The statistical analysis of immunohaematological data

Roberto Reverberi

Servizio di Immunoematologia e Trasfusionale, Arcispedale S.Anna, Azienda Ospedaliera -Universitaria di Ferrara, Italy

Almost 30 years ago, I submitted one of my first research papers on immunohaematology for publication in a prestigious journal. After a few weeks, I received back the referee's comments, with many criticisms, including the fact that I had used statistical tests. The referee had been disappointed because "statistical tests are not usually performed in immunohaematological studies". For that time, the referee was actually right. Even in more recent years, however, immunohaematologists continued to apply statistics very sparingly. For example, consider a table such as table I, which is frequently encountered in immunohaematological articles. It contains rows representing antibody specificities, columns representing two or more techniques for antibody detection, which are being compared, while data are titration end-points. When there are many rows (say, more than six), a fleeting reader would greatly appreciate a summary statistic. However, I remember just a few papers in which such a statistic was presented[1-3], and in only one of them[1] was it the most appropriate (the geometric mean)[4].

The statistical analysis of immunohaematological studies is, however, straightforward, provided adequate consideration is given to the type of measurement and the distribution of the data. In most cases, non-parametric tests are appropriate.

In this paper, I shall briefly review the modern statistical software that is freely available to researchers and then show the use of statistical tools in the most common situations. First of all, however, I shall discuss the contribution of statistics to the presentation and analysis of immunohaematological data.

Table I - Summary data (reciprocals of titre end-points) from a comparison of three techniques (Aggl: standard tube test; ELAT-W: enzyme-linked antiglobulin test; ELAT-G: like ELAT-W but with gradient centrifugation instead of washing) for the indirect antiglobulin test. Three summary statistics are presented: arithmetic mean, median and geometric mean (GM). The geometric mean has the most preferable properties. (Q_1-Q_3: interquartile range; SD: standard deviation; GSD: geometric standard deviation).

Antibody	Technique		
	Aggl	**ELAT-W**	**ELAT-G**
1	32	512	128
2	32	256	128
3	32	64	128
4	32	64	128
5	32	64	128
6	8	64	32
7	32	32	128
8	16	32	64
9	8	16	64
10	64	16	128
11	8	8	256
12	4	8	16
13	4	8	32
14	4	8	32
15	4	8	2
Mean (±1SD)	**21** (3 - 38)	**77** (-58 - 213)	**93** (26 - 160)
Median (Q_1-Q_3)	**16** (4 - 32)	**32** (8 - 64)	**128** (32 - 128)
GM (±1GSD)	**14** (5 - 37)	**31** (8 - 115)	**61** (18 - 206)

What statistics can do for you

As mentioned above, summary statistics are useful to display data in a concise but representative way. Statistical tests are mainly useful for estimating the probability of making an error concluding against the null hypothesis (the null hypothesis is that the observed differences are casual). It should be pointed out that statistical tests cannot

decide for you. The correct interpretation of "p=0.05" is that in 5 cases out of 100, a difference such as that observed arises by chance alone. This is not really different from a 4 or 6% probability.

Significance limits should be set in advance, before results are analysed. The question should be: "what probability of error am I willing to accept?". If concluding against the null hypothesis entails considerable inconvenience and cost, I would accept it only if the chances of error were no more than 1%. In the opposite case, I would probably content myself with a 10% probability. In any case, a statistical test leaves you with probabilities, not certainties.

Statistical significance is different from practical significance: particularly when the number of cases is very high, differences of no practical consequence may be statistically significant. However, once you have decided the minimum difference that you would consider to be practically significant, statistics can provide an estimate of the probability of erring in favour of the null hypothesis, given the number of cases and the observed variability.

Free statistical software

It is not usually practical, or advisable, to resort to a professional statistician for every statistical aspect of your research. Unless the statistician is already well versed in your field, it is usually easier and faster for you to acquire a basic knowledge of statistics, than vice versa. It is also more profitable, because it is a precondition for a correct study design. In fact, with the help of currently available software, calculations are no longer a problem and you may concentrate on the right analysis to perform.

The number and scope of the statistical programmes that are now freely available to the scientist are impressive. They can be divided into general-purpose programmes with built-in statistical functions, such as spreadsheets and database engines, and specialised software.

