International Journal of Psychological Research, 2009. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Beaulieu-Prévost, D.,(2010). Gaining confidence with intervals: Practical guidelines, advices and tricks of the trade to face real-life situations. *International Journal of Psychological Research, 3* (1), 46-57.

Gaining confidence with intervals: Practical guidelines, advices and tricks of the trade to face real-life situations.

Ganando confianza con intervalos: guías prácticas, consejos y trucos para enfrentar situaciones reales de investigación.

Dominic Beaulieu-Prévost
Département de Sexologie, Université du Québec à Montréal, Montréal, Canada

ABSTRACT

Confidence intervals and measures of effect size are gradually becoming the standard way of reporting the results of statistical analyses in research articles, used instead of or in addition to *p* values. However, this shift in research practices barely affected teaching practices up to now. This paper is the third of a series written to serve as a general reference on the use of confidence intervals in quantitative social sciences. Its purpose is to provide guidelines, advices and useful tricks of the trade that will allow readers (a) to face most of the statistical problems emerging in real-life research settings and (b) to improve their understanding of confidence intervals and answer more efficiently their questions of interest. The first part of the article briefly introduces the basic elements of an approach based on confidence intervals: Calculations, interpretation, and hypothesis testing. The second part is an attempt to present some of the most important (but sometimes neglected) advanced issues concerning confidence intervals: Graphic representations, complex distributions, national surveys, the larger family of interval statistics (e.g., prediction intervals), and the Bayesian approach to probabilities.

Key words: Confidence intervals, interval statistics, guidelines, graphic representation, national surveys, Bayesian approach.

RESUMEN

Los intervalos de confianza (IC) y las medidas de tamaño de efecto están convirtiéndose gradualmente en la forma estándar de reportar resultados de análisis estadísticos en artículos de investigación, en lugar de, o además de, los valores *p*. Sin embargo, tal cambio en las prácticas de investigación se ha comunicado poco en la enseñanza de la estadística. Este artículo es el tercero en una serie escritos que sirven como referencia general sobre el use de los IC en las ciencias sociales. Este artículo tiene como propósito proveer guías, consejos, y trucos útiles que le permitan al lector (a) enfrentar la mayoría de problemas estadísticos que suceden en situaciones reales de investigación y (b) mejorar su conocimiento sobre los IC y contestar más eficientemente las preguntas de interés. La primera parte del artículo presenta brevemente los elementos básicos acerca del uso de los IC: cómo computarlos, cómo interpretarlos, y cómo usarlos en las pruebas de hipótesis. La segunda parte presenta algunos de los asuntos más importantes (aunque algunas veces negados) acerca de los IC: representaciones gráficas, distribuciones complejas, encuestas nacionales, la familia de la estadística de los intervalos (e.g., intervalos de predicción), y la aproximación Bayesiana a las probabilidades.

.

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:
Dominic Beaulieu-Prévost, Département de Sexologie, Université du Québec à Montréal, C.P. 8888, succ. Centre-Ville, Montréal, CANADA H3C 3P8
Email: dbprevost@gmail.com

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

Beaulieu-Prévost, D.,(2010). Gaining confidence with intervals: Practical guidelines, advices and tricks of the trade to face real-life situations. *International Journal of Psychological Research, 3* (1), 46-57

Tests of statistical significance, also known as null hypothesis testing (NHT), have been the first established gold standard for reporting statistical results and for deciding upon the scientific value of hypotheses in quantitative social sciences. Due to the major problems inherent to this approach, confidence intervals and measures of effect size are gradually becoming the standard way of reporting the results of statistical analyses in research articles, used instead of or in addition to *p* values. However, this shift in research practices barely affected teaching practices up to now. Indeed, many introductory courses and manuals in statistics for the social sciences still present NHT as the only paradigm and most of those who present confidence intervals and measures of effect size still cover them only briefly and present NHT as the main statistical tool.

This paper is the third of a series that I wrote to serve as a general reference on the use of confidence intervals in quantitative social sciences. This series was written for NHT-trained researchers and students, with the explicit objective of bridging the gap between their NHT training and the actual publication standards in the field. While a basic introduction to confidence intervals is provided in each paper, each one focuses on a specific aspect of the issue and they aim at being complementary. The first paper (Beaulieu-Prévost, 2007) clarifies the major methodological and epistemological problems of NHT and introduces confidence intervals as a solution to these problems. The second paper (Beaulieu-Prévost, 2006) focuses on the basic mechanics of calculating confidence intervals and testing hypotheses with an approach based on confidence intervals. The purpose of this third paper is to provide practical guidelines, advices and useful tricks of the trade that will allow readers (a) to face most of the statistical problems emerging in real-life research settings and (b) to improve their understanding of confidence intervals and answer more efficiently their questions of interest. Since it is impossible to cover every aspect of the subject in a single article, an effort was also made to provide quality references covering certain aspects in more depth.

The first part of the article briefly introduces the basic elements of an approach based on confidence intervals: Calculations, interpretation, and hypothesis testing. Readers less familiar with these aspects are referred to the second paper of the series (Beaulieu-Prévost, 2006) for more details. The second part is an attempt to present some of the most important (but sometimes neglected) advanced issues concerning confidence intervals: Graphic representations, complex distributions, national surveys, the larger family of interval statistics, and the Bayesian (i.e., subjective) approach to probabilities.

CONFIDENCE INTERVALS 101: BASIC INTERPRETATION AND APPLICATIONS

What are confidence intervals?

Confidence intervals are mathematically equivalent to tests of significance. In fact, for every test of significance, an equivalent confidence interval can be constructed. However, instead of providing a *p* value to evaluate if an effect is statistically different from zero, confidence intervals provide information about the effect size in the sample and the precision of the parametric estimation of the effect size.

