Statistics of Two Variables

Functions

Y is a **function** of *X* if variable *X* can assume a correspondence to one or more values of *Y*. If only one value of *Y* corresponds to each value of *X*, then we say that *Y* is a **single-valued function** if *X* (also called a "**well-defined function**"); otherwise *Y* is called a **multivalued function** of *X*. Our secondary school curriculum assumes that we mean a well-defined function, when functions are discussed, and we refer to multivalued functions as **relations**. All of the definitions above also assume that we are referring to binary relations (i.e. relations of two variables). The input variable (or **independent variable**) is usually denoted by ***x*** in mathematics, and the output variable (or **dependent variable**) by ***y***. The set of all values of *x* is called the **domain** and the set of all values of *y*, the **range**. As is the case with one variable statistics, the variables can be **discrete** or **continuous**.

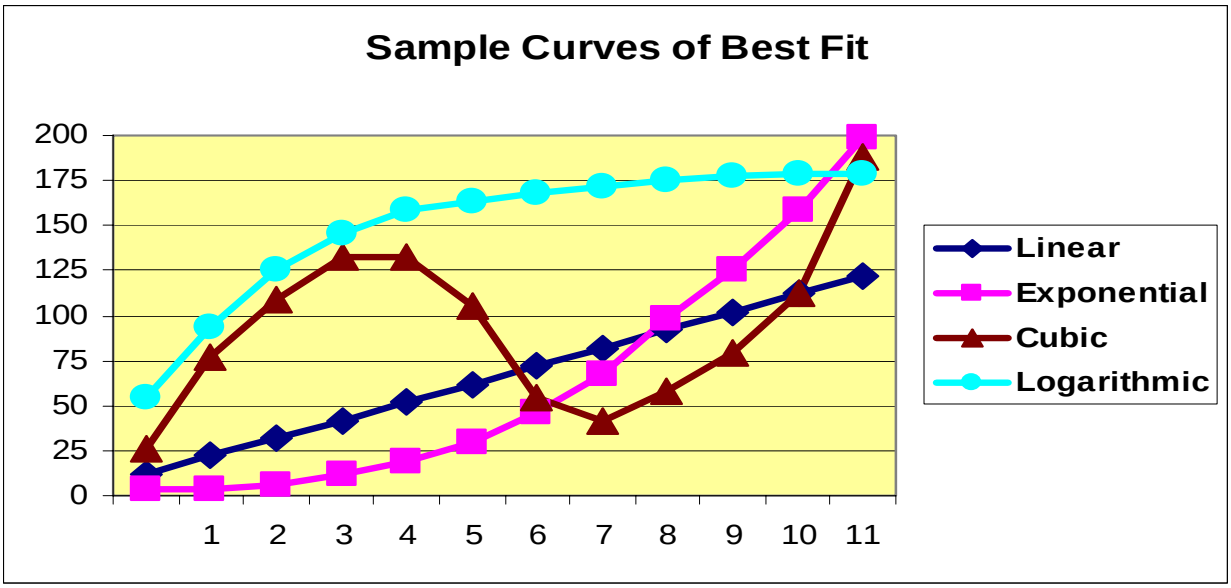
The function dependence or correspondence between variables of the domain and the range can be depicted by **a table**, by **an equation** or by **a graph**. In most investigations, researchers attempt to find a relationship between the two or more variables. We will deal almost exclusively with relations between two variables here. For example the circumference of a circle depends (precisely) on its radius; the pressure of a gas depends (under certain circumstances) on its volume and temperature; the weights of adults depend (to some extent) on their heights. It is usually desirable to express this relationship in mathematical form by finding an equation connecting these variables. In the case of the first two examples, the relationship allows for an exact determination (at least in theory as far as mathematicians are concerned, and within specified error limits as far as scientists are concerned). Virtually all "real life" investigations generate statistical or probability relationships (like the latter example above) in which the resulting function produces only approximate outcomes. Much of our statistical analysis is concerned with the reliability the outcomes when using data to make predictions or draw inferences.

x-variable →	**a "rule" determines what happens here**	→ ***y***-variable
Input →		→ Output
Independent variable →		→ Dependent variable
Control →		→ Response
Cause →		→ Effect

In order to find the **defining equation** that connects variables, a graph (called a **scatter plot**) is constructed and an **approximating curve** is drawn which best approximates the data. This **line of best fit** can be straight (the ideal situation for easy of analysis and computation) or curved. Hence the equation which approximates the data can be linear, quadratic, logarithmic, exponential, periodic or otherwise. Some examples of these are shown on the next page.

Examples of these equations (with *X* the independent variable, *Y* the dependent variable and all other variables as constants) are:

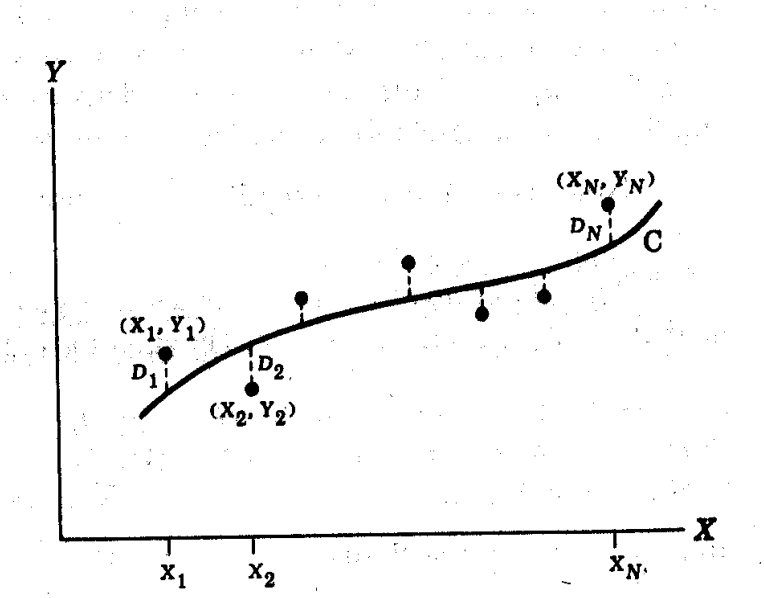
Straight line	$Y = aX + b,\ Y = a_0 + b_0X,\ Y = mX + b,$ etc.
Polynomial	$Y = aX^2 + bX + c,\ Y = aX^3 + b\,X^2 + cX + d,$ etc.
Exponential	$Y = ab^X$ **or** $\log Y = \log a + (\log b)X = a_0 + a_1X$
Geometric	$Y = aX^b$ **or** $\log Y = \log a + b \log X$
Modified exponential	$Y = ab^X + k$
Modified geometric	$Y = aX^b + k$



Statisticians often prefer to eliminate any constant term added to the primary function (as in the last two examples above) through a vertical translation, forcing the curve through the origin. This generally makes for greater ease of analysis – if for no other reason than it eliminates one variable. A wide variety of each of these forms of equations is found in statistical texts so you can expect to see numerous variations of these. The alternate (logarithmic) form of the exponential and geometric functions allow for a useful method of recognizing these relationships when examining data. The use of **semi-log graph paper** and **log-log graph paper** transforms such relations into straight line functions.

