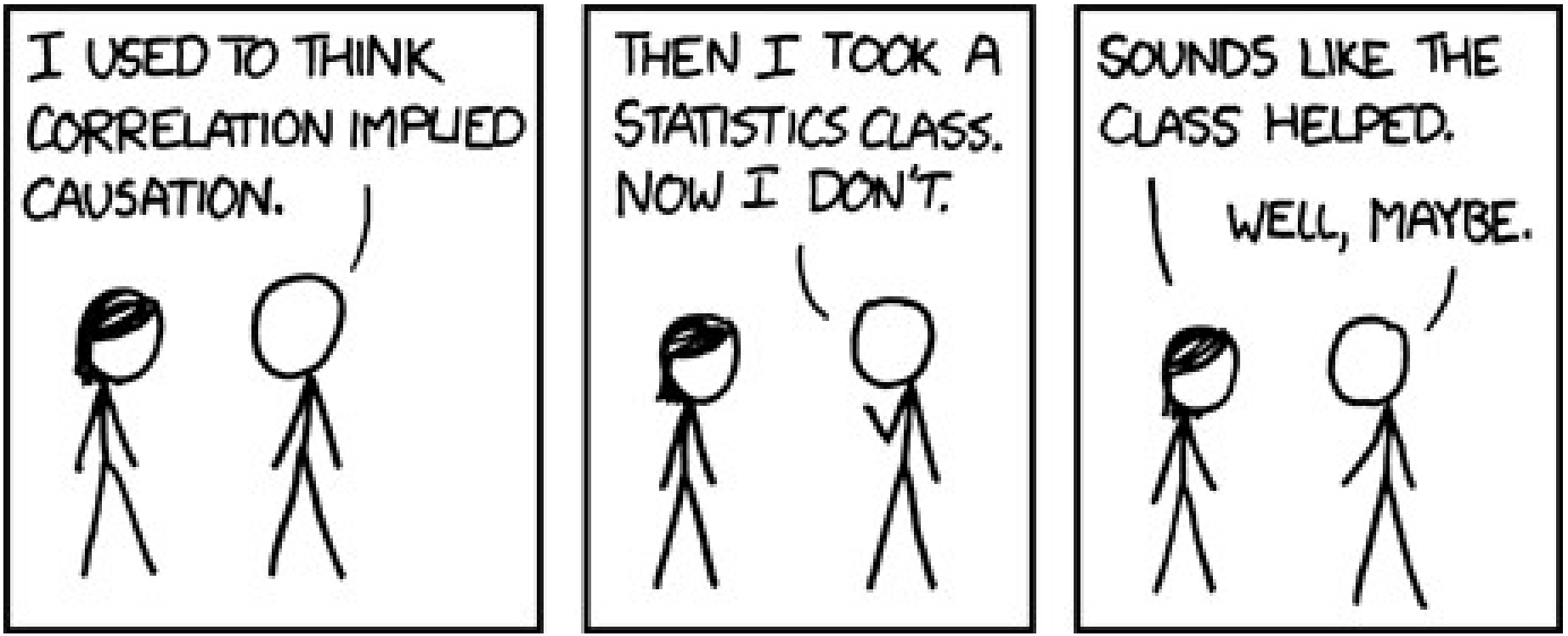
Analyzing Data and Statistics



xkcd.com

Most statistical work can be done, and more easily done, by computer using programs such as:

MS Excel

Open Office

SPSS

SAS

Most statistical work can be done, and more easily done, by computer using programs such as:

MS Excel is the most common.

Available from UT for cheap, ~$30.

If you have not used it, start practicing now.

Most statistical work can be done, and more easily done, by computer using programs such as:

Open Office is a free alternative.

If you have not used it, start practicing now.

The Basics:
Mean, median, and mode

Mean- aka the average.

Sum of all numbers divided by the number of data points.

$(14+17+7+6+4+11+8)/7 = 9.57$

Median- the middle number of a group of ordered numbers

1 17 7 6 4 11 8

4 6 7 8 11 14 17 median is 8

Median- the middle number of a group of ordered numbers

1 17 7 6 4 11 8

2 6 7 8 11 14 17 median is 8

What about 4 6 7 11 14 17?

Median- the middle number of a group of ordered numbers

1 17 7 6 4 11 8

2 6 7 8 11 14 17 median is 8

What about 4 6 7 11 14 17?

Median is 9.

Mode- the most common value in a group.

9, 8, 3, 4, 5, 2, 4, 5, 2, 3, 6, 1, 6, 2, 3, 9, 2, 6

Mode is 2

Why are there 3 ways to analyze a group of numbers?

Why are there 3 ways to analyze a group of numbers?

The mean is the most common form of analysis.

Why are there 3 ways to analyze a group of numbers?

The mean is the most common form of analysis.
2, 3, 2, 4, 2, 7, 2, 5, 3, 2, 5, 4, 3, 5, 6, 121, 130

Mean = 18

Why are there 3 ways to analyze a group of numbers?

2, 3, 2, 4, 2, 7, 2, 5, 3, 2, 5, 4, 3, 5, 6, 121, 130

Mean = 18

Is this an accurate representation of these numbers?

Why are there 3 ways to analyze a group of numbers?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Median can be more accurate when there are a few especially large or small numbers.

What is your favorite color?

What is your favorite color?

Mode can be used with non-numerical data.

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Standard Deviation is a measure of variability.

Standard deviation is a measure of variability. The standard deviation is the root mean square (RMS) deviation of the values from their arithmetic mean.

$$s = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$$

where $\sum$ = Sum of
X = Individual score
M = Mean of all scores
N = Sample size (Number of scores)

(Do not memorize this formula; you will do these calculations via spreadsheet.)

