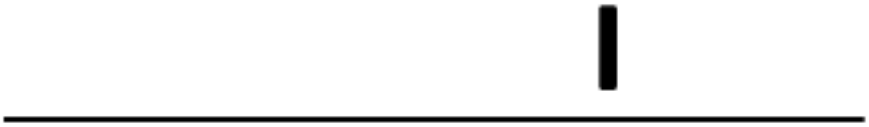


III

STATISTICAL SIGNIFICANCE

Suppose that at the cost of considerable effort, you, the research scientist, have done everything right: you've run a randomized double-blinded trial to test a prechosen hypothesis of interest in terms of a preselected one-dimensional response measurement. Now you face another difficult problem. Chapter 2 spoke loosely of the measurements showing a clear difference between the treatment and control groups, but what does "a clear difference" mean in dealing with noisy data? Statisticians[2] have developed an answer to this question, "significance testing", that has become a standard tool of scientific inference. It is used so often that the words "statistically significant" are creeping into popular usage, threatening to join other favorites like "quantum" and "absolute zero" in the lexicon of familiar but still somehow mysterious terms. This chapter will show you how to run a significance test, and exactly what statistical significance really means.

3.1 THE CHOLOSTYRAMINE EXPERIMENT

Table 3.1 presents the early data from a randomized clinical trial run in the 1970's to test the efficacy of the drug cholestyramine for reducing cholesterol levels. Thirty eight

[2] Most of the ideas discussed here were originally developed by R.A. Fisher, building on older methods, in the period between 1920 and 1940.

subjects have been randomly assigned either to the treatment group, where they recieve 6 packets a day of cholestyramine, a gritty rather unpleasent powder, or to the control group where they recieve 6 packets per day of placebo powder. Neither the subjects (all middle-aged male doctors) nor their physicians know the assignments. The response variable of interest here is the the decrease in cholesterol level over the course of the experiment. Positive decreases are good of course, so we see that the first subject in the treatment group did well, decreasing his cholesterol level by 41 points, while the second treatment subject went the wrong way, getting a decrease of -3 points, that is an 3 point increase.

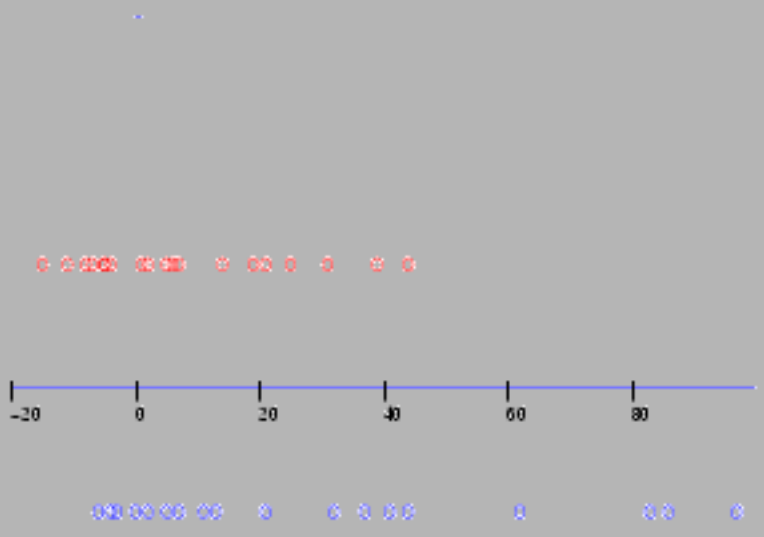
TABLE 3.1 Does cholestyramine reduce blood cholesterol? Cholesterol decreases for 38 subjects in a randomized clinical trial; 18 received cholestyramine, 20 placebo.

