ALTERNATIVES TO A *P*-VALUE IN SIMPLE "*t*-TESTS"

Many practical applications focus on a simple "*t*-test" of a set of observations partitioned by treatment and control groups. Here, as with all experiments, the main issue is the estimation of effect size (E) – the difference *caused* by the treatment. Using traditional methods, one often computes a "*P*-value," which is the probability of a test statistic as large as, or larger than, that observed, *given* that the null hypothesis is true. [Many people error in thinking that a *P*-value is the probability that the null hypothesis is true – this is not the proper meaning of a *P*-value.]

Information-theoretic approaches can be employed to provide more meaningful quantities, such as:

> the likelihood of both the null hypothesis and the alternative hypothesis, given the data, $\mathcal{L}(\text{H}_0|\text{data})$ and $\mathcal{L}(\text{H}_a|\text{data})$,
>
> the probability of both the null hypothesis and the alternative hypothesis, given the data, $\mathcal{P}rob(\text{H}_0|\text{data})$ and $\mathcal{P}rob(\text{H}_a|\text{data})$,
>
> the evidence ratio of the two hypotheses.

The model likelihoods and model probabilities have a rigorous, clean interpretation related to strength of evidence. [A *P*-value is not a measure of strength of evidence.] Model averaged estimates of effect size are also possible for observational studies.

The computation of these quantities is quite easy because a proper residual sum of squares (RSS) is available from the statistics leading to a *t*-statistic and the *P*-value.

Given the RSS for each model, $\mathrm{AICc} = n \cdot \log\left(\frac{\mathrm{RSS}}{n}\right) + 2K + \frac{2K(K+1)}{n-k-1}$, and $\Delta_i = \mathrm{AICc}_i - \mathrm{AICc}_{min}$.

Model likelihoods, model probabilities, and evidence ratios are easily computed from the Δ_i. The procedure for computing the RSS (using the MLEs of the structural parameters) is given below for both the classical unpaired and paired designs. Note that RSS/n is the maximum likelihood estimate (MLE) of the residual variance.

Unpaired Design:

The null hypothesis, H$_0$ Effect size = 0.

μ and σ^2, $K = 2$ parameters. RSS = $\sum\left(x_{ci} - \hat{\mu}\right)^2 + \sum\left(x_{ti} - \hat{\mu}\right)^2$

The alternative hypothesis, H$_a$ Effect size = E = $\mu_c - \mu_t$.

μ_c, μ_t, and σ^2, $K = 3$ parameters.

RSS = $\sum\left(x_{ci} - \hat{\mu}_c\right)^2 + \sum\left(x_{ti} - \hat{\mu}_t\right)^2$

Paired Design: The differences (d_i) are critical; d_i = $x_{ci} - x_{ti}$.

The null hypothesis, H$_0$ Effect size = 0.

σ^2, $K = 1$ parameter. RSS = $\sum^{n}\left(d_i\right)^2$

The alternative hypothesis, H$_a$ Effect size = E = $\bar{d}$.

$\bar{d}$ and σ^2, $K = 2$ parameters. RSS = $\sum^{n}\left(d_i - \bar{d}\right)^2$

These “treatment/control” data, including extensions to ANOVA designs, constitute a single data set.

Consider the example from Snedecor and Cochran (1967),

TABLE 4.3.1
NUMBER OF LESIONS ON HALVES OF EIGHT TOBACCO LEAVES*

Pair No.	Preparation 1 X_1	Preparation 2 X_2	Difference $D = X_1 - X_2$	Deviation $d = D - \bar{D}$	Squared Deviation d^2
1	31	18	13	9	81
2	20	17	3	-1	1
3	18	14	4	0	0
4	17	11	6	2	4
5	9	10	-1	-5	25
6	8	7	1	-3	9
7	10	5	5	1	1
8	7	6	1	-3	9
Total	120	88	32	0	130
Mean	15	11	$\bar{D} = 4$		$s_D^{2} = 18.57$
$s_{\bar{D}}^{2} = 18.57/8 = 2.32,\ s_{\bar{D}} = 1.52$ lesions					

Under the null hypothesis, the RSS is 258 (obtained by squaring and summing the values in the 4th column in the Table above – $13^2+3^2+\ldots+1^2 = 258$) with a sample ($n$) of 8 pairs, and $K = 1$, so

$$\begin{aligned}\text{AICc} &= n \cdot \log\left(\frac{\text{RSS}}{n}\right) + 2K + \frac{2K(K+1)}{n-k-1},\\ &= 8 \cdot \log\left(\frac{258}{8}\right) + 2(1) + \frac{2(1)(2)}{8-1-1},\\ &= 30.4548\,.\end{aligned}$$

Under the alternative hypothesis, the RSS is given in the Table as 130 and $K = 2$, so

$$\begin{aligned}\text{AICc} &= 8 \cdot \log\left(\frac{130}{8}\right) + 2(2) + \frac{2(2)(3)}{8-2-1},\\ &= 28.7047\,.\end{aligned}$$

$$\Delta_i = \mathrm{AICc}_i - \mathrm{AICc}_{min}$$

$$H_0 = 30.4548 - 28.7047 = 1.7501$$

$$H_a = 28.7047 - 28.7047 = 0$$

The best model is the alternative hypothesis H_a; $\bar{d} = 4$.

