AN EXPOSITORY SURVEY OF MULTIPLE COMPARISONS

Erkan ÖNGEL *

I. Introduction :

Very often a researcher's interest is to compare **two** treatments to see if there is a significant difference. In the case of two treatments the necessary statistical test was provided to him long time ago. The early development of F-test faciliated testing the equality of differences among **all** treatments. This seems to be rather unfortunate because this was not exactly what the researcher needed. Sir R. A. Fisher, the originator of the F-test, tried to find a solution to the problem.

In 1935 he introduced the concept of LSD (Least Significant Difference) as a tool for judging individual comparisons following a significant F-test. This is where the complex problem of multiple comparisons started.

Researchers were not satisfied and somewhat disappointed because sometimes the results of the F-test contradicted those of the t-tests, making it difficult to know which to rely on. The statistician was not satisfied either, and found himself responsible for resolving this contradiction. But the **real** dissatisfaction for the statistician was the concern about Type I error. Because everything that he knew about Type I error and power did not seem to be directly applicable to the new testing procedure, it was necassary to develop new terminology for different types of errors.

Starting in 1939, Newman, Tukey, Duncan, Scheffé, Dunn, Dunnett and others developed new methods all of which easily solved the consumers first problem, but introduced a new one, namely, he observed that diffirent procedures resulted in almost entirely different conclusions for the same comparisons. The reason was that

---

* Assistant Professor at the Middle East Technical University, Department of Economics and Statistics, Ankara.

I thank to Dr. J. E. Carlson and Dr. L. A. Pingel for their critical comments and constructive suggestions.

different definitions of error rate were used (or emphasized). Statisticians still have not agreed upon a common definition. In practice, there seems to be a resistance against the recent definitions of error rate and defence of the Comparisonwise Error Rate (Wilson, 1962). As Miller described, "... For them to change now may not be emotionally possible" (1966, p. 93). On the other hand, because the statisticians could not agree upon a single all-purpose definition, they passed the responsibility to the experimenter. "The experimenter is far more familiar with the data, its virtues and vagaries, than the reader, so it is his prime responsibility to draw the main conclusions" (Miller, 1966, p. 34).

Regardless of who was responsible for this state of affairs, the basic problem is clear and straightforward: What definition of significance level is adopted in each procedure? As Waller and Duncan pointed out, "Opinions differ as to whether the customary operational choices of α (.10, .05, .01) should be made in terms of a comparisonwise α, an experimentwise α, or some intermediate form" (1969, p. 1484). Miller describes the same paradox in a simpler form with different terminology: "To this author the principal disagrement seems to revolve around whether the consumer needs a test of significance or a confidence interval" (1966, p. 2).

A clear statement of the problem does not lead to a neat solution. New terms introduced in place of classical Type I error are conceptually difficult to grasp. Even worse, selection and use of possible substitutes are quite arbitrary or based on subjectivity, the statistician can not be of further help. The complexity of multiple comparisons reached to such a point that Ryan's conclusion is understandable, "An adequte solution of the problem might even lead to an abondonment of significance testing in favor of some other method of dealing with the effects of sampling error which would not create the dilemma with which we are now faced" (1962, p. 305).

In the above discussion the existence of a problem has been introduced. It was observed that the problem basiciy centered around the error rates. Therefore without a sound description of error rates, the nature of the dilemma can not be fully explored.

II. The Concept Of Error Rate

It is important to distinguish 'multiple test' procedures from 'single test' produres, since from this difference emerged the new

conceptualization of error rates. A single test, which may not be very useful and efficient in many research designs, may be defined as 'employing a test statistic, only **once,** on the same data". This situation is seldom encountered in actual research but it is often discussed in elemantary statistics texts. Research problems more often require multible testing procedures rather than a single test statistic. Some illustrations of the wide and important usage of multible testing procedures are, the analysis of intercorrelations of tests given to a particular group, the F-ratios of two way (or higher) ANOVA, simultaneous confidence intervals for regression coefficients, and the replication of an experiment. Multiple comparisons are only one instance of general multiple testing theory in which a statistic is compared for more than two conditions[1]. To make the distinction clear, the F-test in one-way ANOVA is a single test procedure, whereas t-tests among all possible pairs of groups means, is a multiple test procedure.

Suppose $\alpha = .05$ is selected for one specified individual comparison when many exist, and for an F-test. The meaning of Type I error is clear for the F-test but not so for an individual comparison when the existence of other individual comparisons is considered. The probability of rejecting a true difference is then not equal to .05. This means that given an α — level significance in any single test of a comparison, the probability of at least one Type I error will be increased as the number of comparisons increases. The reason is that the actual error rate is related to the set of individual comparisons, and "it is quite a different matter to expect the t-test to be valid for determining the signicinace of the difference between the smallest and largest sample means. A t-test applied to the largest contrast takes no account of the number of groups. (Glass and Stanley, 1970, p. 382)

In the case of independant multiple testing procedures, actual error rate (sometimes called 'error risk'[2]), can be readily obtained through the binomial distribution since an individual test can be taken as a Bernoulli trial. (It will be seen later that actual error rate is more difficult to determine in the case of dependence). Given the two parameters for the binomial distribution, $p = \alpha$ and $n = J$ (num-

---

(1))In this paper only the 'means' will be considered, although medians, correlation coefficent, frequencies or proportions could be considered as well. (Renner, 1969; Ryan, 1960)

(2) The complement of this term, 'Protection Level' is more often used.

ber of groups), $[\alpha + (1 - \alpha)]^J$ yields the probability function of errors for J tests. It can be noted that the last term of expansion of this distribution is the probability of no error in J tests. Consequently, the complement of this probability of having at least one Type I error. That is, $1 - (1 - \alpha)^J$ is the actual error rate for multiple testing. (Hays, p. 312; Glass and Stanley, p. 387 also provide a graphic interpretation).

