The What, Why, and How of Biostatistics in Medical Research

1.1 DEFINITION OF STATISTICS AND BIOSTATISTICS

The *Oxford Dictionary of Statistics* (2002, p. 349) defines statistics as "The science of collecting, displaying, and analyzing data." Statistics is important in any scientific endeavor. It also has a place in the hearts of fans of sports, particularly baseball. Roger Angel in his baseball book, *Late Innings*, says "Statistics are the food of love."

Biostatistics is the branch of statistics that deals with biology, both experiments on plants, animals, and living cells, and controlled experiments on humans, called clinical trials. Statistics is classified by scientific discipline because in addition to many standard methods that are common to statistical problems in many fields, special methods have been developed primarily for certain disciplines. So to illustrate, in biostatistics, we study longitudinal data, missing data models, multiple testing, equivalence and noninferiority testing, relative risk and odds ratios, group sequential and adaptive designs, and survival analysis, because these types of data and methods arise in clinical trials and other medical studies. Engineering statistics considers tolerance intervals and design of experiments. Environmental statistics has a concentration in

the analysis of spatial data, and so does geostatistics. Econometrics is the branch of statistics studied by economists, and deals a lot with forecasting and time series.

Statisticians are professionals trained in the collection, display, and analysis of data and the distribution theory that characterizes the variability of data. To become a good applied statistician, one needs to learn probability theory and the methods of statistical inference as developed by Sir Ronald A. Fisher, Jerzy Neyman, Sir Harold Jeffreys, Jimmie Savage, Bruno deFinetti, Harald Cramer, Will Feller, A. N. Kolmogorov, David Blackwell, Erich Lehmann, C. R. Rao, Karl and Egon Pearson, Abraham Wald, George Box, William Cochran, Fred Mosteller, Herman Chernoff, David Cox, and John Tukey in the twentieth century. These are some of the major developers of the foundations of probability and statistics. Of course, when selecting a list of famous contributors like this, many have been unintentionally omitted. In the late twentieth century and early twenty-first century, computer-intensive statistics arose, and a partial list of the leaders of that development are Brad Efron, Leo Brieman, David Freedman, Terry Speed, Jerry Friedman, David Siegmund, and T. L. Lai. In the area of biostatistics, we should mention Thomas Fleming, Stuart Pocock, Nathan Mantel, Peter Armitage, Shein-Chung Chow, Jen-pei Liu, and Gordon Lan. You will be introduced to these and other famous probabilists and statisticians in this book. An applied statistician must also become familiar with at least one scientific discipline in order to effectively consult with scientists in that field.

Statistics is its own discipline because it is much more than just a set of tools to analyze data. Although statistics requires the tools of probability, which are mathematical, it should not be thought of as a branch of mathematics. It is the appropriate way to summarize and analyze data when the data contains an element of uncertainty. This is very common when measurements are taken, since there is a degree of inaccuracy in every measurement. Statisticians develop mathematical models to describe the phenomena being studied. These models may describe such things as the time a bus will arrival at a scheduled stop, how long a person waits in line at a bank, the time until a patient dies or has a recurrence of a disease, or future prices of stocks, bonds, or gasoline.

Based on these models, the statistician develops methods of estimation or tests of hypotheses to solve certain problems related to the data.

Because almost every experiment involves uncertainty, statistics is the scientific method for quantitative data analysis.

Yet in the public eye, statistics and statisticians do not have a great reputation. In the course of a college education, students in the health sciences, business, psychology, and sociology are all required to take an introductory statistics course. The comments most common from these students are "this is the most boring class I ever took" and "it was so difficult, that I couldn't understand any of it." This is the fault of the way the courses are taught and not the fault of the subject. An introductory statistics course can be much easier to understand and more useful to the student than, say, a course in abstract algebra, topology, and maybe even introductory calculus. Yet many people don't view it that way.