Quite frankly, most researchers use a **freehand technique** for constructing the line or curve of best fit. More precise mathematical methods are available but involve extremely long calculations unless electronic devises are used to assist with the process. In general, we require at least as many points on the curve as there are constants in its equation, in order to determine the value of these constants.

To avoid "guessing" in the construction of the best-fitting curve, we require a definition of what constitutes the "best". The most common approach involves finding the squares of the **vertical displacements** from the best line of fit. These distance differences are called **deviations** or **residuals**. The line of best fit will be found when the sum of these squares is a minimum.



In the diagram show, the line or curve having the property that

$D_1^2 + D_2^2 + \ldots + D_N^2$ is a minimum

will be the line or curve of best fit. Such a line or curve is called the best-fitting curve or the least square line. The process for obtaining it is called **least square analysis** and is one part of **regression analysis**. If *Y* is considered to be the independent variable, we obtain a different least squares curve.

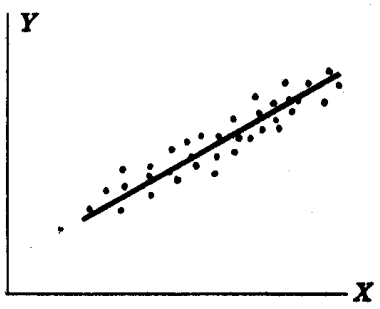
The following formulas are used for determining the straight line of best fit (i.e. to obtain the values for the constants a and b of $Y = a + bX$. We must solve the pair of equations below simultaneously. $\sum Y = aN + b\sum X$ and $\sum XY = a\sum X + b\sum X^2$. Students will usually use graphing calculators or spreadsheets to find this line. Similar equations exist to allow for the determination of least square parabolas, exponential curves, etc., but they are rarely done by hand calculations. The amount of time needed for this process can be reduced by using the transformations $x = X - \overline{X}$ and $y = Y - \overline{Y}$. Here $\overline{X}$ and $\overline{Y}$ are the respective means, and the transformed line will pass through the origin. The least squares line will pass through point $\left(\overline{X}, \overline{Y}\right)$ and this point is called the **centroid** or centre of gravity of the data.

If the independent variable is time, the values of Y will occur at specific times and the data is called a **time series**. The regression line corresponding to this type of relation is called a **trend line** or **trend curve** and is used for making predictions for future occurrences. If more than two variables are involved these can be treated (usually with great difficulty) in a manner analogous to that for two variables. The linear equation for three variables, X, Y and Z is given by: $Z = aX + bY + c$. In a three-dimensional coordinate system, this represents the equation of a plane, called an **approximating plane**, or a **regression plane**.

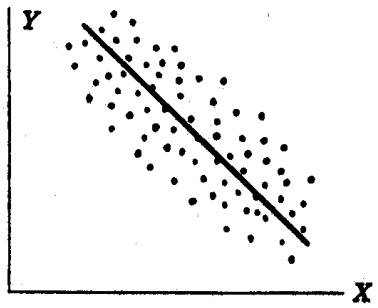
Correlation Theory

Correlation deals with the degree to which two or more variables have a definable relation. The relationship that exists between variables may be precise (perfectly correlated) such as the case with the length of the side of a square and the length of its diagonal or may be uncorrelated (no definable relation) as in the case with the numbers on each die if the dice are tossed repeatedly. Such cases are of great importance in understanding our world and discovering new concepts, but usually bear little resemblance to the analysis of most data sets. When only two variables are involved, the relationship is called **simple correlation** and it is called **multiple correlation** when three or more variables are used. As indicated above, the relationship which best describes the correlation between two variables can be linear, quadratic, exponential, etc. Because of the extreme complexity of the mathematical evaluation techniques, it is most desirable to consider linear correlation with respect to two variables. It is often possible to transform other defining relationships into straight line, simple correlation by means of a variety of mathematical or subjective judgment revisions.

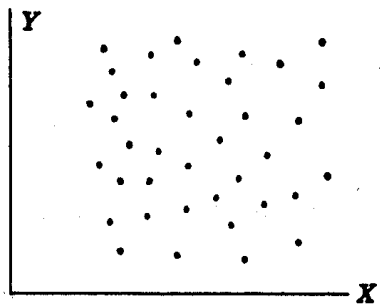
In dealing with linear correlation only, there are three general cases to consider (as outlined below):



positive correlation



negative correlation



no correlation

We could further denote the first two cases illustrated above as depicting **strong** linear correlation and **weak** (or moderate) linear correlation, respectively. Of course, mathematicians require that we define these cases by means of a quantitative rather than qualitative measure.

Indeed, the first reaction which students (and non-mathematicians in general) will reveal when describing data sets is to resort to qualitative analysis rather than quantitative analysis. There are many different types of regression analysis available (most requiring advanced skill levels in mathematics) but the most common is through the **least square regression line** described in the previous section.

We will need to consider both the regression lines of *Y* on *X* and of *X* on *Y* here. These are defined as: $Y = a_0 + a_1 X$ and $X = b_0 + b_1 Y$. The method for determining the values of the constants used here was shown in the previous section given above. It is important to understand that these regression lines are identical *only if* all points from the data set lie precisely on the line of best fit. This would be the case for the example of the relation between the side and the diagonal of a square given earlier, but we would almost never rely on regression analysis for determining such relations.

