2

Optimal Procedures

2.1 Defining Optimal

As we saw in the preceding chapter, the professional statistician is responsible for choosing both the test statistic and the testing procedure. An amateur might hope to look up the answers in a book, or, as is all too commonly done, use the same statistical procedure as was used the time before, regardless of whether it continues to be applicable. But the professional is responsible for choosing the best procedure, the optimal statistic. The statistic we selected in the preceding chapter for testing the effectiveness of vitamin E seemed an obvious, intuitive choice. But is it the best choice? And can we prove it is? Intuition can so often be deceptive.

In this chapter, we examine the criteria that define an optimal testing procedure and explore the interrelationships among them.

2.1.1 Trustworthy

The most obvious desirable property of a statistical procedure is that it be trustworthy. If we are advised to make a particular decision, then we should be correct in doing so. Alas, our observations are stochastic in nature, so there may be more than one explanation for any given set of observations. The result is we never can rely 100% on the decisions we make. At best, they can be like politicians, trustworthy up to a point. We ask only that they confine themselves to small bribes and rake-offs, that they not bankrupt or betray the country.

In the example of the missing labels in the preceding chapter, we introduced a statistical test based on the random assignment of labels to treatments. Knowing in advance that the experiment could have any of $\binom{6}{3} = 20$ possible outcomes, we will reject the null hypothesis only if the obtained value of the test statistic is the maximum possible that could arise from only one permutation of the results. The test we derive is valid under very broad

assumptions. The data could have been drawn from a normal distribution or they could have come from some quite different distribution. To be valid at a given percent level, all that is required of our permutation test is that (under the hypothesis) the population from which the data in the treatment group are drawn be the same as that from which the untreated sample is taken.

This freedom from reliance on numerous assumptions is a big plus. The fewer the assumptions, the fewer the limitations, and the broader the potential applications of a test. But before statisticians introduce a test into their practice, they need to know a few more things about it, namely:

- Is it *exact*? That is, can we make an exact determination of the probability that we might make an error in rejecting a true hypothesis?
- How *powerful* a test is it? That is, how likely is it to pick up actual differences between treated and untreated populations? Is this test as powerful or more powerful than the test we are using currently?
- Is the test *admissible*? That is, is there no other test that is superior to it under all circumstances?
- How *robust* is the new test? That is, how sensitive is it to violations in the underlying assumptions and the conditions of the experiment?

2.1.2 Two Types of Error

It's fairly easy to reason from cause to effect—that is, if you have a powerful enough computer. Get the right formula (Boyle's Law, say), plug in enough values to enough decimal places, and out pops the answer. The difficulty with reasoning in the opposite direction, from effect to cause, is that more than one set of causes can be responsible for precisely the same set of effects. We can never be completely sure which set of causes is responsible. Consider the relationship between sex (cause) and height (effect). Boys are taller than girls. True? So that makes this new $6'2''$ person in our lives ... a starter on the women's volleyball team.

In real life, in real populations, there are vast differences from person to person. Some women are tall and some women are short. In Lake Wobegone, Minnesota, all the men are good looking and all the children are brighter than average. But in most other places in the world there is a wide range of talent and abilities. As a further example of this variation, consider that half an aspirin will usually take care of one person's headache while other people take two or three aspirin at a time and get only minimal relief.

Figure 2.1 below depicts the results of an experiment in which two groups were each given a "pain-killer." The first group got buffered aspirin; the second group received a new experimental drug. Each of the participants then provided a subjective rating of the effects of the drug. The ratings ranged from "got worse," to "much improved," depicted below on a scale of 0 to 4. Take a close look at Figure 2.1. Does the new drug represent an improvement over aspirin?

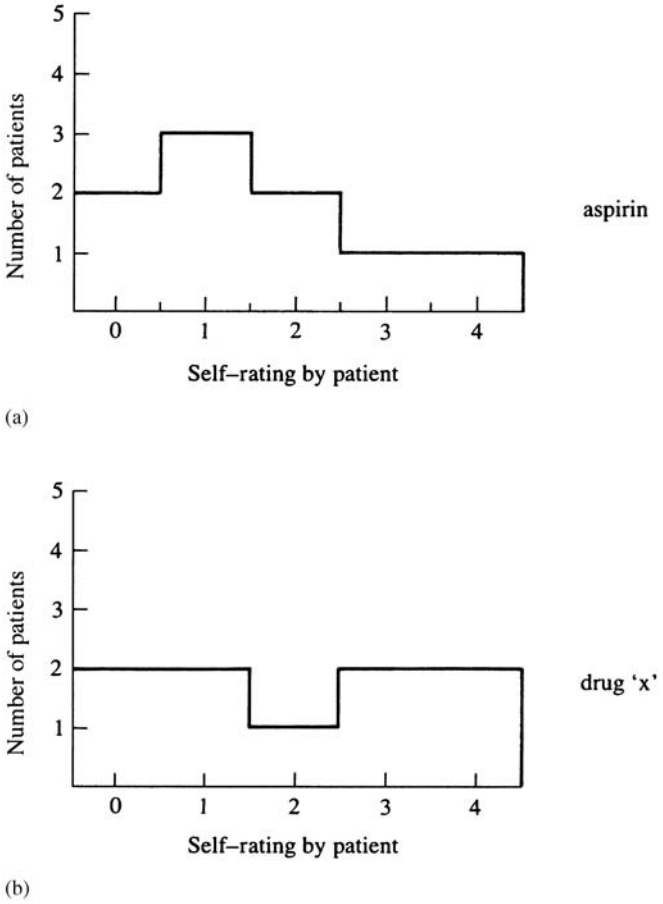

Fig. 2.1. Response to treatment: Self-rating patient in (a) asprin-treated group, (b) drug-'x'-treated group.

