scientific standpoint. What is important is not the type of observation, but whether it matches the question at hand. This is what decides whether an observation counts toward evidence in a given instance. Evidence comes from scientifically guided empirical observations combined with background information, logic, and scientific expertise. Observations may come from both manipulative and observational experiments. Experiments may be either theory driven experiments, designed to test a particular theory, or poke-at-it experiments, designed just to look at something potentially interesting. Because all theories of global import deal with phenomena, mechanisms, and processes at many different scales, all kinds of evidence must be used in building and testing these theories. Scheiner's thesis is that scientific theories are built upon the consilience of the evidence. Maurer and Scheiner express several common themes. One is the importance of consilience in scientific process. Another is that evidence is evidence, whether data were collected before theory was proposed or after. This idea needs to be examined in light of the frequentist ideas of multiple testing and joint versus individual confidence intervals.

1 A Brief Tour of Statistical Concepts

Nicholas Lewin-Koh, Mark L. Taper, and Subhash R. Lele

ABSTRACT

This chapter serves as a tutorial, introducing key concepts of statistical inference. We describe the language and most basic procedures of Fisherian *P*-value tests, Neyman-Pearson tests, Bayesian tests, and the ratio of likelihoods as measures of strength of evidence. We demonstrate each method with an examination of a simple but important scientific question, Fisher's thesis of equal sex ratios. Even within the confines of this simple problem, these methods point toward very different conclusions.

SCIENCE AND HYPOTHESES

In the seventeenth century, Francis Bacon proposed what is still regarded as the cornerstone of science, the scientific method. He discussed the role of proposing alternative explanations and conducting tests to choose among them as a path to the truth. Bacon saw the practice of doing science as a process of exclusions and affirmations, where trial and error would lead to a conclusion. This has been the principal framework under which science has been conducted ever since. Bacon saw science as an inductive process, meaning that explanation moves from the particular to the more general (Susser, 1986).

Nicholas Lewin-Koh thanks his wife Sock-Cheng for the frequent conversations and advice that helped to shape his understanding of statistical concepts, and Paul Marriot for comments, which helped to clarify explanations and the general readability. Mary Towner and Prasanta Bandyopadhyay gave helpful comments on an earlier version of this manuscript.

Karl Popper in the twentieth century argued entirely differently. Popper (1959) argued that science progresses through deduction, meaning that we proceed from the general to the specific. Popper proposed the doctrine of falsification, which defines what is acceptable as a scientific hypothesis: if a statement cannot be falsified, then it is not a scientific proposition.

Both Popper and Bacon promote the idea that we learn through trial and error. Popper's deductive approach stipulates that, through reasoning from logical premises, we can make predictions about how systems should behave, and the hallmark of a scientific statement is that it can be tested.

> I do not think that we can ever seriously reduce by elimination the number of [the] competing theories, since this number remains infinite. What we do—or should do—is hold on, for the time being, to the most improbable of the surviving theories or, more precisely, to the one that can be most severely tested. We tentatively "accept" this theory—but only in the sense that we select it as worthy to be subjected to further criticism, and to the severest tests we can design. (Popper, 1959, p. 419)

Popper's views have been criticized extensively, and it is generally agreed that confirmation of theories plays an important role (Sober, 1999; Lloyd, 1987). What is not controversial is that theories, models, and hypotheses need to be probed to assess their correctness. This creates a need for an objective set of methodologies for assessing the validity of hypotheses. If experimental outcomes were not subject to variation and experiments were controlled to the point where measurement error and natural variation were negligible, then hypothesis testing would be a relatively simple endeavor. However, as any practicing scientist knows, this is not the case; measurements are not always reliable, and nature is not uniform. In this context, scientists have a need for tools to assess the reliability of experimental results and to measure the evidence that the outcome of an experiment provides toward accepting or rejecting a hypothesis. To this end, statistical methods have been developed and applied, as criteria to evaluate hypotheses in the face of incomplete and imperfect data.

