DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

Inference, Statistical

(Article begins on next page)

INFERENCE, STATISTICAL

To perform inference, in layman's terms, is to make an educated or informed guess of an unknown quantity of interest given what is known. *Statistical inference*, again in layman's terms, goes one step further, by making an informed guess about the error in our informed guess of the unknown quantity. To the layman, this may be difficult to grasp—if I don't know the truth, how could I possibly know the error in my guess? Indeed, the exact error—that is, the difference between the truth and our guess—can never be known when inference is needed. But when our data set, or more generally, quantitative information, is collected through a probabilistic mechanism—or at least can be approximated or perceived as such—probabilistic calculations and statistical methods allow us to compute the probable error, formally known as the "standard error," of our guess, or more generally, of our guessing method, the so-called "estimator." Such calculations also allow us to compare different estimators, that is, different ways of making informed guesses, which sometimes can lead to the best possible guess, or the most efficient estimation, given a set of (often untestable) assumptions and optimality criteria.

Consider the following semihypothetical example. Mary, from a prestigious university in Europe, is being recruited as a statistics professor by a private U.S. university. Knowing that salaries at U.S. universities tend to be significantly higher than at European universities, Mary needs to figure out how much she should ask for without aiming too low or too high; either mistake could prevent her from receiving the best possible salary. This is a decision problem, because it depends on how much risk Mary is willing to take and many other factors that may or may not be quantifiable. The inference part comes in because, in order to make an informed decision, Mary needs to know something about the possible salary ranges at her new university.

FROM SAMPLE TO POPULATION

As with any statistical inference, Mary knows well that the first important step is to collect relevant data or information. There are publicly available data, such as the annual salary surveys conducted by the American Statistical Association. But these results are too broad for Mary's purposes because the salary setup at Mary's new university might be quite different from many of the universities surveyed. In other words, what Mary needs is a conditional inference, conditional on the specific characteristics that are most relevant to her goal. In Mary's case, the most relevant specifics include (1) the salary range at her new university and (2) the salary for someone with experience and credentials similar to hers.

Unlike at public universities, salary figures for senior faculty at many private universities are kept confidential. Therefore, collecting data is not easy, but in this example, through various efforts Mary obtained $140,000, $142,000, and $153,000 as the salary figures for three of the university's professors with statuses similar to Mary's. Mary's interest is not in this particular sample, but in inferring from this sample an underlying population of possible salaries that have been or could be offered to faculty members who can be viewed approximately as exchangeable with Mary in terms of a set of attributes that are (perceived to be) used for deciding salaries (e.g., research achievements, teaching credentials, years since PhD degree, etc.). This population is neither easy to define nor knowable to most individuals, and certainly not to Mary. Nevertheless, the sample Mary has, however small, tells her something about this population. The question is, what does it tell, and how can it be used in the most efficient way? These are among the core questions for statistical inference.

DEFINING ESTIMAND

But the first and foremost question is what quantity Mary wants to estimate. To put it differently, if Mary knew the entire distribution of the salaries, what features would she be interested in? Formulating such an inference objective, or estimand, is a critical step in any statistical inference, and often it is not as easy as it might first appear. Indeed, in Mary's case it would depend on how "aggressive" she would want to be. Let's say that she settles on the 95th percentile of the salary distribution; she believes that her credentials are sufficient for her to be in the top 5 percent of existing salary range, but it probably would not be an effective strategy to ask for a salary that exceeds everyone else's.

Mary then needs to estimate the 95th percentile using the sample she has. The highest salary in the sample is $153,000, so it appears that any estimate for the 95th percentile should not exceed that limit if all we have is the data. This would indeed be the case if we adopt a pure nonparametric inference approach. The central goal of this approach is very laudable: Making as few assumptions as possible, let the data speak. Unfortunately, there is no free lunch—the less you pay, the less you get. The problem with this approach is that unless one has a sufficient amount of data, there is just not enough "volume" in the data to speak loudly enough so that one could hear useful messages. In the current case, without any other knowledge or making any assumptions, Mary would have no base to infer any figure higher than $153,000 to be a possible estimate for the 95th percentile.

