



GER 1000

Quantitative Reasoning





Association

∞ Relationship between Two Variables ∞

Introduction



Hi everyone, my name is Peiyi. I'm the presenter for this Chapter, Association. In this Chapter, we will learn how to examine the association between two variables.

Association, sometimes, people call it relationship or correlation depending on the situation. The data that we recorded can be numerical values from taking measurements, such as height, weight, and time; or categorical, for example, ethnic groups, and disease status.

Although there are many questions that we can and will ask about two variables, in most of cases, our primary interest is: **Is there a relationship or an association between two variables?** So, what is association?

About Association

✓ Deterministic Relationship

→ The value of a variable can be determined if we know the value of the other variable.

For example,

$$\text{temperature in } ^\circ\text{F} = ^\circ\text{C} \times 9/5 + 32$$

Well, in quantitative reasoning, the association or so-called relationship is NOT deterministic, for which the value of one variable can be determined precisely from the value of the other variable.

For example, the temperature scale used is usually either in Celsius or in Fahrenheit. There is a conversion formula. If we know the current temperature in Celsius, say 32 degrees, then the temperature in Fahrenheit is precisely determined through this formula. It is 89.6 in Fahrenheit.

So, what exactly is the association that we are interested in here?

About Association

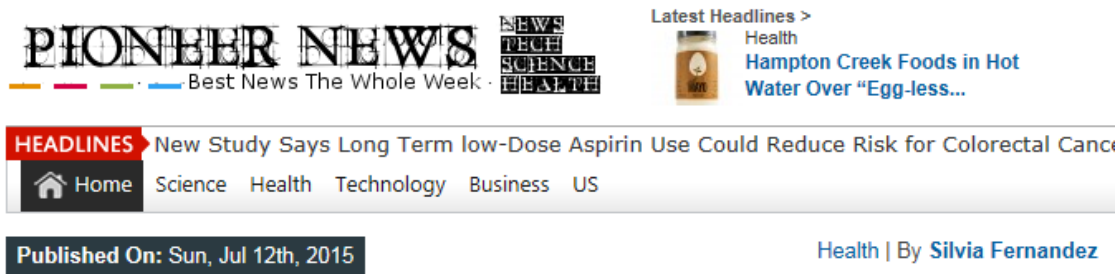
✓ *Statistical Relationship*

- ➔ *Natural variability exists in measurements/outcomes of two variables under study.*
- ➔ *The average pattern of one variable can be described given the value of the other variable.*

What we will focus on is a statistical relationship, for which we consider the variability exists in the measurements or outcomes of the two variables. The average pattern of one variable can be described, given the value of the other variable. **This is the association that we are interested in.**

In fact, we encounter the terms, association or correlation, quite often in our daily life. I shall just mention a few.

Association - In the News Headline



Study Finds Correlation Between Smoking and Education

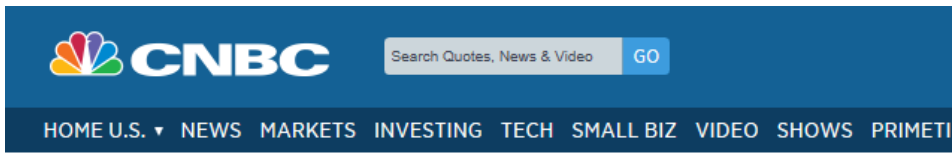
I was just reading some news and articles earlier, and happened to see this news headline: “Correlation between Smoking and Education?” Oh... I have to read on!

Well, it is common understanding that education can get you further in life. It is also common knowledge that smoking and health are associated, as many studies have shown that smoking is related to high risk of many diseases. Now, a new study shows staying in school might also be better for your health—and your life. **How do we interpret this correlation?**

Education makes a person less likely to smoke and healthier? Or maybe the other way around? Smoking less will make a person stay in school? **Or maybe the observed association is related to another variable?** [confounder?](#)

What we want to find out here is not only the measure of association between smoking and education, but also the interpretations of observed association.

Association - Finance



NETNET

Stock market correlation is at a seven-year low

Jeff Cox | @JeffCoxCNBCcom

Thursday, 20 Aug 2015 | 2:08 PM ET

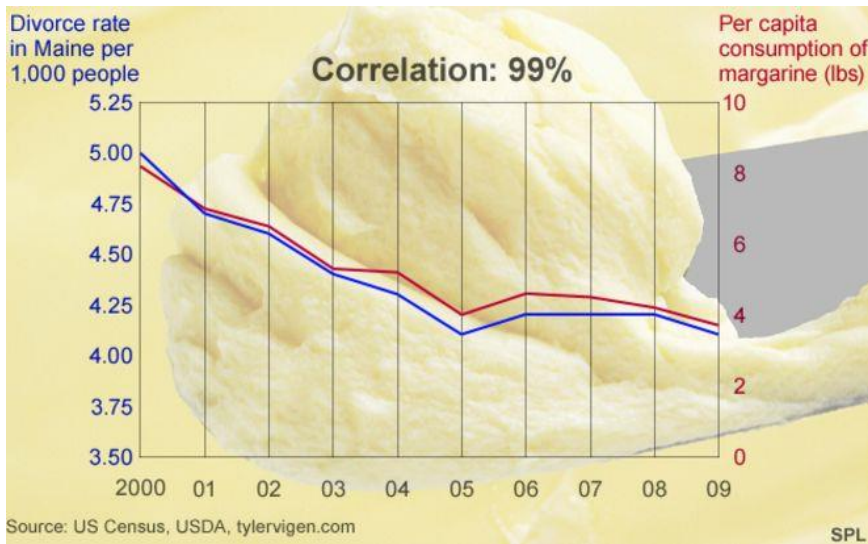


I happened to see this article: “Stock market correlation is at a seven-year low” while watching stock market movement on the other day. The headline looks striking, but does not mean anything clear to me!

If we know the association or so-called correlation good enough, we will know there must be, at least, **two variables for us to talk about “correlation”**. I read on and realized that the claimed declining correlation is observed between individual stock prices and industry sectors.

So, I wonder how this observation will help me in terms of making decision in stock market investment.

“Margarine Consumption Linked to Divorce”



BBC News magazine, 26 May 2014

Wow! "Margarine consumption linked to divorce." I bet you have never thought of such a link! In fact, this was published on BBC News Magazine accompanied by a graph.

Let's look at this compelling graph. It shows the divorce rates and margarine consumption tracking each other closely over almost 10 years! We may be tempted to believe there could be a link... Could it be "when there's more margarine in the house it's more likely to cause divorce;" Or maybe they are linked through something else!

We've seen a lot of headlines, especially sensational ones like – "Scientists find a connection between A and B". In a lot of those situations there might be a correlation, but it's really important for us to be critical about whether the two variables are linked to each other through another variable or have a causal mechanism.

The learning of quantitative reasoning, and in particular, the chapter on association, may help you gain some insight on these concerns.

Chapter Outline



Relationship between two Measurement Variables

- ① Bivariate Data & Scatter Diagram
- ② Exploring Relationship
- ③ Correlation Coefficient
- ④ Some Limitations

For measurement variables, we will learn to display the bivariate data and visualize the relationship through a scatter diagram. It is essential to use a numerical measure, correlation coefficient, to summarize and interpret the observed relationship and understand the limitations when we use such a measure.

Chapter Outline



Relationship between two Measurement Variables

- ⑤ Ecological Correlation
- ⑥ Cautionary Notes
- ⑦ Simple Linear Regression

Ecological correlation is referring to the correlation based on aggregates. It has been used and misinterpreted very often. Therefore, it is necessary for us to gain a better understanding in applying ecological correlation. As correlation coefficient is computed based on bivariate data, we need to be aware of mishandling as well as manipulations of data. Once an association between two variables is observed, it is natural for us to ask if we can use the information of one variable to predict the values of the other variable. Simple linear regression is introduced with the emphasis on interpretation, rather than calculation. That's all for the outline of this Chapter.

QR Framework

A diagram showing the QR Framework steps. A horizontal orange line with seven red dots spans the width of the slide. A vertical orange line descends from the rightmost dot, forming a corner. The five steps are listed to the left of this corner: 1 Frame, 2 Specify, 3 Collect, 4 Analyse, and 5 Communicate. The first three steps are in blue, and the last two are in maroon.

① *Frame*

② *Specify*

③ *Collect*

④ *Analyse*

⑤ *Communicate*

In earlier chapters, “the quantitative reasoning framework”, better known as the QR framework, has been described in detail. This systematic approach points out a path and helps us better understand real-life issues involving quantitative matters. This chapter, Association, fits well in the framework.

Though the first three steps, Frame, Specify, and Collect, are NOT our main focuses in this chapter, it is important to know “the questions to ask”, “what to measure in order to answer the question”, as well as “how to measure” these. These three steps will be implicitly included in contents and examples throughout the chapter as they play vital parts in quantitative reasoning. What we will discuss in this chapter is most relevant to the fourth and fifth steps: *Analyze* and *Communicate*.

Using appropriate methods to *Analyze* the information or data is an essential skill for quantitative reasoning. A sound understanding of the results and a good ability to describe and *Communicate* the results and make inferences are the key focuses. Many examples will be discussed throughout the chapter to illustrate these two steps.

Hope you have gained some basic ideas about what will be discussed in the subsequent units. Let’s move on!



Association

∞ Relationship between Two Variables ∞

Bivariate Data & Scatter Diagram



In this unit, we will focus on the relationship between two variables. In particular, how we can display and examine the relationship.

Bivariate Data

✓ Karl Pearson's Father-and-Son Data:



Karl Pearson
(1857 – 1936)

Looking at the unit title, you may be wondering: what does “bivariate data” mean? Why don’t we illustrate this with an example?

The data set that we will use, throughout the chapter of Association, is Karl Pearson’s Father-and-Son data set.

Karl Pearson had an influential contribution to the discipline of Statistics and establishment of modern statistics. In 1911, he founded the Department of Applied Statistics at University College London. This is the very first university statistics department in the world! Besides, Pearson was instrumental in the development of the correlation and regression, the underlying theory for this chapter!

Like Father, Like Son



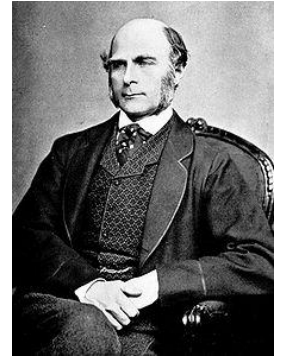
✓ The motivation...

➔ *The degree that children resemble their parents.*

✓ The action plan...

➔ *Gather huge amount of data*

➔ *Quantify the resemblance*



Sir Francis Galton
(1822 – 1911)

There is a saying... Like Father, like son. Have you ever wondered about the degree to which you resemble your parents?

As early as late 19th century, Sir Francis Galton had thought about this and taken action to work on it. He had the idea of gathering data to quantify such resemblance between a father's and his son's heights.

The study was later carried out by his disciple, Karl Pearson. The Father-and-son data set was resulted from this study!

Like Father, Like Son



✓ The study...

- Heights of 1078 father-and-son pairs were collected
- Two variables are collected
The father's height (X)
and his son's height (Y)
- Bivariate data (X, Y)



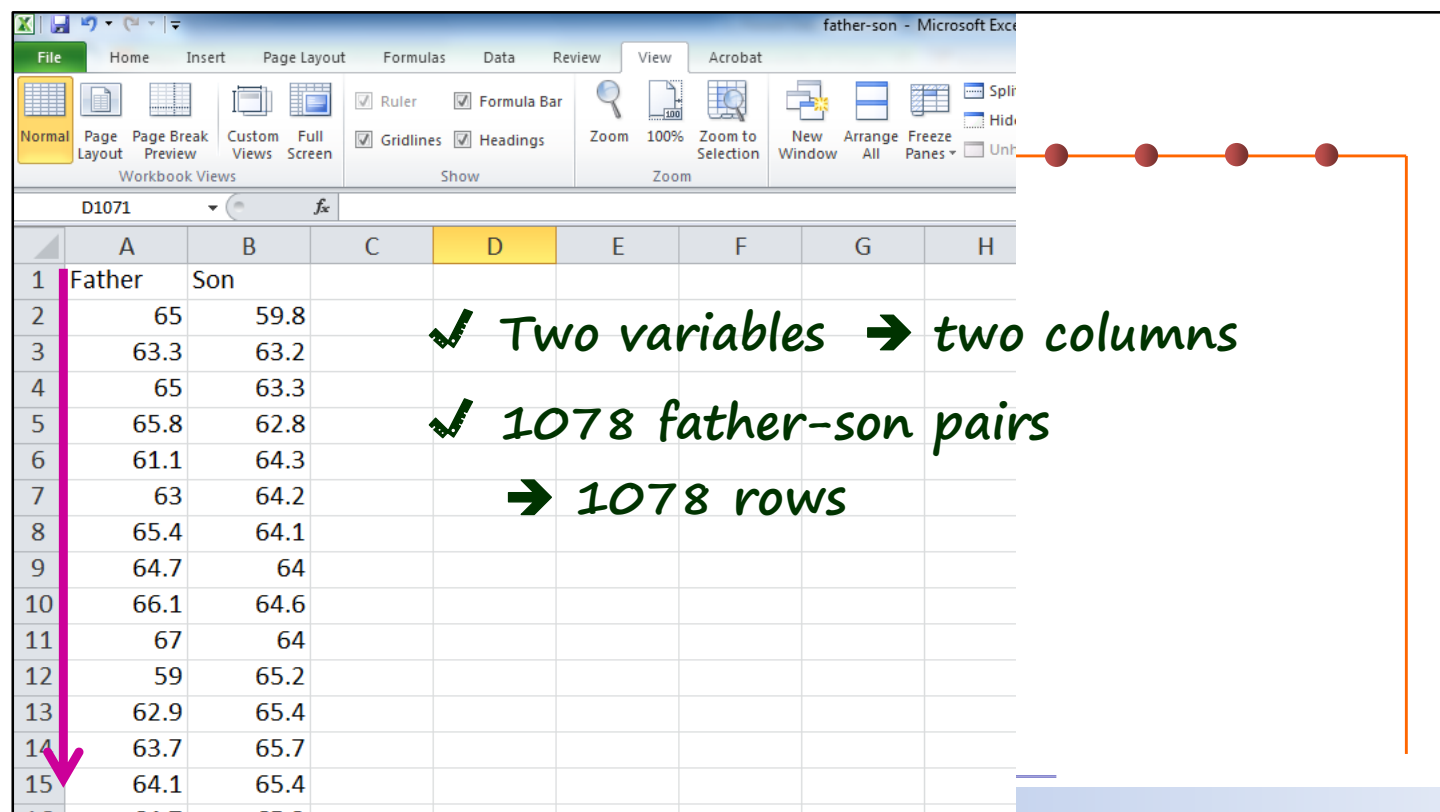
Karl Pearson
(1857 – 1936)

Enough of history. Let's focus on the study and the data set.

Acting on Galton's ideas, Pearson studied one thousand and seventy-eight father-son pairs. Their heights, in inches, were recorded. We may just consider each father-son pair as one unit. So, we have one thousand and seventy-eight (1078) units in this data set. In each unit, two variables are included. That is, the father's height and his "adult" son's height.

We use X to denote the variable representing a father's height; Use Y to denote the variable representing his adult son's height. That is, for each unit, we have two variables. This is "bivariate data". We may examine the relationship between the two variables, that is, a father's height and his son's heights, through such a bivariate data set.

Let's take a look at the data.



	A	B	C	D	E	F	G	H
1	Father	Son						
2	65	59.8						
3	63.3	63.2						
4	65	63.3						
5	65.8	62.8						
6	61.1	64.3						
7	63	64.2						
8	65.4	64.1						
9	64.7	64						
10	66.1	64.6						
11	67	64						
12	59	65.2						
13	62.9	65.4						
14	63.7	65.7						
15	64.1	65.4						
16	64.7	65.2						

Here is the data set presented in Excel format. It is available in the IVLE.

The two variables are represented by two columns. The heights for one thousand and seventy-eight father-son pairs are shown in one thousand and seventy-eight (1078) rows.

So, we know what bivariate data are through this example and learn how the data can be displayed in Excel. Can we try to describe or summarize father's and son's heights?

Like Father, Like Son



✓ The average –

→ The father's average height was 68 inches and the sons' average 69 inches.



✓ The spread/variability –

→ Standard Deviation (sd)

First of all, we can use a simple and intuitive measure to do so, which is the average!

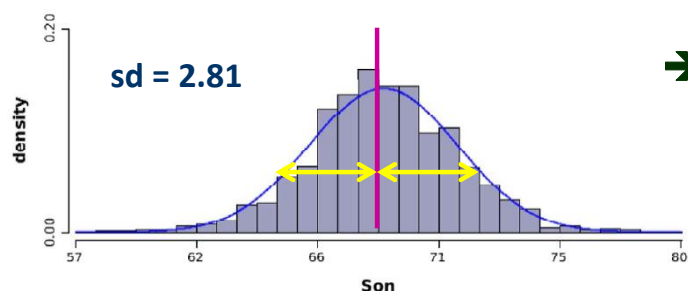
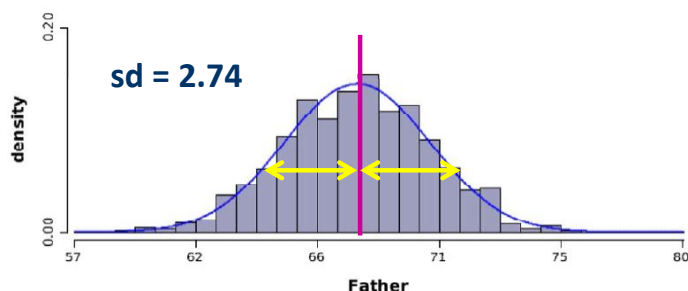
1078 fathers' average height was 68 inches. We will use the unit from the original data. However, just to let you have a better idea, it is about 172 centimeters. The average height for their sons was 69 inches, which is about 174 centimeters, taller than their fathers, on average.

However, this simple measure may not be sufficient, in terms of describing or summarizing the data. We should further describe the fathers' heights and the sons' heights by showing how the data spread out around their average.

