What is a P-value?

Ronald A. Thisted, PhD *
Departments of Statistics and Health Studies
The University of Chicago

8 June 1998, Corrections 14 February 2010

Abstract

Results favoring one treatment over another in a randomized clinical trial can be explained only if the favored treatment really is superior *or* the apparent advantage enjoyed by the treatment is due solely to the working of chance. Since chance produces very small advantages often but large differences rarely, the larger the effect seen in the trial the less plausible chance assignment alone can be as an explanation. If the chance explanation can be ruled out, then the differences seen in the study must be due to the effectiveness of the treatment being studied.

The p-value measures consistency between the results actually obtained in the trial and the "pure chance" explanation for those results. A p-value of 0.002 favoring group A arises very infrequently when the only differences between groups A and C are due to chance. More precisely, chance alone would produce such a result only twice in every thousand studies. Consequently, we conclude that the advantage of A over B is (quite probably) real rather than spurious.

What does it mean when the results of a randomized clinical trial comparing two treatments reports that, "Treatment A was found to be superior to Treatment C ($p = 0.002$)"? How much, and how little, should non-statisticians make of this? The interpretation of the p-value depends in large measure on the design of the study whose results are being reported. When the study is a randomized clinical trial, this interpretation is straightforward.

Conducting the ideal study

The ideal study to compare two treatments — an active drug and a placebo, for example — would be to test each treatment in a laboratory experiment that would be *identical* aside from whether Compound A or Compound C was under

*Professor, Departments of Statistics, Health Studies, and Anesthesia & Critical Care,The University of Chicago, 5841 South Maryland Avenue (MC 2007), Chicago, IL 60637. URL:<http://www.stat.uchicago.edu/~thisted>.

test. We could then attribute any difference in outcome to the only factor that varied, namely, the difference in the effects of the two compounds.

Laboratories are designed so that extraneous factors which *could* affect the outcomes can be tightly controlled. Clinical settings, however, are far from this ideal testing environment. But for medical treatments, the clinical setting is ultimately the only one that matters. Clinical trials are studies of responses to medical treatments by human subjects in the clinical setting, and they are designed to control extraneous factors as well as can be accomplished outside the laboratory.

Controlling extraneous factors in a clinical trial is achieved using two strategies. The first is by constructing a tightly-drawn *protocol,* which is the manual of operation by which the study is carried out. The second is *randomization.*

The Protocol

In writing the protocol, investigators anticipate major factors that can affect patient outcomes. Then the procedures of the trial are fashioned to control them closely. These factors will be different in each clinical trial and for each compound or treatment being studied. For example, a clinical study to determine whether taking aspirin daily can reduce the risk of heart attack needs to take account of the facts that, independent of any treatment, males are at higher risk than females, 70-year-olds are at higher risk than 40-year-olds, and smokers are at higher risk than nonsmokers.

Inclusion and exclusion criteria

There are many ways for the investigators to "hold these factors constant." One approach, for instance, would be to exclude smokers from participating in the study. While this might reduce the generalizability of findings from the study, we could be 100% certain that any differences in rates of heart attack between those receiving aspirin and those receiving placebo were not due to differences in smoking behavior. Through its *inclusion and exclusion criteria,* the protocol will spell out in advance just what patient characteristics will be permitted for participants in the study.

The disadvantages of using inclusion/exclusion criteria to control important factors are (a) that doing so reduces the number of potential subjects who are available to participate, and (b) that a study showing an effect in subjects limited to, say, nonsmoking women in their 40s, may well not apply to older women, or to men, or to smokers. Thus, limiting inclusion also limits generalizability.

Stratification

A second approach would be to make certain that the mix, say of men and women in the study, is the same for both test groups. If both the aspirin

and the placebo groups were composed of 30% men and 70% women, then a lower heart-attack rate in the aspirin group could not be attributed to having a disproportionate share of the (lower-risk) women participating in the study. When major factors are purposely distributed evenly between study groups, the process is called *stratification,* or *blocking.*

The main problems with stratification are (a) that the major factors must be identified in advance of the study and must be ascertained from each potential subject before that subject can be enrolled in the study and assigned to receive one of the two treatments, and (b) that even a small degree of stratification can introduce massive complexity in the procedures needed to carry out the study. For instance, suppose in the aspirin study we wish to stratify on age, sex, and smoking status. That way, each study group would contain the same number, say, of female nonsmokers in their 50s. If we had four age groups (say, 50s, 60s, 70s, and 80s), then we have sixteen different patient categories, or *strata,* within each of which we have to insure balance in assigning subjects to treatments. The age/smoking/sex mix in the pool of potential subjects may result in strata which have only one subject, or perhaps, none at all.

As a practical matter, stratification can deal with only one or two major factors, and always increases the cost and complexity of the clinical trial.

Randomization

The methods of careful protocol design discussed above cannot hope to control for all of the possible factors that might affect the likelihood of having a heart attack. Genetic predisposition to heart disease cannot be measured easily, but is known to vary across individuals. Individuals with high levels of some cholesterols and triglycerides are at much greater risk than those with low levels. Diet matters, as does physical activity. How can we make sure that our study groups are balanced, not only with respect to each of these risk factors that we can name, but also with respect to any other risk factors that we simply haven't known enough to include? And if we can't make certain that the study groups are balanced on every factor, how could we possibly conclude that a lower heart-attack rate in the aspirin group is actually due to getting aspirin, as opposed to some other risk-lowering factor which just happens to be more predominant in the group of subjects who were assigned to receive aspirin?