The basic model of a confidence interval is:

$$CI = ES \pm V_C * SE \qquad (1)$$

where the confidence interval (*CI*) is constructed by adding and subtracting from the observed size of an effect (*ES*) the product of its standard error (*SE*) and the two-tailed critical value at the chosen alpha level of statistical significance (V_C). Every value around the effect size and between the upper and lower limits of the interval is included in the confidence interval. When the CI excludes zero, the equivalent test of significance is statistically significant and vice versa. Although the term *effect size* is now often used to refer strictly to standardized, or metric-free, indexes of the size of an effect as observed in a sample such as the *d* statistic (Rosenthal, 1994), it is used in the present article in its older and simpler form, i.e., to refer to the unstandardized size of an effect as observed in a sample (e.g., a correlation or a difference between means). The term *parameter* will be used to refer to the unstandardized size of an effect in the population.

A confidence interval can be intuitively defined as a range of plausible population values for the corresponding parameter. By using the sample to estimate population values, the central value of a confidence interval (called the *point estimate*) represents both the size of the effect in the sample and the best estimate of the parameter in the corresponding population, while the two limits of the interval represent the estimated lowest and highest probable values of the parameter in that population. The width of the confidence interval is specified by a percentage value equals to one minus the value of the chosen alpha level. Thus, an alpha of 0.05 produces a 95% CI. This percentage represents the level of confidence of the interval (assuming a normally distributed variable).

Calculating confidence intervals in real life

Due to their growing popularity, confidence intervals are more and more provided in general statistical software, either as a part of the basic results or as an option

of the analysis. Also, some software allow an easy access to user-written sub-routines that can be easily integrated (e.g., STATA). This flexibility often allows researchers to find sub-routines designed to calculate an impressive variety of confidence intervals. Thus, most calculations should now be done quite easily in most situations. A third option is to use stand-alone calculators or spreadsheets already designed to calculate the required type of confidence intervals or to create such calculators if one has the skills and the formulae. An example of spreadsheet is available with this article via the journalôs web site to calculate confidence intervals for correlations and for a difference between two correlations (Supplementary Notes). Many researchers also provide downloadable spreadsheets or online calculators on internet (e.g., Hopkins, 2009; Beaulieu-Prévost, 2009). These calculators can generally be found with an internet search using keywords such as ñconfidence intervalsò, either ñspreadsheetò or ñcalculatorò, and a keyword describing the type of confidence interval that you want to calculate (e.g., ñcorrelationsò). Naturally, as with any information downloaded from internet, you have to evaluate the quality of your calculator from the authorôs credibility or at least test the calculator with known data. Readers interested in the specific calculations of basic confidence intervals are referred to the second paper of this series (Beaulieu-Prévost, 2007).

A note on statistical inference and sample representativeness

As for NHT, estimates produced from confidence intervals are based on the assumption that the sample is representative of a specific population. Samples created from a random sampling procedure are the best example of representative samples. In these samples, every individual from the target population has equal chances of being part of the sample. However, pure random sampling is rarely possible in most research fields unless the population is small, clearly defined, cooperative and easily accessible.

In many fields of social sciences, samples are often non-probabilistic, i.e., the probabilities of selection associated to the sampling procedure cannot be determined. For example, most of studies in experimental psychology are done with either volunteer participants (who are thus self-selected) or students who have to participate in studies for course credits. In these cases, the assumption of sample representativeness is potentially breached, and the validity of the conclusions coming from the statistical inferences is undermined. Statistical inferences are still a standard procedure in these situations but one has to keep in mind that when sample representativeness is uncertain, the results of statistical inferences (whether NHT or confidence intervals) have to be taken with a grain of salt because they probably underestimate the standard error, and thus overestimate the precision of the parametric estimations. A full treatment of sampling procedures cannot be done here but researchers should be aware of these issues when interpreting statistical results.

How to interpret confidence intervals

To adequately interpret confidence intervals, one has to clarify the notion of probability on which traditional statistical inference is based. Both NHT and confidence intervals are based on what is called the *frequentist approach* to probability. According to a well-known frequentist mathematician, ñthe essential distinction between the frequentists and the non-frequentists is, I think, that the former, in an effort to avoid anything savouring of matters of opinion, seeks to define probability in terms of the objective properties of a population, real or hypothetical, whereas the latter do notò (Kendall, 1949). Indeed, to differentiate probabilities from subjective expectations, frequentists defined what they called empirical or objective probability as the relative frequency of an event over time, i.e., its relative frequency of occurrence after repeating a process a large number of times (ideally an infinity of times) under similar conditions. More explicitly, if confidence intervals could be calculated for an infinity of random samples coming from the same population, the parameter of the population would be included in [1-alpha] of them. It can also be said that conclusions stating that a parameter lies within a confidence interval will err in [corresponding alpha] of the occasions.

However, when a single confidence interval is interpreted, it is inadequate to say that there is a probability of 95% that the parameter is included in the confidence interval or that the parameter is probably included in the confidence interval. From a frequentist point of view, it makes no sense to speak about probabilities for a specific confidence interval. The interval either includes the parameter or it doesnôt. The only meaning that can be given to a specific confidence interval is as a representation of the amount of sampling error associated with that estimate within a specified level of uncertainty. It is thus said that all the values included in a confidence interval can be considered to be equivalent with a level of confidence of [1-alpha]. These values are considered equivalent because the sensitivity of the statistical analysis (i.e., its statistical power) is not high enough to differentiate them.

Confidence intervals can also be interpreted from a subjective point of view. These subjective confidence intervals are called credible intervals. The details of such an interpretation and its consequences will be presented in the second part of this article.

