Statistical Inference, Statistics Education, and the Fallacy of the Transposed Conditional

Andrew A. Neath
Department of Mathematics and Statistics
Southern Illinois University Edwardsville
Edwardsville, IL 62026
Email: aneath@siue.edu

I. Introduction

The primary approaches to statistical inference presented to students are significance testing and confidence interval estimation. The interpretations of significance tests and confidence intervals play a key role in an introductory statistics course. In spite of the considerable attention paid to these issues, the interpretations which are often presented to statistics students are based on a logical fallacy. This essay introduces readers to how the "fallacy of the transposed conditional" leads to interpretations of statistical inferential procedures which lack proper justification.

II. Issues in Significance Testing

For ease of exposition, we will introduce a simple case of statistical inference. Suppose our data consists of random variables $Y_1, Y_2, \ldots, Y_n$ distributed independent $N\left(\mu, \sigma^2\right)$. The variance σ^2 is known. The mean μ is unknown, but a nominal value μ_o is available. Based on the observed data, the problem is to decide between the competing hypotheses $H_o: \mu = \mu_o$, $H_a: \mu \neq \mu_o$. This case can serve as an introduction to hypothesis testing in a first statistics course. The simplicity of the case results in no loss of generality as an introduction to hypothesis testing since the interpretations generalize to more complicated cases. (It is this ability to generalize to the more complicated which makes teaching introductory statistics an interesting and important endeavor.) The errors which can be committed in a hypothesis testing problem are, of course, denoted as type I (deciding H_a when H_o is true) and type II (deciding H_o when H_a is true). If committing a type I error is considered to be severe, a decision rule can be created where the probability of committing a type I error is controlled at $P\left(\text{type I error}\right) = P\left(\text{decide } H_a | H_o \text{ true}\right) = \alpha$, where α is called the the significance level of the test. Our simple case then results in a decision of H_a when the test statistic $|Z^*| = \left|\sqrt{n}\left(\overline{Y} - \mu_o\right)/\sigma\right|$ exceeds $z\left(\alpha/2\right)$, the upper $\left(\alpha/2\right)^{\text{th}}$ percentile of the standard normal distribution.

Because we are presenting a setting where a type I error carries a higher cost than a type II error, the hypotheses H_o and H_a are said to be asymmetric. Common explanations for taking this asymmetric approach to hypothesis testing are that the alternative hypothesis is "what we hope to establish as true", so that the "burden of proof", or "sufficient evidence" is required before a decision in favor of H_a is reached. A decision rule created with a type I error probability set at significance level α can be reasoned to students as initially believing the null hypothesis as true, and not moving off this belief unless the observed data is so unusual that doubt must be cast on this initial belief. A decision in favor of H_a is commonly interpreted as "rejecting the null hypothesis". A decision in favor of H_o is explained to students as one for which the data is not strong enough to cast doubt on the null hypothesis. It is stressed that the data can not "support" the truth of the null hypothesis. A decision in favor of H_o is commonly interpreted as "failing to reject the null hypothesis".

We can not be sure that a decision reached after performing a hypothesis test is correct. A decision, whether in favor of H_o or H_a, is made with uncertainty. The implication behind the interpretations of the decisions "reject H_o" and "fail to reject H_o" is that the uncertainty behind the decision in favor of H_a (reject H_o) is negligible, whereas the uncertainty behind the decision in favor of H_o (fail to reject H_o) is substantial enough to prevent a strong conclusion. Let's add some rigor to this discussion. Denote the events $B = [H_o \text{ true}]$ and $A = [\text{decide } H_a]$. A significance test is developed so that $P\left(A|B\right) = \alpha$ is small. Aside from the probability of a type I error, no other probabilities are specified in a significance test. The probability of a type II error, $P\left(A'|B'\right)$, need not be specified. It is not required to state how the particular test fits into a broader array of similar test results. So, $P\left(B\right)$ is not specified. The justification behind "rejection of the null hypothesis" as an interpretation of a decision in favor of H_a rests entirely from the idea that $P\left(A|B\right)$ is small.

