Statistics

Variables

Def'n: characteristics which change from person to person within a study

1. quantitative - varying by number
2. qualitative - varying by description
3. discrete - values differ only by a fixed amount
4. continuous - values differ by any amount

Distributions

1. frequency distributions - eg. tables of frequencies
2. class intervals - summary of frequencies into groups
   - allows concise review
3. relative frequencies - each class interval expressed as a percentage of total
4. cumulative relative frequencies
   - summation of preceeding relative frequencies

Graphical Representation

1. histogram - summary of either class interval or relative frequency
2. bar charts - useful for both qualitative & discrete variables

Measurement

1. nominal measurement - name, birthplace
2. ordinal measurement - mild - moderate - severe
3. interval measurement - meaningful distance between values
   - Celcius temperature scale
   - allows study of differences, but ***not*** absolute magnitude
   - eg. 80°C is not twice as hot as 40°C
4. ratio measurement - allows study of absolute magnitude
   - eg. metres length, degrees Kelvin

Summary Statistics

■ ***Measures of Location or Central Tendency***

1. ***mean*** = arithmetic average value, where x is a quantitative variable
   → interval or ratio

$$\bar{x} = \sum \frac{x_i}{n}$$

2. ***median*** = value above & below which ½ the measurements fall
   → interval or ratio, (± ordinal)

3. ***mode*** = most frequently occurring value
   → nominal, ordinal, interval or ratio

■ ***Measures of Dispersion or Variability***

1. ***range*** = difference between minimum & maximum values

2. ***interquartile range*** = (largest value 3rd quartile) - (largest value 1st quartile)
   → 75th - 25th percentiles in paediatrics

3. ***standard deviation***
   - "average deviation", how far variables are, on average, away from their mean

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

   - where, n-1 compensates for small sample sizes (n < 30) and higher probability of falling outside the SD
   - roughly, 68% of a sample will be within 1SD
     95% of a sample will be within 2SD of the mean

Probability

1. all probabilities are between 0 and 1.0, or 0% and 100%

2. ***independent events***
   - the probability of a given outcome remains the same, irrespective of any other outcomes (ie. successive rolls of a dice)
   - the probability of multiple events occurring is given by the ***multiplication rule***
   - this can be modified for events which are not necessarily independent,
     eg. drawing successive aces from a pack of cards (4/52 x 3/51)

     $\rightarrow$ Event A ***and*** Event B happening
   - in general, the multiplication rule for non-independent events is the product of the probabilities of the first event, and the second event given that the first has taken place

3. ***mutually exclusive events***
   - ie. the occurrence of one event precludes the occurrence of the other
   - the probability of ***at least*** one event happening is given by the ***addition rule***
   - this may be modified for non-mutually exclusive events (ie. compatible events),

     $\rightarrow$ Event A ***or*** Event B happening,
     minus the probability of ***both*** happening

4. ***binomial formula***
   - for a chance process, carried out in stages as a sequence of *n* trials
   - the probability *P* of a given number of outcomes
   - where, n = the number of trials, which must be fixed in advance
     k = the number of times the event of interest occurs
     p = the probability that the event will occur in any trial, which remains constant, $\therefore$ only independent events

$$P = \frac{n!}{k!(n-k)!} \cdot p^k (1-p)^{n-k}$$

5. ***standard error***
   - for a given series of events, the actual number of outcomes will vary from the predicted number of outcomes
   - the degree of variation from the predicted value is given by the ***standard error***
   - the method of calculation varies with the chance of the given outcome (see later)

Probability Distributions

■ ***Binomial Distribution***

- applies where there are 2 possible outcomes to an event, eg. heads/tails, boy/girl
- using the binomial formula above, the likelihood of a given number of events from a series with 2 outcomes, forms distribution curve with a mean at the probability of the given event
- eg., the probability of 1, 2,..., 10 boys from 10 consecutive births, forms binomial distribution with the highest probability at 5 boys (p ~ 0.25)
- the ***expected value*** is given by, $E = np$ (10 x 0.5 = 5 boys)
- however, the probability of exactly this number is only 25%

