

# 2

## The Basis of Statistical Reasoning in Medicine

### 2.1 The Cow in the Investigator's House

Consider the following Russian folk proverb:

There is a poor farmer who is preparing for the coming winter. Although he had a miserably small and unproductive plot of land and an old cow, his only possession of any value is his house. In reality, it's just a one room shack, but he is nevertheless proud of it and its warm fireplace.

The winter that comes is especially brutal. On one frigid day, the farmer looks out from his warm room and sees his cow, shivering and mooing in the harsh wind outside. He thinks for a moment, and then lets the cow into his house with him. The farmer doesn't approve of what he has done. He doesn't like the idea of letting an animal into his house. He detests the fact that the animal takes up so much of his room. He hates the smell. But he needs the cow to survive. So, in order to live, he gets used to it.\*

Health researchers with a nonmathematical background “get used to statistics” in order to survive in research. Appearing to be an amorphous mixture of hard unforgiving mathematics and nebulous concerns about “the freak of chance”, statistics can seem to be the worst of everything. A successful businessman relates the following story.

Sure, I enjoyed mathematics in high school and in college. I actually made the mistake of trying to take a second course in college algebra, and did fine, right up until we got to this thing called “ $e$  to the  $x$ ”. When asked about the ingredients

---

\* Taken from a debate between Soviet Premier Nikita S. Khrushchev and members of the Soviet Politburo in 1962, at the height of the Cuban Missile Crisis.

that made up this curious entity, the professor said “I can’t tell you exactly what “ $e$ ” is, but “ $x$ ” can be anything.” That’s when I got up, walked right out of the math building, and over to the business school!

To the many healthcare researchers who have no special training in mathematics, statistical thinking is like entering a hall of mirrors. Investigators are interested in proving what they believe, yet statistics seems to focus on disproving what they don’t believe. Interpreting multiple endpoints in studies can be particularly complex and troublesome, because although some results are generalizable, others are not. These counterintuitive complications deepen the suspicion that many investigators hold about this mysterious field. Most researchers go into research not because of statistics, but in spite of it.

Clinical investigators need not be experts in statistical computation, but they should be experts in statistical reasoning, that is identifying that relatively small set of circumstances that justify applying results from small samples to large populations. The reliability of the results of a clinical trial rest on the investigators’ abilities to separate a true signal of a population effect from the background noise created by the random aggregation of subjects in the sample. A unifying philosophy is critical for interpreting these experiments.

Appropriate statistical reasoning in the presence of interim monitoring is an even more delicate matter. The smaller number of subjects, greater imprecision in the effect size estimates, and the occasional unplanned nature of the analysis combine to complicate, and commonly obfuscate, the best interpretation. Thus, successful implantation of interim monitoring requires not just calculation, but a clear view of that calculation’s meaning.

This chapter provides a brief overview of the salient statistical issues in clinical research, serving as a preamble for the discussion of monitoring guidelines that is the main subject of this text. Useful references for more detailed discussions of these issues are available [1,2,3,4].

## **2.2 Research, Populations, and Samples**

The purpose of research is simply to learn, and learning in healthcare requires that we study patients. However, we are faced with the inescapable observation that we cannot study everyone in the population that we wish to understand.

### ***2.2.1 The Tail Wagging the Dog***

Consider a researcher who wishes to execute a clinical trial to assess the effect of a new intervention on the overall mortality rate for stroke patients. However, when we press her, we discover that she has a much larger, more audacious goal. There is no question of her clear honest intent to learn about the effect of therapy in these 300 patients. However, she is more focused on applying her results to the U.S. stroke population. Specifically, she wishes to take the findings from her 300 patient study and apply them to the 600,000 patients who have a stroke in the United States

each year. If she could have, she would have studied all 600,000 patients, but this was logistically, financially, and ethically\* impossible.

Thus, even though she would like to study the entire population, she cannot. What are the implications of this restriction? We know for certain that  $600,000 - 300 = 599,700$  patients about whom she wants to learn, she in fact will never study. These subjects were never recruited, never randomized, never treated, never followed, and never measured. Yet she claims that she will learn something of value about these unevaluated subjects. This is tantamount to allowing a very small tail to wag a very big dog.

Specifically, studying a small sample drawn from a large population leaves most of the population unstudied and introduces uncertainty. Of course the difficulty is compounded when the researcher ends a study before its scheduled termination time. An early conclusion produces less information available to serve as the basis of generalization to a population. How then can her answer be assured when not just most of the population of interest, but in the case of interim monitoring, much of the sample of interest remains unstudied in her research effort?

The simple and honest response is that there is no guarantee that any sample-based answer is correct. However, there are practices that the investigator can follow that will improve the reliability of the sample-based finding. Specifically, these procedures will make it more likely that the sample's results closely reflects the answer residing in and governing the population from which the sample was drawn. However, even with the use of these modern approaches, the ability to generalize the results from a sample to a population remains limited. Furthermore, the best statistical monitoring effort can not compensate for the weaknesses of a poorly designed and badly executed research effort.

## 2.3 Three Principles of Sample-Based Research

The *primun movens* of sample-based research is to generalize results from a small sample to a large population. Yet the justification for this extension is not in the motivation, but in the research effort's procedures. Good methodology is greatly assisted by mathematics; however, application of the best research efforts requires an appreciation of sampling that is not so easily quantifiable.

The act of selecting a sample of subjects from the population at large is a combination of science and art. The scientific aspect of sampling is simply the mathematical mechanism used to identify the relatively small number of subjects from the population who will comprise the sample. The art of the process is the ability to tailor the sample to provide clear objective answers to the questions that generated the research. This sculpting ability can be sensitively applied only when the investigator understands what a sample can, and cannot provide.

