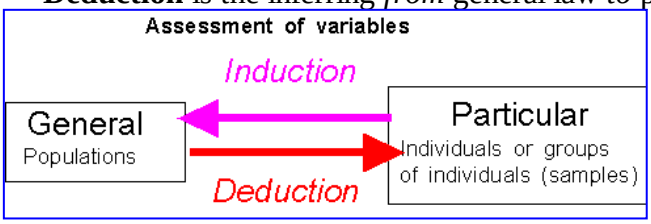
FACTS, FIGURES AND STATISTICS
"Lies, damn lies, and statistics" ***Mark Twain***
"But some damned lies are the truth" PDW
"When in doubt tell the truth" ***also Mark Twain***

Statistics is the systematic selection, presentation and assessment of numerically based data. When information available is less than complete statistics can minimize, but never completely exclude, the risk of erroneous conclusions. Skepticism is important. Statistics has practical limitations: the chances of your existing and reading this book are *almost* zero, but has turned out to be reality: almost incredibly each and every one of your ancestors (starting with unicellular organisms) was fertile!

Assessment in its simplest forms utilizes two main methods.

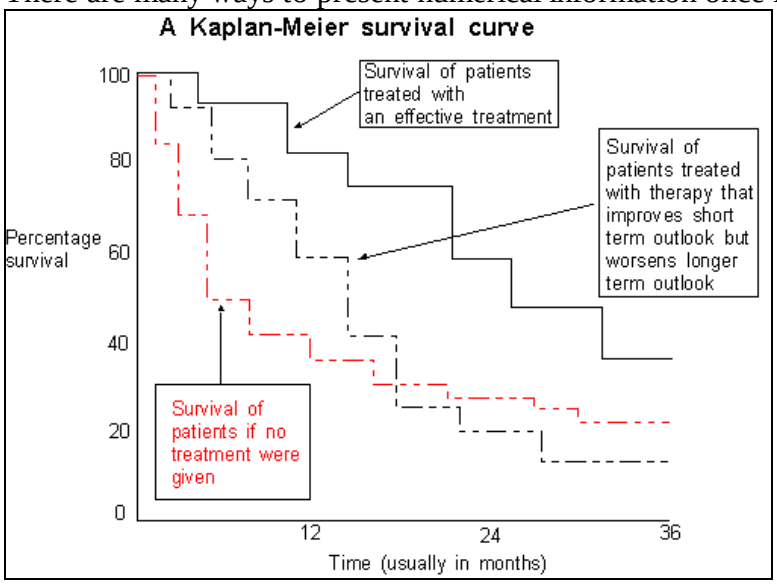
- **Induction** is the inferring of general law *from* particular instances whereas
- **Deduction** is the inferring *from* general law *to* particular instances



People often generalize (a statement which provides evidence of its own existence) without thinking and such assessments may be wrong. "I had a patient like this who responded to X, therefore it should work for other patients" - inductive wishful thinking - or "(I think) my impression is that most of my patients have responded to Y therefore this patient will" is deductive wishful thinking. Inductive wishful thinking is easy to counter but deductive wishful thinking often requires effort to collect data to confirm or deny the claimed associations.

Results from a whole population are called parameters. However it is not necessary to survey whole populations before it is possible to make draw reasonable conclusions and interpretations from a sample (for example mean, mode, median, range, standard deviation) is the business of statistics.

There are many ways to present numerical information once it has been collected.



Data Presentation

Raw data as collected

86	81	79	77	94
73	65	68	83	77
78	87	75	71	72
62	74	89	90	80
81	79	82	80	85

Data in recharted in ascending order

62	65	68	71	72
73	74	75	77	77
78	79	79	80	80
81	81	82	83	85
86	87	89	90	94

= the *median* which splits the number of observations in half (if there are an even number of observations the median is half way between the two middle observations)

= the minimum value observation

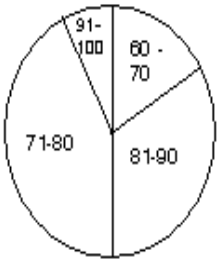
= the maximum value observation

The maximum minus the minimum gives the *range*

Grouped data

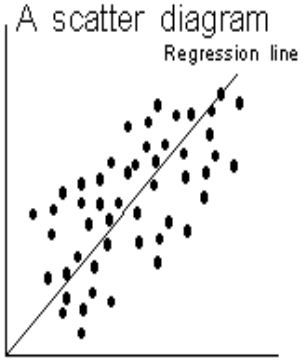
Grouped data	Numbers in each group	%
60-70	3	12
71-80	12	48
81-90	9	36
91-100	1	4

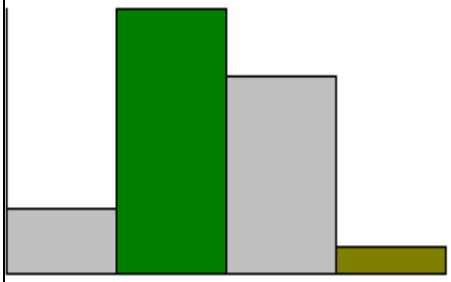
Pie charts



Pie charts are usually more useful for displaying categorical data

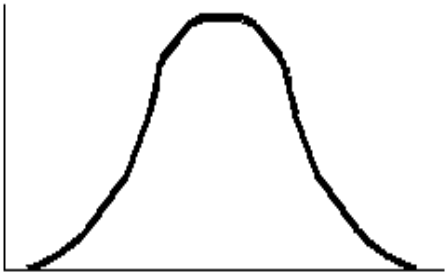
A scatter diagram





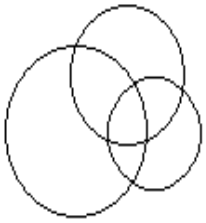
Charting of grouped data as a histogram

Number of observations of each number



Numerical values of observations

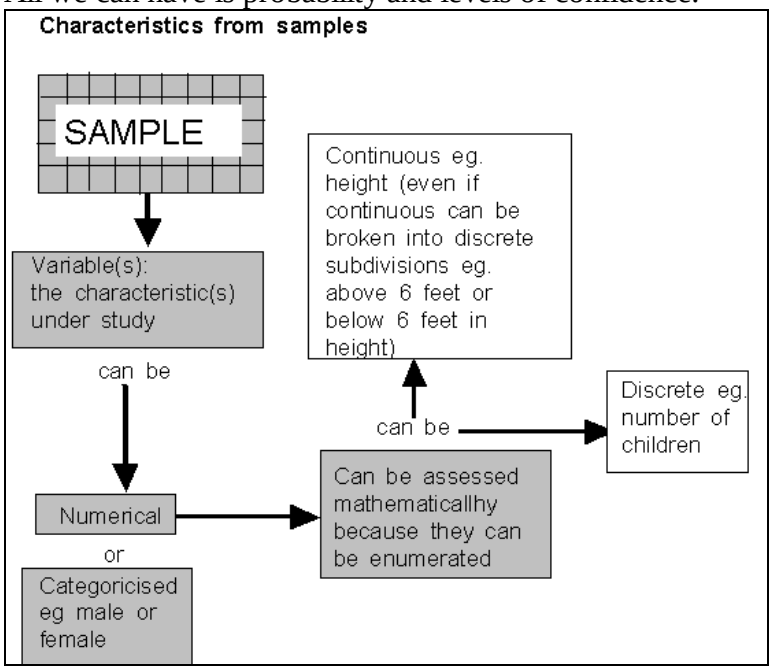
This curve shows the shape of the distribution of the observations. In many biological observations, if the number of observations is large enough, and there is an even distribution, then this "Standard Distribution curve results



A Venn diagram

Samples

Results obtained from samples can *never* guarantee to represent the whole population. All we can have is probability and levels of confidence.



