Beyond p-Values: Hypothesis Testing based on the BIC Model Selection Criterion

Robert F. Bordley* Stanley L. Sclove†

Abstract

In view of recent mounting criticism concerning over-use of p-values, first in the journal *Basic and Applied Social Psychology,* followed by statements by the American Statistical Association and the American Psychological Association, this paper discusses an alternative approach to hypothesis testing, via the BIC model selection criterion.

We do not necessarily advocate the complete abandonment of the p-value approach but rather ask that a logical framework for hypothesis testing and model selection be employed.

In this paper, first, a logical background for hypothesis testing from the viewpoint of Decision Risk Analysis is reviewed.

Then an approach via the BIC model-selection criterion is presented. The earlier AIC (Akaike Information Criterion) is reviewed and compared with BIC.

Those p-values corresponding to some BIC-based tests are computed, with a view toward seeing how much or how little sense the various levels of p-values make. For specific illustration, examples include one- and two-sample tests in Normal distributions, although obviously the methods can be applied in other situations as well.

Key Words: Decision risk analysis, hypothesis testing, p-value, Bayesian information criterion (BIC)

1. Introduction and Background

1.1 Organization

This paper reviews some recent criticism of over-use of p-values, reviews some hypothesis testing situations and the decision-risk analysis approach to them, and discusses an alternative approach based on model-selection criteria.

At the end of the paper, to facilitate reading the paper, there is, for reference, an Appendix listing the notations used.

1.2 Introduction

Hypothesis testing involves choice of null and alternative parameter values, Type I and Type II error rates, and the sample size. The Type I error rate is called $alpha$ and denoted by α. If the **achieved** level of significance p is less than the **prescribed** level of significance α, the null hypothesis is rejected. In recent years there have been several articles critical of the conventional use of p-values in hypothesis testing. There was an article in the journal *Basic and Applied Social Psychology,* followed by statements by the American Psychological Association and the American Statistical Association, concerning over-exclusive use of p-values.

*Booz Allen Hamilton, Troy, MI 48084,

†Department of Information & Decision Sciences, University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607-7124. Writing of this paper was initiated on April 14, 2017

Here are the bibliographic references to some of those recent papers (in chronological order):

David Trafimow and Michael Marks (2015), Editorial, *Basic and Applied Social Psychology,* **37(1),** 1-2. Taylor & Francis (Routledge), Philadelphia, PA.

American Psychological Association. "P-values Under Question." *Psychological Science Agenda,* March, 2016.

Ronald L. Wasserstein and Nicole A. Lazar (2016). The ASAs statement on p-values: context, process, and purpose. *The American Statistician,* **Vol. 70, No. 2,** 120 – 133. June 9, 2016.

Scott Nestler and Harrison Schramm (2016). P-value Primer: P OR (P-values in operations research) M N O P Q R S T You don't need a license to download R, but you should have a good understanding of p-values. *OR/MS Today,* **Vol. 43,** June, 2016.

Some of the best succinct advice we have seen on choosing p is that of Rupert Miller (1966), early in his excellent book: Do the test at the .01 level if it gives reasonable power against reasonable alternatives; otherwise, drop back to a p of .05 or even .10.

Here, in this paper, joint choice of sample size and error rates is reviewed. The situation is discussed from the viewpoint of Decision Risk Analysis, in terms of loss and risk functions. Further, an alternative approach to testing, via model selection criteria, is discussed. In view of the recent criticism of p values, those p-values corresponding to tests based on model selection criteria, especially AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion), are evaluated, with a view toward seeing how much or how little sense the p-values make in that context.

The criterion now called AIC developed by Akaike (1973, 1974) for looking at situations with many variables and a large number of alternative models.

As an example that is related to risk analysis, Sclove (2017) compared simultaneously several different types of models for a dataset. The data was on number of days ill in a year for n persons but in a risk-analysis, actuarial context could be considered alternatively as numbers of accidents in a year for n persons. The alternative models included

- histograms with different numbers of bins,
- clustering by the finite mixture model, and
- a predictive distribution (marginal distribution of number of events, integrating out the prior on the rate parameter λ), which can be considered as an infinite mixture model.

The set-up is that there are K alternative models, indexed by say, $k = 1, 2, \ldots, K$, and we want to rank them and choose the best among them. Let m_k be the number of free parameters in Model k and LL_k the maximized log likelihood for Model k. LL_k is a measure of goodness-of-fit (GOF) of Model k.

The criteria AIC and BIC as often stated, and as we state them here, are *smaller-is-better* criteria. AIC is given by

$$AIC_k = -2LL_k + 2\,m_k.$$

BIC is

$$BIC_k = -2LL_k + (\ln n)m_k.$$

Both of these model-selection criteria (MSCs) are of the form

$$MSC_k = LOF_k + Penalty_k,$$

where $LOF_k = -2\,LL_k$ is the Lack of Fit of Model k and the Penalty is $Penalty(m_k, n) = a(n)\,m_k$, where m_k is the number of free parameters in Model k and $a(n) = \ln n$ for BIC (note that for $n \geq 1, \ln n > 0$) and $a(n) = 2$ for AIC.

BIC is derived from an approximation (Schwarz 1978) to the posterior probability pp_k of Model k,

$$\ln pp_k \;=\; LL_k \;-\; (\ln n)\,m_k/2.$$

These posterior probabilities play an important role in the developments discussed in this paper.

Here, however, we look at some smaller problems, too. One may consider the hypothesis testing problem as a special case with the number K of alternative models being just $K = 2$. Examples here include one- and two-sample tests in Normal distributions. Other examples would include log Normal distributions and negative Exponential distributions for positive random variables, and choosing a subset of explanatory variables for use in a regression model.

2. Test on a Single Mean

Although much of our focus will be on comparing two samples and larger problems, we begin with a one-sample problem. The simplest hypothesis testing problem concerning a parameter θ can be phrased as testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$. about a scalar parameter θ, . where θ_0 and θ_1 are specified values.

Consider the case where θ is the mean μ, H_0: $\mu = \mu_0$, H_1: $\mu = \mu_1$. When the values of μ_0, μ_1 and the variance σ^2 are specified, the test statistic is $z \;=\; (\mu_1 - \mu_0)/\sigma_{\bar{y}}$, where $\sigma^2_{\bar{y}} \;=\; \sigma^2/n$. Assume that $\mu_1 > \mu_0$. Then one rejects the null hypothesis for large z.

2.1 Computation of required n, in terms of α, β and Effect Size

The required minimum sample size is given by

$$n \;=\; \sigma^2\,\frac{(z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2} \;=\; \frac{(z_\alpha + z_\beta)^2}{ES^2}.$$

Here ES is the *effect size* ES = $(\mu_1 - \mu_0)/\sigma$, and z_p is the upper p-th percentage point of the standard Normal distribution, that is, if $\Phi(z)$ is the cumulative distribution function (c.d.f.) of the standard Normal distribution, then $\Phi(z_p) = 1 - p$.

This formula for the required sample size n holds in the Gaussian case and at least approximately in other cases, due to the Central Limit Theorem, provided that n is not small. (The formula could be re-computed to apply to Student's t when the variance is not specified and so is estimated.)

2.2 Choice of error rates and sample size

But where should the Type I and Type II error rates α and β come from? Intelligent choice of a level of significance α and power $1 - \beta$ would involve several factors. One factor is that n, α and β should be decided upon jointly and appropriately. The above formula for n in terms of α and β can be turned around to solve for β given n, α or for α, given n and β.

We have

$$(z_\alpha + z_\beta)^2 \;=\; n(\mu_1 - \mu_0)^2/\sigma^2$$

and, for $\mu_1 - \mu_0, z_\alpha$ and z_β positive, (that is, α and β greater than one-half,)

$$z_\alpha + z_\beta \;=\; \sqrt{n}\,(\mu_1 - \mu_0)/\sigma,$$
$$z_\beta \;=\; \sqrt{n}\,(\mu_1 - \mu_0)/\sigma - z_\alpha,$$

2.3 Tabulation of Power by $n, \alpha,$ and Effect Size

We have, remembering that the power is the complement of the Type II-error rate β,

$$\text{Power } = 1 - \beta = \Phi(z_\beta) \;=\; \Phi(\sqrt{n}\,\frac{\mu_1 - \mu_0}{\sigma} \;- z_\alpha) \;=\; \Phi(\sqrt{n}\,ES \;- z_\alpha).$$

For example, if ES = 0.5, $\alpha = .05$, and $n = 25$, then $\sqrt{n}ES - z_\alpha = 5(0.5) - 1.645 = 2.500 - 1.645 = \; 0.855$ and Power $= 1 - \beta = \Phi(0.855) \approx .804$. The power is tabulated in Table 1 for several combinations of ES, n and alpha.