Treatment Group (mean=29.3)

subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
decrease	41	-3	97	5	-6	13	32	44	7	21	0	11	37	-4	62	2	83	86

Control Group (mean=6.9)

subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
decrease	31	-8	-15	25	19	7	44	-7	-21	-5	-5	2	6	39	14	-4	-11	21	5	1




```
Tment    Controls
     :  -2  :  1
     :  -1  :  51
   643 :  -0  :  87554
  7520 :   0  :  12567
    31 :   1  :  49
     1 :   2  :  15
    72 :   3  :  19
    41 :   4  :  4
     :   5  :
     2 :   6  :
     :   7  :
    63 :   8  :
     7 :   9  :
```

FIGURE 3.1 Visual Comparison of Two groups

Figure 3.1 gives visual comparisons of the results in terms of the values of the two groups.

The cholestyramine researchers, who hope to prove that cholestyramine lowers cholesterol levels, have some preliminary grounds for optimism. The mean (average) decrease in the treatment group is 29.3, considerably bigger than the control group mean of 6.9. Among all 38 subjects, the four largest decreases are in the treatment group while the three

smallest decreases (i.e. biggest increases) are in the control group. An eyeball comparison of the two histograms conveys a general impression of cholestyramine's efficacy.

"Not so fast" says the skeptic. "It isn't as if every cholestyramine subject did better than every placebo subject, there's lots of overlap between the two groups. Four of the 18 treatment decreases were actually increases. And one of the two group means had to be bigger, so how do we know that the next 38 subjects would not show placebo better than cholestyramine?"

These kinds of questions could go on forever. The theory of significance testing was designed to give an objective answer to a difficult question: how certain can we be that the observed difference in favor of the treatment group in Figure 3.1 is genuine, and not just an accidental combination of small samples (only 18 and 20 subjects) and noisy data?

3.2 PERMUTATION TESTS

A summary statistic is a single number that summarizes the results of an experiment or observational study. In the cholestyramine experiment we might use the difference between the treatment and control means as our preferred summarizer. Calling it "D" for short,

$$D = 29.3 - 6.9 = 22.4.$$

D is positive, favoring the efficacy of the treatment compared to placebo, but perhaps, as the skeptic suggested, it is just "accidentally" positive rather then "actually" positive, reflecting random variability in the data rather then a genuine cholestyramine effect. A permutation test provides an objective resolution of this question.

Permutation tests, and significance tests in general, begin with a "Null Hypothesis", a statement of what it means for nothing interesting to be happening. In this case we could state it this way,

Null Hypothesis: cholestyramine is equivalent to placebo in its cholesterol reducing effect.

The Null Hypothesis is a devil's advocate that the researchers hope to disprove. Significance testing is a skeptical court of scientific law, where treatments are considered ineffective until proven effective. As we shall see, the permutation test is required to rather decisively reject the Null Hypothesis before we are entitled to claim that cholestyramine is genuinely effective.

The key idea of permutation testing is that the Null Hypothesis implies all subjects were treated identically. There were 38 cholesterol subjects observed in the cholestyramine

experiment, 18 treatments and 20 controls, but if the `Null Hypothesis` is true than all 38 of them were actually placebo subjects, and the 38 cholesterol decreases should be interchangable with each other. The permutation test consists of comparing the summary statistic we actually have, $D = 22.4$, with the values of D we get by scrambling ("permuting") the data. If D is bigger than most of its scrambled versions we can take that as evidence against the Null Hypothesis and by implic- cation for the efficacy of the treatment. There is a precise methodology for doing all of this, as described next.

The scrambling is carried out using a random number generator. All 38 decreases from Table 3.1 are put into a computer file, from which 18 are randomly selected to act as the permuted treatment group, the remaining 20 acting as the permuted control group. Table 4 shows one such permutation.

Permuted Treatment Group (mean=13.1)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
-7	2	14	44	-5	-8	21	-5	86	25	32	6	21	13	0	-4	7	-6

Permuted Control Group (mean=21.5)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
-3	-11	5	83	37	39	-15	44	-4	1	5	41	2	19	97	11	31	-21	7	62

TABLE 3.2 Permuted data; gives permuted difference statistic $D^* = -8.4$.

Comparing it with the original data in Table 3.1 we see that the first subject in the treatment group, the one who showed a 41 point cholesterol decrease, has become the twelfth subject in the permuted control group, the first of the actual control subjects has become the seventeenth permuted control subject, etc.

