4. Styles of Rhetoric

The investigator's usual desire to have strong results can exert a biasing influence on his or her presentation of a study's outcomes. The indicators of effect magnitude discussed in chapter 3 can often be exaggerated in the telling, and investigators differ in their tendencies toward exaggeration or understatement of results. Stylistic variations are most obvious in the treatment of p values. Despite warning the reader of overattention to ps, this chapter devotes attention to them because they are ubiquitous in the social science literature. It is clear, though, that over- or underexaggeration tendencies also apply to other magnitude indicators. Some researchers, for example, call correlation coefficients of .35 "modest," whereas others label them "substantial." Furthermore, investigators have a great deal of leeway to pick and choose which aspects of the results to emphasize. As Rosenthal (1991) said in reference to the reporting of studies, "a fairly ambiguous result often becomes quite smooth and rounded in the discussion section" (p. 13).

BRASH, STUFFY, LIBERAL, AND CONSERVATIVE STYLES

Consider the situation of the investigator obsessing over prospective significance tests. Aside from a few rare examples such as those given in chapter 2, acceptance of the null hypothesis is usually a dull outcome that seems to brand the research a waste of time. On the other hand, rejection of the null hypothesis suggests that some systematic, explanatory factor is causing the groups to be different. "Yes, there is a group difference favoring Group A over Group B" is clearly a more satisfying statement than "there could be a difference favoring Group A over Group B, but that result cannot be claimed with confidence."

It is only natural for researchers to prefer saying something of substance to saying something vapid. Negative results often do not even get written up. Students may abandon dissertations because null hypotheses cannot be rejected. Faculty members, journal editors, and other readers of manuscripts often react less kindly toward accepted than rejected null hypotheses (see Greenwald, 1975, and Rosenthal, 1979, for a discussion of these attitudes and their consequences). Given all this, it is very tempting for people to try desperately to make their results come out statistically significant. In the peculiar language of American politics, one might say that investigators try to exert "spin control" over the interpretations of their results. This is especially true for outcomes that are almost significant, say, .05 < p < .10. The game then becomes one of somehow pushing the results toward or beyond the conventional p = .05 level.

The Brash Style

What devices are available to the desperate researcher for arguing that the results look good, when a dispassionate observer would say they are marginal or worse? There are at least five:

1. Use a one-tailed test.
2. When there is more than one test procedure available, use the one producing the most significant results.
3. Either include or exclude "outliers" from the data, depending on which works better.
4. When several outcomes are tested simultaneously, focus on the one(s) with the best p value(s)—the "hocus focus" trick.
5. State the actual p value, but talk around it.

Sophisticated readers of research reports are of course aware of these devices, and will recognize blatant attempts by the author to make weak p values sound strong. The net result of the usual built-in motivation to reject null hypotheses is thus to create a temptation to overstate one's results, balanced by the risk that one will be found out, at some loss to one's immediate case and long-run reputation. In terms of our legal analogy, this is the dilemma faced by a lawyer who is offered a sizable fee to defend a client of questionable virtue. With enough inventiveness, a case may be drawn in defense of almost any client; but to manufacture flimsy arguments may eventually justify the attribution that one is a shyster. This is an especially damaging attribution for a scientific researcher, which is why we emphasize that researchers should be like honest lawyers.

We refer to the rhetorical style that overstates every statistical result as brash. Investigators who use the five devices previously listed, freely and inappropriately, invite skepticism and disfavor.
The Stuffy Style

Life would be simple if one could give the unqualified advice to the statistics user, "Never be brash." Among other things, this would imply total abstinence from any of the five devices that enhance the argument for rejection of the null hypothesis. Indeed, some statistics texts and some statistics instructors come close to this flat-out admonition. Taking the five devices one by one, the most complete injunction against brashness would be as follows:

1. Never use one-tailed tests.
2. Only use a single, predetermined analysis for any data set.
3. Never exclude outliers.
4. Avoid special focus on any particular result, especially if it is favorable.
5. Stick strictly to a fixed significance level, for example, .05, and make no distinctions between outcomes that nearly beat it ($p < .06$, say), and those far from significance.

When such proscriptions are packaged together, the net effect is to make statistical analysis into a set of legal or moral imperatives, such as might be announced at a public swimming pool (ABSOLUTELY NO DOGS OR FRISBEE ALLOWED. VIOLATORS WILL BE PROSECUTED.) The teaching and learning of statistics as a series of "don'ts" can be so intimidating that students are often heard to ask the question, "Can I do this analysis?" as though seeking permission from a higher authority who will guarantee their immunity from prosecution. This style of thinking about statistics goes to the opposite extreme from the excesses of brashness. We refer to it as the stuffy style of rhetoric. In Tukey's (1969) terms, this approach views statistics as a ritual of sanctification, destroying the exercise of statistical detective work.

As we see in this and subsequent chapters, there are sometimes good reasons to depart from the stuffy style, and the student who acts on these reasons should not be "left with feelings of dishonesty and guilt at having violated the rules" (Gigerenzer, 1993, p. 326). To the student who asks, "Can I do this?" a reasonable answer is: You can do anything you choose, and ponder the potential meaning of the results for your research. But keep in mind that the way you present the outcome(s) will affect the persuasiveness of the case you make. Usually it's not good to be too brash, but you don't want to be so pompous that you avoid the most cogent analysis. (At this point the student will probably say, "Yes, but can I do this?" whereupon you should send out for pizza and repeat your advice as many times as necessary.)

Liberal and Conservative Styles

The two extremes of unrestrained brashness and stultifying pomposity bracket a dimension along which styles of approach to statistics may vary. A liberal style emphasizes exploration of, and speculation about, data. It is looser, more subjective, and more adventurous. A conservative style is tighter, more codified, and more cautious. In statistics as in politics, either style can be defended, and there are individual differences in preference. Also as in politics, the most successful arguments are those that satisfy both liberals and conservatives. This will happen when the investigator's substantive claims are backed by conservative procedures, because in that case the claims would also be warranted by liberal standards. Contrariwise, when a liberal approach produces null results, a conservative approach will, too.

Debatable cases arise when null hypotheses are rejected according to liberal test procedures, but accepted by conservative tests. In these circumstances, reasonable people may disagree. The investigator faces an apparent dilemma: "Should I pronounce my results significant according to liberal criteria, risking skepticism by critical readers, or should I play it safe with conservative procedures and have nothing much to say?" In the next few sections, we sketch some of the contextual features tilting the resolution one way or the other, key to our five rhetorical devices.