Also, those not well trained in statistics may see articles in medicine that are contradictory but still make their case through the use of statistics. This causes many of us to say "You can prove anything with statistics." Also, there is that famous quote attributed to Disraeli. "There are lies, damn lies and statistics." In 1954, Darrell Huff wrote his still popular book, *How to Lie with Statistics*. Although the book shows how graphs and other methods can be used to distort the truth or twist it, the main point of the book is to get a better understanding of these methods so as not to be fooled by those who misuse them. Statisticians applying valid statistical methods will reach consistent conclusions. The data doesn't lie. It is the people that manipulate the data that lie. Four books that provide valuable lessons about misusing statistics are Huff (1954), Campbell (1974), Best (2001), and Hand (2008).

1.2 WHY STUDY STATISTICS?

The question is really why should medical students, physicians, nurses, and clinicians study statistics? Our focus is on biostatistics and the students we want to introduce it to. One good reason to study statistics is to gain knowledge from data and use it appropriately. Another is to make sure that we are not to be fooled by the lies, distortions, and misuses in the media and even some medical journals. The medical journals now commonly require good statistical methods as part of a research paper, and the sophistication of the methods used is greater.

So we learn statistics so that we know what makes sense when reading the medical literature, and in order to publish good research.

We also learn statistics so that we can provide intelligent answers to basic questions of a statistical nature. For many physicians and nurses, there is a fear of statistics. Perhaps this comes from hearing horror stories about statistics classes. It also may be that you have seen applications of statistics but did not understand it because you have no training. So this text is designed to help you conquer your fear of statistics. As you learn and gain confidence, you will see that it is logical and makes sense, and is not as hard as you first thought.

Major employers of statisticians are the pharmaceutical, biotechnology, and medical device companies. This is because the marketing of new drugs, biologics, and most medical devices must be approved by the U.S. Food and Drug Administration (FDA), and the FDA requires the manufacturers to demonstrate through the use of animal studies and controlled clinical trials the safety and effectiveness of their product. These studies must be conducted using valid statistical methods. So any medical investigator involved in clinical trials sponsored by one of these companies really needs to understand the design of the trial and the statistical implications of the design and the sample size requirements (i.e., number of patients need in the clinical trial). This requires at least one basic biostatistics course or good on-the-job training.

Because of uncontrolled variability in any experimental situation, statistics is necessary to organize the data and summarize it in a way so that signals (important phenomena) can be detected when corrupted by noise. Consequently, bench scientists as well as clinical researchers need some acquaintance with statistics. Most medical discoveries need to be demonstrated using statistical hypothesis testing or confidence interval estimation. This has increased in importance in the medical journals. Simple *t*-tests are not always appropriate. Analyses are getting much more sophisticated. Death and other time-to-event data require statistical survival analysis methods for comparison purposes.

Most scientific research requires statistical analysis. When Dr. Riffenburgh (author of the text *Statistics in Medicine*, 1999) is told by a physician "I'm too busy treating patients to do research," he answers, "When you treat a patient, you have treated a patient. When you do research, you have treated ten thousand patients."

In order to amplify these points, I will now provide five examples from my own experience in the medical device and pharmaceutical

industries where a little knowledge of statistics would have made life easier for some of my coworkers.

In the first scenario, suppose you are the coordinator for a clinical trial on an ablation catheter. You are enrolling subjects at five sites. You want to add a new site to help speed up enrollment. The IRB for the new site must review and approve your protocol for the site to enter your study. A member of the IRB asks what stopping rule you use for safety. How do you respond? You don't even know what a stopping rule is or even that the question is related to statistics! By taking this course, you will learn that statisticians construct stopping rules based upon accumulated data. In this case, there may be safety issues, and the stopping rule could be based on reaching a high number of adverse events. You won't know all the details of the rule or why the statistician chose, it but you will at least know that the statistician is the person who should prepare the response for the IRB.

