International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Cousineau D., Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research, 3* (1), 58-67.

Outliers detection and treatment: a review.

Detección y tratamiento de valores extremos: una revisión.

Denis Cousineau
Université de Montréal
Sylvain Chartier
University of Ottawa

ABSTRACT

Outliers are observations or measures that are suspicious because they are much smaller or much larger than the vast majority of the observations. These observations are problematic because they may not be caused by the mental process under scrutiny or may not reflect the ability under examination. The problem is that a few outliers is sometimes enough to distort the group results (by altering the mean performance, by increasing variability, etc.). In this paper, various techniques aimed at detecting potential outliers are reviewed. These techniques are subdivided into two classes, the ones regarding univariate data and those addressing multivariate data. Within these two classes, we consider the cases where the population distribution is known to be normal, the population is not normal but known, or the population is unknown. Recommendations will be put forward in each case.

Key words: Statistics, outlier detection, outlier treatment.

RESUMEN

Los valores extremos son observaciones o medidas que son sospechosas en tanto que son mucho menores o mucho mayores que el resto de las observaciones. Estas observaciones son problemáticas en tanto que puede que no sean causadas por los procesos mentales que están siendo estudiados o puede que no reflejen la habilidad que se está estudiando. El problema es que unas pocas observaciones extremas son suficientes para distorsionar los resultados (alterando el desempeño medio, incrementando la variabilidad, etc.). En este artículo se revisan varias técnicas diseñadas para detectar observaciones extremas. Estas técnicas se subdividen en dos clases, aquellas relacionadas con datos univariados y aquellas relacionadas con datos multivariados. Dentro de estas dos clases, se consideran casos en que la distribución de la población es asumida como normal, casos en que la distribución es normal pero no conocida, o casos en que la población es desconocida. Para cada escenario se proponen algunas recomendaciones.

.

Palabras clave: intervalos de confianza, estadística de los intervalos, guías, representación gráfica, encuestas nacionales, aproximación Bayesiana.

Article received/Artículo recibido: December 15, 2009/Diciembre 15, 2009, Article accepted/Artículo aceptado: March 15, 2009/Marzo 15/2009
Dirección correspondencia/Mail Address:
Denis Cousineau, Département de psychologie, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal, Québec, Email: denis.cousineau@umontreal.ca
Sylvain Chartier, University of Ottawa, 200, avenue Lees, E-217, Ottawa, Ontario, Canada, K1N 6N5, Email: sylvain.chartier@uOttawa.ca, chartier@uOttawa.ca.

INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH esta incluida en PSERINFO, CENTRO DE INFORMACION PSICOLOGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET y GOOGLE SCHOLARS. Algunos de sus articulos aparecen en SOCIAL SCIENCE RESEARCH NETWORK y está en proceso de inclusion en diversas fuentes y bases de datos internacionales.
INTERNATIONAL JOURNAL OF PSYCHOLOGICAL RESEARCH is included in PSERINFO, CENTRO DE INFORMACIÓN PSICOLÓGICA DE COLOMBIA, OPEN JOURNAL SYSTEM, BIBLIOTECA VIRTUAL DE PSICOLOGIA (ULAPSY-BIREME), DIALNET and GOOGLE SCHOLARS. Some of its articles are in SOCIAL SCIENCE RESEARCH NETWORK, and it is in the process of inclusion in a variety of sources and international databases.

International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Cousineau D., Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research, 3* (1), 58-67.

INTRODUCTION

Studying human behavior is a difficult enterprise for many reasons. One such reason is that when an individual is accomplishing a given task (ranging from an attentional task to a paper-and-pencil questionnaire), there are processes other than those of interest occurring at the same time. These processes can be physiological, neurological and/or cognitive. Most of the time, they operate in the background and have no influence on the measures collected. Other times, they may contaminate the results and occasionally, they may even substitute to the processes being studied. An example of physiological process generally operating in the background is blood circulation. fMRI studies which examine the BOLD response will be influenced by blood circulation which is itself influenced by numerous factors, but these influences are assumed to cancel across trials. Other examples, such as lapses of attention, either caused by fatigue or mentalizations not related to the task, contaminate the results as they may increase the time to perform the task or reduce the accuracy of the response. Lapses do not cancel across trials as they only increase processing times. Finally, fast guesses, either caused by incorrect understanding of the goal or lack of motivation, is an example of process that substitutes to the processes that should normally operate.

All those undesired behaviors nevertheless produce measurable responses that may happen to be correct by chance. Hence, a spurious behavior can go undetected because the response obtained from it resembles an appropriate response. Other responses however may (and should) attract attention due to their unusual aspect. Those last ones are denoted *outliers* in the following. When the measure is one-dimensional (e.g. IQ or response time), outlier responses can be suspiciously small or suspiciously large. Hence, a datum lying to the left (right) of the scale is potentially problematic and called a low-outliers (a high-outliers). However, the problematic responses are more likely to be entangled around appropriate responses, so that detecting them is impossible from a purely data-driven perspective. Therefore, we have to either accept their existence, reduce the impact they may have on our inferences or choose experimental settings that minimize their occurrences. Flawed designs can never be corrected by any given analysis. Therefore much attention must be spent in the design phase before it is too late and a lot of outlier responses are collected.

