

USING BAYESIAN STRENGTH OF BELIEF TO TEACH CLASSICAL STATISTICS

Milo Schield, Augsburg College, Minneapolis, MN USA

Previous papers by the author have argued that the Bayesian strength of belief can be used in interpreting classical hypothesis tests and classical confidence intervals. In hypothesis tests, one's strength of belief in the truth of the alternate upon rejecting the null was argued to be equal to $(1 - p)$ under certain conditions. In confidence intervals, being 95% confident was argued as being operationally equivalent to a willingness to bet on a 95% chance. These interpretations were taught in an introductory class of non-majors. Students found this approach to be extremely natural for confidence intervals. But in hypothesis testing, students had difficulty relating the quality of the test (p -value) to the quality of the decision. The underlying problem is student difficulty with related conditionals. To overcome this problem, we should teach more about conditionality -- not less.

Introduction: Students often treat classical probabilities as though they measured the Bayesian strength of belief in the truth of a claim. For a particular confidence interval, students may treat being 95% confident as a strength of belief that is somehow related to a 95% classical probability. In hypothesis tests, students often assess the truth of either hypothesis using strength of belief. In rejecting the null, they presume that the strength of belief in the null being true is related to the p -value.

To address these intuitions, business majors in introductory statistics were taught classical statistics with three differences. Type I error and alpha were defined differently. The classical p -value was used to calculate the Bayesian strength of belief that the null hypothesis is true when rejecting the null. The classical confidence level was interpreted as a Bayesian strength of belief.

Redefinition of Alpha and Type I Error: Alpha is traditionally defined as "the probability of Type I error." Type I error is traditionally defined as "the rejection of the null when the null hypothesis is true." Yet, Type I error is often illustrated by means of Table 1.

Table 1: Description of Outcomes in Classical Hypothesis Testing

CELLS	----- STATE OF NATURE -----	
DECISION	null is true	null is false
Fail to reject null	OK outcome	Type II error
Reject null	Type I error	OK outcome

In this 2x2 table, Type I error is shown as a single cell. This gives students reason to conclude that Type I error is a simple intersection of two co-equal conditions. And if

alpha is the probability of Type I error, then alpha should be the probability of having a random outcome in that cell. Thus, students may think alpha is just a table percentage or a row percentage. They have little reason to think that alpha is a column percentage.

Schild (1996) recommended that the definitions of Type I error and alpha be refined. Type I error was defined as "the null being rejected and being true" or as "rejecting a true null." This refinement makes the single-cell illustration of Type I error (Table 1) completely accurate. Alpha was defined as "the probability of Type I error when sampling from the null distribution." This refinement explicitly references both the whole and the part rather than having the whole buried inside the traditional definition of Type I error.

Hypothesis Testing: In Schild (1996), the relation between classical statistics (α and p-value) and the Bayesian strength of belief in the truth of the alternate was investigated. Consider a fixed level hypothesis test using alpha with a null ($\mu \leq \mu_o$) and a separated alternate ($\mu > \mu_1$) where $\mu_1 > \mu_o$. By varying the separation ($\mu_1 - \mu_o$) or the sample size (n) one can obtain $\beta = \alpha$ for any value of α . Under certain assumptions (including $\beta = \alpha$),

$$\delta = \alpha\gamma' / (\alpha\gamma' + \alpha'\gamma) \qquad \alpha = \delta\gamma / (\gamma\delta' + \delta\gamma)$$

where γ is the prior strength of belief in the truth of the alternate, where δ is the probability of a Type I error given one has rejected the null, and where $\alpha' = 1 - \alpha$ and $\gamma' = 1 - \gamma$. If $\gamma = 0.5 = \gamma'$, then $\delta = \alpha$. And if $\gamma = \alpha$, then $\gamma' = \alpha'$ and $\delta = 0.5$. When using a p-value to reject the null (a test of significance), simply use the p-value in place of alpha.