For a typical case let $\alpha = .05$ and $J = 5$. If the formula is applied, the Type I risk is found to be 0.23. This is considered too large an error for it is usually believed in research that Type I errors are more dangerous than Type II errors. An alternative procedure would be to test each contrast at a relatively stringent level, say $= .001$, so that the total type one error would remain small. But this approach produces a low degree of power since, everything being equal, decreases in α - level increase Type II error. Another reason for not employing a very stringent test is to avoid the phenomenon, of finding no significant contrasts with multible t-tests, after a significant F ratio. (Dayton, 1970, p. 39); That is lack of consistency between protection levels (Duncan, 1965, p. 178). The preceeding example indicates that Type I error risk can be controlled either for an individual or for a collection of comparisons. As was mentioned earlier, there is not objective criteria for deciding which of the conceptual units should be controlled for. However, relative merits of each unit can be discussed so that the experimenter is conscious of the particular merits of each unit choosen. Before doing this formal definitions of various error rates will be given and illustrated with a numerical example.

Testing a hypothesis means making a statement about the result of a statistical procedure, in favor of or against the null hypothesis. Each statement is either right or wrong with regard to the true value in the population. Any set of statements[1] is denoted as a family in an experimental design. Then error rate for the family is defined as,

$$ER\ \{Fam\} = \frac{N_{I(Fam)}}{N_{T(Fam)}}$$

(1) Sets of statements can be arbitrarily selected for the present discussion without any harm. In practice, a family consists of statements corresponding to a factor in ANOVA. Therefore in a one-way ANOVA, family and experiment are equal. This is not consistent throughout the literature however. Miller for instance, uses family in place of experiment and sub-family for factors.

Where $N_I(_{Fam})$ is the number of incorrect statements and $N_T(_{Fam})$ is the total number of statements in the family. Under the null hypothesis $N_I(_{Fam})$ a random variable, $N_T(_{Fam})$ is a constant (if it is finite) so that ER {Fam} is also a random variable. The mean of this distribution is called the **Expected Family Error Rate** and denoted by

$$E\{F\} = \frac{E\{N_I(_{Fam})\}}{N_T(_{Fam})} \qquad (I)$$

$N_I(_{Fam})$ can be considered as the sum of expected errors for each statement in the family; therefore the equality[1] can be written

$$E\{F\} = \frac{\alpha_1 + \alpha_2 + \ldots + \alpha_{NT\,(FAM)}}{N_T(_{Fam})}$$

The numerator of (I) is also given a special name, **Error Rate Per Family.** When $\alpha_1 = \alpha_2 = \ldots = \alpha_{NT\,(FAM)} = \alpha_P$ this term takes a simplified form $\alpha_{PF} = N_T(_{Fam}) \times \alpha_P$ (2)

In other words, Error Rate Per Family is the long-run average number of erroneous statements made per family. In statistical jargon (Ryan, 1959, p. 29) it is expected NUMBER of errors per family.

The probability of a non-zero family error rate will now be considered. This error rate can be easily controlled and is employed in rather important multible comparison procedures such as, the Turkey and Scheffé methods.

$$P\{F\} = P\{N_I(_{Fam}) / N_T(_{Fam}) > 0\} = P\{N_I(_{Fam}) > 0\}$$

This error rate is also called Error Rate Familywise, and may be expressed as the PROBABİLİTY that one or moro erroneous conciusions will be drawn in a given family. It should be emphasized that the term family is used here as any collection (sub-set) of entire statements in an experiment. Because every set is a subset of itself, the family is not distinghished when it comprises all statements in the experiment. But usually this distinction is made for convenience in introductory text boks and special names are given: Error Rate Per Experiment and Error Rate Experimentwise. Formulas and definitions are the same as the family counterparts except the word family has to be replaced by experiment.

---

(1) From the property of the expected value operator :

$$E(\Sigma X_j) = E(X_1 + X_2 + \ldots + X_N) = E(X_1) + E(X_2) + \ldots + E(X_N)$$

Another important error rate was developed by Duncan and called Error Rate Per Degree of Freedom :

$$\alpha_{pdf} = 1 - (1 - \alpha_c)^{p-1}$$

This is infact a compramise between the two error rates mentioned above and error rete per statement.

Error Rate Per Comparison is the error rate which non-multible comparisonist usually employs, and can be defined as the probability that any particular one of the comparisons will be incorrectly declared significant. That is,

$$\alpha_c = \frac{\text{Number of comparisons falsely declared significant}}{\text{total number of comparisons}}$$

This error rate can be regarded as a special case (when the number of comparisons is equal to one in the family) of the Familywise or/and Per Family Error Rates. (Miller, 1966, p. 9)

So far, definitions of various error rates have been presented. The calculations in a hypothetical situations will now be illustrated and relations among them pointed cut. Consider a situatino in which the same experimental design is repeatedly carried out 590 times with 15 statements in each (i.e., treatment A is greater than B, B is less than E, and so forth) giving a total of 8850 statements. Suppose now that 177 out of 8850 statements are false and that they are found in 140 experiments out of 590. Applying the formulas introduced above, the fallowing actual error rates are calculated:

Errer Rate Per Experiment = 177/590 = .30

Error Rate Experimentwise = 140/590 = .24

Error Rate Per Degree Of Freedom (assuming one-way ANOVA with 6 treatment groups) . . . . . = $1 - 98^5$ = .096

Error Rate Per Comparison = 11/8850 = .02

This numerical example help to indicate the following relationships among different error rates.

