Journal of Research in Nursing
http://jrn.sagepub.com/

# Look before you leap and don't put all your eggs in one basket : The need for caution and prudence in quantitative data analysis

Steven Pryjmachuk and David A. Richards

*Journal of Research in Nursing* 2007 12: 43
DOI: 10.1177/1744987106070260

The online version of this article can be found at:
http://jrn.sagepub.com/content/12/1/43

Published by:
SAGE
http://www.sagepublications.com

**Additional services and information for *Journal of Research in Nursing* can be found at:**

**Email Alerts:** http://jrn.sagepub.com/cgi/alerts

**Subscriptions:** http://jrn.sagepub.com/subscriptions

**Reprints:** http://www.sagepub.com/journalsReprints.nav

**Permissions:** http://www.sagepub.com/journalsPermissions.nav

**Citations:** http://jrn.sagepub.com/content/12/1/43.refs.html

Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

ORIGINAL PAPER

# Look before you leap and don't put all your eggs in one basket

## *The need for caution and prudence in quantitative data analysis*

jrn

**Journal of Research in Nursing**
© 2007
SAGE PUBLICATIONS
*London, Thousand Oaks, New Delhi*
VOL 12(1) 43–54
DOI: 10.1177/1744987106070260

**Steven Pryjmachuk**
*Lecturer*
School of Nursing, Midwifery and Social Work, University of Manchester

**David A. Richards**
*Professor of Mental Health*
Department of Health Sciences, University of York

**Abstract** This paper's aim is to draw attention to the pitfalls that novice and, sometimes, experienced researchers fall into when undertaking quantitative data analysis in the health and social sciences, and to offer some guidance as to how such pitfalls might be avoided.

Many health and social science students are routinely instructed that the procedure for undertaking data analysis in quantitative research is as follows: specify hypotheses; collect data and enter it into a computerised statistical package; run various statistical procedures; examine the computer outputs for *p*-values that are statistically significant. If significant differences are found, jubilation often exists because statistically significant results are deemed to be a clear indicator that something worthwhile (and publishable) has been discovered. This paper argues that this approach has two major oversights: a failure to explore the raw data *prior* to analysis and an overdependence on *p*-values. Both of these oversights are routinely present in much health and social-science research, and both create problems for scientific rigour.

Researchers need to exercise caution ('look before you leap') and prudence ('don't put all your eggs in one basket') when undertaking quantitative data analyses. Caution demands that, prior to full data analysis, researchers employ procedures such as data cleaning, data screening and exploratory data analysis. Prudence demands that researchers see *p*-values for their true worth, which exists only within the context of statistical theory, confidence intervals, effect sizes and the absolute meaning of statistical significance.

**Keywords** statistics, research methods, quantitative approaches, statistical significance, exploratory data analysis



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

## Introduction

In many health and social science departments, students are routinely instructed that the procedure for undertaking quantitative research that involves a degree of statistical analysis is as follows: firstly, a priori hypotheses are specified; then the data are collected and entered into a computerised statistical package such as *SPSS* (see SPSS Inc., 2005) or *SAS* (see SAS Institute, 2005); various statistical procedures are run; and, lastly, the outputs of these procedures are examined for *p*-values that are statistically significant at, at least, the 5% level. On finding significant differences between group means (for comparative studies) or on finding correlations that are significantly different from zero (for relational studies), jubilation often exists among researchers because statistically significant results are — so they are taught — a clear indicator that something worthwhile (and publishable) has, indeed, been discovered.

Two common oversights, however, are evident in this rather mechanistic approach to data analysis: a lack of exploration of the raw data *prior* to analysis and an overdependence on *p*-values. Both of these oversights create problems for scientific rigour: a lack of exploration of the raw data prior to analysis means that important trends in the data may be missed or inappropriate statistical tests used; an overdependence on *p*-values means that credit might be given to findings where credit is not necessarily due. These two oversights form the focus of this paper.

## Data cleaning, data screening and exploratory data analysis

Examining and exploring the raw data prior to any higher-level analyses serves two principal purposes: it ensures the integrity of the data and it helps the researcher to become acquainted with the data. With regard to the integrity of the data, the examination and exploration of raw data introduces a degree of audit into the data-analysis process in that it can to help correct some of the errors that arise during the data collection, tabulation and entry phases. Because initial versions of datasets inevitably contain errors, these datasets are often called 'dirty' (Babbie, 2001), and the process of dealing with such dirty datasets is often, unsurprisingly, referred to as 'data cleaning'.

