principle, elicit your own initial probability density $\alpha(\lambda)$ for the unknown weight of the potato, but in practice the self-interrogation may not work very well and you may be very vague about $\alpha(\lambda)$. There is a temptation, under these circumstances, to say that $\alpha(\lambda)$ does not exist or that you know nothing about the weight of the potato. Actually, it has proved impossible to give a satisfactory definition of the tempting expression ‘know nothing’. Still more, you do know a good deal about your opinions about the weight of the potato, and these can be quite well expressed in terms of partial specifications of $\alpha(\lambda)$. If, for example, it were necessary to mail the potato without weighing it, you could put on enough postage to be reasonably sure that it would not be returned to you nor be 500 per cent overpaid.

More important, you are almost sure to have a certain kind of knowledge about $\alpha(\lambda)$, which has seldom been mentioned explicitly, but which will be very useful after you have a chance to weigh the potato on a good balance. To illustrate, suppose that you found out, as a result of some experiment, that the weight of the potato to the nearest gram was either 146 or 147 gm. Given this knowledge, you would probably be willing to accept odds only slightly more favourable than one-to-one in favour of either of the two possibilities, 146 gm or 147 gm. This may be interpreted to mean that, for you, the average value of $\alpha(\lambda)$ near 146 is almost the same as its average value near 147. Continuing along this line, you might arrive at the conclusion that $\alpha(\lambda)$ varies by at most a few per cent in any 10-gm interval, included between, say, 100 and 300 gm. You might also conclude that $\alpha(\lambda)$ is nowhere enormously greater, say 1000 times greater, than even the smallest value that it attains between the bounds of 100 gm and 300 gm.

Armed with such knowledge, what could you conclude after weighing the potato on a tried and true balance known to have a normally distributed error with a standard deviation of 1 gm? Bayes’s theorem, in this context, can be written

$$\alpha(\lambda | x) \propto \phi(x - \lambda) \alpha(\lambda), \tag{2}$$

where

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}. \tag{3}$$
At first sight, (2) may seem inapplicable, because you know so little about $\alpha(\lambda)$. But suppose, for definiteness, that $x = 174.3$ gm. As Fig. 1 shows, the function $\phi(174.3 - \lambda)$ is almost zero outside the interval $174.3 \pm 5$; the function $\alpha(\lambda)$ varies by at most a few per cent inside that interval and is never enormously larger outside the interval than it is inside of it. Under these circumstances, the product on the right side of (2) is well approximated for many purposes by

$$\phi(174.3 - \lambda) \alpha(174.3).$$

![Graph of $\phi(174.3 - \lambda)$ and $\alpha(\lambda)$](image)

**Fig. 1.** Prior probability density $\alpha(\lambda)$ and likelihood $\phi(174.3 - \lambda)$.

The probability density $\alpha(\lambda)$ and the likelihood $\phi(174.3 - \lambda)$ are not drawn to the same vertical scale. Such quantities need not generally even be of the same dimension; with Poisson data, for instance, one would be probability per unit frequency, the other simply probability. Therefore, $\alpha(\lambda|x)$ is a probability density in $\lambda$ that is well approximated by some constant multiple of $\phi(x - \lambda)$, but the only such multiple that is a probability density in $\lambda$, that is, the one that is suitably normalized, is clearly $\phi(x - \lambda)$ itself. Thus, after the weighing, your opinion about $\lambda$ is expressed to a good degree of approximation by saying that $\lambda$ is normally distributed around 174.3 gm with a standard deviation of 1. Though this is much the kind of conclusion that is usually ridiculed in the statistics classroom, I hope you now feel that, in the presence of reasonable assumptions about your own initial subjective probability, it is not ridiculous but true.
Note well that the example depends on two properties of \(\alpha(\lambda)\), its near-constancy in the neighbourhood of the value of the observed value of \(x\), and the relative moderateness of \(\alpha(\lambda)\) far from that value. Both these properties of course refer to empirical facts about you and the potato. If, for example, you had weighed a potato yesterday that you thought might quite likely be the same potato that you are weighing today, \(\alpha(\lambda)\) has a more or less sharp peak. If this initial peak lies close to \(x\), then the assumption of near uniformity of \(\alpha(\lambda)\) in the neighbourhood of \(x\) is violated; if the peak falls 4 or 5 gm away from \(x\) (and if you are all but certain that this is indeed the potato that you weighed yesterday), then the assumption of moderate behaviour of \(\alpha(\lambda)\) distant from \(x\) fails to be satisfied.