We often use "standard deviation" to describe the spread or variability of data around the average. Conceptually, it shows the average distance from each data point to their average. I'm going to use "sd" to denote standard deviation. The formula used to produce sd will not be discussed in this chapter. Instead, we will look at the meaning of standard deviation through some graphical displays.

s.d meaning

Like Father, Like Son



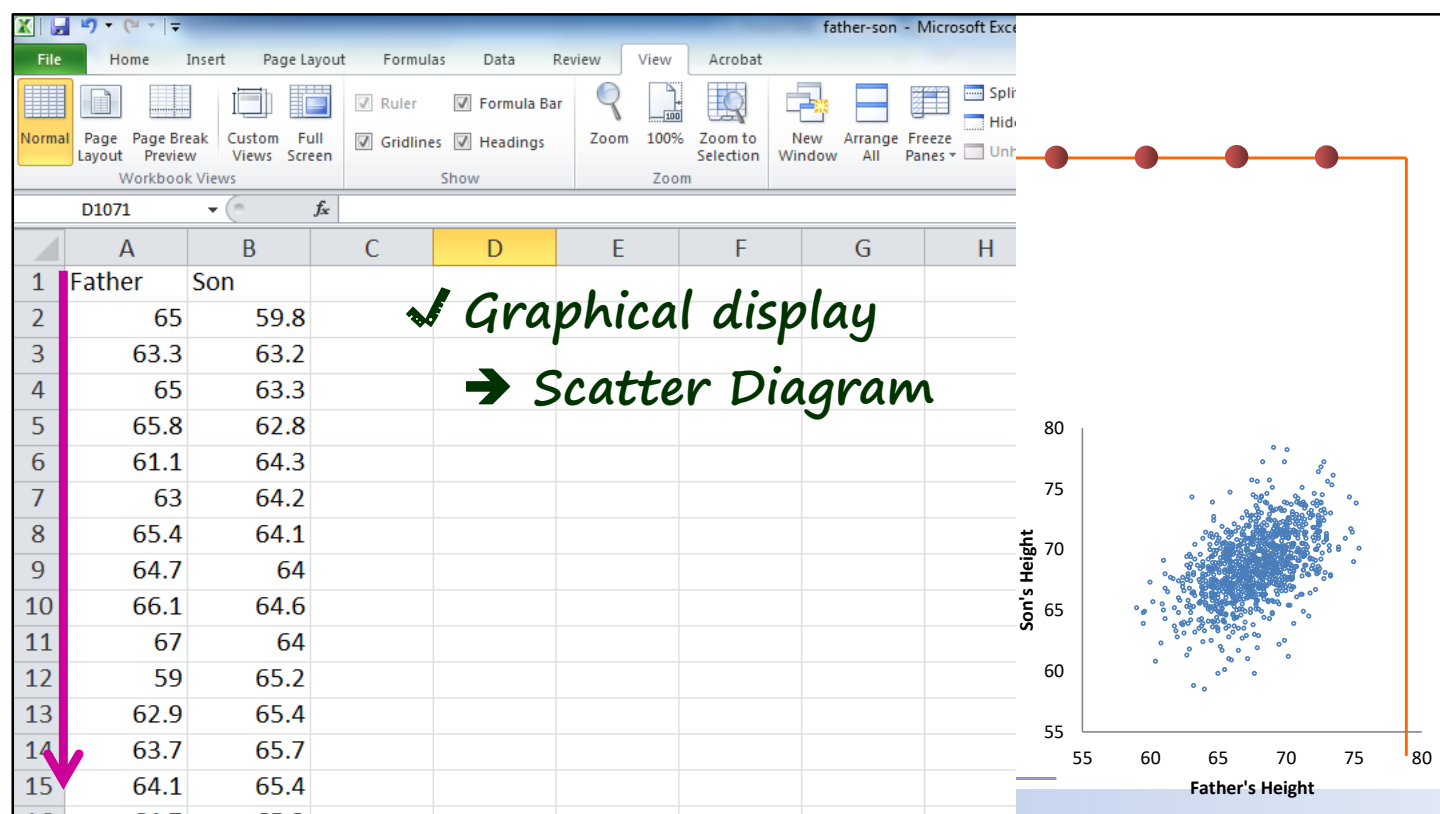
→ Slightly larger sd among sons' heights

What we see here are density histograms. We will use these histograms to show how data spread out.

This is a density histogram for 1078 fathers' height. The average height was 68 inches. Sd indicates the spread around the average. The sd for fathers' heights was 2.74 inches

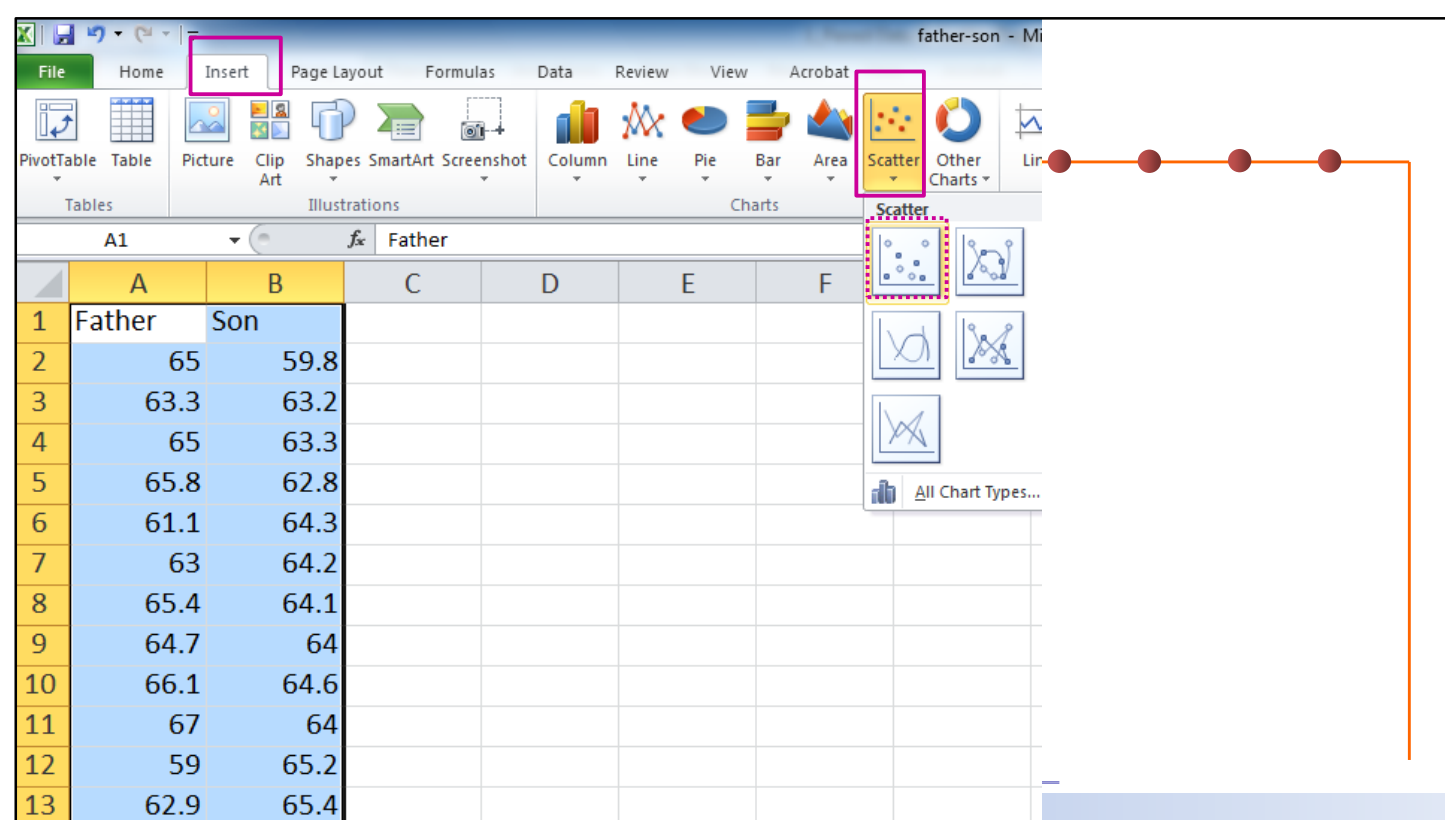
Below it is the density histogram for sons' heights. The average height was 69 inches. If you are really perceptive, you may be able to tell that the spread among sons' heights is slightly larger. The sd for the sons' heights was 2.81 inches, larger by about 0.07 inches.

We have just explored the fathers' heights and their sons' heights separately. Now, the question is how we are going to explore the relationship between the fathers' and sons' heights?



Let's come back to the excel data file. Scrolling down the spreadsheet and looking at all these numerical data probably will not help us much in exploring the relationship between the two variables. 1078 pairs! That's a lot of numbers to consider! It will be hard to visualize relationship without some meaningful tools!

As a rule of thumb, we may consider displaying the data with some graphs. For bivariate data, one of the **most commonly used graphical displays is scatter diagram** (or scatter plot). It looks like this. In fact, the scatter diagram can be produced by Excel!

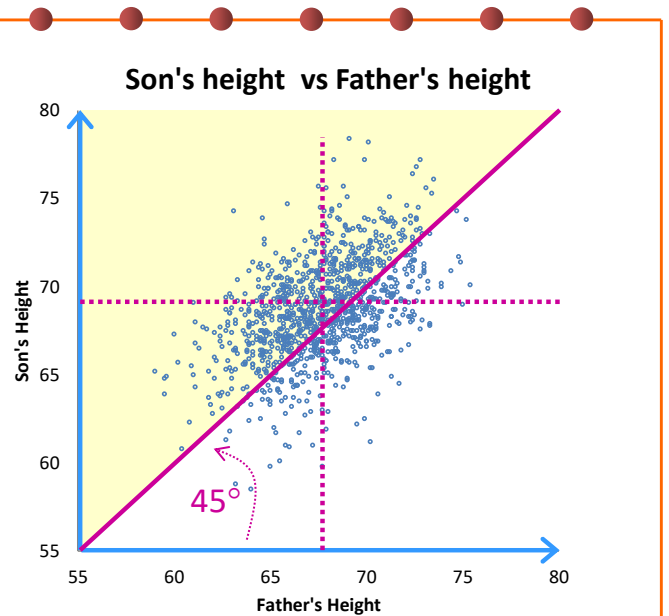


We will highlight the two columns of data.

Under the “insert” tab, select “scatter”.

Scatter Diagram

- ✓ 45° line: $F = S$
- ✓ Above 45° line:
→ A son is taller than his father
- ✓ Majority of sons were taller than their fathers
- ✓ Sons' average height was larger



A scatter diagram is produced. So, how do we make use of this scatter diagram to examine the relationship between fathers' and their adult sons' heights? Here are a few things that we may consider...

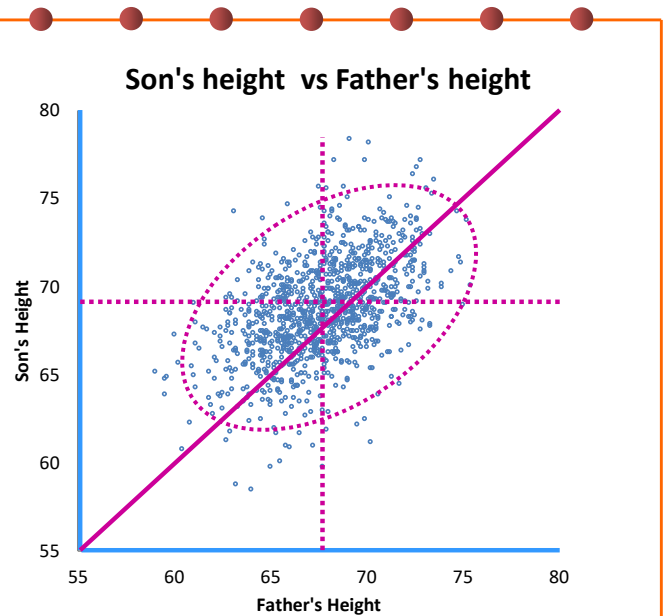
If vertical axis and horizontal axis are of the same scale, we can draw a 45 degree line. Points that lie on the 45 degree line are the cases where a father and his son are of the same height.

How about the points that lie above the 45 degree line? A Son is taller than his father... right?

So it's quite clear that more points lie above 45 degree line. More than 50%, that's majority of the sons in the study were taller than their fathers. We also know that the average height of fathers was 68 inches; however the sons had a larger average of 69 inches

Scatter Diagram

- ✓ Sons' average height was larger
- ✓ Positive association



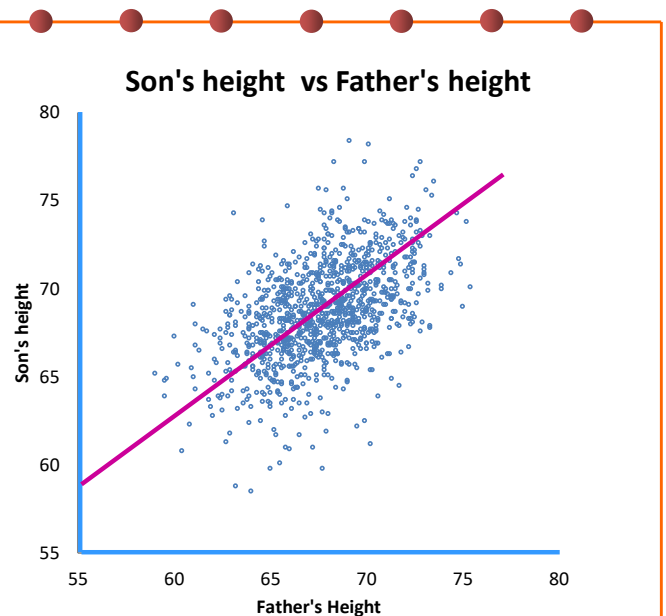
Now, let's remove the 45 degree line and relook at the scatter diagram. The data points form a **distinct oval-shape cloud**.

It seems that taller fathers had taller sons. In other words, the data showed a positive association between fathers' and their sons' heights.

Scatter Diagram

- ✓ Sons' average height was taller
- ✓ Positive association
- ✓ Linear relationship

Quantify the strength of the association



Let's clear up the scatter diagram and look at the data again. The data seemed to cluster around a straight line with a positive slope. That is, a linear relationship may be used to describe the association between the two variables.

Now, I'm just wondering if the observed association is strong. If so, knowing the father's height will much help predict his son's height. If the relationship is weak, knowing the father's height will not help much in gauging his son's height!

In short, we would like to take it one step further - define a measure to quantify the strength of the association between the two variables. We will leave this for the next unit.

Unit: Bivariate Data & Scatter Diagram

- ① *Bivariate Data*
- ② *Visualize bivariate relationship with a scatter diagram*
- ③ *Explore possible relationship between two variables*

As a quick recap...we have discussed the following:

The meaning of bivariate data; and display such data using scatter diagram; we also discussed how to explore the relationship through a scatter diagram.



Association

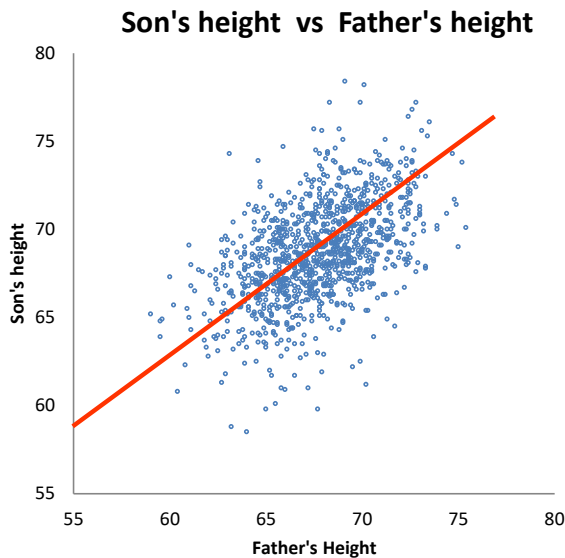
∞ Relationship between Two Variables ∞

Exploring Relationship with Correlation Coefficient



In this unit, we will focus on the linear relationship between two variables. In particular, how we can quantify the linear relationship.

Exploring Association



- ① Nature of relationship
→ Linear
- ② Direction of relationship
→ Positive
- ③ Strength of relationship

Let's look at Pearson's father and son data. Once we plot the scatter diagram, we may visualize the association between fathers and their sons' heights.

Here are few things that we have learned from examining the scatter diagram. First, nature of relationships, whether it is linear or non-linear. It appears that there is a linear relationship between fathers' and their sons' heights.

Scatter diagram will also suggest the direction of linear relationship, whether it is positive or negative. It seems that taller fathers have taller sons. This shows a positive relationship.

One of the most important things that we wish to know would be "the strength" of observed relationship. We would like to know whether it is strong or weak. If the relationship is strong, the data points will cluster closely to the straight line; if the relationship is weak, the data points will scatter loosely around the straight line. In general, the closer the points are to a straight line form, the stronger the correlation between the two variables.

Well, close or loose around the straight line? Sometimes, it is not easy to determine just with our naked eyes. Therefore, it would be necessary for us to use a numerical measure to quantify the strength of the linear relationship.

Correlation Coefficient (r)



- ✓ *A measure of linear association between two variables*
- ✓ *Range between -1 and 1 . ($-1 \leq r \leq 1$)*
- ✓ *It summarizes the direction and strength of linear association.*

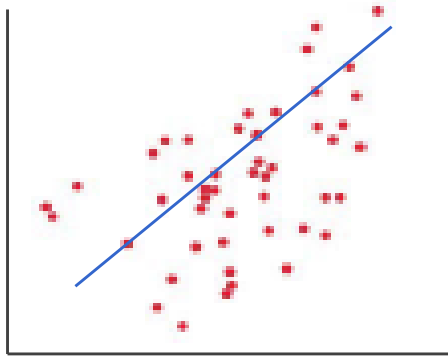
This measure is “correlation coefficient”. It is usually abbreviated as “ r ”. In this unit, our focus is not on calculation of “ r ”. Instead, we will discuss the meaning and interpretations of the correlation coefficient, r .

r is a measure of “linear” association between two variables. It has a value between -1 and 1 . We can use this single numerical measure to summarize the direction, as well as the strength of the linear association between two variables.

Understand Correlation Coefficient (r)

✓ *Direction of the linear association...*

$r > 0 \rightarrow$ *Positive association*



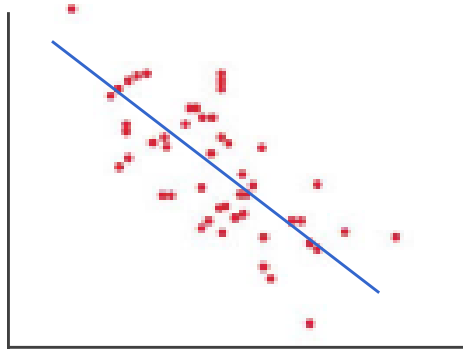
Direction of the linear association can be either positive or negative. Besides visualizing the direction of the relationship from scatter diagram, we may look at the “sign” of r value to determine the direction too.

A positive r value shows a positive association. That is, as one variable increases so does the other.