Because the primary purpose of the research is to learn about a population, the primary purpose of the sample is to *represent* that population. Each individual selected in the sample is selected not just for his or her own attributes, but also to stand in for the many hundreds, thousands, or sometimes millions of patients who

---

\*Many patients, for varied personal reasons, would in all likelihood not have consented to the study.

were never selected for the sample. Therefore, it is simply not enough to measure the germane events that occur in an individual. These occurrences must be evaluated in a way that allows that measurement to represent the unobtainable observations from the unselected members of the population. This difficulty is compounded by the early termination of a study, because the number of subjects on which the termination decision is made is commonly only a fraction of the number of subjects the researchers prospectively declared was the minimum number necessary to complete the effort. In a sense, the “representative value” of a single subject’s measurements increases in the early termination environment.

In addition, the investigator recognizes that there is another phenomenon complicating efforts to generalize his results. Different samples, when selected from a single population, will contain different individuals with different life experiences, producing different data. Although the data can be similar across some samples, other samples will reveal marked differences. This sample-to-sample variability is called sampling error. Because the number of subjects on which an early termination decision rests is smaller than the number required to complete the study, the potential effect of sampling error is greater at the interim monitoring point than at the conclusion of the study.

The presence of sampling error raised the question of how likely it is that, despite the investigator’s best efforts, the population generated an unrepresentative sample. This possibility is always present in sample-based research, and has an important impact on the use and interpretation of monitoring procedures. Therefore, investigators who wish to draw conclusions from this type of research are obligated to report the degree to which sampling error may have influenced their results.

We may summarize these three principles of good methodological\* execution of sample-based research as

- Principle 1. Clearly define your question, then select from the population a sample that is representative of the population and whose study will be responsive to your query.
- Principle 2. Carry out your sample-based measurements in such a way that the findings in your sample can stand for not just the sample results but can also accurately represent the results that occurred if the study had been carried out in every member of the population.
- Principle 3. Accurately measure and report the degree to which sampling error may have misled you.

### **2.3.1 Analysis Triaging**

By their nature, investigators will and must analyze what they believe is illuminating, informative, and interesting. However, they are also obligated to report those findings clearly and in a manner that provides the best interpretation of these re-

---

\* The issue of ethics is central to a productive research effort, but is not the subject of this chapter.

sults. Some analyzes have a clear interpretation and are the most generalizable. Others are not.

Analysis triage can guide the investigators' consideration of generalizable versus hypothesis-generating results. This triage process divides a research effort's evaluations into several clearly defined categories, each of which has a clear interpretation. It is a two-phased process.

The first phase determines if the candidate analysis is confirmatory or exploratory. Is the analysis to be prospectively planned or data driven? The major advantages of prospectively planned analyzes are that the estimates of effect size, confidence intervals, and standard errors are trustworthy. Alternatively, data-driven evaluations, although commonly carried out, and frequently exciting, are less reliable. Post hoc, exploratory results should be executed and reported, but they must be clearly labeled as exploratory. These evaluations require confirmation before they can be integrated into the fund of knowledge of the medical community.

The second level of triage during the design phase of the clinical trial is carried out among the prospectively planned analyzes, dividing them into primary analyzes or secondary analyzes. Primary analyzes are those analyzes on which the conclusions of the trial rest. Each of the primary analyzes will have a prospectively set type I error level attached to it in such a way that the overall (or familywise) type I error rate does not exceed the community accepted level (traditionally 0.05). The trial will be seen as positive, null (no finding of benefit or harm), or negative (harmful result) based on the results of the primary analyzes. It is important to note that a clinical trial can have more than one primary evaluation. If appropriately designed, the study can be judged as positive if any of those primary endpoints is positive.

Secondary endpoints are prospectively declared analyzes in which no attempt is made to control the familywise error rate. Typically, each secondary analysis is typically interpreted at nominal (i.e., judged significant if the  $p$ -value is less than 0.05) levels. Secondary analyzes, being prospectively designed, produce trustworthy estimates of effect sizes, precision, and  $p$ -values. However, because secondary analyzes do not control the familywise error, the risk of a false positive finding to the population is too great for confirmatory conclusions to be based upon them. Therefore, the role of secondary endpoints is to provide support for the primary endpoint findings, and not to serve as independent confirmatory analyzes.

In the typical clinical trial, there are more exploratory analyzes than there are prospectively declared endpoints, and more secondary endpoints than there are primary endpoints. This is a finding that is consistent with the statement that a small number of key questions should be addressed, accompanied by careful deliberation on the necessity and extent of adjustment for multiple comparisons [5].

We will return to these three principles in Chapters Five through Seven and the impact that they have on monitoring clinical research.

## 2.4 The Monitoring Complication

The two issues of (1) a sample's ability to represent a population, and (2) the role of sampling error are twin forces that, if not correctly assessed and balanced, can de-

stroy the utility of a clinical study's results. This is, of course, why clinical research must be carefully designed and executed.

Methodological execution requires that the investigator collect a specified number of patients who meet clearly defined criteria and follow them until the end of the study. Only a result of at least a pre-specified magnitude, measured with appropriate precision would be considered positive. However, an investigator who is tasked with monitoring the research effort often finds herself in the position of working to draw a conclusion about a research effort's results before the study is completed. If the study was designed to randomize a precise number of patients and follow them for a pre-specified time period in order to draw an unambiguous conclusion from the study, then doesn't the early termination of a study undermine this well-considered effort?

