	75	70

a. Which measurement is the independent variable? Why?

Number of breaks is the independent variable because mark depends on the number of breaks.
b. Create a scatter plot using the above data.

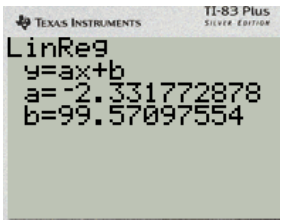


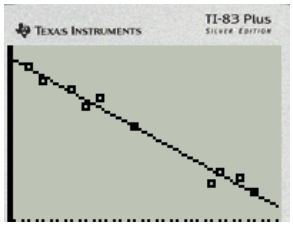


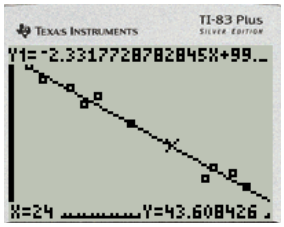
c. Describe the pattern, correlation and trend, if any.

There is a strong negative linear correlation between the number of breaks and marks.

d. Draw the line of best fit and predict your mark if you take 24 bathroom breaks using interpolation.







From the graph we can predict the mark is 43.6 if you take 24 bathroom breaks.
Also from the equation of the line, we can calculate:

y = -2.33x + 99.57
When x = 24, y = -2.33 x 24 + 99.57 = 43.65

e. Determine the equation of the line using two data points that are on the line.

Pick (18, 58) and (35, 18).

Slope: m = 40/-17 = -2.35

Point – slope form: y – 58 = -2.35(x - 18)
➔ y = -2.35x + 100.3

f. Use the equation to predict the bathroom break you can have if you want 95%. What about 100%?

95 = -2.35x + 100.3 → x = 2 breaks
100 = -2.35x + 100.3 → x = 0.13 approximately 0 breaks.

Curve of Best Fit

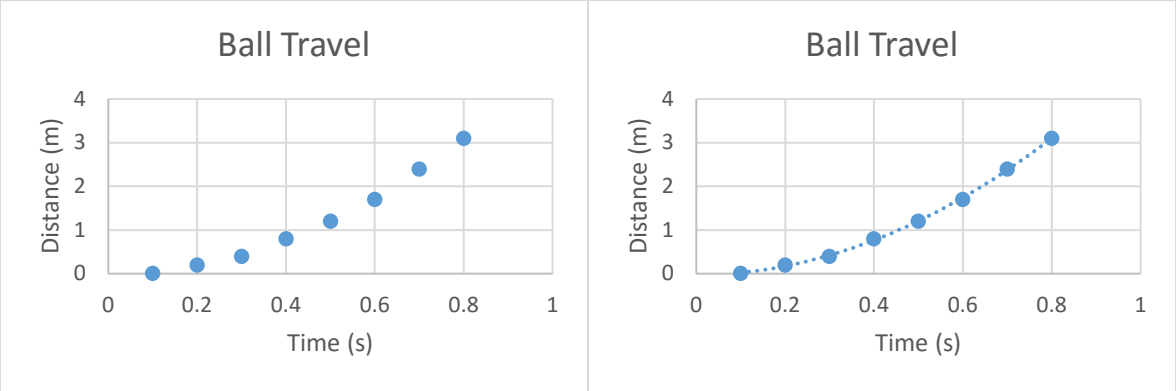
When the scatter plot exhibits a non-linear pattern, it can be approximated by a curve of best fit. We can use the curve of best fit to interpolate and extrapolate values, however it is more challenging to extend the curve for extrapolation. Plugging into the equation of the curve will be a much smarter choice.

Now, we know how to graph it by hand, we can try to use some software such as graphing calculator and Excel to help us plot the points.

Example 1. This table represents the distance that a ball travels when dropped from a 4-m ladder

Time (s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Distance (m)	0.005	0.2	0.4	0.8	1.2	1.7	2.4	3.1

a. Make a scatter plot of the data.



b. Draw a line or curve of best fit.

c. Describe the pattern, relation and the trend of this graph.

The graph is increasing, as time goes, distance increases. The curve of best fit shows a strong positive correlation between time and distance.

d. Do you think this pattern will keep going? Why?

No. The ball will eventually drop to the ground and stop moving. As time goes on, the distance will stay the same after that point.