STATISTICS FOR DATA SCIENCE

Power Test & Simple Linear Regression

Dr. Karthiyayini

Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

Unit 5 : Power Test & Simple Linear Regression

Session : 5

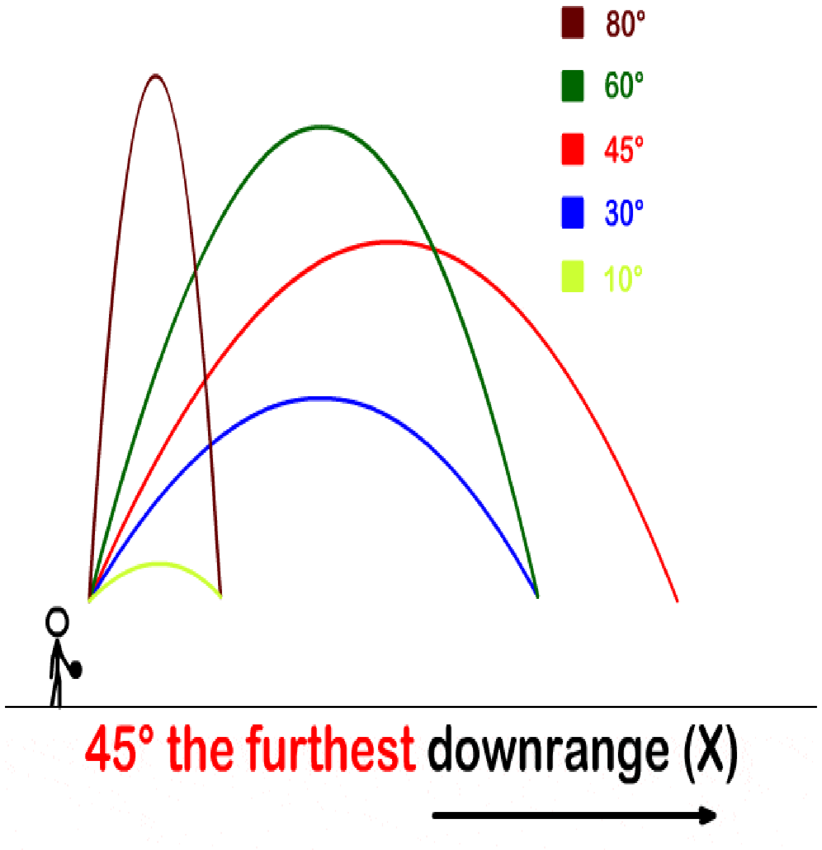
Sub Topic : Correlation & Regression Analysis

Dr. Karthiyayini

Department of Science & Humanities

- More about Correlation coefficient!
- What is Regression Analysis ?
- How to obtain the Least Squares line for a given set of Bivariate data?

Correlation Coefficient measures only Linear Association



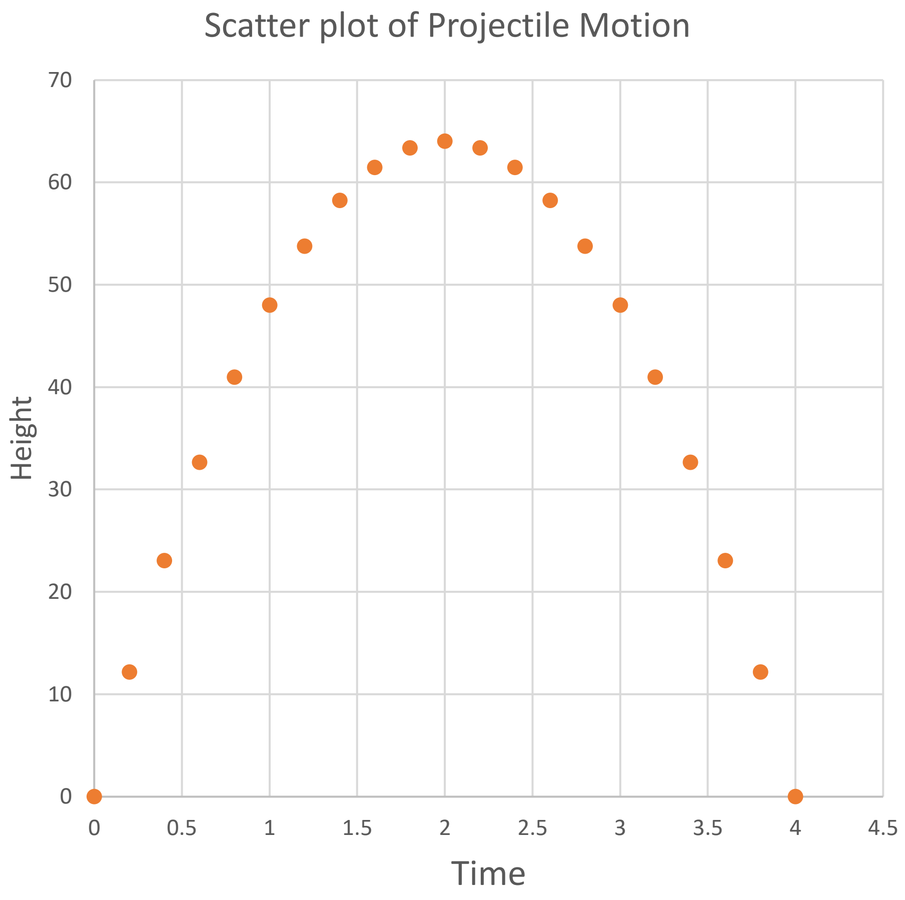
- Consider the projectile motion of an object that is fired in the upward direction from the ground.
- Initial velocity $= 64ft/s$
- Equation of the path traced by the object : $y = 64x - 16x^2$
- Correlation Co-efficient $r = 0$

Source : stickmanphysics.com

Correlation Coefficient measures only Linear Association

Observations :

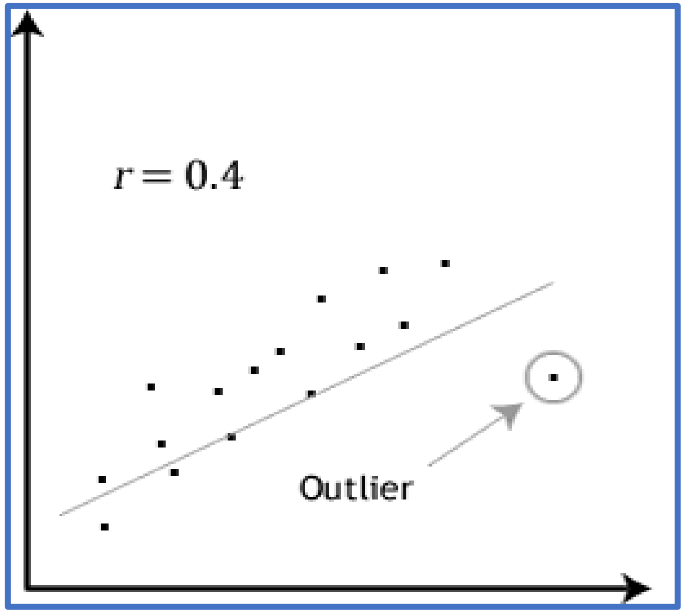
- ❖ $r = 0 \implies$ there is *no linear relationship* between the time and height. But it is a well know fact that in this case the height of the object varies with time.
- ❖ The relationship is non linear in this case.
- ❖ This example indicates that if correlation coefficient is used in non linear relationships, the results obtained will be misleading.
- ❖ Therefore, correlation coefficient measures only Linear relationships.