In many circumstances, the answer is yes. The early termination of a study, if incorrectly managed, will undercut the researchers' efforts to produce a clearly interpretable research result. However, there is a precise set of circumstances in which a clinical research effort may be terminated early and still provide adequate assurances of the result's validity. The clinical investigator's role is to design the research effort so that these circumstances can be created and preserved during the study's execution.

The statistical methodology that governs the monitoring of clinical research must be embedded in the design of a research effort so that its execution, and reliance on its conclusions does not undermine the overall research effort. It must be prospectively detailed, unambiguous, and lead to conclusions that are supported by all of the trial's methodology, in concordance with the three aforementioned principles of sample-based research. These are important constraints, and within these constraints, very few studies can be ended early. Our goal is to understand how the intelligent use of statistical monitoring procedures can aid in the identification of that precise set of circumstances that would lead to the successful and early termination of a clinical research effort.

## 2.5 The Need for Prospective Design

The need for a clear early statement for the design of a research effort has two general motivations. Although the first is self evident the second is hidden. However, like an iceberg, it is the second, submerged component that is commonly the most damaging when not recognized.

### 2.5.1 Sample Vision

The first motivation for the prospective design of a study is administrative. Any enterprise, including scientific endeavors, that requires resources needs careful planning to first obtain and then utilize those resources. If one is going to carry out a study evaluating the effect of a genome-drug interaction on short-term lung function (e.g., forced expiratory volume at 1 second ( $FEV_1$ )) then the necessary logistics must be in place to produce precise, reproducible measures of  $FEV_1$ . If, on the other hand, the purpose of the study is to provide information about the long-term mortality of its participants, then different measurement mechanisms must be in place. These include (1) the legal mechanism to obtain hospital charts and death certifi-

cates,\* (2) the availability of specialists to determine the cause of death, and 3) the expertise to carry out the appropriate analyzes. Each of these designs is feasible, but requires different resources, and time is required to make these resources available. Clearly, knowing what the study will measure allows the investigator to husband the necessary resources for the study.

### 2.5.2 Advantages of the Random Sample

A second reason for a prospective design is that early thought must be given to how the sample will be selected. A sample should be selected that, in general, represents the population. However, more specifically, the sample must allow a clear depiction of the relationship that the investigator wishes to illuminate in the population. Ideally, this will involve a random selection mechanism.

We have discussed one use of random mechanisms in clinical research earlier. In Chapter One, discussion focused on attribution of effect, a property that is most directly produced by the random allocation of therapy.<sup>†</sup> However, the random allocation of therapy is a procedure that is executed after a subject has been selected for sample inclusion. The random selection of subjects is a different mechanism, with a different motivation.

If each population member has the same likelihood of being chosen for the sample, then no member is excluded a priori at the expense of another, and there are no built-in biases against any subject based on that subject's individual characteristics. The procedure that precludes this bias is the *random selection mechanism*, and this process generates a *simple random sample*. The process by which individuals are selected randomly from the population for the sample ensures that every patient in the population has the same opportunity (statisticians say the same constant probability<sup>‡</sup>) of being selected for the sample.

### 2.5.3 Limitations of the Random Sample

There are two caveats that we must keep in mind when considering simple random samples. The first is that they are rarely achieved, due to the operation of a set of exclusion criteria. These exclusion criteria are required for logistical and ethical reasons. Sometimes they are used to identify a cohort or collection of individuals that are most likely to demonstrate the relationship that the research is designed to identify.<sup>§</sup> Because each exclusion restricts a patient from entering the study based on a characteristic of that patient, the body of exclusion criteria makes the sample less representative of the general population. The inability of most clinical trials to

---

\* This aspect of clinical research has become both more important and frustrating as society has become more concerned and restrictive about access to the personal information of individuals.

<sup>†</sup> Discussed on pages 7–10.

<sup>‡</sup> There are more complicated mechanisms that involve random selection. Only the simplest is described here.

<sup>§</sup> A fine example is clinical trials that exclude patients who are believed by the investigator to be (1) unlikely to comply with the intervention if it is self-administered over a period of time or (2) unwilling to complete a rigid follow-up attendance schedule.

achieve a sample that even approximates a simple random sample is an important limitation of this research tool.

A second caveat is the incomplete operation of the random chance mechanism in small samples. Because samples exclude most patients from the population, we would expect that a particular sample is not going to represent each of the innumerable descriptive facets of the population. A sample of 1000 patients from a population of 19 million diabetic patients will not provide representative age–ethnic–educational background combinations.\* It is asking too much of the sampling mechanism to produce a relatively small sample that is representative of each and every property and trait of the individuals in the population. Thus, the sample must be shaped by the investigator so that it is representative of the population for the traits that are of greatest interest. This contouring process represents a compromise. The sample is created to be representative of some aspects of the population, therefore it is not going to be representative of others. Thus, the resulting sample will have a spectrum of representation, accurately reflecting a relatively small number of traits of the population, and producing inaccurate depictions of others.

Investigators who are unaware of this spectrum, and who therefore report every result from their study as though those results were valid and generalizable simply because they were produced by a random sample, can mislead the medical community. This is a dangerous trap for investigators because it is so easy to collect unrelated but “interesting” data from a study that was itself designed to evaluate a separate question.

For example, consider an investigator interested in determining the change in exercise tolerance in patients with congestive heart failure. After she collects a sample of 300 patients and assesses their exercise tolerance over time, she also queries them about the frequency of hospitalizations for heart failure. In the end, this investigator reports not just the rate of change of exercise tolerance, but also the hospitalization rate of her cohort. She thinks that this is appropriate because she believes that the sample was “representative”, and therefore, the hospitalization data are just as reliable as the exercise tolerance data. However, the study was not designed to measure hospitalizations. Patients who were likely to be hospitalized could not meet the entry criteria for exercising, and thus never had the opportunity to enter the study. Hospitalization discharge data was not collected with the same attention to detail as exercise tolerance data. Thus, the ease of collecting data for an evaluation that was not considered during the design of the study, in concert with a sample that was nonrepresentative of hospitalization rates combine to provide a misleading statement about the hospitalization rates for these patients.