Statistical functions of spreadsheets

For a brief introduction to spreadsheets, consult reference 5. Spreadsheets have a limited support for statistics. Their main use in this context is to provide an easy way to store, transform, summarise, display and exchange data. In fact, Gnumeric[6] performs many parametric tests, such as the *t*-test, analysis of variance, covariance, regression, correlation and others, but it lacks non-parametric tests and cannot be a real substitute for a specialised programme. Spreadsheets are also limited as regards the number of data that they can manage: no more than 65,536 rows and 256 columns for a single sheet. It is quite unusual to need more columns, but modern information systems generate a huge amount of data, which may exceed the maximum number of rows. In those cases, a real database is necessary, such as Microsoft Access[7] or its open-source counterpart, OpenOffice.Base[8]. Both have a visual interface and, unlike the most powerful database engines, such as Oracle[9], PostgreSql[10], MySql[11] and similar, they are suitable for personal use. Their built-in statistical support is limited to the most basic summary statistics. Although they can be programmed to perform statistical tests, this is outside the reach of an occasional user.

Specialised statistical software

There are so many free statistical packages that it is impossible to mention all of them[12-14]. Many of them, however, are designed for specialised tasks and are only useful to experts. On the other hand, the most comprehensive and powerful package, the R programming language and environment[15], has a command-line interface and a very steep learning curve. For the occasional user, I recommend a user-friendlier programme. My choice is Instat+[16], which is not open source software, but is free for non-commercial, personal use. In Appendix B, there are a few examples fully worked out with this programme. Instat+ (currently version 3.36) is only available for Windows. It imports/exports data from/to a spreadsheet or a text file. It offers an extensive set of statistical tests, including almost all those useful to a non-statistician, and a limited number of graphs, which can be exported in a variety of formats. The documentation, which is accessible from inside the program itself, both as hypertext and in a printable format (pdf), is particularly useful.

There is also a web site[17] that offers the possibility of performing a good number of statistical tests on line. This is useful when data are not complex and are limited to a few rows and columns. Furthermore, these web pages contain helpful remarks on the assumptions that must be satisfied in order to use each particular test correctly.

Carefully selected graphs are obviously important because they are able to convey the main messages of your study much more effectively than words and numbers. They are also invaluable during the analysis of the data, as an exploratory tool. In many cases, the graphs provided by a spreadsheet, supplemented by those available with Instat+ or a similar programme, will suffice. More complex graphs can be produced with RLPlot[18]. This programme is available for both Windows and Linux, although the Windows version is not very stable and I suggest that you save your work frequently.

In the next section, I shall present some of the most

Blood Transfus 2008; 6: 37-45 DOI 10.2450/2008.0001-08

common cases of statistical treatment of immunohaematological data.

Statistical treatment of immunohaematological data

Data types

Examples of categorical (nominal scale) data are Rh positive/negative, donor/patient/pregnant woman[19], idiopathic/methyldopa/neonate/DAT+ donor (cases of positive direct antiglobulin test)[20]. Agglutination scores (- to ++++) are ordinal data, because +++ is more than ++ but the difference between +++ and ++ is not necessarily the same as that between + and -. Data from more sophisticated scoring systems[21] appear to be measured on an interval scale (such as optical density or fluorescence) but, actually, are fundamentally ordinal and should be analysed with non-parametric tests only. Titre end-points are halfway between ordinal and interval scales: their scale is not continuous but discrete; moreover, the difference between 1:512 and 1:256 is not really lesser or greater than that between 1:4 and 1:2.