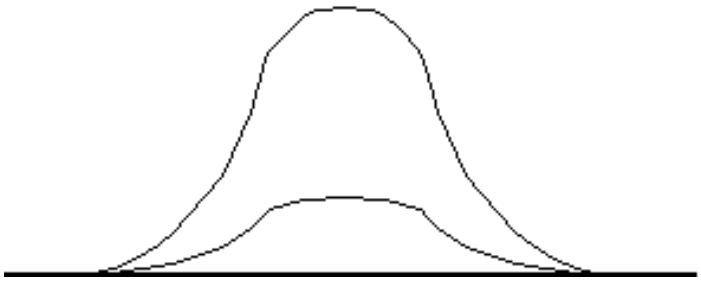
Obviously samples should be representative of the population about which a generalization is sought. Vaguely selected samples, especially those defined in retrospect - "the patients I have seen" - are unlikely to be representative. Samples are more likely to be representative if a researcher selects his sample randomly from the population under study. Randomly does not mean haphazard - haphazard often means that the bias of selection is unrecognized and unacknowledged. Lists of random numbers are available to assist this process but if the same sequence of random numbers is used repeatedly then bias might be introduced.

Random allocation of an intervention should ensure that the intervention and non-intervention groups are similar because three should be no allocation bias. Observer bias should be abolished by having a double blind trial in which the observer does not, and cannot, suspect whether a patient has or has not received the intervention and then a difference in results must be attributable, directly or indirectly, to the intervention.

There is more chance of a sample being representative if:

- A large sample of the population under study is used
- There is an automatic "mechanical" selection process
- Stratification is used if necessary (the population studied might comprise hidden subgroups, membership of which might influence the outcome). Subgroups of the population might give different replies because they are members of a subgroup and this, if unrecognized, may invalidate results. For example failure to stratify a sample according to whether those sampled were male or female might lead to inappropriate generalizations for behavior of the whole population.

One population curve might hide another subpopulation



Samples are surveyed for the presence of the characteristic(s) under study - the variables. There are two types of variable. *Numerical* variables (each variable has an intrinsic numerical value, weight for example) or *category* variable (each variable either is or is not a something, male or female for example).

Definitions
The mean, colloquially "the average," is the most commonly used statistic in everyday conversation. It is an intuitively reassuring, although possibly fallible, measure of what is "normal." The mean is the total of numerical variables divided by the number of such variables. Means tend to remain stable over successive samples

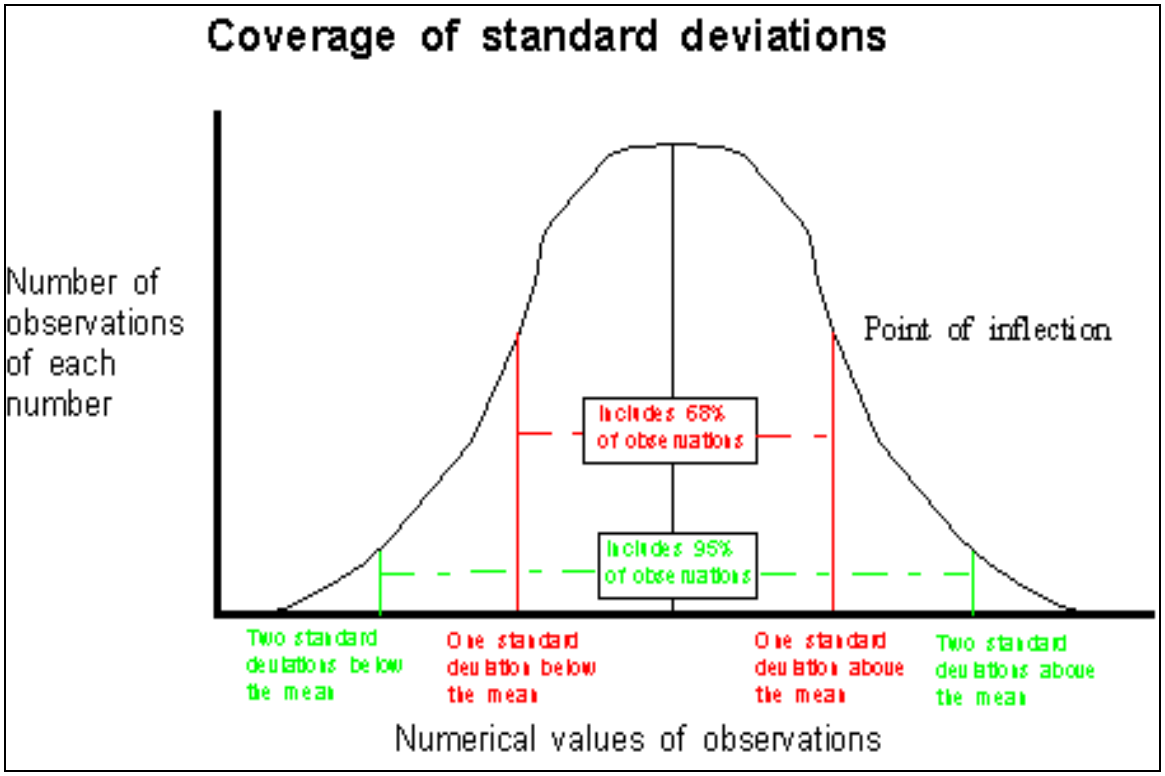
The median is the middle observation in a series (if there is an even number of observations the median is taken to be halfway between the two middle observations)

The mode is the value of the most frequent observation (usually used in assessment of category variables)

The range is the numerical difference between the numerical value of the highest and lowest observation

The mean tells us little of the distribution of the numerical results (the means of 2,2,2,2,2,2 and 8,8,8,8,8,8 is 5 as is the mean of 4,4,4,4,4,4 and 6,6,6,6,6). A measure of scatter, "dispersion," or variability of the numerical variables is required. One commonly used measure is the Standard Deviation.