Table 1: Power, in terms of ES, n and α

ES	n	α	ES$\sqrt{n} - z_\alpha$	Power	ES	n	α	ES$\sqrt{n} - z_\alpha$	Power
0.5	9	0.01	-0.83	0.20	1.5	9	0.01	2.17	1.00
0.5	9	0.05	-0.14	0.44	1.5	9	0.05	2.86	1.00
0.5	9	0.1	0.22	0.59	1.5	9	0.1	3.22	1.00
0.5	25	0.01	0.17	0.57	1.5	25	0.01	5.17	1.00
0.5	25	0.05	0.86	0.80	1.5	25	0.05	5.86	1.00
0.5	25	0.1	1.22	0.89	1.5	25	0.1	6.22	1.00
0.5	100	0.01	2.67	1.00	1.5	100	0.01	12.67	1.00
0.5	100	0.05	3.36	1.00	1.5	100	0.05	13.36	1.00
0.5	100	0.1	3.72	1.00	1.5	100	0.1	13.72	1.00
1.0	9	0.01	0.67	0.75	2.0	9	0.01	3.67	1.00
1.0	9	0.05	1.36	0.91	2.0	9	0.05	4.36	1.00
1.0	9	0.1	1.72	0.96	2.0	9	0.1	4.72	1.00
1.0	25	0.01	2.67	1.00	2.0	25	0.01	7.67	1.00
1.0	25	0.05	3.36	1.00	2.0	25	0.05	8.36	1.00
1.0	25	0.1	3.72	1.00	2.0	25	0.1	8.72	1.00
1.0	100	0.01	7.67	1.00	2.0	100	0.01	17.67	1.00
1.0	100	0.05	8.36	1.00	2.0	100	0.05	18.36	1.00
1.0	100	0.1	8.72	1.00	2.0	100	0.1	18.72	2.00

Perhaps more to the point of the present discussion is a tabulation of alpha in terms of ES, n and power, to see what alpha might be reasonable in terms of the other quantities, rather than looking rather blindly toward an alpha of five percent.

The above formula for n in terms of α and β can be turned around to solve for α given n and β. We have

$$(z_\alpha + z_\beta)^2 \;=\; n(\mu_1 - \mu_0)^2/\sigma^2$$

and, for $\mu_1 - \mu_0, z_\alpha$ and z_β positive,

$$z_\alpha + z_\beta \;=\; \sqrt{n}\,(\mu_1 - \mu_0)/\sigma \;=\; \sqrt{n}\,ES,$$

$$z_\alpha \;=\; \sqrt{n}\,ES \;- z_\beta,$$

$$\Phi^{-1}(1-\alpha) \;=\; \sqrt{n}ES \;- z_\beta,$$

$$1 - \alpha \;= \Phi(\sqrt{n}ES \;- z_\beta).$$

$$\alpha \;=\; 1 - \Phi(\sqrt{n}ES \;- z_\beta).$$

Table 2: Level α, in terms of ES, n and β

ES	n	β	ES$\sqrt{n} - z_\beta$	α	ES	n	β	ES$\sqrt{n} - z_\beta$	α
0.5	9	0.05	-0.145	0.56	1.5	9	0.05	2.855	0.00
0.5	9	0.10	0.218	0.41	1.5	9	0.10	3.218	0.00
0.5	9	0.20	0.658	0.26	1.5	9	0.20	3.658	0.00
0.5	25	0.05	0.855	0.20	1.5	25	0.05	5.855	0.00
0.5	25	0.10	1.218	0.11	1.5	25	0.10	6.218	0.00
0.5	25	0.20	1.658	0.05	1.5	25	0.20	6.658	0.00
0.5	100	0.05	3.355	0.00	1.5	100	0.05	13.355	0.00
0.5	100	0.10	3.718	0.00	1.5	100	0.10	13.718	0.00
0.5	100	0.20	4.158	0.00	1.5	100	0.20	14.158	0.00
1.0	9	0.05	1.355	0.09	2.0	9	0.05	4.355	0.00
1.0	9	0.10	1.718	0.04	2.0	9	0.10	4.718	0.00
1.0	9	0.20	2.158	0.02	2.0	9	0.20	5.158	0.00
1.0	25	0.05	3.355	0.00	2.0	25	0.05	8.355	0.00
1.0	25	0.10	3.718	0.00	2.0	25	0.10	8.718	0.00
1.0	25	0.20	4.158	0.00	2.0	25	0.20	9.158	0.00
1.0	100	0.05	8.355	0.00	2.0	100	0.05	18.355	0.00
1.0	100	0.10	8.718	0.00	2.0	100	0.10	18.718	0.00
1.0	100	0.20	9.158	0.00	2.0	100	0.20	19.158	0.00

In this context, the set of values of ES, n and β for which $\alpha = .05$ would be appropriate is given by

$$.05 = \ \alpha \ = \ 1 - \Phi(\sqrt{n}\, ES \ - z_\beta).$$

For example, when $\beta = .20$ (power = .80), $z_\beta = 0.842$, and

$$.05 = \ \alpha \ = \ 1 - \Phi(\sqrt{n}\, ES \ - 0.842).$$

This gives

$$\Phi(\sqrt{n}\, ES \ - 0.842) \ = \ .95$$

or

$$\sqrt{n}\, ES \ - 0.842 \ = \ \Phi^{-1}(.95) = 1.645,$$

or

$$\sqrt{n}\, ES \ = \ \Phi^{-1}(.95) = 2.487.$$

If ES =1.0, $\sqrt{n} = 2.487$, and the n needed is about 6. If ES = 0.5, $\sqrt{n} = 2(2.487) = 4.974$. and n is about 25. If ES = 0.25, $\sqrt{n} = 4(2.487) = 9.958$ and n is about 100.

2.4 Approach via Decision Risk Analysis

But where should α and β come from? But should the procedure on even be centered on these error probabilities?

Presumably, the prescribed level of significance α is set to a very small probability because the losses associated with a Type I error are greater than those associated with a Type II error. A rational way to make a decision (accept or reject H_0) between the two models, without pre-specifying the error rates, would be by means of decision risk analysis, minimizing posterior expected loss. This approach presupposes specifying prior probs π_0 and π_1 for the two states H_0 and H_1 and a *loss function* $L(a, s)$ giving the loss for each

action a and state of nature, s. For a hypothesis testing problem, the actions a are to accept or reject the null hypothesis, and the states of nature are that the hypothesis is true or that the hypothesis is false.

Let us discuss the approach via decision risk analysis in more detail. An intelligent choice of a level of significance α would involve several factors. One is that n, α and β need to be decided upon jointly and intelligently.

The intelligent choice of how to do a test is illuminated by some concepts from decision risk analysis. This is illustrated below.

The simplest hypothesis testing problem concerning a parameter θ, phrased as testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, can be stated in terms of a two state, two action decision problem, to be based on data represented as a sample vector $\boldsymbol{y}$.

But first take a look at the no-data decision problem. Assume the following set-up. Later, given data, Action 0 will correspond to accepting the null hypothesis; action 1, to rejecting it. The *loss* of taking action a under state s is denoted by $L(a \,|\, s)$.

Table 3: Loss Function, $L(a|s)$

	State of Nature	
	0	1
Prior probabilities	π_0	π_1
Action 0 (accept)	$L(0\|0)$	$L(0\|1)$
Action 1 (reject)	$L(1\|0)$	$L(1\|1)$

Typically, there will be gains instead of losses when the correct decision is made, so $L(0|0)$ and $L(1|1)$ are negative. Also, the decision is obviously to take action 0 if $L(0|0) < L(1|0)$ and $L(0|1) < L(1|\,1)$, and the decision is obviously to take action 1 if $L(1|0) < L(0|0)$ and $L(1|1) < L(0\,|\,1)$, so we assume that neither of these conditions hold.

