Tutorials in Quantitative Methods for Psychology
2009, Vol. 5 (2), p. 40-50.

DOI: 10.20982/tqmp.05.2.p040

Getting the most from your curves: Exploring and reporting data using informative graphical techniques

Fernando Marmolejo-Ramos
The University of Adelaide
Universidad del Valle

Masaki Matsunaga
Waseda University

Most psychological research employs tables to report descriptive and inferential statistics. Unfortunately, those tables often misrepresent critical information on the shape and variability of the data's distribution. In addition, certain information such as the modality and score probability density is hard to report succinctly in tables and, indeed, not reported typically in published research. This paper discusses the importance of using graphical techniques not only to explore data but also to report it effectively. In so doing, the role of exploratory data analysis in detecting Type I and Type II errors is considered. A small data set resembling a Type II error is simulated to demonstrate this procedure, using a conventional parametric test. A potential analysis routine to explore data is also presented. The paper proposes that essential summary statistics and information about the shape and variability of data should be reported via graphical techniques.

Exploratory data analysis (EDA), as an analytical routine, is alarmingly rare in the current normative practice of conducting research in psychology and other related fields (Behrens & Yu, 2003). Kline (2008) reminds us of the importance of EDA with the blunt yet pithy phrase, "garbage in, garbage out". The author argues that "the quality of computer output depends on the accuracy of the input. *Thus, it is critical to check the data for problems before conducting any substantive analyses*" (p. 233; emphasis in original). To address these concerns and offer an analytical tool for researchers, the current paper illustrates the benefits of using graphical techniques in EDA processes.

First, it is shown how simple graphical techniques aid in making statistical decisions regarding Type I and Type II errors. Second, a simulated data set that exemplifies a Type II error is explored using conventional statistical tests, i.e., homogeneity and normality tests. Then, the same data set is inspected using EDA processes, i.e., looking for outliers and using data transformations. Both approaches have valuable properties that when put together can generate a reliable analytical tool. Thus, a tentative routine is proposed to analyse data sets which takes the best of both worlds.

Finally, the role of graphical techniques that present essential and informative summary statistics for given data is discussed. Particularly, it is suggested that graphics reporting results should not only represent summary

Fernando Marmolejo-Ramos, School of Psychology, Faculty of Health Sciences, University of Adelaide, Australia and Institute of Psychology, Universidad del Valle, Colombia. Masaki Matsunaga, Centre for English Language Education in Science and Engineering, School of Science and Engineering, Waseda University, Tokyo, Japan. Email: matsunaga@aoni.waseda.jp.

The authors thank Jake Olivier, Kardi Teknomo, Marc Brysbaert, and the anonymous reviewers for their comments on the content of the paper, and Yihui Xie and Yuka Toyama for their help with the code for some of the graphics. The authors also thank Louise "B.D." Mooney for checking the structure of the paper.

Correspondence concerning this article should be addressed to: School of Psychology, Faculty of Health Sciences, University of Adelaide, Adelaide, South Australia, Australia, 5005. Email: fernando.marmolejoramos@adelaide.edu.au.

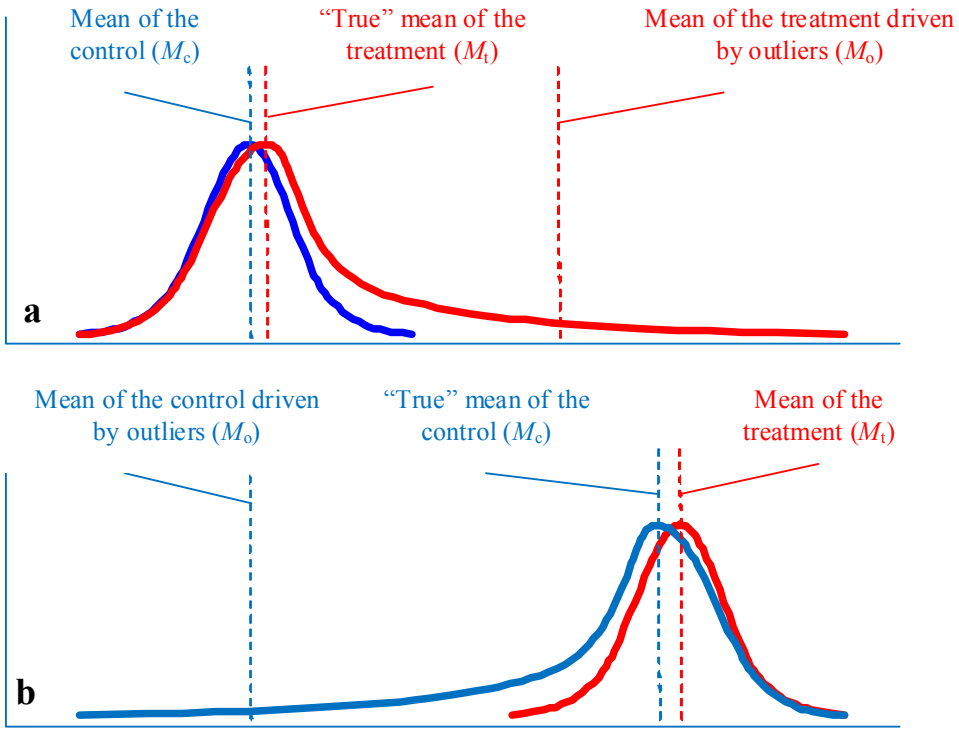

Figure 1. Hypothetical situation that would likely lead to a Type I error. In both cases the treatment is genuinely ineffective, but spurious observations in the treatment group drive its mean away from the control group's mean (a). In other situations, it is the control group that has spurious observations driving its mean away from the treatment group's mean (b). In both situations the null hypothesis is rejected when in fact it is false.

statistics such as the mean and the standard deviation, but also valuable information on the shape and variability of the data's distribution. With this concern in mind, this paper suggests a graphical variation that can provide an effective tool in reporting research results.