Titres

Summary statistics

Titre end-points are usually reported as the reciprocal of the last dilution found positive (i.e., 16 instead of 1:16). This has the undesirable effect of increasing the arithmetic difference between a dilution and the successive one, as the titration progresses. The arithmetic mean is particularly sensitive to this effect. An example is shown in table I. Three techniques for the indirect antiglobulin test were compared[22]: a standard agglutination tube method (Aggl), an enzyme-linked antiglobulin test (ELAT-W) and an ELAT variant in which gradient centrifugation was used instead of washing (ELAT-G). The usual summary statistics, arithmetic mean ± 1 standard deviation (SD), are clearly inappropriate: in the case of ELAT-W, the standard deviation is greater than the mean and the mean –1SD is negative. The median with the first and third quartiles (Q_1-Q_3) is more satisfactory. However, in the case of ELAT-G, the median is equal to Q_3. Dealing with titres, the geometric mean (GM) and the geometric standard deviation (GSD) are the preferred measures of location and dispersion[4]. By definition, GM is:

$$GM = \sqrt[n]{a \times b \times c \times \ldots} \quad (1)$$

where n is the number of observations and a, b, c ... are the observations. When there are many cases, calculating GM through (1) is computationally heavy. The following formula is mathematically equivalent to (1):

$$GM = e^{\left(\frac{\ln(a)+\ln(b)+\ln(c)+\ldots}{n}\right)} \quad (2)$$

where *ln* means the natural logarithm and e is the transcendental number, base of the natural logarithms. This is equivalent to saying that GM is the antilogarithm of the arithmetic mean of the log-transformed data. Similarly, GSD is equal to the antilogarithm of the arithmetic standard deviation of the log-transformed data.

Most often, titrations are obtained by doubling dilutions. In these cases, using the base-2 logarithm is particularly convenient, because it reduces to counting the tubes, excluding the neat sample ($\log_2(4)=2$, $\log_2(8)=3$ etc.) (See Appendix A for the instructions on how to calculate base-2 antilogarithms with a spreadsheet).

At this point it is appropriate to stop for a moment and reflect on the meaning of the summary statistics we have just calculated. Another example will be of help in this context. Let us suppose we are studying the occurrence of anti-Wr^a. Anti-Wr^a is usually a "natural" antibody, but may also be immune. We, therefore, compare its frequency in blood donors and multitransfused patients. We also compare the titres in the two populations of subjects. In this case, we are really interested in the central tendency and its dispersion. In other words, we attribute them a biological meaning and want to discover what is their best measure, whether it is the arithmetic mean and the standard deviation, or the geometric mean and the geometric standard deviation, or the median and the interquartile range. Let us now consider the data in table I again. It is this quite probable that the 15 antibodies were chosen for the study for their characteristics or for other practical reasons: in any case, their average titre has no biological meaning and what we are interested in is just to summarise the data in an unbiased but "expressive" way.

Tests of significance

Judging from the summary statistics, the three techniques seem to yield different results: what is the probability that the differences are casual? The three techniques were tried on the same 15 antibodies. Therefore, two categorical variables are compared at the same time: the techniques and the antibodies. We are not, in fact, interested in the variable "antibody": we already expect there to be differences among them (we could also have chosen them purposefully for that reason: e.g., a selection of high and low titre antibodies). For the integrity of the study, it is only important that we did not select them to influence the comparison between the techniques. In any

case, it is worth taking apart the variability due to the antibodies because, in this way, we decrease the "background noise" and it is easier to discriminate what is due to the techniques.

Given the experimental design, the appropriate test would normally be a two-way analysis of variance ("two-way" refers to the two categorical variables). However, the analysis of variance requires that the data be normally distributed and measured on an interval scale. These requirements are certainly not satisfied by our data. The non-parametric equivalent of the two-way analysis of variance is Friedman's test. This only requires an ordinal scale, i.e. it should be possible to rank the techniques as more or less successful with each of the antibodies. "Ties" or equal values are also possible and admitted, of course. On the other hand, missing values are not allowed: in such a case, the whole row (antibody) has to be excluded from this test. With the data in table I, Friedman's test gives a probability of 0.03% (p=0.0003) for the null hypothesis concerning the techniques [and 0.8% (p=0.008) for the antibodies]. Thus it is quite improbable that the differences observed between the techniques are just casual. Given the assumptions of the test, the results are exactly the same, whether we use raw or log-transformed data.

If the probability of the null hypothesis had been greater than the predefined significance limit, the statistical analysis should have been stopped there. In that case, it is not correct to perform multiple pairwise comparisons between the techniques: particularly when there are many of them, one is almost sure to find "significant" differences just choosing the techniques with the most extreme results. In our case, the null hypothesis is very improbable and we should continue the analysis to determine the contribution of each technique to the statistical significance. As already mentioned, this is obtained comparing the techniques to each other.