The way to make a decision is to minimize expected loss, that is, to choose the action that has the smaller expected loss. The expected loss of Action a given state s is

$$\mathcal{E}[L(a|s)] \;=\; \pi_0\, L(a|0) \,+\, \pi_1\, L(a|1).$$

Each of the following is equivalent to the others.

$$\begin{aligned} \text{Action 0} \quad &\text{is} \quad \text{better than Action 1} \\ \mathcal{E}[\,L(0|s)\,] \quad &< \quad \mathcal{E}[\,L(1|s)\,] \\ \pi_0\, L(0|0) \,+\, \pi_1\, L(0|1) \quad &< \quad \pi_0\, L(1|0) \,+\, \pi_1\, L(1|1). \\ \pi_0\,[L(0|0) \,-\, L(1|0)] \quad &< \quad \pi_1\,[L(1|1) \,-\, L(0|1)] \end{aligned}$$

The difference $L(1|0) - L(0|0)$ is the *regret* when the state is 0. It is the amount by which the loss of the incorrect decision exceeds that of the correct decision when the state is 0. It may be denoted by $R(\cdot|\,0)$. The difference $L(0|1) - L(1|1)$ is the amount by which the loss of the incorrect decision exceeds that of the correct decision when the state is 1. It is the regret when the state is 1 and may be denoted by $R(\,\cdot\,|\,1)$. Continuing, we have

$$\begin{aligned} \pi_0\,[L(0|0) \,-\, L(1|0)] \quad &< \quad \pi_1\,[L(1|1) \,-\, L(0|1)] \\ \pi_0\, R(\,\cdot\,|\,0\,) \quad &< \quad \pi_1\, R(\,\cdot\,|\,1) \\ \pi_1\,/\,\pi_0 \quad &> \quad R(\,\cdot\,|\,0\,)\,/\,R(\,\cdot\,|\,1) \end{aligned}$$

Given data, the prior probabilites are replaced by posterior probabilities $\Pr(0|\boldsymbol{x})$ and $\Pr(1|\boldsymbol{x})$. The posterior expected losses are compared. The expected posterior loss of Action a given state s is

$$\mathcal{E}[L(a|s)\,|\,\boldsymbol{x}] \;=\; \Pr(0\,|\,\boldsymbol{x})\,L(a|0) \;+\; \Pr(1|\boldsymbol{x}\,)\,L(a|1).$$

The optimal rejection region can be described in terms of a set of values of $\boldsymbol{x}$ where the expected posterior loss of action 1 is less than that of action 0. The above inequality becomes: take action 1 if

$$\Pr(\text{State}\,1\,|\,\boldsymbol{x}) \,/\, \Pr\{\text{State}\,0\,|\,\boldsymbol{x}) \;>\; R(\,\cdot\,|\,0\,) \,/\, R(\,\cdot\,|\,1).$$

This is equivalent to

$$\pi_1\, f_1(\boldsymbol{x}) \,/\, \pi_0\, f_0(\boldsymbol{x}) \;>\; R(\,\cdot\,|\,0\,) \,/\, R(\,\cdot\,|\,1).$$

This can be written as

$$f_1(\boldsymbol{x}) \,/\, f_0(\boldsymbol{x}) \;>\; \pi_0\, R(\,\cdot\,|\,0\,) \,/\, \pi_1 R(\,\cdot\,|\,1).$$

The LHS, $f_1(\boldsymbol{x}) \,/\, f_0(\boldsymbol{x})$, is the *likelihood ratio.* Often it reduces to a function of sufficient statistics, such as the sample mean and variance in the Gaussian case. If the null distribution is $\mathcal{N}(\mu_0, \sigma^2)$ and the alternative distribution is $\mathcal{N}(\mu_1, \sigma^2$, then the LR $f_1(\boldsymbol{x})/f_0(\boldsymbol{x})$ reduces to

$$\ln LR \;=\; \ln f_1(\boldsymbol{x}) \,/\, f_0(\boldsymbol{x}) \;=\; n\frac{\mu_1 - \mu_0}{\sigma^2}(\bar{x} \,-\, \frac{\mu_0 + \mu_1}{2}).$$

The probability content of the region where the LR exceeds the RHS above in the space of $\boldsymbol{x}$ under the null hypothesis (that is, in State 0) will be the optimal α. Note that it depends upon the prior probabilities and the four loss values, through the ratio of regrets. The rejection region is $\ln LR \;>\; \ln \pi_0\, R(\,\cdot\,|\,0\,) \,/\, \pi_1 R(\,\cdot\,|\,1)$. This is $n\frac{\mu_1 - \mu_0}{\sigma^2}\,(\bar{x} \,-\, \frac{\mu_0 + \mu_1}{2}) \;> \ln C$, where

$$C \;=\; \pi_0\, R(\,\cdot\,|\,0\,) \,/\, \pi_1 R(\,\cdot\,|\,1).$$

Let us call C the *posterior regret ratio*.

Now, this rejection region reduces to the set of samples $\boldsymbol{x}$ such that the mean $barx$ satisfies

$$\bar{x} \;>\; \mu_0 + \;\frac{\mu_1 - \mu_0}{2} + \frac{\sigma^2}{n(\mu_1 - \mu_0)}\;\ln C.$$

2.5 Optimal α

Now, the rejection region is the set of samples $\boldsymbol{x}$ such that

$$\bar{x} \;-\; \mu_0 > \;\frac{\mu_1 - \mu_0}{2} + \frac{\sigma^2}{n(\mu_1 - \mu_0)}\;\ln C.$$

The optimal alpha is the probability that this inequality holds, where the probability is computed according to the null distribution given by $f_0(\boldsymbol{x})$. The above inequality is equivalent to

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \;>\; \frac{(\mu_1 - \mu_0)/2}{\sigma/\sqrt{n}} + \frac{\sigma}{\sqrt{n}(\mu_1 - \mu_0)}\;\ln C.$$

Note that the RHS is $\sqrt{n}\,ES/2 \;+\; (\ln C) \,/\, \sqrt{n}\,ES$, where $ES \;=\; (\mu_1 - \mu_0)/\sigma$. This is the value of z for the optimal α.

Table 4 lists values of the optimal α corresponding to some values of n, ES and C, where $C \;=\; \pi_0\, R(\,\cdot\,|\,0\,) \,/\, \pi_1 R(\,\cdot\,|\,1)$. The values of $z_{\alpha *}$ can range over the whole axis, so

Table 4: Optimal αs for some values of n, ES and posterior regret ratio $C = \pi_0\, R(\cdot\,|\,0\,) \,/\, \pi_1 R(\cdot\,|\,1)$.

ES	n	C	$z(\alpha)$	α	ES	n	C	$z(\alpha)$	α
0.5	9	0.50	-0.290	0.614	1.5	9	0.50	-0.869	0.808
0.5	9	1.00	0.750	0.227	1.5	9	1.00	2.250	0.012
0.5	9	2.00	1.790	0.037	1.5	9	2.00	5.369	0.000
0.5	25	0.50	-0.483	0.685	1.5	25	0.50	-1.449	0.926
0.5	25	1.00	1.250	0.106	1.5	25	1.00	3.750	0.000
0.5	25	2.00	2.983	0.001	1.5	25	2.00	8.949	0.000
0.5	100	0.50	-0.966	0.833	1.5	100	0.50	-2.897	0.998
0.5	100	1.00	2.500	0.006	1.5	100	1.00	7.500	0.000
0.5	100	2.00	5.966	0.000	1.5	100	2.00	17.897	0.000
1.0	9	0.50	-0.579	0.719	2.0	9	0.50	-1.159	0.877
1.0	9	1.00	1.500	0.067	2.0	9	1.00	3.000	0.001
1.0	9	2.00	3.579	0.000	2.0	9	2.00	7.159	0.000
1.0	25	0.50	-0.966	0.833	2.0	25	0.50	-1.931	0.973
1.0	25	1.00	2.500	0.006	2.0	25	1.00	5.000	0.000
1.0	25	2.00	5.966	0.000	2.0	25	2.00	11.931	0.000
1.0	100	0.50	-1.931	0.973	2.0	100	0.50	-3.863	1.000
1.0	100	1.00	5.000	0.000	2.0	100	1.00	10.000	0.000
1.0	100	2.00	11.931	0.000	2.0	100	2.00	23.863	0.000

$\alpha*$ can be anywhere between 0 and 1. Note that $C = 1$, corresponding for example to equal prior probabilities $\pi_0 = \pi_1 = 1/2$ and equal regrets $R(\cdot\,|\,0\,) = R(\cdot\,|\,1)$, may be in some sense considered a "neutral" value of C.

What would be the values corresponding to .05? Now, $.05 = \Pr\{z > 1.645\}$, that is, the 95th percentile of the standard Normal distribution is 1.645. So the set of input values corresponding to the .05 level can be found (although the point is that there should be nothing special about .05).