The likelihood of each model, given the data, is

$$\mathcal{L}(H_0|\text{data}) = \exp(-\frac{1}{2}\Delta_i) = 0.4168$$

and

$$\mathcal{L}(H_a|\text{data}) = \exp(-\frac{1}{2}\Delta_i) = 1.0000.$$

The model for H_a is more likely.

The probability of each model, given the data, is

$$\mathcal{Prob}\,(H_0|\text{data}) = 0.4169/1.4168 = 0.2942$$

and

$$\mathcal{Prob}\,(H_a|\text{data}) = 1.0000/1.4168 = 0.7058.$$

The model for H_a is more probable.

The evidence ratio is 1.0000/0.4168 or 0.7058/0.2942 = 2.399; again, the evidence supports model H_a over model H_0. The difference is hardly overwhelming; it might be judged to be weak.

These results differ from the traditional approach. Snedecor and Cochran (1967) report the *t*-statistic of 2.63 with 7 df, a *P*-value of "about 0.04," and state that the null hypothesis is rejected.

Several points can be made here. Most importantly, the *P*-value is not the same as the probability of model H_0; they are not really comparable as they mean different things. The "*P*-value" in this case is the probability of a test statistic as large as 2.63, or larger, *given* that the null hypothesis is true. *P*-values are a "tail probability" as they include probabilities for data more extreme than those observed. This approach gauges the probability of the data, or more extreme data, given that the null is true. The *P*-value rests critically on the asymptotic distribution of the test statistic. The *P*-value should not be used as if it were a formal measure of strength of evidence.

The model probabilities, either information-theoretic or Bayesian, provide the probability of the null model, given the data. They also provide the probability for other models; there might be only a single alternative or additional alternative models. There is no test statistic, no assumptions about the theoretical distribution of the test statistic, no concept of a cut-off (α), and no decision about "statistical significance."

In a paired or unpaired observational study, one may want to model average the estimate of effect size. This is easy under an information-theoretic approach and impossible under the traditional null hypothesis testing approach. Finally, the variance of such model averaged estimates can easily incorporate a variance component due to model selection uncertainty.

ANOVA models can be similarly cast into a simple information-theoretic framework. Here again, one needs only the RSS for each model and the sample size. These quantities are always given by statistical software packages. Then one computes

$\text{AICc} = n \cdot \log\left(\frac{\text{RSS}}{n}\right) + 2K + \frac{2K(K+1)}{n-k-1}$, Δ_i, $\mathcal{L}$(model i|data),

$\mathcal{P}rob$(model i|data), and evidence ratios.

Regression models are easily cast into an information-theoretic framework; again one starts with the RSS.

People naturally tend to cling to traditional t-tests and ANOVAs because this is the only thing they have been taught and they are familiar with the procedure. Better methods have been developed since the early methods of the 1920s and 1930s. These methods are superior in virtually all respects and their use is encouraged.

Breeding Spurious Effects

Spurious effects are a threat to valid inductive inference. They arise in such a way that the investigator has no way to realize that a particular effect is, in fact, spurious. Such spurious results come from over-fitting, types of data dredging, model selection bias, and other (sometimes) subtle causes. There are four such causes that seem reasonable to identify:

1. Little theory to guide the hypothesizing and modeling,
2. Small sample size,
3. Large number of variables/factors (say, over a dozen),
4. Large number of models (perhaps 1000s or even millions).

There is a large statistical literature on these issues. These issues are most often associated with descriptive studies or "fishing trips." Sometimes little planning has gone into the effort. Often variables are measured because they are easy to measure. Some people might consider such activities "exploratory data analysis," but such activity was not the focus of John Tukey's work on EDA.

When faced with problems with little theory, small samples, and many variables and models, one must be aware of the very high probability of results that are spurious. Such activities are usually a poor way to advance science to the next level; it is often better to continue thinking hard about the issue and proceed in a more rational manner.

Provided by Dr. David R. Anderson, August 2007
http://www.warnercnr.colostate.edu/~anderson/