Conclusions:  
Good Servants but Bad Masters

Drawing conclusions based on a sample of data is a thinking person’s business. Yet I am often asked for simple rules or guidelines for interpreting medical research. The best rule I can give is this: Use your head! While this deliberative thought process will produce its own share of mistakes, it is my thesis that we are worse off if we let \( p \)-values do our thinking for us.

Correct interpretation of a research effort is not as simple as reading from a rulebook. Every effort is unique, requiring us carefully to sift and weigh information about the prospective nature of the research design, its concordant execution, the state of healthcare in the community, and the appropriateness of the analysis plan before we can make decisions. Nevertheless, the principles of research design, execution, and analysis are simple.

This doctrine begins with the acknowledgment that, for ethical, fiscal, logistical and administrative reasons, researchers can only study a tiny fraction of the targeted population of subjects in which they are interested. However, they are extremely interested in extending their sample results from the relatively small collection of patients they have observed to thousands or millions that they have not. This extension is both necessary and hazardous. It is made more dangerous by the recognition that several competent researchers can each independently evaluate different samples from the same population, generating different “generalizable” results. The sample-to-sample variability weakens the claim that any particular sample is the one whose results are generalizable. Researchers who ignore the effect of sampling error while drawing conclusions from sample-based research are like pilots who ignore the effect of gravity while flying aircraft. The impact of these predictable but unaccounted for forces can be grievous.

The ability to extend a research effort’s results from a sample to a population depends on whether the research was designed to answer the question. \( P \)-values jointly interpreted with effect size are most helpful in determining the influence of sampling error on the confirmatory analyses drawn from research programs.

As important as \( p \)-values are for measuring sampling error, we are reminded that that is all that they do. \( P \)-values are not the sole arbiter for the study. They do not measure effect size, nor do they convey the extent of study discordance. A small \( p \)-value does not in and of itself mean that the sample size was adequate, that the effect size is clinically meaningful, or that there is clear attribution of effect to the exposure or intervention of interest.

Many research findings in healthcare are not generalizable. They occur simply by the random aggregation of subjects in small samples, commonly refereed to as the freak of chance. Commonly these “findings” are unanticipated. It is inap-
appropriate for the researcher to elevate these unanticipated exploratory findings, replacing the prospectively declared evaluations with these “surprise results.” The fisherman is suspect when he returns from a fishing trip not with fish but with boots and claims that he was “fishing for boots all along.” So too is the researcher who claims that he has identified an important answer from his sample when in fact he never intended to ask that research question of his sample. The sample supplied an answer that he never thought to ask.

The tendency among many researchers and their advocates to “analyze everything, and report what is favorable to our belief” is a process that mixes the relatively unstructured “search and exploration” aspect of science with the tight discipline required of research and confirmation. Combining the two confuses the view of the population landscape because it mixes in with the small number of good, prospectively planned evaluations many more false findings commonly generated by a small sample full of sampling error.

The medical community is commonly flummoxed by these reports. Some findings are reliable, most are not, and the investigator provides no methodologically sound discriminator, relying instead only on the \( p \)-value. The readership, confronted with this confusing situation, must work its way through this mixture of reliable an unreliable evaluations like soldiers traversing a mine field with a false map based on misleading \( p \)-values; they learn the hard way (as in ELITE I/II, and PRAISE I/II) which results are trustworthy and which are not.

This problem is avoided by reporting research results at two levels. The highest level is occupied by those questions asked by the investigator before any data were obtained. These questions are often small in number, well considered, and, more specifically, are the inquiries around which the research was designed. The maxim “first say what you plan to do, then do what you say” leads to the clearest extension of the research’s results from the small sample to the large population. It is in this well-defined circumstance that \( p \)-values are most useful, and their joint consideration with effect size, effect size variability and confidence intervals provide the clearest reflection of the population effects in the sample.

The second-tier questions are exploratory or hypothesis-generating questions. The definitive answers provided to this second group of questions must wait until the next research effort, because the research that spawned them was not designed to answer them. \( p \)-values are rarely helpful in these circumstances.

There are other destructive influences that distort research results as well. The wrong exposure duration, inferior inclusion/exclusion criteria, or poor selection of the endpoints can produce non-responsive research programs that shed no useful life on the scientific question at hand. Statistical hypothesis tests are of limited use in this setting; small \( p \)-values cannot save poorly designed studies.

In addition, \( p \)-values cannot rescue a research effort from its own poor execution. Concordantly executed research (i.e., a research effort that is executed according to the prospectively specified protocol) allows the cleanest and clearest interpretation of \( p \)-values, while discordant research (i.e., research efforts that undergo a midcourse change in analysis plan or endpoint) can be impossible to interpret. Because the findings are so difficult for us to integrate into our fund of knowledge, these discordant programs accomplish little more than squandering precious resources. Attempting to view a population effect through a discordant research
Conclusions

275

effort is like trying to learn about your appearance by studying your reflection in muddy water. It is an effort bound to mislead and must ultimately be set aside.

These factors of research design and execution must be considered separately from the statistical analysis of the data. *P*-values do not cover a host of methodologic sins. Bradford Hill (speaking of the chi-square test) described them as good servants, but bad masters. Reflexively responding to *p*-values to the exclusion of these other issues is shortsighted thought evasion.

The appropriate role of *p*-values is to measure the effect of sampling error on the research results. They are most useful when 1) they are the product of a well-designed, well executed research program and 2) are interpreted jointly with effect size, the variability of the effect size, and confidence intervals. In these circumstances, type I and type II error control are an important part of the research design process. Fortunately, their calibration is under the complete control of the investigators, and the acceptable levels of these errors must be chosen with care. I have provided several strategies that produce conservative, tightly controlled type I error levels and lead to clear interpretation of a concordantly executed research program with multiple endpoints. These strategies allow the investigator to choose a level of type I error symmetrically or asymmetrically. They are not the only available strategies, and of course no investigator should be compelled to use them. However, investigators should choose something. They should be plain about their choice, stick with their choice, and report the results of their choice. Readers of clinical research should continue to insist on this standard.

Physician investigators cannot understand alpha error concepts unless they understand the sampling error principles. Statisticians cannot provide the best advice on experimental design unless they understand the medical framework and underlying theory of the intervention or exposure. Statisticians and physician investigators must understand each other, investing the time and effort to communicate effectively, learning to appreciate the nuances of each other’s language. Liking each other is preferable – understanding each other is essential.

In addition, we physicians must control our inclination to explain any research result simply because we read it in a manuscript or heard it in a lecture. We must instead first ask if these results are due to sampling error. If they are, we need not go forward with interesting theories about the research findings; our time could be better spent elsewhere. Just as we would not squander our time on trying to explain why a 7 appeared as the result of throwing a pair of dice, so we should not exert great effort in attempting to explain results that are likely to be due to chance.

