SOME STATISTICAL CONCEPTS TO HELP US LOOK AT DATA.

PLOUGHING
THROUGH
BIOMETRY

INTRODUCTION

- When looking at data we need to know the following:
- How it was collected and sampled; we hope to assume it is unbiased.
    - size of sample, and
    - sampling design
- What is it showing -
    - a comparison,
    - an association,
    - or a more complex model
- Look for an estimate of confidence in the data, or reliability for the estimate

WHAT ARE VARIABLES?

Some variables we measure by doing an experiment, others we collect data using a survey.

- Number v. Text, (called a string)
- some variables are numeric eg weight, height,
- others are text. Eg name, race, country, suburb.

(nb in a spreadsheet such as Excel numbers can be considered 'text' if you want to Eg Postcode.

Make sure you have them in a form that you want)

FOUR TYPES OF VARIABLES

1 Nominal (qualitative) eg Blood type, race, marital status, gender. They *cannot* be ordered.

2 Ordinal (ranked), eg. stage of disease, May not be measurable and may not be equally spaced but *an order* exists. (less severe to more severe).

3 Interval (number) with no absolute zero, ^{O}C

4 Ratio (number) length, time, mass, volume, ^{O}K

We rarely distinguish between the last two, sometimes called scale

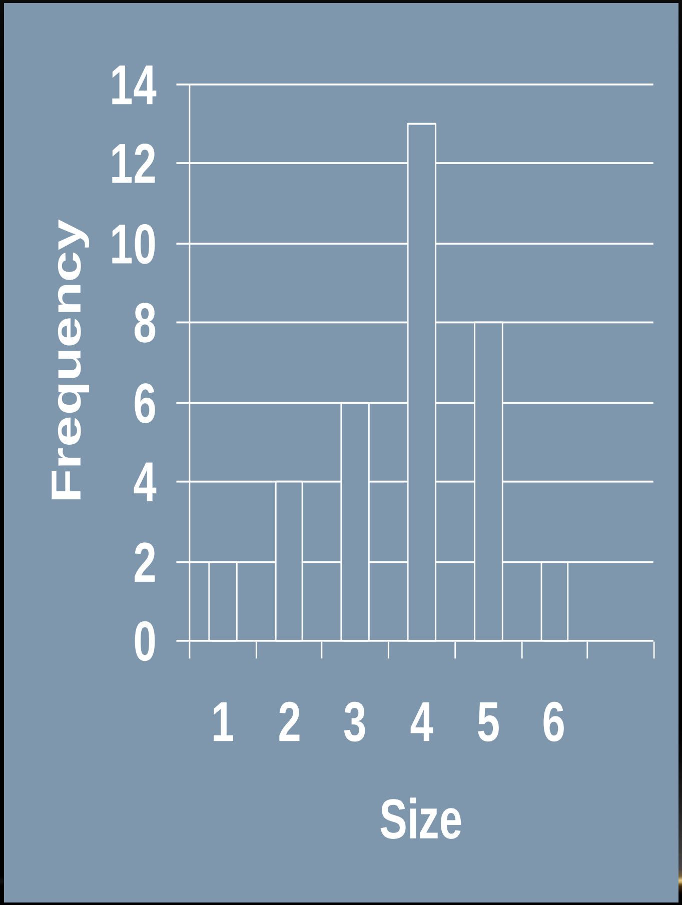
TRY THIS FOR YOURSELF

Type of Data	Examples to think about				
	Height (cm) Weight of Newborn(kg).	Gender. Smoker. Positive for a disease. Race	Suburb. Name. Occupation. Salary	Shoe Size. PostCode. Number of children.	Score for severity of disease.
Numeric or Alpha					
Continuous or discrete.					
Nominal					
Ordinal					
Scale					

POPULATIONS AND SAMPLES

- The set of all possible values for a variable, is the population. We can consider a population mean, or true mean, and a population variance. The mean and variance of a population are called parameters.
- If you don't know the population mean (μ) and standard deviation (σ^2), you can estimate it from a sample.
- Sampling should be random and unbiased.
- If your sample is biased in any way, ie not truly representative you will have an invalid, wrong and biased estimate.
- The mean (xbar) and variance (s^2) from a sample are statistics, they are estimates of the true mean and variance.

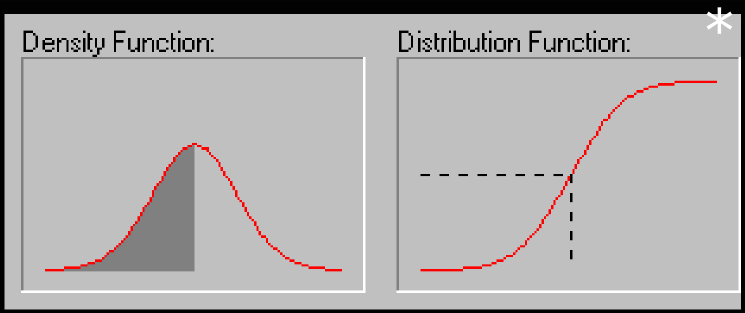
COLLECTION OF DATA



Variable (Size)	1	2	3	4	5
Frequency	2	4	6	13	8

Most data as you collect a large enough sample will approach a normal distribution.