A number of measures are available to help clean datasets. If sufficient resources are available, a precursory measure might be to create two separate first-stage datasets from two independent entries of the same raw dataset, comparing the two first-stage datasets for differences. This 'double-entry' measure deals with direct data-entry errors, whether human (such as an erroneous keystroke) or mechanical (such as an optical scanner misread). If resources are limited, an alternative is to subject a single first-stage dataset to a random 10% check (NCS Pearson Inc., 2004) — in other words, 10% of the cases under investigation are randomly selected, the data recorded in the first-stage dataset being checked carefully against the raw data relating to each case. An additional data-cleaning measure is to undertake a frequency analysis of each of the variables in the first-stage dataset in order to check that the values for each of the variables are in the acceptable (expected, valid) range (Barhyte and Bacon, 1985; Babbie, 2001; NCS Pearson Inc., 2004). Indeed, when using computerised packages like *SPSS* and *SAS*, this measure is so quick and easy that it should be considered routine.

There are numerous criticisms in the literature (see, for example, Afifi and Clark, 1996; Howell, 1997; Wilkinson and the Task Force on Statistical Inference, 1999; Burns and Grove, 2001) regarding the ritual of plunging straight into complex data



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

analyses without first becoming familiar with the data. This is where *data screening* comes into play: once researchers are reasonably certain that the data are clean, they should familiarise themselves with the data via the process that is central to data screening, 'exploratory data analysis' (EDA). (Calls for the routine use of EDA go back at least three decades — see, for example, Tukey (1977) — yet EDA is still rarely reported in research reports. Its omission may well be down to a lack of use by researchers, but it is worth bearing in mind that the pressures on space in the academic journals may be an added complication in the continuation of its absence.) Much of EDA involves the plotting and examination of graphical devices such as histograms, scatterplots, boxplots (box-and-whisker plots) and stem-and-leaf plots. This process helps the researcher to get a feel for the variables under investigation, such as the general shape of a variable's distribution or any notable peculiarities in the data. EDA can also help to check for violations of the assumptions that underpin specific statistical analyses (Afifi and Clark, 1996; Howell, 1997).

## EDA and peculiarities in the data

Specific peculiarities that researchers need to be on the lookout for include *missing data* and *outliers*. Problems can arise if a particular variable has a high proportion of missing responses or values. Afifi and Clark (1996) suggest that variables with a high proportion of missing responses should be deleted and that cases that have missing responses to particular variables should be excluded from analyses involving those variables. Thus, a missing data analysis (or 'missing value analysis' as some computerised packages call it) should be the first stage of EDA.

Outliers — extreme values for a particular variable — are perhaps more problematic than missing data because the influence that outliers can have on the results of analyses is rather more subtle. Outliers can substantially affect the results of statistical analyses (Afifi and Clark, 1996; Howell, 1997; Wilkinson and the Task Force on Statistical Inference, 1999; High, 2000; Hopkins, 2002). High (2000), for instance, notes that outliers can lead to biased population estimates, inflated sums of squares, distorted *p*-values and faulty, even false, conclusions. Sometimes outliers result simply from errors of the data-entry process (in which case, they can be picked up via the data-cleaning process and corrected); on other occasions, they are real, but extreme, values.

A standard approach to the identification of outliers in a univariate distribution (i.e. a single-variable distribution, such as the distribution of the dependent variable, Y, in a comparative research design where the groups being compared are categories of the independent variable, X) is to use the boxplots and stem-and-leaf plots of EDA (Tukey 1977; Afifi and Clark, 1996; Howell, 1997; Burns and Grove, 2001). There is little excuse for not doing this as most computerised statistical packages produce these plots with relative ease. Outliers are relatively easy to spot from a quick visual inspection of these plots. The situation regarding outliers gets more complex for relational research designs (designs employing correlation and regression) as these designs employ bi- or multivariate distributions. Afifi and Clark (1996) argue, however, that extreme values on *both* the *X* and the *Y* variables (for bivariate distributions) or *on any or all* of the *X* variables ($X_1$, $X_2$, $X_3$ . . . $X_i$) *as well as* the *Y* variable (for multivariate distributions) are far more worrying than outliers on solely the *X* or *Y* variables. One of the easiest ways of identifying these so-called 'influential points' is to examine a statistic known as Cook's *D* (both *SPSS* and *SAS* will calculate this on



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

request) and look for responses where $D > 1.00$ (Howell, 1997; Montgomery *et al.*, 2001).