Ziliak and McCloskey (2008) coined the phrase "fallacy of the transposed conditional" to describe the mistake of stating that $P\left(B|A\right)$ is equal to $P\left(A|B\right)$. It is a form of this fallacy which leads one to argue that deciding H_a on the basis of a significance test leads to the strong conclusion of "rejecting" a null hypothesis. But a small conditional probability $P\left(A|B\right)$ is not enough to conclude that the transposed conditional probability $P\left(B|A\right)$ is small. In significance testing, contrary to what is often taught to introductory students, a decision in favor of H_a is not necessarily enough to provide strong evidence that H_a is true and H_o is false. Bayes rule leads to an expression for the probability that a "rejected" null hypothesis

is actually true. We have

$$P(B|A) = \frac{\theta\,\alpha}{\theta\,\alpha + (1-\theta)\,\pi}$$

where $\alpha = P(A|B)$ is the significance level of the test, $\pi = P(A|B')$ is the power of the test, and $\theta = P(B)$ represents to probability of a true null hypothesis before data is observed. Note how the uncertainty involved in "rejecting" the null hypothesis involves not just the significance level α, but the power π and the prior probability θ as well.

Consider an example to illustrate this point. A drug in the early stage of development is not believed to have a strong chance of actually being effective. If the drug if effective, however, it would represent a major breakthrough. So it is worthwhile to conduct a clinical trial investigating the effects of the drug. Suppose that upon completion of the clinical trial, a hypothesis test at significance level $\alpha = .05$ results in the decision to accept the althernative hypothesis that the drug is effective. Introductory statistics students are taught at this point to "reject" the null hypothesis in favor of the conclusion that the drug is effective. For argument's sake, let's set the prior probability of drug effectiveness at .01, so that $P(B) = .99$. Clinical trials are typically designed to achieve power of .8. Let's set $P(A|B') = .8$. A simple calculation using Bayes rule now reveals the posterior probability on the null hypothesis that the drug is not effective as $P(B|A) = .86$. Upon further consideration, the correct interpretation is quite different from a "rejection" of the null hypothesis.

The misunderstanding in the interpretation of significance testing may be explained by the reasons that make the Bayesian approach to inference so compelling. After data is collected, it is natural to want an assessment of what is known about the quantity of interest. In a hypothesis testing problem, it is natural to want an assessment of the belief in the competing hypotheses. However, controlling the type I error rate alone, or controlling the type I error rate and the power, is not enough to allow for such an assessment. Instead of fooling ourselves and our students with interpretations about strong conclusions of "rejecting" a hypothesis on the basis of a significance test, we need to recognize the limitations of frequentist based hypothesis testing.

The intention of this essay is not to promote Bayesian approaches to statistical inference at the expense of frequentist approaches. We only wish to point out the limitations of frequentist approaches which are quite often misunderstood when presented to statistics students. It is important to compare the frequentist and Bayesian approaches to gain an understanding of the benefits and limitations of each. The advantage of frequentist based hypothesis testing is the ability to control the probability of error at important values of the parameter. In significance testing, it is deemed particularly important not to make the error of deciding the alternative when in fact the null is true. Thus, significance testing is an approach with desirable properties when the true parameter is at the null value. Bayesian tests do not necessarily have this property. In a case where the null hypothesis is given small probability a priori, a Bayesian test is more apt to favor the alternative at the expense of controlling the type I error rate. This is not a problem in the mind of the Bayesian, as only small weight is given to the chance that the null hypothesis is true. It does concern the frequentist who considers protection against a type I error to be important regardless of how likely the null hypothesis is to be true. This protection comes at a price. That price is an inability to necessarily achieve a desired posterior level of belief in the resulting decision. To provide a decision with stong belief from a significance test alone would be committing the fallacy of the transposed conditional.

III. Issues in Confidence Interval Estimation

We say the random set $C(Y)$ is a $(1-\alpha)\,100\%$ confidence set for the unknown parameter θ if

$$P_\theta(\theta \in C(Y)) = 1 - \alpha$$

for all θ in the parameter space Θ, where P_θ represents probability computed under parameter θ. If $C(Y)$ is a random interval $[L(Y), U(Y)]$, we refer to the confidence set as a confidence interval. We return to the simple case of random variables $Y_1, Y_2, \ldots, Y_n$ distributed independent $N(\mu, \sigma^2)$, with known variance σ^2. The mean μ is unknown and is to be estimated by a confidence interval. It is easy to see that the random interval $\overline{Y} \pm z(\alpha/2)\,\sigma/\sqrt{n}$ satisfies the conditions of a $(1-\alpha)\,100\%$ confidence interval for μ. The probability $1-\alpha$ holds prior to data collection. It also should be emphasized that the probability holds under all possible μ. This is a nice property that often goes unnoticed. Since the true mean is unknown, one wants the probability of a correct interval to hold across the entire parameter space.