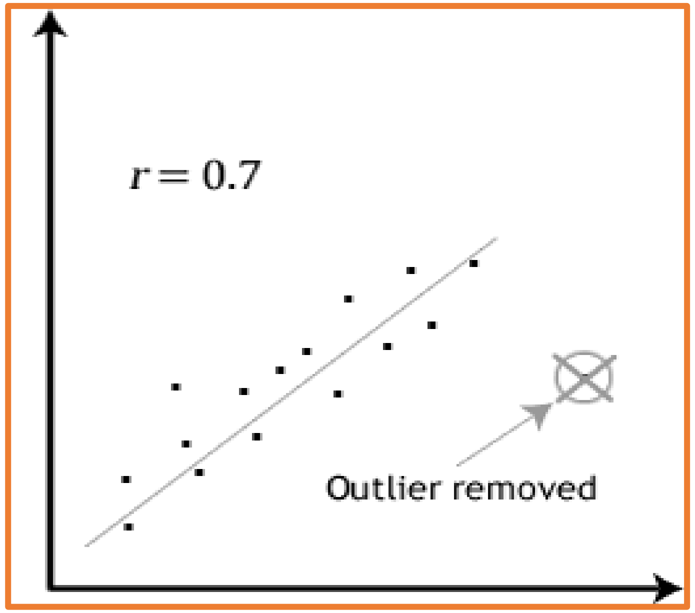


Correlation Coefficient - Misleading when outliers are present

❖ Outliers can have a very large effect on the line of best fit and the Pearson correlation coefficient, which can lead to very different conclusions regarding your data.

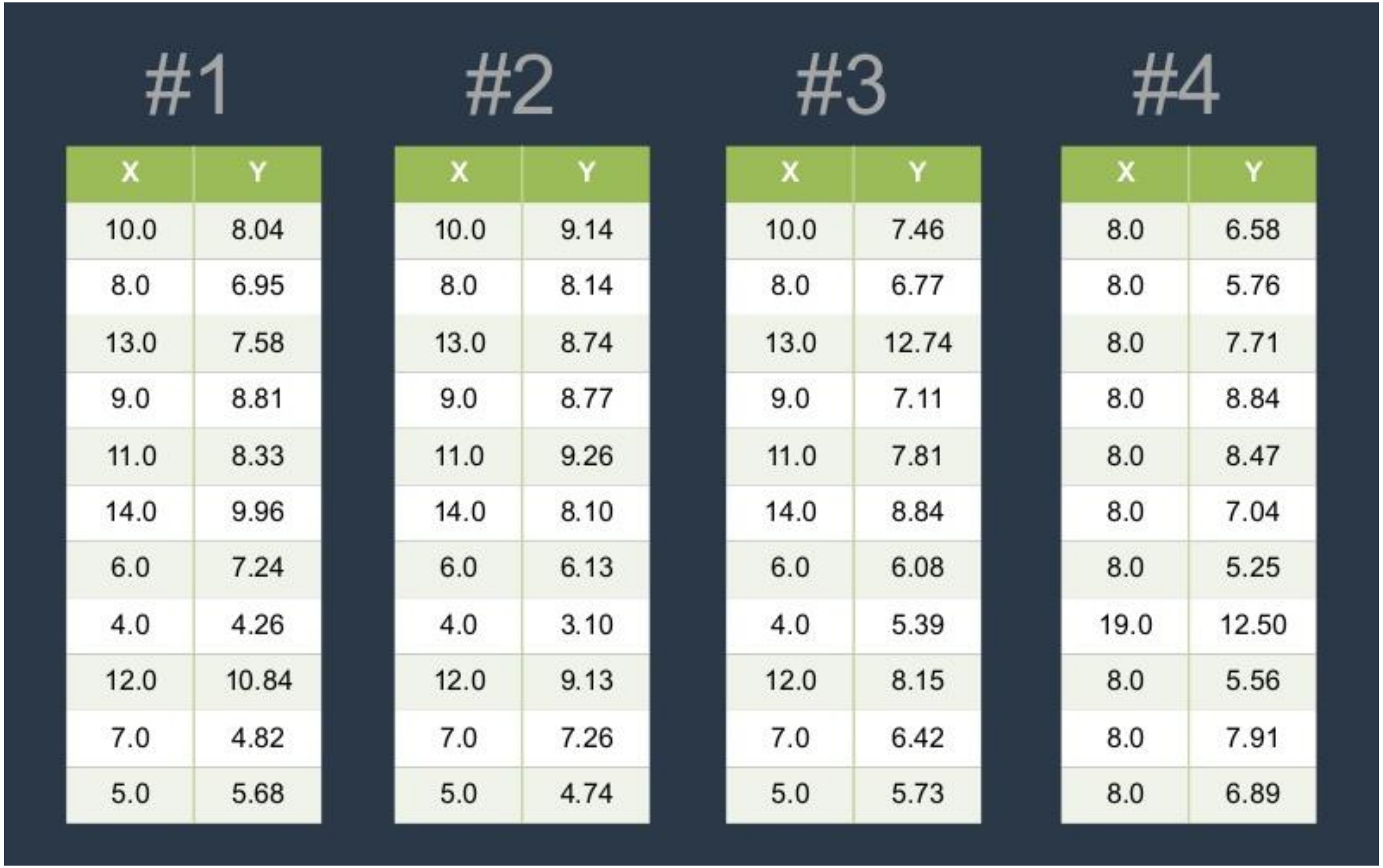
Example :





Source : medium.com

Anscombe's Quartet

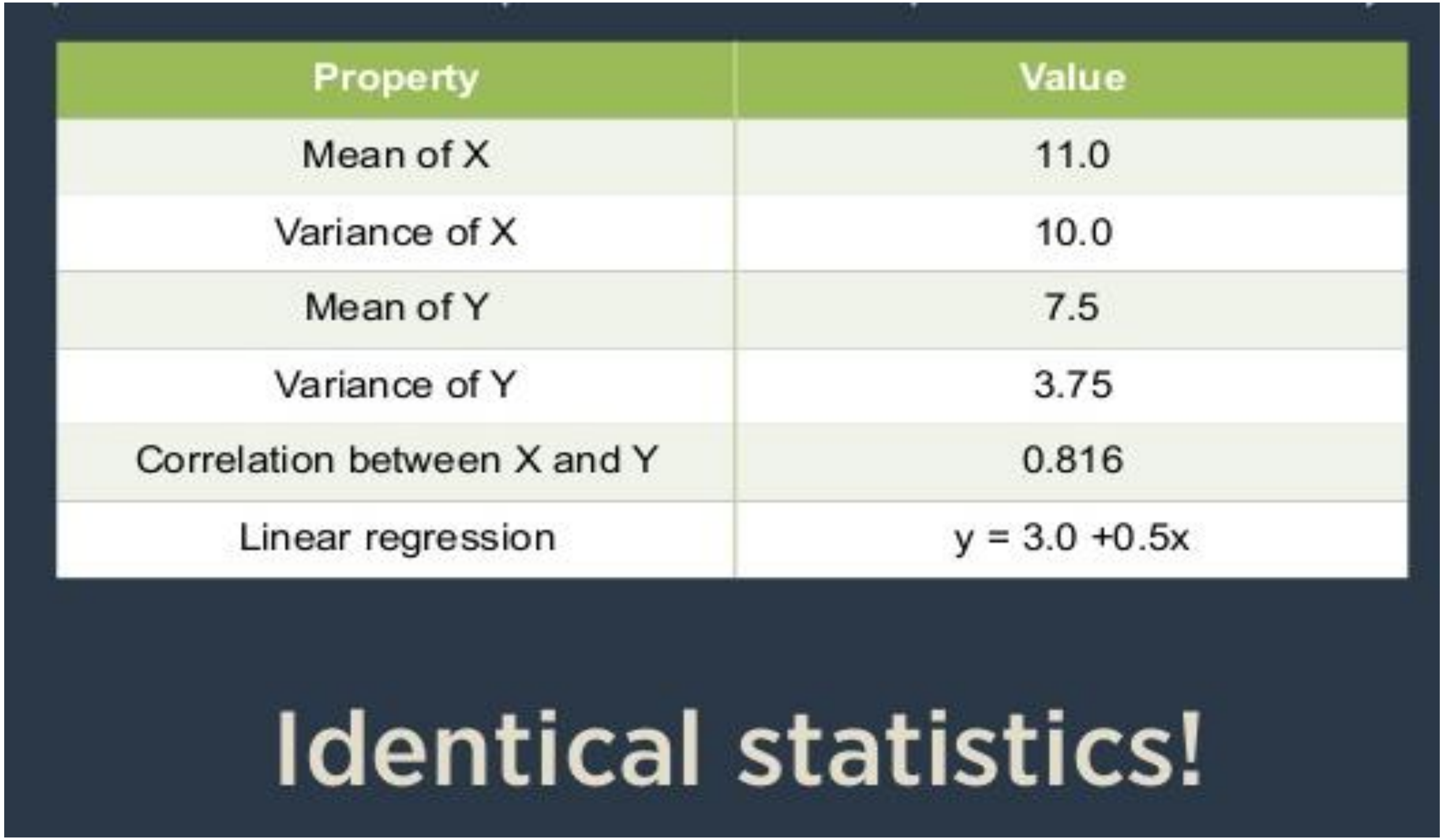


#1		#2		#3		#4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Source : slideshare.net

- Consists of four cleverly constructed dataset having the same Pearson's correlation coefficient.