■ ***Poisson Distribution***

- classically applied where a ***large number*** of individuals are each at a ***small risk*** of a ***rare event***,

  eg. number of fatal road accidents in SA in a day

- generally describes events which occur ***independently*** and ***randomly*** in,

  1. time - at a fixed rate per unit time
     - probability of more than 1 event in a short interval is very small
  2. space - at a fixed density per unit area/volume
     - probability of more than 1 event in a small area is very small

- the chance of k events occuring is given by the formula,

$$P(k) = \frac{e^{-\mu}.\mu^{k}}{k!}$$

where e = 2.718
μ = expected value, usually known from ***empirical data***

■ ***Continuous Distributions***

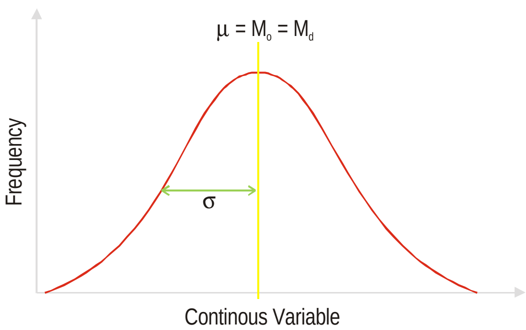
- in both binomial & Poisson distributions, the variable of interest in always a ***discrete*** variable
- variables which can assume an infinite set of values over a given range form continuous probability distributions, the most important being the ***normal*** or ***Gaussian distribution***
- these are represented graphically,

1. the variable of interest on the x-axis
2. the area under the curve represents the probability
3. the total area under the curve is 1.0 by definition
4. the curves are always symmetrical, or bell shaped
5. in a normal distribution, the ***mean, median & mode*** are identical

- with continuous variables, the probability of a given value, eg. height = 156.78 cm, is virtually zero and it is more meaningful to confine probability to a ***specified interval***
- the probability of an event within the specified interval is given by the area under the curve between the 2 specified points of the interval, eg height between 170-180 cm
- a continuous probability distribution may be,

1. symmetrical
2. skewed to the left / negatively skewed
3. skewed to the right / positively skewed

■ ***Normal Distribution***



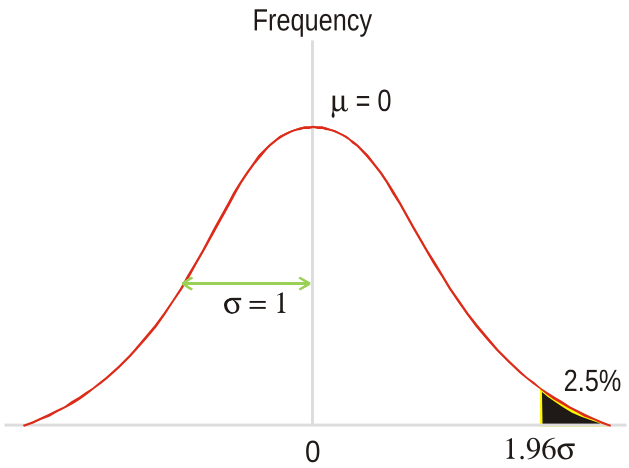
1. distribution of a ***continuous variable***
2. represented graphically as a bell-shaped curve
3. symmetrical about its mean, designated $\boldsymbol{\mu}$
4. the ***mean, median & mode*** are identical
5. described mathematically by 2 quantities,
   - i. the mean $\boldsymbol{\mu}$
   - ii. standard deviation $\boldsymbol{\sigma}$
6. the probability of an event lying within the limits,
   - i. $\boldsymbol{\mu} \pm \boldsymbol{\sigma}$ ~ 0.68
   - ii. $\boldsymbol{\mu} \pm 2\boldsymbol{\sigma}$ ~ 0.95 (NB: 95% actually 1.96σ)
   - iii. $\boldsymbol{\mu} \pm 3\boldsymbol{\sigma}$ ~ 0.99