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Mean = 18

Standard deviation = 40.5

Standard deviation is a measure of variability.

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67

Standard deviation = 1.6

Standard deviation is a measure of variability.

Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7
(121, 130)

Mean = 3.67

Median was 4

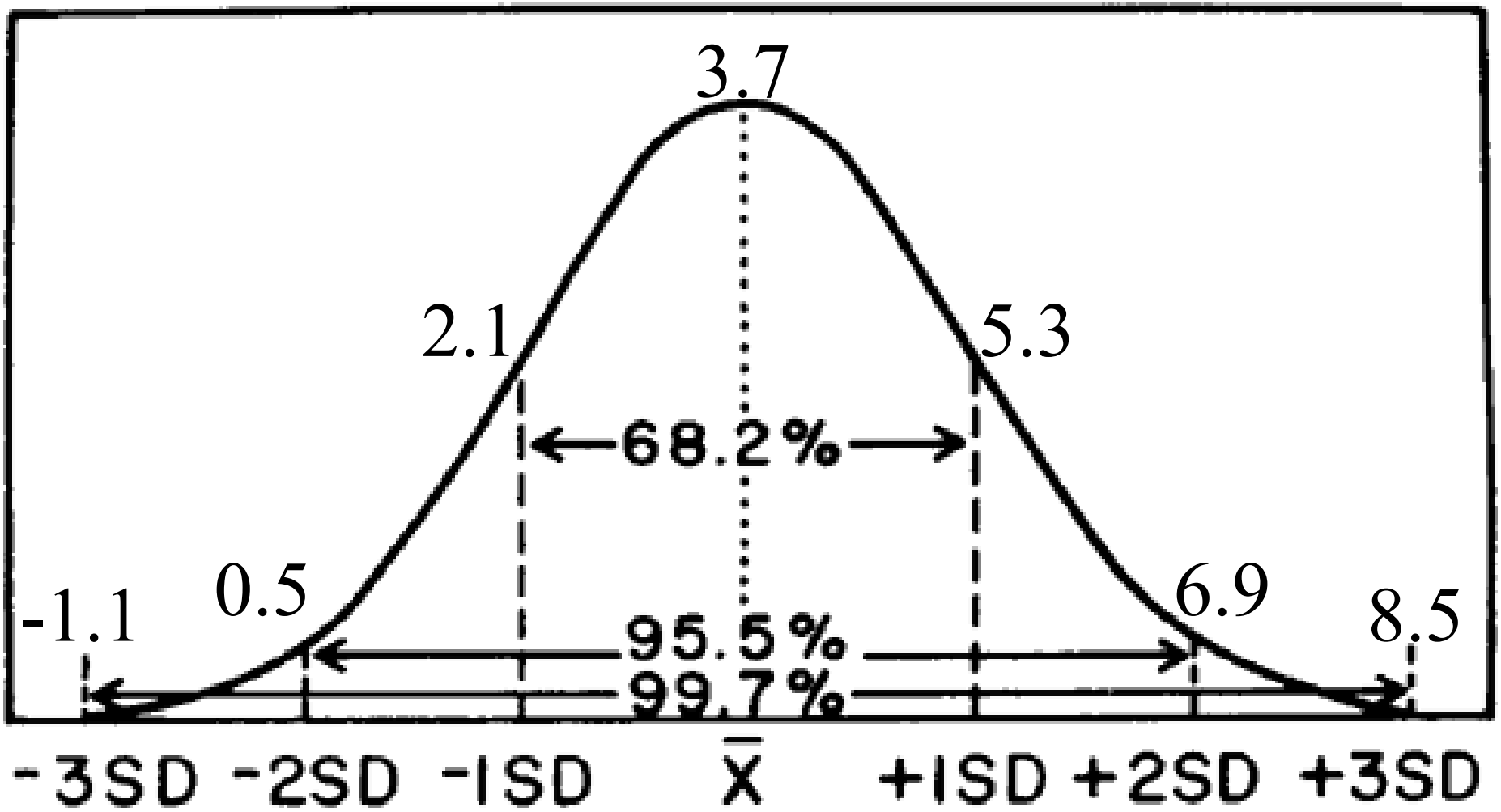
Is there a numerical way to determine the accuracy of our analysis?