Anscombe’s Quartet Summary Statistics



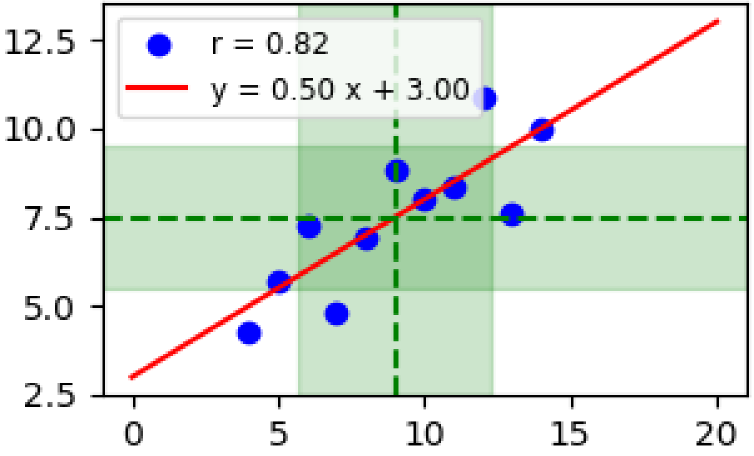
Property	Value
Mean of X	11.0
Variance of X	10.0
Mean of Y	7.5
Variance of Y	3.75
Correlation between X and Y	0.816
Linear regression	y = 3.0 +0.5x

Identical statistics!

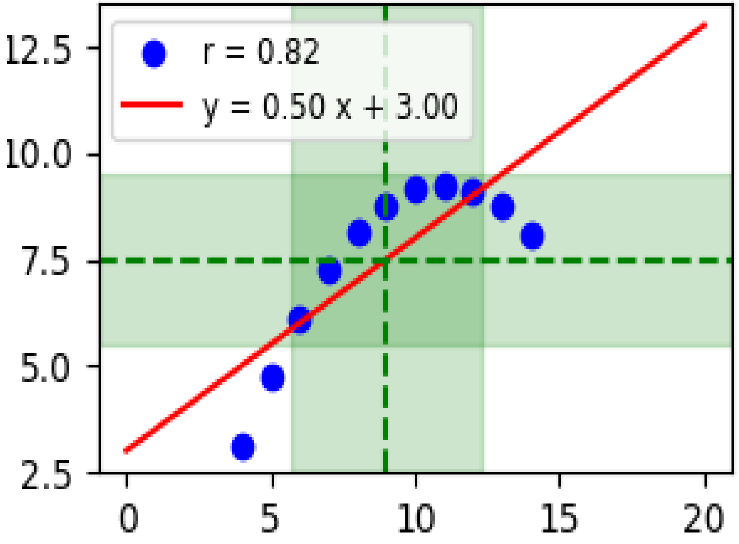
Source : slideshare.net

- The summary statistics (ie., Mean, Variance, Pearson’s correlation coefficient and the Linear Regression) for all the four datasets is the same.
- But the datasets are significantly different and visually distinct. This can be observed in their respective scatter plots.

Visual representation of the Anscombe's Quartet



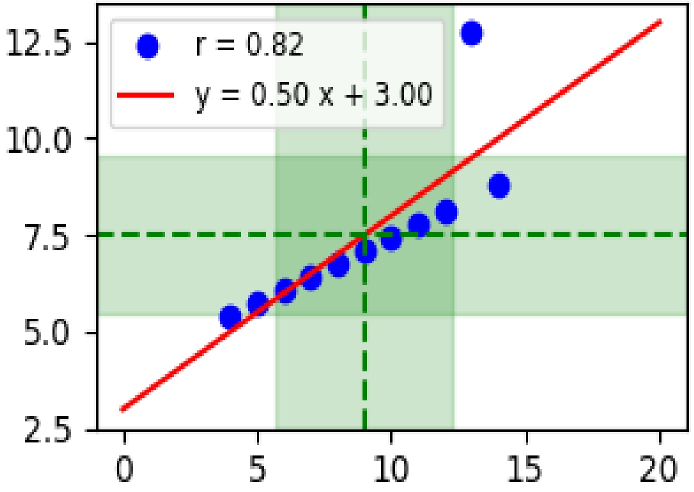
- The first scatter plot appears to be a simple linear relationship, corresponding to two variables



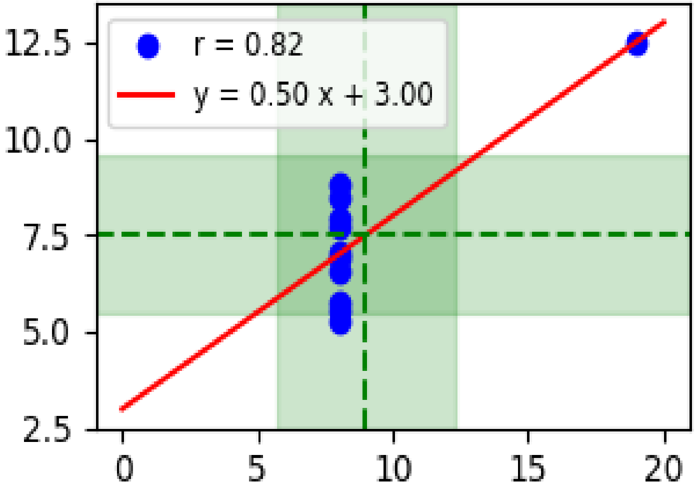
- The second graph is not distributed normally.
- Though a relationship between the two variables is obvious, it is not linear.
- The Pearson correlation coefficient is not relevant.

Source : informatique-python.readthedocs.io

Visual representation of the Anscombe's Quartet



- ❖ In the third graph the linear relationship is perfect.
- ❖ But one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.



- ❖ The fourth example shows another example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

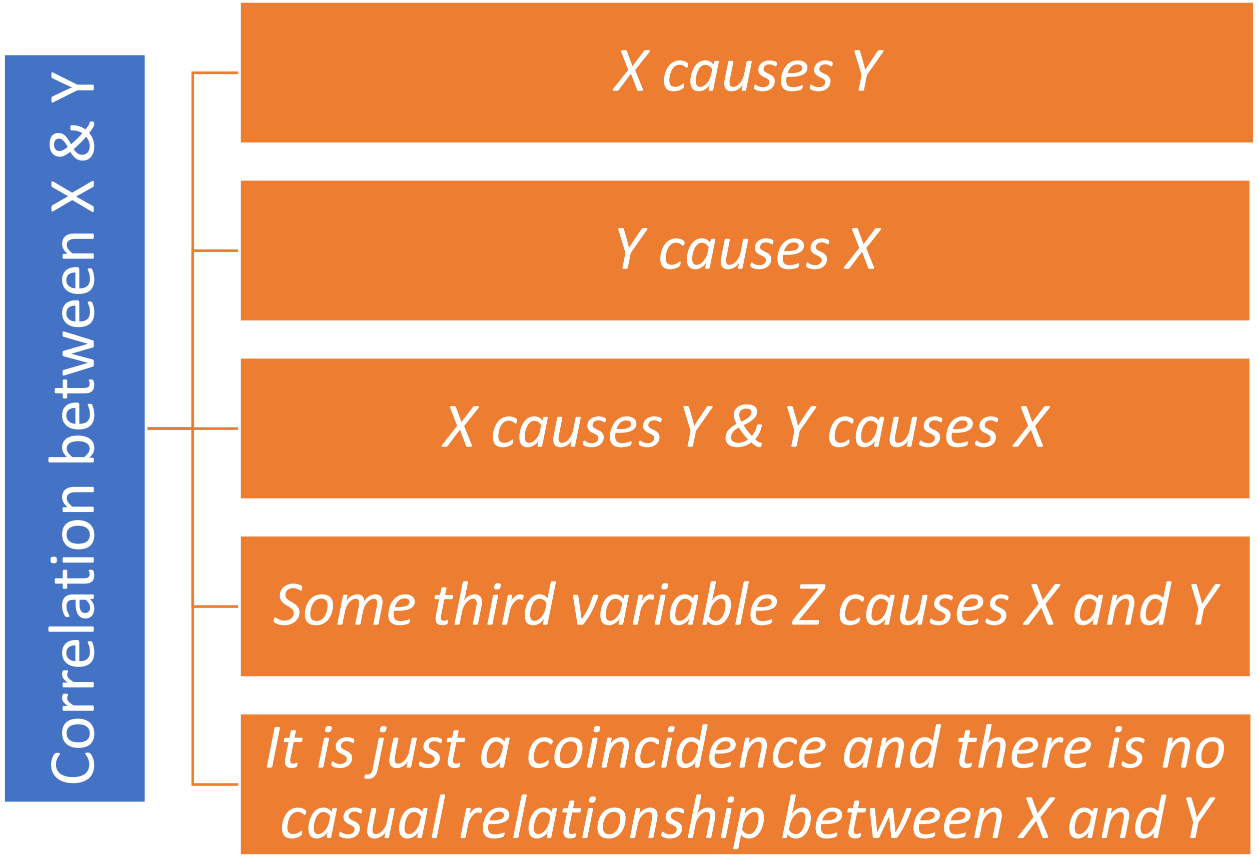
Source : informatique-python.readthedocs.io