Our second example involves you as a regulatory affairs associate at a medical device company that just completed an ablation trial for a new catheter. You have submitted your premarket approval application (PMA). In the statistical section of the PMA, the statistician has provided statistical analysis regarding the safety and efficacy of your catheter in comparison to other marketed catheters. A reviewer at the FDA sent you a letter asking why Peto's method was not used instead of Greenwood's approximation. You do not know what these two methods are or how they apply.

From this course, you will learn about survival analysis. In studying the effectiveness of an ablation procedure, we not only want to know that the procedure stopped the arrhythmia (possibly atrial fibrillation), but also that the arrhythmia does not recur. Time to recurrence is one measure of efficacy for the treatment. Based on the recurrence data from the trial, your statistician constructs a time-to-event curve called the Kaplan–Meier curve.

If we are interested in the probability of recurrence within 1 year, then the Peto and Greenwood methods are two ways to get approximate confidence intervals for it. Statistical research has shown differences in the properties of these two methods for obtaining approximate confidence intervals for survival probabilities. As an example, Greenwood's estimate of the lower confidence bound can be too high in situations where the number of subjects still at risk at the time point of interest is small.
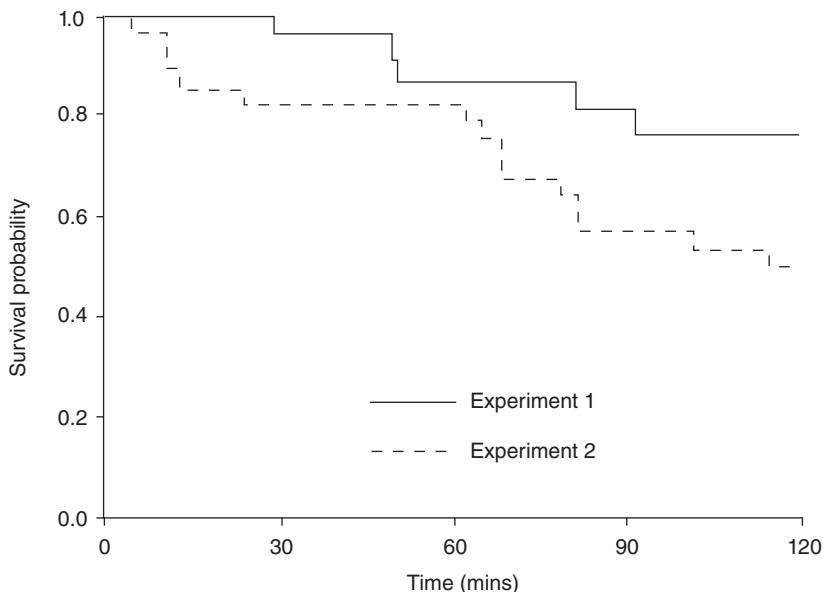
374 Analysis of survival times

Figure 1.1.   Example of a Kaplan–Meier curve. Taken from Altman (1991), *Practical Statistics for Medical Research*. Chapman and Hall/CRC, p. 374.

In these situations, Peto's method gives a better estimate of this lower bound. In general, neither method is always superior to the other. Since the FDA posed this question, the statistician would opt to provide the Peto estimate in addition to Greenwood for the FDA to compare the two lower confidence bounds. Knowing these simple facts would help you deal with the FDA question quickly, effectively, and accurately (Fig. 1.1).

In situation 3, you are in regulatory affairs and are reviewing an FDA letter about a PMA submission. The FDA wants you to report results in terms of confidence intervals, in addition to the *p*-values, before they give final approval to the treatment. You recognize this as a statistical question, but are worried because if it takes significant time to supply the request, the launch date of the new device will be delayed and will upset marketing's plans. You don't even know what a confidence interval is!

In this case, since you have the necessary data to do the binomial test on success probability, you can easily compute an exact confidence

interval. Your statistician can provide this for you in less than 1 day and you are greatly relieved.