The argument that led to \(\phi(174.3 - \lambda)\) as your approximate final probability density for the weight of the potato is given and greatly generalized by Jeffreys (1948, section 3.4). But for the most part, especially in the reference just given, as opposed to Jeffreys (1957), Jeffreys would adopt a somewhat different line of thought, saying that, if you know nothing about the potato, \(\alpha(\lambda)\) should be taken to be literally a constant. This is sometimes objected to on the grounds that such an \(\alpha(\lambda)\) cannot be normalized; it is not literally a probability density. That objection does not seem terribly important to me because there do exist finitely additive probability measures corresponding to the idea of a uniform distribution on the whole line. To me, a more serious objection is that such a uniform distribution does not really represent the initial opinion of anybody, and surely does not represent an opinion that ought to be held by everybody, even apart from our sure knowledge that the weight of the potato is positive. This approach of Jeffreys, in effect, puts forward a valuable approximation as an exact conclusion. An interesting discussion of these ideas is given by Good (1950, p. 51).

Following, and in a few cases extending, Jeffreys (1948), many classical statistical situations can be treated in the spirit of the example about the potato. If a variance \(\sigma^2\) is to be measured on the basis of a statistic \(s^2\) with \(n\) degrees of freedom, your final distribution of \(\sigma^2\) will (under favourable circumstances) be like that of \(ns^2/\chi^2_n\), where the data \(n\) and \(s^2\) are of course regarded as fixed. This can be extended to the corresponding conclusion about a set of covariances in terms of
Wishart's distribution. If the parameters $\mu$ and $\sigma^2$ of a normal distribution are to be jointly estimated from $\bar{x}$ and $s^2$ based on $n$ observations (and therefore $n-1$ degrees of freedom), then $\sigma^2$ is distributed as before with $n-1$ in place of $n$, and $n\hat{b}(\mu - \bar{x})/\sigma$ is, always speaking approximately, distributed independently of $\sigma$ with a standard normal distribution; putting those facts together, the approximate final distribution of $\mu$ is like that of $\bar{x} + n^{-\frac{1}{2}}s\eta_{n-1}$, where $\bar{x}$, $n$, $s$ play the role of constants. This example generalizes to lead to a theory of regression and the analysis of variance formally similar to, but intellectually different from, the usual theory. Standard $F$ distributions arise in connection with the distribution of the ratio of two unknown variances, beta distributions in connection with the estimation of a frequency from Bernoullian data, and gamma distributions in connection with the estimation of a Poisson parameter.

The Behrens-Fisher problem provides a particularly striking illustration of the theory, according to which the final distribution of two normal means when the means and variances are all grossly unknown is that of a certain linear combination of independent $t$-variables: this was shown in effect by Jeffreys (1948).

This is formally just the solution that has always been championed by Fisher. Fisher would say that it gives the exact fiducial probability, but I say only that it gives satisfactory approximate probabilities under suitable circumstances. The solution is not in even approximate agreement with confidence interval ideas; see Wallace (1959). In many cases, the theory of precise measurement tends to coincide in this way with the theory of fiducial probability, but the theory of precise measurement is not coterminus with the theory of fiducial probability (whatever the exact interpretation of the latter may be), because the theory of precise measurement is not dependent on the existence of sufficient statistics in the same sense as the theory of fiducial probability is and because the theory of precise measurement deals as well with discrete data as it does with continuous data.

The theory of precise measurement leads to a good understanding of the problem of estimating the ratio $\mu_1/\mu_2$ of two means from data of the form $\bar{x}_1$, $\bar{x}_2$, and $s$, or the closely related problem of estimating the direction of the vector $(\mu_1, \mu_2)$. The approximate answer adduced here agrees with what I understand to be the one produced by the
fiducial argument. This was given (for the problem of estimating the
direction of the vector) by Creasy and confirmed by Fisher, who were
in disagreement with Fieller; see Fieller et al. (1954).

