Vocational Aqualabs - Vocational Generic Skills for Researchers

Experimental Design

Unit 2a – Fundamentals of Statistics

Trevor Telfer and James Bron
Senior Lecturers
University of Stirling

Education and Culture DG
Lifelong Learning Programme

Basic statistical analysis

- **Why?**
- An understanding of statistics allows one to critically assess the work of others according to a universally approved framework. Statistical techniques provide criteria for detecting differences or relationships between samples or variables and allow the presentation of experimental or observational results to others. **However .....**

- "*An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts - for support rather than for illumination*." Andrew Lang.
- "*There are three kinds of lies: Lies, Damn Lies, and Statistics*." A saying attributed to Mark Twain.

Basic statistical analysis

- "…*epidemic of flawed statistical analysis plaguing today's research literature. From medical "breakthroughs" that prove to be mirages to grand conclusions drawn from tiny samples, the journals are awash with unreliable findings based on faulty statistics*."

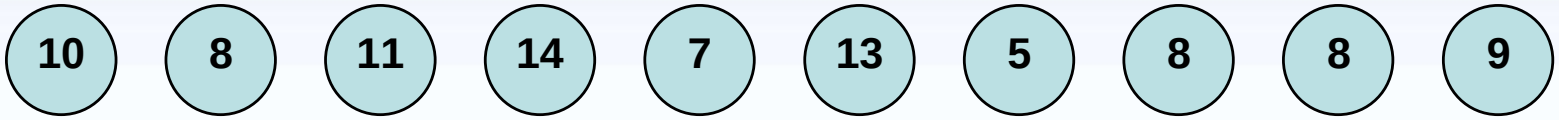
  (Matthews, New Scientist 2385, March 2003)

- "*A recent survey of 141 papers published in the journal "Infection and Immunity" showed ~50% to contain errors in statistical analysis or reporting of results*".

  (Olsen, 2003).

Mean, median, mode and range

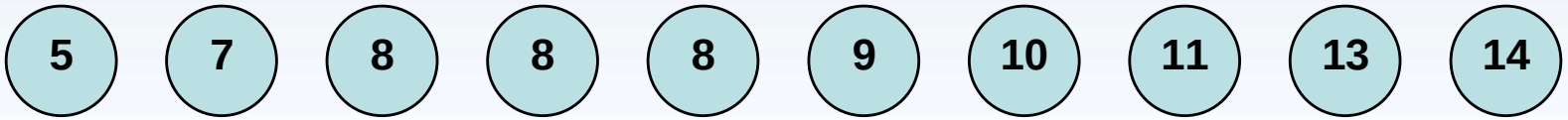
- **Mean** of ten numbers, selected at random



- The mean is the sum of these 10 numbers divided by 10 i.e. 93 / 10 = **9.3**

Mean, median, mode and range

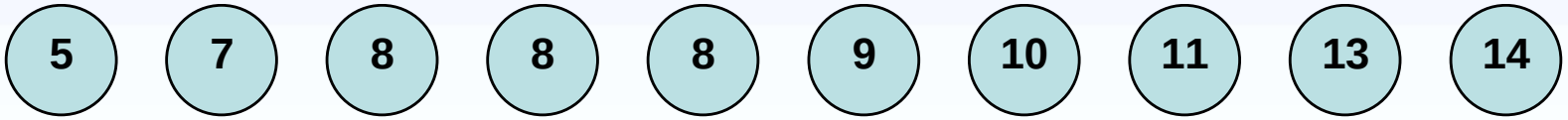
- To calculate the **median**, place all the numbers in order of size, like this:



- Now find the middle number or middle numbers in your series. In the example above there are two middle numbers “8” and “9”, so add these together and divide by 2 = **8.5**

Mean, median, mode and range

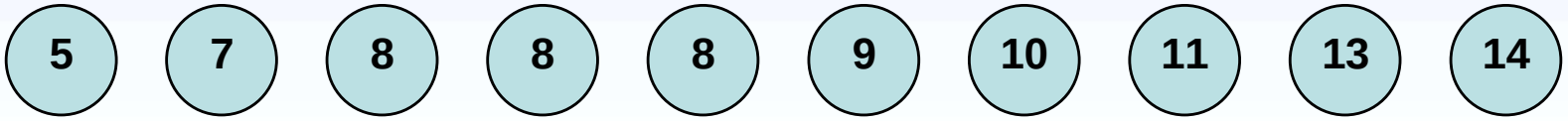
- To calculate the **mode**, look for the number that occurs the most in the series.



