Chapter 5: What is a good inference?

In Chapter 3, we showed that conventional frequentist inferences, even when optimal, are poor when it comes to assessing the evidence for a given hypothesis relative to another. The question of how to make better inferences and whether conditioning improves the quality of an inference is implicit in all three of the papers described in Chapter 4. Cox argued that conditional infernces ought to be regarded as superior because they are less affected by irrelevancies and Welch's example came to be interpreted the same way (though Welch himself argued the opposite on theoretical grounds); Birnbaum argued that a certain kind of strong conditioning is necessary for good inference but showed that this requires us to give up frequentism.

In Chapters 5 & 6, we discuss the main points of contention that have arisen from these papers and the debates that they have generated. In order to add to the criteria on which to base judgements about the quality of inferential theories, we start by considering Buehler's gambling criterion for a good inference and seeing how it applies to the examples used by Welch, Cox and Birnbaum.

5.1 Buehler's gambling criterion.

Using conditional knowledge to win a game.

We have discussed the shortcomings of some unconditional inference methods in terms of the concept of 'relevance'. R. J. Buehler put forward an alternative criterion in 1959. In the paper, "Some validity criteria for statistical inferences", he suggested that we consider a game in which the odds of a particular outcome are based on whichever probability is emphasised by the method in question, and ask whether any betting strategy can win in the face of these odds, and how generally applicable such a strategy is. A strategy that is successful and widely applicable suggests that the conventional probability used in determining the odds is not wholly appropriate for that outcome, and this casts doubt on the validity of the method. In essence, a successful betting strategy can be seen as accessing extra (relevant) information not covered by the conventional probability.

Many, though not all, problems of inference lead to assertions of the type, "The probability that A is true is equal to α," or, "$P(A)=\alpha$". One may ask whether the person making this assertion should be willing to bet that A is true, risking an amount α to win $1-\alpha$, and should be equally willing to bet that A is false, risking $1-\alpha$ to win α, against an opponent who has exactly the same information as he and who is allowed to choose either side of the wager.[1]

The point is not whether $P(A)=\alpha$ is consistent with the data, but rather whether it encapsulates *all* of the relevant information available.

Buehler considered various requirements for good inferences in the form of interval estimation, noting that the conditioning controversy was a point of particular interest:

The appropriate reference set [sample space] has been a subject of controversy in testing situations (i.e. tests of significance and tests of hypothesis) … Some recent examples can be found in Cox [1958] and Cohen. It is to be noted that the present development has been based largely on problems of interval estimation. The usual translation of criteria to testing situations is of course possible in many cases. Thus certain testing situations have been treated implicitly in this work…"[2]

Buehler considered a game played by Peter, the proponent of an inferential rule R, and his challenger Paul.

Unknown conditions of the experiment, for example the value of the population mean μ, are conveniently referred to as the "state of nature" U (for "unknown"). The first player, Peter, has the familiar task of setting confidence

[1] Buehler (1959), p. 845.
[2] Buehler (1959), p. 861.

intervals. It is required that he formulate a rule R which determines the interval as a function of the observations. Then on the basis of the observations he makes a probability assertion

$$"P(A)=\alpha"$$

… The assertions may have validity as "confidence probabilities" or "fiducial probabilities", these being special cases… In order that the second player, Paul, have information equal to Peter's, it is required that he have knowledge of Peter's rule R as well as of the experimental conditions and observations. Paul adopts a strategy S based on R and on the experimental conditions, and consisting in the specification of two subsets C^+ and C^- of the observation space such that

for observations in $\begin{cases} C^+ \\ C^- \end{cases}$ **Paul bets that**

A is $\begin{cases} \text{true,} \\ \text{false,} \end{cases}$ **risking** $\begin{cases} \alpha \\ 1-\alpha \end{cases}$ **to win** $\begin{cases} 1-\alpha. \\ \alpha. \end{cases}$

It is not required that a bet must always be made; thus C^+ and C^- need not be exhaustive. To determine the winner of each bet, we postulate the existence of a referee who knows the true state of nature.[3]

To begin with, Buehler considered an undemanding validity criterion that he called *weak exactness*.

If the model is adequately specified, one should in principle be able to calculate the expected gain to Paul. For any fixed experimental conditions K the expected gain would be a function of (i) the state of nature U, (ii) Peter's rule R, and (iii) Paul's strategy S. Different criteria for the sensibility of Peter's rule might be put forward in terms of this expected gain… Suppose Paul's strategy is to bet consistently that A is false, regardless of the observations. Then if Paul's expected gain is zero for all U, Peter's rule R will be defined to be *weakly exact*.[4]

[3] Buehler (1959), p. 846.

[4] Buehler (1959), p. 846.

(If Paul always bets that A is true regardless of the observations and has *zero* expected gain for all U, R is also weakly exact.) Unconditional methods, such as those of Neyman and Pearson are weakly exact. Since Paul bets the same way every time, he cannot take advantage of any conditional properties (for instance) that R might produce, and the odds $\alpha : 1-\alpha$ accurately reflect the long run success rate of R. Someone who believed that Neyman-Pearson optimality corresponds to a maximal amount of information might bet this way, simply in accordance with the value α, in order not to risk a loss. If a betting strategy S can reliably return Paul a profit then this seems to indicate that he has access to information not accessed by the rule R. (If Paul can reliably make a loss this also suggests that he knows more about what is going on, though he makes strange use of this knowledge.)

How Neyman-Pearson optimality is reflected in the gambling scenario.