The theory contributes to problems like the estimation of the
difference of Poisson parameters or of variances, where convolutions
of gamma distributions are invoked. However, these problems are
often such that the difference is often known to be positive, which
may well come into conflict with the hypotheses about gentle behav-
ior of the initial distribution necessary for the application of the
theory of precise measurement. In such cases the problems must be
considered anew and cannot be expected to have all of the advantages
of precise measurements. Judging from experience with a problem
that seems analogous to me, namely that of inferring an upper bound
on a danger from a perfect safety record, even these less satisfactory
situations can lead to useful inequalities. If a complete approximate
final distribution does not occur in such problems, it is because the
specification of the problem does not justify precise conclusions.
Any theory – confidence intervals, fiducial probability, logical prob-
abilities – that pretends to produce exactness where it is unjustified
is a false servant.

5. Initial and final precision

With our great emphasis on operating characteristics, on how an
experiment can be expected to perform, many of us have tended to
forget the distinction between the precision that was to be expected
from an experiment before it was performed and the precision actually
yielded by it when it was performed. The confusion has been reinforced
by the fact that for certain familiar kinds of experiments, the distinc-
tion really does vanish. The tendency to obliterate the distinction
between initial and final precision is particularly natural to those
objectivistic theories of statistics that officially refuse to discuss final
or terminal opinions. Nonetheless, objectivists of both schools have
pointed out the distinction in certain contexts.

A striking example concerning the estimation of the median of a
uniform distribution of known range is sometimes discussed; see, for
example, Lehmann (1959; p. 7, ex. 7) and the much earlier paper by
Welch (1939).
We have been accustomed to think that if an estimator has a small mean square deviation, then an estimate resulting from the application of this estimator is in some sense trustworthy, but this is not true in general. An example of an estimator that is ordinarily trustworthy is the average of observations from a normal distribution with unit variance. These observations ordinarily constitute what was called in the last section a precise measurement of the unknown mean of the distribution, and there is then strong reason to suppose that the true mean $\mu$ lies close to the sample mean $\bar{x}$; and, in all cases, a sample from this normal family affects the final opinion only through $\bar{x}$.

On the other hand, the natural estimator for the median of a uniform distribution of unit range is the mid-range, that is, the average of the maximum and minimum observations. This estimator has a very small mean square deviation for large sample sizes, about $\frac{1}{2}n^{-2}$, and under usual initial opinions it can well claim to be the best possible estimator. But if, by accident, the range of the sample is very narrow, then the sample leaves grave doubts as to the position of the population median. In fact, the final distribution after such an observation is simply the initial distribution truncated to the interval from the maximum $-\frac{1}{2}$ to the minimum $+\frac{1}{2}$. If, for example, the maximum observation is $3\frac{1}{2}$ and the minimum $2\frac{1}{2}$, the median of the uniform distribution of unit range must be at least $3\frac{1}{2} - 1 + \frac{1}{2} = 2\frac{1}{2}$ and at most $2\frac{1}{2} + 1 - \frac{1}{2} = 3$; study of the likelihood shows that this is in fact all that the sample has to convey about the location of the median. If the initial distribution is sufficiently diffuse, then the final distribution is nearly uniform in the interval $2\frac{1}{2}$ to $3$, no matter how large the sample was. The fact that the mean square deviation of the mid-range estimate is small corresponds to the fact that the final interval of nearly uniform uncertainty is almost always small, but this interval can occasionally be of almost unit length, in which case the sample is a relative failure.

