DAVID FREEDMAN
ROBERT PISANI
ROGER PURVES

Statistics

Fourth Edition

W • W • NORTON & COMPANY
NEW YORK • LONDON

Cartoons by Dana Fradon and Leo Cullum

The text of this book is composed in Times Roman.
Composition by Integre Technical Publishing Company, Inc.
Manufacturing by R. R. Donnelley.

Library of Congress Cataloging-in-Publication Data
Freedman, David, 1938–
Statistics. — 4th ed. / David Freedman, Robert Pisani, Roger Purves.
p. cm.
Rev. ed. of: Statistics / David Freedman . . . [et al.], 3rd ed. ©1998.
Includes bibliographical references and index.
ISBN 0-393-92972-8
ISBN 13-978-0-393-92972-0
1. Mathematical statistics. I. Pisani, Robert. II. Purves, Roger. III. Statistics. IV. Title.
QA276.F683
519.5—dc21

W.W. Norton & Company, Inc., 500 Fifth Avenue, New York, N.Y. 10110
http://www.wwnorton.com

W.W. Norton & Company Ltd., Castle House, 75/76 Wells Street, London W1T 3QT

6 7 8 9 0

To Jerzy Neyman (1894–1981)

Born in Russia, Neyman worked in Poland and England before coming to the United States in 1938. He was one of the great statisticians of our time.

1

Controlled Experiments

Always do right. This will gratify some people, and astonish the rest.
—MARK TWAIN (UNITED STATES, 1835–1910)

1. THE SALK VACCINE FIELD TRIAL

A new drug is introduced. How should an experiment be designed to test its effectiveness? The basic method is *comparison*.[1] The drug is given to subjects in a *treatment group*, but other subjects are used as *controls*—they aren't treated. Then the responses of the two groups are compared. Subjects should be assigned to treatment or control *at random*, and the experiment should be run *double-blind*: neither the subjects nor the doctors who measure the responses should know who was in the treatment group and who was in the control group. These ideas will be developed in the context of an actual field trial.[2]

The first polio epidemic hit the United States in 1916, and during the next forty years polio claimed many hundreds of thousands of victims, especially children. By the 1950s, several vaccines against this disease had been discovered. The one developed by Jonas Salk seemed the most promising. In laboratory trials, it had proved safe and had caused the production of antibodies against polio. By 1954, the Public Health Service and the National Foundation for Infantile Paralysis (NFIP) were ready to try the vaccine in the real world—outside the laboratory.

Suppose the NFIP had just given the vaccine to large numbers of children. If the incidence of polio in 1954 dropped sharply from 1953, that would seem to prove the effectiveness of the vaccine. However, polio was an epidemic disease whose incidence varied from year to year. In 1952, there were about 60,000 cases; in 1953, there were only half as many. Low incidence in 1954 could have meant that the vaccine was effective—or that 1954 was not an epidemic year.

The only way to find out whether the vaccine worked was to deliberately leave some children unvaccinated, and use them as controls. This raises a troublesome question of medical ethics, because withholding treatment seems cruel. However, even after extensive laboratory testing, it is often unclear whether the benefits of a new drug outweigh the risks.[3] Only a well-controlled experiment can settle this question.

In fact, the NFIP ran a controlled experiment to show the vaccine was effective. The subjects were children in the age groups most vulnerable to polio—grades 1, 2, and 3. The field trial was carried out in selected school districts throughout the country, where the risk of polio was high. Two million children were involved, and half a million were vaccinated. A million were deliberately left unvaccinated, as controls; half a million refused vaccination.

This illustrates the method of comparison. Only the subjects in the treatment group were vaccinated: the controls did not get the vaccine. The responses of the two groups could then be compared to see if the treatment made any difference. In the Salk vaccine field trial, the treatment and control groups were of different sizes, but that did not matter. The investigators compared the rates at which children got polio in the two groups—cases per hundred thousand. Looking at rates instead of absolute numbers adjusts for the difference in the sizes of the groups.

Children could be vaccinated only with their parents' permission. So one possible design—which also seems to solve the ethical problem—was this. The children whose parents consented would go into the treatment group and get the vaccine; the other children would be the controls. However, it was known that higher-income parents would more likely consent to treatment than lower-income parents. This design is biased against the vaccine, because children of higher-income parents are more vulnerable to polio.

That may seem paradoxical at first, because most diseases fall more heavily on the poor. But polio is a disease of hygiene. A child who lives in less hygienic surroundings is more likely to contract a mild case of polio early in childhood, while still protected by antibodies from its mother. After being infected, these children generate their own antibodies, which protect them against more severe infection later. Children who live in more hygienic surroundings do not develop such antibodies.