ONE-TAILED, TWO-TAILED, AND LOPSIDED TESTS

Consider the $t$ test for the significance of the difference between the means (A and B) of two groups. This test is usually conducted on a two-tailed basis. If $t$ has a large positive value (with A notably larger than B), or a large negative value (B larger than A), there are grounds for rejecting the null hypothesis. The rejection region—the set of outcomes that will count against the null hypothesis—is divided equally between the positive and negative tails of the $t$ distribution. For the 5% level, each tail contains 2.5% of the total area under the $t$ curve.

When there is a strong theoretical expectation that any group difference will be in a given direction (say, $A > B$), some investigators and textbook writers consider it permissible to use a one-tailed test, that is, to concentrate the rejection region all in the predicted tail (say, the upper 5% of the positive tail). The paradigmatic example of a one-tailed situation is the test of whether some new medical or psychological treatment, or educational improvement plan (A) produces results superior to the standard treatment or practice (B). Because only the outcome $A > B$ is interesting, that is, will lead to a therapeutic or educational reform, it seems sensible to concentrate attention solely on the upper tail of the $t$ distribution.
If the data do fall in the upper tail, the one-tailed test is more liberal than the two-tailed test. The one-tailed 5% level corresponds to the two-tailed 10% level, thus if \( t = 1.80 \), say, with 36 df, the result is significant by a one-tailed test (critical \( t = 1.69 \)), but not with a two-tailed test (critical \( t = 2.03 \)). Investigators with a generally liberal style are thus prone to view one-tailed tests permissively, whereas conservatives frown upon them.

**The One-and-a-Half Tailed Test**

Whatever one's general style, the critical question with a one-tailed test is what you would say if the results fell far out in the "wrong" tail. Suppose that in using a new therapeutic procedure, Group A, the experimental group, does far worse than the control Group B—so much worse, in fact, that using a two-tailed test, one would reject the null hypothesis and declare the new treatment significantly detrimental. If you start with the intention of using a one-tailed test, but switch to a two-tailed test when faced with a surprising data reversal, your rejection region has 5% in the upper tail and 2.5% in the lower tail. With such a rejection region, the probability that a true null hypothesis will be rejected is .075, not the nominal .05. This is a bit too brisk; the stated 5% level is misleading.

We might call this a "one-and-a-half tailed test," and note that it can also arise if you start with the intention of using a two-tailed test, but switch to a one-tailed test argument if the data show A>B, but not quite strongly enough to fall in the upper 2.5% tail. This happens not infrequently, because researchers often find particular directional expectations "obvious" (more stimulus rehearsal improves memory, e.g., or more highly educated citizens show greater political interest, etc.). The potential slipperiness of inducing arguments after the fact has to some extent given the one-tailed test a bad reputation.

What, then, distinguishes the defensible one-tailed test from the misleading one-and-a-half tailed test? Clearly, it is the totally blank lower tail of the former. That is to say, a one-tailed test is only well justified if in addition to the existence of a strong directional hypothesis, it can be convincingly argued that an outcome in the wrong tail is meaningless and might as well be dismissed as a chance occurrence.\(^1\)

This condition is extremely difficult to meet, because researchers are very inventive at concocting potential explanations of wrong-tailed results. Results in the wrong tail are rarely easy to simply ignore, as a pure one-tailed test would require.

**The Lopsided Test**

A compromise between the two brands of \( t \) test is possible. Consider a rejection region of 5% in the expected tail, and .5% in the supposedly wrong tail. Tests based on this region would be liberal in the expected direction, and conservative in the wrong tail. Such a test might be called a lopsided test. The rejection rule would be to declare the result statistically significant if it fell in the expected tail with two-sided \( p < .10 \), or in the unexpected tail with two-sided \( p < .01 \). This lopsided test entails a .055 probability that a true null hypothesis would be rejected, close to what the one- and two-tailed tests specify, but melding the virtues of those two alternatives.

**ALTERNATIVE TESTS APPLIED TO THE SAME DATA SET**

There are many other situations in which alternative test procedures can be applied to the same data set, to the potential confusion of the investigator. Nowadays, such confusion is amplified by the tendency for statistical analysis packages such as SAS (SAS Institute, 1985), to include the results of a plethora of alternative tests, some of which very few people ever heard of. How is one to cope with this unwelcome freedom of choice? To do different tests and pick the one that comes out best smacks of cheating, but if only a single test is to be done, which one should be preferred? We do not try to catalogue every such situation, but we broach the issues involved.

We distinguish three cases: one that involves different ways of expressing the data; the second, different calculation formulas applied to the same data; the third case, different philosophies for framing the analysis of the data configuration.

**Different Modes of Data Expression**

**Parametric Versus Nonparametric Tests.** In the simple situation of two independent samples that could be \( t \)-tested for a difference between means, the argument could be made that because of the potential lack of fit of normal distributions to the data, it is preferable to use the Mann–Whitney (1947) or the (equivalent) Wilcoxon (1945) test—a "nonparametric procedure" that avoids the normality assumption by replacing the original quantitative measures with their rank orders. To make matters still more complicated, there are several other

---

\(^1\)A different consideration applies to one-tailed tests in the application of meta-analysis (Rosenthal, 1991): This technique examines the strength and consistency of direction of a set of results from studies on the same topic. The meta-analyst must keep consistent track of the direction of each study. Therefore, in the amalgamation of \( p \) values, a one-tail orientation for the typical result is appropriate. A null result in any study is entered as \( p = .50 \), and a result at, say, the 5% value for the wrong tail is taken as \( p = .05 \).
nonparametric tests applicable to the same simple data situation (e.g., the exceedance test (Tukey, 1955); see also Cliff, 1993; Madansky, 1988).

The argument in favor of nonparametric tests as protection against non-normal distributions has been undercut by demonstrations (Sawilowsky & Blair, 1992) that the t test is fairly "robust" against (i.e., insensitive to) violations of the assumption of normal distributions. A similar statement can be made for F tests comparing the means of several groups. Besides the weakening of the original motivation for using them, non-parametric tests have also suffered from a lack of ability to articulate detail in presenting results (see chap. 6 on contrasts, comparisons, and beyond). Thus the general preference in the social sciences has been for parametric procedures. 3 A standard exception occurs when the data speak with overwhelming strength to a simple major point, and one merely wants to state an acceptably significant p value in the most straightforward way. In such cases, simple methods like the median test (see, e.g., Siegel, 1956) may be handy.