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

Beaulieu-Prévost, D.,(2010). Gaining confidence with intervals: Practical guidelines, advices and tricks of the trade to face real-life situations. *International Journal of Psychological Research, 3* (1), 46-57

Hypothesis testing with confidence intervals

Hypothesis testing with confidence intervals is quite simple and can basically be done with a glance. The first step is to define a hypothesis. It can either be a point hypothesis, defined by a single value, or a range hypothesis, defined by a range of values. The null hypothesis, stating that an effect equals exactly zero, is an example of point hypothesis. However, a psychologist interested about the clinical significance of a treatment effect might want to test if the average symptom reduction due to the treatment is at least equal to 4 points on a standardized symptom scale. This would be an example of range hypothesis, the range of the hypothesis being defined as ñany value equal or less than -4ò.

As long as the adequate confidence interval has been calculated, testing a hypothesis simply implies verifying which one of the following cases apply: (a) If the confidence interval is completely outside of the range of the values defined by the hypothesis, the hypothesis is infirmed and rejected (i.e., $p < 0.05$ for an alpha of 0.05), (b) if the confidence interval is completely included in the range of values defined by the hypothesis, the hypothesis is confirmed and accepted (i.e., $p > 0.95$ for an alpha of 0.05), (c) if the confidence interval is partly included in the range of values defined by the hypothesis and partly excluded from that range, the hypothesis is considered undetermined due to a lack of statistical power (i.e., $0.95 > p > 0.05$ for an alpha of 0.05). A detailed explanation of the topic is provided in the second paper of the series (Beaulieu-Prévost, 2006).

ADVANCED APPLICATIONS:

PRACTICAL GUIDELINES, ADVISES AND TRICKS OF THE TRADE

As for any statistical approach to inference, knowing the basic elements is often not enough to adequately face practical situations. Real-life data tend not to behave like the idealized situations used in basic examples and some adjustments might be required to adapt the theoretical principles to these concrete situations. The purpose of this second part is twofold: (a) to present solutions to common problems related to the use of confidence intervals in real-life research settings, i.e., the graphic representation of results, complex distributions and complex sampling procedures such as national surveys, and then (b) to present advanced procedures that expand the possibilities offered by an approach based on confidence intervals, i.e., the other types of interval statistics and the subjective interpretation of confidence intervals.

Graphic representation of confidence intervals

An excellent way to help readers to remember the most important results of a study is to present a graphic representation of these results. It is rarely done for measures of associations (e.g., correlations) but frequently done when presenting mean scores and group differences. In a traditional NHT approach, the precision of the estimated parameters is often represented by putting error bars around the point estimations. The function of these bars is to represent the amount of measurement error for the estimation, smaller bars representing a more precise estimation. However, error bars are notoriously difficult to interpret. In fact, many researchers do not even know how to interpret them adequately (Belia, Fidler, Williams, & Cumming, 2005). Since these bars usually define an area equivalent to the value of the point estimate +/- one unit of the standard error of the measurement, they are equivalent to a 68% confidence interval under assumptions of a normally distributed variable, which makes them confusing to use for research purposes unless, for a strange reason, you are using an alpha of 0.32.

What increases even further the challenge of adequately interpreting error bars in research papers is that they do not systematically represent the standard error of the estimate. They can also represent the standard deviation or the limits of a traditional confidence interval (e.g., a 95% CI). For this reason, error bars are occasionally called standard error bars (SE bars), standard deviation bars (SD bars) or confidence interval bars (CI bars) depending on the situation. In fact, one should always specify clearly (e.g., in a note under the graphic) the specific unit represented by these bars.

When using confidence intervals, CI bars are often used in graphic representations. These bars are clearly more useful than SE bars because they allow us to know, simply by looking at the graph, the range of probable values for a parameter, whether or not the value is statistically different from zero (by verifying if zero is included or not between the two limits) and to compare the probable values of two parameters (e.g., the pre- and post-treatment levels of depression in a clinical trial).

One of the ways these CI bars are often used is to test whether or not two values are statistically different. According to the basic logic, the two values are statistically different if and only if they do not overlap. This logic allows to visually evaluate the statistical significance of group differences directly. However, this intuitive procedure is flawed and can easily mislead the user. It is indeed true that if two confidence intervals do not overlap (i.e., they share none of their probable values), the difference between the two parameters will be statistically

Beaulieu-Prévost, D., (2010). Gaining confidence with intervals: Practical guidelines, advices and tricks of the trade to face real-life situations. *International Journal of Psychological Research, 3* (1), 46-57.

significant. However, a group difference can be statistically significant event when the two confidence intervals partially overlap. This might seem counter-intuitive but it is due to the fact that people mistakenly believe that a difference between two confidence intervals is the same as the confidence interval of the difference between two scores. While confidence intervals of mean scores are calculated using the standard error of each mean, group differences (and their associated confidence interval) are calculated using another standard error called the standard error of the difference. This standard error is calculated based partly on the standard errors of each group mean but it also depends on the degree of dependence between the groups (for the specific equations, see Beaulieu-Prévost, 2006). This subtle difference can certainly be a source of confusion for many researchers.

When the purpose of a graphic representation is to present group differences or pre- and post-event scores, researchers are traditionally given two options. The first one is to present the confidence intervals of each parameter as stated above. As mentioned, the problem is that these confidence intervals are inappropriate for inferential purposes and cannot be used to evaluate the statistical significance of the difference between two scores. The second option is to directly present the difference scores and to use the appropriate confidence interval of the difference. This second option is clearly more adequate than the first one since it represents exactly what is intended. However, it does not represent the value of the parameters in each group, condition or measurement time because only the difference between the two parameters is presented. Consequently, an important part of the information is lost in the process.

A third option, called inferential confidence intervals, was recently developed to keep the best of both worlds (Tryon, 2001). Basically, this graphic approach uses the format of the first option (i.e., CI bars centered around the point estimate of each parameter) but adjusts the width of the confidence intervals to exactly represent the statistical significance of the difference. Thus, these confidence intervals will overlap if and only if the difference between the two parameters is statistically significant.