Conclusions from the Anscombe’s Quartet

- The presence of outliers in a dataset has an impact on the evaluated value of the correlation coefficient.
- The basic statistic properties are inadequate for describing realistic datasets.
- It is important to look at a set of datavisually before starting to analyze

Source : iconfinder.com

Remark :



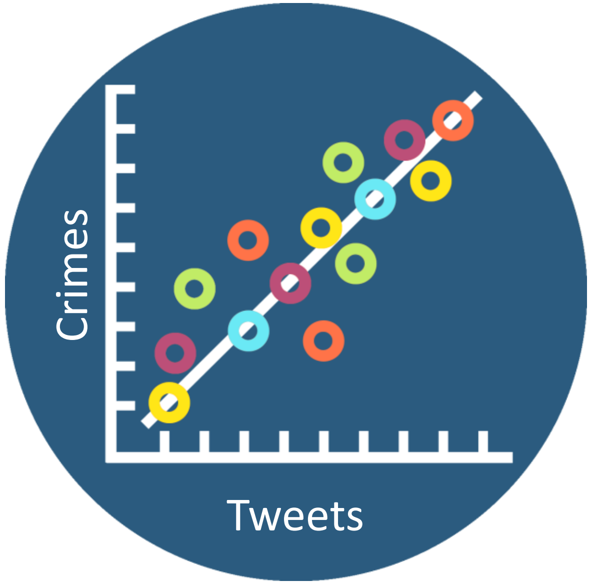
Correlation between X & Y

- *X causes Y*
- *Y causes X*
- *X causes Y & Y causes X*
- *Some third variable Z causes X and Y*
- *It is just a coincidence and there is no casual relationship between X and Y*

Correlation is not causation!!

Example : Relationship between drug related tweets and crime rates (Strong Positive Correlation).

- A strong Positive relationship between tweets and crime has been found.
- But there is no evidence to suggest that *tweets are causing more crime* and *tweets about crime do not necessarily reflect the crime rate*.

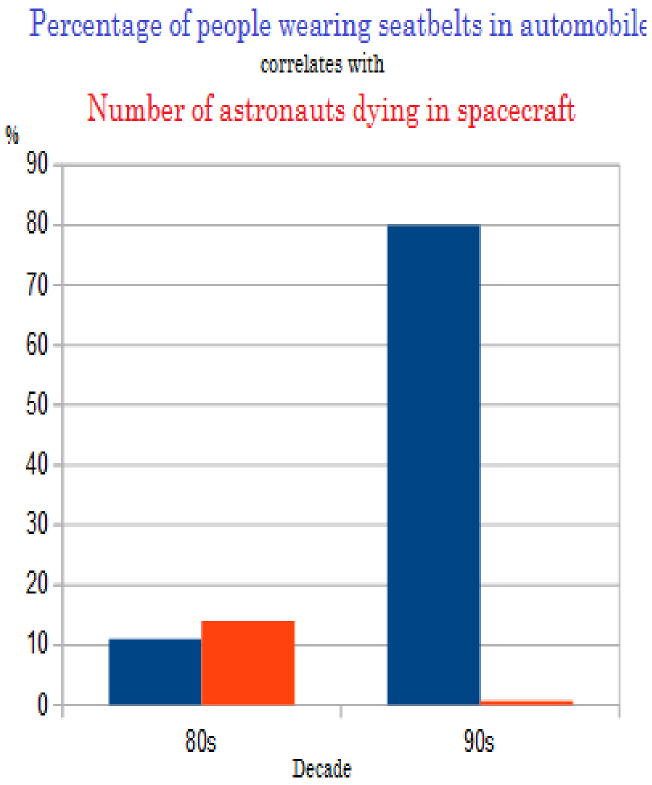


Reference : The Relationship Between Social Media Data and Crime Rates in the United States Yan Wang1 , Wenchao Yu1 , Sam Liu2 , and Sean D. Y Social Media + Society January-March 2019: 1–9 ©

Source :learningspaces.dundee.ac.uk

Correlation is not causation!!

Example 2. : Relationship between wearing seat belt and astraunat deaths (Strong Positive Correlation).



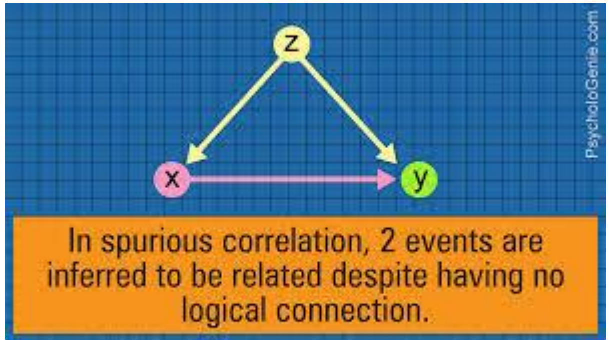
Use your seatbelt and save an astronaut life!

- ❖ The graph shows that an increase in wearing car seat belt results in a lower number of astronaut deaths.
- ❖ Obviously there isn't a real correlation here: putting your seat belt on in a car has nothing to do with the odds of an accident in space.

Sources :statisticshowto.com, pinterest.com, freeart.com

Confounding Variable

- Confounding Variable is a variable that influences both the independent variable as well as the dependent variable causing a spurious correlation.
- This may interfere in your analysis and ruin your experiment by giving useless results.
- Confounding variables can cause two major problems:
  - Increase variance
  - Introduce bias.
- A confounding variable are like extra independent variables that are having a hidden effect on your dependent variables.
- A confounding variable can be what the actual cause of a correlation is, hence any studies must take these into account and find ways of dealing with them.



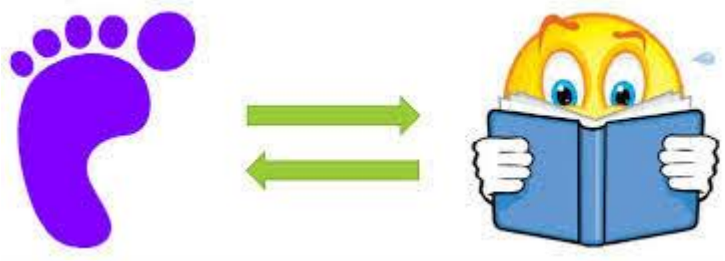
In spurious correlation, 2 events are inferred to be related despite having no logical connection.

Source : psychologenie.com

Confounding Variable!!

Example 1. : Relationship between reading ability and shoe size. (Strong Positive Correlation).

- You collect data on reading ability and shoe size
- You find that bigger the shoe size the better is the reading ability.
- Does that mean bigger shoe size leads to better reading abilities?
- Should children be hence fed growth hormones so that the reading abilities improve?
- Should children start focusing on their reading abilities to increase their shoe size?



