Statistics in review
Part I: graphics, data summary and linear models

John L Moran and Patricia J Solomon

ABSTRACT

Statistics and biomedical literature have historically had an uneasy alliance. A critical approach to the application of statistics is developed. Initially, we survey graphical data display and trace the historical development of the "testing" statistical paradigm, and the contributions of A R Fisher and J Neyman and E Pearson. The nuances of data summary and testing are illustrated by way of population versus sample estimation. The importance of the normality assumption is stressed, and the recurring contrast of parametric (*t* test) versus non-parametric (Mann–Whitney) approaches to summary statistics is discussed. The *t* test is found to be adequate. Effect measures are outlined, and we demonstrate the utility of the unpaired *t* test for binary data analysis. The theory of linear models is introduced, and the underlying assumptions of the standard ordinary least squares regression are presented. The implications of transformations, in particular log transformation, are detailed, and we conclude with an overview of the principles of model selection.

Crit Care Resusc 2007; 9: 81–90

Why another review of statistics? Presumably because the conduct of "statistics" in the medical literature has been found to be consistently poor,[1] the transfer of innovative statistical techniques into the medical literature has been characterised by significant time-lags,[2] and statistical input into medical research and publication, although "widely recommended … [is] … inconsistently obtained".[3] It may also reflect an undervaluation of statistical contributions to medicine, as articulated by the doyen of biostatistics, Norman Breslow.[4] He observed that the work of econometricians Daniel McFadden and James Heckman on discrete choice models and selection bias received a Nobel Prize in 2000, but that similar contributions to medicine by statisticians and epidemiologists remain, as yet, unrecognised. Thus, we must "grapple" with statistics in the same manner as Appleby urged with respect to health economics.[5] To this extent, the now dominant evidence-based medicine movement has mandated "critical appraisal", which incorporates, to varying degrees, statistical methods,[6] and we reiterate that the discipline of statistics is increasingly engaged with "front-line science".[7]

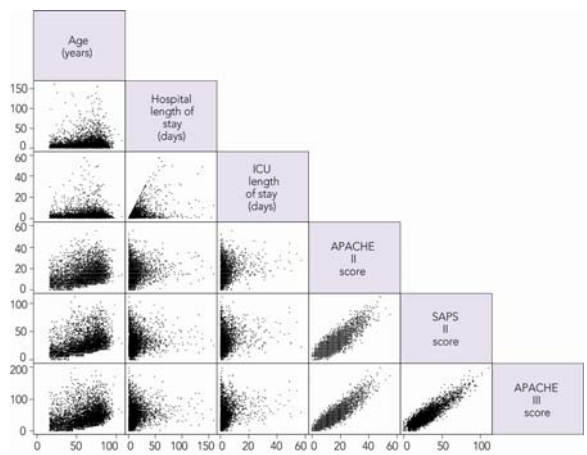
Statistics as we know and practise it today had its foundation in the first half of the 20th century and was established — more particularly, the "testing" paradigm of *P* values and Type I and II errors[8] — by two dominant figures, R A Fisher (1890–1962), who was born in England and died in Adelaide, South Australia (see http://digital.library.adelaide.edu.au/coll/special/fisher/index.html),[9] and Jerzy Neyman (1894–1981), who was born in Russia and died in Berkeley, California, USA.[10]

This discursive introduction does not portend a compression-in-miniature of the standard textbook presentation of medical statistics;[11] rather, the purpose is to highlight directions for those conducting their own analyses or reviewing the literature "critically".

Graphical display

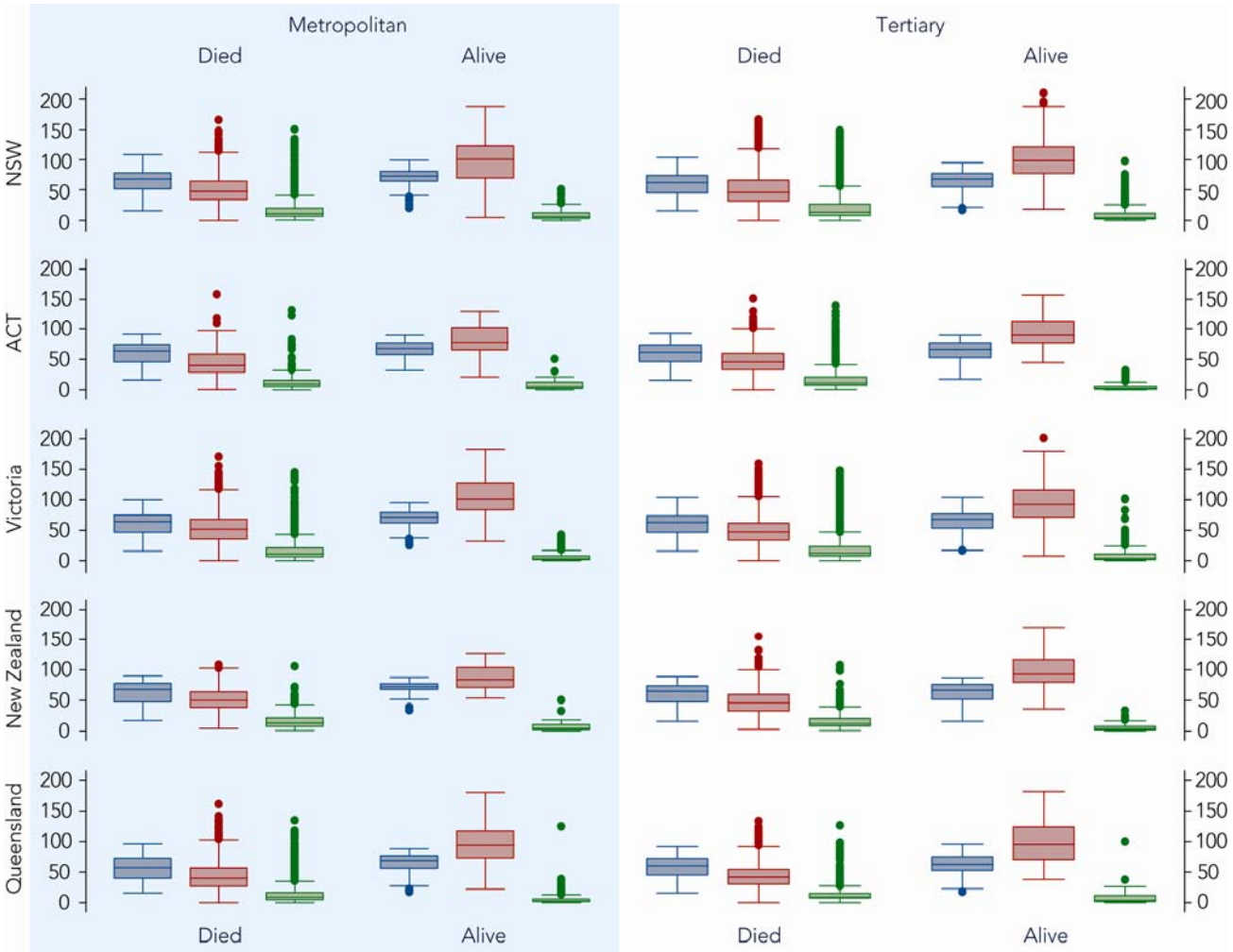
Graphical display[12-14] is fundamental to our interpretation of data,[15] especially in large multi-variable data sets, where, for instance, a scatter-plot matrix is useful (Figure 1).

