

Statistics for Research Methods

Workshop 2

Correlation and Regression

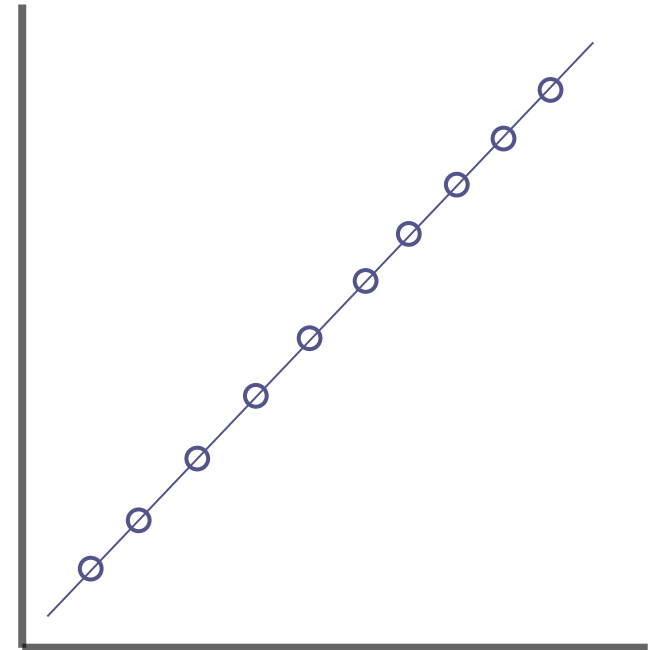
Introduction

- Today we'll look at
 - Correlation Coefficient
 - Pearson Correlation Coefficient
 - Interpreting Correlation Coefficients
 - Spearman Rank Correlation
 - Linear Regression
 - Standard Error
- Until now we've been looking at data that can be described using a single variable
- Today we will start to look at the relationship between two variables
 - Is module attendance related to the final module result?

Correlation Coefficients

- We can graphically relate the two variables *Attendance* and *Result* using a scatterplot
- Suppose we plot attendance against result
 - If the points all fall along a straight line then we say that the two variables are *perfectly correlated*
 - If we know the value of one variable then we can calculate the value of the other variable
- Graphs like this are useful, but a numerical measure is better