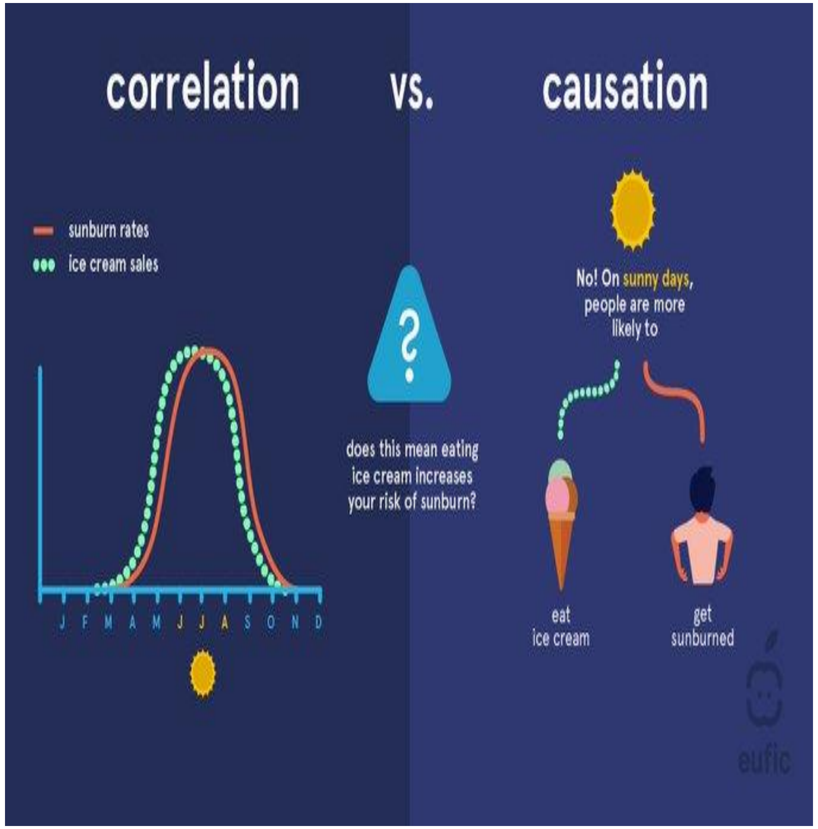
Confounding Variable : There is a third variable—a confounding variable—which causes the increase in both reading ability and shoe size.

Age : As the child's age increase, the foot size increases and also the reading ability increases since the child goes to higher classes.

Source : outsmartstatistics.weebly.com

Confounding Variable!!

Example 2. : Relationship between sun burns and ice – cream consumption. (Strong Positive Correlation).


correlation
vs.
causation
sunburn rates
ice cream sales
J F M A M J J A S O N D
?
does this mean eating ice cream increases your risk of sunburn?
No! On sunny days, people are more likely to
eat ice cream
get sunburned
eufic

Source : twitter.com

- ❖ You collect data on sunburns and ice cream consumption.
- ❖ You find that higher ice cream consumption is associated with a higher probability of sunburn.
- ❖ Does that mean ice cream consumption causes sunburn?

Possibility #1: Sun burns cause purchase of ice cream.

Possibility #2: Eating ice cream causes sun burns.

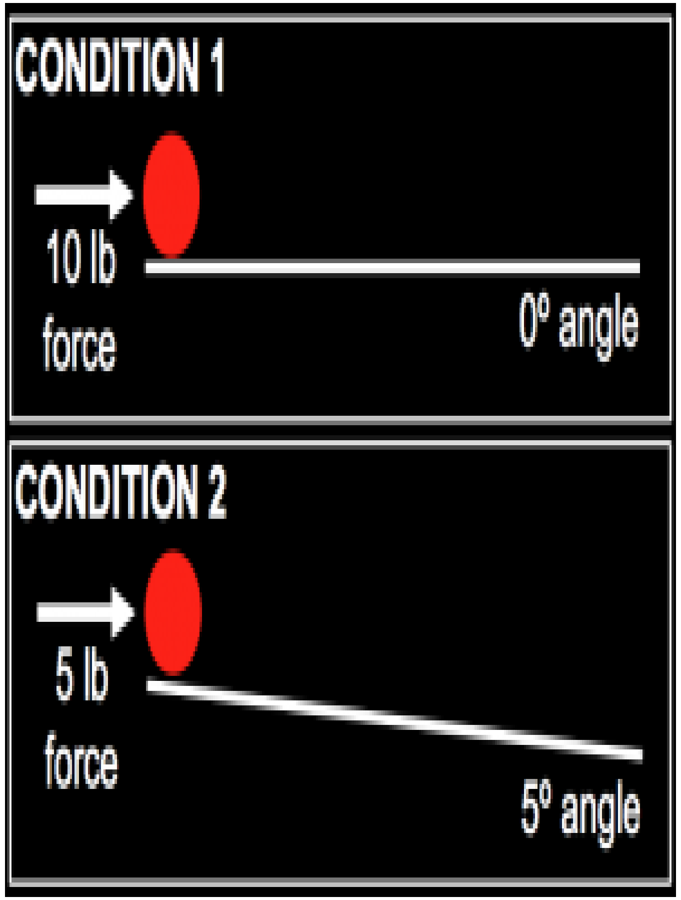
Possibility #3: There is a third variable—a confounding variable—which causes the increase in both ice cream sales and sun burn.

Confounding Variable : Hot temperatures

Confounding Variable!!

Example 3. : Relationship between the force you apply to a ball and the distance the ball travels.

- ❖ Naturally, you predict that the more force you apply, the further the ball will travel.
- ❖ After you run your experiment, you observe that the ball travels further in Condition 2 than it does in Condition 1.
- ❖ In other words, you find that the less force you apply, the further the ball travels.
- ❖ Confounding variable : the angle of the slope.



Source : psychologyinaction.org

Example Problem 1.

- An environmental scientist is studying the rate of absorption of a certain chemical into skin.
- She places differing volumes of the chemical on different pieces of skin and allows the skin to remain in contact with the chemical for varying lengths of time.
- She then measures the volume of chemical absorbed into each piece of skin.
- She obtains the results shown in the following table.

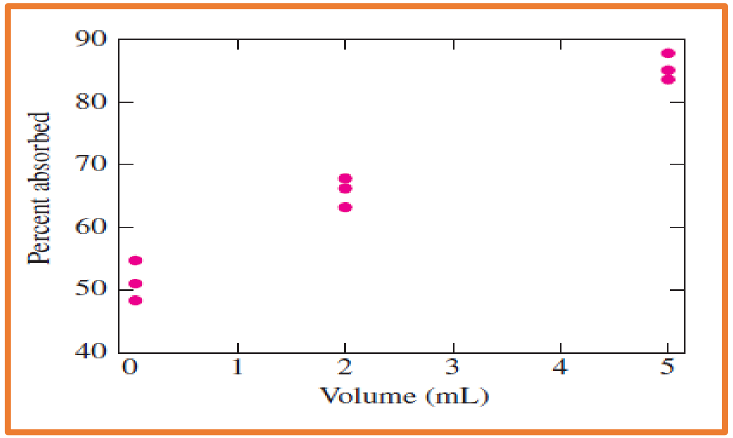
Volume (ml)	Time (h)	Percent Absorbed
0.05	2	48.3
0.05	2	51.0
0.05	2	54.7
2.00	10	63.2
2.00	10	67.8
2.00	10	66.2
5.00	24	83.6
5.00	24	85.1
5.00	24	87.8

Source : Internet

Example Problem

Correlation between Volume & Percent Absorbed

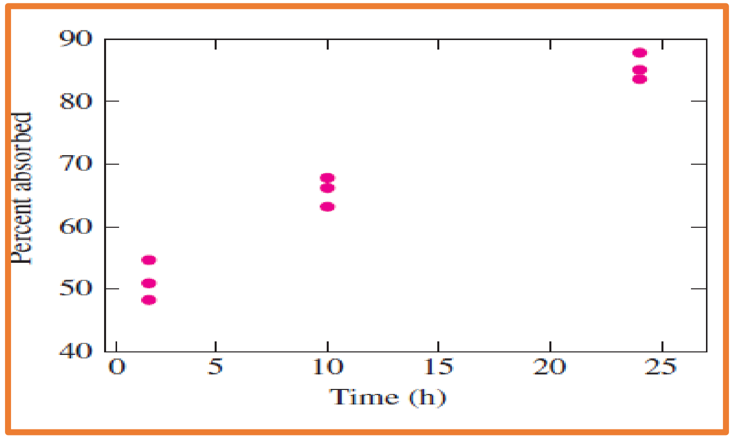
- Scatter Plot :



- Correlation , $r = 0.988$
- Positive Correlation
- Increasing the volume causes the percentage absorbed to increase.

