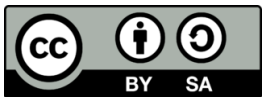
THE UNIVERSITY *of* EDINBURGH

Data Science in Medicine

Lecture 4: Hypothesis Testing – Part 1

Dr Areti Manataki

Usher Institute
The University of Edinburgh



Usher Institute
Population Health Sciences & Informatics

In the previous lecture

- Visualising data
  - Qualitative data: bar charts, pie charts
  - Quantitative data: histograms, box plots
  - Bivariate: scatter plots, line graphs
- Introduction to R
  - R essentials
  - Calculating summary statistics
  - Drawing simple graphs

In this lecture

- Correlation
  - Multidimensional data
  - Correlation and causation
  - Scatterplots revisited
  - Correlation coefficient
- Hypothesis testing
  - Main idea
  - Reflecting on significance
  - Correlation coefficient as a statistical test

Multidimensional data

- Most datasets contain several pieces of information about each of many individuals.

Weekly hours of study	Weekly hours of exercise	Daily hours of sleep	Grades
8	2	7	75
7.4	1.2	6.2	70
8.3	4.5	5.4	86
6.2	4	7.2	48
6.3	1	7.3	54
7	2.3	8.4	62
8.8	5	6	87
6.1	5.2	8.9	49