% Result



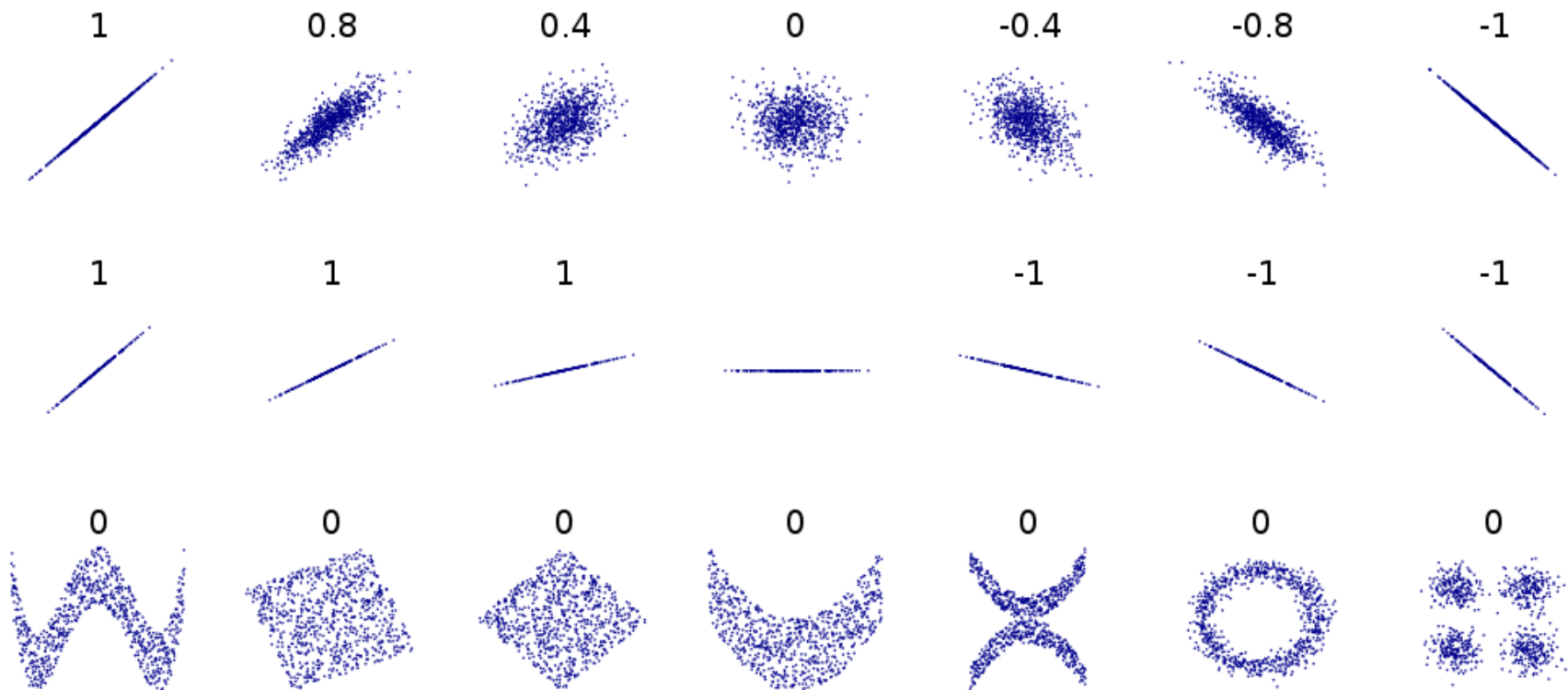
% Attendance

Correlation Coefficients

- Most methods of correlation were designed to measure the degree by which two variables are linearly related
 - High-dimensional correlation techniques exist but are beyond the scope of this module
- The correlation coefficient r_{xy} shows the degree of association between x and y and can take values from -1.0 to +1.0
- The sign of r_{xy} shows how the variables are related
 - Positive values indicate that low values in x are related to low values in y , and high values in x are related to high values in y
 - Negative values indicate that low values in x are related to high values in y , and high values in x are related to low values in y
- When $r_{xy} = 0$ no relationship exists between x and y
- The strength of the relationship is proportional to the absolute value of r_{xy}
 - The higher the value, the more strongly related the two variables

Correlation Coefficients

- Correlation DOES NOT IMPLY causation
- It can also provide some weird results...
- A value of $r_{xy} = \pm 0.5$ is the minimum needed to say that a relationship exists between x and y



Pearson Correlation Coefficient

- The most commonly-used measure of correlation is the *Pearson product-moment correlation coefficient*, r_{xy}
- This describes the relationship between two variables x and y

$$r_{xy} = \frac{Cov_{xy}}{SD_x SD_y}$$

$$\text{Where } Cov_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1} \quad \text{or} \quad Cov_{xy} = \frac{\sum xy}{N - 1}$$

