CONTROVERSIES IN THE FOUNDATION OF STATISTICS (REPRINT)

Bradley Efron

Controversies in the Foundations of Statistics

by Bradley Ephron

This lively and wide-ranging article explores the philosophical battles among Bayesians, classical statisticians (frequentists), and a third group, termed the Fisherians. At this writing, no clear winner has emerged, although the frequentists may currently have the upper hand.

The article gives examples of the approach to estimation of the mean of a distribution by each camp, and some problems with each approach. One section discusses Stein's estimator more rigorously than the Scientific American article by Ephron and Morris. Ephron speculates on the future of statistical theory.

This article will give you insight regarding the fundamental problems of statistics that affect your work (in particular, as regards credibility). The bases of some common actuarial methods are still controversial.

This article is presented as part of a program of reprinting important papers on the foundations of casualty actuarial science. It is reprinted with the generous permission of the Mathematical Association of America. It originally appeared in the American Mathematical Monthly, Volume 85, Number 4, April 1978, pages 231 to 246.

CONTROVERSIES IN THE FOUNDATIONS OF STATISTICS

BRADLEY EFRON

1. Introduction. Statistics seems to be a difficult subject for mathematicians, perhaps because its elusive and wide-ranging character mitigates against the traditional theorem-proof method of presentation. It may come as some comfort then that statistics is also a difficult subject for statisticians. We are now celebrating the approximate bicentennial of a controversy concerning the basic nature of statistics. The two main factions in this philosophical battle, the Bayesians and the frequentists, have

---

Bradley Efron received his Ph.D. in Statistics from Stanford in 1964 under the direction of Rupert Miller. He holds professorships at Stanford in both the Statistics Department and the Department of Preventive Medicine. His interests cover most of theoretical and applied statistics, with special emphasis on the application of geometrical methods to statistical problems. — *Editors*

alternated dominance several times, with the frequentists currently holding an uneasy upper hand. A smaller third party, perhaps best called the Fisherians, snipes away at both sides.

Statistics, by definition, is uninterested in the special case. Averages are the meat of statisticians, where "average" here is understood in the wide sense of any summary statement about a large population of objects. "The average I.Q. of a college freshman is 109" is one such statement, as is "the probability of a fair coin falling heads is 1/2." The controversies dividing the statistical world revolve on the following basic point: just *which* averages are most relevant in drawing inferences from data? Frequentists, Bayesians, and Fisherians have produced fundamentally different answers to this question.

This article will proceed by a series of examples, rather than an axiomatic or historical exposition of the various points of view. The examples are artificially simple for the sake of humane presentation, but readers should be assured that real data are susceptible to the same disagreements. A counter-warning is also apt: these disagreements haven't crippled statistics, either theoretical or applied, and have as a matter of fact contributed to its vitality. Important recent developments, in particular the empirical Bayes methods mentioned in Section 8, have sprung directly from the tension between the Bayesian and frequentist viewpoints.

2. The normal distribution. All of our examples will involve the normal distribution, which for various reasons plays a central role in theoretical and applied statistics. A normal, or Gaussian, random variable x is a quantity which possibly can take on any value on the real axis, but not with equal probability. The probability that x falls in the interval $[a, b]$ is given by the area under Gauss' famous bell-shaped curve,

$$\text{Prob}\{a \leqq x \leqq b\} = \int_a^b \phi_{\mu,\sigma}(x)dx, \tag{2.1}$$

where

$$\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]. \tag{2.2}$$

For convenience we indicate such a random variable by

$$x \sim \mathcal{N}(\mu, \sigma^2), \tag{2.3}$$

with σ^2 instead of σ as the second argument by convention.

Figure 1 illustrates the normal distribution. The high point of $\phi_{\mu,\sigma}(x)$ is at $x = \mu$, the curve falling off quickly for $|x - \mu| > \sigma$. Most of the probability, 99.7%, is within ± 3 σ-units of the central value μ. We can write $x \sim \mathcal{N}(\mu, \sigma^2)$ as $x = \mu + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$; adding the constant μ merely shifts $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ μ units to the right.

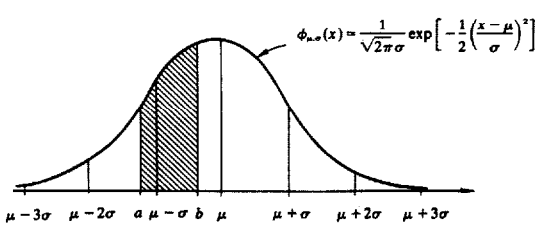

FIG. 1. The normal distribution. The random quantity $x \sim \mathcal{N}(\mu, \sigma^2)$ occurs in $[a, b]$ with probability equal to the shaded area. 68% of the probability is in the interval $[\mu - \sigma, \mu + \sigma]$, 95% in $[\mu - 2\sigma, \mu + 2\sigma]$, 99.7% in $[\mu - 3\sigma, \mu + 3\sigma]$.

The parameter μ is the "mean" or "expectation" of the random quantity x. Using "E" to indicate expectation,

$$\mu = E\{x\} = \int_{-\infty}^{\infty} x\phi_{\mu,\sigma}(x)dx. \tag{2.4}$$

The reader may wish to think of $E\{g(x)\}$ for an arbitrary function $g(x)$ as just another notation for the integral of $g(x)$ with respect to $\phi_{\mu,\sigma}(x)dx$,

$$E\{g(x)\} = \int_{-\infty}^{\infty} g(x)\phi_{\mu,\sigma}(x)dx. \tag{2.5}$$

Intuitively, $E\{g(x)\}$ is the weighted average of the possible values of $g(x)$, weighted according to the probabilities $\phi_{\mu,\sigma}(x)dx$ for the infinitesimal intervals $[x, x+dx]$. In other words, $E\{g(x)\}$ is a theoretical average of an infinite population of $g(x)$ values, where the x's occur in proportion to $\phi_{\mu,\sigma}(x)$.

It is easy to see, by symmetry, that μ is indeed the theoretical average of x itself when $x \sim \mathcal{N}(\mu, \sigma^2)$. A more difficult calculation (though easy enough for friends of the gamma function) gives the expectation of $g(x) = (x-\mu)^2$,

$$E\{(x-\mu)^2\} = \int_{-\infty}^{\infty} (x-\mu)^2\phi_{\mu,\sigma}(x)dx = \sigma^2. \tag{2.6}$$

The parameter σ, called the "standard deviation," sets the scale for the variability of x about the central value μ, as Figure 1 shows. A $\mathcal{N}(1, 10^{-6})$ random variable will have almost no perceptible variability under repeated trials, 997 out of 1000 repetitions occurring in [.997, 1.003], since $\sigma = 10^{-3}$. A $\mathcal{N}(1, 10^{6})$ random variable is almost all noise and no signal, in the evocative language of communications theory.

The normal distribution has a very useful closure property that makes it as easy to deal with many observations as with a single one. Let $x_1, x_2, x_3, \ldots, x_n$ be n independent observations, each of which is $\mathcal{N}(\mu, \sigma^2)$, μ and σ being the same for all n repetitions. Independence means that the value of x_1, say, does not affect any of the other values: observing $x_1 > \mu$ does not increase or decrease the 34% probability that $x_2 \in [\mu, \mu+\sigma]$, etc. A familiar (non-normal) example of independent variables $x_1, x_2, x_3, \ldots$ is given by successive observations of a well-rolled die.