Since the written and visual media quickly convey new research findings for consumption to an increasingly health-conscious community, there is even less time and room for error. We should uphold community health by ensuring that we do not expose it to the corrosive influences of misleading results. Even with adequate precautions, sampling error still produces misleading conclusions from well-designed research efforts. By standing now for community protection through tight type I error control, we can help to shape the future rather than be stampeded in the wrong direction by results that are themselves driven by sampling error.

Finally, I believe there are three important principles of research. They are simple, and, by and large, nonmathematical.
1. Provide for the general welfare of your patients.
This should not be reduced to a truism. Patients volunteer their time, their energy, and quite frankly, their health by taking part in the studies that we researchers design and execute. Researchers are honor bound and obligated to ensure that the participants in their studies receive the best available treatment. These research subjects are a precious resource that must never be taken for granted or squandered.

I have pointed out several areas where the ethics and the mathematics of clinical experiments collide, but this is not one of them. Consider, a 4,000 patient clinical trial, designed to test the effect of therapy to reduce total mortality. If we assume the cumulative event rate for the placebo group is 20%, then we expect $(2,000)(0.20) = 400$ deaths in the placebo group. If we anticipate that the cumulative mortality rate in the treatment group is 0.16, we would expect $(2,000)(0.16) = 320$ patients to die during the course of the trial in the treatment group. Out of the four thousand patient trial, the entire measure of therapy effectiveness is reduced to the experiences of $400 - 320 = 80$ patients. The results on this large, multimillion dollar clinical trial hinge on the findings in only 80 participants (Figure 13.1).

Fig. 13.1. The results of a multimillion dollar study teter precariously on the findings in 80 patients.

Clearly, if the investigators knew who these 80 patients were at the trial’s inception, these patients would be singled out for very special care. For example, they would ensure that these participants understood the importance of the protocol. The investigators would see to it that these patients attended each follow-up visit. Any complaints these patients had would be addressed at once. Of course the investigators do not know the identities of these 80 patients. Therefore the investigator
must treat each patient in the study as if that patient is the patient that will make the difference in the trial.

2. **Promote collegial relationships with co-investigators**

Issues of publication policy, doses of medication, and the characteristics of the patients to be included in the study are critical, and strong-willed scientists can vehemently disagree on these fundamental trial characteristics. Investigators should expect this. However, they must actively work to ensure that the communication between investigators does not become choked with anger, resentment, or hostility. If these are permitted to occur, the research effort is weakened and its survival threatened.

Each investigator makes a unique contribution of personality, intellect and perspective and deserves full expression and consideration. Investigators must remember that their quiet answers to wrathful questions from their strongly opinionated colleagues can blunt this anger, turning it aside.

Research efforts experience external, centrifugal forces (e.g., opposing points of few in the medical community, politically charged scientific issues, contrary findings in other research programs) that threaten to tear the research effort apart. These destructive forces are counterbalanced by the centripetal force of investigators who are able to put their differences of opinion aside and hold their common effort together.

3. **Preserve, protect, and defend the protocol**

Prospective statements of the questions to be answered by the research effort and rejection of tempting data-based changes to the protocol preserve the best estimates of effect size, standard errors, and measures of sampling error. In a clinical trial, carrying out a well-designed protocol as it was initially and carefully planned is one of the best ways to protect vulnerable patient communities from exposure to harmful placebos. By following the enunciated principles of this book, investigators will be able to recognize the menace of type I error to both the patient and scientific communities, while simultaneously controlling and accurately reporting it. Adjustments to type I error are unavoidable. However, we must be ever vigilant, avoiding the corrupting influences of having the data determine the analysis plan.

Following these principles promotes the prosecution of a successful research program, i.e., the construction and protection of a research environment that permits an objective assessment of the therapy or exposure being studied. If there is any fixed star in the research constellation, it is that sample-based research must be hypothesis-driven and concordantly executed to have real meaning for both the scientific community and the patient populations that we serve.
Appendix A

Standard Normal Probabilities

The standard normal distribution is perhaps the most commonly used distribution in probability and statistics. Its ubiquity is due to its wide spread applicability and its ease of use, despite the fact that one must have access to a table in order to compute the probabilities associated with it.

Assume that an observation follows a normal distribution. We must specify its mean $\mu$ and its standard deviation $\sigma$ in order to identify its location and variability. In this case, we say that $x$ follows a $N(\mu, \sigma^2)$. This is enough for us to know that, for example, $P[\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma] = 0.95$. The value of this probability is the same regardless of what $\mu$ and $\sigma$ happen to be.

However, other probabilities will depend on the values of these parameters. For example, if $x$ is blood glucose level, and we wish to compute the probability that $x$ is less than 60 mg/dl, then the solution will depend on whether we think the blood sugar level follows a normal distribution with $\mu = 70$ and $\sigma = 20$ or if its parameters are $\mu = 220$ and $\sigma = 30$.

The implication of this observation is that we would need to have a normal distribution for each combination of $\mu$ and $\sigma$. Fortunately this is not the case. As pointed out in Chapter Three, each normally distributed random variable can be converted to a standard normal distribution. This standard normal distribution has $\mu = 0$ and $\sigma = 1$. This permits one table to be used to compute the values of events whose probabilities are governed by this distribution.

For example, in order to compute the probability that $x < 60$, we write

$$P[x < 60] = P \left[ \frac{(x - 70)}{20} < \frac{(60 - 70)}{20} \right] = P[z < -0.5].$$

We write the last probability in terms of $z$, where $z = \frac{(x - 70)}{20}$ follows a standard normal distribution. Using the table we find that $P[z < -0.5] = 0.480$.

In a similar computation of this probability, if $\mu = 220$ and $\sigma = 30$, then

$$P[x < 60] = P \left[ \frac{(x - 220)}{30} < \frac{(60 - 220)}{30} \right] = P[z < -5.33].$$

Using the same table from this appendix we see that the probability of this value is quite small (<0.001).