Where EDA Helps With Statistical Decisions: Type I and Type II Errors

In scientific research it is commonplace for researchers to face two possible errors when making statistical decisions: To reject the null hypothesis when it is in fact true or failing to reject the null hypothesis when it is in fact false. These possible errors are known as Type I and Type II errors, respectively [1]. In most sciences, researchers are cautious about committing a Type I error when performing a statistical analysis and guard against it by setting a pre-established alpha level (usually .05). However, it is less frequent to encounter situations where researchers guard against Type II errors (Sato, 1996) [2]. The next section graphically exemplifies what happens in each type of error, emphasises the situation where a Type II error occurs, and where EDA can help prevent these errors.

Type I Error Case

Let us assume there are two groups, a control group and a treatment group. When their means are compared, the test reveals that there is a statistically significant difference. A potential pitfall here is that a significant difference can occur because some observations in one of the groups drive its group mean away from the other group's mean, whereas the groups, as a whole and without those outliers, are equivalent.

[1] In drawing a conclusion, there also can be a Type III error and a Type IV error: The former refers to the error of "*correctly rejecting the null hypothesis for the wrong reason*" (Mosteller, 1948, p. 61), whereas the latter refers to "*the incorrect interpretation of a correctly rejected hypothesis*" (Marascuilo & Levin, 1970, p. 398). These errors are, however, not directly related to EDA. Given the scope of the current paper, we therefore chose to focus our attention to Type I and Type II errors.

[2] Sato (1996) argues that researchers are typically less concerned with Type II error as they are with Type I error, perhaps because of the assumption that false-positive conclusions would do more harm than false-negative ones. Nonetheless, neither error is acceptable because they both mark threats to the validity of a conclusion (Kline, 2008).

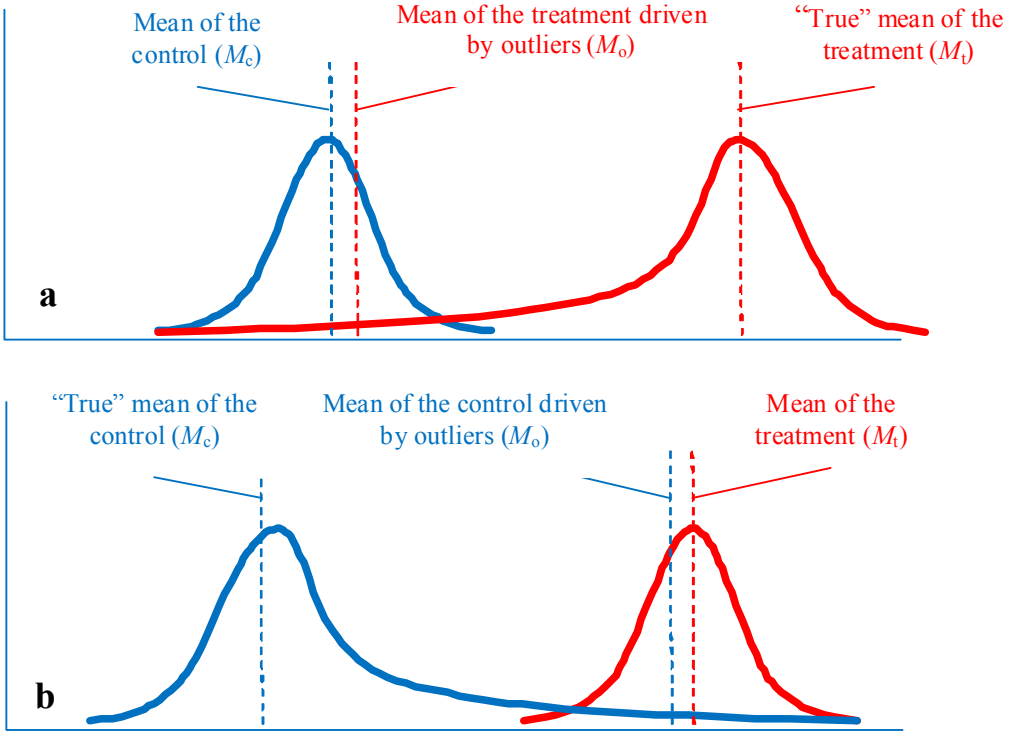

Figure 2. Hypothetical situation that would likely lead to a Type II error. In both cases the treatment is genuinely effective, but spurious observations in the treatment group drive its mean near the control group's mean (a). In other situations, it is the control group that has spurious observations driving its mean near the treatment group's mean (b). In both situations the null hypothesis is not rejected when in fact it is false.

This can happen more often than a researcher may expect. For example, even if a treatment—a newly developed speech-training program, say—has little effect, some individuals assigned to the treatment group might already have the ability of interest, such as a verbal skill, and score high on the administered test. Although, theoretically, random assignment should distribute those innately skilled individuals evenly across groups, it might fail especially when the given sample size is small (Kline, 2008). Additionally, sampling error might occur when some subjects in the control group do not belong to the target population but are included in the sample. Note that all these realistic scenarios could produce outliers in the collected data set. If those observations had been identified earlier, however, researchers could have taken countermeasures, including the removal of those outliers from the data set, and prevented the erroneous rejection of a genuine no difference between the groups' means.

Figure 1 illustrates one such situation. Simply comparing the groups' means, i.e., M_c (the observed mean of the control group) and M_t, (the observed mean of the treatment group including outliers), yields a statistically significant result, although the "true" score mean of the treatment group (i.e., M_t) is actually no different from M_c. Hence, concluding that the treatment had a significant effect is false, leading to a Type I error (Figure 1a). Alternatively, the control group might include some observations whose scores are lower than what is expected in the respective population. Again, failing to detect such cases can result in a Type I error (i.e., a false-positive conclusion that the treatment was effective) (Figure 1b).