Original Versus Transformed Data (Between-Group Analyses).
For better articulation (chap. 6), or to repair heterogeneous variances among t distributions, data are sometimes transformed to a different scale; for example, every score may be replaced by its logarithm. Significance tests performed both on the original data and on the transformed data typically yield very modest differences in p values, with the transformed version yielding a more conservative result. The safest procedure is to use the transformed scale if it homogenizes variance, although this creates a stylistic irony: The conservative analyst naturally prefers the more conservative significance test, but may be uncomfortable using a transformation, because the units of the new scale may not seem as "real" as those of the original scale. It is difficult to get a feel for the logarithm of the time taken to solve a problem, for example, or the square root of the number of errors on a task. By refraining from the transformation, however, the conservative researcher not only passes up the other advantages of the transformation, but is stuck with a generally more liberal significance test. Thus he is hoist with his own petard. 3

Original Versus Transformed Data (Repeated Measures Analyses).
In repeated measures designs, that is, situations in which subjects are exposed to more than one experimental treatment or are observed for

more than one trial, the usefulness of transformations is sharply increased. In contrast to the case of between-groups analyses, F tests in repeated measures designs are highly vulnerable to violations of underlying assumptions, and in cases with such violations, a transformation may be imperative. Here, unlike the simple between-groups case, the p values on the original and transformed data can be wildly discrepant. If one is unfamiliar with this pathology in repeated measures designs, such an outcome can seem bizarre. Here again, the transformed scale that roughly equalizes variances is to be preferred, more urgently than in the between-groups case.

Tunes Versus Speeds. A prototypical case in which a transform is vital arises in experiments in which time to complete a simple, unfamiliar task is measured over a number of learning trials. Times tend to be highly variable from subject to subject in early trials, but become less variable as subjects learn the task. A transformation that often tends to equalize variances for different trials is the reciprocal transformation— which turns times into speeds. As a case in point, I have seen a data set of times for rats to run an alley for different sized rewards, in which the variances were quite heterogeneous (a ratio of 200 to 1 for the first trial vs. the last trial). When the null hypothesis of no reward effect was tested on these data, the p value was > .05; when the times were transformed to speeds and the data reanalyzed, the p value was < .01.

In this case, speed is the proper measure, and time is inappropriate. For designs involving replications of conditions within individuals who have different variances, see Bush, Hess, and Wohlford (1993).

Absolute Versus Relative Effects

Example: Negative Persuasion Using Insults. A common type of conundrum arises when an effect can be regarded in either absolute or relative terms. This case gives rise to a difficult stylistic issue.

We begin with an example, a slightly simplified version of a field experiment conducted by Abelson and Miller (1967). Those investigators wanted to test for "boomerang effects" in persuasion, that is, circumstances in which a speaker not only fails to persuade his audience, but actually drives their attitudes further away from his position. This phenomenon almost never occurs in laboratory studies with captive, polite audiences. In the field experiment, the situation was engineered so that the speaker presented his arguments in an insulting fashion.

The hypothesis was that this would produce a boomerang effect.

4This experiment was performed during a period in which misgivings about the ethics of psychological experiments were not as widespread as they are nowadays.
The persuasion situation was prepared by having a confederate seat himself next to a potential subject on a park bench, whereupon the experimenter arrived as a roving reporter seeking the spontaneous opinions of ordinary citizens on a public issue. After obtaining agreement to participate, the experimenter invited the discussants to check their initial opinions on a 21-point rating scale and then to take turns offering arguments for their positions. In the Insult condition, the confederate ridiculed the subject's opinions before giving his own prepared views. In the Control condition, the insulting remarks were omitted. After six turns by each participant, ratings on the opinion scale were again solicited by the experimenter.

Table 4.1 presents the mean opinion scores before and after debate, and their differences, for each condition. A negative sign for the difference indicates a change away from the confederate's side of the issue, that is, a boomerang effect; a positive sign indicates change toward the confederate, an accommodation to his expressed views.

The Insult group indeed demonstrates a boomerang effect; a $t$ test of the mean change score of -1.81 against the zero chance null hypothesis yields the significant value $t_{(24)} = -2.16$. But what about the No Insult control group? The conventional wisdom is that when change is measured in an experimental group and a control group, one should test the experimental change relative to the control change. The difference between the mean change scores of the Insults and No Insults groups is -2.25, bigger than the mean change score of -1.81 for the Insults group alone. Yet the $t$ test for this bigger difference is now smaller than the previous $t$, and not significant ($t_{(20)} = -1.90$)!

5For those who wish to verify this numerical result, we supply the additional fact that the pooled within group mean square was 16.86, which with $n = 24$ per group yields the standard error of each mean change as .838. The rest follows from standard $t$-test formulas.

What's going on here? At first blush, an outcome like this appears to violate common sense, because a bigger difference yields a smaller $t$. The sophisticated reader who has seen this sort of thing before will realize that the paradox arises because the difference between two changes has a bigger standard error than either one alone. The relative effect is statistically more unstable than the absolute effect, and therefore requires a stronger result to achieve the same level of significance.

The 42% Rule. Let us examine this phenomenon in more detail, in a more general context. Consider a $2 \times 2$ design, with the columns representing some consequential comparison or difference, and the rows, the presence or absence of a key experimental condition. The investigator is interested primarily in the comparative difference (i.e., column effect) in the presence of the key condition (the first row); she has included the second row as a control, but expects it to show no column difference.

Now suppose that the experimenter $t$ tests the comparison separately for the two rows and finds a significant effect for the first row, but no effect whatsoever for the second row. It is tempting to stop there, declare victory, and write it up for publication. But that would contradict the logic of including a control comparison in the first place. The point of running the control condition is to test the relative claim that the effect in the presence of the experimental factor exceeds the effect in its absence. The appropriate test seems to be a test of the interaction between the rows and the columns.