Statistical inferences are often based on tests of means in which the standard deviation is used as a measure of the normal fluctuation of the examined processes. Hence, for such parametric tests, a few high-outliers (or a few low-outliers) can greatly influence the mean response. As a result, the compared means have more chances to be dissimilar if the outliers are not uniformly spread across the various conditions, increasing the likelihood of a Type-I error. Similarly, the presence of both low-outliers and high-outliers will increase the standard deviation, reducing the chances of detecting a significant difference and thus increasing the likelihood of a Type-II error.

The influence of outliers is more important if the sample size is small. It is also more important if the statistic examined is less robust. The sample mean is a moderately robust estimate of the population central tendency so that one outlier among a large sample will have a limited impact (Daszykowksi, Kaczmarek, Vander Heyden, & Walczak, 2007). A more robust estimate of the population central tendency is the median. A few outliers will have a limited impact on this statistic. However, other statistics are much less robust. For example, if a researcher compares standard deviations or coefficients of variation across conditions, the measures will be considerably more influenced by the presence of outliers than the mean. In this case, the quartile deviation may be a better statistic. The skewness of a distribution is a highly unstable statistic; just one outlier in a fairly large sample can distort this statistic completely; the Pearson-2 measure of skewness may be preferable in this case. For a discussion of robust estimates, see Tukey (1977), Mosteller and Tukey (1977) or Ratcliff (1993).

One difficulty with treatments of outliers is that there is no unanimously accepted theoretical framework for the treatment of outliers. Various fields have developed various approaches and rare are the approaches that can be formulated with the concepts of another approach. The readerôs first glance at the literature on this theme is awed by the large number and discordance of the concepts put forth by these techniques. Indeed, very few papers make explicit the context in which these techniques have been developed. In the present review, we will distinguish univariate outliers from multivariate outliers.

Within the univariate cases, we will examine both the situation where the population of scores is assumed to be normally distributed and the situation where the population of scores is of an unknown distribution with a notable asymmetry (skewness). The situation where the population is from an unknown but symmetrical distribution has never been examined, but the techniques developed for the normally distributed population presumably apply; the situation where the population is not normal but has a known distribution will be discussed at the end of the review, as the univariate and multivariate cases are handled in the same manner in this situation.

Within the multivariate cases, we will consider the situation where the population is assumed to have a multinormal distribution. Within this situation, we distinguished cases where there is one dependant variable and many predictors from the cases where there are many

dependant variables. We will end this review with a more general consideration: Should outliers be examined within subject? within conditions?

Throughout this review, we have selected from the vast array of possible techniques the ones that are the most agreed upon or the more promising. Some of these techniques are rather elaborate. In these cases, the algorithm is outlined without much detail, concentrating on the strengths and limitations of the techniques. Nevertheless, it is our feeling that the definite techniques have not been found .This review should therefore simply be considered as a milestone in the continued research for dealing properly with outliers. We hope that it will encourage readers to be aware of the problems caused by outliers and to look for appropriate remedies. As G. Thompson (2006, p. 346) wrote: ñIn a field in which all statistically significant effects [é] are considered interpretable, it is clear that a naïve approach to outlier screening can be costly.ò

1. The univariate domain

1.1- The data are assumed to follow a normal distribution

When the data follow a normal distribution, the most salient characteristic of the data is the near symmetry. If outliers are equally likely to be low-outliers or high-outliers, their influence on the mean may be minimal because they counteract each other. In this case, a Type-II error may be the most likely consequence of the presence of outliers, a consequence that can be alleviated if the sample size can be increased markedly.

The most common method to detect outliers in this condition is to use a criterion based on *z*-scores. For example, by excluding all observations that are four standard deviations away from the sample mean, we would eliminate 3 valid observations every 100,000 observations. Of course, this is true only if the presence of outlier is a random process. Hence, if an observation is indeed eliminated from a much smaller sample, it was more probably an outlier than a valid observation (this reasoning will be used more formally in the last section). In the case of biases induced by the experiment (e.g. lack of motivation, fatigue), the number of outliers can be quite high and they should be expected even in small samples.

The criterion to choose (e.g. 4 standard deviations away from the mean) is a matter of debate. A good starting point is to first set a decision criterion α which will indicate how severe we wish to be before tagging an observation as being an outlier (and divide α by two as outliers can be at either end of the distribution). This decision criterion should be small so that the decision stays very conservative (i.e. it has a bias toward keeping the data). To that end, some authors suggest to use a Bonferroni correction based on the sample size, n, so that the decision criterion is $1 - \alpha/(2n)$. The level 0.01 is often chosen. Table 1 lists some critical *Z* as a function of the decision criterion and whether a Bonferroni correction was used (and the sample size in this eventuality).