Those who took the new experimental drug do seem to have done better on average than those who took aspirin. Or are the differences we observe in Figure 2.1 simply the result of chance? If it's just a *chance* effect—rather than one caused by the new drug—and we opt in favor of the new drug, we've made an error. We also make an error if we decide there is no difference, when, in fact, the new drug really is better. These decisions and the effects of making them are summarized in Table 2.1a below.

We distinguish between the two types of error because they have quite different implications. For example, Fears, Tarone, and Chu [1977] use permutation methods to assess several standard screens for carcinogenicity. Their Type I error, a false positive, consists of labeling a relatively innocuous compound as carcinogenic. Such an action means economic loss for the

Table 2.1a. Decision making under uncertainty.

The Facts	Our Decision	
No difference	No difference	New drug is better Type I error
New drug is better	Type II error	

Table 2.1b. Decision making under uncertainty.

The Facts	Fears et al's Decision	
	Not a carcinogen	Compound a carcinogen
Not a carcinogen (Alternative)		Type I error: manufacturer misses opportunity for profit; public denied access to effective treatment
Carcinogen (Hypothesis)	Type II error: patients die; families suffer; manufacturer sued	

manufacturer and the denial of the compound's benefits to the public. Neither consequence is desirable. But a false negative, a Type II error, would mean exposing a large number of people to a potentially lethal compound.

Because variation is inherent in nature, we are bound to make the occasional error when we draw inferences from experiments and surveys, particularly if, for example, chance hands us a completely unrepresentative sample. When I toss a coin in the air six times, I can get three heads and three tails, but I also can get six heads. This latter event is less probable, but it is not impossible. Variation also affects the answer to the question, "Does the best team always win?"

We can't eliminate risk in making decisions, but we can contain risk through the correct choice of statistical procedure. For example, we can require that the probability of making a Type I error not exceed 5% (or 1% or 10%) and restrict our choice to statistical methods that ensure we do not exceed this level. If we have a choice of several statistical procedures, all of which restrict the Type I error appropriately, we can choose the method that leads to the smallest probability of making a Type II error.

2.1.3 Losses and Risk

The preceding discussion is greatly oversimplified. Obviously, our losses will depend not merely on whether we guess right or wrong, but on how far our

guesstimate is off the mark. For example, suppose you've developed a new drug to relieve anxiety and are investigating its side effects. You ask, "Does it raise blood pressure?" You do a study and find the answer is "no." But the truth is your drug raises systolic blood pressure an average of one millibar. What is the cost to the average patient? Negligible, one millibar is a mere fraction of the day-to-day variation in blood pressure.

Now, suppose your new drug actually raises blood pressure an average of 10 mb. What is the cost to the average patient? to the entire potential patient population? to your company in law suits? Clearly, the cost of a Type II error will depend on the magnitude of that error and the nature of the losses associated with it.

Historically, much of the work in testing hypotheses has been limited to zero or one loss function while that of estimation has focused on losses proportional to the square of the error. The result may have been statistics that were suboptimal in nature with respect to the true, underlying loss (see Mielke [1986], Mielke and Berry [1997]).

Are we more concerned with the losses associated with a specific decision or those we will sustain over time as a result of adhering to a specific decision procedure? Which concerns our company the most: reducing average losses over time or avoiding even the remote possibility of a single, catastrophic loss? We return to this topic in Section 2.2.

2.1.4 Significance Level and Power

In selecting a statistical method, statisticians work with two closely related concepts, significance level and power. The *significance level* of a test, denoted throughout the text by the Greek letter α (alpha), is the probability of making a Type I error; that is, α is the probability of deciding erroneously on the alternative when, in fact, the hypothesis is true.

To test a hypothesis, we divide the set of possible outcomes into two or more regions. We accept the primary hypothesis and risk a Type I error when our test statistic lies in the *rejection region* R; we reject the primary hypothesis and risk a Type II error when our test statistic lies in the *acceptance region* A; and we may take additional observations when our test statistic lies in the boundary *region of indifference* I. If H denotes the hypothesis, then

$$\alpha = \Pr\{X \in R|H\}.$$

The power of a test, denoted throughout the text by the Greek letter β (beta), is the complement of the probability of making a Type II error; that is, β is the probability of deciding on the alternative when the alternative is the correct choice. If K denotes the alternative, then

$$\beta = \Pr\{X \in R|K\}.$$

The ideal statistical test would have a significance level α of zero and a power β of 1, or 100%. But unless we are all-knowing, this ideal cannot be realized. In practice, we will fix a significance level $\alpha > 0$, where α is the largest value we feel comfortable with, and choose a statistic that maximizes or comes closest to maximizing the power for an alternative or set of alternatives important to us.

2.1.4.1 Power and the Magnitude of the Effect

The relationship among power, significance level, and the magnitude of the effect for a specific test is summarized in Figure 2.2, provided by Patrick Onghena. For a fixed significance level, the power is an increasing function of the magnitude of the effect. For a fixed effect, increasing the significance level also increases the power.

2.1.4.2 Power and Sample Size

As noted in Section 2.1.3., the greater the discrepancy between the true alternative and our hypothesis, the greater the loss associated with a Type II error.

