XXXXXXXXXXXXXXXX

Sequential Measurement Schemes, prequential inference and the Dickey–Fuller Test.

Summary. A classical "paradox" in statistical inference concerns a situation where a measurement is performed by an instrument selected at random among two with different error variances. Orthodox Neyman–Pearson theory fails in this case, and the example is a standard argument for conditioning on an ancillary statistic whenever this is possible.

More complex sequential versions of this scenario are discussed. Measurements are made according to schemes where the precision of each measurement depends (in a known or unknown way) of earlier measurements. It is argued that such situations should be handled according to the prequential principle (Dawid 1984), i.e. by ignorance of their sequential nature. Our arguments are based on (imagined but) very concrete scenarios, where ignorance of the prequential principle would force us to behave in absurd ways.

The most general of these scenarios turns out to contain a simplified version of a wellknown problem in econometrics, namely that of testing for an autoregression coefficient equal to 1 in an AR(1) process with known error variance and expectation 0. This results in an extremely simple solution to this problem.

However, recommendations based on the prequential principle must necessarily be given with some reservation, because this principle — when followed strictly — results in other "paradoxes". This is discussed in section 4.

American Mathematical Society 1980 subject classification.
Primary 62A99; secondary 62A10, 62L99.

Key words and phrases.
Statistical inference, ancillarity, autoregression, prequential inference, unit roots, sequential sampling plan, cointegration, Dickey–Fuller test.

0. Introduction.

The purpose of this article is to indicate how some of the wellknown foundational problems in statistics, namely some of the "paradoxes" associated with sequential sampling plans, have a very simple and direct relation to a class of problems in econometrics (key words AR(1), Dickey–Fuller test, cointegration). This relation suggests that it may be possible and perhaps even desirable to replace some of the methods developed in the context of cointegration by simpler, classical methods.

The ideas discussed are closely related to prequential inference as introduced by Dawid (1984). If there is any difference at all between the principle promoted in the present article and Dawid's prequential principle, it has to do with generality. We focus strictly on situations where inference about a parameter in a well–defined statistical model is the issue. A secondary purpose of this work is to contribute to the clarification of some of the concepts related to prequential inference.

It would be an exaggeration to say that a unified conclusion or set of recommendations can be deduced from this. Apart from my personal opinions, the only conclusion coming out of all this is that we have a foundational problem here. But in a wider perspective, I hope that this way of making foundational problems visible, not only in imagined and more or less theoretical situations, but also in statistical practice, will contribute in a constructive way to the necessary clarification.

We begin by recalling some of the foundational problems in statistical inference. The following is a slightly modified version of an example due to Cox (1958).

Scenario no. 1

A normal measurement of an unknown quantity μ is performed with one of two instruments, a very inaccurate instrument with error variance 1*, and a very accurate one with error variance* 1/1000*. The instrument to be used is selected at random by a coin–toss.*

This example has become widely known because it illustrates so very clearly how orthodox Neyman–Pearson theory breaks down when it is used blindly. It would be an exaggeration to call this a paradox, since there is essentially no disagreement today about how to handle a situation like this. As anyone who is not a theoretical statistician can see immediately, the circumstance that we might have used another instrument is irrelevant for our conclusions from the measurement that we actually made. The example is often taken as a standard argument for the principle that one should condition on an ancillary statistic (in this case the outcome of the coin toss) whenever one is present.

It is well known how this *conditionality principle*, together with a principle of similar intuitive appeal, the *sufficiency principle*, can be shown to imply the socalled *(strong) likelihood principle* (see Birnbaum 1962) which in the present context can be expressed as follows. If two experiments for determination of an unknown quantity result in proportional likelihood functions, then the conclusions from these two experiments should be identical.

Unfortunately, this principle questions most of the activities in which theoretical and applied statisticians are involved. Only strictly subjective Bayesian methods are consistent with the likelihood principle, and since these methods have a tendency to question themselves by their dependence on a more or less arbitrary prior distribution, the statistical science really has a problem here. The philosophical justification of what we are actually doing is so weak and self–contradictory that one would probably tend to give up the whole idea, if it wasn't for the fact that statistical methods are so useful and unavoidable in practice.

The purpose of the present paper is not to review the long discussion of the many obvious principles that one can set up for statistical inference and their tendency to contradict each other. This has been done by many other authors, and we would like in particular to draw attention to the rather complete review by Berger and Wolpert (1984). Our aim is merely to indicate (section 1 and 2) — without much discussion of abstract principles — how some of these "paradoxes" come very close to the surface in a class of simple sequential measurement settings, where the interpretation of a given situation as "sequential" is sometimes made impossible by the fact that it would force us to do absurd things in conflict with common sense. Furthermore, as we shall see in section 3, one such situation comes up in a simplified version of a problem from econometrics where the presence of the "paradox" questions the relevance of established econometric methods.