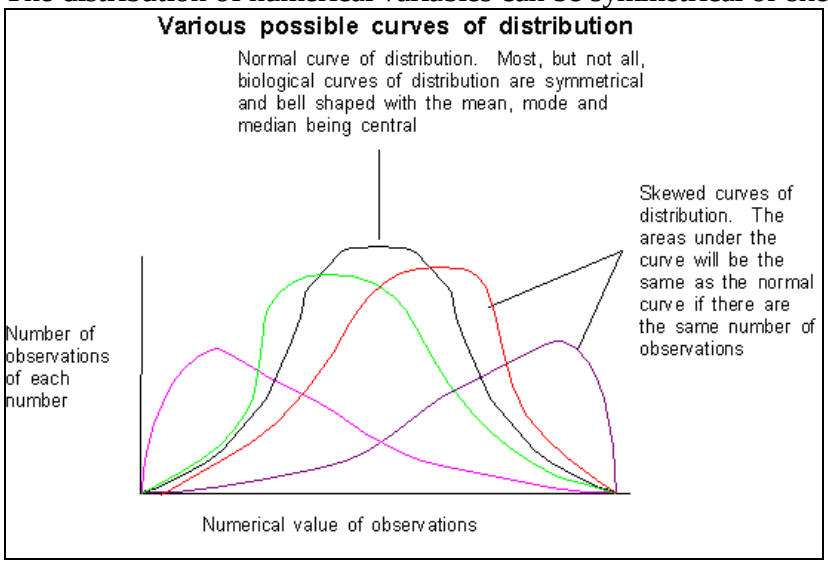
COMMON MISAPPREHENSIONS
The mean "average" = the median. Consider 1,2,3,4,7,8,10. The mean is 35/7 =5 yet the median is 4 (this constitutes a skewed distribution)
The mode = the mean "average" Consider 1,222,3,7,11. The mode is 2 and the mean is 4
It is meaningful to calculate means or median for category variables. Is the mean member of the population 56 percent female?



A useful gauge of scatter would be the average distance from the average. However to confuse matters statisticians count differences above the mean as positive and below as negative and thus if the differences each side of the mean are added the result is zero. To avoid this difficulty statisticians square each difference from the mean (a negative squared then becomes positive) and statisticians then take the square root of the sum of the squares which is designated as the Standard Deviation*). This is not the same as average distance from the average.*

Observations outwith two standard deviations either side of the mean (equivalent to 5 percent of observations) are unusual and are often regarded, somewhat arbitrarily, as abnormal and worthy of comment and/or investigation.

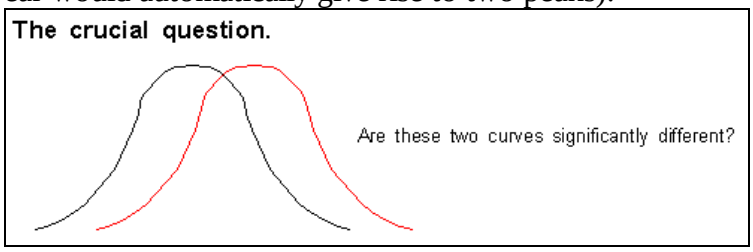
The distribution of numerical variables can be symmetrical or skewed.



Most but not all biological curves of distribution are of normal distribution (being symmetrical and bell shaped) with the mean being central. With more observations the curves of distribution become smoother. Any one value can be rated as "so many standard deviations from the mean." The distance between the mean and the point of inflection "bending back" of a normal curve is one standard deviation from the mean (see Standard Deviation diagram. Sixty-eight percent (roughly two thirds) of observations in a normal curve of distribution lie within the two standard deviations straddling the mean.

Standard deviations above (or below) the mean can be used to compare single results from two sets of observations "Jim's anatomy exam result was 1.25 standard deviations above the mean whereas in physiology he scored 1.1 standard deviations below the mean. This tells us that Jim was better than his peers at anatomy and worse in *physiology even though he may have scored higher marks in physiology than anatomy* (as would occur if most candidates had high marks in physiology but only a few had high marks in anatomy").

When there are two or more separate peaks well away from the mean the standard deviation (which after all is a single numerical value) alone will not reveal the two-peaked shape of the curve. If there are two or more peaks in curves obtained from biological observations suspect two subgroups in the sample or a quirk in data collection (for example measuring the distance of the right and left eye from the right ear would automatically give rise to two peaks).



Even if the curve of distribution of the population is not normal (is skewed) the means of several different samples from the same population can be obtained and the curve of distribution of these means will be a normal curve and the standard deviation of the samples from their mean can be obtained. Then 68 percent (one standard deviation each side of the mean) of the sample**s** will lie within one standard deviation each side of the sample**s** mean. This is known as the standard error. The importance of this is that a sample which yield a mean outwith this value become more likely to be unreliable. The extent of a standard error is mostly related to the proportion of the population covered by the samples, variability of the population, or size of the samples

The smaller the standard error of a particular sample, the more likely it is that a particular sample mean is close to the population mean. There is a 68 percent probability that a particular sample mean plus or minus one standard error contains the population mean (*also a 32 percent chance that it does not*). This is the 68 percent *confidence interval*. A confidence interval of 95 percent covers (approximately) plus or minus two standard errors and is the same as saying that there is less than one chance in 20 that the results are caused by chance. A 99 percent confidence interval covers (approximately) plus or minus two and half standard errors and is the same as saying that there is less than one chance in a hundred that the results are caused by