NB: In practice the probability distributions of the variables we observe are unknown; however, if their distribution is ***bell-shaped*** and reasonably ***symmetrical*** about the mean, then the properties of a normal distribution can be applied in analysing probability

■ ***Standard Normal Distribution***

- an infinite number of curves are possible depending upon **μ** & **σ**
- instead calculations are done referring to a ***standard normal distribution*** which has,

1. a mean, $\mu = 0$
2. a standard deviation, $\sigma = 1$



- intervals are then calculated in multiples of **σ** from the mean, ie. **σ** becomes the unit of measure
- any value (x) is represented by the ***standard normal deviate, z*** given by,

$$z = \frac{x - \mu}{\sigma}$$

- in any normal distribution with mean **μ** and standard deviation **σ**, the probability that an observation will lie between x_1 and x_2 is the same as the probability that a standard normal deviate lies between z_1 and z_2
- published tables calculate the area under the curve to the left of any value of z
- from these areas between any 2 values can be calculated, or the area to the right

Populations & Samples

Def'n: ***population*** refers to any collection of people, objects, events, or observations; this is usually too large & cumbersome to study, so investigation is usually restricted to one or more ***samples*** drawn from the study population

- however, to allow true inferences about the study population from a sample there are a number of conditions,

1. the study population must be clearly defined, even if it cannot be enumerated
2. every individual in the population must have an equal chance of being included in the sample, ie. there should be a ***random sample***
    - random does not refer to the sample, but the manner in which it was selected
    - the opposite of random sampling is ***purposive sampling***, ie. every 2nd patient

NB: studying samples cf. populations results in loss of ***precision***

■ ***Statistics & Parameters***

Def'n: a ***statistic*** is an index descriptive of a ***sample***
a ***parameter*** is an index descriptive of a ***population***

■ *Sampling Errors*

1. sampling errors
   - the smaller the sample ***size***, the greater the error
   - the greater the ***variability*** of the observations, the greater the error
2. non-sampling errors
   - these do not necessarily decrease as the sample size increases
   - result in ***bias***,or systematic distortion of the results

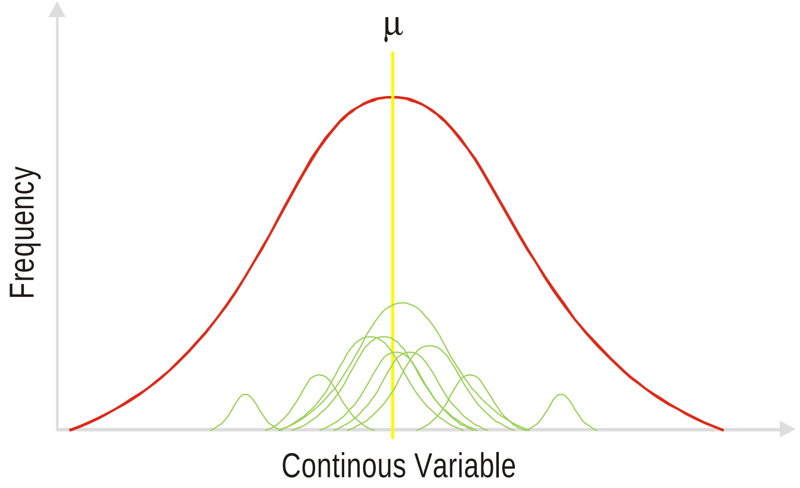
• if a large number of samples are drawn from a population, each sample with its own mean, then these sample means will,

1. tend to be distributed normally, even if the population distribution is markedly non-normal
2. form a normal distribution, with a mean equal to the true population mean, **μ**
3. have a standard deviation, the ***standard error of the mean***, given by,

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

   where **σ** is the standard deviation of x in the ***population***

4. as the sample means are normally distributed, then the means of 95% of the samples will lie within the range: $\mu \pm 1.96\ (\sigma/\sqrt{n})$



• usually only one sample is taken from the population, of size *n* and mean *x′*
• the accuracy with which *x′* predicts the true population mean, **μ** is determined by,

