Chapter 12

Exploration of Old Ideas: A Critique of Classical Statistics

I can see clearly now, the pain is gone
I can see all obstacles in my way
Gone are the dark clouds that had me blind
Gonna be a bright (bright), bright (bright) sun-shiny day!

—Johnny Nash

12.1 Introduction

A volume on statistics would be remiss if it failed to comment on sampling theories, since they occupy so much space in many statistics books and journals. The distinction between the sampling theory and Bayesian viewpoint is stark. It comes down to the issue of what is to be considered fixed and what is to be considered random.

The Bayesian viewpoint is quite simple. All the quantities of interest in a problem are tied together by a joint probability distribution. (Often this joint probability distribution is expressed as a likelihood (*i.e.*, a probability distribution of the data given the parameters) times a prior distribution on the parameters.) This probability distribution reflects the beliefs of the person doing the analysis. Since these beliefs are not necessarily shared by the intended readers, the reasoning behind the beliefs should be explained and defended. Any decisions that are to be made before new data are available are made by maximizing expected utility, where the expectation is taken with respect to the probability distribution specified.

When new information becomes available, in the form of data or otherwise, that new information is conditioned upon, leading to a posterior distribution. And that posterior distribution is used as the distribution with respect to decisions that are made after the data become available. Thus, the probability distributions reflect the uncertainty of the author, both before and after data are observed.

Sampling theory reverses what is random and what is fixed. The parameter is taken to be fixed but unknown (whatever that might mean). The data are taken to be random, and comparisons are made between the distribution of a statistic before the data are observed, and the observed value of the statistic. It further assumes that likelihoods are known (because they are objective or by consensus) while priors are highly suspect (because they are subjective). In the sections below we'll look at examples of reasoning of this kind.

Chapter 9 discusses how to handle missing data in a Bayesian framework. From a sampling theory framework, it is unclear whether missing data are *(i)* fixed parameters that become random when they are observed, *(ii)* "data" that are to be treated as random when they are observed, or *(iii)* a third kind of quantity with its own set of rules.

A simple example can show how difficult it is to adhere to sampling theory. Suppose I observe a random sample of n from a normal distribution with mean μ and variance σ^2. Then the likelihood function is

$$
\begin{aligned}
f(x,\ldots,x_n \mid \mu,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp \quad -\frac{1}{2} \quad \frac{x_i-\mu}{\sigma} \quad ^{2} \\
&= \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{\sigma^n} \exp \quad -\frac{1}{2} \sum_{i=1}^{n} \quad \frac{x_i-\mu}{\sigma} \quad ^{2} \quad . \qquad (12.1)
\end{aligned}
$$

We'll adopt the popular sampling theory estimation using maximum likelihood. It is easily shown that the maximum likelihood estimators are

$$
\hat{\mu} = \bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} \text{ and } \hat{\sigma}^2 = \frac{{}^{n}_{i=1}(x_i-\bar{x})^2}{n}. \qquad (12.2)
$$

Now suppose that in fact there were originally $2n$ observations, of which the n observed are chosen at random. How should $X_{n+1},\ldots,X_{2n}$ be treated?

It would seem legitimate to think that from $x_1\ldots,x_n$ I have learned something about $X_{n+1},\ldots,X_{2n}$. So perhaps I can treat them as parameters. If I do, the maximum likelihood estimates are now

$$
\begin{aligned}
\hat{\mu} &= \bar{X} = \sum_{i=1}^{n} x_i/n, \\
\hat{X}_{n+1} &= \hat{X}_{n+2} = \ldots = \hat{X}_{2n} = \bar{X} \qquad (12.3) \\
\hat{\sigma}^2 &= \frac{{}^{2n}_{i=1}(x_i-\bar{x})^2}{2n} = \frac{{}^{n}_{i=1}(x_i-\bar{x})^2}{2n}.
\end{aligned}
$$

Hence by imagining another n data points I never saw, the estimate of the variance is now half of what it was. And of course if I imagine kn normal random variables, I get

$$
\hat{\sigma}^2 = \frac{{}^{n}_{i=1}(x_i-\bar{x})^2}{kn} \qquad (12.4)
$$

so by dint of a great imagination ($k \to \infty$), the maximum likelihood estimate of the variance vanishes!

Of course what should be done with $X_{n+1},\ldots,X_{2n}$ is to integrate them out. But there is no real distinction between unobserved data and a parameter. And to integrate out a parameter means it must have a distribution, which is where Bayesians were to begin with.

For more on this, see Bayarri et al. (1988).

