What do I need to know first?

- **You need a question**

- **To answer interesting questions you need data.**
- **That is why you need statistics.**

Hypothesis

- **A hypothesis is a proposed explanation for a phenomenon.**
- **Scientific hypothesis: one can test it!**

Non-hypothesis statements can be altered to become hypothesis statements

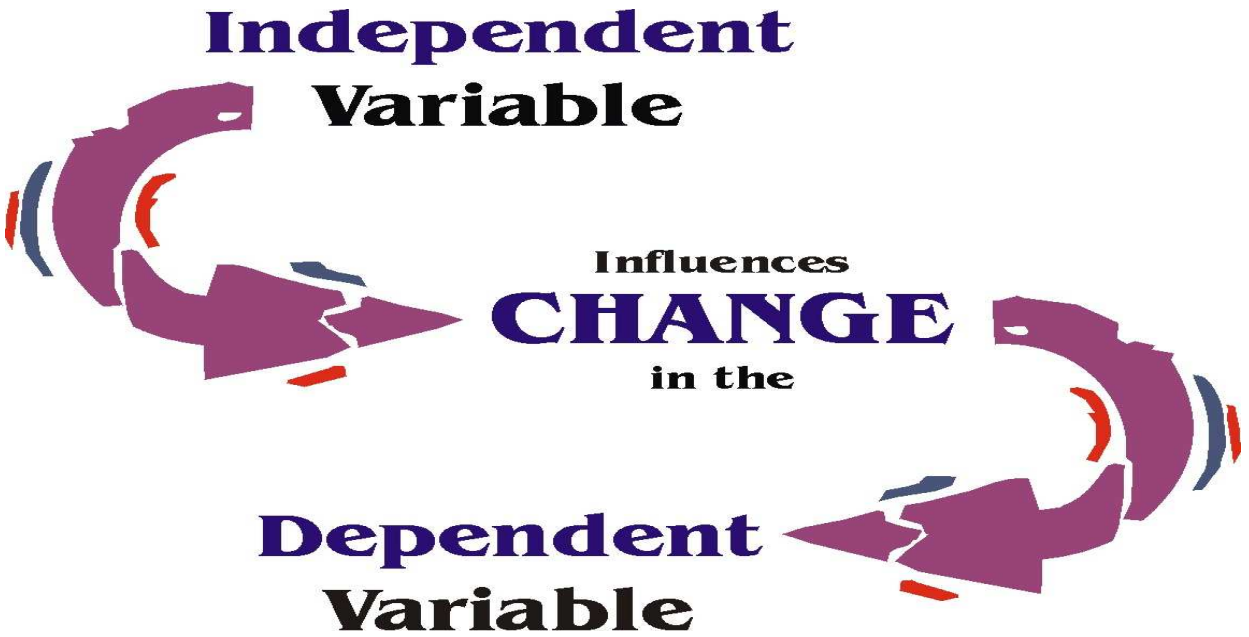
- **The Beatles were the most influential and ever.**

can be restructured to:

- **The Beatles were the best-selling band ever.**

Variables

- **things that can change or vary.**



Independent X dependent

- **Hypothesis can be expressed in terms of two variables:**
- **Proposed cause.**
- **Proposed outcome.**

Coca cola is an effective spermicide (Umpierre, Hill and Anderson, 1985).

Proposed cause = coca cola

Proposed effect = dead sperm

Cause = independent variable = its value does not depend on any other variable.

Effect = dependent variable = value depend on the cause.

Levels of measurement

- **Categorical and continuous**

- **Categorical: entities are divided into distinct categories.**
- **Continuous: within the limits the variable ranges, any value is possible.**
- Number of minutes to finish a problem.
- Number of correct answers.

Exercises

- **What is the level of measurement of the follow variables?**

a. **The number of downloads of different band's songs on iTunes.**

b. **The names of the bands that were downloaded.**

c. **The position in the iTunes download chart.**

d. **The money earned by the bands from the downloads.**

e. **The instruments played by the band members.**

f. **The time they have spent learning to play their instruments.**

Hypothesis

- **Null hypothesis: states that an effect is absent.**
- **Alternative hypothesis: states that an effect is present.**

- H0: Chocolate do not cause pimples.

- H1: Chocolate cause pimples.

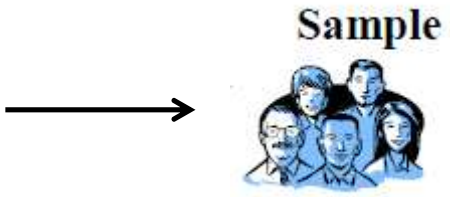
Exercises

- **Write the hypothesis for the questions:**

- Does salt in soil may affect plant growth?
- Does ultra violet light cause skin cancer?
- Does temperature cause leaves to change color?
- Do taller people have larger hand spans?