Type II Error Case

Similar to the case of Type I error illustrated above, suppose there are two groups, a control group and a treatment group. This time, however, the test reveals that there is no statistically significant difference. The problematic potential in such cases is that a non-significant difference can occur when some observations in one of the groups drive its mean near the other group's. Perhaps some individuals in the treatment group might have failed to follow the instructions during the experiment and their scores unexpectedly deviate from those of the other subjects who properly followed the procedure. Or, subjects in the control group somehow manage to "receive" the treatment by communicating with those assigned to the treatment group, contaminating the experiment.

Figure 2 illustrates those situations. Figure 2a shows a

case where, although the treatment per se has an appreciable effect, the outliers in the treatment group pull down the overall treatment group mean score; consequently, testing the difference between M_c and M_t may result in a failure to reject the null hypothesis, or a Type II error. Alternatively, when extreme observations in the control group boost its mean, the true significant effect of the treatment is therefore concealed (Figure 2b).

In all these cases, drawing a false conclusion (either Type I or Type II error) can be prevented through EDA and visual inspections, particularly graphical techniques that enable the researcher to know more about how the data are distributed. To discuss this utility of EDA, we demonstrate below how EDA helps avoid a Type II error.

Simulated Data

In order to illustrate how the use of data exploration enables us to grasp the essential features of the data distribution for further analysis, two groups of data were generated using R (R Project for Statistical Computing, 2007). Both groups were generated from a normal distribution, but in one of the groups two normal observations were replaced with two outliers. The whole data set consists of two groups, each with a sample size of 20; Group A has a mean of 50.00 (SD = 12.49), whereas Group B's mean is 54.50 (SD = 20.93).

These parameters were specified so that the statistical computations and the graphs accompanying every step are simple enough to keep track of every single observation and reach clarity. Next, a research scenario in psychology is used to put the data in context so that statistical computations and graphical techniques used are readily interpretable. The hypothetical research design presented here is commonly utilized in many areas in psychology like social (e.g., Moorehouse & Sanders, 1992), developmental (e.g., Valenzuela, 1997) and health psychology (e.g., Frisch, Shamsuddin, & Kurtz, 1995), and thus, provides a realistic case. Indeed, the following research situation resembles a study carried out some years ago (see Guy & Cahill, 1999).

Imagine that 40 individuals are recruited to participate in an experiment to test human memory for emotional events. Researchers divide these subjects into two experimental groups, Group A and Group B. Subjects in Group B are presented with video clips which are considered to arouse happiness (i.e., treatment condition), whereas subjects in Group A are presented with video clips that do not evoke any particular emotional state (i.e., control condition). After one week participants are given a free recall test of all the clips viewed. Based on previous emotion research, it is expected that subjects in Group B have higher scores in the free recall test than those in Group A.

Conventional Data Exploration

Initially, we apply two tests that are conventionally utilized for preliminary analyses: normality and homogeneity of variance tests. Normality tests are almost routinely applied because most parametric analyses such as independent-samples t test, analysis of variance, and regression invoke the assumption that the given sample is drawn from a normally distributed population. Given the relatively small sample size (each group's n < 30), the Lilliefors (also known as Kolmogorov-Smirnov) normality test is considered (Lilliefors, 1967). The results of the Lilliefors normality test applied to our simulation data suggest that both groups are drawn from normally distributed populations: for Group A, D (20) = 0.09, p = 0.95; for Group B, D (20) = 0.16, p = 0.20. Thus, researchers drawing on the outcomes of this test would conclude that these data have no problem regarding their normality.

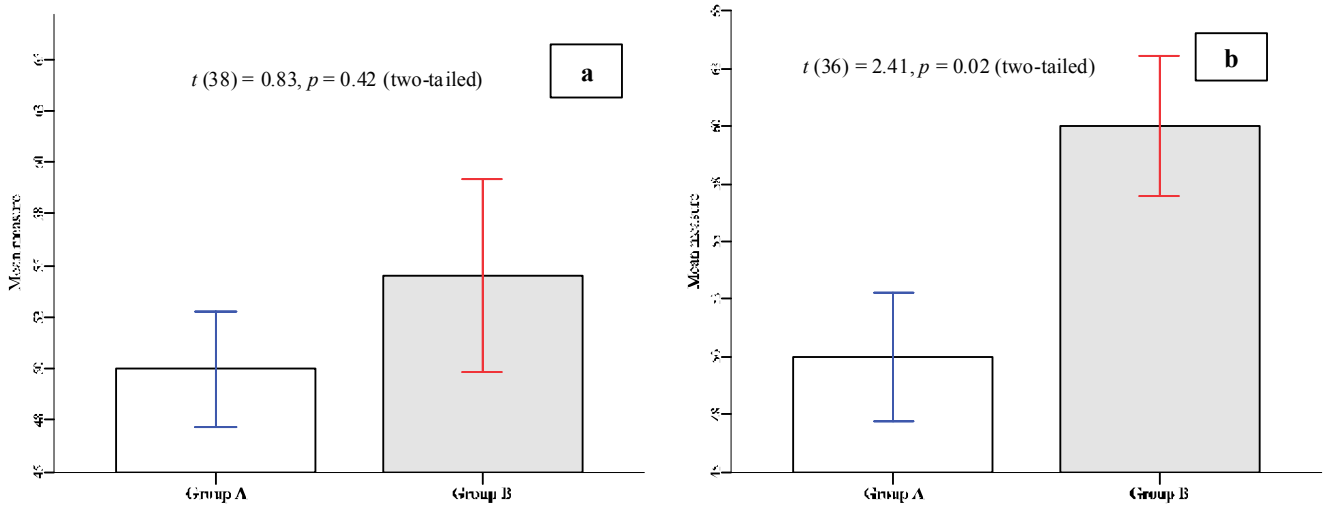

Figure 3. Bar plots with error bars representing the simulated groups of data before (a) and after (b) outlier removal. Error bars represent one standard error.

