JOURNAL OF MATHEMATICAL PSYCHOLOGY 8, 444–451 (1971)

Statistical Inference with Dependent Observations: Extensions of Classical Procedures[1]

M. FRANK NORMAN

University of Pennsylvania, Philadelphia, Pennsylvania 19104

Some standard statistical procedures for independent observations are extended to long stationary time series in which observations widely spaced in time are practically independent. A sequence of many successive observations on a single subject "at asymptote" will often have this property. A generalized t statistic permits inferences concerning a subject's mean performance level. And a generalization of the one-way analysis of variance yields an F test for differences in stable performance levels between several subjects working under the same experimental condition. This test can also be used to compare several performances of a single subject serving as his own control under different experimental conditions, provided there is no carry-over between conditions. In the final section the F test is used to demonstrate substantial individual differences in asymptotic performances in a probability learning experiment, contrary to the predictions of various learning models.

INTRODUCTION

Much of classical statistics is concerned with estimating and testing hypotheses about the mean $\mu = E(X_j)$ of a sequence $X_1, X_2, \ldots, X_n$ of independent and identically distributed random variables. A related problem is that of testing whether the means $\mu_i = E(X_{ij})$ of several such sequences $X_{i1}, X_{i2}, \ldots, X_{in_i}$, $i = 1, \ldots, I$ are equal. When the number n of available observations is small the standard t and F tests assume that the X's are normal, but the central limit theorem obviates this assumption when n is large.

Within psychology independent random variables usually represent observations on different subjects, and the procedures mentioned above permit inferences about the expected performance produced by an experimental condition, or comparison of corresponding expectations for several conditions, when a number of subjects are observed within each condition.

[1] Supported by National Science Foundation grants GP 7335 and GB 7946X. The raw data for the experiment considered in the last section were kindly made available by Prof. W. K. Estes. I am grateful to Prof. Estes and to Profs. N. H. Anderson, F. W. Irwin, and J. Pickands for their comments on earlier drafts of this paper.

Frequently in psychology one obtains instead a sequence $X_1, \ldots, X_n$ of successive responses from the same subject under one experimental condition, or I such sequences X_{ij}. In the latter case there may be I subjects performing under a single condition, or a single subject serving as his own control under I conditions. The assumption that $X_1, \ldots, X_n$ are independent will not, in general, be appropriate, though X_j and $X_{j'}$ will often be nearly independent if $| j - j' |$ is large. If these variables represent an asymptotic performance, so that the process $X_1, \ldots, X_n$ can be regarded as, in some sense, stationary, the same sorts of questions about $\mu = E(X_j)$ and $\mu_i = E(X_{ij})$ are of interest as in the classical case. One may be interested, for example, in whether μ is the probability matching value (Estes, 1964) in a two choice experiment ($X_j = 1$ or 0 depending on whether A_1 or A_2 occurred on trial j). Or one may want to know whether μ_i is the same for all of the subjects run under a certain experimental condition. This might be predicted by a certain model on a parameter-free basis (as is probability matching in several models), and it will certainly be predicted by any model if the same parameter values are assumed to be applicable to all subjects.

Statistical inference for stationary sequences of correlated observations is a facet of the rapidly developing field of *time series analysis*, which includes *statistical spectral analysis* (Jenkins and Watts, 1968; Parzen, 1967). Work in this area has provided a basis for modification of some standard statistical procedures to take account of dependencies between observations. The main purpose of this paper is to describe some extensions of this sort that are applicable when n is large. A completely different aspect of time series analysis is discussed by Gottman, McFall, and Barnett (1969).

In the final section the generalized F test described in the next to last section is used to demonstrate large differences in μ_i for different subjects in a probability learning experiment where such differences were not obvious and where, moreover, group data showed excellent probability matching. Much work remains to be done before it will be possible to use methods like those presented in this paper routinely, but, as this example shows, these techniques are certainly capable of yielding decisive results in favorable circumstances.

Variance of a Subject's Mean

Suppose that $X_1, X_2, \ldots, X_n$ is a sequence of random variables for which the expectation $E(X_j) = \mu$ and *autovariance of lag k*

$$\operatorname{cov}(X_j, X_{j+k}) = c(k)$$

do not depend on absolute time j. Such a sequence is called a *second order stationary time series*. Note that $c(0) = \operatorname{var}(X_j)$. Generalizing the independence assumption of

classical statistics, it is assumed that observations separated by large lags are practically uncorrelated, in the sense that