In situation 4, you are a clinical research associate in the middle of an important phase III trial. Based upon a data analysis done by the statistics group and an agreement with the FDA prior to the trial, the primary endpoint can be changed from a condition at the 6-month follow-up visit to that same condition at the 3-month follow-up visit. This is great news, because it means that the trial can be finished sooner!

There is a problem though. The protocol only required follow-up visits at 2 weeks and 6 months, and the 3-month follow-up was optional. Unfortunately, some sites opted not to conduct the 3-month follow-up. Your clinical manager now wants you to have all the patients that are past the 3-month time point since the procedure was done and did not have the 3-month follow-up to come in for an unscheduled visit. When you requested that the investigators do this, a nurse and one investigator balked at the idea and demanded to know why this is necessary. You need an answer from your statistician!

To placate the investigator, the statistician tells the investigator that they could not use the 3-month follow-up initially because the FDA had not seen data to indicate that a 3-month follow-up would be enough to determine long-term survival. However, during the early part of the trial, the statistician was able to find relevant survival curves to indicate the survival probability flattens out at 3 months' duration. This was enough to convince the FDA that the 3-month endpoint was sufficient to determine long-term survival. If we now have the unscheduled visits, these could be the subjects' last visit, and many subjects will not need a 6-month follow-up, allowing a shorter accrual time and a chance to get the product to market faster.

This explanation helped, but the problem could have been avoided had the clinician had the foresight to see the importance of making the 3-month follow-up mandatory in the protocol. The investigator was pleased because although it would cost more to add these unscheduled visits, this would be more than compensated by the dropping of the 6-month follow-up, for those getting the unscheduled visit, and possibly some others.

In the last situation (situation 5), imagine you are the VP of the Clinical and Regulatory Affairs Departments at a medical device company. Your company hired a contract research organization (CRO) to run a blinded randomized control phase III clinical trial. You have a

statistics group, but the CRO is tasked to handle the data collection, processing, and analysis, so as to keep your company blinded and thus maintain greater integrity for the data and to avoid any presumption of bias.

The CRO can view the data in an unblinded fashion as they prepare their report. You are very curious to see the results, since a successful trial outcome is of paramount importance. Now, as the report is complete, you are the only representative of the company who can see the report. As you look at the report, you see p-values for statistical tests. You recall only a little statistics but remember to look for p-values below 0.05 because those were indicative of statistical significance. You are alarmed, when looking at a demographic comparison of treatment and control groups by age and gender, to see high p-values. One p-value was 0.56. You would like to show this to your statistician, but cannot, because he must remain blinded.

If you had taken a course like this one, you would know that for efficacy variables, the hypotheses are set up to be rejected, and low p-values are good. But we want the demographic factors to be nearly the same for both groups. For demographics, we do not want to reject the null hypothesis, and a high p-value is actually good news!!

The main reason for similarity between the groups with respect to all these demographic factors is randomization. Fisher originally suggested randomization in experiments because of confounding of effects. Perhaps unknown to the investigators, the treatment is more effective in women than men. Suppose we have 100 patients in each group. In the control group, 30 are women and 70 are men. In the treatment group, 80 are women and 20 are men, and we see a statistically significant effect. Is it due to the treatment or the fact that so many more women are in the treatment group than in the control group? Unfortunately, we do not know! This is what is called confounding.

Randomization overcomes this problem because it tends to balance out factors that we are not interested in. Simple random sampling will proportion the men and women nearly in the proportions that they occur in the patient population. This too avoids bias and confounding. In situation 5, the high p-value shows that the randomization is doing its job!

We now summarize what we have learned in this section.

1. Statistics and statisticians played an important role in research. Their role in medical research and particularly randomized

controlled clinical trials continues to increase rapidly, as the demand for finding new and better treatments for severe diseases increases.

2. The regulatory agencies and pharmaceutical companies continue to emphasize controlled clinical trials for the evaluation of efficacy and safety for a new drug.