There are a variety of statistical approaches to hypothesis validation. As background for this book, we will briefly introduce without advocacy the ones most influential in modern science. We will relate these methods to the general framework of scientific practice that we have outlined. In the interests of concision, many practical considerations and fine details will be glossed over.

HYPOTHESES AND MODELS

Hypotheses play a key role in scientific endeavors. A hypothesis stems from theory in that a good theory suggests some explicit statements that can, in some sense, be tested (Pickett, Kolasa, and Jones, 1994; Sober, 1999). However, we must differentiate between a *scientific* hypothesis and a *statistical* hypothesis and understand the relationship of both to statistical models.

We usually start a study or set of experiments within the context of a body of knowledge and thought built from previous studies, observations, and experience. From this context arises a set of questions. We try to organize these questions into a coherent order with explanatory properties. To do this, we construct a model of how the system works. This is a *scientific model* or scientific hypothesis. To be scientifically testable, this model must be connected to observable quantities.

The expression of the scientific hypothesis in the form of a model leads to a statistical hypothesis and a corresponding statistical model. A statistical model is an explicit quantitative model of potential observations that includes a description of the uncertainty of those observations due to natural variation,[1] to errors in measurement, or to incomplete information, such as when observations are a sample from a larger or abstract population. A statistical hypothesis is a statement about the attributes of a statistical model whose validity can be assessed by comparison with real observations. If a statistical model is found to be inadequate to describe the data, the implication is that the scientific model it represents is also inadequate. Thus, a scientist may compare scientific hypotheses through the analysis of statistical models and data. Figure 1.1 shows these relationships diagrammatically. It is important to realize that, in this book, when we talk of testing hypotheses and evidence, we are referring to the statistical hypotheses, and model selection refers to the selection of statistical models. The scientific inference derived from any statistical analysis can be no better than the adequacy of the models employed.

In the next section, we demonstrate the process of transformation of a scientific hypothesis into a statistical hypothesis with a simple example (see Pickett, Kolasa, and Jones, 1994, for a more complete discussion of translation and transformation in ecological theory).

1. Natural variation occurs when entities within a class are not homogeneous, i.e., when the differences among entities are not extreme enough to class them separately but cause measurements of an attribute to vary within the class.

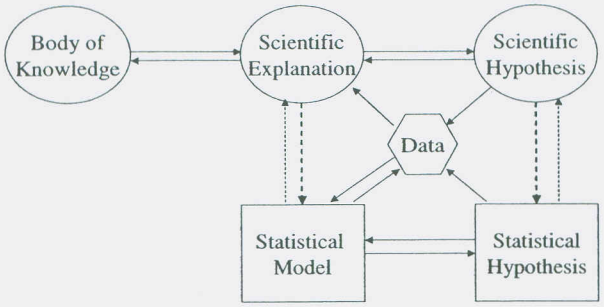

FIGURE 1.1 The relationships between statistical model and scientific explanation: The solid lines are direct translation paths, while the dotted lines are transformation paths. In this model of science, we translate an existing body of knowledge into an explanation and derive hypotheses from the explanations. An explanation is transformed into a statistical model, and hypotheses are generated and tested from the model.

SIMPLE BIOLOGICAL QUESTION

Fisher (1958, 158–60) showed that, in a sexually reproducing population with random mating and equal parental investment in the sexes, natural selection should favor equal proportions of male and female offspring. When these assumptions hold, most studies have shown that there are indeed equal proportions of the sexes in natural populations.