Using the symmetry of the normal distribution, we can compute $P[z > a] = P[z < -a]$ when $a$ is a positive number. Additionally we can also write $P[a < z < b] = P[z < b] - P[z < a]$. These relationships increase the utility of the following table.
Figure A. The standard normal distribution.
<table>
<thead>
<tr>
<th>$z$</th>
<th>$P(Z &lt; z)$</th>
<th>$z$</th>
<th>$P(Z &lt; z)$</th>
<th>$z$</th>
<th>$P(Z &lt; z)$</th>
<th>$z$</th>
<th>$P(Z &lt; z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.90</td>
<td>0.0000</td>
<td>-1.95</td>
<td>0.0256</td>
<td>0.00</td>
<td>0.5000</td>
<td>1.95</td>
<td>0.9744</td>
</tr>
<tr>
<td>-3.85</td>
<td>0.0001</td>
<td>-1.90</td>
<td>0.0287</td>
<td>0.05</td>
<td>0.5199</td>
<td>2.00</td>
<td>0.9772</td>
</tr>
<tr>
<td>-3.80</td>
<td>0.0001</td>
<td>-1.85</td>
<td>0.0322</td>
<td>0.10</td>
<td>0.5398</td>
<td>2.05</td>
<td>0.9798</td>
</tr>
<tr>
<td>-3.75</td>
<td>0.0001</td>
<td>-1.80</td>
<td>0.0359</td>
<td>0.15</td>
<td>0.5596</td>
<td>2.10</td>
<td>0.9821</td>
</tr>
<tr>
<td>-3.70</td>
<td>0.0001</td>
<td>-1.75</td>
<td>0.0401</td>
<td>0.20</td>
<td>0.5793</td>
<td>2.15</td>
<td>0.9842</td>
</tr>
<tr>
<td>-3.65</td>
<td>0.0001</td>
<td>-1.70</td>
<td>0.0446</td>
<td>0.25</td>
<td>0.5987</td>
<td>2.20</td>
<td>0.9861</td>
</tr>
<tr>
<td>-3.60</td>
<td>0.0002</td>
<td>-1.65</td>
<td>0.0495</td>
<td>0.30</td>
<td>0.6179</td>
<td>2.25</td>
<td>0.9878</td>
</tr>
<tr>
<td>-3.55</td>
<td>0.0002</td>
<td>-1.60</td>
<td>0.0548</td>
<td>0.35</td>
<td>0.6368</td>
<td>2.30</td>
<td>0.9893</td>
</tr>
<tr>
<td>-3.50</td>
<td>0.0002</td>
<td>-1.55</td>
<td>0.0606</td>
<td>0.40</td>
<td>0.6554</td>
<td>2.35</td>
<td>0.9906</td>
</tr>
<tr>
<td>-3.45</td>
<td>0.0003</td>
<td>-1.50</td>
<td>0.0668</td>
<td>0.45</td>
<td>0.6736</td>
<td>2.40</td>
<td>0.9918</td>
</tr>
<tr>
<td>-3.40</td>
<td>0.0003</td>
<td>-1.45</td>
<td>0.0735</td>
<td>0.50</td>
<td>0.6915</td>
<td>2.45</td>
<td>0.9929</td>
</tr>
<tr>
<td>-3.35</td>
<td>0.0004</td>
<td>-1.40</td>
<td>0.0808</td>
<td>0.55</td>
<td>0.7088</td>
<td>2.50</td>
<td>0.9938</td>
</tr>
<tr>
<td>-3.30</td>
<td>0.0005</td>
<td>-1.35</td>
<td>0.0885</td>
<td>0.60</td>
<td>0.7257</td>
<td>2.55</td>
<td>0.9946</td>
</tr>
<tr>
<td>-3.25</td>
<td>0.0006</td>
<td>-1.30</td>
<td>0.0968</td>
<td>0.65</td>
<td>0.7422</td>
<td>2.60</td>
<td>0.9953</td>
</tr>
<tr>
<td>-3.20</td>
<td>0.0007</td>
<td>-1.25</td>
<td>0.1056</td>
<td>0.70</td>
<td>0.7580</td>
<td>2.65</td>
<td>0.9960</td>
</tr>
<tr>
<td>-3.15</td>
<td>0.0008</td>
<td>-1.20</td>
<td>0.1151</td>
<td>0.75</td>
<td>0.7734</td>
<td>2.70</td>
<td>0.9965</td>
</tr>
<tr>
<td>-3.10</td>
<td>0.0010</td>
<td>-1.15</td>
<td>0.1251</td>
<td>0.80</td>
<td>0.7881</td>
<td>2.75</td>
<td>0.9970</td>
</tr>
<tr>
<td>-3.05</td>
<td>0.0011</td>
<td>-1.10</td>
<td>0.1357</td>
<td>0.85</td>
<td>0.8023</td>
<td>2.80</td>
<td>0.9974</td>
</tr>
<tr>
<td>-3.00</td>
<td>0.0013</td>
<td>-1.05</td>
<td>0.1469</td>
<td>0.90</td>
<td>0.8159</td>
<td>2.85</td>
<td>0.9978</td>
</tr>
<tr>
<td>-2.95</td>
<td>0.0016</td>
<td>-1.00</td>
<td>0.1587</td>
<td>0.95</td>
<td>0.8289</td>
<td>2.90</td>
<td>0.9981</td>
</tr>
<tr>
<td>-2.90</td>
<td>0.0019</td>
<td>-0.95</td>
<td>0.1711</td>
<td>1.00</td>
<td>0.8413</td>
<td>2.95</td>
<td>0.9984</td>
</tr>
<tr>
<td>-2.85</td>
<td>0.0022</td>
<td>-0.90</td>
<td>0.1841</td>
<td>1.05</td>
<td>0.8531</td>
<td>3.00</td>
<td>0.9987</td>
</tr>
<tr>
<td>-2.80</td>
<td>0.0026</td>
<td>-0.85</td>
<td>0.1977</td>
<td>1.10</td>
<td>0.8643</td>
<td>3.05</td>
<td>0.9989</td>
</tr>
<tr>
<td>-2.75</td>
<td>0.0030</td>
<td>-0.80</td>
<td>0.2119</td>
<td>1.15</td>
<td>0.8749</td>
<td>3.10</td>
<td>0.9990</td>
</tr>
<tr>
<td>-2.70</td>
<td>0.0035</td>
<td>-0.75</td>
<td>0.2266</td>
<td>1.20</td>
<td>0.8849</td>
<td>3.15</td>
<td>0.9992</td>
</tr>
<tr>
<td>-2.65</td>
<td>0.0040</td>
<td>-0.70</td>
<td>0.2420</td>
<td>1.25</td>
<td>0.8944</td>
<td>3.20</td>
<td>0.9993</td>
</tr>
<tr>
<td>-2.60</td>
<td>0.0047</td>
<td>-0.65</td>
<td>0.2578</td>
<td>1.30</td>
<td>0.9032</td>
<td>3.25</td>
<td>0.9994</td>
</tr>
<tr>
<td>-2.55</td>
<td>0.0054</td>
<td>-0.60</td>
<td>0.2743</td>
<td>1.35</td>
<td>0.9115</td>
<td>3.30</td>
<td>0.9995</td>
</tr>
<tr>
<td>-2.50</td>
<td>0.0062</td>
<td>-0.55</td>
<td>0.2912</td>
<td>1.40</td>
<td>0.9192</td>
<td>3.35</td>
<td>0.9996</td>
</tr>
<tr>
<td>-2.45</td>
<td>0.0071</td>
<td>-0.50</td>
<td>0.3085</td>
<td>1.45</td>
<td>0.9265</td>
<td>3.40</td>
<td>0.9997</td>
</tr>
<tr>
<td>-2.40</td>
<td>0.0082</td>
<td>-0.45</td>
<td>0.3264</td>
<td>1.50</td>
<td>0.9332</td>
<td>3.45</td>
<td>0.9997</td>
</tr>
<tr>
<td>-2.35</td>
<td>0.0094</td>
<td>-0.40</td>
<td>0.3446</td>
<td>1.55</td>
<td>0.9394</td>
<td>3.50</td>
<td>0.9998</td>
</tr>
<tr>
<td>-2.30</td>
<td>0.0107</td>
<td>-0.35</td>
<td>0.3632</td>
<td>1.60</td>
<td>0.9452</td>
<td>3.55</td>
<td>0.9998</td>
</tr>
<tr>
<td>-2.25</td>
<td>0.0122</td>
<td>-0.30</td>
<td>0.3821</td>
<td>1.65</td>
<td>0.9505</td>
<td>3.60</td>
<td>0.9998</td>
</tr>
<tr>
<td>-2.20</td>
<td>0.0139</td>
<td>-0.25</td>
<td>0.4013</td>
<td>1.70</td>
<td>0.9554</td>
<td>3.65</td>
<td>0.9999</td>
</tr>
<tr>
<td>-2.15</td>
<td>0.0158</td>
<td>-0.20</td>
<td>0.4207</td>
<td>1.75</td>
<td>0.9599</td>
<td>3.70</td>
<td>0.9999</td>
</tr>
<tr>
<td>-2.10</td>
<td>0.0179</td>
<td>-0.15</td>
<td>0.4404</td>
<td>1.80</td>
<td>0.9641</td>
<td>3.75</td>
<td>0.9999</td>
</tr>
<tr>
<td>-2.05</td>
<td>0.0202</td>
<td>-0.10</td>
<td>0.4602</td>
<td>1.85</td>
<td>0.9678</td>
<td>3.80</td>
<td>0.9999</td>
</tr>
<tr>
<td>-2.00</td>
<td>0.0228</td>
<td>-0.05</td>
<td>0.4801</td>
<td>1.90</td>
<td>0.9713</td>
<td>3.85</td>
<td>0.9999</td>
</tr>
</tbody>
</table>
Appendix B
Sample Size Primer