Pairwise comparisons

Two non-parametric tests are available: the sign test and Wilcoxon's matched-pairs signed-rank test. The sign test is less potent (this means that it is less likely to find a significant difference even when it should) but it is also suitable for a very rough measurement scale. Wilcoxon's matched-pairs test uses not only the sign but also the width of the difference between the results of the two techniques that are being compared. It, therefore, requires an interval measurement scale. It is doubtful that titre end-points meet this requirement. In any case, they should be log-transformed, lest the differences between the higher dilutions be considered greater than those between the lower ones. The results of the pairwise comparisons of the three techniques, presented in table I, are shown in table II. Both ELAT-G and ELAT-W give significantly higher titres than the Agglutination technique (the probability of a casual occurrence is equal or less than 1%). ELAT-G seems better than ELAT-W, but here the probability of the null hypothesis is around 10%.

As expected, the probabilities calculated by the sign test are generally, but not always, higher than those calculated by Wilcoxon's matched pairs test. In our case, the differences are very small. Greater differences in the probabilities should only be expected when the techniques compared give really different results, but in a few cases only: in the majority of them, the results are equivalent but not precisely equal. The wider differences in the titre end-points are, therefore, all in the same direction, while the smaller ones are more uniformly distributed and constitute the majority. In this situation, the sign test may well yield a misleadingly high probability for the null hypothesis. However, it is almost impossible that this could happen when data are titration end-points, which are measured on a discrete scale. Thus, I generally recommend the sign test because it yields a more prudent estimate and its assumptions are nearly always satisfied.

Table II - Multiple pairwise comparisons between the techniques of table I. Wilcoxon's matched-pairs signed-rank test is more sensitive than the Sign test and generally, but not always, gives lower p values.

Comparison	Statistical Test	
	Sign	**Wilcoxon's matched pairs**
Aggl/ELAT-W	p=0.003	p=0.012
Aggl/ELAT-G	p=0.001	p=0.0008
ELAT-W/ELAT-G	p=0.118	p=0.08

Adjustments for multiple comparisons

In rare cases, the strategy outlined above (first check the overall significance, then continue with pairwise comparisons) cannot be followed. This happens when there are only a few antibodies tested with all the techniques, but a good number of them have been tested with at least two techniques. When there are too few rows, Friedman's test is not sufficiently potent and there is the risk of not detecting a real difference. In this case, multiple pairwise comparisons are necessarily the first step. However, as

Blood Transfus 2008; 6: 37-45 DOI 10.2450/2008.0001-08

mentioned above, "significant" differences are likely to appear by chance alone, when testing many hypotheses at the same time. The significance limit should be lowered to take this into account. The simplest way is to divide the predefined significance limit by the number of comparisons (Bonferroni's correction)[23]: e.g., to maintain the overall significance level at 5% (p=0.05) when performing five comparisons, lower the limit to 1% (p=0.01). However, this correction is considered too conservative and prone to incur false negative results. An improved algorithm is the following (the Holm-Bonferroni method)[24]: apply Bonferroni's correction and consider the lowest p obtained; if it is lower than the limit, reject the null hypothesis and exclude this p from further analysis; repeat the procedure from the beginning, only considering the remaining p values; stop when the lowest p is higher than the corrected limit; this and the remaining p values are not significant. More sophisticated adjustments maximise power without increasing the false positive rate[25], but this is a field of current active research and it should be left to professional statisticians. Finally, some critics believe that no adjustment is necessary[26,27]. This opinion has been hotly debated[28-30]. It is certainly true that a mindless application of the adjustment approach may lead to absurd consequences, such as the *a posteriori* revision of p values of published papers, once new tests are performed on the same study[26,27]. Some propose relaxing the methodological requirements of exploratory studies, as opposed to confirmatory studies[29]. Others suggest reporting p values without the usual comment "statistically (not) significant"[31]. Probably, a Bayesian approach would solve this, as many other problems[32] (see below).