Neyman-Pearson hypothesis testing methods are weakly exact. In a game in which the odds are based on the appropriate failure rates (α or β), someone betting always-for or always-against the success of the method will break even in the long run. This is also true if they bet sometimes for and sometimes against the method in a random or arbitrary way not based on the data. This indicates that the error probabilities are appropriate in what we might call a ‘blindfolded’ sense, but can we say any more than this? The ‘optimality’ of the method is revealed by the fact that it can be used as the basis of a betting strategy that will be successful against any approach which is similar, in the sense that it is (i) weakly exact, and (ii) does not assign varying probabilities to any (proper) subsets of the sample space of the experiment – in other words, it is unconditional.

Imagine a non-optimal method, for instance, a hypothesis test in which the parameter of interest is the mean of a Normal population with known variance and the test statistic is the sub-optimal sample median. The proponents of this method find a rejection region that yields a significance level of α and a power of κ. They can use these values to give you appropriate odds for a gamble on whether their test results are right or wrong (the odds will vary depending on whether it is actually H or K which is true). If you were to bet in a blindfolded manner, either ‘always right’, ‘always wrong’ or ‘randomly’, you would break even in the long run; this shows that their method, while sub-optimal, is weakly exact. However, you could create a winning strategy based on the optimal Neyman-Pearson method, as follows. Find the optimal rejection region with the same significance level, α; this will have power $\kappa^+ > \kappa$ (since it is the most powerful test). If you bet that H is true whenever the data is outside this region and that K is true whenever the data is inside this region, then when H is true you will (in the long run) break even but when K is true you will tend to make a profit. If the test statistic is continuous, you can modify this method so that it delivers a profit even if H is always true. Suppose $\kappa^+ = \kappa + \varepsilon$, for some $\varepsilon > 0$, then instead of using an (optimal) rejection region with a power of κ^+, use one with a power of $\kappa + (\varepsilon / 2)$; your power will still be greater than the value on which the odds are based, but by reducing the power you will have raised the significance level which is (for your region) now equal to $\alpha^+ > \alpha$. If you make bets based on this rejection region, you will tend to win in the long run for any proportions p and $1 - p$ in which H and K occur, $0 \le p \le 1$. Thus, Buehler’s gambling criterion can be interpreted as showing that the approach of Neyman and Pearson is the *optimal unconditional* approach.

Stronger criteria.

Bueher’s *strong exactness* is a much more demanding criterion than weak exactness. A rule R is **strongly exact** if Paul’s expected gain is zero regardless of the state of nature and for all possible strategies, S. “In other words: Whatever the true state of nature and whatever strategy Paul may use, the expected gain to Paul is zero. It is essentially equivalent to write $P(A \mid C) = \alpha$ for all U, C.”[5]

Buehler regarded strong exactness as unrealistic:

[5] Buehler (1959), p. 856.

It appears to be impossible to satisfy the very stringent condition of strong exactness except in rather special cases, e.g., in Bayesian estimation where the condition "all U" is in fact no requirement at all since U is not a variable. Thus strong exactness is not so much a practical requirement as a goal toward which one might strive even though it cannot actually be reached.[6]

R is strongly exact only if there is no betting strategy that will deliver a non-zero expected gain for *any* state of nature (for instance, any value of the parameter of interest). We can define a requirement that is weaker than this, but still substantially stronger than *weak exactness*, as follows.

There should be no ***single*** *strategy that can deliver a strictly positive expected gain to Paul for* ***all*** *values of* θ.

A rule does not need to be strongly exact in order to satisfy this requirement. This criterion will discredit Peter's rule based on "$P(A)=\alpha$" whenever there exists an event C such that $P_\theta(A\,|\,C)<\alpha<P_\theta(A\,|\text{ not }C)\;\;\forall\theta\in\Theta$. Thus, conditional probabilities can play a key role in the application of this criterion. (Buehler defined the concepts *biased relevant subset* and *biased semi-relevant subset*, which have since been widely used in discussions of conditional issues, and noted that such subsets could be used as the basis of betting strategies.)

The examples of Welch and Cox are now regarded as instances where the optimal (unconditional) approach is inferior to a conditional approach. In the following sections we show that the above gambling criterion confirms this, since a gambling strategy, based on the conditional probabilities, can be used to beat the error probabilities defined by the unconditional approach. We also look at a Binomial example (due to Birnbaum) that has surprising conditional features, and confirm that conventional inferences on the Bernoulli parameter, $p=P(success)$, are vulnerable to a betting strategy, which, therefore, casts doubt upon the validity of the usual inference.

[6] Buehler (1959), p. 857.

Gambling on the Uniform example.

Let us use the gambling criterion to examine the Uniform case. We can construct Neyman-Pearson optimal confidence intervals for the centre of the Uniform distribution (θ), based on the same approach as used in the previous chapter to find an optimal hypothesis test. When $n = 2$, the formula for a $100(1-\alpha)\%$ confidence interval is:

$$\begin{cases} (m \pm (1-r)/2), & r \geq \sqrt{\tfrac{\alpha}{2}} \\ (m \pm [(1-r)/2 - (\sqrt{\tfrac{\alpha}{2}} - r)]), & r < \sqrt{\tfrac{\alpha}{2}}. \end{cases}$$

The conditional significance levels of the optimal hypothesis test vary with r, and so do the conditional coverage properties of the optimum confidence interval. Formerly we argued that the 'sub-optimal', conditional results are more relevant than the optimal results; if this is true, we might expect a betting strategy, based on the conditional results, to beat the probabilities defined by the optimal method.

