Five Minute Science Lesson: Using the Statistical Test of Proportions to Unveil Hidden Disparities

Raymond Nelson

Science is a systematic process used to solve complex problems in the natural world. Science is also used to understand the human behavior. Statistics and probabilities are the mathematical language of science, and anyone who wishes to engage in, make use of, or understand the results from scientific experiments or scientific tests will face an obligation to become familiar and fluent with probabilistic discussion, probabilistic procedures, and probabilistic concepts. Scientific decisions emerge quantitatively from rigorous mathematical logic and the preponderance or consistency of information. This approach acknowledges the inherent uncertainty in many scientific conclusions, and relies on rigorous evaluation of the strength, variability, and reproducibility of the evidence that may support different possible conclusions. In this way, probabilistic information provides a foundation for informed choices in a world where certainty is rare.

Whereas an intuitive or intellectual approach to science involves the formulation of new knowledge or conclusions based on abstract thinking and previous knowledge or held assumptions, an empirical approach to science relies on observation, experimentation, and the observation and analysis of data. Empirical science begins with a hypothesis – often in the form of a question or conjecture based on either abstract thinking or inconsistencies in observed or expected data – followed by data collection and analysis.

In the paradigm of null hypothesis significance testing (NHST), an experimenter will systematically obtain evidence for an experimental condition and then evaluate the strength of the evidence for the different possible conclusions. The different possible conclusions are often described as the *null hypothesis* (H0) which states essentially that there are no real differences between the observed evidence,

such as the frequency of some event, and the *alternative hypothesis* (H1 or HA) which states that there is a real difference in the observations. The HA is an explicit statement of expectation that the observed data, when transformed into numerical quantities, will be statistically significant – that the observed data will be inconsistent with information driven by random chance. In other words, in the NHST paradigm, when the probability of obtaining the observed data due to random variation alone is sufficiently small it can then be assumed that the observed data is due to a real difference in the experimental conditions. In this case the H0 can be rejected, and the HA can be said to be *supported by evidence*[1].

Introduction to the Test of Proportions

One very simple and useful method to evaluate the probabilistic strength of evidence for different possible conclusions is a statistical *Test of Proportions* (TOP). TOPs can be calculated in different ways, including using the z-test approximation for two sample groups and via the chi-squared (χ^2) test for three sample groups. The TOPs are used to evaluate the observed data or evidence as to whether differences in the frequency of observed events are consistent with random variation. Regardless of whether there are two groups or more, there are always two possible conclusions – that the observed differences between the groups are within the expected range that can be attributed to random variation, or they are statistically significant. TOPs can be essential tools in data analysis whenever there is a need to investigate different possible conclusions about differences in observed proportions between groups or categories.

Z-test Approximation for the Test for Proportions of Two Samples

A z-test is a common statistical test that uses the standard normal distribution (Gaussian distribution or bell-curve[2]) which has a mean of 0 and a standard deviation of 1. Many naturally occurring phenomena conform to this distribution. Because the mathematical properties of this distribution are well known, for any sample of normally distributed data it is fairly easy to calculate the proportion of data values that are greater than or less

---

[1] However, a statement that an HA that is supported by evidence is not the same as stating that the HA has been proven or is supported by mathematical and logical proof. Much of the available scientific knowledge exists in the form of theories that are supported by evidence, and for which it is acknowledge that our present knowledge remains incomplete. There is nearly always more to learn about reality, the universe, and human behavior.

[2] This distribution is also sometimes referred to as the Quetelet distribution, for Adolphe Quetelet, a Belgian statistician who used it to describe human physical and social characteristics during the early half of the 19th century. It is commonly referred to as Gaussian because Carl Friedrich Gauss did extensive work with the distribution in the context of astronomical observations in the early 19th century. It is also referred to a bell-curve because the characteristic bell-shape of the distribution. Approximately 68% of values are within +/1 standard deviations of the mean, with approximately 95% of values within +/- 2 standard deviations, and approximately 99% of values within +/-3 standard deviations from the mean. And while a number of other distribution types are also highly useful, the normal distribution has been used extensively in virtually all areas of science and data analysis.

than any individual value. A very convenient aspect of this distribution is that tables of z-values have been published extensively in numerous statistics textbooks. An even more convenient aspect of this distribution is that virtually every microcomputer today (including, mobile devices, tablets, laptops, and desktop computers) includes well-formulated calculation tools with built-in functions to calculate z-values.