Inferential statistics

- **Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample.**
- **Sample HAS TO BE representative of the group to which it is being generalized.**



The mean: a very simple statistical model

- **Hypothetical value that does not have to be a value that is observed in the data.**

Ex.: take five UC lecturers and measure the number of friends that they had. We might find the following data:

$$(1 + 2 + 3 + 3 + 4) / 5 = 2.6$$

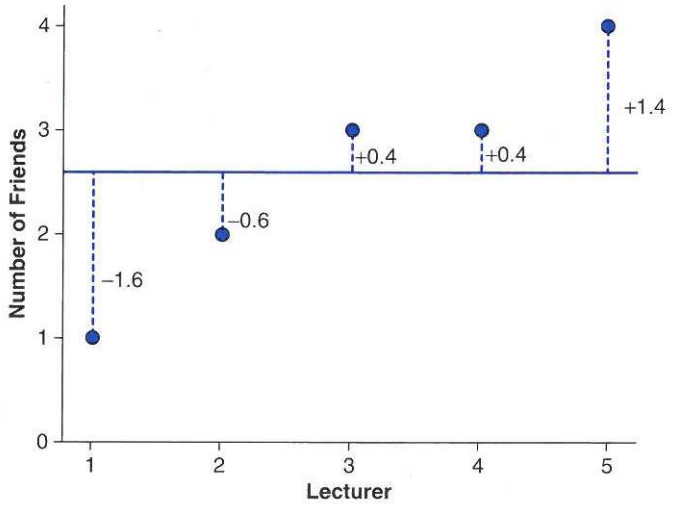
It is impossible to have 2.6 friends (unless you chop someone).

So the mean value is a hypothetical value.

Mean is a model created to summarize our data.

Assessing the fit of the mean

- **We have to assess the fit to know how much our sample model resembles the population reality.**
- **Look at the difference between the data observed and the model fitted.**



- **Deviance = error in model**

Total error:

Add all the total deviances and get the total error.

Total error = sum of deviances

$$= \sum (x_i) = (1.6) + (-0.6) + (0.4) + (0.4) + (1.4) = 0$$

Is the mean a perfect representation of the data?

No, there were errors, but some are + and some are -, and they have simple cancelled each other!

- **How to avoid this problem?**

Rather than calculating the total error, we square each error.

Sum of squared error (SS) = $\sum (x_i).(x_i)$

$(-1.6)^2 + (-0.6)^2 + (0.4)^2 + (0.4)^2 + (1.4)^2 = 5.20$

Is SS a good measure of accuracy of our model?

Yes, but it is dependent of the amount of data.

- **To overcome this problem, we can find the average error by dividing SS by N-1.**

Welcome variance!!!!!

- **Variance (S^2) = $\frac{SS}{N-1} = \frac{\sum (x_i - \bar{x})^2}{N-1} = \frac{5.20}{4} = 1.3$**

But Larissa, why -1????

Welcome degrees of freedom!!!!

DF relates to the number of observations that are free to vary.

Ex.: Rugby game example.

- if values in a sample are 8, 9, 11, 12 (mean = 10) and we change three of these values to 7, 15, and 8, then the forth value must be 10 to keep the mean constant.

So, we know that variance is the average error between the mean and the observations made.

It is a measure of how well the model fits the actual data.

Perfect???

No, because it gives us a measure in units squared.

The average error in our data is 1.3 friends squared!

