Journal of Fertilization: In Vitro - IVF-Worldwide, Reproductive Medicine, Genetics & Stem Cell Biology

Walters et al., JFIV Reprod Med Genet 2016, 4:3
DOI: 10.4172/2375-4508.1000186

Review Article

Explanatory Notes on Some Important Statistical Topics

Eurof Walters*

Churchill College, Cambridge, UK

Abstract

This paper describes and elaborates on some statistical topics that will be of importance to research workers in ART (Assisted Reproduction Technology). The aim is to provide fairly elementary and easily accessible descriptions of those topics. The topics included will of course have been well covered in statistical texts, but in a manner that may be alien to biologists. Simplicity and accessibility will therefore be the keynote.

The selection of topics for ventures of this sort will of course be a personal matter, but those selected in this paper appear to the writer to be the most important, and in need of a simple elaboration. The anticipated readership will be research workers in the field of ART (Assisted Reproduction Technology).

Keywords: Statistical analysis; Assisted reproduction technology (ART)

Introduction

The purpose of this paper is to discuss, at a fairly elementary level, some topics that are crucially important in the statistical analysis of data. The expected readership will be research scientists working the field of Assisted Reproduction Technology (ART), but the content would be appropriate for many other biological research workers.

By following the study of biological subjects at school and at university, the mature biological researcher often experiences a deficiency in his knowledge of the mathematical or statistical disciplines. Although statistical packages now enable the researcher to employ, at a practical level, some rather sophisticated techniques, those packages very often provide little guidance in the philosophical background of those techniques. It is hoped that this paper will go some way to fill this gap in understanding the rationale of statistical evaluation.

This paper will not follow the customary pattern of papers published in this journal. Indeed it may seem more like a lexicon of statistical terms rather than a scientific paper. It is hoped that this is not perceived as a flaw, but as a useful addition to the readers' knowledge and appreciation of statistical methods. The author has selected, for this exercise, a few topics that it is hoped will throw light on issues that often cause problems for the practical research worker.

The Central Limit Theorem

Although not widely known outside the statistical/mathematical fraternity the Central Limit Theorem occupies a pivotal role in the statistical evaluation of data. It is in fact crucial for the reliable use of statistical methods, at least for those methods that demand a normal distribution of errors. Put into simple language the Central Limit theorem states that if several, say 'n' independent random observations are drawn from a population, of whatever finite distribution, the distribution of the mean values will tend to normality as 'n' tends to infinity. What this means is that if one is worried that the distribution of a variable being analysed departs from normality, taking the mean value of several independent observations will improve the situation immensely. The theorem was first proposed by Laplace, but a formal proof was due to the Russian mathematician Liapounoff in 1901. Various manifestations of the Central Limit Theorem have been described by Feller.

The effect of the Central Limit Theorem can be demonstrated quite dramatically by generating the mean values of as few as four randomly drawn observations from a rectangular distribution. The distribution of the resulting mean values is displayed in Figure 1, and it already has the familiar bell-shaped appearance of a normal distribution.

Although ART research workers will not be over-concerned with the theoretical aspects of the Central Limit theorem, they will surely appreciate its value in statistical inference.

Randomisation

For the statistician perhaps the most disturbing aspect of a good deal of biological research is how infrequently randomisation is applied in experimental design. Now this procedure was 'invented' by perhaps the most famous of all statisticians R. A. Fisher, and despite Fishers'

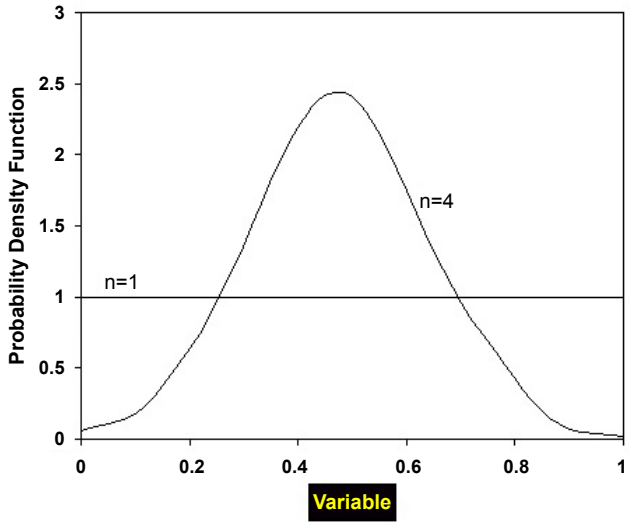

Figure 1: The central limit theorem in action.