Figure 1

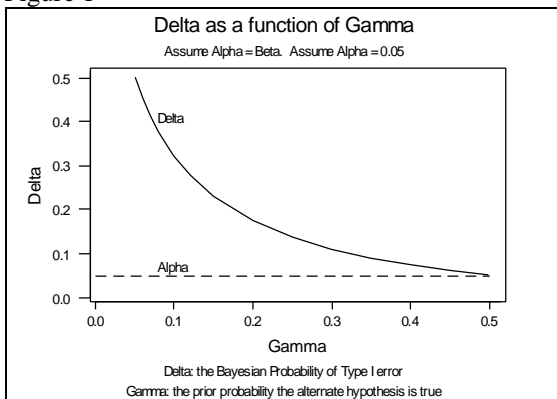


Figure 2

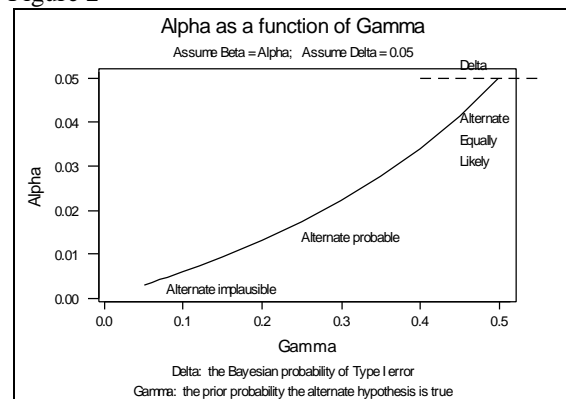


Figure 1 shows how one's strength of belief in the truth of the null (δ) increases as γ decreases given a rejection of the null. Recall that γ is one's strength of belief in the truth

of the alternate prior to the test. If $\alpha \ll 1$ and $\gamma \ll 1$, then $\delta \cong \alpha / (\alpha + \gamma)$. Thus, when rejecting the null, the posterior strength of belief in the truth of the alternate is given by $\delta' = 1 - \delta \cong \gamma / (\alpha + \gamma)$. As one's prior strength of belief in the truth of the alternate (γ) decreases, so does one's posterior strength of belief (δ') for a given level of α (or p-value). Remember that these formulas assume that α is *always* equal to β .

Figure 2 shows how to select α in order to have a 95% posterior strength of belief in the truth of the alternate given one has rejected the null. If $\alpha \ll 1$ and $\gamma \ll 1$, then $\alpha \cong \gamma(\delta/\delta')$ and $\alpha_{95} \cong \gamma/19$. As γ decreases (as one's prior strength of belief in the truth of the alternate decreases), one must decrease α to maintain a 95% "confidence" (δ') in the truth of the alternate given a rejection of the null. The more unlikely the condition (γ) the more powerful the test must be (the smaller α) for a given level of posterior "confidence" (δ').

To avoid burying the influence of a subjective prior inside the posterior, one should calculate what prior strength of belief in the alternate would be necessary to achieve a specified posterior strength of belief. For $\alpha \ll 1$, $\gamma \cong \alpha(\delta'/\delta)$, so $\gamma_{95} \cong 19\alpha$. If the p-value for a test of ESP is 10^{-6} and one rejects the null, one needs a prior strength of belief in ESP of at least 1.9×10^{-5} in order to obtain a 95% posterior "confidence" in ESP.

Confidence Level. In Schield (1997) confidence was viewed as measuring an objectively calibrated strength of belief. This viewpoint agrees with the Bayesians that confidence is psychological (a measure of one's strength of belief in the truth of a claim) and that psychological confidence is usually subjective. It agrees with the traditionalists that fixed parameters are not random variables and that statistical confidence should be objective. This viewpoint unites the Bayesian strength of belief (typically subjective) with the classical (traditional) approach to probability ("objective"). Table 2 illustrates the argument.

The first row shows probability from a classical perspective. The second row illustrates confidence as a strength of belief. The three columns indicate different situations. The first column is prior to sampling; the second column is after a particular sample is obtained but prior to knowing the statistic; the third column is after the statistic is known. The difference between the first two columns is metaphysical: potential versus actual. The difference between the last two columns is epistemic: unknown versus known.

Table 2. Claims about large-sample confidence intervals from a normal population where the standard deviation is known to be σ and where $SE = \text{Std. Error} = \sigma/\sqrt{n}$.

Confidence Intervals	Context of uncertainty (what)		
	Sample has not been drawn. μ is unknown \tilde{x} is variable and unknown	Sample is drawn; Statistic is unknown μ is unknown \bar{x}_γ is constant but unknown	Sample is drawn; Statistic is known μ is unknown \bar{x}_0 is constant and known
Description of uncertainty			
Objective Frequentist	Classical/traditional $P[(\mu_0 - 2 SE) \leq \tilde{x} \leq (\mu_0 + 2 SE)] = .95$	$P[(\mu_0 - 2SE) \leq \bar{x}_\gamma \leq (\mu_0 + 2SE)] = 0 \text{ or } 1$	$P[(\mu_0 - 2SE) \leq \bar{x}_0 \leq (\mu_0 + 2SE)] = 1$
Confidence Strength of belief that $ \bar{x} - \mu_0 \leq 2SE$	B ① 95% confidence 95% confident $ \tilde{x} - \mu_0 \leq 2 SE$	P 95% confidence 95% confident $ \bar{x}_\gamma - \mu_0 \leq 2 SE$	P 95% confidence 95% confident $ \bar{x}_0 - \mu_0 \leq 2 SE$

The three numbered steps form a sequential argument whereby the circled 95% confidence after step 3 is based on the circled 95% probability at the beginning of Step 1. Step 1 relies on the Principal Principle (Howson and Urbach, 1993, p. 240). “The principle states that if the objective, physical probability of a random event (in the sense of its limiting relative frequency in an infinite sequence of trials) were known to be r , and if no other relevant information were available, then the appropriate subjective degree of belief that the event will occur on any particular trial would also be r .” The Principal Principle is normative. Prior to sampling, one *should* be 95% confident that the population parameter will be included in the next random 95% confidence interval. Prior to flipping a fair coin, one *should* be 50% confident that the next flip will be heads.