Imagine the two $t$ tests, $t_{(1)}$ and $t_{(2)}$, performed respectively on the two rows. Suppose further that the interaction, conventionally reported with an $F$ value, is converted to a $t$ value, call it $t^{(*)}$. (The square root of the $F$ yields the $t^{(*)}$.) Assuming that the same pooled error term is used for all $t$ tests, it can be shown that:

$$t^{(*)} = \frac{t_{(1)} - t_{(2)}}{\sqrt{2}}$$

(1)

This innocent equation produces the paradoxical type of result. The interaction test can lead to acceptance of the null hypothesis of no relative difference in the column comparisons for the two rows, even when the test for the first row, $t_{(1)}$, comes out significant, and for the second row, $t_{(2)}$, yields a flat zero. In this case, the equation tells us that $t^{(*)}$, the interaction test statistic, is smaller than $t_{(1)}$, the simple effect for the first row, by a factor of the square root of two. Further tinkering with the equation creates the 42% Rule.5 When a mean difference in one

5The critical percentage is actually 41.43%, but I have rounded it up to 42%. Readers familiar with A Hitchhiker's Guide to the Galaxy (Adams, 1980) will recognize the cosmic status of the number 42. In Adams's most powerful computer in the universe is set working on the question of the meaning of life, the universe, and everything. After millions of years of contemplation, the computer—to everyone's dismay—prints out "42." This interesting number was first noticed (Feynman, 1965) as the exponent of the ratio of certain fundamental physical quantities, and more recently (“Egad!” in Adams, 1992) has turned up in the vegetable garden. It was also Jackie Robinson’s uniform number. (But don't take any of this too seriously.)
subsample is contrasted with a mean difference in a second, equally sized subsample, the t statistic for the interaction is larger than the t for the first difference if and only if the second difference is in the opposite direction, and is at least 42% as big as the first.

The Insult/No Insult data provide a pernicious example of the 42% rule. Look back at Table 4.1. The No Insult mean change score was in the opposite direction from the Insult mean change score, all right, but it was less than 42% as big in absolute size \(|t(44/1.81) = 243|\), forcing \(t^*\) to be less than \(t(1)\).

The Stylistic Conundrum. I faced a stylistic conundrum when submitting this study for publication. Using the .05 level, the Insult group by itself showed a significant boomerang effect, but when the Insult group was compared to the No Insult group, the boomerang finding was not significant. This was despite the fact that the No Insult group, taken alone, demonstrated nothing that would weaken the claim of a boomerang effect from insults. It yielded a banal and nonsignificant tendency of moving toward the speaker’s position. As we have seen, the villain of the piece is the statistical fact that every time you introduce new observations for purposes of comparison, the variability of the summary statistic (here, the mean difference) goes up, by a factor of the square root of 2 (if the \(n\) of additional cases = the \(n\) of old cases.)

Well, what did I do? I could have brashly tried to argue that this comparison with a No Insult control group was irrelevant. Editors and readers, however, don’t generally like experiments with no control group. My eclectical solution was to give the comparisons of the Insult group mean change score with both the zero baseline \((t = 2.16, p < .03)\), and the No Insult control group mean change score \((t = 1.90, p < .07; \text{one-tailed } p < .04)\). I argued for the one-tailed test on the Insult/No Insult difference, because the entire conception of the experiment revolved around the expectation of a boomerang effect under Insult. (Had a positive mean change score been found in the Insult group, I would have gone back to the drawing board.) There are other choices here, such as being content with the two-tailed \(p\) of .07, or if you want to be really adventurous, claiming that the \(t\) of 1.90 is significant at the .05 level by the lopsided test (see the discussion of one-tailed tests, this chapter).

I have the sense that a compromise statistical test could be invented that would resolve the apparent paradox in this type of situation. In any case, it is important to understand the mischievous potential consequences of the 42% Rule.

Different Calculation Formulas

There are a number of pairs of significance tests with different formulas for achieving the same purpose. For some pairs it is mathematically guaranteed that they will yield the same \(p\) value when both are applied to the same data set. For other pairs, the two formulas are based on alternative conceptions, and the results may diverge.

Difference Between Two Means, \(t\) Test and \(F\) Test. When testing for a difference between the means of two independent groups, it is immaterial whether a (two-tailed) \(t\) test or an \(F\) test is used. Both are based on the same assumptions—homogeneous variances and normal distributions—and it would be disastrous if they came out different. Fortunately they don’t, even though their formulas look rather dissimilar. Given \(k\) degrees of freedom for the \(t\) test (where \(k\) is the total number of cases in the two groups, minus 2), the corresponding \(F\) test will have 1 and \(k\) degrees of freedom. Given the well-known mathematical relationship \(F = t^2\), it will always be the case that the square of the calculated \(t\) value will equal the \(F\) value for the equivalent comparison on the same data. The tables of significant values reflect this, and therefore the two tests yield exactly the same \(p\) value. It is a simple matter to inspect \(t\) and \(F\) tables to see the equivalence between \(F\) (with 1 \(df\) in the numerator) and \(t^2\). A useful exercise for the student is to verify that the calculating formulas for \(F\) and \(t\) lead algebraically to the well-known relationships just described. Another simple case in which two formulas look different but yield the same result is the critical ratio test of a difference between proportions (say, the proportion of men who agree with an opinion item minus the proportion of women who agree), and the ordinary, uncorrected chi-square test of association in the 2 \(\times\) 2 table (here, sex by agreement).

Sometimes, of course, when two significance tests have formulas that look different, they are different. There are several situations in which the different significance test formulas can yield somewhat different outcomes. Some may involve “pathology” in the data, that is, violation of some critical restriction.

Different Chi-Square Formulas in Log Linear Analysis. We mention here an important case that is relatively unfamiliar to psychologists, but familiar to sociologists and others who deal with categorical data arranged as frequency counts in multiway cross-classification tables. (In two dimensions, this is the familiar contingency table.) The general method for analyzing associations among the ways of the table
is log-linear analysis (Fienberg, 1980; Wickens, 1989). The standard significance test for goodness of fit of log-linear models is a chi-square test, a small chi-square indicating a good fit. But there are two different methods for calculating a chi-square: the familiar Pearson chi-square as used on simple two-way contingency tables, and the chi-square derived from the maximum likelihood method. These two methods usually yield only slightly different values.

There are two circumstances in which the difference between the two methods can be notable. One case is when both chi-squares are very large and highly significant, but they differ in size. This happens because the approximations to an exact chi-square gradually break down as the model being tested fits more and more badly. Here the discrepancy is rather harmless, because the conclusion is the same in both cases: The model being tested is resoundingly rejected.