Comparing volunteers to non-volunteers biases the experiment. The statistical lesson: the treatment and control groups should be as similar as possible, except for the treatment. Then, any difference in response between the two groups is due to the treatment rather than something else. If the two groups differ with respect to some factor other than the treatment, the effect of this other factor might be *confounded* (mixed up) with the effect of treatment. Separating these effects can be difficult, and confounding is a major source of bias.

For the Salk vaccine field trial, several designs were proposed. The NFIP had originally wanted to vaccinate all grade 2 children whose parents would consent,

leaving the children in grades 1 and 3 as controls. And this design was used in many school districts. However, polio is a contagious disease, spreading through contact. So the incidence could have been higher in grade 2 than in grades 1 or 3. This would have biased the study against the vaccine. Or the incidence could have been lower in grade 2, biasing the study in favor of the vaccine. Moreover, children in the treatment group, where parental consent was needed, were likely to have different family backgrounds from those in the control group, where parental consent was not required. With the NFIP design, the treatment group would include too many children from higher-income families. The treatment group would be more vulnerable to polio than the control group. Here was a definite bias against the vaccine.

Many public health experts saw these flaws in the NFIP design, and suggested a different design. The control group had to be chosen from the same population as the treatment group—children whose parents consented to vaccination. Otherwise, the effect of family background would be confounded with the effect of the vaccine. The next problem was assigning the children to treatment or control. Human judgment seems necessary, to make the control group like the treatment group on the relevant variables—family income as well as the children's general health, personality, and social habits.

Experience shows, however, that human judgment often results in substantial bias: it is better to rely on impersonal chance. The Salk vaccine field trial used a chance procedure that was equivalent to tossing a coin for each child, with a 50–50 chance of assignment to the treatment group or the control group. Such a procedure is objective and impartial. The laws of chance guarantee that with enough subjects, the treatment group and the control group will resemble each other very closely with respect to all the important variables, whether or not these have been identified. When an impartial chance procedure is used to assign the subjects to treatment or control, the experiment is said to be *randomized controlled*.[4]

Another basic precaution was the use of a *placebo*: children in the control group were given an injection of salt dissolved in water. During the experiment the subjects did not know whether they were in treatment or in control, so their response was to the vaccine, not the idea of treatment. It may seem unlikely that subjects could be protected from polio just by the strength of an idea. However, hospital patients suffering from severe post-operative pain have been given a "pain killer" which was made of a completely neutral substance: about one-third of the patients experienced prompt relief.[5]

Still another precaution: diagnosticians had to decide whether the children contracted polio during the experiment. Many forms of polio are hard to diagnose, and in borderline cases the diagnosticians could have been affected by knowing whether the child was vaccinated. So the doctors were not told which group the child belonged to. This was *double blinding*: the subjects did not know whether they got the treatment or the placebo, and neither did those who evaluated the responses. This randomized controlled double-blind experiment—which is about the best design there is—was done in many school districts.

How did it all turn out? Table 1 shows the rate of polio cases (per hundred thousand subjects) in the randomized controlled experiment, for the treatment group and the control group. The rate is much lower for the treatment group, decisive proof of the effectiveness of the Salk vaccine.

Table 1. The results of the Salk vaccine trial of 1954. Size of groups and rate of polio cases per 100,000 in each group. The numbers are rounded.

The randomized controlled double-blind experiment			*The NFIP study*		
	Size	*Rate*		*Size*	*Rate*
Treatment	200,000	28	Grade 2 (vaccine)	225,000	25
Control	200,000	71	Grades 1 and 3 (control)	725,000	54
No consent	350,000	46	Grade 2 (no consent)	125,000	44

Source: Thomas Francis, Jr., "An evaluation of the 1954 poliomyelitis vaccine trials—summary report," *American Journal of Public Health* vol. 45 (1955) pp. 1–63.

Table 1 also shows how the NFIP study was biased against the vaccine. In the randomized controlled experiment, the vaccine cut the polio rate from 71 to 28 per hundred thousand. The reduction in the NFIP study, from 54 to 25 per hundred thousand, is quite a bit less. The main source of the bias was confounding. The NFIP treatment group included only children whose parents consented to vaccination. However, the control group also included children whose parents would not have consented. The control group was not comparable to the treatment group.

The randomized controlled double-blind design reduces bias to a minimum—the main reason for using it whenever possible. But this design also has an important technical advantage. To see why, let us play devil's advocate and assume that the Salk vaccine had no effect. Then the difference between the polio rates for the treatment and control groups is just due to chance. How likely is that?

With the NFIP design, the results are affected by many factors that seem random: which families volunteer, which children are in grade 2, and so on. However, the investigators do not have enough information to figure the chances for the outcomes. They cannot figure the odds against a big difference in polio rates being due to accidental factors. With a randomized controlled experiment, on the other hand, chance enters in a planned and simple way—when the assignment is made to treatment or control.