Let

$$\bar{x} = \sum_{i=1}^{n} x_i/n \tag{2.7}$$

be the observed average of the n independent $\mathcal{N}(\mu, \sigma^2)$ variables. It is easy to show that

$$\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n). \tag{2.8}$$

The distribution of $\bar{x}$ is the same as that for the individual x_i except that the scaling parameter has been reduced from σ to $\sigma/\sqrt{n}$. By taking n sufficiently large we can reduce the variability of $\bar{x}$ about μ to an arbitrarily small level, but of course in real problems n is limited and $\bar{x}$ retains an irreducible component of random variability.

In all of our examples σ will be assumed known to the statistician. The unknown parameter μ will be the object of interest, the goal being to make inferences about the value of μ on the basis of the data $x_1, x_2, x_3, \ldots, x_n$. In 1925 Sir Ronald Fisher made the fundamental observation that in this situation *the average $\bar{x}$ contains all possible information about μ*. For any inference problem about μ, knowing $\bar{x}$ is just as good as knowing the entire data set $x_1, x_2, x_3, \ldots, x_n$. In modern parlance, $\bar{x}$ is a "sufficient statistic" for the unknown parameter μ.

It is easy to verify sufficiency in this particular case. Given the observed value of $\bar{x}$, a standard

probability calculation shows that the random quantities $x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, \ldots, x_n - \bar{x}$ have a joint distribution which does not depend in any way on the unknown parameter μ. In other words, what's left over in the data after the statistician learns $\bar{x}$ is devoid of information about μ. (This deceptively simple principle eluded both Gauss and Laplace!)

3. Frequentist estimation of the mean. The statistician may wish to estimate the unobservable parameter μ on the basis of the observed data $x_1, x_2, x_3, \ldots, x_n$. "Estimate" usually means "make a guess $\hat{\mu}(x_1, x_2, x_3, \ldots, x_n)$ depending on $x_1, x_2, \ldots, x_n$, with the understanding that you will be penalized an amount which is a smooth increasing function of the error of estimation $|\hat{\mu} - \mu|$." The usual penalty function, which we shall also use here, is $(\hat{\mu} - \mu)^2$, the squared-error loss function originally introduced by Gauss.

Fisher's sufficiency principle says that we need only consider estimation rules which are a function of $\bar{x}$. The most obvious candidate is $\bar{x}$ itself,

$$\hat{\mu}(x_1, x_2, \ldots, x_n) = \bar{x}. \tag{3.1}$$

This estimation rule is "unbiased" for μ; no matter what the true value of μ is,

$$E\bar{x} = \mu. \tag{3.2}$$

Unbiasedness is by no means a necessary condition for a good estimation rule, as we shall see later, but it does have considerable intuitive appeal as a guarantee that the statistician is not trying to slant the estimation process in favor of any particular μ value.

The expected penalty for using $\hat{\mu} = \bar{x}$ is, according to (2.6) and (2.8),

$$E(\hat{\mu} - \mu)^2 = \sigma^2/n. \tag{3.3}$$

Gauss showed that among all unbiased estimation rules $\hat{\mu}(x_1, x_2, \ldots, x_n)$ which are linear in $x_1, x_2, x_3, \ldots, x_n$, the rule $\hat{\mu} = \bar{x}$ uniformly minimizes $E(\hat{\mu} - \mu)^2$ for every value of μ. In the early 1940's this result was extended to include any unbiased estimator at all, linear or nonlinear. The proof, which depends on ideas Fisher developed in the 1920's, was put forth separately by H. Cramér in Sweden and C. R. Rao in India.

If we agree to abide by the unbiasedness criterion and to use squared-error loss, $\bar{x}$ seems to be the best estimator for μ. It is helpful for the statistician to provide not only a "point estimator" for μ, $\bar{x}$ in this case, but also a range of plausible values of μ consistent with the data. From (2.8) and Figure 1 we see that

$$\text{Prob}\{|\bar{x} - \mu| \leqq 2\sigma/\sqrt{n}\} = .95, \tag{3.4}$$

which is equivalent to the statement

$$\text{Prob}\{\bar{x} - 2\sigma/\sqrt{n} \leqq \mu \leqq \bar{x} + 2\sigma/\sqrt{n}\} = .95. \tag{3.5}$$

The interval $[\bar{x} - 2\sigma/\sqrt{n}, x + 2\sigma/\sqrt{n}]$ is called a "95% confidence interval" for μ. The theory of confidence intervals was developed by J. Neyman in the early 1930's. As an example, suppose $n = 4$, $\sigma = 1$, and we observe $x_1 = 1.2$, $x_2 = 0.3$, $x_3 = 0.7$, $x_4 = 0.2$. Then $\bar{x} = 0.6$ and the 95% confidence interval for μ is $[-.04, 1.6]$.

All of this seems so innocuous and straightforward that the reader may wonder where the grounds for controversy lie. The fact is that all of the results presented so far are "frequentist" in nature. That is, they relate to theoretical averages with respect to the $\mathcal{N}(\mu, \sigma^2/n)$ distribution of $\bar{x}$, with μ assumed fixed at its true value, whatever that may be. Unbiasedness itself is a frequentist concept; the theoretical average of $\hat{\mu}$ with μ held fixed, $E\hat{\mu}$, equals μ. Results (3.3) and (3.5), and the Cramér–Rao theorem, are frequentist statements. For example, the proper interpretation of (3.5) is that the interval $[\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n}]$ covers the true value of μ with frequency 95% in a long series of independent repetitions of $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$.

Nobody doubts that these results are true. The question raised by Bayesians and Fisherians is whether frequentist averages are really relevant to the process of inference scientists use in reasoning from noisy data back to the underlying mathematical models. We turn next to the Bayesian point of view.

4. Bayesian estimation of the mean. So far we have considered μ to be a fixed, albeit unknown, quantity. Suppose though that μ itself is a random variable, known to have the normal distribution with mean m and standard deviation s,

$$\mu \sim \mathcal{N}(m, s^2), \tag{4.1}$$

m and s being constants known to the statistician. For example, if μ is the true I.Q. of a person randomly chosen from the population of the United States, (4.1) holds with $m = 100$ and $s = 15$ (approximately). About 68% of I.Q.'s are between 85 and 115, about 95% between 70 and 130, etc. Information like (4.1), a "prior distribution for μ" in the language of the Bayesians, changes the nature of the estimation process.

Standard I.Q. tests are constructed so that if we test our randomly chosen person to discover his particular μ value, the overall test score*, say $\bar{x}$, is an unbiased normally distributed estimator of μ as in Section 3,

$$\bar{x} \mid \mu \sim \mathcal{N}(\mu, \sigma^2/n), \tag{4.2}$$

with $\sigma/\sqrt{n}$ about 7.5. We can expect $\bar{x}$ to be within 7.5 I.Q. points of μ 68% of the time, etc. The notation "$\bar{x} \mid \mu$" emphasizes that the $\mathcal{N}(\mu, \sigma^2/n)$ distribution for $\bar{x}$ is *conditional* on the particular value taken by the random quantity μ. The reason for this change in notation will be made clearer soon.