***Corresponding author:** Eurof Walters, Churchill College, Cambridge, United Kingdom, Tel: 01223891558; E-mail: dew1@hermes.cam.ac.uk

Received May 23, 2016; **Accepted** June 17, 2016; **Published** June 24, 2016

Citation: Walters E (2016) Explanatory Notes on Some Important Statistical Topics. JFIV Reprod Med Genet 4: 186. doi:10.4172/2375-4508.1000186

massive contribution to mathematical statistics, there are some who believe that conceiving this simple yet beautiful idea was perhaps his greatest achievement. If two or more treatments are allocated to the experimental material by a random process, it guarantees the absence of systematic bias, and furthermore ensures the reliability of experimental error. R. A. Fisher first promoted this idea in his early book, 'The Design of Experiments' Fisher [1] where one of the sub-headings was Randomisation: the physical basis of the validity of the test.

Unfortunately, the statistical rigour enshrined in the randomisation procedure is often relaxed, particularly in medical investigations, where ethical matters as well as practical considerations may be in conflict with the pure objectivity of randomisation. Even so the (ART) analyst would be wise to ponder on whether any systematic bias has entered the system due to the lack of randomisation. There may be factors that are confounded with the main effects of interest, by which is meant that due to the structure of the design the two factors may be linked in such a way as to make disentanglement difficult. When analysing data which are derived from an existing database, this often being the situation obtained in ART research, the analyst needs to assure himself, as far as is possible, that there is no reason to believe that the absence of randomisation has caused a built-in bias.

However, a recent paper [2] was, in this writer's view, quite irresponsible to argue, and I quote that an 'obsession with prospective randomised trials' was misplaced. It is surely relevant that although seeming to want to destroy one of the cornerstones of statistical practice, the reference list of 29 citations in Gleicher's paper does not contain a single paper from a mainstream statistical journal.

Randomisation in experimental design is a beautiful and rigorous procedure and experimenters should strive to use the technique whenever possible. Even if randomisation is not possible an author would be wise to note its importance, and make the point that the arrangement under review may well correspond in some way to a random process. At the very least one needs to assume that no bias has been introduced by the absence of randomisation.

I fear that, due to ethical and practical reasons, the randomisation principle is not often applied in ART research, Even so, the research worker should really regard it as 'The gold Standard' in the research process

Null Hypothesis

The Null Hypothesis is a concept that is a part of what may be called classical statistical testing. The rationale here is to assume, initially, that the effect of interest to the analyst does not exist and then to calculate the probability of such an extreme set of results as that obtained occurring in the absence of the effect.

To illustrate these ideas consider a coin tossing experiment. Suppose that in 10 tosses the coin a 'tail' occurs 8 times. The Null hypothesis here is that the coin is perfectly balanced so that the probability of a tail is 0.5, and the expected number of tails in 10 throws is 5. In a two-sided test therefore the probability of SUCH AN EXTREME departure from expectation, that is less than 3 and greater than 7 tails is made up as follows.

Probability=$P_0+P_1+P_2+P_8+P_9+P_{10}$=0.00098+0.00978+ 0.04395+0.04395+0.00978+0.0.00098=0.1094

Note that the adoption of a two-sided test involves taking both tails of the distribution. We thus see that if the coin was perfectly balanced, such an extreme result as 8 or more tails (or 2 or fewer tails) would occur with probability of 0.1094. Since this is a good deal greater than the customary value of 0.05, we conclude that there is no firm statistical evidence that the coin is unbalanced.

In this simple example the probability of such an 'extreme' event is calculated directly. A more frequent situation obtained in experimental work is for a test statistic, for example a Student's 't' value or a Variance Ratio (F statistic) to be computed, and the evidence of departure from expectation assessed by referring the test statistic to tables of the distribution of that statistic in the null case.

Although the numerical example used here to describe the relevance of the Null Hypothesis may be very far removed from ART research work, the reader should be able to translate those ideas into that setting.

The P Value

The P value is a quantity used in the classical mode of statistical inference. If an investigation involves estimating a certain treatment difference the P value represents the probability of such an extreme difference as that obtained occurring by chance, under the 'Null Hypothesis' (q.v.).

The rationale therefore is to assume that the effect of interest is not present, and then to calculate the probability of such an extreme event occurring under those circumstances.