1. There is a direct relatioship between Error Rate Per Experiment and Error Rate Per Comparison. The formula (2) presented earliner holds true, that is, 30. = 15 x .02

2. The relationship between Error Rate Experimentwise and Error Rate Per Comparison cannot be stated with the same simpli-

city. When Miller gives a numerical example he warns the reader not to derive one from another, nor to use any table (1966, pp. 13-14). Duncan, however, illustrates the relation more extensively in the case of three means. He shows that when $\alpha_c = .05$ the protection level for testing the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ can be as low as .878 and this in calculated by taking the intergral over the acceptance region of the bivariate distribution. (1955, pp. 25-26) It can be noted that the there means case is an over simplification in terms of calculations, since β_1, β_2, β_3 constitute a hexagonal region on a two dimensional space and the third dimension for densities in the bivariate distribution. Obviously, the problem becomes more and more complex as the number of comparisons increases.

3. Error Rate Per Degrees Of Freedom lies between Error Rate Experimentwise and Error Rate Per Comparison, a kind of compramise. In the preceeding example the same author reports the protection level as .9025 when based on degrees of freedom versus .878 when multible t-tests were employed.

4. Error Rate Per Experiment is quite close to Error Rate Experimentwise, and these two will be closer as α_c decreases. This should be intuitively obvious; for example in our illustration, if a total of 9 erroneous statements had been observed instead of 177 ($\alpha_c \cong .001$) it would be reasonable to think that those 9 false statements were scattered in 9 or 8 different experiments. In other words, it is very unlikely that the majority of these statements will fall in only one or two experiments. As a matter of fact, this relation can be formally defined for both dependant and independant cases. For examle, if there are 10 comparisons in an experiment and Experimentwise Error Rate is controllled at the .05 level, then the upper limit of Error Rape Per Experment occur in the case when all statements are wrong (10 x .05 = .50). The other extreme, or the lower bound, accurs in the case of only one erroneous and 9 true statements.[1] Therefore a range for the Expected NUMBER of errors per experiment (Error Rate Per Experiment) can be written as :

Experimentwise ER $\leq$ ER Per Experiment $\leq$ N x Eperimentwise ER

or $.05 \leq$ ER Per Experiment $\leq .50$

In the case of independance, given Experimentwise Error Rate, the Error Rate Per Experiment can be calculated exactly by using the

---

(1) This follows from the definiton of Experimentwise Error Rate which gives the probability of one or more (up to10 in the example) false stetements.

binomial distribution, the mean of the distribution yielding the desired value. That is,

$$\text{ER Per Experiment} = N\left(1 - \sqrt[N]{1 - \text{Experimentwise ER}}\right)$$
$$= 10\,(1 - \sqrt[10]{.95}) = 10\,(1 - .99488)$$
$$= .0512$$

Given an Experimentwise Error Rate, it is usually not possible to derive the Error Rate Per Comparison and this is the place where an average researcher is stymied. From the discussion presented above it is possible to find fairly good approximation by combining 1 and 4, if he really wants to (!). As illustrated through examples, this difficulty arouse because of dependance among the tests. This is considered important and many authors treat dependant and independant multible comparisons under separate headings. The dependance can be best shown by the expansion of the probability of a union in terms of the probabilities of the intersections and in the following illustration a case of three comparisons will be considered,

Given the following events and probabilities,

Event A : Error in the first comparisons; $Pr(A) = \alpha_1$
Event B : Error in the second comparison; $Pr(B) = \alpha_2$
Event C : Error in the third comparison; $Pr(C) = \alpha_3$

We can write the probability of at least one error, Experimentwise ER

$$\text{Experimentwise ER} = Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(A \cap C) - Pr(B \cap C) + Pr(A \cap B \cap C)$$
$$= \alpha_1 + \alpha_2 + \alpha_3 - [Pr(A \cap B) + Pr(A \cap C) + Pr(B \cap C)] + Pr(A \cap B \cap C)^1 \qquad (3)$$

(1) It is interesting to note the case of independance again. Probability of intersections can be expressed as the product of individual probabilities in this case,

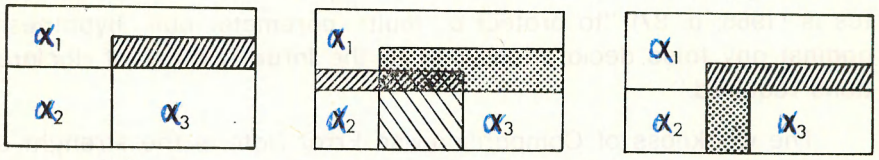
$$\text{Experimentwise ER} = \alpha_1 + \alpha_2 + \alpha_3 - [Pr(A)Pr(B) + Pr(A)Pr(C) + Pr(B)Pr(C)] + Pr(A)Pr(B)Pr(C)$$
$$= \alpha_1 + \alpha_2 + \alpha_3 - [\alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3] + \alpha_1\alpha_2\alpha_3$$

By seting $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$ we have,

$= 3\alpha - 3\alpha^2 + \alpha^3$
$= 1 - 1 + 3\alpha - 3\alpha^2 + \alpha^3$
$= 1 - (1 - 3\alpha + 3\alpha^2 - \alpha^3)$
$= 1 - (1 - \alpha)^3$ and the result is same in the case of binomial approach as shown earlier.