MAKING ASSUMPTIONS

But as a professional statistician, Mary knows better. She knows that she needs to make some distributional assumptions before she can extract nontrivial information out of merely three numbers, and that in general, log-normal distribution is not a terrible assumption for salary figures. That is, histograms of the logarithm of salary figures tend to be shaped like a "bell curve," also known as the Gaussian distribution. This is a tremendous amount of information, because it effectively reduces the "infinitely unknowable" distribution of possible salary figures to only two parameters, the mean and the variance of the log of the salary. Mary can estimate these two parameters using the sample size of three if the three log salary figures (11.849, 11.864, 11.938) she obtained can be regarded as a probabilistic sample. This is a big "if," but for now, let us assume this is approximately true. Then the sample mean 11.884 and sample standard deviation 0.048 provide valid estimates of the unknown true mean μ and true standard deviation σ. Because for the normal distribution $N(\mu, \sigma^2)$ the 95th percentile is $z_{95} = \mu + 1.645\sigma$, Mary's estimate for the 95th percentile for the log salary distribution is $11.884 + 1.645 \times 0.048 = 11.963$. Because the log transformation is strictly monotone, this means that Mary's estimate for the 95th percentile for the salary distribution is exp $(11.963) = \$156,843$, about 2.5 percent higher than the observed maximal salary of \$153,000!

ASSESSING UNCERTAINTY

With a sample size of three, Mary knows well that there is large uncertainty in estimating the mean μ, as well as in estimating σ. But how do we even measure such error without knowing the true value? This is where the probabilistic calculation comes in, if the sample we have can be regarded as a probabilistic sample. By probabilistic sample, we mean that it is generated by a probabilistic mechanism, such as drawing a lottery. In Mary's case, the sample was clearly not drawn randomly, so we need to make some assumptions. In general, in order for any statistical method to render a meaningful inference conclusion, the sample must be "representative" of the population of interest, or can be perceived as such, or can be corrected as such with the help of additional information. A common assumption to ensure such a "representativeness" is that our data form an independently and identically distributed (i.i.d.) sample of the population of interest. This assumption can be invalidated easily if, for instance, faculty members with higher salaries are less likely to disclose their salaries to Mary. This would be an example of selection bias, or more specifically, a nonresponse bias, a problem typical, rather than exceptional, in opinion polls and other surveys that are the backbone of many social science studies. But if Mary knew how a faculty's response probability is related to the faculty member's salary, then methods do exist for her to correct for such a bias.

Mary does not have such information, nor does she worry too much of the potential bias in her sample. To put it differently, she did her best to collect her data to be "representative," being mindful of the "garbage-in-garbage-out" problem; no statistical analysis method could come to rescue if the data quality is just poor. So she is willing to accept the i.i.d. assumption, or rather, she does not have strong evidence to invalidate it. This is typical with small samples, where model diagnosis, or more generally, assumption checking is not directly feasible using the data alone. But contrary to common belief, just because one does not have enough data to check assumptions, this does not imply one should shy away from making parametric assumptions. Indeed, it is with small samples that the parametric assumptions become most valuable. What one does need to keep in mind when dealing with a small sample is that the inference will be particularly sensitive to the assumptions made, and therefore a sensitivity analysis—that is, checking how the analysis results vary with different assumptions—is particularly necessary.

Under the i.i.d. assumption, we can imagine many possible samples of three drawn randomly from the underlying salary population, and for each of these samples we can calculate the corresponding sample mean and sample standard deviation of the log salary. These sample means and sample standard deviations themselves will have their own distributions. Take the distribution of the sample mean as an example. Under the i.i.d. assumption, standard probability calculations show that the mean of this distribution retains the original mean μ, but its variance is the original variance divided by the sample size n, σ^2/n. This makes good intuitive sense because averaging samples should not alter the mean, but should reduce the variability in approximating the true mean, and the degree of reduction should depend on the sample size: The more we average, the closer we are to the true mean, probabilistically speaking. Furthermore, thanks to the central limit theorem, one of the two most celebrated theorems in probability and statistics (the other is the law of large numbers, which justifies the usefulness of sample mean for estimating population mean, among many other things), often we can approximate the distribution of the sample mean by a normal distribution, even if the underlying distribution for the original data is not normal.