Approximation, in this usage, refers to the use of a convenient distribution in place of a less convenient one. In this case, the TOP data is binomial – involving the numbers of observations and possibilities. With a sufficiently sized dataset the discrete values of the binomial distribution become asymptotic with or approximate the standard normal distribution.[3] For this reason, the convenient computations of the normal distribution are often substituted for the calculation of actual binomial probabilities. Formally, the TOP involves the mathematical and statistical comparison of the difference between two sample proportions. The formulae for the TOP is shown in Figures 1, 2 and 3.

Figure 1. Z-test Approximation for the Test of Proportions.

$$z = \frac{(p_1 - P_2)}{SE}$$

Figure 2. Calculation of the pooled standard error.

$$SE = \sqrt{(p(1-p) * (1/n_1 + 1/N_2))}$$

Figure 3. Calculation of the pooled proportion.

$$p = \frac{p_1 * n_1 + p_2 * n_2}{n_1 + n_2}$$

3Asymptotic, in this usage, means "in the limit" or "as n goes to infinity." So, the binomial distribution is asymptotically normal as n approaches infinity. But for any fixed n, it's still a discrete distribution (with no meaningful or useful fractions in between the discrete values) whereas the normal distribution is continuous (with an infinite number of potentially continuously smaller fractional values in between each scale item).

Where:

- p1 and p2 are the sample proportions (the number of observed events compared to the number of possible events)

- n1 and n2 are the number of possible events (the sample group sizes)

The z-value (from the formula in Figure 1) can be used to obtain a p-value from the standard normal distribution. And the resulting p-value can be interpreted as the likelihood or probability of obtaining the observed data under the H0 (if the H0 were correct).

If the p-value is sufficiently small – less than an alpha level, or tolerance for type-1 error, that was stated before obtaining the data and before completely the calculations, then the H0 can be rejected and the data can be said to support the HA. Conversely, the H0 cannot be rejected if the p-value exceeds the alpha level, in which case, the data do not support the HA. However, this states only that the H0 cannot be rejected, and is not to say that the data have proved the H0 to be correct.

X^2 Test of Proportions for Three Samples

The TOP can be extended to three or more samples using the chi-squared (X^2) test. Formally, the X^2 test is used to evaluate the independence of the sample categories. Figure 4. shows the formula for the X^2 TOP.

Figure 4. X^2 Test of Proportions.

$$X^2 = \sum \frac{(O-E)^2}{E}$$

Where:

- *O* is the observed frequency of occurrence, and

- *E* is the expected frequency under H0

Historical Application of the TOP: Addressing Gender Bias in College Admissions

A historical example of the application of the TOP involved addressing gender

bias in college admissions during the 1970s. During this period, concerns arose regarding potential gender bias in the admission practices of colleges and universities in the United States – specifically at the University of California at Berkeley. It was suspected that women were facing discrimination, leading to unequal representation in higher education. Statisticians and researchers subsequently applied the TOP, specifically the z-test for two samples, to assess whether there were statistically significant differences in the proportions of male and female applicants admitted to educational institutions.

Initial results using data from all academic departments showed that differences in college admission rates were statistically significant with men being admitted more frequently. However, results for individual departments show a reversal of this trend due to the differences in the frequency (number) of male and female applicants in different departments. Females applied more frequently to more highly competitive departments, where rejection was more likely, while males applied more frequently to departments with less competitive admissions[4]. Although overall 44% of male applicants were admitted, compared to 35% of female applicants, when individual departments were analyzed no department was discriminating against female applicants, with a small but statistically significant bias in favor of female applicants. The trend of females outpacing males in college admissions and graduations has been observed consistently since the late 1970s, though females are underrepresented in some areas of study.

Application of the TOP to Polygraph Countermeasure Analysis

The comparison question technique (CQT) is a form of scientific credibility assessment (lie detection) test that relies on the comparison physiological responses to relevant questions (RQs) about the event in question with responses to comparison questions (CQs). All forms of scientific testing are concerned with the construct validity of the test data and test results. Polygraph testing does not detect or measure deception per se, simply because deception is an amorphous construct). Instead, polygraph tests rely on autonomic signals that are correlated with deception in the CQT, and for which statistical models can be developed using combinations of different signals. Because scientific tests of all types are used to quantify phenomena that cannot be subject to perfect deterministic observation (immune to human behavior and unaffected by random variation) or physical measurement (which requires a physical phenomena, and would be subject only to random measurement error), scientific test results are inherently probabilistic and rely on the statistical relationship between a phenomena of interest that cannot be measured physically and proxy signals that are available for recording and measurement. To the degree

[4] This reversal of statistical trends for whole group and sub-group data is referred to as Simpson's paradox.

that all physiological activity is associated with multiple forms of human activity, polygraph tests, like many scientific tests, may include some vulnerabilities to strategic faking or manipulation, and statistical methods can be used to monitor and reduce such vulnerabilities.

