USING BAYESIAN INFERENCE IN CLASSICAL HYPOTHESIS TESTING

Milo Schield, Augsburg College
Dept. of Business & MIS. 2211 Riverside Drive. Minneapolis, MN 55454

KEY WORDS: Critical thinking, statistical inference, Type I error, teaching, epistemology.

ABSTRACT:
When students obtain a statistically significant sample at a 5% level of significance, they may conclude they can be 95% confident that the alternate hypothesis is true. From a classical perspective, this conclusion is unwarranted. Two elements may inadvertently support this unwarranted conclusion: the traditional definitions of Type I error and alpha, and the silence about confidence. From a Bayesian perspective, this conclusion might be warranted. Bayes rule can relate the Bayesian and classical probabilities of Type I error if classical hypotheses are treated as point masses and if one can treat degrees of belief about the truth of a state of nature as a probability. If the truth of the null and alternate are equally likely, if $\beta = \alpha$, and if the sample statistic triggers a rejection of the null, then the Bayesian probability of Type I error is numerically equal to alpha and the Bayesian probability the alternate is true equals $1 - \alpha$. The Bayesian probability of Type I error increases as the alternate becomes more improbable for a given level of alpha. Using this technique to select alpha and to interpret p-values may improve understanding of classical tests and decrease statistical opportunism.

INTRODUCTION

Students in introductory statistics are usually introduced to the classical fixed-level hypothesis test. Given an alpha of 5% and a statistically significant random sample they may conclude they can be 95% confident that the alternate hypothesis is true.

I. CLASSICAL EVALUATION

From a classical perspective, this conclusion is unwarranted and in error. *However, there are two aspects of the classical approach that might encourage this error: the traditional definitions of Type I error and alpha, and the silence about confidence in hypothesis testing.*

A. DEFINITIONS OF ALPHA & TYPE I ERROR

In presenting the classical hypothesis test, alpha is traditionally defined as the probability of Type I error. Type I error is often illustrated as being an intersection of two conditions as illustrated in Table 1.

Table 1: Figurative Description of 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

Students may think as follows. On the one hand, one might take 50 samples from the Null distribution and perhaps 2 of them fall in the reject region. On the other hand, one might take 50 samples from the alternate distribution and perhaps 38 of them fall in the reject region. Thus, students might create the following table.

Table 2: Table of hypothetical counts

COUNT	STATE OF NATURE		
	Null is true	Null is false	Total
Fail to reject	48	12	60
Reject	2	38	40
Total	50	50	100

Now there may be some errors in this. First, in reality either the null is true or it is not. In reality, you cannot have counts in both columns. Second, alpha is the criteria by which the rejection region was defined – prior to sampling. Alpha is not obtained by sampling – except in the limit of large numbers. These relative frequencies are just estimates of alpha.

Problem:
But, given these counts and the aforementioned definitions of alpha and Type I error, some students would say that alpha is a table percentage since Type I error is a single cell. Given this data, those students might estimate alpha as 2% (2/100).

Some students may calculate alpha as a row percentage since rejecting the null is a necessary condition for Type I error. Given this data, these students would estimate alpha as 5% (2/40).

Actually, alpha is a column percentage since alpha = P(null is rejected | null is true). In this case, alpha is properly estimated as 4% (2/50). Unfortunately, the traditional definition of alpha may prevent students from seeing alpha as a column percent.

Traditional Solution:
To avoid this problem, most authors traditionally define – not just describe – Type I error as being conditional on a column-based process. Typical examples include:

- Type I error: rejecting H_O when H_O is true
- Type I error: rejecting a null hypothesis that is true
- Type I error: rejecting H_O given that H_O is true.
- Type I error: occurs if H_O is rejected when it is true

The appendix contains examples of conditional definitions taken from Smith, Kitchens, Iman, Moore, Moore and McCabe, Hogg and Tanis, Neter, Wasserman and Whitmore, and by Mendenhall, Sheaffer and Wackerly. Note that in all of these conditional definitions of Type I error, the stipulated condition is that the null is true.

Thus, these traditional definitions imply that Type I error occurs only within a column-based process – a classical test of significance. By making Type I error meaningful only if one is sampling from the null, then – and only then – is the short-form statement of alpha (the probability of Type I error) a proper definition.