These inferential confidence intervals can also be used to improve the meaningfulness of difference scores. Indeed, some difference scores are quite difficult to interpret. A classical example is a difference between two correlations. Since a difference between two correlations is not a correlation (and not even an easily understandable unit), it is hard to know if such a difference can be considered big or not. Inferential confidence intervals can be used in such a case to transform the units of the confidence interval of the difference between two correlations into difference in explained variance, a clearly more meaningful unit. An example of this procedure used in a context of a meta-analysis can be seen in Beaulieu-Prévost & Zadra (2007). An improvement of Tryonôs method has also recently been published (Tryon & Lewis, 2008).

There is presently no gold standard for the graphic representation of statistical inferences (except for the fact that intervals are generally represented as some kind of bars around a point estimate) and it is still a debated issue. Consequently, two main elements should be used as guidelines: (a) make sure that the graphic representation exactly represents the inferential statistics that you use and (b) clearly indicate in your graph what exactly is represented by the bars. To avoid incorrect interpretations of the graph, you can also present a written interpretation of your graph in your result section, such as ñ*As can be seen in Figure 1é* ò. That way you guide the readerôs interpretation and reduce the chances of faulty inferences.

Dealing with complex distributions

As with a NHT approach, the calculations used to estimate a confidence interval for a continuous variable are traditionally based on the assumption that the variables involved are normally distributed. The correspondence between the alpha level of the confidence interval and the values of its lower and upper limits will thus generally be inferred from a *Z* distribution (based on the normal law), a *t* distribution or another standard statistical distribution. However, as researchers quickly learn from experience, continuous variables coming from real-life data do not behave exactly like theoretical distributions and will even often breach the assumption of normality. In these situations, the confidence intervals (and their associated NHT tests) can be biased because they were based on incorrect distributional assumptions.

The first step to deal with potential problems of non-normality is to evaluate if some of the tested variables have problematic distributions. There are many ways to verify for breaches of normality and only the basic ones will be covered here but researchers have to know that it is an important issue every time statistical inferences are done. Most statistical programs offer indicators of skewness and kurtosis when asked for descriptive statistics. Traditionally, an index of skewness or kurtosis that is more than twice its standard error is considered to represent a statistically significant breach of normality (two standard errors being an approximation of an alpha of 0.05). A visual inspection of each distribution is also highly suggested to have a clear picture of the observed distributions. If some variables are found to be too far from normality, a solution

has to be found to insure the validity of the statistical inferences.

The basic solutions used to deal with complex distributions will be presented in the following paragraphs.

Data transformation. Data transformation is a common way of dealing with problems of normality. It basically implies applying a mathematical transformation to the data that keeps the original order between the values but stretches or compresses the distribution to make it more normal. A classical example of such a transformation is the construction of confidence intervals for correlations based on the Fisherôs *Z* transformation (for details, see Beaulieu-Prévost, 2007). Common transformations are the square, the radical, the natural log and the exponential function of the original value. However, any transformation that keeps the order and brings the distribution close enough to normality requirements is adequate. A constant might also need to be added to each value before the transformation to avoid transforming negative values. Confidence intervals and statistical tests are then computed with these transformed data. To transform the limits and the point estimate of the resulting confidence interval into meaningful units, each value simply has to be transformed back using the inverse transformation (e.g., using a radical if the original value was squared) and presented with the original scale. The two limits will not be equally close to the central value if the original distribution was skewed. However, if the confidence interval of a difference is calculated with transformed data, a simple inverse transformation is not adequate because difference scores do not behave exactly like scores. In that case, one way to transform the confidence interval into original units is to calculate the inferential confidence intervals of the difference (see Tryon, 2001), to transform the values of the inferential intervals into original units by using the inverse transformation (as above) and to recalculate the confidence interval of the difference from these inferential confidence intervals. The meta-analysis cited in the section about the graphic representation of confidence intervals (Beaulieu-Prévost & Zadra, 2007) presents a detailed example of such calculations for scores of difference between correlations. A calculator using this approach to build meaningful confidence intervals for differences between correlations is also available via the journalôs web site (Supplementary Notes).

Some distributions cannot be normalized. For example, frequencies or counts of rare events (e.g., number of nightmares per week or number of suicide attempts in the population) have a very high proportion of zeros and will stay skewed independently of the transformation. In that case, a more radical transformation can be used, at a certain cost. Indeed, such variable can be transformed into a proportion by sacrificing a part of the information and reducing it to a binary variable. The variable can then be treated like any proportion and problems of normality are avoided. However, the resulting confidence interval cannot be transformed back into original units and this transformation often reduces the statistical power of the statistical procedure because it reduces the variance of the data.

Robust estimators. A second solution to deal with problems of normality (and other departures from distributional assumptions) is to use statistical procedures that are adapted to the situation. Some complex procedures like linear regressions can now be estimated with *robust* standard errors. This so-called robust estimator for regressions (also called sandwich estimator) was developed to allow a certain departure from the basic assumptions in the data without biasing the inferences and is implemented in some statistical software (e.g., STATA, SAS). Thus, confidence intervals can be built with these robust standard errors if necessary. Similar robust estimators, based on a method developed by Satorra and Bentler (1994), also exist for structural equation modeling (SEM). They are available in Mplus (MLM and MLR), Lisrel (ML Robust), EQS (ML Robust) and AMOS (Robust ML). Since robust procedures continue to be improved and adapted to new situations, readers interested in using them are advised to at-least read the user guide of their software or a similar documentation to verify if the robust option offered is adequate with their type of data.

