Probable Error of a Mean, The ("Student") [Biometrika, 6, 1-25]

Initially appreciated by only a handful of brewers and statisticians, “The Probable Error of a Mean” is now, 100 years later, universally acclaimed as a classic by statisticians and behavioral scientists alike. Written by William Sealy Gosset under the pseudonym “Student”, its publication paved the way for the statistical era that continues today, one focused on how best to draw inferences about large populations from small samples of data.

Gosset and “Student”

Schooled in mathematics and chemistry, Gosset was hired by Arthur Guinness, Son, & Co., Ltd. to apply recent innovations in the field of statistics to the business of brewing beer. As a brewer, Gosset analyzed how agricultural and brewing parameters (e.g., the type of barley used) affected crop yields and, in his words, the “behavior of beer”. Because of the cost and time associated with growing crops and brewing beer, Gosset and his fellow “experimental” brewers could not afford to gather the large amounts of data typically gathered by statisticians of their era. Statisticians, however, had not yet developed accurate inferential methods for working with small samples of data, requiring Gosset to develop methods of his own. With the approval of his employer, Gosset spent a year (1906-1907) in Karl Pearson’s biometric laboratory, developing “The Probable Error of a Mean” as well as “Probable Error of a Correlation Coefficient”.

The most immediately striking aspect of “The Probable Error of a Mean” is its pseudonymous author: “Student”. Why would a statistician require anonymity? The answer to this question came publicly in 1930, when fellow statistician Harold Hotelling revealed that “Student” was Gosset, and that his anonymity came at the request of his employer, a “large dublin Brewery”. At the time, Guinness considered its use of statistics a trade secret and forbade its employees from publishing their work. Only after negotiations with his supervisors was Gosset able to publish his work, agreeing to neither use his real name nor publish proprietary data.

The Problem: Estimating Sampling Error

As its title implies, “The Probable Error of a Mean” focuses primarily on determining the likelihood that a sample mean approximates the mean of the population from which it was drawn. The “probable error” of a mean, like its standard error, is a specific estimate of the dispersion of its sampling distribution, and was used commonly at the start of the 20th century. Estimating this dispersion was then and remains today a foundational step of statistical inference: To draw inference about a population parameter from a sampled mean (or, in the case of null hypothesis significance testing, infer the probability that a that a certain population would yield a sampled mean as extreme as the obtained value), one must first specify the sampling distribution of the mean. The Central Limit Theorem provides the basis for parametrically specifying this sampling distribution, but does so in terms of population variance. In nearly all research, however, both population mean and variance are unknown. To specify the sampling distribution of the mean, therefore, researchers must use the sample variance.

Gosset confronted this problem with using sample variance to estimate the sampling distribution of the mean, namely that there is error associated with sample variance. And because the sampling distribution of the variance is positively skewed, this error is more likely to result in the underestimation than the overestimation of population variance (even when using an unbiased estimator of population variance). Furthermore, this error, like the error associated with sampled means, increases as sample size decreases, presenting a particular (and arguably exclusive) problem for small sample researchers such as Gosset. To draw inferences about population means from sampled data, Gosset could not – as large sample researchers did – simply calculate a standard *z* statistic and rely on a unit normal table to find the corresponding *p* values. The unit normal table does not account for either the estimation of population variance or the fact that the error in this estimate depends on sample size. This limitation inspired Gosset to write "The Probable Error of a Mean" in a self-described effort to 1) determine at what point sample sizes become so small that the above method of normal approximation becomes invalid, and 2) develop a set of valid probability tables for small samples sizes.

The Solution: "z"

To accomplish these twin goals, Gosset derived the sampling distribution of a new statistic he called "*z*". He defined *z* as the deviation of the mean of a sample ($\overline{X}$) from the mean of a population (u) divided by the standard deviation of the sample (*s*), or $(\overline{X} - u)/s$. In his original paper, Gosset calculated *s* with the denominator *n* (leading to a biased estimate of population variance, s^2) rather than the unbiased $n-1$, likely in response to Karl Pearson's famous attitude that "only naughty brewers take n so small that the difference is not of the order of the probable error!'' To determine the sampling distribution of z, Gosset first needed to determine the sampling distribution of s. To do so, he derived the first four moments of s^2, which allowed him to make an informed guess concerning its distribution (and the distribution of s). Next, he demonstrated that $\overline{X}$ and s were uncorrelated, presumably in an effort to show their independence. This independence – in conjunction with equations to describe the distribution of s – allowed Gosset to derive the distribution of z.

This first portion of "The Probable Error of a Mean" is noteworthy for its speculative, incomplete, and yet ultimately correct conclusions. Gosset failed to offer a formal mathematical derivation for the sampling distribution of s, despite the fact that, unbeknownst to him, such a proof had been published 30 years earlier by the German statistician Friedrich Robert Helmert. Nor was Gosset able to prove that the sampling distributions of s^2 and $\overline{X}$ were completely independent of each other. Nevertheless, Gosset was correct on both counts, as well as his ensuing derivation of the sampling distribution of z, leading many to note that his statistical intuition more than compensated for his admitted mathematical shortcomings.

Pioneering Use of Simulation

"The Probable Error of a Mean" documents more than Gosset's informed speculation, however: it presents one of the first examples of simulation in the field of statistics. Gosset used simulation to estimate the sampling distribution of z non-parametrically, and then compared this result to his parametrically-derived distribution. Concordance between the two sampling distributions, he argued, would confirm the validity of his parametric equations.