$x = (X - \bar{X})$, the deviation scores for X

$y = (Y - \bar{Y})$, the deviation scores for Y

$\sum xy$ = the sum of the products of the paired deviation scores

N = the number of pairs of scores

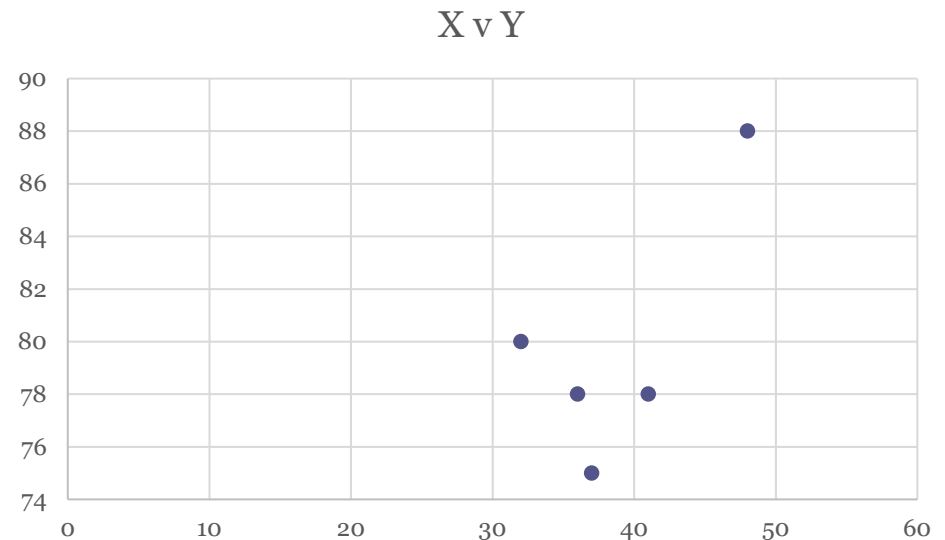
SD_x = the standard deviation of X

SD_y = the standard deviation of Y

Example

Person	X	Y
Alice	37	75
Bob	41	78
Charlotte	48	88
David	32	80
Eve	36	78

- First, draw a scatterplot to get a feel for the data



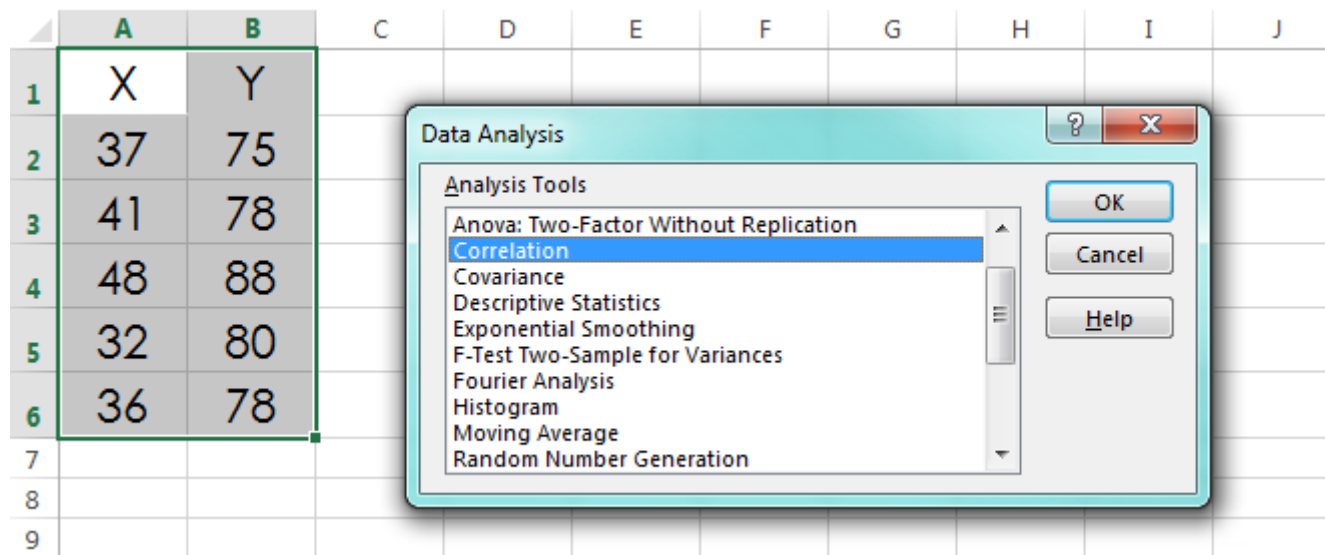
Example

Persons	X	Y
Alice	37	75
Bob	41	78
Charlotte	48	88
David	32	80
Eve	36	78

- Now calculate the correlation coefficient
 - This can be done by hand, but in general nowadays spreadsheets or statistical packages are used instead
- Correlation calculations are available via the Excel Analysis ToolPak

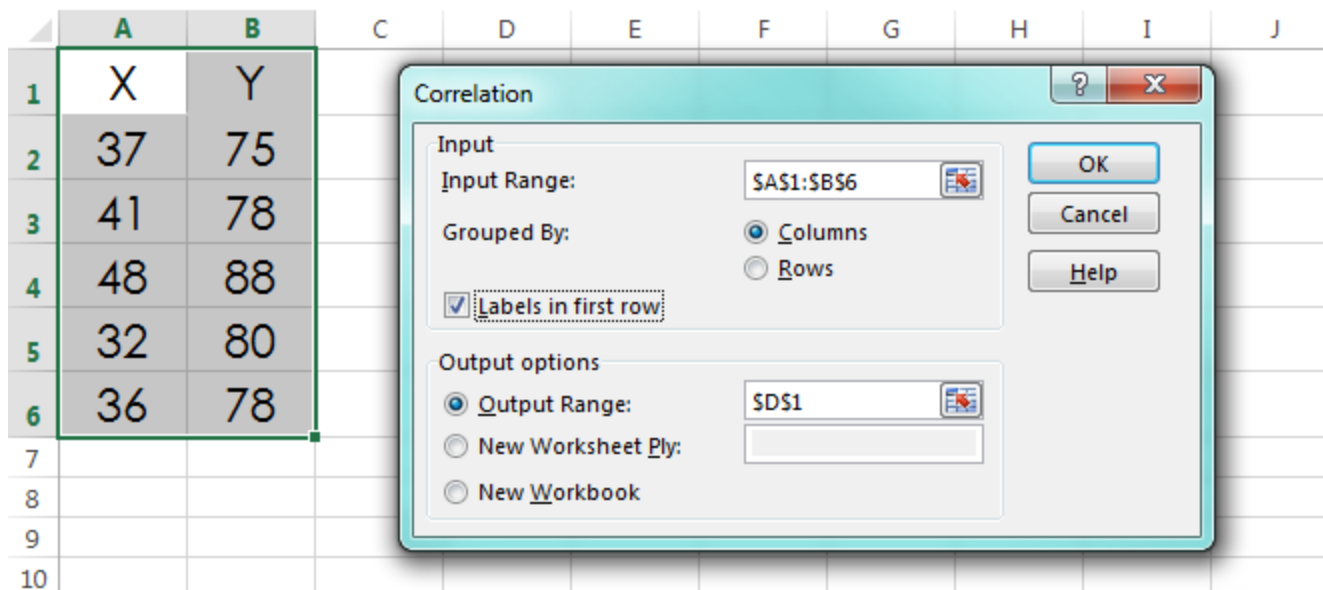
Correlations in Excel

- Once the Analysis ToolPak has been added it will appear at the end of the Data tab of the ribbon
- Select 'Correlation' from the data analysis tool



Correlations in Excel

- Input the data range for analysis, input a location for the data to be output, and tick the labels box if appropriate