This problem of the sex proportions gives us a framework to demonstrate some key concepts in statistics. What does the statement "the ratio of investment between males and females should be equal for sexually reproducing species" really mean? Certainly, one does not expect that a litter of 3 offspring should have 1.5 males and 1.5 females. Instead, what is being specified is the probability that a given offspring will be male or female is equal for the entire population. If we combine this idea with the ancillary assumption that the sexes of individuals within a litter are independent, then a statistical model describing the proportion of sexes is that the number of males in a litter of a given size is binomially distributed.[2]

2. The binomial distribution arises as the sum of independent trials where each trial can be either a "success" or a "failure" (also known as Bernoulli trials). If we imagine a sequence of independent coin flips with probability θ of landing heads, if we flip the coin n times, then the probability of x successes (defined in this case as landing heads) is $\binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 1, 2, \ldots, n$. The term $\binom{n}{x}$ is the binomial coefficient, which counts the number of possible sequences of n events with x successes.

A statistical hypothesis is a statement about the parameters of a distribution. If θ is the probability of an offspring being born male, then $\theta = .5$, $\theta = .7$, $\theta > .5$, and $\theta \neq .5$ are possible statistical hypotheses. The scientific hypothesis of equal parental investment now corresponds to the statistical hypothesis $\theta = .5$.

To validate or test a statistical hypothesis corresponding to a scientific hypothesis, an investigator must gather real-world data. We utilize a classic data set from the literature on sex ratios in pig litters. The data comes from pig births registered in the British National Duroc-Jersey Pig Record, vol. 67 (Parks, 1932). There are 7929 male and 8304 female births recorded for a total of 16233 births.

How can one investigate whether these data are consistent with Fisher's scientific hypothesis of a 50% proportion of males? In the following sections, we will use these data to illustrate the mechanics of several different statistical approaches to this problem that have been very influential in science. However, before we can do this, we need to introduce a few critical statistical concepts.

THE SAMPLE SPACE, RANDOM VARIABLES, AND THE PARAMETER SPACE

A basic concept essential for understanding statistical inference is the notion of a sample space. The sample space S is the space of all possible outcomes that can occur when an experiment/observation is made. As a simple demonstration, consider our sex proportion example. For a family with three offspring, there are eight ($2^3 = 8$) possible birth sequences. Thus, the sample space is: $S = \{\{F, F, F\}, \{F, F, M\}, \{F, M, F\}, \{M, F, F\}, \{F, M, M\}, \{M, F, M\}, \{M, M, F\}, \{M, M, M\}\}$, where F indicates a female offspring and M a male.

However, our interest is in the proportion of males (or females) in the population, meaning that order is not important. Let X be the number of male piglets in a litter. Then $X(\{F, M, F\}) = 1$ and $X(\{M, M, F\}) = 2$, and so on. Corresponding to each possible litter in our sample space, we can assign a specific number to X. Such an X is called a random variable. For the case of a family size of three, the possible values of X are 0, 1, 2, or 3. The set of possible values of X is often called the range or support of the random variable.

Assuming that the probability of an offspring being male (or female) is one-half, $\theta = \frac{1}{2}$, sexes of littermates are independent, and θ is constant for

all litters, then the probability of any specific litter of 3 being entirely composed of males is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. Because only one birth sequence can occur at a time, the probability that the random variable X takes on a specific value is equal to the sum of the probabilities of all birth sequences with that number of males. Thus, as only a single birth sequence will produce all females, $P(X = 0) = \frac{1}{8}$, similarly only a single sequence will produce all males and $P(X = 3) = \frac{1}{8}$. However, there are three possible sequences with one male and also three sequences with two males; thus $P(X = 1) = \frac{3}{8}$ and $P(X = 2) = \frac{3}{8}$. The set of probabilities for each possible value of a random variable is known as the probability distribution of the random variable. If the random variable is discrete, as in this example, we refer to the distribution as a probability mass function (*pmf*) and if the random variable has continuous support we refer to the distribution as a probability density function (*pdf*).