While the identification of outliers and influential points is reasonably straightforward, dealing with them is another matter — especially since there is no clear-cut answer in the literature on what to do with outliers and influential points once identified. A common recommendation (see, for example, Afifi and Clark, 1996; Howell, 1997) is to run the analysis twice: on the full dataset and on the dataset with any outliers and/or influential points removed. In these circumstances, both sets of results need to be presented. However, it is extraordinarily rare see any discussion of outliers and influential points in the empirical health and social science literature, let alone see two sets of results (full dataset vs. outliers removed).

## EDA and the assumptions underlying statistical tests

The choice of statistical procedure or test within a specific research design is dependent on several factors. For a start, whether *comparisons* or *relationships* between variables are being examined will have a bearing on the choice of test, as will the type of variable (e.g. categorical vs. scale) and the number of variables being studied. Flowcharts, decision trees and tables aiding the choice of test abound in the literature (see, for example, Burns and Grove, 2001; Hawkins, 2005), and it is rare for the experienced researcher to choose a wholly inappropriate test. There is some controversy, however, in choosing between *parametric* and *non-parametric* tests (see, for example, MacDonald 1999), although there is still a general consensus that parametric tests are superior to non-parametric tests, and that they should be employed in preference to non-parametric procedures unless there are strong reasons for not doing so (Howell, 1997; Hopkins, 2004). 'Strong reasons for not doing so' may, indeed, be the very crux of the controversy. One purportedly strong reason for employing non-parametric tests is that they can provide straightforward and relatively quick answers. However, these answers are provided at the expense of precision, and they are quick only with relatively small sample sizes. With large sample sizes, non-parametric tests are difficult to compute, although recent (and future) advances in high-powered computing are likely to change this position (to the extent that 'quick' may need to be redefined). Another strong reason for preferring non-parametric tests is that real-life populations are very often *not* normally distributed (Micceri, 1989). In all, it may be that the parametric/non-parametric debate is somewhat artificial and, as such, the advice of Wilkinson and the Task Force on Statistical Inference (1999) to choose a minimally sufficient analysis (i.e. avoid complex procedures when simples ones will do) is eminently sensible advice.

Still, the point remains that if parametric procedures are to be employed, then researchers must ensure that the underlying assumptions of each test are adhered to. Perhaps the strongest reason for *not* employing a parametric procedure is having data that are incompatible with the underlying assumptions of the particular test. EDA has a particularly useful place in checking the compatibility of the data against the underlying assumptions of a specific statistical test, as will become apparent in the ensuing discussion on the two main statistical approaches in the health and social sciences. The *comparative* approach includes parametric procedures such as the t-test and analysis of variance (ANOVA), as well as non-parametric procedures such as the Mann–Whitney and Kruskal–Wallis tests, and is concerned with differences on one or more dependent variables, $Y_1$, $Y_2$, $Y_3$, etc., across the categories of some independent



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

variable, X (one-way analysis), or group of variables, $X_1$, $X_2$, $X_3$, etc. (two-way, three-way analysis, etc.). The *relational* approach includes procedures such as correlation and regression, and is concerned with the relationship between a pair of variables, X and Y, or between Y and a number of (independent) variables, $X_1$, $X_2$, $X_3$ . . . $X_i$.

## Assumptions underlying parametric test of comparison

Although non-parametric procedures are not entirely without underlying assumptions (both the Mann–Whitney and Kruskal–Wallis tests require distributions with similar shapes, for example), the prerequisites for parametric procedures are generally more exacting. For tests such as the t-test and ANOVA, there are three main assumptions about the distributions being compared that need to be met. First, these tests assume that the distributions being compared are *normal* and one of the best ways to check for normality is to visually examine the plots produced by the EDA functions in the computerised statistical packages (Howell, 1997; Wilkinson and the Task Force on Statistical Inference, 1999). The distributions do not have to be absolutely normal: the t-test is robust, and relatively minor deviations from normality do not appear to influence the results unduly (Afifi and Clark, 1996; Howell 1997). ANOVA can similarly cope with deviations from normality as long as the distributions being compared are similar in shape (Howell, 1997).