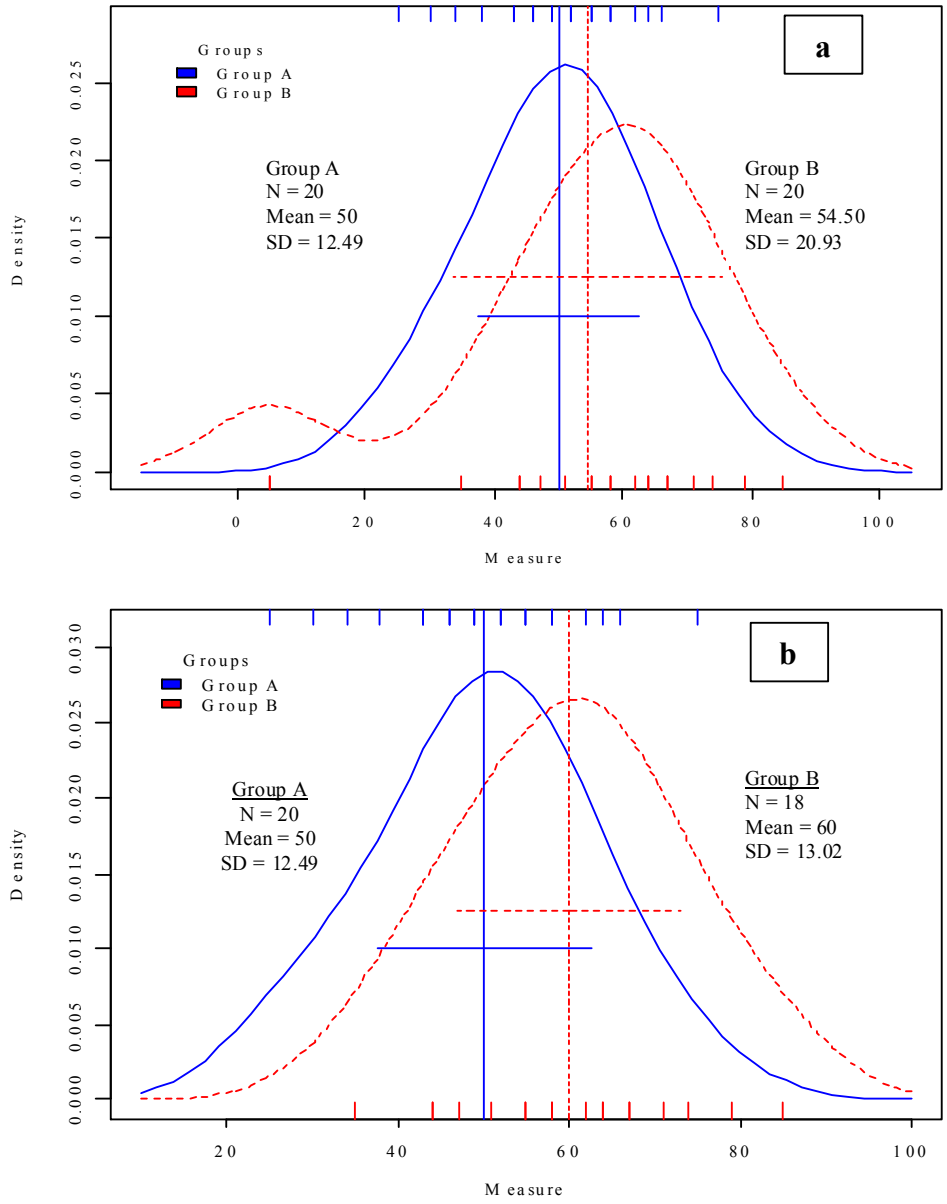

Figure 4. Kernel densities of the simulated groups of data before (a) and after (b) outlier removal. In (a) both groups of data have similar variances but different distributions, whereas in (b) both groups have similar distributions and variances. The rugs (short vertical lines) are added to highlight the actual observations for each data set. The horizontal lines represent the groups' standard deviations. The vertical lines represent the groups' means.

The second step that is typically taken before statistically examining the difference between two independent groups is to verify the homogeneity of variance. The accuracy of a statistical test is adversely affected if the groups under study have different variances (see Zimmerman, 1998). In such cases, some data manipulations are called for in order to render the data amenable to a parametric test. To explore this possibility, we subjected our simulation data to the Levene's test (Levene, 1960), which provides a robust test of homogeneity of variance between relatively small groups (i.e., $n < 50$) (see Correa, Iral, & Rojas, 2006). With the current data, the results of the Levene's test suggest that the two groups have homogeneous variances, F (1, 38) = 2.23, p = 0.14. Thus, again, no problem is detected by the conventionally utilized preliminary test.

Accordingly, researchers examining these data would feel content and ready to take the final step of performing a parametric test. Given the nature of the hypothetical research scenario described above, they would most appropriately run an independent-samples t test to see if there s a statistically significant difference between the two groups' mean scores. To their disappointment, however, the

results of the *t* test would show that the groups are not significantly different, t (38) = 0.83, p = 0.42 (two-tailed) (see Figure 3a).

In terms of the research hypothesis, these results suggest that the video clip shown in the treatment condition did not sufficiently arouse the participants' happiness to claim theoretically appreciable effects. This conclusion, however, leads the researchers to commit a Type II error. As illustrated below, the current data sets were devised to mimic a situation where the observed mean of the treatment group is indistinguishable from that of the control group, even though the treatment per se is effective. We demonstrate how the EDA procedures we propose help detect this problem (which are undetectable through conventionally utilized techniques) in the following section.

Innovative Approaches to Data Exploration

In this section, we propose and illustrate the EDA procedures to inspect data not only using numerical computations but also graphic-based inspections. The core purpose of the EDA procedures is to find patterns in the data, non-admissible observations (via outlier detection), and adjust data to generate a balanced data set (via data transformation) (see Tukey, 1969). More specifically, EDA serves to spot problematic features of the data that may not be detectable via conventional approaches (see Behrens & Yu, 2003). Additionally, the combination of graphical explorations of the data with confirmatory calculations is discussed (see Gelman, 2004).