The probability θ of a piglet being male need not be 1/2. It could conceivably be any number between 0 and 1. For a fixed value of θ, the probability distribution is given by $\{(1 - \theta)^3, 3(1 - \theta)^2\theta, 3(1 - \theta)\theta^2, \theta^3\}$. We can further generalize this to the probability distribution for any θ and any litter size as

$$\left\{\binom{n}{0}(1-\theta)^n(\theta)^0, \binom{n}{1}(1-\theta)^{n-1}(\theta)^1, \ldots, \binom{n}{n}(1-\theta)^0(\theta)^n\right\}.$$

This is called the binomial distribution with parameters θ and n and is denoted by Bin(n, θ). Often the number of observations, n, is considered known and fixed, and is thus not a parameter.

Finally, the parameter space Θ is defined as the collection of all possible values that the parameter θ could conceivably take. In this case where X is a random variable from a binomial distribution with parameter θ, where $0 < \theta < 1$, then Θ is the set $\{\theta : 0 < \theta < 1\}$.

With this background in hand, we proceed to the problem of testing statistical hypotheses.

THE MECHANICS OF TESTING STATISTICAL HYPOTHESES

There are several basic approaches to testing statistical hypotheses. In this section, we give a gentle introduction to those approaches and couch them in the framework we have presented so far. We also give examples of how these approaches work using our sex proportions example. The basic approaches that we discuss are (1) Fisher's *P*-value-based tests, (2) Neyman-Pearson tests, and (3) Bayesian tests.

Fisher's *P*-Value Tests

Fisherian significance testing focuses on quantifying the support that can be found in the data for a single hypothesis (Fisher, 1935). The question asked is "If the hypothesis under consideration were true, in replicate experiments, what would be the frequency of observing these data or more extreme data?" The data, or the data summary, are determined extreme in terms of the probability of observing the data summary under the considered hypothesis.[3] The hypothesis is considered unsupported if this frequency or probability is low. This probability is presented as a measure of evidence for or against the single hypothesis under consideration (Cox, 1977).

We proceed with a Fisherian test of the equal sexual proportions hypothesis using the data given above. The hypothesis under consideration is that the fraction of males is .5. Under the assumed binomial distribution, the probability of observing 7,929 males is .0000823. Any observation with fewer than 7,930 males or more than 8,303 males will have a probability less than or equal to .0000823 and will thus be considered an extreme event (Royall, 1997, p. 67). Summing the probability of all extreme events, we find that probability of observing an event as extreme as or more extreme than the observed 7929 males is .003331. Since this probability is quite small, this unusual result is taken as evidence against the statistical hypothesis that piglet births are distributed as independent Bernoulli trials with $\theta = .5$.

Neyman-Pearson Tests

The Neyman-Pearson approach to inference is constructed as a decision rule for deciding between two alternative hypotheses. One of these hypotheses is commonly given primacy in that it represents the status quo. This is called the null hypothesis and is generally labeled H_0. The other is called the alternate hypothesis and is commonly labeled H_1. The Neyman-Pearson approach divides the set of all possible outcomes, the sample space, into two regions, one called the acceptance region and the other the rejection region. The outcome of the test depends on which region the observed data fall into. If the data are in the acceptance region, the decision is to accept the null hy-

3. The definition of extreme is not consistent in the statistical literature. Sometimes, extremeness is defined in terms of the value of the test statistic (e.g., Cox and Hinkley, 1974, 66). However, such a definition can present problems when considering distributions that are asymmetric or multimodal.

TABLE 1.1 The table shows the different outcomes of actions in relation to the "true" state of nature.

	H_0 true	H_1 true
reject H_0	type I error	correct decision
reject H_1	correct decision	type II error

pothesis as true.[4] If, on the other hand, data fall in the rejection region, the decision is to reject the null hypothesis and accept the alternative hypothesis as true.

However, even if applied correctly, this process is bound to make some erroneous decisions in the long run. There are two kinds of errors that can be made in this process. First, if the null hypothesis were true, the investigator could mistakenly conclude that the null hypothesis should be rejected. Second, if the null hypothesis were false, the investigator could mistakenly conclude that the null hypothesis should be accepted. These two errors have been named type I and type II errors respectively. Table 1.1 illustrates the four possible outcomes in a testing situation.