Bayes' theorem, originally discovered by the remarkable Reverend Thomas Bayes around 1750, is a mathematical formula for combining (4.1) and (4.2) to obtain the conditional distribution of μ given $\bar{x}$. In this case the formula gives

$$\mu \mid \bar{x} \sim \mathcal{N}(m + C(\bar{x} - m), D), \tag{4.3}$$

where

$$C = \frac{n/\sigma^2}{1/s^2 + n/\sigma^2} \quad \text{and} \quad D = \frac{1}{1/s^2 + n/\sigma^2}. \tag{4.4}$$

For example, if $\bar{x} = 160$ (and $m = 100$, $s = 15$, $\sigma/\sqrt{n} = 7.5$) then

$$\mu \mid \bar{x} \sim \mathcal{N}(148, (6.7)^2). \tag{4.5}$$

Expression (4.5), or more generally (4.3), is the "posterior distribution for μ given the observed value of $\bar{x}$." It is possible to make such a statement in the Bayesian framework because we start out assuming that μ itself is random. In the Bayesian framework the averaging process is reversed; the data $\bar{x}$ is assumed fixed at its observed value while it is the parameter μ which varies. In (4.5) for example, the conditional average of μ given $\bar{x} = 160$ is seen to be 148. If we randomly selected an enormous number of people, gave them each an I.Q. test, and considered the subset of those who scored 160, this subset would have an average true I.Q. of 148; 68% of the true I.Q.'s would be in the interval $[148 - 6.7, 148 + 6.7]$, etc.

How should we estimate μ in the Bayesian situation? It seems natural to use the estimator $\mu^*(\bar{x})$ which minimizes the conditional expectation of $(\mu - \mu^*)^2$ *given the observed value of* $\bar{x}$. From (4.3) it is

* The symbols $\bar{x}$ for the test score and $\sigma/\sqrt{n}$ for its standard deviation are chosen to agree with our previous notation, even though real I.Q. scores aren't actually the average of n independent test items. Perfect normality, as expressed in (4.2), is an ideal only approximated by actual test scores.

easy to derive that this "Bayes estimator" is

$$\mu^*(\bar{x}) = m + C(\bar{x} - m), \tag{4.6}$$

the mean of the posterior distribution of μ given $\bar{x}$. Having observed $\bar{x} = 160$, the Bayes estimate is 148, not 160. Even though we are using an unbiased I.Q. test, so many more true I.Q.'s lie below 160 rather than above that it lowers the expected estimation error to bias the observed score toward 100. Figure 2 illustrates the situation.

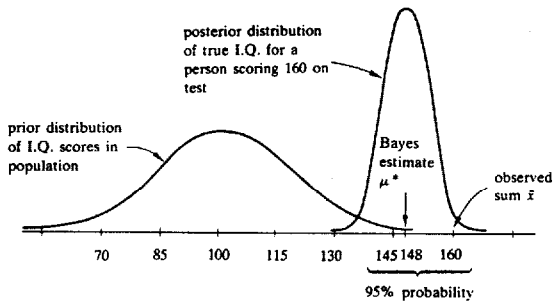

FIG. 2. I.Q. scores have a $\mathcal{N}(100, (15)^2)$ distribution in the population as a whole. A randomly selected person scoring 160 on a normal unbiased I.Q. test with standard deviation 7.5 points is estimated to have a true I.Q. of 148. The probability is 95% that the person's true I.Q. is in the interval [134.6, 161.4].

Confidence intervals have an obvious Bayesian analogue, from (4.3),

$$\text{Prob}\{\mu^*(\bar{x}) - 2\sqrt{D} \leqq \mu \leqq \mu^*(\bar{x}) + 2\sqrt{D} \mid \bar{x}\} = .95. \tag{4.7}$$

The notation $\text{Prob}\{\cdot \mid \bar{x}\}$ indicates probability conditional on the observed value of $\bar{x}$. In the I.Q. example, $\text{Prob}\{134.6 \leqq \mu \leqq 161.8 \mid \bar{x} = 160\} = .95$.

Nobody (well, almost nobody) disagrees with the use of Bayesian methods in situations like the I.Q. problem where there is a clearly defined and well-known prior distribution for μ. The Bayes theory, as we shall see, offers some striking advantages in clarity and consistency. These advantages are due to the fact that Bayesian averages involve only the data value $\bar{x}$ actually seen, rather than a collection of theoretically possible other $\bar{x}$ values.

Difficulties and controversies arise because Bayesian statisticians wish to use Bayesian methods when there is no obvious prior distribution for μ, or going even further, when it is clear that the unknown μ is a fixed constant with no random character at all. (For example, if μ is some physical constant, such as the speed of light, being experimentally estimated.) It is not perversity that motivates this Bayesian impulse, but rather a well-documented casebook of unpleasant inconsistencies in the frequentist approach.

As an example of the kind of difficulties frequentists experience, let us reconsider the I.Q. estimation problem, but without assuming knowledge of the prior distribution (4.1) for μ. In other words, assume only that we observe $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$, $\sigma/\sqrt{n} = 7.5$, and wish to estimate μ. Having observed $\bar{x} = 160$, the results of Section 3 tell us to estimate μ by $\hat{\mu} = 160$, with 95% confidence interval $[\hat{\mu} - 2\sigma/\sqrt{n}, \hat{\mu} + 2\sigma/\sqrt{n}] = [145, 175]$.

Suppose now that the frequentist receives a letter from the company which administered the I.Q. test: "On the day the score of $\bar{x} = 160$ was reported, our test-grading machine was malfunctioning. Any score $\bar{x}$ below 100 was reported as 100. The machine functioned perfectly for scores $\bar{x}$ above 100."

It may seem that the frequentist has nothing to worry about, since the score he received, $\bar{x} = 160$, was correctly reported. However, the reason he is using $\hat{\mu} = \bar{x}$ to estimate μ is that it is the best unbiased estimator. The malfunction of the grading machine implies that $\hat{\mu}$ is no longer even unbiased!

If the true value of μ equals 100, the machine functioning as described in the letter produces $E\bar{x} = 103$, a bias of $+3$ points. To regain unbiasedness the frequentist must replace the estimation rule $\hat{\mu} = \bar{x}$ with $\hat{\mu}' = \bar{x} - \Delta(\bar{x})$, where the function $\Delta(\bar{x})$ is chosen to remove the bias caused by the machine malfunction.

The correction term $\Delta(\bar{x})$ will be tiny for $\bar{x} = 160$, but it is disturbing that any change at all is necessary. The letter from the grading company contained no new information about the score actually reported, or about I.Q.'s in general. It only concerned something bad that might have happened but didn't. Why should we change our inference about the true value of μ? Bayesian methods are free from this defect; the inferences they produce depend only on the data value $\bar{x}$ actually observed, since Bayesian averages such as (4.6), (4.7) are conditional on the observed $\bar{x}$.

How can a Bayesian analysis proceed in the absence of firm prior knowledge like (4.1)? Two different approaches are in use. The "subjectivist" branch of Bayesian statistics attempts to assess the statistician's subjective probability distribution for the unknown parameter μ, before the data is collected, by a series of hypothetical wagers. These wagers are of the form "would you be willing to bet even money that $\mu > 85$ versus $\mu \leqq 85$? Would you be willing to bet two-to-one that $\mu < 150$ versus $\mu \geqq 150$?..." The work of L. J. Savage and B. deFinetti shows that a completely rational person should always be able to arrive at a unique (for himself) prior distribution on μ by sufficiently prolonged self-interrogation.