- In this example, it is “8” and its **frequency** (i.e. the number of times it appears), is “3” (i.e. 3 times).

Mean, median, mode and range

- The **range** of your data, is simply the upper and lower limits of your observations



- In this series the range is 5 to 14

Standard deviation

- Standard deviations (***σ***) are used to explain how widely spread the values in your collected data are, normally in relation to the mean of your data.
- For example, “How hairy are university lectures?” Here are two lecturers.

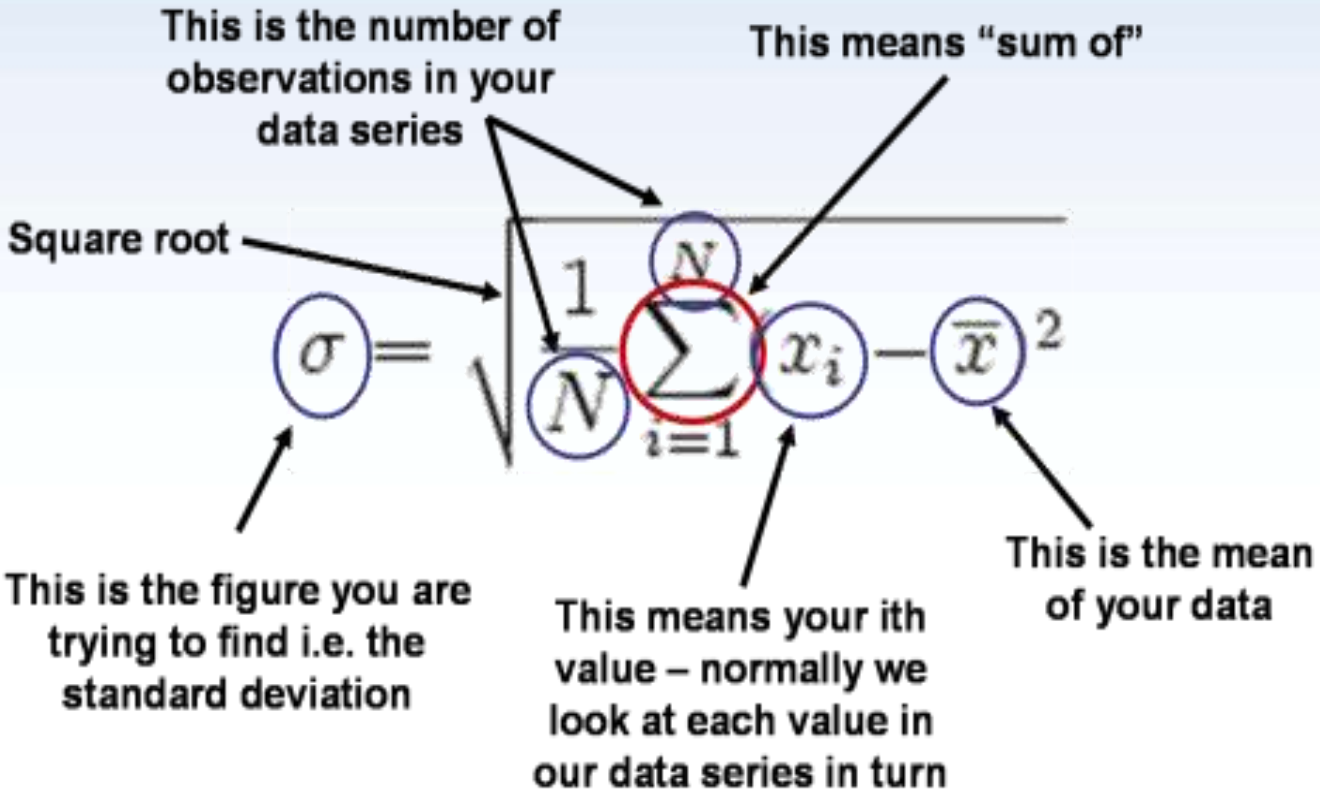
Dr Smith only gets a score of “4” on a scale of 1-10. However, Dr Brown gets a score of “8. Consider the **mean** of their “hairiness” scores (4 + 8 / 2) = 6.The standard deviation of these scores is “2”