The confidence set definition above meets a frequentist criterion. As in our discussion of frequentist inference from the last section, the frequentist approach to confidence estimation has the advantage of controlling the probability of error at important parameter values. A confidence set, in fact, protects the error probability across the entire parameter space. But a mistake in interpretation, similar to what we saw in the section on significance testing, also occurs in the presentation of confidence intervals to statistics students. After the data is observed and the interval is computed, it is natural to want a degree of certainty placed on the computed interval. It is explained to students that one can place the same confidence on a computed interval as on the process itself. The implication behind this interpretation is that the uncertainty in the correctness of the interval prior to data collection is passed to the computed interval after data collection.

Consider an example to illustrate the flaw in this line of reasoning. The level of a marker in a patient's body depends on the stage of the disease. Let's denote the three stages of the disease as parameter space $\Theta = \{I, II, III\}$. Consider the sample space for the observed marker to be the discrete set $\{0, 1, 2, 3, 4, 5, 6\}$. Suppose the sampling distribution of Y

for each stage can be displayed as

		Θ		
		I	II	III
	0	.04	.025	.01
	1	.2	.025	.02
	2	.5	.2	.02
Y	3	.2	.5	.2
	4	.02	.2	.5
	5	.02	.025	.2
	6	.01	.025	.04

Table 1 : Probability distribution for marker level

A 90% confidence set can be formed by inverting a set of .90 probability for Y on each stage in the parameter space. The result of this inversion can be described as a set function of the random variable Y, as displayed in Table 2.

		confidence set
	0	$\emptyset$
	1	$\{I\}$
	2	$\{I, II\}$
Y	3	$\{I, II, III\}$
	4	$\{II, III\}$
	5	$\{III\}$
	6	$\emptyset$

Table 2 : Confidence set function for disease stage

It is easy to verify that the above function satisfies the requirement for a 90% confidence set. Prior to data collection, the probability of a correct set is equal to .90 under each of the possible disease stages. It is also clear that the definition of a $(1-\alpha)\,100\%$ confidence set does not negate the possibility of an empty interval, or an interval consisting of the entire parameter space. Suppose one observes $y = 0$ from one of the distributions in Table 1. The interpretation put forth by the approach taught to statistics students would be "We are 90% confident that the true mean lies in the empty set." Such a quantification of confidence on an estimate that is certain to be incorrect is not justifiable. If one observes $y = 3$, the confidence set consisting of the parameter space is certain to be correct. Unfortunately, the interpretation we are passing on to students is that we can only be 90% confident in this certain event.

A variation of the "fallacy of the transposed conditional" is being committed when we teach students that a probability of a correct interval conditional on the parameter becomes a quantification of confidence on an interval estimate conditional on the observed data. As in the interpretation put forth in significance testing, it is natural here to want to think like a Bayesian. A measure of the certainty in a computed interval estimate is desired. Such a measure, however, requires a Bayesian approach to the estimation problem. Frequentist confidence intervals do have desirable properties that Bayesian intervals lack. As mentioned earlier, frequentist intervals have the advantage of controlling the probability of error at all possible values of the parameter. Bayesian intervals do not necessarily have this property. Bayesians will not protect the a priori error rates at parameter values considered unlikely.

IV. Concluding Remarks

The interpretations of significance testing and confidence interval estimation often presented to statistics students are based on the "fallacy of the transposed conditional". This essay is not the first attempt to bring attention to the problem. Previous attempts, most notably those of Ziliak and McCloskey (2008), have focused on the usage and applications of statistics. The focus here is on the education of statistics students. I am not proposing a drastic overhaul of statistics education. The view presented in this essay is that significance testing and confidence interval estimation under frequentist viewpoints are statistical methods with solid justification. Frequentist methods can be designed to control, before data collection, the probability of error at those values of a parameter where such control is deemed to be important. We can think of frequentists as prudent for their willingness to protect the error rate under these scenarios. We are wrong as statistics educators, however, in giving the impression that frequentist methods allow for a posterior measure of belief. There is no question that the use of terms such as "rejection" and "confidence", purposefully or not, are misleading. This sort of posterior assessment is desirable, but requires a Bayesian approach to statistical inference. A major advantage to a Bayesian approach is the ability to assess the certainty in an outcome. A disadvantage is that a Bayesian will not necessarily control the probability of error conditional on important possible parameter values. If we think of frequentists as being prudent, perhaps we can think of Bayesians as being impetuous for not protecting against these scenarios.

The proposal here is quite modest. Both frequentist and Bayesian methods should be presented to students in an introductory course on statistical inference. An understanding of both approaches is necessary to understand the pros and cons of each. Furthermore, students will not leave with a statistics education predicated on a logical fallacy.

V. Reference

Ziliak, S. and McCloskey D. (2008). The Cult of Statistical Significance. University of Michigan Press.