To depict this situation, the area of a rectangle is assumed to stands for Experimentwise ER. Out of the infinite number of dependant cases only three will be considered.

FIGURE I



(1) (2) (3)

As was stated before the diffuculty arises from the undeterministic behaviour of the joint probabilities of the incorectness and, acording to Miller (1966, p. 7), these are "generally mathematically unobtainable in statistical problems". The same author provides more rigorous treatment of the subject matter presented above (pp. 7-8).

So far various kinds of Error Rates and the relations among them have been surveyed. Then the question of choosing between them may be raised. This cannot bet answered properly until the implications of any error rate is determined in the theoretical and practical areas of a discipline, ie psyhology, economy, which may reasonably take decades. However, the relative merits of each one will be discussed briefly.

There is not much to say about Comparisonwise Error Rate[1]. its major advantage is the simplicity and the popular usage in elementary works. In addition to this it yields more powerful tests as compared to others, that is to say through use of this error rate, real differences among comparisons are more likely to be detected. This is, of course, an important factor in many experimental works. As Glass and Stanley state (1970, p. 388), "After all, isn't it a contrast that is primary? And should not the probability of erring in concluding that $\mu_j - \mu_j{}^*$ is or is is not different from zero be uneffoted by how many and which other groups one happens to include in the 'experiment' ". Taking the same position, Wilson (1962, p. 299) expressed his feelings: "Traditional practice apparently has choosen the hypothesis as the unit and this paper maintains that this is the correct choice. It sems that the hypothesis is psychologically

(1) Distinction is not made between this error rate and Hypothesiswise ER for this discussion although they may well be separated (Wilson 1962 p. 297)

more logical unit". The objection to this error rate with respect to Type I error has been cited several times in this short acount. An interesting real-life Monte Carlo Experiment (as oppossed to simulation) was conducted by Kenyon (1965) to show the fallacy of the Comparisonwise Error Rate. But the real objection is that this error rate is against the spirit of multible comparisons, which Miller states is (1966, p. 87) "to protect a multi - paremeter null hypothesis against any false declarations due to the large number of declarations required."

The weakness of Comparisonwise Error Rate is the strenght of the Experimentwise Error Rate. The Type I error is fixed for the whole experiment and is not effected by the number of comparisons to be made. This feature of the error rate facilitates comparison of different experiments carried out under different conditions, but introduces another problem. To make a meaningful comparison depends upon how well the experiment as a unit is defined. What constitutes an experiment may not be adequately answered. To give some aid Miller (1966, p. 34) has introduced the concept of a "Natural Family" by which he means conceptual clarity and applicability of the law of large numbers. This point is very important because, as was stated earlier, when Experimentwise Error Rate is used a single experiment is considered as a Bernoulli trial and the probability statement (error rate) is meaningful only if trials are repeated.[1]

Another popular objection to this error rate is that it generally yields low power in terms of detecting the real differences among contrasts. This is not quite true, because power can be adjucted to any desired level. As Ryan (1959, p. 37) explains, "The issue is not a more or less powerful test, but simply how we are going to evaluate the Type I error." It can be reasonably claimed that the whole purpose is not to detect significant contrasts but find significant contrasts which are meaningful. In a typical experimental design with two or three factors and several levels for each factor, it may be possible to find between 20 and 30 significant differences. Then the question is how these differences will be interpreted within a limited frame of deference in social science theory. On the other hand, using conventional levels of significance α (.01, .05) may result in overlooking some important differences among contrasts

(1) This is why earliler it was stated that to determine the usefulness of an Error Rate may take decades.

and wasting experimenters' efforts. Because of this Scheffe suggests (1959, p. 71) employing the 10 percent significance level.

Error Rate Per Experiment is affected by the number of comparisons while Experimentwise ER is insensitive to this. Another weakness of Experimentwise ER, is that it does not distinguish between experiments with one incorrect statement and experiments with more than one incorrect statements. This is not a shortcoming for ER per Experiment, since it is defined as the expected number of erroneous statements. On the other hand, as the previous numerical example illustrated this error rate is unnecessarily conservative because intersections of individual error rates are not subtructed (formula 3) in calculating total error for simultaneous inference. However, in many instances the two error rates will be very close to each other and in many instances choosing one of them is a matter of individual preference. Common practice favors Experimentwise ER, the reason being quite subjective. Miller (1966, p. 10) expressed this faith as follows "... to the author's clients the thought that all of their statements are correct with high probability seems to afford them a greater serenity and tranquility of mind than a discourse on their expected number of mistakes."

