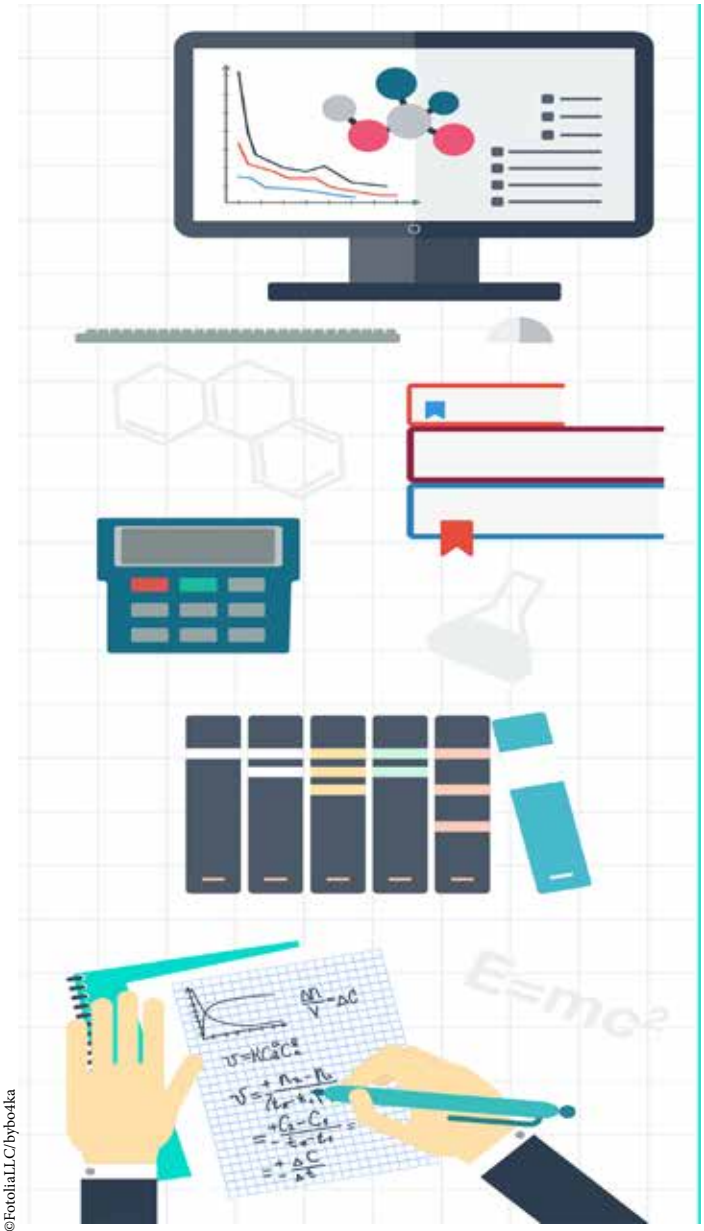
Polygraph Testing as a Single-Subject Scientific Experiment

By Raymond Nelson



©FotoliaLLC/byb04ka

Scientific testing is a process of both classification and inference. Scientific tests can be thought of as a form of single-subject research or experiment where the test data, test scores and probabilistic test result are the evidence to support a professional opinion or conclusion to classify the test as belonging to one of multiple possible outcome categories. Professional opinions and scientific conclusions are, by definition, statements that are supported by evidence.

Evidence in scientific testing and research is commonly expressed in the form of statistical or probabilistic information based on a replicable numerical and quantitative analysis of the test data. Conclusions without objective and quantifiable supporting evidence are mere personal opinions, including when offered by a professional. Perfect deterministic solutions - immune to human behavior and immune to the effects of random chance - often do not exist for many interesting and important real-world phenomena (e.g., personality, intellectual functioning, interpersonal rapport, and the discrimination of deception and truth-telling). A statistically significant test result is supportive of a categorical or professional conclusion, and is analogous to a statistically significant result from a scientific experiment.

Discussion

Before proceeding further it will be useful to remember that scientific conclusions are

always relative to some alternative. Scientific conclusions begin with a suggestions or hypothesis. A hypothesis is a form of conjecture, speculation or explanation for some phenomena or observation. Scientific hypothesis are, in practicality, questions that must be either supported or refuted by evidence. Personal opinion, including professional opinions that are offered without replicable quantitative analysis, can be thought of as a form of un-researched hypothesis.

Null hypothesis significance testing.

The tradition of hypothesis testing – formally called null-hypothesis significance testing – involves the comparison of the strength of evidence that a conclusion or hypothesis is correct against the strength of evidence that the conclusion or hypothesis is erroneous or that the observed data could have occurred due to random or uncontrolled causes. The formal name for the hypothesis is alternative hypothesis. The purpose of an alternative hypothesis is to attempt to provide a systematic explanation for some observed data or phenomena (i.e., how the universe works). In contrast to this is the null hypothesis which says that the observed data or phenomena may have been the result of random or uncontrolled factors. As a corollary to the notion that observed data or phenomena are due to random or non-systematic causes, the alternative hypothesis cannot be supported as a systematic explanation of the observed data or phenomena.