2.6 Posterior Probabilities

Now the posterior probabilities of the two models are $pp_0 = \pi_0\, f_0(\boldsymbol{x})/f(\boldsymbol{x})$ and $pp_1 = \pi_1\, f_1(\boldsymbol{x})/f(\boldsymbol{x})$, where

$$f(\boldsymbol{x}) \;=\; \pi_0 f_0(\boldsymbol{x}) \;+\; \pi_1\, f_1(\boldsymbol{x}).$$

That is to say, the decision rule is to take action 1 if

$$\text{posterior odds,}\;\; \Pr(\text{State}\,1\,|\,\boldsymbol{x}) \,/\, \Pr\{\text{State}\,0\,|\,\boldsymbol{x}) \;>\; R(\cdot\,|\,0\,) \,/\, R(\cdot\,|\,1),$$

in short,

$$\text{posterior odds,}\;\; pp_1/pp_0 \;>\; R(\cdot\,|\,0\,) \,/\, R(\cdot\,|\,1).$$

In turn, this can be approximated as

$$\pi_1\, \exp(LL_1)\, n^{m_1/2} \,/\, \pi_0\, \exp(LL_0)\, n^{m_0/2} \;> R(\cdot\,|\,0\,) \,/\, R(\cdot\,|\,1),$$

that is

$$\pi_1\, M_1\, n^{m_1/2} \,/\, \pi_0\, M_0\, n^{m_0/2} > R(\cdot\,|\,0\,) \,/\, R(\cdot\,|\,1),$$

where $M_k = \exp(LL_k)$ is the maximized likelihood of Model k. This can be conveniently considered as

$$(\pi_1\, \pi_0)\, (M_1/M_0)\, n^{(m_1-m_0)/2} \;>\; R(\cdot \,|\, 0\,) \,/\, R(\cdot \,|\, 1),$$

that is,

$$\text{prior odds } \times \text{likelihood ratio } \;\times n^{(m_1-m_0)/2} \;>\; R(\cdot \,|\, 0\,) \,/\, R(\cdot \,|\, 1).$$

The decision-risk-analysis framework obviously extends to more than two actions. Here there is some focus on the case of two actions because it corresponds to that form of hypothesis testing.

Note that a prescribed significance level α (such as .05) is not an input to the problem but rather could be computed as a function of the inputs. Often these inputs would be hard to come by, so another approach, just focusing on the values of model-selection criteria, is also discussed here.

3. Interval Estimation

What might be a model-selection approach to interval estimation (confidence intervals, or Bayesian credibility intervals)?

3.1 Bayesian Credibility Intervals

A **credibility interval** is an interval of plausible parameter values, in the sense that these values have relatively high posterior probability, that is, the set of parameter values θ for which $pp(\theta \,|\, \boldsymbol{x}) > C$. The posterior probability is

$$pp(\theta \,|\, \boldsymbol{x}) \;=\; f_\Theta(\theta)\, f(\boldsymbol{x}|\theta)/f(\boldsymbol{x}),$$

where $f_\Theta(\theta)$ is the p.d.f. of the prior distribution on the parameter Θ. More generally, a **credibility region** for a vector parameter $\boldsymbol{\theta}$ is a set of plausible parameter values, in the sense that these values have relatively high posterior probability, that is, the set of parameter values $\boldsymbol{\theta}$ for which $pp(\boldsymbol{\theta} \,|\, \boldsymbol{x}) > C$. Using the Schwarz approximation, $pp(\boldsymbol{\theta}|\boldsymbol{x}) \;> C$ is approximately $LL(\boldsymbol{\theta}) - \;(\ln n)\, m/2 \;> C$, where m is the number of free parameters estimated. BIC (Schwarz 1978) as defined here is $(\,-2\, LL(\boldsymbol{\theta}) \,+\, (\ln n)\, m)$, where $LL(\boldsymbol{\theta})$ denotes the maximum log likelihood over $\boldsymbol{\theta}$. This is the Schwarz (1978) approximation to $-2\, \log$ posterior probability, so BIC is applicable to the formation of credibility intervals. (That is, the Schwarz approximation ot the posterior probability of a model with m parameters is $pp \;=\; LL(\boldsymbol{\theta}) \,-\, (\ln n)\, m/2$.)

So, large values of posterior probability correspond to small values of BIC as it is expressed here (with the multiplier of -2). That is, a credibility region will be a set of parameter values with sufficiently good (low) BIC. For Gaussian models with error variance σ^2, the statistic $-2LL = \; n \ln \hat{\sigma}^2$; thus the credibility region is the set such that $n \ln \hat{\sigma}^2 \,+$ $(\ln n)\, m$ is sufficiently small. Recall that $n \ln \hat{\sigma}^2$ is the so-called **deviance.** So, in such Gaussian models, the credibility region is equal to the set of parameter values for which the deviance is sufficiently small.

3.2 Confidence Intervals

A *confidence interval* for a mean is a set of values μ_0 that are plausible in the sense that they would be accepted if they were values specified in the null hypothesis. For example, in testing the null hypothesis H_0: $\mu = \mu_0$ against two-sided alternatives, based on a sample $y_1, y_2, \ldots, y_n$ of n from a Normal distribution with specified standard deviation σ. The value μ_0 is accepted if $|z(\mu_0)| \leq z^*$, where $z^* = z_{\alpha/2}$ and z_p denotes the upper p-th percentile of the standard Normal distribution and $z(\mu_0) = (\bar{y} - \mu_0)/\sigma/\sqrt{(n)}..$ This interval of "acceptable" μ_0 is the set $\{\mu_0 : \bar{y} - z^*\sigma/\sqrt{n} < \mu_0 < \bar{y} - z^*\sigma/\sqrt{n}\}$.

The confidence interval comes from the probability

$$\Pr\{\mu_0 - z^*\,\sigma_{\bar{y}} < \bar{Y} < \mu_0 + z^*\sigma_{\bar{y}} \mid \mu = \mu_0\}.$$

Here $\sigma_{\bar{y}}$ is the standard deviation of the mean: $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. If z^* is the upper $\alpha/2$ percentile of the standard Normal distribution, then this probability is $1 - \alpha$. This probability is called the *coverage probability* of the interval because the probability statement above can be rewritten as

$$\Pr\{\bar{Y} - z^*\,\sigma_{\bar{y}} < \mu_0 < \bar{Y} + z^*\sigma_{\bar{y}} \mid \mu = \mu_0\} = 1 - \alpha.$$

So the probability that the interval $(\bar{Y} - z^*\sigma_{\bar{y}},\ \bar{Y} + z^*\sigma_{\bar{y}})$ covers the true parameter value is $1 - \alpha$.

Because a confidence interval is a set of "acceptable" values μ_0, this could, in the model-selection context, be taken to mean a set of values for which the value of the MSC is better than that of the MLE. Since the MLE has the highest value of LL, this means that for a value to be in the confidence interval, the difference MSC for that value minus MSC for the MLE must be less than the penalty amount:

Let $\text{MSC}(\hat{\mu}) = -2LL(\hat{\mu}) + a(n)\,m_k$, where $LL(\hat{\mu})$ is the maximum log likelihood at $\hat{\mu}$, $a(n) = 2$ for AIC and $= \ln n$ for BIC, and m_k is the number of parameters estimated in Model k, remembering that such MSCs have the form MSC = Lack-of-fit + Penalty, where the Lack-of-Fit term is $-2\,LL_k$ and the Penalty for the number of parameters used is of the form $a(n)\,m_k$.