To resolve the power problem, Duncan developed a new concept of protection level. He believes that as the number of comparisons increases more real differences are expected and the power of the test must be high to detect them. He developed the Error Rate Per Degree Of Freedom, the rational for it being is opposite that of Experimentwise ER. The criticism made about ER Per Comparison also apply to Error Rate Per Degree of Freedom, although to a lesser degree. The major objection to Duncan's approach is that statistical considerations rather than the material being tested determine the significance. Usually not all levels of a factor are uqually important for the researcher and those with little importance or irrelevant conditions will spuriously and unnecessarily increase the power but still lower the protection level. The premise of Duncan's argument is based on the researcher's use of an α - level test for each independant comparison, which may or may not be true. In the case where comparisons are related theoritically to each other and combined comparisons would yield better protection, he may not wish to use indepedant test. If the only problem is power it has been mentioned earlier that power can be adjusted to any desired level when using Experimentwise ER. This procedure is not commonly followed in practice, too. For example, the F-test is applied

at the same probability level regardless of the number of treatments. Although it was widely used in mid 1950's, this method did not get much appreciation from statisticians and disappeared from use in the field of multible comparisons practice.

In the literature of ANOVA and experimental designs, discussion of error rates can be found at various levels. To give some examples for brief exposure: Bancroft (1968, pp. 105-106); Dayton (1970, pp. 38-39); Fryer (1966, pp. 264-265); Glass and Stanley (1970, pp. 386-388); Myers (1966, pp. 332-333); Steel and Torrie (1960, pp. 108-109); Winer (1962, pp. 68-69). The subject is more extensively treated in Duncan (1955, pp. 11-19); Harter (1957, pp. 516-521); Kirk (1968, pp. 82-86); and fully presented in Miller (1966, pp. 5-12; 31-35; 89) and Ryan (1959, pp. 28-40)

III. An Outline Of Some Multible Comparison Procedures :

Because of the problematic nature of multible comparisons this fields of statistics has become a very dynamic one and a number of different procedures have been suggested. One new development came from Duncan and his colleagues (Duncan 1965; Waller and Duncan, 1969) who defended the Bayesian approach as a solution. Another procedure proposed by Naik (1969) was based on Arrow's 'Equal Probability' test and heavily dependant on Duncan's 'Loss Functions'. These two developments however were not widely accepted by statisticians. For example Miiler (1966, p. 88) is not willing to accept the additivity of loss functions and says "the statistician wants to prevent **any** false statements under nullity, and one or several incorrect statements should be regarded with just about the same amount of disfavor". Meanwhile, advocates of Experimentwise Error Rate continued their persuations and extended their procedures in terms of coverage. For example, Hodges and Lehman (1968) prepared a table for power of the Tukey test, Sen (1969) showed the generalization of the same test for interactions and Scheffé (1970) extended his procedure to cover ratios as well as means.

Discussion of recent developments in the field and substantial account of all procedures cannot be given in this elementary presentation. However, some of the procedures will be presented in brief, in hope that the previous discussion of error rates will make the following presentation more meaningful.

A. Planned Multible Comparisons: In some research activity there may be specific questions to be answered, and the experiment so designed to serve this purpose. Statistical tests usually are more powerful in these 'a priori' comparisons than in 'data snooping' procedures. Post-hoc methods, may provide useful information for further research, but because of the low power under conventional levels in post-hoc comparison this usefulness is limited. A good discussion of planned versus post-hoc comparisons can be found in Hays (1963) section 14.17.

Planned comparisons appear in two forms; orthogonal and non-orthogonal:

A — I. Planned Orthogonal Comparisons: A comparison or contrast is any linear combination of means in which the sum of the coefficients is zero. That is, for the i^{th} comparison:

$$\hat{\psi}_i = C_{i1}(\bar{X}_2) + C_{i3}(\bar{X}_3) + \ldots + C_{iJ}(\bar{X}_J) \text{ such that } \sum^{J} C_{iJ} = 0$$

Two comparisons (ψ_i, ψ_i') are said to be orthogonal if the product of the coefficients sum to zero. That is, the i'^{th} comparisons are orthogonal if,

$$\sum_{j=1}^{J} \frac{C_{i'j}\ C_{ij}}{n_j} = 0 \text{ or } \sum_{j=1}^{J} C_{i'j}\ C_{ij} = 0 \text{ when n's are equal.}$$

Orthogonal comparisons provide conceptually clear information to the researcher. Interpretation of the analysis is usually straightforward, because each comparison gives non-redundant, non-overlaping information; test results do not produce internal contradiction in the data analysis[1].

The test statistic used in planned orthogonal comparison is the usual t-test and the procedure can be referred to as Multible t-test. For the i^{th} comparison

$$t_i(\upsilon) = \frac{C_{i1}(\bar{X}_1) + C_{i2}(\bar{X}_2) + \ldots + C_{iJ}(\bar{X}_J)}{\sqrt{MS_e\left[\frac{C^2_{i1}}{n_1} + \frac{C^2_{i2}}{n_2} + \ldots + \frac{C^2_{iJ}}{n_J}\right]}}$$

(1) This may not be the case in other procedures. For example in a one-way ANOVA with three groups it is possible to conclude that $\mu_1 = \mu_2, \mu_2 = \mu_3$ but $\mu_1 \neq \mu_3$.

Where C_{ij} is the coefficient of $\bar{X}_j^{th}$ mean for the $_i^{th}$ comparison; n_j is the sample size of the $_j$th group and ME_{error} is the pooled within-group variance. A contrast is declared significant if the absolute value is greater than A where.