The purpose of this appendix is to provide a brief discussion of the underlying principles in sample size computations for a clinical trial. In the process, one of the simplest and most useful formulas for the sample size formulations will be reproduced. These basic formulas are the source of several of the calculations in Chapters Four through Eleven. First we will provide the solution and proceed to a discussion that both motivates and derives the sample size and power formulas.

B.1 General Discussion of Sample Size

Assume that a clinical trial has been designed to measure the effect of a randomly allocated intervention on a prospectively defined primary endpoint. Let $\theta_c$ be the cumulative incidence rate of the primary endpoint in the control group and let $\theta_t$ be the cumulative incidence rate of the primary endpoint in the treatment group. Then the statistical hypothesis for the primary endpoint in this clinical trial is

$$H_0 : \theta_c = \theta_t \quad \text{versus} \quad H_a : \theta_c \neq \theta_t. \quad (B.1)$$

Let $Z_a$ be the $a^{th}$ percentile from the standard normal distribution. The investigators have chosen an a priori test-specific type I error level $\alpha$, and the power of the statistical hypothesis test is $1 - \beta$. The hypothesis test will be two-sided. Let $p_c$ be the cumulative incidence rate of the primary endpoint in the control group of the research sample, and let $p_t$ be the cumulative incidence rate of the active group in the research sample. Then the trial size, or the sample size of the clinical trial, $N$ may be written as

$$N = \frac{2 \left[ p_c (1-p_c) + p_t (1-p_t) \right] \left[ Z_{1-a/2} - Z_\beta \right]^2}{(p_c - p_t)^2}. \quad (B.2)$$

Analogously, the power of the study may be calculated as a function of $N$,

$$1 - \beta = P \left[ N(0,1) > Z_{1-a/2} - \frac{p_c - p_t}{\sqrt{\frac{p_c (1-p_c)}{N/2} + \frac{p_t (1-p_t)}{N/2}}} \right]. \quad (B.3)$$

*This is the total number of patients in the study (number of patients in the placebo group plus the number of patients in the control group).
There are many different treatises on sample size calculations in clinical trials. A representative group is [1–5]. Several of these sources discuss important and useful nuances of the sample size computation that are useful in complex clinical trial design. The focus of the discussion here, however, will be on the most basic sample size computation, since that formula demonstrates most clearly the influence of the design parameters of the study (cumulative primary endpoint event rate in the control group, the anticipated effect of the intervention, the magnitude of the statistical errors, and test sidedness) on the resulting sample size.

For these discussions, assume that patients are randomized to receive either a new intervention or to receive control group therapy. In this example, there is one primary endpoint that occurs with a cumulative event rate \( \theta_c \). In the intervention group the cumulative event rate for the primary endpoint is \( \theta_t \). The investigator does not know the value of \( \theta_c \), since he does not study every patient in the population. He therefore selects a sample from the population and uses that sample to compute \( p_c \), which will serve as his estimate of \( \theta_c \). If the clinical trial has been executed concordantly, then \( p_c \) is a good estimator of \( \theta_c \); this means that the investigator can expect that \( p_c \) will be close to the value of \( \theta_c \). Analogously \( p_t \) is the estimate from the investigator’s sample of the cumulative incidence of the endpoint in the population \( \theta_t \).

Thus, if the trial was executed according to its protocol (and not subject to the destabilizing influences of random research), then \( p_c - p_t \) can be expected to be an accurate estimate of \( \theta_c - \theta_t \). If the null hypothesis is true, then \( \theta_c - \theta_t \) will be zero and we would expect \( p_c - p_t \) to be small. If the alternative hypothesis is correct, and the investigator’s intuition that the therapy being tested in the clinical trial will reduce the cumulative event rate of the primary endpoint is right, then \( \theta_c \) is much greater than \( \theta_t \), and \( p_c - p_t \), the best estimate of \( \theta_c - \theta_t \) will be large as well.

A key point in understanding the sample size formulation is the critical role played by the number of endpoint events produced by the sample. The research sample produces primary endpoints—the rate at which these endpoints are accumulated is directly linked to the cumulative event rate in the control group. This cumulative event rate therefore plays a central role in the sample size calculations. If the primary endpoint of a clinical trial is total mortality, then recruiting 1,000 patients into the study provides no useful information for the evaluation of the effect of therapy on total mortality if at the end of the study none of the 1,000 recruited patients die.

Therefore, the more primary endpoint events that occur during the course of the trial, the greater the volume of germane data available to answer the scientific question of whether the occurrence of those endpoint events are influenced by the intervention being studied. It follows that the larger the cumulative control group event rate is, the greater the number of primary endpoint event rates that will be generated. The greater the rate at which primary endpoints are produced, the smaller the required sample size for the clinical trial will be, assuming that everything else (effect of therapy, test sidedness, magnitude of the statistical errors) is equal, (or ceteris paribus).