The choice of how to split the sample space is made so as to control these two kinds of errors. The split is designed to satisfy two constraints. The first constraint is that the probability of making a type I error should be, at most, some predetermined but arbitrary value. This value is known as the size of the test and is typically denoted by α. Values often chosen for α are .05 or .01. There may be more than one way to partition S so that the test has the specified size α. When this is the case, then the partition of size α that minimizes the probability of a type II error is selected. The probability of type II error is designated as β. The probabilities α and β are conceived as the frequencies of each kind of error occurring if the experiment were to be replicated a large number of times.

We now apply the Neyman-Pearson approach to testing Fisher's theory of equal sex proportions. We will consider two different formulations of the alternative statistical hypotheses:

1. The first case compares two simple hypotheses about θ. A simple hypothesis completely specifies the distribution of the random variable.

4. Sometimes, instead of talking about accepting H_0, we speak in terms of failure to reject H_0. This is because the null hypothesis is often set up as an overly precise specific model that is almost surely unbelievable (Lindgren, 1976, p. 278).

2. The second case compares a simple hypothesis with a composite hypothesis. A hypothesis is composite if it is not simple, i.e., it does not uniquely specify the distribution of the random variable.

First, we consider two simple hypotheses: H_0: $\theta = .5$ against H_1: $\theta = .45$, where θ is the true unobserved population sex proportion. Under the Neyman-Pearson paradigm, we need to choose between H_0 and H_1, as to which is the correct specification of the parameter. The first step is to fix the probability of type I error, say, at $\alpha = .05$. Next, we need to define the sample space. The sample space for this problem is the set of all possible values for the number of male piglets in the study, that is $\{0, 1, \ldots, 16233\}$. The Neyman-Pearson lemma (Casella and Berger, 1990) provides an optimal partition of this sample space so that the probability of type I error is equal to α (.05) and the probability of type II error is minimized. For this problem, the acceptance region[5] can be found (Casella and Berger, 1990) to be $\{8012, 8013, \ldots, 16233\}$ and the rejection region to be $\{0, 1, \ldots, 8011\}$. The observed number of male piglets is 7929. This is in the rejection region; thus, we reject H_0, $\theta = .5$ and accept H_1: $\theta = .45$ as true.

Many times in practice, we do not wish to be constrained to a simple hypothesis but, instead, wish to test H_0 against all possible alternatives. This is called testing a composite hypothesis. A composite hypothesis is not represented by a single distribution such as Bin (16233, .5), but instead is represented by a collection of distributions e.g., $\{\text{Bin}(16233, \theta): \theta \neq .5\}$. The Neyman-Pearson lemma does not apply to the testing of composite hypotheses. However, a reasonable test can be constructed using similar ideas (see Casella and Berger, 1990, for details). Now the acceptance region is the set $\{7992, \ldots, 8241\}$ and the rejection region is the union of the sets $\{0, 1, \ldots, 7991\}$ and $\{8242, \ldots, 16233\}$. Again, H_0 is rejected, but in this case we conclude only that θ is something other than .5.

Bayesian Tests

The previous two approaches, Fisherian P-values and Neyman-Pearson testing, both construct rules for decision making. These rules are designed to optimize the decision-making process over many repeated identical experi-

5. Strictly speaking, for the discrete distributions, it is generally not possible to find an acceptance region that has size exactly equal to α. One could use what is called a randomized test. However, for the sake of simplicity, we ignore this subtlety and provide an approximate acceptance region.

ments. Bayesians reject the concept of optimizing the decision process for hypothetically repeated experiments. They believe that the analysis and decisions should be made only on the basis of the observed data and not on the basis of what we could have observed had we repeated the experiment over and over again. The Bayesian framework is not formulated in terms of decision, but in terms of belief. Before any data are observed, the scientist quantifies his or her belief about competing hypotheses in terms of probabilities. If the scientist believes strongly that a hypothesis is true, then that hypothesis is given a high probability.