This situation is complicated when the study is being monitored for efficacy and safety. During the interim evaluations of this study, the investigator examines both exercise tolerance data and hospitalization rates. As the dataset grows, trends appear and disappear in the dataset for both exercise tolerance and hospitalizations. However, the hospitalization rate interim results can be misleading. The

---

\* For example, if the proportion of subjects who are Hispanic, greater than 65 years of age, and have a graduate school education is less than 1 in 1000, then a sample of 1000 patients is not likely to select any members of the population with these characteristics at all, and the sample will be unresponsive to any questions about this subgroup.



combination of (1) a set of conditions that deselected patients likely to be hospitalized, (2) the inability to obtain good quality data for hospitalization rates, and (3) the random aggregation of subjects in a sample combine to create trends that are not representative and would provide only a misleading perspective on what trends in sample hospitalization rates actually imply for the population at large.

Thus, simple random samples are only representative of the aspect of the population that they were explicitly and overtly designed to measure. Observing a population through a sample is like viewing a complicated and intricately detailed landscape through glasses. It is impossible to grind the glass lens so that every object in the landscape can be viewed with the same sharp detail. If the lens is ground to view near objects, then the important features of the distant objects are distorted. On the other hand, if the lens is ground for the clear depiction of distant objects, then near objects are blurred.

This is a major motivation for concentrating a research effort on a small number of inquiries. By focusing on this short list of questions, investigators are able to choose a sample containing patients with the desired characteristics. However, the wise investigator understands that, by focusing on a small number of prospectively stated questions and selecting a sample that provides representative views of these issues, the sample may not be representative of other characteristics of the population. The sample's results will most likely be generalizable for the questions that it was designed to answer, but not for much else.

### 2.5.4 Trustworthy Estimators

Measurements on research subjects are combined into estimators that are known by specific names (e.g., means, standard deviations, odds ratios, and relative risks). However, they all have the same function—to provide a reliable estimate of a quantity in the population.

It is easy in this modern computational era to take the reliability of these estimators for granted. However, decades of work were required to identify them and to gain consensus on their use [6].\* These estimators have pleasing mathematical properties and are designed to work well in the sampling error environment. It is important to note that they were not designed to remove sampling error. Instead, they channel it into both effect size estimates (e.g., means) and the variability of these estimates (e.g., standard deviations and confidence intervals). If the researcher is also interested in inference (i.e., statistical hypothesis testing), then statistical procedures will channel sampling error into  $p$ -values and power. Thus, when used correctly, statistical methodology will appropriately recognize and transmit sampling error into familiar quantities that researchers can interpret.

The estimators were designed to perform well in the presence of sampling error. However, for them to function effectively, there can be no other source of

---

\*The idea of repeating and combining observations made on the same quantity appears to have been introduced as a scientific method by Tycho Brae towards the end of the 16<sup>th</sup> century. He used the arithmetic mean to replace individual observations as a summary measurement. The demonstration that the sample mean was a more precise value than a single measurement did not appear until the end of the 17<sup>th</sup> century, and was based on the work of the astronomer Flamsten.

random variability. When other, nonsystematic error is present in the research, the estimators on which we rely become unreliable; that is, they no longer measure what they were intended to measure. This most commonly happens in the paradigm of “random research”.<sup>\*</sup> Specifically, random research is the circumstance when a sample-based dataset produces an answer to a question that the investigator did not prospectively think to ask.

We already know that one difficulty with the random research paradigm is that the sample was not created to answer the nonprospectively asked question whose answer is suggested by the data analysis. However, a second difficulty lies in the environment in which the estimator is now expected to operate. A sample will, if interrogated often enough, suggest a provocative answer to a question that it was not designed to address. The difficulty with accepting this solution is that other samples from the same population will suggest (1) a different and most times, less provocative answer to this question, and (2) other provocative answers to other questions. These surprise results, being random, appear and disappear across samples; it is not the population transmitting its “signal” through the sample but instead is merely the sampling error “noise” making itself heard.

Estimators function appropriately when they incorporate random data that is gathered in response to a fixed question. They do not perform so well when the selection of the research question is itself random, that is, left to the data. Operating like blind guides, these estimators mislead us about what we would see in the population based on our observations in the sample. The result is a wavering research focus, leaping from one provocative finding to the next, careening wildly about on the random waves of sampling error. Therefore, a primary purpose of the prospective design is to fix the research questions, so that their analysis is well anchored.[7]<sup>†</sup> This distinction between confirmatory and exploratory analyzes will be particularly important (if not troublesome) in the discussion of interim monitoring for safety in Chapter Seven.

## 2.6 The Role of Probability and Statistics

The requirement of a sample with its consequent sampling error complicates the interpretation of healthcare research. The implications of sampling error are profound, forcing the investigator to predict or estimate what would happen in the population based on what is observed in the sample. Sampling error can distort the view of the population, and if the research questions are not insulated from the effects of sampling error, this source of variability can wreck the ability of the estimators to provide any useful information at all.

Because the researchers’ efforts to predict and estimate population effects from sample findings (1) are quantitative, and (2) must acknowledge and incorporate the notion of random error, it is only natural that they incorporate statistics. Statistics focuses on the ability to first estimate a population quantity based on data that is obtained from a sample, and then, if necessary, infer a population relation-

---

<sup>\*</sup> This issue is discussed in Chapter 2 of [4].