A more interesting circumstance that may yield a large discrepancy is the presence in the contingency table of one or more cells with very small expected frequencies (less than 1.0 or even .5). Many elementary statistics texts warn of inexact p values when expected frequencies are small, but specification of what small means depends on when and by whom the text was written. The point of such warnings is that the calculated chi-square may not approximate well the exact distribution given in the chi-square tables. When the two distinct chi-square formulas—the Pearson and the maximum likelihood—give discrepant (but not huge) values, it is a sign that some of the expected values are indeed too small. In this event, one recourse is to recalculate a chi-square leaving out the cells with small expected frequencies. An interesting example of this device can be found in Duncan, Sloane, and Brody (1982). For an excellent discussion of small cell problems, see Wickens (1989).

**Different Ways to Frame an Analysis**

A rather different kind of case arises when alternative formulas signify differences in the guiding philosophy of the analysis.

**Combining p Values from Multiple Experiments.** In the procedure of meta-analysis, discussed in chapter 7, it may be desired to combine p values from multiple, independent studies of the same null hypothesis, to create a single, omnibus test: for example, of departures of mental empathy scores from chance baseline scores (Honorton et al., 1990). The most popular procedure for combining p values is the Stouffer test popularized by Mosteller and Bush (1954), but there are at least six alternatives (Rosenthal, 1991). To make a point, we focus on one of these, a little-used alternative (Fisher, 1946) that involves adding logarithmically transformed p values. The Stouffer and Fisher tests yield predictably different results. The Stouffer test statistic is sensitive to consistent, even if mild, departures from the null hypothesis in separate studies (say, four results at p = .15), whereas the Fisher procedure is most sensitive to occasional, extreme departures (say, three results at p = .10, and one at p = .01). Table 4.2 contrasts the results of the two tests for these two data situations. (The Stouffer test is based on critical ratios, the Fisher on a chi-square statistic with degrees of freedom = twice the number of p values.)

The table gives the p value attaching to the aggregate, that is, a significance test of the omnibus null hypothesis, for each of the two tests on each of the two sets of results. When all four studies yield p < .15, the Stouffer test rejects the omnibus (at .02), but the Fisher test does not. When three of the individual results produce p < .05, but one reaches p < .001, the tests reverse roles. The Fisher is significant (at p < .02), but the Stouffer is not. If one were always to choose the test yielding the most extreme p value, that would be too harsh. The choice could be made resistant to criticism by using the conventional test (here, the Stouffer).

However, there might be theoretical reasons to use the Fisher test for certain applications. For example, if one is to believe the rhetoric of proponents of ESP, success at mental telepathy comes and goes, but when present, is impressive. Multiple tests of mental telepathy therefore might be a rare candidate for p value aggregation by the Fisher test.

**Summary of Alternative Test Procedures**

A breezy summary of the various cases of alternative tests applied to the same data set might go something like this: Often the choice between alternatives makes no practical difference. Occasionally (as in the log-linear case discussed earlier) what is important is not so much the choice between alternative methods, as what their discrepancy signifies about the nature of the data. In the remaining cases, when different tests do come out somewhat differently, one of them is usually conventional. As in the situation with two-tailed tests versus one-tailed or lopsided tests, a prudent policy is to choose the more conventional approach, but remain willing to deviate from it when an intelligent rationale based on appropriate special conditions can be put forward.
DEFECTIVE OBSERVATIONS

"Klinkers"

Almost all data sets contain defective observations. Recording equipment may fail; respondents may blatantly misperceive, misunderstand, or misrespond; samples of subjects may include one or more obviously inappropriate people (such as those who don't understand English); experimenters may pollute the experimental procedure. The more obvious the circumstances producing klinkers, the less problematic it is for the investigator simply to eliminate these observations from the data set, or to replace flawed subjects with new subjects.

Circumstances may not be obvious, however. The investigator may be unsure whether or not Subject 17 was drugged out, whether Subject 34 wrote his intended response on the wrong line of the questionnaire, and so on. The usual solution for ambiguous cases is to run data analysis both ways—that is, with and without inclusion of the questionable data points. The happiest outcome would be that it does not make (much of) a difference. In marginal cases when it does make a difference, a conservative policy would dictate that possible klinkers helping you to reject the null hypothesis should be thrown out, but those that hurt your ability to reject the null hypothesis should be kept in. This is appropriately cautious advice, but there is a catch. The research audience cannot notice possible peculiarities of individual observations, because they are not available to public scrutiny. Thus it is left up to the investigator to notice oddities, take them seriously, and nobly throw away klinkers even when they help you. Consider the following legend, possibly apocryphal but instructive nonetheless.

An experienced experimenter, in carefully reviewing an experiment testing a hypothesis about overweight people, eliminated the data of one subject because it turned out that he was a wrestler, who should be

1The omission of observations from a carefully designed data structure may sometimes create awkwardness for the statistical analysis. Problems caused by missing data, and recommended solutions, are discussed by Cohen and Cohen (1983) and by Kirk (in press), among others.

2Typically, when "both ways" of analysis are run, one way is presented in the text of the research report, and the alternative way in a footnote. To the best of my recollection, such footnotes always say that the alternative choice made no difference. This, of course, is what the investigator would like to say, and I am made somewhat suspicious by the unanimity of happy accounts in these footnotes. The exact wording may be a tip-off to the convenient suppression of unpleasant details. Thus, when the footnote reads, "inclusion/exclusion of these observations did not affect the conclusions of the study," it could be the case that inclusion (exclusion) changed a p value from p < 0.05 to p > 0.15, but that the investigator still drew the same conclusions. Stilistically, the conservative investigator—as opposed to the brash investigator—will use more precise wording of such footnotes (e.g., "...inclusion/did not weaken the p level[s] by more than such and such...").

considered not really overweight, merely "bulked up." The wrestler's data had gone strongly against the experimenter's prediction. Without faulting the experimenter for deliberate bias, one may nevertheless question whether the distinction between being overweight and being bulked up would have been constructed had the wrestler's data been supportive of the experimental hypothesis. The experimenter would probably not have been vigilant enough to scrutinize the details of favorable data, and might never even have discovered that one subject was a wrestler. The moral is as easy to preach as it is hard to carry out: Be your own toughest critic.