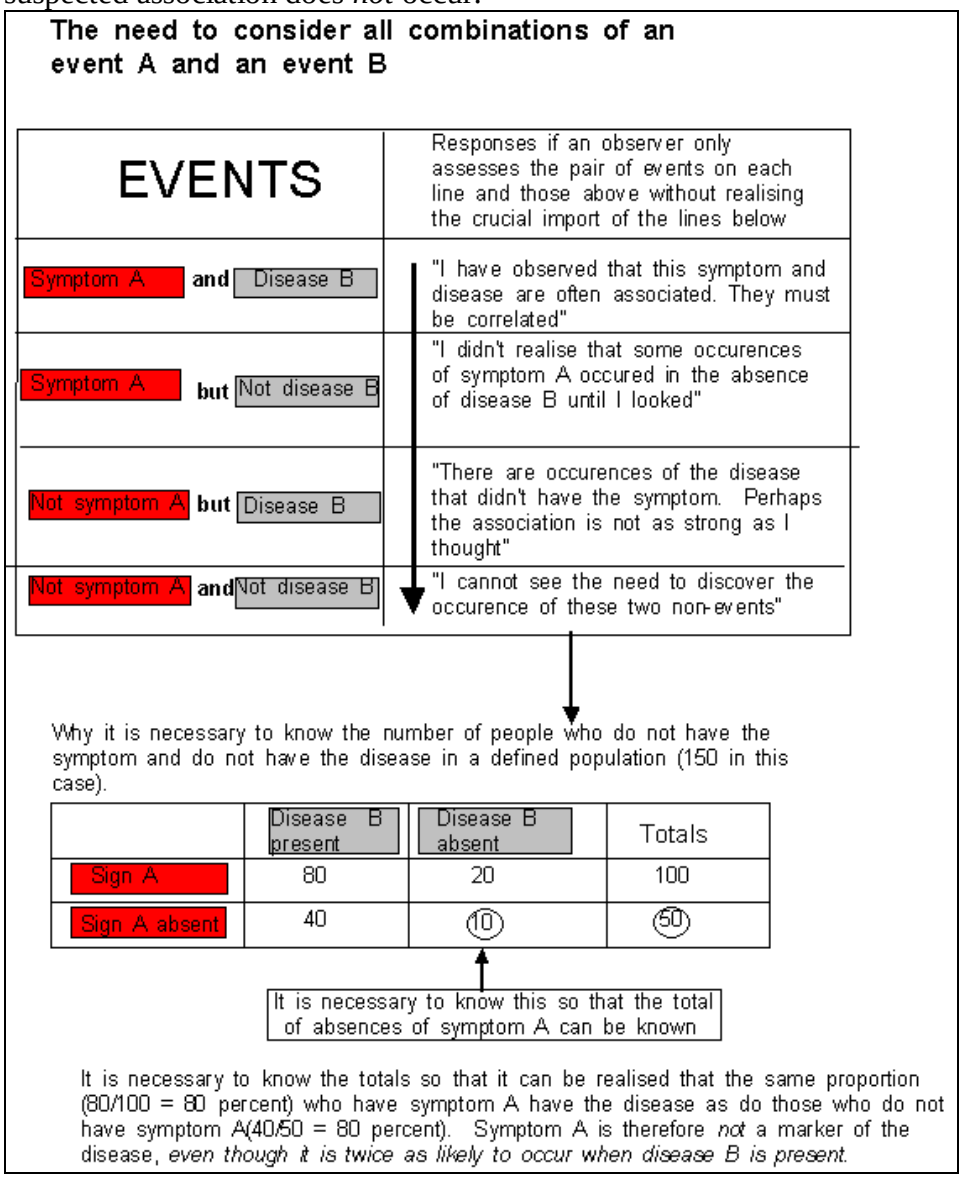
chance. The larger the difference in standard error between two samples then the more it is likely that the difference between the two samples is significant.

When comparing two sets of results the null hypothesis (that there is no difference between two sets of results) has to be proved wrong. There are two types of error:

- Type 1 error is accepting a difference as significant when it isn't (falsely rejecting the null hypothesis)
- Type 2 error is failing to accept a difference as significant when it is significant (incorrectly accepting the null hypothesis)

Associations

Not all associations are causal. When assessing associations it is important to be aware of the influence of base rates and the relevance of the number of cases in which the suspected association does *not* occur.



The P value (the derivation of which will not be explained here) expresses how likely a difference in results between two arms of a trial would have been likely to have arisen

by chance. P=<0.05 is arbitrarily taken to be statistically significant (the observed result could have arisen by chance in less than one in twenty similar trials). P=<0.01 is arbitrarily taken to be highly statistically significant (the observed result could have arisen by chance in less than one in a hundred similar trials). A statistically significant result may have no clinical relevance and non-significant P values suggest that there is no difference between groups assessed or that there were too few subjects for conclusions to be drawn. Worse, a P value of <0.05 means that one trial in 20 will report a spurious statistically significant result. Do a trial assessing 100 putative risk factors and 5 will be positive at P=0.05. Do a few more than 19 trials and the chance is that all risk factors will be shown to be statistically significant at least once.

Controversially the plausibility of any statistically significant result should be ascertained before any assessment of meaningfulness. The problem is that what you think is plausible, I may think is nonsense. For example my observations of Scotland "Who's Who" reveals that people whose surname has an "a" as the second letter are statistically significantly more mentioned in "Who's Who." This is fine but the explanation is that MacSomebodies are disproportionately represented in any analysis using names.

A causative association between a disease and a putative cause is likely if:

- The overlap of disease distribution and the putative cause is large
- Altering the putative cause affects the disease
- There are a large number of observations
- Other possible risk factors appear to be minimal
- Bias of the observers is absent or minimal
- There are no obvious confounding factors e.g. Smoking in the association between alcohol and lung cancer
- There is a dose response relationship between the putative cause and the disease
- Removal or reduction of the putative cause results in a reduction of the disease
- The geographical or other distribution of the disease varies with the geographical distribution of the putative cause
- The association is a constant finding in several studies (unconfirmed studies, even if yielding statistically significant results, might well require verification in other populations
- A similar population can be identified who have differing patterns of the disease and observing that the same putative cause is present and that the incidence of the disease is similar
- Laboratory evidence supports the hypothesis that the association is causal
- The association is statistically unlikely to be caused by chance
- Exposure to the putative cause preceded the disease
- There is no other explanation

Having said all this, a cause may not be single or may contain within itself two or more factors.

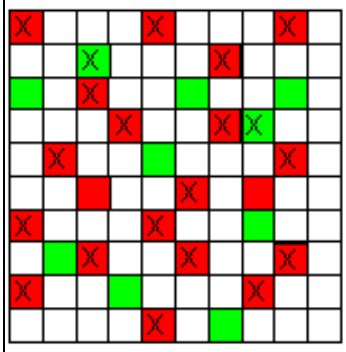
A BRIEF DIGRESSION CONCERNING LOGIC AND MEDICINE

Most clinical practice involves decisions which are based on the ability to recognize significant associations and thereby predict occurrence of disease. We are bad at this (intuitive thinkers, such as myself, are worse still because unless we are careful we allow imagination to run ahead of logic). Some examples.

Rain is forecast and forecasts are 80 percent accurate. If you go out for an hour should you take an umbrella? You should not even attempt to answer this question because you do not have the complete information on which to make an assessment. The answer is not 80 percent of the time but is probably about 30 percent. Why?

The crucial realization is that the forecast does *not* say it will rain 80 percent of the time and what is important (but not included in the information) is the likelihood of it raining *at all* for the time you are out. This is the base rate. In the UK if you go out for one hour there is a one in ten chance that it will be raining (the base rate) and a nine in ten chance that it will not be raining. In 100 trips out, each of one hour, it will rain on 10 trips (but only eight of these will have been predicted) and it will be fine on 90 trips but rain would have been predicted on 18 of these trips (the 20 percent inaccurate forecast). You will have to carry your umbrella on 10 + 18 = 28 trips (about 30 percent) to avoid rain on 8 occasions (plus you will receive an unpredicted soaking on two occasions).

The importance of base rates in evaluating forecasting success rates



You have to go out 100 times, each time for one hour. How many times should you take your umbrella if there is an 80% success rate in forecasting rain? In 100 hours there will be 10 rainy hours. Eight of the 10 will have been predicted ( ) because of the 80 percent success in forecasting but two ( ) will have not been forecast. Rain will have been incorrectly forecast in 18 ( ) of the remaining 90 hours (because of the 20 percent *failure* rate in forecasting). Therefore an 80 percent success rate in forecasting suggest that you will have to take your umbrella out in + = 8 +18 = 26 occasions if you follow the forecasters' predictions. You will be rained on 10 times.