$$\sum_{k=0}^{\infty} |\, c(k)| < \infty. \tag{1}$$

Let

$$S = \sum_{j=1}^{n} X_j\,,$$

$$X. = S/n,$$

and

$$Z = \sqrt{n}\,(X. - \mu).$$

It is easy to show that

$$\operatorname{var}(S) = nc(0) + 2\sum_{k=1}^{n-1} (n - k)\, c(k).$$

If $c(k) = 0$ for $k \geqslant 1$, this yields the well known fact that $\operatorname{var}(S) = nc(0)$ or $\operatorname{var}(Z) = c(0)$. The following result serves the same purpose in the general case:

$$\operatorname{var}(Z) = c(0) + 2\sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) c(k) \to c(0) + 2\sum_{k=1}^{\infty} c(k) = \sigma^2 \tag{2}$$

as $n \to \infty$. The quantity σ^2 will play the same role below that $c(0)$ plays when the X_j's are independent. The case $\sigma^2 = 0$ is quite degenerate, so it is assumed that $\sigma^2 > 0$.

Estimation of $c(k)$ and σ^2

The standard estimator of $c(k)$ is the corresponding *sample autocovariance*

$$\hat{c}(k) = \frac{1}{n}\sum_{j=1}^{n-k} (X_j - X.)(X_{j+k} - X.);$$

$\hat{c}(0)$ is the familiar sample variance. It will be consistent (Brunk, 1965, p. 140) if, for example, the time series $X_j' = X_jX_{j+k}$ is second order stationary and satisfies (1).

If it is known that $c(k) = 0$ for $k > K$ (e.g., $K = 0$, the case of uncorrelated variables) then, referring to (2), the obvious estimator of σ^2 is

$$\hat{\sigma}^2 = \hat{c}(0) + 2\sum_{k=1}^{K} \hat{c}(k). \tag{3}$$

In the general case one might consider this estimator with K as large as possible, that is, $K = n - 1$. However, a simple computation shows that this choice leads to $\hat{\sigma}^2 = n^{-1}[\sum_{j=1}^{n}(x_j - x.)]^2 = 0$. Consistent estimation can be achieved if K is selected large enough to include the bigger covariances, but small in comparison to n. An intuitively appealing principle is that K should be chosen as small as possible, subject to the constraint that most of the $\hat{c}(k)$ for $k > K$ appear to be negligible relative to $\hat{c}(0)$.

Much techical information relevant to the choice of K is implicit in the extensive recent literature on estimation of the *spectral density function*

$$f(t) = \frac{1}{2\pi}\left(c(0) + 2\sum_{k=1}^{\infty}\cos(kt)\, c(k)\right),$$

since $\sigma^2 = 2\pi f(0)$. Other estimators defined by various weightings of the terms in (3) have also been considered. The interested reader should consult Jenkins and Watts (1968, Chaps. 6 and 7), Parzen (1967, Papers 5–9), and Pickands (1970). For an example of spectral estimation in a psychological context, see Weiss, Laties, Siegel, and Goldstein (1966). In their study the X_j's are successive interresponse times in a free-operant experiment. Such experiments should provide fertile grounds for applications of the methods presented in this paper.

Asymptotic Normality of Z

It has been shown that $\text{var}(Z) \to \sigma^2$ as $n \to \infty$, and, of course, $E(Z) = 0$. The additional requirement is now imposed that Z/σ have asymptotically the standard normal distribution. Several criteria for this generalized central limit theorem are known (Ibragimov, 1962). They involve replacing (1) by one of several stronger but still reasonable conditions which mean that $X_1, \ldots, X_j$ and $X_{j+k}, X_{j+k+1}, \ldots$ approach independence as $k \to \infty$. Moreover, asymptotic normality of Z is predicted by standard mathematical learning models (Norman, 1968).

If $\hat{\sigma}^2$ is a consistent estimator of σ^2, it follows that the generalized t statistic

$$t = Z/\hat{\sigma}$$

is asymptotically standard normal as $n \to \infty$. Confidence intervals for μ, as well as one-and two-tailed tests of hypotheses of the form $H_0 : \mu = \mu_0$, are then constructed in the usual way. For example, the interval with end points

$$X. \pm 1.96\hat{\sigma}/\sqrt{n}$$

is a 95% confidence interval for μ. If Y_j and Y_j' are comparable responses of two different subjects in some sort of yoked control experiment, inferences can be made about $\mu = E(Y_j) - E(Y_j')$ by applying this theory to $X_j = Y_j - Y_j'$.

Suppose now that X_{1j}, $j = 1,..., n_1$ and X_{2j}, $j = 1,..., n_2$ represent performances of two nonyoked subjects, or of one subject serving as his own control in two experimental conditions, between which there is no carry-over. If each response sequence satisfies our assumptions with expectations $\mu_i = E(X_{ij})$ and estimators $\hat{\sigma}_i^2$ of σ_i^2, then

$$t = \frac{(X_{1.} - X_{2.}) - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2}}$$

is asymptotically standard normal as n_1 and $n_2 \to \infty$, and can serve as a basis for inferences about $\mu_1 - \mu_2$. The next section considers the problem of comparing I subject (or condition) means μ_i, assuming that the corresponding σ_i^2 are equal.