Figure 1. Scatter-plot matrix of patient database variables*



* Data for this and other figures and examples were obtained from the Australian and New Zealand Intensive Care Society (ANZICS) Adult Patient Database (1993–2003) with permission of the ANZICS Database Management Committee.

Figure 2. Trellis box plots of age, APACHE III score and hospital length of stay for ICU–hospital level (metropolitan or tertiary) against geographic locality for patients alive and dead*



* ***Key:*** age = blue; APACHE III score = pink; hospital length of stay = green.

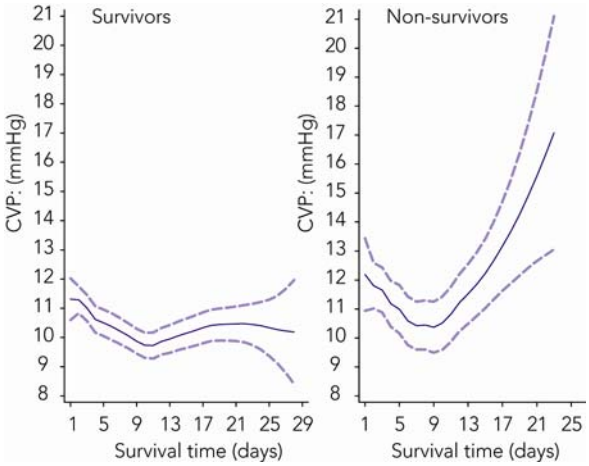
This may be variously extended to, say, a trellis plot (Figure 2) based on a formula with the structure

$$y \sim x \mid a \times b$$

where y is a continuous or factor variable, x is continuous, and a and b are factors.[16] Smoothing techniques are also useful for time profiling, and we illustrate this using a running line smoother[17] for time change of central venous pressure in survivors and non-survivors after acute lung injury (Figure 3).

Although there are standard tests for "non-normality" (Shapiro–Wilk and Shapiro–Francia),[18,19] graphical display is of value, particularly quantile-normal plots (which emphasise the distribution tails, normal-probability plots (which emphasise the distribution centre),[20,21] and kernel density plots. The latter, a useful general graphical tool, are a modification of the histogram (a "smoothed" histogram), where densities are the continuous analogues of proportions (derivatives of the cumulative distribution function, so that areas under the density function read off as probabilities). The data are divided into intervals (which may overlap), and estimates of the density at the interval centres are produced; the "kernel" is the function (a number are available) that weights the observations by the distance from the centre of the interval.[22] Figure 4 shows normal-probability and quantile-normal plots (a "normal" distribution approximates to the 45° line), and a conventional histogram of patient APACHE III scores[23] (n = 4408) from a number of Australian intensive care units, with

Figure 3. Central venous pressure (CVP) profile for survivors and non-survivors after acute lung injury



Profiles show running line smooth (solid line) and 95% confidence limits (dotted lines).

superimposed normal (solid line) and kernel (dashed line) density plots.