At this point, where we are considering inference about a single mean, we can consider that there are two models, indexed by $k = 0, 1$, 0 for the mean specified to be μ_0 and 1 for the mean to be estimated. With specified variance, the number m_k of parameters estimated is $m_0 = 0$ for μ specified to be μ_0 and $m_1 = 1$ parameter estimated when μ estimated by the MLE, $\bar{y}$. A confidence interval could be defined according to the model-selection criterion MSC as the set of values μ_0 such that

$$\text{MSC}(\mu_0) \leq \text{MSC}(\bar{y}).$$

Recall that for Gaussian models, the likelihood L is

$$L = L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp(-(1/2\sigma^2) \sum (y_i - \mu)^2,$$

as a function of the parameters μ and σ^2. The log likelihood l is

$$l = -(n/2)\,\ln(2\pi) - (n/2)\ln\sigma^2 - (1/2\sigma^2)\,\sum (y_i - \mu)^2.$$

This gives

$$-2\,l = n\ln(2\pi) + n\ln\sigma^2 + \sum (y_i - \mu)^2\,/\sigma^2.$$

Denoting the maximum of this l over σ^2 by LL, that is, with

$$\hat{\mu} = \bar{y} \text{ and } \hat{\sigma}^2 = \sum (y_i - \bar{y})^2/n,$$

we have

$$-2\,LL \;=\; n\ln(2\pi) + n\ln\hat{\sigma}^2 \;+\; n.$$

Under $H_0 : \mu = \mu_0$, the estimate of σ^2 is $\sum(y_i - \mu_0)^2/n$. This gives

$-2\,LL(\mu_0) \;=\; n\ln 2\pi \;+\; n\ln\sum(y_i - \mu_0)^2/n \;+\; n.$

The confidence interval could be taken as the set of μ_0 such that $MSC(\mu_0) \;\leq MSC(\bar{y})$. This is

$$-2\,LL(\mu_0) \;+a(n)m_0 \;\leq -2\,LL(\bar{y}) \;+a(n)m_1.$$

Since in this case $m_0 = 0$ and $m_1 = 1$, this is simply

$$-2\,LL(\mu_0) \;\leq -2\,LL(\bar{y}) \;+a(n).$$

For AIC, $a(n) = 2$ and this is $-2\,LL(\mu_0) \;\leq -2\,LL(\bar{y}) + 2$. For BIC, $\;a(n) = \ln n$, so this criterion for inclusion of μ_0 in the confidence interval is $-2\,LL(\mu_0) \;\leq -2\,LL(\bar{y}) \;+\; \ln n$.

Alternatively, a confidence interval for μ could be taken to be a set of values for which the value of the MSC is within an amount C of that of the MLE (the preceding case being $C = 0$):

$$MSC(\mu_0) \leq MSC(\bar{y}) + \;C.$$

This is

$$-2\,LL(\mu_0) \;\leq -2\,LL(\bar{y}) \;+a(n) + \;C.$$

This becomes

$$n\ln 2\pi + n\ln\textstyle\sum_{i=1}^{n}(y_i - \mu_0)^2/n \;+\; n$$

$$\leq\; n\ln(2\pi) + n\ln\sum(y_i - \bar{y})^2/n \;+\; n + a(n) \;+C.$$

This is

$$n\ln\sum(y_i - \mu_0)^2/n \;\leq n\ln\sum(y_i - \bar{y})^2/n + a(n) \;+\; C.$$

This in turn is

$$n\ln[\textstyle\sum(y_i - \mu_0)^2/\sum(y_i - \bar{y})^2] \;\leq a(n) \;+\; C$$

or

$$\textstyle\sum(y_i - \mu_0)^2/\sum(y_i - \bar{y})^2 \;\leq\; \exp\{[a(n) \;+\; C]/n\}$$

Now, the lefthand side is

$$[\textstyle\sum(y_i - \bar{y})^2 + n(\mu_0 - \bar{y})^2]\,/\,\sum(y_i - \bar{y})^2 \;=\; 1 + \dfrac{(\mu_0 - \bar{y})^2}{(n-1)s^2/n}\,.$$

The inequality becomes

$$1 + \frac{(\mu_0 - \bar{y})^2}{(n-1)s^2/n} \;\leq\; \exp\{[a(n) \;+\; C]/n\}$$

This is

$$\frac{(\mu_0 - \bar{y})^2}{(n-1)s^2/n} \;\leq\; \exp\{[a(n) \;+\; C]/n\} \;-1$$

or

$$(\mu_0 - \bar{y})^2 \;\leq\; [\exp\{[a(n) \;+\; C]/n\} \;-1](n-1)s^2/n.$$

This is

$$|\,\mu_0 - \bar{y}\,| \;\leq\; \sqrt{[\exp\{[a(n) \;+\; C]/n\} \;-1]}\,\sqrt{n-1}\,s/\sqrt{n}.$$

Focusing on μ_0, this is the confidence interval

$$\bar{y} - m.e. \leq \mu_0 \leq \bar{y} + m.e.,$$

where the margin of error m.e. is

$$m.e. = \sqrt{[\exp\{[a(n) \ + \ C]/n\} \ - 1]}\, \sqrt{n-1}\, s/\sqrt{n}.$$

This is

$$m.e. \ = \ \exp\{(1/2)[a(n) \ + \ C]/n\} \ - 1]\, \sqrt{n-1}\, s/\sqrt{n}.$$

This compares with the m.e. $t_{\alpha/2} s/\sqrt{n}$, where t_p denotes the upper p-th percentile of the Student's t distribution with $n-1$ degrees of freedom, for a $100(1-\alpha)\%$ confidence interval.

4. Posterior Probabilities for K Alternative Models

Approximate posterior probabilities can be obtained from BIC, as BIC is the first terms of an expansion of $-2 \ln pp_k$, where pp_k is the posterior probability of Model k. We have $\text{BIC}_k \ = \ -2\, LL_k + (\ln n)\, m_k$. The posterior probability pp_k is proportional to $\exp(-BIC_k/2)$. If there are prior probabilities π_k, then this is adjusted to $\pi_k \ \exp(-BIC_k/2)$. That is, $pp_k = \ C\, \pi_k \ \exp(-BIC_k/2)$. Thus,

$$1 = \sum_k^K \ pp_k = \ C \sum_k^K \pi_k \exp(-BIC_k/2),$$

so that $C \ = \ 1/\sum_k^K \ \pi_k \ \exp(-BIC_k/2)$.

5. Comparison of Means of Two Gaussian Distributions

Consider the two-sample problem for means μ_1 and μ_2, with null hypothesis $\text{H}_0 : \ \mu_1 = \mu_2$ vs. $\text{H}_{alt} \ : \ \mu_2 > \mu_1$, given samples of sizes n_1 and n_2 from Gaussian distributions with specified common variance σ^2. The test is to reject H_0 if $z > z_\alpha$, where z_α is the upper α-th percentage point of the standard Normal distribution and

$$z \ = \ (\bar{y}_1 - \bar{y}_2)/\sigma_{\bar{y}_1 - \bar{y}_2},$$

where $\sigma^2_{\bar{y}_1 - \bar{y}_2} = \sigma^2\, (1/n_1 + 1/n_2)$. (The concepts discussed here would of course apply to two-sided alternatives as well.)

Given values of the error rates α and β, one can determine the needed sample size from

$$n_1 + n_2 \ = \ \sigma^2\, (z_\alpha + z_\beta)^2/(\mu_1 - \mu_2)^2 \ = \ (z_\alpha + z_\beta)^2/ES^2,$$

where here the effect size ES is given by $\text{ES} \ = (\mu_1 - \mu_2)/\sigma$ and z_p denotes the upper p-th percentile of the standard Normal distribution.

5.1 Approach via Decision Risk Analysis

But where should α and β come from? But should the procedure on even be centered on these error probabilities?

Presumably, the prescribed level of significance α is set to a very small probability to demand a sufficient level of "proof". Beyond that, it could be because the losses associated

with a Type I error are greater than those associated with a Type II error. A rational way to make a decision (accept or reject H_0) between the two models, without pre-specifying the error rates, would be by means of decision risk analysis, minimizing posterior expected loss. This approach presupposes specifying prior probs π_0 and π_1 for the two states H_0 and H_1 and a *loss function* $L(a, s)$ giving the loss for each action a and state of nature, s. For a hypothesis testing problem, the actions a are to accept or reject the null hypothesis, and the states of nature are that the hypothesis is true or that the hypothesis is false.

Let us discuss the approach via decision risk analysis in more detail. For another thing, intelligent choice of a level of significance α would involve several factors. One is that n, α and β need to be decided upon jointly and intelligently.

The intelligent choice of how to do a test is illuminated by some concepts from decision risk analysis. This is illustrated below.

The simplest hypothesis testing problem concerning a parameter θ, phrased as testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, can be stated in terms of a two state, two action decision problem, to be based on data represented as a sample vector $\boldsymbol{y}$.

But first take a look at the no-data decision problem. Assume the following set-up. Later, given data, Action 0 will correspond to accepting the null hypothesis; action 1, to rejecting it. The *loss* of taking action a under state s is denoted by $L(a \mid s)$.

Table 5: Loss Function, $L(a|s)$

	State of Nature	
	0	1
Prior probabilities	π_0	π_1
Action 0 (accept)	$L(0\|0)$	$L(0\|1)$
Action 1 (reject)	$L(1\|0)$	$L(1\|1)$

Typically, there will be gains instead of losses when the correct decision is made, so $L(0|0)$ and $L(1|1)$ are negative. Also, the decision is obviously to take action 0 if $L(0|0) < L(1|0)$ and $L(0|1) < L(1|1)$, and the decision is obviously to take action 1 if $L(1|0) < L(0|0)$ and $L(1|1) < L(0|1)$, so we assume that neither of these conditions hold.