1. A simple sequential experiment.

We proceed with an example which is essentially Cox's example once more, but this time in a sequential dress.

Scenario no. 2 (a sequential version of Cox's example)

A measurement of an unknown quantity μ is performed with error variance 1. A coin is flipped. If head comes up, 999 additional measurements (also with variance 1) are performed, otherwise no further measurements are taken.

(Remark: Here and in the following, measurements are independent and normally distributed).

Clearly, this is essentially equivalent to our first scenario. The equivalence becomes even more obvious if the coin–flip is assumed to be done before the first observation, so that we are merely selecting the sample size (1 or 1000) at random. However, this apparently innocent change of the order in which things are done is a crucial point in the examples to follow. An important assumption in the following is that if we really want to let coin–tosses determine our actions, then there is no need to perform the tosses before their outcomes are required.

In our next scenario, again we observe either 1 or 1000 $N(\mu, 1)$ random variables. But the coin–toss of scenario 2 is now replaced with a decision based on the first measurement.

Scenario no. 3 (a simple sequential scheme)

A measurement Y_1 *of an unknown quantity* μ *is performed with error variance 1. If* Y_1 *is less than a certain constant* c*, 999 additional measurements (also with variance 1) are performed, otherwise no further measurements are taken.*

This is a proper sequential situation, and things are less transparent here. However, it is possible to argue that an experiment like this should also — once it is performed — be interpreted as if the number of measurements had been decided in advance. A Bayesian fundamentalist would accept this immediately as a consequence of the likelihood principle. For the rest of us, a more direct and rather convincing argument goes as follows. Consider

Scenario no. 4

We intend to do as follows. A coin is flipped. If head comes up, we simply perform 1000 measurements. If tail comes up, we follow the sequential scheme of scenario 3.

However, since we cannot find a coin right away, and since we are going to make at least one measurement anyway, we decide to perform the first measurement while somebody else is taking care of the search for a coin. This results in a value $Y_1 < c$*. A coin is still not available, but since the next 999 measurements are to be performed anyway, we proceed with these.*

Having performed the 1+999 = 1000 measurements, we proceed with the final task, which is to find a coin and flip it.

This situation is, of course, absurd. No person with his or her common sense in behold could possibly be persuaded to regard the final coin–tossing as an important or informative matter. The problem is that the coin is supposed to tell us how to interpret our 1000 measurements. If we are not willing to flip the coin, we are forced to admit that the outcome of this is irrelevant, hence that the distinction between the two interpretations of our 1000 measurements is irrelevant. In the language of Dawid (1991), the *production model* — which is a rather complicated thing here, involving the constant c and the distribution of various statistics under the sequential scheme — can be replaced with the *inferential model* which simply assumes the 1000 measurements to be i.i.d. $N(\mu, 1)$.

2. A general sequential scheme.

Scenario 3 (and in fact any of the scenarios considered) is a special case of a more general scheme, which can be explained in terms of an "instrument manager". The instrument manager is the person — or rule, or algorithm, if you wish — who decides for us which instrument to use next and when to stop the sequence of measurements. The decisions of the instrument manager may depend on previous measurements. In principle, it is also possible to let the decisions depend on other things belonging to the past, including external randomization, but this is not an important point here and is therefore disregarded in the following. With this simplification, we can formalize the situation as follows.

Scenario no. 5 (a general sequential scheme)

A finite sequence

$$\begin{aligned} Y_1 &\sim N\left(\mu, \sigma_1^2\right) \\ Y_2 &\sim N\left(\mu, \sigma_2^2(Y_1)\right) \\ Y_3 &\sim N\left(\mu, \sigma_3^2(Y_1, Y_2)\right) \\ &\cdots \\ Y_n &\sim N\left(\mu, \sigma_n^2(Y_1, \ldots, Y_{n-1})\right) \end{aligned}$$

of measurements are performed. The variance of each measurement is a function of previous measurements, and the normal distributions specified are conditional, given all previous observations. The number of observations n is a stopping time.

The last condition can formally be build into the functions σ_1^2, σ_2^2, ... by the assumption that we have $\sigma_i^2 = +\infty$ from a certain stage with probability 1. However, if this appears too complicated, it suffices to think of the case where n is fixed.

