

## SIMULTANEOUS STATISTICAL INFERENCE

L. N. Gray

January 1, 2004

### HYPOTHESIS TESTING

In testing hypotheses the researcher may either make a correct decision, an incorrect decision, or judgment may be suspended. If a completely randomized experiment, for example, contains two treatment levels and, correspondingly, one comparison among means, the probability of committing a Type I Error if the (null) specific statistical hypothesis,  $H_0: \mu_1 - \mu_2 = 0$ , is true (and all assumptions are correct) is determined by the alpha level adopted. In the case of regression analysis, the same situation arises when there is a single independent variable, i.e.,  $H_0: \beta_1 = 0$ .

In both cases the (null) hypothesis is examined via a  $t$ -test of the form

$$\Pr\left(\frac{|\hat{H}_0|}{\hat{\sigma}_{H_0}} \geq t_{\alpha/2; N-2}\right) = \alpha$$

where

$\frac{|\hat{H}_0|}{\hat{\sigma}_{H_0}}$  is the estimated absolute value of  $\mu_1 - \mu_2$  or  $\beta_1$  divided by its estimated standard error, and  $t_{\alpha/2; N-2}$  is the value of Student's  $t$  at  $N - 2$  degrees of freedom and two-tailed level of significance,  $\alpha/2$ . If we had used an  $F$ -test instead of a  $t$ -test, we would have used 1 and  $N - 2$  degrees of freedom and the one-tailed level of significance,  $\alpha$ .

### MULTIPLE INDEPENDENT VARIABLES

With more than two treatment levels and situations involving multiple independent variables, the meaning of "significance"

becomes more confusing. Suppose we had a completely randomized experiment with three levels, or a regression with two independent variables. If we employed a conventional ANOVA approach to the experiment, our overall hypothesis would be  $H_0: \mu_1 = \mu_2 = \mu_3$ . In the regression situation the analogous hypothesis would be  $H_0: \beta_1 = \beta_2 = 0$ . In both cases these general hypotheses would be tested using the  $F$ -distribution and the test statistic would be computed as

$$\begin{aligned}
 F_{2, N-3} &= \frac{SS(\text{Model}) / df(\text{Model})}{SS(\text{Residual}) / df(\text{Residual})} = \frac{MS(\text{Model})}{MS(\text{Residual})} \\
 &= \frac{(N - 3) \times SS(\text{Model})}{2 \times SS(\text{Residual})} .
 \end{aligned}$$

When we use this  $F$ -value in an hypothesis test we are actually testing two comparisons or parameters in each of the applications. In the regression situation we are testing both  $H_{01}: \beta_1 = 0$  and  $H_{02}: \beta_2 = 0$  and questioning whether or not *one or more* of these parameters is different from zero. In the completely randomized experimental design we are questioning whether or not *one or more* of the means is different from the others. In each case the overall  $F$ -test actually represents a "family" error rate, where the "family" consists of two sub-hypotheses. In all cases  $df(\text{Model})$  indicates the number of (orthogonal) hypotheses that can be examined.

In both of these cases we have 2 degrees of freedom in the numerator for the test of the general hypothesis using the  $F$ -distribution. Thus a significant (rare) probability level refers to both hypotheses simultaneously and can be interpreted to mean that at least one of the two component null hypotheses is unlikely to be true.

Contrast this with the situation of a single degree of freedom test. There, we could assign the probability directly to a

specific hypothesis and/or parameter. With two degrees of freedom we can't do that; we are only told that a summary "model" relationship is rare or not.

#### MULTIPLE HYPOTHESIS TESTING

For a simple experiment (one-way ANOVA) it is possible to perform  $k - 1$  independent tests of significance among the  $k$  means. The probability that at least one of the comparisons or contrasts will show spurious significance (Type I Error) is

$$1 - (1 - \alpha)^{k - 1}.$$

This is approximately equal to  $(k - 1)\alpha$  for small  $\alpha$  levels. Even if fewer than  $k - 1$  tests are made, the probability of spurious significance is

$$1 - (1 - \alpha)^c,$$

where  $C$  is the number of contrasts or comparisons actually made.