The second assumption of these tests is that the variances of the distributions being compared are roughly equal. This 'homogeneity of variance' assumption is not a great problem with the t-test as most computerised statistical packages output two results: one for homogeneous variances and an adjusted result where heterogeneous variances are evident (whether or not the variance is homogeneous can be gleaned from the Levene test, included in the output from most of the statistical packages). With ANOVA, Howell (1997) argues that heterogeneous variances are not particularly problematic as long as the groups being compared have roughly equal sample sizes (as with the t-test, homogeneity of variance can be checked via the Levene test). Howell adds, however, that an ANOVA with unequal sample sizes together with heterogeneous variances produces a serious violation of the underlying assumptions and any results obtained will be suspect. In these circumstances, transformations of the dependent variable(s) to a form that yields homogenous variances should be considered (Howell, 1997). Alternatively, a non-parametric test such as the Kruskal–Wallis test can be employed.

An additional assumption concerns the *independence* of observations, i.e. whether the observations in one of the comparison groups (the categories of the independent variable) are influenced by observations in any or all of the other comparison groups. Afifi and Clark (1996) argue that when data are collected from people (as they often are in health and social science research), it is frequently safe to assume independence of observations collected from different people. The only potential problem arises when *repeated* measures are employed; however, given that specific versions of the t-test and ANOVA exist for repeated measures designs, the independence of observations is rarely a problem in health and social science research.

## Assumptions underlying parametric tests of relationship

The main statistical approaches available when exploring relationships between variables are correlation and regression. Correlation is a measure of the relationship between two variables, X and Y; regression (a closely related technique) is concerned



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

with how one variable, the dependent variable (Y), might be *predicted* from one or more independent (X) variables. Correlation and (linear) regression scenarios in the health and social sciences are typically subject to three basic assumptions (Afifi and Clark, 1996; Howell, 1997). The first assumption concerns 'linearity': when correlating two variables, X and Y, or using regression to predict Y from any number of X variables, there is an assumption that the relationship between the X variable(s) and Y is *linear*, i.e. it can be represented graphically by a straight line. The second assumption concerns normality. This assumption is analogous to the normality assumption in the comparative approach, although different normality assumptions exist depending on whether a 'fixed-*X*' or 'variable-*X*' model is employed (the distinction between the two lies in whether the researcher fixes the values of the X variable(s) prior to data collection). In the variable-X scenario, the normality assumption is that the data to be analysed are taken from the bivariate (X, Y) normal distribution or, in the case of multiple regression, from the multivariate ($X_1$, $X_2$, $X_3$ ... $X_i$, Y) normal distribution. In the fixed-X scenario, it is the 'conditional distribution(s)' of Y (the distribution(s) of Y for specific values of $X_i$) that need to be normal. The third assumption is that of 'homoscedasticity', or a similarity in 'scatteredness'. Homoscedasticity exists where the variances for each value of $X_i$ are similar (cf. homogeneity of variance with the comparative approach).

Checking these assumptions can be tricky, especially when complex correlation and regression models are employed, although some relatively simple checks are available for straightforward bivariate correlation. For example, linearity and bivariate normality can be checked by visually inspecting the EDA scatterplot of X against Y and looking for a roughly elliptical shape (Afifi and Clark, 1996; Howell, 1997). Furthermore, in most health and social science scenarios, homoscedasticity is rarely a problem if bivariate or multivariate normality is demonstrated (Afifi and Clark, 1996). Using scatterplots to check for linearity and normality is extraordinarily difficult in multivariate scenarios, however, as employing more than two independent (X) variables requires an ability to visualise in four or more dimensions.

Where violations of the assumptions for bivariate correlation are evident, non-parametric procedures such as Spearman's *R* or Kendall's _ should be used in preference to the parametric Pearson r. It is worth noting, however, that meeting the underlying assumptions for correlation are only important if the procedure is being used to make *inferences* about some target population on the basis of a discrete sample. If the purpose of the correlation is merely to *describe* some quality of the sample, the assumptions do not need to be met (Cohen, 1988; Howell, 1997).