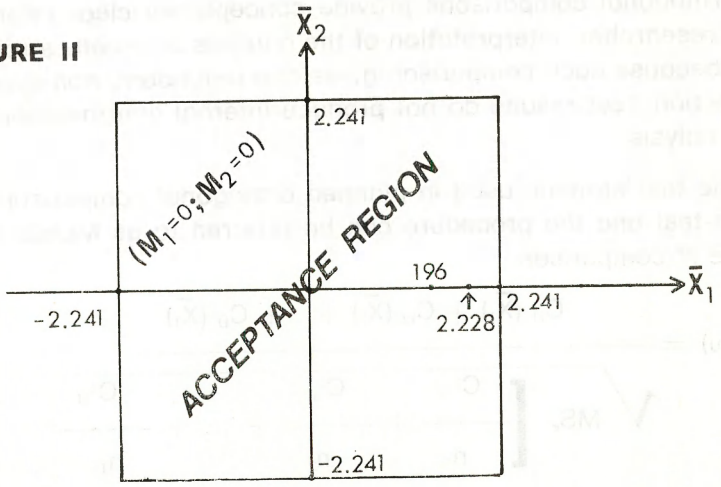
Prob $(\hat{\psi} - A \leqslant \psi \leqslant \hat{\psi} + A) = 1 - \alpha$ and A is defined as

$$A = t\alpha/_2 , \nu \sqrt{MS_e \left[\frac{C^2_{i1}}{n_1} + \frac{C^2_{i2}}{n_2} + \ldots + \frac{C^2_{iJ}}{n_J} \right]}$$

The conceptual unit for error rate in multible t-test is the contrast and Comparisonwise Error Rate is used. A graphical presentation of the case $H_o : \mu_1 = \mu_2 = 0$ is given by Miller (1966, pp. 13-14) and detailed numerical examples can be found in Hays (1963, pp. 478-483), and Kirk (1968, pp. 74-77).

A — 2. Planned Non-orthogonal Comparisons : The method used for this kind comparison is called Bonferroni t or Dunn procedure.[1] The conceptual unit for Error Rate is any collection of contrasts (for example, all comparisons in an experiment) and the type of error rate used is Error Rate Per Experiment or collection of contrasts). Because of the error rate adopted, this procedure gives a less conservative test than the post-hoc procedures. For the special case, $H_o : \mu_1 = \mu_2 = 0$, this can be readily shown by a figure (Miller, 1966, p. 15).

FIGURE II



(1) The Dunn procedure was treated under planned comparisons because, in practice this is the usual place it is used. But there is no theoritical reason why it should not be considered as a podt-hoc comparison.

In order to have a difference declared significant at $\alpha = 0.05$, 1.96, 2.228, 2.241 standart deviations are necessary respectively normal, t_{10}, and Dunn test procedures. A confidence interval for any comparison $\hat{\psi}_i$, Prob $(\hat{\psi} - A \leqslant \psi \leqslant \hat{\Psi} + A) = 1 - \alpha$ where α is Error Rate Per Experiment and A is defined as:

$$A = t'D\alpha/_2 ; C, \nu \sqrt{MS_e \left[\frac{C^2 i_1}{n_1} + \frac{C^2 i_2}{n_2} + \ldots + \frac{C^2 iJ}{n_J} \right]}$$

where C is the number of comparisons, and $t'd\alpha/_2$ is tabulated by Dunn (1961).

B. Post-hoc Methods: These methods can be also summarized in two parts: Significance tests and confidence intervals. Fisher's LSD will not be covered here, since it was discussed at the beginning.

B — I — a. Student-Newman-Keuls: The test employs a layer approach to the set of ordered means. For a difference to be declared significant in this procedure two means must differ by at least.

$$A_r = q_r;r,\nu\sqrt{MS_e/n}$$

where q_r is the Studentized range statistic, and r is the number of steps separating ordered means. The diffuculty with this method is that A_r changes for different r's, and consequently none of the error rates discussed earlier are directly applicabable. The mechanics of testing and numerical examples may be found in Kirk (1968, pp. 92-93) and Winer (1962, pp. 82-84).

B — I — b. Duncan's New Multible Range Test: The testing procedures are similar to those of the previous test but Error Rate Per Degree of Freedom is employed. A difference is declared to be significant it if exceeeds A_r in absolute value.

$$A_r = q_r;r,\nu\sqrt{MS_e/n}$$

where q_r is tabulated by Duncan[1]. A detailed illustration can be found in Edwards (1968, pp. 131-134).

B — 2 — a. Comparison of Control Group With Treatment Groups: Dunnett (1955) has developed a test statistic for this special purpose. The probability associated with the joint confidence inter-

(1) Scheffé (1959, p. 78) points out a pitfall in Duncan's justification, but Duncan (1965, p. 178) claims that there are no mistakes in derivations except for small errors of computational approximation.

val, which is Experimentwise Error Rate, is set. The difference A that a comparison must exceed in order to be declared significant is defined as follows:

$$A = tD\alpha/_2;J,\nu\sqrt{MS_e\ (2/n)}$$

where $tD\alpha/_2$ values were tabulated by Dunnett (1955) and revised by the same author (1964).