Stated formally, a hypothesis is supported when the null hypothesis can be rejected as not likely to have caused the observed data. Hypothesis that are supported by evidence can be retained for further consideration as potentially useful systematic conclusions or explanations of the observed data or phenomena (i.e., how the universe works). A hypothesis is not supported if the evidence is not sufficient to reject the null hypothesis. The basic concepts of hypothesis testing can be applied to a single test observation in a manner similar to the way they are applied to a research sample. The classical, frequentist, standard of inference is that a conclusion or hypothesis is supported when the probability is sufficiently low that the observed data occurred due to causes other than the systematic conclusion, having decided a maximum tolerance for error prior to testing.

The principles of science and testing obligate us to acknowledge that there is no such thing as a deterministic or perfect test for which there will be no potential for error. The goal of a scientific test or experiment is to quantify the margin of error surrounding a possible conclusion. Hypothesis tests, regardless of whether a single case or a group of cases, begin with an explicit declaration of our tolerance for error. Neglecting this would encourage misguided expectations for perfection.

Alpha, p-value, and statistical significance.

The alpha boundary (a) is the common statistical term used to describe the level tolerance for error. Alpha is often set at a = .05, which refers to an error tolerance of 5%. Some circumstances may warrant a more restrictive alpha boundary; in these cases a = .01 can be used to constrain the observed error rate to a 1% level. Of course, there will be a decrease in the classification rate for

cases that are statistically significant at this level. Other circumstances may benefit from a reduction of results that are not statistically significant or inconclusive, in which case alpha can be set to a = .10 for which the observed error rate will can be expected to be constrained to less than 10% while increasing the number of cases that are classified as statistically significant.

Alpha boundaries are always set a priori (i.e., before initiating a test or experiment). Otherwise there is an increased risk for both test errors and manipulated test outcomes. In the polygraph context, alpha is analogous to the cut-score, for which there is a cut-score for deceptive classifications along with a different cut-score for truthful classifications. Numerical cut-scores for scientific polygraph tests are calculated as a function of the desired alpha boundary and the reference model or reference data that describe the expected distributions of scores of deceptive and truthful persons. In practice, alpha boundaries and polygraph cut-scores will most often be established as a matter of agency policy.

The p-value (probability value) is the common statistical term used to describe the calculated estimate of the probability that the observed data are the result of random or uncontrolled causes. Formally, a p-value is the proportion of cases under the null hypothesis that are expected to produce a result that is similar or more extreme than the observed result. As a practical matter, results from scientific testing and research are also described categorically as to whether results are significant (i.e., statistically significant) or non-significant (i.e., not statistically significant). A result is said to be significant at the alpha level (e.g., statistically significant at the .05 level) when the p-value (probability of error) is less than alpha (tolerance for error).

Classification: categorical results.
When probabilistic results are statistically significant we are permitted to categorically classify test results as positive or indicative of the presence of the condition or issue that is being tested. When the result is not statistically significant the result we are permitted to classify the result as negative or not indicative of the presence of the condition or issue being tested. These terms are considered objective and neutral abstractions compared to more emotionally laden terms such as pass and fail.

It is through this process of statistical inference (i.e., that observed data are not likely to be due to chance) that we can make professional conclusions based on probabilistic evidence while remaining accountable for the fact that results from scientific tests and scientific experiments are conclusions about amorphous phenomena for which neither simple and perfect deterministic observation nor direct physical/linear measurement can be achieved. Conceptually, the p-value is a statistical term that is analogous to the polygraph test score, and is calculated from the numerical test score and the statistical reference model that can be derived either empirically (i.e., either sampling or population data) or theoretically (i.e., through facts based on mathematical proof and formal logic).

Categorical results of polygraph tests have often involved descriptive terms specific to the polygraph context. A categorical conclu-

sion for positive polygraph results might be that there is deception indicated when the data differ at a statistically significant level from the statistical reference distribution for truthful persons. In polygraph screening contexts we often state more generally that there are significant reactions. Similarly, a categorical conclusion might be that there is no deception indicated when the polygraph data differ at a statistically significant level from the reference model or reference data for deceptive persons. In polygraph screening context we often state that there are no significant reactions for the sake of both simplifying and clarifying communication with others. The scientific implications are the same regardless of which labelling scheme is selected to describe the categorical test result: there is no perfect test, and all test results are fundamentally probabilistic. We use the language of statistical significance and scientific research to communicate the probabilistic and categorical polygraph test result. Categorical results of polygraph exams are supported by probabilistic calculations based on the test score and the p-value.

