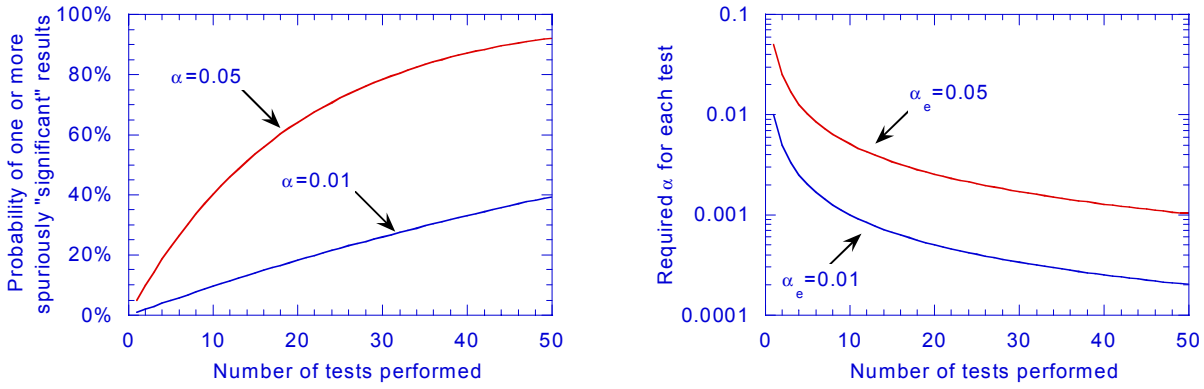
Researchers can control the risk of "false positives" (that is, the risk of rejecting the null hypothesis when it is in fact true) by setting α at an appropriate value. It is crucial to realize that α expresses the risk of a "false positive" occurring in a *single* test, *not* the risk of a false positive in *multiple* statistical tests. When each of the statistical tests has been designed in advance, and when each tests a specific *a priori* hypothesis, then the risks of false positives are accurately quantified by α in each case. But one often sees *multiple unplanned* statistical tests being analyzed as if the overall risk of a false positive in *any* of the tests is α. In fact, it is much higher. In statistical "fishing expeditions", α can drastically underestimate the risk of false positives.

Statisticians distinguish between two different "Type I" (or "false positive") error rates. The *comparisonwise* error rate is the risk of a "false positive" occurring in an *individual* statistical test. This is expressed by the familiar α or statistical significance level. The *experimentwise* error rate, in contrast, is the risk that one or more "false positives" will occur *somewhere* among the many statistical tests that make up an experiment. This is known by many different symbols; we will use α_e. The experimentwise error rate obviously depends on both the comparisonwise error rate (the error rate per comparison or test) and the number of comparisons that make up an experiment. If α is the comparisonwise error rate in one test, then the probability that a Type I error will *not* occur in that test is $1-\alpha$. If we conduct these tests a total of k times, then the probability that a Type I error will not occur in *any* of the k comparisons is $(1-\alpha)^k$. Thus, the experimentwise error rate (the chance of one or more Type I errors) is,

$$\alpha_e = 1-(1-\alpha)^k \qquad (1)$$

As a first approximation, the experimentwise error rate is approximately the error rate per test, times the number of tests; that is, $\alpha_e \approx k\alpha$ for $k\alpha<0.25$ or so (see table on other side). Statistical "fishing expeditions" can easily have experimentwise error rates that greatly exceed α, as the figure on the left shows:



Multiple tests can arise in many different ways:

a) Testing many different attributes, to see whether any of them exceed standards, or differ between groups. (For example, testing whether levels of any of k different organochlorides are elevated in drinking water, or whether incidence of any of k different cancers are higher in an exposed group than a control group.)

b) Testing many different groups to see whether any of them look unusual. (For example, examining k different cities to see whether any of them exceed air pollution standards.)

c) Testing many different groups to see whether any of them differ from the overall average. (This should be done by ANOVA, followed by the Honestly Significant Difference ("Tukey Test"), not by repeated application of the t test).

d) In case-control studies, testing a large number of characteristics to see whether any of them explain the difference between cases and controls. (For example, testing whether any of k lifestyle characteristics are found more often in cancer victims than in the population at large).

e) Fishing for "significant" correlations in large correlation matrices. (For example, looking for "significant" relationships between any of, say, 50 different environmental pollutants and incidence of 10 different childhood diseases, whereupon k=50*10=500!).

f) Meta-analyses that review the results of large numbers of published studies on a given subject, and summarize the "significant" results (which could arise spuriously if a large enough number of studies were undertaken). "Publication bias" (negative results are rarely published) exaggerates this problem.