Bootstrapping. Bootstrapping is another sophisticated way of constructing robust confidence intervals. The basic bootstrapping method is a computer-intensive procedure in which an impressive number of artificial samples (sometimes more than 5000) are constructed from the original dataset by random sampling with replacement. The procedure subsequently produces a distribution of the means of the bootstrap samples. A robust confidence interval can then be constructed by identifying the percentiles in the distribution of bootstrap means. For example, a 95% CI can be derived by using the 2.5 and the 97.5 percentiles. Traditional bootstrapping (also called percentile bootstrapping) deals well with problems of kurtosis but is less efficient for problems of skewness (Efron, 1982). However, newer bootstrap procedures such as the bias-corrected accelerated bootstrap (Davidson & Hinkley, 1997) can now also deal with problems of skewness.

Noncentrality interval estimation. Noncentrality interval estimation (Steiger & Fouladi, 1997) is another alternative for constructing confidence intervals for statistics with complex distributions. The traditional (i.e., central) distributions of *t*, *F*, and *X2* can be considered special cases of the noncentral distributions of the same test statistics. These noncentral distributions are similar to their

central counterpart but include an additional noncentrality parameter that indicates the degree of departure from centrality of the distribution. However, this type of procedure is computer-intensive and still not implemented in most statistical software, which makes it a rarely used solution for now (see Kline, 2004).

Keeping the status quo. When researchers have good reasons not to use the preceding solutions, they sometimes calculate the confidence intervals as if the distributions of the variables involved were normal. Reason occasionally given to act that ways are that the departure from normality is not thought to affect the results of the statistical inference or that the results are easier to interpret if the original scales are used. This option might seem quite simple but it should be considered only as a last resort because it denies the problem (or at least its consequences) instead of offering a solution. If a researcher chooses to do so, the problems of non-normality should still be explicitly assessed and discussed and the reasons for not treating these problems should also be well argued and stated explicitly. This will allow the readers to do their own assessment of the adequacy of the researcherôs decision. A compromise between the status quo and the other solutions is to (a) do the analyses both with the original variables and with an alternate solution, (b) present the results for the analyses with the original variables, (c) verify if the results and conclusions stemming from the two analyses are equivalent, (d) mention in the article whether the two sets of analyses produced equivalent results or not without showing the details of the second set of analyses and (e) discuss the implications if the two sets of analyses produce different results. This way, the results of the analyses will satisfy both those who considered that the problems of non-normality had to be treated and those who thought that a standard analysis was adequate. An example of such a compromise using both original and transformed data in a linear regression can be seen in Beaulieu-Prévost and Zadra (2005).

The case of national surveys and complex sampling procedures

In most national surveys, the sampling procedures are complex and semi-random. Thus, confidence intervals directly calculated from them tend to be biased. For such surveys, a statistical procedure, called weighting, is used to correct the distortions between the sample and the target population as best as possible. In a nutshell, weighting procedures give a different weight to each individual in the sample to represent how many individuals in the population are represented by this individual. These population weights are based on the specific sampling method used and on what is already known of the population because of previous national censuses and surveys. Thus, the point estimates of weighted confidence intervals from national surveys are generally considered representative of their target population.

With longitudinal surveys, a weighting variable is available for each cycle to compensate for attrition and consequently to insure that each cycle stays representative of the population originally sampled. In these cases, researchers should be careful to use the appropriate weight when doing longitudinal analyses. A general guideline when choosing the weighting variable is to choose the weighting which is the most representative of the sample analyzed. For example, in the case of a longitudinal analysis in which cases are selected for analysis only when they were present in all the cycles, the population weights for the last cycle (or final weight) are generally the most representative. When funnel weights (i.e., weights for the sample of cases that were presents at every cycle) are available for the survey, these weights are even more representative. However, in a survival analysis in which all the cases present at cycle 1 are included at the beginning, the population weights for the first cycle are generally more appropriate.

Bootstrap weights. To estimate the standard error of the confidence intervals for these weighted surveys, a special procedure called bootstrap weights has to be used. The basic problem is that even when population weights are used to calculate confidence intervals (or the statistical significance of the estimates), the calculated standard error tends to have a systematic downward bias, i.e., it is generally smaller than it should be. This phenomenon is called the *design effect*. Furthermore, there is actually no known formula that can adequately estimate standard errors with these complex sampling designs. Bootstrap weights offer an empirical solution to this problem by estimating the error of the estimate from the variance in the estimates produced from a series of bootstrap samples (e.g., 500 bootstrap samples) taken from the original sample. Although the procedure used to create bootstrap weights is beyond the scope of this article, it is rarely necessary to create bootstrap weights because they are generally pre-calculated and available for national surveys requiring them.

The specific syntax used to produce a weighted confidence interval with bootstrap weights depends both on the type of analysis done and on the statistical software used. But even though the specific details of the syntax vary, the general procedure is generally as follow: (a) The same syntax is generally used as for an equivalent unweighted analysis except that it must be mentioned that a population weight and bootstrap weights will be used and (b) the population weight and the series of bootstrap weights have to be identified. Bootstrap weights are not implemented in every statistical software and are not

available for every type of statistical procedure. For example, STATA can handle bootstrap weights for many types of analyses, especially with the *svyset* (survey set) protocol, which enables it to specify the characteristics of a survey and apply the required procedures (e.g., bootstrap weights) to every subsequent analysis done with that survey. SAS can also handle bootstrap weights for many statistical procedures. However, bootstrap weights are still not implemented for some types of analyses such as structural equation modeling. In these cases, population weights can still be used to calculate the point estimates but an alternative approach has to be taken to calculate the standard error of the estimates.

Alternatives to bootstrap weights. Since bootstrap weights cannot always be implemented to take the design effect of a survey into account, the documentation provided with national surveys occasionally comes with estimations of the average size of the design effect for that survey that can be used as the best alternative to bootstrap weights. These estimations are often produced for different sub-populations (e.g., different regions) and can be used as a general estimation of the amount that should be multiplied to the calculated variation of error of the estimate to approximate the ñrealò variation of error for the population. The exact calculations needed to use these approximations of the design effect will not be detailed because they can be specific to the survey. However, researchers should know that when approximate values are provided for the design effect, the procedure required to calculate standard errors from these values are also specified in the documentation.