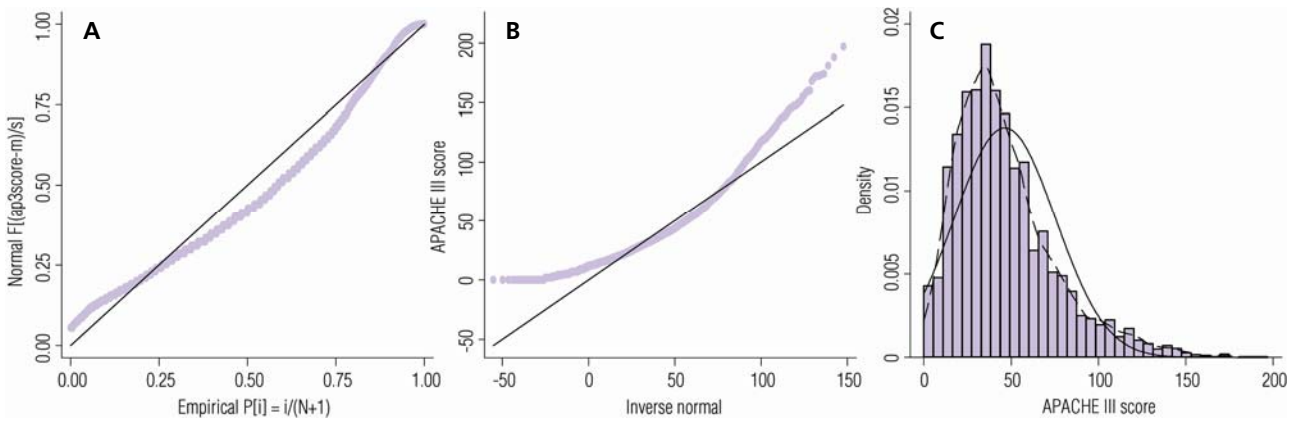
P values and confidence intervals

Since the establishment of the "testing" paradigm in the 1920s and 1930s by A R Fisher, Jerzy Neyman and Egon Pearson,[24] the status of *P* values in the scientific literature has been problematic.[25] As previously reviewed,[26] the Fisher significance test derives from inductive inference to establish a null hypothesis (H_0) and to use data discrepancies to reject this hypothesis. The associated *P* value was deemed the probability of obtaining a result equal to or more extreme than what was actually observed. The deductive revision of Fisher's position by Neyman and Pearson formulated the now familiar two competing hypotheses paradigm (effectively rules for making decisions): the null (H_0) and alternate (H_A) hypotheses; and the probability of committing two kinds of errors: false rejection (Type I or α error) and false acceptance (Type II or β error). The α error probability thus had the interpretation that a series of α level tests will reject no more than 100α% of true H_0 (in the long run). Confidence intervals (CIs), introduced by Neyman in 1937, were considered integral to the overall theory of hypothesis testing, and the interpretation was *not* that of a probability interval. Rather, in an infinite number of repetitions of a study, an exact proportion (say, 95%) of all such intervals would enclose the parameter θ for a 95% CI. Importantly, once the data had been collected, and a single 95% CI had been calculated, the probability that θ lay within this CI was now 0 or 1.[27]

Data summaries and tests

Initially, we stress the importance of the normality assumption.[28] If the population mean (location parameter) is μ, and the standard deviation (SD, or scale parameter) is σ, then the value of a normal curve (the

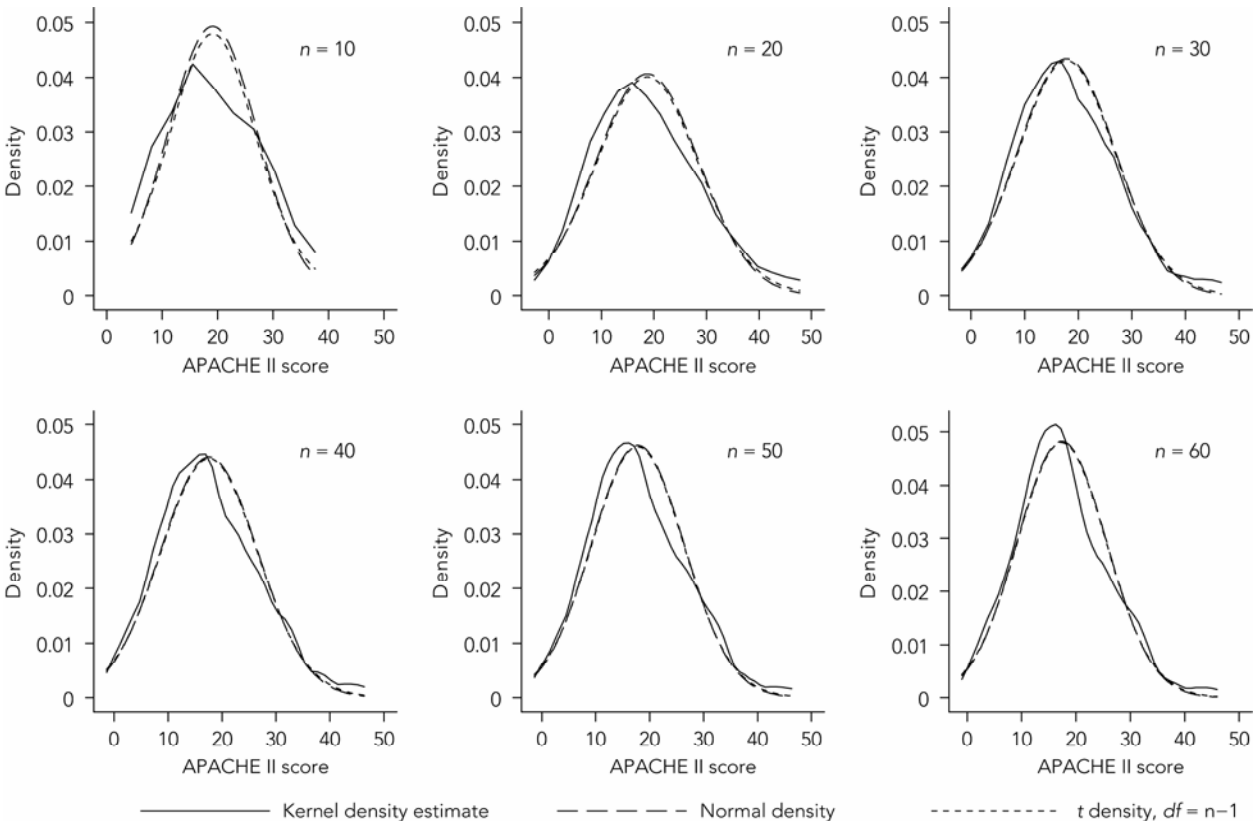
Figure 4. Normal-probability, quantile-normal and kernel density plots of patient APACHE III scores



A and B. Normal-probability and quantile-normal plots of patient APACHE III scores (n = 4408) from Australian intensive care units.[23] A "normal" distribution approximates to the 45° line.