Correlations in excel

	A	B	C	D	E	F	G
1	X	Y			X	Y	
2	37	75		X	1		
3	41	78		Y	0.702981	1	
4	48	88					
5	32	80					
6	36	78					

- On completion, Excel will calculate the Pearson's correlation for the data

Interpreting Correlation Coefficients

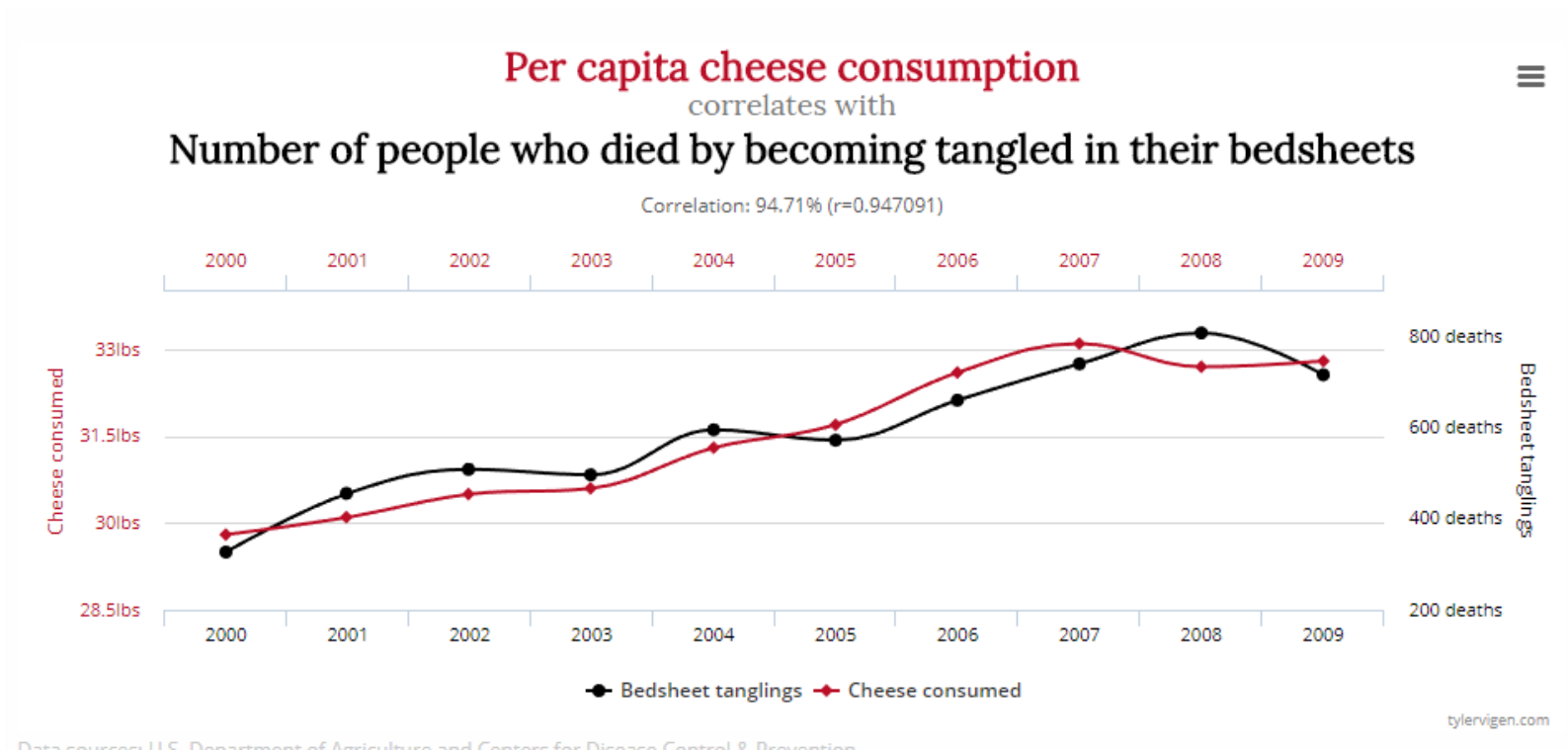
- Correlation does not imply causation
 - It's not easy to prove causality so you must have reasons for doing so
- If causation exists then you can do a lot with it
 - Promote the relationship to provide benefits
 - Prevent the cause to avoid harmful consequences
 - Prepare for unavoidable harmful consequences
 - Prosecute the perpetrator of the cause
 - Pontificate about what might happen in the future if the relationship holds over time
 - Probe for knowledge simply out of curiosity

Hill's Criteria for Causation

- Medical statistician Austin Bradford Hill defined a set of minimal conditions necessary to accept that a causal relationship between two variables is likely:
 1. Strength: The correlation coefficient should be large and statistically significant
 2. Consistency: The relationship can be replicated
 3. Specificity: If there is no other likely explanation
 4. Temporality: The effect always occurs after the cause
 5. Gradient: If a greater exposure leads to a greater effect
 6. Plausibility: A plausible mechanism exists between cause and effect
 7. Coherence: The relationship is compatible with current theory
 8. Experiment: If the relationship can be verified experimentally
 9. Analogy: There are proven relationships between similar causes and effects