A 'P-value' of 5% (0.05) is often adopted as a critical value in making inferences. Thus if the P value associated with an effect is calculated as 0.05 the inference is that such a large effect would only occur by chance with a probability of 0.05, or once in 20. Generally this would be taken as evidence of the presence of the effect in question. It is important to realise however that under those circumstances there is a 1 in 20 chance of being wrong in the conclusion. See also Sensitivity and Specificity.

Unfortunately the P value is a statistic that is widely misused and mis-interpreted in a great deal of experimental work. Very often a very large number of tests are cited in a publication but very little attention is paid to the probability of spurious significant results. After all if 50 similar tests are carried out in a single investigation, one would expect more than 2 significant results at the 5% level simply by chance and random variation. Clearly for large investigations some allowances are needed to deal with this troublesome feature in order to guarantee reliable inferences.

One of the earliest attempts to deal with this problem was by Bonferroni [3] who recommended that the nominal P value, often set at 0.05, should be scaled down by the number of tests envisaged. Thus if there are 10 tests in an investigation, the nominal significant P value for each test should be reduced from 5% to 0.5%, in order to retain an experiment-wise Type 1 error of 5%.

The rationale for this very simple solution to the problem, proceeds as follows. If 'n' is the number of independent tests envisaged in an investigation, and we wish to retain an 'experiment-wise' probability of a Type 1 error of say 0.05 then the P value applied to each individual test α_t is the solution of the equation

$$1 - (1 - \alpha_t)^n = 0.05$$

Thus $\alpha_t = 1 - (1 - 0.05)^{1/n}$

More generally, if α_e is the chosen 'experiment-wise' Type 1 Error and α_t is the corresponding individual test P value then

$$\alpha_t = 1 - (1 - \alpha_e)^{1/n} \approx \alpha_e/n$$

In fact the Bonferroni approach is rather conservative, and numerous alternative procedures have been suggested over the years,

important early contributions having been made by Scheffe [4] and Duncan [5]. A more recent contribution to the topic is by Benjamini and Hochberg [6].

On the Performance of Statistical Tests

When performing a statistical test to detect the presence of a particular 'effect', there are of course two types of erroneous conclusions. One could conclude that the effect is genuine when in fact it is absent, or one could fail to detect the (genuine) presence of the effect. In statistics the terms Sensitivity and Specificity are used to deal with these situations.

Sensitivity defines the probability that a test procedure will correctly identify an 'effect' that is present. It therefore represents the 'Power' of a statistical test, which is the more familiar term used in the standard statistical literature. The difference of the sensitivity value from unity, which represents the failure to detect a condition that is present, is often referred to as 'false negative', error of the second kind, or **β** type error. Power therefore is denoted by (1-**β**).

Specificity defines the probability that the test procedure will correctly fail to detect an 'effect' that is absent. The difference of the specificity value from unity, which represents the erroneous detection of a condition when it is absent is sometimes called 'false positive'. This statistic is variously described as the error of the first kind, Alpha (**α**) type error, P value, or the size of a test. The following diagram displays the circumstances giving the relevant probabilities.

	Test	
	Present	**Absent**
Condition Present	(1-β)	β
Condition Absent	α	(1-α)

The Sensitivity or the Specificity may be changed (perhaps improved) by modifying the test conditions, but the other will then be automatically adjusted. The two statistics are inextricably linked. The analyst generally decides on the acceptable value of one or other of the probabilities, the second probability then being fixed, automatically.

Robustness

The property of robustness is another feature that contributes to the reliability of many statistical procedures. Generally, and in order to be absolutely reliable, a statistical procedure such as the Analysis of Variance, or Student' t test demands some validating assumptions. In this context, the property of robustness means that the procedure behaves fairly reliably even when the assumptions are not satisfied fully. One typical example is that in an analysis variance the 'errors' should be independent and follow a Normal Distribution. Even if these conditions are not satisfied fully, the analysis will often remain fairly reliable. In fact Efron [7] demonstrated that in the extreme case of data consisting entirely of 'zeroes' and 'ones', inferences derived from a resulting analysis of variance were still quite reliable.

The relative reliability of statistical evaluation, even in what appear to be the most unpromising of circumstances, owes a great deal to the robustness of the methods allied to the Central Limit Theorem in action.