A second measure that is critical in sample size considerations is the effectiveness of the therapy. This is often measured by the difference between the cumulative incidence rate of the primary endpoint in the population \( \theta_c \) and the cumulative
incidence of the primary event rate in the population if everyone in the population were to receive the treatment being studied in the clinical trial, $\theta_c$. This difference is commonly referred to as “delta” or $\Delta = \theta_c - \theta_t$. The greater the difference between $\theta_c$ and $\theta_t$, then the fewer the number of patients required to obtain a reliable estimate of that difference.

To understand this principle, it may be helpful to think of the two primary sources of variability involved in the estimation of the treatment effect in a clinical trial. The test statistic used to test the statistical hypothesis that $\theta_c = \theta_t$ versus the alternative hypothesis that these events are not equal is

$$\frac{p_c - p_t}{\sqrt{\text{Var}[p_c - p_t]}}.$$  \hspace{1cm} (B.4)

The first source of this variability is systematic; it is induced by the intervention being studied by the clinical trial and is an estimate of $\Delta$, the difference between the treatment and control group event rates that are seen in the sample. This variability is estimated by $p_c - p_t$ and resides in the numerator of (B.4). This is the “signal.”

The denominator of (B.4) is the second source of variability or the “noise”; it is an expression of the fact that, since the research is sample-based, estimates of $p_c - p_t$ will vary from sample to sample. Since this sampling variability “noise” should not be confused with the systematic, intervention-induced “signal” measured by $p_c - p_t$, this noise must be removed from the estimate of the therapy’s effect. Therefore using these characterizations, the greater the signal–to–noise ratio, the larger the expression in (B.4) will be.

The greater the signal–to–noise ratio as represented by (B.4), the easier it is to detect a genuine population effect of the intervention. If the magnitude of the difference $\theta_c - \theta_t$ is small in the population, then $p_c - p_t$ is also likely to be small. In this circumstance where the magnitude of the signal is small, the noise must be coincidentally reduced to detect the weak signal with precision. One useful tool the investigator has to reduce the background noise is to increase $N$, the sample size of the clinical trial. Part of the genius of choosing the reliable estimate $p_c - p_t$ of $\theta_c - \theta_t$ is that this estimate’s sampling variability decreases as the sample size increases.†

### B.2 Derivation of Sample Size

It is useful to consider the sample size computation as a three phased calculation. For each demonstration, there will be three phases of the computation.

Phase I – under the null
Phase II – under the alternative

* Sometimes it is useful to refer to the percent reduction in events attributable to the therapy, otherwise known as the therapy’s efficacy.
† This indispensable property of the estimates of effect size can be lost if the experiment is not executed concordantly (see Chapter Two).
Phase III – consolidation

We will step through each of these phase as we compute the sample size for the clinical trial as outlined earlier in this appendix.

**B.2.1 Phase 1: Under the Null Hypothesis**

Note that the test statistic

\[
\frac{p_c - p_t - (\theta_c - \theta_t)}{\sqrt{\text{Var}[p_c - p_t]}}
\]  

follows a normal distribution. Under the null hypothesis that \( \theta_c - \theta_t = 0 \) reduces to

\[
\frac{p_c - p_t}{\sqrt{\text{Var}[p_c - p_t]}}.
\]  

One useful way to think of this test statistic is as a normed effect size. Under the null hypothesis, we expect this normed effect size to have a mean of zero and a variance of one. It will follow the normal or bell shaped distribution. Then, the null hypothesis will be rejected when*

\[
\frac{p_c - p_t}{\sqrt{\text{Var}[p_c - p_t]}} > Z_{1-\alpha/2}
\]

or,

\[
p_c - p_t > Z_{1-\alpha/2} \sqrt{\text{Var}[p_c - p_t]}.
\]  

**B.2.2 Phase 2: Under the Alternative Hypothesis**

We now consider what should have if the alternative hypothesis was true. In this case, we start with the definition of statistical power.

\[ \text{Power} = \text{Probability [the null hypothesis is rejected | alternative hypothesis is true]} \]

* This is not the only circumstance under which the null hypothesis will be rejected. It will also be rejected when harm is caused by the intervention or when \( p_t - p_c \) is very much less than zero. However, in the sample size computation, attention is focused on the tail of the distribution in which the investigators are most interested.
The null hypothesis is rejected when the test statistic falls in the critical region or when \( p_c - p_t > Z_{\alpha/2} \sqrt{Var[p_c - p_t]} \). The alternative hypothesis is true if \( \theta_c - \theta_t = \Delta \geq 0 \). This allows us to write

\[
\text{Power} = 1 - \beta = P \left[ p_c - p_t > Z_{\alpha/2} \sqrt{Var[p_c - p_t]} \mid \theta_c - \theta_t = \Delta \right]. \tag{B.9}
\]

We now standardize the argument in the probability statement of (B.9) so that the quantity on the left follows a standard normal distribution. This requires subtracting the population mean effect under the alternative hypothesis (i.e., \( \Delta \)) and dividing by the square root of the variance of \( p_c - p_t \). These operations must be carried out on both sides of the inequality in the probability expression in (B.9) as follows.

\[
1 - \beta = P \left[ \frac{p_c - p_t - \Delta}{\sqrt{Var[p_c - p_t]}} > \frac{Z_{\alpha/2} \sqrt{Var[p_c - p_t]} - \Delta}{\sqrt{Var[p_c - p_t]}} \right]
\]

\[
= P \left[ \frac{p_c - p_t - \Delta}{\sqrt{Var[p_c - p_t]}} > Z_{\alpha/2} - \frac{\Delta}{\sqrt{Var[p_c - p_t]}} \right] \tag{B.10}
\]

\[
= P \left[ N(0,1) > Z_{\alpha/2} - \frac{\Delta}{\sqrt{Var[p_c - p_t]}} \right].
\]

By the definition of a percentile value from a probability distribution, we can now write

\[
Z_\beta = Z_{\alpha/2} - \frac{\Delta}{\sqrt{Var[p_c - p_t]}}. \tag{B.11}
\]

### B.2.3 Phase 3: Consolidation

We are now ready to conclude this computation, by solving for \( N \), the size of the trial. The sample size is embedded in the variance term in the denominator of expression (B.11).