C. A conventional histogram of the scores, with superimposed normal (solid line) and kernel (dashed line) density plots.

Figure 5. Effect of sample size (*n*) on density distributions



probability density function) is[29]

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

and the familiar *z* score (standard normal deviate, written N (0, 1)) is defined as

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

and, correspondingly,

$$x = \mu + z\sigma\ .$$

One SD corresponds to $z = 1$ (68.26% of the distribution within 1 SD of the mean), and the familiar 95% of the distribution lies within 1.96 SDs of the mean. By the central limit theorem, the sampling distribution of a mean is normally distributed; the mean of this sampling distribution is therefore μ, and the SD (or the standard error [SE] of the sample mean) is $\sigma/\sqrt{n}$, where n is the sample size. Therefore for "large" samples (say, $n \geqslant 60$,[30] although this may be a "generous" n if the distribution is fairly symmetric with tails that decay rapidly, when the more often quoted $n \leqslant 30$–35 seems sufficient[31]), the 95% confidence limits of the sample mean

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

are given by

$$\overline{x} \pm 1.96 \times \text{SE},$$

where the SE of $\overline{x}$ is now estimated by $s\sqrt{n}$, s being the sample standard deviation. The difference between means (where variances are known), using the *z* test, is given by

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}}$$

(if variances are estimates, the statistic is a *t* statistic).

For "small" samples, the *t* distribution (introduced in 1909 by William Gosset) is appropriate, and the larger tails (for smaller n) of a bell-shaped distribution are

Table 1. Estimates of central tendency for ICU length of stay

Variable	Estimate	*n*	Mean	95% CLs	
				Lower	Upper
ICU length of stay (days)	Arithmetic	4408	3.05	2.91	3.19
	Geometric	4408	1.76	1.71	1.81
	Harmonic	4408	1.15	1.12	1.19
	Quadratic	4408	5.58	5.15	5.98

CL = confidence limit.

determined by the degrees of freedom (*df*, $n - 1$). The 95% confidence limits are given by

$$\bar{x} \pm \left(t' \times \left(\frac{s}{\sqrt{n}} \right) \right)$$

where t' gives the percentage points of the t distribution.[30] The difference between means (using the pooled variance), using the t test, is

$$t = \frac{x_1 - x_2}{s\sqrt{1/n_1 + 1/n_2}}, df = n_1 + n_2 - 2$$

where s is given by

$$\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \right)}.$$

For unequal variances, the t statistic requires the calculation of "approximate" degrees of freedom.

The effect of sample size (n) on density plots for random samples of APACHE II scores (total n = 223 129), with normal and t density distributions overlaid, is shown in Figure 5.

In reports of observational studies or controlled trials, there are invariably initial data summaries[32] with, perhaps, group comparisons reporting the (sample) mean of variables of interest ($\bar{x}$) in tabular or graphical form. Other measures of central tendency are also available:

the geometric mean ($\sqrt[n]{\prod x_n}$, where $\prod$ is the "product of all the *x*s", and $\sqrt[n]{\ }$ is the nth square root[33]);

the harmonic mean ($n / \frac{1}{\sum x_n}$);

the quadratic mean (root mean square) $\sqrt{\frac{\sum x_i^2}{n}}$; and

formal trimming of the arithmetic mean.[34,35]

Note that the mean of log-transformed values is the geometric mean of the original data. For ICU length of stay, the four estimates of central tendency are shown in Table 1.

Two sometimes vigorous debates have attended such initial data presentations and tests:

The use of the standard deviation (SD)

$$\sqrt{\frac{\sum \left(x - \bar{x}\right)^2}{n - 1}}$$

versus the standard error (SE) $s/\sqrt{n}$ (for binary data, SE of a proportion is

$$\sqrt{\frac{p\,(1 - p\,)}{n}}).$$[36]

The SE, reflecting the variability of the mean (if the study were repeated a large number of times), is "not particularly useful", and the SD, reflecting the variability of the original data, should be reported[37] and has been mandated for some journals.[38]

The use of parametric or non-parametric summary statistics and tests when biomedical data, such as length of stay, length of ventilation, costs and risk of death are "non-normally" distributed with kurtosis ("peakedness" of the distribution) and/or skewness (usually a long right tail).[39,40] The t test has demonstrated remarkable robustness in the face of small n non-normal data, as opposed to the often recommended (non-parametric) Mann–Whitney test[41] (which is not a test of medians, but rather a test of the equality of group mean ranks[42] — that is, location and shape), and under most circumstances the t test is to be preferred.[43] Permutation and bootstrap techniques which relax the assumptions of normality and equal variances (respectively and cumulatively) may be utilised, but are computer-intensive, especially for more-than-small studies.[44] When examining the statistical significance of the difference between two means, the method of "95% CI overlap" (mean ± 1.96 × SE) is noted to be conservative and cannot be substituted for formal hypothesis testing of the difference; however, the abutting of the 83% CI (approximately equal to mean ± 1.4 × SE) corresponds with a P of approximately 0.05.[45,46]

Group comparisons normally proceed by conventional tests: for continuous data, the t test (theory should dictate when to use equal or unequal variances); and for

Table 2. 2 × 2 table analysing an imaginary trial of therapy versus control with respect to mortality outcome

	Dead	Alive	Total
Therapy	57 ("a")	93 ("b")	150
Control	80 ("c")	80 ("d")	160

categorical data, Fisher's exact (a permutation test[47]) or the χ^2 test (introduced by Karl Pearson[48]). With respect to the use of the χ^2 versus Fisher exact test, the general rule of thumb is that the χ^2 approximation works well, provided all cell frequencies are > 5; if any cell frequencies are $\leqslant 5$, Fisher's exact test is indicated, although claims have been made for the preferential use of the latter test for inference under conditions of randomisation.[49] For the approximation of the binomial distribution by the normal, the following guides have been formulated: $np(1-p) > 9$ and $np > 5$ for $0 < p < 0.5$ and $n(1-p) > 5$ for $0.5 < p < 1$ (where p is the binomial probability $[0 < p < 1]$).[50] The unpaired t test may also be used for binary data,[51] and we illustrate this little appreciated fact by contrasting the probability of success (scored 1) of two binary series (scored 0 or 1). The sample estimates are: mean of A = 0.5, and mean of B = 0.6.

