A significant $F$ in an analysis of variance is simply an indication that not all the population means are equal. It does not tell us which means are different from which other means. As a result, the overall analysis of variance often raises more questions than it answers. We now face the problem of examining differences among individual means, or sets of means, for the purpose of isolating significant differences or testing specific hypotheses. We want to be able to make statements of the form $\mu_1 = \mu_2 = \mu_3$, and $\mu_4 = \mu_5$, but the first three means are different from the last two, and all of them are different from $\mu_6$.

Many different techniques for making comparisons among means are available; here we will consider the most common and useful ones. A thorough discussion of this topic can be found in Miller (1981), Hochberg and Tamhane (1987), and Toothaker (1991). The papers by Games (1978a, 1978b) are also helpful, as is the paper by Games and Howell (1976) on the treatment of unequal sample sizes.

## 12.1 Error Rates

The major issue in any discussion of multiple-comparison procedures is the question of the probability of Type I errors. Most differences among alternative techniques result from different approaches to the question of how to control these errors.\(^1\) The problem is in part technical; but it is really much more a subjective question of how you want to define the error rate and how large you are willing to let the maximum possible error rate be.

We will distinguish two basic ways of specifying error rates, or the probability of Type I errors.\(^2\) In doing so, we shall use the terminology that has become more or less standard since an extremely important unpublished paper by Tukey in 1953. (See also Ryan, 1959; O’Neil and Wetherill, 1971.)

### Error Rate per Comparison (PC)

We have used the error rate per comparison (PC) in the past and it requires little elaboration. It is the probability of making a Type I error on any given comparison. If, for example, we make a comparison by running a $t$ test between two groups and we reject the null hypothesis because our $t$ exceeds $t_{0.05}$, then we are working at a per comparison error rate of 0.05.

---

\(^1\) Some authors choose among tests on the basis of power and are concerned with the probability of finding any or all significant differences among pairs of means (any-pairs power and all-pairs power). In this chapter, however, we will focus on the probability of Type I errors and the way in which different test procedures deal with these error rates.

\(^2\) There is a third error rate called the error rate per experiment (PE), which is the expected number of Type I errors in a set of comparisons. The error rate per experiment is not a probability, and we typically do not attempt to control it directly. We can easily calculate it, however, as $PE = ce$ where $c$ is the number of comparisons and $e$ is the per comparison error rate.
**Familywise Error Rate** ($FW$)

When we have completed running a set of comparisons among our group means, we will arrive at a set (often called a *family*) of conclusions. For example, the family might consist of the statements

$$
\mu_1 < \mu_2
$$

$$
\mu_3 < \mu_4
$$

$$
\mu_1 < (\mu_3 + \mu_4)/2
$$

The probability that this family of conclusions will contain *at least* one Type I error is called the *familywise error rate* ($FW$). Many of the procedures we will examine are specifically directed at controlling the $FW$ error rate, and even those procedures that are not intended to control $FW$ are still evaluated with respect to what the level of $FW$ is likely to be.

In an experiment in which only one comparison is made, both error rates will be the same. As the number of comparisons increases, however, the two rates diverge. If we let $\alpha'$ represent the error rate for any one comparison and $c$ represent the number of comparisons, then

- Error rate per comparison ($PC$): $\alpha = \alpha'$
- Familywise error rate ($FW$): $\alpha = 1 - (1 - \alpha')^c$

(if comparisons are independent)

If the comparisons are not independent, the per comparison error rate remains unchanged, but the familywise rate is affected. In most situations, however, $1 - (1 - \alpha')^c$ still represents a reasonable approximation to $FW$. It is worth noting that the limits on $FW$ are $PC \leq FW \leq ca$ and in most reasonable cases $FW$ is in the general vicinity of $ca$. This fact becomes important when we consider the Bonferroni tests.

**The Null Hypothesis and Error Rates**

We have been speaking as if the null hypothesis in question were what is usually called the *complete null hypothesis* ($\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$). In fact, this is the null hypothesis tested by the overall analysis of variance. In many experiments, however, nobody is seriously interested in the complete null hypothesis; rather, people are concerned about a few more restricted null hypotheses, such as ($\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, $\mu_6 = \mu_7$), with differences among the various subsets. If this is the case, the problem becomes more complex, and it is not always possible to specify $FW$ without knowing the pattern of population means. We will need to take this into account in designating the error rates for the different tests we shall discuss.