Figure 3a shows bar plots of both groups' means with their standard errors (SE). The large overlap between the SE visually suggests that both groups are not significantly different (as the *t* test confirmed). Unfortunately, those graphics do not reveal density estimates in order to visualize critical differences between the groups' distributions and see if there could be observations affecting the distribution of the data.

Figure 4a shows the distributions of both groups of data. The densities were estimated using a kernel (and denoted hereafter as kernel densities; see Silverman, 1986; Wilcox, 2004). It can be noticed that whereas Group A seems to be normally distributed, Group B seems to be not. Not all normality tests show similar results regarding the normality of a group. Normality tests are heavily dependent on the given sample size and therefore not entirely reliable. It has been argued that more robust tests of normality are used to check whether a data set dramatically departs from normality. Some researchers argue that the Lilliefors test is not very sensitive and instead the more sensitive Shapiro-Wilk test should be used (Field, 2005). According to this test, Group A is still within normal parameters, W (20) = 0.99, p = 0.99; however, Group B departs from normality, W (20) = 0.88, p = 0.02.

On data transformation

One of the techniques used by statisticians to normalise skewed distributions and heterogeneous variances is via data transformation. There are several transformations available, but Box-Cox, logarithmic, square-root, and inverse transformations are broadly used (see Bland & Altman, 1996, 1996a, 1996b, 1996c; Osborne, 2002). The core idea behind data transformation is to ensure that the data set meets the assumptions of normality and homogeneity of variance (Osborne, 2002). Also, in practical terms, it permits researchers not to discard valuable data.

When the data set of Group B is submitted to a transformation process, it does not benefit its distribution. For example, it is known that the log-transformation can be used to deal with highly skewed distributions (see Olivier, Johnson, & Marshall, 2008), but it only works effectively when distributions are positively skewed. Given that Group B is negatively skewed, the logarithmic transformation just exacerbated this problem. Also, other transformations showed similar results.

On outlier identification

It is common practice in social sciences to regard observations with less than 5% frequency as "outliers" and, on its flipside, 95% as the "acceptable" confidence level (see Cowles & Davies, 1982), i.e., it is common to use a 2 standard deviations (SD) cut-off. By looking again at the kernel density of Group A (see Figure 4a), it can be noticed that there are no observations too distant from the mean. However, there are in Group B a couple of observations which seem to be below 2 SD from the mean, which in turn create a high variance and a significant departure from normality (according to the Shapiro-Wilk test). Those observations could be causing the distribution to skew to the left (*skewness* = -1.22, SE_{sk} = 0.512). The important issue to note here is that this sort of graphic representation of the data permits the researcher to visually pinpoint the actual observations that might be distorting the distribution of the data, despite some normality tests suggesting otherwise.

If the Lilliefors test had been the sole normality test used, there could have been no grounds to perform any data manipulation despite the visual representation of the data set suggesting otherwise. Given that the graphic indicated that Group B did not seem to be normally distributed, it was a reason to check its normality with another test. Again, this situation suggests that some normality tests are not totally reliable and that visual inspection is highly recommended.

As mentioned earlier, it is essential to confirm any visual inspection with formal statistical tests. Using the standardised values of the dependent measure, it can be

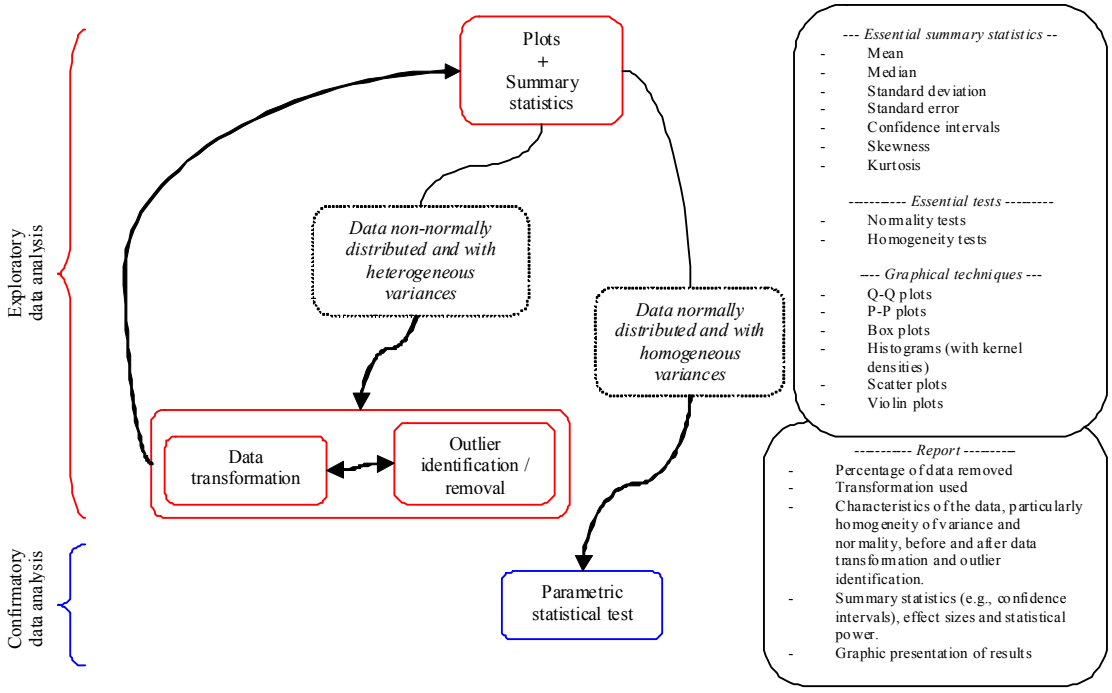

Figure 5. Suggested routine for parametric data analysis. This flow chart takes into account the considerations about data screening outlined by Tabachnick and Fidell (2007), therefore it can be applied to univariate and multivariate data. Data sets that are non-normally distributed and that present heterogeneous variances can be submitted to a non-parametric test. However, such an issue is not tackled here since it goes beyond the scope of the paper.

determined which observations are above or below 2 SD from the mean in each group. In Group A, there are no observations 2 SD below or above the group mean. However, there are two observations in Group B (with the value of 5) which are 2 SD below the group mean. Outlier *z* tests showed that both observations are significant outliers, both z (20) = -2.36, p = 0.009 (see Shiffler, 1988). Once these observations are removed, the distribution and shape of Group B look more normal-like (see Figure 4b). This fact is again corroborated using the sensitive Shapiro-Wilk test, W (18) = 0.99, p = 0.99. Also, the homogeneity of variances between the two groups was highly improved by removing those spurious observations, F (1, 36) = 0.067, p = 0.797.