\[
Var[p_c - p_t] = \frac{p_c (1 - p_c)}{n_c} + \frac{p_t (1 - p_t)}{n_t}. \tag{B.12}
\]

where \( n_c \) is the number of patients to be recruited to the control group in the clinical trial and \( n_t \) is the number of patients to be recruited to the active group. The sample size or trial size is the total number of patients required for the experi-
ment = \( N = n_c + n_t \). If we assume that the number of patients in the control group will equal the number of patients in the treatment group, then \( n_c = n_t = n \) and \( N = 2n \). Then (B.11) can be rewritten as

\[
Z_p = Z_{1-\alpha/2} - \frac{\Delta}{\sqrt{\frac{p_c(1-p_c) + p_t(1-p_t)}{n}}},
\]

We only need solve this equation for \( n \)

\[
n = \frac{\left[p_c(1-p_c) + p_t(1-p_t)\right] \left[Z_{1-\alpha/2} - Z_\beta \right]^2}{\Delta^2}.
\]

The trial size \( N = 2n \) may be written as

\[
N = \frac{2\left[p_c(1-p_c) + p_t(1-p_t)\right] \left[Z_{1-\alpha/2} - Z_\beta \right]^2}{\Delta^2}.
\]

To compute the power we only need to adapt the following equation from the last line of expression (B.10),

\[
1 - \beta = P \left[N(0,1) > Z_{1-\alpha/2} - \frac{\Delta}{\sqrt{\text{Var}[p_c - p_t]}}\right]
\]

and rewrite the \( \text{Var}[p_c - p_t] \) to find

\[
1 - \beta = P \left[N(0,1) > Z_{1-\alpha/2} - \frac{\Delta}{\sqrt{\frac{p_c(1-p_c) + p_t(1-p_t)}{N/2}}}ight].
\]

**B.3 Example**

If the experiment is designed for a two–sided \( \alpha \) of 0.05, 90 % power (\( \beta = 0.10 \)), \( p_c = 0.20 \), and \( \Delta = 0.03 \), then \( p_t = 0.17 \) (corresponding to a \((0.20 - 0.17)/0.20 = 0.15\), or a 15% reduction in events attributable to the intervention. The trial size can be computed from

\[
N = \frac{2\left[p_c(1-p_c) + p_t(1-p_t)\right] \left[Z_{1-\alpha/2} - Z_\beta \right]^2}{[p_c - p_t]^2}.
\]
B.4 Continuous Outcomes

Inserting the data from this example reveals
\[
N = \frac{2 \left[ (0.20)(0.80) + (0.17)(0.83) \right] \left[ 1.96 - (-1.28) \right]^2}{(0.20 - 0.17)^2} = 7024 \quad (B.19)
\]
or 3,512 subjects per group. If only 2,000 subjects per group can be identified, the power can be formulated from
\[
\text{Power} = P \left[ N(0,1) > Z_{\alpha/2} - \frac{\Delta}{\sqrt{\frac{p_c(1-p_c)}{N/2} + \frac{p_t(1-p_t)}{N/2}}} \right] \quad (B.20)
\]
and including the data from this example
\[
\text{Power} = P \left[ N(0,1) > 1.96 - \frac{0.03}{\sqrt{\frac{(0.20)(0.80)}{2000} + \frac{(0.17)(0.83)}{2000}}} \right] = 0.69. \quad (B.21)
\]

B.4 Continuous Outcomes

Many clinical trials have outcome measures that are continuous. Consider a clinical experiment that is designed to test the effect of an intervention on the change in left ventricular end diastolic volume (EDV). Patients are recruited using a random-sampling plan and have their baseline EDV measured. They are then randomized to receive placebo care or the intervention, and followed for three months, at the end of which they have their EDV measured again. The investigator assumes that the EDVs will be normally distributed, and wishes to analyze the change in EDV over time across the two groups. He believes that there will be a large increase in EDV in the placebo group, reflecting the natural progression of the disease. It is his hope that the EDV change will be smaller in the treatment arm of the experiment.

Let \( \mu_d(c) \) be the population mean change in the end diastolic volumes for the placebo group and \( \mu_d(t) \) be the population mean change in the end diastolic volume in the active group. Let’s begin with the null hypothesis,
\[
H_0 : \mu_d(c) = \mu_d(t) \quad \text{versus} \quad H_a : \mu_d(c) \neq \mu_d(t).
\]
Clearly, the investigator does not believe the alternative hypothesis as stated, he believes that $\mu_d(c)$ the population mean change in EDV in the placebo group, will be greater than $\mu_d(t)$, the population mean change in EDV in the active group. However, since he recognizes that he does not know the effect of therapy, he states the alternative hypothesis as two–sided.* However, his true belief in the ability of the treatment to affect the change in EDV will be reflected in phase II.

### B.4.1 Phase I: The Null Hypothesis

The purpose of phase I is simply to construct the test statistic and identify its critical region. The distribution of the test statistic is the distribution under the null hypothesis, i.e., under the assumption that there is no treatment effect on the mean change in EDV. As was stated before, the investigator believes the difference in EDVs will follow a normal distribution. Let $d_c$ be the sample mean change in the placebo group, and $d_t$ is the sample mean change in the active group, and $\text{Var}[d_c - d_t]$ be the variance of the difference in change of the EDVs. We note that under the null hypothesis the quantity

$$\frac{d_c - d_t}{\sqrt{\text{Var}[d_c - d_t]}}$$

follows a normal distribution. Then the null hypothesis will be rejected when

$$\frac{d_c - d_t}{\sqrt{\text{Var}[d_c - d_t]}} > Z_{1-\alpha/2},$$

where $Z_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile value from the standard normal distribution with mean zero and variance one. We may rewrite equation (B.23) to see that we will reject the null hypothesis in favor of the alternative if

$$d_c - d_t > Z_{1-\alpha/2} \sqrt{\text{Var}[d_c - d_t]}.$$

This ends phase I.

### B.4.2 Phase II: The Alternative Hypothesis

This next phase incorporates the result of phase I with the notion of power. Begin with the definition of power:

$$\text{Power} = \text{Prob}[\text{The null hypothesis is rejected | the alternative hypothesis is true}]$$

* This notion of test sidedness is discussed in Chapter Five.
The null hypothesis is rejected when the test statistic falls in the critical region. The alternative hypothesis is true if \( d_c - d_i = \Delta \geq 0 \). This quantity \( \Delta \) is the difference that the investigator hopes to see between the changes in the two groups. This consideration is not two-sided at this point, and is the opportunity for the investigator to state precisely state the magnitude of efficacy he believes this treatment will produce.