Outliers

We have labeled as klinkers those observations deemed inappropriate by external criteria such as failed equipment or errors in the selection or handling of subjects. Also vexing are outliers (Tukey, 1977)—observations obtained under seemingly normal circumstances, but that turn out to be extremely deviant from the main body of observations. To illustrate, consider a reaction time study in which almost all of the reaction times fall, say, between 1.2 and 2.3 seconds, but for no apparent reason there suddenly arises a time of 8.5 seconds. Was the subject daydreaming? Did he miss the ready signal? We will never know, but it is clear that such an outlier will distort the mean and standard deviation of any group of observations within which it is included.3 This problem is so common in reaction time studies that cognitive psychologists have developed a variety of methods to handle it, for example, using "trimmed means" (Bush et al., 1993; Wilcox, 1992) instead of ordinary means, or simply eliminating outliers entirely from the data set. An authoritative treatment of the various possibilities was given by Ratcliffe (1993).

General Advice

Discussion of these alternative fixes for the outlier problem would carry us too far afield. A key observation here is that except for those who study reaction times, most psychological and other social science researchers have not confronted the problem of what to do with outliers—but they should. A generalized conservative suspicion against doing anything, lest it seem brash, is not a good solution. Doing nothing is as much a choice as is doing something.

3The vulnerability of means and standard deviations to being overinfluenced by one or more wildly discrepant observations is a phenomenon much analyzed by statisticians under the heading of robustness. Measures of central tendency (or variability, etc.) that are not vulnerable to distortion by outliers are called robust statistics (Hoaglin et al., 1983; Tukey, 1962).
The general effect of doing nothing about outliers is to tolerate more noisiness in data, that is, to have lower power, rejecting null hypotheses less often. In any particular case, however, the do-nothing policy could either help or hurt the investigator wanting to reject the null, depending on the location of the outlier(s). Thus, what one wants to avoid is ad hoc treatment of outliers, differing by whim from one study to the next. Instead, one wants to develop a consistent policy applicable to all studies of a given type. This advice gives rise to Abelson's Third Law: Never flout a convention just once. In other words, either stick consistently to conventional procedures, or better, violate convention in a coherent way if informed consideration provides good reason for so doing. This advice in favor of stylistic consistency applies not only to treatment of outliers, but in general.

**MULTIPLE TESTS WITHIN THE SAME DATA SET**

When there are multiple tests within the same study or series of studies, a stylistic issue is unavoidable. As Diaconis (1985) put it, “Multiplicity is one of the most prominent difficulties with data-analytic procedures. Roughly speaking, if enough different statistics are computed, some of them will be sure to show structure” (p. 9). In other words, random patterns will seem to contain something systematic when scrutinized in many particular ways. If you look at enough boulders, there is bound to be one that looks like a sculpted human face. Knowing this, if you apply extremely strict criteria for what is to be recognized as an intentionally carved face, you might miss the whole show on Easter Island.

**Setting the Error Rate**

We discuss multiplicity issues from the standpoint of the Type I error rate, that is, the proportion of times that hypothetically true null hypotheses would be rejected. The conventional .05 level signifies that for every 100 tests of a true null hypothesis, false significance will be claimed for 5 of them—an error rate of 5 in 100. This sounds precise, but avoids the question of the scope of each null hypothesis. A null hypothesis could be specific to a test on a single mean, between a pair of means, or among several means—or it could apply to every test whatsoever within a study, or over a whole series of studies.

Because there is a community of researchers in any given field, the development of a new statistical convention requires an authoritative explicit declaration, or the implicit negotiation of a consensus. The former usually comes via methodological articles in prestige journals (e.g., Clark, 1973; Green & Tukey, 1960). The latter could arise from the gradual diffusion of illustrative applications of a given policy decision, perhaps marked by occasional public debates over the advantages and disadvantages of alternatives.

---

Manny Powers, the Psychic. Suppose, for example, that an investigator interested in extrasensory perception comes across an individual with a reputation as a psychic (call him Manny Powers). The investigator carries out 40 studies on Powers, one study per day. Each day, Powers is tested on five different ESP tasks (telepathy, precognition, psychokinesis, etc.). The goal is to find the circumstances under which Powers is psychic, operationally defined as performing better than chance on a given task on a given day.

How should the investigator specify what is meant by the .05 significance level? A skeptic who doesn't believe in any manifestation of ESP whatsoever would point out that if a significance level of .05 is established for each task on each day, there would be 200 separate opportunities to reject a null hypothesis. Thus, assuming that the skeptic's nonbelief is warranted, nevertheless (.05 x 200) = 10 rejections of a null hypothesis will be expected on average. In other words, there will be about 10 occasions on which extrasensory skill will be claimed for Powers because he performed "better than chance" on some particular task. A cunning or self-deluded investigator might focus on these 10 or so successes, setting aside the approximately 190 failures—the *hocus pocus* trick. (The lack of public knowledge of the number of failures is called the file drawer problem; Iyengar & Greenhouse, 1988; Rosenthal, 1979.)

Considering the whole series of studies and tasks as a single unit, the error rate is 200 times too high. To achieve an error rate of only 5 per 100 hypothetical repetitions of the whole series of studies, the investigator would have to set the significance level per individual significance test at .05/200, or .00025. This procedure of adjusting the significance level according to the number of tests is called the Bonferroni method (Emerson, 1991a; Miller, 1981). Obviously, it would make it much harder to claim ESP for any specific occasion.

However, a side effect of such a conservative procedure, as Duncan, 1955) and others have complained, is to penalize the investigator who is ambitious enough to have run so many repeated tests. A lazier investigator who ran Manny Powers on five tasks for only 1 day would by the same reasoning be able to set the significance level per task at .05/5, or .01. And the laziest investigator of all, using one task on 1 day, could simply use the .05 level without modification.

The Policy Dilemma. The stylistic debate pitting stuffiness against brashness is very stark here. When contemplating a significance testing policy for a series of studies, with multiple tests to be performed within each study, the most stiflingly conservative thing to do is to set the error rate at 5 potentially false claims per 100 *series*. This forces an extremely high threshold for making a claim of significance from any single test. At the other extreme, the most loosely liberal policy is to
establish the error rate at 5 per 100 tests, allowing a multitude of potentially false claims to fall where they may. Neither extreme policy is particularly attractive; the best argument for each lies in the potential folly of the other.

