Five Minute Science Lesson: Statistical Confidence Intervals and Point Estimates

Raymond Nelson

Confidence intervals are used in inferential statistics to describe an interval estimate, in contrast to a point estimate, for an unknown population parameter. The purpose of any scientific test or experiment is to measure or quantify some phenomena that cannot be subject to either perfect deterministic observation or direct physical/linear measurement . Perfect deterministic observation would be immune from any influence from human behavior or random variation, while direct physical/linear measurement requires both a physical substance and a physical unit of measurement. Tests achieve the goal of quantifying amorphous phenomena through the use of statistical measurement and the use of proxy data signals that are correlated to the phenomena of interest though they are not themselves the phenomena of interest. Very often a single proxy signal will provide only an insufficient level of precision, instead multiple proxy signals are often combined to increase or optimize test performance.

Adding additional data can potentially increase the effectiveness of a testing model, as long as the additional data is itself sufficiently correlated with the phenomena of interest, is not redundant with other data, and can be combined in an optimal structural model with the other proxy signals. The effectiveness of a scientific test depends on the structural combination of the proxy data, the representativeness of the normative or reference sample, the structural or statistical combination of

the different signal data. Ultimately, because test data are not themselves the phenomena of interest, all test results are statistical estimations of the phenomena of interest.

All scientific test results are probabilistic approximations, for which there is no realistic expectation of 100% accuracy. Scientific tests are not expected to be infallible. Instead, scientific tests are expected to quantify the level of confidence or margin of uncertainty that can be realistically and reproducibly assigned to a conclusion.

Reproducibility is among the most important objectives for any probabilistic test result and scientific conclusion. However, because scientific tests are not deterministic, and are subject to potential influence from random variation and from human behavior, some normal variation is expected for all scientific test results. Reproducibility goals therefor involve the calculation and estimation of the expected range of normal variability, within which a result or conclusion can be reasonably expected to be observed for a desired portion of experiments or trials. That range of normally expected variability is referred to as the confidence interval, often expressed as the 95% confidence interval though any desired percentile range can be used.

Population parameters and sample statistics

A parameter is a number that describes a measurable characteristic of a population. A population consists of all members of a group. In contrast, a statistic is a number that describes a measurable characteristic of a sample group. A sample is a subset of a population. It is often not realistic to attempt to study an entire population, and many scientific studies involve sample groups instead. Knowledge gained from a sample, and insights about the sample group characteristics, can be representative or informative of the population if the sample is selected randomly (i.e., wherein every member of the population as an equal chance of being included in the sample). Of course, knowledge and insights gained from samples based on non-random sampling methods can be expected to provide biased or misleading information about the population. The tradition of frequentist inference is the practice of gaining knowledge from a sample group based on the proportions or frequencies of observable characteristics

Point estimation and interval estimation

Interval estimation, first described by Neyman (1937), involves the calculation of an interval of possible or likely values of an unknown population parameter based on sampling data. In contrast, point estimation is the calcu-

lation of a single value, a statistic, that is used as a best guess regarding an unknown population parameter. For example, a sample mean can be used an as a point estimate for an unknown population mean.

Confidence intervals can also be calculated for sampling proportions of correct classification. For example, the effectiveness of credibility assessment tests such as the polygraph test can be described using confidence intervals. Table 1 shows the criterion accuracy for event specific polygraph techniques that satisfy the American Polygraph Association requirements for evidentiary testing, as reported in the meta-analytic survey of validated polygraph techniques (American Polygraph Association, 2011), including the means, standard errors, and 95% confidence intervals.

Effective development and use of scientific tests can be thought of as an optimization problem, which can involve either the maximization of a goal such as the correct classification of deception and truth-telling, or the minimi-

Table 1. Criterion accuracy of evidentiary polygraph techniques (APA, 2011) including proportional means (standard errors) and {95% confidence intervals}.