The subjectivist approach can be very fruitful in cases where the statistician (usually in collaboration with the experimenter, of course) has some vague prior opinions about the true value of μ, which he is trying to update on the basis of the observed data $\bar{x}$. Because it is subjective, the method is not much used where objectivity is the prime consideration, for example in the publication of controversial new scientific results.

Another line of Bayesian thought, which might be (but usually isn't) called "objective Bayesianism," attempts, in the absence of prior knowledge, to produce a prior distribution that everyone would agree represents a completely neutral prior opinion about μ. In the I.Q. problem, such a "flat" prior might take the form $\mu \sim \mathcal{N}(0, \infty)$, whereby we mean $\mu \sim \mathcal{N}(0, s^2)$ with s^2 going to infinity. From (4.3), (4.4) we get

$$\mu \mid \bar{x} \sim \mathcal{N}(\bar{x}, \sigma^2/n). \tag{4.8}$$

This result has a lot of appeal. The Bayes estimator μ^* equals the frequentist estimator $\hat{\mu} = \bar{x}$. The 95% Bayes probability interval (4.7) is the same as the 95% frequentist confidence interval (3.5). Moreover, because (4.8) is a Bayesian statement, the letter from the I.Q. testing company has no effect on it. We seem to be enjoying the best of both the frequentist and Bayesian worlds.

An enormous amount of effort has been expended in codifying the objective Bayesian point of view. Bayes himself put forth this approach (apparently with considerable reservations—his paper appeared posthumously and only through the efforts of an enthusiastic friend) which was adopted unreservedly by Laplace. It fell into disrepute in the early 1900's, and has since been somewhat revived by the work of Harold Jeffreys. One difficulty is that a "flat" prior distribution for μ is not at all flat for μ^3, say, so expressing ignorance seems to depend on which function of the unknown parameter one is interested in. A more pernicious difficulty is discussed in Section 8; in problems involving the estimation of several unknown parameters at once, what appears to be an eminently neutral prior distribution turns out to imply undesirable assumptions about the parameters.

5. Fisherian estimation of the mean. Ronald Fisher was one of the principal architects of frequentist theory. However, he was a lifelong critic, often vehemently so, of the standard frequentist

approach. His criticisms moved along the same lines as those of the Bayesians: why should we be interested in theoretical averages concerning what happens if infinitely many $\bar{x}$ values are randomly generated from $\mathcal{N}(\mu, \sigma^2/n)$, with μ fixed? We only have one observed value of $\bar{x}$ in any one inference problem, and the inference process should concentrate on just that observed value.

Fisher was also opposed to the Bayesian approach, perhaps because the type of data analysis problems he met in his agricultural and genetical work were not well suited to the assessment of prior distributions. With characteristic ingenuity he produced another form of inference, neither Bayesian nor frequentist.

The relation $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$ may be written

$$\bar{x} = \mu + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2/n). \tag{5.1}$$

We obtain the observation $\bar{x}$ by adding normal noise, $\varepsilon \sim \mathcal{N}(0, \sigma^2/n)$, to the unobservable mean μ. Expression (5.1) can also be written as

$$\mu = \bar{x} - \varepsilon. \tag{5.2}$$

It is obvious, or at least was obvious to Fisher, that in a situation where we know nothing a priori about μ, observing $\bar{x}$ tells us nothing about ε. As a matter of fact, said Fisher, if we can learn something about ε from $\bar{x}$ then model (5.1) by itself must be missing some important aspect of the statistical situation. We shall see this argument again, in more concrete form, in the next section.

If $\varepsilon \sim \mathcal{N}(0, \sigma^2/n)$ then $-\varepsilon \sim \mathcal{N}(0, \sigma^2/n)$ because of the symmetry of the bell-shaped curve about its central point. Fisher's interpretation of (5.2) was

$$\mu \,|\, \bar{x} \sim \mathcal{N}(\bar{x}, \sigma^2/n). \tag{5.3}$$

This looks just like the objectivist Bayesian statement (4.8), but has been obtained without recourse to prior distributions on μ. The interval statement following from (3.3) is

$$\text{Prob}\{\bar{x} - 2\sigma/\sqrt{n} \leqq \mu \leqq \bar{x} + 2\sigma/\sqrt{n} \,|\, \bar{x}\} = .95. \tag{5.4}$$

This is a "fiducial" probability statement, in Fisher's terminology.

In the fiducial argument randomness resides neither in the data $\bar{x}$, as in frequentist calculations, nor in μ, as in Bayesian calculations. Rather it lies in the mechanism which transforms the unobservable μ to the observed $\bar{x}$. (In the case at hand, this mechanism is the addition of $\varepsilon \sim \mathcal{N}(0, \sigma^2/n)$ to μ.) Fiducial statements such as (5.4) are obtained as averages over the random transformation mechanism.

The fiducial argument has fallen out of favor since its heyday in the 1940's. Most, though not all, contemporary statisticians consider it either a form of objective Bayesianism, or just plain wrong. Applied to the simultaneous estimation of several parameters, the fiducial argument can lead to disaster, as shown in Section 8.

Lest the reader feel sorry for Fisher, two other of his novel ideas on averaging, conditional inference and randomization, are still very much in vogue, and are the subjects of the next two sections.

6. Conditional inference. We return to the frequentist point of view, but with a twist, "conditioning," introduced by Fisher in 1934. Conditional inference illustrates another major source of ambiguity in the frequentist methodology, the choice of the collection of theoretically possible data values averaged over to obtain a frequentist inference.

Suppose again that we have independent normal variables $x_1, x_2, x_3, \ldots, x_n$, each $x_i \sim \mathcal{N}(\mu, \sigma^2)$, but that before observation begins the number n is randomly selected by the flip of a fair coin,

$$n = \begin{cases} 10 & \\ & \text{with probability} \\ 100 & \end{cases} \begin{matrix} 1/2 \\ \\ 1/2. \end{matrix} \tag{6.1}$$

We still wish to estimate μ on the basis of the data $x_1, x_2, x_3, \ldots, x_n$, and n with σ a known constant as before.

The conditional distribution of $\bar{x}$ *given the observed value of* n is

$$\bar{x} \mid n \sim \mathcal{N}(\mu, \sigma^2/n) \tag{6.2}$$

as at (2.8). The observed average $\bar{x}$ by itself is not a sufficient statistic in this situation. We also need to know whether n equals 10 or 100. Without this knowledge we still have an unbiased estimator of μ, namely $\hat{\mu} = \bar{x}$, but we don't know the standard deviation of $\hat{\mu}$.

What is the expected squared error of $\hat{\mu} = \bar{x}$ in this situation? Averaging (3.3) over the two values of n gives

$$E(\hat{\mu} - \mu)^2 = \frac{1}{2}\frac{\sigma^2}{10} + \frac{1}{2}\frac{\sigma^2}{100}. \tag{6.3}$$

Fisher pointed out that this is a ridiculous calculation. It is obviously more appropriate to assess the accuracy of $\hat{\mu}$ conditional on the value of n actually observed,

$$E\{(\hat{\mu} - \mu)^2 \mid n\} = \begin{cases} \sigma^2/10 & \text{if } n = 10 \\ \sigma^2/100 & \text{if } n = 100. \end{cases} \tag{6.4}$$

There is nothing wrong with (6.3), except that the average squared error it computes is irrelevant to any particular value of n and $\bar{x}$ actually observed! If $n = 100$ then (6.3) is much too pessimistic about the accuracy of $\hat{\mu}$, while if $n = 10$ it is much too optimistic.