Exercise

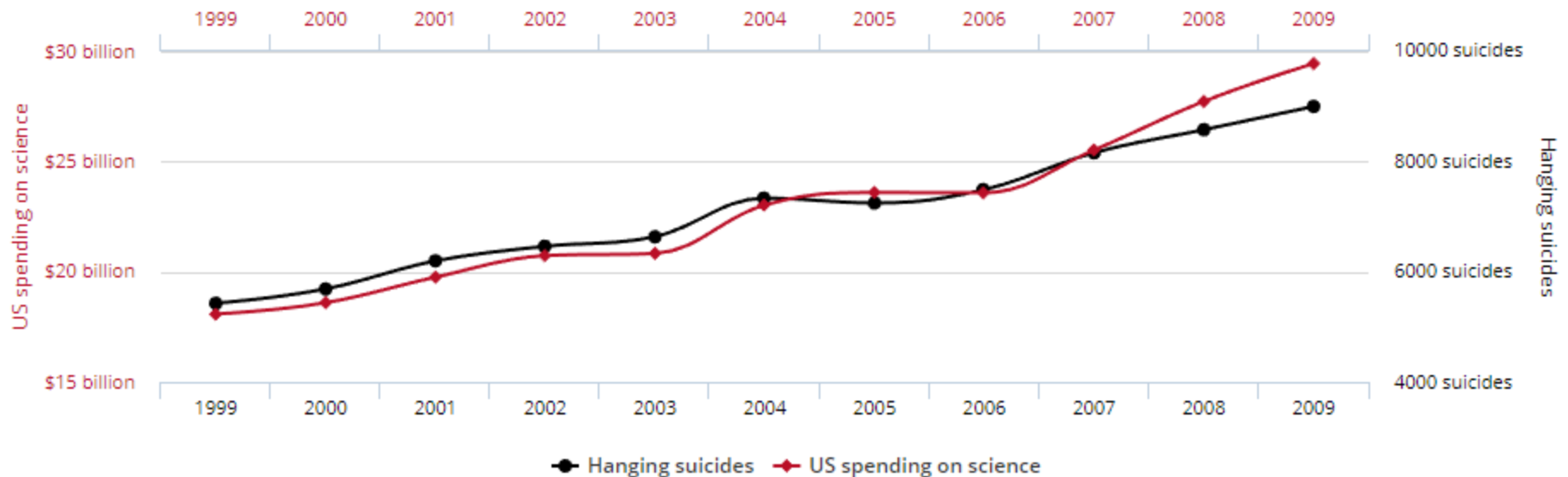
- Given the relationships below, discuss how far they meet Hill's criteria:
 - A correlation between IQ and the number of exams passed
 - A correlation between a person carrying a particular gene and the likelihood of them getting a cancer mediated by that gene



Spurious Correlations

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

Correlation: 99.79% ($r=0.99789126$)



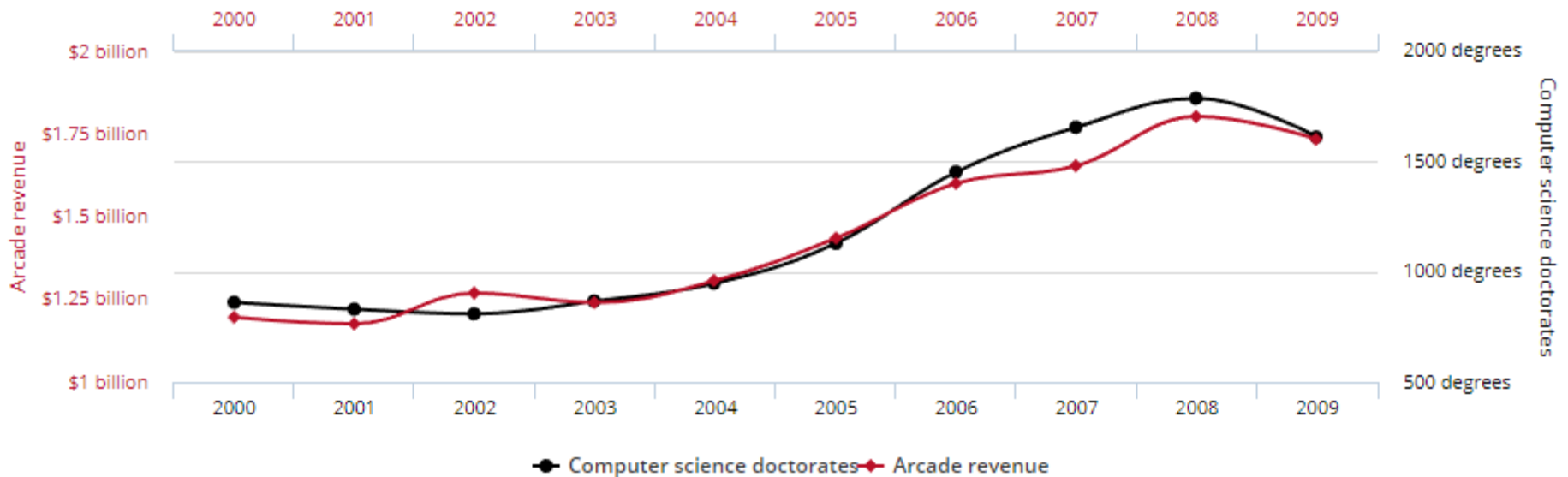
tylervigen.com

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Spurious Correlations

Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US

Correlation: 98.51% ($r=0.985065$)