Disadvantages
The traditional approach has several disadvantages. It requires students to ignore the simple definition of Type I error as an intersection that is readily illustrated by means of a 2x2 table. It requires students to consider Type I error as being conditional and thus meaningful only within a test of significance. It requires students to consider alpha as being conditional without using the normal keywords for conditionality. This process-oriented, conditional definition of Type I error is extremely subtle and very easy to misinterpret.

Explanation
Given these disadvantages, why do authors traditionally present alpha as being unconditional and Type I error as being conditional on a column-based process?

Authors may be using these unusual definitions to introduce a hidden premise: Bayesian conditional probabilities (a row-based process) are meaningless in a hypothesis test involving a state of nature. Since the hypotheses are about a state of nature and since a state of nature is, in fact, either true or false, *they argue it follows that* a Bayesian row probability of Type I error is meaningless since is it either 0 or 1. Once Type I error is defined as a column-based process (rejecting the null given the null is true), then one cannot use this concept in any row-based process (calculating a Bayesian probability of Type I error).

R. A. Fisher regarded probability theorems involving "psychological tendencies" (Bayesian reasoning) as "useless for scientific purposes" (Fisher, 1947). One wonders if the conditional wording of Type I error reflects this conviction. Authors who define Type I error conditionally may simply be following his practice without intending any claim about Bayesian inference.

This process-based definition makes highly assertive and highly disputable claims about the epistemological status of probability. This claim is much stronger than saying that a classical test of significance simply ignores predictions of the Bayesian probability of error.

This hidden premise attempts to reduce knowledge from being contextual to being intrinsic. To understand the contextual - intrinsic distinction, consider having tossed a fair coin in a situation where no one yet knows the true state of nature about this coin. In reality (metaphysically or intrinsically), the probability of heads is either 0 or 1; but in our minds (epistemologically or contextually) we do not know this reality. So to us the state of the coin is still a random variable with a probability of 0.5 of being heads. Students, in placing counts in both columns in Table 2, are acting as though the state of nature is like the state of this coin: determined in reality but uncertain in our context of knowledge. Their action is consistent with the view that knowledge is contextual – not intrinsic.

In summary, the traditional column-based definition of Type I error hides a highly disputable assertion about the nature of probability. By enclosing this assertion in a definition, students, teachers and even authors may have difficulty recognizing that a highly disputable philosophical argument is being made. Teachers may bypass this problem by referring to alpha as P(reject null | null is true). But this correct outcome hides the difficulties in using the traditional definitions of alpha and Type I error. Bayesians avoid this problem by simply avoiding the use of Type I error in determining P(null is true | null is rejected).

Recommendation
Authors and teachers should abandon the traditional definitions and use definitions that are more general:

- *Define Type I error as an intersection of two logically co-equal conditions:* Type I error occurs whenever the null is true and the null is rejected.
- *Define alpha conditionally as a column-based process:* alpha is the probability of Type I error if the sample is drawn from the null distribution.

This definition of Type I error makes it descriptively neutral rather than being disputably assertive (See Kelley's review of definitions in *The Art of Reasoning*.)

This general approach has several advantages. It presents Type I error simply as a single cell in a 2x2 table. It makes explicit the conditional nature of alpha: P(null will be rejected | null is true). It permits Bayesians to talk about the probability of Type I error given the null is rejected. Most importantly, in terms of Table 1, it should decrease the chance of mistaking alpha (a column percentage) for the Bayesian probability of Type I error (a row percentage).

B. SILENCE ABOUT CONFIDENCE

Within a classical approach, confidence is never mentioned in discussing hypothesis testing. The traditional explanation is that a particular hypothesis describes a state of nature. As such, the hypothesis is either true or false. One has no choice about which distribution one samples from. Among the statistically significant samples, either all or none will result in Type I error.

But suppose students are interested in confidence. In confidence intervals, students were told there is a complementary relation between alpha (the probability of error) and confidence level. This may generate certain expectations in hypothesis testing. And since most texts and teachers are resolutely silent about confidence in hypothesis testing, students presume confidence applies to what they are interested in as decision makers -- the confidence that a decision is correct. Thus, they conclude that an alpha of 5% means they can be 95% confident that a decision to reject the null is correct.

The solution to the problem of silence is to be explicit about the inability of the classical approach to speak of confidence, to present the Bayesian approach and then to present the strengths and weaknesses of each approach. For as Berger (1980) 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.)".