Where there are doubts about whether the data collected will meet the assumptions that underlie a specific multivariate correlation or regression procedure, a pragmatic resolution might be to employ a procedure requiring fewer assumptions, or one where specific assumptions do not matter. For example, logistic regression (where the dependent variable, Y, is predicted from a sigmoid curve rather than a straight line) has advantages over linear regression in that it can handle categorical independent (X) variables and in that neither normality nor homoscedasticity are issues (Hosmer and Lemeshow, 1989; Afifi and Clark, 1996; Garson, 1999). There is always a trade-off with pragmatism, however, and with logistic regression, it manifests with the necessity of having a dichotomous dependent (Y) variable.

A more important issue to be aware of in any multivariate regression procedure, however, is that of 'multicollinearity'. Multicollinearity occurs when several of the independent variables are highly intercorrelated (Rawlings *et al.*, 1988; Hosmer and



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

Lemeshow, 1989; Afifi and Clark, 1996; Garson, 1999; Montgomery *et al.*, 2001). In *linear* regression, multicollinearity can be identified by a statistic called the 'variance inflation factor' (VIF), or by its inverse, known as 'tolerance' (again, most computerised statistical packages produce these statistics with relative ease). For any particular independent variable, high VIF/low tolerance values imply that that variable can itself be predicted from one or more of the other independent variables in the regression equation. VIF is typically seen as being high when it exceeds a value of 10.0 (Rawlings *et al.*, 1988; Montgomery *et al.*, 2001) and, in these circumstances, the variable should be considered for removal from the equation (Afifi and Clark, 1996). Although multicollinearity is as much an issue in non-linear regression as it is in linear regression, there are, unfortunately, no established procedures for identifying VIF and tolerance in non-linear models (Afifi and Clark, 1996; Garson, 1999).

## Confidence intervals, effect sizes and the meaning of significance

At this point, the reader should be aware of how the robustness of any proposed data analyses can be improved via some simple preparatory processes (involving the conventions of EDA together with a few additional procedures, all of which are well within the capabilities of most computerised statistical packages). It is pointless having robust analyses, however, if the results of those analyses are reported in a ritualistic manner that fails to acknowledge the debate surrounding the meaning and interpretation of statistical test results or if the results are detached from the philosophical context in which they are set. This debate is far from recent (see Cohen, 1994) yet, despite many decades of criticism, null-hypothesis significance testing and the associated fixation with *p*-values — particularly the notion of significance at the 5% level — still feature heavily in the health and social science literature. The problem with the *p*-value approach is that it merely tells the individual that the two (or more) means are not the same. As Cohen (1994) points out, differences among groups always exist at some level of precision: attaining statistical significance is merely a matter of sample size. With large sample sizes, very small differences may be statistically significant but have little real clinical or practical importance. Indeed, Cohen has noted that very small, yet statistically significant, differences have erroneously led to the establishment of theory. Moreover, the logic of the scientific method implies that the results of statistical testing can only have meaning if a priori hypotheses have been established. Such hypotheses guide the researcher and help maintain objectivity; without them, the search for Truth becomes a haphazard data trawl and scientific rigour falters.

Many disciplines have moved to address the widespread overdependence on *p*-values, with alternatives such as confidence intervals, decision theory or Bayesian approaches being proposed (Bailar and Mosteller, 1988; Johnson, 1999; Wilkinson and the Task Force on Statistical Inference, 1999). It is beyond the scope of this paper to discuss each of these alternatives in detail, but the calls from the medical and psychological professions (Bailar and Mosteller and Wilkinson *et al.*, respectively) are perhaps the most relevant to health and social science researchers. To address the widespread overdependence on *p*-values, medicine and psychology recommend that alpha-values (the threshold at which statistical significance is accepted, e.g. for 5% significance, $\alpha = 0.05$) be adjusted to take account of the sample size, that confidence intervals be employed as an adjunct to (or, indeed, instead of) *p*-values and that



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

effect sizes be routinely reported alongside the test statistics and corresponding p-values.