Again, we can argue that statistical inference from this experiment should be performed exactly as if the number of observations n and the variances σ_1^2, σ_2^2, ..., σ_n^2 were known and fixed. An argument, similar to our argument for the same principle in scenario no. 3, goes as follows. Consider

Scenario no. 6

A coin is flipped. If head comes up, we perform ten measurements by pre–determined instruments with error variances $\sigma_1^2, \ldots, \sigma_{10}^2$. If tail *comes up, we proceed as in scenario 5, following the scheme of the instrument manager.*

However, a coin is not available right away. We ask the instrument manager what his first choice would be (just in case ...). Most surprisingly, he claims that his first choice would be the instrument with error variance σ_1^2. We decide — since a measurement with that instrument is to be performed anyway — to perform that measurement, while others are trying to find a coin.

After the observation of Y_1 a coin is still not available. We ask the instrument manager what his next choice would be. Most surprisingly ...

... and so on and on and on until ...

we observe Y_{10}, and the instrument manager claims that this is where he would like to say stop, if he was asked.

It remains to find a coin and flip it.

Again, we find ourselves in the totally absurd situation of being forced to let a coin decide for us how to interpret our ten measurements. While others are continuing the search for a coin, we might even proceed with two parallel statistical analyses and the writing of two final reports, being willing, of course, to drop the irrelevant one in the paper basket when the coin has told us which one it is. It is tempting to take the attitude that this kind of behaviour has no relevance in the scientific world. But if we take this attitude we are forced to accept that the final coin toss is irrelevant. Hence, we are forced to accept that our final analysis of the measurements should be independent of the coin toss. Hence we are forced to admit that the sequence of measurements may as well be interpreted as if head had come up, i.e. as if the ten instruments had been selected in advance.

It is wellknown that this attitude creates other problems. This is actually why we are using the word "paradox". These other problems will be recalled in section 4. But let us first take a look at one of the consequences of this "principle of treating a sequentially determined measurement scheme as if the scheme had been decided in advance".

3. Testing for an autoregression coefficient equal to 1 in the AR(1).

Consider the following problem, which is a simplified version of a problem first discussed (I believe) by Dickey and Fuller (see e.g. Dickey and Fuller 1981) and later studied intensively by econometricians in the context of "cointegration" (see e.g. Engle and Granger 1987, Johansen 1991).

Let $(X_0, X_1, \dots, X_n)$ be an autoregressive process of order 1 with mean 0 and known prediction error variance σ^2. By this we mean the following.

$X_0 = x_0$ can be regarded as fixed, since we are going to condition on it anyway. For convenience, we assume $x_0 \neq 0$. $X_1, \ldots, X_n$ are assumed to be generated recursively as

$$X_i = \alpha X_{i-1} + \sigma U_i$$

where the "normalized prediction errors" $U_1, \ldots, U_n$ are i.i.d. $N(0, 1)$. Our concern is estimation of the unknown autoregression coefficient α and, in particular, test of the hypothesis $\alpha = 1$.

It is easy to transform this to a special case of the "instrument manager scheme" (scenario no. 5). If we define $Y_i = X_i/X_{i-1}$, we have, conditionally on previous observations $Y_1, \ldots, Y_{i-1}$ (or $X_1, \ldots, X_{i-1}$)

$$Y_i \sim N\left(\alpha, \frac{\sigma^2}{X_{i-1}^2}\right).$$

Thus, the transformed series $Y_1, \ldots, Y_n$ can be interpreted as a sequence of measurements of the autoregression coefficient α with variances determined by the earlier measurements. Following the principle that these variances should be regarded as predetermined, we obtain (by ordinary averaging of our measurements with their inverse variances as weights) the estimate

$$\begin{aligned} \hat{\alpha} &= \frac{X_0^2 Y_1 + X_1^2 Y_2 + \cdots + X_{n-1}^2 Y_n}{X_0^2 + X_1^2 + \cdots + X_{n-1}^2} \\ &= \frac{X_0 X_1 + X_1 X_2 + \cdots + X_{n-1} X_n}{X_0^2 + X_1^2 + \cdots + X_{n-1}^2} \end{aligned}$$

which is just the OLS estimate of α obtained by regression (with intercept $= 0$) of the variate $X_1, \ldots, X_n$ on its first lag $X_0, \ldots, X_{n-1}$. Formally, the variance of this estimate is obtained by the rule for addition of precisions,

$$\operatorname{var}(\hat{\alpha})^{-1} = \left(\frac{\sigma^2}{X_0^2}\right)^{-1} + \cdots + \left(\frac{\sigma^2}{X_{n-1}^2}\right)^{-1},$$

i.e.

$$\operatorname{var}(\hat{\alpha}) = \frac{\sigma^2}{X_0^2 + X_1^2 + \cdots + X_{n-1}^2}$$

and a test for $\alpha = 1$ can thus be based on the statistic

$$U = \frac{\hat{\alpha} - 1}{\sqrt{\operatorname{var}(\hat{\alpha})}}$$

which is (formally) $N(0, 1)$ under the hypothesis.