Understand Correlation Coefficient (r)

✓ *Direction of the linear association...*

$r < 0 \rightarrow$ *Negative association*

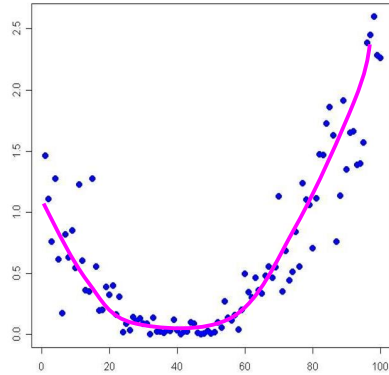
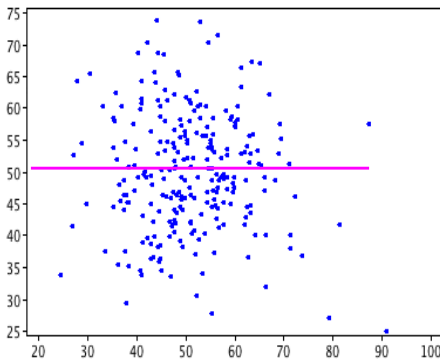


A negative r value shows a negative association. That is, as one variable increases the other variable decreases.

Understand Correlation Coefficient (r)

✓ *Direction of the linear association...*

$r = 0 \rightarrow$ No linear association



How about an r value of zero? Does this mean there is no association? Remember, r is used to measure “linear” association. **Therefore, a zero r value simply means no “linear” association!**

For example, if we examine this scatter diagram (on the left), it appears that there is no association... that is, r is nearly zero.

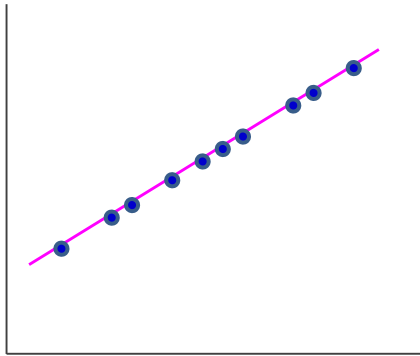
How about this diagram (on the right)? It seems to show some association or relationship between the two variables. However, the association does not look linear... Instead, it looks like a curve! In this case, r is nearly zero too, as correlation coefficient detects only “linear” association between two variables.

Either one of these two displays showed that there is no linear association between two variables.

Understand Correlation Coefficient (r)

✓ *Strength of the linear association...*

① $r = 1 \rightarrow$ *perfect positive association*



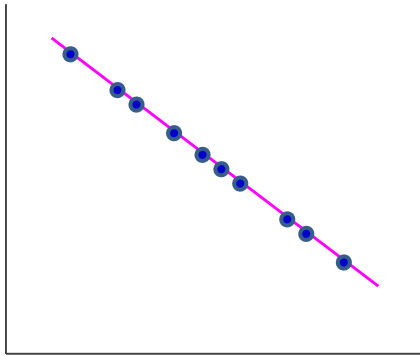
We can use this measure to quantify the strength of the linear association between the two variables too.

We know that correlation coefficient has a value between -1 and 1. When r equals 1, all data points lie on a straight line like this. We say that there is a perfect positive linear association between the two variables.

Understand Correlation Coefficient (r)

✓ Strength of the linear association...

② $r = -1 \rightarrow$ perfect negative association



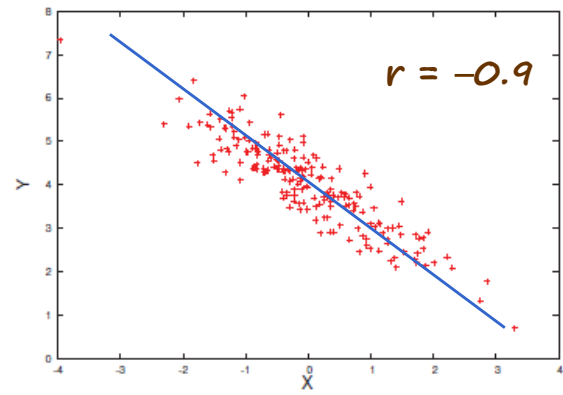
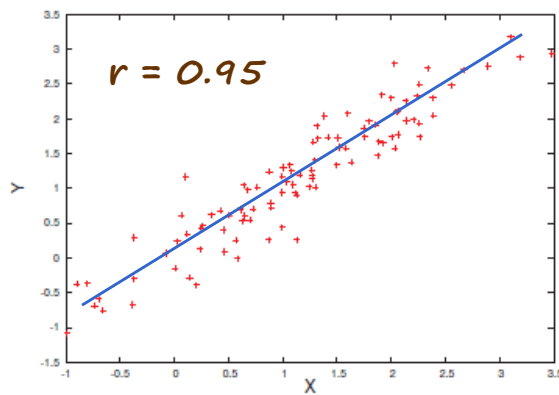
When r equals -1 , all data points lie on a straight line like this. We say that there is a perfect negative linear association between the two variables.

Besides “perfect” linear association, the strength can be described as strong, moderate or weak depending on the r value.

Understand Correlation Coefficient (r)

✓ Strength of the linear association...

③ r value is close to ± 1 → Strong association



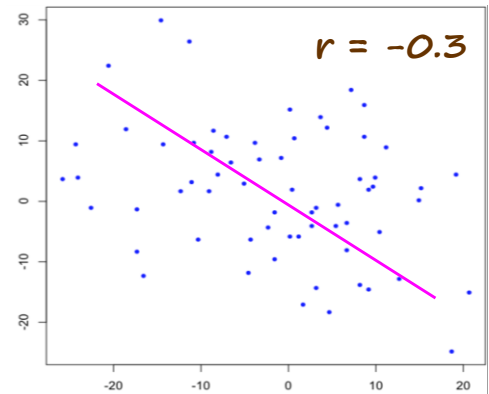
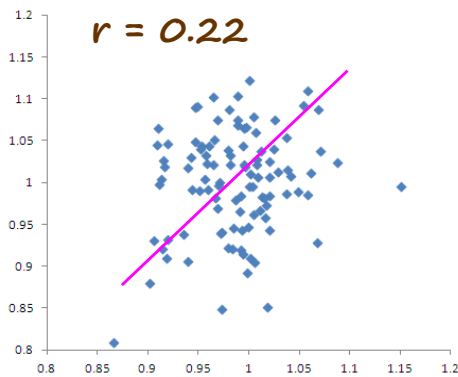
When r value is close to positive or negative 1, the two variables have a strong linear association. In such cases, all data points lie closely to a straight line. For example, this scatter diagram shows data with r equals 0.95. This one shows data with r equals negative 0.9.

If there is a strong association between two variables, then the information that we have on one variable will help us a lot in predicting the value of the other variable.

Understand Correlation Coefficient (r)

✓ Strength of the linear association...

④ r value is close to 0 → Weak association



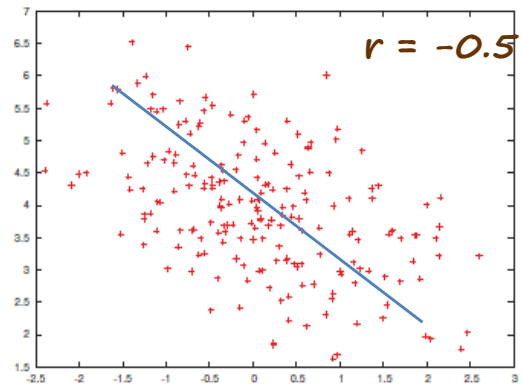
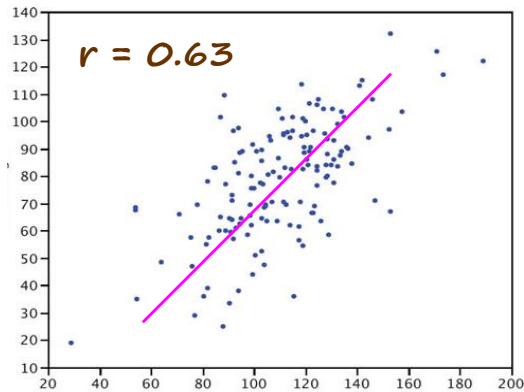
When r value is close to 0, the two variables have a weak linear association. In such cases, all data points lie loosely along the straight line. For example, this scatter diagram shows data with r equals 0.22. This one shows data with r equals -0.3.

We have learned that, as a rule of thumb, the linear association is considered to be strong if r value is close to positive or negative one; the linear association is considered to be weak if r value is close to zero.

Understand Correlation Coefficient (r)

✓ Strength of the linear association...

⑤ r value is close to $\pm 0.5 \rightarrow$ moderate association

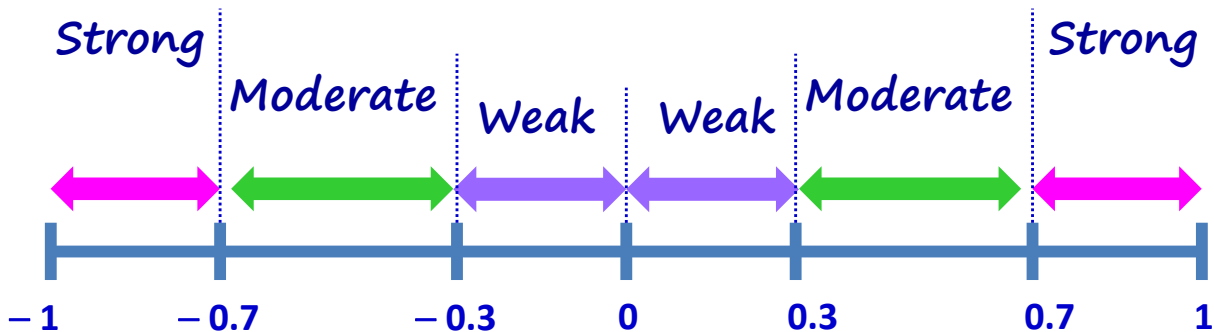


How if the r value is close to positive and negative 0.5? Strong or weak?

In such a case, the two variables are said to have a moderate linear association. For example, this scatter diagram shows data with r equals 0.63. This one shows data with r equals -0.5.

Understand Correlation Coefficient (r)

✓ *Strength of the linear association... rule of thumb*



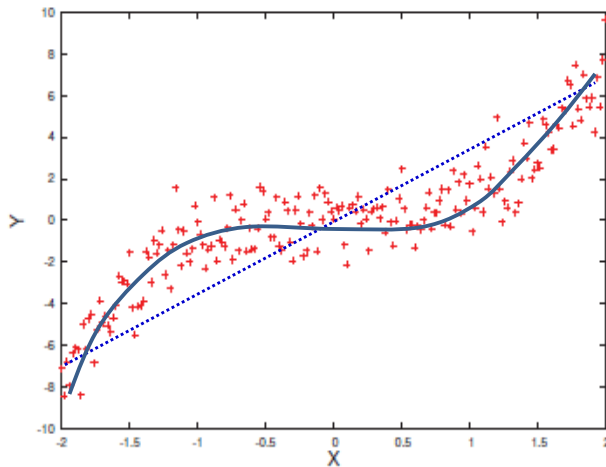
To sum up, the strength of the linear association can be classified as strong, moderate or weak, based on the value of correlation coefficient.

We know when r value is 1 or negative 1, the two variables have a perfect linear association. When r value is 0, we say there is no linear association

As a rule of thumb, when r value is between -1 and -0.7 or between 0.7 and 1, the two variables are strongly associated. When r value is between -0.7 and -0.3 or between 0.3 and 0.7, the association between the two variables is moderate. When r value is between -0.3 and 0 or between 0 and 0.3, the two variables have a weak association.

Bear in mind what we show here is a rule of thumb. Some of the textbooks may have a slightly different indication. However, we should know that the closer the r value is to "zero", the weaker the linear association. The further the r value is to zero, the stronger the linear association. In other words, an r value that is very close to negative or positive 1, suggests a very strong linear association between two variables.

Scatter Diagram & Correlation Coefficient



→ $r \approx 0.85$

*Is a linear association
best describe the
relationship between
X & Y?*

So, we know that correlation coefficient is a very useful measure when we want to learn the direction and strength of linear association between two variables... However, it is important that we draw a scatter diagram first.

Let's look at this scatter diagram. Is there linear association between two variables? Or you think such a curve will better describe the relationship. If you say some curve may be more appropriate, then you are right! A curve seems to describe the association between the two variables better! The curve you see is a "cubic curve".

However, such a non-linear association is, somehow, quite close to a linear association. In fact, the correlation coefficient, r , can be computed. It is about 0.85.

Even though a cubic curve described the association better, the computed correlation coefficient, r , indicated quite a sizable correlation! A "strong" positive linear association!

The lesson to learn: a scatter diagram is necessary to help us visualize some details that may be missed out from a simple measure, r !

We shall discuss more about this in other units.

Unit: Exploring Relationship with Correlation Coefficient (r)

- ① *Introduce correlation coefficient (r)*
- ② *Use r to understand the direction & strength of linear association*
- ③ *Scatter diagram and r*

As a quick recap...we have discussed:

How we can further explore the association between two variables with a simple measure, correlation coefficient, r . This is a single numerical value used to indicate the direction as well as strength of the linear association. However, it is prudent to draw a scatter diagram first before computing the r value.



Association

∞ Relationship between Two Variables ∞


About the Correlation Coefficient



In this unit, we will further discuss the correlation coefficient by introducing the calculation as well as some basic properties. We will also show how we can make use of Excel to obtain the correlation coefficient.

Computing the Correlation Coefficient (r)

✓ *Father and son data set...*

- 
- *Two variables are: the father's height, X and his son's height, Y*
 - *Fathers' average height, $\bar{X} = 68$ inches standard deviation, $sd_x = 2.74$ inches*
 - *Sons' average height, $\bar{Y} = 69$ inches standard deviation, $sd_y = 2.81$ inches*



How can we compute the correlation coefficient, r ? Let's take Pearson's "Father and Son" data set as an example. We shall briefly state the basic statistics of this data set before illustrating the calculation of the correlation coefficient.

There are two variables in this data set. The father's height, which is represented by X ; and his son's height, which is represented by Y . Fathers' average height was 68 inches. Here we use \bar{X} to indicate the average. The standard deviation was 2.74 inches. The notation "sd" is used for standard deviation and the "x" is the corresponding variable, father's height. The sons in this data set had an average of 69 inches, with a standard deviation of 2.81 inches.

Computing the Correlation Coefficient (r)

✓ The steps:

① Convert each variable to standard unit (SU)


$$\Rightarrow 65 \text{ inches} \Rightarrow SU = \frac{65 - 68}{2.74} = -1.1$$

$$\Rightarrow 71 \text{ inches} \Rightarrow SU = \frac{71 - 69}{2.81} = 0.71$$

Now we shall look at the steps for computing the correlation coefficient between the fathers and sons heights. First, we will convert father's height and son's height, to standard units. Let's use the short-hand notation "su" here.

So, what is "standard unit"? For example, a father's height is 65 inches. Then the standard unit can be computed as 65 minus the fathers' average, 68 inches, and then divided by the standard deviation, 2.74 inches. So the standard unit for the father is negative 1.1 (-1.1). It shows that the father's height is 1.1 units of standard deviation below the average.

How about the standard unit for his son, whose height is 71 inches? Why don't you try to take out your calculator and compute the standard unit? Remember, the sons' average height was 69 inches. So, 71 minus 69, divided by standard deviation, 2.81 inches. You should get 0.71, a positive number. That is, the son's height is "above" average by 0.71 units of standard deviation.

In conclusion, a positive standard unit shows an observation is above average; a negative standard unit shows an observation is below average.

Computing the Correlation Coefficient (r)

✓ The steps:

① Convert each variable to standard unit (SU)



$$\rightarrow SU = \frac{X - \bar{X}}{sd_x}$$



$$\rightarrow SU = \frac{Y - \bar{Y}}{sd_Y}$$

Here is a general expression for a father's standard unit. We express the difference between a father's height and the fathers' average, in terms of units of standard deviation.

Similarly, here is a son's standard unit.

Computing the Correlation Coefficient (r)

✓ The steps:

- ① Convert each variable to standard unit
- ② Take the product of the standard units for each father-son pair

➔ For this pair... -1.1×0.71

➔ Do this for all 1078 father-son pairs

Next, we will get the product for each father-son pair by multiplying the father's and his son's standard units. That is, negative 1.1 times 0.71 for the father-son pair in the example.

We will do steps 1 and 2 for all 1078 father-son pairs in the data set.

Computing the Correlation Coefficient (r)

✓ *The steps:*

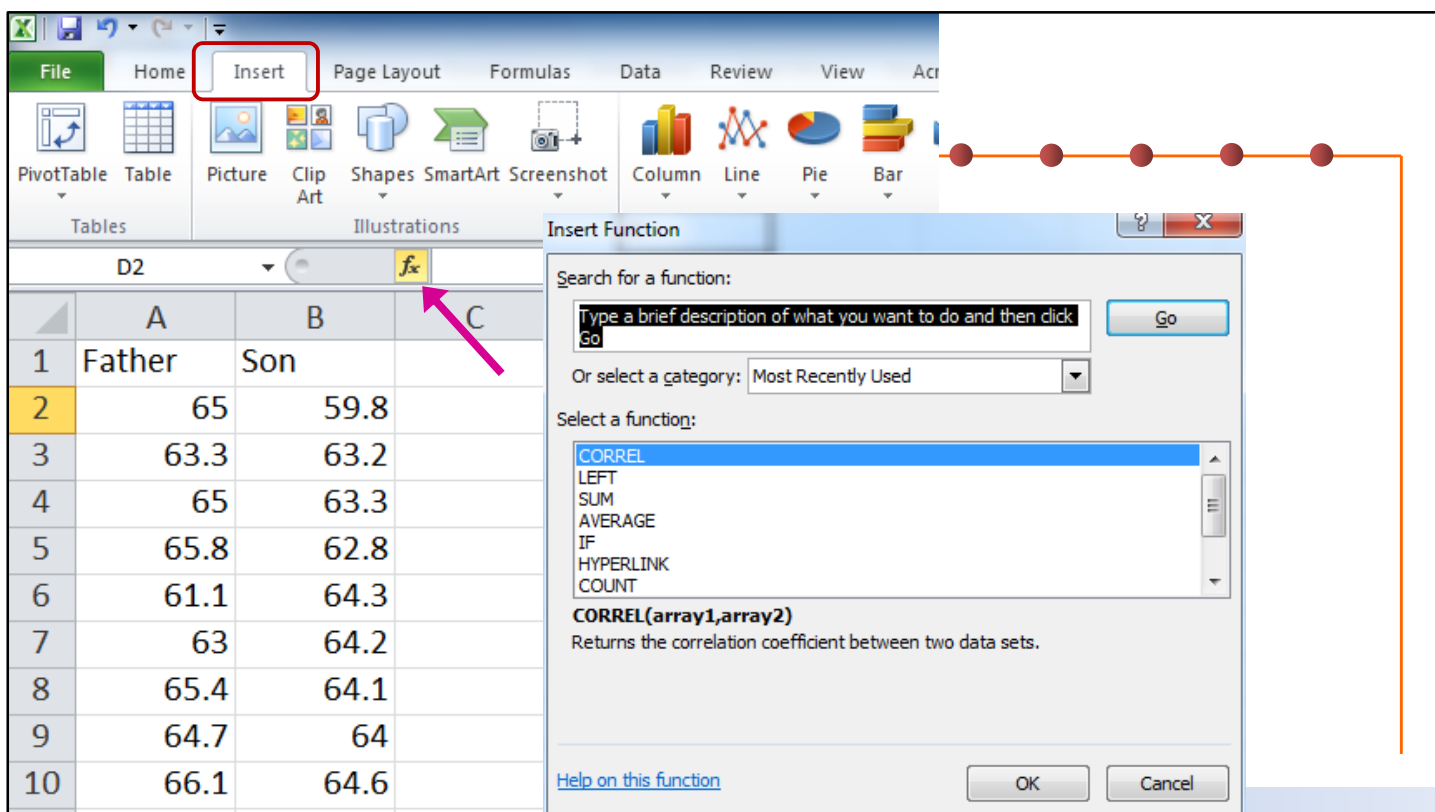
- ① *Convert each variable to standard unit*
- ② *Take the product of the standard units for each father-son pair*
- ③ *$r \rightarrow$ Average of the 1078 products*

$$r = \frac{1}{1078} \sum_{i=1}^{1078} \left(\frac{X_i - \bar{X}}{sd_X} \right) \left(\frac{Y_i - \bar{Y}}{sd_Y} \right)$$

The last step is to average all 1078 products that we have obtained from steps 1 and 2. This is the correlation coefficient, r .