Permuting the data changes the means in the two groups,

$$\begin{aligned} &\text{permuted treatment group mean} &&= 13.1 \\ &\text{permuted control group mean} &&= 21.5 \end{aligned}$$

and of course changes our summary statistic too,

$$D^* = 13.1 - 21.5 = -8.4.$$

Here we have put a "star" on D (pronounced "D star") to distinguish it from the actual value D=22.4.

Notice that D^* is smaller (less positive) than D. This is what we would expect to happen if the Null Hypothesis were false and cholestyramine was generally effective in decreasing cholesterol levels. In that case the bigger decreases should preferentially appear in the treatment group, an effect that is destroyed by data scrambling. Permutation destroys any real cholestyramine effect, but the Null Hypothesis says that there is no such effect,

and so D should not be systematically bigger than D^*. Our first result, that the D^* in Table 3.2 is less than the actual D, is a small bit of evidence against the Null Hypothesis and for the efficacy of cholestyramine.

We can get more evidence by doing more permutations, that is by creating more versions of Table 3.2. Figure 3.2 goes whole hog in this direction. The histogram represents 2000 D^* values, each generated by an independent random permutation of the original data. Among the 2000 D^* values only 13 exceeded the actual value D=22.4. This is summarized by saying "The permutation test attained significance level 13/2000 =.0065." The ratio 13/2000=.0065 is also called a "p-value".

The strongest possible evidence against the Null Hypothesis would be a p-value of 0/2000=0, where all 2000 D^*'s were less than D. Our evidence is less strong than that, but still seems impressive. We will see next that a p-value as small as .0065 would usually be considered strong evidence against the Null Hypothesis. Now we can relay good news to the anxious researchers, "the Null Hypothesis that cholestyramine is equivalent to placebo has been strongly rejected by the permutation test. There is statistically significant evidence for the efficacy of cholestyramine in lowering cholesterol levels."

3.3 FISHER'S SCALE OF EVIDENCE

If scientific contributions are judged by usage then our next topic is a contender for the heavyweight championship of the twentieth century. Pausing from his ground-breaking mathematical labors in the development of statistical theory, R.A. Fisher rather informally proposed a scale for the interpretation of significance levels or p-values. Fisher's scale has been invoked literally millions of times since its debut in the 1920's. When you read that something is "statistically significant" in a scientific journal or the newspaper it almost always means exactly what Fisher proposed, that a significance test has attained a p-value of less than the magic level .05.

TABLE 3.3 Fisher's scale of evidence for interpreting p-values (attained sig- nificance levels) in significance testing. The smaller the p-value the stronger the evidence against the Null Hypothesis. Fisher's scale of evidence for interpreting p-values (attained sig- nificance levels) in significance testing. The smaller the p-value the stronger the evidence against the Null Hypothesis.

p-value:	.10	.05	.025	.01	.005	.001
strength of evidence against Null Hypothesis:	borderline	Moderate	substantial	strong	very strong	over- whelming

Fisher's scale of evidence appears in Table 3.3. This is a current in- terpretation, but differs only in detail from the the original proposal. The key point on the scale is .05,