For the example we have been using (three means) the probability of at least one Type I Error in two independent tests is .0975 if both tests are made at the  $\alpha = .05$  level. The ordinary  $F$ -test for a one-way ANOVA with two degrees of freedom in the numerator, would lead us to believe, if significant, that there's only a 5% chance of a Type I Error. Which is correct?

For a multiple regression with  $k$  independent variables (in a model of full-rank) it is possible to perform  $k + 1$  independent tests of significance among the regression coefficients (including the intercept). The probability that at least one of the coefficients will show spurious significance (Type I Error) is

$$1 - (1 - \alpha)^{k + 1}.$$

This is approximately equal to  $(k + 1)\alpha$  for small  $\alpha$  levels. Even if fewer than  $k + 1$  coefficients are tested, the probability of spurious significance is still

$$1 - (1 - \alpha)^C,$$

where  $C$  is the number of independent coefficients tested.

For the example we have been using (two independent variables) the probability of at least one Type I Error in the two tests is .0975 if both tests are made at the  $\alpha = .05$  level. The ordinary  $F$ -test for an  $R$ -square with two degrees of freedom in the numerator, would lead us to believe, if significant, that there's only a 5% chance of a Type I Error. Which is correct?

#### SIMULTANEOUS INFERENCE

The answer to both questions is that they both are correct, depending on how we define the base for the probability in question. With overall tests, based on  $SS(\text{Model})$  and  $df(\text{Model})$  in comparison with  $SS(\text{Residual})$  and  $df(\text{Residual})$ , we have a  $100\alpha\%$  chance of a Type I Error when we assert that some aspect of the model (e. g., any mean or regression coefficient) is different or non-zero. The base for the probability statement is the set of overall tests, i.e., the probability statement means that  $100\alpha\%$  of the overall  $F$ -values in exact replications of the study would exceed the criterion value when the composite hypothesis ( $H_0: \mu_1 = \mu_2 = \mu_3$  or  $H_0: \beta_1 = \beta_2 = 0$ ) is true.

In contrast, when we are testing specific hypotheses regarding parameters (e.g.,  $H_{01}: \beta_1 = 0$  or  $H_{02}: \beta_2 = 0$  for regression or analogous hypotheses for an experiment) the base for the probability statement is the individual hypothesis itself, i.e., the probability statement means that  $100\alpha\%$  of  $t$ -values in replications of the study and the specific hypothesis test

within the study would exceed the criterion value when the hypothesis is true, granting all assumptions of course.

#### A *PRIORI* AND A *POSTERIORI* HYPOTHESES

There are two general situations in which hypotheses are tested: (1) situations in which the hypotheses are specified prior to data collection and analysis, i.e., a *priori* hypotheses; and (2) situations in which hypotheses are tested during a process of "data snooping", i.e., specified after data collection and initial analysis (*a posteriori*).

In the first case we need not report the results of an overall analysis of the data, except for its descriptive value or as implied support for our *a priori* hypotheses. In the second case the first analytical step is the testing of the overall hypothesis and, where applicable, its associated measure of association. After this outcome is shown to be rare (significant), a *posteriori* hypotheses can be tested. If we've found our overall model "significant", we can be sure that at least one of the embedded hypotheses is significant at the same probability level.

#### A *PRIORI* TESTING OF MULTIPLE HYPOTHESES

If a significant overall test suggests that some effect may exist with a  $100\alpha\%$  chance of a Type I Error, we might prefer that the same level of probability statement to apply to the set of subsidiary hypotheses that are embedded in the overall hypothesis. When a researcher has specific hypotheses that are advanced prior to data collection or analysis, such a *priori* hypotheses are, in sociology, ordinarily tested in the usual way so that multiple simultaneous hypotheses have a combined Type I Error probability of

$$1 - (1 - \alpha)^C,$$

where  $C$  is the number of independent hypotheses tested. Since there is no difficulty in computing this probability, it has been felt that this approach was not misleading. Currently, fields such as psychology, biometrics, and neuroscience have moved in the direction of reporting probabilities that take into account the simultaneous nature of hypothesis testing and making this fact clear to their audiences.