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6

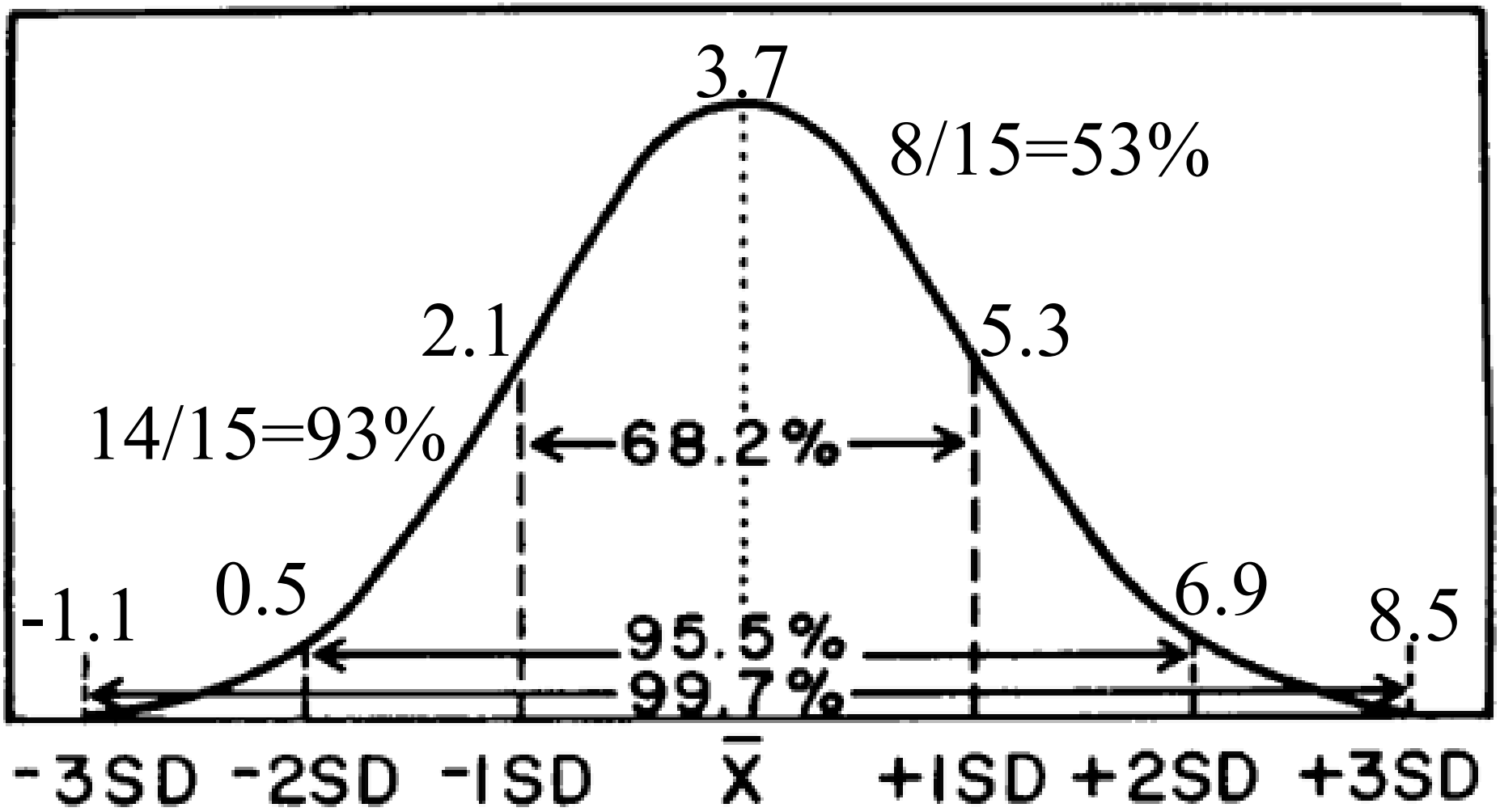
Standard deviation is a measure of variability.

Percent of data at 1, 2, or 3 standard deviations from the mean



http://www.westgard.com/lesson34.htm

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7



15/15=100%

http://www.westgard.com/lesson34.htm

How significant of a difference is this?

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6 range = 2.07 to 5.27

And

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

Mean = 7.25 ± 1.48 range = 5.77 to 8.73

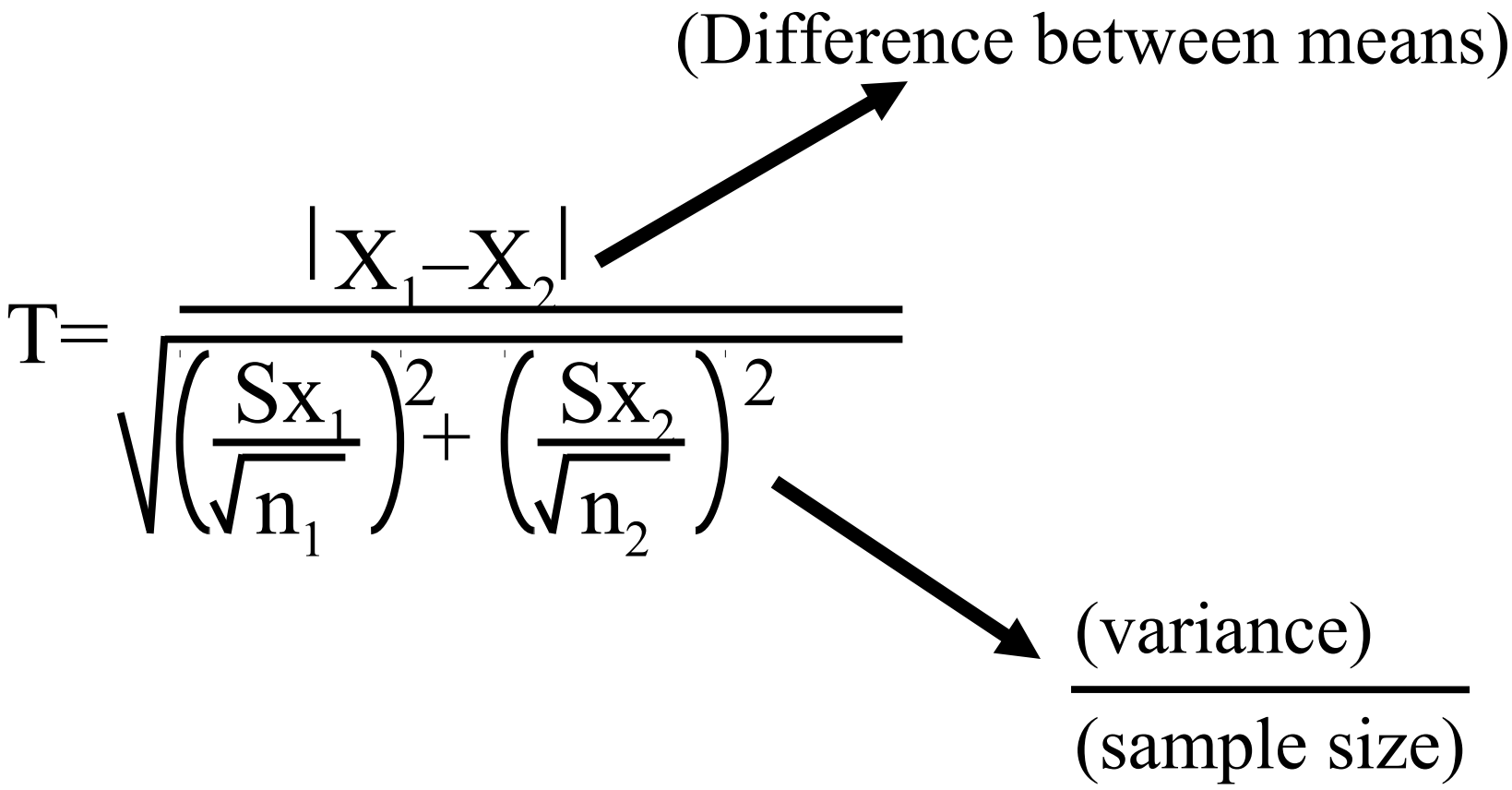
The ‘Students’ T-test is a method to assign a numerical value of statistical difference.