Table 1. *z-score beyond which a datum (sign ignored) is considered an outlier as a function of the decision criterion and whether a Bonferroni correction is used or not.*

Decision criterion	No Bonferroni correction	Sample size if corrected				
		10	20	30	50	100
0.10	1.645	2.576	2.807	2.935	3.090	3.291
0.05	1.960	2.807	3.023	3.144	3.291	3.481
0.01	2.576	3.291	3.481	3.588	3.719	3.891
0.005	2.807	3.481	3.662	3.765	3.891	4.056

Although such criteria are often used, they are problematic in the case of small samples sizes. Consider for example a sample of size 3. Whatever the data are, the z scores will never exceed (ignoring the sign) the value 1.16 (try using two IQ of 100 and one IQ of one billion!). Hence, with a sample size of 3, even with a liberal decision criterion of 10%, no datum will ever be extreme enough to be an outlier. Shiffler (1988) has computed the maximum possible *z*-core that can be obtained as a function of sample size. In his Table 1, we see that for a sample of size 10, no z scores will ever be larger than 2.8, so that using a decision criterion of 5% and a Bonferroni correction, no outlier will ever be detected. Hence, a screening performed only on the basis of *z*-scores will lead to an erroneous confidence that all the data are legitimate when sample sizes are small. This is a reminder that visual inspection should always be performed.

1.2- The data are from an unknown but asymmetrical population

Measures that have an asymmetric distribution include frequencies of uncommon events (e.g. number of nightmares in a week) and measures that have a definite lower bound but no upper bound (e. g. salary). However, the most commonly studied measure which is definitely not symmetrical is the response time to complete a task. Numerous studies have verified this fact and it is not disputed (Cousineau & Shiffrin, 2004; Rouder, Lu, Speckman, Sun & Jiang, 2005; Ratcliff, 1993). In the simplest of task, the same-different task (Bamber, 1969), response times are typically between 260 ms and 1000 ms

with a standard deviation of approximately 120 ms.[1] With these figures, there remains very little room to place a fast-guest, low-outlier (fast guess can hardly be faster than 160 ms); however, a high-outlier caused by, say, a lapse of attention, can easily be larger than 1500 ms. The former outlier is 2.8 standard deviation below the mean whereas the latter is more than 8 standard deviation above the mean! As seen with this example, asymmetry is a major concern, and it has never been addressed by the techniques created to work through response times.

A transformation approach

One approach would be to first make the data symmetrical by the use of a non-linear transformation. On the transformed data, the outliers could be located using a technique taken from the normal case described above. Presumably with such an approach, outliers would also be more symmetrically away from the central tendency, providing an equal chance of locating low- and high-outliers. However, there is no single technique that makes the data symmetrical when they have been contaminated with outliers.

We have explored the three commonly used transformations (Log-transform, square-root transform, arcsin transform) and have found that the following modification of the square root transformation is very apt to locate outliers at either side of the distribution for response time data:

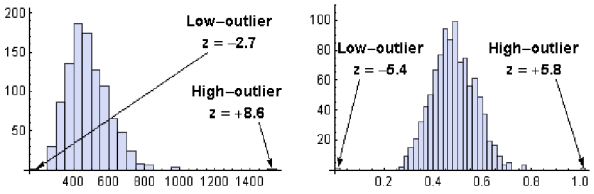
$$y = \sqrt{\frac{x - X_{(1)}}{X_{(n)} - X_{(1)}}} \qquad (1)$$

in which $X_{(1)}$ denotes the smallest item of the sample X, and $X_{(n)}$ denotes the largest. Dividing by the range (the largest observations minus the smallest) normalizes the data so that they are located between 0 and 1, with the smaller at exactly 0 and the largest at exactly 1. This step bounds the data into the range [0..1]. The square root then enlarges observations that are the smallest, pushing the lower part of the distribution towards a more central location. Once this transformation is completed, z scores of the transformed data can be computed ($z = (y - \overline{Y})/s_Y$).

With this square root transform, a score of 160 has a *z*-score on y of -5.4 and a score of 1500 has a *z*-score on y of +5.8. Figure 1 shows simulated data that were used and the result of the transformation. Despite the asymmetry of the raw data, the *z*-scores of the outliers are reasonably symmetrical. This approach can therefore locate outliers at both ends of the distribution with equal chance.

The square-root transform works well for response times and for other measures whose population distribution is moderately and positively skewed. It will not work in cases of extreme asymmetry (e.g. salary). However, for such population, it is merely impossible to detect high outliers since any extremely large measure is always possible in such populations.

Figure 1: *Examples of simulated data resembling response times. Left: raw data with two outliers on either side of the distribution; right: same data after the square root transformation of Eq. (1) was used.*



Recursive and non-recursive approaches using adaptive criterion

Van Selst and Jolicoeur (1994) have taken a different approach. Their argument starts with the observation that low-outliers may have a smaller impact on the mean. Hence, they were primarily interested in locating high-outliers. However, as noted by Ulrich and Miller (1994), removing valid data from only the upper end of the distribution will reduce the mean relative to the true population mean. Hence, as the sample size is larger, the odds that valid data lay above a criteria increase, data that will be removed, resulting in a smaller observed mean. Such procedure applied to an asymmetrical distribution therefore introduces a bias: the larger the sample size, the smaller the observed mean will be after the procedure is applied. Since it is not always possible to have conditions with equal number of observation (either for methodological reasons or because erroneous responses are removed from the analyses), the potentially significant effects could be the result of the manipulation or the result of the unequal number of observations.