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Vocational Aqualabs

Standard deviation

- Where:



Vocational Aqualabs

Standard error

- Whereas σ gives an estimate of the population mean
- What we really want to know is “how good is our estimate of the mean?”
- Here we calculate the standard error
- Where:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Variance

- **Variance** describes the amount of dispersion that there is in our data and how it is spread about the expected mean value
- The variance is σ^2 or S^2

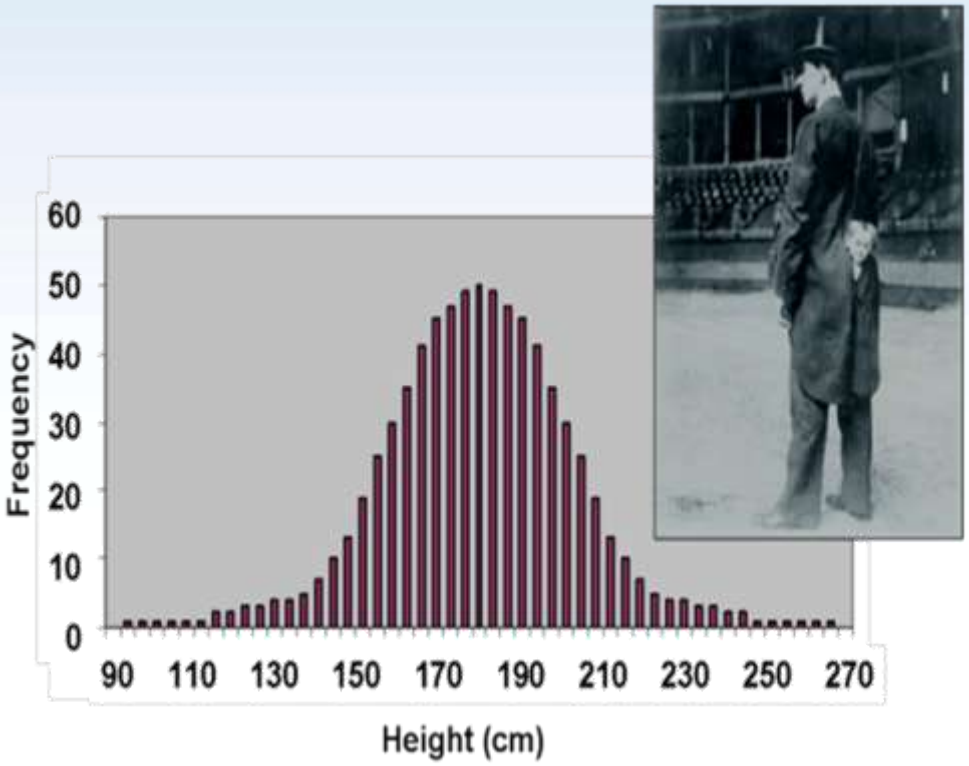
$$\sum d^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$\text{sample variance } (S^2) = \frac{\sum d^2}{n-1}$$

Where: ***Ex*** = the sum of all our obs
n = the number of our obs

Data distribution – “Normal”

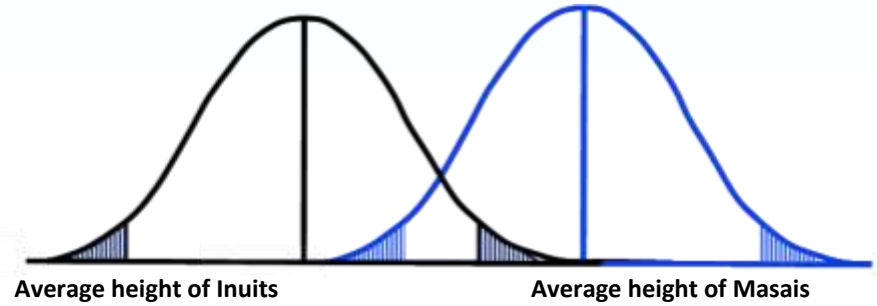
- Description of how the variation in data is distributed in statistical terms
- For example, if you were measure the height of all the humans in the world and then plot the number of people you found of a certain height into categories as a histogram ...