II. BAYESIAN JUSTIFICATION

From a Bayesian perspective, one can evaluate the Bayesian probability of Type I error associated with a classical hypothesis test by following a three step process. The first step involves the use of Bayes rule in comparing the quality and predictive power of medical tests. This step is not controversial so long as each subject can be either diseased or disease free.

The Bayesian approach to medical tests is featured by Ellisor and Morrel in *Statistics for Blood Bankers.* The Bayesian approach to acceptance testing is presented by Moore and McCabe in *Introduction to the Practice of Statistics* and by Neter, Wasserman and Whitmore in *Applied Statistics.* The Bayesian approach to medical tests and acceptance testing is reviewed at length by Hamburg in *Statistical Analysis for Decision Making.*

Step 1: Evaluating Medical Tests on Individuals

Medical tests on individuals can be evaluated using a 2x2 table involving two contradictory states of disease for each individual and two test outcomes.

Table 3: the four cells in a 2x2 table:

CELLS	DISEASE STATUS	
TEST RESULT	Disease-free	Diseased
Negative	OK outcome	Type II Error
Positive	Type I Error	OK Outcome

In Table, 4, the following row probabilities are used:

- δ is the Bayesian probability the subject is disease-free given that the test is positive (Type I error)
- ε is the Bayesian probability the subject is diseased given the test is negative (Type II error).
- γ is the prior probability that the subject is diseased

Let $\delta' = 1 - \delta$, $\varepsilon' = 1 - \varepsilon$, and $\gamma' = 1 - \gamma$.

- δ' is the Bayesian probability that the subject is diseased given that the test is positive. This is called the Positive Predictive Value (PPV).
- ε' is the Bayesian probability that the subject is not diseased given the test is negative. This is called the Negative Predictive Value (NPV).

PPV and NPV are used by Kolins in *Statistics for Blood Bankers* edited by Ellisor and Morel (1983). Kolins references Galen and Gambino (1975) *Beyond Normality* John Wiley and Sons as a primary source.]

Table 4: the quality of a prediction (row percents)

ROW %	DISEASE STATUS		
Outcome	Disease Free	Diseased	
Negative	$\varepsilon' = 1 - \varepsilon$	ε	1
Positive	δ	$\delta' = 1 - \delta$	1
Incidence	$\gamma' = 1 - \gamma$	γ	1

Table 5 illustrates the column probabilities associated with sensitivity and specificity in medical tests. The symbol α is used to identify the probability of a positive test among those who are disease free. At this point this alpha has no relation to the alpha used in classical hypothesis testing. But this choice foreshadows what will come.

Table 5: the quality of a test (column percents)

COL %	DISEASE STATUS	
TEST RESULTS	Disease Free	Diseased
Negative	specificity (α')	β
Positive	α	sensitivity (β')
TOTAL	1	1

When α, β and γ are known, we can generate the counts in a 2x2 table for any test involving N subjects. Note that $\alpha' = 1-\alpha$, $\beta' = 1-\beta$, and $\gamma' = 1-\gamma$.

Table 6: the counts for each cell:

COUNT	SUBJECT STATUS		
	Disease-free	Diseased	
Negative	$\alpha' \gamma' N$	$\beta \gamma N$	$(\alpha'\gamma' + \beta\gamma)N$
Positive	$\alpha \gamma' N$	$\beta' \gamma N$	$(\alpha\gamma' + \beta'\gamma)N$
Prevalence	$\gamma' N$	γN	N

ROW VERSUS COLUMN PROBABILITIES

Row probabilities can be generated given column probabilities using counts in Table 6 or by using Bayes rule: δ = P(Positive & Disease Free) / P(Positive).

$$\delta = \alpha\gamma' / (\alpha\gamma' + \beta'\gamma) \qquad \text{Eq. 1a}$$
$$\varepsilon = \beta\gamma / (\alpha'\gamma' + \beta\gamma) \qquad \text{Eq. 1b}$$

Column probabilities can be generated given row probabilities by solving 1a and 1b for alpha and beta:

$$\alpha = \delta\gamma\beta' \ / \gamma'\delta' \qquad \text{Eq. 2a}$$
$$\beta = \varepsilon\gamma'\alpha' / \gamma\,\varepsilon' \qquad \text{Eq. 2b}$$

In summary, γ is the probability that a random patient has the disease – *prior* to (before) the test. If the patient tests positive, then δ' is the revised probability the patient is diseased – *posterior* to (after) the test. The symbols δ and ε are reversed from those used in Ellisor and Morel's *Statistics for Blood Bankers*. This reversal links the alphabetic sequence (δ and ε) with Type I and 2 errors respectively just like with α and β.