1. the sample size - as n increases, the SE(x) decreases
2. the standard deviation of the population values, or the variability of x, as **σ** increases, so SE(x) increases

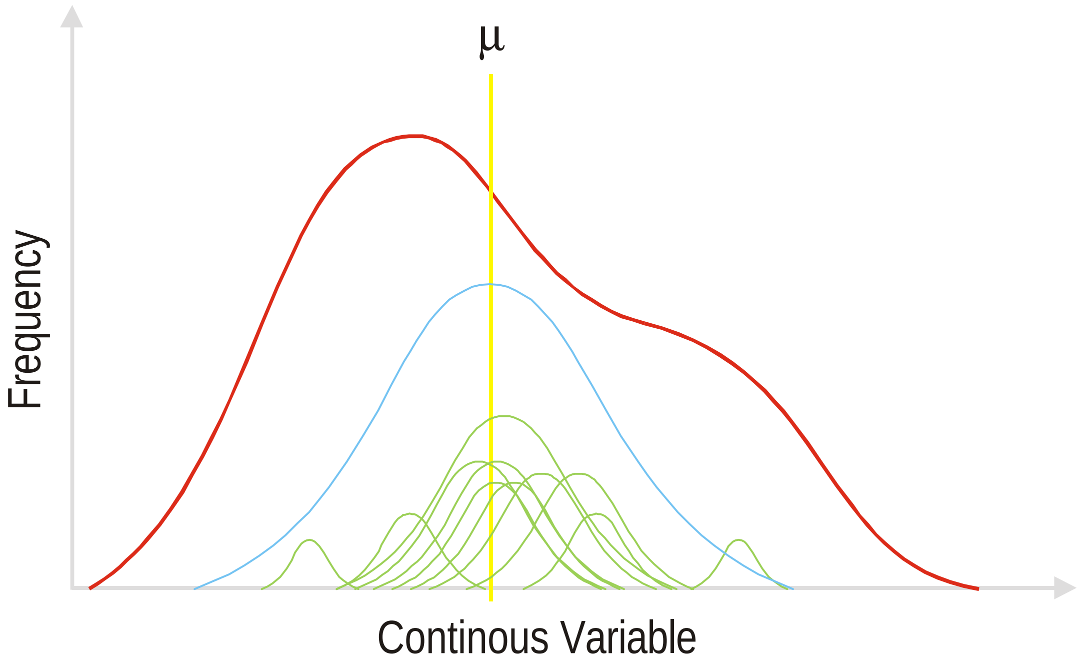
■ ***Confidence Intervals & Limits***

- in real life population means are unknown and have to be estimated from sample means
- from one sample of size *n* there is a 95% chance of getting the sample mean within the interval,

$$\boldsymbol{\mu} \pm 1.96\ \mathrm{SE}(x) \qquad \mathrm{SE}(x) = (\boldsymbol{\sigma}/\sqrt{n})$$

- this is termed the ***95% confidence interval*** for **μ**
- the upper and lower values the ***95% confidence limits***
- if greater certainty is required, a larger standard normal deviate is chosen $\pm$ 2.58 SE(x) > 99%

NB: the "normal" distribution of sample means remains even for distributions which are themselves non-normal



Tests of Signifcance

Def'n: ***signifcance*** refers to the likelihood of an observed outcome being due to chance
the null hypothesis is that the observed difference reflects chance variation

the alternative hypothesis is that the observed difference represents a real deviation from the population mean, due to some additional factor

Test Statistics

• a statistic derived from ***sample data***, used to measure the difference between the observed data and what would be expected under the null hypothesis,

i. z-statistics - apply the principal of the standard normal deviate
ii. t-statistics - small samples, with limited degrees of freedom
iii. χ^2-statistics - categorical or ***qualitative*** variables

Significance Levels

• usually denoted by the letter ***p***, and represents the probability of the observed value being due solely to chance variation
• the smaller the value of ***p*** the less likely the variation is to be due to chance and the stronger the evidence for rejecting the null hypothesis
• most scientific work, by ***accepted convention***, rejects the null hypothesis at $p < 0.05$
• this means that we shall reject the null hypothesis on 5% of occasions, when it is in fact true, ie. there was simply a chance variation
• the level of probability accepted will depend on the inherent variability of the variable being measured, as well as the nature of the alternative hypothesis