The coverage of the $100(1-\alpha)\%$ confidence interval, conditional on the observed ancillary statistic r, is:

$$\begin{cases} 100\%, & r \geq \sqrt{\tfrac{\alpha}{2}} \\ 100\left\{1 - \dfrac{(\sqrt{\tfrac{\alpha}{2}} - r)}{2(1-r)}\right\}\%, & r < \sqrt{\tfrac{\alpha}{2}} \end{cases}$$

We will use 90% confidence intervals, i.e. $\tfrac{\alpha}{2} = \tfrac{1}{20}$. We could base a betting strategy on whether or not $r < \sqrt{\tfrac{\alpha}{2}}$, but have instead opted for a slightly more sophisticated strategy. The conditional coverage of the optimal interval is exactly 90% only when $r = r' = \tfrac{(\sqrt{20}-1)}{19}$. When r is less than this value the conditional coverage is less than 90%, and when r is more than this value the conditional coverage is more than 90%.

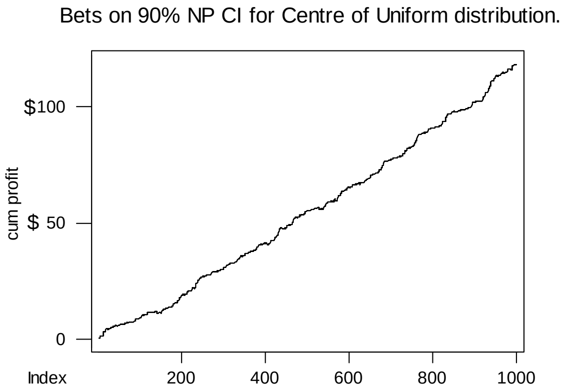
With this in mind, we use the following betting strategy:

$$\begin{cases} \text{Bet that } \theta \in \text{CI, whenever } r > r' \\ \text{Bet that } \theta \notin \text{CI, whenever } r < r'. \end{cases}$$

Since the value 90% is associated with the event '$\theta \in \mathrm{CI}$', the defender of the unconditional method should be prepared to let us bet on this event, risking 90c to win 10c, or bet against this event, risking 10c to win 90c.

We ran a simulation of 1000 samples, using the above strategy (θ was *zero*). The possible profits for any single bet are (in dollars) -0.9, -0.1, +0.1, and +0.9. The cumulative profits over 1000 bets are shown in the graph below.

Figure 5.1



The total profit at the end of the 1000 bets was $119.60, giving an average profit on each bet of 11.96c. We can see how this strategy worked by comparing the betting strategy with the performance of the confidence intervals. The following table shows our 1000 confidence intervals, generated by random samples, cross-tabulated according to (i) whether or not they were successful (contained θ) and (ii) which type of bet was made in that case, i.e. whether r was greater or less than r'.

Table 5.1

	$\theta \in \mathrm{CI}$	$\theta \notin \mathrm{CI}$	
Bet that $\theta \in \mathrm{CI}$	678	5	683
Bet that $\theta \notin \mathrm{CI}$	230	87	317
	908	92	1000

Of the 1000 intervals generated, 908 contained θ giving an observed coverage rate of 90.8%, so the performance of the confidence intervals was slightly better on this occasion than it would be in the long run (90%). However, when r was 'large' and we bet in favour of the confidence interval, the success rate was 678/683 > 99%, whereas when r was small, the success rate was only 230/317 $\approx$72%, reflecting the fact that the coverages of the intervals, *conditional* upon r being greater or less than r', are not 90%. It is still true that 90% of the intervals overall contain θ but it is not true that we know no more than this; for some of the intervals (those where $r > \sqrt{\frac{\alpha}{2}}$) we know with certainty that the interval contains θ (though we did not use this knowledge in the above strategy) but even when we do not have this level of certainty, we can site the interval in a smaller subset with a different identifiable coverage. This game demonstrates a way in which such knowledge can be useful. However, were someone to use the two conditional coverage values as the appropriate ones for judging the success of the confidence intervals (splitting the sample space into the two sub-spaces $\{(m,r): r < r'\}$ and $\{(m,r): r > r'\}$) their contention would be vulnerable to precisely the same attack via a betting strategy, since we could split either of these subspaces up into two more subsets and use these as the basis for bets which would make a profit even in a game in which the odds are based on the original conditional values. It is therefore clear that the value of r can be used to produce a hierarchy of betting strategies that become increasingly powerful as we home in on the *exact* value of r. These strategies work regardless of the value of θ (Buehler's U) and even regardless of whether θ remains the same throughout the sampling process or varies. It is apparent that the Neyman-Pearson intervals infringe the requirement defined in the previous section, even though it is much less stringent than *strong exactness*. A strategy that worked for just one possible value of θ would have shown that the method was not strongly exact, but here we have a technique that works for all (or varying) θ.

Gambling on the two-stage example.

Let us look at Cox's two-stage example and consider a gambling strategy to beat the optimal method. This time we will consider a test of two simple hypotheses in order to use the exact values we gave in *Example 4.2*. Recall that $X_a \sim N(\mu, \sigma_a^2)$, the test was carried out using a 5% level of significance, the hypotheses were H: $\mu = 0$ versus K: $\mu = 5$ and the ancillary statistic A takes the values 1 or 2 with probability of ½ each, and

$$\sigma_a^2 = \text{var}(X_a) = \begin{cases} 4, & a = 1 \\ 1, & a = 2, \end{cases}$$

indicating that 'machine 2' is more accurate than 'machine 1'.

A test of the form *Reject H in favour of K whenever*

$$x > \begin{cases} 2.62906, & a = 1 \\ 2.53227, & a = 2 \end{cases}$$

has a significance level of $\alpha = 5\%$ and power of $\kappa = 93.7643\%$ and is the most powerful 5% test.