There is no universal right answer to this policy dilemma, and different conventional answers have grown up around different types of analysis. In multiway analyses of variance designed to test two or more main effects and some number of interactions, for example, the standard thing to do has been to apply the same significance level, usually 0.05, to each effect, ignoring the proliferation of tests. This extremely liberal policy is inconsistent with more conservative practice in other cases, but most researchers follow this convention without second thoughts. If a defense for this policy had to be spelled out, it would make two points: First, multiway designs are generally only run after preliminary research has established that there is a phenomenon to be studied; thus, the overall null hypothesis of no real effects whatever is even less of a serious possibility than it typically is. Second, the investigator is usually interested in a small handful of particular effects; if the analysis of variance table has many lines to it, scattered effects besides the interesting ones might well beat the 0.05 level by chance, but the investigator should not go to great lengths trying to interpret them. A clear analytic focus, established prior to running the study, serves to diminish the vagaries of multiplicity.

Focused Tests in the Many-Group Study. Conflicting standards also obtain for multiple tests applied to the several means of a many-group main effect. When there are several degrees of freedom for an effect, the ordinary $F$ test is called an omnibus test. It tests the overall null hypothesis that none of the means truly differ, against the vague alternative hypothesis that some of them differ to some extent from some of the others. The null hypothesis might come to be rejected in many ways, none of which are specified in advance. Specification of particular alternatives to the null hypothesis may involve planned contrasts—planned because they should be specified before looking at the data, and contrasts because they involve patterns of differences among the means (e.g., the linear contrast, whereby the means of groups A, B, C, etc., increase from left to right in equal sized steps). Because planned contrasts are usually more powerful than omnibus tests, they tend to yield stronger $p$ values for rejecting the null hypothesis. They are also better articulated (as is discussed in chap. 7).

An issue of style can nonetheless arise when an unusual contrast is chosen, because the stronger $p$ value may come at a cost in rhetorical credibility—the critic may not be convinced that the quirky-looking contrast was really planned. Instead, it may seem like a brash attempt to make something out of whatever pattern happened to appear in the data. In line with our Third Law, then, the reasoning behind the choice of a contrast must be clear and used often to be convincing. As a fallback, there is also a significance test appropriate for unplanned contrasts—the Scheffé (1959) test—but it is extremely conservative.

Moderately conservative orientations characterize some of the many methods of multiple comparisons, in which each of a set of means is compared with every other to determine which pairs of means should be declared significantly different. The bellwether of the family of possible multiple comparisons procedures is the Tukey (1953) test. If in a series of experiments, each with several groups, the overall null hypothesis were always true, the Tukey test would on average make five false declarations of significance per 100 experiments. The per-experiment error rate seems a reasonable compromise between per-comparison and per-series error rates.

The two most ultraliberal procedures—multiple $t$ tests and Duncan's (1955) test—which base the error rate on the number of false claims per 100 comparisons, have very bad reputations among statisticians. Scheffé (1959), for example, in his authoritative book on the analysis of variance, had the scathing footnote, "I have not included the multiple comparison methods of D. B. Duncan because I am unable to understand their justification" (p. 78). Even the middle-of-the-road Newman–Keuls test (Keuls, 1952; Newman, 1937), which sets the error rate at five experiments out of 100, offends conservative sensibilities (Ramsey, 1981). A reason for this is that when the Newman–Keuls does produce false claims in a given experiment, it may tend to produce several of them—and thus a higher total number of errors in 100 experiments, compared to the Tukey test.

Different multiple comparison tests have proliferated like raspberry bushes (see Hochberg & Tamhane, 1987). Perhaps because of the confusion attaching to the steady introduction of new tests, different subfields in psychology differ in the degree of liberality they are willing to tolerate in multiple comparisons tests. Social psychologists in particular, for reasons that escape me, tend to be rather profligate in the use of multiple $t$ tests. My own view is that the decision to use multiple comparison procedures should bring with it a somewhat conservative attitude. That is the price of multiplicity. If the investigator finds the Tukey (1953) test, or something like it, too conservative, she should ponder whether complete testing of every mean against every other is really to the point. Perhaps she really only cares about a couple of particular differences. Or better, perhaps she could apply a meaningful contrast. Conceptual focusing not only helps ease the stylistic dilemma, it leads to cleaner studies and clearer theories.

---

11 Some sociologist of knowledge should study this! Differences in conventions are diagnostic of boundaries between different research subcultures.
STATING AND INTERPRETING \( p \) VALUES

Word Play

The desperate investigator with an almost significant result, say \( p = .07 \), may try to talk the result across the conventional .05 boundary, or to rationalize the failure to reach the boundary.

Typical Rhetorical Flourishes. Among the phrases used to attempt such rhetorical feats are these:

- The result was significant at the .07 level ...
- The result was marginally significant \((p = .07)\) ...
- Although the result did not reach the conventional .05 level, it is nevertheless highly suggestive ...
- Because of the limited number of subjects (or low power), the result just missed the .05 level. Nevertheless ...

These verbal devices, though not entirely unreasonable, make for somewhat defensive rhetoric, suggesting a penchant for weaseling. (Note especially the gambit in the reference to the conventional .05 level, as though to blame the convention rather than the result.) On the other hand, the .05 level is admittedly an arbitrary standard, and there is not much real difference between results at the \( p = .07 \) and \( p = .05 \) levels. What is the author to do?

Common sense suggests being straightforward—not trying to make too much out of the situation. Give the test statistic and the \( p \) value of .07. Leave it to intelligent readers to appreciate for themselves that your result just missed.

Results That "Lean" and "Hint." Interestingly, John Tukey, the developer of moderately conservative procedures for multiple comparisons, recently (Tukey, 1991) came out with what seems a shockingly radical proposal. He recommended coining new words for results at certain benchmark levels weaker than the .05. His starting point is the assertion that the null hypothesis is never literally true (see also Cohen, in press; Schmidt, 1992). Thus in testing the difference between two means, the so-called acceptance of the null hypothesis merely signifies a reluctance to bet on the direction of the true mean difference. If the two-tailed \( p \) were greater than .05, but less than, say, .15, one might at least be tempted to bet that the observed direction of difference were the true one. Tukey suggested saying in this case that the difference between A and B leans in the positive direction. For \( .15 < p < .25 \), he proposed stating that there is a hint about the direction. Imagine using a level as scorned as .25 to enable a substantive statement about one's results! Will wonders never cease?