<sup>†</sup> The problems with midstream changes in analysis plans occasionally rises to the level of public awareness. This most recent was a statement by an FDA scientist.

ship (e.g., treatment-induced lower cumulative fatal and nonfatal stroke rate) from observations in the population. We will now turn our attention to the basic idea behind statistical inference.

## 2.7 Statistical Inference

The preceding discussion of the use of the sample has prepared us for the notion of statistical inference. The purpose of drawing the sample is to learn something of value about the population from which the sample was obtained, that is, to infer from the sample to the population. Statistical inference is the process by which that inference is carried out.

We already understand some of the pivotal steps in the inference process. The researcher must choose the appropriate estimators. The environment in which these estimators operate must not be shaken by the perturbations generated by changes in a research protocol that are induced by findings in the data. These concepts are central to the ability of these estimators to estimate what they were designed to measure.

Statistical inference focuses on what to do with these estimators once they have been obtained. Although the concept of statistical inference in healthcare is well accepted, there have been important and continuous disagreements as to how this inference should be carried out. The use of formal hypothesis testing, a tradition that has strong roots in the medical research community, took root in the 1940s and remains a central component of healthcare research. This paradigm involves the construction of null and alternative hypotheses, and ultimately the generation of a  $p$ -value. The groundswell of enthusiasm for this perspective has been tracked and discussed [8]. In fact, the hypothesis testing scenario has become so popular that the notion of statistical inference and statistical hypothesis testing have become synonymous in healthcare research. However, there are other approaches to drawing conclusions from a sample to a population that do not involve formal hypothesis testing that have demonstrated themselves to be worthy competitors.

### 2.7.1 Confidence Intervals

Although there has been a 60 year tradition of carrying out formal statistical hypothesis testing, an influential community of epidemiologists has developed a continuous and formidable resistance to its application to healthcare research. The appearance of misleading research results from studies that have abused the hypothesis testing scenario, in concert with the combination of sample size, effect size, and variability into one number has caused many in epidemiology to eschew the  $p$ -value for the confidence interval[9].

The confidence interval provides important and useful information about the role that sampling error plays in the generation of the result. Incorporating the point estimate and its standard error, the confidence interval provides a readily interpretable assessment of the point estimate's precision.

This concept can be illustrated by an example from a recent clinical study. The Heart Outcomes Prevention Evaluation (HOPE) trial was designed to assess the effect of the ACE-i therapy ramipril on clinical measures of cardiovascular disease [10]. It was well-designed, and executed in accordance with its protocol (i.e., the

research effort was *concordantly executed*). At its conclusion, one of its major findings was the effect of ramipril on the combined measure of myocardial infarction (MI), stroke, or cardiovascular (CV) death. The relative risk for this effect was 0.78\* and the 95% confidence interval was 0.70 to 0.86[11].

This 95% confidence interval draws attention to the range of possible values of the relative risk in the population from which the HOPE sample was drawn. A common interpretation of the confidence interval is conveyed by saying that it is likely that the value of the true relative risk lies somewhere in this 0.70 to 0.86 range.†

Although useful, sole reliance on the confidence interval has its opponents. One criticism of the confidence interval approach is that it does not easily lead to dichotomous decisions (e.g., is the therapy effective). Some workers rely on whether the confidence interval contains the value 1 (signifying no therapy effect) as evidence that the therapy is unlikely to be effective in the population; therefore this is tantamount to carrying out a hypothesis test, a procedure that the worker may be attempting to avoid. Finally, the confidence interval, much like the estimate of the relative risk itself, is only accurate insofar as the research environment is a concordant one. Data-based changes in the protocol undermine the confidence interval estimate as easily as they destabilize the estimate of the relative risk.

## 2.7.2 Bayesian Procedures

An alternative to the traditional hypothesis testing paradigm is the implementation of Bayes procedures. Their underlying philosophy is distinct enough from the standard (or frequentist approach) to statistical inference that many now view the two perspectives as polar opposites, and the literature is replete with vigorous debates between the zealots of each philosophy. However, here we will steer well clear of these controversies, contenting ourselves with a brief review of each approach.

### 2.7.2.1 Classical Statistics (the “Frequentists”)

Classical statistics is the collection of statistical techniques and devices that evaluate the accuracy of a technique in terms of its long-term, repetitive accuracy [12]. The conclusion from any particular research effort may be wrong. However, if the experiment were repeated many times, the application of classic hypothesis testing procedures would produce the correct answer most times. This concept of the over-

---

\* The relative risk of the effect of therapy demonstrated a  $1 - 0.78 = 0.22$  or 22% reduction in the incidence of the combined endpoint associated with the use of ramipril.

†This is not the most accurate definition, because it suggests that the variability is associated with the population relative risk which, in this paradigm, is constant. The sample-to-sample variability is associated with the location of the 95% confidence interval and whether it contains the population relative risk. The most accurate interpretation of the HOPE-generated confidence interval is as follows. If there were 100 samples obtained (in this case, this would mean that 100 HOPE studies were performed), each with its own confidence interval, then 95% of these confidence intervals would contain the true population relative risk. Of course, with only one study, and one confidence interval, we do not know one way or the other whether this confidence interval contains the true value of the relative risk.

all accuracy of the procedure buoys the confidence of the classical statistician, even though the wrong conclusion may be obtained from any particular experiment. However, for researchers who have much of their time (and sometimes, the taxpayers or stockholders' money) bound up in an important research effort, this observation can produce a small shock. The realization that sampling error can lead to an erroneous result in even an expensive, well-designed, and well-executed research effort is not very comforting.