Some hint of the havoc caused by the doctrine that parameters do not have distributions can be seen in the classical treatment of fixed and random effects, as in Scheffé (1999). Take for example an examination given to each child in a class, and several classes in a school, as discussed in Chapter 9. If you are interested in each individual child's performance, classical doctrine says to use a fixed effect model. However, if you are interested in how the classes compare, you should use a random effects model. This is a puzzle on several grounds:

1) According to the classical theory, the model represents how the data were in fact generated. But the above account has the model dependent on the interest of the investigator, the children or the classes, which is a utility matter.

2) To have a random effects model means to use a prior on the parameters for each child, and to integrate those parameters out of the likelihood. So here a classical statistician apparently feels OK about using such a prior. If that's OK for random effects, why not elsewhere?

3) An investigator might be interested in both each child and the classes. What model would classical statistics recommend then?

Because of this fundamental difference about what is fixed and what is random, attempts to find compromises or middle grounds between Bayesian and sampling theory statistics have failed. For example, Fisher (1935) proposed something he called fiducial inference. An instance of it looks like this:

$$X \sim N(\theta, 1) \tag{12.5}$$

$$X - \theta \sim N(0, 1) \tag{12.6}$$

$$\theta - X \sim N(0, 1) \tag{12.7}$$

$$\theta \sim N(X, 1) \tag{12.8}$$

This looks plausible if one isn't too precise about what $\sim$ means. A more careful version would write

$$X \mid \theta \sim N(\theta, 1). \tag{12.9}$$

Then one can proceed through analogs of (12.6) to get an analog of (12.7),

$$\theta - X \mid \theta \sim N(0, 1), \tag{12.10}$$

from which (12.8) does not follow. Barnard (1985) also attempted to find compromises essentially having to do with what he, following Fisher, called pivotals, like $X - \theta$ above, which have the distribution $N(0, 1)$ whether regarding X or θ as random. Fraser's structural inference (1968, 1979) is yet another attempt to find cases that can be interpreted either way. But at best these are examples of a coincidence that holds only in special cases. As soon as there is divergence, the issue must be addressed of which is fundamental and which is not. Hence, each reader has to decide for themselves what path to take.

12.1.1 Summary

The key distinction between Bayesian and sampling theory statistics is the issue of what is to be regarded as random and what is to be regarded as fixed. To a Bayesian, parameters are random and data, once observed, are fixed. To a sampling theorist, data are random even after being observed, but parameters are fixed. Whether missing data are a third kind of object, neither data nor parameters, is a puzzle for sampling theorists, but not an issue for Bayesians.

Some standard modern references for sampling theory statistics are Casella and Berger (1990) and Cox and Hinkley (1974).

12.1.2 Exercises

1. Show that $\hat{\mu}$ and $\hat{\sigma}^2$ given in (12.2) maximize (12.1) with respect to μ and σ^2. HINT: maximize the log of f, first with respect to μ, and then substitute that answer in and then maximize with respect to σ.
2. Show that $\hat{\mu}, \hat{\sigma}^2$ and $\hat{X}_{n+1}, \ldots, \hat{X}_{2n}$ given in (12.3) maximize the analogue of (12.1) with respect to μ, σ^2 and $X_{n+1}, \ldots, X_{2n}$. Same hint.
3. Explain what is random and what is fixed to (a) a Bayesian and (b) a sampling theorist.

12.2 Testing

There are two flavors of testing that are part of sampling theory, Fisher's significance testing and the Neyman-Pearson testing of hypotheses. We'll consider them in that order.

Suppose that $X_1, \ldots, X_n$ are a random sample (*i.e.*, independently and identically distributed, given the parameter) from a normal distribution with mean μ and variance 1, which can be written

$$X_1 \ldots, X_n \sim N(\mu, 1). \tag{12.11}$$

Then we know that

$$\frac{(\bar{X} - \mu)}{\sqrt{n}} \sim N(0, 1). \tag{12.12}$$

Suppose that we wish to test the hypothesis that $\mu = 0$. (Such a hypothesis is called "simple," reflecting the fact that it consists of a single point in parameter space. A "composite" hypothesis consists of at least two points.) If $\mu = 0$,

$$P\{\mid \bar{X}_n \mid > 1.96/\sqrt{n}\} = 0.05, \tag{12.13}$$

so Fisher would say that the hypothesis that $\mu = 0$ is rejected at the .05 level. This is (more's the pity) the most common form of statistical inference used today.

Of course, the number 0.05 (called the size of the test) is arbitrary and conventional, but that's not the heart of the difficulties with this procedure.