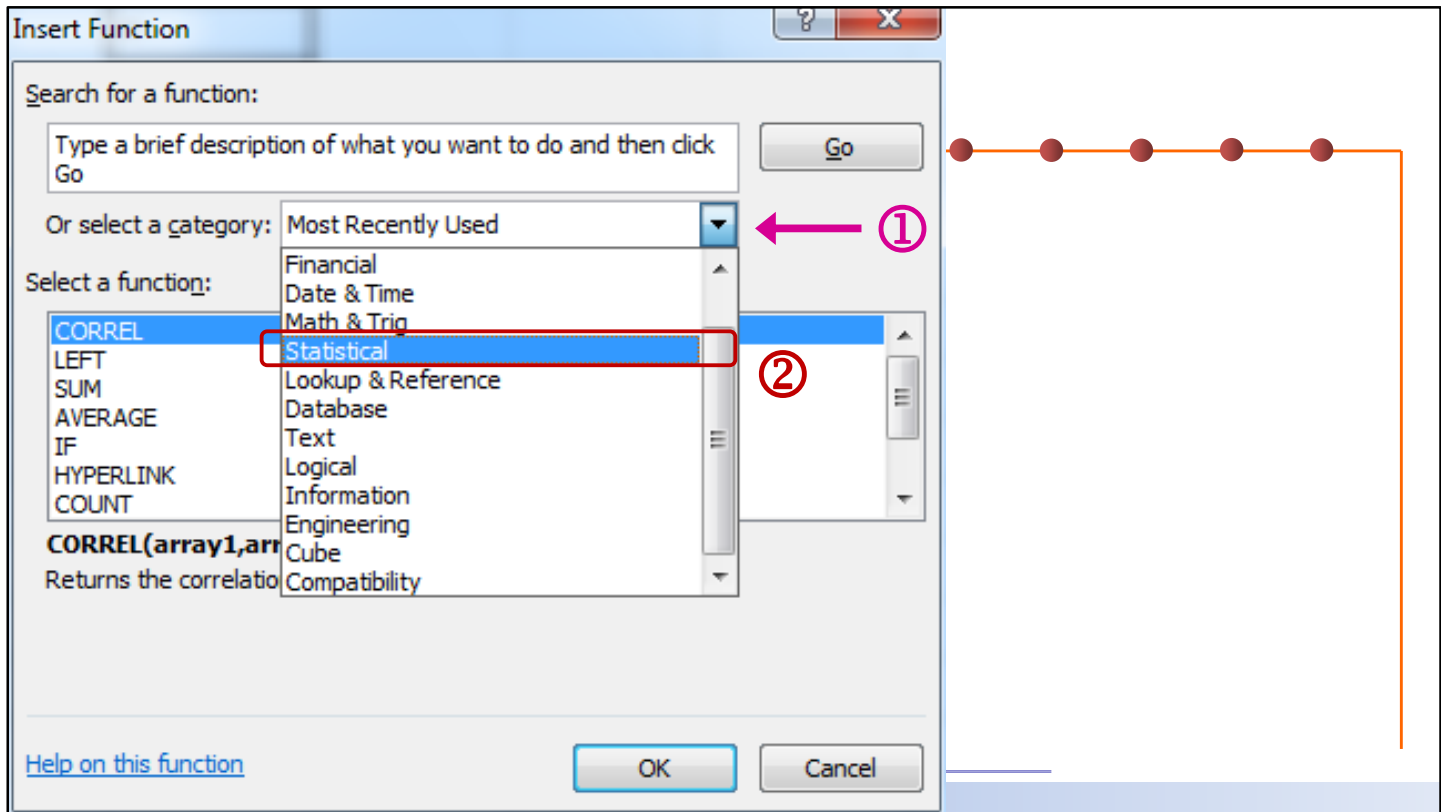
We have learned the steps for computing the r value. Let's try to put the formula together based on the three steps: Compute the standard units for each father-son pair. Then take the product of father-son standard units. Sum up all 1078 products. Take the average... that is, divided by 1078. Here, we have the correlation coefficient.

Well, it seems to be very tedious to compute the correlation coefficient. Especially, we have 1078 father-son pairs in the data set! It is necessary for us to obtain this measure with some help. Excel would work well for this purpose.



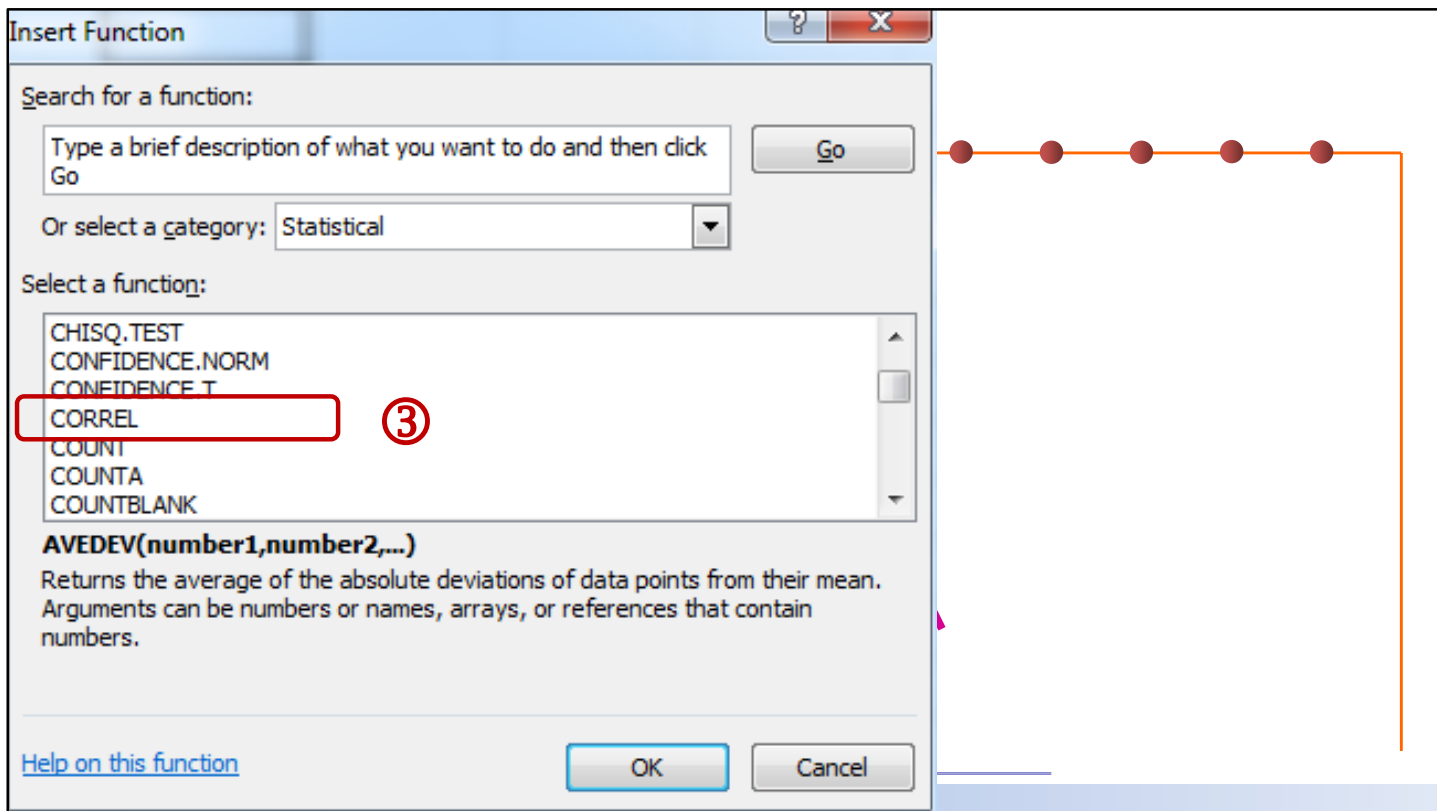
Once we open the data file in Excel, click on the “insert” tab. A function dial is displayed right here.

Click on this dial, an “insert function” window will pop up.



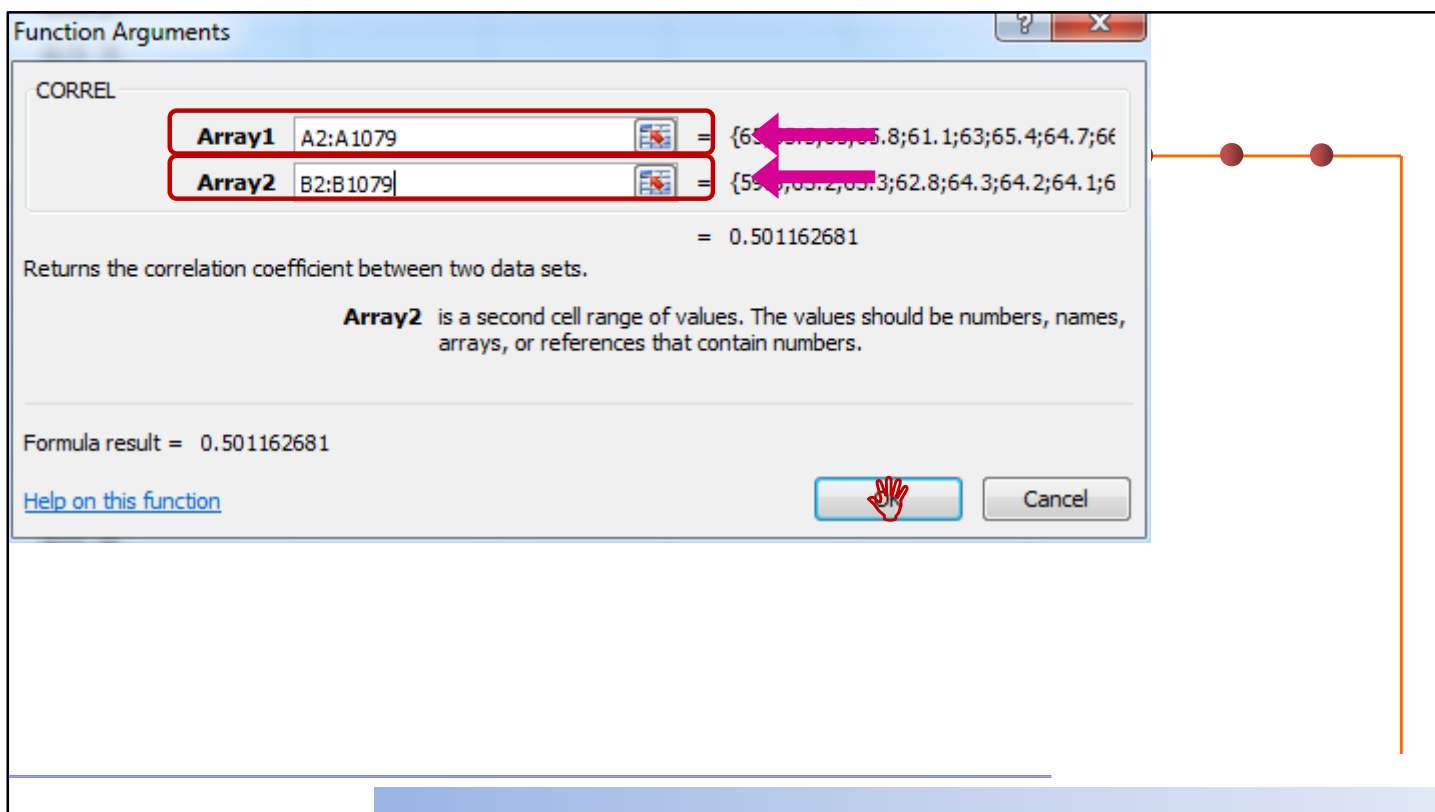
Let's focus on the insert function pop-up window.

1. Click this triangle to select a category. A drop-down is displayed.
2. Scroll down to select "statistical" function.

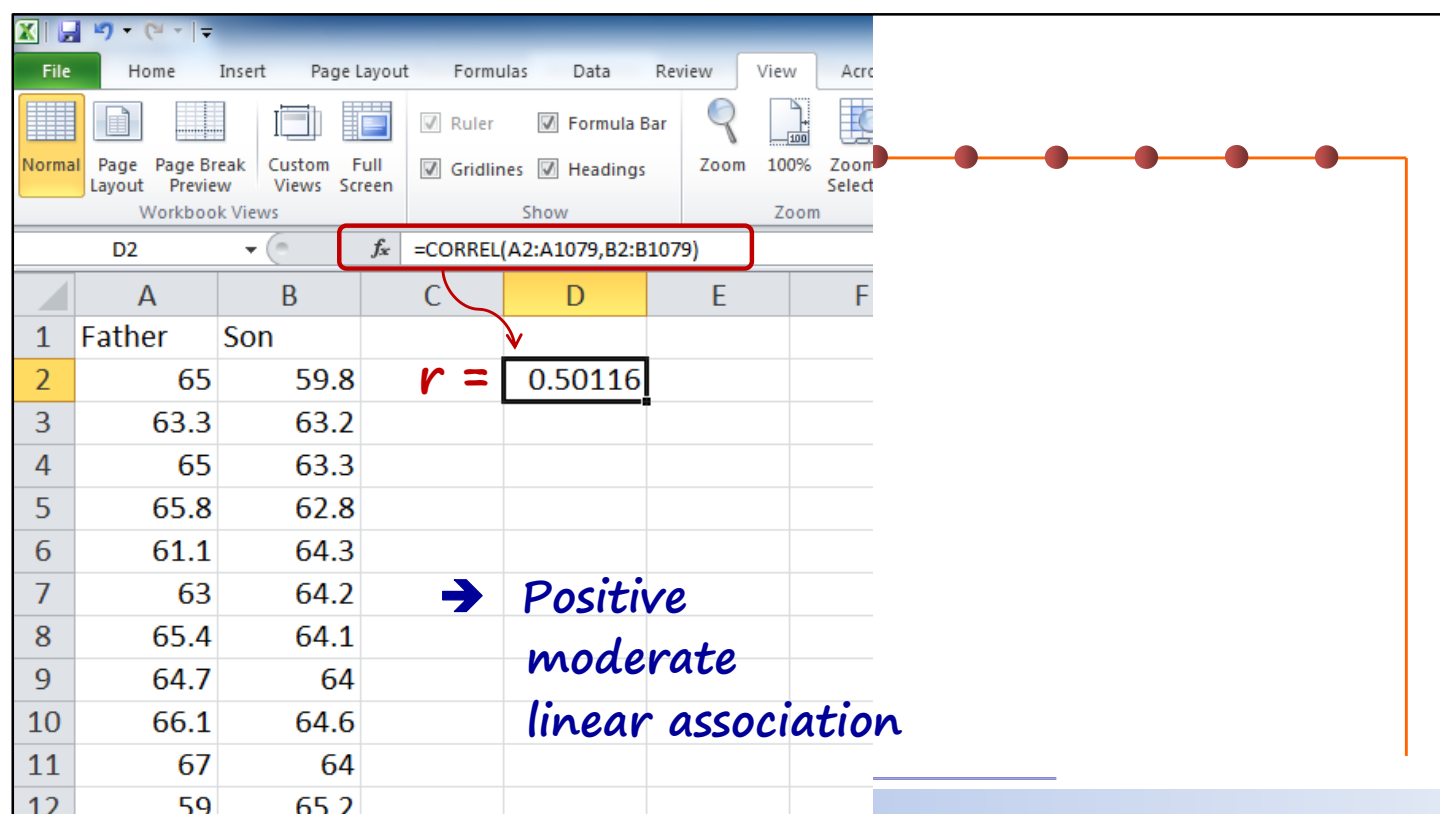


All the available statistical functions are listed in alphabetical order. Scroll down to look for “correlation” function. You will see this function.

3. This function is for correlation coefficient.

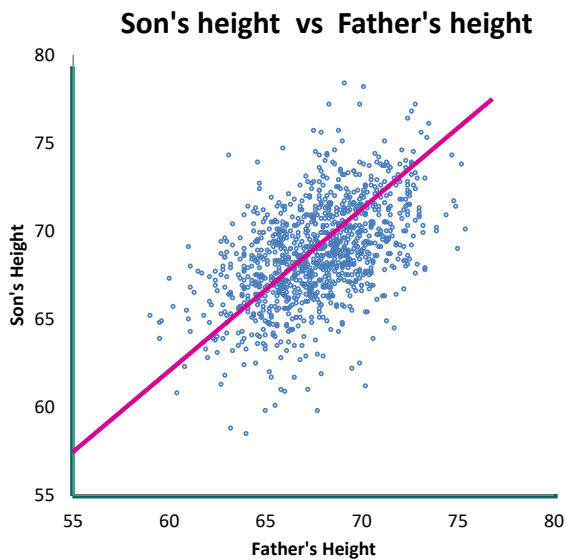


A new pop-up will show up for us to select data range for computing correlation coefficient. Enter the data range for fathers' heights; enter the data range for sons' heights. Now we are ready to move on.