The above was a non-medical example as an introduction to medical problems.

There is a test for a disease that affects one in a thousand people. The false positive rate is low, only five percent. What is the chance that someone who tests positive will have the disease? The answer is one in fifty (Fig 00). If a disease occurs rarely then a test has to be highly specific (identifying accurately without false positives or false negatives). The sensitivity of a test is a measure of how many people with the disease will be missed. If a test is highly sensitive there will be no missed disease but there might well be numerous false positives.

There are several sources for error in assessment:

- Availability error. The most memorable occurrence affects your judgment. During the era of dramatic jet highjacking deterred air travelers at a time when air travel was still safer than any other form of travel
- Primacy error. Assessment of later material is affected by prior exposure "First come, first served."
- Halo error. Readily available good points outweigh less readily available counterbalancing bad points
- Unrepresentative error. "I had a patient once who..... and therefore I now....." Make sure that you have sufficient numbers of significant observations. Under some circumstances one observation may be sufficient but such circumstances are rare
- Coincidence error. Coincidences are bound to happen as a purely statistical phenomenon. Striking they may be but significant they may not be. Even random numbers exhibit totally meaningless runs
- Total logic failure error. "A high proportion of patients given AZT for AIDS die. Therefore I will not take AZT" I have heard this from patients far too often
- Conformity error. Make up your own mind and never accept anyone else's diagnosis uncritically
- Ignoring evidence error. Just one result may be sufficient to discredit a theory. Take care not to dismiss inconvenient and irritating results
- Concentration error. Restricting attention to positive findings
- Superficial logic error. If an association seems logical that is not proof of an association
- Measuring the wrong parameter error. Doctors measure the blood levels of many compounds when what is important is the relevant organ levels. Even if they realize this they often presume a directly proportional relationship between the blood and relevant organ level
- Reasoning from cause to effect error. Doctors see the effects. to pick a possible cause, no matter how reasonable, and then reason towards the effect observed means that other possible causes will be ignored
- Confusion error. Failure to exclude information that is irrelevant and thus confusing
- Analysis failure error. If results are negative be careful that the results might contain a subgroup that does have a correlation
- Projection error. The tendency to project personal views onto a situation. Whether you would prefer to accept a certain loss of £100 or a loss of £200 with a probability of 0.5 (and thus a 0.5 chance of loosing nothing) depends on whether you have a bank balance of zero and an overdraft limit of £175
- Over confidence error. More than 50 percent of people think they are more intelligent than average

Probabilities

Probabilities are graded from 0 (impossibility) to 1 (certainty). The probability that two or more arbitrary events will occur together is the sum of the possibility that each individual event will occur **minus the probability that both events will occur.** If 60 percent of doctors read the BMJ, 20 percent read the Lancet and 10 percent read both then the probability that at least one is read is 70 percent (60+20-10) = a probability of 0.7 The probability that neither is read is.3 (1.0-0.7)

Coincidences are bound to occur with, to the numerically naive, surprising frequency. The most publicized example is the birthday coincidence. The chances of two people sharing the same birthday in a gathering of 23 people is more than half (this is *not* the same as saying that two people will have the same birthday on a *particular* date)

Finally there is logic in hypocrisy! Whooping cough is a childhood infectious disease with rare but serious complications and for which there is an effective vaccine which (so it was thought) had rare but serious side effects which were much less common than the disease complications. A doctor should advise everyone to have their children vaccinated but omit to vaccinate his own children. The disease would not be available for them to catch it and they would be spared the risks of vaccination. Politicians advise about the importance of family life but with disappointing regularity are found to be having extramarital dalliances.

Correlation: an association which is causal and not coincidental

Associations. Causal or coincidental?

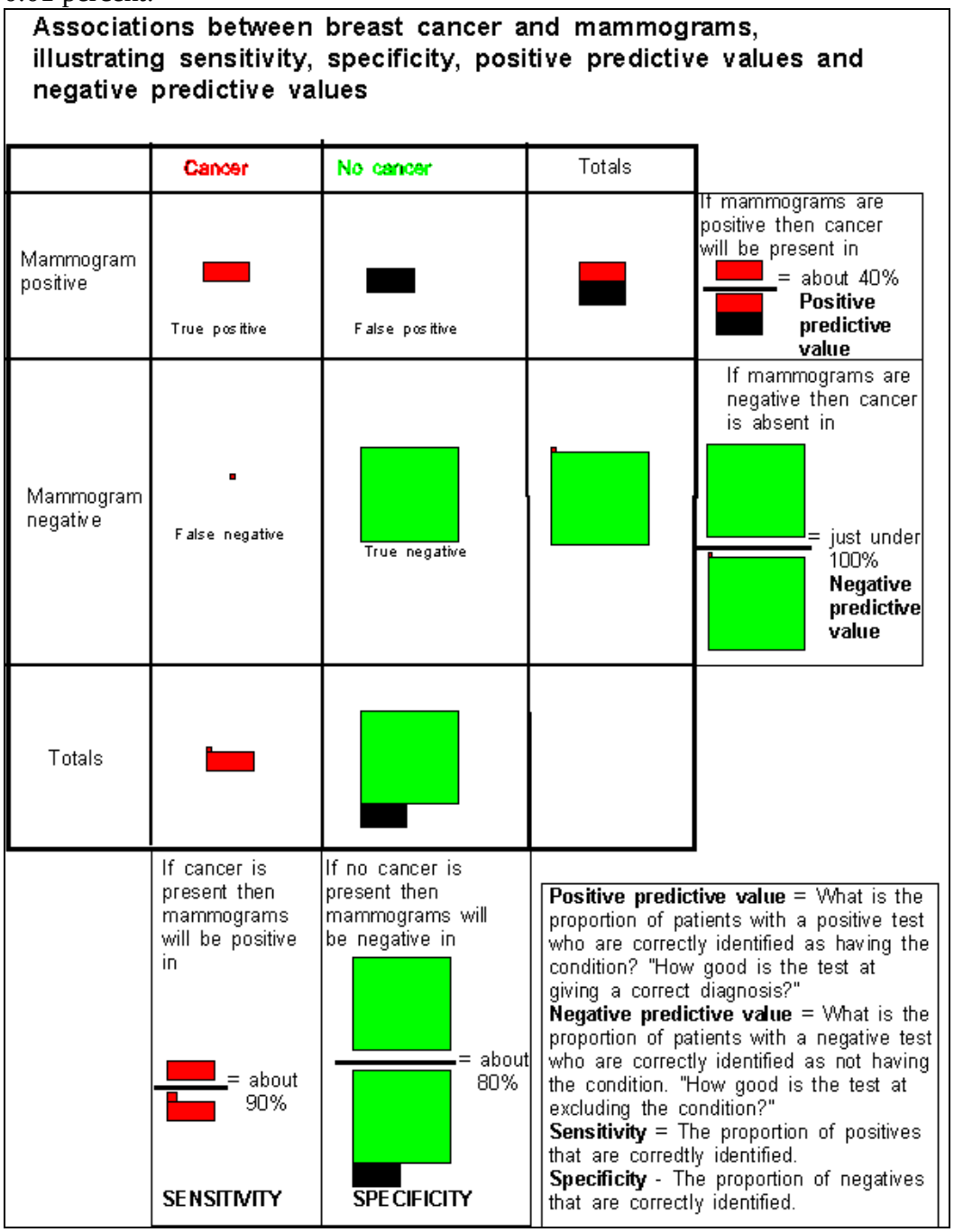
It is a widely unrecognized fact that doctors are continually acting as bookmakers. They see symptoms, signs, and results and have to make a diagnosis which may evoke (sometimes unpleasant) investigations. Never consult a doctor who gambles on games of chance, such as roulette, when the odds are against long-term gain. For a proper assessment of association between two entities it is necessary to appreciate all possibilities.