What does it mean to reject such a hypothesis? Fisher (1959a, p. 39) says that it means that either the null hypothesis is false or something unusual has happened. However this theory does not permit one to say which of the above is the case, nor even to give a probability for which is the case. If the null hypothesis is not rejected, nothing can be said. Furthermore, one may reject a true null hypothesis, or fail to reject when the null hypothesis is false.

The biggest issue with significance testing, however, is a practical one. It is easy to see (and many users of these methods have observed) that when the sample size is small, very few null hypotheses are rejected, while when the sample size is large, almost all are rejected. This is because of the $\sqrt{n}$ behavior in (12.12). Thus, while significance testing purports to be addressing (in some sense) whether $\mu = 0$, in fact the acceptance or rejection of the null hypothesis has far more to do with the sample size than it does with the extent to which the null hypothesis is a good reflection of the truth.

This lesson was driven home to me by some experiences I had early in my career. I was coauthor of a study of participation in small groups (Kadane et al. (1969)). There was a simple theory we were testing. The theory was rejected at the .05 level, the .01 level, indeed at the 10^{-6} level. I had to think about whether I would be more impressed if it were rejected at say the 10^{-13} level, and decided not. The issue was that we had a very large data set, so that any theory that isn't exactly correct (and nobody's theory is *exactly* correct) will be rejected at conventional levels of significance. A simple plot showed that the theory was pretty good, in fact.

Sometime later I was working at the Center for Naval Analyses. A study had been done comparing the laboratory to the field experience on a new piece of equipment. The draft report said that there was no significant difference. On further scrutiny, it turned out that, while the test was correctly done, there were only five field-data points (which cost a million dollars apiece to collect). Indeed, the machine was working roughly 75% as well in the field, which seemed a far more useful summary for the Navy.

These experiences taught me that with a large sample size virtually every null hypothesis is rejected, while with a small sample size, virtually no null hypothesis is rejected. And we generally have very accurate estimates of the sample size available without having to use significance testing at all!

Significance testing has been criticized for years because of its binary character (a data set, model and null hypothesis are either "significant" or "not significant".) Most recently, Amrhein, Greenland and McShane, together with over 800 signatures have called for "it's time for statistical significance to go" (Amrhein et al., 2019).

A more general view of significance testing relies on p-values., the probability under the null hypothesis of seeing data as or more discrepant than that observed. P-values have become so widely misused that the American Statistical Association, for the first time in its history, undertook to issue a statement on a methodological matter. The ASA's six points are:

"1. P-values can indicate how incompatible the data are with a specified statistical model.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.."

Significance testing violates the likelihood principle, which states that, having observed the data, inference must rely only on what happened, and not on what might have happened but did not. The Bayesian methods explored in this book obey this principle. But the probability statement in (12.13) is a statement about $\bar{X}_n$ before it is observed. After it is observed, the event $\mid \bar{X}_n \mid > 1.96/\sqrt{n}$ either happened or did not happen, and hence has probability either one or zero.

There's one other general point to make about significance testing. As discussed in section 1.1.2, it is based on a limiting relative frequency view of statistics. The interpretation is that if μ were zero and $\bar{X}_n$ were computed from many samples of size n, the proportion of instances in which $\mid \bar{X}_n \mid$ would exceed $1.96/\sqrt{n}$ would approach .05. But the application of this method is to a single instance of $\bar{X}_n$. Thus, a theory that relies on an arbitrarily large sample for its justification is being applied to a single instance.

Consider, for example, the following trivial test. Flip a biased coin that comes up heads with probability 0.95, and tails with probability 0.05. If the coin comes up tails, reject the null hypothesis. Since the probability of rejecting the null hypothesis if it is true is 0.05, this is a valid 5% level test. It is also very robust against data errors; indeed it does not depend on the data at all. It is also nonsense, of course, but nonsense allowed by the rules of significance testing.

A Bayesian with a continuous prior on μ (any continuous prior) puts probability zero on the event $\mu = 0$, and hence is sure, both prior and posterior, that the null hypothesis is false. It is an unusual situation in which a hypothesis of lower dimension than the general setting (here the point $\mu = 0$ on the real line for μ) is so plausible as to have a positive lump of probability on exactly that value.

Neyman and Pearson (1967) modify significance testing by specifying an alternative distribution, that is, an alternative value (or space of values) for the parameter. Thus, they would test (using (12.11) again) the null hypothesis $H_0 : \mu = 0$ against an alternative hypothesis, like $H_a : \mu = \mu_0 > 0$, for a specific chosen value of μ. In this case Neyman and Pearson would choose to see whether the event

$$\bar{X}_n > 1.68/\sqrt{n} \tag{12.14}$$

occurs, because, under the null hypothesis, this event has probability 0.05 and, under the alternative hypothesis, it has maximum probability. The emphasis on this probability, which they call the power of the test, is what distinguishes the Neyman-Pearson theory of testing hypotheses from Fisher's tests of significance.