Here you go. The range of data that we have just entered is displayed. The correlation coefficient is computed and shown on a cell that we have designated. The correlation coefficient between fathers' and their sons' heights are about 0.5, a positive and moderate linear association...

Computing the Correlation Coefficient (r)

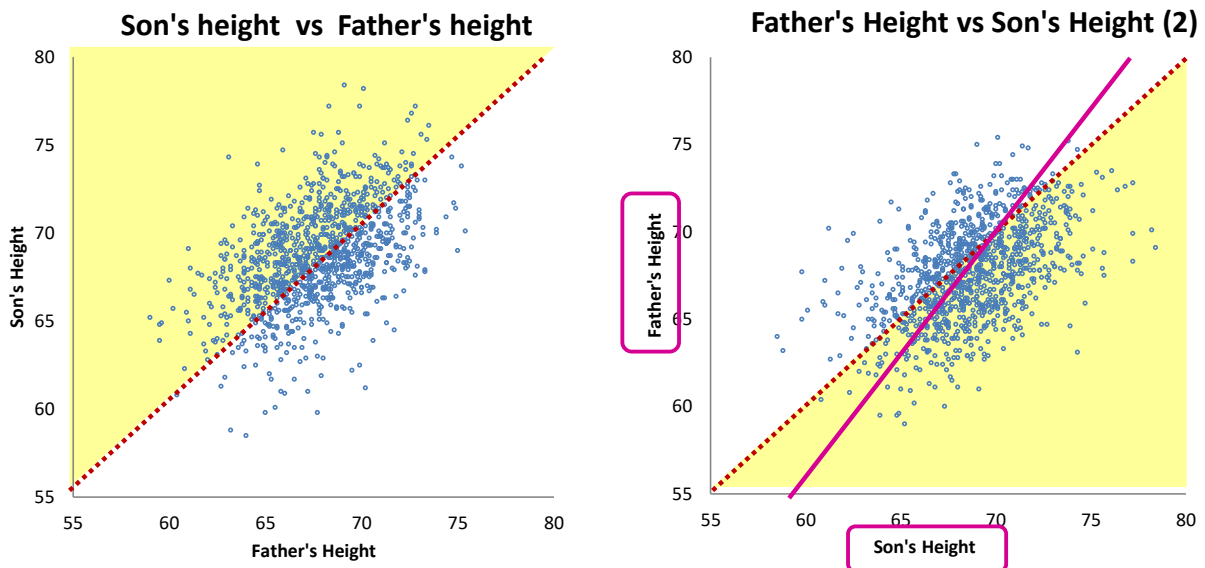


This seems to be in-line with what we have seen in the scatter diagram.

You may have noticed, we have been always placed fathers' heights on x-axis, the horizontal axis; and the sons' heights on y-axis, the vertical axis.

Can we interchange the two variables? What will happen to the correlation coefficient if we put sons' heights on x-axis and fathers' heights on y-axis?

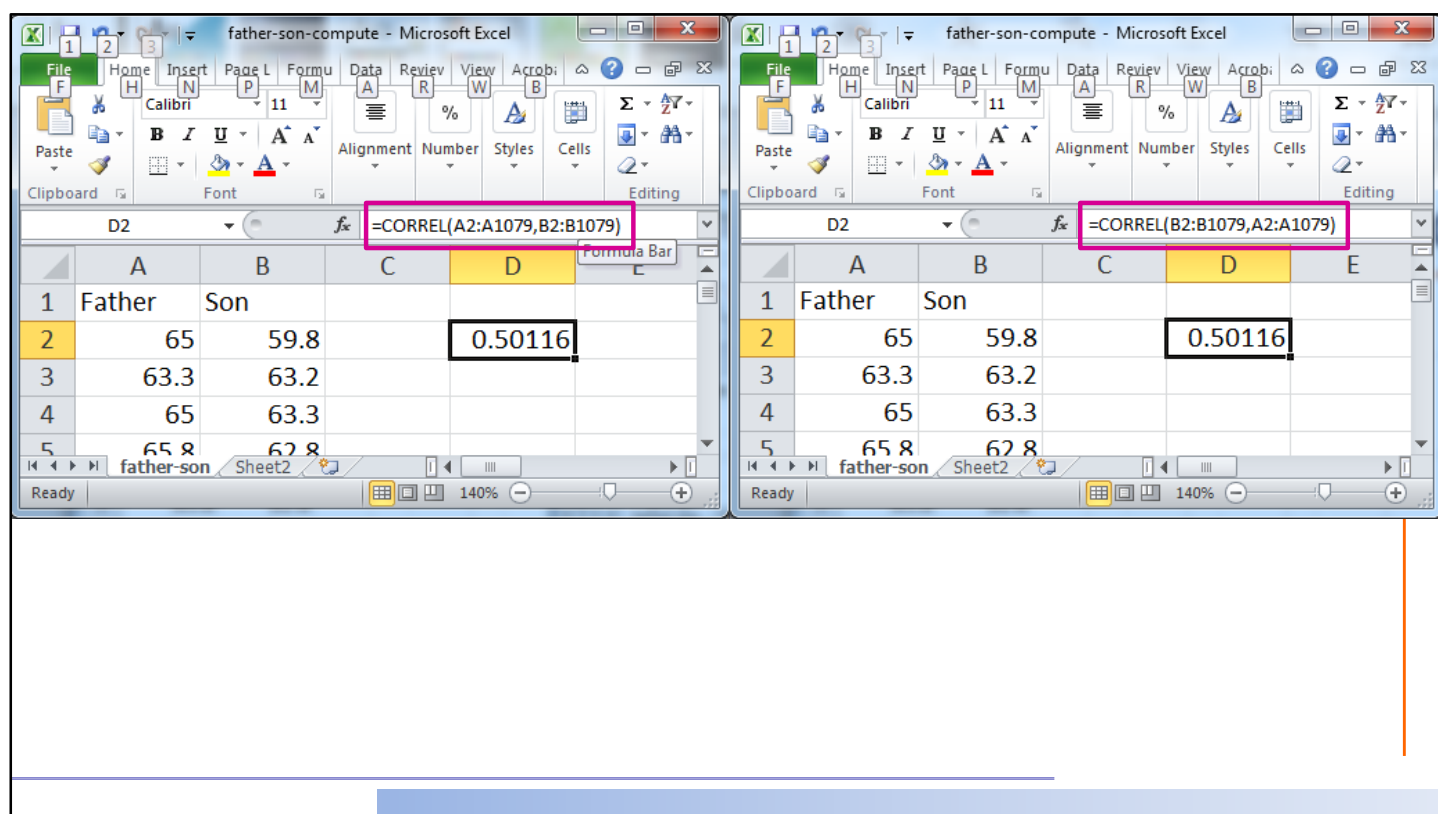
Computing the Correlation Coefficient (r)



Maybe we can put two scatter diagram side-by-side. This is the scatter diagram with sons' heights being x-axis and fathers' heights being y-axis. The two diagrams seem to look different.

Let's draw a diagonal through the first diagram as well as the second diagram. The image of data points along the upper diagonal in the first diagram would look like the data along the lower diagonal in the second diagram. They look like mirror images of each other. In the second diagram, we still observe a positive linear association.

How about strength of the linear association? Think about this: We are looking at the same set of two variables. The nature, the direction and the strength should not be different even we interchange the X and Y. That is, correlation coefficient, r , stays unchanged.



You can verify this from Excel worksheet. We interchange X and Y variables, the correlation coefficient stays unchanged.

About Correlation Coefficient (r)...

✓ r is a pure number without units

✓ r will not be affected by

① Interchange of the two variables

② Adding a number to all values of a variable

③ Multiplying a positive number to all values of a variable

Change of Scale

Let's summarize the properties of the correlation coefficient.

It is a pure number without unit. For example, fathers' heights and sons' heights in the data set were measured in inches. However, correlation coefficient r is about 0.5, a number without units.

It will not be affected by interchange of the two variables. We have seen this. Even we place son's height in x-axis and father's height in y-axis, the correlation coefficient r stays unchanged, which is 0.5.


In fact, even we add the same value to all fathers' heights, and adding the other value to all sons' heights, the correlation coefficient will still be 0.5.

Even if we multiply all values of fathers' heights with a positive number, and multiply all values of sons' heights with another positive number, the correlation coefficient will still be the same number, 0.5.

Points 2 and 3 actually tell us that the r value is not affected by the change of scale.



Unit: About the Correlation Coefficient

- ① *Computing correlation coefficient*
 - ② *Work with Excel*
 - ③ *Basic properties*
- 

As a quick recap...In this unit, we have discussed:

The steps for computing correlation coefficient, r , and shown how it can be achieved by Excel. We have also reviewed basic properties of correlation coefficient r .



Association

∞ Relationship between Two Variables ∞

Some Limitations with Correlation

In this unit, we will discuss some key limitations about correlation through examples.

Some Limitations with Correlation



✓ *Three key limitations...*

- ① *Causality*
- ② *Outliers in the data set*
- ③ *non-linear association*

We will talk about three key limitations with regard to the use of correlation, including

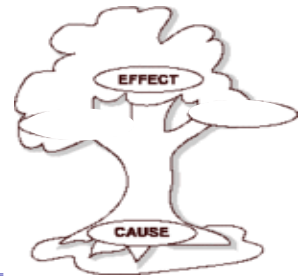
- 1) correlation and causation,
- 2) impact of outliers in the data set, as well as
- 3) non-linear association.

Correlation & Causation



What is “causation”?

- ✓ A change in one variable “produces” or “causes” a change in the other variable

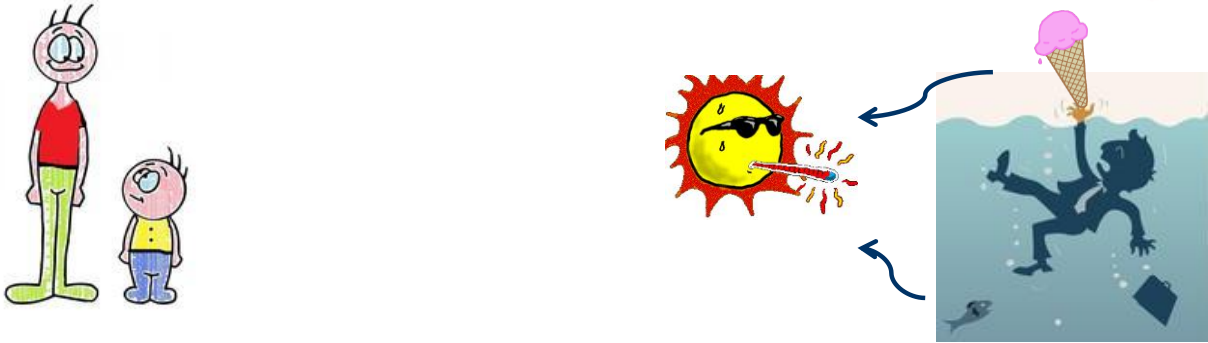


First, let's talk about correlation and causation. We have been talking about correlation. It indicates the association between two variables.

However, what is causation? Causation means that two variables are not only just associated, but that a change in one variable actually “produces” or “causes” a change in the other variable. This is a **cause-and-effect relationship between two variables**

Correlation & Causation

Correlation does NOT imply causation



It is important to note that correlation does not imply causation! It is impossible to prove the cause-and-effect relationship by just looking at correlation. A **strong correlation merely indicates a “strong evidence” of linear association** between two variables.

For example, there is a strong correlation between height and weight of a person. But this does not mean heavy weight “causes” a person to be “tall” or vice versa.

Another example, strong association between ice cream sales and number of death by drowning. If this is a cause-and-effect relationship, maybe we shall stop having ice cream!

It is obvious; these two variables are associated through a third variable -- the weather!

hot weather = more ppl buy ice cream + more ppl swim

So, moral of the story, correlation demonstrated association, not causation. Remember, association is not the same as causation!

A randomized experiment is necessary to establish the claim of cause-and-effect relationship.

Oral health link to heart trouble

US tests find that heart patients had more bacteria in their mouths

WASHINGTON: People with the most germ-infested mouths are the most likely to have heart attacks, United States researchers reported yesterday.

A study that compared heart attack victims to healthy volunteers found the heart patients had higher numbers of bacteria in their mouths, the researchers said. Their findings add to a growing body of evidence linking oral hygiene with overall health.

Dr Oelsa Andriankaja and colleagues at the University at Buffalo in New York

were trying to find out if any species of bacteria might be causing heart attacks.

Their tests on 386 men and women who had suffered heart attacks and 340 people free of heart trouble showed that two types of bacteria – *Tannerella forsythensis* and *Prevotella intermedia* – were more common among the heart patients.

But more striking, the people who had the most bacteria of all types in their mouths were the most likely to have had heart attacks, they told a meeting of the International Association of Dental Research in Miami.

"The message here is that even though some specific periodontal pathogens have been found to be associated with an increased risk of coronary heart disease, the total bacterial pathogenic burden is more important than the type of bacte-

ria," said Dr Andriankaja, who is now at the University of Puerto Rico.

"In other words, the total number of 'bugs' is more important than one single organism," she said.

Doctors are not sure how bacteria may be linked with heart attacks but several studies have shown associations between gum disease and heart disease.

In Singapore, while cardiac experts said that no local research had been carried out on the link between oral hygiene and heart problems, they said a connection could not be ruled out.

Dr Stanley Chia, associate consultant at the National Heart Centre's cardiology department, said that bacterial infection in general can cause changes to blood vessels, raising the risk of heart attacks. So, bacteria from the mouth could enter the bloodstream and possibly trigger a chain reaction leading to heart attacks, he said.

And with four out of five adult Singaporeans believed to have some form of gum disease, the latest study may act as a reminder to brush up on oral hygiene.

As Dr Koh Chu Guan, senior consultant at the periodontics unit of the National Dental Centre, observed: "A lot of Singaporeans don't visit a dentist and do not know about dental diseases."

REUTERS

Additional reporting by April Chong in Singapore

Straits Times, 2 April 2009

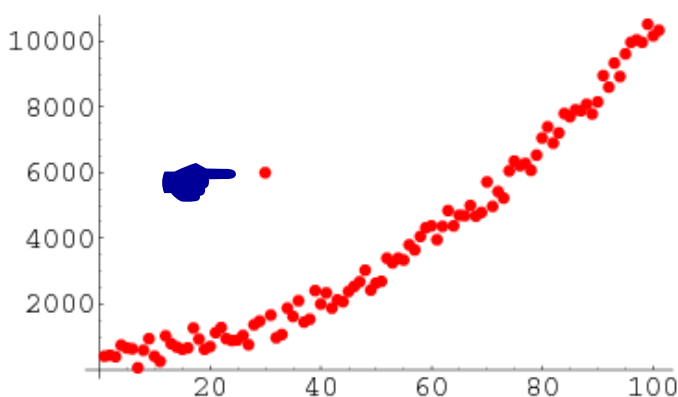
Let's look at this news clip. It reported a study outcome. Many studies have shown the link between oral health and heart diseases.

Does it mean that oral health problems may "cause" heart disease in the future? Maybe. But we can't draw such conclusion through an observational study. We only establish the association!

Impact of Outliers on Correlation

Outliers

Data points that are 'unusually' far away from the bulk of the data



Next, we will talk about how the outliers in a data set influence the correlation...

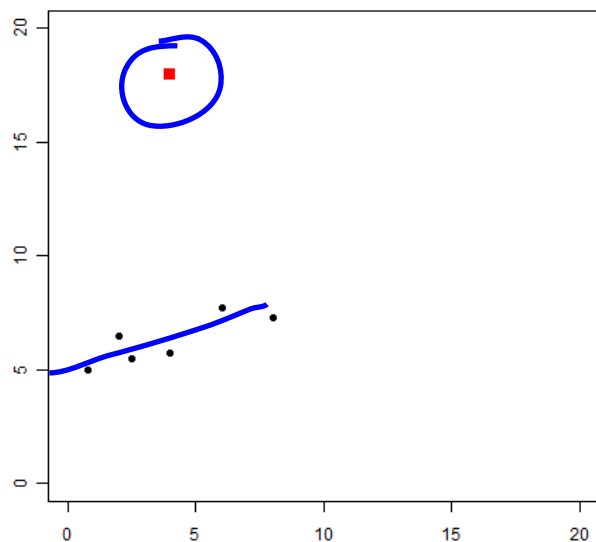
What is an outlier in a data set? In general, the data points that are unusually far away from the bulk of data are thought to be outliers.

Though there is no rigid definition of an outlier, few commonly used methods are available to detect potential outliers. We will not further discuss these methods here. However, we may be able to identify an outlier or outliers visually from a scatter diagram.

In this diagram, it is quite clear that this data point is, potentially, an outlier. Some outliers may be more influential than others in determining correlation. Without this outlier in the diagram, there seems to be a strong association between two variables. Shall we just remove it?

It is dangerous to exclude outliers from the analysis, for the sake of "creating" an association, without understanding the causes of occurrence. Outliers must be handled with care in any analysis, as they may be telling us something very valuable about the association between two variables.

Impact of Outliers on Correlation

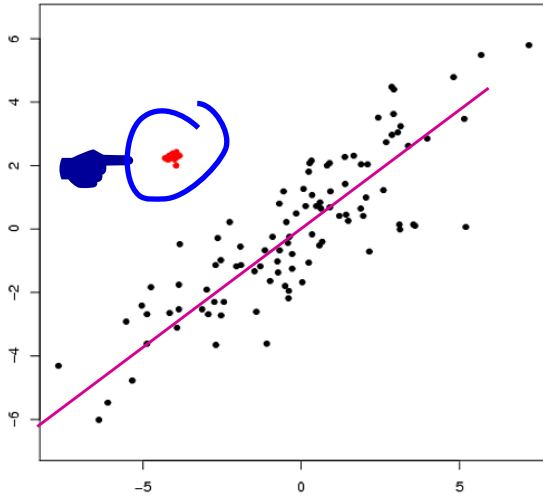


→ $r = ?$

Now, let look at a few examples to understand the impact of outliers on correlation.

Let's look at this simple scatter diagram. There seems to be a strong linear association between the two variables. Correlation coefficient, r , is about 0.86. However, there is an outlier (in red) in the original data set that was removed. This outlier is quite influential! It drastically decreases the correlation coefficient to 0.22.