CONSTRUCTING CONFIDENCE INTERVALS

Consequently, we can assess the probable error in the sample mean, as an estimate of the true mean, because we can

use the sample standard deviation to estimate σ, which can then be used to form an obvious estimate of the standard error $\sigma/\sqrt{n}$. For Mary's data, this comes out to be $0.048/\sqrt{3} = 0.028$, which is our estimate of the probable error in our estimate of μ, 11.884. In addition, we can use our distributional knowledge to form an interval estimate for μ. Typically, an interval estimator is in an appealing and convenient form of "sample mean $\pm$ 2 $\times$ standard error," which is a 95 percent confidence interval when (1) the distribution of the sample mean is approximately normal; and (2) the sample size, n, is large enough (how large is large enough would depend on problems at hand; in some simple cases, $n = 30$ could be adequate, and in others, even $n = 30,000$ might not be enough). For Mary's data, the assumption (1) holds under the assumption that the log salary is normal, but the assumption (2) clearly does not hold. However, there is an easy remedy, based on a more refined statistical theory. The convenient form still holds as long as one replaces the multiplier 2 by the 97.5th percentile of the t distribution with degrees of freedom $n - 1$. For Mary's data, $n = 3$, so the multiplier is 4.303. Consequently, a 95 percent confidence interval for μ can be obtained as $11.884 \pm 4.303 \times 0.028 = (11.766, 12.004)$. Translating back to the original salary scale, this implies a 95 percent confidence interval ($128,541, $163,407). This interval for the mean is noticeably wider than the original sample range ($140,000, $153,000); this is not a paradox, but rather a reflection that with sample size of only three, there is a tremendous uncertainty in our estimates, particularly because of the long tail in the log-normal distribution.

So what is the meaning of this 95 percent confidence interval? Clearly it does not mean that (11.766, 12.004) includes the unknown value μ with 95 percent probability; this interval either covers it or it does not. The 95 percent confidence refers to the fact that among all such intervals computed from all possible samples of the same size, 95 percent of them should cover the true unknown μ, if all the assumptions we made to justify our probabilistic calculations are correct. This is much like when a surgeon quotes a 95 percent success chance for a pending operation; he is transferring the overall (past) success rate associated with this type of surgery—either in general, or by him—into confidence of success for the pending operation.

By the same token, we can construct a confidence interval for σ, and indeed for Mary's estimand, a confidence interval for the 95th percentile $z_{95} = \mu + 1.645\sigma$. These constructions are too involved for the current illustration, but if we ignore the error in estimating σ (we shouldn't if this were a real problem), that is, by pretending $\sigma = 0.048$, then constructing a 95 percent confidence interval for $z_{95} = \mu + 1.65\sigma$ would be the same as for $\mu +$

$1.645 \times 0.048 = \mu + 0.079$, which is $(11.766 + 0.079, 12.004 + 0.079) = (11.845, 12.083)$. Translating back to the original salary scale, this implies that a 95 percent confidence interval for z_{95} would be ($139,385, $176,839). The right end point of this interval is about 15 percent higher than the maximal observed salary figure, $153,000. As Mary's ultimate problem is making a decision, how she should use this knowledge goes beyond the inference analysis. The role of inference, however, is quite clear, because it provides quantitative information that has direct bearing on her decision. For example, Mary's asking salary could be substantially different knowing that the 95th percentile is below $153,000 or could go above $170,000.

LIMITATIONS

One central difficulty with statistical inference, which also makes the statistical profession necessary, is that there simply is no "correct" answer: There are infinitely many incorrect answers, a set of conceivable answers, and a few good answers, depending on how many assumptions one is willing to make. Typically, statistical results are only part of a scientific investigation or of decision making, and they should never be taken as "the answer" without carefully considering the assumptions made and the context to which they would be applied. In our example above, the statistical analysis provides Mary with a range of plausible salary figures, but what she actually asks for will depend on more than this analysis. More importantly, this analysis depends heavily on the assumption that the three salary figures are a random sample from the underlying salary distribution, which is assumed to be log-normal. Furthermore, this analysis completely ignored other information that Mary may have, such as the American Statistical Association's annual salary survey. Such information is too broad to be used directly for Mary's purposes (e.g., taking the 95th percentile from the annual survey), but nevertheless it should provide some ballpark figures for Mary to form a general prior impression of what she is going after. This can be done via Bayesian inference, which directly puts a probabilistic distribution on any unknown quantity that is needed for making inference, and then computes the posterior distribution of whatever we are interested in given the data. In Mary's case, this would lead to a distribution for z_{95}, from which she can directly assess the "aggressiveness" of each asking salary figure by measuring how likely it exceeds the actual 95th percentile. For illustration of this more flexible method, see Gelman et al (2004).