We consider what happens under both states of nature: $\mu = 0$ and $\mu = 5$. The punter bets on whether the 'test result' (Accept H or Reject H) is right or wrong in terms of the truth about μ. The punter may bet that the result is right risking 95c to win 5c if H is true and risking (\$) κ to win (\$)$1-\kappa$ if K is true. We keep the winnings and losses secret from the punter during the game in order to give them no clue about the true state of nature. We first ran 1000 trials where H was always true and then 1000 trials where K was always true. Since X_2 is a more precise estimator of μ than X_1 (which has a larger variance), the punter used the following strategy:

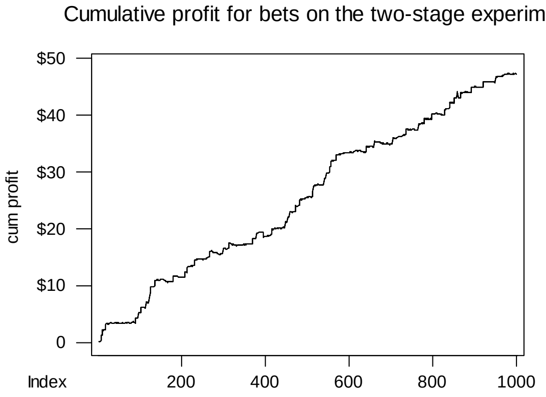
$$\begin{cases} \text{Bet that the test result is 'right' if } a = 2 \\ \text{Bet that the test result is 'wrong' if } a = 1. \end{cases}$$

.

This strategy should work under both H and K since $\alpha_2 < \alpha < \alpha_1$ and $\beta_2 < \beta < \beta_1$.

When H is true, the profits that can be made on each bet are -95c, -5c, 5c, and 95c. The graph below shows the cumulative profit over the 1000 trials where H was true.

Figure 5.2


Cumulative profit for bets on the two-stage experiment.

The final profit after 1000 bets was $47.40 giving an average gain per bet of 4.74c. The breakdown of test performance was as follows; the profit on a single bet of each kind is shown in brackets in each cell.

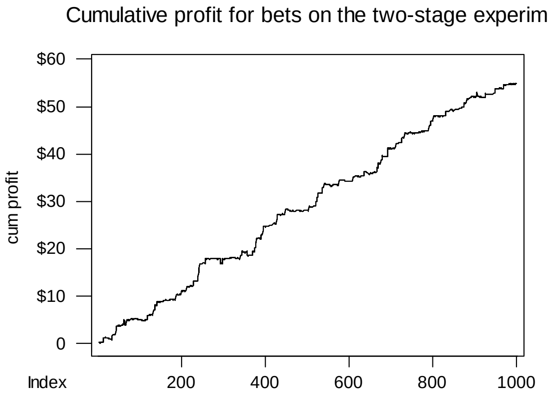
Table 5.2

	Accept H (Right)	Reject H (Wrong)	
Bet result right $(a=2)$	512 [+5c]	2 [-95c]	514
Bet result wrong $(a=1)$	438 [-5c]	48 [+95c]	486
	950	50	1000

Overall 50/1000 = 5% of the test results were wrong, exactly mimicking the long run error rate, but when $a=2$ and the punter always bet that the result was right, the error rate was 2/514 $\approx 0.4\%$ (conditional long run rate: 0.567%) and when $a=1$ and the punter always bet that the result was wrong, the error rate was 48/486 $\approx 9.876\%$ (conditional long run rate: 9.433%). Thus the punter managed to utilise the fact that the test result is more likely to be correct when $a=2$.

When K was true this strategy also worked.

Figure 5.3



In this case the final profit after 1000 bets was $55.0023 (we used κ accurate to 6dp). The test outcomes are shown below, as are the profits for each type of bet.

Table 5.3

	Accept H (Wrong)	Reject H (Right)	
Bet result right $(a = 2)$	4 [-93.76c]	488 [+6.24c]	492
Bet result wrong $(a = 1)$	60 [+93.76c]	448 [-6.24c]	508
	64	936	1000

Overall 936/1000 = 93.6% of the tests yielded the right result consistent with the power of 93.76%, but for $a = 2$ the success rate was 488/492 $\approx 99.19\%$ (conditional long run rate: 99.32%) and for $a = 1$ the success rate was only 448/508 $\approx 88.19\%$ (conditional long run rate: 88.21%). Insight into how the success rates varied between the two subsets allowed the punter to win against the Neyman-Pearson based odds. Since the same strategy works when either H or K is true, it will also work in a long series of trials where sometimes H and sometimes K is true. Neither here, nor in the Uniform case, have we needed to avoid making a bet in any instance, although Buehler argued that this latitude should be granted to the punter.

Gambling on the Binomial example.

We turn now to an example given by Birnbaum in the same paper (1962) that contained his theorem. Birnbaum identified an ancillary statistic for certain hypothesis tests on the Bernoulli parameter $p = P(success)$ where the test statistic has a Binomial distribution (i.e. the number of trials is fixed in advance). One of the interesting features of this example is that the ancillary statistic is ‘internal’, that is, unlike the ‘coin toss’ statistic in Cox’s two-stage scenario, it is *not* easily identifiable as a separate part of the experiment. In this respect it is similar to the range (R) in the Uniform case, but the Uniform example is somewhat artificial (and, as we shall see, not well understood) whereas Binomial test statistics are very frequently used to test hypotheses about probabilities. The other interesting point is that, despite the long history of this method, Birnbaum was the first to notice that it sometimes has this structure.

Let $X \sim Bin(n, p)$ where p is the parameter of interest. For instance, if we carry out n independent tosses of a coin where $p = P(head)$ is unknown but constant for all trials, we base our inference about p on $X = \#(heads)$.

Let $n \geq 2$ and suppose we are interested in testing two simple hypotheses of the form H: $p = \theta$ versus K: $p = 1 - \theta$ for some value $0 < \theta < 1$. For instance, we could imagine that at some time in the past we have generated data (x) from a Binomial set-up where the two probabilities were θ and $1 - \theta$ (θ known), but cannot now recall which of the two events we defined as ‘success’ and which as ‘failure’. In this case, if we confine ourselves to looking at two specific hypotheses of this form, the parameter space of p is given by $\{\theta, 1 - \theta\} = \Theta_B$, where θ is some fixed value in $(0,1)$.