The way to make a decision is to minimize expected loss, that is, to choose the action that has the smaller expected loss. The expected loss of Action a given state s is

$$\mathcal{E}[L(a|s)] \;=\; \pi_0\, L(a|0) \;+\; \pi_1\, L(a|1).$$

Each of the following is equivalent to the others.

$$\begin{aligned} \text{Action 0} \quad &\text{is} \quad \text{better than Action 1} \\ \mathcal{E}[\,L(0|s)\,] \quad &< \quad \mathcal{E}[\,L(1|s)\,] \\ \pi_0\, L(0|0) \;+\; \pi_1\, L(0|1) \quad &< \quad \pi_0\, L(1|0) \;+\; \pi_1\, L(1|1). \\ \pi_0\,[L(0|0) \;-\; L(1|0)] \quad &< \quad \pi_1\,[L(1|1) \;-\; L(0|1)] \end{aligned}$$

The difference $L(1|0) - L(0|0)$ is the *regret* when the state is 0. It is the amount by which the loss of the incorrect decision exceeds that of the correct decision when the state is 0. It may be denoted by $R(|0)$. The difference $L(0|1) - L(1|1)$ is the amount by which the

loss of the incorrect decision exceeds that of the correct decision when the state is 1. It is the regret when the state is 1 and may be denoted by $R(\,|\,1)$. Continuing, we have

$$\begin{aligned} \pi_0\,[L(0|0) - L(1|0)] &< \pi_1\,[L(1|1) - L(0|1)] \\ \pi_0\,R(\,|\,0\,) &< \pi_1\,R(\,|\,1) \\ \pi_1\,/\,\pi_0 &> R(\,|\,0\,)\,/\,R(\,|\,1) \end{aligned}$$

Given data, the prior probabilites are replaced by posterior probabilities $\Pr(0|\boldsymbol{x})$ and $\Pr(1|\boldsymbol{x})$. The posterior expected losses are compared. The expected posterior loss of Action a given state s is

$$\mathcal{E}[L(a|s)\,|\,\boldsymbol{x}] \;=\; \Pr(0\,|\,\boldsymbol{x})\,L(a|0) \;+\; \Pr(1|\boldsymbol{x}\,)\,L(a|1).$$

The optimal rejection region can be described in terms of a set of values of $\boldsymbol{x}$ where the expected posterior loss of action 1 is less than that of action 0. The above inequality becomes: take action 1 if

$$\Pr(\text{State}\,1\,|\,\boldsymbol{x})\,/\,\Pr\{\text{State}\,0\,|\,\boldsymbol{x}) \;>\; R(\,|\,0\,)\,/\,R(\,|\,1).$$

This is equivalent to

$$\pi_1\,f_1(\boldsymbol{x})\,/\,\pi_0\,f_0(\boldsymbol{x}) \;>\; R(\,|\,0\,)\,/\,R(\,|\,1).$$

This can be written as

$$f_1(\boldsymbol{x})\,/\,f_0(\boldsymbol{x}) \;>\; \pi_0\,R(\,|\,0\,)\,/\,\pi_1 R(\,|\,1).$$

The LHS, $f_1(\boldsymbol{x})\,/\,f_0(\boldsymbol{x})$ is the *likelihood ratio.* Often it reduces to a function of sufficient statistics, such as the sample mean and variance in the Gaussian case. The probability content of the region where the LR exceeds the RHS above in the space of $\boldsymbol{x}$ under the null hypothesis (that is, in State 0) will be the optimal α. Note that it depends upon the prior probabilities and the four loss values, through the ratio of regrets.

The decision risk analysis framework obviously extends to more than two actions. Here the case of two actions has been discussed because it corresponds to hypothesis testing.

Note that α is not an input to the problem but rather could be computed as a function of the inputs. Often these inputs would be hard to come by, so another approach, via model-selection criteria, is discussed here.

5.2 Approach via Model Selection Criteria

Model selection criteria were really developed mainly for choosing among models with a somewhat large number of parameters. But it is interesting to see what they imply about problems with smaller numbers of parameters. In particular, it is interesting to develop model selection criterion-based tests as alternatives to conventional hypothesis testing, which is coming under so much criticism.

5.2.1 BIC

Another approach is via model comparison by menas of a model selection criterion such as BIC. (Schwarz 1978, Kashyap 1982). Given K alternative models, indexed by $k = 1, 2, \ldots, K$, the criterion BIC is

$$\text{BIC}_k \;=\; -2\,\text{LL}_k \;+\; m_k\,\ln(n) \;=\; \text{LOF}_k \;+\; \text{Penalty}_k,$$

where m_k is the number of free parameters estimated in Model k. The term LOF_k is the Lack of Fit of Model k, namely, -2 LL_k, where LL, for "log likelihood", is the log of the maximized likelihood.. The penalty of Model k due to the number m_k of parameters fit is $m_k \ln n$. BIC is a smaller-is-better criterion. BIC_k is the leading terms of a Taylor series expansion of $-2 \ln pp_k$, where $\text{pp}_k = \Pr(\text{Model } k \mid data)$ is the posterior probability of Model k, given the data. The posterior probability pp_k is approximately $pp_k \approx C \exp(-BIC_k/2)$, where the constant C is determined so that the sum of the probabilities over the models is 1.

An advantage of BIC is that, in a single criterion, it considers both model fit, via LOF, and sample size, n, via the coefficient $\ln n$ in the penalty factor. In AIC, this coefficient is 2, independent of sample size.

In the two-sample problem, the number of models is $K = 2$, the number of parameters in Model 1 is $m_1 = 1$, the common mean, and the number of parameters in Model 2 is $m_2 = 2$ means. Model 1 corresponds to the null hypothesis H_0, Model 2, to the alternative hypothesis H_1. Let $n = n_1 + n_2$. Indexing the two models by $k = 1, 2$. we have

$\text{BIC}_1 = -2\,\text{LL}_1 + m_1 \ln(n) = -2\,\text{LL}_1 + 1 \ln(n) = -2\,\text{LL}_1 + \ln(n)$,
$\text{BIC}_2 = -2\,\text{LL}_2 + m_2 \ln(n) = -2\,\text{LL}_2 + 2 \ln(n)$.

In a notation familiar from the Analysis of Variance (ANOVA), the decomposition of the sum of squares is

$$\text{SST} = \text{SSB} + \text{SSW},$$

where

$$\text{SST} = \text{SS(Total)} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \bar{y})^2,$$

$$\text{SSB} = \text{SS(Between Groups)} = \sum_{g=1}^{G} n_g (\bar{y}_g - \bar{y})^2,$$

$$\text{SSW} = \text{SS(Within Groups)} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g)^2.$$

The maximized likelihoods come out in terms of SST and SSW and are
$\max L_1 = (2\pi\sigma^2)^{-n/2} \exp[(-1/2\sigma^2)\, SST]$,
$\max L_2 = (2\pi\sigma^2)^{-n/2} \exp[(-1/2\sigma^2)\, \text{SSW}]$,
and for the log maximum likelihoods

$$l_1 = (-n/2) \ln(2\pi\sigma^2) - \text{SST}/2\sigma^2$$

$$l_2 = (-n/2) \ln(2\pi\sigma^2) - \text{SSW}/2\sigma^2;$$

this gives

$$-2\text{LL}_1 = -2\, l_1 = n \ln(2\pi\sigma^2) + SST/\sigma^2 = n \ln(2\pi\sigma^2) + \text{SSB}/\sigma^2 + \text{SSW}/\sigma^2$$

$$-2\text{LL}_2 = -2 l_2 = n \ln(2\pi\sigma^2) + \text{SSW}/\sigma^2$$

So

$$\text{BIC}_1 = -2LL_1 + \ln n = n \ln(2\pi\sigma^2) + \text{SSB}/\sigma^2 + \text{SSW}/\sigma^2 + \ln n,$$

$$\text{BIC}_2 = -2LL_2 + 2 \ln n = n \ln(2\pi\sigma^2) + \text{SSW}/\sigma^2 + 2 \ln n.$$

The test is to reject if $\text{BIC}_1 > \text{BIC}_2$, that is, if $\text{BIC}_1 - \text{BIC}_2 > 0$. This difference is $\text{BIC}_1 - \text{BIC}_2 = [(\text{SSB} + \text{SSW}) - \text{SSW}]/\sigma^2 - \ln n = \text{SSB}/\sigma^2 - \ln n$. Thus the test is to reject if

$$\text{SSB}/\sigma^2 > \ln n.$$

Now note that SSB $= n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2$, where $\bar{y}$ is the overall mean, $(n_1\bar{y}_1 + n_2\bar{y}_2)/(n_1 + n_2)$. This gives SSB $= n_1 n_2(\bar{y}_1 - \bar{y}_2)^2/(n_1 + n_2)$. Note that this means that the test statistic SSB$/\sigma^2$ is equal to $n_1 n_2(\bar{y}_1 - \bar{y}_2)^2/\sigma^2(n_1 + n_2) = n_1 n_2(\bar{y}_1 - \bar{y}_2)^2/\sigma^2(1/n_1 + 1/n_2) = z^2$. Note further that

$$z^2 = (\bar{y}_1 - \bar{y}_2)^2 / \left[\sigma^2 (1/n_1 + 1/n_2)\right] = (\bar{y}_1 - \bar{y}_2)^2/\sigma^2_{\bar{y}_1 - \bar{y}_2}.$$

Performing the test requires only comparing the two BICs. It does not directly involve the distribution of the test statistic. Probabilities are already taken into account by the BICs. However, it is interesting to note that in this comparison of two means, the BIC-based test statistic is the square of a Normally distributed random variable and thus has a chi-square distribution with one degree of freedom, χ^2_1.