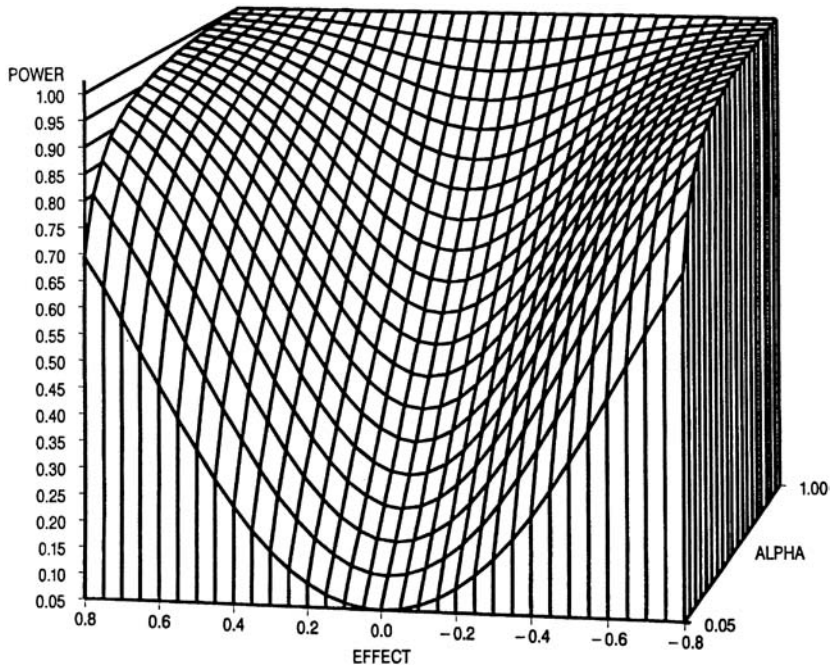

Fig. 2.2. Power of the two-tailed t-test with sample sizes of $n_1 = n_2 = 20$ as a function of the effect size (EFFECT) and the significance level (ALPHA) under the classical parametric assumptions.

Fortunately, in most practical situations, we can devise a test where the larger the discrepancy, the greater the power and the less likely we are to make a Type II error.

The relationship among power, effect magnitude, and number of observations for a specific test is summarized in Figure 2.3, provided by Patrick Onghena.

Figure 2.4a depicts the power as a function of the alternative for two tests based on samples of size 6. In the example illustrated, the test φ_1 is uniformly more powerful than φ_2, hence, using φ_1 in preference to φ_2 will expose us to less risk.

Figure 2.4b depicts the power curve that results from using these same two tests, but for different size samples; the power curve of φ_1 is still based on a sample of size 6, but that of φ_1 now is based on a sample of size 12. The two new power curves almost coincide, revealing the two tests now have equal risks. But we will have to pay for twice as many observations if we use the second test in place of the first.

Moral: A more powerful test reduces the costs of experimentation, and it minimizes the risk.

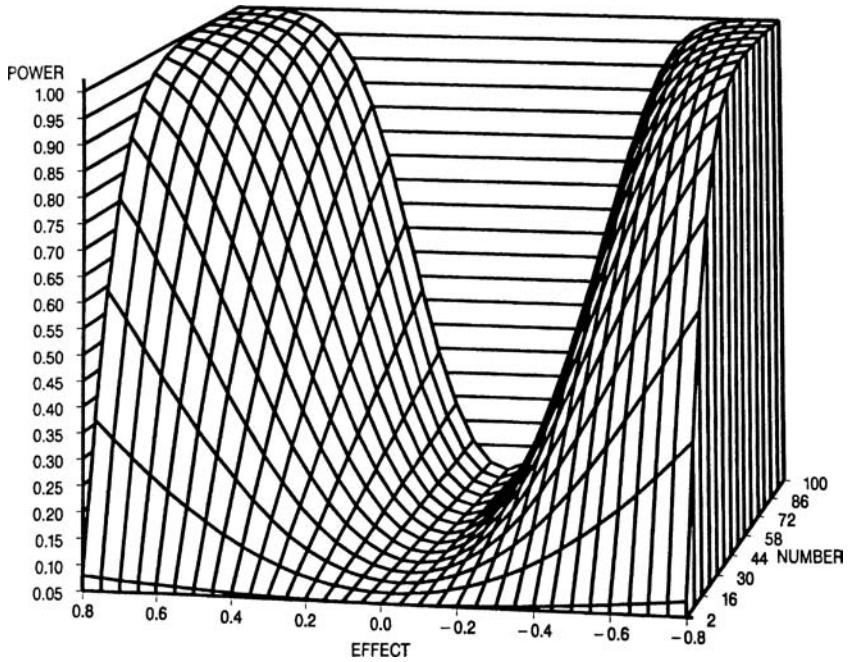

Fig. 2.3. Power of the two-tailed t-test with $p = 0.05$ as a function of the effect size (EFFECT) and the number of observations (NUMBER, $n_1 = n_2$) under the classical parametric assumptions.

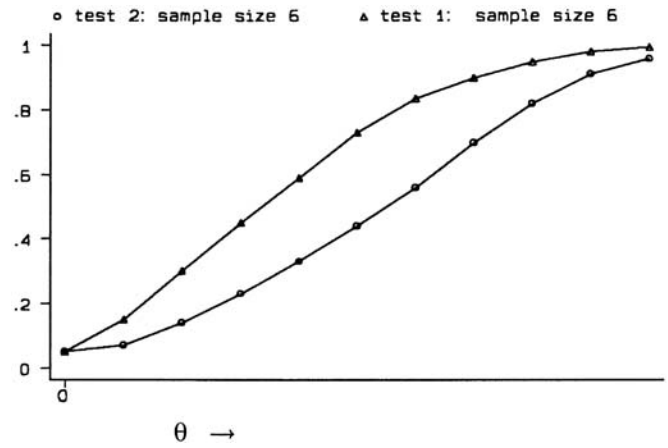

Fig. 2.4a. Power as a function of the alternative. Tests have the same sample size.