The variables *X* and *Y* (the left hand sides of the two linear equations given above) can be better described as estimates of the value predicted for *X* and for *Y*. For this reason they are also referred to as Y_{est} and X_{est} for the purpose of determining the error limits defining the reliability of the data (or the predicted values resulting from the data). We now define the **standard error estimate** of *Y* on *X* as: $s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{est})^2}{N}}$. An analogous formula is used for the standard error estimate of *X* on *Y*. Once again, it is important to note that $s_{X.Y} \neq s_{Y.X}$ here.

The quantity $s_{X.Y}$ has properties that are analogous to those of the standard deviation. Indeed, if we construct lines parallel to the regression line of *Y* on *X* at respective vertical distances $s_{Y.X}$, $2s_{Y.X}$ and $3s_{Y.X}$ we find that approximately 68%, 95% and 99.7% of the sample points lie between these three sets of parallel lines. These formulas are only applied to data sets for which $N > 30$. Otherwise the factor *N* in the denominator is replaced with $N - 2$ (as in single variable analysis). The "2" is used here because there are two variables involved. The total variation of *Y* on *X* is defined as $\sum (Y - \bar{Y})^2$ (i.e. the sum of the squares of the deviations of the values of *Y* from the mean $\bar{Y}$).

One of the theories in correlation study contends that $\sum (Y - \bar{Y})^2 = \sum (Y - Y_{est})^2 + \sum (Y_{est} - \bar{Y})^2$. Here the first term on the right is called the **unexplained variation** while the second term on the right is called the **explained variation**. This results from the fact that the deviations $(Y_{est} - \bar{Y})$ have a definite (mathematically predictable) pattern while the deviations $(Y - Y_{est})$ behave in a random (unpredictable) pattern.

The ratio of the explained variation to the total variation is called the **coefficient of determination**. If there is zero explained variation, (i.e. all unexplained) the ratio is zero. Where there is zero unexplained variation (i.e. all explained) the ratio is one. Otherwise the ratio will lie between zero and one. Since this ratio is always positive, it is denoted by $\boldsymbol{r}^2$. The resulting quantity ***r***, called the **coefficient of correlation**, is given by: $r = \pm\sqrt{\frac{\text{explained variation}}{\text{total variation}}} = \pm\sqrt{\frac{\sum (Y_{est} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}}$. This quantity must always lie between 1 and −1. Note that ***r*** is a dimensionless quantity (independent of whatever units are employed to describe *X* and *Y*).

This quantity can be defined by various other formulas. By using the definition of $s_{Y.X}$ above, and the standard deviation for *Y*, $s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}}$, we also have $r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}}$. While the regression lines of *Y* on *X* and *X* on *Y* are not the same (unless the data correlates perfectly) the values of ***r*** are the same regardless of whether *X* or *Y* is considered the independent variable. These equations for the correlation coefficient are general and can be applied to non-linear relationships as well. However, it must be noted that the Y_{est} is computed from non-linear regression equations in such cases, and by custom, we omit the ± symbols for non-linear correlation.

The **coefficient of multiple correlation** is defined by the extension of the formulas above. In the most common case (one dependent variable *Y*, and two independent variables, *X* and *Z*) the coefficient of multiple correlation is: $R_{Y.XZ} = \sqrt{1 - \frac{s_{Y.XZ}^2}{s_Y^2}}$. Again, this can apply to non-linear cases, but the computations involved in generating the required mathematical expressions and the best curve itself are frightening to say the least! It is much more common to approach the situation by considering the correlation between the dependent variable and one (primary) independent variable while keeping all other independent variables constant. We denote $r_{12.3}$ as the correlation coefficient between X_1 and X_2 while keeping X_3 constant. This type of analysis is defines the correlation coefficient as **partial correlation**.

When considering data representing two variables, *X* and *Y*, we generally use a (shorter) computation version of the formula for finding *r*, namely: $r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\left[N\sum X^2 - (\sum X)^2\right]\left[N\sum Y^2 - (\sum Y)^2\right]}}$.

This formula can also appear as $r = \frac{\sum XY - \frac{1}{N}(\sum X)(\sum Y)}{(N-1)s_X s_Y}$ when $N < 30$ (the sample size).

The equation of the least square line $Y = a_0 + a_1 X$ (the regression line of *Y* on *X*) can be written as: $Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X})$. Similarly the regression line of *X* on *Y* is $X - \bar{X} = \frac{rs_X}{sY}(Y - \bar{Y})$. The slopes of these two lines are equal if and only if $r = \pm 1$. In this case the lines are identical and this occurs if there is perfect linear correlation between *X* and *Y*. If $r = 0$ the lines are at right angles and no linear correlation exists between *X* and *Y*. Note that if the lines are written as $Y = a_0 + a_1 X$ and $X = b_0 + b_1 Y$ then $a_1 b_1 = r^2$.

Another useful method for finding the correlation of the variables is to consider only their position if ranked (either in ascending or descending order). Here the **coefficient of rank correlation** is given by $r_{rank} = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$. Here, *D* represents the differences between the ranks of corresponding values of *X* and *Y* and *N* is the number of pairs of data. If data are tied in rank position, the mean of all of the tied positions is used. This formula is called **Spearman's Formula for Rank Correlation**.

It is used when the actual values for *X* and *Y* are unknown or to compensate for the fact that some values for *X* and/or *Y* are extremely large or small in comparison to all others in the set.

The least squares regression line, $Y = a + bX$, can also be determined from the value for *r* as follows: $b = r\left(\frac{s_Y}{s_X}\right)$ and $a = \overline{Y} - b\overline{X}$. Note that *b* is the slope and *a* is the *y*-intercept in this case.

When the dependent variable indicates some degree of correlation with two or more independent variables, most researchers will make a definitive attempt to reduce the relationship to one which relates to just one independent variable. This can be accomplished in several ways, including:

- Ignore all but one independent variable (somewhat justifiable if $|r| < 0.7$) for all but one independent variable – the one which is thus used for the investigation.
- Gather data only from subjects that have roughly the same characteristics or values for all but the one variable being studied.
- Combine two or more of the significant (as defined by the "rule of thumb" above) independent variables to form one new formula (and one new independent variable) and hence reduce the relation to a simple correlation analysis.
- Assume that all but one of the independent variables are constants, but adjust (through transformations) the value of this one *x*-variable in terms of these (assumed) constants.