A: 1 1 0 1 0 1 0 0 1 1 0 1 0 0 1 1 0 1 1 1

B: 1 1 0 1 0 0 1 0 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 1 1 1 0 0 0 0

The t test: $t = -0.6871$; $df = 48$; $P = 0.4953$;

Pearson's χ^2 test with Yates' continuity correction: $\chi^2 = 0.1644$; $df = 1$; $P = 0.6852$; and

Pearson's χ^2 test without Yates' continuity correction: $\chi^2 = 0.4831$; $df = 1$; $P = 0.487$ (note that this is also apparent comparing the actual statistics: $t^2 = 0.472$).

Effect measures

The treatment effects of a randomised trial are variously reported.[52] In Table 2, we analyse an imaginary trial of innovative "therapy" versus control (placebo) with respect to mortality outcome in a condition with a baseline control mortality (risk) of 50%, and visualise the results in a familiar 2 × 2 table with the cells also classified using the "a, b, c, d" terminology.

The risk of therapy is a/(a+b) = 0.38, and of control is c/(c+d) = 0.5. The odds of death are a/b = 0.61 for therapy, and c/d = 1.0 for control, and the risk = odds/(1 + odds). The risk ratio (RR) is [a/(a+b)]/[c/(c+d)] = 0.38/0.5 = 0.76 (95% CI, 0.59–0.98), and the odds ratio (OR) is (a/b)/(c/d) = 0.61/1.0 = 0.61 (95% CI, 0.39–0.96); RR = OR/[1 + I_c(OR – 1)] where I_c is the event incidence in the control group (under the "rare" event assumption). The risk difference (RD) is [a/(a+b)] – [c/(c+d)] = 0.38 – 0.5 = – 0.12 (95% CI, – 0.23 to – 0.01). The one-sided Fisher exact test gives a P value of 0.02, and the two-sided 0.04. The number needed to treat (NNT) = 1/RD = 1/0.12 = 8 (number of avoided events per 1000 population: 10–230, for baseline control risk of 0.5). We adjudge our innovative therapy as efficacious.

Odds ratios have better statistical properties than RRs or RDs and are the key parameter in the linear logistic regression model.[53,54] The (log) odds scale is unbounded in both directions, but is numerically greater than the risk ratio when underlying event rates are frequent.[55] Risk ratio has been found to be more intuitive than OR,[56] but is bounded above in a manner dependent on the control group risk. Risk difference is immediately intuitive and expresses the consequences of no therapy (unlike both ORs and RRs), but is constrained from – 1 to 1, and may suffer from bias with variable time to follow-up. An advantage of RD is that it enables an NNT and its confidence interval to be conveniently estimated. However, the NNT is also affected by the baseline risk, and recent cautions have been expressed about the properties of this statistic.[57]

Linear regression models

Fundamental to data analysis, beyond "simple" descriptive statistics and graphical display, are linear models, the general form of which is

$$Y = f\left(X_1, \mathrm{K}, X_p\right) + \varepsilon$$

Where $f\left(X_1, \mathrm{K}, X_p\right)$ is an expectation function (of the dependent variable Y), and $\boldsymbol{\varepsilon}$ is the error term.[58] The most well-known linear model is the simple linear regression model:[59]

$$y = \alpha + \beta x + \varepsilon$$

where y is the dependent variable, $\boldsymbol{\alpha}$ the intercept, x the independent (predictor) variable(s) with associated

parameter(s) β, and ε the "error" term(s). We can further distinguish additive (multiple) linear models

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and multiplicative linear models, the simplest form of which is the interaction effect

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \left(x_1 . x_2 \right) + \varepsilon$$

Non-linear models (which we do not further discuss) are similar to linear models, but the expectation function does not have to be linear in all parameters; thus, the Michaelis–Menton model from pharmacokinetics

$$y_i = \theta_1 x_i / \left(\theta_2 + x_i \right) + \varepsilon_i$$

where θ is used (instead of β) to indicate this distinction.

Simple linear regression

Simple linear regression, a form of the ordinary least squares model (OLS),[60] minimises the sum of squared errors:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where Y_i is the observed value, and $\hat{Y}_i$ is the predicted value (this may also be expressed as

$$\sum_{i=1}^{n} \left(Y_I - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right)^2$$

where $\hat{\beta}$, say, is the *estimate* of β). The assumptions of OLS are conventionally:

- Dependent observations are assumed to be independent with a common constant variance;
- Independent variable(s) are measured without error (the statistical evaluation of measurement error in the dependent variable addresses the various concepts associated with reliability studies,[61,62] including Deming regression which incorporates errors in the dependent and independent variables[62]);
- The mean value of the dependent variable is a linear function of the (combinations of) independent variables;
- There is a lack of perfect (multi-)collinearity;
- For the error term(s): the variance is constant for dependent variable combinations (homoscedasticity), they are normally distributed with mean = 0, with constant variance (these assumptions are termed iid: independent and identically distributed; that is the variables have the same probability distribution and are mutually independent), and no serial correlation nor correlation with the independent variables; and
- There is no substantive effect of outliers.[63]

Various methods have been developed to compensate for the common violations of these requirements, such as:

- "robust" variance and clustering adjustments to overcome observation non-independence;
- so-called "Newey–West" errors to adjust for serial correlation;[64] and
- measurement error models which model covariate measurement error (for an additive measurement error, the effect is to bias the estimated coefficient ($\hat{\beta}$) towards the null[65]).