Many other translation families exhibit much the same phenomenon. Consider the double exponential (or Laplace) distribution with density $\frac{1}{2}\exp(-|x-\mu|)$. The maximum likelihood estimate of the parameter $\mu$ is here the median of the sample. For large samples, its variance decreases nearly proportionally to $1/n$, and the Pitman (1939) estimate has a slightly smaller mean square deviation. Once
more, this does not mean that an observation of a large sample for
this distribution necessarily gives a sharp indication of the position
of the median $\mu$. If, for example, that there were two hundred obser-
vations of which one hundred fell beneath $-1$ and the other hundred
above $+1$, the likelihood function would be a constant throughout
the interval $-1$ to $+1$, and while the data may give strong evidence
that the true value of $\mu$ is somewhere between the 100th and 101st
observation, it gives no clue at all as to where in that interval $\mu$ is.
The corresponding example (Fisher, 1934) for an odd sample size is
not much less striking. The Cauchy distribution produces similar
phenomena, and extensive observation of a Cauchy distribution can,
very rarely to be sure, leave us with a sharply bimodal final distribu-
tion with the modes far separated from one another.

The examples just given are closely related to, and can be reformu-
lated in terms of, a certain phenomenon about confidence intervals.
For example, for the uniform distribution, the usual analogy between
the theory of testing and confidence intervals suggests that the mid-
range of the sample $\pm \epsilon_n$ would be a good confidence interval for the
median $\mu$, with the constant $\epsilon_n$ suitably chosen. This is indeed a
confidence interval, but it has distressing properties. If the range
of the sample is in excess of $1 - 2\epsilon_n$, then we know with certainty that
the interval contains the true value of $\mu$, which was more than was
bargained for. If the mid-range is unusually small, there is good
evidence that the interval fails to cover $\mu$. This evidence is meaningful
even to objectivists of the Neyman-Pearson school, for it is easy to
calculate that anyone who offers to pay 19 to 1 if the confidence inter-
val misses the true value conditional on the mid-range's being un-
usually small stands to lose money. Even the system of confidence
intervals based on what is technically called the best unbiased test of a
null hypothesis about the mid-range has much the same unreasonable
properties (Welch, 1939).

There is some literature, mostly objectivistic, devoted to this
phenomenon of confidence intervals that are, so to speak, not condi-
tional confidence intervals. One reference, leading to others, is
Wallace (1959). So far as problems of translation are concerned,
confidence intervals proposed by Pitman (1939) meet the situation,
and, in fact, agree with intervals that would be generated from the
theory of precise measurement by taking intervals from the centre of the approximate final distribution with specified probability.

A prominent instance in which the outlook of the Neyman-Pearson school led to neglect of the difference between precision promised and precision delivered is the Stein two-sample procedure (Stein, 1945) for producing a confidence interval of fixed length, say one, for normal distributions of unknown mean and unknown variance, as has been mentioned by Lindley (1958). The general idea here is to use a pilot sample with mean \( \bar{x}' \) and sample standard deviation \( s' \) and then to adopt a total sample size \( N \) that promises to be large enough to ensure the required precision. If \( \bar{x} \) is the mean of the whole sample, then (except for slight approximations) \( \bar{x} \pm \frac{1}{2} \) is a confidence interval at the required confidence level and the problem is formally solved. Stein actually proposed certain refinements to take account of the discreteness of the integers and of the possibility that the first sample is already more than adequate. These refinements need not detain us here, but in common sense they are steps in the wrong practical direction, as Stein points out, for they waste data to avoid justifying a conclusion of more precision than is required. Now, if the Stein procedure is carried out, and if it happens, as it occasionally will, that the standard deviation \( s \) of the whole sample is much larger than \( s' \), then there is good evidence that the interval \( \bar{x} \pm \frac{1}{2} \) has missed its mark. Once more, the evidence can be called objective. The ingenious procedure that was widely acclaimed in the statistical climate of 1945 has since been seen by frequentists, including its author, and by personalists to have the serious defect just pointed out.

Incidentally, the concept of precise measurement immediately suggests a good practical solution to the kind of practical situation that gave rise to Stein’s problem. If you want to weigh a potato on a balance of highly unknown standard deviation sufficiently frequently to be able justifiably to give odds of 99 to 1 that the true weight \( \mu \) of the potato lies in an interval 0.1 gm long, the natural and I believe the right thing to do is this. Simply weigh the potato repeatedly until you find that the middle 99 per cent of your final (or better, intermediate) distribution has a length of 0.1 gm. This is practical, because after the first four or five weighings, your intermediate distribution will typically be well approximated with the aid of the \( t \)-distribution of
$n-1$ degrees of freedom. An interesting objectivist discussion of this and related methods is given by Anscombe (1954).