Impact of Outliers on Correlation



→ $r = ?$

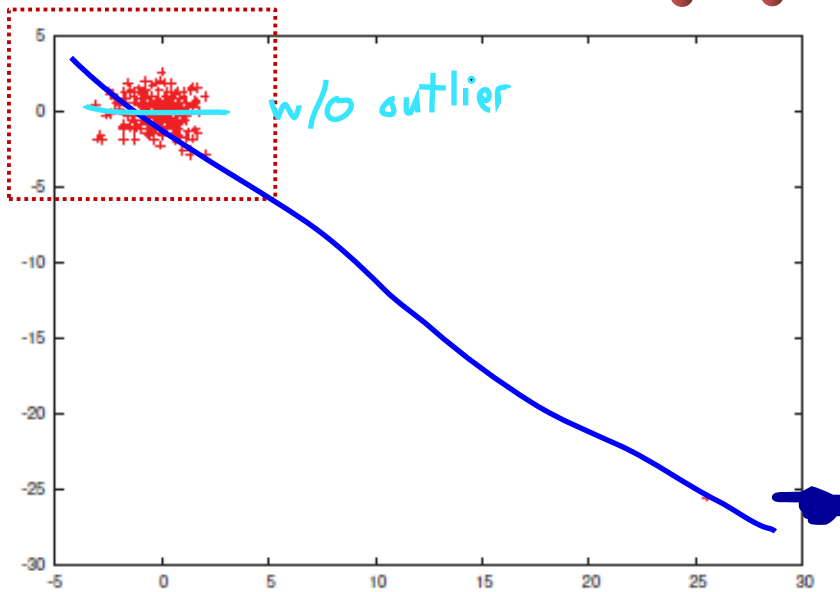
0.5 → 0.84
after removing

The data presented in this scatter diagram show that there are few data potential outliers (pointed, in red). The data points seem to lie quite closely to a straight line, aside from these few outliers. The correlation coefficient is about 0.5.

If we remove these few outliers, the correlation coefficient will change to 0.84.

From these two examples, the presence of outliers seemed to decrease the correlation. However, there are situations where the outliers will, actually, increase the correlation!

Impact of Outliers on Correlation

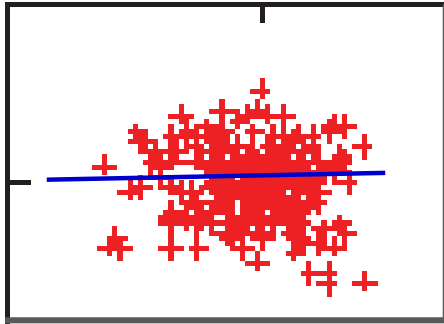


→ $r = ?$

In this diagram, it is clear that this data point is an outlier (pointed by an index finger). The correlation coefficient is about negative 0.75 (-0.75).

This is, actually, quite an influential outlier. With this outlier, the linear association is quite strong. However, if we remove the outlier and only focus on these data points...

Impact of Outliers on Correlation

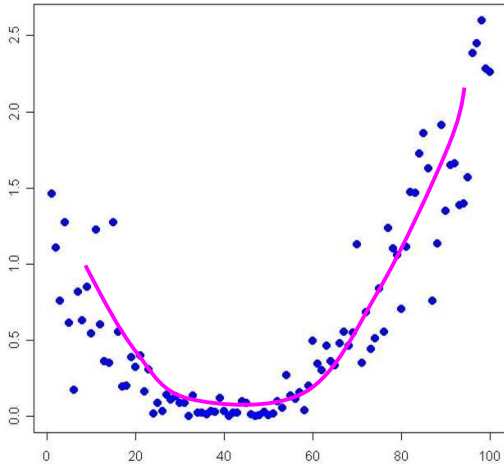


$$\rightarrow r = 0.01$$

The linear association is extremely weak. The correlation coefficient is only 0.01, very close to zero! The strength of correlation is actually increased when we consider this outlier...

From the above examples, we learn that correlation coefficient can be very sensitive to outliers. Existence of outliers may either inflate or deflate the value of a correlation coefficient.

r and Non-Linear Association



Zero correlation ($r = 0$)

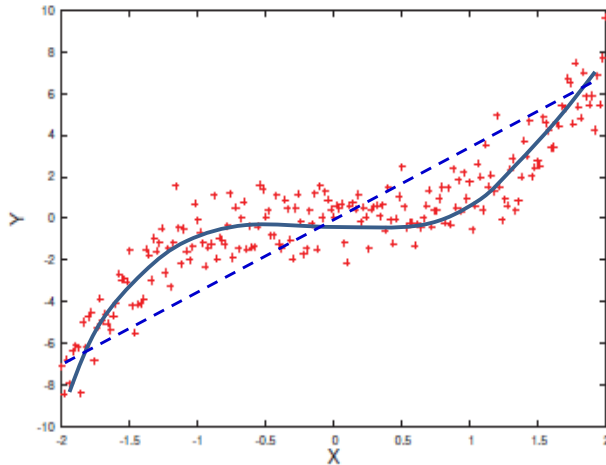
➔ No linear association

Correlation coefficient describes the direction and strength of linear association. It does not measure non-linear association between two variables.

For example, correlation coefficient for the two variables illustrated in this diagram is about zero. However, the association between the two variables can be described by a curve. This is a non-linear association.

That is, zero correlation only tells us no “linear association”.

r and Non-Linear Association



→ A non-linear association
(but $r \approx 0.85$)

Let's look at this scatter diagram. How would you describe the nature of the association between the two variables? Linear or non-linear? It may be better described by such a curve, rather than a straight line.

However, the data seemed to lie closely to a straight line. If we just use the data to compute the correlation coefficient, we will have the r value of 0.85 -- A strong and positive linear association? I don't think so!

The key lesson -- Always look at the scatter diagram first before computing and interpreting the correlation coefficient.

Unit: Some limitations with Correlation

- ① *Correlation does not imply causation*
- ② *Outliers may increase or decrease the correlation*
Be aware of removal of outliers from the analysis
- ③ *Correlation does NOT measure non-linear associations!*

Just a quick recap... Three key limitations of correlation were discussed in this unit: First, correlation does not imply causation. Next, presence of outliers may increase or decrease the correlation. It is dangerous to remove outliers in a data set just to “improve” or “manipulate” the correlation. The last one, correlation measures linear association, not non-linear association.

Don’t be contented with the r value; be aware of the misuse and the inappropriate interpretations of correlation.

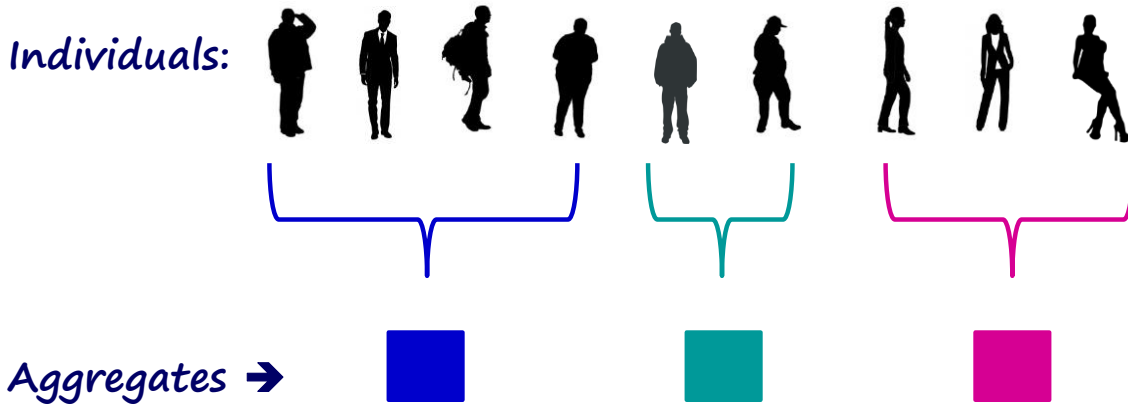
Association

∞ Relationship between Two Variables ∞

Ecological Correlation

Aggregated data are usually easier to obtain than data on individuals, especially in sociology, epidemiology and political science. In this unit, we will discuss the correlation coefficient computed based on such aggregate data. It is called Ecological correlation.

What is Ecological Correlation?



What is Ecological correlation?

We compute the correlation coefficient of the two variables based on individuals in the data set. For example, in Pearson's father and son data, we use the fathers' heights and sons' heights from each father-son pair to get the correlation coefficient. However, an ecological correlation is computed based on aggregated data, such as group averages or rates.

First, let's visualize the structure and understand the correlation computed at individual level and ecological correlation based on group aggregates. Here we have individuals selected from a population, some males and some females. Correlation coefficient can be computed based on all these individual data. However, we can look at the aggregated data according to certain grouping factors, such as country, ethnic group, and so on. An ecological correlation is computed based on the aggregates.

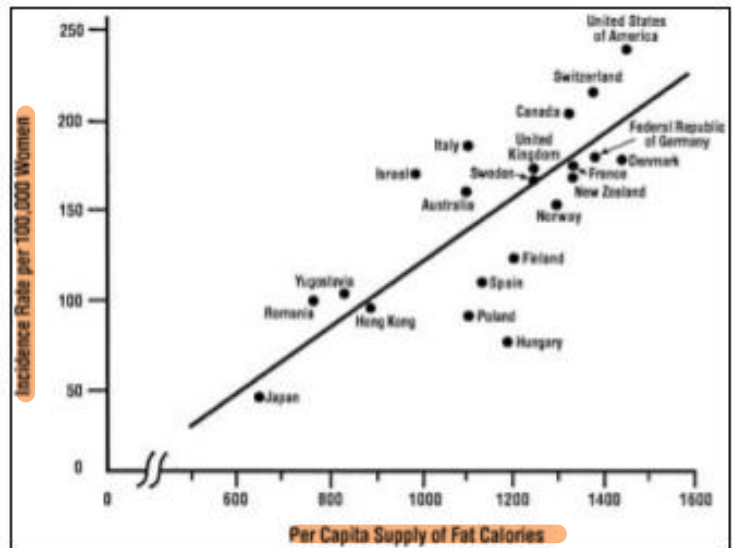
Now, we have realized that the ecological correlation is distinguished by the fact that unit used is not an individual; rather it is a group aggregate.

About Ecological Correlation...

Ecological Correlation:

Correlation based on aggregated data, such as group averages or rates

An ecological study of the relationship between dietary fat intake and breast cancer mortality. Sasaki, Horacek, Kesteloot. Prev Med 1993, 22(2), 187-202.



Correlation between dietary fat intake and breast cancer by country.

In short, correlation computed based on aggregated data, such as group averages or rates, is called “ecological correlation”. Ecological correlation is often useful. Especially, aggregated data are usually easier to obtain than data on individuals in reality.

For example, we may use ecological correlation to examine the association between the average exposures to a risk factor in various countries with the overall disease rates within the countries.

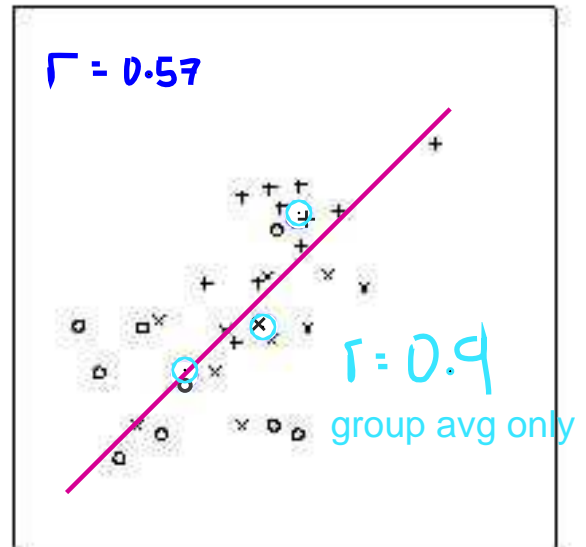
In this diagram, we look at the association between average fat intake of countries and overall breast cancer incidence rates within the countries. It shows a positive linear association. In other words, countries with higher average fat intake have higher breast cancer incidence rates among women. However, this diagram does not tell us the association between fat intake and breast cancer incidence among women in a certain country.

We must applied ecological correlation with caution.

About Ecological Correlation...

Ecological Correlation:

Correlation based on aggregated data, such as group averages or rates



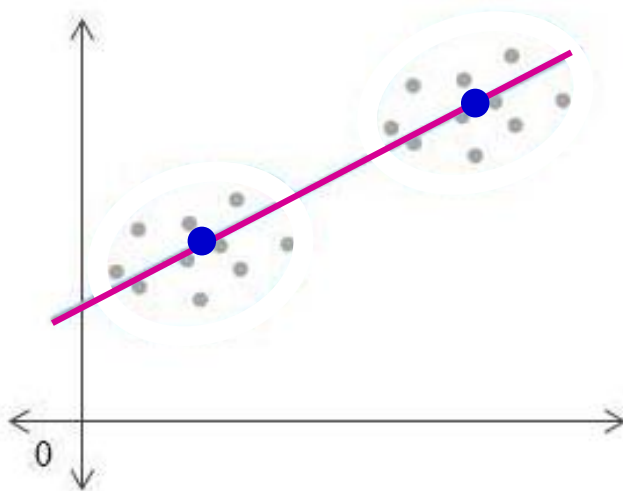
Let's look at this example. The data were collected from individuals from three different groups, circle, cross and plus are used to show the group an individual was selected from. The correlation based on individual seems to indicate a moderate and positive linear association, with r equals 0.57.

If we only use the averages for each of the three groups, they are represented by blue dots; we can compute correlation based on these group average values too. It appears that three aggregates lies very closely to the straight line, with an r values of 0.9.

In other words, this data set showed that the ecological correlation, which is computed based on group averages, appears to indicate a stronger linear association as compared to correlation computed based on individuals.

Is this always the case?

The Anatomy of a Ecological Correlation



→ Association is “overstated”
based on aggregated data

Graph: inoyo.net/wp-content/uploads/2012/09/ecological-fallacy.jpg

We shall answer this question with some graphical illustrations showing structural connections between correlations among individual and ecological correlation in specific situations.

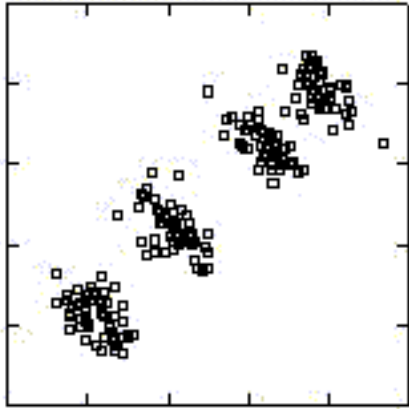
From this scatter diagram, the data points seem to have a positive linear correlation. The strength may be moderate to strong. However, it is clear to us that there may be two distinctive groups in the data set. We can compute the average for each group. When we examine the correlation based on these group averages, strength of linear association has become stronger. So, if we draw conclusions about the linear association between two variables based on an ecological correlation, the strength of such association tends to be overstated!

In general, when the associations for both individuals and aggregates are in the same direction, the ecological correlation, based on aggregates, will typically overstate the strength of the association in individuals. That's because the variability among individuals will be eliminated once we use the group aggregates.

In general, when the associations for both individuals and aggregates are in the same direction, the ecological correlation, based on aggregates, will typically overstate the strength of the association in individuals.

That's because the variability among individuals will be eliminated once we use the group aggregates.

The Anatomy of a Ecological Correlation



Graph: www.jerrydallal.com/lhsp/pix/ecorr2.gif

Ecological Fallacy:

Deduce the inferences on correlation about individuals based on aggregated data

Let's look this scatter diagram.

Assume that we investigating the association between two variables, a certain food consumption and cancer risk. It is quite obvious that there are four distinctive groups, say, they represent 4 different countries. There seems to be a negative linear association among the two variables within each country.

Based on this, I would think this particular food consumption is associated with low cancer risk! However, if we look at the average of each country, the correlation based on averages seems to show a very strong, actually, almost a perfect, positive linear association. That is, a country with higher average of such food consumption seems to show a higher cancer risk.

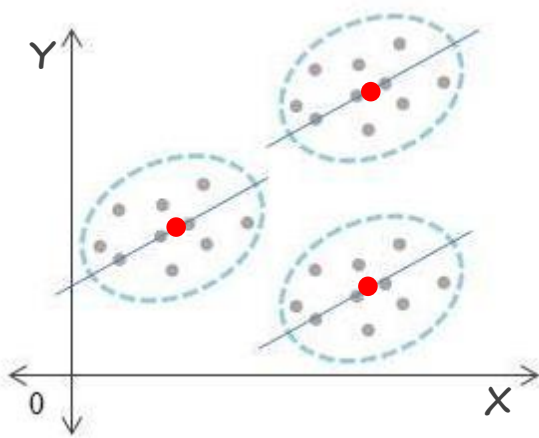
What a contradiction!

Moral of the story, we should not assume that correlations based on aggregates will hold for individuals. **Remember, an ecological correlation and correlation based on individuals are not the same.**

A common pitfall that we often encounter is that **people draw conclusions about individuals based on aggregated data. This is an "ecological fallacy"!**

The ecological correlation is misleading. We should have studied individuals rather than group aggregates to see the correlation between the two variables among individuals.

The Anatomy of a Ecological Correlation



Atomistic Fallacy:

Generalize the correlation based on individuals towards the aggregate-level correlation

Graph: inoyo.net/wp-content/uploads/2012/09/ecological-fallacy.jpg

Here is another example. This diagram actually presents a picture which is a complete opposite of ecological fallacy.

It is quite clear that there are three distinctive groups of individuals, and we observed positive linear association between the two variables within each of the three groups.

However, is it appropriate to generalize the correlation observed based on individuals towards the aggregate level correlation? As we plot the aggregates for each group, no clear correlation can be inferred from the diagram.

Applying correlation based on individuals towards the aggregates is "Atomistic Fallacy".

Unit: Ecological Correlation

→ *What is ecological correlation?*

→ *Ecological correlation*

vs.

Correlation at individual level

→ *Ecological fallacy*

vs.

Atomistic fallacy

Just a quick recap... We have discussed the followings:

The meaning of ecological correlation: Ecological studies can be very valuable in areas such as sociology, epidemiology and political science.

We have also talked about the conclusions drawn based on ecological correlation, as well as the differences between ecological correlation and the correlation at individual level.

Ecological correlation tends to overstate the correlation at individual level under certain circumstances. Two research fallacies in studying ecological correlations were introduced. Ecological fallacy tells us that it is not appropriate to draw conclusion about individuals based on aggregated data; while Atomistic fallacy shows that the correlation observed among individuals may not apply to aggregated data.

The bottom line is – Handle the ecological correlation with care; A scatter diagram is always helpful!



Association

∞ Relationship between Two Variables ∞

Some Cautionary Notes on Correlation

In this unit, we draw attention to two phenomena and situations with the use of correlation for measuring linear association. Examples are given in order to demonstrate these phenomena and situations.