Correlation between Time & Percent Absorbed

- Scatter Plot :



- Correlation , $r = 0.987$
- Positive Correlation
- Increasing the time that the skin is in contact with chemical causes the percentage absorbed to increase.

Are these conclusions Justified???

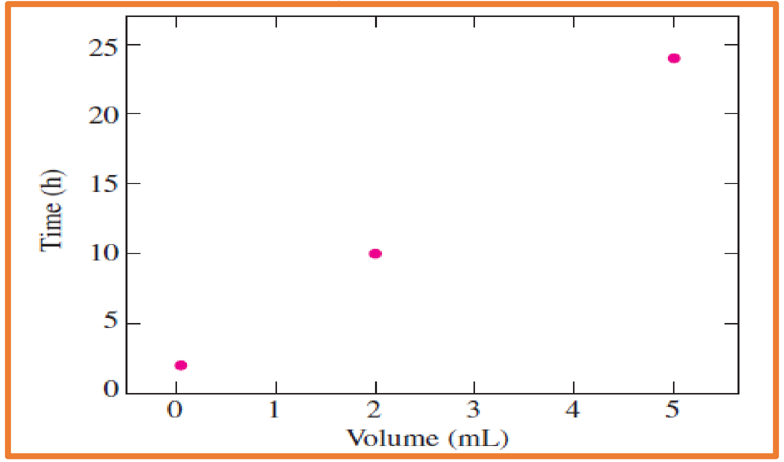
Sources : slideteam.net,
Statistics for Engineers and Scientists, William Navidi

Example Problem

No! The conclusions are not justified!

Suggested Solution :

- The correlation between time & volume has to be explored.
- The Scatter plot :



- The correlation, $r = 0.999$
- Conclusion : These 2 variables are completely confounded.

Observations :

- Since, time and volume are highly correlated with each other if either time or volume affects the percentage absorbed, both will appear to do so.
- Hence it is impossible to determine whether it is the time or the volume that is having an effect.
- This relationship between time and volume resulted from the design of the experiment and should have been avoided

Source : Statistics for Engineers and Scientists, William Navidi

Example Problem 2.

- The Scientist in Example Problem 1 has repeated the experiment, this time with a new design.
- The results are presented in the table.
- She then calculates the correlation between volume and absorption and obtains $r = 0.121$
- She concludes that increasing the volume of the chemical has little or no affect on absorption.
- She then calculates the correlation between time and absorption and obtains $r = 0.952$
- She concludes that increasing the time that the skin is in contact with the chemical will cause the percentage absorbed to increase.

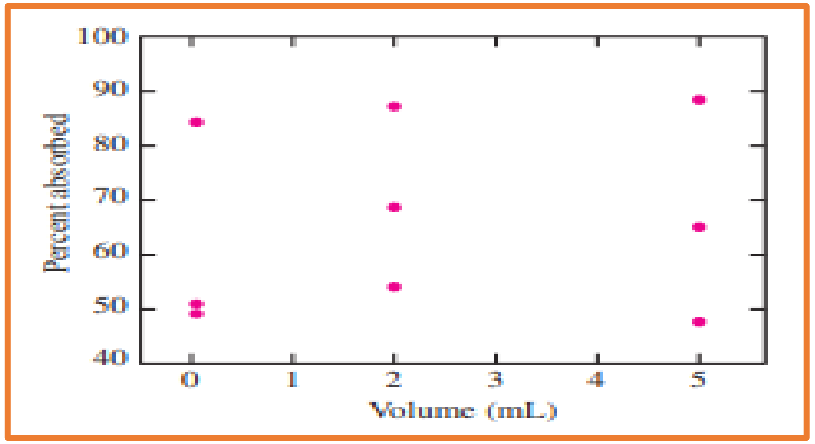
Volume (ml)	Time (h)	Percent Absorbed
0.05	2	49.2
0.05	10	51.0
0.05	24	84.3
2.00	2	54.1
2.00	10	68.7
2.00	24	87.2
5.00	2	47.7
5.00	10	65.1
5.00	24	88.4

Source : Internet

Example Problem

Correlation between Volume & Percent Absorbed

❖ Scatter Plot :



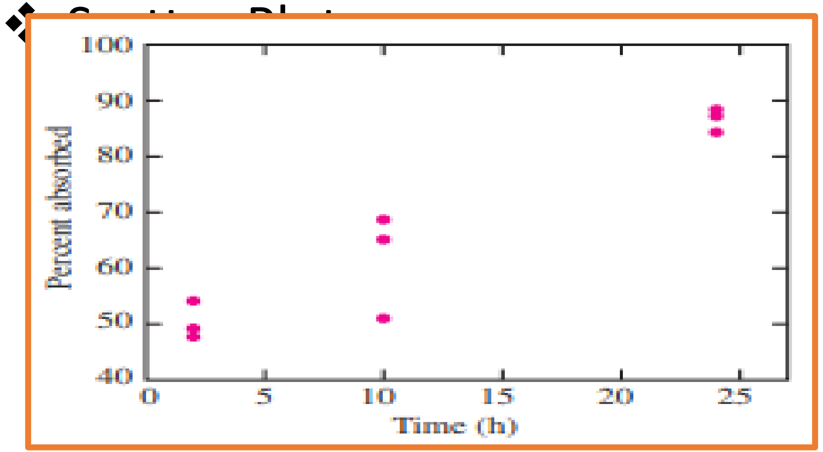
❖ Correlation, $r = 0.121$

❖ Weak Positive Correlation

❖ Hence increase of volume has little or no effect on the percentage absorbed.

Correlation between Time & Percent Absorbed

❖ Scatter Plot :



❖ Correlation , $r = 0.952$

❖ Strong Positive Correlation

❖ Increasing the time that the skin is in contact with the chemical will cause the percentage absorbed to increase.

Are these conclusions Justified???

Sources : slideteam.net,
Statistics for Engineers and Scientists, William Navidi

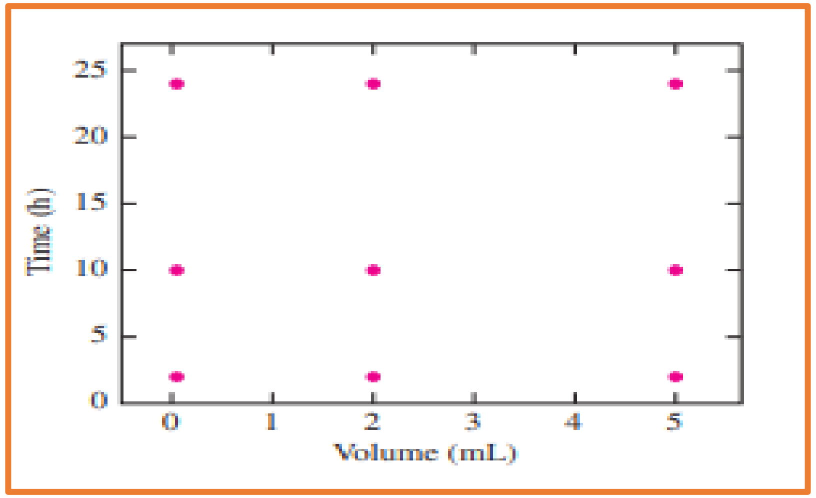
Example Problem

These conclusions are justified as

- The experiment has been designed in such a way that there is no correlation between time and volume.
- This can be observed in the scatter plot.
- Hence there is no possibility of confounding due to time or volume.
- From the correlation coefficient and scatter plot for the relationship between time and percent absorbed it is clear that in this case time, but not volume has an affect on the percentage absorbed.