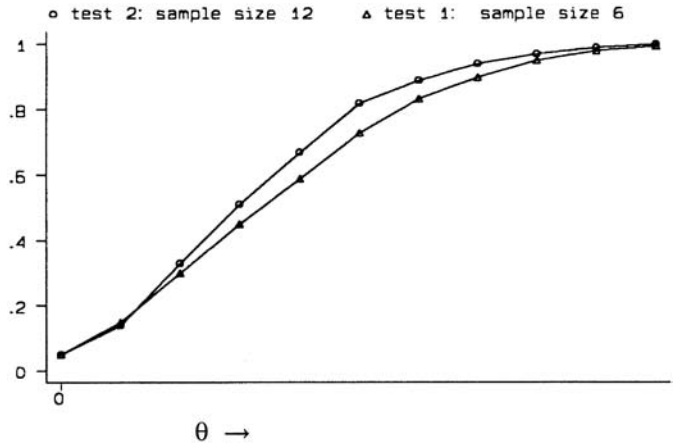

Fig. 2.4b. Power as a function of the alternative. Tests have different sample sizes.

2.1.4.3 Power and the Alternative

If a test at a specific significance level α is more powerful against a specific alternative than all other tests at the same significance level, we term it *most powerful*. But as we see in Figure 2.5, a test that is most powerful for some alternatives may be less powerful for others. When a test at a specific significance level is more powerful against *all* alternatives than all other tests at the same significance level, we term such a test *uniformly most powerful*.

We term a test *admissible*, providing either a) it is uniformly most powerful or b) no other test is more powerful against all alternatives.

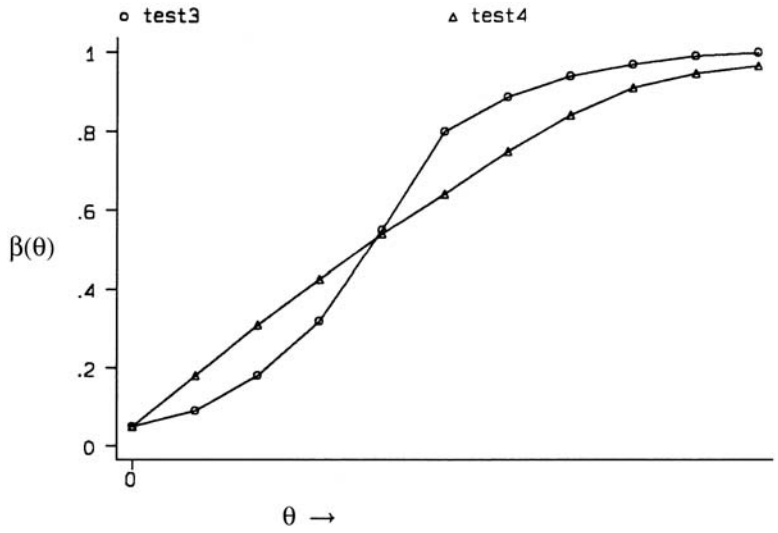

Fig. 2.5. Comparing power curves: For near alternatives, with θ close to zero, test 4 is the more powerful test; for far alternatives, with θ large, test 3 is more powerful. Thus, neither test is uniformly most powerful.

Note: One can only compare the power of tests that have the same significance level. For if the test $\varphi_1[\alpha_1]$ is less powerful than $\varphi_2[\alpha_2]$, where the significance level $\alpha_1 < \alpha_2$, then it may be that the power of $\varphi_1[\alpha_2]$ is greater than the power of $\varphi_2[\alpha_2]$.

The significance level and power may also depend upon how the variables we observe are distributed. For example, does the population distribution follow a bell-shaped normal curve with the most frequent values in the center, as in Figure 2.1a? Or is the distribution something quite different? To protect our interests, we may need to require that the Type I error be less than or equal to some predetermined value for all possible distributions. When applied correctly, permutation tests always have this property. The significance levels of parametric tests and of tests based on the bootstrap are dependent on the underlying distribution.

2.1.5 Exact, Unbiased, Conservative

In practice, we seldom know either the distribution of a variable or the values of any of the distribution's *nuisance* parameters.[1] We usually want to test a *compound hypothesis*, such as H: X has mean value 0. This latter hypothesis includes several *simple hypotheses*, such as H_1: X is normal with mean value

[1] A good example of nuisance parameters is a distribution's unknown means when variances are being compared (see Section 3.7.1).

0 and variance 1, H_2: X is normal with mean 0 and variance 1.2, and H_3: X is a gamma distribution with mean zero and four degrees of freedom.[2]

A test is said to be *exact* with respect to a compound hypothesis if the probability of making a Type I error is exactly α for each and every one of the possibilities that make up the hypothesis. A test is said to be *conservative* if the Type I error never exceeds α. Obviously, an exact test is conservative, though the reverse may not be true.

The importance of an exact test cannot be overestimated, particularly a test that is exact regardless of the underlying distribution. If a test that is nominally at level α is actually at level c, we may be in trouble before we start: If $c > \alpha$, the risk of a Type I error is greater than we are willing to bear. If $c < \alpha$, then our test is suboptimal, and we can improve on it by enlarging its rejection region.