Attenuation Effect

✓ Range restriction –

- ➔ a bivariate data set are formed based on certain criteria on one variable
- ➔ the data for the other variable only available for a limited range



The phenomenon that we will discuss here is attenuation effect. The occurrence of such an effect in correlation is closely related to range restriction of available data. In general, we would like to have a full range of data to account for the variability in computing correlation coefficient.

“Range restriction” occurs when the units in bivariate data set are formed based on the selection of one variable. Say, X. As a result, the values of the other variable, Y, only available for a restricted or limited range.

I think it may be easier if we use a familiar example to demonstrate the term, “range restriction”.

Attenuation Effect

Like Father, Like Son

- ✓ Fathers' heights: 59 ~ 75.4 inches
- sons' heights: 58.5 ~ 78.4 inches
- Correlation coefficient (r) → About 0.5



Let's look at Karl Pearson's father-and-son data set. Fathers' heights vary from 59 to 75.4 inches (That is, 150 to 191 centimeters). Their sons' heights extend from 58.5 to 78.4 inches (That is, 148.5 to 199 centimeters). The correlation coefficient is about 0.5, a moderate and positive linear association.

Attenuation Effect

Like Father, Like Son

✓ Select data with fathers → At least 70 inches tall
(227 father-son pairs)

Their sons' heights: 61.2 ~ 78.2 inches

Correlation coefficient (r) → 0.24

How about data with the fathers who are at least 70 inches tall? That's at least 177.8cm. I would say, tall fathers. There are 227 father-son pairs with fathers being at least 70 inches in height.

How about the range of their sons' heights among the 227 pairs? It varies from 61.2 to 78.2 inches. That's 155.5 to 198.6 centimeters.

The 227 father-son pairs form a range restricted data set.

So, what's the correlation coefficient, r , for these 227 father-son pairs? About 0.5 too? The same nature and strength of association as what we observed among the 1078 father-son pairs?

Or smaller than 0.5? Or larger than 0.5? What do you think?

Range restriction tends to have a diminishing influence on the strength of association!

The correlation coefficient for these 227 father-son pair is 0.24, quite a decrease from 0.5!

With such range restriction, the strength of the association between fathers, at least 70 inches tall, and their sons' heights has decreased to a weak association! Such an effect is "attenuation" effect.

Attenuation Effect

Due to range restriction in one variable, the correlation coefficient obtained tends to “understate” the strength of association between two variables

In short, the attenuation effect refers to the phenomenon:

Due to range restriction in one variable, such as fathers' heights were restricted to the range of at least 70 inches tall. The correlation coefficient computed based on this range restricted data set tends to be smaller in magnitude. Therefore, it understates the strength of association between the two variables.

Removal of Some Data

3	Temperature °F	Damage Index
4	53	11
5	57	4
6	58	4
7	63	2
8	66	0
9	67	0
10	67	0
11	67	0
12	68	0
13	69	0
14	70	4
15	70	0
16	70	4
17	70	0
18	72	0
19	73	0
20	75	0
21	75	4
22	76	0

We are moving on to a commonly asked question about data used for analysis. For example, in a data set showing damage indices at certain temperatures, there are some data points with seemingly unimportant information, such as “no damage”. Can we remove or exclude these data points from analysis, such as data presentation or computation of correlation coefficient? Well, removal of data from analysis may change the whole picture of the analysis and result in a different or even opposite inferences. It could be quite dangerous! In fact, we have learned this from history.

Space Shuttle: Challenger



Photos: NASA

This is a piece of sad history, the journey of space shuttle Challenger. Almost 30 years ago, on the bitter cold morning of January 28, 1986, Space Shuttle Challenger disintegrated 73 seconds after lift-off.

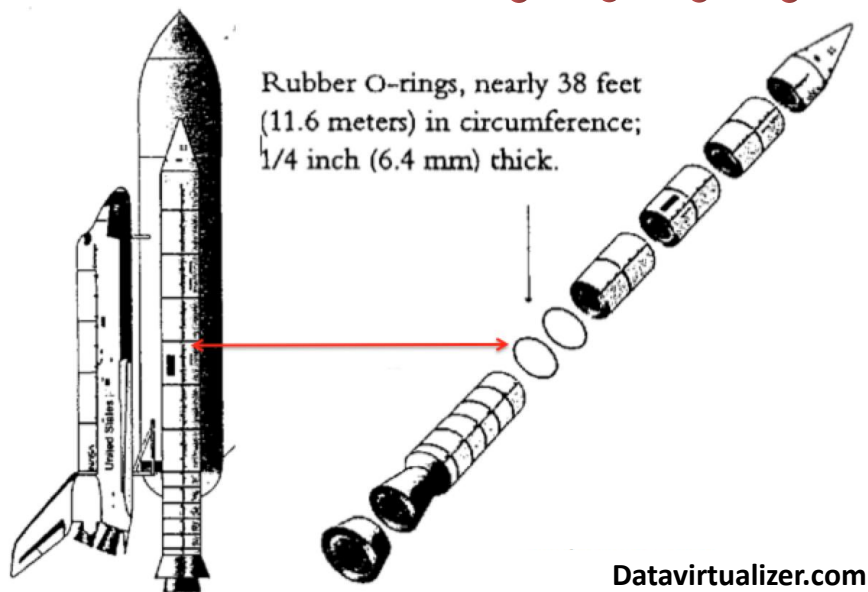
What went wrong?

The night before, the engineers held a conference call with NASA management from Kennedy Space Center and Marshall Space Center. During the call, the engineers warned NASA to cancel the launch due to the cold weather and the possibility of rubber O-rings losing their ability to seal the joints of the solid rocket boosters.

The engineers failed to convince NASA management with their presentations.

Could the engineers prove, without a doubt, that the forecasted cold temperature posed a high enough risk to the shuttle that NASA should postpone liftoff? Many believed the history could have changed if the complete data were presented and analyzed in a more appropriate way.

Space Shuttle: Challenger



This is a piece of sad history, the journey of space shuttle Challenger. Almost 30 years ago, on the bitter cold morning of January 28, 1986, Space Shuttle Challenger disintegrated 73 seconds after lift-off.

What went wrong?

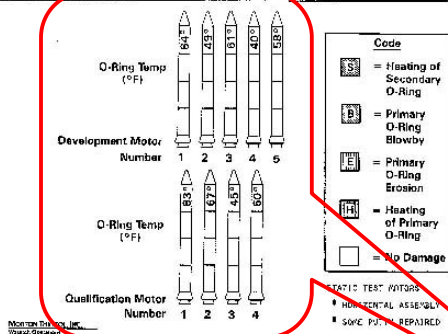
The night before, the engineers held a conference call with NASA management from Kennedy Space Center and Marshall Space Center. During the call, the engineers warned NASA to cancel the launch due to the cold weather and the possibility of rubber O-rings losing their ability to seal the joints of the solid rocket boosters.

The engineers failed to convince NASA management with their presentations.

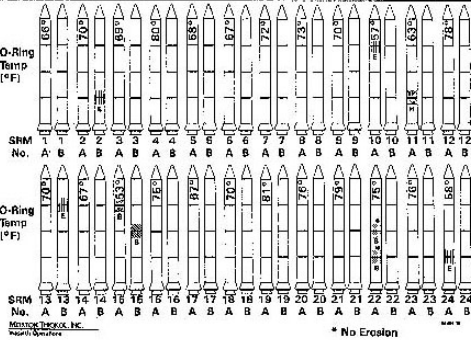
Could the engineers prove, without a doubt, that the forecasted cold temperature posed a high enough risk to the shuttle that NASA should postpone liftoff? Many believed the history could have changed if the complete data were presented and analyzed in a more appropriate way.

Space Shuttle: Challenger

History of O-Ring Damage in Field Joints



History of O-Ring Damage in Field Joints (Cont)

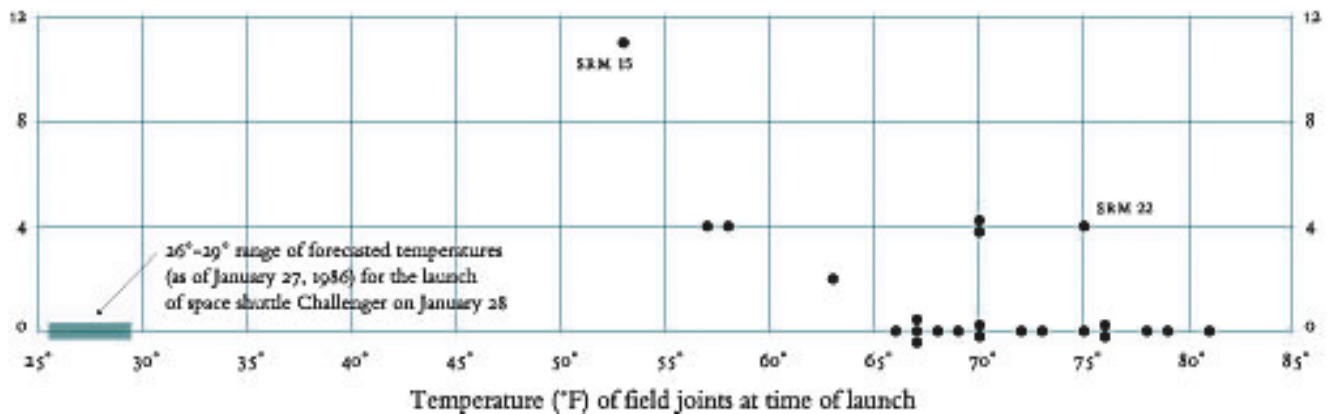


Let's try to understand the incidence from the data collected and presented to the NASA. Actually, you are looking at the piece of information presented to NASA the night before.

Out of all data available, the focus was only given to 7 data points with O-ring failure occurrences.

Space Shuttle: Challenger

O-ring damage
index, each launch



Graph by Prof Edward Tufte

If we look at this illustration done by Professor Edward Tufte, a statistician, later after the Challenger explosion, there seems to be a negative association between temperature and damage index.

With this graphical presentation, it is quite a clear the risk of O-ring failure is high at low temperature! But these points were omitted from the presentation to NASA. Without these points, the picture on the O-Ring damage index versus temperature looks different! By excluding these data, the whole dynamics of the information has been diluted!

Lesson to learn: Some valuable information may be eliminated by removing data!

Unit: Some Cautionary Notes on Correlation

- *Attenuation effects*
- *Ignore or remove some data from analysis*

Just a quick recap, we have discussed the correlation for range restricted data and attenuation effects. Data with range restriction may exert a deteriorating influence on the strength of association.

Next, we have learned about the danger if we ignore or remove some seemingly unimportant data from analysis. In such situation, some valuable information may be left out and the results of analysis can be skewed.



Association

∞ Relationship between Two Variables ∞

Linear Regression



In this unit, we will learn how the linear regression line are fit to the data and see how we can predict the value of one variable with the information on the other variable.

Like Father, Like Son



✓ Fathers' heights (X): 59 ~ 75.4 inches

average height, $\bar{X} = 68$ inches

standard deviation, $sd_x = 2.74$ inches

✓ Sons' heights (Y): 58.5 ~ 78.4 inches

average height, $\bar{Y} = 69$ inches

standard deviation, $sd_y = 2.81$ inches

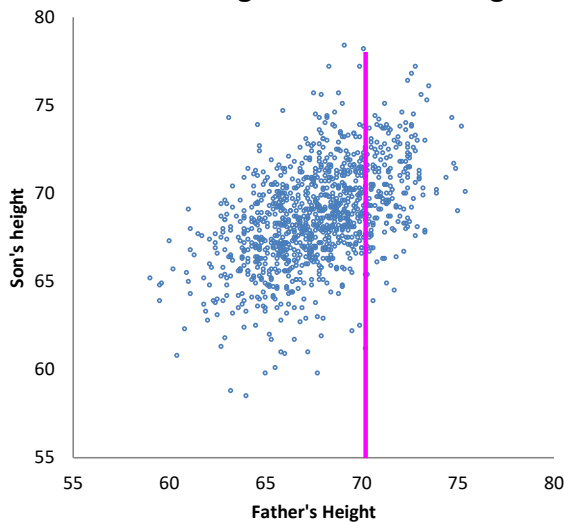
✓ Correlation coefficient (r) → About 0.5

Recall Karl Pearson's father-and-son data set. Here is summary information about fathers' heights in the data set. Fathers' average height is about 68 inches with a standard deviation of 2.74 inches. The sons have a taller average height and wider spread. The difference between father's and son's average height is about 1 inch. The correlation coefficient for the 1078 father-son pairs is about 0.5.

With a moderate and positive linear association, can we make use of information that we have on a father's height to predict his adult son's height?

Prediction

Son's height vs Father's height



Fathers' average = 68 inches

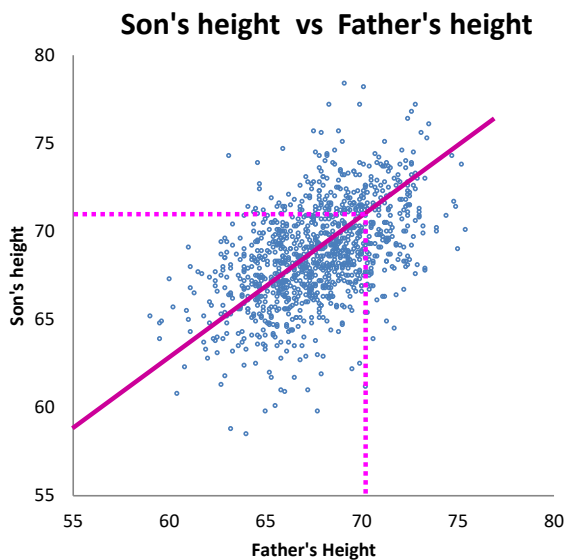
Sons' average = 69 inches

Father + 1 inch = Son

If a father is 70 inches tall, what is the best predicted height for his adult son? As the sons' average is taller than fathers' average by 1 inch, can I just assume a son should be taller than his father by 1 inch? So, how about 71 inches?

Probably not. The relationship between a father's and his son's is NOT deterministic! We need to take the variability of the sons' heights into account!

Prediction with a Regression Line



→ Use a Father's height (X) to predict a son's height (Y)

→ X: independent (Predictor)
Y: dependent

→ $Y = a + bX$

As we know there is a moderate linear association between fathers' and their sons' heights. If we could just find a straight line that best describes the linear relationship between a father and his son's height, we shall be able to predict sons' average height given that the father is 70 inches.

How do we find this straight line?

Conceptually, this is the line which best fits the data, and we call it the regression line. To write the equation for the regression line, we need to determine the predictor and the predicted outcome.

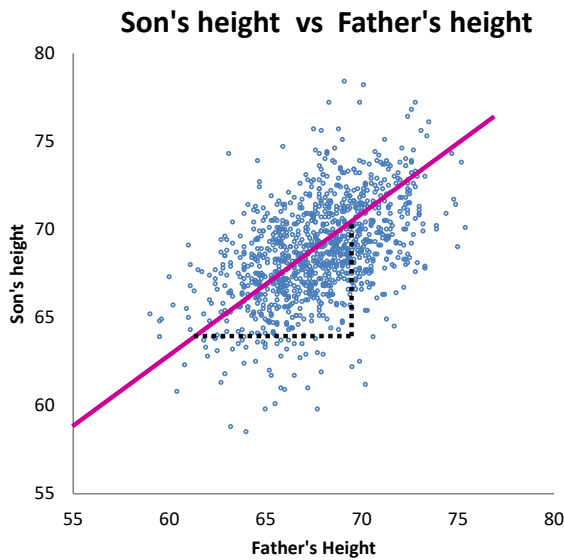
As we know father's height and wish to use the father's height to predict his son's height. Therefore, **father's height is the predictor, or so-called independent variable.**

With the information about the father's height, we may predict the son's height. Therefore, the **son's height is a dependent variable to be predicted.**

In general, this straight line can be expressed by such an equation, $Y = a + bX$ (Note: In Statistics, we use \hat{Y} to indicate a predicted value in the regression equation, rather than Y). Once we have a father's height, we can just plug in the equation to predict the son's height.

The only problem is that we need to find out the values for a and b in the equation.

Linear Regression



$$Y = a + bX$$

→ *a: intercept*

→ *b: slope*

'a' is the intercept of the regression line, which is the Y value when X value equals zero.

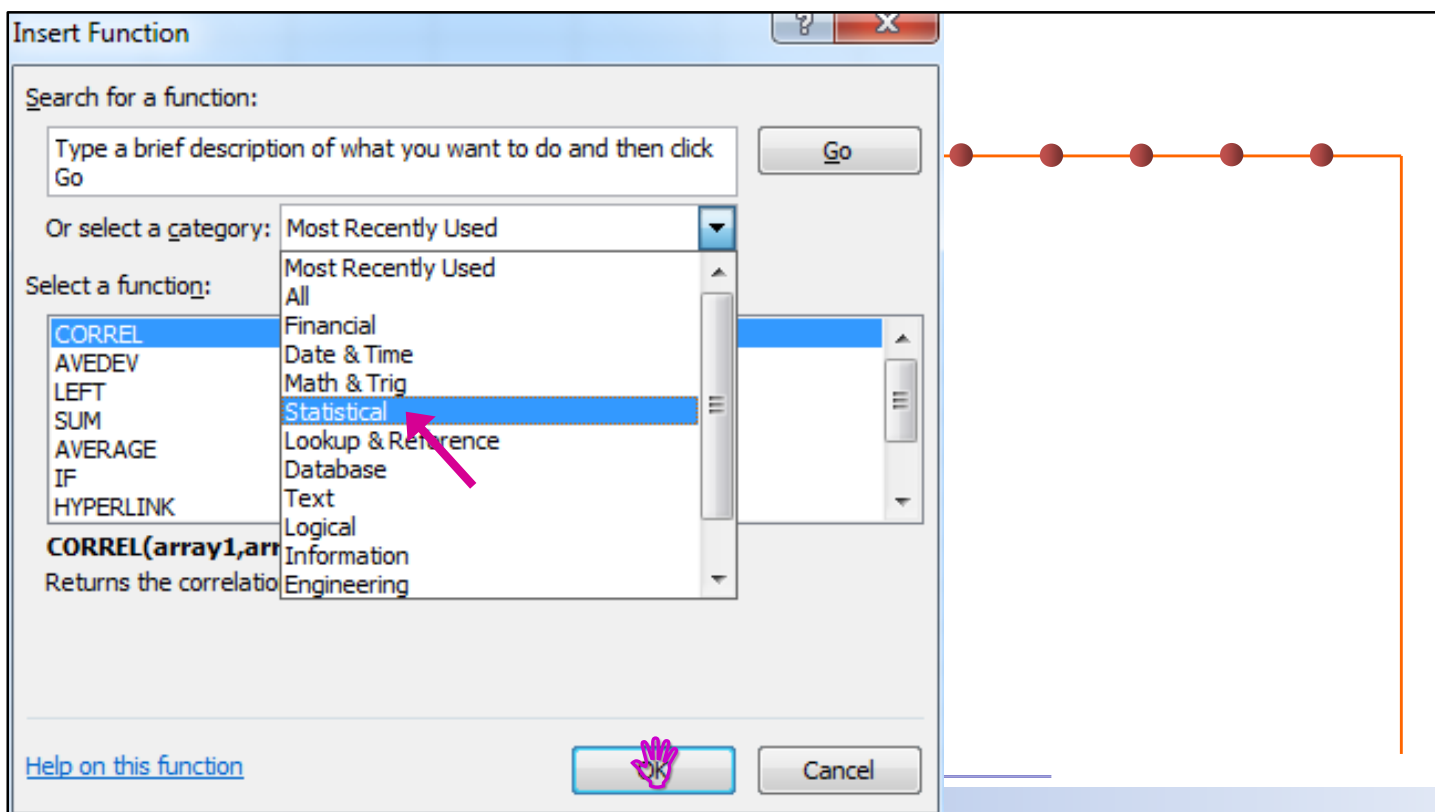
'b' is the slope of the regression line. It can be interpreted as "the amount of change in Y, when X value increased by one unit".