This may all seem so obvious that it is hardly worth saying. Fisher's surprise was to show that exactly the same situation arises, more subtly, in other problems of statistical inference. We will illustrate this with an example involving the estimation of two different normal means, say μ_1 and μ_2, on the basis of independent unbiased normal estimates for each of them,

$$\bar{x}_1 \sim \mathcal{N}(\mu_1, 1), \qquad \bar{x}_2 \sim \mathcal{N}(\mu_2, 1), \tag{6.5}$$

$\bar{x}_1$ and $\bar{x}_2$ independent of each other. (For simplicity we have assumed that both estimates have $\sigma^2/n = 1$.) The two dimensional data vector $(\bar{x}_1, \bar{x}_2)$ can take on any value in the plane, but with high probability lies no more than a few units away from the vector of means (μ_1, μ_2).

Given no further information we would probably estimate (μ_1, μ_2) by $(\bar{x}_1, \bar{x}_2)$. (But see Section 8!) However, we now add the assumption that (μ_1, μ_2) is known to lie on the circle of radius 3 centered at the origin,

$$(\mu_1, \mu_2) = 3(\cos\theta, \sin\theta) \qquad -\pi < \theta \leqq \pi. \tag{6.6}$$

The statistical problem, as illustrated in Figure 3, is to estimate the unknown parameter θ on the basis of $(\bar{x}_1, \bar{x}_2)$.

Let us indicate the polar coordinates of $(\bar{x}_1, \bar{x}_2)$ by

$$\hat{\theta} = \arctan(\bar{x}_2/\bar{x}_1), \qquad r = \sqrt{\bar{x}_1^2 + \bar{x}_2^2}. \tag{6.7}$$

Then $\hat{\theta}$ is the obvious estimator of θ. It is unbiased, $E\hat{\theta} = \theta$, with expected squared error

$$E(\hat{\theta} - \theta)^2 = .12 \tag{6.8}$$

(obtained by numerical integration; (6.8) makes the convention that $\hat{\theta} - \theta$ ranges from $-\pi$ to π for any value of θ, the largest possible estimation error occurring if $(\bar{x}_1, \bar{x}_2)$ is antipodal to (μ_1, μ_2). This convention is unimportant because the probability of $|\hat{\theta} - \theta| > \pi/2$ is only .0014).

The unobvious fact pointed out by Fisher is that r plays the same role as did "n" in examples (6.1)–(6.4).

(i) The distribution of r does not depend on the true value of θ. (For readers familiar with the

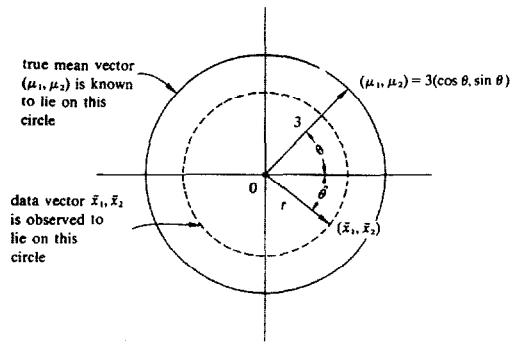

FIG. 3. The model $\bar{x}_1 \sim \mathcal{N}(\mu_1, 1)$ independent of $\bar{x}_2 \sim \mathcal{N}(\mu_2, 1)$, with (μ_1, μ_2) known to lie on a circle of radius 3 centered at the origin. We wish to estimate the angular location θ of (μ_1, μ_2) on the circle. The data vector $(\bar{x}_1, \bar{x}_2)$ is observed to have polar coordinates $(\hat{\theta}, r)$.

bivariate normal density, this follows from the circular symmetry of the distribution (6.5) of $(\bar{x}_1, \bar{x}_2)$ about (μ_1, μ_2).)

(ii) If r is small, then $\hat{\theta}$ has less accuracy than (6.8) indicates, while if r is large then $\hat{\theta}$ has greater accuracy that (6.8) indicates. Table 1 shows the conditional expected squared error $E\{(\hat{\theta} - \theta)^2 | r\}$ as a function of r.

In Fisher's terminology, r is an "ancillary" statistic. It doesn't directly contain information about θ, because of property (i), but its value determines the accuracy of $\hat{\theta}$. It now seems obvious that we should condition our assessment of the accuracy of $\hat{\theta}$ on the observed value of r. If $r = 2$, as in Figure 3, then $E\{(\hat{\theta} - \theta)^2 | r\} = .18$ is more relevant to the accuracy of $\hat{\theta}$ than is the unconditional expectation $E(\hat{\theta} - \theta)^2 = .12$.

r	1.5	2	2.5	3	3.5	4	4.5	5	Unconditional Value $E(\hat{\theta} - \theta)^2$
$E\{(\hat{\theta} - \theta)^2 \mid r\}$	.26	.18	.14	.12	.10	.09	.08	.07	.12

TABLE 1. The conditional expected squared error of estimation in the circle problem, $E\{(\hat{\theta} - \theta)^2 | r\}$, as a function of the ancillary statistic $r = \sqrt{\bar{x}_1^2 + \bar{x}_2^2}$. The accuracy of $\hat{\theta}$ improves as r increases. Fisher argued that $E\{(\hat{\theta} - \theta)^2 | r\}$ is a more relevant measure of the accuracy of $\hat{\theta}$ than is the unconditional expectation $E(\hat{\theta} - \theta)^2$.

Many real statistical problems have the property that some data values are obviously more informative than others. Conditioning is the intuitively correct way to proceed, but few situations are as clearly structured as the circle problem. Sometimes more than one ancillary statistic exists, and the same data value will yield different accuracy estimates depending on which ancillary is conditioned upon. More often no ancillary exists, but various approximate ancillary statistics suggest themselves. What the circle example reveals is that frequentist statements like (6.8) may be true but irrelevant. Fisher's point was that the theoretical average of $(\hat{\theta} - \theta)^2$ should be taken not over all possible data values, but only over those containing the same amount of information for θ. So far it has proved impossible to codify this statement in a satisfactory way.

A Bayesian would agree that it is correct to condition one's opinion of the accuracy of $\hat{\theta}$ on the

observed value of r, but would ask why not go further and condition on the observed value of $(\bar{x}_1, \bar{x}_2)$ itself. This is impossible in the frequentist framework, since if we reduce our averaging set to one data point, there is nothing left to average over. Bayesian inferences are always conditional on the data point actually observed. In the circle problem the natural flat prior is a uniform distribution on $\theta \in [-\pi, \pi]$. With this prior distribution it turns out that $E\{(\theta - \hat{\theta})^2 | (\bar{x}_1, \bar{x}_2)\}$ equals $E\{(\hat{\theta} - \theta)^2 | r = \sqrt{\bar{x}_1^2 + \bar{x}_2^2}\}$ as given in Table 1, so in this particular case the objective Bayesian and conditional frequentist points of view agree. (Notice that in the first expectation "θ" is the random quantity, while in the second it is "$\hat{\theta}$" which varies.)