"Hypercholesterolaemia is associated with heart attacks. All you need to do to realize this is to work on a coronary care unit." As it happens hypercholesterolaemia is significantly associated with heart attacks but working on a coronary care unit will not confirm this as it will not tell you how many people have hypercholesterolaemia but do not have heart attacks. Additionally it could be argued, somewhat deviously, that hypercholesterolaemia reduces the chances of dying from a myocardial infarction and that is why hypercholesterolaemic patients are more likely to be seen in coronary care units.

Probability versus inverse probability

Abnormal breath sounds can be heard in, say, one in five patients. If a doctor carries a stethoscope at all times then, given a perfect world, the doctor should always hear abnormal breath sounds if they are present. The probability that he will have his stethoscope when there are abnormal breath sounds to be heard is 1. The (inverse) probability that abnormal breath sounds will be present if he uses his stethoscope is 0.2. That is rather obvious and no one would confuse the two. However consider that is a woman has breast cancer then mammograms will be positive in a large proportion (usually the probability is about 0 .9) whereas in women who do not have breast cancer the mammograms will be negative in a similarly large proportion (typically just below .9). These results may be academically interesting but are of limited clinical use as it does not tell patients and their doctors what they want to know in practice - "How often do women with positive mammograms have cancer?" and "How often do women with negative mammograms not have breast cancer?" Putting aside the

reasons why various women have mammograms the initial information does not tell us about base rates. The probability of breast cancer after a negative test is low but the probability of a cancer giving a negative test is also low. The two probabilities are not the same and neither are the same as the academic information initially given. If a doctor does not understand these differences then many more breast biopsies will be done because the significance of mammography would have been undervalued if the doctor tells a woman that a positive mammogram indicates breast cancer in 90 percent rather than 40 percent. The doctor will also tell her that a negative mammogram misses breast cancer in about 7 percent of mammograms whereas the true figure is a 0.01 percent.



TRIALS

It is easy to show that something is better than nothing. In the real world it is important to show that something new is better than what was available before.

Trials have been published from general practice that report that giving a certain drug was highly effective because patients did not reconsult. Cyanide would have been equally efficacious!

In a *randomized trial* the intervention or lack of intervention is randomly allocated to comparable individuals. Random allocation should ensure there is no allocation bias. The statistical analysis of randomized trials should ideally be based on *intention to treat* which entails comparison of outcome in all individuals originally randomly allocated *including those who subsequently dropped out or did not comply for any reason* (excluding those who dropped out or who did not comply would introduce a retrospective bias). Such trials are more likely to reflect the clinical picture "No matter what the theory, in real life how did patients respond to the intervention?" The alternative form of trial based *on treatment* ignores those who dropped out of the trial runs the risk of serious bias because people who were failing on the treatment or who developed side effects would be likely to drop out (and thus lead to over optimistic results). However if there are many factors that may influence the result of a randomized trial the will almost inevitably be differences in the groups despite randomization. Detailed selection of patients to minimize the number of such potentially interfering factors (minimization) is a useful technique.

Randomized trials allow rigorous scientific evaluation of a single variable in a precisely defined group of patients. The disadvantages are that such trials are expensive (both financially and in time required), and thus such trials tend to use as few patients as is statistically justifiable and tend to be of brief duration. Endpoints tend to be attainments of laboratory criteria rather than clinical outcome.

Ideally trials should be comparative with at least two interventions, one of which is known to be effective for the same reason that the other interventions is known to be effective. A new drug designed to reduce heart dysrhythmias should be known to be advantageous when compared to other known anti-dyrythmic drugs.

A *single blind trial* is a trial in which either the observer or patient does not know which treatment is being given.

In a *double blind* trial neither the observer nor any of the observed knows which of the observed has received an active intervention.

In *paired or matched comparison trials* patients receive different treatments are matched to allow for confounding variables (eg age and sex) and the results are analyzed from differences between subject pairs.

A *controlled trial* is one in which utilizes a control group of groups who had not received the intervention to compare outcome. Controls can be historical (patients with the disease who, in the past, had not received the intervention) or geographical (patients surveyed elsewhere where the intervention was not available).

A case-controlled trial is a trial in which groups of individual patients who have received an intervention are matched with an "identical" patients who did not receive the intervention.

A *placebo controlled trial* is a trial in which one group of those studied ("arm" in trial jargon) should receive a inactive dummy "the placebo" identical (in appearance, taste etc.) to the active treatment. That will reveal how effective an intervention is *compared to an imitation treatment* which may have psychological or other effects but these effects should be the same in both arms of the trial. There are problems with this in that it may be unethical to give someone nothing; in such circumstances it might be better to give all patients something known to be effective if the intention is to find out which is better (a *comparative trial*).

In *crossover trials* a group of patients receives one intervention then a different intervention (in a random sequence). Often there is a"washout period" with no treatment so that the effect of the first intervention does not affect the second. In its simplest those in one arm of the trial receive intervention A then intervention B whereas those in the other arm receive intervention B then intervention A.

Whichever type of trial is being performed do not confuse efficacy with effectiveness. Efficacy is the outcome of an intervention *in a controlled setting* whereas effectiveness is the effect *in the setting for which it is intended.*

Chi-Square test

This compares the frequency with which certain observations would occur, if chance alone were operating, with the frequency that actually occurs.

The Chi-squared test.

Simply, the Chi-squared test gives a numerical guide to the difference between a series of actual observations from a series of theoretically predicted observations.

120 tosses of an unbiased dice should (very theoretically) yield 20 of each face

TABLE1						
Actual results	16	15	26	13	23	27
Predicted results	20	20	20	20	20	20
Difference	-4	-5	+6	-7	+3	+7

Theoretically the sums of all the pluses and negatives (if the dice may be biased) should be 0

Statsticians, to get rid of - numbers (you've seen this trick before, and don't ask why they don't just forget the positive and negative signs and add the numbers together) square each number (so that negatives becomes positives) and add them all together. Applied to the table above this gives 184 (-4^2 plus -5^2 plus $+6^2$ plus -7^2 plus $+3^2$ plus $+7^2$) which is a numerical measure of the difference of actual results from theoretically predicted results.