The ‘Students’ T-test is a method to assign a numerical value of statistical difference.

$$T = \frac{|X_1 - X_2|}{\sqrt{\left(\frac{Sx_1}{\sqrt{n_1}}\right)^2 + \left(\frac{Sx_2}{\sqrt{n_2}}\right)^2}}$$

(Do not memorize this formula; you will do these calculations via spreadsheet.)

The ‘Students’ T-test is a method to assign a numerical value of statistical difference.



The ‘Students’ T-test is a method to assign a numerical value of statistical difference.

$$T = \frac{|X_1 - X_2|}{\sqrt{\left(\frac{Sx_1}{\sqrt{n_1}}\right)^2 + \left(\frac{Sx_2}{\sqrt{n_2}}\right)^2}}$$

T is then used to look up the P-value from a table. Also need ‘degrees of freedom’ $= (n_1+n_2)-1$.

Partial table for determining P from T

	P-value		
Df	**0.05**	**0.02**	**0.01**
1	12.71	31.82	63.66
2	4.303	6.965	9.925
3	3.182	4.541	5.841

T

How significant of a difference is this? Using a speadsheet to get a P value = 3.44×10^{-6}.

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6

And

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

Mean = 7.25 ± 1.48

How significant of a difference is this?

P value = 3.44×10^{-6}. So the chance that these 2 sets of data are **not** significantly different is 3.44×10^{-6}

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ± 1.6

And

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

Mean = 7.25 ± 1.48

How significant of a difference is this?

P value = 3.44×10^{-6}. So the chance that these 2 sets of data are significantly different is

$1 - 3.44 \times 10^{-6}$ or 0.999996559

We can be 99.9996559% certain that the difference is statistically significant.

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 9, 5

Mean = 7.25 ± 1.48

In this data set, the range of +/- one standard deviation overlaps, but the T-test shows a very significant difference between these sets of numbers.

Set 1= 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7

Mean = 3.67 ±1.6 range = 2.07 to 5.27

Set 2= 8, 6, 7, 8, 9, 5, 6, 7, 9, 8, 4, 5

Mean = 6.83 ± 1.64 range = 5.19 to 8.47

P-value = 4.41×10^{-5}

Generally a P-value of 0.05 or less is considered a statistically significant difference.

20% random difference : 80% confidence

10% random difference : 90% confidence

5% random difference : 95% confidence

1% random difference : 99% confidence

0.1% random difference : 99.9% confidence

T-test is one valid and accurate method for determining if 2 means have a statistically significant difference, or if the difference is merely by chance.

Outliers…

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Outliers: When is data invalid?

Outliers: When is data invalid?

Not simply when you want it to be.

Outliers: When is data invalid?

Not simply when you want it to be.

Dixon's Q test can determine if a value is statistically an outlier.

Dixon's Q test can determine if a value is statistically an outlier.

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = |(700 - 766)| \div |(789 - 700)|$$

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = |(700 - 766)| \div |(789 - 700)| = 0.742$$

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = |(700 - 766)| \div |(789 - 700)| = 0.742$$ **So?**

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

You need the critical values for Q table:

Sample #	Q critical value
3	**0.970**
4	**0.831**
5	**0.717**

If Q calc > Q crit rejected

From: E.P. King, J. Am. Statist. Assoc. 48: 531 (1958)

You need the critical values for Q table:

Sample #	Q critical value
3	**0.970**
4	**0.831**
5	**0.717**

If Q calc > Q crit than the outlier can be rejected

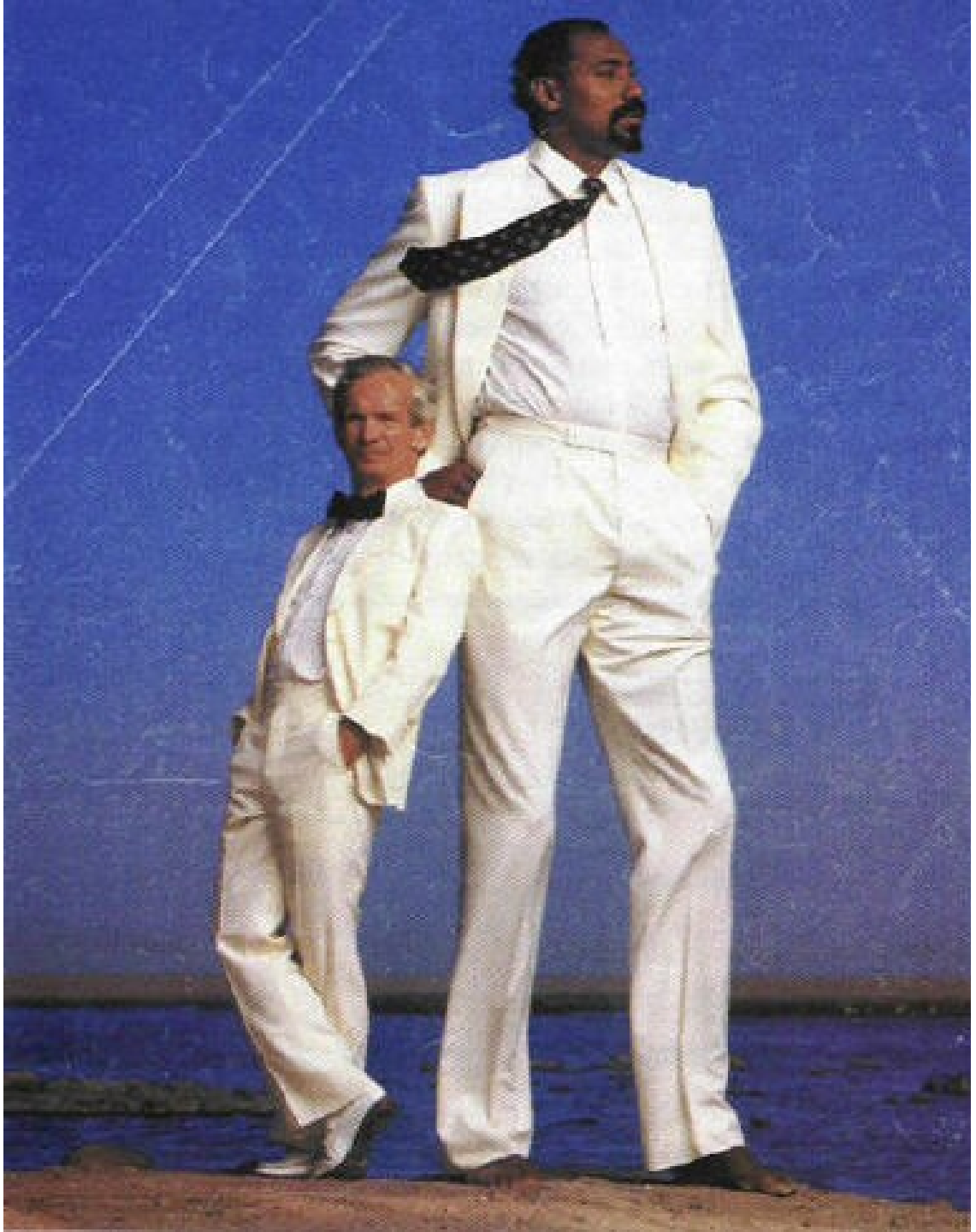
Q calc = 0.742

Q crit = 0.717

= rejection

From: E.P. King, J. Am. Statist. Assoc. 48: 531 (1958)

What can outliers tell us?