Using the result of Phase I we can write the power equation as

\[
\text{Power} = P \left[ \frac{d_c - d_i - \Delta}{\sqrt{\text{Var}[d_c - d_i]}} > Z_{1-\alpha/2} \frac{\sqrt{\text{Var}[d_c - d_i]}}{\sqrt{\text{Var}[d_c - d_i]}} - \frac{\Delta}{\sqrt{\text{Var}[d_c - d_i]}} \right].
\]

We now standardize this so that the quantity on the left follows a standard normal distribution. Under phase II, the alternative hypothesis the mean of the treatment difference \( d_c - d_i = \Delta \geq 0 \). This leads to

\[
\text{Power} = P \left[ \frac{d_c - d_i - \Delta}{\sqrt{\text{Var}[d_c - d_i]}} > Z_{1-\alpha/2} \frac{\Delta}{\sqrt{\text{Var}[d_c - d_i]}} \right],
\]

which can be simplified to

\[
= P \left[ N(0,1) > Z_{1-\alpha/2} - \frac{\Delta}{\sqrt{\text{Var}[d_c - d_i]}} \right],
\]

These steps are simply algebra. At this point, we can use the fact that \( P \left[ N(0,1) \geq Z_{\beta} \right] = 1 - \beta \) to write

\[
Z_{\beta} = Z_{1-\alpha/2} - \frac{\Delta}{\sqrt{\text{Var}[d_c - d_i]}}.
\]

This concludes Phase II

### B.4.3 Phase III: Consolidation

Phase II concluded with an equation, that we must now solve for \( n \). We assume that there were be an equal number of subjects in the treatment group and the intervention group. The sample size \( n \) is embedded in the variance term in the denominator of equation (B.28).

\[
\text{Var}[d_c - d_i] = \frac{\sigma_D^2}{n} + \frac{\sigma_D^2}{n} = \frac{2\sigma_D^2}{n},
\]

(B.29)
where \( \sigma_{D}^{2} \) is the variance of an intrasubject difference. The trial size (i.e. the total number of subjects needed for the experiment) = \( N = 2n \). Replacing the denominator of the expression on the right-hand side of equation (B.28) with the right hand size of (B.29), we have

\[
Z_{\beta} = Z_{1-a/2} - \frac{\Delta}{\sqrt{\text{Var}[d - d_i]}}.
\]

(B.30)

We need only solve this equation for \( n \)

\[
n = \frac{2\sigma_{D}^{2} \left[ Z_{1-a/2} - Z_{\beta} \right]^{2}}{\Delta^{2}},
\]

(B.31)

and the trial size \( N \), is

\[
N = \frac{4\sigma_{D}^{2} \left[ Z_{1-a/2} - Z_{\beta} \right]^{2}}{\Delta^{2}}.
\]

(B.32)

To compute the power one need only adapt the following equation from Phase II,

\[
1 - \beta = P \left[ N(0,1) > Z_{1-a/2} - \frac{\Delta}{\sqrt{\text{Var}[d_s - d_i]}} \right],
\]

(B.33)

and rewrite the variance to find

\[
1 - \beta = P \left[ N(0,1) > Z_{1-a/2} - \frac{\Delta}{\sqrt{2\sigma_{D}^{2}/n}} \right]
\]

(B.34)

**B.4.4 Example**

If, for this experiment, the investigator chooses a two-sided alpha of 0.05, 90% power (beta = 0.10), delta = 10 and \( \sigma_{D} = 18 \), the trial size is

\* The trial size is the total number of patients required for the experiment; here it is the number of patients randomized to the placebo group plus the number of patients randomized to the intervention group.
\[ N = \frac{4\sigma_D^2 [Z_{1-\alpha/2} - Z_{\beta}]^2}{[\Delta]^2} = \frac{4(18)^2 [1.96 - (-1.28)]^2}{10^2} = 136. \quad (B.35) \]

Or 68 subjects per group. If the delta of interest is 5 rather than 10, the power is

\[
1 - \beta = P \left[ N(0,1) > Z_{1-\alpha/2} - \frac{\Delta}{\sqrt{\frac{2\sigma^2}{n}}} \right] = P \left[ N(0,1) > 1.96 - \frac{5}{\sqrt{\frac{2(18)^2}{68}}} \right] = P[N(0,1) > 0.34] = 0.37. \quad (B.36)
\]

**References**

Daubert hearings are discussions that place before the court in cases involving the determination of causation between an exposure and a disease. To assist the court, the federal court system provided the following guidance.

**C.1 The Daubert Factors**

The expert's general and specific causation methods should be scrutinized pursuant to amended FRE Rule 702 and the *Daubert* factors. Amended Rule 702 provides additional areas of inquiry and each factor applicable to the testimony should be included in the motion. The *Daubert* factors are —

- **Testing:** Has the theory or methodology been tested or can it be?
- **Rate of Error:** What is the known or potential rate of error in an expert's methodology? High rates of error detract from the reliability of the methodology and conclusions. See *United States v. Dorsey*, 45 F.3d 809, 815 (4th Cir.), cert. denied, 515 US 1168 (1995).
- **Peer Review:** Has the theory or technique been published in a peer reviewed journal? Such publication is a relevant consideration in accessing the scientific validity of a particular technique or methodology on which an opinion in premised.
- **General Acceptance:** Has the theory or methodology gained "general acceptance" in the relevant scientific community?

**C.2 The 702 Factors**

1. Are the experts proposing to testify about matters growing naturally and directly out of research they have conducted?
2. Has the expert unjustifiably extrapolated from an accepted premise to an unfounded conclusion?
3. Has the expert adequately accounted for obvious alternative explanations (other causes)?
4. Is the expert being as careful as he/she would be in his regular professional work outside of his paid litigation consulting?

Federal Rule of Evidence 702 provides some general standards the trial court must use to access the reliability and helpfulness of proffered expert testimony:

If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact
in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

Failure of plaintiff's expert to meet any one of the above factors is fatal to the plaintiff's expert, and hence the case. Plaintiffs often argue that the amendments to Rule 702 mean that the courts should be more flexible and not rely on Daubert exclusively. See Michael H. Graham, The Expert Witness Predicament: Determining "Reliable" Under the Gatekeeping Test of Daubert, Kumho and Proposed Amended Rule 702 of the Federal Rules of Evidence, 54 U. Miami L. Rev. 317 (2002).
Index

A
aampling error
regression model building, 240
absolute risk, 151
adjusted analysis. See regression analysis
Adjusting for covariates, 237
Alpha allocation, 247
alternative hypothesis, 290
amlodipine
and clinical trials, 206
analysis plans
traige, 59
Arrowsmith, 28
attribution of effect, 93
Avicenna, 1

B
backward study, 105
Bayes approach
absolute update strength, 258
alpha allocation, 264
conditional distribution, 255
hypothesis testing, 267
posterior distribution, 255
relative update strength, 258
significance testing and posterior
distribution assessment, 266
bell shaped curve
description. See normal distribution
Berkson, Joseph, 125
biostatistics
and sampling error, 120
Bonferroni inequality
table of computations, 187