Step 2: Reducing Continuous Hypothesis

The second step in evaluating the Bayesian probability of Type I error in a classical hypothesis test is to reduce a continuous quantitative variable to a point mass. Reducing a null hypothesis from a range (H_O: $\mu \le \mu_O$) to a point mass (H_O: $\mu = \mu_O$) is standard procedure within the classical approach. Specifically, the point mass is situated so as to maximize the associated error.

Step 3: Using Degrees of Belief and States of Nature

The third step is to give prior probabilities about states of nature based on degrees of belief (Bayesian) the same epistemic status as prior probabilities about individual subjects based on relative frequencies (frequentist). This entails treating the existence of both null and alternate as simultaneously possible in thought even though in reality only one is true. It means treating Table 1 as being conceptually similar to Table 2. For more on this very important and highly disputable step, read *Scientific Reasoning* by Howson and Urbach.

Summary:
If we allow degrees of belief about a state of nature, then the dichotomous model used in medical testing can encompass classical hypothesis tests involving a quantitative variable. The Bayesian approach to classical hypothesis testing is mentioned in *Statistical Reasoning* by Smith. It is discussed in *Statistical Decision Theory and Bayesian Analysis* by Berger and in *Statistical Analysis for Decision Making* by Hamburg.

General Case

With a fixed sample size and specific values for alpha and beta, the Bayesian probability of Type I error can be deduced using Eq. 1a for any given value of the prior probability. As the alternate becomes more unlikely, alpha must decrease for a fixed level of Bayesian confidence. As Hamburg (1983) noted: "Prior knowledge concerning the likelihood of truth of the competing hypotheses also helps the investigator in establishing the significance level. Hence, if it is considered likely that the null hypothesis is true, we will tend to set α at a very low figure in order to maintain a low probability of erroneously rejecting that hypothesis."

Beta equals Alpha

If H_o: $\mu \le \mu_o$, H_A: $\mu > \mu_1$ and $\mu_1 > \mu_o$, then by varying the separation ($\mu_1 - \mu_o$) or the sample size (n), one can obtain $\beta = \alpha$ for any value of alpha. If $\beta = \alpha$, then

$$\delta \;=\; \alpha\gamma' / (\alpha\gamma' + \alpha'\gamma) \qquad \text{Eq. 3a}$$
$$\alpha \;=\; \delta\gamma \ / (\gamma'\delta' + \delta\gamma) \qquad \text{Eq. 3b}$$

These relations are illustrated in Figures 1 and 2.

Figure 1: δ as a function of γ for a fixed value of α.

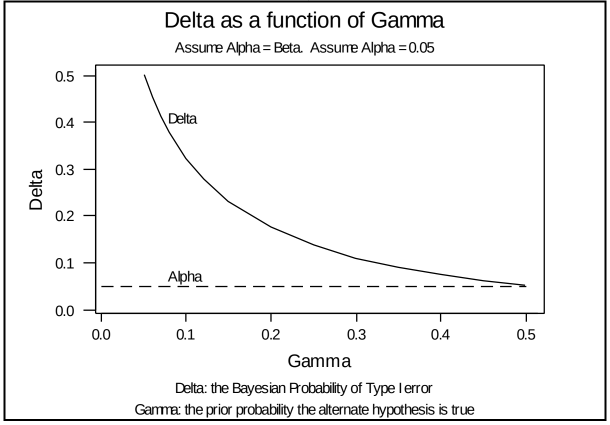
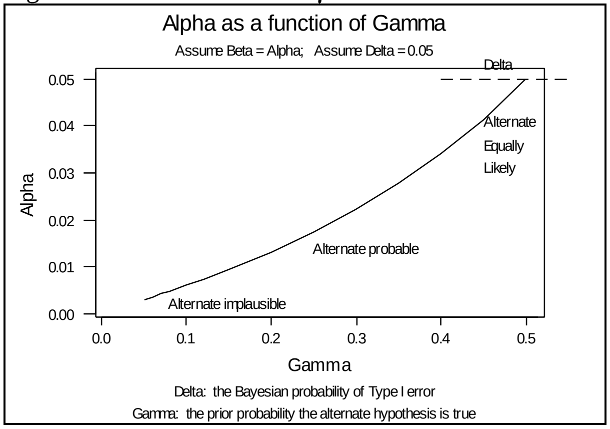


Figure 1 shows that as the alternate becomes more unlikely (γ decreases), the Bayesian probability of Type I error (δ) increases for a fixed classical probability of Type I error (α). Thus the Bayesian confidence (δ') that the alternate is true (given rejection of the null) decreases as γ decreases – for a fixed value of alpha.