Depending on the effect size, the number of "subjects" required in OLS is usually around four to five per predictor variable, although other larger requirements have been formulated (from $50 + m$ up to $50 + 8m$ [where m is the number of predictors]).[66]

The dependent variable is often transformed to achieve linearity of effect, rather than "normality" (a common misconception), which is not often achieved;[67] the most common transformation is logarithmic.[68] Under this transformation, the interpretation of (continuous) predictor (independent) variables is that they show the percentage change in the untransformed dependent variable per one unit change in the predictor variable (if in original units); if the predictor has also been log transformed, the coefficient is interpreted as the percentage change in the untransformed dependent variable for a 1 per cent change in the untransformed predictor variable. For categorical variables, this translation is somewhat biased.[69] With log transformation, recovery of predicted values (of the dependent variable) in the original metric is not transparent; simple exponentiation is biased (it recovers the geometric mean). Alternative methods are available as:

- exponential of the sum of prediction + one half of the square of the regression root mean square ($\exp[\text{prediction} + 0.5*(\text{RMSE})^2]$); and
- Duan's smearing estimate.[70,71]

Model selection

We may ask of any regression analysis that:

The covariate selection process is specified; we may be suspicious of (automated) stepwise selection as resulting in potentially biased and unstable models;[72,73]

The functional form of continuous variables is investigated; categorisation (eg, median splitting) is not necessarily a good idea (it wastes data, inflates Type I error rates and produces biased increments in coefficient estimates);[74] continuous variables may have non-linear effects which may be best displayed by fractional polynomials,[75] splines or other smoothing techniques;[76]

Interaction(s) as reflecting entirely plausible model mechanisms[58,77,78] should be formally addressed, albeit there is an increased sample size required for their detection compared with main effects, and an associated decreased power of finding interaction terms with categorical variables;[79]

Centring ($x - \overline{x}$) of continuous covariates may aid in interpretation and reduce collinearity (ie, reference to the mean covariate values rather than the default regression reference value of "0");

Variable (multi-)collinearity should be reported, using the variance inflation factor (VIF) and condition number (CN) (where VIF < 10 and CN < 30 are apposite[80]), and should be addressed;

Model "fit" should be explored using:

- residual analysis (eg, normal distribution of residuals and no evidence of heteroscedasticity[71]); the functional form of continuous predictors may be revealed via analysis of specific residuals;
- specific indices, such as
  - mean absolute error, RMSE[68] and R^2 in OLS and logistic[81] regression, noting that transformations of the dependent variable may affect the magnitude of R^2 (increased with log transformation[82]);
  - for logistic regression, the conventional criteria of discrimination (area under receiver operating characteristic [ROC] curve[83] and calibration [Hosmer–Lemeshow $\hat{C}$ statistic];[84] the latter test should be interpreted with caution for large data sets as the P value is invariably "significant", at $P << 0.1$ for a $\hat{C}$ statistic >> 15.99[85]);
  - for Cox hazard regression;[86] Harrell's C statistic;[73] the May–Hosmer goodness-of-fit statistic, testing for proportional hazards and approximation of cumulative Cox–Snell residuals to (– log) Kaplan–Meier estimates;[87]
  - for parametric (accelerated failure time [AFT]) survival models, approximation of cumulative Cox–Snell residuals to (– log) Kaplan–Meier estimates, plotting log(time) against a linear function of the cumulative (Nelson–Aalen) hazard rate;[88]
- likelihood ratio tests and information criteria, such as the Akaike information criterion (AIC: – 2*(model log-likelihood) + 2*p, where p = number of parameters) and Bayesian information criterion (BIC: – 2*(model log-likelihood) + log(n)*p, where scalar model differences of $\geqslant 10$ are adjudged as meaningful);[89] these may be used for model comparisons. Likelihood ratio and AIC assume nested comparisons (those in which covariates in one model form a subset of the covariates in a larger model, and formal goodness of fit to the data can be compared using standard tests, such as likelihood ratio tests), whereas BIC does not assume nested models.

Standard individual parameter tests reported by statistical software programs ("Wald" tests: estimate/SE) are fallible and not invariant under transformations of the parameters,[90] and, in logistic and time-to-event regression models, likelihood ratio tests are preferable;[91]

Predictive models generated on a single data set are known to have inferior performance when applied to independent data sets,[92,93] and methods to reduce such bias have been developed, such as cross validation (data splitting) and bootstrap validation;[94,95]

In general, there is no independence of model selection and inference,[96] although any "dependence" may be a function of limited data size;

The impact of a covariate or risk factor (the measure of association indicated by β) may be insufficient to determine, say, the public health implication of the covariate, without knowledge of the prevalence, as realised in the "attributable risk".[97]

Conclusion

The principles of linear modelling, having been established, will be further illustrated in Part 2 of this review, which will appear in the next issue. Part 2 will discuss generalised linear models, time-to-event and time-series analysis.

Acknowledgements

We thank the Australian and New Zealand Intensive Care Society (ANZICS) Database Management Committee for permission to use data from the ANZICS Adult Patient Database (1993–2003) (http://www.anzics.com.au).