Two simple techniques are commonly employed in order to control the overall  $\alpha$  level when multiple *a priori* hypotheses are tested: (1) the *Bonferroni* approach, and (2) the *Sidak* (1967) approach. Using the *Bonferroni* approach, if  $C$  *a priori* hypotheses are to be tested, each should be tested at the  $\alpha/C$  level. Thus if five *a priori* hypotheses were to be tested simultaneously at the .05 level, the researcher would test each hypotheses at the .01 level (they sum to .05).

Similarly we can apply this approach to the obtained significance of an ordinary  $t$ -test: the "real" significance, adjusted for the  $C$  simultaneous tests, is  $Cp$ , where  $p$  is the level associated with the ordinary  $t$ -test and  $C$  is the number of *a priori* hypotheses tested. The *Bonferroni* approach is quite conservative and is not often employed when large numbers of hypotheses are tested simultaneously. Nevertheless, it is a clear approach to the problem and should make the difficulty of sorting out complicated systems of hypotheses clear to an audience (especially to policy makers).

The *Sidak* (1967) adjustment is slightly different and provides somewhat narrower confidence intervals. The adjusted probability value is given by

$$p' = 1 - (1 - p)^C$$

where  $p'$  is the adjusted value,  $C$  is the number of simultaneous *a priori* tests and  $p$  is the probability level associated with an ordinary  $t$ -test. Ideally  $C$  should be the number of independent hypotheses simultaneously tested, but still can be useful when the hypotheses are not independent.

#### A POSTERIORI TESTING OF MULTIPLE HYPOTHESES

There are several ways of modifying our procedures in "data snooping" situations, but one of the most general was developed by Sheffe (1959). Using this approach, each test of a specific parametric hypothesis within an overall ANOVA would employ the criterion value of

$$F'_{\alpha:k-1, N-k} = (k - 1)F_{\alpha:k-1, N-k}$$

instead of the usual value. This approach effectively projects the  $F$ -distribution across the orthogonal dimensions the data space so that all probability levels are consistent.

A closely related conservative approach that holds our error rate at the desired  $\alpha$  level for all the regression coefficients in an equation is another extension of the Sheffe (1959) approach, and is given by an adjustment to the usual  $t$ -test

$$t'_{\alpha:N-(k+1)} = \sqrt{(k + 1)F_{\alpha:k+1, N-(k+1)}} \cdot$$

The Sheffe (1959) approach is one of the most generally useful *a posteriori* techniques, but there other, more specific, techniques that can be effectively applied in particular research situations.

An adjusted probability, using the Scheffe approach, can be calculated for *a posteriori* simultaneous hypotheses by using

$$p' = F\left(\frac{t^2}{C}, C, v\right),$$

where  $p'$  is the adjusted probability,  $t$  is the value of the conventional  $t$ -statistic,  $C$  is the number of hypotheses tested,  $v$  is the appropriate residual degrees of freedom, and  $F$  represents the  $F$ -distribution. While this technique is sometimes suggested for *a priori* hypothesis testing, it tends to be very conservative and much better suited to *a posteriori* data snooping.

#### SUMMARY

The point in using these approaches is to avoid misstating the likelihood of an outcome and advancing outcomes as "important" under questionable circumstances. Simultaneous techniques are rarely employed in sociology, though they are often used in many other fields, particularly in Europe. The use of a simultaneous approach results in a lower likelihood of finding "significant" outcomes and, consequently, fewer publications and/or papers. On the other hand, this approach also reduces the likelihood of wasting our time on marginal phenomenon; its unfamiliarity in sociology may be a serious handicap.

#### BASIC REFERENCES

- Miller, Rupert G., Jr. 1966. *Simultaneous Statistical Inference*. New York: McGraw-Hill Book Company.
- Scheffe, Henry. 1959. *The Analysis of Variance*. New York: John Wiley & Sons, Inc.
- Sidak, Z. 1967. "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association* 62: 626-633.