In order to avoid such bias for reaction time data, they proposed to use a criterion for exclusion based on the sample size. In a series of Monte Carlo simulation, they have estimated what would be the appropriate criterion for various sample sizes. These values are reproduced in Table 2. The procedure was automatized by G. Thompson (2006) within SPSS. These adaptive criterion were labeled ñmoving criterionò by Van Selst and Jolicoeur (1994). The

[1] For the purpose of this illustration, we generated simulated data with a Weibull distribution and a shape parameter of 2.0, a scale parameter of 250 and a location parameter of 260. This distribution has a skew of 0.63, not an extreme skew but visible on a frequency plot.

difficulty with this table is that it is not possible to alter the decision criterion. These decision criterions were chosen to mimic the bias that would occur with a criterion of 2.5 applied to a sample of size 100. This represents roughly a criterion decision of 1 %.

The authors also presented a recursive version of the above technique in which the *z*-scores are computed by first excluding the most extreme datum. The process is repeated after each datum is removed until no more responses are removed. The rational for that recursive procedure is that high-outliers increase the standard deviation, which results in unrealistically small *z*-scores. Accompanying this method is a new set of criterion adapted to the recursive method (also obtained from Monte Carlo simulations). These criteria are larger for the reason noted above. As pointed by the authors, the adaptive non-recursive and adaptive recursive methods both avoid introducing biases in the resulting mean (means are not influenced by the sample size). However, it is fairly easy to show that both methods are equivalent. Hence, the non-recursive is to be preferred over the recursive.

Table 2. *z-score criterion for excluding an observation as being an outlier, ignoring the sign (from Van Selst and Jolicoeur, 1994).*

Sample size	criterion
100	2.500
50	2.480
35	2.450
30	2.431
25	2.410
20	2.391
15	2.326
12	2.246
10	2.173
9	2.120
8	2.050
7	1.961
6	1.841
5	1.680
4	1.458

With this square root transform, a score of 160 has a *z*-score on y of -5.4 and a score of 1500 has a *z*-score on y of +5.8. Figure 1 shows the simulated data that were used and the result of the transformation. Despite the asymmetry of the raw data, the *z*-scores of the outliers are reasonably symmetrical. This approach can therefore locate outliers at both ends of the distribution with equal chance.

The square-root transform works well for response times and for other measures whose population distribution is moderately and positively skewed. It will not work in cases of extreme asymmetry (e.g. salary). However, for such population, it is merely impossible to detect high outliers since any extremely large measure is always possible in such populations.

Recursive and non-recursive approaches using adaptive criterion

Van Selst and Jolicoeur (1994) have taken a different approach. Their argument starts with the observation that low-outliers may have a smaller impact on the mean. Hence, they were primarily interested in locating high-outliers. However, as noted by Ulrich and Miller (1994), removing valid data from only the upper end of the distribution will reduce the mean relative to the true population mean. Hence, as the sample size is larger, the odds that valid data lay above a criteria increase, data that will be removed, resulting in a smaller observed mean. Such procedure applied to an asymmetrical distribution therefore introduces a bias: the larger the sample size, the smaller the observed mean will be after the procedure is applied. Since it is not always possible to have conditions with equal number of observation (either for methodological reasons or because erroneous responses are removed from the analyses), the potential statistically significant effects could be the result of the manipulation or the result of the unequal number of observations.

In order to avoid such bias for reaction time data, they proposed to use a criterion for exclusion based on the sample size. In a series of Monte Carlo simulation, they have estimated what would be the appropriate criterion for various sample sizes. These values are reproduced in Table 2. The procedure was automatized by G. Thompson (2006) within SPSS. These adaptive criterion were labeled ñmoving criterionò by Van Selst and Jolicoeur (1994). The difficulty with this table is that it is not possible to alter the decision criterion. These decision criterions were chosen to mimic the bias that would occur with a criterion of 2.5 applied to a sample of size 100. This represents roughly a criterion decision of 1 %.

2. The multivariate domain

To simplify the presentation, the multivariate domain will be covered through the general linear model. Therefore, focus will be put on the multiple regression case (ANOVA being a special case) in the first part, then on mutivariate multiple regressions (including canonical correlation, discriminant analysis and MANOVA) in the last part.

International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

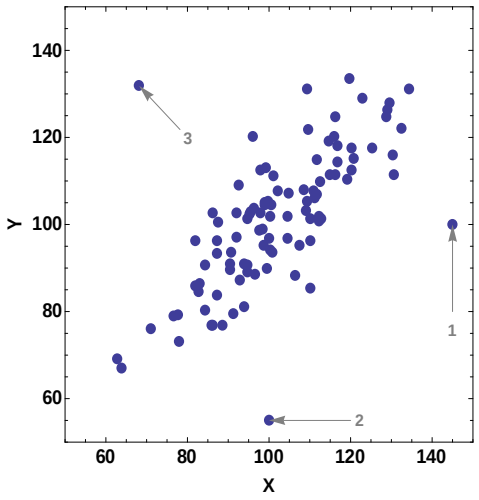
2.1- Multiple regression

This section deals with multiple regressions in which one dependant variable **Y** is predicted by a set of predictor variables **X**. The standard model assumes normality of the scores as well as a linear relationship between the predictors.