Author details

John L Moran, Senior Consultant[1]
Patricia J Solomon, Associate Professor[2]
1 Department of Intensive Care Medicine, Queen Elizabeth Hospital, Adelaide, SA.
2 School of Mathematical Sciences, University of Adelaide, Adelaide, SA.
Correspondence: john.moran@adelaide.edu.au

References

1 Altman DG. Statistics in medical journals: some recent trends. *Stat Med* 2000; 19: 3275-89.
2 Altman DG, Goodman SN. Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. *JAMA* 1994; 272: 129-32.
3 Altman DG, Goodman SN, Schroter S. How statistical expertise is used in medical research. *JAMA* 2002; 287: 2817-20.
4 Breslow NE. Are statistical contributions to medicine undervalued? *Biometrics* 2003; 59: 1-8.
5 Appleby JL. Why doctors must grapple with health economics. *BMJ* 1987; 294: 326.
6 Morris RW. Does EBM offer the best opportunity yet for teaching medical statistics? *Stat Med* 2002; 21: 969-77.
7 Efron B. Bayesians, frequentists, and scientists. *J Am Stat Assoc* 2005; 100: 1-5.
8 Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (a's) in classical statistical testing. *Am* . 2003; 57: 171-82.
9 Ludbrook J. R.A. Fisher's life and death in Australia, 1959-1962. *Am Stat* 2005; 59: 164-5.
10 Lehmann EL, Reid C. In Memoriam: Jerzy Neyman, 1894-1981. *Am Stat* 1982; 36: 161-2.
11 Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. 4th ed. Oxford: Blackwell Scientific Publications, 2002.
12 Cleveland WS. Visualizing data. Summit, NJ: Hobart Press, 1993.
13 Cleveland WS. The elements of graphing data. Summit, NJ: Hobart Press, 1994.
14 Wilkinson L. Graphical displays. *Stat Methods Med Res* 1992; 1: 3-25.
15 Gelman A, Pasarica, C, Dodhia, R. Let's practice what we preach: turning tables into graphs. *Am Stat* 2002; 56: 121-30.
16 Heiberger RM, Holland B. Statistical analysis and data display. New York, NY: Springer Science+Business Media Inc, 2004.
17 Royston P, Cox NJ. A multivariable scatterplot smoother (gr0017, "mrunning"). *Stata Journal* 2005; 5: 405-12.
18 Miller RGJ. Beyond ANOVA: basics of applied statistics. Boca Raton, Fl: Chapman and Hall/CRC, 1997.
19 Fox J. Describing univariate distributions. In: Fox J, Long JS. Modern methods of data analysis. Newbury Park, Calif: Sage Publications, 1990: 58-125.
20 Gan FF, Koehler KJ. Goodness-of-fit tests based on P-P probability plots. *Technometrics* 1990; 32: 289-303.
21 Gan FF, Koehler KJ, Thompson JC. Probability plots and distribution curves for assessing the fit of probability models. *Am Stat* 1991; 45: 14-21.
22 Wilcox RR. Kernel density estimators: an approach to understanding how groups differ. *Understanding Statistics* 2004; 3: 333-48.
23 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13: 818-29.
24 Gigerenzer G, Swijtink Z, Porter T, et al. The empire of chance: how probability changed science and everyday life. New York: Cambridge University Press, 1989.
25 Sterne JA, Davey SG. Sifting the evidence – what's wrong with significance tests? *BMJ* 2001; 322: 226-31.
26 Moran JL, Solomon PJ. A farewell to P-values? *Crit Care Resusc* 2004; 6: 130-8.
27 Neyman J. Silver Jubilee of my dispute with Fisher. *J Oper Res Soc Japan* 1961; 3: 145-54.
28 Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. *Annu Rev Public Health* 2002; 23: 151-69.
29 Evans M, Hastings N, Peacock B. Statistical distributions. 3rd ed. New York: John Wiley and Sons, 2000.
30 Kirkwood BR, Sterne JA. Medical statistics. 2nd ed. Malden, Mass: Blackwell Sciences, 2003.
31 Moore DS, McCabe GP. Introduction to the practice of statistics. 5th ed. New York, NY: WH Freeman and Co, 2005.
32 Chatfield C. The initial examination of data. *J R Stat Soc [Ser A]* 1985; 148: 214-53.
33 Parkhurst DF. Arithmetic versus geometric means for environmental concentration data. *Environ Sci Technol* 1998; 32**:** 92A-8A.
34 Lee AH, Xiao J, Vemuri SR, Zhao Y. A discordancy test approach to identify outliers of length of hospital stay. *Stat Med* 1998; 17: 2199-206.
35 Micceri T. The unicorn, the normal curve, and other improbable creatures. *Psychol Bull* 1989; 105: 156-66.
36 Brown GW. Standard deviation, standard error. Which 'standard' should we use? *Am J Dis Child* 1982; 136: 937-41.
37 Streiner DL. Maintaining standards: differences between the standard deviation and standard error, and when to use each. *Can J Psychiatry* 1996; 41: 498-502.
38 Bartko JJ. Rationale for reporting standard deviations rather than standard errors of the mean. *Am J Psychiatry* 1985; 142: 1060.
39 Yuan KH, Bentler PM, Zhang W. The effect of skewness and kurtosis on mean and covariance structure analysis: the univariate case and its multivariate implication. *Sociol Methods Res* 2005; 34: 240-58.
40 Buchner DM, Findley TW. Research in physical medicine and rehabilitation. VIII. Preliminary data analysis. *Am J Phys Med Rehabil* 1990; 69: 154-69.
41 Bergmann R, Ludbrook J, Spooren WP. Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *Am Stat* 2000; 54: 72-7.
42 Ludbrook J. Statistics in physiology and pharmacology: a slow and erratic learning curve. *Clin Exp Pharmacol Physiol* 2001; 28: 488-92.
43 Barber JA, Thompson SG. Open access follow up for inflammatory bowel disease. Would have been better to use t test than Mann-Whitney U test. *BMJ* 2000; 320: 1730-1.
44 Moran JL, Solomon P. Worrying about normality. *Crit Care Resusc* 2002; 4: 316-9.
45 Austin PC, Hux JE. A brief note on overlapping confidence intervals. *J Vasc Surg* 2002; 36: 194-5.
46 Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. *Am Stat* 2001; 55: 182-6.
47 Ludbrook J, Dudley H. Why permutation tests are superior to t and F tests in biomedical research. *Am Stat* 1998; 52: 127-32.
48 Barnard GA. Must clinical trials be large? The interpretation of P-values and the combination of test results. *Stat Med* 1990; 9: 601-14.