An Analysis of Variance

Let X_{ij} represent the jth observation on subject i, where $i = 1,..., I$ and $j = 1,..., n_i$. It is assumed that, for each i, the sequence $X_{i1}, X_{i2},..., X_{in_i}$ satisfies the assumptions of the previous sections, so that

$$Z_i = \sqrt{n_i}\,(X_{i.} - \mu_i)$$

is asymptotically normal with mean 0 and variance

$$\sigma_i^2 = c_i(0) + 2\sum_{k=1}^{\infty} c_i(k) > 0$$

as $n_i \to \infty$. It is further assumed that the σ_i^2 are all equal, $\sigma_i^2 = \sigma^2$ for all i, which is certainly the case if $c_i(k) = c(k)$ for all i and k. Under these assumptions we wish to test $H_0 : \mu_i = \mu$, $i = 1,..., I$.

If $W_1,..., W_I$ are independent standard normal variables, and $d_1,..., d_I$ are constants such that $\sum_{i=1}^{I} d_i^2 = 1$, Fisher's lemma (Brunk, 1965, p. 294) shows that

$$f(W_i, d_i) = \sum_{i=1}^{I} (W_i - d_i Y)^2,$$

where $Y = \sum d_i W_i$, has distribution χ^2_{I-1}. Since f is a continuous function, it follows that the distribution of

$$\frac{1}{\sigma^2}\sum_{i=1}^{I} n_i(X_{i.} - X_{..})^2 = f\left(\frac{Z_i}{\sigma}, \sqrt{\frac{n_i}{n}}\right)$$

converges to χ^2_{I-1} as $n_1, \ldots, n_I \to \infty$, where $n = \sum n_i$ and

$$X.. = \sum_{i=1}^{I} \frac{n_i}{n} X_{i.} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} X_{ij} .$$

If U has the distribution χ_N^2, then U/N has the distribution $F_{N,\infty}$. And if $\hat{\sigma}_i^2$ is a consistent estimator of $\sigma_i^2 = \sigma^2$ based on $X_{i1}, \ldots, X_{in_i}$,

$$\hat{\sigma}^2 = \sum_{i=1}^{I} \frac{n_i}{n} \hat{\sigma}_i^2 \tag{4}$$

is a much more stable consistent estimator of σ^2. Consequently the distribution of

$$F = \frac{1}{\hat{\sigma}^2} \frac{1}{I-1} \sum_{i=1}^{I} n_i (X_{i.} - X..)^2$$

converges to $F_{I-1,\infty}$ as $n_i \to \infty$.

Any departure from H_0 should tend to inflate the numerator of F without affecting the denominator. Thus F will tend to be large. Therefore H_0 is rejected at level α if F exceeds the upper α point of the distribution $F_{I-1,\infty}$.

If $X_{i1}, X_{i2}, \ldots$ are independent and $\hat{\sigma}_i^2 = \hat{c}_i(0)$, $(1 - I/n)F$ is the standard F statistic for the one way analysis of variance. When the X_{ij} are normal, its distribution is $F_{I-1,n-I}$. Thus our test is equivalent to the standard F test when the assumptions of the latter test are met and the n_i are large. We remark that Winer's (1962, Chap. 4) discussion of "single factor experiments having repeated measures on the same elements" is addressed to totally different situations than those considered here.

If the $\hat{\sigma}_i^2$ are given by (3), and the same K is used for all subjects, $\hat{\sigma}^2$ in (4) takes the form

$$\hat{\sigma}^2 = \hat{c}.(0) + 2 \sum_{k=1}^{K} \hat{c}.(k), \tag{5}$$

where

$$\hat{c}.(k) = \sum_{i=1}^{I} \frac{n_i}{n} \hat{c}_i(k)$$

$$= \frac{1}{I} \sum_{i=1}^{I} \hat{c}_i(k)$$

if $n_i = J$ for all i. By analogy with the suggestion given previously for selecting K in the case of a single time series, K is chosen as small as possible consistent with the requirement that most $\hat{c}.(k)$ be negligible for $k > K$. When $n_i = J$ a plausible interpretation of "negligible" is "not significantly different from 0" according to a t test applied to $\hat{c}_1(k), \ldots, \hat{c}_I(k)$. It is not claimed that this criterion rests on a firm statistical

foundation. It is used in the next section, though, as we shall see, the decision reached there is quite insensitive to K.