It is understood that the research reviewer (or teacher assessor) will make every effort to confirm that the unused independent variables have been disposed of in an adequate fashion and also identify any other independent variables that might have been overlooked by the researcher.

It is essential to understand that even when the correlation between variables is very strong (approaching 1 or −1) that this does **not** mean that a causal relationship exists between the variables. This is often stated philosophically as **"correlation does not imply causation"**.

A correlation relationship simply says that two things perform in a synchronized manner. A strong correlation between the variables may occur only by coincidence, or as a result of both variables sharing a common cause (another variable, often designated as a hidden or lurking variable). Several types of reasons are often stated for apparent causation relationships between variables demonstrating moderate or strong correlation other than a direct cause and effect relationship. These include:

- **Reverse causation** (the effect of smoking tobacco vs. the incidence of lung cancer)
- **Coincident causation** (the size of children's feet vs. their ability to spell)
- **Common-cause causation** (the price of airline tickets and baseball players' salaries)
- **Confounding causation** (often denoted as "presumed", "unexplained", "placebo effect", etc.)

Another type of problem in determining correlation coefficients involves outliers. **Outliers** are atypical (by definition), infrequent observations. Because of the way in which the regression line is determined (especially the fact that it is based on minimizing the sum of *squares of distances* of data points from the line), outliers have a profound influence on the slope of the regression line and consequently on the value of the correlation coefficient. A single outlier is capable of considerably changing the slope of the regression line and, consequently, the value of the correlation coefficient for those data points. Some researchers use quantitative methods to exclude outliers. For example, they exclude observations that are outside the range of ±2 standard deviations (or even ±1.5 standard deviations) around the group or design cell mean. In some areas of research, such "cleaning" of the data is absolutely necessary.

We will discuss (and take up some of) the following problems at the session on Saturday. You can work them out in advance if you like. You may wish to have a graphing calculator (the TI-83 is fine here) or a lap-top with a spread-sheet software (Excel is the most common) with you on Saturday, to follow along with the calculations. We will refer only briefly to the definitions outlined above.

1.) For the data in the table at right:

Person	Height (inches)	Self Esteem*
A	68	4.2
B	71	4.4
C	62	3.6
D	74	4.7
E	58	3.1
F	60	3.3
G	67	4.2
H	68	4.1
I	71	4.3
J	69	4.1
K	68	3.7
L	67	3.8
M	63	3.5
N	62	3.3
O	60	3.4
P	63	4.4
Q	65	3.9
R	67	3.8
S	63	3.4
T	61	3.3

* using a survey rating between 1 to 5 where 1 is very low and 5 is very high

a) Determine the correlation coefficient using:

i) height as the independent variable and self-esteem as the dependent variable

ii) self-esteem as the independent variable and height as the dependent variable

Construct two scatter plots to illustrate the data (using height and then self-esteem as the independent variable).

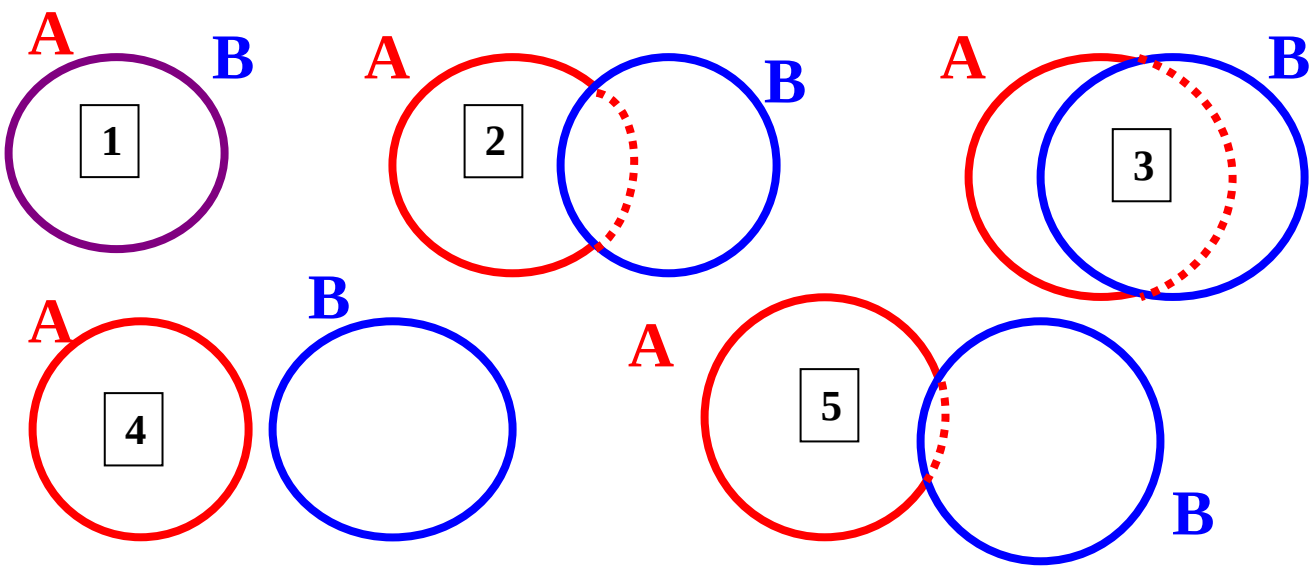
b) Identify any outlier(s) in the data set. Remove these and re-calculate the correlation coefficient, *r*.

c) Use these calculations to identify the mathematical relationship between height and self-esteem.