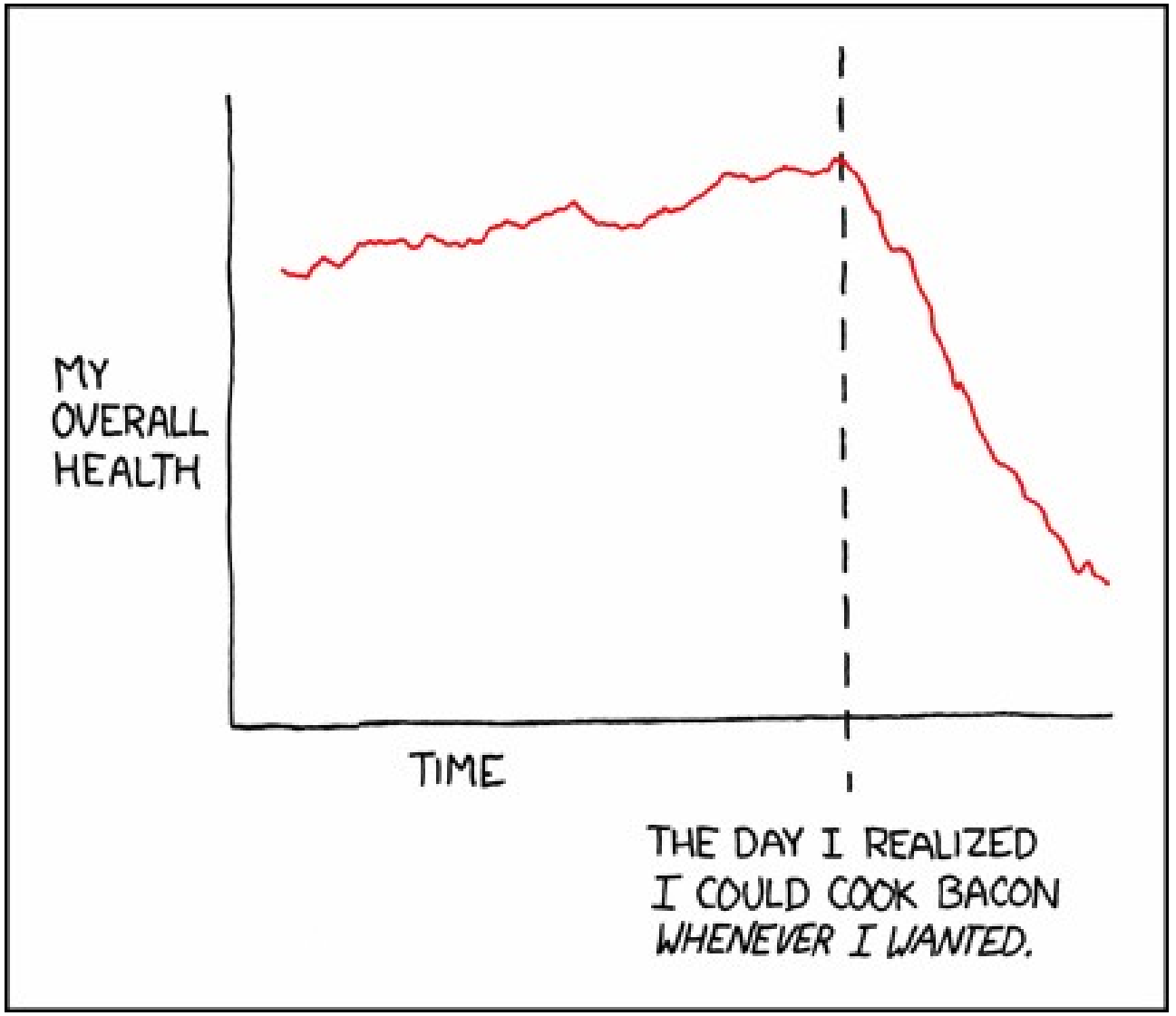


If you made a mistake, you <u>should have</u> already accounted for that.

Outliers can lead to important and fascinating discoveries.

Transposons "jumping genes" were discovered because they did not fit known modes of inheritance.

What about relating 2 variables?



XKCD.com

What about relating 2 variables?

R^2 gives a measure of fit to a line.

If $R^2 = 1$ the data fits perfectly to a straight line

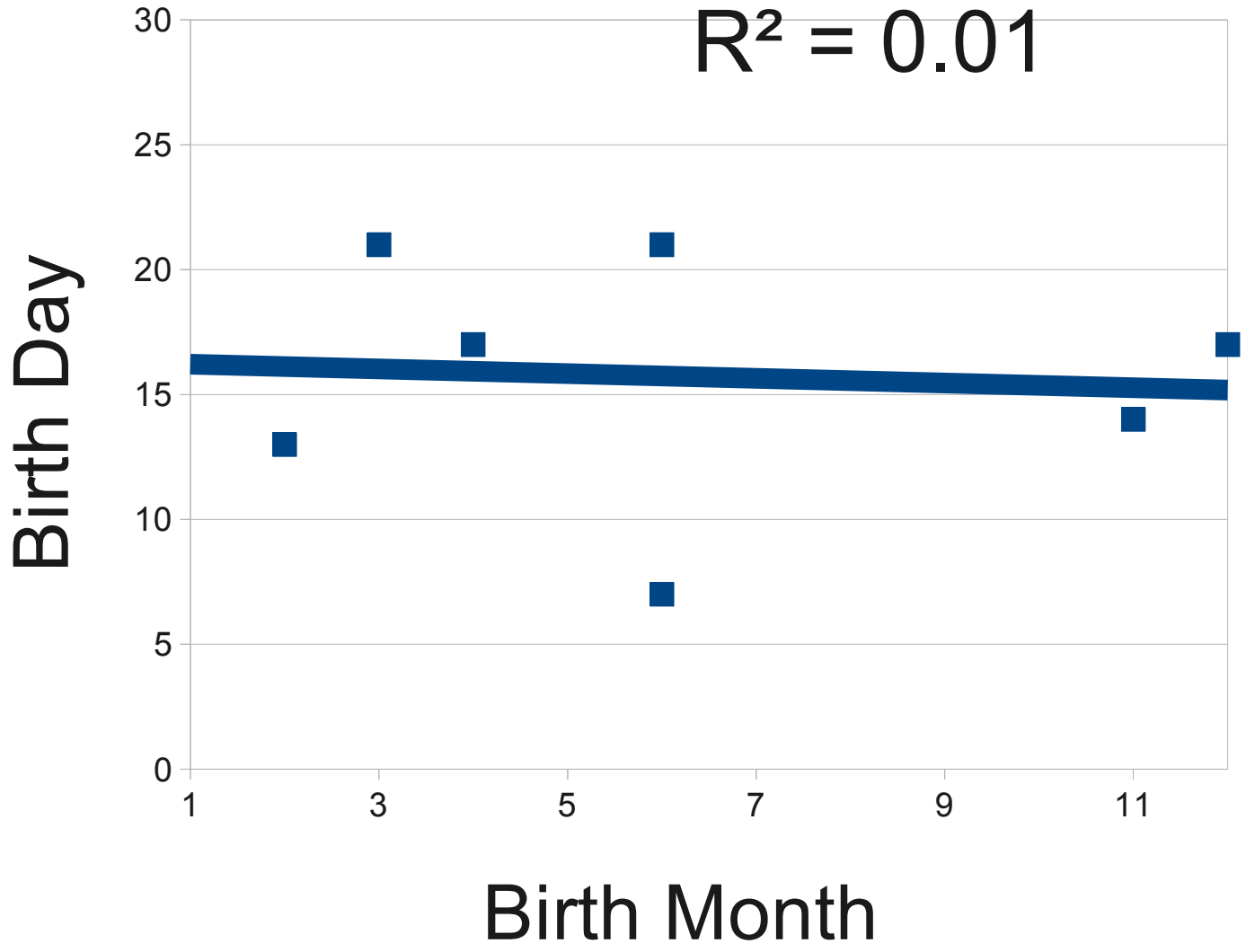
If $R^2 = 0$ there is no correlation between the data

R^2 gives a measure of fit to a line.