Consequently, the test has Type I error rate $\alpha = \Pr\{\chi^2_1 > \ln n\}$. The α for various sample sizes are given below.

Table 6: Alpha in terms of sample sizes

$n = n_1 + n_2$	ln n	$\alpha = \Pr\{\chi^2_1 > \ln n\}$
10	2.303	0.129
20	2.996	0.083
30	3.401	0.065
47	3.850	0.050
50	3.912	0.048
100	4.605	0.032

5.2.2 Another Criterion: AIC

Another well-known model selection criterion, actually preceding BIC, is AIC, Akaike's Information Criterion (Akaike 1973). This is AIC$_k = -2\,\text{LL}_k + 2m_k,\ k = 1, 2, \ldots,\ K$ models; that is, the factor $\ln n$ in BIC is replaced by 2 in AIC. Both model selection criteria (MSCs) AIC and BIC are "penalized likelihood criteria" and take the form

$$\text{MSC}_k = \text{LOF}_k + \text{Penalty}_k = -2\,LL_k + m_k\,a(n),$$

where LOF is the Lack of Fit: LOF$_k = -2\,\text{LL}_k$; the penalty term is $m_k\,a(n)$, the penalty constant is $a(n) = 2$ for AIC and $a(n) = \ln n$ for BIC.

So, AIC does not include the direct dependence upon n that BIC does. AIC is derived (Akaike 1973) as -2 times the estimated entropy of the distributions in the various models. BIC is derived (Schwarz 1978) as the first terms of an expansion of -2 times the posterior probability of Model k.

5.2.3 Discussion in Terms of the Likelihood Ratio Test

To test H$_0$ against H$_1$, one can maximize the likelihood L under H$_0$ and H$_1$ and compare the results. In our indexing of models by $k = 1, 2, \ldots, K$, here $K = 2$ with $k = 1$ corresponding to H$_0$ and $k = 2$ corresponding to H$_1$. The **likelihood ratio** for testing a hypothesis is L_1/L_2. The Likelihood Ratio Test is to reject H$_0$ if $LR = \max L_1/\max L_2 < c$, where c is determined by α. That is, $\Pr(\text{Type I error}) = \Pr(\text{Reject H}_0 \text{ when H}_0 \text{ is true}) = \Pr\{\max L_1/\max L_2 < c \text{ when H}_0 \text{ is true}\} = \alpha$. We have

$$\text{MSC}_1 = -2\ln\max L_1 + m_1 a(n),\ \text{MSC}_2 = -2\ln\max L_2 + m_2 a(n)$$

The difference in MSC values is $\text{MSC}_1 - \text{MSC}_2 = -2\ln(\max L_1/\max L_2) + (m_1 - m_2)a(n) = -2\ln LR - (m_2 - m_1)a(n)$. So one rejects H_0 if

$$\text{MSC}_1 - \text{MSC}_2 > 0;$$

This condition is $-2\,LR - (m_2 - m_1)a(n) > 0$, or $-2\,LR > (m_2 - m_1)a(n)$.

The hypothesis H_0 restricts the parameter to a subset of the parameter space. The *generalized likelihood ratio statistic,* often denoted by Λ, is the ratio of the restricted maximum of the likelihood to the unrestricted maximum. Thus it is defined as

$$\Lambda = \max_{S_0} L(\boldsymbol{\theta}) \,/\, \max_{S_0 \cup S_1} L(\boldsymbol{\theta}).$$

Here $\boldsymbol{\theta}$ is the vector of parameters, S_0 is the set over which $\boldsymbol{\theta}$ varies when H_0 is true, and $S_0 \cup S_1$ is the set over which $\boldsymbol{\theta}$ varies when it is not restricted by the null hypothesis. In the two-sample problem for means, $\boldsymbol{\theta} = (\mu_1, \mu_2)$, S_0 is the set in the parameter space where $\mu_1 = \mu_2$, and $S_0 \cup S_1$ is the whole (μ_1, μ_2) plane.

If the max of L occurs in S_0, then $\lambda = 1$. The test is to reject for small values of the test statistic, that is, to reject if

$$\Lambda < c,$$

where c is a suitably chosen constant that in the classical setting would be dependent upon the level of significance, α.

Under some regularity conditions, the asymptotic (large-sample) distribution of $-2\ln\Lambda$ is chi-square with $m_2 - m_1$ degrees of freedom. (Presumably, this is why Akaike (1973) used a multiplier -2 in defining his criterion.) Note that if $\max L_1 \geq \max L_2$ then $\Lambda = L_1/L_1 = 1$ and $\text{LR} = L_1/L_2 \geq 1$. If $\max L_1 < \max L_2$ then $\Lambda = L_1/L_2 = LR < 1$.

6. Discussion

In this paper, tests based on model-selection criteria are proposed as replacements for testing based on p-values. We have shown how tests based on model selection criteria involve the usually relevant statistics.

We note that the MSCs can be modified to take account of prior probabilities π_k, $k = 1, 2, \ldots, K$ on the models. The likelihood L_k then changes to $\pi_k\, L_k$. The log likelihood l_k is changed to $\ln\pi_k + \ln L_k = \ln\pi_k + l_k$. The rejection criterion $\text{MSC}_1 - \text{MSC}_2 > 0$ becomes

$$[-2\ln\pi_1 - 2\,\text{LL}_1 + 1\,a(n)] - [-2\ln\pi_2 - 2\,\text{LL}_2 + 2a(n)] > 0,$$

$$[-2\ln(\pi_1/\pi_2) - 2\,\text{LL}_1 + a(n)] - [-2\text{LL}_2 + 2a(n)] > 0,$$

$$-2[\text{LL}_1 - \text{LL}_2] > 2\ln(\pi_1/\pi_2) + a(n)$$

The focus here has been on MLEs, often leading to a focus on means. It can also be of interest to consider the extent to which comparison of *medians*, and effect size defined in terms of them, makes more practical sense.

7. Appendix: Review of Some Model-Selection Criteria

Prominent among model-selection criteria are *likelihood based* model-selection criteria such as AIC and BIC.

Consider alternative models indexed by $k = 1, 2, \ldots, K$. Let m_k = number of explanatory variables used in Model k. Let L_k be the maximized likelihood for Model k and LL_k be the natural log of L_k. The "deviance" for Model k is] $-2LL_k$ = LOF$_k$. Here LOF means lack-of-fit. For Gaussian models, Deviance$_k$ = $n \ln(SSE_k \,/\, n)$. Note that SSE$_k/N$ is the maximum-likelihood estimate of the error variance in Model k. Note also that $\ln(SSE_k/n) = \ln SSE_k - \ln n$, and in the present context the term $-\ln n$ can be dropped for purposes of comparing models because it is the same for all models.

Penalized likelihood MSCs take the form MSC$_k$ = Deviance$_k$ + penalty term$_k$. The penalty term$_k$ = is $a(n)\, m_k$, where m_k = number of free parameters used in Model k, $a(n)$ is a function of n or a constant.

7.1 AIC and BIC

MSCs are an alternative to a sequence of hypothesis tests for choosing a model.

AIC: $a_N = 2$ Akaike's Information Criterion: AIC$_k$ = Deviance$_k$ + $2\, m_k$

BIC: $a_N = \ln N$ Bayesian Information Criterion: BIC$_k$ = Deviance$_k$ + $(\ln N)\, m_k$

AIC and BIC involve the LOF (via the Deviance) and the number of parameters used to achieve the fit.