One & Two Sided Tests

• the null hypothesis is that there is no significant difference, and chance has occurred, it makes no assumption about the direction of change or variation
• the alternative hypothesis states that the difference is real, further that it is due to some specific factor,

1. where no direction of change is specified, ie. that there is simply a difference between the population mean and the observed data, both ends of the distribution curve are important, and the test of significance is ***two-sided***, or two-tailed
2. where the direction is specified, then only one tail of the curve is relevant, and the test of significance is ***one-sided***, or one-tailed

• the ***critical value*** is the value of a test statistic at which we decide to either accept or reject the null hypothesis
• the critical value for a one-sided test at significance p, will be equivalent to that for a two-sided test at $2p$ (one-sided $p = 0.025$ / two-sided $p = 0.05$)
• thus, it is tempting to use one-sided tests as the significance is greater, but the decision should be made ***before*** the data is collected, not after the direction of change is observed

The t-distribution & Degrees of Freedom

- the ***z-test*** requires that,
    1. the sample size is ***large*** ($n > 30$)
    2. the population standard deviation, **σ** is known
    3. the variable could be assumed to be normally distributed in the population

NB: commonly the population σ is unknown,
however it is possible to use the sample standard deviation, ***s*** as an estimate of **σ**

the type of test to be used then depends upon the ***sample size***

■ ***Large Samples***

- if the sample size is large, $n > 30$, then the sample standard deviation, ***s*** is considered to be an adequate estimate of the population **σ**
- thus, the standard error of the sample mean becomes,

$$SE(x) = (\boldsymbol{s}/\sqrt{n})$$

- under these circumstances the ***z-test*** can again be used to test the significance of the difference between the population mean **μ** and the sample mean ***x'***
- this assumes that the population fits a normal distribution

■ ***Small Samples***

- if the sample size is small, $n < 30$, then the sample standard deviation, ***s*** is ***not*** considered to be an adequate estimate of the population **σ**
- to test small samples, ***Student's t-test*** is employed (actually Gosset in 1908)
- the ***t-distribution*** actually describes a series of curves,
    1. dependent upon the number of ***degrees of freedom*** (ν - nu)
        - as opposed to standard deviation, and is integrally related to the sample size
    2. as for the normal distribution, these are symmetrical with a mean **μ = 0**

- degrees of freedom refers to the number of observations completely free to vary
- often conditions exist which restrict freedom and the **ν** is less than the actual sample size
- the corresponding tables are in effect ***more conservative*** in rejecting the null hypothesis, due to the added uncertainty of using ***s*** as the population standard deviation, where,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where, there are ***n-1*** degrees of freedom

NB: the corresponding ***critical value*** for the ***t-test***,

at the ***p = 0.05*** level of significance, is 2.26, cf. 1.96 for the z-test
at the ***p = 0.01*** level of significance, is 3.25, cf. 2.58 for the z-test

- tables for t-tests vary from z-tables,

1. the listed area (α) is to the ***right*** of the t-statistic and is ***one-sided***
   - therefore for two-sided tests the values need to be doubled
2. the values for α are listed under verious values of ν
3. as for z-tests, if the value of the t-statistic exceeds the critical value, then the null hypothesis may be rejected

■ ***Confidence Limits***

- as for significance tests, when the population σ is unknown, the sample ***s*** may be used as an approximation
- if sample sizes are large (n > 30), then ***s*** is an adequate estimate
- if sample sizes are small, then instead of using the standard normal deviate to estimate the confidence interval, the t-distribution at (*n-1*) degrees of freedom is used

Comparison of Sample Means

Def'n: ***variance*** is the standard deviation squared, and is a measure of ***dispersion***,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

where, **s** = sample standard deviation
$\mathbf{s^2}$ = sample variance
σ = population standard deviation
σ^2 = population variance