A difficulty for those who attempt to voluntarily manipulate or fake their polygraph test data will be that there is some likelihood that voluntary faking activity may produce data that is qualitatively or quantitatively distinct from normal autonomic activity. Unusual or atypical data signals are referred to by polygraph examiners as "artifacts" or more generally as "atypical physiological activity." A related difficulty is that data artifacts may sometimes be the result of involuntary or innocent causes, and the mere presence of data artifacts may lead to suspicions or accusations of faking under some high-risk or high-value circumstances. Because the CQT is a standardized and systematic procedure, faking efforts must also be systematic if they are to be effective. That is, successful faking in the CQT must systematically reverse the loading of changes in physiological activity that occur in response to different types of test stimuli under the analytic theory of the polygraph test. So, while consistent systematic activity may have the greatest potential to alter the resulting numerical values it will also have the greatest vulnerability to be easily observed. More sophisticated faking efforts might involve the use of a variety of strategic activities that are executed in an inconsistent or pseudo-random manner that may be less easily observed, but which, if they insufficiently systematic, may also fail to achieve the desired reversal of loading of the physiological data.

A difficulty for polygraph field practitioners and is that is while the mere observation of unusual activity may not be difficult, attempts to make accurate attributions about the motivation or intent of such activity requires accurate insight into an examinee's motivation or state of mind – whether observed artifacts are systematic and deliberate or random and involuntary. Many polygraph artifacts can produce similarly atypical data regardless of whether they are voluntary and systematic or involuntary. If it were possible to gain accurate insight into the examinee's state of mind it might preclude the need for polygraph testing. Moreover, when considering the presently insurmountable complications surrounding the possibility of mind-reading at the present time, it becomes problematic, and therefor incorrect, to attempt to make attributions that observed artifacts are voluntary or systematic when attribution to a plausible innocent cause has not been ruled out in an acceptable scientific manner.

The common scientific approach to the need to make decisions under circumstances that are subject to inherent uncertainty is to use statistical methods to quantify the margin of uncertainty or level of confidence that the data can provide in support of the different possible conclusion. The TOP provides a viable means to investigate the possibility of systematic faking by evaluating the frequency and proportions of data artifacts that occur in response to different types of test stimuli. If the TOP indicates a significant difference in the proportions data artifacts during CQs compared to RQs, it

suggests that countermeasures may be in use. This information can be valuable in evaluating the reliability of polygraph test results and adjusting interrogation strategies accordingly.

Conclusion

Science is a structured process that helps solve complex issues in the natural world and understand human behavior. It uses statistics and probabilities as its mathematical language. Scientific conclusions, with inherent uncertainties, are derived from rigorous mathematical logic and consistent information. Probabilities form the bedrock of informed decisions in a world where certainty is elusive. While intuitive science is based on abstract thinking and assumptions, empirical science relies on observation and data analysis. Null Hypothesis Significance Testing (NHST) is a method where evidence from experiments is used to evaluate the validity of null and alternative hypotheses. The Test of Proportions (TOP) is a statistical method to analyze the strength of evidence. It can be executed through z-tests for two samples or the chi-squared (X^2)test for three or more samples. Historically, the TOP has been used to address gender biases in college admissions and has applications in polygraph countermeasure analysis.

References

Agresti, A. (2002). *Categorical Data Analysis (2nd ed.).* John Wiley & Sons, Inc.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119-126.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science, 187(4175)*, 398-404.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Bruning, J. L, & Kintz, B. L. (1997) *Computational Handbook of Statistics*. (*4th Ed.*). Pearson.

Cochran, W. G. (1954). Some methods for strengthening the common X2 tests.*Biometrics, 10(4)*, 417-451.

Cohen, J. (1994). The Earth is round (p < .05). American Psychologist, 49(12), 997.

Conover, W. J. (1999). *Practical Nonparametric Statistics (3rd ed.)*. John Wiley & Sons.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science, 25(1),* 7-29.

Everitt, B. S. (1977). *The Analysis of Contingency Tables.* Chapman & Hall.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Gentle, J. E. (2009). *Computational Statistics*. New York: Springer.

Honts, C. R. (2004). The comparison question test. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (pp. 14-40). Cambridge University Press.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694-706)*, 289-337.