Neyman and Pearson use language different from Fisher's to explain the consequences of such a test. If the event (12.14) occurs, they would reject the null hypothesis and accept the alternative. Conversely, if it does not they would accept the null hypothesis and reject the alternative. The probability of rejecting the null hypothesis if it is true is called the type 1 error rate; the probability of rejecting the alternative if it is true is called the type 2 error rate.

Again, Neyman-Pearson hypothesis testing violates the likelihood principle, because the event (12.14) either happens or does not, and hence has probability one or zero. Again, the behavior of the test depends critically on the sample size, particularly when it is used with a fixed type 1 error rate, as it most typically is. And again, a single instance of $\bar{X}_n$ is being compared to a long-run relative frequency.

The trivial test that relies on the flip of a biased coin that comes up heads with probability 0.95 is again a valid test of the null hypothesis within the Neyman-Pearson framework, but it has disappointingly low power.

Often in practice the Neyman-Pearson idea is used, not with a simple alternative (like $\mu = \mu_0$) in mind, but with a whole space of alternatives instead. This leads to power (one minus the type 2 error) that is a function of just where in the alternative space the power is evaluated.

From a Bayesian perspective, it would make more sense to ask for the posterior probability of the null hypothesis, as a substitute for significance testing, or for the conditional posterior probability of the null hypothesis given that either the null or alternative hypothesis is correct, as a substitute for the testing of hypotheses.

12.2.1 Further reading

The classic book on testing hypotheses is Lehmann and Romano (2005). More recent developments have centered on the issue of maintaining a fixed size of test when simultaneously testing many hypotheses (see, for instance, Miller (1981)). Still more recently, literature has sprung up concerning limiting the false discovery rate (Benjamini and Hochberg (1995)). A recent defense is Mayo (2018).

For a detailed comparison of methods in the context of an application, see Kadane (1990).

There have been various attempts to square testing with the Bayesian framework. For example, Jeffreys (1961) proposes to put probability 1/2 on the null hypothesis. This is unobjectionable if it is an honest opinion an author is prepared to defend, but Jeffreys presents it as an automatic prior to use in a testing problem. Thus, Jeffreys would change his prior depending on what question is asked, which is incoherent. Bayes factors are yet another way to try to fit testing hypotheses into the Bayesian framework. See (2.14).

For the context of the ASA statement, see Wasserstein and Lazar (2016). For the statement itself, and the commentary on the six points, see American Statistical Association (2016).

12.2.2 Summary

Although widely used in statistical practice, testing, whether done using the Fisher or the Neyman-Pearson approach, rests on shaky foundations.

12.2.3 Exercises

1. Vocabulary. State in your own words the meaning of:

(a) test of significance

(b) test of a hypothesis

(c) null hypothesis

(d) alternative hypothesis

(e) type I and type II error

(f) size of a test

(g) power of a test

(h) the likelihood principle

12.3 Confidence intervals and sets

The rough idea of a confidence interval or, more generally, a confidence set, is to give an interval in which the parameter is likely to be. However the fine print that goes with such a statement is crucial.

There is a close relationship between testing and confidence intervals. Indeed a confidence set can be regarded as the set of simple null hypotheses which, had they been tested, would not have been rejected at the (say) 0.05 level. More formally, it is a procedure (*i.e.*, an algorithm) for producing an interval or set having the property that (say) 95% of the time it is used it will contain the parameter value. Recall, however, that this is part of sampling theory, in which the data are random and the parameters fixed but unknown. Therefore, what is random about a confidence interval (or set) is the interval, not the parameter.

It is appealing, but wrong, to interpret such an interval as a probability statement about the parameter, because that would require a Bayesian framework in which parameters have distributions. There are such intervals and sets, called credible intervals and credible sets, which contain, say, 95% of the (prior or posterior) probability.

Like their testing cousins, confidence intervals and sets violate the likelihood principle. Also, like them, such sets rely on a single instance in a hypothetical infinite sequence of like uses for their justification. The trivial flip-of-a-biased coin example of the preceding section has the following confidence set equivalent: if the coin comes up heads (which it will with 95% probability) take the whole real line. Otherwise (with probability 5%) take the empty set. Such a random interval has the advertised property, namely that 95% of the time it will contain the true value of the parameter, whatever that happens to be. Therefore, this is a valid confidence interval. It is also useless, since we know immediately whether this is one of the favorable instances (the 95% of the time we get the whole real line), or one of the 5% of the time we get the empty set.