Vocational Aqualabs

Data distribution – “Bimodal”

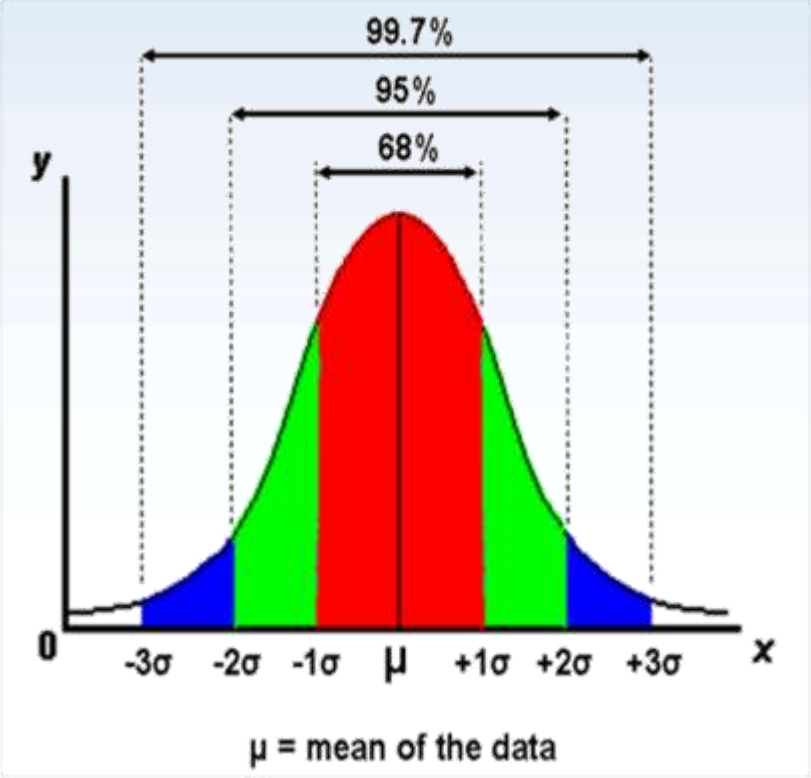
- A sub-set of this data though could illustrate another type of distribution – biomodal
- For example - take the height distribution of Masai people in Africa and the height distribution of adult Inuit people, we would get ……
- The reason we get a bimodal distribution is because Masais are known to be a very tall race of people while the Inuit are known to be quite short in stature.



Vocational Aqualabs

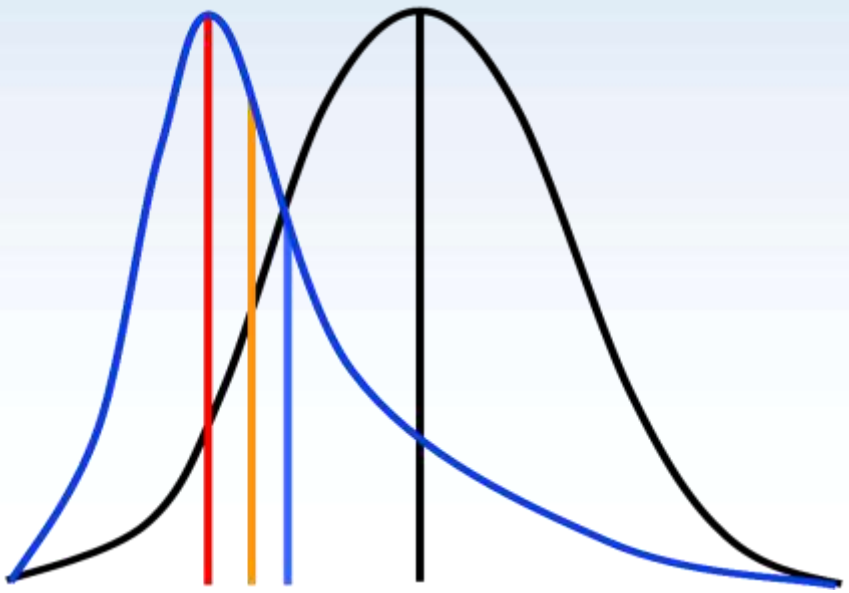
Data distribution

- For normally distributed data:
  - One standard deviation from the mean accounts for about 68% of the observations (p = 0.42)
  - Two standard deviations from the mean accounts for 95% of the observations (p = 0.05)
  - Three standard deviations account for about 99.7% of the observations (p = 0.003)