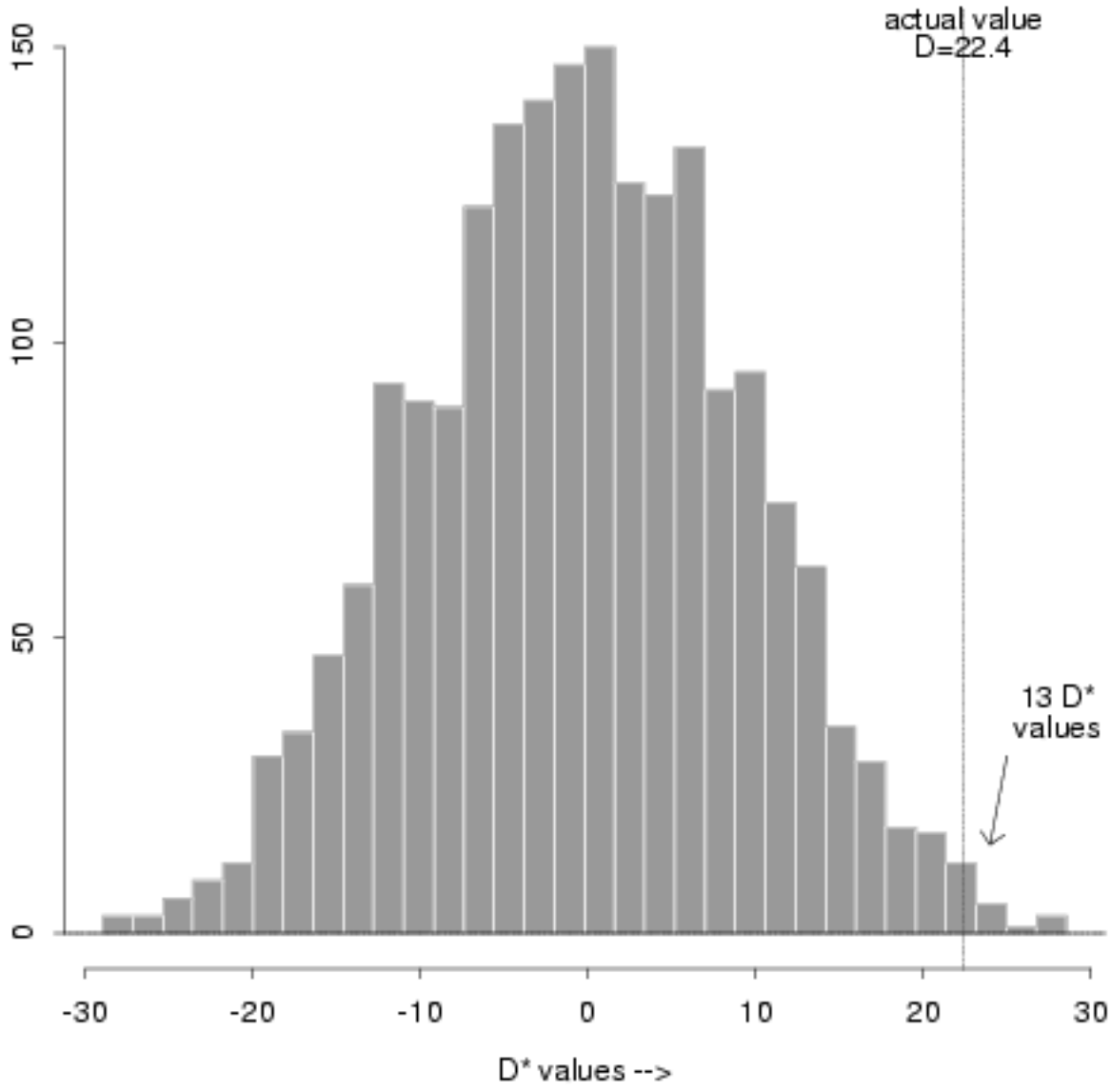


FIGURE 3.2 Histogram of the 2000 permutation values D* for the cholesterol decrease difference. Only 13 of the 2000 exceeded the actual value D=22.4. This gives significance level .0065=13/2000 , strong evidence against the Null Hypothesis that cholostyramine is no different than placebo, and for the conclusion that cholostyramine is effective in lowering cholesterol levels.

Fisher's threshold for rejection of the Null Hypothesis. It has acquired so much clout that major research projects, notably in the pharmaceutical industry[3], can live or die on the difference between a .06 p-value, "not significant", and a .04, "significant!".

This is putting much too fine a point on things. Fisher intended his scale as an aid to the research scientist trying to decide whether or not an avenue of investigation was worth further pursuit. The cholostyramine researchers should be greatly encouraged by their .0065 p-value, almost down to the "very strong" evidential level, and they have good reason to believe that cholestyramine will fare well in future investigations. (It did, as we will describe soon.) A big p-value, say .30, would probably have ended any further cholestyramine work. This is what happened, or at least should happen, with the secretin trial in chapter 2.