The intercept and slope can be estimated by the data. We will not show the computational formula for determine the regression equation. Instead, we will use Excel to determine the intercept and slope of the linear regression equation

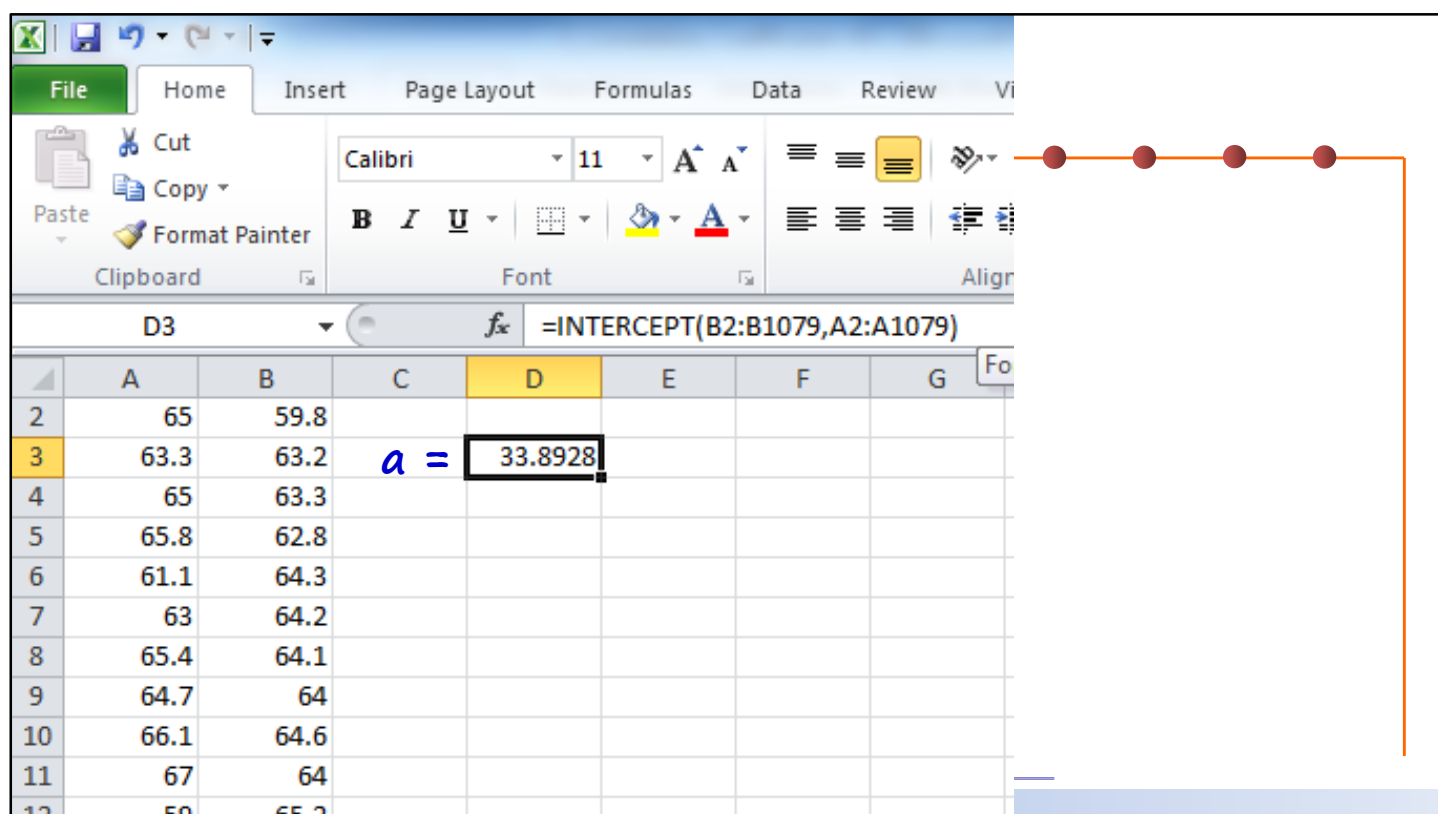
The screenshot shows the Microsoft Excel interface. The 'Insert Function' dialog box is open, displaying a list of functions. The function 'CORREL' is selected. The dialog box includes a search bar, a category dropdown set to 'Most Recently Used', and a list of functions including AVEDEV, LEFT, SUM, AVERAGE, IF, and HYPERLINK. The 'CORREL' function is highlighted, and its description, 'Returns the correlation coefficient between two data sets.', is visible. The background shows a spreadsheet with data for 'Father' and 'Son' scores.

	A	B
1	Father	Son
2	65	59.8
3	63.3	63.2
4	65	63.3
5	65.8	62.8
6	61.1	64.3
7	63	64.2
8	65.4	64.1
9	64.7	64
10	66.1	64.6
11	67	64

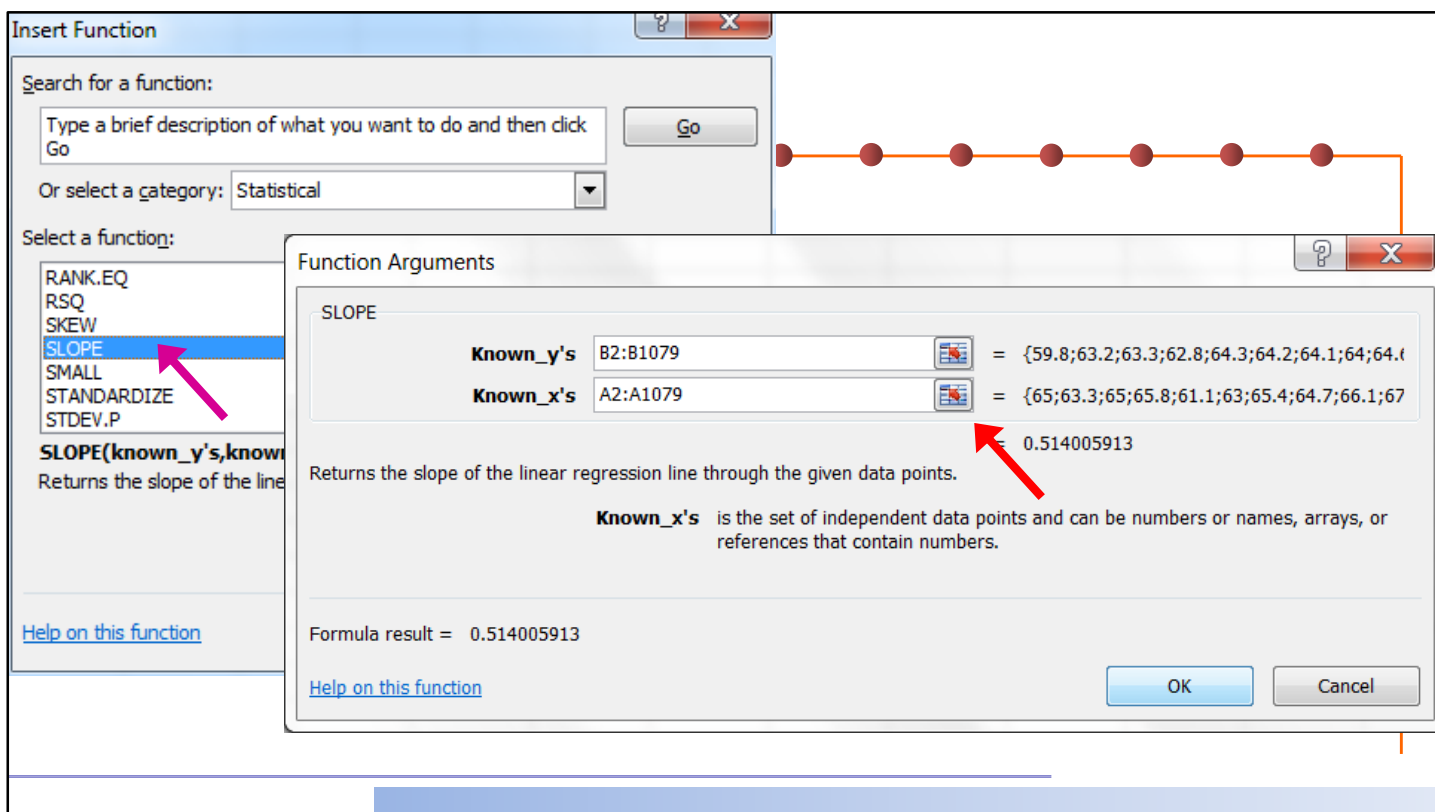
Here is the data set and we will click on this function dial to get the 'insert function' pop-up.



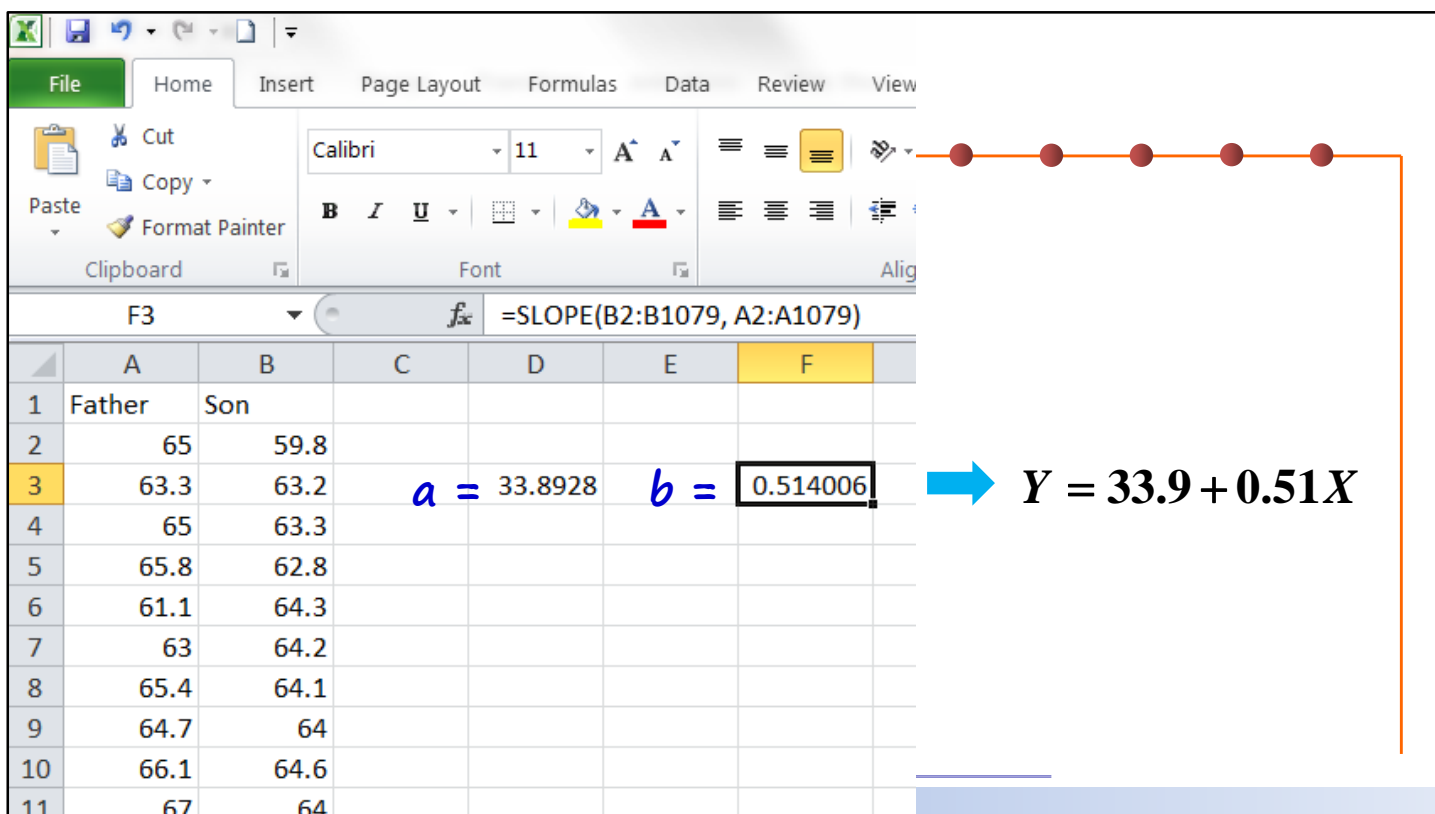
Select statistical function, and then click OK.



Here is the intercept.



Similarly, choose slope, then enter the data range in the pop-up window.

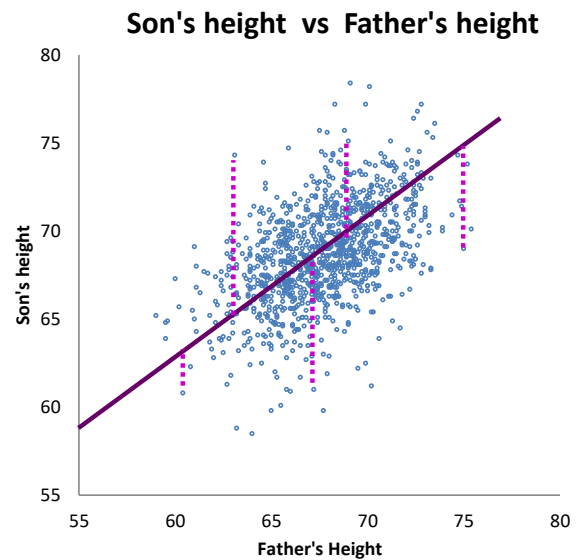


The slope will be computed. Here is the regression equation

More about Linear Regression

✓ Determine the best-fit regression line

→ Least-square method:



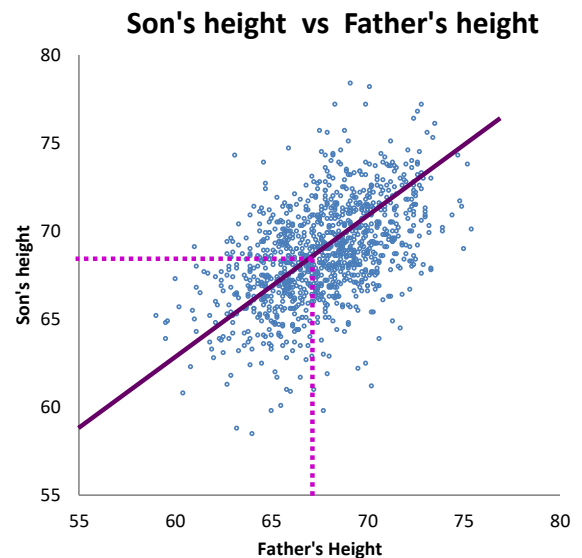
Now we have some basic ideas about linear regression and learn how the regression line, in particular, the slope and intercept of the linear equation, can be obtained by using Excel.

Have you ever wondered how the best-fit regression line is determined? It is determined by a method called least-square method, which we will not discuss further here.

The important thing is to know that the best-fit regression line is determined by least square method, so that the overall distance from the data points to the line is minimized.

More about Linear Regression

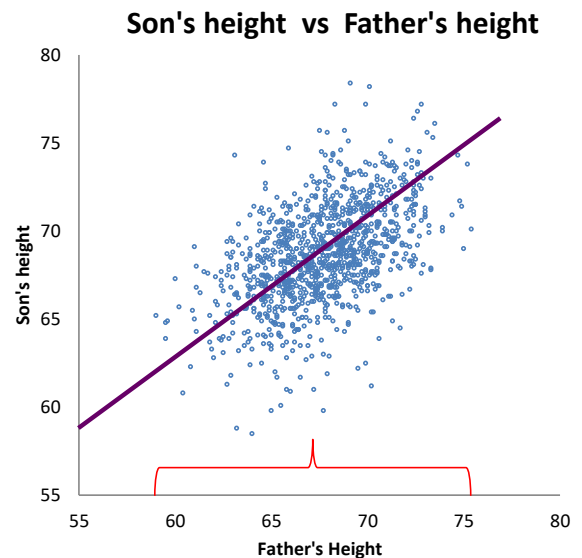
- ✓ Prediction of son's height with the regression line
 - ➔ Given a father's height the sons' average height can be estimated



Here is something that we need to pay attention to when we use the regression line for prediction or estimation. For example, given a father with height, say 67 inches, what we are predicting is NOT the exact height of his son. Instead, we are predicting the average height of the sons whose fathers are 67 inches.

More about Linear Regression

- ✓ *Prediction of son's height with the regression line*
 - ➔ *Prediction of sons' heights beyond observed range of father's heights is dangerous!*



Predicting the sons' heights given their fathers' heights, we also need to take note of the range of the fathers' heights in the data set.

You may recall the observed range for fathers' heights. It varies from 59 to 75.4 inches. The prediction of sons' heights based on the regression line is only applicable to those fathers with heights within such observed range.

Any prediction of sons' heights given fathers' heights beyond this range, will be dangerous. With wider range of fathers' height collected, the best-fit regression line may change!



Unit: Linear Regression

- ① *Introduce linear regression*
- ② *Regression equation*
- ③ *Linear regression with Excel*

As a quick recap...we have introduced:

The basic idea of linear regression; and the intercept and slope of the regression equation; we have also shown how we can obtain the regression equation using Excel.