Independent samples

Let us forget the real origin of the data in table I and imagine the antibodies were different across the three techniques: e.g., they had been allocated to one of the techniques by consulting a table of random numbers. In this situation, the appropriate test of significance is the Kruskall-Wallis test. This is analogous to the one-way analysis of variance, but it is suitable for ordinal data. For this test, the columns need not have the same number of rows. Applied to the data in table I, the Kruskall-Wallis test gives a probability of 0.37% (p=0.0037) for the null hypothesis, i.e. about 12 times higher than with Friedman's test. This result is not surprising: the matched design allows Friedman's test to separate the variability due to the techniques from that due to the antibodies. This is the quantitative counterpart of the intuition that is preferable to use the same antibodies to compare different techniques. Kruskall-Wallis test does not use the raw data points but their ranks: the results are, therefore, exactly the same with raw or log-transformed data.

Pairwise comparisons can be performed by means of Wilcoxon's two-sample test, which is often called the Mann-Whitney U test. This version for independent samples only requires an ordinal scale. The results are the following: Aggl/ELAT-W: p=0.097, Aggl/ELAT-G: p=0.013, ELAT-W/ELAT-G: p=0.069. These probabilities are generally higher than those calculated with the test for matched pairs: the same considerations as above apply.

Scores

Scores are obviously ordinal data and they should be dealt with using the same statistical tests described for titres. In many cases, titration results are expressed as the sum of the scores obtained with each dilution. This type of data resembles interval-level data, in that its distribution is continuous. The use of Wilcoxon's matched-pairs signed-rank test with such data should not raise any objection.

Frequencies

A typical example of frequency data is shown in table III. The occurrence of anti-Wr^a, a common "natural" antibody directed against a rare antigen, was sought in blood donors, hospital patients (not thalassaemics), and multitransfused thalassaemic patients. We hypothesised that in patients, and particularly in multitransfused patients, anti-Wr^a could also be an immune antibody and, consequently, we expected a higher frequency in these patients. The appropriate statistical test is the chi-square test for equality of distributions. With the data in table III, this test gives a probability of 1.7% (p=0.017) for the null hypothesis. This is generally considered "significant". However, a closer look at the data shows that the highest frequency is in non-thalassaemic patients. Therefore, the null hypothesis seems not very probable, but at the same time the data do not correspond to the biological expectations of the alternative hypothesis. This is an important point. Every data set is compatible with infinite hypotheses ("theories")[33] and, even though it does not lend support to the null hypothesis, this should not automatically be equated with a confirmation of the alternative hypothesis.

The chi-square test requires that the observations are independent. In the case of the data of table III, this means that the subjects' subdivisions must be mutually exclusive. But what if the observations are repeated on the same

subjects? In this case the appropriate test is McNemar's test. "Repeated" observations typically mean before and after a "treatment", but the test is also suitable for a binary response to two different treatments on the same subjects/ samples. McNemar's test requires that data be presented as in table IV: we compared the false positive rate of two techniques for the indirect antiglobulin test, on 2,017 samples from hospital patients. PLIS gave 18 (0.9%) false positive results, while ID gave 14 (0.7%). According to McNemar's test, there is a probability of 28.9% (p=0.289) that the difference is due to chance alone.

McNemar's test can be extended to response patterns that are more complex than a simple binary response, or to more than two treatments. In such cases, I suggest recourse to a professional statistician.

Table III - Frequency of anti-Wr^a in three groups of subjects. The probability that the frequencies do not differ between the groups is 1.7% (p=0.017; chi-square test with continuity correction).

	Anti-Wr^a	
	+	-
Blood donors	3	97
Patients (not thalassaemics)	16	84
Thalassaemics	28	251

Table IV - Frequency of false positive results with two techniques for the indirect antiglobulin test. The probability that the two techniques do not differ is 28.9% (p=0.289; McNemar's test). (PLIS: tube test with PEG 1000 as a potentiator; ID: gel test).