Experienced research workers don't consider .05 to be anything more than moderate evidence against the Null Hypothesis; .01 is much more reassuring. Going the other direction, a p-value of .10 is just on the borderline between no evidence at all and perhaps a hint of something interesting. Scientists operating under the impetus of a particularly strong hunch might persevere in the face of a .10 p-value, but they should brace themselves for disappointment.

[4]. Figure 3.3 shows histograms comparing cholesterol decreases in the two groups. Now cholestyramine's efficacy seems much clearer, and the permutation test using the mean difference, shown in Figure 3.4, gives overwhelming evidence against the Null Hypothesis: the p-value is 0, the strongest possible result, and in fact none of the D^*'s comes anywhere near the actual difference D=24.43.

Take a moment to compare Figure 3.3, all 337 subjects, with Figure 3.1, just the first 38. We can see that the early data was very sketchy, giving a noisy picture at best of the eventual results. Nevertheless our permutation test provided a clear indication of cholestyramine's effect. Significance testing is a powerful tool, designed to yield an early answer to a rather crude question: is there or is there not something promising about the treatment? Powerful tools aren't necessarily the sharpest ones, and we will be discussing other statistical methods that answer more subtle questions.

One scarcely needs a significance test for the full data set of Figure 5, the effect of the treatment being almost obvious to the eye, but it was a necessity in the small-sample context of Figure 3.1. What is a "small sample", and when do we need significance tests. Unfortunately there is no general answer to that question. If cholestyramine had

[3] The Food and Drug Administration requires .05 significance from two independent randomized clinical trials to qualify a new drug for approval.

[4] This trial, conducted at Stanford University, was one "arm" of twelve similar studies carried out at various sites in the United States. All twelve verified the efficacy of cholestyramine for lowering cholesterol levels.