This seems perfect but the sum of the squares could give the same result if there had been 1,200 tosses

TABLE 2						
Actual results	196	195	206	193	203	207
Predicted results	200	200	200	200	200	200
Difference	-4	-5	+6	-7	+3	+7

The two tables give the same sum of squares (184) but the result is derived from a much smaller proportion of observations in first table. To remedy this the summed differences have to be related to the expected frequency by dividing each square in the first table by 20 and by 200 in the second table. This gives the summed values from those predicted (Chi-squared) values for Table 1 as 9.05 and Table 2 as 0.905. The smaller the Chi-squared number the more closely the observed and predicted frequencies agree. There are tables which tell us which chi-squared values lie outside 95 percent of the values (statistically speaking we choose to regard values that lie outside the 95 percent as being outwith the limits of chance and thus suggesting that the dice was biased).

The power of a trial is the probability that a trial will produce a significant result at a desired significance level. For a specific trial this will depend on the difference between the populations compared, the samples size(s), and the significance level desired.

Is all this clear? If not, ensure that at least you have absorbed the principles!

The correlation coefficient (whose derivation will not be explained) reflects the strength of relationship between values of two variables. If a disease has a low prevalence then, because sampling errors are proportionally greater in effect, larger samples are required. A correlation coefficient of +1 indicated perfect correlation "tall men always marry tall women" whereas a correlation coefficient of -1 a perfect lack of correlation "tall men always marry short women."

Regression analysis determines the nature of the relationship. The amount of scatter in a scatter diagram gives a subjective feeling for the strength of a correlation and this process can be assisted if "lines of best fit" - *regression lines* (link) can be drawn.

The confidence interval around a trial result indicates the limits within which the real difference between two interventions is likely to be found, and thus reflect the strength of the inference that can be drawn from the result.

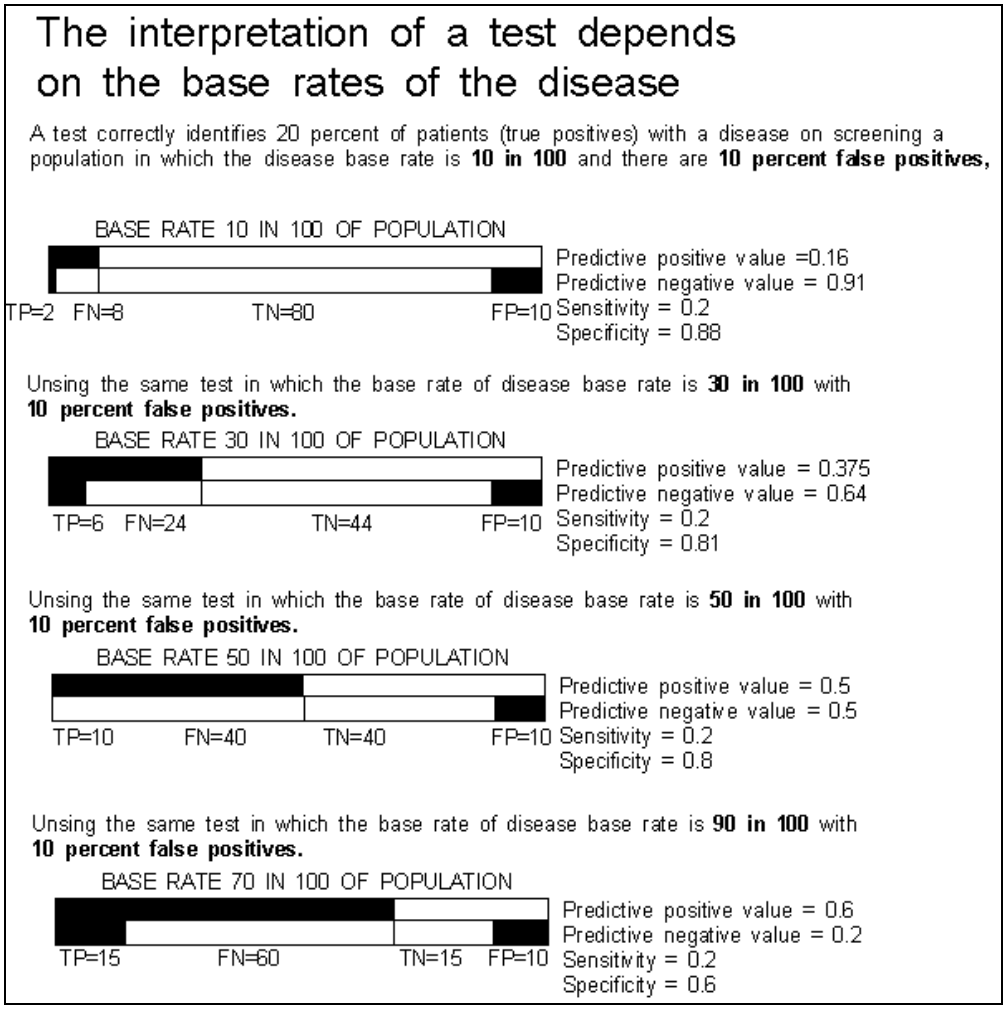
A valuable numerical result, rarely mentioned, is the number (of people) needing to be treated to avoid a condition. If for example a serious complication occurred after bacterial sore throats with an incidence of 1 in 1,000,000 would anyone give 999,999 people penicillin solely to prevent that one complication? Almost certainly no. But if the number needed to treat were 1 in 10,000? Or 1 in 1,000. Or 1 in 100?

MEDICAL EPIDEMIOLOGY

Epidemiology is the study of the incidence, distribution and determinants of diseases in human populations.

Clinical medicine is mostly concerned with *individuals* who have a disease. Epidemiology concentrates on *populations* with diseases and usually attempts to identify associations and causes, and (optimistically) prevention of diseases - "What, why, and what might be done?"

The following diagram illustrates the importance of understanding the importance of base rates.



Epidemiological data can be obtained by many means including:

- Surveys
- Morbidity data
- Mortality data
- Registers of disease prevalence and incidence
- General practice or hospital attendance
- Inspection of centrally held records

Screening

Usually a screening test should be highly selective and highly sensitive. Problems arise when screening large populations for rare diseases when false positive tests can be more common than true positives.

The major problem of epidemiology is the accurate identification of the causes of what is observed, especially when there are several candidate causes and several potential underlying cofactors. For example the effect of Social class on many conditions may be obvious, or hidden, or controversial. Are the higher mortality rates in social class V *associated* with unemployment, but is the association *causal* and directly caused by unemployment?

SOCIAL CLASSES
I. Professional
II. Semi-professional
III. Skilled. Non-manually skilled IIIN
manually skilled IIIM
IV. Semi-skilled
V. Unskilled

Population studies deal with:

- Individuals who have a specific disease
- Individuals who do *not* have the disease
- Changes in frequency of specific diseases
- Patterns of disease
- Incidence rates of specific disease
- Prevalence rates of specific diseases

The incidence of a condition is the number of new cases of a disease in a population occurring in a defined time = "How many *new* cases each year?" Prevalence is the frequency of the condition is at a specified point in time in a defined population "How many cases are there altogether?"