SEE ALSO *Classical Statistical Analysis; Degrees of Freedom; Distribution, Normal; Errors, Standard; Inference, Bayesian; Selection Bias; Standard Deviation; Statistics; Statistics in the Social Sciences*

BIBLIOGRAPHY

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* 2nd ed. Pacific Grove, CA: Thompson Learning.

Cox, D. R. 2006. *Principles of Statistical Inference.* Cambridge, U.K.: Cambridge University Press.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics.* London: Chapman and Hall.

Gelman, Andrew, J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian Data Analysis.* Boca Raton, FL: Chapman and Hall/CRC.

Xiao-Li Meng

INFERENCE, TRAIT

SEE *Trait Inference.*

INFERIORITY COMPLEX

Although Sigmund Freud is best known for his influence on the field of psychology, he was also a renowned teacher. Alfred Adler (1870–1937), a student of Freud, broke from Freud's teachings, criticizing his focus on the sexual. Adler established an approach he called *individual psychology*, which focused on the individual's need for fulfillment and power; he is credited with developing concepts such as birth order, quest for significance, mental life, and a variety of complexes including the *inferiority complex.* Adler suggested that the two needs children have to master are inferiority (or the will for power) and the need for social approval. According to Adler, people are constantly striving to be powerful, and feelings of inferiority (or weakness) often pull them into a consuming state of self-interest. It is important to note that, for Adler, inferiority itself is not negative; rather, it is a normal and even motivating force in life. All humans have some feelings of inferiority and are striving to overcome them. It is when one becomes fully consumed in his or her pursuit of power, or by feelings of inferiority to the point of paralysis, that inferiority becomes a burden. It is at this point that one shifts from having feelings of inferiority to having what Adler called the *inferiority complex.*

Adler initially conceptualized inferiority with regard to what he termed *organ inferiority.* In 1907 Adler wrote *Study of Organ Inferiority and Its Physical Compensation,* in which he theorized that organ inferiority occurred when one bodily organ was significantly weaker than another organ, causing the surrounding organs to compensate for the weakness in the inferior organ and make up for the deficiency in another way. Similarly, Adler suggested that individuals have or perceive that they have areas in which

they are deficient—whether physical or psychological. Adler thought that, beginning in childhood, a deep feeling of inferiority is instilled in each individual as a result of the child's physical stature. In contrast to an adult, a child sees himself or herself as inferior in both physical and psychological abilities. However, Adler thought that the degree to which the child feels inferior is largely the result of the child's environment and interpretation of that environment. That is, some children perceive that they have more deficiencies or greater weaknesses because of the challenges they face, the way they interact with the adults in their lives, or the negative messages they get about their abilities. These children come to believe that they are inferior based on their perceptions of themselves and their life, not based on measurable or concrete criteria.

As adults, individuals also perceive areas of deficiency or weakness. These perceived weaknesses may result from life experiences (e.g., receiving a low test score) or from critical statements made by important others (e.g., being called stupid). Regardless of how the perceived deficit is brought to the individual's awareness, once the individual identifies an area perceived to be a weakness, he or she tries to compensate for those feelings of inferiority and achieve power. However, if this compensation does not have the desired result, the individual may become fully focused on the inferiority and develop what Adler called the *inferiority complex.*

According to Adler, the inferiority complex is a neurosis; the individual is fully consumed in their focus on the inferiority. It is a magnification of the normal feelings of inferiority, and results when strivings to overcome inferiority are greatly hindered. Individuals who struggle with feelings of inferiority may rate themselves in some area important to them as a 5 on a scale of 1 to 10, when they would aspire to a 6 or 7. In contrast, those with an inferiority complex may rate themselves as a 2 on a scale of 1 to 10 when they aspire to a 9. Those with an inferiority complex may also believe that there is no hope of ever reaching 9. The perception of one's shortcomings is an important aspect of this complex. That is, it matters more where individuals perceive themselves to be than it does where they actually are.

An individual with an inferiority complex is often overwhelmed, and as a result, the inferiority complex can become as consuming as an ailment or disease. Individuals may become manipulative in order to try to get others to give them the affirmation they are looking for, or they may try to use their deficiencies to get special attention or accommodation for circumstances that they are actually capable of handling or overcoming on their own. Those with inferiority complexes may be self-centered, depressed, incapable of development, compliant, shy, insecure, timid,