C
Cardiac Arrhythmia Suppression Trial,
170
CARE study, 215, 216

and gender effects, 222
carvedilol, 51
case reports
definition, 100
examples, 101
causality, 119
and case reports, 101
causation without p-values, 100
definition of causation, 92
direction of the relationship, 92
Hill tenets, 96
versus association, 92
ceteris paribus fallacy, 112
Christopher Columbus, 48
clinical significance. See statistical
significance
clinical trial
and secondary endpoints, 192
clinical trials
expense, 109
Cohen, Stanley
nobel prize, 22
co-investigators
and collegial relationships, 277
combined endpoints
and epidemiology, 195
and sample size, 195
ascertainment, 200
Lipid Research Clinics, 200
therapy homogeneity, 198
concordant research
definition, 57
vs. discordant research, 57
conditional distribution. See Bayes
approach
confidence intervals
as substitutes for p-values, 130
confirmatory analysis
definition, 45
difference from exploratory, 45
critical region
definition, 74
cross sectional, 105
definition, 105
necessity of, 48
tools of, 50
versus confirmatory, 44
exposure disease relationships, 119
familywise error rate
community protection issue, 188
FDA advisory committees, 51
Federal Food and Drug Administration (F. D. A.), 98, 110
Fisher R.A., 13
frequentist, 250
long-term accuracy, 250
Significance testing, 250
Frye test, 158
general acceptance, 158
Galton, 14
gambling. See probability
generalization obstacles, 38
generalizations and random analyses, 38
Gosset Student pseudonym, 13
Graunt, John. See demography
Hill, Bradford, 232
and causality tenets, 127
hypothesis generating analyses definition. See exploratory analysis
hypothesis testing and community responsibility, 90
definition and examples, 64
definition and examples, 64
definition, 70
in intervention studies, 89
manipulation, 118
null hypothesis testing, 70
one-sided testing, 75
one-sided versus two-sided testing, 86
preconceived ideas, 64
problems with one-sided testing, 83
science versus statistics, 64
two-sided testing, 83, 84
hypothesis testings
motivation for multiple analyses, 182

I
industrial revolution
rise of intellectualism, 6
intention to treat
vs. "as treated", 218
interaction analysis. See regression analysis

J
Johnson, James, 224

L
Lentz W., 114
let the data decide. See exploratory analysis
likelihood principle, 251
Lipid Research Clinics
misinterpretation of results, 173
Lipid Research Clinics sample size, 145

M
Madam Curie, 48
manuscripts
design manuscript, 56
Merrell Dow, 160, 161, 164
middle ages, 1
absolute kings, 3
poverty, 2
religious wars, 3
taxation, 4
urban versus rural life, 2
Mina Mata disease, 101
model building. See regression analysis
MRFIT, 211, See exploratory analysis
multiple analyses
definition and examples, 182

N
Napolean Bonaparte, 91
Neyman, Jerzy, 125
normal distribution
definition, 65
description, 65
formula, 65

O
observational studies
and directionality, 104
definition, 94
historical cohorts, 107
versus experiments, 92
odds ratio, 96

P
past occurrence of disease, 106
patient perspective
physicians and difficulties with research principles, 18
patient welfare, 276
Pearson, Egon, 125
Pearson, Karl, 125
Phase 1. See sample size computations
Phase 2. See sample size computations
Phase 3. See sample size computations
physicians
and the patient perspective, 18
clinical vs. research skills, 21
integrating perspectives, 20
theory vs. evidence, 20
political arithmetic, 8
and religion, 9
population perspective
vs. patient perspective, 19
Posicor (mefibridil), 111
posterior distribution. See Bayes approach
power, 141
definition, 77
PRIASE I and II, 26
probability
Abraham de Moivre, 10
eyear denial of its need, 5
rise of gambling, 5
prospective, 105, 106
distinction from retrospective, 105
prospective designs, 106
flexibility, 106
prospective plans, 50
protocols, 55
anchored vs. random, 40
protopathic bias
and temporality, 97
p-values
and American Journal of Public Health, 128
and causation, 100
and medical community confusion, 274
and small effect sizes, 149
and two-sided testing, 83
appropriate role in research, 131
criticism of the 0.05 level, 118
definition, 75
dimensionality reduction, 132
incorrect use, 273
joint interpretation, 108
mistaken identity, 91
misuse, 127
nominal testing, 185
reporting effect size and variability, 111
resampling procedures, 190
strength of evidence, 137
sum of their parts, 91
summary statement of appropriate use, 275
versus alpha level, 75

effect modification, 241
effect modification and complications, 244
error terms, 231
explainer variables, 229
goal, 229
model building, 239
underlying assumptions, 230
variance partitioning, 233
Regression models, 237
relative risk, 151
research efforts
the great compromise, 23
retrospective, 105, 106
study directionality, 104
risk factor – disease relationship, 106
Rule 402, 158
relevance, 159

S

sample size
and Lipid Research Clinics, 145
and primary endpoints, 284
and test sidedness, 174
balancing type I and type II errors, 82
calculations for continuous outcomes, 289
computations for dichotamous outcomes, 286
continuous outcome example, 292
dichotamous outcome example, 288
general discussion, 144
introduction and general discussion, 283
mimicking the sample size of others, 147
sources of variability, 285
three phase approach, 285
samples
absence of certainty, 24
clear view of population, 33
definition of random sample, 29
generalization to population, 29
limitations of random samples, 30
vs. population, 24
sampling error
and PRAISE I-II, 28
and variability, 26
definition, 24
definition, 24
examples, 24

R

radium girls, 99
random allocation of therapy, 93
random research. See generalizations
randomization
stratified, 216
use of, 92
randomized experiments
definition, 92
regression analysis
adjustments, 237
automatic versus manual model building, 241
building in the coherency argument, 242
building in the consistency argument, 242
cross sectional versus longitudinal analysis, 233

Index

regression model building, 241
Seldane (terfenadine), 110
signal of effect and research designs, 110
Society of the Living Dead, 99
statistical estimators
  perturbed, 38
statistical reasoning
  critical question, 22
  definition, 18
statistical significance
  early use, 122
  first use of the 0.05 level, 122
  pre-Fisher examples, 123
  Ronald Fisher and controversy, 122
  versus clinical significance, 153
statistics
  early dispute over inference, 11
Student’s t distribution, 14
subgroups
  confirmatory analyses, 220
  definition, 208
  difficulties with block interpretation, 224
  effect domination principle, 219
  interaction and effect modification, 221
  interpretation difficulties, 208
  random effects, 209
  within stratum effects, 222
successful research program
definition, 277
Supreme Court, 159, 161, 162, 165

t
  test sidedness
    and sample size, 174
    attraction of one-sided test, 167
    belief and research design, 168
    belief versus knowledge, 167
    non-symmetrical testing, 176
    uniformly most powerful test, 169
  thalidomide, 98, 101, 113
  Timing (retrospective vs. prospective), 105
  type I error
    balancing type I and type II errors, 80
    definition, 74
  type II error
    definition, 77

V

variable adjustment in observational Studies, 107
variance partitioning. See regression analysis
vesnarinone, 53