Data distribution

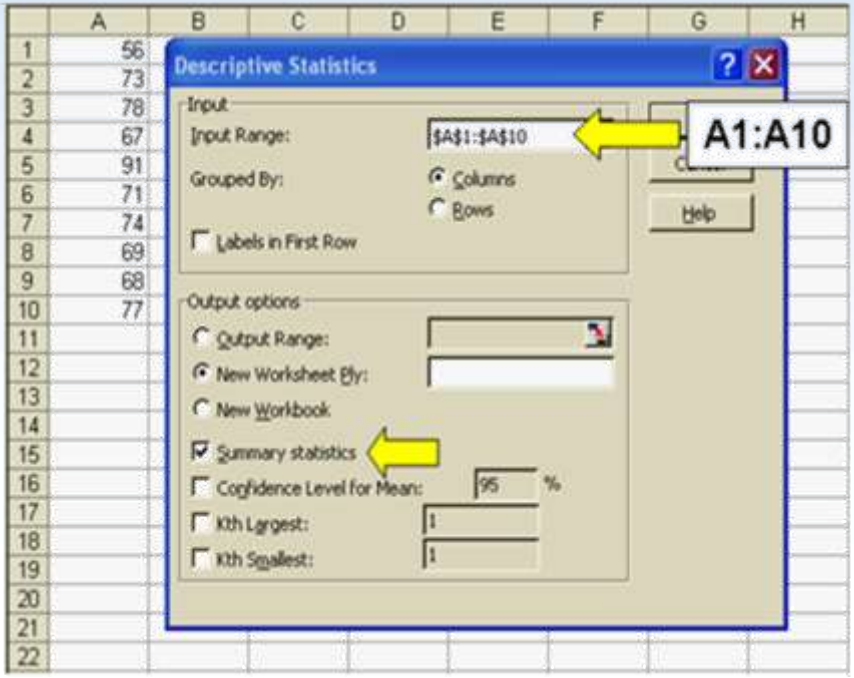
- The use of parametric statistical tests depend on normally distributed data
- These may give wrong conclusions if the sample distribution is skewed
- This can be corrected



Black = normally distributed
Blue = skewed left

Data distribution

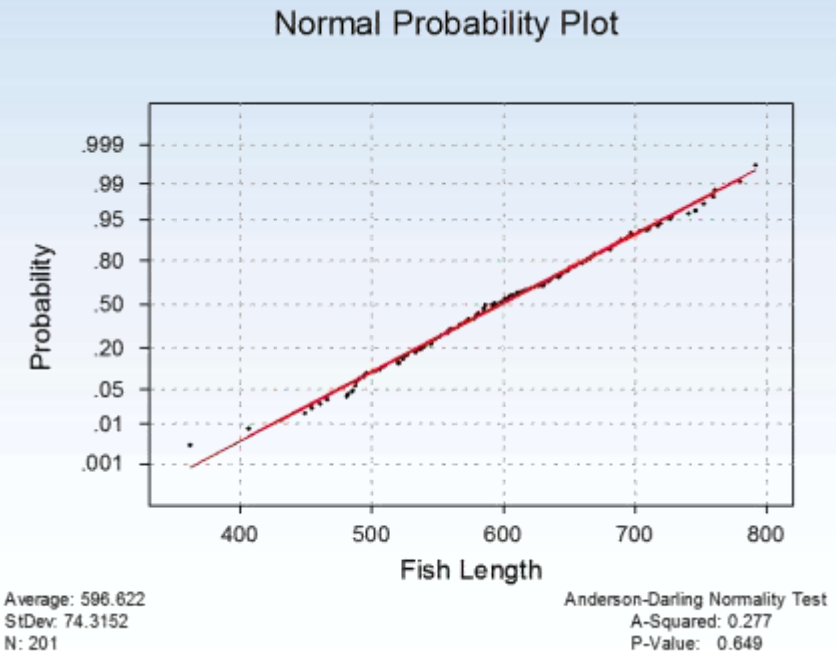
- The distribution of data can be tested by:
  - Plotting
  - Descriptive statistics – measure of skewness
  - Statistical tests such as the Anderson Darling method



Vocational Aqualabs

Data distribution

- The distribution of data can be tested by:
  - Plotting
  - Descriptive statistics – measure of skewness
  - Statistical tests such as the Anderson Darling method