---

1. This error rate is frequently referred to, especially in older sources, as the “experimentwise” error rate. However, Tukey’s term “familywise” has become more common. In more complex analyses of variance, the experiment often may be thought of as comprising several different families of comparisons.
A Priori versus Post Hoc Comparisons

It is often helpful to distinguish between a priori comparisons, which are chosen before the data are collected, and post hoc comparisons, which are planned after the experimenter has collected the data, looked at the means, and noted which of the latter are far apart and which are close together. To take a simple example, consider a situation in which you have five means. In this case, there are 10 possible comparisons involving pairs of means (e.g., $\bar{X}_1$ versus $\bar{X}_2$, $\bar{X}_1$ versus $\bar{X}_3$, and so on). Assume that the complete null hypothesis is true but that by chance two of the means are far enough apart to lead us erroneously to reject $H_0$: $\mu_i = \mu_j$. In other words, the data contain one Type I error. If you have to plan your single comparison in advance, you have a probability of 0.10 of hitting on the 1 comparison out of 10 that will involve a Type I error. If you look at the data first, however, you are certain to make a Type I error, assuming that you are not so dim that you test anything other than the largest difference. In this case, you are implicitly making all 10 comparisons in your head, even though you perform the arithmetic for only the largest one. In fact, for some post hoc tests, we will adjust the error rate as if you literally made all 10 comparisons.

This simple example demonstrates that if comparisons are planned in advance (and are a subset of all possible comparisons), the probability of a Type I error is smaller than if the comparisons are arrived at on a post hoc basis. It should not surprise you, then, that we will treat a priori and post hoc comparisons separately. It is important to realize that when we speak of a priori tests, we commonly mean a relatively small set of comparisons. If you are making all possible pairwise comparisons among several means, for example, it won’t make any difference whether that was planned in advance or not.

Significance of the Overall $F$

Some controversy surrounds the question of whether one should insist that the overall $F$ on groups be significant before conducting multiple comparisons between individual group means. In the past, the general advice was that without a significant group effect, individual comparisons were inappropriate. In fact, the rationale underlying the error rates for Fisher’s least significant different test, to be discussed in Section 12.4, required overall significance.

The logic behind most of our post hoc tests, however, does not require overall significance before making specific comparisons. First of all, the hypotheses tested by the overall test and a multiple-comparison test are quite different, with quite different levels of power. For example, the overall $F$ actually distributes differences among groups across the number of degrees of freedom for groups. This has the effect of diluting the overall $F$ in the situation where several group means are equal to each other but different from some other mean. Second, requiring overall significance will actually change the $F$W, making the multiple-comparison tests conservative. The tests were designed, and their significance levels established, without regard to the overall $F$.

Wilcoxon (1987a) has considered this issue and suggested that “there seems to be little reason for applying the (overall) $F$ test at all” (p. 36). Wilcoxon would jump straight to multiple comparisons without even computing the $F$. Others have said
12.2 Multiple Comparisons in a Simple Experiment on Morphine Tolerance

In discussing the various procedures, it will be helpful to have a data set to which each of the approaches can be applied. We will take as an example a study similar to an important experiment on morphine tolerance by Siegel (1975). Although the data are fictitious and a good deal of liberty has been taken in describing the conditions, the means (and the significance of the differences among the means) are the same as those in Siegel's paper. It will be necessary to describe this study in some detail, but the example is worth the space required. It will be to your advantage to take the time to understand the hypotheses and the treatment labels.

Morphine is a drug that is frequently used to alleviate pain. Repeated administrations of morphine, however, lead to morphine tolerance, in which morphine has less and less of an effect (pain reduction) over time. (You may have experienced the same thing if you eat spicy food very often. You will find that the more you eat it, the hotter you have to make it to taste the way it did when you started.) A common experimental task that demonstrates morphine tolerance involves placing a rat on an uncomfortably warm surface. When the heat becomes too uncomfortable, the rat will lick its paws, and the latency of the paw-lick is used as a measure of the rat's sensitivity to pain. A rat that has received a morphine injection typically shows a longer paw-lick latency, indicating a reduced pain sensitivity. The development of morphine tolerance is indicated by a progressive shortening of paw-lick latencies (indicating increased sensitivity) with repeated morphine injections.

Siegel noted that there are a number of situations involving drugs other than morphine in which conditioned (learned) drug responses are opposite in direction to the unconditioned (natural) effects of the drug. For example, an animal injected with atropine will usually show a marked decrease in salivation. If, however, after repeated injections of atropine, physiological saline (which should have no effect whatsoever) is suddenly injected (in the same physical setting), the animal will show an increase in salivation. It is as if the animal were compensating for the anticipated effect of atropine. In such studies, it appears that a learned compensatory mechanism develops over trials and counterbalances the effect of the drug. (You experience the same thing if you leave the seasoning out of food that you normally add seasoning to. It will taste unusually bland, though the Grape Nuts you eat for breakfast does not taste bland.)

Siegel theorized that such a process might help to explain morphine tolerance. He reasoned that if you administered a series of pretreatments in which the animal was injected with morphine and placed on a warm surface, morphine tolerance would develop.