The devil's-advocate hypothesis says that the vaccine has no effect. On this hypothesis, a few children are fated to contract polio. Assignment to treatment or control has nothing to do with it. Each child has a 50–50 chance to be in treatment or control, just depending on the toss of a coin. Each polio case has a 50–50 chance to turn up in the treatment group or the control group.

Therefore, the number of polio cases in the two groups must be about the same. Any difference is due to the chance variability in coin tossing. Statisticians understand this kind of variability. They can figure the odds against a difference as large as the observed one. The calculation will be done in chapter 27, and the odds are astronomical—a billion to one against.

2. THE PORTACAVAL SHUNT

In some cases of cirrhosis of the liver, the patient may start to hemorrhage and bleed to death. One treatment involves surgery to redirect the flow of blood through a *portacaval shunt*. The operation to create the shunt is long and hazardous. Do the benefits outweigh the risks? Over 50 studies have been done to assess the effect of this surgery.[6] Results are summarized in table 2 below.

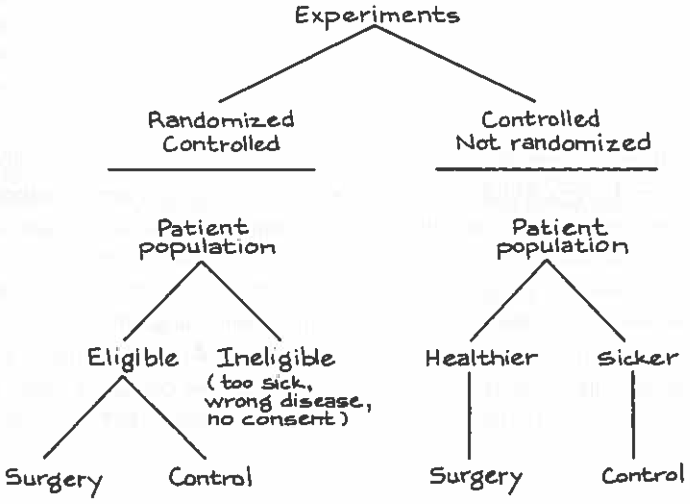
Table 2. A study of 51 studies on the portacaval shunt. The well-designed studies show the surgery to have little or no value. The poorly-designed studies exaggerate the value of the surgery.

	Degree of enthusiasm		
Design	*Marked*	*Moderate*	*None*
No controls	24	7	1
Controls, but not randomized	10	3	2
Randomized controlled	0	1	3

Source: N. D. Grace, H. Muench, and T. C. Chalmers, "The present status of shunts for portal hypertension in cirrhosis," *Gastroenterology* vol. 50 (1966) pp. 684–91.

There were 32 studies without controls (first line in the table): 24/32 of these studies, or 75%, were markedly enthusiastic about the shunt, concluding that the benefits definitely outweighed the risks. In 15 studies there were controls, but assignment to treatment or control was not randomized. Only 10/15, or 67%, were markedly enthusiastic about the shunt. But the 4 studies that were randomized controlled showed the surgery to be of little or no value. The badly designed studies exaggerated the value of this risky surgery.

A randomized controlled experiment begins with a well-defined patient population. Some are eligible for the trial. Others are ineligible: they may be too sick



to undergo the treatment, or they may have the wrong kind of disease, or they may not consent to participate (see the flow chart at the bottom of the previous page). Eligibility is determined first; then the eligible patients are randomized to treatment or control. That way, the comparison is made only among patients who could have received the therapy. The bottom line: the control group is like the treatment group. By contrast, with poorly-controlled studies, ineligible patients may be used as controls. Moreover, even if controls are selected among those eligible for surgery, the surgeon may choose to operate only on the healthier patients while sicker patients are put in the control group.

This sort of bias seems to have been at work in the poorly-controlled studies of the portacaval shunt. In both the well-controlled and the poorly-controlled studies, about 60% of the surgery patients were still alive 3 years after the operation (table 3). In the randomized controlled experiments, the percentage of controls who survived the experiment by 3 years was also about 60%. But only 45% of the controls in the nonrandomized experiments survived for 3 years.

In both types of studies, the surgeons seem to have used similar criteria to select patients eligible for surgery. Indeed, the survival rates for the surgery group are about the same in both kinds of studies. So, what was the crucial difference? With the randomized controlled experiments, the controls were similar in general health to the surgery patients. With the poorly controlled studies, there was a tendency to exclude sicker patients from the surgery group and use them as controls. That explains the bias in favor of surgery.

Table 3. Randomized controlled experiments vs. controlled experiments that are not randomized. Three-year survival rates in studies of the portacaval shunt. (Percentages are rounded.)