We have already seen this principle in operation. The use of the confidence interval to estimate a relative risk does not guarantee that any particular experiment will generate a confidence interval that contains the true value of the relative risk. Instead, the underlying principle provides an assurance that, in the overwhelming number of samples (95% of these samples for a 95% confidence interval), the confidence interval will contain the population relative risk.

Another frequentist characteristic is the focus on not just what has occurred, but what has not occurred in a research program. How these non-occurrences are handled can have a dramatic impact on the answer to a scientific question, and preoccupation with them can bedevil the investigator.

### 2.7.2.2 The Bayesians

Like classical statistics, Bayes theory is applicable to problems of parameter estimation and hypothesis testing. The Bayesian formulation is based on a principle, termed the likelihood principle, which states that a decision should have its foundation in what has occurred, not in what has not occurred.

Like frequentists, Bayesians are interested in parameter estimation and hypothesis testing. Bayesians estimate the population parameter  $\theta$  of a distribution (just as frequentists do). However, unlike frequentists who believe that the parameter is constant, Bayesians treat the parameter as though it itself has a probability distribution. This is called the *prior distribution*, signified as  $\pi(\theta)$ .

Once the prior distribution is identified, the Bayesian works forward, next identifying the probability distribution of the data given the value of the parameter. This distribution is described as the *conditional distribution* (because it is the distribution of the data conditional on the value of the unknown parameter) and is denoted as  $f(x_1, x_2, x_3, \dots, x_n | \theta)$ . This step is not unlike that of the frequentist. When attempting to identify the mean change in blood pressure for a collection of individuals, both the frequentist and the Bayesian may assume that the distribution of blood pressures for this sample of individuals follows a normal distribution with an unknown mean, whose estimation is the goal. However, the frequentist treats this unknown mean as a fixed parameter. The Bayesian assumes that the parameter is not constant, but changes over time. Its distribution is called the *prior distribution*.

The Bayes process continues by combining the prior distribution with this conditional distribution to create a *posterior distribution*, or the distribution of the parameter  $\theta$  given the observed sample, denoted as  $\pi(\theta | x_1, x_2, x_3, \dots, x_n)$ . From the Bayes perspective, the prior distribution reflects knowledge about the location and behavior of  $\theta$  before the experiment is carried out. After the experiment is executed, the researcher has new information in the form of the conditional distribution.

These two sources of information are combined to obtain a new estimate of  $\theta$ . To help in interpreting the posterior distribution, some Bayesians will construct a loss function, which identifies the penalty that they pay for underestimating or overestimating the population parameter. Bayesian hypothesis testing on the value of the parameter is based on the posterior distribution.

The Bayesian approach to statistical analysis makes unique contributions. It explicitly considers available prior distribution information, and allows construction of a loss function that directly and clearly states the loss (or gain) for each decision. However, the requirement of a specification of the prior distribution can be a burden if there is not much good information about the parameter to be estimated. Similarly, the choice of the loss function can be difficult to justify from a clinical perspective.

Several interim monitoring procedures have been developed that reflect the Bayes perspective about which we will have more to say in Chapter Eight.

### 2.7.3 Hypothesis Testing Paradigm

The scientific method, easily recognized as the driving force motivating research efforts, begins with an idea. Investigator conceived and formulated, this idea is commonly an affirmative one; for example, a new class of drug is an acceptable alternative to coumadin for the prevention of stroke in patients with atrial fibrillation (AF). This clinical postulate either represents scientific truth or it does not. In order to determine the accuracy and applicability of the researcher's concept, the investigators carry out an experiment. During the design of this research, the clinical hypothesis is converted into one (or a collection of) statistical hypotheses.

The scientific method begins with a hypothesis or initial idea that the researcher hopes to prove. If  $\theta_A$  is the cumulative stroke rate in the active group, and  $\theta_C$  is the cumulative stroke rate in the control group, then the investigator believes and states his clinical hypothesis as  $\theta_A < \theta_C$ . However, statistical hypothesis testing commonly begins with a hypothesis that the researcher hopes to disprove or nullify. Thus, the investigator who believes the new class of drugs is beneficial will commonly state a *null hypothesis*. In the current example, the null hypothesis is that patients who are assigned to the new class of drugs will have the same stroke rate as those patients who were assigned to coumadin. Thus, the statistical null hypothesis is not that  $\theta_A < \theta_C$  but that  $\theta_A = \theta_C$ . It is this null hypothesis that the investigator wishes to disprove or nullify with the experiment's results.

The reason for this change of emphasis from a positive clinical hypothesis to a null statistical hypothesis deserves some discussion. The investigator cannot be blamed for his first impression that, by being forced to turn away from proving an affirmative hypothesis to disproving a null one, he has lost the intellectual initiative. However, the investigator must recognize that he himself has chosen to be involved in an act of nullification. Specifically, he has chosen to nullify the current approach that is used to prevent post-AF thrombotic events.

Prior to the investigator's research, the current, accepted standard of care in the medical community is that coumadin is the best outpatient therapy available to reduce the stroke rate in patients with AF. By believing that the new class of drugs is better than coumadin, clearly an affirmative concept, the investigator an-

nounces his nonacceptance of the assertion that the accepted standard of care is optimum. He wishes to nullify this belief, and he will do that by designing a trial that demonstrates the effectiveness of the new class of drug.