While such an example is extreme, the same kind of thing happens in more real settings. Consider a random sample of size two, Y_1 and Y_2, from a distribution that is uniform on the set $(\theta - 1/2, \theta + 1/2)$ for some θ (fixed but unknown). First, we do some calculations:

$$\begin{aligned} P\{\min(Y_1, Y_2) > \theta \mid \theta\} &= P\{Y_1 > \theta \text{ and } Y_2 > \theta \mid \theta\} = \\ P\{Y_1 > \theta \mid \theta\}P\{Y_2 > \theta \mid \theta\} &= 1/2 \cdot 1/2 = 1/4. \end{aligned} \tag{12.15}$$

Similarly,

$$\begin{aligned} P\{\max(Y_1, Y_2) < \theta \mid \theta\} &= P\{Y_1 < \theta \text{ and } Y_2 < \theta \mid \theta\} = \\ P\{Y_1 < \theta \mid \theta\}P\{Y_2 < \theta \mid \theta\} &= 1/2 \cdot 1/2 = 1/4. \end{aligned} \tag{12.16}$$

Therefore

$$P\{\min(Y_1, Y_2) < \theta < \max(Y_1, Y_2) \mid \theta\} = 1/2 \text{ for all } \theta, \tag{12.17}$$

so the interval $(\min(Y_1, Y_2), \max(Y_1, Y_2))$ is a valid 50% confidence interval for θ. If the length of this interval is small, however, it is less likely to contain θ than if the interval has length approaching one. Indeed if the interval has length one, we would know that θ lies within the interval, and, even more, we would know that θ is the midpoint of that interval. Thus, in this case the length of the interval gives us a very good hint about whether this is one of the favorable or unfavorable cases for the confidence interval, which is like the previous example. Because whether a procedure yields a valid confidence interval is a matter of its coverage over many (a limiting infinite number!) uses and not its character in this particular use, examples like this cause embarrassment. (This example is discussed in Welch (1939) and DeGroot and Schervish (2002, pp. 412-414).)

What property might make a particular confidence interval desirable among confidence intervals? Presumably one would like it to be short if it contains the point of interest, and wide otherwise. The standard general method is to minimize the expected length of the interval, where the expectation is taken with respect to the distribution of possible samples at a fixed value of the parameter. However this criterion is challenged by Cox (1958), who discusses the following example: Suppose the data consist of the flip of a fair coin, which is coded as $X = 0$ for heads and $X = 1$ for tails.

If $X = 0$, we see data

$$Y \sim N(\theta, \sigma_0).$$

If $X = 1$, however, we see data

$$Y \sim N(\theta, 100\sigma_0).$$

In this case, urges Cox, doesn't it make sense to offer two confidence intervals, one of $X = 0$ and a different one of $X = 1$, each having the standard structure? An interval with shorter average length can be found by making the interval, conditional on $X = 1$, a lot shorter at the cost of making the interval conditional on $X = 0$ a bit longer. See also the discussion in Fraser (2004) and in Lehmann (1986, Chapter 10). A statistic such as X is called ancillary, because its distribution is independent of the parameter. Cox and Fraser advocate conditioning on the ancillary statistic. However, Basu (1959) shows that ancillary statistics are not unique, which calls into question the general program of conditioning on ancillary statistics.

As teachers of statistics, it is common that, no matter how carefully one explains what a confidence interval is, many students misinterpret a confidence interval as if it were a (Bayesian) credible interval, that the probability is α that the parameter lies in the interval specified, where what is random and hence uncertain, is the parameter. Credible intervals and sets can be seen as a part of descriptive statistics, that is, as a quick way of conveying where the center of a distribution, prior or posterior, lies.

A confidence interval (say a 95% confidence interval) is precisely the set of null hypotheses which, had they been tested, would not have been rejected at the 5% level. Consequently, these are exactly the null hypotheses about which nothing can be said, according to Fisher. Thus, the Fisherian and Neyman-Pearson viewpoints are far apart on how to interpret a confidence interval.

An attempt to square confidence intervals with Bayesian ideas is through the concept of highest posterior density regions. Such regions depend critically on the parameterization, unlike Bayesian decision theory, which does not. I see highest posterior density regions as an attempt by early Bayesian writers not to seem too radical to their frequentist colleagues. But Bayesian thought is radically different from frequentism.

12.3.1 Summary

Like the theory of testing, the basis of confidence intervals is weak.