tylervigen.com

Data sources: U.S. Census Bureau and National Science Foundation

Exercise

- X and Y are scores obtained by a group of participants in two separate tests
- What is the correlation between them, and what does this mean?

X	Y
140	42
130	44
120	35
119	30
115	23
114	27
114	25
113	20
112	16
111	12

Spearman Rank Correlation Coefficient

- The Spearman rank correlation (r_s) aka Spearman's rho can be used when we are looking at data ranks rather than at actual data
- The calculation is the same as for Pearson's coefficient, but the data has to be ranked first, then the correlation calculated on the ranks

Example

Persons	X	X_r	Y	Y_r
Alice	30	1	100	6
Bob	40	2.5	80	3
Charlotte	40	2.5	70	2
David	50	4	90	5
Eve	60	5	88	4
Frank	70	6	60	1

- ▶ In this case we first generate the ranks from lowest to highest
- ▶ If two or more scores have the same values then give them the mean of the ranks
 - ▶ Two marks at equal second would take ranks 2 and 3, so would get a rank score of 2.5
 - ▶ Three marks at equal fifth would take ranks 5, 6 and 7, so would get rank 6

Exercise

- X and Y are scores obtained by a group of participants in two separate tests
- Calculate the ranked correlation
- If this differs from the Pearson's correlation then why might this happen?

X	Y
140	42
130	44
120	35
119	30
115	23
114	27
114	25
113	20
112	16
111	12

Making Predictions

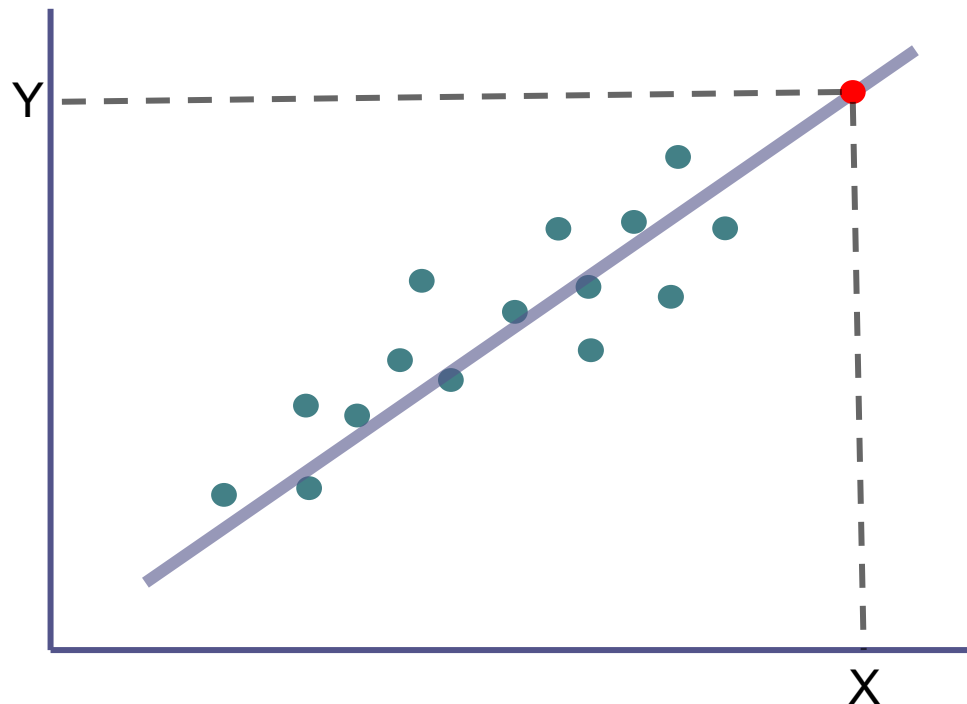
- If we know nothing about a relationship between X and Y we can only predict that a value will be close to the mean of that data range
- Suppose we also know that there is a strong positive correlation between scores in test X and test Y
 - People who score high in X tend to score high in Y, while people who score low in X tend to score low in Y
- This allows us to improve our prediction
 - If Bob scored 70 in test X we might predict that he will score 70 in test Y
 - If Bob scored 30 in test X we might predict that he will score 30 in test Y

Making Predictions

- When two variables are correlated it is possible to use a *regression equation* to predict with greater than chance accuracy the score on one variable given the score on the other variable
- The higher the correlation, the better the prediction will be
- If the correlation is low then the prediction will be closer to chance