Distribution-Free Methods

In view of the important statistical properties already discussed, that is Robustness and the Central Limit Theorem, the application of classical statistical procedures will often be quite reliable even in what appears to be unpromising circumstances. However, if the analyst remains uneasy about the application of those methods a plausible alternative is to use 'Distribution-Free' methods. As the title suggests these methods do not demand the customary distributional assumptions. One class of distribution-free methods involves replacing the data points with the corresponding rankings, whereas a second class involves examining all possible permutations of the data points.

Data Structure	Normal Theory	Rank Test	Randomisation Test
Unmatched two sample	Student's t	Mann-Whitney	Yes
Matched Two Sample	Student's t	Wilcoxon Test	Yes
Two Factor Design	F Test	Kruskal Wallace	Yes
Contingency table	Chi-Squared	N/A	Yes
Two Variable	Corr. Coeff.	Kendal/Spearman	Yes

Table 1: Table listing three possible tests for some simple data structures.

Now rank methods have a long history and have been easily accessible to the statistical analyst. However, methods involving permutations of the data structure although known in principle, were found to be impractical before the advent of computers. Those difficulties have now been removed and methods involving a study of permutations, (see [8]) and methods involving a re-sampling of the data structure (See [9]) may be carried through quite expeditiously. Table 1 lists for some simple data structures three possible methods of analysis, being classical methods, rank methods and methods adopting the randomisation/permutation principle.

The Concept of Orthogonality

In Mathematic/Statistics the word orthogonality is used to describe several situations that are of course closely linked. When applied to two lines or vectors in a geometric setting orthogonality simply means that those lines/vectors are at right angles.

When used in connection with experimental design and the subsequent analysis of the data orthogonality means that the structure is such that the several factors are independent of each other. For example in a two way classification of rows and columns, with an equal number of observations in the cells, the variation in the dependent variable calculated between the row totals does not contain any inter-column variation and vice-versa. If the rows and columns represent the levels of two factors, then estimation of the factor effects follow directly from the corresponding row or column totals. If there are one or more missing observations in the cells, or if the degree of replication in the cells varies, orthogonality is lost and the analysis of the data becomes more troublesome.

Before the widespread availability of computers the property of orthogonality was crucial in that the analysis of experimental data was then fairly simple and straightforward and could be carried out on a desk calculator. The loss of orthogonality however meant that the analysis of the data would typically involve the inversion of fairly large matrices, a task that would not really be feasible on a calculator. This really explains why early books on experimental design lay great emphasis on orthogonality, or certainly some measure of 'balance'. When the degree of non-orthogonality was small, involving perhaps a few missing observations, ingenious simple methods were promoted to carry out the analysis. A large number of missing data however usually led to the need for matrix methods.

Although the absolute need for orthogonality has now diminished and computer programs can deal quite happily with non-orthogonal designs, the desirability of orthogonality remains.

In statistics the word confounding is used to describe the situation

where two factors in a design are linked. If the linkage is high the analyst's ability to extract and estimate the impact of the effects is severely diminished. This would be called partial confounding. If however two factors are linked in a complete sense so that it is impossible to disentangle and estimate the separate effects, this situation would be called complete or total confounding. It would in general be impossible to determine which of the two confounded factors caused the resulting variation.

Generalised Linear Modelling (GLM)

Before the advent of computers the 'Analysis of Variance' and 'Regression Analysis' were regarded as essentially different techniques. Regression analysis usually involved continuous explanatory variables whereas the analysis of variance used categorical or qualitative explanatory variables, with perhaps one or two continuous covariates. Both however generally assumed errors that were normally distributed. Now however, thanks to the computer, the two techniques have been subsumed into a single flexible procedure; Generalised Linear Modelling (GLM). Qualitative factors and continuous explanatory variables can easily be included in the same analysis, and the error distribution may be selected by the analyst.

This development has a claim to being the most radical and beneficial advance in applied statistics in the last 50 years. Further, many computer packages, which should be available to ART researchers, offer GLM modelling as an option.

Meta Analysis

Although the essential ingredients of Meta-Analysis have been known and applied for very many years in the research area of Agricultural Field Trials, this new term has more recently been proposed. It describes a procedure whereby several independent studies on the same subject matter are pooled to provide a composite finding. However, a meaningful interpretation of Meta Analyses requires a careful handling of the hierarchical error structures (See [10-13]).

Although the actual calculations for a Meta Analysis can now be carried out quite simply using the Cochrane algorithm [14], there is still plenty of scope for erroneous and misleading analysis and interpretation. See for example Walters [13] for a critique of the way that Meta Analyses are often mis-applied.