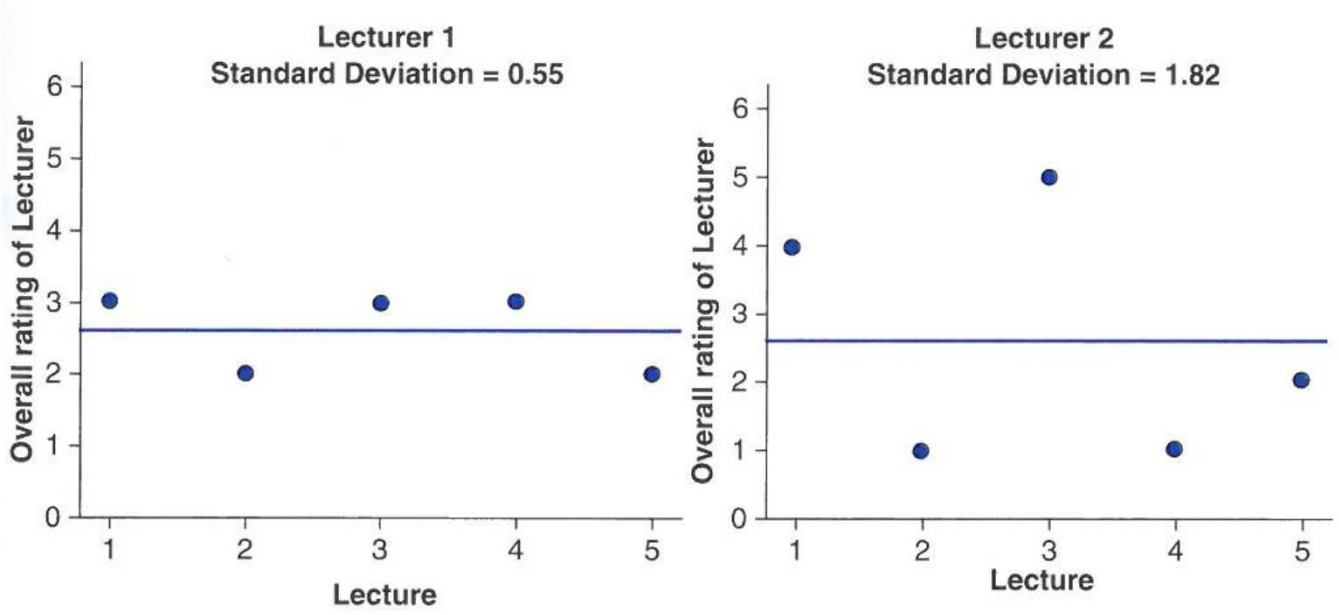
Weird…..

Welcome standard deviation!!!

It is simple the square root of the variance

- **What does large standard deviation mean?**

It indicates that the data points are distant from the mean.



- **But how can I know how representative a sample is likely to be of the population?**
- **Usually (not always) a large sample is defined as greater than 30, leading to a normal distribution.**

Welcome confident interval!!!!

- **CI for the mean is a range of scores constructed such that the population mean will fall within this range in 95% of samples.**
- **The CI interval is NOT an interval within which we are 95% confident that the population mean will fall.**

Analyzing data

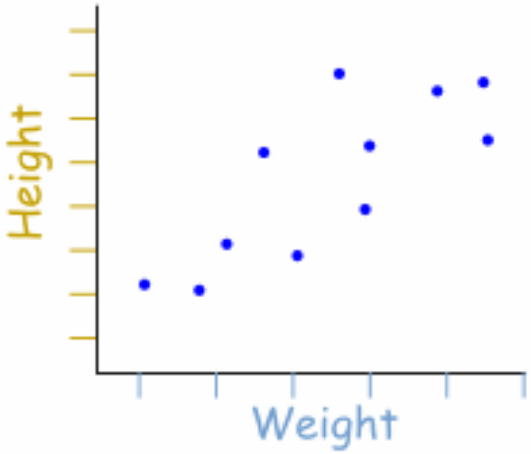
- **Final stage**
- **Look at your data graphically.**
- **Check what the general trends in the data are.**
- **Fit a statistical model to the data.**