Graphical Interpretation

- Note that although we've made a prediction, we're not likely to be exactly right
 - We just want to be closer than chance
- We are effectively fitting a line through our correlated data and making a prediction along that line



Regression Equations

- The regression equation is the equation of a line fitted through all of the data points to minimise the distance between the data points and the line
 - Generally a least-squares approach
- If we have such an equation then we can use it to predict the value of X if we know Y , and vice versa
- Excel lets us do this by adding a *trendline* to a scatterplot and displaying its equation
 - This equation can be used to calculate X or Y

Regression Equations

- $Y = \beta_0 + \beta_1 X + \epsilon$
- β_0 - Intercept (Value of Y When $X = 0$)
- β_1 - Slope coefficient (represents the change in Y for a one-unit change in X.)
- ϵ - Error term

=====

Result interpretation

R (correlation coefficient)

Strength and direction of the linear relationship between independent and dependent variables

R-squared (coefficient of determination)

How much of the variation in the dependent variable (Y) is explained by the independent variable(s) (X).

Example: If R-squared is 0.80, it means 80% of the variation in Y is explained by X.

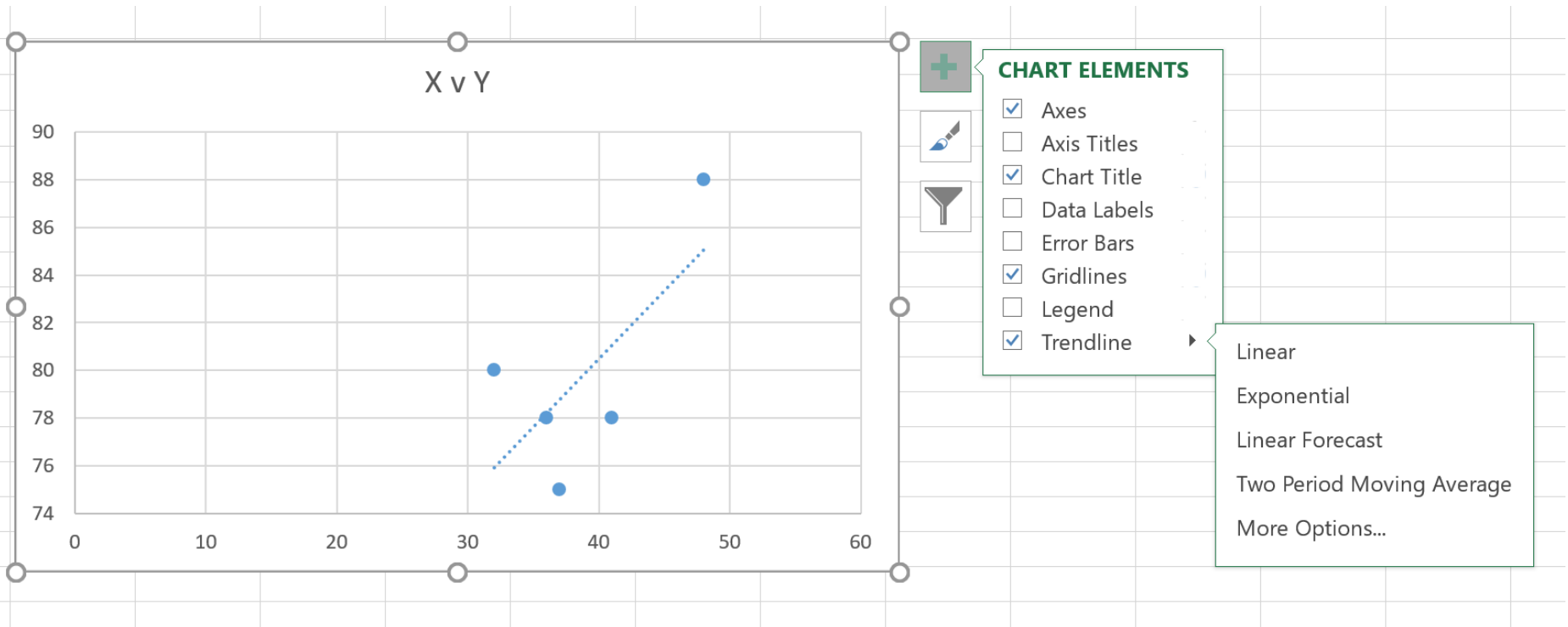
Standard error

The standard error tells us how much the predicted values of Y differ from the actual values on average.