Correlation between Volume & Percent Absorbed

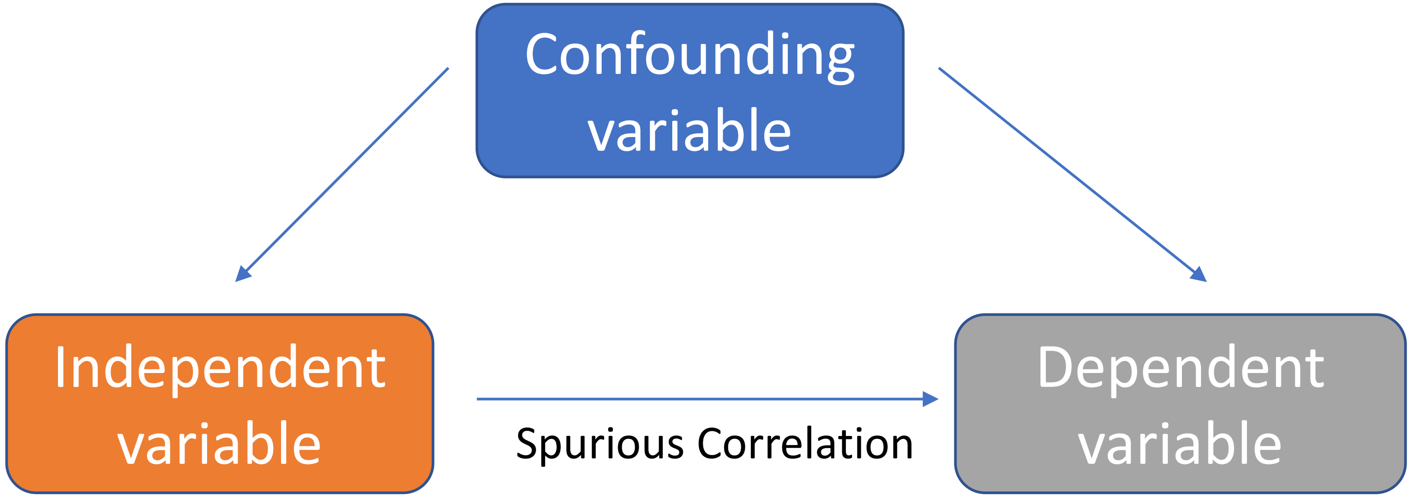
- Scatter Plot :



- *Time & Volume are not correlated in this case*

Source :
Statistics for Engineers and Scientists, William Navidi

Controlled Experiments reduce the risk of Confounding



- ❖One of the ways by which confounding can be avoided in controlled experiments by choosing values for certain factors in such a way that there exists no correlation between those factors.
- ❖ For instance in Example Problem 1. & 2. the environmental scientist reduced confounding by assigning values to volume and time such that they were uncorrelated.
- ❖But this is not possible in all cases.

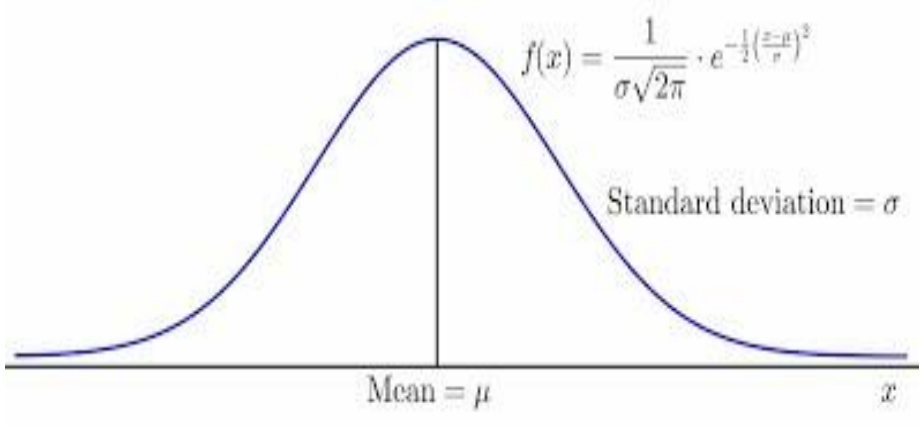
Controlled Experiments reduce the risk of Confounding

- ❖The values of factors cannot be chosen by the observer in case of observational studies/ experiments.
- ❖For instance, in studies involving public health issues like impact of environmental pollutants on human health, the observer cannot assign values to any of the factors.
- ❖Hence it becomes difficult to avoid confounding.
- ❖Example : People who live in areas with higher level pollutants may tend to have lower socio-economic status, which may affect their health.
- ❖ In observational studies to avoid or reduce confounding the study must be repeated a number of times under a variety of conditions before drawing reliable conclusions !!!

Then how can confounding be avoided in such cases ???

Bivariate Normal Distribution

❖The “regular” normal distribution has one random variable.

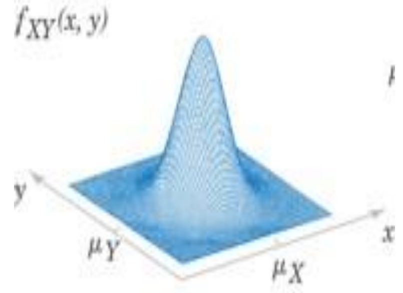


- ❖ **A bivariate normal distribution is made up of two independent random variables.**
- ❖ The two variables in a bivariate normal are both normally distributed, and they have a normal distribution when both are added together.
- ❖ Two random variables $X\ and\ Y$ are said to be bivariate normal, or jointly normal, if $aX + bY$ has a normal distribution for all $a, b\ \in\ R$

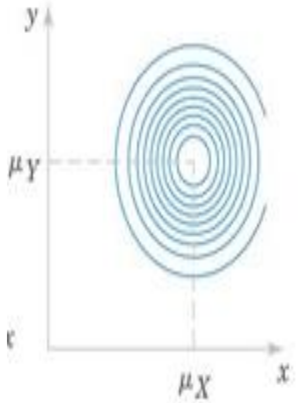
Source : Internet

Bivariate Normal Distribution

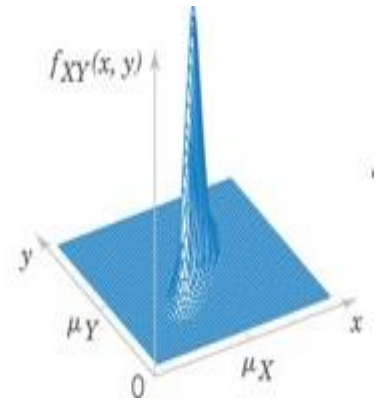
Visually, the bivariate normal distribution is a three-dimensional bell curve.



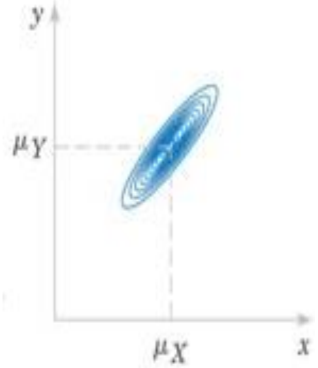
This is the Bivariate Normal distribution of 2 independent random variables $X \,\&\, Y \ (\rho = 0)$



By cutting Bivariate Normal distribution horizontally we obtain the contour circles with center at the point of means (μ_x, μ_y)



This is the Bivariate Normal distribution of 2 dependent random variables $X \,\&\, Y \ (\rho > 0)$