B — 2 — b. Tukey Procedure: In the literatarute this method is also called the T-Method or HSD (Honestly Significant Difference). The conceptual unit employed for the test is the experiment, and Experimentwise Error Rate is adopted. This method was originally developed in 1953 for pairwise comparisons among means, then extended for all possible contrasts. The proof of the latter case can be found in Miller (1966, pp. 74-75), or Scheffe 1959, pp. 74-75)

A comparison involving two means is declared significant if it exceeds A, where A is defined as:

$$A = q\alpha,\nu\sqrt{MS_e/n}$$

Alternatively, a confidence interval can be contructed for any pairwise comprasion, $_j$ and $_{j'}$

$$\bar{X}_j - \bar{X}_{j'} \pm q,\nu\sqrt{MS_e/n}$$

The significance test also can be performed by utilizing a Studentized Range test, which may help to show the nature of the difference betweed this procedure and the following one.

$$q = \frac{C_j\bar{X}_j + C_j\bar{X}_{j'}}{\sqrt{MS_e / n}}$$

In case of unequal n's an approximate value can be obtained by using the harmonic mean, given in Bancroft (1968, p. 109), Kirk (1968, p. 90), Steel and Torrie (1960, p. 114). A detailed numerical example for the test is provided by Guenther (1964, pp. 55-57).

B — 2 — c. Scheffé method : In recent years this method has been used frequently because of "its simplicity and versatility over a wide variety of situations" (Hays, 1963, p. 484). Scheffé has proven that the probability is 1 - α that the statement

$$\hat{\psi}_i - A \leqslant \psi_i \leqslant \hat{\psi}_i + A$$

is true simultaneously for all ψ_i where A is defined as :

$$A = \sqrt{(J - 1) F\alpha} \sqrt{MS_e \sum_{j=1}^{J} \frac{(C_j)^2}{n_j}}$$

Proof of this statement canbe found in Scheffe's original paper (1953, pp. 89-90), in his text book (1959, pp. 68-72), or in Miller (1966, pp. 63-66).

Because this method covers all possible contrasts among means, confidence intervals are unnecessarily larger than those of other methods. For example, to compare the method with some others (shown in figure 2) for the on testing of same hypothesis, a difference of 2.87 is required in order to have a contrast be declared significant. In the case of unknown variances the Scheffe test can be also expressed as a ratio

$$Q = \left[\frac{[C_{j1}\bar{X}_1 + C_{j2}\bar{X} + \ldots + C_{jj}\bar{X}_j]^2}{MS_e [C^2_{ji}/n_1 + C^2_{j2}/n_2 + + C^2_{jj}/n_J} \right] (J - 1)$$

where (J — 1) is a constant and the term in the major bracket is an F-ratio. That is, a difference is significant if it exceeds the product of (J — 1) and a theoritical F-value. Numerical illustrations can be found in many text books such as Guenther (1966, pp. 57-59), and Hays (1963, pp. 484-87).

IV. A Brief Comparison Of Various Methods For Muitible Comparions :

It must be emphasized that a direct comparison can not be made because the rational and error rates underlying the procedures vary. Hopking and Chadbourn (1967, p. 409) developed a useful algorithm relating to this problem which is partly reprinted in Glass and Stanley (1970, p. 396).

Hays (1963, pp. 487-489) compares planned orthogonal comparisons with the usual post-hoc comparison methods and shows that the confidence interval for any given comparison is shorter when the comparison is planned than when it is post-hoc.

Two multible range tests, the Duncan and the Student-Newman-Keuls test were treated together in Miller (1966), Winer (1962) and the latter shown to be more conservative[1].

The Dunn procedure since it uses Error Rate Per Experiment can be compared with the Tukey and Scheffe procedures which use Experimentwise Error Rate since these two error rates are very close to each other under typical conditions. Kirk (1968, p. 81) summarizes Dunn's work and indicates that when there are many groups and few comparisons the Dunn procedure yields shorter confidence intervals than the Tukey and Scheffé methods, but when the number of groups is small and a large number of comparisons are to be made than the latter two methods provide shorter intervals. This conclusion makes sense considering the formulas presented before. Namely, confidence intervals in the Tukey and Scheffe methods depend on the number of groups, whereas in the Dunn procedure they are affected by the number of comparisons rather than the number of groups.

Because the Tukey and Scheffé procedures use the same error rate, comparison of these two may make more sense. Scheffé (1953, pp. 91-93 and 1959, pp. 75-77) furnishes the most rigorous discussion. An outline of the results of this study may be found in many introductory text books, e.g. Myers (1966, p. 336). Following is a summary of the comparison of the two methods:

(a). The Tukey method gives a shorter interval for pairwise comparisons, while the Scheffé procedure is more powerful for more complex comparisons.

(b). The Scheffé method does not require equal n's. In an unequal n's case the Tukey method is approximate.

(c). The Scheffé procedure is less sensitive to violations of the normality assumption than is the Tukey method.[2]

.V Conclusions :

In some of the references cited in this paper the authors prefer to use the same data for numerical examples of the different procedures, so that they may be compared more directly. This is useful, of

(1) It should be noted that in the case of two group means, the two procedures give the same result, also equivalent to a t-test.

(2) A latter study by Öngel (1971) has shown that the q statistic used in Tukey Procedure is also roboust against normality assumption.

course, but the lack of emphasis on the various error ratas may be very much misleading for an average reader who wants to learn the subject matter.

For the same reason, it is not meaningful in a research report to state the significant differences without making clear what approach has been employed. Assuming a statistically naive reader has a fixed idea about Type I error (i.e., Comparisonwise Error Rate), interpretation of the results may be completely misleading; for the more knowledgable reader, results may not be interpretable at all.