12.4 Estimation

An estimator of a real-valued parameter is a real-valued function of the data hoped to be close, in some sense, to the value of the parameter. As such, it is an invitation to certainty-equivalence thinking, neglecting the uncertainty about the value of the parameter inherent in the situation. Sometimes certainty-equivalence is a useful heuristic, simplifying a problem so that its essential characteristics become clearer. But sometimes, when parameter uncertainty is crucial, such thinking can lead to poor decisions. Thus, estimation is a tool worth having, but not one to be used automatically.

In order to think about which estimators might be good ones to use, it is natural to have a measure of how close the estimator $\hat{\theta}(\mathbf{x})$ is to the value of the parameter. The most commonly used measure of loss (*i.e.*, negative utility) is squared error,

$$(\hat{\theta}(\mathbf{x}) - \theta)^2. \tag{12.18}$$

When uncertainty is taken with respect to a distribution on θ (prior or posterior) the optimal estimator is

$$\hat{\theta}(\mathbf{x}) = E(\theta) \tag{12.19}$$

and the variance of θ (prior or posterior) is the resulting loss. (Indeed this estimator is called in some literature "the Bayes estimate," as if squared error were a law of nature, rather than a statement of the user's subjective utility function.)

However, when (12.18) is viewed from a sampling theory point of view, the expectation must be taken over $\mathbf{x}$ with θ regarded as fixed. The result is an expected loss that depends, with rare exceptions, on the value of θ. The two candidate estimators can have expected loss functions that cross, meaning that for certain values of the parameter one would be preferred, and for other values of the parameter, a different one would be preferred. Since the sampling theory paradigm has no language to express the idea that certain parts of the parameter space are more likely (and hence more important) than others, an impasse results. A plethora of principles then ensues, with no guidance of how to choose among them except for the injunction to use something "sensible," whatever that might mean.

One criterion often used by sampling theory statisticians is unbiasedness, which requires that

$$E(\hat{\theta}(\mathbf{X})) = \theta \tag{12.20}$$

for all θ, where the expectation is taken with respect to the sampling distribution of $\mathbf{X}$. And among unbiased estimators, one with minimum (sampling) variance is to be preferred. Of course this violates the likelihood principle, since it depends on all the samples that might have been observed but were not. Nonetheless, I can see some attractiveness to this idea in the case in which the same commercial entities do business with each other repetitively. Each can figure that whatever such a rule may cost them today will be balanced out over the long run. And here there is a valid long run to consider, unlike most other applications of statistics.

However, unbiased estimates don't always exist, and many times minimum-variance unbiased estimates exist only when unbiased estimates are unique. Consider estimating the function $e^{-2\lambda}$ where X has a Poisson distribution with parameter λ.

An unbiased estimate is

$$I\{X\text{is even}\} - I\{X \text{ is odd}\},$$

which has expectation

$$E(I\{X\text{is even}\} - I\{X\text{is odd}\}) =$$

$$e^{-\lambda}(1 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \ldots) - e^{-\lambda} \quad \frac{\lambda}{1!} + \frac{\lambda^3}{3!} + \ldots \quad = e^{-2\lambda}, \tag{12.21}$$

and indeed it can be shown (Lehmann (1983, p. 114)), that this is the only unbiased estimator, and hence a fortiori the minimum variance unbiased estimator. But this estimator is either +1 or −1. Of course −1 is surely too small since $e^{-2\lambda}$ is always positive, and +1 is too big, since $e^{-2\lambda}$ is always less than 1.

Another popular method is maximum likelihood estimation. Were the likelihood multiplied by the prior, what would be found is the mode of the posterior distribution. Under some circumstances, maximum likelihood estimation can thus be a reasonable general method for finding an estimate, if it is necessary to find one. However, the example discussed in section 12.1 shows that even maximum likelihood estimates can have problems when the parameter space is unclear.

That's not all of the story, however. Consider the following example: with probability p, we observe a normal distribution with mean μ and variance 1; with probability $1 - p$, we observe a normal distribution with mean μ and variance σ^2. Thus, the likelihood is

$$p\phi(x-\mu) + \quad \frac{1-p}{\sigma} \quad \phi \quad \frac{x-\mu}{\sigma} \tag{12.22}$$

for a single observation x, and the product of these for a sample of size n is:

$$f(x \mid \mu, \sigma, p) = \prod_{i=1}^{n} \quad p\phi(x_i - \mu) + \frac{1-p}{\sigma}\phi \quad \frac{x_i - \mu}{\sigma} \qquad . \tag{12.23}$$

Maximizing (12.23) with respect to μ, σ and p yields the following: if $\hat{\mu} = x_i$ for some i, $\sigma \rightarrow 0$ and $\hat{p} = 1/2$, the likelihood goes to infinity! Thus, for a sample of size n there are n maximum likelihood estimates for μ. And this example has only 3 parameters and independent and identically distributed observations.