Example

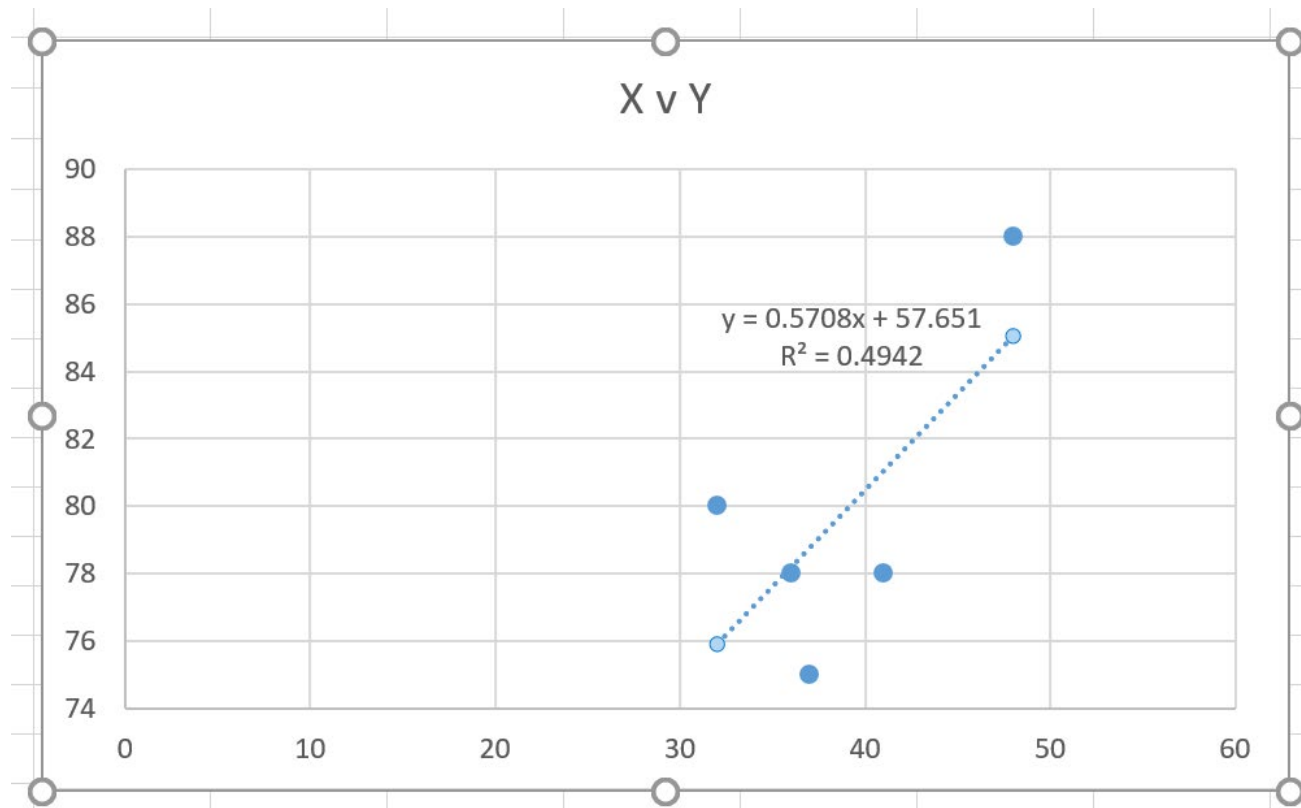
If the standard error is small, it means our predictions are usually close to the actual values.

Regression Equations



- Once the trendline has been added, the equation itself can also be shown via the More Options... dialog

Regression Equations



- This can then be extracted to use in Excel to perform any appropriate calculations
- R^2 represents the *coefficient of determination*
 - Goodness of fit parameter
 - 95% is good

Standard Error

- R^2 provides a general measurement of good fit, but a more useful value is the *Standard Error*
- How likely is it that we are going to be correct in our prediction?
- Suppose X and Y in the last question had been test scores
 - If Bob got 80 on test Y will he definitely get 40? in test X? No
- Because the correlation is based on numbers that are spread about a 'best-fit' line, it is likely that all of the people who were predicted to get 80 got a range of different scores
 - They would form a normal distribution around 80 (mean of 80)
- If they form a normal distribution then we can calculate the standard deviation of their distance from the mean
 - This is called the *standard error*
- About 68% of all people will fall between the standard error and the predicted value (mean $\pm$ 1 standard deviation)

Standard Error

- Use the Analysis Toolpak regression tool to calculate this
- Bob's score could then be predicted as 40 ± 5.0

Correlation 
Coefficient 
Of Determination
Standard 
Error

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.702981
R Square	0.494182
Adjusted R Square	0.325576
Standard Error	4.975076
Observations	5

Exercise

- X and Y are scores obtained by a group of participants in two separate tests
- Calculate the linear regression function and the standard error for this relation
- If X is 123, what is Y?
- If Y is 21, what is X?

X	Y
140	42
130	44
120	35
119	30
115	23
114	27
114	25
113	20
112	16
111	12

Summary

- Topics covered
 - Correlation Coefficient
 - Pearson Correlation Coefficient
 - Interpreting Correlation Coefficients
 - Spearman Rank Correlation
 - Regression Equations
- Next session we will cover
 - The Sampling Experiment
 - Sample Values (statistics) and Population Values (Parameters)
 - The Null Hypothesis
 - Type I and Type II Errors
 - Statistical Significance and Type I Errors
- Inferential statistics allow us to make predictions based upon the sampling of a large population

Coming soon

- **Next session:** Inferential Statistics