An Application to Probability Learning

In the experiment of Friedman, Burke, Cole, Keller, Millward, and Estes (1964), 80 college students attempted to predict which of two lights (1 or 2) would be illuminated on each of a sequence of trials. Response A_k is prediction of light k, and event E_k is illumination of this light. Different phases of the experiment were distinguished by different values of $\pi = P(E_1)$. In the third experimental session there were 288 trials at $\pi = .8$. During the middle $J = 144$ trials of this series the proportions of A_1 responses for all subjects in successive 12 trial blocks show only small unsystematic deviations from the probability matching asymptote of .8 (see Friedman *et al.*, 1964, Table 9 and Fig. 5). Only these trials are considered below. Subjects' responses were lost on a very small proportion of the 80×144 subject-trials, but 23 subjects had at least one response lost and were excluded from the analysis that follows. Thus $I = 57$ subjects were used. The variable X_{ij} is 1 or 0, depending on whether or not subject i made A_1 or A_2 on trial j.

TABLE 1

Mean Covariances $\hat{c}\,.\,(k)$ and Their Standard Errors $D(k)$ for the Friedman *et al.* (1964) Study

k	$\hat{c}\,.\,(k)$	$D(k)$
0	.1468	
1	.0154	.0032
2	.0043	.0028
3	.0018	.0023
4	.0018	.0018
5	.0038	.0021

For these subjects $X.. = .7967$, which represents excellent probability matching for the group. Table 1 gives the first 6 $\hat{c}.(k)$, together with the corresponding standard errors $D(k)$ for $k = 1,..., 5$. The ratio $t = \hat{c}.(k)/D(k)$ is significant at the 5% level (2 tailed) only for $k = 1$. Consequently we take $K = 1$ in (5), so that $\hat{\sigma}^2 = .178$. The other factor

$$\frac{J}{I-1}\sum_{i=1}^{I}(X_{i.} - X_{..})^2$$

of F is 2.217, so $F = 12.5$. This is to be compared with 1.66, the upper .001 point of $F_{60,\infty}$. Thus the assumption that all subjects have the same $\mu_i = P(A_1)$ appears to be seriously in error.

The choice of K is not critical. If, for example, $K = 40$ and $\hat{c}.(k) \leqslant .005$ for $5 < k \leqslant K$, then $\hat{\sigma}^2 \leqslant .551$ and $F > 4.0$.

Of course, a significant F could mean that some of the auxiliary assumptions of the F test (e.g., $\sigma_i^2 = \sigma^2$) are wrong instead of, or, more likely, in addition to, the null hypothesis. Whatever its source, the significant F is not without interest. For all of the assumptions of the F test will be met by standard learning models (e.g., the linear model considered by Friedman *et al.*, 1964, and the pattern model considered by Atkinson and Estes, 1963, pp. 153–181) under the conventional assumption that all relevant parameters (θ for the linear model, c and N for the pattern model) are the same for all subjects. The large F suggests that this extrapsychological and rather implausible assumption is a source of serious error in many applications of learning and other mathematical models. One expects to see this assumption relaxed with increasing frequency in the years to come.

REFERENCES

ATKINSON, R. C., AND ESTES, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. II. New York: Wiley, 1963. Pp. 121–268.

BRUNK, H. D. *An introduction to mathematical statistics*. (2nd ed.) Waltham, Mass.: Blaisdell, 1965.

ESTES, W. K. Probability learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press, 1964. Pp. 89–128.

FRIEDMAN, M. P., BURKE, C. J., COLE, M., KELLER, L., MILLWARD, R. B., AND ESTES, W. K. Two-choice behavior under extended training with shifting probabilities of reinforcement. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford: Stanford University Press, 1964. Pp. 250–316.

GOTTMAN, J. M., MCFALL, R. M., AND BARNETT, J. T. Design and analysis of research using time series. *Psychological Bulletin*, 1969, **72**, 299–306.

IBRAGIMOV, I. A. Some limit theorems for stationary processes. *Theory of Probability and Its Applications*, 1962, **7**, 349–382.

JENKINS, G. M., AND WATTS, D. G. *Spectral analysis and its applications*. San Francisco: Holden-Day, 1968.

NORMAN, M. F. Mathematical learning theory. In G. B. Dantzig and A. F. Veinott, Jr. (Eds.), *Mathematics of the decision sciences*. (Part 2). Providence: American Mathematical Society, 1968. Pp. 283–313.

PARZEN, E. *Time series analysis papers*. San Francisco: Holden-Day, 1967.

PICKANDS, J. Spectral estimation with random truncation. *Annals of Mathematical Statistics*, 1970, **41**, 44–58.

WEISS, B., LATIES, V. G., SIEGEL, L., AND GOLDSTEIN, D. A computer analysis of serial interactions in spaced responding. *Journal of the Experimental Analysis of Behavior*, 1966, **9**, 619–626.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

RECEIVED: June 1, 1970

Printed in Belgium