With regard to adjusting _ to account for sample size, a protocol that deals with this in an indirect yet pragmatic manner is for researchers to report the actual p-values rather than use asterisks or the generic '$p < 0.05$' to mark significance and '$p > 0.05$' or the ubiquitous 'NS' to mark non-significance. Presenting the actual p-value together with the sample size gives readers of a research report the opportunity to make up their own minds about appropriate alpha-values. Reaching statistical significance should be seen as a starting point for discussing *potentially* interesting results rather than as an absolute indication of discovery.

## Confidence intervals

The confidence interval (CI) is the range of observations in which a researcher can be certain that the true mean of population lies, delimited by an upper and a lower 'confidence limit'. Confidence limits of 95% is the normal convention: this means that, for a given variable, the researcher can be 95% confident that the true population mean will lie somewhere between the upper and lower confidence limits. The utility of CIs is two-fold. First, when a specific variable is being compared across two or more groups, plotting the variable's CI for each of the groups side-by-side can give an almost immediate visual indication of whether there is a difference between the groups (in most cases, the respective CIs do not overlap). For example, in the hypothetical scenario contained in Figure 1, group A's CI does not overlap with the CI of either group B or group C. Given that there is an overlap between the CIs of groups B and C, the implication is that group A is (significantly) different from groups B and C on activity scores, and that groups B and C do not differ significantly from each other. On the other hand, if only the means were plotted, then it would

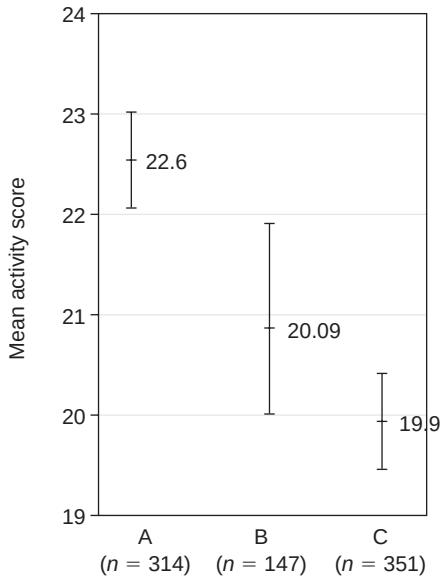


**Figure 1** Mean hypothetical 'activity' scores (with 95% confidence intervals) for three groups.



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

be understandable if the observer thought that there might also be a difference between group B and group C.

An additional utility of CIs occurs when CIs for the *difference* between two means is calculated. Such a CI can act as an adjunct — if not an alternative — to significance testing in that, if the CI for the difference between two means spans zero, then the difference will not be statistically significant. To put it another way, the researcher cannot be confident that there is a difference between the means.

## Effect size

Effect size (ES) statistics are important because they give an indication of something that has rarely been reported in empirical studies — the *magnitude* of the effect (or difference) found. The use of ES statistics can guard against Cohen's complaint that very small, yet statistically significant, differences are often implicated in the establishment of theory. They can also help practitioners and researchers in the health and social sciences to understand the importance of the debate regarding clinical vs. statistical significance (see, for example, Ottenbacher, 1995) in that very small effect sizes (clinical insignificance) often accompany *p*-values much smaller than 0.05 (statistical significance), especially when large sample sizes are present.

Wilkinson and the Task Force on Statistical Inference (1999) argue that effect sizes should be routinely reported alongside statistically significant findings. A number of effect sizes (ES) statistics are available, depending in the statistical test employed (Table 1).

The three effect size thresholds listed in Table 1 give an indication of the values required for an effect (difference or relationship) to be qualitatively described as 'small', 'medium' or 'large' (these figures are often used in power/sample size calculations when researchers have no a priori effect sizes available) . Two threshold sets are quoted for Cohen's *d* (a standardised measure of the magnitude of the difference between two means) mainly because Hopkins (2004) argues that Cohen's values are too low to define the thresholds for medium and large effects. Hopkins also challenges the utility of a three-descriptor effect size 'scale', preferring instead to think of a seven-descriptor scale (trivial–small–moderate–large–very large–nearly perfect–perfect), ranging from zero (trivial) to one (perfect) for r, from zero (trivial) to infinity (perfect) for *d* and from one (trivial) to infinity (perfect) for relative risk and the odds ratio. The odds ratio (the exponential of the logistic regression coefficient, B) for a specific predictor (i.e. independent) variable is the factor by which the odds of being in the disease-present category of the dependent variable (as opposed to the disease-absent category) increase as the scores on the predictor variable rise by