BIC is usually preferred nowadays.

Note that BIC incorporates not only the Deviance and the number of parameters but also the sample size.

Some modifications of these criteria follow.

CAIC. CAIC (Bozdogan 1987) or "consistent" AIC is

$$CAIC_k \;=\; -2LL_k + (\ln n + 1)m_k.$$

It is the same as BIC, but with $\ln n$ in BIC replaced by $\ln n + 1$ in CAIC.

The acronym CAIC also stands for Corrected AIC (also called :bias-corrected AIC): In a multivariate regression problem, let p be the number of Ys and k be the number of Xs. Then

$$\begin{aligned} CAIC &= np \ln 2\pi + n \ln |\hat{\Sigma}| + n(n+k)p/(n-p-k-1) \\ &= AIC + (p+k+1)(p+2k+1)p \,/\, (n-p-k-1-1), \end{aligned}$$

where $AIC = np(\ln 2\pi + 1) + n \ln |\hat{\Sigma}| + 2pk + p(p+1)$.

AICc is AIC with a correction for finite sample size. It is

$$AICc \;=\; AIC \;+\; 2k(k+1)/(n-k+1),$$

where n denotes the sample size and k denotes the number of parameters. It is AIC but with a greater correction, by an amount $2k(k+1)/(n-k+1)$, for extra parameters.

7.2 Mallows' C_p

Mallows' C_p is a model selection criterion for alternative multiple regression models. The statistic C_p is an estimate of the Mean Squared Error of Prediction (MSEP), so it is a smaller-is-better criterion. Note that MSEP is the sum of the variance and squared bias.

Given K alternative models] $k = 1, 2, \ldots, K,$ denote the value of C_P for Model k by $C_P(k)$. Then $C_P(k) = (N - m_k - 1)(\frac{MSE_k}{MSE_{full}} - 1) + (m_k + 1)$. Here MSE_{full} is MSE_K, the MSE of the full model with the maximum number K of predictors. Note that the second term $m_k + 1$ increases with the number m_k of variables in Model k. The first term compares the MSE of Model K with that of the full model. For the full model, this first term equals 0. The statistic MSE_k includes both the error variance and the square of the bias of Model k; the bias reflects the contribution of omitted variables. Note that this criterion also is of the form, LOF + penalty term.

7.3 Development of BIC

We are viewing hypothesis testing as a particular case of model selection. Schwarz (1978) considers the problem of selecting one of a number of models of different dimensions is treated by finding its Bayes solution, and evaluating the leading terms of its asymptotic expansion. These leading terms in the expansion are a valid large-sample criterion beyond the Bayesian context, since they do not depend on the prior distribution.

Kashyap (1982) reviews and applies this development. His proof has a few more comments along with it than does Schwarz'.

Qualitatively, both Schwarz' procedure and Akaike's give "a mathematical formulation of the principle of parsimony in model building." Quantitatively, since Schwarz' procedure differs from Akaike's only in that the dimension is multiplied by $1/2 \log n$, Schwarz procedure leans more than Akaike's towards lower-dimensional models (when there are 8 or more observations). For large numbers of observations the procedures differ markedly from each other. If the rather minimal assumptions made by Schwarz are accepted, Akaike's criterion cannot be asymptotically optimal. According to Schwarz, no such proof of optimality seems to have been published before his work, and the heuristics of Akaike(1974) and of Tong (1975) do not seem to lead to any such proof.

Following Akaike (1973, 1974), we multiply the criteria by -2, obtaining the expressions

$$\text{AIC} \;=\; -2LL + 2m_k$$

and

$$\text{BIC} \;=\; -2LL + (\ln n)m_k$$

for the two criteria. In this form, the criteria are of the form LOF + Penalty, where LOF is lack of fit. Apparently, Akaike was motivated to introduce the multiplier of -2 because of the analogy with $-2 \ln \Lambda$ being asymptotically distributed according to a chi-square distribution, where the statistic Λ is the generalized likelihood ratio test statistic. Also, $-2LL$ in Gaussian models is n times the log of the MLE of the error variance σ^2, $n \ln \hat{\sigma^2}$. This statistic $-2LL \;=\; n \ln \hat{\sigma^2}$ is often called the *deviance.*

8. Notation

Here, the acronyms and symbols used are listed.

8.1 Acronyms

AIC Akaike's Information Criterion

BIC Bayesian Information Criterion

GOF goodness of fit

LHS lefthand side

LL maximized log likelihood; LL_k, maximized log likelihood of Model k

LOF lack of fit

MSC Model selection criterion (such as AIC or BIC)

RHS righthand side

SIC Schwarz' Information Criterion

8.2 Symbols

D dimension of the sufficient statistic Y in the expression for an Exponential (Koopman-Darmois) distribution

$f_X(x))$ value of the p.d.f. of the r.v. X, evaluated at x

K number of alternative models, indexed by $k = 1, 2, \ldots, K$

$L(a, s)$ loss of taking action a when s is the true state of nature

M maximized likelihood; M_k, maximized likelihood of Model k

n sample size

pp_k posterior probability of Model k

$R(\cdot \mid s)$ regret when the state is s

θ value of a parameter

Θ parameter, considered as a r.v. in a prior distribution

X, Y random variables (r.v.s)

$x_1, x_2, \ldots, x_n$ observations: observed values of a r.v. X

$y_1, y_2, \ldots, y_n$ observations: observed values of a r.v. Y

z_p the upper p-th percentile of the standard Normal distribution

9. Bibliography

9.1 References on p-values

American Psychological Association. "P-values Under Question." *Psychological Science Agenda,* March, 2016. http://www.apa.org/science/about/psa/2016/03/p-values.aspx

Miller, Rupert G., Jr. (1966), *Simultaneous Statistical Inference.* McGraw-Hill Book Company, New York. (1981), Springer, New York.

Nestler, Scott, and Schramm, Harrison (2016), "P-value Primer: P OR (P-Values in Operations Research) M N O P Q R S T You don't need a license to download R, but you should have a good understanding of p-values", *OR/MS Today,* **Vol. 43,** June, 2016.

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology,* **36(1)**, 1-2. Taylor & Francis (Routledge), Philadelphia, PA.

Trafimow, D., and Marks, M. (2015), Editorial, *Basic and Applied Social Psychology,* **37(1),** 1-2. Taylor & Francis (Routledge), Philadelphia, PA.

Wasserstein, Ronald L., and Lazar, Nicole A. (2016), "The ASAs Statement on p-Values: Context, Process, and Purpose". *The American Statistician,* **Vol. 70, No. 2,** 120 – 133.

9.2 References on Model Selection Criteria

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle", *Proceedings of the 2nd International Symposium on Information Theory* (eds. B.N. Petrov and F. Csaki), 267-281, Akademia Kiado, Budapest.

Akaike, H. (1974), " A New Look at the Statistical Identification Model", *IEEE Trans. on Automatic Control,* **19,** 716 – 723.

Bozdogan, H. (1987), "Model Selection and Akaike'sIinformation Criterion (AIC): the General Theory and its Analytical Extensions", *Psychometrika,* **52,** 345 - 370.

Burnham, Kenneth P., and Anderson, David R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer.

Imori, Shinpei, Yanagihara, Hirokzu, and Wakaki, Hirofumi (2014), . "Simple Formula for Calculating Bias-corrected AIC in Generalized Linear Models", *Scandinavian Journal of Statistics,* **41,** 535–555.

Kashyap, R. L. (1982), "Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models", *IEEE Trans. Pattern Analysis and Machine Intelligence,* **4,** 99-104.

Schwarz, Gideon (1969), "A Second Order Approximation to Optimal Sampling Regions", *Annals of Mathematical Statistics,* **40,** 313 – 315.

Schwarz, Gideon (1971), "A Sequential Student Test", *Annals of Mathematical Statistics,* **42,** 1003 – 1009.

Schwarz, Gideon (1978), "Estimating the Dimension of a Model", *Annals of Statistics,* **6,** 461-464.

Sclove, Stanley L. (2017), . "Levels of Granularity: from Histograms to Clusters to Predictive Distributions". Presented at the 2016 Annual Meeting of the Classification Society, University of Missouri, June 1-4, 2016. To appear in the *Journal of Statistical Theory and Applications*.

Tong, Howell (1975), "Determination of the Order of a Markov Chain by Akaike's Information Criterion", *Journal of Applied Probability,* **12,** 488 – 497.