Multivariate Analysis

The term multivariate analysis has a claim to being the most mis-used expression in the entire statistical lexicon. The strict statistical definition would be an analysis that considers several dependent variables viewed 'COLLECTIVELY'. Unfortunately it is often used to describe analyses when several dependent variables are analysed in turn, but individually, according to some model.

Suppose the variables height, weight, girth etc. are noted for two groups of individuals. Inter-group differences may be examined in the usual way for each variable in turn, but a multivariate analysis would look at the inter-group differences on all the variables viewed together. Typically a multivariate analysis on 'p' variates would involve working with (p x p) matrices, whereas univariate analyses do not require the use of matrix arithmetic, at least not for balanced data.

Multivariate procedures which the analyst will sometimes find to be of value in processing complicated data structures include Multivariate Analysis of Variance, Discriminant Analysis. Canonical Variate Analysis, Principal Component Analysis and Factor Analysis. There are numerous excellent books that describe these topics in detail. For example, Krzanowski [15].

The Value of Negative Inference

The purpose of this section of the paper is to emphasize the care that is needed in interpreting 'negative inference'; that is a failure to detect a significant effect in a particular situation. A typical manifestation of this phenomenon is when the analyst concludes that various nuisance variables do not exert an effect; this often being based on a test of very low power. This finding is often found to be to the analyst's advantage, so that there is a vested interest not to detect the effect in question. After all if you don't really want to find something the best plan is not to look too diligently for it.

A favourite source of controversy regarding negative inference, in the early days of IVF, was when the replacement of numerous fertilized embryos was promoted by many clinicians. The claim was that they had failed to detect a significantly higher incidence of undesirable multiple pregnancies, whereas the overall pregnancy rate was increased substantially by increasing the number of replacements. Thus they would then be using 'negative inference' to their own advantage. A good example of this unsatisfactory rationale will be found in an early paper by Azem et al. [16], which prompted a response by the present author Walters [17].

The controversy regarding the number of fertilized embryos to be replaced was recently given another airing by Johnson et al. [18], Gleicher [19], Bissonnette et al. [20].

If authors need to make important claims based on the absence of a certain effect, they should really provide figures giving the power of the tests applied for plausible experimental situations. The reader would then be able to assess the appropriate weight of evidence in support of the writer's conclusions.

Publication Bias

Publication Bias may be defined as the phenomenon by which scientific papers reporting statistically significant results are more likely to be accepted and published than other perfectly good papers reporting inconclusive results. This tendency will certainly distort the perceived view of the subject matter, and further will cause severe distortion when the published papers are included in a systematic review (Meta-Analysis). The sense of the distortion will almost certainly be to exaggerate the importance of a new, perhaps novel, finding.

A renewed interest in Publication Bias, and its impact of scientific research, seems to have resulted from the paper by Ravnskov [21], since when there have been a large number of papers dealing with this phenomenon.

Ravnskov, considering the impact of lowering cholesterol levels on coronary heart disease, found that papers reporting positive (i.e., statistical significant results) were six times more likely to be published than papers reporting inconclusive results.

As far as assisted reproduction is concerned one paper giving a more recent perception of the problem is by Polyzos et al. [22] who found an odds ratio of 2.5 in favour of the publication of significant results. Although the ratio is not as great as that for the cholesterol study, the bias is clearly still present.

Bayesian Methods

Although what is called the classical mode of statistical inference is the one used almost exclusively by statistical packages, there is an alternative system that has been promoted by an enthusiastic band of followers; that is the Bayesian approach. This approach follows

an original idea by the 18th century non-conformist clergyman and amateur mathematician Thomas Bayes (1702-1761).

The essence of the system is that the analyst should combine any prior information about a particular phenomenon with current information, perhaps in the form of experimental data, to produce a final result.

The most controversial aspect of the Bayesian approach, and indeed it has generated a great deal of controversy over the years, concerns the quantification of the prior information by the analyst. Analysts have exercised considerably latitude in their choice of prior information, and changing the prior information of course has an impact on the final result. This subjective element in the process assumes great importance when there is no prior information at all and the analyst has to decide on a plausible prior reflecting complete ignorance.

Perhaps it should be mentioned that in Bayes's original paper which involved the binomial distribution and its parameter (p), it was quite logical to assume a rectangular distribution as prior information for 'p', reflecting complete ignorance as to its value.