Unweighted Average Accuracy	.921 (.028) {.865 to .977}
Unweighted Average Inconclusives	.088 (.030) {.029 to .147}
Sensitivity	.699 (.053) {.596 to .802}
Specificity	.717 (.055) {.610 to .824}
FN Errors	.063 (.035) {.001 to .130}
FP Errors	.059 (.037) {.001 to .131}
D Inc	.091 (.041) {.010 to .172}
T Inc	.086 (.044) {.001 to .173}
PPV	.927 (.036) {.856 to .999}
NPV	.915 (.043) {.830 to .999}
D Correct	.917 (.039) {.841 to .993}
T Correct	.925 (.040) {.846 to .999}

zation of costs associated with errors such as false-positive or false-negative classifications, or both. With optimization problems it is often the worst case scenario that is of greatest usefulness. This will mean paying attention to the lower limit of the confidence intervals for correct classifications, and the upper limit of the confidence intervals for test errors. Confidence intervals, because they express a range of likely values for an unknown population parameter, are inherently more reproducible and more generalizable than point estimates.

Whereas the means of sampling proportions can be taken as a point estimates, confidence intervals can provide information that is much more generalizable and therefore more useful. This is because sampling statistics cannot perfectly represent an unknown population parameter. In the strictest sense, point estimates based on sampling data are always incorrect and not generalizable. Sampling statistics are not expected to precisely or perfectly indicate the unknown population parameters, and some variability is expected when observing or comparing results from different random samples. In other words, different random samples drawn from the same population can be expected to have different values for the sampling mean.

Standard errors

The measurement of variability for sampling statistics is referred to as the standard error (SE) of a sampling statistic, or the standard error of the mean estimate (SEM). Standard errors for sampling statistics are similar to standard deviations for the values of a population. Whereas a standard deviation describes the variability of the members of a population, a standard error describes the variability of a sampling statistic. With many samples we will can construct a distribution of sampling statistics.

When dealing with a sampling proportion ($\hat{p}$, pronounced "p-hat") as a point estimate for an unknown population proportion (p), the standard deviation is np(1-p), where n = the number of samples and p = the proportion of the sample that has a certain characteristic (e.g., correct or incorrect classification). An interesting and useful phenomena is that the sampling distribution of a sample proportion will be approximately normally distributed when np(1-p) > 10. Similarly, when the number of samples is greater than 30 the distribution of sampling statistics will be approximately normally distributed regardless of the shape of the underlying data distribution.

Another interesting useful phenomenon, referred to as the law of large numbers (LLN), is that the mean of a sampling statistic (e.g. the sampling means or any other statistic) will converge towards the population mean

upon numerous repeated sampling experiments. In other words, the sampling distribution of the sample means is the probability distribution of all possible sampling means, and has a mean equal to the population mean μ (pronounced "mu") with a standard error (i.e., the SEM) equal to σ / n ^ (.5), where σ (pronounced "sigma") is the standard deviation of the sampling means. A number of scientific tests rely on this phenomenon to compare a test result to an estimate of an unknown population parameter.

The standard error (SE) of a statistic can also be used to calculate statistical a confidence interval, often in the form of a 90%, 95%, or 99% range. Confidence intervals are useful because they remind us to avoid incorrect and simplistic expectations that a sampling statistic is a perfect representation of the population statistics. Use of confidence intervals can also deter us from another simplistic error of rejecting scientific and statistical results because they are inherently probabilistic and therefor imperfect.

What do confidence intervals tell us?

Confidence intervals allow us to understand the probability relationship between an observed sample point estimate and an unknown population parameter. A 95% confidence interval shows the range of possible population parameters that do not differ significantly from the sample statistic at the .05 level. Similarly, a 99% confidence interval shows the range of possible parameters that do not differ significantly from the sample statistic at the .01 level, while a 90% confidence interval shows the range of possible values for which the difference between an unknown population parameter and the observed point estimate are not statistically significant at the .10 level.

Using the 95% confidence interval as an example, if the true value of the unknown population parameter is outside a sample confidence interval then it can be said that an event has occurred for which probability of occurrence due to random chance is less than or equal to 5%. This same kind of statement can be made about the 90% and 99% confidence intervals.