If neither bootstrap weights nor approximate values for the design effect are available for a survey using a complex sampling procedure, a last resort alternative is to use a smaller alpha to approximately compensate for the design effect. Since this method is highly approximate there is no specific formula available to decide which value to use and the size of the reduction of the alpha is simply based on a judgment call. Researchers facing such a situation should consult published studies using the same survey to know if specific guidelines are offered. Basic suggestions can be to divide the alpha level by two (e.g., from 0.05 to 0.025) or to use the next traditional alpha level (e.g., from 0.05 to 0.01) but these choices are highly subjective and should ultimately be based on risk management: The first option produces more liberal estimates than the second but both options are blind guesses since the size of the design effect unknown in these cases.

Other uses for the interval logic: The broad family of interval statistics

Although confidence intervals are the most well known types of interval statistics, other types of intervals can be quite useful to researchers. This section proposes a brief description of some of the most useful of them. Readers should also know that, because of their theoretical and methodological similarities with confidence intervals, most of the topics discussed in this articles (e.g., graphic representation, solutions for problems of normality,é) can also be applied to these interval statistics with little or no modifications.

Empirical intervals. The simplest variants of confidence intervals are sometimes called empirical intervals. These intervals are mainly relevant for continuous variables and are simply expected to cover a proportion of the observations from the sample equivalent to a pre-established level of coverage. These intervals are calculated using the same equation as for confidence intervals except that the standard deviation of the sample is used instead of its standard error. To decide the appropriate critical value to use, the level of coverage is simply treated as a confidence level. By using a rule of thumb, it can thus be considered that, for normally distributed variables, empirical intervals with limits that are one, two or three standard deviations from the mean respectively cover approximately 68%, 95% or 99% of the sample. Another simple but extremely conservative rule called the Bienaymé-Chebyshev inequality states that the proportion of observations contained within a distance of *k* standard deviations of the sample mean is at least equal to $100(1-1/k^2)\%$. The advantage of this rule is that it can also be applied regardless of the shape of the distribution. A last note about empirical intervals is that because the sample distribution only approximates the population distribution, these intervals are intended to cover a proportion of the sample distribution but not of the population. Readers interested in estimating the latter will want to read about tolerance intervals two sections below.

Prediction intervals. These intervals represent, with a certain confidence level, the probable range in which a future observation will fall, i.e., the distribution of individual future observations. The main difference with confidence intervals is that prediction intervals estimates possible scores at the individual level while the former estimates parameters at the population level. Another characteristic is that while the width of confidence intervals becomes closer to zero as sample size increases, the width of prediction intervals becomes closer to a fixed value as the sample size increases. From a mathematical point of view, prediction intervals are quite similar to confidence intervals and their basic model is:

$$PI = \widehat{\mu} \pm t_c * \widehat{\sigma}_p \qquad (2)$$

where the prediction interval (*PI*) is constructed by adding and subtracting to the predicted mean ($\widehat{\mu}$), the product of

critical *t* value of the corresponding alpha and the estimated standard deviation of the predicted scores ($\hat{\sigma}_p$).

Since prediction intervals take into account both the measurement error of the population parameter (i.e., the standard error) and the random variations of individual scores around that parameter, the general model to calculate the standard deviation of the predicted scores is:

$$\hat{\sigma}_p = \sqrt{\hat{\sigma}^2 + \hat{\sigma}_f^2} \qquad (3)$$

where $\hat{\sigma}^2$ is the residual variance and $\hat{\sigma}_f^2$ is the standard error of the model. Due to the usefulness of prediction intervals in research, the details of their calculation are presented in the following paragraphs. To construct a prediction interval for a sample of scores, the formula can be reduced to:

$$PI = \overline{X} \pm t_c * S_x \sqrt{1 + 1/n} \qquad (4)$$

where $\overline{X}$ is the mean score, S_x is the standard deviation and *n* is the sample size. Prediction intervals are also frequently done to for a linear regression to assess the probable range of the scores for the outcome. In that case, the basic formula takes the following form:

$$PI = b_0 + b_1 x_i \pm t_c \sqrt{MS_{res}\left\{1 + \frac{1}{n} + \frac{\left(x_i - \overline{X}\right)^2}{SS_x}\right\}} \qquad (5)$$

where b_0 is the intercept, b_1 is the slope of *x*, x_i is the value of *x* for which the predicted outcome is assessed, MS_{res} is the mean square of the residuals and SS_x is the sum of squares of *x*. The required information can generally be found in the regressionôs output for most statistical software.

Tolerance intervals. These intervals are expected to cover a fixed proportion of the population with a stated confidence. Thus, tolerance intervals are defined both by their level of confidence (as for confidence intervals) and by a level of coverage corresponding to the proportion of the population that should be included in the interval. For example, a 95% tolerance interval for 80% of the population implies that the interval includes at least 80% of the population with a confidence level of 95%. As for prediction intervals, the width of tolerance intervals becomes closer to a fixed value (related to the desired coverage) as the sample size increases.

If the population distribution parameters are known (which is rarely the case), the confidence level is automatically 100% (i.e., there is absolutely no measurement error) and tolerance intervals can be constructed as empirical intervals. However, in the usual case in which the population parameters are estimated from a sample distribution, the calculations become more complex because they have to take into account both the level of coverage required and the level of confidence desired. A good introduction for readers interested to know more about tolerance intervals is the *e-Handbook of Statistical Methods* (NIST/SEMATECH, 2009) available on the internet.