A test is said to be *unbiased* and of level α providing its power function β satisfies the following two conditions:

- β is conservative; that is, $\beta \leq \alpha$ for every distribution that satisfies the hypothesis;
- $\beta \geq \alpha$ for every distribution that is an alternative to the hypothesis.

That is, a test is unbiased if you are more likely to reject a false hypothesis than a true one when you use such a test. I find unbiasedness to be a natural and desirable principle, but not everyone shares this view; see, for example, Suissa and Shuster [1984].

Faced with some new experimental situation, our objective always is to derive a uniformly most powerful unbiased test if one exists. But if we can't derive a uniformly most powerful test (and Figure 2.5 depicts just such a situation), then we will look for a test that is most powerful against those alternatives that are of immediate interest.

2.1.6 Impartial

Our methods should be *impartial*. Decisions should not depend on the accidental and quite irrelevant labeling of the samples; nor should decisions depend on the units in which the measurements are made nor when they are made.

To illustrate, suppose we have collected data from two samples and our objective is to test the hypothesis that the difference in location of the two populations from which the samples are drawn is less than or equal to some value. (This is called a *one-tailed* or *one-sided test*.) Suppose further that the first sample includes the values a, b, c, d, and e and the second sample the values f, g, h, i, j, k. If the observations are completely reversed, that is, if the first sample includes the values f, g, h, i, j, k and the second sample the values a, b, c, d, and e, then, if we rejected the hypothesis in the first instance, we ought to reject it in the second.

[2] In this example the variance is an example of a nuisance parameter.

The units we use in our observations should not affect our decisions. We should be able to take a set of measurements in feet, convert to inches, make our estimate, convert back to feet, and get absolutely the same result as if we'd worked in feet throughout. Similarly, where we locate the zero point of our scale should not affect the conclusions. Measurements of temperature illustrate both these points.

Finally, if our observations are independent of the time of day, the season, and the day on which they were recorded (facts which ought to be verified before proceeding further), then our decisions should be independent of the order in which the observations were collected.

Such impartial tests are said to be *invariant* with respect to the transformations involved (the conversion of units or the permutation of subscripts).

2.1.7 Most Stringent Tests

Let $\beta_\varphi(\theta)$ denote the power of a test φ against the alternative θ. Let the envelope power function $\beta^*_\alpha(\theta)$ be the supremum of $\beta_\varphi(\theta)$ over all level-α tests of the hypothesis. Then $\beta^*_\alpha(\theta) - \beta_\varphi(\theta)$ is the amount by which a specific test φ falls short of the maximum power attainable. A test that minimizes its maximum shortcoming over all alternatives θ is said to be *most stringent*.

2.2 Basic Assumptions

The parametric and bootstrap tests considered in this text rely on the assumption that successive observations are *independent* of one another. The permutation tests rely on the less inclusive assumption that they are exchangeable. We provide formal definitions of these concepts in this section and in Section 15.5.

2.2.1 Independent Observations

If you and I each flip separate coins, the results are independent of one another. But if the two of us sit together at a table while a poll taker asks us about our preferences, our responses are unlikely to be independent if you or I modify our responses in an effort to please or placate one another.

We say that two observations X_1 and X_2 are *independent* of one another with respect to a collection of events $\mathcal{A}$ if

$$\Pr\{X_1 \in A \text{ and } X_2 \in B\} = \Pr\{X_1 \in A\}\Pr\{X_2 \in B\}$$

where A and B are any two not necessarily distinct sets of outcomes belonging to $\mathcal{A}$.[3]

[3] We formalize this definition of independence in Section 15.1.

Suppose I choose to have my height measured by several individuals. These observations may well have a normal distribution with mean μ_{phil}, my height as measured by some "perfect" measuring device. With respect to the set of events leading to such observations, the various measurements on me are independent.

Suppose instead that several individuals including myself are selected from a larger population, one that has a distribution F centered about the value μ. Observations on me may be viewed as including two random components, one that results from selecting me from F and the other the observational error described in the previous paragraph. The result is to generate a much larger set of events with respect to which the observations on my height are no longer independent.[4]

Some additional examples of independent and dependent observations are given in Exercise 5. Some additional properties of independent observations are given in Section 4.1.

2.2.2 Exchangeable Observations

A sufficient condition for a permutation test such as the one outlined in the preceding chapter to be exact and unbiased against shifts in the direction of higher values is the *exchangeability* of the observations in the combined sample.[5] Let $G\{x; y_1, y_2, \ldots, y_{n-1}\}$ be a distribution function in x and symmetric in its remaining arguments—that is, if the remaining arguments were permitted, the value of G would not be affected. Let the conditional distribution function of x_i given $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ be G for all i. Then the $\{x_i\}$ are exchangeable.

Independent, identically distributed observations are exchangeable. So are samples without replacement from a finite population, termed *Polya urn models* [Koch, 1982]. An urn contains **b** black balls, **r** red balls, **y** yellow balls, and so forth. A series of balls is extracted from the urn. After the ith extraction, the color of the ball X_i is noted and k balls of the same color are added to the urn, where k can be any integer, positive, negative, or zero. The set of random events $\{X_i\}$ form an exchangeable sequence.[6]

Also exchangable are dependent normally distributed random variables $\{X_i\}$ for which the variance of X_i is a constant independent of i and the covariance of X_i and X_j is a constant independent of i and j. An additional example of dependent but exchangeable variables is given in Section 3.7.2.

Sometimes a simple transformation will ensure that observations are exchangeable. For example, if we know that X comes from a population with

[4] The student in search of greater clarity will find it in the formal exposition of Section 15.1.