In multiple regressions, three types of outliers can be encountered (Figure 2). An outlier can be an extreme case with respect to the independent variable(s) **X** (case 1), the dependant variable **Y** (case 2), or both (case 3). Not all outliers will have an impact on the regression line. In the case where there is only one predictor, detecting an outlier is straightforward using a scatterplot (e.g. Figure 2).

However, such representation is harder for two predictors and impossible for three and more. In addition, a univariate outlier may not be extreme in the context of multiple regressions, and a multivariate outlier may not be detectable in a two-variable or a one-variable analysis. First, the focus will be given to identifying cases with outlying **Y** observation. These methods apply for example to ANOVA analyses (in which case **X**s are only used for identifying the conditions). Second, we will see techniques that identify **X** outliers. To assess if a given score is an outlier, the procedures below proceed by removing it from the data set and see how a target estimate changes. Finally, the influence of those outliers will be assessed to determine whether they should be removed or not.

Figure 2. *Examples of various outliers found in regression analysis. Case 1 is an outlier with respect to **X**. Case 2 is an outlier with respect to **Y**. Case 3 is an outlier with respect to **X** and **Y**. All three outliers are at the same distance from the population mean (45 units), which corresponds to three population standard deviations.*



Identifying Outlying Y observations

The idea is to perform outlier detection based on the examination of the residuals. Residuals (*e*) can be obtain as a vector of size *n*, the number of data, with

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \qquad (2)$$

where **I** represents the identity matrix (of size $n \times n$) and **H** the hat matrix obtained using

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}. \qquad (3)$$

in which **X** is a $n \times p$ matrix containing the *p* predictors for the *n* observations.

In order to test for outliers, remove a particular data and "look" for the effect of deletion on the regression line. If the predicted **Y** deteriorates a lot, then we can affirm that the deleted data was an outlier. On the other hand, if following the deletion of a given datum, the residual has not changed significantly, then the data is not an outlier with respect to **Y**. A new regression line is not required for each deleted data under investigation because the deleted residuals can be obtained from **e** and **H**. It is further possible to studentize the deleted residual with the following

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}. \qquad (4)$$

where *n* represents the number of observations, *p* is the number of predictors, *SSE*, a scalar, is the sum of square of the error ($SSE = \mathbf{e}^{\mathrm{T}}\mathbf{e}$), h_{ii} is the i^{th} diagonal element of the hat matrix and, e_i, is the i^{th} residual (Neter and Wasserman, 1974). Since the deleted residual follows the Student *t* distribution, we can use a critical value based on Bonferroni correction $t_{1-\alpha/(2n)}(n-p-1)$.

Identifying Outlying X observations

Going back to the hat matrix **H**, from Equation 3, we see that the matrix is determined using the predictors **X** alone. Therefore, the hat matrix is useful to indicate whether or not a given set of predictor includes outliers. More precisely, h_{ii} measures the role of the **X** values that will influence the fitted value $\hat{Y}_i$. In this context, the diagonal of h_{ii} is called the leverage. As a rule of thumb, a leverage value h_{ii} will be considered large if it is more than twice as large as the mean leverage value ($\bar{h} = p/n$). This rule applies if the number of observation is large relative to the number of predictors (Kutner, Nachtseim, Neter, & Li, 2004).

International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Cousineau D., Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research, 3* (1), 58-67.

Once outlying observations have been identified, we need to establish if they are influential, in other words, if their exclusion will have a significant impact on the fitted regression function as a whole.

Identifying Influential Cases

To assess the influence of possible outliers, three common tests are used: The DFFITS (Belsley, Kuh, & Welsch, 1980), Cook's Distance (Cook, 1977) and the DFBETAS (Belsley, Kuh, & Welsch, 1980). The DFFITS is useful to measure the influence of a single case *i* has on the fitted value $\check{Y}_i$. It is therefore quite similar to identifying outlying **Y** (two sections above). This measure is determined by

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}} \qquad (5)$$

where t_i represents the studentized deleted residual (Equation 4) and h_{ii} the leverage value. It is suggested (Kutner, Nachtseim, Neter, & Li, 2004) that a case is influential if its DFFITS exceeds 1 for small to medium data (less than 30 observations) sets and $2\sqrt{p/n}$ for large one (more than 30 observations).

In contrast to the DFFITS, Cook's distance (D_i) will consider the influence of a given case *i* on all the *n* fitted value. This distance is obtained by

$$D_i = \frac{e_i^2}{pMSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right) \qquad (6)$$

where e_i represents the residual of the i^{th} datum (Equation 2) h_{ii}, its leverage and *MSE* is the mean square error, given by *SSE / (n - p)*. Therefore, if e_i, or h_{ii}, or both have high values, the Cook's distance will be high. To assess if a given D_i is influential or not, select a criterion using the percentile value of a Fisher Ratio distribution $F(p, n\text{-}p)$. If the percentile is less than about 20%, the case under investigation has little apparent influence on the fitted value.