7. Randomization. Randomization is yet another form of inferential averaging introduced by R. A. Fisher. In order to discuss it simply we must change statistical problems, from estimation theory to "hypothesis testing." The data are now in the form of $2n$ independent normal observations $x_1, x_2, x_3, \ldots, x_n, y_1, y_2, y_3, \ldots, y_n$,

$$x_i \sim \mathcal{N}(\mu_1, \sigma^2), \qquad y_i \sim \mathcal{N}(\mu_2, \sigma^2) \qquad i = 1, 2, \ldots, n, \tag{7.1}$$

with σ known, μ_1 and μ_2 unknown. We wish to test the "null hypothesis" that $\mu_2 = \mu_1$ versus the "alternative hypothesis" that $\mu_2 > \mu_1$, often written

$$H: \mu_2 = \mu_1 \quad \text{versus} \quad A: \mu_2 > \mu_1. \tag{7.2}$$

(For our purposes, $\mu_2 < \mu_1$ is assumed impossible.)

In hypothesis testing the null hypothesis H usually plays the role of a devil's advocate which the experimenter is trying to disprove. For example, the x's may represent responses to an old drug and the y's responses to a new drug that the experimenter hopes is an improvement. Because there is a vested interest in discrediting H, conservative statistical methods have been developed which demand a rather stiff level of evidence before H is declared invalid. The frequentist theory, which is dominant in hypothesis testing, accomplishes this by requiring that the probability of falsely rejecting H in favor of A, when H is true, be held below a certain small level, usually .05. A test satisfying this criterion is said to be ".05 level" for testing H versus A.

With the data as in (7.1) it seems natural to compute $\bar{x} = \Sigma_1^n x_i / n$, $\bar{y} = \Sigma_1^n y_i / n$, and reject H in favor of A if

$$\bar{y} - \bar{x} > c. \tag{7.3}$$

The constant c is chosen so that if H is true then $\text{Prob}\{\bar{y} - \bar{x} > c\} = .05$. Standard probability calculations show that $c = 2.326 \cdot \sigma / \sqrt{n}$ is the correct choice. The theory of optimal testing developed by J. Neyman and E. Pearson around 1930 shows that (7.3) is actually the best .05 level test of H versus A, in the sense that if A is actually true then the probability of rejecting H in favor of A is maximized.

The x's and y's we observe are actually measurements on some sort of experimental units, perhaps college freshmen or white mice or headache victims. Let us denote these units by $U_1, U_2, U_3, \ldots, U_{2n}$. The opportunity for randomization arises when we have an experiment in which we can decide beforehand which n of the units are to be x's, and which n are to be y's. If we are lazy we can just give the first n units we happen to have at hand the x treatment and the last n the y treatment. This is begging for disaster! The first n headache victims may be those with the worst headaches, the first n mice those in the cage with the heavier animals, etc. An experiment done in the lazy way may have probability of falsely rejecting the null hypothesis much greater than .05 because of such uncontrolled factors.

In his vastly influential work on experimental design, Fisher argued that the choice of experimental units be done by randomization. That is, the assignment of the n units to the x treatment group and the n units to the y treatment group be done with equal probability for each of the $(2n!)/(n!)^2$ such assignments. A random number generating device is used to carry out the randomization process.

Fisher pointed out that randomized studies were likely to be free of the type of experimental biases discussed above. Suppose for example that there is some sort of "covariate" connected with the experimental units, by which we mean a quantity which is thought to affect the observation on that unit no matter which treatment is given. For example, weight might be an important covariate for the white mice. Heavy mice might respond less well to the stimulus than light mice. If n is reasonably large, say 10, it is very unlikely that the randomized experiment will have all the heavy mice in the x group and the light mice in the y group. This statement applies equally to every covariate, whether or not we know it affects the response, and even if we are unaware of its existence.

None of this has anything to do with averaging. The connection comes through Fisher's next suggestion: that we compute theoretical averages not over the hypothesized normal distributions, but instead over the randomization process itself. Suppose that if all $2n$ experimental units had received treatment x, the observations would have been $X_1, X_2, \ldots, X_{2n}$, X_i being the observation on unit U_i. The capital letters indicate that these are hypothetical observations and not necessarily the observed data. Under the null hypothesis H, treatment y is the same as treatment x, so we can indeed consider all $2n$ units to have received treatment x. In this case the observed data $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_n$ coincide with the theoretical values $X_1, X_2, \ldots, X_{2n}$. Let $\mathcal{S}(x)$ be the indices of those units actually assigned to the x treatment and $\mathcal{S}(y)$ those assigned to the y treatment. Then, *if H is true*,

$$\bar{x} = \sum_{i \in \mathcal{S}(x)} X_i / n, \qquad \bar{y} = \sum_{i \in \mathcal{S}(y)} X_i / n. \tag{7.4}$$

If the study has been randomized then $\bar{x}$ is merely the average of n randomly selected X's and $\bar{y}$ the average of the remaining n X's.

The randomization (or "permutation") test of H analogous to (7.3) is constructed as follows:

(i) Given the observed data $x_1, x_2, \ldots, x_n$, $y_1, y_2, \ldots, y_n$, define $u_1 \equiv x_1$, $u_2 \equiv x_2, \ldots, u_{n+1} \equiv y_1, \ldots, u_{2n} \equiv y_n$. (Notice that, if H is true, the u's coincide with the X's of the previous paragraph.)

(ii) For each partition $\mathcal{P} \equiv \{\mathcal{S}_1, \mathcal{S}_2\}$ of $\{1, 2, \ldots, 2n\}$ into two disjoint subsets of size n, calculate

$$(\bar{y} - \bar{x})_{\mathcal{P}} \equiv \sum_{i \in \mathcal{S}_2} u_i / n - \sum_{i \in \mathcal{S}_1} u_i / n. \tag{7.5}$$

(iii) List all $(2n!)/(n!)^2$ values of $(\bar{y} - \bar{x})_{\mathcal{P}}$ in ascending order.

(iv) Reject H in favor of A if the value of $\bar{y} - \bar{x}$ actually observed is in the upper 5% of the list.

The randomization test has a .05 chance of falsely rejecting H, *where the probability .05 now refers to an average taken over all $(2n!)/(n!)^2$ random assignments of treatment types to experimental units.* The test is still of the form "reject H in favor of A if $\bar{y} - \bar{x} > c$," except that c no longer equals the constant $2.326 \cdot \sigma / \sqrt{n}$. Instead c is a function of the set of values $\{u_1, u_2, \ldots, u_{2n}\}$ constructed in (i). For each set $\{u_1, u_2, \ldots, u_{2n}\}$, c is selected to satisfy (iv).

The randomization test has one big advantage over test (7.3). Its .05 probability of falsely rejecting H remains valid under any null hypothesis that says the $2n$ x's and y's are generated by the same probability distribution, normal or otherwise. As a matter of fact, no randomness at all in the observations need be assumed. We can just take the null hypothesis to be that each unit U_i has a fixed response X_i connected with it, no matter whether it is given the x or y treatment. This last statement reemphasizes that the randomization test must involve a non-frequentist form of averaging.

Randomization, or at least inference based on randomization, appears heretical to a Bayesian statistician. The true Bayesian must condition on the assignment $\{\mathcal{S}(x), \mathcal{S}(y)\}$ of units to treatments actually used, since this is part of the available data, and not average over all possible partitions that might have been. (Fisher's arguments on ancillarity seem to point in exactly the same direction, which is to say directly opposite to randomization!)