[5] If an observation O consists of a deterministic part D and a stochastic part S, $O = D + S$, only the stochastic parts need be exchangeable. We use this more precise definition in Chapter 6.

[6] See, also, Dubins and Freedman [1979].

mean μ and distribution $F(x-\mu)$ and an independent observation Y comes from a population with mean v and distribution $F(x-v)$, then the independent variables $X' = X - \mu$ and $Y' = Y - v$ are exchangeable.

In deciding whether your own observations are exchangeable and a permutation test applicable, the key question is the one we posed in the very first chapter: Under the null hypothesis of no differences among the various experimental or survey groups, can we exchange the labels on the observations without significantly affecting the results?

2.3 Decision Theory

A statistical problem is defined by three elements:

1) the class $F = (F_\theta, \theta \in \Omega)$ to which the probability distribution of the observations belongs; for example, we might specify that this distribution is either unimodal, or symmetric, or normal;
2) the set D of possible decisions $\{d\}$ one can make on observing the sample $X = (X_1, \ldots, X_n)$;
3) the loss $L(d, \theta)$, expressed in dollars, persons' lives or some other quantifiable measure, that results when we make the decision d when θ is true.

The problem is a statistical one when the investigator is not in a position to say that X will take on exactly the value x, but only that X has some probability $P\{A\}$ of taking on values in the set A.

So far in this chapter we've limited ourselves to two-sided decisions in which either we accept a hypothesis H and reject an alternative K or we reject the hypothesis H and accept the alternative K.

One example is H: $\theta \leq \theta_0$ K: $\theta > \theta_0$. In this example we would probably follow up our decision to accept or reject with a confidence interval for the unknown parameter θ. This would take the form of an interval $(\theta_{\min}, \theta_{\max})$ and a statement to the effect that the probability that this interval covers the true parameter value is not less than $1 - \alpha$. This use of an interval can rescue us from the sometimes undesirable all-or-nothing dichotomy of hypothesis vs. alternative.

Our objective is to come up with a decision rule D, such that when we average out over all possible sets of observations, we minimize the associated risk or expected loss,

$$R(\theta, D) = EL(\theta, D(X)).$$

In the first of the preceding examples, we might have

$$\begin{aligned} L(\theta, d) &= 1 \quad \text{if } \theta \in K \text{ and } d = H \text{ (Type II error)}, \\ L(\theta, d) &= 10 \quad \text{if } \theta \in H \text{ and } d = K \text{ (Type I error)}, \\ L(\theta, d) &= 0 \quad \text{otherwise.} \end{aligned}$$

Typically, losses L depend on some function of the difference between the true (but unknown) value θ and our best guess θ^* of this value, the *absolute*

deviation $L(\theta, \theta^*) = |\theta^* - \theta|$, for example. Other typical forms of the loss function are the *square deviation* $L(\theta^* - \theta)^2$, and the *jump*, that is, no loss occurs if $|\theta^* - \theta| < \delta$, and a big loss occurs otherwise.

Unfortunately, a testing procedure that is optimal for one value of the parameter θ might not be optimal for another. This situation is illustrated in Figure 2.5 with two decision curves that cross over each other. The risk R depends on θ, and we don't know what the true value of θ is! How are we to choose the best decision? This is the topic we now discuss by considering Bayes, mini-max, and generalized decisions.

2.3.1 Bayes' Risk

One seldom walks blind into a testing situation. Except during one's very first preliminary efforts, one usually has some idea of the magnitude and likelihood of the expected effect. This is particularly true of clinical trials that are usually the culmination of years of experimental effort, first on the computer to elicit a set of likely compounds, and then in the laboratory in experiments with inbred mice and, later, dogs or monkeys. The large scale Phase III clinical trial takes place only after several years. And even then after small numbers of humans have been exposed to determine the maximum safe dose and the minimum effective dose.

In the case of a simple alternative, we may start with the idea that the prior probability that the null hypothesis is true is close to 1, while the probability of the alternative is near 0. As we gain more knowledge through experimentation, we can assign posterior odds to the null hypothesis with the aid of Bayes' theorem:

$$\Pr\{(H)E_1, \ldots, E_n, E_{n+1}\} = \frac{\Pr\{E_{n+1}|H\}\Pr\{H|E_1, \ldots, E_n\}}{\Pr\{E_{n+1}|H\}\Pr\{H|E_1, \ldots, E_n\} + \Pr\{E_{n+1}|K\}\Pr\{K|E_1, \ldots, E_n\}},$$

where $E_1, \ldots, E_{n+1}$ are the outcomes of various experiments.

We may actually have in mind an a *prior* probability density $\rho(\theta)$ over all possible values of the unknown parameter, and so we use our experiment and Bayes' theorem to deduce a posterior probability density $\rho'(\theta)$.

Here is an example of this approach, taken from a report by D.A. Berry[7]:

> A study reported by Freireich et al.[8] was designed to evaluate the effectiveness of a chemotherapeutic agent 6-mercaptopurine (6-MP) for the treatment of acute leukemia. Patients were randomized to therapy in pairs. Let

[7] The full report titled "Using a Bayesian approach in medical device development" may be obtained from Donald A. Berry at the Institute of Statistics & Decision Sciences and Comprehensive Cancer Center, Duke University, Durham NC 27708–025.

[8] *Blood* 1963; **21**: 699–716.