The discrete statistic $A = | X - \frac{n}{2} |$ is ancillary with respect to the parameter space Θ_B since it has the same distribution under H and under K; this is true regardless of the value of θ, as shown below.

$A = a$ if and only if X equals $\frac{n}{2} - a$ or $\frac{n}{2} + a$, hence

$$
\begin{aligned}
&P(A=a)=P(X=\tfrac{n}{2}-a)+P(X=\tfrac{n}{2}+a)\\
&=\binom{n}{\frac{n}{2}-a}p^{(n/2)-a}(1-p)^{n-[(n/2)-a]}+\binom{n}{\frac{n}{2}+a}p^{(n/2)+a}(1-p)^{n-[(n/2)+a]}\\
&=\binom{n}{\frac{n}{2}-a}\{p^{(n/2)-a}(1-p)^{(n/2)+a}+p^{(n/2)+a}(1-p)^{(n/2)-a}\}.
\end{aligned}
$$

Since this expression is symmetric in p and $1-p$, it makes no difference to its value whether we let $p=\theta$ or $p=1-\theta$ and thus $P_H(A=a)=P_K(A=a)$, showing that A is ancillary.

Note, as an aside, that the likelihood ratio statistic in this case is

$$LR(X)=\frac{\theta^X(1-\theta)^{n-X}}{(1-\theta)^X\,\theta^{n-X}}$$

and the statistic A can be written as $c\cdot|\ln LR(X)|$, where $c=\{2|\ln(\frac{\theta}{1-\theta})|\}^{-1}$ is fixed for any given Θ_B and, hence, $|\ln LR(X)|$ is itself ancillary for $p\in\Theta_B$. In future chapters we will observe a number of cases where $|\ln LR(\underset{\sim}{X})|$ is an ancillary statistic.

Example 5.1

We look at the case $n=20$, with hypotheses H: $p=0.7$ versus K: $p=0.3$. (The features identified in this case occur for all values of n no matter how large or small (≥ 2).) The support of X is $\{0,1,2,...20\}$ and $A=|X-10|\in\{0,1,2,...10\}$. In Cox's two-stage example, the ancillary statistic A is a measure of the reliability (sometimes called a 'precision index'[7]) of X. This is also true here; the larger a is, the more clear-cut the evidence we get regarding H versus K from x; for example, if $a=1$ then x is either 9 or 11 ($\hat{p}=\frac{x}{n}$ is either 0.45 or 0.55), this is not very helpful when it

[7] Throughout this work we will use 'A is a precision index' to draw attention to the fact that A, in some sense, orders the data according to how informative it is for choosing between the two hypotheses. In many cases with a high level of symmetry, it is easy to argue that A partitions the sample space into subsets, each of which contains all $\underset{\sim}{x}$-values with a certain *precision*, i.e. ability to distinguish between the hypotheses. For a much more rigorous tretment of 'precision index', see Buehler (1982).

comes to deciding whether p is 0.7 or 0.3; on the other hand, when $a=8$ then x is either 2 or 18 ($\hat{p}$ is either 0.1 or 0.9) and both these values are much more consistent with one of the hypotheses than the other.

A standard test of H versus K might use the rule *Reject H whenever* $x \le 9$ since this rule gives a significance level of $\alpha = P_H(X \le 9) = 1.71448\%$ and a power of $\kappa = P_K(X \le 9) = 95.2038\%$ and is Neyman-Pearson optimal for this significance level.

Since A can take eleven different values, we could split the sample space up into eleven subsets with different levels of reliability, however for the purpose of creating a betting strategy to test the relevance of α and κ, we only need to distinguish two different levels of reliability (since we can only bet *for* or *against* the test result), thus we define $A^* = \begin{cases} 0, & \text{if } A \le 4 \quad (\text{i.e. } X \in \{6,7,\ldots 14\}) \\ 1, & \text{if } A \ge 5 \quad (\text{i.e } X \in \{0,\ldots 5,15,\ldots 20\}). \end{cases}$

Since A^* is a function of A, it is also ancillary with respect to Θ_B. When $A^*=0$ we have less informative data than when $A^*=1$, and so it makes sense to bet that the test result is wrong when we see $a^*=0$ and that it is right when we see $a^*=1$. We randomly generated 1000 samples under each hypothesis with the following results; the appropriate profit for each result/bet combination is shown (rounded-off) in brackets.

a) H is true $(X \sim Bin(20, 0.7))$.

Some relevant probabilities for this scenario are:

- $\alpha = P(\text{Result wrong}) = 1.71448\%$
- $P(A^* = 0) = 58.3586\%$
- $P(\text{Result wrong} \mid A^* = 0) = 2.93048\%$
- $P(\text{Result wrong} \mid A^* = 1) = (1.03119 \times 10^{-2})\%$.