6. Sharp null hypotheses

To give you an illustration of the application of subjective probability outside of precise measurement, I shall say something about testing sharp null hypotheses, mainly in the form of an allegory, though I still know too little about this application of subjective probability.

At least three different situations are commonly crowded into a common Procrustean bed in the name of testing a null hypothesis. There are still other situations that used to be confused with these but that are less often now (Bahadur and Robbins, 1950). These three situations have been poorly distinguished, when distinguished at all, because the real differences among them are largely differences in initial probabilities, about which objectivists have no adequate vocabulary. Let me illustrate by telling three versions of the legend of King Hiero’s crown.

In all three versions the king knows or suspects that his goldsmiths have adulterated the gold of his new crown. Archimedes, under delightful circumstances, hits on the idea of determining the density of the crown by weighing it and a specimen of pure gold in air and water. It does no important violence to the story to suppose that Archimedes has in effect measured a number $\lambda$, with the error of measurement normally distributed with standard deviation $\sigma$ about $\lambda$. The crown is either unadulterated, denser than pure gold, or less dense than pure gold according as $\lambda = 0$, $\lambda > 0$, or $\lambda < 0$. I choose to forget here that the presumed adulterant was silver, which would lighten the crown; retaining that feature would lead to one-sided tests.

It is odd, though not unthinkable, that Archimedes and the king should know $\sigma$, but this helps keep the example down to essentials. Archimedes may in fact have made and averaged many measurements. Why he made just the number he did is another story, but once he stops, the average is in effect a single measurement as postulated.

Depending on his source of information and on his objectives, Hiero might be imagined to have one of the following sorts of initial attitudes, among others.
Version 1. The king is sure that there has been cheating, and his opinion about its extent is diffuse with respect to Archimedes's measurement. The king would like to decide whether the crown is denser, or less dense, than gold. But he is willing and free to abstain from deciding if the evidence is inconclusive.

Version 2. The king is sure that there has been cheating, but his opinion is not diffuse. Rather, he feels with considerable confidence that $|\lambda| < 2\sigma$. Once more he would like, if possible, to decide whether the crown is denser or less dense than gold. This version is the most difficult and the one of most prominence in realistic statistical practice and thinking.

Version 3. The king attaches some credence, be it large or small, to the possibility that there has been no cheating ($\lambda = 0$), and his opinion about the extent of the cheating, conditional on there being some, is diffuse with respect to the measurement. He wants to hang the goldsmiths if they are at all guilty, otherwise not.

According to many objectivist textbooks, the king's response to a measurement $x$ should be about the same in all three versions. He should, it is implied, select some small probability $\alpha$, say $\alpha = 0.05$, 0.01, or 0.001, at his discretion. He should then compute the probability that $r = x/\sigma$ would be at least as large as the observed value if $\lambda$ were 0, that is

$$1 - \int_{-|r|}^{+|r|} \phi(z) \, dz = \Phi(-|r|) + (1 - \Phi(|r|)) = 2\Phi(-|r|),$$

If this value is less than his $\alpha$, he should reject the null hypothesis, otherwise accept it. For Versions 1 and 2, rejection means to take the sign of $x$ seriously as an indication of the sign of $\lambda$. For Version 3, it means to hang the goldsmiths.

An alternative textbook doctrine that might be offered is not to reject at any fixed $\alpha$ but to regard $\alpha(r) = 2 - 2\Phi(|r|)$ as some kind of measure of the doubt the king should have were he to reject the null hypothesis. You shall see that both objectivist doctrines are inappropriate to Version 1 and especially inappropriate to Versions 2 and 3. No version for which they are appropriate is known to me.