The design of the clinical trial is central to this process. Patients are randomized to active or control group therapy in order to minimize differences between the groups. Investigators endeavor to ensure that patients are treated similarly across the two groups. They work to reduce differences in compliance with medication between the groups. The investigators determine the occurrence of clinical endpoints without knowledge of the patient's assigned therapy. This system is constructed so that, if the current standard of care is correct, there will be no differences in the cumulative stroke rates between the two groups. Thus, if there are important differences in the stroke incidence rates, they can be due to only two reasons: (1) the freak of chance produced by the random aggregation of patients in the research sample, or (2) the therapy actually made a difference.

Therefore, the null hypothesis is merely a mathematical characterization of the current practice of medicine. It is consistent with an underlying theme of the clinical trial; that is, if the therapies being tested have the same effect on the clinical endpoint, then the data and the trial support the state of the art, or the null hypothesis.

When the findings are more extreme, they are commonly described as "unlikely to be due to chance alone". This means that the sample-to-sample variability is too small to serve as the only explanation for the large difference in the stroke rates between the two groups. This statement has come to be encapsulated in the  $p$ -value.

### 2.7.4 $P$ -Values

The  $p$ -value is a measure of sampling error. It is simple in concept, but its long and complex history is undeniable. Before we put their use in context, let's first discuss what  $p$ -values are supposed to be, and then acknowledge what they have become.

When the positive conclusions of a well-designed, concordantly executed research program are placed in her hand, the researcher must acknowledge two possible explanations for the results. The first is that the sample truly represents the findings of the larger population. However, the second explanation is that the sample's results are due to sampling error and do not represent the population.

This second explanation is motivated by what sampling error can produce in the absence of a real treatment effect in the population. In this situation, the positive research finding does not accurately reflect relationships in the population. Instead, the positive sample findings are unique to the sample. They are not seen in, nor are they representative of, the population. Just as a "population" of 1000 fair coins when flipped could produce a "sample" of five coins that all showed "heads", a population that has no effect can produce a sample which contains, just through the play of chance, an important "treatment effect".

Because sampling error is always present in sample-based research, we can never know for sure whether a sample's results are representative of the population or just due to the play of chance. The  $p$ -value simply measures the likelihood that sampling error has produced a positive result in the research sample of the in-

investigator. If sampling error produced the research result, the researcher would be wrong in concluding that the effect seen in her sample represents a true finding in the population. The  $p$ -value is the probability that a population in which there is no effect would produce a sample that demonstrates an effect.\* The smaller the  $p$ -value, the less satisfactory is the explanation that sampling error explained the results, and the more likely a truly representative research result was identified.

The idea of the  $p$ -value and significance testing is based on the work of the agricultural statistician Ronald Fisher.† As he worked through the design and analyzes of agrarian experiments in the 1920s, he stated that, if there was a greater than five percent chance that a population that had no positive findings produced a sample with positive findings, the positive findings in the sample should be discarded because the likelihood that they were due to the random, meaningless aggregation of events was too great [13,14].

This was the beginning of “significance testing”, and the “ $p < 0.05$ ” concept. There is no deep mathematical theory that points to 0.05 as the optimum type I error level—only tradition. The rise and pre-eminence of the 0.05 level has its roots less in science and more in the “sociology of science”, as 1940s journal editors and senior grant reviewers struggled with differentiating worthy scientific results from second-tier ones [15,16].

Unfortunately, many researchers have substituted the 0.05 criterion for their own thoughtful, critical review of a research effort, and this replacement has led to uninformed research interpretation. Poole [17] pointed out that the mechanical reflexive acceptance of  $p$ -values at the 0.05 level is the nonscientific, easy way out of critical and necessary scientific discussions. For example, highly statistically significant effects (i.e., results associated with small  $p$ -values) have been produced by small, inconsequential effect sizes. In other research efforts, small  $p$ -values themselves were rendered meaningless when the assumptions on which they had been computed were violated. In addition, there is the observation that statistical significance may not indicate true scientific, biological, clinical, or economic significance [18,19, 20, 21, 22].

The reduction of a complex research endeavor’s result to a single  $p$ -value is perhaps at the root of the inappropriate role of significance testing. This condensation effort may be due to the fact that the  $p$ -value is itself constructed from several constituents. Sample size, effect size, and effect size variability are important components of the  $p$ -value and are directly incorporated into the  $p$ -value’s formulation. However, in reality, what is produced is not a balanced measure of these important contributory components, but only a measure of the role of sampling error as a possible explanation for the results observed in the research sample. Thus,  $p$ -values are deficient reflections of the results of a research effort, and must be supplemented with additional information (the research effort’s concordance,‡ sample

---

\* The  $\alpha$  error rate is the type I error rate that is set before the research begins. The  $p$ -value is the measure of  $\alpha$  that is based on the result of the research.

† This is the same Ronald Fisher whose contribution of the tool of randomization was discussed in Chapter One.

‡ Concordance is the desirable property of research that derives from the tight match between the research execution and the plans for its execution as stated in the research protocol.



size, effect size, and effect size precision) in order for the study to receive a fair and balanced interpretation [23]. The investigator must jointly consider these measures when interpreting a research endeavor's results.

### 2.7.5 Statistical Power

Consider a research effort that is designed to identify the effect of a therapeutic intervention in a sample of patients. As we saw previously, sampling error can produce from a population in which there is no therapeutic effect of interest a sample in which there is a significant treatment effect. This misleading sample is generated by the random and unpredictable selection of subjects in the sample, that is, by chance alone.

However, the influence of sampling error can be equally insidious when the study results are not positive. After all, it is quite possible that a population in which there is a treatment effect of interest may produce a sample in which there is no effect. In this case, the researcher is compelled to conclude that, because his research sample produced no effect of interest, the therapy is not effective for its studied use. However, this would be a false result because this sample that produced the null finding was produced by chance alone. This is called a *type II error*. The probability that a population in which there is an important treatment effect also produces a sample containing that same effect is the power of the study, and may be computed as simply  $Power = 1 - P[Type II error]$ .