When each individual comparison is planned in advance (that is, when there is a specific hypothesis for each comparison), researchers generally agree that you should set α=0.05 (or whatever your desired Type I error rate is) for each statistical test, regardless of how many there might be. That is because there is a specific hypothesis under test in each case. Therefore, each hypothesis is meaningful and interesting in its own right, and your risk of falsely accepting each null hypothesis should be α.

The situation is *fundamentally different* when the tests are unplanned (that is, when the hypotheses are made up after looking at the data and picking out the relationships that look strongest or most interesting). If you do that, whether you realize it or not, you are actually testing *only one overall hypothesis* that states that *some one or more* of the (many) possible relationships is non-zero. To properly control the risk of "false positives" here, the relevant alpha is the experimentwise alpha, α_e, which describes the risk of *any* relationships spuriously appearing, from among all your possible relationships.

So, how do you set α for each relationship tested, such that the risk of *any* false positives among the whole set of k possible tests is no more than α_e? There are many different specialized techniques for answering this question; see Sokal and Rohlf or Light et al. (references below), or any good book on experimental design for all the details. But in many situations, there is a fast-and-easy general rule that's a good (though sometimes conservative) approximation. Just invert equation (1) so that,

$$\alpha = 1 - \sqrt[k]{1-\alpha_e} \approx \frac{\alpha_e}{k} \qquad (2)$$

For any $\alpha_e << 1$, (say, α_e=0.05), the k^{th} root of 1-α_e is to a very good approximation simply 1-α_e/k, so the approximation in equation (2) is nearly exact (see table below). As the figure on the right above shows, the significance level α required in each comparison will often be much, much smaller than the overall experimentwise error rate α_e. The table below shows, for different numbers of unplanned comparisons, the experimentwise error rate that results if each test is conducted at α=0.05, and conversely, the α that must be used in each individual test in order to keep the experimentwise error rate at α_e=0.05 for the whole series of comparisons:

number of tests (k)	experimentwise error rate if α=0.05 (α_e)	comparisonwise error rate needed to keep $\alpha_e \le 0.05$ (α)
1	0.05	0.05
5	0.22	0.010
10	0.40	0.0051
15	0.54	0.0034
20	0.64	0.0026
50	0.92	0.0010
100	0.994	0.00051

Now for two bits of bad news. Bad news item #1: the very small α values required for large numbers of tests can result in a *severe loss of power* (particularly when sample sizes are small).

Bad news item #2: in the equations and figures above, k is the total number of relationships you *could have* tested, *not* the number of formal statistical tests you actually *performed*. If you sift through, say, 50 different relationships, but only five of them look "important" or "interesting", then the relevant k is 50 (not five), and α will need to be $\le$0.001 (not $\le$0.01) in order to keep $\alpha_e \le 0.05$. The reason is that when you look at the data and visually filter out the most "interesting" relationships, you implicitly perform a large number of tests, even though you only formally test the few relationships that look strongest. If you've scanned a large enough number of possibilities, one or more of them could look "significant" entirely by chance--which is what you're trying to guard against--even though you can probably come up with a plausible explanation for the "significant" relationship after you know which one it is!

For further reference:

Light, R., J. Singer and J. Willett, *By Design*, Harvard University Press, Cambridge, MA, 1990.

Sokal, R. R. and F. J. Rohlf, *Biometry*, W. H. Freeman, New York, 1981, pp. 241 ff.

Stevens, J., *Applied multivariate statistics for the social sciences*, Lawrence Erlbaum Assoc's., Hillsdale, New Jersey, 1992, pp. 6-9.