- this is useful for comparing sample means from two different populations ($x_1' - x_2'$)
- ie. is any difference real, ie. due to a difference $\mu_1 - \mu_2$, or due simply to chance
- if multiple samples are taken from the 2 populations, then the differences between the means ($x_1' - x_2'$) will also form a normal distribution, with a mean about the true difference between the two population means ($\mu_1 - \mu_2$)
- analagously, the ***standard error*** of the ***difference of two means*** is given by,

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Tests of Significance

NB: these formulae are generated algebraically by the assumtpion of the null hypothesis that there is no difference between the means, ie. $\boldsymbol{\mu_1 - \mu_2 = 0}$

■ ***Population Standard Deviations Known***

- require calculation of the z-statistic, and use of the normal distribution tables
- where, z is given by,

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

■ ***Population Standard Deviations Unknown Sample > 30***

- accept calculation of the z-statistic, with use of the sample ***s*** as the population **σ**
- where, z is given by,

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\mathbf{s}_1^2}{n_1} + \frac{\mathbf{s}_2^2}{n_2}}}$$

■ ***Population Standard Deviations Unknown Sample < 30***

- require calculation of the t-statistic, which relies on the weighted average of s_1^2 and s_2^2
- such that t is given by,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\mathbf{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- if it cannot be assumed that the population variances σ_1^2 and σ_2^2 are ***equal***, then the t-test cannot be used
- the equality of population variances, as estimated by sample variances, can be measured by the ***variance ratio test***, with reference to the continuous probability ***F-distribution***

NB: if the t-test cannot be used, then a statistical test not dependent on any underlying ***probability distribution*** is used,
tests of this type being termed, ***distribution-free*** or ***non-parametric***

The Chi-Squared Test

■ ***Contingency Tables***

- used when investigations concern ***categorical*** or ***qualitative*** variables
- the t-test can in fact be used to compare 2 categories
- the advantage of the χ^2 test is that it allow comparison of many more categories, drawn-up into a ***contingency table***
- the ***null hypothesis*** is that any number of categories have equal chance of any other factor
- from this, tables of ***expected frequencies*** can be calculated by cross-multiplication
- the test then concernes itself with wheather the gap between ***observed & expected*** frequencies is too large to have airsen simply by chance
- the χ^2***-statistic*** is given by,

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- like the t-distribution, the χ^2 distribution is actually a family of distributions, depending upon the number of differences involved
- generally, the number of ***degrees of freedom*** for such a table is given by,

$$\nu = (r - 1).(c - 1)$$

- ie., for a 2x2 table there is 1 degree of freedom
- calculation of the ***expected frequency*** for each cell of a table is given by,

E = (row total).(column total) / overall total

Expected Frequencies			
Table	**A**	**B**	
C	$(x_C \times y_A)/n$	$(x_C \times y_B)/n$	x_C
D	$(x_D \times y_A)/n$	$(x_D \times y_B)/n$	x_D
	y_A	y_B	Total = n

■ ***Goodness of Fit***

- apart from using contingency tables, the χ^2-statistic can be used to see if an observed set of observations follows a particular distribution
- eg. by calculating the expected frequencies from say a Poisson distribution, and then comparing these with the observed data
- alternatively the distribution may be a genetic model
- the degrees of freedom will be the number of observations, ***n - 1***

- ***General Features***

1. the formula: $$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

   is valid only for comparing observed and expected ***frequencies***,
   it cannot be used for comparing percentages or proportions directly,
   nor can it be used for derived statistics, eg. means, rates etc.