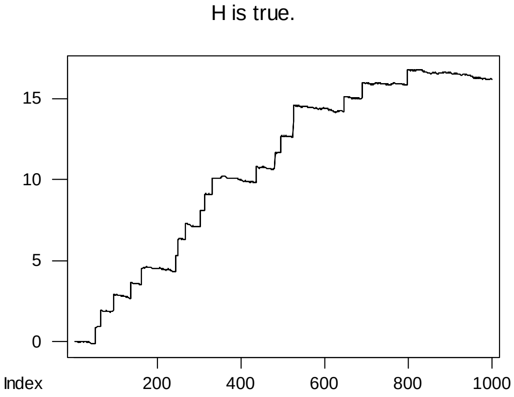
Table 5.4

	Result wrong $(x \leq 9)$	**Result right** $(x \geq 10)$	**Total**
Bet 'wrong' $(A^* = 0)$	19 [+0.983]	561 [-0.017]	**580**
Bet 'right' $(A^* = 1)$	0 [-0.983]	420 [+0.017]	**420**
Total	**19**	**981**	**1000**

The four relative frequencies corresponding to the probabilities cited above are (respectively): 19/1000 = 1.9%, 580/1000 = 58%, 19/580 = 3.3% and 0/420 = 0%; these values show that there is nothing particularly unusual about the 1000 samples we took. The total profit made by the end of 1000 bets is +16.214 units; the strategy is very successful because of the variation in error probability for the two different cases described by A^*.

The cumulative profit on the 1000 bets is shown below.

Figure 5.4



The same strategy works when K is true.

b) K is true ($X \sim Bin(20, 0.3)$).

Some relevant probabilities for this scenario are:

- $\beta = P(\text{Result wrong}) = 4.7962\%$
- $P(A^* = 0) = 58.3586\%$
- $P(\text{Result wrong} \mid A^* = 0) = 8.21112\%$
- $P(\text{Result wrong} \mid A^* = 1) = (1.03119 \times 10^{-2})\%$

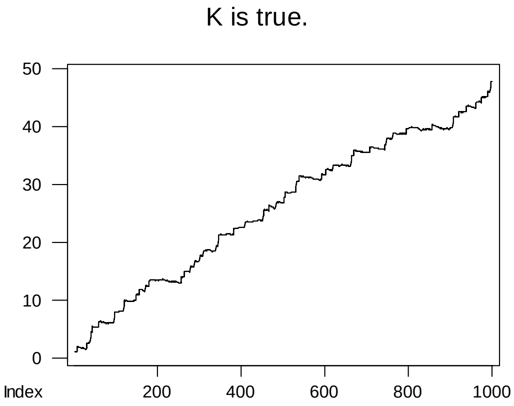
Table 5.5

	Result wrong $(x \leq 9)$	**Result right** $(x \geq 10)$	**Total**
Bet 'wrong' $(A^* = 0)$	56 [+0.952]	528 [-0.048]	**584**
Bet 'right' $(A^* = 1)$	0 [-0.952]	416 [+0.048]	**416**
Total	**56**	**944**	**1000**

The four relative frequencies corresponding to the above probabilities are (respectively): 56/1000 = 5.6%, 584/1000 = 58.4%, 56/584 = 9.6% and 0/416 = 0%.

The cumulative results of the 1000 bets are shown below.

Figure 5.5



Again, since the strategy works under both H and K, it will work for the case where the true state of nature varies between the two hypotheses. Although the optimal test is good in terms of the overall error probabilities, it fails to utilise the fact that some data are more informative than others.

The fact that this ancillary statistic only exists in cases where the two hypothesised values of p sum to *one*, means that it is of very limited application, particularly since tests often involve the null hypothesis $p = \frac{1}{2}$. We will show later that the structure utilised in this example exists much more generally in the case of continuous variables.

5.2 Principles and controversies.

Cox (1958) argued that attention to relevance should lead us to perform (frequentist) inferences conditional on the value of an ancillary statistic. He advocated a principle that we will call the *restricted conditional principle* although his argument, based on the notion of 'relevance', was sufficiently strong to justify the *unrestricted conditional principle* (see below). Birnbaum (1962) showed that the latter version of the CP, when coupled with the *sufficiency principle*, entails the *likelihood principle* and thereby rules out frequentist inference of any kind. The publication of these works, within four years of each other, provoked considerably discussion about inferential principles in the context of frequentism. In this section we examine some of the most contentious points, including: the validity of Birnbaum's theorem and the light it throws upon the connection between various principles; the differences between the two versions of the conditional principle; problems that arise from using the restricted conditional principle; and the suggestion that the unrestricted conditional principle is at odds with the sufficiency principle. Finally we give a brief survey of the confused attitudes towards conditional inference today.

The proof of Birnbaum's theorem.

Birnbaum's theorem showed that the sufficiency and conditional principles together entail the likelihood principle. This result is often viewed as surprising because the sufficiency principle is of long standing and the conditional principle is compelling and not obviously at odds with frequentist inference, whereas the likelihood principle is well known to conflict with (any) frequentist approach. Cox had argued that Neyman-Pearson theory should be modified and take on a more Fisherian character through conditioning but there was no suggestion that this would lead to the abandonment of frequentist inference altogether. In order to understand why this theorem is true and what misconceptions have lead many to be surprised by it, we need to look at the proof of the theorem. Our discussion is modelled on the proof given by Berger & Wolpert and will be put in the context of a particular example.

The 'two stopping rules' example.

Suppose we are interested in the Bernoulli parameter, $p = P(head)$, in a series of coin tosses where the result of the i^{th} coin toss is given by[8] $V_i \sim Bern(p)\ (\forall i = 1,..n)$ and $V_1,\ldots,V_n$ are independent. Now consider two different experiments.

Experiment 1 (E_1).