Not all researchers are fond of outlier identification (see Orr, Sackett, & DuBois, 1991). However, we support the idea that identifying outliers is an important procedure that avoids reporting biased results (see Judd, McClelland, & Culhane, 1995). The method to identify outliers used here (standardized residuals) is just one of the possible options to use. Other useful techniques include the shifting z score criterion (see Thompson, 2006; van Selsts & Jolicoeur, 1994), Cooks' and Mahalanobis distances, leverage values as well as multivariate outliers detection via kurtosis (see Peña & Prieto, 2001). Finally, outlier identification is a very debatable issue which has no consensus among researchers. This situation renders this topic a quite interesting one and worthwhile of further investigation.

A final comment on data manipulation

Using data transformation can improve homogeneity and normality of a data set, but it is not always the case as was explained earlier. Finding a fine balance between homogeneity of variance and normality implies a trade-off between data transformations and identification of outliers. Some researchers place the identification of outliers as a previous step to data transformation (e.g., Behrens, 1997), while other researchers favour the other-way-around procedure (e.g., Tabachnick & Fidell, 2007). Here, we suggest a negotiable use of data transformations and outliers' identification. If the first procedure chosen does not contribute much to meet the assumptions which parametric tests are based on, then begin with the other procedure and continue the process. Figure 5 illustrates this situation.

The role of graphs in exploring data

When the spurious observations in Group B are removed, not only does the distribution become more normal-like but also the homogeneity of variances between Group A and Group B improves (see formal tests above). Under these conditions, the data set has ideal levels of normality and homogeneity that make both groups' means comparable. A two-tailed *t*-test indicates that both groups have different means, t (36) = 2.41, p = 0.02. In terms of the hypothetical research scenario, the video clips did arouse

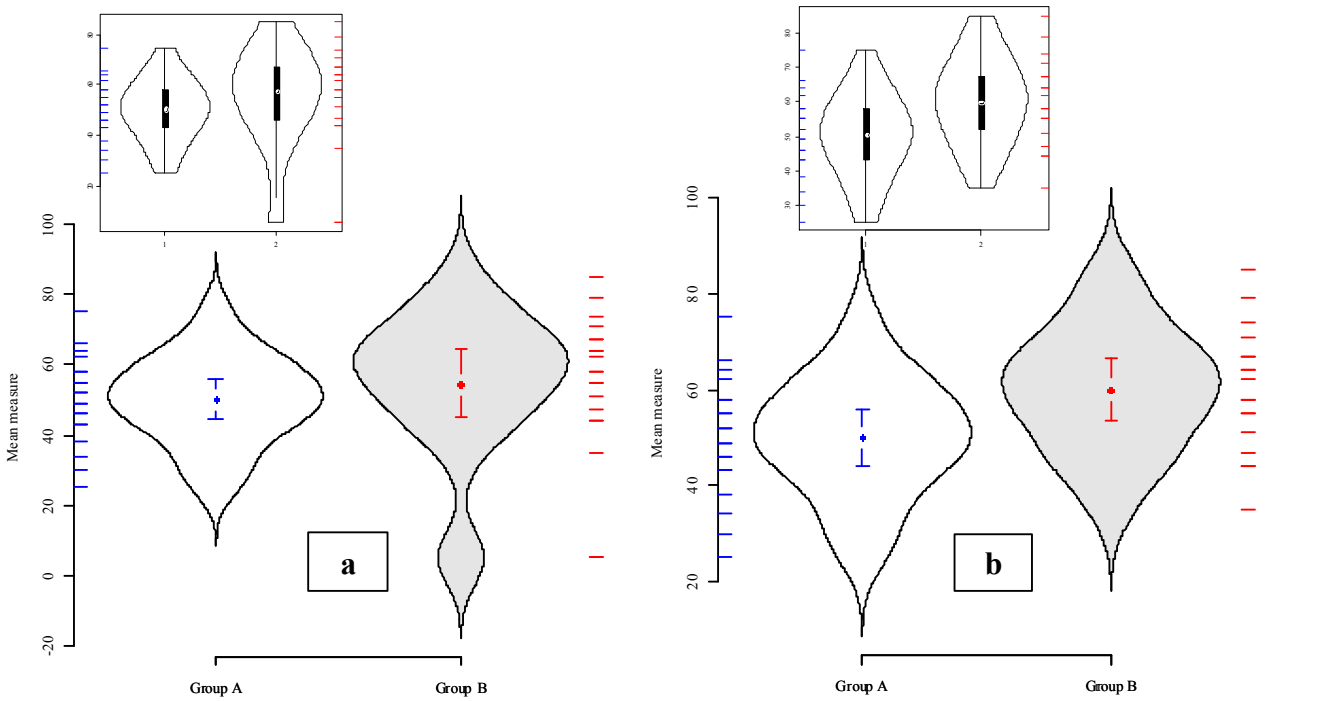

Figure 6. Violin plots representing the mean and the 95% CI of groups of data before (a) and after (b) outlier removal. The rugs next to each data set are added in order to highlight the actual observations. Insets show the traditional violinplot representing the groups of data. The black bars represent the first and third quartile and the white dots represent the median. Note that the traditional violinplot and the modified violinplot use different types of kernel densities to represent the density estimate for each data set. Hence, the obvious visual difference between the two violinplots.

participants' happiness and their free recall test scores were significantly different from the scores given by the participants who watched videos with neutral emotional content (see Figure 3b).