In Step 2, a particular sample is obtained but the statistic is unknown. The classical probability becomes either zero or one. But since the outcome is unknown, one's confidence has no basis for being changed either up or down. If one "trusts in the process," one's confidence after selection should be the same as that before selection.

In Step 3, the sample statistic and the associated confidence interval are known. If this particular confidence interval includes the population parameter, then the classical probability is 1. But generally that fact of reality is unknown. Simply knowing the value of the sample statistic typically gives no reason to change one's strength of belief that the associated 95% confidence interval includes the fixed parameter.

In summary, one should be indifferent between (1) betting that a particular 95% confidence interval includes the unknown population parameter and (2) betting that one will draw a red ball from an urn containing 19 red balls and one blue ball.

Teaching Strength of Belief: There is considerable debate about the utility of teaching anything Bayesian in an introductory course. Berger (1980, p. 120) concluded, "...most such users (and probably the overwhelming majority) interpret classical measures in the direct probabilistic [Bayesian] sense. (Indeed the only way we have had even moderate success, in teaching elementary statistics students that an error probability is not a probability of a hypothesis, is to teach enough Bayesian analysis to be able to demonstrate the difference with examples.)" In opposition, David Moore (1992) has argued "There are, I think, good reasons not to stress Bayesian methods in beginning instruction about inference. First, they require a firm grasp of conditional probabilities.... This [difference between conditionals] is fatally subtle."

Teaching Results: Fifty business majors in two full-semester introductory statistics courses were taught using these new ideas: confidence as a objectively calibrated strength of belief, the redefinition of α and Type I error, and the calculation of the strength of belief in the truth of the alternate when rejecting the null. Conditional probabilities and Bayes Rule were presented by means of tables of counts and percents.

Confidence was taught by saying, "If you have confidence that the sample is random, then *operationally* one should treat a 95% confidence interval the same way one would treat a 95% chance." Operationally, one's willingness to bet on the result of flipping a fair coin should be independent of whether or not the coin has been flipped (provided the outcome is unknown). Confidence in the process justifies the same behavior. The result was a quiet success. Students had no difficulty accepting the equivalence. No extended argument was needed. A top student wondered why anyone would think otherwise.

Hypothesis testing was taught classically and then strength of belief was introduced using tables and counts. Students didn't seem noticeably better able to understand the meaning of alpha. It is much easier to simply say that alpha is "the probability of Type I error" than to have to add the phrase "when sampling from the null distribution." Students quickly realized the difference between the quality of the test (p-value) and the quality of

the prediction (strength of belief that the alternate is true given the null is rejected). Still, the students had problems: (1) They had difficulties describing Type I error (cf., "rejecting the null") and p-value (cf., "probability of rejecting the null" or "probability of Type I error"). (2) They had difficulty envisioning a 'trial'. In medical tests, multiple trials involved different subjects. In hypothesis tests about a state of nature, how could there be multiple trials with only one world? In general, students seemed overwhelmed by the large number of concepts involved in interpreting hypothesis tests from two different perspectives: classical and Bayesian.

Conclusion: David Moore is absolutely right; students do have considerable difficulty with conditional probability. Differences between related conditionals are "fatally subtle." But if we want our students to really understand hypothesis testing then we should teach students more about conditionality -- not less.

References

Berger, James O. (1980). *Statistical Decision Theory and Bayesian Analysis* 2nd Ed. Springer-Verlag.

Howson, Colin and Peter Urbach, (1993). *Scientific Reasoning* 2nd Ed. Open Court.

Moore, David (1993). *ICOTS Introduction to the Practice of Statistics*, 2nd ed., Freeman.

Schild, Milo (1996). *Using Bayesian Inference in Classical Hypothesis Testing*. 1996 ASA Proceedings of the Section on Statistical Education. p. 274.

Schild, Milo (1997). *Confidence: A Neo-classical Interpretation*. 1997 ASA Proceedings of the Section on Statistical Education.

Acknowledgments: Valuable feedback on the 1996 paper was received from Eric Sowey at the SISC-96 conference in Sydney; from Peter Holmes, Anne Hawkins and Tony O'Hagan at the RSS Centre for Statistical Education; from Colin Howson at the London School of Economics; and from Jeff Witmer at the Midwest Conference on Teaching Statistics. Dr. Schild can be reached at schild@augsborg.edu.