The key to dealing with all of these other factors in a clinical trial is to use a *randomization* procedure when assigning subjects to the study treatments. Typically, the randomization procedure will make each subject as likely to be assigned to receive aspirin as to receive placebo. [1] While this doesn't guarantee that exactly half of the smokers will be taking aspirin and half placebo, it does guarantee that there are no factors, *known or unknown,* that are systematically

[1] Under some circumstances, it is desirable to allocate patients to treatments in a ratio that is different from 1:1. In a 2:1 randomization plan, for instance, each subject would have a two-thirds chance of getting treatment A and a one-third chance of getting treatment B. The allocation ratio is determined as part of the study design, and is spelled out in the protocol.

being assigned to one of the treatment groups as opposed to the other.

Here is how randomization works to achieve balance. Suppose that of the 1600 subjects who will enroll in our study, 20 of these subjects have an abnormality of the coronary arteries that makes them very highly likely to have a heart attack during the study. With random allocation to treatments — in effect, tossing a perfect coin for each subject then assigning the subject to receive aspirin if the coin comes up "heads" — the most likely outcome would be for 10 of these high-risk subjects to be in the aspirin group and 10 in the placebo group. This perfect 10/10 split occurs about 18% of the time. Half of the time, the split will be no worse than 11/9, and almost 90% of the time this procedure will produce splits no worse than 13/7. In other words, randomization doesn't guarantee perfect balance, but nearly all of the time it keeps things nearly balanced.

While it is possible for the randomization procedure to assign all 20 of these high-risk patients to just one of the treatment groups — after all, it is *possible* for the coin to come up "heads" twenty times in a row — it would be unusual in the extreme for that to happen. In order to see twenty coin tosses result in twenty heads, you would have to repeat this coin-tossing procedure about half a million times!

The p-value

In the laboratory with all factors carefully controlled, any differences we observe between treatments can logically be said to be "real." If the active treatment has a lower pH than the control treatment in the lab, we are convinced that the active treatment is more acidic. Our degree of certainty can't be quite so great in a clinical setting, but it can be close.

In a carefully done randomized clinical trial, if we observe fewer heart attacks in the aspirin group than in the placebo group, then we must conclude *either* that aspirin really is effective in lowering heart-attack risk, *or* that despite the randomization, fewer subjects at risk of heart attack found their way into the aspirin group. Of course, the latter is possible for exactly the same reason that it is possible to observe many more heads than tails in a series of coin tosses.

Let's suppose that 170 subjects of the 800 assigned to receive placebo had a heart attack during the study, and that 118 of the 800 assigned to receive aspirin had a heart attack. If aspirin really is no better than placebo, then that means that 288 subjects were destined to have a heart attack during the study, and the randomization plan just happened to assign 170 to placebo and 118 to aspirin. Is this scenario plausible? We don't expect perfect balance (144 assigned to each group), but this seems a bit lopsided; how rarely does true randomization produce at least this much imbalance? The answer, calculated using the binomial distribution, is about one time in every four hundred. The p-value for this study is $p = 0.002593 \approx 1/400$. While not ruling out the "chance explanation" as a *theoretical* possibility, most people would say that for practical purposes "chance" as the sole explanation could be dismissed.

Interpreting the p-value

Had the study ended somewhat differently, with 150 and 138 heart attacks in placebo and aspirin groups, respectively, the p-value would have been $p = 0.52$. In other words, one out of every two randomization results would have produced at least this much imbalance. This is so frequent that most people would say that the "chance explanation" could not really be ruled out on the basis of this study.

With a p-value of 0.0025, we feel comfortable dismissing the "chance explanation," which means that we are comfortable ascribing the imbalance seen to the effect of the superior drug treatment. With a p-value of 0.52, we can't dismiss chance. But failing to rule out an explanation (such as "chance") hardly means that the explanation is true—only that it is not obviously false.

Consequently, small p-values permit investigators to draw a conclusion about whether one drug is more effective than another by allowing them to eliminate all but that possibility. Large p-values simply do not permit investigators to eliminate *any* possibilities.

With a p-value of 0.52, it would be incorrect to conclude that aspirin and placebo are equally effective. Rather, we would have to conclude that aspirin's effectiveness remains unproven.

Where do we draw the line between unproven and effectively established? For instance, 160 and 128 heart attacks in placebo and aspirin groups, respectively, give $p = 0.07$. One in fourteen randomizations lead to at least this much imbalance. One in fourteen is neither remote nor common.

How small does the p-value have to be before the evidence is considered convincing? The answer to this question will vary from one individual to the next, and will depend on many factors. If adopting a new drug would be very expensive, or expose patients to severe side effects, then one might demand a very high standard of evidence (that is, a very small p-value) before taking actions based on the superior results from a clinical trial. On the other hand, the downside to adopting the new treatment may be quite low, and may offer advantages over an existing treatment, in which case the evidence barrier could be quite low.

What does it mean to be "statistically significant"?

It has become scientific convention to say that p-values exceeding 0.05 (one in twenty) just aren't strong enough to be the sole evidence that two treatments being studied really differ in their effect. When the term *statistically significant* is used to describe the results from a clinical trial, it means that the p-value is less than this conventional reference point.

It is possible to have very strong evidence to show that even a very small effect is nonetheless a real one. Consequently, it is possible to have a small p-value even though the *size* of the effect is not worth bothering with from the

clinical standpoint. Thus, *statistical significance and clinical importance must be completely separate assessments.*

Even when one treatment is superior to another in clinically important ways, it is possible for the study not to end in statistical significance. The study may have been too small to detect any but the largest of differences [2]. And of course chance can work against a treatment as easily as work in its favor. Consequently, *not achieving statistical significance must not be interpreted as having shown that the treatments studied are equally effective.*

[2] Such studies are said to be *underpowered*.