As mentioned earlier, the hypothetical situation presented here exemplifies a Type II error (see Figure 2a) where the treatment group's mean *was* higher than that of the control group even before the spurious observations were removed. So, if the researcher of this hypothetical situation looked only at the groups' means, (s)he would likely fail to detect the real reason for the non-significant result (i.e., presence of outliers) and falsely conclude that the treatment was not effective enough.

Graphical techniques that represent spread and shape of data

Earlier, kernel densities were presented to highlight changes in data spread and shape given that this information is impossible to extract from bar plots like those representing the results of the *t* test. Unfortunately, kernel densities are graphical techniques which are not commonly reported in research papers unless the paper is focused on the study of distributions. More critically, it is quite rare to find papers reporting parameters of the data's underlying distributions like skewness, kurtosis, normality, and homogeneity values. Kernel densities and summary statistics are of great importance since they throw light on how groups of data differ which in turn provides a basis for further hypothesis testing (see Wilcox, 2004).

Fortunately, there are graphical methods that communicate more about the shape and spread of the data than bar plots do. A very useful graphical technique that enables researchers to have some information about the spread of the data is the boxplot (McGill, Tukey, & Larsen, 1978). Although the summary statistics presented in a boxplot centre around the median, the boxplot enables identification of potential outliers. Nevertheless, it is still difficult to note the data's spread and shape without further visual scrutiny. A graphical technique that keeps properties of the box plot but also plots the underlying distribution of the data is called the violin plot (Hintze & Nelson, 1998). Violin plots are a recent technique used to report data in other sciences, e.g., biology (e.g., Julenius & Pedersen, 2006), economics (Chumpitaz, Kerstens, Paparoidamis, & Staat, in press), and politics (Kastellec & Leoni, 2007), but it has not been used in exploring or reporting data in psychological research. The core feature of the violin plot is that it presents the same information given in a boxplot plus a smoothed histogram - a density estimate - of the groups of data (see insets in Figure 6).

A variation of the violin plot based on the mean

Violin plots are a very informative graphical technique since they show the spread and shape of a data set based on statistics around the median. This fact might, though, discourage many researchers to implement it since summary statistics and computations around the mean are the usual currency. However, using software for statistical computing and graphics, violin plots can be customised to represent summary statistics around the mean.

The variation on the violin plot presented here is implemented in order to show the density estimate of the data plus 95% confidence intervals (CI) around the mean. The core advantage of plotting the mean is that it is a very frequent statistic reported in most scientific research. Also, it is advantageous to report the 95% CI since it indicates where the true mean might fall (see Cumming, Fidler, & Vaux, 2007) and gives a visual opportunity to note if groups of data might have significant differences between their means (see Masson & Loftus, 2003). Figure 6 presents Groups A and B, before and after data treatment, as violin plots together with their means and 95% CIs around them.

A brief note on the computation of confidence intervals and their interpretation

The type of 95% CIs assumed here are not those computed using *z* scores (i.e., the 1.96 value, $\overline{x} \pm z_{95\%} \times \frac{\sigma}{\sqrt{n}}$) but the *t* critical values for two-tailed tests ($\overline{x} \pm t_{95\%,df} \times \frac{s}{\sqrt{n}}$). This assumption is adopted on the basis that the computation based on *z* scores applies to situations when population variance is known (which usually is never known) or the sample size is large (see Cumming, 2007). Cumming (2007) presents some recommendations on how to interpret confidence intervals when they are reported graphically. The essential idea is that the closer the confidence interval of one of the groups gets to the mean of the other group, the closer the *p* value gets to the significance level (i.e., 0.05). In other words, a rule of thumb to visually estimate when two groups have significantly different means is when there is less than 50% of overlap between the CIs of the groups. Note that these recommendations are straightforward only when groups have homogeneous variances which are graphically denoted by CIs of similar length. However, as in most cases groups should have homogeneous variances for a parametric test to be performed (see above), the rule of thumb proposed here holds.

Conclusions

This paper stresses the need to use graphical techniques to explore and report data by exploring and analysing a simulated small data set. Recommended procedures for data analysis are presented and an educated routine is proposed in order to fit data to parametric tests' assumptions. Although the procedures and the routine are well-founded, they are by no means exhaustive and raise questions that deserve further empirical investigation.

Combining various graphical techniques permits researchers to know more about the data and have access to relevant information about the spread and shape of the data. This information is supported also by essential summary statistics like the confidence intervals around the mean. Given that the mean is the statistic most frequently reported in psychological research and other sciences, future work should propose graphical techniques which represent essential summary statistics around the mean and give information about the data's distribution (e.g., Marmolejo-Ramos & Tian, 2009).

The use of violin plots should start to take place in the report of psychological research given the qualities they have. As mentioned earlier, the advantages of using violin plots is that they display the actual spread and shape of the data and can be customised to show basic summary statistics. It was also graphically demonstrated that violin plots can be altered to show the mean and the confidence intervals around it in conjunction with the kernel density of the data. The code that produces the modified violinplot can be obtained from the journal's website [3].

References

Behrens, J. T. (1997). *Principles and procedures of exploratory data analysis*. Psychological Methods, 2, 131-160.

Behrens, J. T., & Yu, C. H. (2003). *Exploratory data analysis*. In J. A. Schinka & W. F. Velicer, (Eds.), Handbook of psychology Volume 2: Research Methods in psychology (pp. 33-64). New Jersey: John Wiley & Sons, Inc.

Bland, J. M., & Altman, D. G. (1996) *Transforming data. British* Medical Journal, *312*, 770.

Bland, J. M., & Altman, D. G. (1996a). *The use of transformations when comparing two means.* British Medical Journal, *312,* 1153.

Bland, J. M., & Altman, D. G. (1996b). *Transformations, means, and confidence intervals*. British Medical Journal, *312,* 1079.