> p be the population proportion of pairs in which the 6-MP patient stays in remission longer than the placebo patient. (To distinguish probability p from a probability distribution concerning p, I will call it a population *proportion* or a *propensity*.) The null hypothesis H_0 is $p = 1/2$: no effect of 6-MP. Let H_1 stand for the alternative hypothesis that $p > 1/2$. There were 21 pairs of patients in the study, and 18 of them favored 6-MP.
>
> Suppose that the prior probability of the null hypothesis is 70 percent and that the remaining probability of 30 percent is on the interval (0,1) uniformly.... So under the alternative hypothesis H_1, p has a uniform(0,1) distribution. This is a mixture prior in the sense that it is 70 percent discrete and 30 percent continuous.
>
> The uniform(0,1) distribution is also the beta(1,1) distribution. Updating the beta(a, b) distribution after s successes and f failures is easy, namely, the new distribution is beta$(a + s, b + f)$. So for $s = 18$ and $f = 3$, the posterior distribution under H_1 is beta(19,4).

If our decision procedure is $\delta(X)$ and our loss function is $L(\theta, \delta(X))$, our *risk* when θ is true is $R(\theta, \delta) = L(\theta, \delta(X))$, and our overall average loss is $r(\rho, \delta) = R(\theta, \delta)\rho(\theta)d\theta$. A decision procedure d that minimizes $r(\rho, d)$ is called a *Bayes' solution* and the resultant r, the *Bayes' risk*.

Suppose Θ, the unobservable parameter, has probability density $\rho(\theta)$, and that the probability density of X when $\Theta = \theta$ is $p_\theta(x)$. Let $p(x) = \rho(\theta')$ $p_{\theta'}(x)d\theta'$. Let $\pi(\theta|x)$ denote the a posteriori probability density of Θ given x, which by Bayes' theorem is $\rho(\theta)p_\theta(x)/p(x)$. Then Bayes' risk can also be written as $L(\theta, \delta(x))\pi(\theta|x)d\theta]p(x)dx$.

In the case of testing a simple alternative against a simple hypothesis, let the cost of each observation be c. This cost could be only a few cents (if, say, we are testing the tensile strength of condoms) or more than $10,000 in the case of some clinical trials. Let c_1 and c_2 denote the costs associated with Type I and Type II errors, respectively. Then the Bayes' risk of a procedure d is

$$r(\rho, d) = \pi[\alpha c_1 + cE_0N] + (1 - \pi)[(1 - \beta)c_2 + cE_1N].$$

2.3.2 Mini-Max

An insurance company uses the expected risk in setting its rates, but those of us who purchase insurance use a quite different criterion. We settle for a fixed loss in the form of the insurance premium in order to avoid a much larger catastrophic loss. Our choice of procedure is the decision rule d, here the decision to pay the premium that *minimizes the maximum risk* for all possible values of the parameter.

Other possible criteria fall somewhere in between these two. For example, we could look for the decision rule that minimizes the Bayes' risk among the class of all decision rules for which $R(\theta, d)$ never exceeds some predetermined upper bound.

2.3.3 Generalized Decisions

The simple dichotomy of hypothesis versus alternative and the associated set of decisions, accept or reject, covers only a few cases. More often, we will have a choice among many decisions.

Recently, a promising treatment was found for a once certain fatal disease. Not all patients were cured completely; for some, there was a temporary remission of the disease, which allowed other cures to be tried, while other patients could only report that they felt better, and, alas, there were still many for whom the inevitable downward progress of the disease continued without interruption. The treatment was expensive and carried its own separate risks for the patient. A university laboratory had come up with a predictive method that could be employed prior to starting the treatment. Still, this method wasn't particularly reliable. The small company for whom I worked as a consultant felt sure its technology would yield a far superior predictive measure. The question for the statistician was how the company could turn this feeling into something more substantial, something that could be used to convince both venture capitalists and regulatory agencies of the new method's predictive value.

A committee was formed consisting of two physicians—specialists in the disease and its treatment, a hospital administrator, and a former senior staff member of a regulatory agency. Each was asked to provide their estimates of the costs or losses, relative or absolute, that would be incurred if a measure predicted one response, while the actual outcome was one of the three alternatives. The result, after converting all the costs to relative values and then averaging them, was a loss matrix that looked like this:

	Cured	Remission	Slight relief	No effect
Cured	0	−1	−3	−6
Remission	−2	0	−2	−4
Slight relief	−5	−2	0	−1.2
No effect	−10	−5	−1	0

We already had records for a number n of patients, including samples of frozen blood that had been drawn prior to treatment. These were tested by each of the proposed prediction methods. For each method, we then had an overall risk given by the formula $\sum_i L[d_i, \delta_i]$, where the sum was taken over the entire sample of patients. Since there were only four outcomes, we might also have written this sum as $\sum_{k=1}^{4} \sum_{j=1}^{4} f_n[k, j] L[d_k, \delta_j]$, where $f_n[k, j]$ is the empirical frequency distribution of outcomes for this sample of patients.

In situations where the objective is to estimate the value of a parameter θ, the further apart the estimate θ^* and the true value θ, the larger our losses are likely to be. Typical forms of the loss function in such a case are the absolute deviation $|\theta^* - \theta|$; the square deviation $(\theta^* - \theta)^2$; and the jump, that is, no loss if $|\theta^* - \theta| < \delta$; and a big loss otherwise. Or the loss function may resemble the

square deviation but take the form of a step function increasing in discrete increments.

Where estimation is our goal, our objective may be one of two: either to find a decision procedure $d[X]$ that minimizes the risk function $R(\theta, d) = E_\theta [L(\theta, d[X]]$ or the average loss as in our prediction example, or to find a procedure that minimizes the maximum loss. Note that the risk is a function of the unknown parameter θ, so that an optimal decision procedure based on minimizing the risk may depend upon that parameter unless, as in the example of hypothesis testing, there should exist a uniformly most powerful test.