Unlike the case of the  $p$ -value, where there has been a strong tradition of setting the threshold at 0.05, the minimal acceptable power for a study has been a standard that has changed over time. Acceptable power levels can extend up to and sometimes exceed 95%. Rarely, however, is a study acceptable that is based on a power level of less than 80%.

Which one of the type I error or type II error the reader of a well-designed and concordantly executed\* study should track depends on the findings of the research. If the study results are positive, then the reader must focus on the likelihood that a false positive study could be produced through sampling error. This is a concern that focuses on the  $p$ -value. If, on the other hand, the study findings are null, the reader turns her attention to the power of the study. If the power of the study is high, she may assume that it is unlikely that a population in which the research effect was important would produce by chance alone a sample in which the research effect was absent.

During the design phase of the study, the researcher will not know which of these two errors may occur, so he must design the study with a priori concern for each of these errors.

---

\* Of course, if the research is not designed well, or is executed poorly, then these measures of sampling error can be corrupted and therefore, inaccurate. One learns how well the research was designed and executed from an examination of the methodology section of the manuscript. See Chapter 4, Section 3.2 Systematic Reviews from Moyé [22].

## 2.8 Conclusions

Study results from undisciplined research efforts can produce provocative findings with no lasting value. The same tendency, left unchecked, will also plague the results produced from the interim monitoring on clinical research. These intermediate results, based on smaller samples, powered by small  $p$ -values, can with loud voice point the medical community in the wrong direction.

As long as research is based on drawing a sample from a population, sampling error will play a role in the product of that research. Two important precepts to follow are (1) clearly define your question, (2) select from the population a sample that is representative of the population, and (3) develop a clear a priori protocol and follow that protocol during the conduct of the research. Adhering to these principles will contain the extent and the limit the role of sampling error as an explanation of the research results, be they interim results or the final results of the study.

## References

1. Friedman L, Furberg C, Demets D. (1996) *Fundamentals of Clinical Trials*. 3<sup>rd</sup> Edition. New York, Springer.
2. Meinert CL. (1986) *Clinical Trials Design, Conduct, and Analysis*, New York, Oxford University Press.
3. Piantadosi S. (1997) *Clinical Trials: A Methodologic Perspective*. New York, John Wiley.
4. Moyé L. (2003). *Multiple Analyzes in Clinical Trials: Fundamentals for Investigators*. New York. Springer.
5. Proschan MA, Waclawiw MA. (2000) Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials* **21**:527–539.
6. Plackett RL. (1958) The principle of the arithmetic mean. *Biometrika* **45** 130–135.
7. Harris G. (2004) Merck says it will post the results of all drug trials. *New York Times*. September 6, 2004. C4.
8. Gigerenzer G, Swijtink Z, Porter T, Dasxton L, Beatty J, Kruger L. (1989) *The Empire of Chance* Cambridge. Cambridge University Press.
9. Moyé L. (2000). *Statistical Reasoning in Medicine. The Intuitive P-value Primer*. New York, Springer.
10. The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators. (1996) The HOPE (Heart Outcomes Prevention Evaluation) Study: The design of a large, simple randomized trial of an angiotensin-converting enzyme inhibitor (ramipril) and vitamin E in patients at high risk of cardiovascular events. *Canadian Journal of Cardiology* **12**:127–137.
11. The Heart Outcomes Prevention Evaluation Study Investigators. (2000) Effects of angiotensin-converting enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *New England Journal of Medicine* **342**:145–53.
12. Berger JO. (1980). *Statistical Decision Theory. Foundations, Concepts and Methods*. New York, Springer-Verlag.
13. Fisher RA. (1926) *Statistical Methods for Research Workers*. Edinburg. Oliver and Boyd.

- 
14. Fisher RA. (1933) The arrangement of field experiments. *Journal of the Ministry of Agriculture*. 503 - 513.
  15. Goodman SN. (1999). Towards evidence-based medical statistics. 1: The *p*-value fallacy. *Annals of Internal Medicine* **130**:995–1004.
  16. Gigerenzer G, Swijtink Z, Porter T, Dasxton L, Beatty J, Kruger L. (1989) *The Empire of Chance*. Cambridge, Cambridge University Press.
  17. Poole C. (1987) Beyond the confidence interval. *American Journal of Public Health* **77**:195–199.
  18. Walker AM. (1986) Significance tests [sic] represent consensus and standard practice (Letter). *American Journal of Public Health* **76**:1033. (See also journal erratum **76**:1087.
  19. Fleiss JL. (1986) Significance tests have a role in epidemiologic research; reactions to AM Walker. (different views) *American Journal of Public Health*. **76**:559–560.
  20. Fleiss JL. (1986) Dr. Fleiss response (Letter) *American Journal of Public Health* **76**:1033–1034.
  21. Walker AM. (1986) Reporting the results of epidemiologic studies. *American Journal of Public Health* **76**:556–558.
  22. Thompson WD. (1987) Statistical criteria in the interpretation of epidemiologic data (different views) *American Journal of Public Health* **77**: 191–194.
  23. Moyé L. (2004) *Finding Your Way in Science: How You Can Combine Character, Compassion, and Productivity in Your Research Career*. Vancouver. Trafford Press.



<http://www.springer.com/978-0-387-27781-3>

Statistical Monitoring of Clinical Trials  
Fundamentals for Investigators

Moyé, L.A.

2006, XXII, 254 p., Softcover

ISBN: 978-0-387-27781-3