One aspect of randomization makes both frequentists and Bayesians uneasy. Suppose, just by bad luck, that the randomization process does happen to assign all heavy mice to the x treatment and all light mice to the y treatment. Can we still use the .05 level randomization test to reject H in favor of

A? The answer seems clearly not, but it is difficult to codify a way of avoiding such traps. To put things the other way, suppose we know the weights $w_1, w_2, w_3, \ldots, w_{2n}$ of the mice before we begin the experiment. Under reasonable frequentist assumptions there will be a unique best way $\{\mathcal{S}(x), \mathcal{S}(y)\}$ of assigning the mice to the treatments for the purpose of testing treatment x versus treatment y, one that optimally equalizes the weight assignments to the two groups. Statisticians trained in the Fisherian tradition find it difficult to accept such "optimal experimental designs" because the element of randomization has been eliminated.

8. Stein's Phenomenon. The reader may have noticed that the controversies so far have been more academic than practical. All philosophical factions agree that in the absence of prior knowledge $[\bar{x} - 2 \cdot \sigma/\sqrt{n}, \bar{x} + 2 \cdot \sigma/\sqrt{n}]$ is a 95% interval for μ, the disagreement being over what "95%" means. This situation changes, for the worse, when we consider the simultaneous estimation of many parameters.

Suppose then that we have several normal means $\mu_1, \mu_2, \ldots, \mu_k$ to estimate, for each one of which we observe an independent, unbiased normal estimate

$$\bar{x}_i \sim \mathcal{N}(\mu_i, 1) \quad \text{independently} \quad i = 1, 2, \ldots, k. \tag{8.1}$$

(Once again we have taken the variance σ^2/n equal to 1 for the sake of convenience.) The natural analogue of squared error loss when there are several parameters to estimate is Euclidean squared distance. To simplify notation, let $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$ be the vector of observed averages, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ the vector of true means, and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)$ the vector of estimates. Then the squared error misestimation penalty is

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{k} (\hat{\mu}_i - \mu_i)^2. \tag{8.2}$$

Before pursuing the problem of estimating $\boldsymbol{\mu}$ on the basis of $\mathbf{x}$, we note an elementary but important fact. This fact, which can be proved in one line by readers familiar with the multivariate normal distribution, is that for every parameter vector $\boldsymbol{\mu}$ we have

$$\text{Prob}\{\|\bar{\mathbf{x}}\| > \|\boldsymbol{\mu}\|\} > .50. \tag{8.3}$$

That is, the data vector $\bar{\mathbf{x}}$ tends to be farther away from the origin than does the parameter vector $\boldsymbol{\mu}$, no matter what $\boldsymbol{\mu}$ is. Table 2 shows that for $k = 10$ the probability is actually quite a bit greater than .50 for moderate values of $\|\boldsymbol{\mu}\|$.

Suppose that $k = 10$, and we observe a data vector $\bar{\mathbf{x}}$ with squared length $\|\bar{\mathbf{x}}\|^2 = 12$. Assume also that we have no prior knowledge about $\boldsymbol{\mu}$. Looking at Table 2, it seems to be a very good bet that $\|\boldsymbol{\mu}\|^2 < 12$. For $\|\boldsymbol{\mu}\|^2$ in the range $[0, 40]$, which is almost certainly the case if $\|\mathbf{x}\|^2 = 12$, more than 75% of the time we have $\|\mathbf{x}\| > \|\boldsymbol{\mu}\|$. However, this is a frequentist "75%," calculated with $\boldsymbol{\mu}$ fixed and $\bar{\mathbf{x}}$ varying randomly according to (8.1). The analogue of the objective Bayesian argument presented in Section 4 gives quite different results.

$\|\boldsymbol{\mu}\|^2$	0	6	12	18	24	30	40	60
$\text{Prob}\{\|\bar{\mathbf{x}}\| > \|\boldsymbol{\mu}\|\}$	1.00	.967	.904	.857	.822	.795	.762	.719

TABLE 2. The probability that $\|\bar{\mathbf{x}}\| \geq \|\boldsymbol{\mu}\|$ is always greater than .5. For the case $k = 10$ the probabilities are much greater than .5 for moderate values of $\|\boldsymbol{\mu}\|$.

Given our complete prior ignorance about the parameter vector $\boldsymbol{\mu}$, it seems natural to use a flat prior of the form $\mu_i \sim \mathcal{N}(0, \infty)$ (that is, $\mu_i \sim \mathcal{N}(0, s^2)$ with $s^2 \to \infty$) independently for $i = 1, 2, \ldots, k$. This leads to the posterior distribution (4.8) for each parameter μ_i,

$$\mu_i \mid \bar{x}_i \sim \mathcal{N}(\bar{x}_i, 1) \tag{8.4}$$

independently for $i = 1, 2, \ldots, k$. This of course is a Bayesian statement, with the $\bar{x}_i$'s fixed at their observed values and the μ_i's varying randomly according to (8.4). Reversing the names of the fixed and random quantities in Table 2 gives

$$\text{Prob}\{\|\boldsymbol{\mu}\| > \|\bar{\mathbf{x}}\| \,|\, \|\bar{\mathbf{x}}\|^2 = 12\} = .904. \tag{8.5}$$

It now seems to be a very good bet that $\|\boldsymbol{\mu}\| > \|\mathbf{x}\|$. As a matter of fact,

$$\text{Prob}\{\|\boldsymbol{\mu}\| > \|\bar{\mathbf{x}}\| \,|\, \bar{\mathbf{x}}\} > .50 \tag{8.6}$$

for every observed data vector $\bar{\mathbf{x}}$! Fisher's fiducial argument of Section 5 also leads to (8.4)–(8.6).

Equations (8.3) and (8.6) show a clear contradiction between the frequentist and Bayesian points of view. Which is correct? There is a most surprising and persuasive argument in favor of the frequentist calculation (8.3). This was provided by Charles Stein in the mid 1950's and concerns the estimation of $\boldsymbol{\mu}$ on the basis of the data vector $\bar{\mathbf{x}}$ (or equivalently the estimation of the parameters $\mu_1, \mu_2, \ldots, \mu_k$ on the basis of $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$).

The obvious estimator is

$$\hat{\boldsymbol{\mu}}(\bar{\mathbf{x}}) = \bar{\mathbf{x}}, \tag{8.7}$$

which estimates each μ_i by $\bar{x}_i$, as at (3.1). This estimate has expected squared error loss

$$E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = k \tag{8.8}$$

for every parameter vector $\boldsymbol{\mu}$. What Stein showed is that if k, the number of means to be estimated, is ≥ 3, then the estimator

$$\tilde{\boldsymbol{\mu}}(\bar{\mathbf{x}}) = \left[1 - \frac{k-2}{\|\bar{\mathbf{x}}\|^2}\right]\bar{\mathbf{x}} \tag{8.9}$$

has

$$E\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 < k \tag{8.10}$$

for every $\boldsymbol{\mu}$! (This particular form of $\tilde{\boldsymbol{\mu}}$ was developed jointly with W. James in 1960.) From a frequentist point of view, $\tilde{\boldsymbol{\mu}}$ estimates $\boldsymbol{\mu}$ uniformly better than does $\hat{\boldsymbol{\mu}}$. It is also better from a Bayesian point of view: given any prior distribution on $\boldsymbol{\mu}$, estimating by $\tilde{\boldsymbol{\mu}}$ rather than $\hat{\boldsymbol{\mu}}$ results in a lower overall expected squared error of estimation (averaging now over the randomness in $\boldsymbol{\mu}$ and the randomness in $\mathbf{x}$).