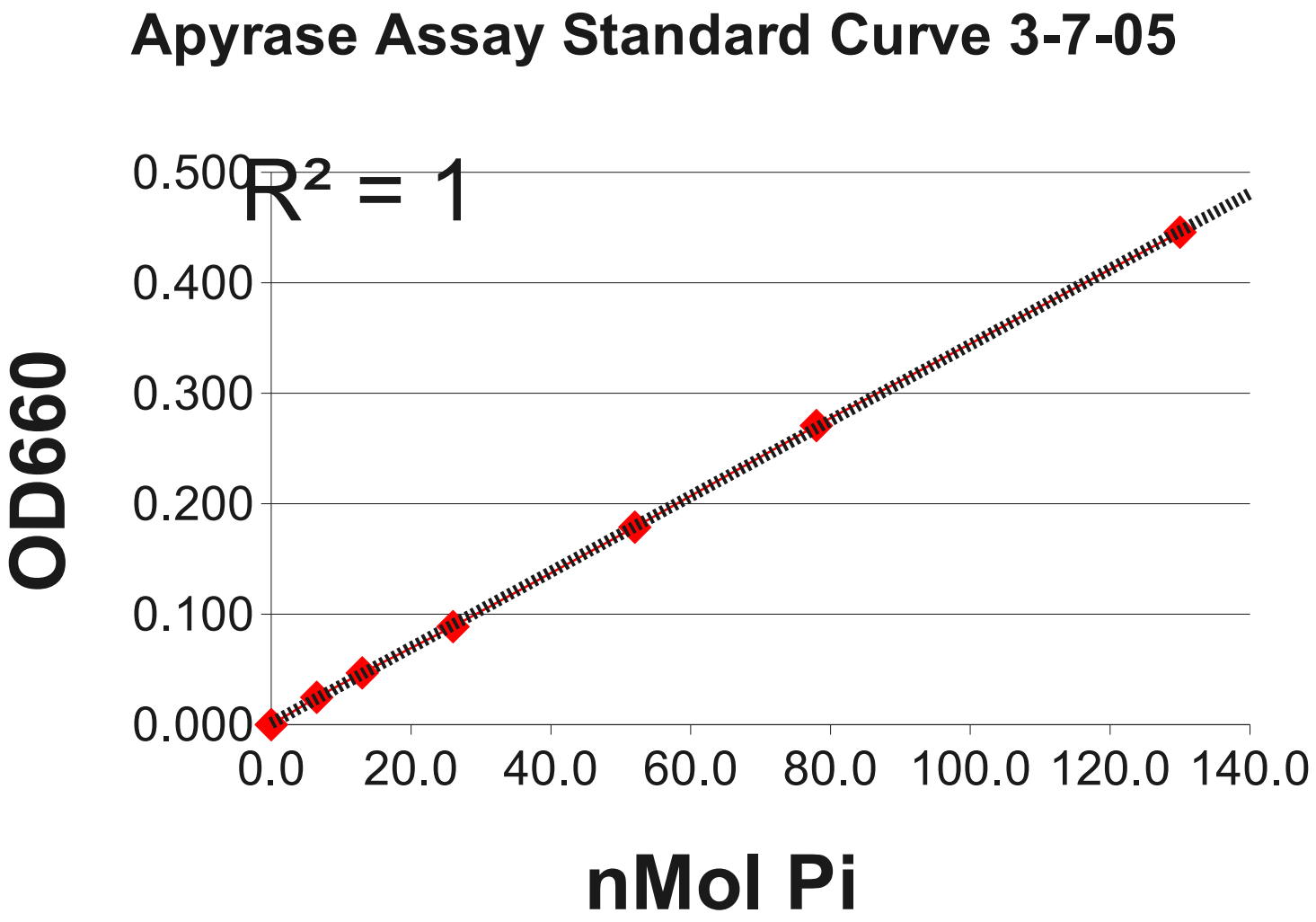
birth month vs birth day

4	17
11	14
6	7
12	17
2	13
6	21
3	21

birth month vs birth day



phosphate quantity vs absorbance



What about relating 2 variables?

- **To use R^2 the data must be continually variable...**

R^2 gives a measure of fit to a line.