2. when dealing with a continuous variable, the range must be divided into suitable ***intervals***, and the observed & expected frequencies in each interval compared

3. the χ^2 should ***not*** be calculated when the expected frequency in a cell is **< 5**

4. χ^2 is actually a probability distribution, whereas observed frequencies are discrete, when frequencies are small a ***continuity correction*** for **ν = 1** should be added,

$$\chi^2_{(1)} = \Sigma \frac{(|O-E| - 0.5)^2}{E}$$

   this correction is of little consequence unless the frequencies are ***small***

- ***Other Non-Parametric Tests***

1. Wilcoxon's Rank Sum Test
   i. paired data - the signed rank test
      - two groups matched for other confounding factors prior to treatment
      - the differences between the pairs is calculated, then ranked in order
      - 2 pairs having the same difference are given the mean of what would have been their successive ranks (ie. 2 & 3 $\rightarrow$ 2.5 & 2.5)
      - these ranks are then given the ***sign*** of the actual difference between the pairs
      - each of the (+) & (-) ranks is totalled, & the smaller referred to a table

   ii. unpaired data - the two-sample test
      - all results are ranked in order, but the two groups distinguished
      - the ranks for the two samples are then added seperately & the smaller total used
      - requires greater numbers to produce a significant result cf. pairing`

2. Mann-Whitney U Test
   - similar aproach to (1) and entirely comparable results

Correlation & Regression

■ ***The Scatter Diagram***

• this is used to study 2 variables which are ***quantitative***, cf. qualitative variables using χ^2
• classically these are represented by a series of dots, each representing a ***pair*** of data, ie. one x and one y coordinate for the two variables being studied
• the resulting graph is termed a ***scatter diagram***
• in studies of relationships of 2 variables, usually one is labelled the ***independent*** variable (x-axis), and the other the ***dependent*** variable (y-axis)

■ ***Correlation Coefficient***

• for any cluster of points, the following ***summary statistics*** are readily calculated,

1. mean & standard deviation of x values x' and s_x
2. mean & standard deviation of y values y' and s_y

• however, these tell nothing about the association between the two variables
• to assess this the ***correlation coefficient (r)*** is calculated,

$$r = \frac{\sum (x-\bar{x}).(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \times \sum (y-\bar{y})^2}}$$

■ ***General Features***

1. correlations are always between -1 and +1
   - a positive correlation means as one value increases, so does the other
   - a negative correlation means as one value increases, the other decreases
   - a value of zero means there is no linear correlation
2. ***r*** = ±1 is referred to as a ***perfect correlation***,
   in that there is a perfect ***linear*** relationship between variables
3. ***r*** does not always give a true indication of clustering,
   the two main exceptions are,
   i. non-linearity - ***r*** only measures linear association
   ii. outliers - result in a dramatic approach of ***r*** $\rightarrow 0$
   - these should only be excluded for sound reasons
4. important to remember, ***correlation does not equal causation***
   - further there may be an indirect association, or a confounding factor

5. the following is a rough guide to the ***magnitude*** of ***r***,

    i. 0.8 - 1.0 strong

    ii. 0.5 - 0.8 moderate

    iii. 0.2 - 0.5 weak

    iv. 0.0 - 0.2 negligible

6. the ***variance*** of ***r***, given by $\boldsymbol{r}^2$ measures the dependence of one variable on the other

    - ie. if $r^2 = 0.72$, then 72% of the value of dependent variable (y) is due to the independent variable (x)

7. the ***significance of r***, in determining wheather the variation from zero is real or simply due to chance can be measured by the t-test, where,

$$t = \mathbf{r}\sqrt{\frac{n-2}{1-\mathbf{r}^2}}$$

and the table is entered at ***n - 2*** degrees of freedom

Regression

- correlation predicts whether one variable is related to another, but not ***how***
- regression constructs the ***line of best fit***, for a linear correlation, such that,

$$y = a + bx$$

- where ***b*** is the ***regression coefficient***, and describes the slope of the line
- this is achieved by looking at the difference between the observed & expected values,

$$d = y - (a + bx) \qquad \text{or,}$$

$$d^2 = [y - (a + bx)]^2$$

- the charteristics of the regression line is that the ***sum*** of values for ***d^2*** is a minimum
- this may be achieved by calculus, and is the ***least squares method***
- fortuitously it turns out that,

$$b = \frac{\sum (x-\bar{x}).(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

- this allows prediction of y for a given value of x, within the range of values of x, extrapolation from lines of regression is usually risky

▪ ***Multiple Regression***

- frequently multiple factors are implicated in disease processes
- simple one-to-one cause effect is rare
- some of these factors may be interrelated, others may be independent
- this involves constructing a ***multiple regression equation***,