A rather unusual set of events concerns Bayes's original work and how it came to public notice. It was not published during his lifetime, and the radical thinker Richard Price who was a friend of Thomas Bayes was given the task of sorting out Bayes's papers after the latter's death in 1761. Price eventually prepared the manuscript for presentation at the Royal Society [23]. The contents of the paper generated little interest until the 20th century when the ideas were absorbed and vastly extended. The result of course was a branch of statistics that we now know as Bayesian methods.

Discussion

The main body of this paper consists of a list of statistical topics, together with a short explanation, and where applicable a brief discussion on the topic. Clearly this list could be extended substantially to result in something that resembles a dictionary of statistical terms. See for example Upton and Cook [24]. That is not the intention however. It is hoped that the short explanatory passages provided for the selected topics will enable the reader to better understand the important features of that topic, and the impact on the statistical evaluation of data. These topics have been selected as being particularly relevant to the research area of Assisted Reproduction, and will hopefully assist the research worker in an area that may be outside his everyday experience.

Although ethical and practical considerations in ART research means that the strict rigorous application of statistical methods may not always be possible, the analyst should nevertheless aim for a strict adherence to good statistical principles.

References

1. Fisher RA (1935) The design of experiments. Oliver & Boyd, London.
2. Gleicher N, Barad DH (2010) Misplaced obsession with prospectively randomized studies. Reprod Biomed Online 21: 440-443.
3. Bonferroni CE (1935) Il calcolo delle assicurazioni su gruppi di teste, In Studi in Onore del Professore Salvatore Ortu Carboni, Rome, Italy.
4. Scheffe H (1953) A method for judging all contrasts in the analysis of variance. Biometrika 40: 87-100.
5. Duncan DB (1955) Multiple range and multiple F tests. Biometrics 11: 1-42.
6. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B 57: 289-300.
7. Efron B (1978) Regression and ANOVA with zero-one data; Measures of random variation. J Amer Statist Ass 73: 113-121.
8. Walters DE (1981) Sampling the randomisation distribution. Journal of the Royal Statistical Society Series D 30: 289-295.
9. Efron B (1980) The jack-knife, the bootstrap and other re-sampling plans. Division of Statistics, Stanford University.
10. Walters DE (2000) The need for statistical rigour when pooling data from a variety of sources. Hum Reprod 15: 1205-1206.
11. Daya S, Gunby J (2001) Potential dangers in the customary methods of conducting meta-analyses. Hum Reprod 16: 2250-2252.
12. Walters E (2002) "Uneasy Science" - the pooling of heterogeneous data. Fertil Steril 77: 1308-1309.
13. Walters E (2013) A new title to an old song: some observations on the conduct of meta-analyses. Reprod Biomed Online 27: 562-567.
14. Higgins JPT, Green S (2011) Cochrane Handbook of systematic reviews of intervention. Version 5.1.0, The Cochrane Collection.
15. Krzanowski WJ (1990) Principles of Multivariate Analysis. Clarendon Press, Oxford.
16. Azem F, Yaron Y, Amit A, Yovel I, Barak Y, et al. (1996) Transfer of six or more embryos improves success rates in patients with repeated in vitro fertilization failures. Fertil Steril 63: 1043-1046.
17. Walters DE (1996) The statistical implication of the 'number of replacements' in embryo transfer, Hum Reprod 11: 10-12.
18. Johnson M, Cohen J, Grudzinskas G (2011) Who should control how many embryos to transfer: the state or the patient? Reprod Biomed Online 23: 399-400.
19. Gleicher N (2011) Eliminating multiple pregnancies; an appropriate target for government intervention? Reprod Biomed Online 23: 403-406.
20. Bissonnette F, Phillips SJ, Gunby J, Holzer H, Mahutte N et al (2011) Working to eliminate multiple pregnancies; a success story in Quebec. Reprod Biomed Online 23: 500-504.
21. Ravnskov U (1992) Cholesterol lowering trials in coronary heart disease; frequency of citation and outcome. BMJ 305: 15-19.
22. Polyzos NP, Valachis A, Patavoukas E, Papanikolaou EG, Messinis IE, et al. (2011) Publication bias in reproductive medicine: From the European Society of Human Reproduction and Embryology annual meeting to publication. Hum Reprod 26: 1371-1376.
23. Bayes T, price (1763) An essay towards solving a problem in the doctrine of chances. Phil Trans 53: 370.
24. Upton GJG, Cook I (2008) A Dictionary of Statistics. Oxford University Press.