Tue Tjur

Some Paradoxes Related to Sequential Situations

Summary. A classical "paradox" in statistical inference concerns a situation where a measurement of a quantity (a physical constant, say) is performed by an instrument selected at random among two — an extremely accurate intrument and an extremely inaccurate one. Orthodox Neyman–Pearson theory fails in this case, and the example is a standard argument for conditioning on ancillary statistics whenever this is possible.

More complex sequential versions of this situation are discussed. Several measurements are performed, and the accuracy of each measurement is a function of previous measurements. It is argued that such situations should be handled by ignorance of their sequential nature.

One such situation is equivalent to a simplified version of a problem from econometrics, that of testing autoregression coefficient = 1 in an AR(1) process with mean 0. Here the "non–sequential" approach represents a technical simplification, when compared to existing methods.

However, if the "principle of ignoring the sequential sampling plan" is followed strictly, other wellknown paradoxes appear.

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

Key words and phrases.
Statistical inference, ancillarity, sequential plan, cointegration.

0. Introduction.

The following is a slightly modified version of an example due to Cox (1958).

Situation no. 1

A measurement of an unknown quantity μ is performed with one of two instruments, a very inaccurate instrument with (normal) error variance 1*, and a very accurate one with (normal) 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 if 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. Clearly, as anyone who is not a theoretical statistician can see immediately, the circumstance that we might have used another instrument is entirely irrelevant for our conclusions from the measurement that we actually performed. The example is often cited as a standard argument for the principle that one should condition on an ancillary statistic whenever such 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 the same likelihood function, 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 strong likelihood principle, and since these methods have a tendency to question themselves by their dependence on a prior distribution, the statistical science really has a problem here. The philosophical justification of what we are actually doing is so week 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 one can set up for statististical 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 — without much discussion of abstract principles — how some of these "paradoxes" come up 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 this would force us to do absurd things in conflict with common sense. Furthermore, we shall show how one such situation comes up in a (simplified version of a) problem from econometrics, where the presense of the "paradox" questions the relevance of existing methods.

By making foundational inconsistencies visible, not only in imagined situations but also in statistical practice, we hope to contribute to the general confusion, which we assume to be the necessary driving force behind the thought–breaking ideas that someone is hopefully going to come up with sooner or later.

1. A simple sequential experiment.

We proceed with an example which is merely Cox's example once more in a sequential dress.

Situation 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 are performed, otherwise no further measurements are taken.

(Remark: Here and in the following, "measurements" are subsumed to be normally distributed).

Clearly, this is essentially our first situation. The equivalence becomes even more clear if the coin–tossing is assumed to be done before the first observation, so that we are merely selecting the sample size at random, by external randomization. However, this apparently innocent change of the order in which things are done is a crucial point in the examples to follow.

In our next scenario, again we perform a number of measurements of a quantity μ with variance 1. But the coin–tossing of situation 2 is replaced with a decision based on the first measurement.

Situation 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 (pre–determined) constant c, 999 additional measurements are performed, otherwise no further measurements are taken.

This is a proper sequential situtation, 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 (1 or 1000) had been decided in advance. One of the more convincing arguments goes as follows. Consider

Situation 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 scheme of situation 3.

However, since we cannot find a coin right away, we decide to perform the first measurement (which is to be done anyway), 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.

It should be noticed, that situation 3 is a simplified version of situations that are not at all artificial or irrelevant. In medical trials, the idea of a small pilot study before the big mashinery is turned on, is commonly accepted. Also situations like no. 4 could occur in practice. Suppose we know that a research fund will certainly support the long and expensive part of the study if the pilot investigation indicates a positive result, whereas other funds may or may not be willing to support the entire study without asking for a pilot investigation. Then, impatience could easily force us into a situation very close to no. 4.

2. A general sequential scheme.

Situation 3 is a special case of a more general situation, which we may explain in terms of an "instrument manager". The instrument manager is the person (or mashine, or algorithm) who decides for us which instrument to use next and when to stop the sequence of measurements. The decisions of the instrument manager may be based on our previous measurements, and also, if desired, on external randomization (but not on knowledge about μ, of course; assume that the instrument manager knows no more about μ than we do). We can formalize this situation as follows (disregarding the possibility of external randomization, which is not an important point).

Situation no. 5 (general sequential scheme)

A sequence

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

of measurements are performed. The variance of each measurement is a (known) function of previous measurements, and the normal distributions specified are conditional on 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, $\ldots$ by the assumption that we have $\sigma_i^2 = +\infty$ from a certain stage with probability 1. However, if this seems too complicated, it suffices to think of the case where n is fixed.

Again, we can argue that inference from this experiment should be performed exactly as if the variances σ_1^2, σ_2^2, $\ldots$ and the number of observations n were known in advance. An argument similar to our argument for the same principle in the special case of situation 3 goes as follows. Consider

Situation 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 situation 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 meaurement with that instrument is to be performed in any case — 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.

Finally, a coin is found and flipped.

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 then 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 must 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 well known that this attitude creates other problems. This will be recalled briefly in section 4. However, let us first take a look at one of the consequences of this "principle".

3. Testing autoregression coefficient = 1 in the AR(1).

Consider the following problem, which is a simplified version of a problem studied intensively by econometricians in the context of "cointegration", see e.g. 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 $x_0 \neq 0$. $X_1, \dots, X_n$ are generated recursively as

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

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

It is easy to transform this to a special case of situation 5. If we define $Y_i = X_i / X_{i-1}$, we have (conditionally on previous observations $X_1, \dots, X_{i-1}$)

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

Thus, we can think of each Y_i as a measurement of α with a variance $\sigma^2 / X_{i-1}^2 = \sigma^2/(x_0 Y_1 Y_2 \dots Y_{i-1})^2$ which is a function of the previous observations. Following the principle that these variances should be regarded as predetermined, we obtain (by ordinary weighted 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 usual least squares estimate of α, obtained by regression 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 be based on the statistic

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

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

However, this is not the way things are usually done. First of all, in a real example σ^2 would usually be unknown, and the autoregression equation would probably contain (at least) a constant term. The extra parameter σ^2 can be dealt with. In fact, we could have assumed in all our examples that variances were known only up to a common scale factor, that would not have made much difference. A constant term (and, perhaps, terms corresponding to periodic trends or covariates) can also be dealt with, though it makes things more complicated. The important difference comes from the fact that the sample distribution of U above, even under our assumptions (i.e. σ^2 known, no additional terms), 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”. But this is not a conditioning in the usual sense of this word, since that would involve (more or less) a conditioning on the observations themselves.

4. Discussion.

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 (cfr. Vovk 1993) where the formation of Cox's 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 sample distribution of the test statistic for $\alpha = 1$. However, it must not be forgotten that the idea of analysing any sequential experiment by non–sequential methods has its own traps or "paradoxes". The standard warning goes something like this. Consider

Situation no. 7

I.i.d. measurements $Y_1, Y_2, \dots \in N(\mu, 1)$ *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 for $\mu = \mu_0$. But 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.

The kind of paradoxes discussed here are not new. Some references related to this kind of problems and attempts to solve them are Oden (1977) and Dawid (1984, 1991). The purpose of the present paper is to isolate and clarify the nature of this kind of phenomena. Unfortunately, we can not present anything like a conclusion.

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.

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.

Tue Tjur
Institute of Mathematical Statistics
University of Copenhagen
Universitetsparken 5
DK-2100 Copenhagen
DENMARK