An examiner using a comparison question polygraph test format will actually evaluate one of two hypothesis, deception or truth-telling, depending on whether greater changes in physiological activity are loaded onto relevant or comparison stimuli. Numerical scores of positive test results of a comparison question polygraph, indicative of deception, will reject the null hypothesis that the data are not significantly different the reference model for truth-telling. Test scores that are said to be negative, indicative of truth-telling, will reject the null hypothesis that the data are not significantly different from the reference model for deception. Formally, both deceptive and truthful conclusions are statistically significant (i.e., positive), though in practice classifications of truth-telling are described pragmatically as negative or indicating an absence of deception the regarding problem or issue that is being tested.

In the context of the comparison question polygraph test, the alternative hypothesis that an examinee has been deceptive is evaluated against the null hypothesis that the examinee's data do not differ significantly from the reference model or reference data for truthful persons. This is especially obvious when considering the polygraph screening context for which there is no known incident or allegation regarding the behavioral issues under investigation. Similarly, the alternative hypothesis that an examinee has been truthful will be evaluated against the null hypothesis that the observed test data do not differ significantly from the reference data or reference model for deceptive persons. A classical application of the principles of hypothesis testing to the comparison question polygraph test might evaluate the two-tailed hypothesis using a theoretical reference distribution calculated from mathematical and logical facts. In this application, the null hypothesis might be that the observed pattern of responses does not differ from the theoretical reference distribution. The alternative hypothesis might state that the observed pattern of physiological changes are systematically, non-randomly, loaded differentially onto relevant or comparison stimuli in a manner that can be shown to discriminate deception and truth-telling.

Conclusion

The principles of scientific decision making have been applied to the polygraph context for many years, ever since the first time an examiner offered a conclusion or opinion that a result was significant. Today, however, we are mindful of the fact that significance cannot be inferred visually or intuitively. Statements or conclusions about significance cannot be offered without the mathematical calculation of the level of statistical significance. Significance cannot be determined via a single presentation of a test stimulus question, and cannot be determined without first obtaining all test data required for its calculation. Discussions of significance in the scientific context belongs to the realm of quantitative and probabilistic analysis and categorical conclusion that is intimately connected with the tradition of hypothesis testing.

Categorical and probabilistic conclusions can be made for both event-specific diagnostic polygraphs, using both grand total and subtotal scores, and for multiple issue screening polygraphs, for which grand total scores are generally not used. In general, statements and conclusions of significance will become more precise when they are based on more data (i.e., more test stimuli or more iterations of the test stimuli), but will become less precise as more probabilistic conclusions are made at one time (i.e., when using subtotal scores). All computations and categorical conclusions of significance will be subject to the laws of probability. It will help the polygraph profession to advance if field practitioners can become more conversant with the basic principles of probability, scientific testing and scientific experiments. The practical goals of scientific research and scientific testing will continue to involve both classification and statistical inference for as long as tests are needed and useful. Scientific tests are not expected to be perfect. They are expected to quantify the margin of error and uncertainty such that observed error rates will be within our established tolerance for error.

Although some complexities and controversies have surrounded the practice of null hypothesis significance testing, the basic principles have remained embedded in the scientific tradition such that every student in statistics from high-school through graduate school will be required to study the practical application of these basic principles. Virtually every student in every introductory statistics course will also learn about omnibus tests (i.e., F-tests and other methods) that evaluate a number of statistical hypothesis as a single hypothesis without the need for statistical corrections to manage and account for multiplicity effects, and the use of statistical corrections for multiplicity effects such as those expected when using subtotal scores on comparison question polygraphs.

Although field practitioners will never be required to complete the actual mathematical and statistical calculations, any professional who wishes to become expert in the testing an analytic aspects of the polygraph should be expected to become conversant with the language and concepts of scientific testing. A well-developed understanding of the polygraph test as a contextual application of the principles of statistical decision making can help to more effectively convey test results in ways that satisfy and educate others about the scientific basis of

the polygraph and the potential usefulness of polygraph results in legal proceedings that are in some ways constructed around similar burden of proof concepts.

In addition to a basic understanding of concepts of frequentist inference (e.g., p-value, alpha, statistical significance, etc.), it will also be useful and important for polygraph professionals to become familiar with the concepts of Bayesian inference. Bayesian inferential methods have proven their usefulness in a variety of contexts, however different a priori assumptions can lead to different a posteriori results. In practice Bayesian inference can provide probability results that some people may find to be more intuitive than p-values and alpha boundaries. Unlike frequentist inference, Bayesian inference does not involve the use of an arbitrary alpha boundary of .05, or .01, or .10, but does involve an explicit a priori declaration of our assumption regarding the prior probability or base-rate (i.e., before testing) of each alternative conclusion. Polygraph examinations can also be thought of as a single subject Bayesian experiment, but that will have to be the subject of another publication.