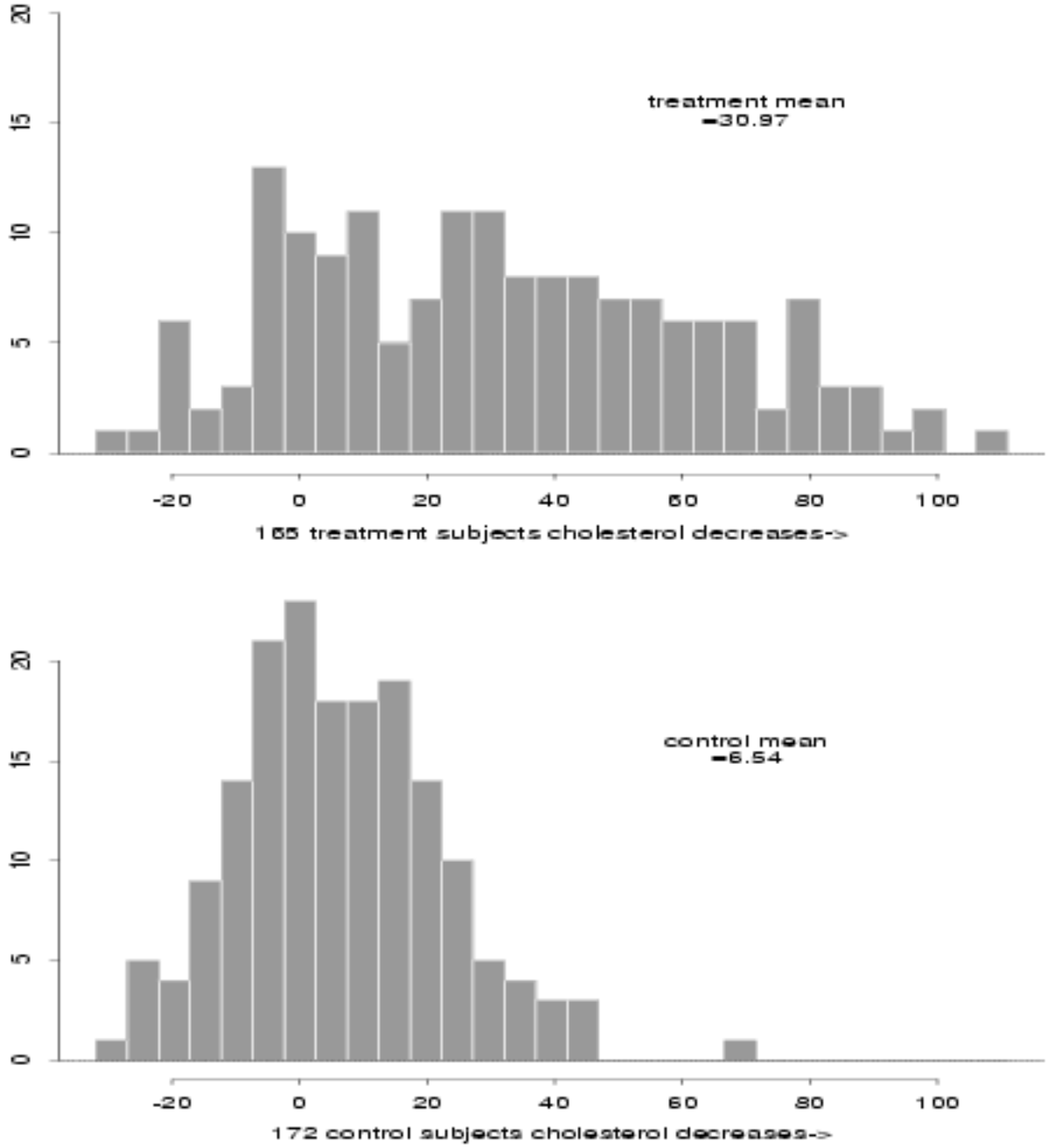


FIGURE 3.3 The cholestyramine experiment went on to enroll a total of 337 subjects, randomly assigning 165 to the cholestyramine treatment and 172 to the placebo control

been less effective, reducing chol- esterol levels by an average of say 8 points instead to the 24+ points shown in Figure 3.3, then we would have required all 337 subjects, and maybe more, to verify a significant difference. A much bigger cholestyramine effect, say 50 points reduction on average, would have made the results eyeball obvious even with only 38 subjects. Researchers usually make a preliminary guess of the effect size before planning the size of an experiment, as discussed briefly with the concept of statistical "power" in the next chapter.

3.4 HOW MANY PERMUTATIONS?

We used 2000 permutations for the test of Figure 4. This was excessive. Table 6 shows the significance levels we would have attained if we had stopped the computer earlier. Even after just 100 permutations it was clear[5] that the attained significance level was going to fall somewhere near the "strongly significant" mark.

This didn't matter much in our situation because the basic computation, taking the difference of two means, is so fast on a modern computer. All of Figure 4, including the randomizations, took less than a second of computer time. However permutation tests are often applied to much more involved pro- blems where evaluating the summary statistic of interest, the equivalent of our "D", can take a while even on a fast machine. If one evaluation of D or D^* takes a minute then we probably won't do 2000 of them. As a rough rule of thumb, 100 permutations are usually enough to get a reasonable idea of where we are on Fisher's scale; as few as 25 can be informative in truly difficult situations. It is important to remember that running more permutations doesn't create new information about the question of interest (only increasing the number of subjects does that), but it does help tell us what that information is.

TABLE 3.4 We could have stopped the permutation test in Figure 4 much earlier. Even after 100 permutations it was clear that we were going to obtain a small p-value.

Number of Permutations:	100	250	500	1000	2000
Number of D^* values exceeding D=22.4:	1	1	3	8	13
p-value:	.01	.004	.006	.008	.0065

[5] Just how clear must await Part II, where we discuss the accuracy of statistical estimates.