**Table 1** **Small, medium and large effects for typical statistical procedures (after Hopkins, 2000).**

| Effect size measure | Effect size thresholds | | |
|---|---|---|---|
| | Small | Medium | Large |
| Correlation coefficient, *r* | 0.10 | 0.30 | 0.50 |
| Cohen's *d* (Cohen, 1988) | 0.20 | 0.50 | 0.80 |
| Cohen's *d* (Hopkins, 2000) | 0.20 | 0.60 | 1.20 |
| Relative risk | 1.2 | 1.9 | 3.0 |
| Odds ratio (risk) | 1.5 | 3.5 | 9.0 |



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

one unit. Thus if the odds ratio = 5 for a particular predictor, a one unit increase on that predictor means that the odds of having the disease increase five-fold. An odds ratio = 0.50 means the odds of having the disease halve or, conversely, the odds of not having the disease double. An odds ratio = 1.00 means no change as the scores on the predictor increase or decrease (hence Hopkins' descriptor of 'trivial' for an odds ratio = 1). When CIs are presented for an odds ratio, a CI that includes the value of 1.00 suggests a nil effect in the same way that there is a nil effect when the CI for the difference between two means includes zero.

Both Cohen and Hopkins also talk about the use of 'variance explained' (a measure with a range of 0% to 100%) as an additional or alternative ES measure in statistical procedures, like regression and ANOVA, where variance or the sum-of-the-squares are central concepts. Thus, $R^2$ is often used as an ES measure in multivariate regression, $r^2$ as an ES measure in bivariate regression and eta-squared ($\eta^2$) and omega-squared ($\omega^2$) as ES measures in ANOVA. In logistic regression, where no true $R^2$ is available, 'pseudo' $R^2$ measures such as the Nagelkerke $R^2$ have been developed, although the odds ratio is a suitable alternative.

## Conclusion: a robust procedure

To summarise, when undertaking quantitative research, a robust data-analysis procedure has the following essential elements. First, hypotheses need to be set and suitable data collected. The raw dataset collected then needs to be cleaned, after which exploratory data analysis should be employed to screen the data both for peculiarities (outliers and missing data) and the extent to which the variables under investigation are concordant with the underlying assumptions of the statistical procedures proposed. Where peculiarities exist or violations of the underlying assumptions are identified, appropriate remedial action should be taken: running a double set of analyses (full dataset vs. outliers removed) or utilising transformed variables or non-parametric procedures, for example. Following analysis of the data, readily obtainable measures such as sample size, confidence intervals and effect size should be routinely reported, certainly as an adjunct, perhaps even as an alternative, to the conventional coupling of a single test-specific statistic with its *p*-value (and where *p*-values are reported, these should be the *actual* *p*-values).

The procedure does not stop here, however. Any findings need to be put into context — in other words, *interpreted* — and further pitfalls arise at this stage. Even if the analytical and theoretical procedures are rigorous and the researcher has stumbled upon something interesting in relation to a specific hypothesis (a moderate-to-large effect size, a small confidence interval that does not cross zero, tiny *p*-values and an appropriate sample size), there is still no guarantee of discovery. Before the researcher can be reasonably confident of discovery, there is a need for consistency, both within the investigation itself and with the wider body of research in the same area. Fenwick (1997) points out this need for consistency and the danger of relying on isolated studies in advising his readers to 'treat new study reports as what they are . . . a first impression. If the topic interests you, watch for future developments and, unless they're consistent, you can probably safely ignore them' (Fenwick, 1997: 226).

In the search for Truth, quantitative methods can be powerful tools. However, as with any tool, those using such methods need to be aware of not only their strengths and advantages but also their potential dangers and limitations.



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

## Key points

- The routine, mechanistic approach to quantitative data analysis taught in many health and social science departments creates problems for scientific rigour.
- Failure to explore the raw data prior to analysis means important trends in the data may be missed or inappropriate statistical tests used.
- An overdependence on *p*-values means that credit can be given to findings where credit is not necessarily due.
- A systematic approach to quantitative data analysis involving data cleaning, data screening and exploratory data analysis helps increase scientific rigour.
- The use of confidence intervals and effect size statistics as adjuncts to the *p*-value improves the credibility and weight of any findings.