If $R^2 = 1$ the data fits perfectly to a straight line

If $R^2 = 0$ there is no correlation between the data

Samples vs populations

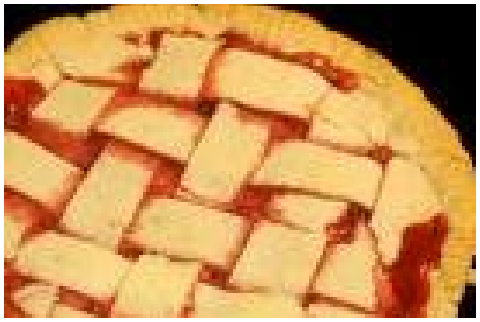








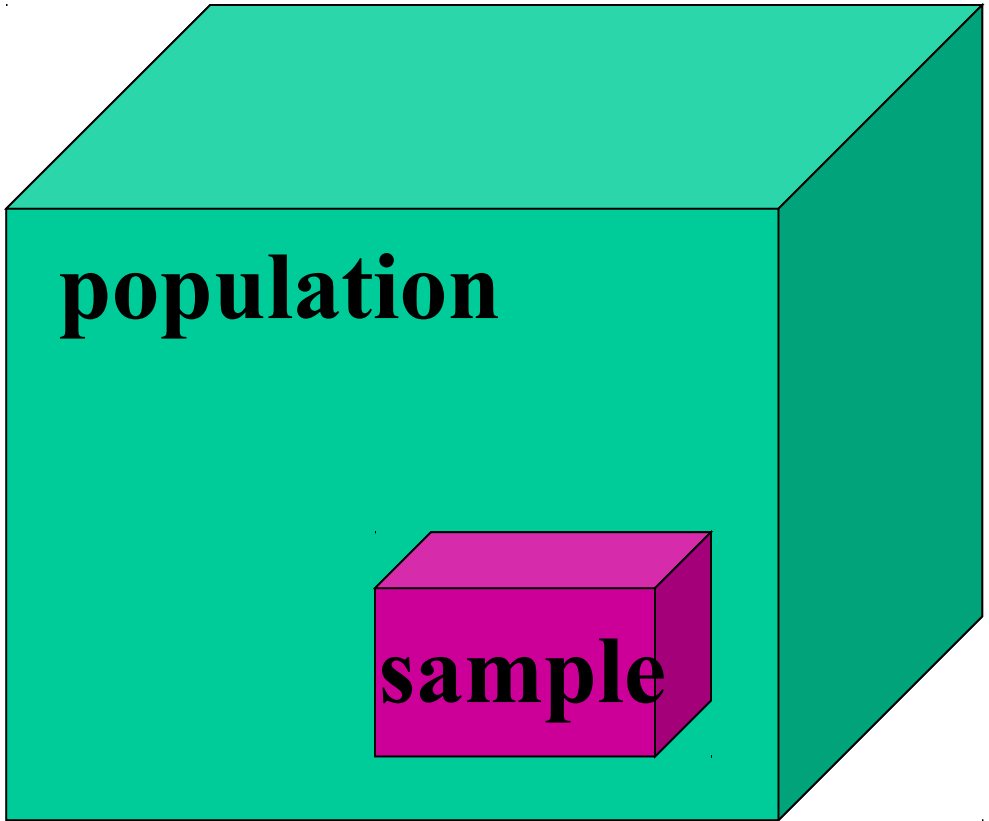





Samples vs populations

Population- everything or everyone about which information is sought

Sample- a subset of a population (that is hopefully representative of the population)



Population-

- U.S. census
- Dogs
- 1 – infinity

Sample-

- Travis county
- Poodles
- Prime numbers

Why use a sample instead of a population?

Why use a sample instead of a population?

- Logistics

Why use a sample instead of a population?

- Logistics
- Cost

Why use a sample instead of a population?

- Logistics
- Cost
- Time

Samples:

Random- each member of population has an equal chance of being part of the sample.

or

Representative- ensuring that certain parameters of your sample match the population.

Replicates:

Technical vs Experimental

Technical replicate- one treatment is divided into multiple samples.

Experimental replicate- different, replicate, treatments are done to different samples.

Testing blood sugar levels after eating a Snickers:

Testing blood sugar levels after eating a Snickers:

Divide a participants blood into 3 samples and test blood sugar in each sample.

Technical or Experimental replicate?

Testing blood sugar levels after eating a Snickers:

Test 3 different people.

Technical or Experimental replicate?

Testing blood sugar levels after eating a Snickers:

Test the same person on 3 different days.

Technical or Experimental replicate?

What sample size do you need?

What sample size do you need?

It depends on the error you expect.

To determine an appropriate sample size, you need to estimate a few parameters.

- Means
- Standard Deviation

- Power:

The probability that an experiment will have a significant (positive) result, that is have a p-value of less than the specified significance level (usually 5%).

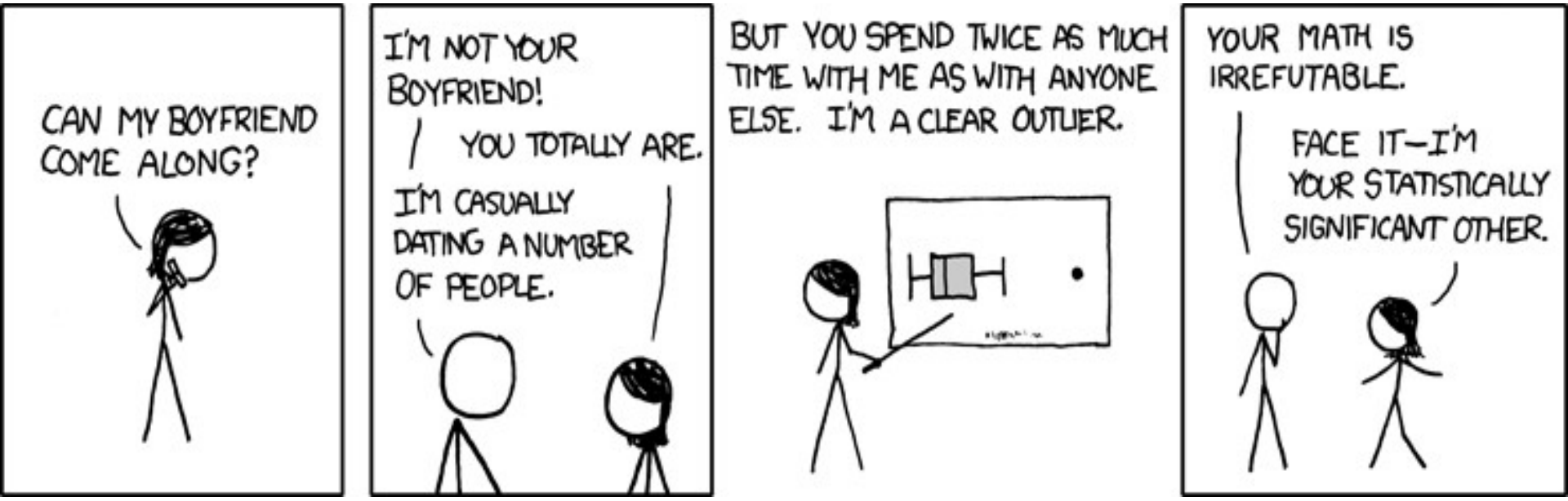
This calculator will help you determine the appropriate sample size:

http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

What sample size do you need?

It depends on the error you expect.

(So it is impossible to predict with 100% accuracy before the experiment is carried out.)


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

xkcd.com