The difference between incidence and prevalence

	Prevalence high	Prevalence low
Incidence high	Common persisting conditions	Commonly occurring brief infections
	eg. osteoarthritis in the elderly	eg. the common cold
Incidence low	Persisting diseases which are uncommon	Rare brief duration diseases
	eg. rheumatoid arthritis	eg. malaria in the United Kingdom

The incidence rate is the incidence divided by the total population and the prevalence rate is the prevalence divided by the total population

Principals of epidemiological inquiry

First there must be an idea, probably derived from a casual observation, then a hypothesis which can be investigated to provide description, quantification, then analysis and finally interventional studies to test the hypothesis.

Description involves reporting:

- The variation of disease and the putative cause
- The geographical or situational variation
- Variation of putative causes
- Variation of the disease
- Variation in the disease in those affected

Often factors, such as age, affect disease incidence and prevalence. A useful assessment as to whether there is a change in serious disease is the standardized mortality ratio which is the actual number of deaths each year compared with the expected number of deaths. This should be done for whole populations and subsections of populations (age often affects diseases and it would be important to find the standardized mortality ratio in those of various ages).

Life expectancy is the mean number of years that individuals drawn from a specified population can expect to live.

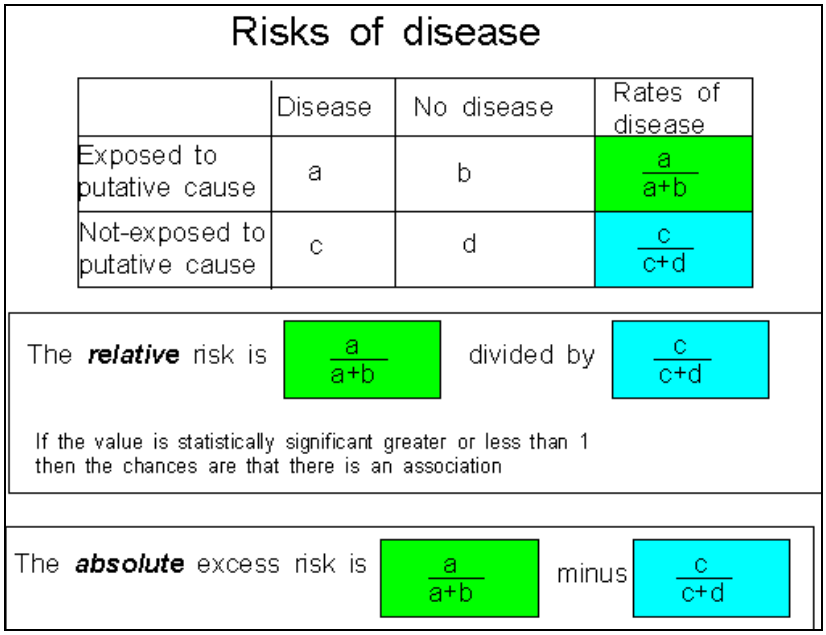
Descriptive studies are studies of the variation in incidence of a disease according to time, place, or person and may:

- Suggest causes of disease
- Enable quantification of the health problem
- Enable quantification of the financial implications

Analytical studies are designed to quantify the risk of disease associated with a putative risk factor.

The *relative risk*, the amount of disease that a putative factor might cause, is the incidence amongst those exposed to the putative factor **divided** by the incidence amongst those not exposed to the putative risk factor.

The absolute excess risk is the incidence amongst those exposed to the putative factor **minus** the incidenc amongst those not exposed to the putative risk factor.



The odds ratio is the ratio of the probability that an event of interest occurs to the probability that it does not (the odds ratio approximates to the relative risk if the risk of disease is low).

Prospective studies

Prospective studies (also known as longitudinal or cohort studies) entail following up individuals who are exposed to a putative risk factor and discovering how many develop the disease in question.
The advantages of prospective observations are:

- Knowing that exposure to the putative risk factor antedates the disease
- Knowing that there should be more accurate observations "looking for the disease at the time rather than relying on adequacy of notes made at the time"
- Unexplained associations may become apparent if varied observations are made

There are problems with prospective studies:

- They may take many years to produce results - and thus are often not career-enhancing for those in training
- They may require a large number of observations (especially if the disease in question is rare)

Retrospective studies

Retrospective studies can be case controlled (in which there is a comparison of samples of individuals who have developed the disease with a sample who have not). The relative risk cannot be identified because the proportions of the population from which each group are drawn are usually not known.
The advantages of retrospective studies include:

- Potential speed of completion
- Smaller numbers are often required
- They are usually cheaper

- They are useful for rare diseases (waiting for rare diseases to occur can be very boring)

The disadvantages of retrospective studies include:
- Uncertainty as to whether the putative risk factor preceded the disease (as it should)
- Reliance in part on the individual memories of events, or notes of events made at the time
- Identification of relevant individuals are often unfocussed because of failure to look for risk factors or the disease *at the time*
- No straightforward assessment of the excess risk is possible (although indirect assessments for rare diseases are possible)

Intervention studies

Intervention studies assess the effect of planned interventions and thus can be planned and costed within a defined timescale.

Statistical analysis can then reveal the degree that changes observed happened by chance or occurred because of the intervention.

Control groups who had not received the intervention are used to compare outcome. These can be:
- Historical controls - patients with the disease who, in the past, had not received the intervention
- Geographical controls - patients surveyed elsewhere where the putative risk factor may be different
- Randomized controls - the intervention or lack of intervention is randomly allocated to comparable individuals

Historical or geographical controls might not compare like with like. Patients who are more severely affected by a disease might gravitate towards researching centers and thus the results may not be representative of the population as a whole. Trials can be designed using a control group and other groups, each group receiving a different intervention.

Not all associations are causal. The best illustration is the known association of lung cancer with alcohol intake. Alcohol does not cause lung cancer. It is the increased frequency of smoking in those with a high alcohol intake that is the cause. Establishment that an association is causal entails:
- the identification of associations
- discovering if the pattern of disease can be altered by intervening or altering the putative cause
- by discovering a similar population who have differing patterns of the disease and observing if the same risk factor(s) are present and what the incidence of the disease is in that group.

Prevention

There are three types of prevention:

- Primary prevention. Prevention of future occurrence in unaffected individuals by removing a cause. Possible causes include environmental, economic, social, educational, and dietary factors. Interventions include remedying adverse causes and vaccination programmes
- Secondary prevention. Prevention of clinical disease by screening, early detection and/or treatment
- Tertiary prevention (in theory) prevention of disease by treating clinical cases

Prophylaxis usually refers to prevention using drugs. Primary prophylaxis is using a drug to prevent a disease before it occurs whereas secondary prophylaxis is used once a disease has occurred in an individual in an attempt to prevent it recurring.