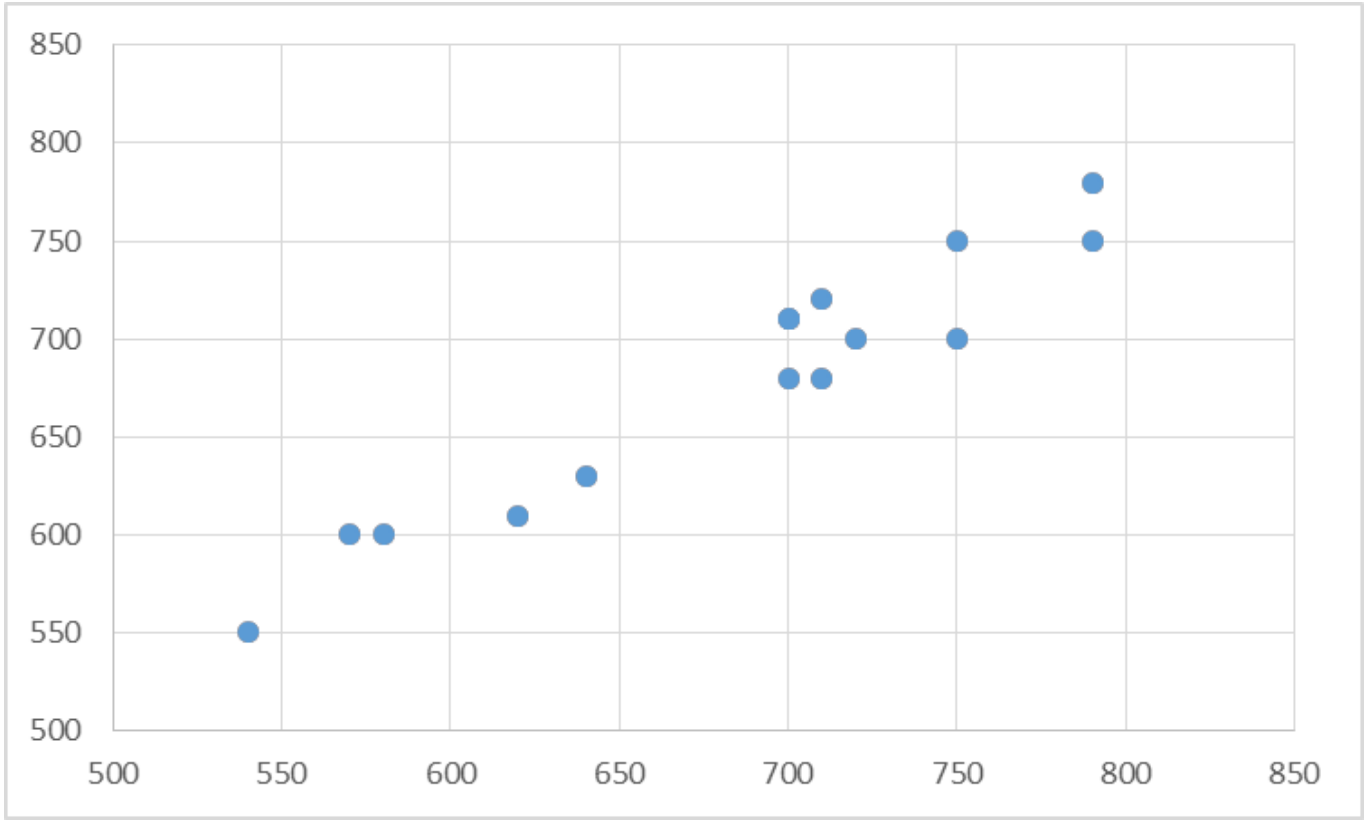
Correlation

- We can ask whether there is any observed relationship between the values of two different variables
  - If there is no relationship, then the variables are said to be independent.
  - If there is a relationship, then the variables are said to be correlated.

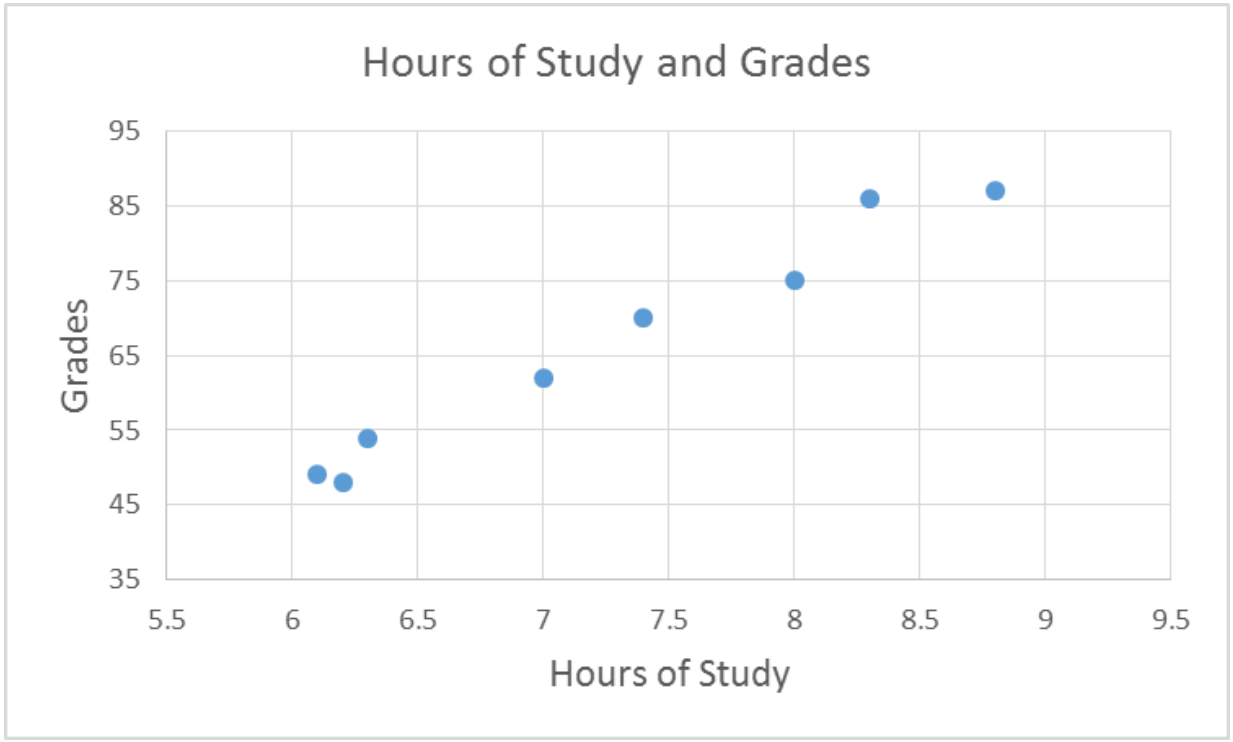
Correlation vs causation

- Two variables are causally connected if variation in the first causes variation in the second. If this is so, then they will also be correlated.
- However, the reverse is not true:
  Correlation Does Not Imply Causation!
- If we do observe a correlation between variables X and Y, it may due to any of several things:
  - Variation in X causes variation in Y
  - Variation in Y causes variation in X
  - Variation in X and Y is caused by some third factor Z
  - Chance