Confidence intervals can also be described in term of repeated sampling experiments in this manner: if the experimental procedure were repeated with numerous different samples then 95% of the confidence intervals for those different samples would include the unknown population parameter. The notion of confidence intervals can also be applied to a single future experiment, in which case the confidence interval is simply an expression of the probability that the future cal-

culation of the confidence interval will cover the unknown population parameter.

What confidence intervals cannot mean

Confidence intervals are potentially misunderstood, and it can be useful to clarify what they cannot be taken to mean. One potential misunderstanding of confidence intervals would attempt to interpret them as a probability estimate for a sample statistic of a repeated experiment.

Confidence intervals cannot be interpreted as a probability measurement of the unknown population parameter, or as an estimate of the probability that the unknown population parameter exists within a specified interval. Attempts at this type of interpretation are fundamentally incorrect because the unknown population parameter is a constant value for which probability statements are not warranted. Whether a calculated confidence interval does or does not include the population parameter is not a matter of random chance. Instead the confidence interval itself is the random variable of interest.

The purpose of a confidence interval is to describe the range of plausible values for an unknown population parameter based on the sample data. In this case, the 95% probability describes the reliability or the repeatability of the estimation procedure. In other words, the confidence interval describes the probability that the data have occurred due to random chance if the actual population parameter is not within the confidence interval.

Calculation of confidence intervals

Calculation of a confidence interval is often accomplished using critical values of z for the standard normal distribution. Table 2 shows z-values for commonly used confidence intervals. The lower limit and upper limit of the desired confidence interval are calculated using the sampling mean $\bar{x}$ (pronounced "x-bar") using two equations: lower limit = $\bar{x} - z * SE$ and upper limit = $\bar{x} + z * SE$. Often it is the lower limit or worst case scenario that is the most useful and informative value for

Table 2. Values of z for commonly used confidence intervals.	
99%	z = 2.576
98%	z = 2.326
95%	z = 1.959
90%	z = 1.645

risk evaluation and risk management decisions.

Conclusion

This paper has described the basic concept of confidence intervals in inferential statistics, including basic conceptual vocabulary and rudimentary calculations. In the realm of Bayesian statistics, the allegorical concept is referred to as a credible interval, for which a more complete description will have to be the topic of another paper.

Scientific test results are ultimately probability statements. Professional polygraph examiners who wish to attain or claim a level of expertise beyond the mere execution of procedural rules for test data acquisition and procedural rules for test data analysis will want to develop their ability to understand and converse on these topics. Many field practitioners can develop subject matter and interviewing expertise that can provide great practical and informational value without expertise and familiarity with the underlying test theory and statistical formulae. If the polygraph test result is not regarded with any usefulness, then the role of the polygraph examiner is simply that of an interviewer or interrogator. In that case there will be no expectation that polygraph field practitioners develop or possess any expertise in understanding the meaning, nuances and limitations of scientific and probabilistic classification and decision models.

If polygraph test results of themselves are to ever be regarded as a useful form of scientific and probabilistic information then it may become necessary for polygraph experts to become more familiar with the conceptual vocabulary and use of probabilistic models, including the correct use and understanding of statistical confidence intervals. Of course, some field practitioners may have little or no interest in developing their expertise in areas of statistical abstractions and probabilistic thinking, and may prefer instead to emphasize the role of subject matter expert in various topical areas of interviewing and interrogation, leaving the details of science and probabilistic classification to persons with expertise in those areas. Field practitioners will never be expected to execute the mathematical calculations themselves. For those who are interested in developing a level of professional expertise in the science of polygraph, lie detection, and credibility assessment it is hoped that this document may be a useful resource.

References

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal So-

ciety A. (236) 333–380.

Rosner, B. (2010). Fundamentals of Biostatistics. Brooks/Cole: Boston, MA.

Smithson, M. (2003). Confidence intervals. Quantitative applications in the Social Sciences Series, No. 140. SAGE Publications.

Mayo, D. G. (1981). In defence of the Neyman-Pearson theory of confidence intervals. Philosophy of Science, 48(2), 269–280.