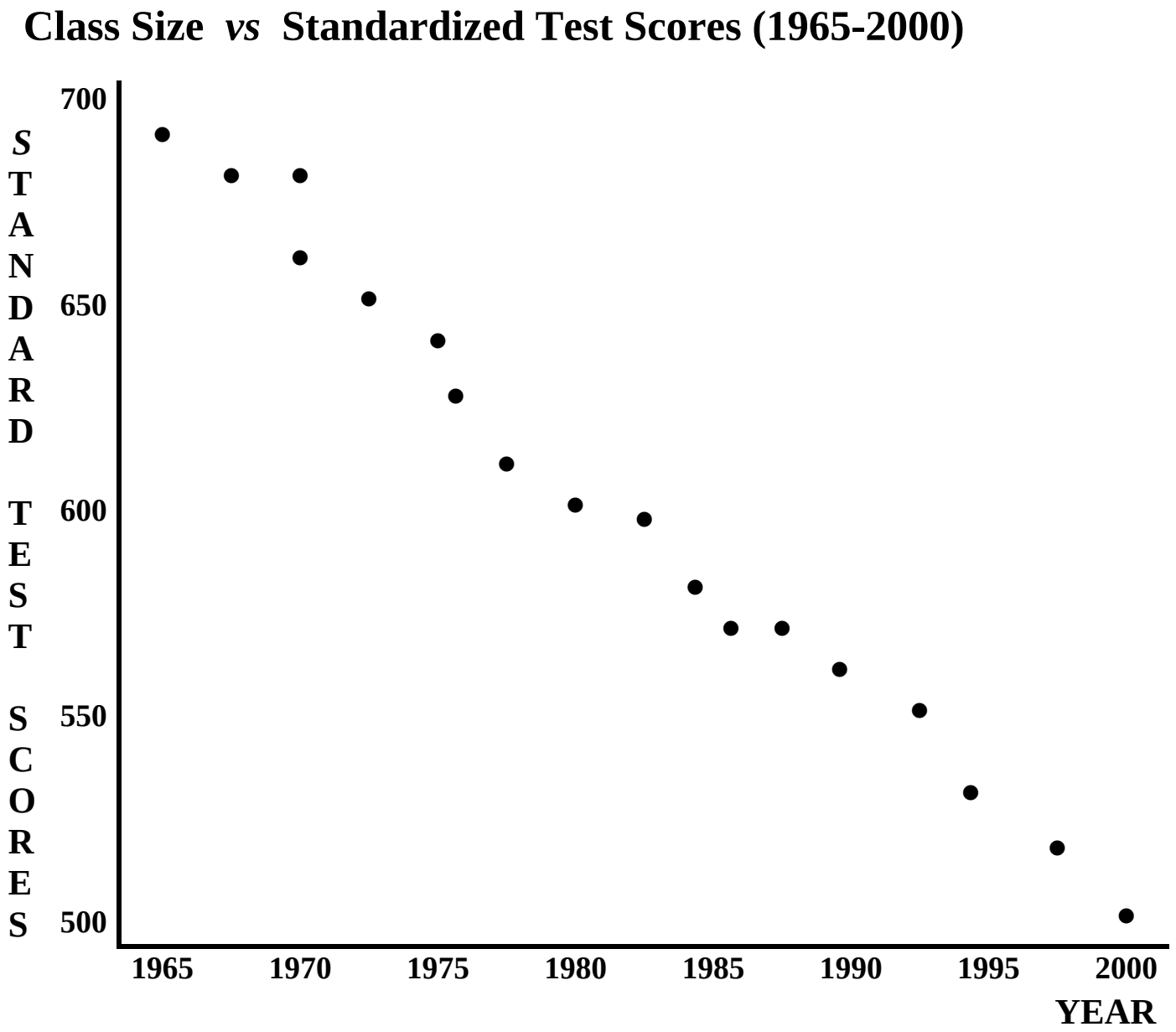
d) Speculate on the causal relationship involved here.

2.) Categorize each of the following Venn diagrams (1. to 5. shown) as representing:

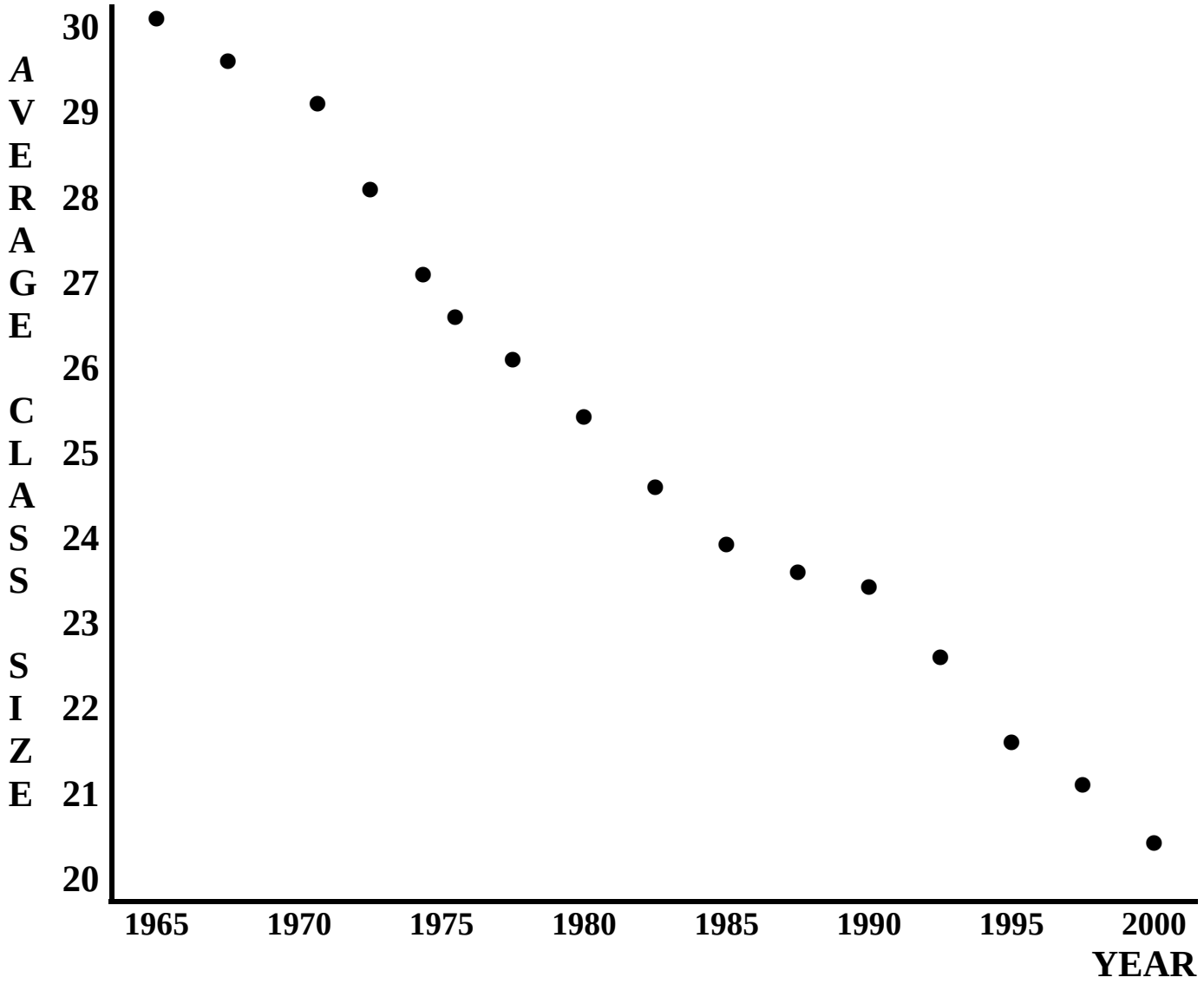
a) independent, dependent or mutually exclusive events
b) strong positive, strong negative or no correlation between A and B