Visualising correlation: scatterplots

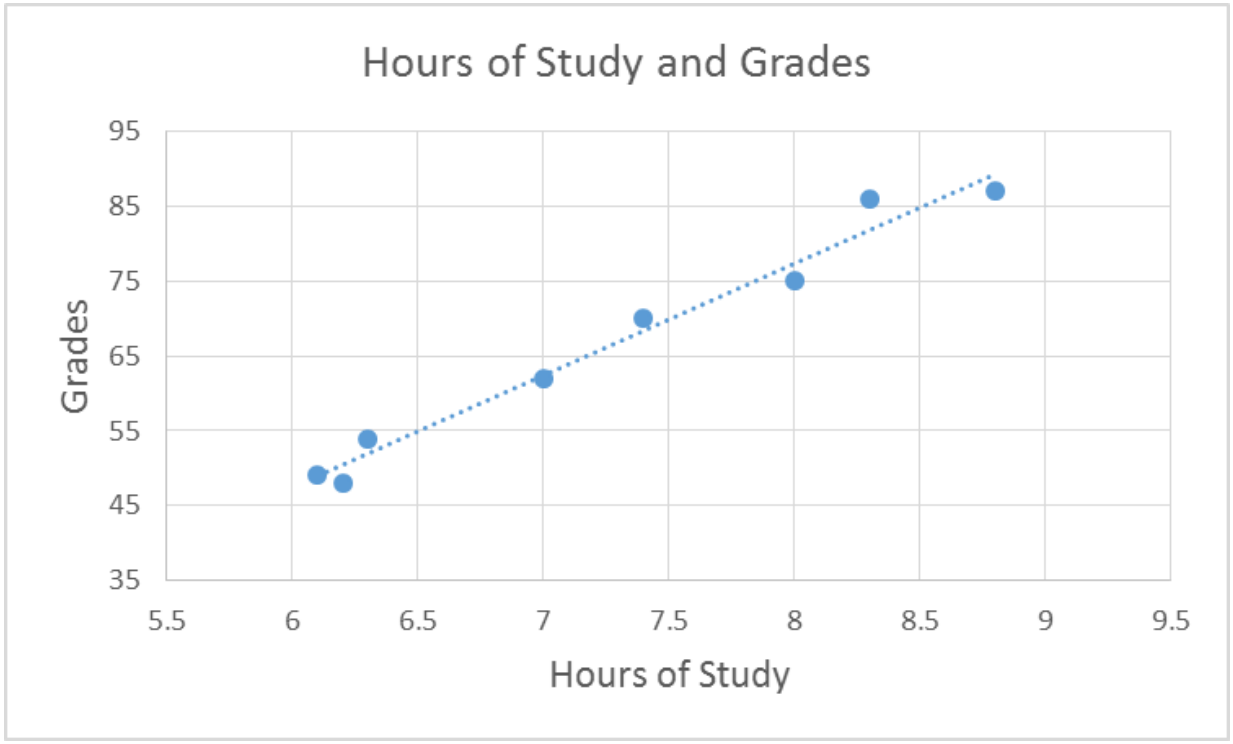


Visualising correlation: scatterplots



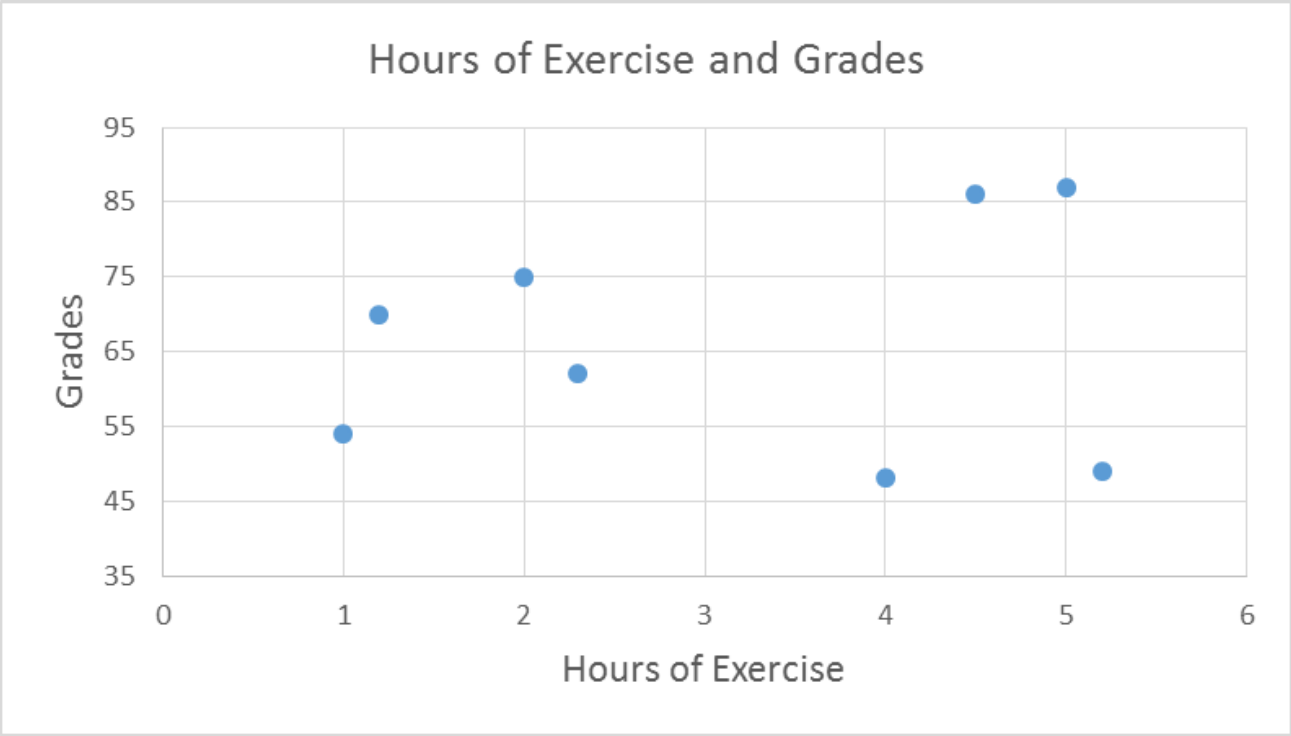
This graph indicates a strong, positive, linear relationship between the two variables.