PLIS	ID	
	+	-
+	12	6
-	2	1997

Limitations of the current statistical approach

What I have briefly outlined above is the prevalent view of statistics in (medical) research. However, it is fair to say that this view has attracted criticism from widely different standpoints. One contentious issue is what seems to be the introduction of subjective aspects such as the perspective of the investigator. According to the prevalent view, the scientist should agree to a sort of statistical code of ethics, avoiding any retrospective subgroup analyses ("data dredging" or "data torturing")[34,35]. Critics object that data do not change because of the researcher's motivations or according to when the hypothesis is generated[26]. In my opinion, this criticism is ill founded, as it is based on the misconception that hypotheses originate directly from data. In fact, data are compatible with infinite theories[33] and the risk to avoid is the *post hoc* selection of that exactly corresponding to the observation: it is the familiar error of assuming what should be demonstrated.

A more serious criticism is that common statistical tests of significance ignore the weight of previous evidence (biological plausibility, strength of previous results, etc.)[32,36,37] supporting the competing hypotheses. This is, in fact, true: they look at the data as if they are isolated in a vacuum. This is a defect inherent to the traditional (frequentist) statistical approach. Bayesian statistics combines in a single formula *a priori* probability (pre-existing information) and new evidence (the contribution of current data) to calculate an *a posteriori* probability. The Bayesian approach should, therefore, potentially solve this problem[32,36-38] but, at the present time, its diffusion seems to be largely limited to professional statisticians.

Keywords: statistics, statistical tests, statistical software, immunohaematology.

References

1) Ahn JH, Rosenfield RE, Kochwa S. Low ionic antiglobulin tests. Transfusion 1987; **27**: 125-33.
2) De Man AJM, Overbeeke MAM. Evaluation of the polyethylene glycol antiglobulin test for detection of red blood cell antibodies. Vox Sang 1990; **58**: 207-10.
3) Schrem A, Flegel WA. Comparison of solid-phase antibody screening tests with pooled red cells in blood donors. Vox Sang 1996; **71**: 37-42.
4) Taylor RN. Measurement of variation and significance in serologic tests. Ann N Y Acad Sci 1983; **420**: 13-21.
5) Reverberi R, Reverberi L. Removal kinetics of exchange transfusion. Blood Transfusion 2007; **5**: 93-101.
6) http://www.gnome.org/projects/gnumeric/
7) http://office.microsoft.com/en-us/access/default.aspx
8) http://www.openoffice.org/product/base.html
9) http://www.oracle.com/technology/software/products/database/index.html
10) http://www.postgresql.org/
11) http://www.mysql.com/
12) http://statistiksoftware.com/free_software.html.
13) http://freestatistics.altervista.org/en
14) http://statpages.org/
15) http://www.r-project.org/
16) http://www.rdg.ac.uk/ssc/software/instat/instat.html
17) http://www.fon.hum.uva.nl/Service/Statistics.html
18) http://rlplot.sourceforge.net/
19) Sallander S, Shanwell A, Aqvist M. Evaluation of a solid-

phase test for erythrocyte antibody screening of pregnant women, patients and blood donors. Vox Sang 1996; **71**: 221-5.

20) Garratty G, Nance SJ. Correlation between in vivo haemolysis and the amount of red cell-bound IgG measured by flow cytometry. Transfusion 1990; **30**: 617-21.
21) Marsh WL. Scoring of haemagglutination reactions. Transfusion 1972; **12**: 352-3.
22) Galluccio L, Reverberi R, Castellani A, et al. Il test dell'antiglobulina con metodica immunoenzimatica. Quad Sclavo Diagn 1985; **21**: 62-70.
23) Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. Br Med J 1995; **310**: 170.
24) Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am J Public Health 1996; **86**: 726-8.
25) Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol 2002; **64**: 479-98.
26) Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. Am J Epidemiol 1995; **192**: 904-8.
27) Perneger TV. What's wrong with Bonferroni adjustments. Br Med J 1998; **316**: 1236-8.
28) Aickin M. Other method for adjustment of multiple testing exists. Br Med J 1999; **318**: 127 [letter].
29) Bender R, Lange S. Multiple test procedures other than Bonferroni's deserve wider use. Br Med J 1999; **318**: 600 [letter].
30) Perneger TV. Adjusting for multiple testing in studies is less important than other concerns. Br Med J 1999; **318**:1288 [letter].
31) Sterne JAC, Davey Smith G. Sifting the evidence-what's wrong with significance test? Br Med J 2001; 322:226-31.
32) Goodman SN, Royall R. Evidence and scientific research. Am J Public Health 1988; 78:1568-74.
33) Popper KR. *The logic of scientific discovery*. New York, NJ, Harper and Row, 1959.
34) Oxman DA, Guyatt GH. A consumer's guide to subgroup analysis. Ann Intern Med 1992; **116**:78-84.
35) Mills JL. Data torturing. New Engl J Med 1993; **329**:1196-9.
36) Goodman SN. Toward evidence-based medical statistics. 1: the p value fallacy. Ann Intern Med 1999; **130**:995-1004.
37) Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med 1999; **130**:1005:13.
38) Gurrin LC, Kurinczuk JJ, Burton PR. Bayesian statistics in medical research: an intuitive alternative to conventional data analysis. J Eval Clin Pract 2000; **6**:193-204.