The belief that the scientist has in the hypotheses before observing the data is termed the prior probability distribution. The Bayesian analysis is a formulation of the influence the data have on prior belief. The belief after the data are taken into consideration is called the posterior probability distribution. Once data have been observed, Bayesian scientists' beliefs are changed according to the following rule.

Let $P(H_0)$ and $P(H_1)$ be the *prior* beliefs in H_0 and H_1, and let X be the observed data. To be considered a proper probability distribution one must impose the following constraints: $0 \leq P(H_0) \leq 1$, $0 \leq P(H_1) \leq 1$, and $P(H_0) + P(H_1) = 1$. Then the probability of H_0 after the data X are observed, the *posterior* probability, is given as

$$P(H_0|X) = \frac{P(H_0)P(X|H_0)}{P(X|H_0)P(H_0) + P(X|H_1)P(H_1)}.$$

Similarly, the posterior probability of H_1 is given as

$$P(H_1|X) = \frac{P(H_1)P(X|H_1)}{P(X|H_0)P(H_0) + P(X|H_1)P(H_1)}.$$

Notice that these posterior probabilities satisfy the constraints given above and thus constitute a probability distribution. Notice also that the posterior probabilities are contingent on the prior probability distribution, and thus may vary somewhat from researcher to researcher.

We now illustrate the Bayesian approach using our sex proportion example. Consider the same two simple hypotheses about θ that we used in the Neyman- Pearson example. H_0: $\theta = .5$ and H_1: $\theta = .45$. First, we need to specify the prior distribution. If we strongly believe that the true sex proportion is .5, a possible specification of the prior distribution might be

$P(H_0) = \frac{2}{3}$ and $P(H_1) = \frac{1}{3}$. This distribution represents a stronger belief in H_0 than in H_1. How do we change our belief in the light of the observation of 7929 males in 16233 piglets? Using the above formula,

$$P(H_0|X) = \frac{\frac{2}{3}\binom{16233}{7929}(.5)^{7929}(.5)^{16233-7929}}{\frac{2}{3}\binom{16233}{7929}(.5)^{7929}(.5)^{16233-7929} + \frac{1}{3}\binom{16233}{7929}(.45)^{7929}(.55)^{16233-7929}} \approx 1$$

Similar calculations show that $P(H_1|X)$ is approximately 0. Thus, after observing these data, our belief in H_0 is vastly strengthened.

PRODUCTS OF TEST PROCEDURES

The basic statistical paradigms presented above give very different products. Fisherian P-values evaluate a single hypothesis, not in relation to an alternative. The Neyman-Pearson procedure is designed to dictate a decision between two hypotheses: one is rejected and the other is accepted. Bayesian approaches result in a changed degree of belief after viewing the data.

Although these are the paradigms under which the majority of statistical analysis in science has always been conducted, there has been a history of discomfort with all of these approaches. The Neyman-Pearson paradigm does not provide a measure of the strength of evidence; the Fisher P-value paradigm purportedly has a measure for the strength of evidence, but only considers a single hypothesis; and the Bayesian posterior includes subjective elements due to the choice of prior.

AN EVIDENTIAL PARADIGM

A series of statisticians, philosophers, and scientists (Barnard, 1949; Hacking, 1965; Edwards, 1972; Royall, 1997) have felt that an explicit concept of the comparative strength of evidence is essential. To date, the primary statistical tool used in quantifying the strength evidence has been the likelihood ratio. We now explain the concepts of likelihood and likelihood ratio and illustrate their use in measuring the strength of evidence in the sex proportions example.