In E_1 the number of coin tosses, n, is fixed in advance, and hence if $X = \sum_{i=1}^{n} V_i = \# heads$, then $X \sim Bin(n,p)$ and the likelihood of X is given by:

$$L_1(x;p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \forall p \in (0,1), \forall x \in \{0,1,...n\}.$$

[8] That is, $V_i = \begin{cases} 1, \text{ if } i^{\text{th}} \text{ toss results in a 'head'} \\ 0, \text{ if } i^{\text{th}} \text{ toss results in a 'tail'.} \end{cases}$

Experiment 2 (E_2)**.**

In E_2, we use a different stopping rule: the coin continues to be tossed until t tails have occurred, where t is a known positive integer fixed prior to the experiment. Hence $Y = \#heads$ has the likelihood:

$$L_2(y;p) = \binom{y+t-1}{t} p^y (1-p)^t, \quad \forall p \in (0,1), \forall y \in \{0,1,2,3,....\}.$$

Now consider the case where we observe exactly the same numbers of heads, tails, and tosses $(n-t,\ t,$ and n, respectively) from the two experiments E_1 and E_2. Then $L_1(x = n-t; p)$ is proportional to $L_2(y = n-t; p)$ with the same constant of proportionality for all p since

$$\begin{aligned} L_2(n-t;p) &= \binom{n-1}{t} p^{n-t}(1-p)^t \\ &= \frac{(n-t)}{n} \times L_1(n-t;p). \end{aligned}$$

Thus, by the likelihood principle, we should infer the same about p from observing $n-t$ heads out of n tosses from E_1 as from observing the same number of heads out of the same number of tosses from E_2 regardless of the difference between the two stopping rules.

This is a point of contention between frequentists and supporters of the likelihood principle. When we consider two different stopping rules where the probability of a given result – eg. (#heads, #tails) – is the same up to a constant of proportionality unaffected by the value of the parameter of interest, then according to the likelihood principle, we should make the same inference regardless of the stopping rule. This position is also sometimes argued from the common sense perspective that a result such as *14 heads & 6 tails out of 20 tosses* should be able to give us an inference about $P(head)$ independent of the stopping rule used.

Frequentist inferences are highly sensitive to the stopping rule because the constant of proportionality and the 'more extreme' unobserved values critically affect tail-area measures such as the p-value or significance level. For instance, if we want to test the null hypothesis $p = \frac{1}{2}$ against any larger alternative value, then the p-value of the result '14 heads, 6 tails' from E_1 is:

$$\sum_{i=14}^{20} \binom{20}{i} \left(\frac{1}{2}\right)^{20} \approx 5.77\%,$$

whereas the p-value of the same result coming from E_2 is:

$$\sum_{j=14}^{\infty} \binom{j+5}{5} \left(\frac{1}{2}\right)^{j+6} \approx 3.18\%.$$

Thus the likelihood principle is breached by the use of p-values.

If Birnbaum's theorem is true and the sufficiency and conditional principles imply the likelihood principle, then we can only satisfy the sufficiency and conditional principles by inferring the same thing from '14 heads & 6 tails' regardless of the stopping rule used[9]. Thus, using p-values must breach either the conditional or the sufficiency principle.

The sufficiency principle can be applied only to a given, single experiment. It is easy to show that the conventional inference does not breach the sufficiency principle for either E_1 or E_2. Now define a third experiment, E^*, as follows.

The experiment, E^*, has two stages:

i. Observe one roll a fair die, and
ii. If the die produces an odd number, perform E_1 with $n = 20$, and if the die produces an even number, perform E_2 with $t = 6$.

Let A record the result of stage (i) as follows:

$$A = \begin{cases} 1, \text{ if die result is } \textit{odd} \\ 2, \text{ if die result is } \textit{even.} \end{cases}$$

[9] Within that set of stopping rules that produce the same likelihood ratio.

The outcome of this experiment is then of the form (a,u) where $u=\begin{cases} x, a=1 \\ y, a=2. \end{cases}$

Now consider the two possible outcomes $(1,14)$ and $(2,14)$. The first indicates that the die result was *odd* and E_1 was performed resulting in 14 heads out of 20 tosses, the second indicates that the die result was *even* and E_2 was performed also resulting in 14 heads and 6 tails. The likelihoods of these two outcomes are:

$$L^*((1,14);p)=\tfrac{1}{2}L_1(14;p)=\tfrac{1}{2}\binom{20}{14}p^{14}(1-p)^6, \forall p$$
$$L^*((2,14);p)=\tfrac{1}{2}L_2(14;p)=\tfrac{1}{2}\binom{19}{5}p^{14}(1-p)^6, \forall p.$$

The likelihoods of these two outcomes are proportional. Whenever two outcomes of an experiment have proportional likelihood values, it follows that the value of the (minimal) sufficient statistic is the same for both outcomes, and hence, by the *sufficiency principle*, the same inference should be drawn from both results (see footnote for more detail[10]). Therefore, for experiment E^*, the outcomes (1,14) and (2,14) should, by the *sufficiency principle*, produce the same inference. What about experiments E_1 and E_2?

The composition of the ‘umbrella’ experiment E^* is precisely that described in the preamble to the conditional principle (see §4.4). The conditional principle states that under certain circumstances[11], embedding an experiment such as E_1 or E_2 in a larger experiment should not affect the way in which we interpret the results of the sub-experiment, i.e. we should interpret the data the same way regardless of whether E_i is embedded or not. Thus the outcome of E_1 (or E_2) should give rise to the same inference regardless of whether E_i is performed in isolation or as part of an umbrella

[10] To confirm this, note that the statistic $s(a,u)=\begin{cases} (a,u), u\neq 14 \\ u, \quad u=14 \end{cases}$ is minimal sufficient, since the likelihood ratio of any (a,u) (for any two values of p) is a function of $s(a,u)$ only. The likelihood function is as follows:

$$L^*((a,u);p)=\tfrac{1}{2}\{(2-a)\binom{20}{u}p^u(1-p)^{20-u}+(a-1)\binom{u+5}{u}p^u(1-p)^6\}.$$

[11] Where, $\forall i$, $P(E_i)$ is the same $\forall\theta\in\Theta$.

experiment such as E^*. From this it follows, by the *conditional principle*, that the outcome $x = 14$ from E_1 should lead to the same inference about p as the outcome $(a,u) = (1,14)$ from E^*, and also that the outcome $y = 14$ from E_2 should lead to the same inference about p as the outcome $(a,u) = (2,14)$ from E^*. The sufficiency and conditional principles give the following requirements (where '$\equiv_p$' denotes 'requires the same inference about p as:'):

a) {(1,14) from E^*} $\equiv_p$ {(2,14) from E^*} by the SP,

b) {(1,14) from E^*} $\equiv_p$ {$x = 14$ from E_1} by the CP,

c) {(2,14) from E^*} $\equiv_p$ {$y = 14$ from E_2} by the CP.