A LOT OF DATA HAS A SIMILAR SHAPE



μ the mean

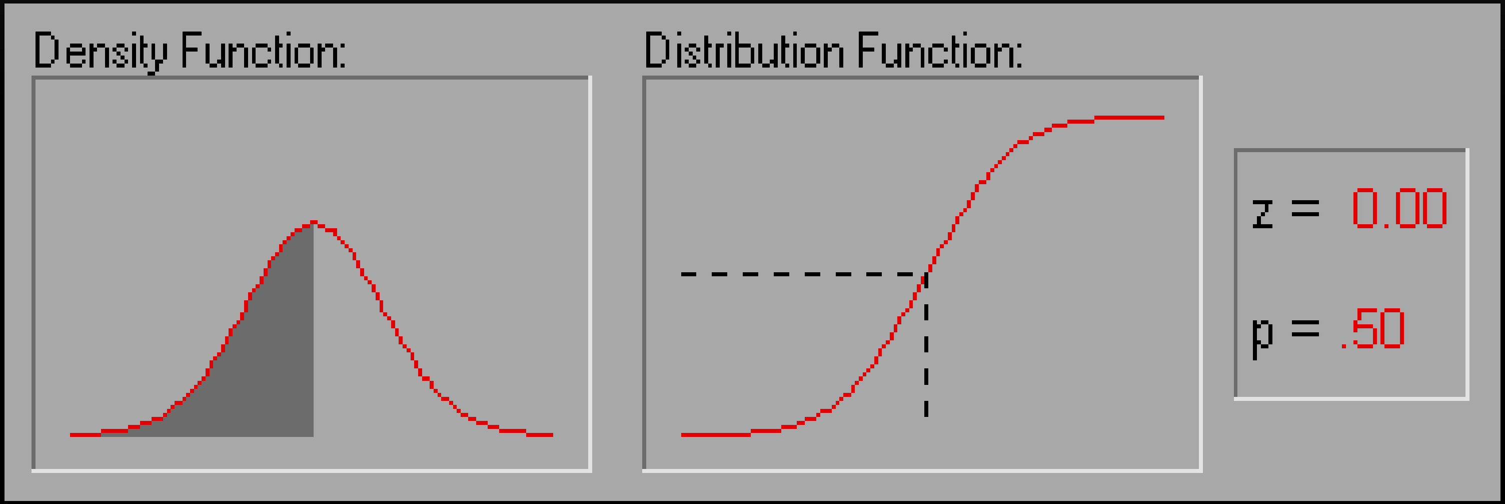
σ Standard deviation

- Shape is a symmetrical bell-shape curve.
- Bell shaped curves can be standardised.
- Frequency plots.
- Cumulative density

* Image is from http://www.statsoft.com/textbook/stathome.html

- It is dynamic on the internet.

NORMAL DISTRIBUTION SHAPE



The probability of value being under the curve is 1.0. (all data is under the curve.)

The mean splits the data into to symmetrical areas of 0.5 or 50% of the data.

PROPERTIES OF THE NORMAL DISTRIBUTION

- Symmetrical, it has a particular ‘bell shaped curve’ or distribution.
- The middle point is called the mean
- The standard deviation, is a measure of variation *at the same scale* as the mean.
- How can we use this? All we need is the mean and standard deviation to describe the data, and for parametric tests statistical tests (eg. t test).

MEANS, VARIANCES AND SD

- The mean (xbar) is the estimate of the population mean.
- The variance (s^2) is the sum of squared deviations from the mean.

- The standard deviation (s) is the square root of the variance; which is a measure of variability taken back to the original scale of the data.

CONFIDENCE INTERVALS

- A mean plus or minus 1.96 * sd gives us the 95% confidence interval. So in Rule of thumb 95% CI = mean ± 2 * sd
- This means if we sampled 100 times, in 95 of the times our estimate of the mean would lie within that ‘confidence interval’.
- Most tables or graphs should show a standard error bar or or 95% CI.

CORRELATION AND REGRESSION

- Don’t confuse these two.
- Correlation is examining the *linear* association between two variables
- Regression, is used to describe the equation for the straight line, and the fit of the data.
- We can fit curved lines (curvilinear), or several variables (multiple regression). Other powerful techniques are available

ASSOCIATIONS AND FIT OF DATA

- Correlation, can find statistical significance of the Correlation Coefficient (ρ).
- R squared in regression, is a measure of the fit of the data.
- Lines of best fit and 90% or 95% CI along a plot; which one is closer in to the line. This is most accurate at the mean points, and curves away.
- Caution on extrapolation outside the range of the data.

EXTRAPOLATION OF DATA

- If your data has been collected over a particular range it *is not good* practise to extrapolate outside of that range.
- %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%