One can interpret 1-α as the Bayesian confidence the alternate is true given rejection of the null when the prior probability of a false null is 50% ($\gamma = 0.5$). In this case $\delta = \alpha$ and thus $\delta' = \alpha'$. This was noted by 1965 by John Pratt in *Bayesian Interpretation of Standard Inference Statements.* "Now consider a problem involving two acts and two simple hypotheses H_O and H_1. (Classification problems are sometimes of this type and all problems of this type can be restated as classification problems.) Any procedure in such a problem may be regarded as a test of the null hypothesis H_O; then α is the probability, given H_O, of taking the less desirable action [rejecting H_O]. Let β be the probability, given H_1, of taking the less desirable action [rejecting H_1]. This problem can be viewed as a choice among the available (α,β)-points. If the loss attributable to taking the wrong action is the same for both kinds of error, then the orthodox framework rather suggests choosing that test with α=β among the admissible (here essentially also most powerful and likelihood-ratio) tests. This will coincide with the Bayesian procedure if H_O and H_1 are equally likely *a priori* and are suitably symmetric with respect to one another so that the admissible (α,β)-curve is symmetric in α and β." [Section 8.4; underscore added]

Figure 2 shows how alpha should be decreased as the plausibility of the alternate (gamma) decreases in order to maintain a fixed level of Bayesian confidence.

Figure 2: α as a function of γ for a fixed value of δ



In certain cases, Equation 3a and 3b can be simplified:

$\delta \cong \alpha(\gamma'/\gamma)$ when $\alpha << 1$ Eq. 4a

$\alpha \cong \delta(\gamma/\gamma')$ when $\delta << 1$ Eq. 4b

Beta = 1 - Alpha

If H_o: $\mu \le \mu_o$ and H_a: $\mu > \mu_o$, then $\beta = 1 - \alpha = \alpha'$ regardless of sample size. If $\beta = \alpha'$, then

$\delta = \gamma' = 1 - \gamma$ and $\delta' = \gamma$ for all α Eq. 5

Equation 5 is similar to equation 4a in that δ increases as γ decreases. The difference is that the magnitude of the Bayesian probability of Type I error is independent of the value of alpha. In this case, decreasing the value of alpha has no effect on the Bayesian probability of Type I error! This result may seem counter-intuitive. But one should not expect that decreasing the value of alpha will overcome the difficulty in distinguishing between two hypothesis which are separated only infinitesimally ($\mu = \mu_o$ vs. $\mu > \mu_o$). One might say that such tests are worthless since they give no new knowledge from a Bayesian perspective. A better alternative is to treat this situation as though one were testing an alternate with $\beta = \alpha$ and evaluate the Type I error using the results derived in the previous section.

Objections

Frequentists may object to including any Bayesian reasoning since frequency is "objective" while the probability of an alternate hypothesis being true is generally "subjective". But the choice of alpha is not "objective" Having social agreement on the significance associated with certain values of alpha does not give alpha the same objectivity as has the p-value (which is determined by the data and the Null).

Many might argue that fixed-level hypothesis tests are outdated and one should simply report p-values. But a p-value functions as a sample-dependent alpha if we use the sample statistic to reject the null. As such all the comments made about alpha in this paper apply with equal force to p-values. If β = p-value, then P(H_A is true | p-value of statistic) = $p'\gamma / (p\gamma' + p'\gamma)$ where p indicates the p-value, and $p' = 1 - p$. If $\beta = p$ and $\gamma = 0.5$, then P(H_A is true | p-value) = 1 - p.

Bayesians might argue that one should simply bypass the classical test of significance and use a full-fledged Bayesian approach. This approach ignores the stature of classical hypothesis tests and ignores the goal of this paper which is to help users of classical tests better understand the meaning of alpha and p-value.

III. CONCLUSIONS

Users of classical tests must realize that the value of alpha is not “objective” and should consider setting a value for alpha based on Bayesian inference. Alpha should decrease as the alternate becomes less likely.