4. Discussion.

Our solution to the unit–root testing problem would be considered very controversial by many econometricians, not only because we have over-simplified the situation (assuming σ^2 known, intercept = 0, no covariates etc.), but especially because the distribution of the test statistic U above is *not* a normalized normal distribution when $\alpha = 1$, not even in the limit as $n \to \infty$. This is so because the random walk behaviour of the AR(1) for $\alpha = 1$ implies a random behaviour of the denominator in the expression for $\hat{\alpha}$ which is not compensated by the law of large numbers. In the sequential measurement setting, we can explain this random variation as a variation of the total information (= the sum of the inverse variances), and our interpretation of the variances as pre–determined implies a sort of "conditioning on the information" which, conceptually, is very similar to the conditioning on the coin toss in Cox's classical example. But here we are not talking about a conditioning in the usual sense of this word, since that would involve (more or less) a conditioning on the observations themselves.

Conceptually, this kind of "conditioning", or whatever it is, is wellknown in time series analysis. The interpretation of a lagged variable as fixed when it occurs on the right hand side of a regression equation, even though it occurs as the random response on the left side of the equation just above, is an example. A similar idea is known from survival analysis where the formation of Cox's partial likelihood involves a similar recursive conditioning on previous events, including previous responses.

It is tempting to conclude from all this that econometricians are making life unnecessarily difficult for themselves when they focus on the complicated distribution of the test statistic for $\alpha = 1$. In fact, my personal opinion is that they are. But we must not forget, in this context, that the idea of analysing any sequential experiment by non–sequential methods has its own traps or "paradoxes". The classical warning goes something like this. Consider

Scenario no. 7

I.i.d. measurements $Y_1, Y_2, \dots \sim N(\mu, 1)$ *are taken until* $|\bar{Y} - \mu_0| \times \sqrt{n} \geq 3$*, where* μ_0 *is a (pre–determined) constant.*

Thus, we are sampling until the usual estimate of μ is at least three standard deviations from μ_0. This happens sooner or later with probability 1, even when $\mu = \mu_0$. Thus, regardless of whether $\mu = \mu_0$ or not, the standard (non–sequential) test for $\mu = \mu_0$ results in a highly significant rejection ($|U| \geq 3$) with probability 1.

When theory breaks down — as it certainly seems to do here — we are forced to rely on our intuition and common sense. For my own part, I can say that when I analyse AR or VAR models I do it unscrupulous by reference to the standard regression framework, with lagged responses taking the same role as other explanatory variables that are relevant for the prediction, and without much concern about whether the estimated autoregression coefficient(s) corresponds to a stationary time series, a series with unit–roots or an exponentially exploding series. To me, the usual dynamic description of the next observation as a sum of a linear predictor and an independent error term, provides sufficient justification for making inference as in an ordinary regression situation. In particular, if I needed to test the hypothesis $\alpha = 1$ in the AR(1), I would do it without hesitation by a standard T–test. I am aware that this attitude somehow ignores the fact that the test statistic is not T–distributed in the usual concrete sense. But to me, this problem is not much different from the "problem" that a test for a simple hypothesis in Cox's example (scenario no. 1) should also be based on a distribution which is different from the marginal distribution of the test statistic. Other researchers, I know, would feel very uncomfortable with this attitude and prefer to use the more complicated methods developed in the framework of cointegration, see e.g. Johansen (1991).

Perhaps the only general and quite safe recommendation that one can give at present is to try both methods. If they more or less agree, everything is fine. If they don't, we may have a problem. My impression is that the methods tend to agree in most relevant situations (implying that scenario 7 is somehow pathological), but that is a quite different story.