## References


Afifi, A.A., Clark, V. (1996) *Computer-Aided Multivariate Analysis*, 3rd edn. London: Chapman-Hall.

Babbie, E. (2001) *The Practice of Social Research*, 9th edn. Belmont, CA: Wadsworth.

Bailar, J.C., Mosteller, F. (1988) Guidelines for statistical reporting in articles for medical journals. *Annals of Internal Medicine* **108**: 266–273.

Barhyte, D.Y., Bacon, L.D. (1985) Approaches to cleaning data sets: a technical comment. *Nursing Research* **34**:1, 62–64.

Burns, N., Grove, S.K. (2001) *The Practice of Nursing Research: Conduct, Critique and Utilisation*, 4th edn. Philadelphia, PA: W.B. Saunders.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994) The earth is round ($p < 0.05$). *American Psychologist* **49**: 997–1003.

Fenwick, J.H. (1997) *Studies Show: a Popular Guide to Understanding Scientific Studies*. Amherst, NY: Prometheus.

Garson, G.D. (1999) PA 765 statnotes: an online textbook. Available from: www2.chass.ncsu.edu/garson/pa765/statnote.htm (accessed September 2005).

Hawkins, D. (2005) *Biomeasurement: Understanding, Analysing and Communicating Data in the Biosciences*. Oxford: Oxford University Press.

High, R. (2000) Dealing with 'outliers': how to maintain your data's integrity. Available from: cc.uoregon.edu/cnews/spring2000/outliers.html (accessed September 2005).

Hopkins, W.G. (2002) Dimensions of research. *Sportscience* [e-journal] Available from: sportsci.org/jour/0201/wghdim.htm (accessed September 2005).

Hopkins, W.G. (2004) A new view of statistics. Available from: sportsci.org/resource/stats/index.html (accessed September 2005).

Hosmer, D.W., Lemeshow, S. (1989) *Applied Logistic Regression*. New York: John Wiley and Sons.

Howell, D.C. (1997) *Statistical Methods for Psychology*, 4th edn. Belmont, CA: Duxbury Press.

Johnson, D.H. (1999) The insignificance of statistical significance testing. *Journal of Wildlife Management* **63**:3, 763–772.

MacDonald, P. (1999) Power, type I, and type III error rates of parametric and nonparametric statistical tests. *The Journal of Experimental Education* **67**:4, 367–379.

Micceri, T. (1989) The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* **105**: 156–166.

Montgomery, D.C., Peck, E.A., Vining, G. . (2001) *Introduction to Linear Regression Analysis*, 3rd edn. Chichester: John Wiley and Sons.

NCS Pearson Inc. (2004) *Survey Research Notes: Cleaning Your Data* [pdf document]. Available from: www.pearsonncs.com/research/download/analyzing_cleaning.pdf (accessed September 2005).

Ottenbacher, K.J. (1995) Why rehabilitation research does not work (as well as we think it should). *Archives of Physical Medicine and Rehabilitation* **76**:2, 123–129.

Rawlings, J.O., Pantula, S.G., Dickey, D.A. (1988) *Applied Regression Analysis*, 2nd edn). London: Springer.

SAS Institute (2005) Products and solutions. Available from: www.sas.com (accessed September 2005).

SPSS Inc. (2005) Software and solutions. Available from: www.spss.com (accessed September 2005).

Tukey, J.W. (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wilkinson, L. and the Task Force on Statistical Inference (1999) Statistical methods in psychology journals: guidelines and explanations. *American Psychologist* **54**:8, 594–604.




Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011

**Steven Pryjmachuk** (PhD) is a lecturer and research associate in the School of Nursing, Midwifery and Social Work at the University of Manchester. As well as being a mental health nurse, he is also a chartered psychologist. His specialist area is mental health, with research interests in primary mental health and in child and adolescent mental health. E-mail: steven.pryjmachuk@manchester.ac.uk

**David Richards** (PhD) is Professor of Mental Health at the University of York. He is at the forefront of efforts to improve access to treatment for those suffering from common emotional distress, through the development and testing of new ways of organising low-intensity psychological treatments for people with mental health problems in primary care.



Downloaded from jrn.sagepub.com at SAGE Publications on May 18, 2011