In Version 1, the king feels rather sure before making the measure-
ment that he will obtain practically unequivocal information about the sign of \( \lambda \), which is what he wants to know in this version. He might, for example, be willing to bet 100 to 1 that \(|x|\) will be found to exceed 4\( \sigma \). If \(|x|\) is in fact that large, the principles of precise measurement will leave the king little doubt about the sign of \( \lambda \), though (under typical circumstances) it would be an abuse of those principles to attempt thus to measure this small doubt by other than a rough upper bound. If, however, \( x \) does happen to fall within, say, 3\( \sigma \) or 4\( \sigma \) of 0, the king may, under favourable circumstances, find that \( \Phi(-|t|) \) is about the probability for him that the sign of \( \lambda \) is not that of \( x \). Note that \( \Phi(-|t|) \) is not \( \alpha(t) \) but \( \frac{1}{2} \alpha(t) \), so that neither textbook doctrine is quite appropriate to Version 1.

In Version 2, the king rather expects to find \( x \) within 2\( \sigma \) of zero. If this does in fact happen, his terminal distribution of \( \lambda \), and in particular his terminal probability that \( \lambda \) is positive, depends quite sensitively on the behaviour of his initial distribution of \( \lambda \) near \( \lambda = 0 \). This distribution was not adequately specified in the description of the problem, and, in particular, it may be very asymmetrical. Still more, the king may not know his own mind well enough to specify adequately the behaviour of his initial distribution of \( \lambda \). In this case, the experiment will inescapably leave him in a quandary. Any theory that pretends to do more, to reach a conclusion about the sign of \( \lambda \) without using the king’s initial opinion, goes too far. What can be hoped for is useful inequalities.

Version 3 has an interesting theory for which we are indebted to Jeffreys (1948), who presents both special cases and general theory. Lindley also discovered a broad generalization of this theory some years ago, which will be published shortly (Lindley, 1961).

Let \( I \) be the king’s initial probability that \( \lambda = 0 \). Let \( \pi(\lambda) \) be his (diffuse) initial density for \( \lambda \) given that \( \lambda \neq 0 \). The corresponding final quantities \( I' \) and \( \pi'(\lambda) \) are determined through Bayes’s theorem by

\[
I' = \frac{C}{\sigma} \phi(\frac{x-\lambda}{\sigma}) I = \frac{C}{\sigma} \phi(t) I
\]

and

\[
I' \pi'(\lambda) = \frac{C}{\sigma} \phi\left(\frac{x-\lambda}{\sigma}\right) I \pi(\lambda).
\]
Integrate (2), recalling that $\pi$ is supposed to be diffuse relative to $\phi$, to conclude that

$$I' = C\pi(x) I,$$

(3)

provided $t$ is not enormous; otherwise, there is strong but ill-measured evidence of guilt.

The division of (1) by (3) yields an approximation for the final odds in favour of innocence ($\lambda = 0$),

$$\frac{I'}{I'} \approx \frac{\phi(t)}{\sigma} \cdot \frac{1}{\pi(x) I'},$$

(4)

It is gratifying to find that the terminal odds are a multiple of the initial odds. The one other aspect of the king’s initial opinion about $\lambda$ that enters is $\pi(x)$. Objectivistic doctrine would suggest that the only thing important about the measurement for the king’s decision should be the ‘double tail-area’ $\alpha(t)$. Actually, it is not through $\alpha(t)$ but through the density $\phi(t)$ that $t$ enters, and $\sigma$ plays a role as well as $t$.

Precedence for the importance of $\sigma$ here can probably be found in the writings of objectivists, but, broadly speaking, it is in contrast with objectivistic theorizing.

The king will hang the goldsmiths if $I'/I'$ is large enough. He may be hard pressed to tell himself how large is large enough or to evaluate the personal factor $I/(\pi(x) I')$. This lack of self-knowledge may leave him in a quandary, but often $t$ and $\sigma$ will be such that very rough values of the personal factor and the critical odds suffice. Also, in real life, the king might be able to make a new measurement if in doubt or to accept the risk of excusing slight guilt as relatively unimportant or to do still other things that would ameliorate his dilemma.

The general theory developed by Jeffreys and Lindley for problems like Version 3 is adequate to deal with nuisance parameters. Various double dichotomy problems now normally treated by $\chi^2$ with one degree of freedom are a good testing ground for this theory; some exploration is carried out by Jeffreys (1948, sections 5.11–5.14; 1957, section 3.6).