3.) Is there a causal relationship between the variables depicted in the two scatter plots given below?



Class Size *vs* Standardized Test Scores (1965-2000)



4.) The winning times of the men's 400 m race at the Olympic Summer Games from 1900 to 2000 are listed in the table below. The times are measured in seconds.

Year	1896	1900	1904	1906	1908	1912	1920	1924	1928
Time	54.2	49.4	49.2	53.2	50.0	48.2	49.6	47.6	47.8
Year	1932	1936	1948	1952	1956	1960	1964	1968	1972
Time	46.2	46.5	46.2	45.9	46.7	44.9	45.1	43.86	44.26
Year	1976	1980	1984	1988	1992	1996	2000	2004	
Time	44.26	44.60	44.27	43.87	43.50	43.49	43.84	44.00	

a) Identify the independent and the depend variables.
b) Construct a scatter gram to illustrate the data given above. Label fully.
c) Draw the line of best fit in your graph.
d) Which points in your scatter gram lie the greatest distance from the line of best fit? Give possible reasons for the fact that the winning times for the men's 400 m for those years are so far from their expected values.
e) Determine the values for m and b in the equation $y = mx + b$ that represents your line of best fit drawn in part b. above. Show all calculations.
f) What does the slope represent in your graph?
g) What does the y-intercept represent in your graph?
h) Explain why the Olympic Summer Games were not held in 1916, 1940 and 1944.
i) Predict what the winning time would have been in 1944, had the Olympic Games been held that year. Show all calculations.
j) Toronto lost its bid to host the Olympic Summer Games for both 1996 and 2008. The city plans to put forth a bid for the 2020 games. Predict what the winning time will be in the men's 400 m at the 2020 Olympic Summer Games. Show all calculations.

5.) The table given below lists the average cost for producing a motion picture (in the USA) for selected years since 1935. The figures are in millions of dollars (US funds). (Source: International Motion Picture Almanac)

Year	1972	1974	1976	1978	1980	1982	1984	1986
Average Cost	1.89	2.01	4.0	4.4	5.0	11.8	14.0	20.0
Year	1988	1990	1992	1994	1996	1998	2000	
Average Cost	18.1	26.8	26.9	29.2	33.6	52.0	55.6	

a) Identify the independent and the depend variables.
b) Construct a scatter gram to illustrate the data given above. Label fully.
c) Draw the line of best fit in your graph.
d) Determine the values for m and b in the equation $y = mx + b$ that represents your line of best fit drawn in part b. above. Show all calculations.
e) What does the slope represent in your graph?
f) Does the y-intercept of your graph have any meaning?
g) Describe the type of correlation represented by this data.
h) From your equation found in part c. above, predict what the average cost of producing a motion picture will be in the year 2005. Show all calculations.
i) From your equation found in part c. above, predict the year when the average cost of producing a motion picture will reach 200 million dollars. Show all calculations.
j) Give some reasons why the cost of producing a motion picture has risen in recent years.

6.) The table given below lists the population of the world for the years 1950 to 2005 (the last year is estimated. The figures are in billions. (Source: U.S. Bureau of the Census, Intern'al Data Base)

Year	1940	1945	1950	1955	1960	1965	1970	1975
Population	2.074	2.208	2.555	2.780	3.040	3.346	3.708	4.087
Year	1980	1985	1990	1995	2000	2005		
Population	4.454	4.853	5.285	5.696	6.085	6.450		

a) Identify the independent and the depend variables.
b) Construct a scatter gram to illustrate the data given above. Label fully.
c) Draw the line of best fit in your graph.
d) Determine the values for m and b in the equation $y = mx + b$ that represents your line of best fit drawn in part b. above. Show all calculations.
e) What does the slope represent in your graph?
f) Does the y-intercept of your graph have any meaning?
g) Describe the type of correlation represented by this data.
h) From your equation found in part c. above, predict the year when the population of the world will reach 8 billion people. Show all calculations.
i) From your equation found in part c. above, predict what the population of the world will be in the year 2050. Show all calculations.

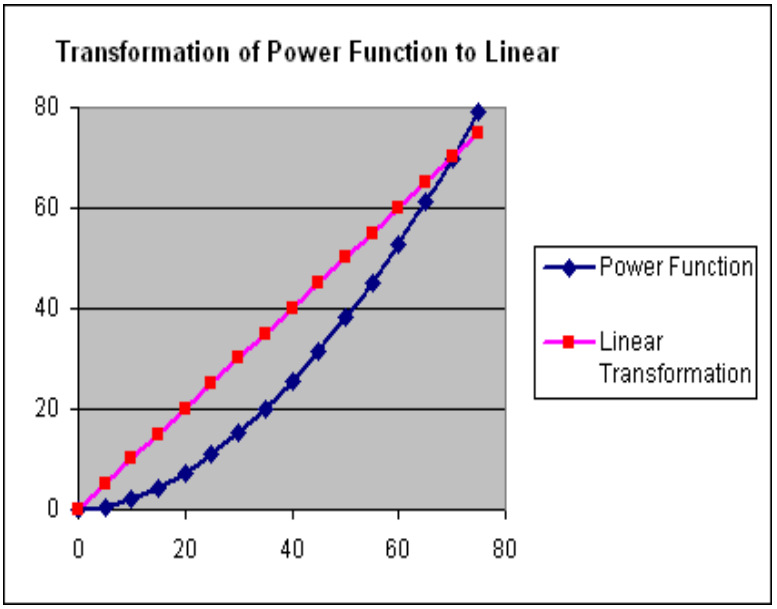
7.) The table given below lists the number of days that the ground level ozone reading exceeded the acceptable safety level for all centres in Canada for which measurements are taken. The acceptable ozone level is considered to be 82 parts per billion for one-hour average levels during the day from May to September. (Source: Statistics Canada)