Finally, influence on the regression coefficients can be measured using the DFBETAS. Since the DFBETAS value is a difference between the estimated regression coefficient and the coefficient when the i^{th} case has been omitted, its value can be positive or negative. A positive value indicates an increase in the estimated coefficient, while a negative value indicates a decrease. Formally, the DFBETAS is obtained following

$$DFBETAS_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}_{jj}}}. \qquad (7)$$

in which b_k is the k^{th} regression coefficient, $b_{k(i)}$ is the same coefficient after datum *i* has been deleted and $MSE_{(i)}$ is the mean square error after deletion of datum *i*. A case will be considered influential if the absolute value of DFBETAS exceeds 1 for small to medium data sets (less than 30 observations) and $2\sqrt{n}$ for large data sets (more than 30 observations). Some authors suggest as rule of thumb the value of 2 (Kutner et al., 2005). As a final note, outliers might not be influential, a high leverage value might not be influential and an influential score might not be an outlier. It is therefore recommended to remove only outliers that are influential as well.

We tested the above methods using simulated data sets generated from a multinormal distribution with mean (100, 100), standard deviations of 15 for both variables and a population coefficient of correlation of .80. Three outliers (presented in Figure 2) were also included in the distribution.

The studentized residual technique identified only the case 3 as being an outlier. Both the DFFITS and Cookôs recognized this item as influential. The studentized residual technique failed to notice case 2, even though it is aimed at identifying such type of outliers and therefore should be avoided. The leverage method identified case 1 as an outlier, but only Cookôs D detected this item as influential. Of the methods reviewed here, none could identify case 2 as an outlier despite its clear symmetry with case 1.

2.2- Nonlinear regression

Until now, only the linear model has been considered under the assumption that the processes underlying the data were normally distributed. In the case of nonlinear regressions, methods for detecting outliers generally donôt exist, and when they do, they differ markedly from those of linear regressions (e.g. in the case of logistic regressions or Poisson regressions; McCullagh and Nelder, 1999; Kutner, Nachtseim, Neter, & Li, 2004).

2.3- Mutivariate multiple regression

Multivariate multiple regressions address the cases in which there is no predictor variable and *q* dependant variables **Y** or the more general case with *p* predictor variables and *q* dependant variables.

If the population is multinormal (a multidimensional normal distribution) and the data under

study are simple (dummy coding variable as predictors, like MANOVA), then the Mahalanobis distance can be used (for a review, see Meloun & Militký, 2001). This measure is given by

$$M_i = (\mathbf{Y}_i - \overline{\mathbf{Y}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{Y}_i - \overline{\mathbf{Y}}). \quad (8)$$

where **Y** is the matrix containing the *q* measures for the *n* subjects, $\overline{\mathbf{Y}}$ is the mean across the subjects (a vector of size *q*), and **S** is the variance-covariance matrix of size *q* × *q*. A decision criterion can be chosen since this distance follows a χ^2 distribution with parameter *q* ï 1 (e.g. $\chi^2_{1-\alpha/(2n)}(q$ ï $1)$). This approach is very reliable if there is only one outlier or if the outliers are scattered all around the legitimate data.

The Mahalanobis distance (*M*) is related to the leverage (Equation 3) since

$$h_{ii} = \frac{M_i}{n-1} + \frac{1}{n} \text{ or likewise, } M_i = (n-1)h_{ii} - \frac{n-1}{n} \quad (9)$$

if the **Y** matrix used to obtain **H** has one extra column dummy coded with 1s. We computed the Mahalanobis distance on the items of the data set illustrated in Figure 2. The critical value using a Bonferroni correction is given by $\chi^2_{1-5\%/(2n)}(p-1)$ in which *q* equals 2 and *n*, 100 equals 13.41. The three outliers (cases 1 to 3) have a Mahalanobis distance of 13.5, 15.2 and 23.9. They are the only items exceeding the critical value.

However, the situation is more complicated when the analysis involves continuous multiple dependent and independent variables. In those cases, masking effects may occur and thus may affect the variance-covariance matrix as well as the detection of other outliers in the data sets. Many techniques have been proposed for detection of multiple outliers of various influential degrees. Each solution has its own particular implementation and the details of each technique are beyond the scope of this paper. For a review see for example Meyer, (2003); Walczak & Massart (1998); Wisnowski, Montgomery & Simpson (2001); Daszykowski, Kaczmarek, Heyden & Walczak (2006).