Another example shows just how unintuitive maximum likelihood estimation can be. An urn has 1000 tickets, 980 of which are marked 10θ and the remaining 20 are marked θ, where θ is the parameter of interest. One ticket is drawn at random, and the number x on the ticket is recorded. The maximum likelihood estimate $\hat{\theta}$ of θ is $\hat{\theta} = x/10$, and this has 98% probability of being correct.

Now choose an $\epsilon > 0$; think of ϵ as positive but small. Let $a_1, \ldots, a_{980}$ be 980 distinct constants in the interval $(10 - \epsilon, 10 + \epsilon)$. Suppose now that the first 980 tickets in the urn are marked $\theta a_1, \ldots, \theta a_{980}$, while the last 20 continue to be marked θ. Again, we choose one ticket chosen at random, and observe the number x marked. Then the likelihood is

$$L(\theta|x) = \begin{cases} .02 & \theta = x \\ .001 & \theta = x/a_i \ i = 1, 2, \ldots, 980 \ . \\ 0 & \text{otherwise} \end{cases}$$

Hence the maximum likelihood estimator in this revised problem is $\hat{\theta} = x$, which has only a 2% probability of being correct. We know that there is a 98% probability that θ is in the interval $(x/(10 + \epsilon), x/(10 - \epsilon))$, but maximum likelihood estimation is indifferent to this knowledge.

12.4.1 Further reading

The classic book on estimation is Lehmann (1983). An excellent critique of estimation from a Bayesian perspective is given by Box and Tiao (1973, pp. 304-315). The second example of peculiar behavior of a maximum likelihood estimate is a modification of one given in Basu (1975).

12.4.2 Summary

Estimation is useful (sometimes) as a way of describing a prior or posterior distribution, particularly when it is concentrated around a particular value. As such, for Bayesians it is part of descriptive statistics.

12.4.3 Exercise

1. Let $\epsilon > 0$. Show that there are 980 distinct numbers between $10 - \epsilon$ and $10 + \epsilon$.

12.5 Choosing among models

Model choice is estimation applied to the highest level in the hierarchical model specified in section 9.4. Under what circumstances is it useful to choose one particular model and neglect the others? One circumstance might be if one model had all but a negligible amount of the probability. This case corresponds to estimation where a posterior distribution is concentrated around a particular value.

As a general matter, I would think it is sounder practice to keep all plausible models in one's calculations, and hence not to select one and exclude the others.

12.6 Goodness of fit

There is a burgeoning literature in classical statistics examining whether a particular model fits the data well. However, the assumptions underlying goodness of fit are rarely questioned.

Typically, fit is measured by the probability of the data if the model were true. As such, the best fitting model is one that says that whatever happened had to happen. Such a model is useless for prediction of course, but fits the data excellently. Why do we reject such a model out of hand? Because it fails to express our beliefs about the process generating the data. Also it is operational only after seeing the data, and hence is prone to hindsight bias (see section 1.1.1).

Generally, goodness of fit has to do with how regular (or well-understood) the process under study is, compared to some, often unexpressed, independence model. I think a better procedure is to be explicit about what alternative is contemplated, and then use the methods outlined in section 9.4.

12.7 Sampling theory statistics

A general issue for sampling theory statistics goes under the name of "nuisance parameters," which roughly are parameters not of interest, those that do not appear in a utility or loss function. But "nuisance" hardly describes the havoc such parameters wreak on many sampling theory methods. Bayesian analyses are undisturbed by nuisance parameters: you can integrate them out and deal only with the marginal distribution of the parameters of interest, or you can leave them in. Either way the expected utility of each decision, and hence the expected utility of the optimal decision, will be the same.

As you can see, I find serious foundational problems with each of these methods. But to voice these concerns is not to denigrate the authors cited or the many others who have contributed to sampling theory. Quite the contrary: I stand in awe and dismay at the enormous amount of statistical talent that has been devoted to work within, and try to make sense of, a paradigm with such weak foundations.

To find that the foundations of frequentism are weak is not the same as to hold that the inferences drawn using it are necessarily incorrect. Indeed, it may be that the very messiness of frequentism is a strength, in that it can encourage an adventurous attitude, in which some justification (and frequentism offers many competing justifications) can be found for a method that seems intuitively reasonable. Perhaps there's something to be learned from the history of shrinkage. It was originally approached from a frequentist perspective, but later was found to have a simple Bayesian interpretation (see section 8.2.1).