Year	1980	1981	1982	1983	194	1985	1986	1987
Number of days	15.9	22.2	13.2	17.5	24.8	8.8	9.4	10.9
Year	1988	1989	1990	1991	1992	1993	1994	1995
Number of days	26.0	10.9	9.4	17.3	6.4	5.5	5.4	8.3
Year	1996	1997	1998	1999	2000	2001	2002	
Number of days	4.5	4.6	4.0	3.5	4.2	3.3	3.1	

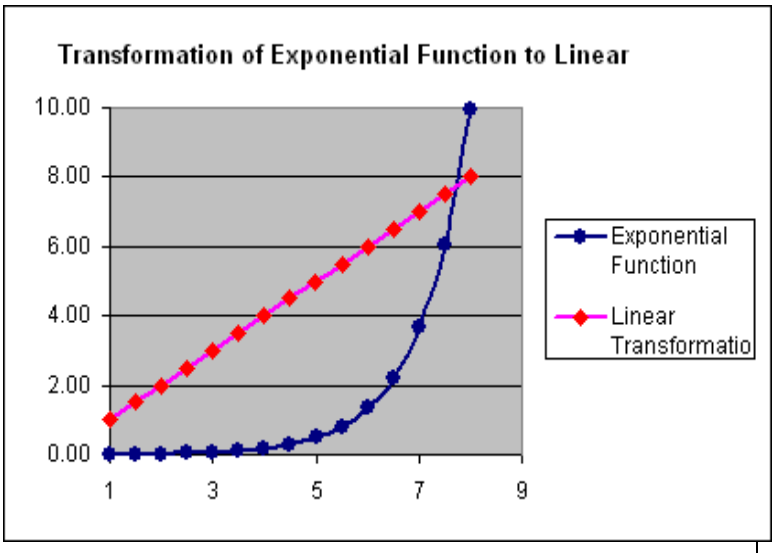
a) Identify the independent and the depend variables.
b) Construct a scatter gram to illustrate the data given above. Label fully.
c) Draw the line of best fit in your graph.
d) Determine the values for m and b in the equation $y = mx + b$ that represents your line of best fit drawn in part b. above. Show all calculations.
e) What does the slope represent in your graph?
f) What does the y-intercept represent in your graph?
g) From your equation found in part c. above, predict the number of days that the ground level ozone readings will exceed the acceptable minimum levels in Canada for the year 2008. Show all calculations.
h) Which points in your scatter gram lie the greatest distance from the line of best fit? Give possible reasons for the significant variation of the ozone readings for these years from the line of best fit.
i) Describe the type of correlation represented by this data.
j) From the readings in this graph, do you think that Canada is doing enough to reduce the level of ozone emissions over the last few decades?

Transforming Curves To Linear Functions For Analysis

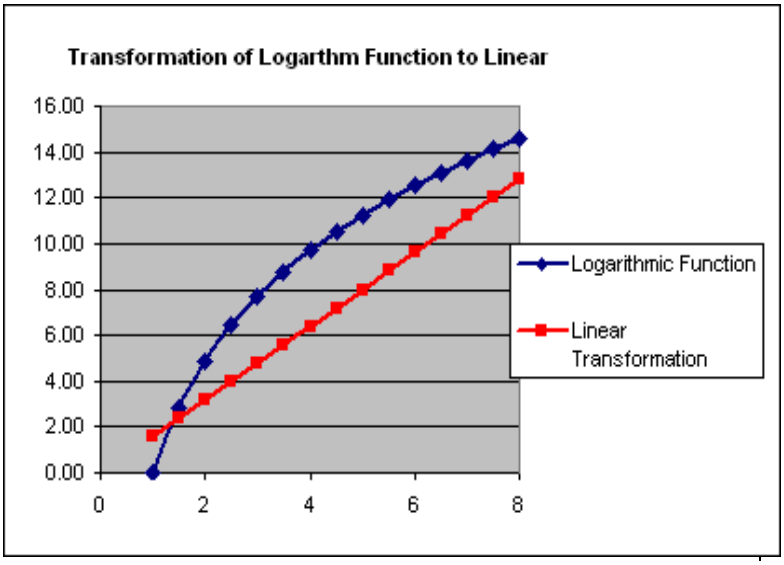
X	*Y*	*Y-revised*
0	0.00	0.0
5	0.60	5.0
10	2.10	10.0
15	4.36	15.0
20	7.32	20.0
25	10.94	25.0
30	15.19	30.0
35	20.05	35.0
40	25.50	40.0
45	31.53	45.0
50	38.11	50.0
55	45.24	55.0
60	52.91	60.0
65	61.11	65.0
70	69.83	70.0
75	79.07	75.0



X	*Y*	*Y-revised*
1	0.01	1.0
1.5	0.01	1.5
2	0.02	2.0
2.5	0.04	2.5
3	0.07	3.0
3.5	0.11	3.5
4	0.18	4.0
4.5	0.30	4.5
5	0.49	5.0
5.5	0.82	5.5
6	1.34	6.0
6.5	2.22	6.5
7	3.66	7.0
7.5	6.03	7.5
8	9.94	8.0



X	*Y*	*Y-revised*
1	0.00	1.6
1.5	2.84	2.4
2	4.85	3.2
2.5	6.41	4.0
3	7.69	4.8
3.5	8.77	5.6
4	9.70	6.4
4.5	10.53	7.2
5	11.27	8.0
5.5	11.93	8.8
6	12.54	9.6
6.5	13.10	10.4
7	13.62	11.2
7.5	14.10	12.0
8	14.56	12.8