Stein's estimator is based on (8.3). Since $\|\hat{\boldsymbol{\mu}}\| = \|\bar{\mathbf{x}}\|$ tends to be greater than $\|\boldsymbol{\mu}\|$ with high probability, a shrinking factor $[1 - (k-2)/\|\bar{\mathbf{x}}\|^2]$ is used to give an estimate nearer $\boldsymbol{\mu}$. The shrinking factor is more drastic when $\|\bar{\mathbf{x}}\|^2$ is small. With $k = 10$, $\|\bar{\mathbf{x}}\|^2 = 12$, we have $\tilde{\boldsymbol{\mu}} = [.333]\bar{\mathbf{x}}$. If instead $\|\bar{\mathbf{x}}\|^2 = 800$ then $\tilde{\boldsymbol{\mu}} = [.99]\bar{\mathbf{x}}$. Figure 4 gives a schematic illustration.

Notice that the origin $\mathbf{O}$ plays a special role in the construction of $\tilde{\boldsymbol{\mu}}$, even though there is nothing in the statement of the estimation problem that favors $\mathbf{O}$. As a matter of fact, we can change the origin to any other point in k dimensional space, $\mathbf{O}'$ say, and obtain a different Stein estimate,

$$\tilde{\boldsymbol{\mu}}' = \mathbf{O}' + \left[1 - \frac{k-2}{\|\bar{\mathbf{x}} - \mathbf{O}'\|^2}\right](\bar{\mathbf{x}} - \mathbf{O}'), \tag{8.11}$$

which is also uniformly better than $\hat{\boldsymbol{\mu}}$.

Stein's result has created a host of difficulties for frequentists and Bayesians alike, which we can't pursue here. The implications for objective Bayesians and fiducialists have been especially disturbing. The seemingly flat prior distribution leading to (8.4) isn't flat at all: it forces the parameter vector to relatively far away from any prechosen origin $\mathbf{O}'$. If a satisfactory theory of objective Bayesian inference exists, Stein's estimator shows that it must be a great deal more subtle than previously expected.

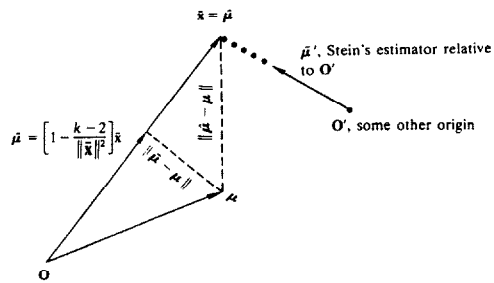

FIG. 4. Stein's estimate $\hat{\mu}$ is obtained by shrinking the obvious estimate $\hat{\mu} = \hat{x}$ toward the origin O. The shrinking factor is more extreme the closer $\|\hat{x}\|$ lies to O. Stein and James showed that $E\|\hat{\mu} - \mu\|^2 < E\|\hat{\mu} - \mu\|^2$ for every μ. We can choose any other origin O' and obtain a different Stein estimate, μ', which also dominates $\hat{\mu}$.

The trouble with the multiparameter estimation problem is not that it is harder than estimating a single parameter. It is easier, in the sense that dealing with many problems simultaneously can give extra information not otherwise available. The trouble lies in finding and using the extra information. Consider the Bayesian model (4.1). With just a single μ to estimate this model must be taken on pure faith (or relevant experience). However, if we have several means to estimate, $\mu_1, \mu_2, \ldots, \mu_k$, each drawn independently from an $\mathcal{N}(m, s^2)$ population, the data $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ allows us to *estimate* m and s^2, instead of postulating their values. Plugging the estimated values into (4.6) gives an "empirical Bayes rule" very much like the Stein rule (8.11). Empirical Bayes theory, originally developed by Herbert Robbins in the early 1950's, offers some hope of a partial reconciliation between frequentists and Bayesians.

9. Some last comments. The field of statistics continues to flourish despite, and partly because of, its foundational controversies. Literally millions of statistical analyses have been performed in the past 50 years, certainly enough to make it abundantly clear that common statistical methods give trustworthy answers when used carefully. In my own consulting work I am constantly reminded of the power of the standard methods to dissect and explain formidable data sets from diverse scientific disciplines. In a way this is the most important belief of all, cutting across the frequentist-Bayesians divisions: that there do exist more or less universal techniques for extracting information from noisy data, adaptable to almost every field of inquiry. In other words, statisticians believe that statistics exists as a discipline in its own right, even if they can't agree on its exact nature.

What does the future hold? At a recent conference Dennis Lindley, of University College, London, gave a talk entitled, "The future of statistics—A Bayesian 21st century." My personal subjective probability is .15 on that eventuality. The big advantage of subjective Bayesianism, which is what Professor Lindley was referring to, is its logical consistency. Philosophers who investigate the foundations of scientific inference usually wind up being repelled by frequentism and attracted to the Bayesian argument.

But consistency isn't enough. Subjective Bayesianism must face the challenge of scientific objectivity. This is the ultimate stronghold of the frequentist viewpoint. If the 21st century is Bayesian, my guess is that it will be some combination of subjective, objective, and empirical Bayesian, not significantly less complicated and contradictory than the present situation. The complexity of the problems statisticians are asked to deal with is increasing at an alarming rate. It is not unusual these days to deal with data sets of a million numbers, and models with several thousand parameters. As Section 8 suggests, this trend is likely to exacerbate the difficulties of producing a logically consistent theory of statistics.

Annotated References

V. Barnett, Comparative Statistical Inference, Wiley, New York, 1973. [A clear discussion of the frequentist viewpoint as compared with Bayesian methods.]

A. Birnbaum, On the foundations of statistical inference (with discussion). J. Amer. Statist. Assoc., 57 (1962) 269–326. [A much deeper discussion of foundational controversies. The discussion is excellent in its own right. I stole Pratt's meter man example for Section 4.]

B. DeFinetti, Foresight: Its logical laws, its subjective sources. Studies in Subjective Probability, ed. by M. Kyburg and H. Smokler, 93–158, Wiley, New York, 1964. [The most extreme, and with Savage the most influential, subjectivist of our time wrote this seminal work in 1935. This volume also contains essays by Venn, Boral, Ramsey, Koopman, and Savage.]

B. Efron, Biased versus unbiased estimation. Advances in Math., No. 3, 16 (1975) 259–277. [Stein's estimator in theory and practice.]

R. A. Fisher, Statistical Methods and Scientific Inference. Oliver and Boyd, London, 1956. [Fisher's last major work. Fiducial and conditional arguments are persuasively advanced. Must be read with caution!]

H. Jeffreys, Theory of Probability, 3rd Edition. Clarendon Press, Oxford, 1967. [The most important modern work on objective Bayesianism.]

D. V. Lindley, Bayesian Statistics—A Review. SIAM Monographs in Applied Mathematics, SIAM, Philadelphia, (1971). [A good reference for the Bayesian point of view, both subjective and objective.]

———, The future of statistics—a Bayesian 21st century, Proceedings of the Conference on Directions for Mathematical Statistics, (1974). Special supplement to Advances in Applied Probability, September 1975. [The essays by P. J. Huber and H. Robbins also relate to the future of statistics.]

L. J. Savage, The Foundations of Statistics. Wiley, New York, 1954. [This book sparked the revival of interest in the subjectivist Bayesian point of view.]

Department of Statistics, Stanford University, Stanford, CA 94305.