Visualising correlation: scatterplots



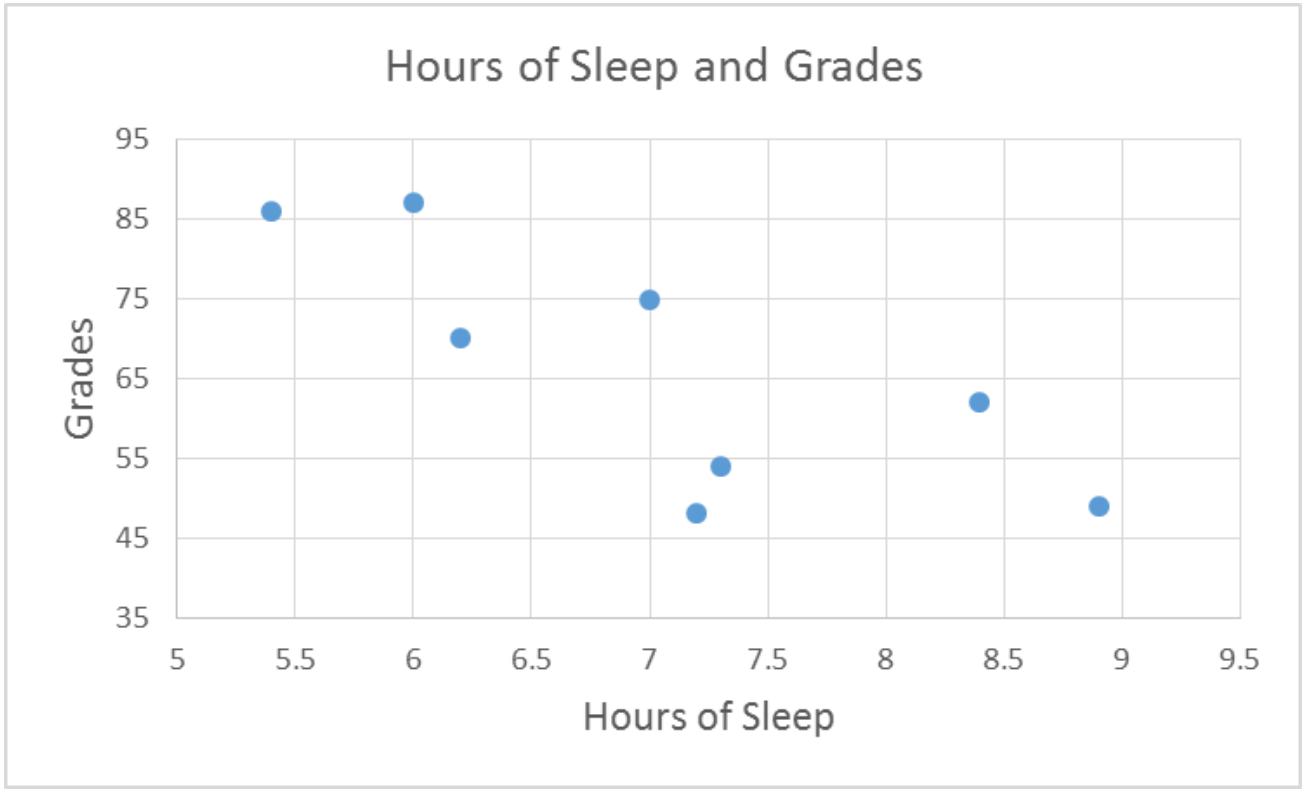
This graph (with the line of best fit) indicates a strong, positive, linear relationship between the two variables.