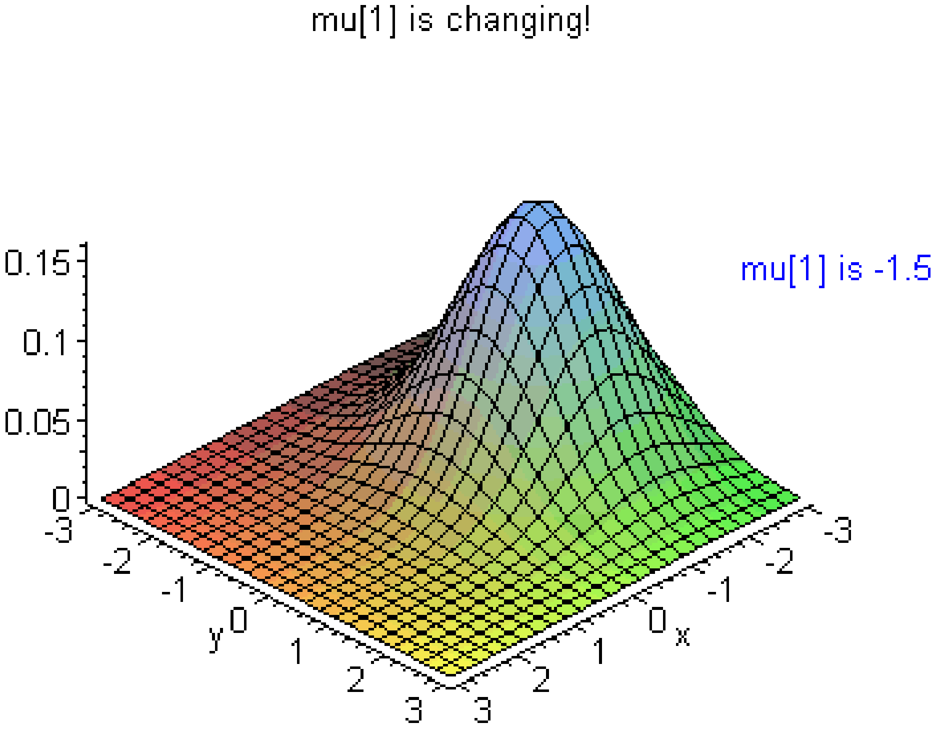


By cutting Bivariate Normal distribution horizontally we obtain the contour ellipses with center at the point of means (μ_x, μ_y)

Source : Internet

Bivariate Normal Distribution

Visually, the bivariate normal distribution is a three-dimensional bell curve.



This animated graph helps you to visualize the Bivariate normal distribution when the centre changes. Here one of the means is varied resulting in a change in the center.

Source : statisticshowto.com

Inference on the population Correlation

If the random variables X and Y have a certain joint distribution called a Bivariate normal distribution, then the sample correlation coefficient can be used

1. To construct a confidence interval on the population correlation coefficient ρ.
2. To test the null hypothesis on the population correlation ρ.

Inference on the Population Correlation

Let,

- X & Y : Random variables with the bivariate normal distribution
- $(x_1, y_1), \ldots, (x_n, y_n)$: Random sample from the joint distribution of X & Y.
- r : Sample correlation of the n points.
- ρ : Population correlation between X & Y.

The Fisher transformation

$$W = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \text{........................(1)}$$

$$W \sim N(\mu_w, {\sigma^2}_w)$$

where the mean,

$$\mu_w = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) \text{........................(2)}$$

and variance,

$$\sigma^2{}_w = \frac{1}{n-3} \text{..................................(3)}$$

Confidence Interval

Computing confidence interval for ρ

- Obtain the Confidence Interval for μ_W as, $W \pm z\sigma_W$
- Use upper and lower confidence bounds of μ_W to construct the confidence interval for ρ using $\rho = \frac{e^{2\mu_W - 1}}{e^{2\mu_W + 1}}$ which is obtained from $\mu_w = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)$

Example Problem 1.

- In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 10 subjects.
- Times were measured in ms.
- The results are presented in the following table.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

- Find a 95% confidence interval for the correlation between the two reaction times.

Example Problem 1.

Solution : We need to obtain the following :

1. Compute the Sample correlation $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$
2. Obtain W using the Fischer's formula $W = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right)$.

   Then find $W \sim N(\mu_w, \sigma^2{}_w)$ where $\mu_w = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)$ and $\sigma^2{}_w = \frac{1}{n-3}$
3. Compute the confidence interval for μ_w using W
4. Finally, convert the confidence interval back to 'ρ' using the relation $\rho = \frac{e^{2\mu_w - 1}}{e^{2\mu_w + 1}}$ which is obtained from $\mu_w = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)$

Solution :

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
161	159	-36.2	-33.7	1310.44	1135.69	1219.94
203	206	5.8	13.3	33.64	176.89	77.14
235	241	37.8	48.3	1428.84	2332.89	1825.74
176	163	-21.2	-29.7	449.44	882.09	629.64
201	197	3.8	4.3	14.44	18.49	16.34
188	193	-9.2	0.3	84.64	0.09	-2.76
228	209	30.8	16.3	948.64	265.69	502.04
211	189	13.8	-3.7	190.44	13.69	-51.06
191	169	-6.2	-23.7	38.44	561.69	146.94
178	201	-19.2	8.3	368.64	68.89	-159.36
197.2	192.7	0.00	0.00	4867.6	5456.1	4204.6

1. $$r = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$$
$$= \frac{4204.6}{\sqrt{4867.6\ \sqrt{5456.1}}}$$
$$= 0\ .8159$$

2. $$W = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2} ln\left(\frac{1+0.8159}{1-0.8159}\right)$$
$$= 1.1444$$

- W is normally distributed with standard deviation$\sigma_W = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{10-3}} = 0.3780$

Solution :

3. A 95% confidence interval for μ_w is given by $W - z\sigma_w < \mu_w < W + z\sigma_w$

That is, $1.1444 - 1.96(0.3780) < \mu_w \; 1.1444 + 1.96(0.3780)$

{We know that for 95% confidence interval for μ_w would be -1.96 & $1.96.$}

$$\Rightarrow \; 0.4036 < \mu_w < 1.8852$$

4. Now to obtain the 95% confidence interval for ρ, we consider

$$\rho = \frac{e^{2\mu_w - 1}}{e^{2\mu_w + 1}}$$

$$\frac{e^{2(0.4036) - 1}}{e^{2(0.4036) + 1}} < \frac{e^{2\mu_w - 1}}{e^{2\mu_w + 1}} < \frac{e^{2(1.8852)\; - 1}}{e^{2(1.8852)\; + 1}}$$

$$\Rightarrow \; 0.383 < \rho < 0.955$$

Hypothesis testing

- If $\rho = \rho_0, \rho \leq \rho_0$ and $\rho \geq \rho_0$ where $\rho_0 \neq 0$ is a constant, for performing the Hypothesis testing, we can use the same transformation W.

- If $\rho = 0, \rho \leq 0$ and $\rho \geq 0,$ for performing the Hypothesis testing, we use the test statistic $U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ which follows a student t distribution with $n - 2$ degrees of freedom.

Example Problem 2.

- In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 10 subjects.
- Times were measured in ms.
- The results are presented in the following table.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

i. Find the P – value for testing $H_0\colon\ \rho \leq 0.3$ versus $H_1\colon\ \rho > 0.3$

ii. Test the hypothesis $H_0\colon\ \rho \leq 0$ versus $H_1\colon\ \rho > 0$

Solution :

i. Under H_0, we take $\rho = 0.3$

- We know that for W,

  the mean $\mu_w = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) = \frac{1}{2} ln\left(\frac{1+0.3}{1-0.3}\right) = 0.3095$

  and the standard deviation $\sigma = \sqrt{\frac{1}{10-3}} = 0.3780$
- Therefore, $W \sim N(\mu_w, \sigma^2{}_w) \Rightarrow W \sim N(0.3095, 0.3780^2)$
- From the Example Problem 1. the observed value of $W = 1.1444$
- Therefore, $z - score$ is, $z = \frac{1.1444-0.3095}{0.3780} = 2.21$
- From the Normal table, we have the P – value as 0.0136.
- Since the P – value is less than the significance level, we reject the null hypothesis.
- Hence we conclude that $\rho > 0.3$

Solution :

ii. Under H_0, we take $\rho = 0$

- So we need to take the test statistic U given by $U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
- From the Example Problem 1. the value of $r = 0.8159$
- Therefore, $U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.8159\sqrt{10-2}}{\sqrt{1-0.8159^2}} = 3.991$
- Since U follows the Student's t distribution with $n-2$ degrees of freedom, using the Student's t distribution table for 8 degrees of freedom we find that the P – value is between 0.001 and 0.8159.
- Since the P – value is less than the significance level α, we reject the null hypothesis.
- Hence we conclude that $\rho > 0$

PES UNIVERSITY ONLINE

THANK YOU

Dr. Karthiyayini

Department of Science & Humanities

Karthiyayini.roy@pes.edu

+91 80 6618 6651