One thing that I have tried to convey in this section is what seems to become of the theory of testing hypotheses when it is studied through
subjective probability, but this is not a topic that I could cover fully here, even if there were space. In particular, it would be challenging to offer a clear analysis of ‘shotgun tests’. Friends have emphasized that real life seems to offer few problems at all like Version 3; I agree and conclude, with those friends, that hypothesis testing is not nearly of such widespread appropriateness as many who routinely use statistics seem to think.

7. Conclusion

Every topic in statistics ought to be reviewed in the light of the concept of subjective probability, but certain broad problems seem particularly important at present.

First, there are many situations like precise measurement, except that an infinite (or at least unlimited) number of parameters are involved, and typically we do not expect to have enough information to measure all of them precisely but only to measure precisely a few functions of the infinite set of parameters. The problem of five-year survival mentioned in section 3 is a case in point. Another, and a typical problem of non-parametric statistics, is to measure the median or other percentage points of a largely unknown distribution function. Whittle (1957, 1958) has recently applied subjective probability to curve-fitting and to the estimation of the autocorrelation function of a stochastic process. Both these problems are further illustrations of what I meant to suggest by an unlimited number of unknown parameters, and it would be good to see the impetus given by Whittle followed up.

Many feel intuitively that there are circumstances that do call for application of the traditional tests, like $F$ tests and $\chi^2$ tests. Where, if anywhere, this intuition is justified, I am confident that it will not be found at variance with the concept of subjective probability, but the situation has not yet been properly analysed and explored.

Again, though we all feel sure that randomization is an important invention, the theory of subjective probability reminds us that we have not fully understood randomization. It is not enough to say something like this: ‘If you randomize here, you are very unlikely to make a mistake.’ It can happen that when the randomization is done, the
experimenter sees that he has made a mistake, that is, sees that the experiment called for by the randomization is inappropriate. This point has been studied by several who probably think of themselves as objectivists (Grundy and Healy, 1950; Jones, 1958; Yates, 1951a, b). In particular, randomization could accidentally closely correlate any variable that has not been controlled by stratification or some such device, with one of the treatments. For example, we might accidentally choose at random a Latin square with the treatments running in regular slanting lines across the field. It would usually be most ill-advised to carry out such an experiment in which treatment is highly correlated with a possible gradient in fertility merely because this bad design had arisen at random.

The problem of analysing the idea of randomization is more acute, and at present more baffling, for subjectivists than for objectivists, more baffling because an ideal subjectivist would not need randomization at all. He would simply choose the specific layout that promised to tell him the most. The need for randomization presumably lies in the imperfection of actual people and, perhaps, in the fact that more than one person is ordinarily concerned with an investigation. The imperfections of real people with respect to subjective probability are vagueness and temptation to self-deception, as has been explained, and randomization properly employed may perhaps alleviate both of these defects.

Other problems that seem to concern self-discipline are those of dealing with outlying observations and with empirical surprises generally. Here again, subjective probability gives a simple solution in principle. For example, an observation ought to be regarded as due to a gross error if and only if Bayes's theorem tells us that it probably is, but vagueness and temptation make it particularly difficult to apply this simple maxim. Again, self-discipline is often enforced by using part of the data to suggest ideas and the rest of it to confirm these ideas. This process is not really easy to understand and appraise. In particular, there seems to be little cogent advice as to what fraction of the data should be used for exploration and what fraction for confirmation. Half and half is often suggested, but without particularly good reason. One interesting discussion of this problem of hindsight is by Simon (1953).
In this lecture, I have emphasized applications of subjective probability that do not depend sensitively on many details of the initial probabilities. Such robustness is important when it occurs, but we must remember that statistics presents many problems that are not robust in this sense. The problem of choosing a sample size (and some other design parameters) is an especially conspicuous example. Even in these cases, it is, I believe, far better to use the concept of subjective probability to order our thoughts than to try to make the necessary choices by unformalized intuition.

I hope that even though you may not yet fully share my enthusiasm, you have come to feel that subjective probability promises to make important contributions to statistical theory. The improvements are so simple and far-reaching that they are by no means confined to academic theorizing but should have an immediate impact on our teaching and consulting.