Bland, J. M., & Altman, D. G. (1996c). *Logarithms*. British

[3] The graphs in this article were created using R programming language (v. 2.7.1), except Figures 1, 2, 3, and 5. R is a freeware and can be downloaded from http://www.r-project.org (click on CRAN, look for the nearest mirror country, select your platform (Linux, Mac, or Windows), click on "base", and download the .exe file. The supporting packages can be obtained from the CRAN page and clicking on "packages". An easier way to install packages can be done from the R program window: look for "packages" option, and then "install package(s)").

Medical Journal, *312*, 700.

Chumpitaz, R., Kerstens, K., Paparoidamis, N., & Staat, M. (in press). *Hedonic price function estimation in economics and marketing: revisiting Lancaster's issue of "noncombinable" goods*. Annals of Operations Research.

Correa, J. C., Iral, R., & Rojas, L. (2006*). Estudio de potencias de prueba de homogeneidad de varianza [A study on homogeneity of variance tests]*. Revista Colombiana de Estadística, *29*, 57-76.

Cowles, M. & Davis, C. (1982). *On the origins of the .05 level of statistical significance*. American Psychologist, *37*, 553-558.

Cumming, G. (2007). *Inference by eye: Pictures of confidence intervals and thinking about levels of confidence*. Teaching Statistics, 29, 89–93.

Cumming, G., Fidler, F., & Vaux, D. L. (2007). *Error bars in experimental biology*. Journal of Cell Biology, *177*, 7-11.

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage.

Frisch, A. S., Shamsuddin, K., & Kurtz, M. (1995). *Family factors and knowledge: Attitudes and efforts concerning exposure to environmental tobacco among Malaysian medical students*. Journal of Asian and African Studies, *30*, 68-79.

Gelman, A. (2004). *Exploratory data analysis for complex models.* Journal of Computational and Graphical Statistics, *13* (4), 755-779.

Guy, S. C., & Cahill, L. (1999). *The role of overt rehearsal in enhanced conscious memory for emotional events*. Consciousness and Cognition, *8*, 114-122.

Hintze, J. L., & Nelson, R. D. (1998). *Violin plots: A box plot-density trace synergism.* American Statistician, *52*, 181-184.

Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). *Data analysis: continuing issues in the everyday analysis of psychological data*. Annual Review of Psychology, *46*, 433-465.

Julenius, K., & Pedersen, A. G. (2006). *Protein evolution is faster outside the cell*. Molecular Biology and Evolution, *23*, 2039-2048.

Kastellec, J. P., & Leoni, E. L. (2007). *Using graphs instead of tables in political science*. Perspectives on Politics, *5* (4), 755-771.

Kline, R. B. (2008*). Becoming a behavioral science researcher: A guide to producing research that matters*. New York: Guilford.

Levene, H. (1960). *Robust tests for equality of variances. In I. Olkin* (Ed.), Contributions to Probability and Statistics (pp. 278-292). Palo Alto, CA: Stanford University Press.

Lilliefors, H. (1967). *On the Kolmogorov-Smirnov test for normality with mean and variance unknown*. Journal of the American Statistical Association, 62, 399-402.

Marascuilo, L. A. & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type-IV errors. American Educational Research Journal, 7, 397-421.

Marmolejo-Ramos, F., & Tian, S. (2009). The shifting boxplot. A variation of the boxplot that represents informative summary statistics around the mean. Manuscript in preparation.

Masson, E. J., & Loftus, G. R. (2003). *Using confidence intervals for graphically based data interpretation*. Canadian Journal of Experimental Psychology, 57 (3), 203-220.

McGill, R., Tukey, J., & Larsen, W. A. (1978). *Variations of boxplots*. American Statistician, 32, 12-16.

Moorehouse, M. J., & Sanders, P. E. (1992). *Children's feelings of school competence and perceptions of parents' work in four sociocultural contexts*. Social Development, 1, 185-200.

Mosteller, F. (1948). *A k-sample slippage test for an extreme population*. Annals of Mathematical Statistics, *19*, 58-65.

Olivier, J., Johnson, W. D., & Marshall, G. D. (2008). *The logarithmic transformation and the geometric mean in reporting experimental IgE results: What are they and when and why to use them?* Annals of Allergy, Asthma & Immunology, 100 (4), 333-337.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). *Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration*. Personnel Psychology, 44, 473- 486.

Osborne, J. (2002). Notes on the use of data transformations. Practical Assessment, Research & Evaluation, 8 (6).

Peña, D, & Prieto, F.J. (2001), *Multivariate outlier detection and robust covariance matrix estimation*, Technometrics, 43 (3), 286-310.

R Project for Statistical Computing. (2007). R [Computer Software]. Retrieved April 29, 2009, from http://www.r-project.org

Sato, T. (1996). *Type I and Type II errors in multiple comparisons*. Journal of Psychology, 130 (3), 293-302.

Shiffler, R. E. (1988). *Maximum z scores and outliers*. American Statistician, 42, 79-80.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. London: Chapman and Hall.

Tabachnick, B. G, & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston: Allyn & Bacon.

Thompson, G. L. (2006). *An SPSS implementation of the non-recursive outlier deletion procedure with shifting z-score criterion* (Van Selst & Jolicoeur, 1994). Behavior Research Methods, 38 (2), 344-352.

Tukey, J. W. (1969). *Analysing data: sanctification or detective work?* American Psychologist, *24* (2), 83-91.

Valenzuela, M. (1997). *Maternal sensitivity in a developing society: The context of urban poverty and infant chronic under nutrition.* Developmental Psychology, 33, 845-855.

Van Selst, M., & Jolicoeur, P. (1994). *A solution to the effect of sample size on outlier elimination*. The Quarterly Journal of

Experimental Psychology, 47A (3), 631-650.

Wilcox, R. R. (2004). *Kernel density estimators: An approach to understanding how groups differ*. Understanding Statistics, 3, 333-348.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. The Journal of Experimental Education, 67 (1), 55-68.

Article received September 1st, 2009
Article accepted September 13th, 2009