Confidence bands. While confidence intervals represent the measurement error for the estimation of a single numerical value, confidence bands represent the measurement error for the estimation of a curve or a function. A common use for them is to graphically represent the error for a linear regression. Two types of confidence bands should be distinguished: (a) Pointwise confidence bands represent, for each point of X taken separately, the value of the confidence interval of the function at that specific point, while (b) simultaneous confidence intervals are wider and are intended to cover an area including all the parametric values of the function with a certain degree of confidence. Pointwise and simultaneous prediction bands can also be calculated for curves and their relation to confidence bands is equivalent to the relation between prediction intervals and confidence intervals. For example, equation 5 can be used to create a pointwise confidence interval for a regression by considering *x* as a variable instead of a fixed value.

Since confidence and prediction bands require complex calculations, are less frequently used and cannot be summarized by a finite set of numerical values, they are less often implemented in statistical software. However, they can occasionally be found in the options of statistical analyses such as linear regression or in the options of scatter plots or other related graphs. Users should still verify if the intervals calculated by the software are pointwise or simultaneous. When the required option is not available, researchers will have to rely on their own wit and research skills to find the specific formulae relevant to their problem.

Credible intervals: The subjective approach to confidence intervals

As mentioned in the first section of the article, confidence intervals are traditionally based on a frequentist approach to probability and are used to produce inferences about objective probabilities, i.e., the relative frequency of an event on the long run. This approach has the advantage of producing inferences about the so-called ñobjectiveò properties of a population. However, this approach also faces the problem of being relatively non-informative when one wants to speculate about the probabilities that the parameter is included or not in a specific confidence interval. From a frequentist point of view, confidence intervals make sense mainly on the long run but can

exclusively be interpreted as an abstract representation of sampling error for a specific confidence interval.

The main problem with the frequentist approach to statistical inference is that it avoids answering the main question of interest of most researchers. Indeed, researchers and decision makers are often more interested to know the probability that a specific confidence interval includes the related parameter than to measure the sampling error of their study. What they crave for is the probability from a *subjective point of view* or, more simply, a reasonable estimation of the odds of being correct if they conclude that the parameter is included in a specific confidence interval. Using that definition, probability takes place in the eye of the beholder, not in the empirical world. It is an assessment of the uncertainty of a statement based on what is known about the situation. In fact, although traditional confidence intervals are not based on a subjective approach, subjective interpretations are quite appealing and intuitive to the human mind. Indeed, traditional confidence intervals are often wrongly interpreted in a subjective way in published articles. Expressions such as ñ*The statistical significance of the tests suggests that the results were not due to chance*ò or ñ*The test shows that the difference is probably present in the population*ò are basic examples of such a faulty subjective interpretation of a frequentist confidence interval or test of statistical significance. For more details on the issue, readers are referred to the first paper of the series (Beaulieu-Prévost, 2007). These faulty interpretations of the results of statistical tests are in fact extremely common among researchers (Lecoutre, Poitevineau & Lecoutre, 2003). For example, a survey of academic psychologists showed that only 11% of them were able to adequately interpret the results of tests of statistical significance (Oakes, 1986).

Since the problem basically comes from the fact that researchers estimate objective probabilities but are generally interested by subjective probabilities, the easiest solution might simply be to provide them with the proper conceptual tools needed to adequately estimate these subjective probabilities. And a well-developed subjective approach to probability does indeed exist. This subjective approach to probability is generally called the *Bayesian approach* because it is mathematically based on Bayesô theorem. Instead of simply aiming at assessing the relative frequency of an event, this approach directly aims at assessing the confidence, or degree of belief, that one can put in an estimation and in the range of probable values for a parameter, given the evidence. In a way, using a subjective approach to probabilities is essentially trying to incorporate the subjective component of probabilities to the estimation instead of avoiding it. If you are interested in subjective probabilities, the most appropriate thing to do is probably to acknowledge it and integrate a formal approach to subjective probabilities to your statistical inferences instead of simply subjectively interpreting your frequentist confidence intervals in an intuitive way.

As mentioned, Bayesian confidence intervals can be made for which it can be reasonably assumed that there is [1-alpha] chances that the parameter is included. To differentiate them from traditional confidence intervals, these Bayesian intervals are called *credible intervals*. The basic mechanics of credible intervals is similar to that of confidence intervals. However, credible intervals take into account the fact that the degree of belief that can be put in an estimation depends on both the results of the experiment itself (like for traditional confidence intervals) but also on the prior knowledge that one already has about that parameter.

To calculate a credible interval, one has first to quantify the prior knowledge or expectations about the parameter, called the prior probability. When estimating a continuous parameter, this knowledge is formalized as a prior distribution with a point estimate representing the most probable value of the parameter, given the actual evidence, and a standard error representing the precision provided by the evidence, much like the distributions on which traditional confidence intervals are based. More precisely, the precision of the estimation is measured by the inverse of the conditional variance, i.e., by the inverse of the squared standard error. As new results (e.g., studies) provide estimations of the parameter, these estimates are weighted by their precision and combined with the prior distribution to create an updated distribution (called posterior distribution) representing the new state of knowledge. This process can continually be updated with additional results coming from subsequent studies to continually represent the most up to date state of knowledge. As the number and the precision of the results included in the analysis increases, the impact of the prior distribution on the posterior distribution decreases. In other words, the more you accumulate new information about a parameter, the less your prior expectations have an impact on your actual estimation of that parameter.

The posterior distribution resulting from the process can be understood as the distribution of the probable values of the parameter, according to the present state of knowledge, and it can directly be used to calculate a Bayesian credible interval, using its point estimate and standard error as for traditional confidence intervals. This process is extremely similar to the basic meta-analytic procedure, in which the resulting confidence interval is calculated by combining the results of all the previous relevant studies and by weighting each result by its precision (i.e., by the inverse of its squared standard error). The main difference between a traditional meta-analytic

confidence interval and a basic Bayesian credible interval in terms of calculations is that the latter takes into account prior knowledge about the parameter while the former does not.