	Randomized	*Not randomized*
Surgery	60%	60%
Controls	60%	45%

3. HISTORICAL CONTROLS

Randomized controlled experiments are hard to do. As a result, doctors often use other designs which are not as good. For example, a new treatment can be tried out on one group of patients, who are compared to "historical controls:" patients treated the old way in the past. The problem is that the treatment group and the historical control group may differ in important ways besides the treatment. In a controlled experiment, there is a group of patients eligible for treatment at the beginning of the study. Some of these are assigned to the treatment group, the others are used as controls: assignment to treatment or control is done "contemporaneously," that is, in the same time period. Good studies use contemporaneous controls.

The poorly-controlled trials on the portacaval shunt (section 2) included some with historical controls. Others had contemporaneous controls, but assignment to the control group was not randomized. Section 2 showed that the design of a study matters. This section continues the story. Coronary bypass surgery is a widely used—and very expensive—operation for coronary artery disease. Chalmers and associates identified 29 trials of this surgery (first line of table 4). There were 8 randomized controlled trials, and 7 were quite negative about the value of the operation. By comparison, there were 21 trials with historical controls, and 16 were positive. The badly-designed studies were more enthusiastic about the value of the surgery. (The other lines in the table can be read the same way, and lead to similar conclusions about other therapies.)

Table 4. A study of studies. Four therapies were evaluated both by randomized controlled trials and by trials using historical controls. Conclusions of trials were summarized as positive (+) about the value of the therapy, or negative (−).

Therapy	*Randomized controlled*		*Historically controlled*	
	+	−	+	−
Coronary bypass surgery	1	7	16	5
5-FU	0	5	2	0
BCG	2	2	4	0
DES	0	3	5	0

Note: 5-FU is used in chemotherapy for colon cancer; BCG is used to treat melanoma; DES, to prevent miscarriage.
Source: H. Sacks, T. C. Chalmers, and H. Smith, "Randomized versus historical controls for clinical trials," *American Journal of Medicine* vol. 72 (1982) pp. 233–40.[7]

Why are well-designed studies less enthusiastic than poorly-designed studies? In 6 of the randomized controlled experiments on coronary bypass surgery and 9 of the studies with historical controls, 3-year survival rates for the surgery group and the control group were reported (table 5). In the randomized controlled experiments, survival was quite similar in the surgery group and the control group. That is why the investigators were not enthusiastic about the operation—it did not save lives.

Table 5. Randomized controlled experiments vs. studies with historical controls. Three-year survival rates for surgery patients and controls in trials of coronary bypass surgery. Randomized controlled experiments differ from trials with historical controls.

	Randomized	*Historical*
Surgery	87.6%	90.9%
Controls	83.2%	71.1%

Note: There were 6 randomized controlled experiments enrolling 9,290 patients; and 9 studies with historical controls, enrolling 18,861 patients.
Source: See table 4.

Now look at the studies with historical controls. Survival in the surgery group is about the same as before. However, the controls have much poorer survival

rates. They were not as healthy to start with as the patients chosen for surgery. Trials with historical controls are biased in favor of surgery. Randomized trials avoid that kind of bias. That explains why the design of the study matters. Tables 2 and 3 made the point for the portacaval shunt; tables 4 and 5 make the same point for other therapies.

The last line in table 4 is worth more discussion. DES (diethylstibestrol) is an artificial hormone, used to prevent spontaneous abortion. Chalmers and associates found 8 trials evaluating DES. Three were randomized controlled, and all were negative: the drug did not help. There were 5 studies with historical controls, and all were positive. These poorly-designed studies were biased in favor of the therapy.

Doctors paid little attention to the randomized controlled experiments. Even in the late 1960s, they were giving the drug to 50,000 women each year. This was a medical tragedy, as later studies showed. If administered to the mother during pregnancy, DES can have a disastrous side-effect 20 years later, causing her daughter to develop an otherwise extremely rare form of cancer (clear-cell adenocarcinoma of the vagina). DES was banned for use on pregnant women in 1971.[8]

4. SUMMARY

1. Statisticians use the *method of comparison*. They want to know the effect of a *treatment* (like the Salk vaccine) on a *response* (like getting polio). To find

out, they compare the responses of a *treatment group* with a *control group*. Usually, it is hard to judge the effect of a treatment without comparing it to something else.

2. If the control group is comparable to the treatment group, apart from the treatment, then a difference in the responses of the two groups is likely to be due to the effect of the treatment.

3. However, if the treatment group is different from the control group with respect to other factors, the effects of these other factors are likely to be *confounded* with the effect of the treatment.

4. To make sure that the treatment group is like the control group, investigators put subjects into treatment or control at random. This is done in *randomized controlled experiments.*

5. Whenever possible, the control group is given a *placebo*, which is neutral but resembles the treatment. The response should be to the treatment itself rather than to the idea of treatment.

6. In a *double-blind* experiment, the subjects do not know whether they are in treatment or in control; neither do those who evaluate the responses. This guards against bias, either in the responses or in the evaluations.