2.4 Exercises

1. a) Power. Sketch the power curve $\beta(\theta)$ for one or both of the two-sample comparisons described in this chapter. (You already know one of the points for each power curve. What is it?)
   b) Using the same set of axis, sketch the power curve of a test based on a much larger sample.
   c) Suppose that without looking at the data you
      i) always reject;
      ii) always accept;
      iii) use a chance device so as to reject with probability α.

   For each of these three tests, determine the power and the significance level. Are any of these three tests exact? unbiased?
2. Suppose that we are testing a simple hypothesis H against a simple alternative K.
   a) Show that if $\alpha_1 \leq \alpha_2$ then $\beta_1 \leq \beta_2$.
   b) Show that if the test $\varphi_1[\alpha_1]$ is less powerful than $\varphi_2[\alpha_2]$ where the significance level $\alpha_1 < \alpha_2$, it may be that $\varphi_1[\alpha_2] > \varphi_2[\alpha_2]$.
3. a) The advertisement reads, "Safe, effective, faster than aspirin." A picture of a happy smiling woman has the caption, "My headache vanished faster than I thought possible." The next time you are down at the pharmacy, the new drug is there at the same price as your favorite headache remedy. Would you buy it? Why or Why not? Do you think the ad is telling the truth? What makes you think it is?
   b) In the United States, in early 1995, a variety of government agencies and regulations would almost guarantee the ad is truthful—or, if not, that it would not appear in print a second time. Suppose you are part of the government's regulatory team reviewing the evidence supplied by the drug company. Looking into the claim of safety, you are told only "we could not reject the null hypothesis." Is this statement adequate? What else would you want to know?

4. Unbiasedness. Suppose a and m denote the arithmetic mean and median of a random variable Y, respectively. Show that
   a) For all real b, c such that $a \le b \le c$, $E(Y-b)^2 \le E(Y-c)^2$.
   b) For all real d, e such that $m \le d \le e$, $E|Y-d| \le E|Y-e|$.
5. Do the following constitute independent observations?
   a) Number of abnormalities in each of several tissue sections taken from the same individual.
   b) Sales figures at Eaton's department store for its lamp and cosmetic departments.
   c) Sales figures at Eaton's department store for the months of May through November.
   d) Sales figures for the month of August at Eaton's department store and at its chief competitor Simpson-Sears.
   e) Opinions of several individuals whose names you obtained by sticking a pin through a phone book, and calling the "pinned" name on each page.
   f) Dow Jones Index and GNP of the United States.
   g) Today's price in Australian dollars of the German mark and the Japanese yen.
6. To check out a new theory regarding black holes, astronomers compare the number of galaxies in two different regions of the sky. Six non-overlapping photographs are taken in each region, and three astronomers go over each photo with each recording his counts. Would the statistical method described in Section 1.3 be appropriate for analyzing this data? If so, how many different rearrangements would there be?
7. a) *Decisions*. Suppose you have two potentially different radioactive isotopes with half-life parameters λ_1 and λ_2, respectively. You gather data on the two isotopes and, taking advantage of a uniformly most powerful unbiased permutation test, you reject the null hypothesis H: $\lambda_1 = \lambda_2$ in favor of the one-sided alternative $\lambda_1 > \lambda_2$. What are you or the person you are advising going to do about it? Will you need an estimate of $\lambda_1 > \lambda_2$? Which estimate will you use? (Hint: See Section 3.2 in the next chapter.)
   b) Review some of the hypotheses you tested in the past. Distinguish your actions after the test was performed from the conclusions you reached. (In other words, did you do more testing? rush to publication? abandon a promising line of research?) What losses were connected with your actions? Should you have used a higher/lower significance level? Should you have used a more powerful test or taken more/fewer observations? Were all the assumptions for your test satisfied?
8. a) Your lab has been gifted with a new instrument offering 10 times the precision of your present model. How might this affect the power of your tests? their significance level? the number of samples you'll need to take?

b) A directive from above has loosened the purse strings so you now can test larger samples. How might this affect the power of your tests? their significance level? the precision of your observations? the precision of your results?

c) A series of lawsuits over silicon implants you thought were harmless has totally changed your company's point of view about the costs of sampling. How might this affect the number of samples you'll take? the power of your tests? their significance level? the precision of your observations? the precision of your results?

9. Give an example (or two) of identically distributed observations that are not independent.
10. Are the residuals exchangeable in a regression analysis? an analysis of variance?
11. Suppose a two-decision problem has the loss matrix $\begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}$. Show that any mini-max procedure is unbiased.
12. Bayes' solutions. Let Θ be an unobservable parameter with probability density $\rho(\theta)$ and suppose we desire a point estimate of a real-valued function $g(\theta)$.
    a) If $L(\theta, d) = (g(\theta) - d)^2$, the Bayes' solution is $E[g(\Theta)|x)$.
    b) If $L(\theta, d) = |g(\theta) - d|$, the Bayes' solution is median $[g(\Theta)|x)$.
13. Many statistical software packages now automatically compute the results of several tests, both parametric and nonparametric. Show that, unless the choice of test statistic is determined before the analysis is performed, the resultant p-values will not be conservative.

http://www.springer.com/978-0-387-20279-2

Permutation, Parametric, and Bootstrap Tests of Hypotheses

Good, P.I.

2005, XX, 316 p. 14 illus., Hardcover
ISBN: 978-0-387-20279-2