49 Ludbrook J, Dudley H. Issues in biomedical statistics: analysing 2 x 2 tables of frequencies. *Aust N Z J Surg* 1994; 64: 780-7.
50 Schader M, Schmid F. Two rules of thumb for the approximation of the binomial distribution by the normal distribution. *Am Stat* 1989; 43: 23-4.
51 D'Agostino RB, Chase W, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial populations. *Am Stat* 1988; 42: 198-202.
52 Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol* 1994; 47: 881-9.
53 Moran J, Solomon P, Warn D. Methodology in meta-analysis: a study from critical care meta-analytic practice. *Health Serv Outcomes Res Methodol* 2004; 5: 207-26.
54 Senn S, Walter S, Olkin I. Odds ratios revisited. *Evid Based Med* 1998; 3: 71-2.
55 Zhang J, Yu K. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 1998; 280: 1690-1.
56 Sackett DL, Deeks JJ, Altman DG. Down with odds ratios. *Evid Based Med* 1996; 1: 164-6.
57 Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses – sometimes informative, usually misleading. *BMJ* 1999; 318: 1548-51.
58 Chatterjee S, Hadi AS. Regression analysis by example. 4th ed. Hoboken, NJ: John Wiley and Sons, 2006.
59 Godfrey K. Simple linear regression in medical research. *N Engl J Med* 2006; 313: 1629-36.
60 deLaubenfels R. The victory of least squares and orthogonality in statistics. *Am Stat* 2006; 60: 315-21.
61 Dunn G. Statistical evaluation of measurement errors. Design and analysis of reliability studies. 2nd ed. London: Hodder Arnold, 2004.
62 Moran JL, Peter JV, Solomon PJ, et al. Tympanic temperature measurements – are they reliable in the critically ill? A clinical study of measures of agreement. *Crit Care Med* 2007; 35: 155-64.
63 Hoffman JP. Generalized linear models: an applied approach. Boston, Mass: Pearson Education, 2004.
64 Baum CF. Panel-data models. In: Baum CF. An introduction to modern econometrics using Stata. College Station, Tex: Stata Press, 2006: 219-46.
65 Kipnis V, Carroll RJ, Freedman LS, Li L. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. *Am J Epidemiol* 1999; 150: 642-51.
66 Green SB. How many subjects does it take to do a regression analysis. *Multivariate Behav Res* 1991; 26: 499-510.
67 Kilian R, Matschinger H, Loeffler W, et al. A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. *J Ment Health Policy Econ* 2002; 5: 21-31.
68 Diehr P, Yanez D, Ash A, et al. Methods for analyzing health care utilization and costs. *Annu Rev Public Health* 1999; 20: 125-44.
69 Giles DEA. The interpretation of dummy variables in semilogarithmic equations. *Econ Lett* 1982; 10: 77-9.
70 Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ* 2001; 20: 461-94.
71 Moran JL, Solomon PJ, Peisach AR, Martin J. New models for old questions: generalized linear models for cost prediction. *J Eval Clin Pract* 2007; 13. In Press.
72 Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol* 2004; 57: 1138-46.
73 Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15: 361-87.
74 Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994; 86: 829-35.
75 Moran JL, Peisach AR, Solomon PJ, Martin J. Cost calculation and prediction in adult intensive care: a ground-up utilisation study. *Anaesth Intensive Care* 2004; 32: 787-97.
76 Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Stat Med* 2000; 19: 1831-47.
77 Siemiatycki J, Thomas DC. Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol* 1981; 10: 383-7.
78 Botto LD, Khoury MJ. Commentary: facing the challenge of gene-environment interaction: the two-by-four table and beyond. *Am J Epidemiol* 2001; 153: 1016-20.
79 Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med* 1983; 2: 243-51.
80 Belsley DA. Conditioning diagnostics, collinearity and weak data in regression. New York: John Wiley and Sons, 1991.
81 Zheng B, Agresti A. Summarizing the predictive power of a generalized linear model. *Stat Med* 2000; 19: 1771-81.
82 Scott A, Wild C. Transformations and R2. *Am Stat* 1991; 45: 127-9.
83 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29-36.
84 Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: John Wiley and Sons, 2000.
85 Rowan KM, Kerr JH, Major E, et al. Intensive Care Society's APACHE II study in Britain and Ireland. I: Variations in case mix of adult admissions to general intensive care units and impact on outcome. *BMJ* 1993; 307: 972-7.
86 Tibshirani R. A plain man's guide to the proportional hazards model. *Clin Invest Med* 1982; 5: 63-8.
87 Hosmer DW Jr, Lemeshow S. Applied survival analysis: regression modeling of time to event data. New York: John Wiley and Sons, 1999.
88 Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. 2nd ed. New York: Springer-Verlag, 2003.
89 Kuha J. AIC and BIC. Comparisons of assumptions and performance. *Sociol Methods Res* 2005; 33: 188-229.
90 Fears TR, Benichou J, Gail MH. A reminder of the fallibility of the Wald statistic. *Am Stat* 1996; 50: 226-7.
91 Singer JB, Willett JB. Fitting basic discrete-time hazard models. Applied longitudinal analysis. New York: Oxford University Press, 2003: 357-406.
92 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; 130: 515-24.
93 Charlson ME, Ales KL, Simon R, MacKenzie CR. Why predictive indexes perform less well in validation studies. Is it magic or methods? *Arch Intern Med* 1987; 147: 2155-61.
94 Picard RR, Berk KN. Data splitting. *Am Stat* 1990; 44: 140-7.
95 Schumacher M, Hollander N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building? *Stat Med* 1997; 16: 2813-27.
96 Buckland ST, Burnham KP, Augustin NH. Model selection: an integral part of inference. *Biometrics* 1997; 53: 603-18.
97 Chen YQ, Hu C, Wang Y. Attributable risk function in the proportional hazards model for censored time-to-event. *Biostat* 2006; 7: 515-29.