LIKELIHOOD, LIKELIHOOD RATIO, AND STRENGTH OF EVIDENCE

Fisher developed the concept of likelihood in 1921. The likelihood is conceived of as a measure of support for a value of θ given the data x. The likelihood function is proportional to the probability of observing the data under particular values of the parameters and is written $L(\theta; x)$. Let X be a random variable with probability density function $f(x; \theta)$, where θ is a parameter or vector of parameters and x is a vector of observations. The likelihood for the parameter θ is given by $L(\theta; x) \propto f(x; \theta)$. The likelihood might be seen as equivalent to the pdf or pmn of X, but there is a critical difference in what is considered fixed. When utilizing a pdf or pmf, θ is considered fixed and the pdf or pmf a function of x, but when utilizing a likelihood, x is considered fixed and the likelihood considered a function of θ (Casella and Berger, 1990). It is clear that $L(\theta; x)$ is not a probability because, in general, $L(\theta; x)$ integrated over all values of θ does not necessarily equal one.

Consider two hypotheses, $H_0 = \theta_0$, $H_1 = \theta_1$; the likelihood ratio for these two hypotheses is defined as $L(\theta_0; x)/L(\theta_1; x)$.

Ian Hacking (1965) suggested that likelihood ratios greater than 1 indicate support for H_0 while ratios less than 1 indicate support for H_1. If the ratio is exactly equal to 1, neither hypothesis is supported over the other. Further, the magnitude of the likelihood ratio quantifies the strength of evidence for H_0 over H_1.

We return to our example of sex ratios in pigs. The competing hypotheses are H_0: $\theta = .5$ and H_1: $\theta = .45$. The ratio of likelihoods is

$$\frac{L(H_0|X)}{L(H_1|X)} = \frac{\binom{n}{x}.5^x(.5)^{n-x}}{\binom{n}{x}.45^x(.55)^{n-x}}$$

where x is the number of males observed. The ratio is $>e^{43}$, indicating a very high degree of support for H_0 in relation to H_1. We can interpret this result as meaning that H_0 is over a quintillion times better supported than H_1, due to the size of the ratio.

INTO THE FRAY

We have described fundamental statistical concepts of Neyman-Pearson testing, Fisherian testing based on P-values, and Bayesian inference. We have

also introduced the idea of the likelihood function and the ratio of likelihoods as a possible measure of the strength of evidence for one hypothesis over another. Notice that in our examples Fisherian P values and Neyman-Pearson testing reject H_0, while the Bayesian test and the ratio of likelihoods strongly support H_0 over H_1. These inference procedures do not necessarily lead to the same conclusions. The rest of this volume largely debates the relationship between these approaches to statistical inference and what constitute appropriate measures of evidence in science.

APPENDIX

In section 1.3, we first introduced the Duroc-Jersey pig data, which we used as an example to show how different approaches to inference work. We collapsed the data over litter size to keep our examples simple. The data are far more complex than what we used for our analysis, and a more complete representation of the data is given in table 1.2. The reader will note that the values tend to center around an equal sex ratio for all litter sizes, but that there is a great deal of heterogeneity in the responses, which is common when the values are discrete and counts not very large. Regardless of which statistical approach is adopted, sophisticated techniques are available for dealing with

TABLE 1.2 The complete Duroc pig data from Parkes (1932). The table shows the frequency of each litter category defined by the size of the litter and the number of male piglets in the litter. For instance, the entry in the upper left hand corner of the table indicates that there are two litters of size two with no males. Where the number of males indicated by the row is greater than the litter size indicated by the column, the cell is empty.