3. Physicians and nurses cannot ignore statistics. It is everywhere, and is mandated by the FDA to provide proof of safety and efficacy of new drugs, devices, and combination therapies.

1.3 THE MEDICAL LITERATURE

Chapter 6 of Doug Altman's book, Altman (1991), discusses statistical methods in the medical literature. He quotes the famous statistician, Sir David Cox, who in 1983 said: "One does feel that statistical techniques both of design and analysis are sometimes adopted as rituals designed to assuage the last holders of absolute power (editors of journals) and perhaps also regulatory agencies, and not because the techniques are appreciated to be scientifically important." I agree with this statement not only as it applied in 1983, but even to a large extent, still today, 27 years later!

Altman uses very strong language regarding problems with the medical literature. He claims "Examples of substandard design and incorrect analysis can be seen in almost any issue of any medical journal." He goes on to say: "The importance of sound design and analysis cannot be overemphasized. Clearly the conclusions from a study must rely on the methods having been correct. If conclusions are unreliable because of faulty methodology, then the study cannot be clinically worthwhile. Worse, it may be clinically harmful by reason of the conclusions being misleading, and a clinically harmful study is surely unethical."

Evidence of the growth of the use of statistical methods in medical research is given in this table about the journal *Pediatrics*, taken from table 16.1, page 479 of Altman's book. The number of papers is on an increasing trend, the percentage of papers without statistics is decreasing, and the percentage with more sophisticated techniques is increasing over the three decades (Table 1.1).

Table 1.1

Use of Statistical Procedures in the Journal *Pediatrics*

	Year			
	1952	1962	1972	1982
No. of papers	67	98	115	151
% with no statistical procedures	66%	59%	45%	30%
% with procedures other than *t*, chi-square, or *r*	3%	5%	12%	35%

From Altman (1991) with permission.

Table 1.2

Errors Found in *Arthritis and Rheumatism* 1967–1968 Compared With 1982 (Continued)

Year of publication	1967–1968	1982
Number of papers	*n* = 47	*n* = 74
Error type		
Undefined method	14 (30%)	7 (9%)
Inadequate description of measure of location or dispersion	6 (13%)	7 (9%)
Repeated observations treated as independent	1 (2%)	4 (5%)
Two groups compared on more than 10 variables at 5% level	3 (6%)	4 (5%)
Multiple *t*-tests instead of ANOVA	2 (4%)	18 (24%)
Chi-squared tests used when observed frequencies are too small	3 (6%)	4 (5%)
At least one of these errors in the paper	28 (60%)	49 (66%)

From Altman (1991), with permission.

From 1982 to 2010, this trend has continued, and fortunately, the quality of the statistical refereeing has improved as well. Altman also looked at errors in a particular journal, *Arthritis and Rheumatism*, comparing the late 1960s to 1982 (Table 1.2).

We see from the tables that the medical literature was notorious for incorrect use of statistical methods. Trends from the late 1960s to the

early 1980s show an increase in the use of statistical methods and particularly the more sophisticated ones. The frequency of occurrence of elementary-type errors declined over this period. Because statistics is used more frequently and with more sophistication, there is an increase in the percentage of papers that have at least one error, as well as an increase in the percentage of papers that contain the more recent type of errors from multiple testing and the use of multiple *t*-tests instead of the analysis of variance.

1.4 MEDICAL RESEARCH STUDIES

Medical research studies involving human subjects can be put into four categories.

1. Cross-sectional studies
2. Retrospective studies
3. Prospective studies (other than clinical trials)
4. Controlled clinical trials, including pharmacokinetic and phar- macodynamic studies

While the controlled clinical trial falls under the category of pro- spective studies, we choose to separate it out because of its clear importance in the evaluation of new drugs and medical devices.

1.4.1 Cross-Sectional Studies Including Surveys

Definition: A cross-sectional study is one that is taken at a given point in time.