5. Prequential inference.

The type of paradoxes discussed here are not new. Some references related to this kind of problems and attempts to solve them are Oden (1977), Berger and Wolpert (1984) (with several references to earlier contributions) and Dawid (1984, 1991). What we have presented here is just one of the many ways of running into self–contradictions and paradoxes when foundational issues are discussed. As we have seen (scenario no. 6), the "principle of ignoring the sequential sampling plan", which is a version of Dawid's prequential principle, is unavoidable. And as we have also seen (scenario no. 7), this principle does, under certain circumstances, imply almost sure rejection of a simple hypothesis, even if it is true.

Our "principle of analysing sequential experiments by non–sequential methods" refers to a situation where a well–defined sampling scheme is combined with a well–defined parametric model. Whether the algorithm determining the sampling scheme — the production model, in Dawid's terminology — is known or unknown is irrelevant, because the statistical inference made from the inferential model depends on this scheme only through the (formally independent) observations and their (conditional) distributions. Thus, all we need is the concrete sequence of observations and the distributions that came out of the instrument manager's algorithm. What would have come out of it in other cases is irrelevant.

So far, we are following Dawid's prequential principle closely, in the sense that our principle can be considered a consequence of Dawid's prequential principle. It is not quite clear to me whether the converse is true or not. Dawid's prequential principle in its original version referred to a more general framework. The primary issue was prediction, not estimation of a parameter or test of a hypothesis. A parametric model was not a necessary ingredient, and at least to begin with there was no reference to the production model either. Later on, it seems to be an assumption that the distributions specified by the inferential model can somehow be derived from a production model as the conditional distributions of the observations, given the previous observations and (perhaps) other things from the past. In this sense, the prequential principle is — perhaps — more or less equivalent to the principle I have referred to as the "principle of analysing sequential experiments by non–sequential methods".

The best argument for this principle is probably the question "what can we conclude from the sequence of measurements if we don't know the instrument manager's algorithm?" (that is, we know the sequence of variances, but not how it was constructed). It is fairly obvious that it must be possible to do *something*, and by far the simplest — and probably the only — thing to do is to behave as if the number of measurements and their variances are determined in advance. But if this is somehow the correct thing to do, the next immediate question is "would it help us at all, and would it enable us to do something better, if we knew the algorithm?"

Vovk (1993) goes a step further in this (or a very close) direction by explicitely pointing out and attempting to solve a problem which, in our framework, can be stated as follows. What if the instrument manager does not use a welldefined algorithm, but makes his choices according to obscure ideas coming to him from who knows where? One thing is that we do not need to *know* the instrument managers algorithm, but if we cannot even *talk* about it, we don't have a production model, and how can we interpret the marginal distributions of the inferential model if not as conditional distributions in the production model? This opens a deep (perhaps interesting, perhaps fruitless) philosophical discussion of whether or not the behaviour of a conscious person can be assumed to be controlled by a (possibly randomized) algorithm. To avoid this discussion, Vovk introduced an entire new logic of probability, which in turn resulted in a further development of the ideas related to prequential inference (Dawid and Vovk 1999).

However, this discussion does not necessarily interfere with the ideas presented here. In our universe, the production model is assumed to exist. Our picture of the instrument manager as a person is only a trick, to emphasize that his decisions are beyond the experimenters control.

References.

Berger, J. O. and Wolpert, R. L. (1984).
The Likelihood Principle.
IMS Lecture Notes — Monograph Series, Vol. 6.

Birnbaum, A. (1962).
On the Foundations of Statistical Inference.
J.A.S.A. **57**, 269–326.

Cox, D. R. (1958).
Some Problems Connected with Statistical Inference.
Ann.Math.Stat. **29**, 357–372.

Dawid, A.P. (1984).
Statistical Theory. The Prequential Approach.
J.R.Statist.Soc. A **147**, 278–292.

Dawid, A.P. (1991).
Fisherian Inference in Likelihood and Prequential Frames of Reference.
J.R.Statist.Soc. B **53**, 79–109.

Dawid, A.P. and Vovk, V. G. (1999).
Prequential probability: principles and properties.
Bernoulli **5**, 125–162.

Dickey, D. A. and Fuller, W. A. (1981).
Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root.
Econometrica **49** *No. 4,* 1057–1072.

Engle, R. F. and Granger, C. W. J. (1987).
Co–integration and Error Correction: Representation, Estimation, and Testing.
Econometrica **55** *No. 2,* 251–276.

Johansen, S. (1991).
Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models.
Econometrica **59** *No. 6,* 1551–1580.

Oden, A. (1977).
Some Principles of Statistics. Ph.D. thesis, University of Gothenburg.

Vovk, V. G. (1993).
A Logic of Probability, with Applications to the Foundations of Statistics.
J.R.Statist.Soc. B **55**, 317–351.