Users of classical tests must realize that very small p-values are not necessarily stronger evidence for the truth of the alternate hypothesis.

Thesis advisors and reviewers of journal articles should apply these principles in accepting statistical support for highly disputable conclusions. If alpha is not adjusted for the likelihood of the alternate, then statistical opportunists will be encouraged to use classical hypothesis testing as evidence in support of unrepeatable findings and the credibility of classical hypothesis testing as a scientific method will suffer.

In teaching students about classical hypothesis tests, we want them to understand the meaning of alpha and p-value. If using Bayes rule in classical hypothesis tests helps students better understand the meaning of alpha and p-value, then this approach should be of value to both Bayesians and non-Bayesians.

APPENDIX

The following quotes concern Type I error and alpha. The underscore and bold are added for emphasis.

“A Type I error rejects a null hypothesis **that** is true.” “α = P[Type I error] = P[reject H_O | H_O is true].” Page 362 in *Statistical Reasoning* by Smith.

“A Type I error is rejecting a null hypothesis **that** is true. The probability of committing a Type I error is denoted as α.” “The level of significance is α, the probability of making a Type I error.” Pages 326 and 328 in *Exploring Statistics* by Kitchens.

“Type I error .. is .. rejecting H_O **when** H_O is true.” “The probability of making a Type I error … is called the level of significance and is denoted by α.” Page 302 in *A Data-based Approach to Statistics* by Iman.

“.. Type I error, reject H_O **when** H_O is true.” “..the probability of the Type I error is denoted by α..” Page 324 and 326 in *Probability and Statistical Inference* by Hogg and Tanis.

“**If** we reject H_O (accept H_A) **when** in fact H_O is true, this a Type I error.” “The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject the null hypothesis H_O when H_O is in fact true.” Pages 482 and 484 in *Introduction to the Practice of Statistics* by Moore and McCabe. Pages 388 and 390 in *The Basic Practice of Statistics* by Moore.

“A Type I error is made **if** conclusion H_1 is selected as being correct **when**, in fact, H_O is the correct conclusion.” “The probability of Type I error will be denoted by α …”. Pages 259 and 266 in *Applied Statistics* by Neter Wasserman and Whitmore.

“A Type I error is made **if** H_O is rejected **when** H_O is true. The probability of a type I error is denoted by α.” Page 374 in *Mathematical Statistics* by Mendenhall, Sheaffer and Wackerly.

REFERENCES

Berger, James O. (1980). *Statistical Decision Theory and Bayesian Analysis* 2nd Ed. Springer-Verlag. p. 120

Ellisor, Sandra and Phyllis Morel (1983). *Statistics for Blood Bankers* American Association of Blood Banks.

Fisher, R. A. (1947). *Design of Experiments*, 4th ed. p 6-7. Edinburgh: Oliver and Boyd. First published in 1926.

Hamburg, Morris (1983). *Statistical Analysis for Decision Making*, 3rd Ed. Harcout Brace Janovich, Publ.

Howson, Colin and Peter Urbach, (1993). *Scientific Reasoning* 2nd Ed. Open Court Publishing 1992.

Kelley, David (1994). *The Art of Reasoning*. 2nd Ed.

Moore, David and George McCabe (1993). *Introduction to the Practice of Statistics* 2nd Ed. p.481.

Neter, Wasserman and Whitmore (1978). *Applied Statistics* 2nd Ed. Allyn and Bacon. ISBN 0-205-05982-1.

Pratt, John W. (1965). *Bayesian Interpretation of Standard Inference Statements.* Journal of the Royal Statistical Society (Ser B) 27, 169-203. Section 8.4

Smith, Gary (1985). *Statistical Reasoning.* Allyn and Bacon. ISBN 0-205-11274-9 hc.

ACKNOWLEDGMENTS

Gerald Kaminski prompted Part I of this paper; Linda Marcu, an Augsburg College student, prompted Part II.

Gerald Kaminski made many cogent comments about this paper. Dr. Julie Naylor encouraged additional research on Part II. Margaret Helgeson provided *Statistics for Blood Bankers.*

A draft of Part II of this paper (*Selecting Alpha in Classical Hypothesis Testing using Bayesian Inference)* was presented in Sydney at SISC-96 in July.

Dr. Schield may be reached at schield@augsburg.edu