Visualising correlation: scatterplots



There is no evident correlation between the two variables, neither positive nor negative.

Visualising correlation: scatterplots



This graph indicates a negative linear relationship between the two variables, but a rather weak one.

Correlation coefficient

- The correlation coefficient is a statistical measure of how closely one set of data values $\{x_1, x_2, \ldots, x_N\}$ are correlated with another $\{y_1, y_2, \ldots, y_N\}$.

$$\rho_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

- Here, μ_x and σ_x are the mean and standard deviation of the x_i values, while μ_y and σ_y are the mean and standard deviation of the y_i values.
- The correlation coefficient measures the strength of a *linear relationship* between two variables.

Interpreting correlation coefficient values

- $\rho_{x,y}$ has a range of [-1,1]
- If $\rho_{x,y}$ is close to 0 then this suggests that there is no correlation.
- If $\rho_{x,y}$ is nearer +1 then this suggests that x and y are positively correlated.
- If $\rho_{x,y}$ is closer to -1 then this suggests that x and y are negatively correlated.

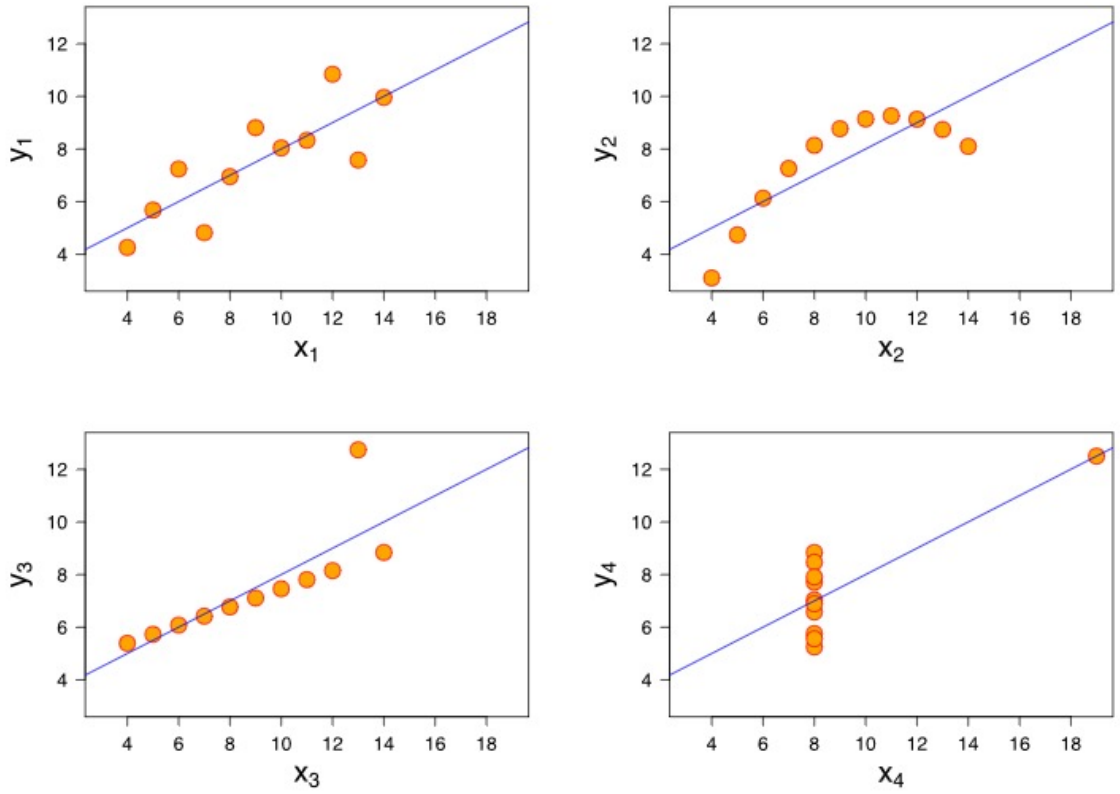
Anscombe's Quartet

- These four datasets have the same summary statistics:
  - $\mu_x = 9$
  - $\sigma^2_x = 11$
  - $\mu_y = 7.5$
  - $\sigma^2_y = 4.125$
  - $\rho_{x,y} = 0.816$

Anscombe's quartet

I		II		III		IV	
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

Anscombe's Quartet



- But they look really different!
- Always draw a scatterplot to visualise your data!

Estimating the correlation coefficient from a sample

- Suppose that we have sample data $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$ drawn from a much larger population of size N, so $n << N$.
- Pearson's correlation coefficient is calculated like this:

- Here, m_x and m_y are the estimates of the population means, and s_x and s_y are the estimates of the population standard deviations.

Hypothesis testing

- Scatterplots and the correlation coefficient may suggest possible correlations between variables.
- Any such suggestion of a correlation is a hypothesis.
- Statistical tests provide the mathematical tools to assess evidence and carry out hypothesis testing.

Hypothesis testing

- We differentiate between research hypothesis and statistical hypothesis.
- Research hypothesis:
  - This is the one you state in your research paper. You want to test it mathematically.
- Statistical hypothesis:
  - H0: there is nothing out of the ordinary in the data: no correlation, no effect, nothing to see.
  - H1: there is indeed something going on…

Hypothesis testing example

- Research hypothesis:
  - Hypertensive patients treated with the new drug X will show greater lowering of their blood pressure than hypertensive patients treated with the currently available drug Y.
- Statistical hypothesis:
  - H0: $\mu_1 \leq \mu_2$
  - H1: $\mu_1 > \mu_2$
- Here μ_1 is the mean lowering of blood pressure in the group treated with drug X, and μ_2 is the corresponding metric for patients treated with drug Y.

How hypothesis testing works

- We state our research and statistical hypotheses.
- We collect data for the test and calculate an appropriate statistic from the data; call this R.
- The hypothesis test is then to investigate how likely it is that we would see a result like R if the null hypothesis were true.
- This chance is called a p-value, with $0 \leq p \leq 1$.
- If p is small enough, then we conclude that the null hypothesis is a poor explanation for the observed data. Based on this we may reject the null hypothesis.