Correspondence: Dr. Roberto Reverberi
Servizio di Immunoematologia e Trasfusionale Arcispedale S.Anna
Corso Giovecca 203
44100 Ferrara - Italy
e-mail: sitfe@ospfe.it

Appendix A –Programming a spreadsheet*

(The reader will find a brief introduction to spreadsheets in reference 5).

Base-2 logarithms and antilogarithms

Spreadsheets have built-in functions for calculating logarithms in any base but it is possible to obtain directly only the natural antilogarithm. However, the base-2 antilogarithm is very easily calculated: open a new sheet; enter text, values and formulae listed in table V. Cells B4 and B5 contain the base-2 logarithm and antilogarithm, respectively.

Table V - Instructions to calculate base-2 logarithms and antilogarithms

CELL	TEXT TO BE ENTERED
A1	Base-2 logarithms and antilogarithms
A3	Enter the number:
A4	Base-2 logarithm:
A5	Base-2 antilogarithm:
	VALUE TO BE ENTERED
B3	The number on which to calculate base-2 logarithm and antilogarithm, e.g. 512
	FORMULA TO BE ENTERED
B4	=LOG(B3,2)
B5	=EXP(B3*LN(2))

Geometric mean and geometric standard deviation

Open a new sheet. Enter the data into column A, starting from cell A3. Enter text, values and formulae listed in Table VI. Copy the formula in B3 along column B down to the last row that contains data. Enter into cell D4 the address of this last cell (e.g., B17, when there are 15 data points). Column B contains the natural logarithm of the data. Cell D5 contains the geometric mean; cells D6 and D7 contain the geometric mean minus/plus one geometric standard deviation, respectively.

* Italian readers using the localized (Italian) versions of the spreadsheets should follow the instructions in the Italian translation of this paper, which is available on line at http://www.bloodtransfusion.it. Briefly, "media" should be substituted for "average", "indiretto" for "indirect", "dev.st" for "stdev" and ";"for ",".

Table VI - Instructions to calculate the geometric mean ± 1 geometric standard deviation

CELL	TEXT TO BE ENTERED
A1	Geometric mean and geometric standard deviation
C3	First row:
C4	Last row:
C5	Geometric mean:
C6	GM-1GSD:
C7	GM+1GSD:
	VALUE TO BE ENTERED
D3	The first cell containing the natural logarithm of the data: B3
D4	The last cell containing the natural logarithm of the data: e.g., B17
	FORMULA TO BE ENTERED
B3	=LN(A3)
D5	=EXP(AVERAGE(INDIRECT(D3): INDIRECT (D4)))
D6	=EXP((AVERAGE (INDIRECT (D3): INDIRECT (D4)) - STDEV(INDIRECT (D3): INDIRECT (D4))))
D7	=EXP((AVERAGE (INDIRECT (D3): INDIRECT(D4)) + STDEV(INDIRECT (D3): INDIRECT (D4))))

Appendix B –Using Instat+

Instat+ can be downloaded from the web page of reference 16. The installation is straightforward but may require the Microsoft Windows Installer and/or the Microsoft Data Access Component, if not already installed.

On opening, Instat+ shows two panels, one containing a spreadsheet-like list of columns and rows, and the other a space in which commands are recorded and results shown. Commands are recorded even if you interact only with the menu.