3. The case (univariate or multivariate) in which the population distribution is known

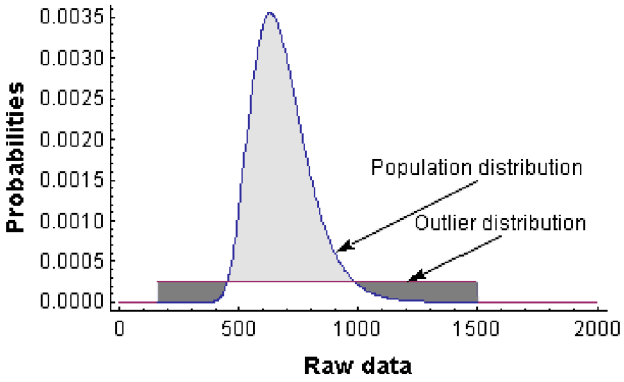
In the case where the population distribution is known, it is possible to adopt a rational approach and compute the odds that a datum was sampled from the population instead of from the distribution of outliers.

To achieve this, letôs assumes that the outliers are uniformly spread out over the whole range of data, that is, from the smallest to the largest datum. Let us define the following probability function:

$$P_{\mathcal{U}[X_{(1)},X_{(n)}]}(x) \quad (10)$$
$$P_{\mathcal{D}}(x) \quad (11)$$

in which $\mathcal{D}$ denotes the distribution of the population assumed known and $\mathcal{U}[X_{(1)}, X_{(n)}]$ denotes the uniform distribution over the range extending from the smallest observation $X_{(1)}$ to the largest observation in the sample $X_{(n)}$. In the case of a uniform distribution, the probability of a certain score x is equal to a constant $1/(X_{(n)} - X_{(1)})$. Figure 3 illustrates the two distributions.

Figure 3. *Two distributions, one representing the population distribution $\mathcal{D}$, the other representing the distribution of spurious observations $\mathcal{U}[X_{(1)}, X_{(n)}]$ extending over the whole range of observed data (here, we assume that the smallest observation $X_{(1)}$ is 160 and the largest, $X_{(n)}$, is 1500. The population distribution depicted here is a Log-normal distribution.*



By computing the ratio

$$\frac{P_{\mathcal{D}}(x)}{P_{\mathcal{U}[X_{(1)},X_{(n)}]}(x)} \quad (12)$$

we get the odds ratio that a certain datum x was sampled from the population relative to it being sampled from a spurious process. An observation x for which this ratio is 99 to 1 or higher represents a datum which is 99 times more likely to be an outlier than a valid observation. Since a ratio of 99 : 1 represents a probability of $\frac{99}{99+1} = 99\%$, it means that the decision criterion is 1%. For a decision criterion of 0.1%, we would look for observations that are 999 times more likely to be a spurious response than a valid response.

Referring to Figure 3, an observation with a ratio of 999 is a datum for which the probability of the outlier distribution is 999 times smaller that the corresponding probability assuming the population distribution.

As an illustration, assuming that the population has for distribution a Log-normal distribution with parameters μ = 6, σ = 0.29, ξ= 260 (reasonable parameters for response times: Heathcote Brown, and Cousineau, 2004), a datum of 160 has a ratio near infinity whereas a datum of 1500 has a ratio of 1211. Both would be rejected as being more plausibly outliers with a confidence level of 0.1%.

In many researches, the population distribution has a known form but the precise parameters are not known. Let the population distribution $\mathcal{D}$ be a function of those parameters denoted collectively by θ, noted $\mathcal{D}(\theta)$. The spurious distribution is given limits taken from the smallest and the largest datum. The observed distribution is therefore a mixture of the two distributions, where a certain proportion, π, of the data are sampled from the population distribution, and the remaining, 1 ï π, are sampled from the uniform distribution. Putting it all together,

$$P(x) = \pi P_{\mathcal{D}(\theta)}(x) + (1-\pi) P_{\mathcal{U}(X_{(1)},X_{(n)})}(x)$$

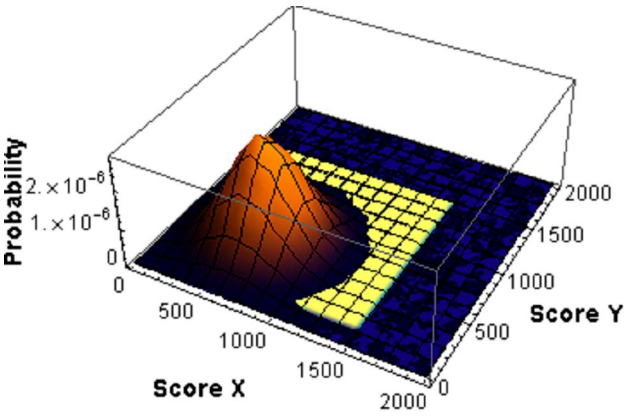
Such probability density function can be fitted to the data by maximizing the likelihood function over the parameters θ and π. Cousineau, Brown and Heathcote, 2004, show how this can be done with more details. Once the parameters θ are obtained, the population distribution is completely determined and the ratio of Equation (12) can be computed.

In the multivariate case, the same equations are valid except that instead of being applied to a single datum (a univariate observation), it is applied to a vector of observations. Figure 4 illustrates the case for a bivariate case in which the population distribution is binormal. Both cases assume that the outliers follow a uniform distribution. This is probably unrealistic, however, for rare low and high outliers, it may be sufficient to characterize them.