Exploring your data before analysis

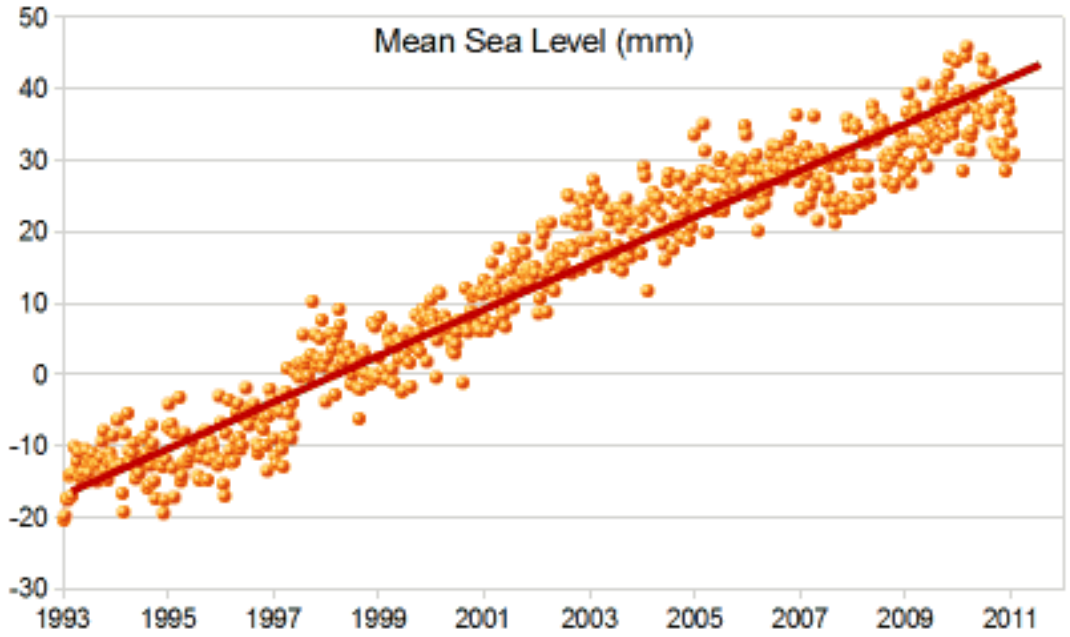
- **How??? Graphs!!!!**

1) Scatter Plot

A graph of plotted points that show the relationship between two sets of data.

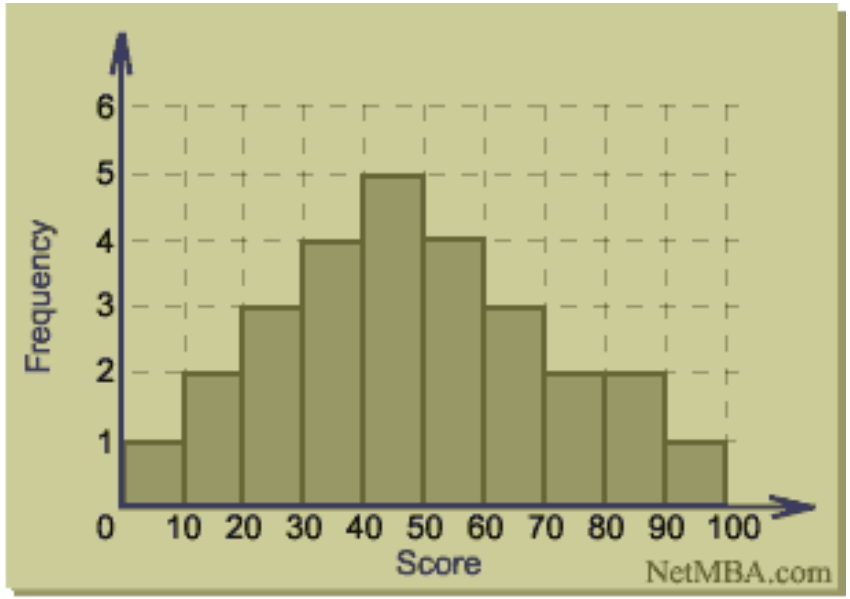


- Adding a funky line:



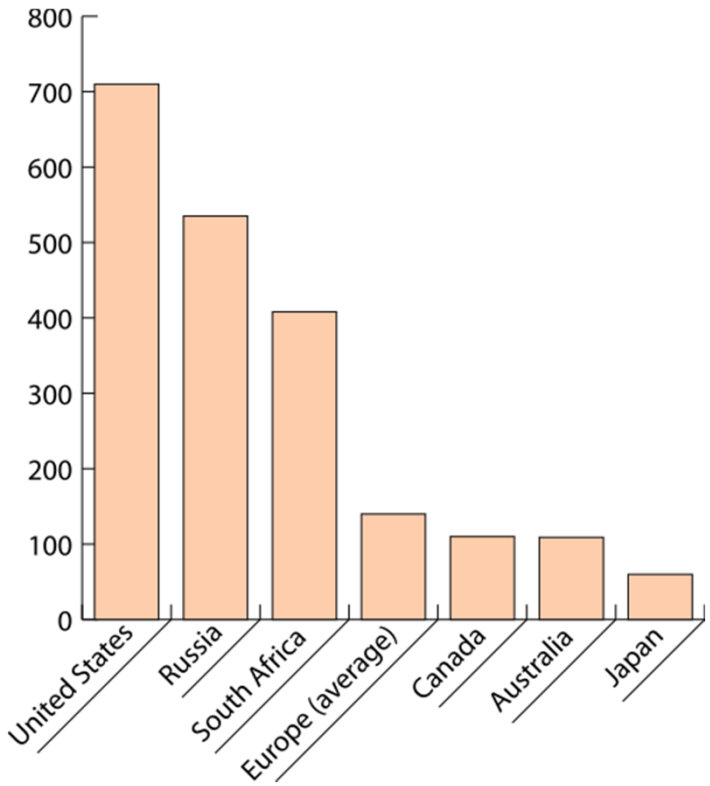
2) Histograms

Graphical representation of the distribution of data. It is an estimate of the probability distribution of a continuous variable.



2) Bar graph

A chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value.



3) Box plot:

- **Depicts groups of numerical data through quartiles.**
- **Whiskers indicates variability outside the upper and lower quartiles**
- **Outliers may be plotted as individual points.**