2006 American League Baseball Batting Statistics

Team	*Y* Runs Scored	X_1 Home Runs	X_2 Bases On Balls	X_3 Total Bases	X_4 On Base Percentage	X_5 Stolen Bases	X_6 Sacrifice Bunts	X_7 Slugging Average	X_8 Batting Average	*Formula 1* Slugging *times* Total bases	*Formula 2* On-base Ave. *plus* Slugging
Minnesota	**801**	143	490	2380	.347	101	31	.425	.287	**1011.500**	**.7720**
New York	**930**	210	649	2607	.363	139	34	.461	.285	**1201.827**	**.8240**
Toronto	**809**	199	514	2590	.348	65	16	.463	.284	**1199.170**	**.8110**
Cleveland	**870**	196	556	2569	.349	55	30	.457	.280	**1174.033**	**.8060**
Chicago	**868**	236	502	2625	.342	93	44	.464	.280	**1218.000**	**.8060**
Texas	**835**	183	505	2523	.338	53	18	.446	.278	**1125.258**	**.7840**
Baltimore	**768**	164	474	2376	.339	121	40	.424	.277	**1007.424**	**.7630**
Los Angeles	**766**	159	486	2383	.334	148	31	.425	.274	**1012.775**	**.7590**
Detroit	**822**	203	430	2531	.329	60	45	.449	.274	**1136.419**	**.7780**
Seattle	**756**	172	404	2406	.325	106	38	.424	.272	**1020.144**	**.7490**
Kansas City	**757**	124	474	2296	.332	65	52	.411	.271	**943.656**	**.7430**
Boston	**820**	192	672	2445	.351	51	22	.435	.269	**1063.575**	**.7860**
Oakland	**771**	175	650	2264	.340	61	25	.412	.260	**932.768**	**.7520**
Tampa Bay	**689**	190	441	2298	.314	134	35	.420	.255	**965.160**	**.7340**
Correlation (Y vs X_N)		**0.5650**	**0.5035**	**0.8389**	**0.8180**	**-0.1960**	**-0.1396**	**0.8095**	**0.6990**	**0.8291**	**0.9195** ***Best Formula***

The least squares regression line is given by:

$Y = -13.4 + 5500X$

(Note that $S_y = 60.43$, $S_x = 0.0279$, mean(Y) = 804.4 and mean(X) = 0.7762)

GCF and LCM – another technique to try:

2	**48**	**56**	**84**	**2**	**36**	**54**		**280**	**420**	**630**
2	24	28	42	**3**	18	27				
3	12	**14**	21	**3**	6	9				
2	4	14	**7**		**2**	**3**				
7	**2**	7	7							
	2	**1**	**1**							
GCF	= 2 × 2 = **4**			**GCF**			**GCF**			
LCM	= 2×2×3×2×7×2 = **336**			**LCM**			**LCM**			

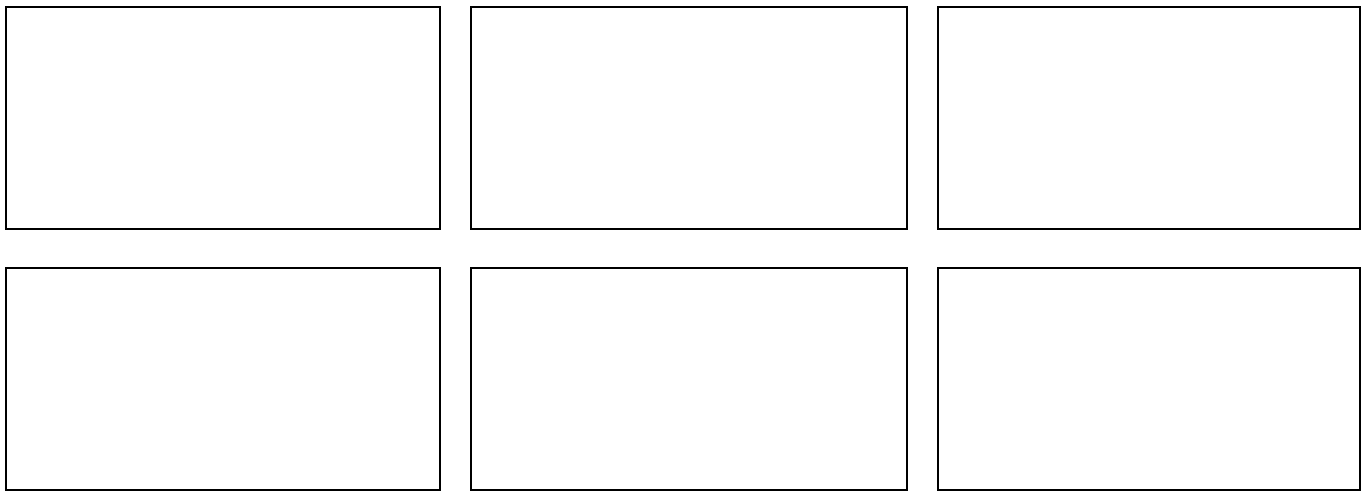
If $\mathbf{P} \Rightarrow \mathbf{Q}$ it does not follow that $\mathbf{Q} \Rightarrow \mathbf{P}$

Secondary School Science Fair Team: Science, Statistics, Computers, Art

Split Credits: Data Management & Writers' Craft

Grade 9 Linear Regression Group Assignments: See examples on attachment

Venn Diagrams: See diagrams on attachment



Probability and Language

A new family moved in next door to me the other day. I knew that this family had two children and one dog. The day after they moved in, I looked outside and saw a boy (obviously one of the children in the new family) playing with his dog. Find the probability that both of the children in this family were boys.

This means: Find the probability that

Equations of the Straight Line

The line in standard form (or the general equation of the line):

$Ax + By + C = 0$

$ax + by + c = 0$

$ax + by =$ c

The Slope, *y*-intercept form $y = mx + b$

The Slope, *x*-intercept form $y = m(x - a)$

The Slope-Point form

$y - y_1 = m(x - x_1)$ or $y = m(x - p) + q$

The Two-Point form $y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$

The Two-Intercept form $\frac{x}{a} + \frac{y}{b} = 1$

The horizontal line y = b

The Vertical line $x = a$

Slope is given by $m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$

The *x*-intercept is *a* and the *y*-intercept is *b*

The point of intersection of the line and the *x*-axis is (*a*, 0)

The point of intersection of the line and the *y*-axis is (0, *b*)

The *x*-coordinate is called the abscissa (technically the distance from the vertical axis to the point)

The *y*-coordinate is called the ordinate (technically the distance from the horizontal axis to the point)