4. What to do with outliers?

Outliers that are clearly the result of a spurious activity should be removed. However, in multivariate designs, doing so may result in removing too many participants to the point that the analysis can no longer be performed. Tabachnick and Fidell (2007) suggested replacing the missing data with the mean of the remaining data in the corresponding cell. However, this procedure will tend to reduce the spread of the population, make the observed distribution more leptokurtic, and possibly increase the likelihood of a type-I error. A more elaborate technique, multiple imputations, involves replacing outliers (or missing data) with possible values (Elliott & Stettler, 2007; Serfling & Dang, 2009).

Figure 4. *A mixture of a population distribution with a spurious distribution. The former is Multinormal with a mean of {600, 600}; the latter is uniform within the area delimited by the corners of the smallest and largest values (here assumed equal for both measures only for the purpose of illustration) {160, 160} and {1500, 1500} and shown with a lighter color.*



5. In what cells do we search for outliers?

When the measure has been replicated a large number of times inside a given condition for each participant, as is often the case in experiments involving simple tasks, the search for outliers should be done across the replications. In this case, the scenario is univariate. However, in a large number of research context, it is not possible to replicate the measure (whether the measure is a questionnaire or an observation taken in an ecological setting). It this case, the outliers should first be searched for within condition across the participants. This is also a univariate search for outliers. There may be some chance that the between subject variability follows a normal distribution so that a simple *z*-score cut-off criterion may be used. Finally, if the participants were measured more than once, a multivariate search for outliers is the final step. This step will locate multivariate outliers, that is, subjects who may not be outlier on a single measure, but for which the observed values of a conjunction of two or more measures are suspicious.

As can be seen, there is no single solution to the problem of outlier detection. In the multivariate cases, current research turns around the notion of clustering which assumes that isolated clusters are probably composed of outliers. The final word is therefore yet to come.

REFERENCES

Bamber, D. (1969). *Reaction times and error rates for "same"-"different" judgments of multidimensional stimuli*. Perception and Psychophysics, 6, 169-174.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics : identifying influential data and sources of collinearity*. Wiley series in probability and mathematical statistics. New York: John Wiley & Sons.

Cook, R. D. (1977). Detection of influatial observation in linear regression. *Technometrics, 19*, 15-18.

Cousineau, D., & Shiffrin, R. M. (2004). *Termination of a visual search with large display size effect*. Spatial Vision, 17, 327-352.

Cousineau, D., Brown, S., & Heathcote, A. (2004). *Fitting distributions using maximum likelihood: Methods and packages*. Behavior Research Methods, Instruments, & Computers, 36, 742-756.

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). *Robust statistics in data analysis-a review basic concepts*, Chemometrics and intelligent laboratory systems, 85, 203-219.

Elliott, M. R. & Stettler, N. (2007). Using a mixture model for multiple imputation in the presence of outliers: the 'Healthy for life' project. Applied Statistics, 56, 63-78.

Heathcote, A., Brown, S., & Cousineau, D. (2004). *QMPE: Estimating Lognormal, Wald and Weibull RT distributions with a parameter dependent lower bound*. Behavior Research Methods, Instruments, & Computers, 36, 277-290.

Kutner, M. H., Nachtseim, C. J., Neter, J., & Li, W. (2004). Applied linear statistical models. New York: McGraw-Hill.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models (2nd ed.)*. London: Chapman and Hall.

Meloun, M., & Militký, J. (2001). *Detection of single influential points in OLS regression model building*. Analytica Chimica Acta, 439, 169-191.

Meyer, D. (2003). *Diagnostics for canonical correlation*, Research Letters in the Information and Mathematical Sciences, 4, 79-89.

Mosteller, F., & Tukey, J. W. (1977). Data analysis and regression. Reading: Addison-Wesley.

Neter, J., & Wasserman, W. (1974). Applied Linear Statistical Models. Homewood, Ill.: Richard D. Irwin.

Ratcliff, R. (1993). *Methods for dealing with reaction time outliers*. Psychological Bulletin, 114, 510-532.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). *A hierarchical model for estimating response time distributions*. Psychonomic Bulletin & Review, 12, 195-223.

Serfling, R., & Dang, X. (submitted, 2009). A numerical study of multiple imputation methods using nonparametric multivariate outlier identifiers and depth-based performance criteria with clinical laboratory data. Journal of Statistical Computation and Simulation, 29 pages.

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

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

Tukey, J. W. (1977). Exploratory data analysis. Reading: Addison-Wesley.

Ulrich, R., & Miller, J. (1994). *Effects of truncation on reaction time analysis*. Journal of Experimental Psychology: General, 123, 34-80.

Van Selst, M., & Jolicoeur, P. (1994). *A solution to the effect of sample size on outlier estimation*. The Quarterly Journal of Experimental Psychology, 47A, 631-650.

Walczak, B., & Massart, D.L. (1998). *Multiple outlier detection revisited*, Chemometrics and intelligent laboratory systems, 41, 1-15.

Wisnowski, J. W., Montgomery, D. C., & Simpson, J. R. (2001). *A comparative analysis of multiple outlier detection procedure in the linear regression model*, Computational Statistics & Data Analysis, 36, 351-382.