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots \text{ etc.}$$

where, $b_1, b_2, b_3,..$ are the ***partial regression coefficients***
$x_1, x_2, x_3,..$ are the multiple factors being examined

- the dependent variable, y, would represent the diesease process in question
- each partial regression coefficient predicts the amount the dependent variable will change, for a given change in that factor
- this allows an estimate of the relative importance of multiple factors in disease causation

Significance Tests & Errors

• with a test of significance, with a value of $p < 0.05$, there is a 5% chance that the variation obsereved is in fact due to random variation
• rejection of the null hypothesis at this level means that we will, in 5% of cases reject the null hypothesis when it is actually true

Def'n: incorrectly ***rejecting*** the null hypothesis → ***type I error***, or ***a-error***

incorrectly ***accepting*** the null hypothesis → ***type II error***, or ***b-error***

• it is inherent in statistical concept, that attempts to quantify the probability that you are correct, carries with it the probability that you are wrong
• in general there are 2 factors which result in investigators failing to show a real difference, ie. where one actually exists,

1. chance alone
   - an unusual data sample which does not support a difference
   - this is ***type II error***
   - statistical methods can produce incorrect conclusions
2. too small a sample size
   - the smaller number of individuals included in a study, the greater must be the real difference before statistical difference may be shown
   - in extreme, no matter how small the real difference, this may be shown statistically if a large enough sample is drawn
   - the probability that a study can predict a difference, when a real difference actually exists, is termed the statistical ***power*** of the study
   - the higher the power of the study, the smaller the difference which may be detected

Statistical Test Errors		**Real Difference**	
		No	**Yes**
Significance Tests	No Effect Demonstrated	$1 - \alpha$	Type II Error **β-error**
	Effect Demonstrated	Type I Error **α-error**	Power $1 - \beta$

Def'n: by definition, if β is the probability of accepting the null hypothesis,
ie. saying that no real difference has occurred, where a real difference exists,
then the probability of finding a real difference, when one exists,
is given by, $\mathbf{1 - \beta}$ = the ***power*** of the study

■ ***How Many Subjects***

• the standard error of estimation of a population mean, from a sample mean,

$$SE(\overline{x}) = \frac{\sigma}{\sqrt{n}}$$

• thus, as the number in the sample *n* increases, the error decreases
• if we choose a ***95% confidence interval***, then then true value is given by,

$$\overline{x} \pm 1.96 \times SE(\overline{x})$$

• if we then state the desired mean difference we wish to demonstrate, this can be rearranged to give the value of *n*
• similar derivations may be made for comparisons between two sample means etc.
• what is required for these calcuations is,

1. a designated ***power*** of the study - how accurate do we require it to be ?
   - what chance of success do we want ?
2. the amount of difference we wish to show, and
3. at what level of ***significance*** we wish to demonstrate this

• when comparing two samples, the minimum total sample size is achieved when $n_1 = n_2$

Prediction Models

Event (eg Death) Continency Table		Prediction by Test: No	Prediction by Test: Yes
True Occurrence "Gold Standard"	**No**	TN	FP
	Yes	FN	TP

Def'n: ***Sensitivity*** — of those who ***actually*** die, how many did the test predict

= TP / (TP + FN)

Specificity — of those who ***did not*** die, how many did the test predict

= TN / (TN + FP)

Predictive Value — of those ***predicted*** to die, how many actually died

= TP / (TP + FP)

Discrimination — overall ***correct classification rate***, ie. how well the model separates those who will & will not die,

= (TP + TN)/n

False Classification Rate = (FP + FN) / n
= 1 - Discrimination

- ***Example*** ***Papazian AJRCCM 1995***

BAL VAP Continency Table		BAL[1]: No	BAL[1]: Yes
Post-Mortem "Gold Standard"	**No**	95[2]	5
	Yes	45	552

[1] Predictive value ~ 90% Discrimination ~ 75%

[2] Specificity ~ 95% Sensitivity ~ 55%