Entering data

It is easier to use a normal spreadsheet for data entry. You should enter the data in columnar format, with columns representing variables. When you are ready, just select the cells containing your data, choose **Edit → Copy**, clic on the top left cell of the Instat+ worksheet, choose **Edit → Paste** in the menu of Instat+ and you are done. Columns are identified by labels X1 ... Xn from left to right, but you can enter a meaningful name for the variable in the cell just below the column label. Instat+ can also import data from Excel or Access automatically, but in my experience those functions are not reliable. Once your data have been imported, save the worksheet for future use.

Friedman's test

We will use the data in Table I. There are three column variables (Aggl, ELAT-W, and ELAT-G) and 15 row variables (antibodies), numbered 1 to 15. For the test, data need to be rearranged in this way: put all titre end-points in a single column and add two more columns: "Tech" and "Case". For each row, enter the technique (Aggl or ELAT-W or ELAT-G) in the cell under column "Tech", and the antibody (1 to 15) in the cell under column "Case". At the end, there are three columns of 45 rows. Select this area and paste it into Instat+ (see above for instructions). You have to inform Instat+ that columns "Tech" and "Case" are factor columns: choose **Manage → Column Properties → Factor**, select the column and press **Apply**.

Choose **Statistics → Non parametric → Two way ANOVA**. Select the column containing the titre end-points in the field **Counts**, "Case" in the field **Row factor**, "Tech" in the field **Column factor**, and press **OK**. The results will appear in the **Command and Output** panel. Two probabilities are calculated, one for

Blood Transfus 2008; 6: 37-45 DOI 10.2450/2008.0001-08

the row factor ("Case") and the other for the column factor ("Tech"). With our data the probabilities are 0.008 and 0.0003, respectively.

Sign test

We will compare Aggl and ELAT-W from Table I. In your preferred spreadsheet, enter the data under columns "Aggl" and "ELAT-W", leaving them in two separate columns. In a third column ("Diff"), enter the difference between the values of columns "Aggl" and "ELAT-W". Select the cells of column "Diff" and paste them into Instat+. Choose **Statistics → Non parametric → One and two samples**. Choose **Sign test** and select column "Diff" in the field **Data column**. Press **OK**. The result will appear in the **Command and Output** panel.

Wilcoxon's matched-pairs signed-rank test

We will reuse the same data of the Sign test. However, for the reasons explained in the main text, we should apply the test to log-transformed data. Put the logarithms (any base will serve the purpose) of the titre end-points into two separate columns ("Log_Aggl" and "Log_ELAT-W"). Select this area and paste it into Instat+. Choose **Statistics → Non parametric → One and two samples**. Choose **Wilcoxon**. Select **Two data columns** in the **Layout** field. Select column "Log_Aggl" in the field **Data Column** and column "Log_ELAT-W" in the field **2nd data column**. Choose **Paired** and **Continuity correction**. Press **OK**. The probability is expressed as a percentage: in our case it is 1.2% (for a two-sided test), which is equivalent to p=0.012. The one-sided probability should only be chosen if you had very good reasons to expect, when you planned the study, that one technique should be better or equal, but not worse, than the other.

The Kruskall-Wallis test

Data should be arranged as for Friedman's test, but in this case there is only one factor column. Choose **Statistics → Non parametric → One way ANOVA**. Select the column containing the data in the field **Y variable** and the factor column in the field **Factor**. Press **OK**.

Wilcoxon's two-sample test (the Mann-Whitney U test)

Instructions are the same as for the paired version, only do not choose **Paired**. In this case it is not necessary to use log-transformed data (the results would be the same, anyway).

Chi-square test

For this and McNemar's test, we will use the web page of reference 17 instead of Instat+. Choose **Chi-square test for equality of distributions**. Clean the box in which you can enter the data. Enter data, including row and column names (use only normal alphabetic characters; avoid composite names; the alignment is not important). Press **Submit**. The probability is calculated with the correction for the continuity, which yields a more prudent estimate. Instat+ offers the option of the continuity correction for 2 x 2 tables only.

McNemar's test

Choose **McNemar's test**. Enter data arranged as in Table III, i.e. with discordant results in the bottom-left and top-right cells. Press **Submit**.