To conduct his simulation, he relied on a biometric database of height and finger measurements collected by British police from 3000 incarcerated criminals; this database served as his statistical population. Gosset randomly ordered the data – written individually on pieces of cardboard – then segregated them into 750 samples of 4 measurements each (i.e., $n = 4$). For every sample, he calculated z for height and finger length, then compared these two z distributions with the curves he expected from his parametric equations. In both cases, the empirical and theoretical distributions did not differ significantly, thus offering evidence that Gosset's preceding equations were correct.

Tables and Examples

Gosset dedicated the final portion of "The Probable Error of a Mean" to tabled probability values for z and illustrative examples of their implementation. To construct the tables, he integrated over the z distributions (for sample sizes of 4-10) to calculate the probability of obtaining certain z values or smaller. For purposes of comparison, he also provided the p-values obtained via the normal approximation to reveal the degree of error in such approximation. The cumbersome nature of these calculations deterred Gosset from providing a more extensive table.

In further testament to his applied perspective, Gosset concluded the main text of

"The Probable Error of a Mean" by applying his statistical innovation to four sets of actual experimental data. In the first and most famous example, Gosset analyzed data from a 1904 experiment that examined the soporific effects of two different drugs. In this experiment, researchers had measured how long patients (n = 15) slept after treatment with each of two drugs and a drug-free baseline. To determine whether the drugs helped patients sleep, Gosset tested the mean change in sleep for each of the drug conditions (compared to the baseline) against a null (i.e., zero) population mean. To test whether one drug was more effect than the other drug, he tested the mean difference in their change values against a null population mean. All three of these tests – as well as the tests used in the three subsequent examples – correspond to modern-day one-sample t tests (or equivalent paired t tests).

Postscript: From *z* to *t*

With few exceptions over nearly twenty years following its publication, "The Probable Error of a Mean" was neither celebrated nor appreciated. In fact, when providing a expanded copy of the "Student" z tables to the then little known statistician Ronald Fisher in 1922, Gosset remarked that Fisher was "the only man that's ever likely to use them!" Fisher ultimately disproved this gloomy prediction by championing Gosset's work and literally transforming it into a foundation of modern statistical practice.

Fisher's contribution to Gosset's statistics was threefold. First, in 1912 and at the young age of 22, he used complex n-dimensional geometry (that neither Gosset nor Pearson could understand) to prove Gosset's equations for the z distribution. Second, he extended and embedded Gosset's work into a unified framework for testing the significance of means, mean differences, correlation coefficients, and regression coefficients. In the process of achieving this unified framework (based centrally on the concept of degrees of freedom), Fisher made his third contribution to Gosset's work: he multiplied z by $\sqrt{n-1}$, transforming it into the famous t statistic that now inhabits every introductory statistics textbook.

During Fisher's popularization, revision, and extension of the work featured in "The Probable Error of a Mean", he corresponded closely with Gosset. In fact, Gosset is responsible for naming the t statistic, as well as calculating a set of probability tables for the new t distributions. Despite Gosset's view of himself as a humble brewer, Fisher considered him a statistical pioneer whose work had not yet received the recognition it deserved.

Historical Impact

The world of research has changed greatly in a century, from a time when only "naughty brewers" gathered data from samples sizes not measures in hundreds, to an era characterized from small sample research. "The Probable Error of a Mean" marked the beginning of serious statistical inquiry into small sample inference, and its contents today underlie behavioral science's most frequently used statistical tests. Gosset's efforts to derive an exact test of statistical significance for such samples (as opposed to one based on a normal approximation) may have lacked in mathematical completeness, but their relevance, correctness and timeliness shaped scientific history.

Samuel T. Moulton

See also Central Limit Theorem, Distribution, Nonparametric Statistics, Sampling Distributions, Sampling Error, Standard Error of the Mean, Student's *t* Statistic, *t* Test (Independent Samples), *t* Test (One Sample), *t* Test (Paired Samples)

Further Readings

Fisher, R. A. (1925). Applications of 'Student's' distribution. *Metron, 5*, 90-104.

Pearson, E. S. (1939). 'Student' as a statistician. Biometrika, 30, 210-250.

Boland, P. J. (1984). A biographical glimpse of William Sealy Gosset. *American Statistician, 38,* 179-183.

Box, J. F. (1981). Gosset, Fisher, and the *t*-distribution. *American Statistician, 35,* 61-66.

Eisenhart, C. (1979). Transition from Student's *z* to Student's *t. American Statistician, 33,* 6-10.

Hanley, J. A., Julien, M., & Moodie, E. E. M. (2008). Student's *z*, *t*, and *s*: What if Gosset had R? *American Statistician, 62,* 64-69.

Lehmann, E. L. (1999). "Student" and small-sample theory. *Statistical Science, 14*, 418-426.

Pearson, E. S. (1968). Studies in history of probability and statistics XX: Some early correspondence between W. S. Gosset, R. A. Fisher, and K. Pearson with note and comments. *Biometrika, 55,* 445-457.

Zabell, S. L. (2008). On Student's 1908 article - "The Probable Error of a Mean". *Journal of the American Statistical Association, 103,* 1-7.