12.8 "Objective" Bayesian methods

> The notion of a reasonable degree of belief must be brought in before we can speak of a probability.
>
> —H. Jeffreys (1963, p. 402)

This volume would also be incomplete if it failed to address "Objective Bayesian" views (Bernardo (1979); Berger and Bernardo (1992)). For example, suppose a Bayesian wants to report his posterior to fellow scientists who share his model and hence his likelihood. Objective Bayesians search for priors that have a minimal effect on the posterior, in some sense. Some comments are in order:

1. It is not an accident that this hypothetical framework is exactly that of classical, sampling theoretical statistics. From the viewpoint of this book, this framework exaggerates the general acceptability of the model, and also exaggerates the lack of general acceptability of the prior. The likelihood is rarely so universally acclaimed, and often there is useful prior information to be gleaned. If you accept the argument of this book, likelihoods are just as subjective as priors, and there is no reason to expect scientists to agree on them in the context of an applied problem. Yet another difficulty with this program is ambiguity in hierarchical models of just where the likelihood ends and the prior begins.
2. The purpose of an algorithmic prior is to escape from the responsibility to give an opinion and justify it. At the same time, it cuts off a useful discussion about what is reasonable to believe about the parameters. Without such a discussion, appreciation of the posterior distribution on the parameters is likely to be less full, and important scientific information may be neglected.
3. The literature is replete with various attempts to find a unifying way to produce "low information" priors. Often these depend on the data, and violate the likelihood principle. Some make distinctions between parameters of interest and nuisance parameters, which implicitly depends on the utility function of an unstated decision problem. Some are disturbed by transformation: if a uniform distribution on $[0, 1]$ is ok for p, is the consequent for the distribution of $1/p$ also ok? Jeffreys' (Jeffreys (1939, 1961)) priors do not suffer from this, but do violate the likelihood principle. The fact that there are many contenders for "the" objective prior suggests that the choice among them is to be made subjectively. If the proponents of this view thought their choice of a canonical prior were intellectually compelling, they would not feel attracted to a call for an internationally agreed convention on the subject, as have Berger and Bernardo (1992, p. 57) and Jeffreys (1955, p. 277). For a general review of this area, see Kass and Wasserman (1996), and, on Jeffreys' developing views, *ibid.* (pp. 1344 and 1345).

And finally, there is the issue of the name. A claim of possession of the objective truth has been a familiar rhetorical move of elites, whether political, social, religious, scientific, or economic. Such a claim is useful to intimidate those who might doubt, challenge or debate the "objective" conclusions reached. History is replete with the unfortunate consequences, nay disasters, that have ensued. To assert the possession of an objective method of analyzing data is to make a claim to extraordinary power in our society. Of course it is annoyingly arrogant, but, much worse, it has no basis in the theory it purports to implement.

I write sometimes about legal problems, in which it is critical that both the likelihood and the prior are fair to the parties involved. I think of myself as modeling an impartial arbitrator. This is different from scientific reporting, in which I am comfortable with likelihoods and priors that convey real information (provided I explain the considerations that led to the modeling choices made).

12.9 The Frequentist Guarantee

This is a slogan used to justify procedures from a frequentist viewpoint. It says that, with specified long-run frequency, the procedure in question will be "correct." It says nothing about the specific use of the procedure on a specific dataset, however.

Suppose there is such a procedure P with a frequentist guarantee. Then I can create a new procedure P' as follows: Let A be an alternative procedure, and let N be a (large) integer. The procedure P' will use A for the first N times this model is encountered, and will use procedure P, starting a time $N + 1$. The procedure P' has the same frequentist guarantee as does P. Note that procedure A is completely arbitrary, and N can be, say, a billion. Thus, I can do anything with the data before me, and claim a frequentist guarantee. In this sense, the frequentist guarantee is meaningless.

12.10 Final word about foundations

Debate about foundations became very fraught and personal, particularly in the 1930's. Since then, as a reaction, I believe, a modern consensus developed in which foundations are not discussed in polite company. Everybody does their thing, and nobody is to raise questions about what it means.

I think that's a mistake for the profession. For statistics to thrive, there should be no questions disallowed. We can discuss foundations without getting nasty, and we should. I feel that I would be doing my readers a disservice if I failed to give my take on foundational questions.

It is no secret that my view is that of a personalistic Bayesian. A more balanced view, if you want one, can be found in Barnett (2009). When I started, the most bothersome issues in Bayesian thought were identification and randomization. I think those have now been answered. The looming question for me is to find an adequate theory for group decision-making, given the result of Theorem 11.8.1.