Number of Males	Size of Litter												
	2	3	4	5	6	7	8	9	10	11	12	13	14
0	2	4	1	2	3	0	1	0	0	0	0	0	0
1	5	7	14	20	16	21	8	2	7	1	0	0	0
2	2	9	23	41	53	63	37	23	8	3	1	1	0
3		4	14	35	78	117	81	72	19	15	8	0	0
4			1	14	53	104	162	101	79	15	4	2	0
5				4	18	46	77	83	82	33	9	1	0
6					0	21	30	46	48	13	18	9	1
7						2	5	12	24	12	11	4	5
8							1	7	10	8	15	2	1
9								0	0	1	4	0	0
10									0	1	0	0	0

the inherent problems in the data. And in fact, Parks (1932) points out some of the pitfalls of using a binomial test on these data due to the problems of overdispersion introduced by ratios of whole numbers.

REFERENCES

Barnard, G. A. 1949. Statistical Inference. *J. Roy. Stat. Soc.*, ser. B., 11:115–139.

Casella, G., and R. L. Berger. 1990. *Statistical Inference.* Belmont, CA: Duxbury.

Cox, D. R. 1977. The Role of Significance Tests. *Scand. J. Statistics.* 4:49–70.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics.* London: Chapman and Hall.

Edwards, A. W. F. 1972. *Likelihood: An Account of the Statistical Concept of Likelihood and its Application to Scientific Inference.* Cambridge: Cambridge University Press.

Fisher, R. A. 1935. *Statistical Methods for Research Workers.* 5th ed. London: Oliver and Boyd.

Fisher, R. A. 1958. *The Genetical Theory of Natural Selection.* New York: Dover.

Hacking, I. 1965. *Logic of Statistical Inference.* Cambridge: Cambridge University Press.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses.* New York: Wiley.

Lindgren, B. W. 1976. *Statistical Theory.* New York: Macmillan.

Lloyd, E. A. 1987. Confirmation of Ecological and Evolutionary Models. *Biol. Phil.* 2: 277–293.

Parks, A. S. 1932. Studies on the Sex Ratio and Related Phenomena. *Biometrika* 15: 373–381

Pickett, S. T. A., J. Kolasa, and C. G. Jones. 1994. *Ecological Understanding.* New York: Academic Press.

Popper, K. R. 1959. *The Logic of Scientific Discovery.* New York: Basic Books.

Royall, R. M. 1997. *Statistical Evidence: A Likelihood Paradigm.* New York: Chapman and Hall.

Sober, E. 1999. Testability. *Proc. Addresses Amer. Phil. Assn* 73:47–76.

Susser, M. 1986. The Logic of Sir Karl Popper and the Practice of Epidemiology. *Am. J. Epidemiol.* 124:711–718.

2 Models of Scientific Inquiry and Statistical Practice: Implications for the Structure of Scientific Knowledge

Brian A. Maurer

ABSTRACT

Practitioners of science often go about their craft in a manner that is unique to each discipline. Statistics must serve different purposes defined by the nature of the subject matter and maturity of a given discipline. Ecology operates under a mixture of techniques, philosophies, and goals. I examine two complementary models of scientific inquiry within ecology, each of which serves unique functions in the discovery of ecological knowledge. Inductive science begins with the accumulation of observations with the intent of discovering patterns. Explanations for patterns are generated post hoc. Repeatable patterns are sought in order to develop generalizations and subsequently theories. For this kind of science, parameter estimation is more useful than formal hypothesis testing. When used, null hypotheses focus on distinguishing "real" patterns from "random" ones. Bayesian statistics are helpful because information on existing patterns can be used to inform estimation procedures seeking to detect additional patterns. Deductive science begins with proposed explanations deduced from formal theories. From these, specific predictions are made about patterns that might arise from data. Data are used to design "strong tests" for the predictions, with the intention of exposing possible errors in a theory. Hypothetico-deductive experimental designs are used to maximize the chance of detecting theoretical flaws by falsification of predictions. Statistical hypothesis tests with a priori choice of significance levels are used. Bayesian statistics can obfuscate formal tests by including information not specifically contained within the experiment itself. These two kinds of science occur in different situations: generally, inductive science is more useful when the field of inquiry or the investigator is "young," while deductive science emerges as the better alter-