Significance

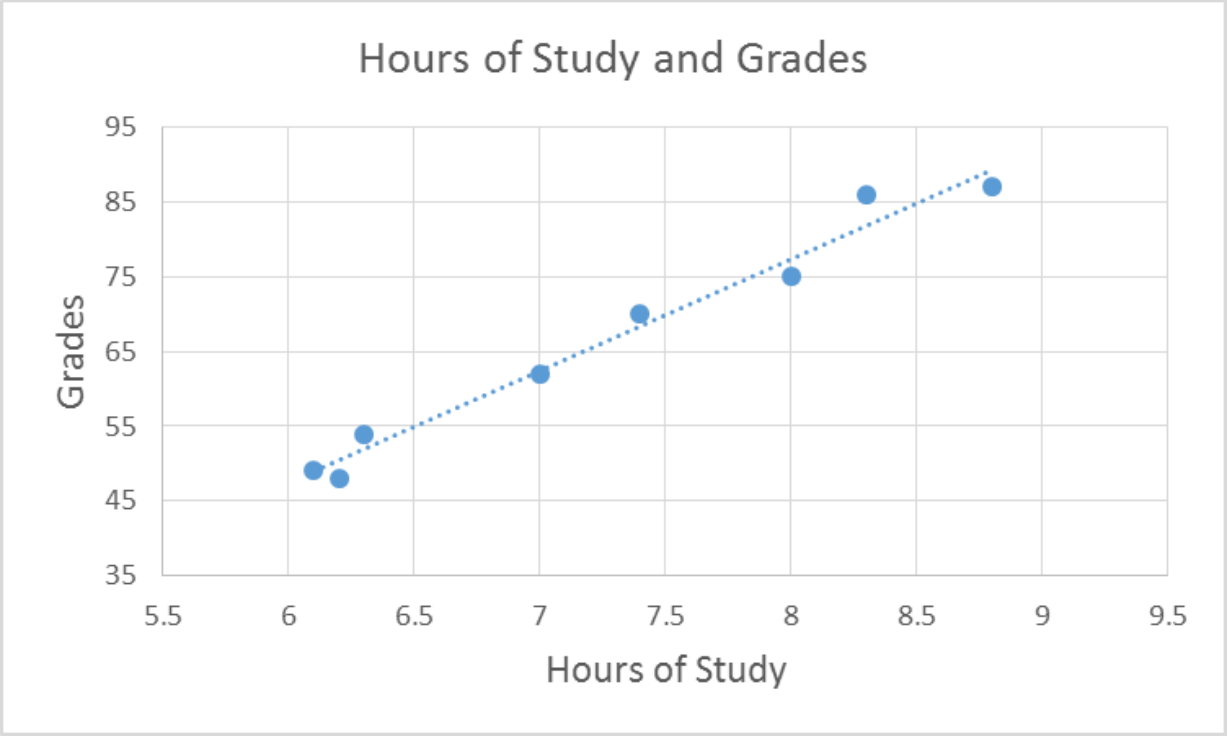
- So when is p “small” enough?
- Standard thresholds are:
  - $p < 0.05$, meaning that there is less than 1 chance in 20 of obtaining the observed result by chance, if the null hypothesis is true; or
  - $p < 0.01$, meaning less than 1 chance in 100 ; or
  - $p < 0.001$, meaning less than 1 chance in 1000.
- An observation that leads us to reject the null hypothesis is described as statistically significant.

Correlation Coefficient as a Statistical Test

- The null hypothesis is that there is no correlation.
- We calculate the correlation coefficient $\rho_{x,y}$ and then do one of two things:
  - Look in a table of critical values for this statistic, to see whether the value we have is significant;
  - Compute directly the p-value for this statistic, to see whether it is small.
- Depending on the result, we may reject the null hypothesis.

Example: correlation between hours of study and final grade

Weekly hours of study	Grades
8	75
7.4	70
8.3	86
6.2	48
6.3	54
7	62
8.8	87
6.1	49



Example: correlation between hours of study and final grade

- $\rho_{x,y} \simeq 0.988$
- Hypothesis testing:
  - H0: There is no correlation between weekly hours of study and final exam grades in Statistics.
  - H1: There is a correlation between weekly hours of study and final exam grades in Statistics.

ρ	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
$N = 7$	0.669	0.754	0.875	0.951
$N = 8$	0.621	0.707	0.834	0.925
$N = 9$	0.582	0.666	0.798	0.898
$N = 10$	0.549	0.632	0.765	0.872

Example: correlation between hours of study and final grade

- The correlation coefficient 0.988 is well above the critical value 0.925 for $p < 0.001$.
- We can reject the null hypothesis.
- Our data strongly indicate a positive correlation between weekly hours of study and final exam grades in Statistics.

Recap on investigating correlation

- Are two (numerical) variables correlated?
  - Draw a scatter plot. Does it look as though there is a relationship between the two variables?
  - Calculate the correlation coefficient $\rho_{x,y}$. Is it close to -1, 0 or 1?
  - Perform hypothesis testing:
    - State the null and alternate hypotheses.
    - Look in a table of critical values to see whether $\rho_{x,y}$ is large, given the number of data points. If $\rho_{x,y}$ is above the critical value for some chosen *p*, say 0.05 or 0.01, then this may be judged statistically significant and lead us to reject the null hypothesis.
    - OR Compute directly the p-value, to see whether it is smaller than the significance level chosen, leading us to reject the null hypothesis.

Conclusions

- All these steps are important:
  - Relying solely on the correlation coefficient without visualising data is a bad idea. Remember Anscombe's Quartet!
  - Arguing about correlation based on the correlation coefficient without carrying out hypothesis testing is only half the story: this can tell us how strong a linear correlation might be, but not how significant.
- There is debate over what an appropriate significance level is.
- Correlation does not imply causation!