Because Experimentwise Error Rate has more merit than some other error rates for many research purposes, in some recent publications the authors present only the Tukey and Scheffé methods for multible comparisons. These methods are not sufficient because in these two procedures the confidence intervals are so constructed that all possible contrasts fall within given limits. As the number of comparisons to be made in a given experiment decreases the large confidence interval becomes more and more unnecessary. Research work is very expensive in general and loss of information may be more costly than these writers might think.

This paper has introduced and defined various types of error rates and the most common multible comparisons produres have been outlined using a non-mathematical approach. It is hoped that such an exposure will direct the researcher's attention to the multiple comparison problem which frequently arises in applied research. It is the author's observation that in the published research reports in our country, the multiple comparison testing is either neglected or completely overlooked.

REFERENCE

Bancroft, T. A. **Topics in intermediate statistical methods.** Volume one, Iowa State Iniversity Press, 1968.

Dayton, C. **Design of educational experiments.** New York: McGrow-Hill, 1970.

Duncan, D. B. Multible range and multible F tests. **Biometrics,** 1955, II, I - 42.

Duncan, D. B. A Bayesian approach to multible comparisons. **Technometrics,** 1965, 7, 171 - 222.

Dunn, O. J. Multible comparisons among means. **Journal of the American Statistical Association,** 1961, 56, 52 - 64.

Dunnett, C. W. A multible comparison procedure for comparing several treatments with a control. **Journal of the American Statistical Association,** 1955, 50, 1096 - 1121.

Dunnett, C. W. New tables for multible comparisons with a control. **Biometrics,** 1964, 3, 482 - 491.

Edwards, A. L. **Experimental design in psychological research.** Third edition. New York: Holt, Rinehart and Winston, 1968.

Glass, G. V., and Julian C. S. **Statistical methods in education and psychology.** Englewood Cliffs, N. J.: Prentice-Hall, 1970.

Guenther, W. C. **Analysis of variance.** Englewood Cliffs, N. J.: Prentice-hal, 1964.

Harter, H. L. Error rates and sample sizes for range tests in multible comparisons. Biometrics, 1957, 13, 511 - 536.

Hays, W. L. **Statistics for psychoolgist.** New York: Holt, Rinehart and Winston, 1963.

Hodges, J. L., and Lehman, E. L. A compact table for power of the T-test. **The Annals of Mathematical Statistics,** 1968, 5.

Hopkins, K. D., and Chadbour, R. A. A scheme for proper utilization of multibie comparisons in research and a case study. **American Eductional Research Journal,** 1967, 4, 407 - 412.

Kenyon, G. S. Multible comparisons and the analysis of variance: An emperical illustration. **Research Quarterly of the American Association for Health, Physical Education, and Recreation,** 1965, 36, 413 - 419.

Kirk, R. E. **Experimental Design: Procedures for the behavioral sciences.** California: Brooks/Cole, 1968.

Miller, R. G. **Simultaneous statistical inference.** New York: McGrow-Hill, 1966.

Myers, J. L. **Fundamentals of experimental design.** Boston: Allyn and Bacon, 1966.

Naik, U. D. The equal probability test and its applications to some simultaneous inference problems. **Journal of the American Statistical Association,** 1969, 327, 986 - 998.

Öngel, E., **A Study of the Studentized Range Statistic Under Nonnormality,** Unpublished Ph. D. Thesis, University of Pittsburgh, 1971.

Renner, S. M. A. graphical method for making multible comparisons of frequencies. **Technometrics,** 1969, 2, 321 - 329.

Ryan, T. A. Multible comparisons in psychological research. **Psychological Bulletin,** 1959, 56, 26 - 47.

Ryan, T. A., Significance tests for multible comparison of proportions, varianc-s, and other statistics. **Psycological Bulletin,** 1960, 57, 318 - 328.

Ryan, T. A. The experiment as the unit for computing rates of error. **Psychological Bulletin,** 1962, 59, 301 - 315.

Scheffé, H. A method for judging all contrasts in the analysis of variance. **Biometrika,** 1953, 40, 87 - 104.

Scheffe, H. **The analysis of variance.** New York: Wiley, 1959.

Scheffe, H. Multible testing versus multible estimation. Improper confidence sets. Estimations of directions and ratios. **The Annal of Mathematical Statistics,** 1970, I.

Sen, P. K. A. generalization of the T-method of multible comparisons for interactinos. **Journal of the American Statistical Association,** 1969, 325, 290 - 296.

Steel, R. G., and Torrie J. H. **Principles and procedures of statistics.** New York: McGraw-Hill, 1960.

Waller, A. R., and Duncan, D. B. A Bayes rule for the symmetric multible comparisons problem. **Journal of the American Statistical Association,** 1969, 328, 1484 - 1504.

Wilson, W. R. A note on the inconsistency inherent in the necesity to perform multible comparisons. **Psychological Bulletin,** 1962, 59, 296 - 300.

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

ÖZET

Ülkemizde yayınlanan uygulamalı araştırma raporlarında, uygun hallerde dahi, çoklu-karşılaştırma test yönteminden yararlanılmadığı bir noksanlık olarak değerlendirilebilir. Bu yazıda, değişik çoklu-karşılaştırma yöntemleri betimlenmiş, sayısal örnekler için kaynaklar gösterilmiş ve bu yöntemler için zorunlu geliştirilmiş bulunan yeni hata kavramlarının tanımlanmasına önem verilmiştir.