Surveys including election polls and censuses are both examples of cross-sectional studies. These studies are conducted when only one point in time is relevant to the question at hand (e.g., censuses, public opinion polls, election polls, and marketing surveys). Here, only the current opinion matters. Not interested in looking far into the future. But often in medicine, we are interested in changes over time after a medical intervention. This goes for both efficacy variables and quality of life variables. So we do not see many cross-sectional studies in medical research except in epidemiological studies.

1.4.2 Retrospective Studies

Definition: A retrospective study is one that examines relationships based on past data.

One important example of a retrospective study is the case-control study (these could also be prospective). Such studies are intended to be similar to what prospective clinical trials are intended to do. The cases are the subjects with the outcome of interest (like a treatment group in a clinical trial). The control subjects are similar demographically or otherwise to their matched case subjects, but for which the outcome did not occur.

A particular example might be a situation where subjects who contracted a particular disease such as lung cancer are asked about their past exposure to a risk factor. The same questions are administered to control subjects who did not get lung cancer. In this case, the risk factor is cigarette consumption.

1.4.3 Prospective Studies Other Than Clinical Trials

Definition: A prospective study is one that is planned in the present and takes place in the future.

Examples include cohort studies and clinical trials. Clinical trials are particularly important to us, as we have already mentioned. So we consider them as a category of their own.

An example of a cohort study is a study that follows a group of disease-free subjects who have a certain risk factor for a disease to see if they eventually develop the disease. The subjects could be young college students, the disease could be emphysema, and the risk factor could be smoking. From cohort studies, statisticians and epidemiologists determine relative risks based on exposure levels to risk factors.

1.4.4 Controlled Clinical Trials

In the context of clinical trials, an experiment is a study that assigns subjects to treatment groups in order to assess differences among treatments. A randomized experiment is one in which randomization is used for the selection process.

Definition: An experiment performed to evaluate the effect of intervention(s) or treatment(s) for a group of human subjects against a control group that does not get a treatment(s) (placebo) or gets different treatment(s).

The purpose is to see if the difference in treatment creates differences in outcomes for the treatment group versus the control group. The gold standard for clinical trials is the double-blinded randomized controlled trial. When constructed properly, these trials provide good statistical information about the differences between two groups or several groups (often there can be more than one treatment). The control group could be on a drug that is an active competitor to the study drug or on placebo, or, more generally, a different treatment protocol, a different medical device or surgical procedure, and so on.

The use of randomization and blinding is to protect the study from biases that could invalidate the results. Not all clinical trials are *blinded, randomized, or completely prospective*. Sometimes in device trials, historical controls or objective performance criteria (OPCs) are used for comparison with the treatment. This makes the comparator retrospective while the treatment is done prospectively. Since the trial only has one arm, there is no blinding or randomization in this type of trial.

1.4.5 Conclusions

1. There are several types of studies in medical research.
2. Each study has its advantages and disadvantages.
3. Cross-sectional studies only look at one point in time.
4. Most medical research and particularly clinical trials are concerned with how patients improve or get worse over time as a function of alternative treatments.
5. Because of (4), cross-sectional studies are not common in medical research other than in some epidemiologic studies.
6. Double-blind randomized control clinical trials provide the gold standard for evaluating a new treatment versus current standard care and/or placebo when done properly. But they are also the most costly and difficult to implement studies.

1.5 EXERCISES

1. What is a Kaplan–Meier curve?
2. For what kind of data do you compute Kaplan–Meier curves?
3. Why is randomization important in clinical trials?
4. What does Greenwood's method refer to?
5. Why do we compute *p*-values? When is it good for *p*-values to be small and when is it all right if they are large?
6. What are cross-sectional studies and why are they uncommon in medical research?
7. What are retrospective studies?
8. What are prospective studies?
9. What are controlled clinical trials and why is blinding important?