The preceding explanations only provides a summary of the basic elements of Bayesian estimations and many aspects, such as the adequate way to define the prior distribution, could certainly be covered in more details. Due to space constraints, these aspects will not be covered in the present article and interested readers are referred to Kline (2004) for an introduction to the basic concepts of Bayesian estimations and a comparison with the meta-analytical approach.

Bridging the gap with confidence intervals. There is a final aspect of credible intervals that is highly relevant to the subject treated in this article and it is worth discussing. Although credible intervals and confidence intervals can produce quite different intervals (especially depending on the prior distribution on which the credible interval is based), there is one situation in which a Bayesian credible interval coincides exactly with its frequentist counterpart. The interest of that situation is that it can be used to understand more clearly in which context and to what extent a subjective interpretation can be given to traditional confidence intervals.

A credible interval will coincide exactly with a traditional confidence interval when the credible interval is based on an agnostic (i.e., a noninformative) prior and takes into account the results of exclusively one study. An agnostic prior is a judgment that one has no useful prior knowledge about the parameterôs probable value and it is represented by a prior distribution with a precision asymptotically close to zero. In that situation, the prior distribution loses its impact on the posterior distribution and the latter becomes completely defined by the results of the study. It is rare that a researcher has absolutely no prior knowledge about the probable value of a parameter. However, this situation can help us understand the value of the results represented by a traditional confidence interval. It can thus be said that when only the experimentôs data are taken into account to estimate a parameter (i.e., when an agnostic prior is postulated), a traditional confidence interval represents an interval for which it is reasonable to assume that there is [1-alpha] chances that the parameter is included. By extension, the distribution related to the confidence interval can be understood, from a Bayesian point of view, as the distribution of the probable values of the parameter according to an agnostic prior. This represents the extent to which a subjective interpretation can be done of a traditional confidence interval. This type of interpretation can also be extended to the other types of interval statistics presented previously by using the same principles.

CONCLUSION

This paper is the third in a series written to facilitate the transition from an approach based on significance testing to one based on confidence intervals. As could be seen, the parallels between significance testing and confidence intervals are numerous and a transition from one to the other can be done smoothly. In addition, an approach based on confidence intervals reveals new ways to answer research questions and improves the usefulness of the results.

The specific purpose of this paper was to provide guidelines, advices and tricks for researchers in the social sciences who want to (a) develop and improve their understanding of confidence intervals and (b) be able to use an approach based on confidence intervals with real-life data. However, this paper should not be considered as the final reference but as a starting point for those deciding to adopt an approach to statistical inference based on confidence intervals. Every statistical inference is an estimation process and estimation methods are continually improving. As such, some of the methods presented herein might be replaced by more efficient alternatives in the near future. Improving oneôs skills in statistical inference is thus best viewed as a continual process and an integral part of a researcherôs activities. It is hoped that this paper can help facilitate this learning process.

ACKNOWLEDGEMENTS

I am greatful to Jean-Sébastien Fallu and Denis Cousineau for their helpful comments and suggestions concerning this paper.

REFERENCES

Beaulieu-Prévost, D. (2006). From tests of statistical significance to confidence intervals, range hypotheses and substantial effects. *Tutorials in Quantitative Methods for Psychology, 2,* 11-19.

Beaulieu-Prévost, D. (2007). Statistical decision and falsification in science: Going beyond the null hypothesis. In B. Hardy-Vallée (Ed.). *Cognitive decision-making: Empirical and foundational issues*. Cambridge : Cambridge Scholar Publishing. [A previous version of the paper also appears at http://eradec.teluq.uquebec.ca/IMG/pdf/CIC_2005_05.pdf, page visited october 15th, 2009]

Beaulieu-Prévost, D. *Professional web site: Statistical resources*, [Online].

http://www.memoryproject.info/stat.html (Page visited october 15th, 2009).

Beaulieu-Prévost, D., and Zadra, A. (2005). Dream Recall Frequency and Attitude Towards Dreams : A reinterpretation of the relation? *Personality and Individual Differences, 38*, 919-927.

Beaulieu-Prévost, D., and Zadra, A. (2007). Absorption, thinness of boundaries and attitude towards dreams as correlates of dream recall frequency: Two decades of research seen through a meta-analysis. *Journal of Sleep Research, 16*, 51-59.

Belia, S., Fidler, F., Williams, J., and Cumming, G. (2005). Researchers misunderstand confidence intervals and standard bars. *Psychological Methods, 10*, 389-396.

Davidson, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, PA: Society of Industrial and Applied Mathematics.

Hopkins, Will G. *New view of statistics: Confidence limits*, [Online]. http://www.sportsci.org/resource/stats/generalize.html (Page visited october 15th, 2009).

Kendall, M. G. (1949). On the reconciliation of theories of probability. *Biometrika, 36*, 101-116.

Kline, R. B. (2004). *Beyond significance testing*. Washington: American Psychological Association.

Lecoutre, M.-P., Poitevineau, P., and Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Test. *International Journal of Psychology 38*, 37-45.

NIST/SEMATECH. *e-Handbook of Statistical Methods*, [Online]. http://www.itl.nist.gov/div898/handbook/ (Page visited october 15th, 2009).

Oakes, M. (1986). *Statistical inference*. New York: Wiley.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russel Sage Foundation.

Satorra, A., and Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C. C. Clogg (Eds), Latent variables analysis: Applications for developmental research (p 399-419). Thousand Oaks, CA: Sage.

Steiger, J. H., and Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik and J.H. Steiger (Eds), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

Tryon, W. W., and Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryonôs (2001) reduction factor. *Psychological Methods, 13*, 272-277.