If the probability value (p) for the test is GREATER THAN 0.05 then your sample is NOT significantly different from a normal distribution i.e. ***it is normally distributed***

Vocational Aqualabs

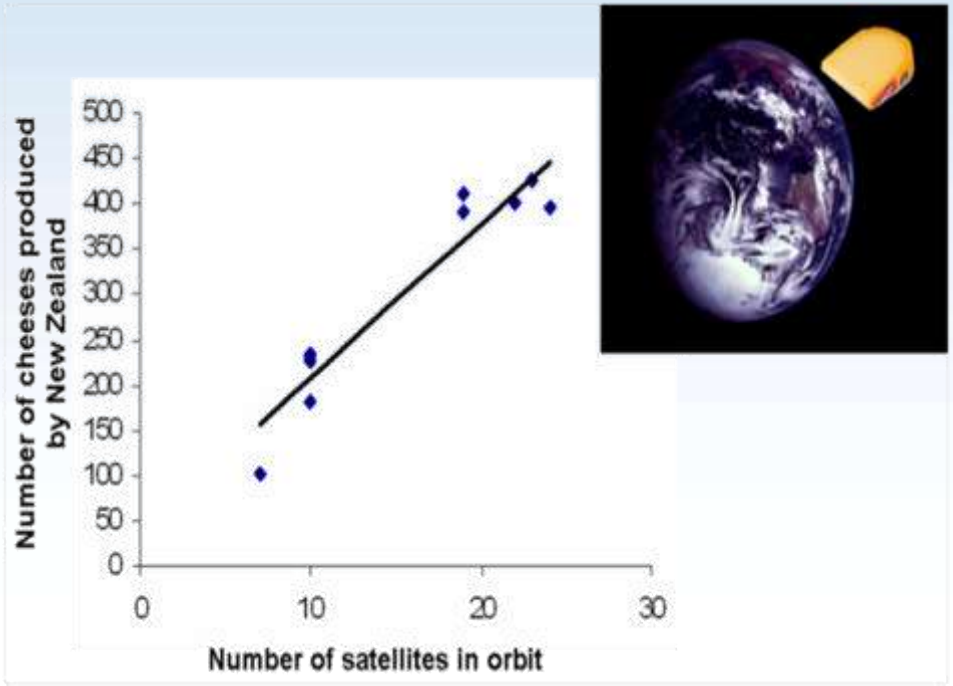
Statistical “traps”

Statistical analyses must be used sensibly and have clear causality for example:

- “A statistical test is not an infallible guide and can never [absolutely] *prove* a particular hypothesis. There is always a possibility that an alternative hypothesis exists” Elliot (1993)
- Causal associations may be two way – e.g. small fish may be more susceptible to high infection levels **or** high infection levels may give rise to smaller fish.
- A correlation between two variables **does not** necessarily mean that they are associated.

Statistical “traps” - causality

- The graph shows a significant correlation between the number of satellites orbiting the Earth in the period 1994-2003 and the number of cheeses produced in New Zealand over the same period.
- Clear relationship?
- Of course not.



Vocational Aqualabs

Identifying causality

Determine causality, these are (after Hill, 1965):

- **Strength of Association**
- **Consistency** with other knowledge
- **Specificity** e.g. only one recognised cause
- **Temporality** – Exposure to risk before effect
- **Biologic gradient** – e.g. Dose response
- **Biologic plausibility** – Scientific theory
- **Biologic coherence** – No conflict with knowledge
- **Experimental evidence** – Test cause and effect
- **Analogy** – Similarity to known association

Probability vs Predition

- Statistics allows evaluation of the probability of a given event - it is **impossible** to predict the outcome of a given event **for certain** using statistics
- It is possible to predict the probable outcome to a certain pre-determined level, e.g:
  - For the biological sciences a probability of **95%** is by convention accepted as **significant** i.e. $p<0.05$.
- By accepting $p<0.05$ investigators are acknowledging that their statistical proof of a tested hypothesis will be erroneous one time out of every twenty.

Vocational Aqualabs

Thank you

Trevor Telfer/James Bron
Institute of Aquaculture
University of Stirling

AQUATT

UNIVERSITY OF ABERDEEN

UNIVERSITY OF STIRLING

INNOVAMAR

Aquark

Education and Culture DG
Lifelong Learning Programme

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained herein.