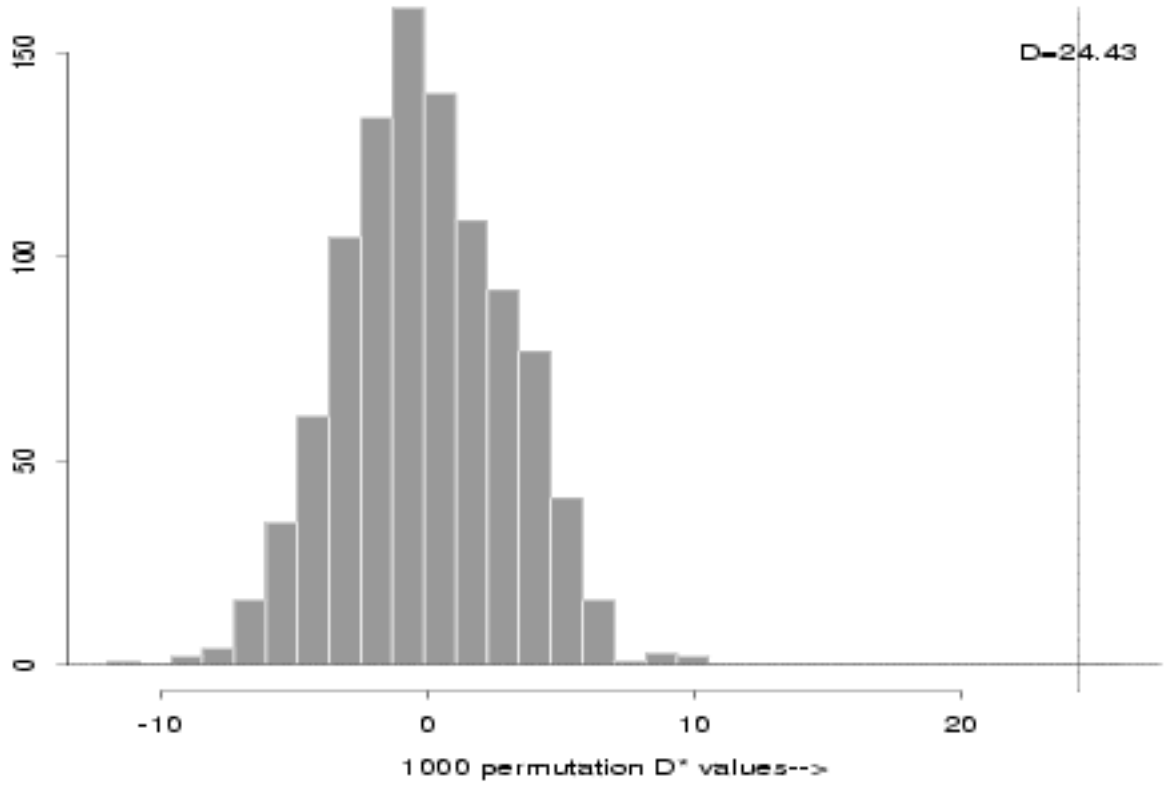


FIGURE 3.4 The permutation test on all 337 subjects isn't even close: all 1000 D* values are much than the actual value D, giving a p-value 0.

3.5 WHY .05?

The choice of .05 as the crucial point on Fisher's scale of evidence reflects a stong desire to protect the scientific literature from misleading "results". Followed scrupulously, the scale guarantees that only one out of twenty claimed effects will turn out to be false alarms. Of course Fisher could have been more conservative by choosing .01 instead of .05, or more liberal by choosing .10; however .05 has worn well with the scientific community.

Various trade-offs are at work here. The little boy who crys wolf has to be balanced with not crying wolf often enough. In the real world scientific experiments are expensive to run, both in time and money, and an overly strict standard of evidence would price smaller teams out of the research market. The choice of .05 is a convention, not a physical constant like the speed of light, but it is a convention that has become an integral part of the modern scientific method.

In actuality Fisher's choice of .05, and the other points on his scale, had a lot to do with the computational equipment of the 1920's. Mechanical calculators, loud, heavy, and slow, were the state of the art. "Computers" wete the people who ran them. Roomfuls of computers, majority of them women, labored for years to do a second's worth of electronic calculation. The computational extravagence of Figure 4 was beyond imagining, and statisticians relied on standard tables of the type discussed in chapter 4. Even these had to be quite limited in scope: rather than tabulating complete probability distributions for standardized significance tests, Fisher cut corners by publishing only a few crucial percentiles, the upper 10% point, 5% point, 2% point, and 1% point. Under these circumstances if was natural for him to seize on one of these as the basic milepost of evidence.