Let us take pause, however, before cheering that now we can publish all those lousy studies that couldn't beat the .05 level. In using a new term, Tukey (1991) was not authorizing the use of .25 as a new significance level for rejecting null hypotheses. A hint is just a hint is just a hint. There is nothing very definitive about it. Instead, what he was telling us is to stop treating statistical testing as a two-valued decision procedure, and instead to use shades of wording to indicate different degrees of uncertainty.

Reproducibility and Power: The Real Issues

One can also envisage a semantics of confidence rather than of doubt. In a challenging article by Greenwald, Gonzalez, Harris, and Guthrie (1993), the authors showed that if a two-group study achieves \( p < .005 \), the probability that an exact replication of the study will yield \( p < .05 \) is approximately .8. Greenwald et al. therefore suggested that the label replicable might be attached to a \( p \) level of .005. (This result is insensitive to the group ns, but does depend on the assumption that the effect size in the first study is the best estimate of the effect size in the replication.) Yes, I know that the whole thing sounds weird, because if the original study came out as strongly as \( p < .005 \), how could the chance of getting a replication to yield merely \( p < .05 \) be as modest as .8?

The Replication Fallacy. Readers who are baffled by this assault on their intuitions are suffering from the replication fallacy (Gigerenzer, 1993), an overconfidence in the repeatability of statistically significant results. The following thought experiment may help to correct the fallacy. Imagine an experimenter who has run a two-group study, and has found by \( t \) test the result \( p = .05 \). What is the chance that if she exactly repeated the study with a new sample of subjects (and the same \( n \) per group) that she would again get a significant result at the .05 level? First give an intuitive answer, and then study the analysis that follows.

Half the time, the observed effect size from the second study ought to be bigger than that of the first study, and half the time, smaller. Because the first observed effect size was just big enough to obtain a \( p \) value of .05, anything smaller would yield a nonsignificant \( p > .05 \). This analysis thus yields an expected repeatability of 50-50, much lower than the usual intuition. Psychologically, overconfidence arises because once you find a significant result, you say to yourself, "Ah, now I've got it. The real thing is there. No more problems!" But this falls into the trap of thinking that just because you have made a categorical assertion about a significant result, the influence of chance has disappeared. Remember that when we discredit a chance explanation, we are merely saying that
an entirely chance account is inadequate. We then claim a systematic effect in addition to chance effects.

**Power and the Wishful Experimenter.** The probability that a significance test will reject the null hypothesis is called its power (Cohen, 1988). Power can be increased by running more subjects, or by increasing the effect size via an increase in the cause size (chap. 3) and/or by a decrease in the influence of chance factors. Long ago, Cohen (1962) criticized the psychological research community for running studies with much too little power (about .42 on the average), and the situation has not improved since (Cohen, 1990). This raises a stylistic issue different from liberal versus conservative result presentations. It concerns the style of designing experiments, and we might call the poles of this dimension vigilant versus wishful. The experimenter who estimates how many subjects are needed to ensure adequate power (Cohen, 1988) is vigilantly guarding against the worst-case scenario of indecisive results. The wishful experimenter, on the other hand, assumes that a divine hand is guiding his research, and plunges ahead to frequent disappointments and blind alleys.

**Silly Significance Tests**

The ethos of doing significance tests as the hallmark of an appropriately conservative style is now so deeply ingrained that tests are sometimes used even when they need not be. Indeed, there are several contexts in which it is really silly to carry out a significance test (Cohen, in press), much less to present its result. For example, if a sample is divided at the median into high scorers and low scorers, there is no point in showing by a t test that the high scorers differ significantly from the low scorers. A somewhat subtler case arises when a trustworthy procedure for random assignment of subjects to experimental conditions seems to go awry, yielding a cluster of smarter or faster (or whatever) subjects in one particular group. In this situation, students are prone to test the significance of the difference between the groups. But because the null hypothesis here is that the samples were randomly drawn from the same population, it is true by definition, and needs no data. What has happened is that one of those fluky outcomes that arise from time to time has arisen this time. As Abelson's First Law says, chance is lumpy. The investigator might want to adjust for the lucky advantage of the one group (say, by analysis of covariance), but a significance test has no relevance to this decision.

**Needless Clutter with p Values**

Meanwhile, there are some subfields of specialization within psychology (and other social and natural sciences) where the prevailing ethos is that experimental results should be so clear that statistical tests are totally unnecessary. The reader need only look at a graph of the results, and the pertinent trends or skips or blips will be obvious. As behaviorist B. F. Skinner (1968) wrote, "...in the experimental analysis of behavior... statistical methods are unnecessary... When a variable is changed and the effect on performance observed, it is for most purposes idle to prove statistically that a change has occurred." (p. 508)

I have some sympathy for this position. Nothing is more tedious than a research report cluttered with obligatory p < .001s for every conceivable claim, including such blatantly obvious assertions as that performance improves with practice, or is correlated with ability. On the other hand, occasionally it happens that "obvious" conclusions are false. And it is easy to overreact to appearances in graphs. For example, there is a strong tendency to perceive approximately regular cycles in the ups and downs of a random time series (Abelson, 1959), or to find mysterious coordination between separate time series, as in the "Maharishi effect" (Orme-Johnson et al., 1988).

Research reports can in various ways avoid belaboring the obvious with p value clutter. The most suspect way is to give a couple of p values at the outset for major results, and then to focus on other data features merely with a look-and-see-the-graph approach. The perceptive reader will be wary of this potentially misleading practice. The prudent investigator can anticipate objections by asking the self-critical question, "How can I demonstrate the statistical significance, if I have to, of each assertion I want to make?" Even a rough-and-ready answer would be salutary, and there is almost always a way to choose an approximate significance test. The text of the report need not be littered with p values.

Sections in which all claims are judged significant can be covered by an opening remark or footnote stating that every claim has p < .05, or better.

**IN THE LAST ANALYSIS**

Most of the noisy stylistic battles fought over the anxious pursuit of p < .05 are needless. Though contention over style will never completely disappear, a low level of stylistic disagreement between investigators is quite tolerable—perhaps even adaptive. Research conclusions arise not from single studies alone, but from cumulative replication. In this cumulative process (which we discuss in chaps. 7 and 9), liberal versus conservative stylistic differences will tend to cancel out, and if the community learns to be selective in its research designs and focused in its statistical tests, cumulation will be more rapid.