Hence it follows (from the transitivity of this relation) that:

d) {$x = 14$ from E_1} $\equiv_p$ {$y = 14$ from E_2} as required by the LP.

The general proof of Birnbaum's theorem follows the pattern for this example. For a single experiment the sufficiency and likelihood principles are identical but the likelihood principle can extend over more than one experiment. By defining an umbrella experiment we can bring the sufficiency principle into operation on outcomes that have proportional likelihoods but are associated with different experiments now defined as sub-experiments under the umbrella of a single two-stage experiment. By making the umbrella experiment consistent with the requirements of the conditional principle, we can use that principle to insist that the outcome of any experiment give rise to the same inference even when it is (as it were) removed from under the umbrella experiment, i.e. when it is performed in isolation. From this it follows that outcomes with proportional likelihoods must lead to the same inference even when they are associated with different experiments; this principle is the likelihood principle. Frequentist inference forces the outcome of an experiment to change its interpretation once it becomes part of an umbrella experiment like E^*. In the case of (unconditional) Neyman-Pearson inference, this is because the aim is to maximise the power of the umbrella experiment rather than that of the relevant sub-experiment; in fiducial inference, it is because Fisher will only condition on a statistic which is part of the minimal sufficient statistic, which the variable A is not.

Two conditional principles.

Several conditional principles have been put forward over time; much of our discussion in this chapter will centre on the version favoured by Birnbaum, but there is another that has also attracted support. Conditional principles require us to condition on the observed value of a particular statistic, usually called 'ancillary' – a term coined by Fisher in 1925. Suppose that the mathematical structure of an experiment is equivalent to that of a two-stage experiment where the result of the first stage is a (the observed value of A) and the values in the support of A correspond to different sub-experiments performable in the second stage; a conditional principle requires us to interpret the result of the two-stage experiment the same way that we would have interpreted the result of the second stage sub-experiment, had it stood alone. Different conditional principles arise from different definitions of ancillarity. Fisher's original concept was somewhat vague; Buehler (1982) refers to the "…characteristic trail of intriguing concepts but no definition."[12] What Fisher apparently had in mind[13] was that $(A,\hat{\theta})$ be a minimal sufficient statistic for $\theta \in \Theta$ where $\hat{\theta}$ is the maximum likelihood estimator of $\theta \in \Theta$ and the distribution of A is the same for all $\theta \in \Theta$, thus the value of a is ancillary (hence the name) *to* the maximum likelihood estimate *in* the minimal sufficient statistic. Most definitions of ancillarity require that A have the same distribution for all θ so that, alone, A tells us nothing about θ. Basu (1959) adopted this as the sole criterion of ancillarity as do Berger & Wolpert[14] and Birnbaum. However Cox (1958) (and later Cox & Hinkley (1974)) required that A should also be part of a minimal sufficient statistic (MSS) without requiring that the other part necessarily be the maximum likelihood estimator.

Fisher has formalized this notion in his concept of ancillary statistics … it is convenient to state a slight modification of the original definitions. Let m be a

[12] Buehler (1982), p. 581.
[13] See Fisher (1956), pp. 156-159 and Basu (1964).
[14] Berger & Wolpert, p. 13.

minimal set of sufficient statistics for the unknown parameter of interest, θ, and suppose that m can be written (t, a) where the distribution of a is independent of θ, and that no further components can be extracted from t and incorporated in a. That is, we divide, if possible, the space of m into sets each similar to the sample space, and take the finest such division, assumed here to be unique subject to regularity conditions. Then a is called an ancillary statistic and we agree to make inferences conditionally on the observed a.[15]

Fisher's definition satisfies Cox's requirements but not vice versa, since Cox's statistic t need not be the maximum likelihood estimator of θ.

Definitions of conditional principles.

The **unrestricted conditional principle (CP)** states that we should condition on a statistic that is ancillary in the sense that **it has the same distribution for all** $\theta \in \Theta$.

The **restricted conditional principle (RCP)** states that we should condition on a statistic that is ancillary in the sense that **it has the same distribution for all** $\theta \in \Theta$ ***and* is a function of the minimal sufficient statistic for** $\theta \in \Theta$.

In the next two sections we will address the questions:

I. How much difference does our choice of conditional principle make?

II. What are the arguments for choosing one principle or the other?

Each version is implicitly associated with one of two competing inferential theories about which writers often have strong views. It is therefore frequently the case that the implications of choosing one principle or the other act as a motivation for the choice itself so that there is a circularity to many of the arguments on this topic even though most of the discussions purport to be addressing the issue directly (i.e. choosing a principle on its own merits). Few writers are as candid about their motives as Helland (1995):

[15] Cox (1958), p.361.

…because the universal validity of the [unrestricted] CP under weak assumptions implies the universal validity of the LP, and (1) the latter has some strange consequences and (2) Berger & Wolpert argue that it nearly leads one to be a universal Bayesian, I have tried to collect some examples, aiming to convince myself and other people that it is not natural to assume that the CP has universal validity.[16]

I. Minimal sufficiency and Birnbaum's theorem.

If we aim to modify Neyman-Pearson inference to take into account the issue of relevance raised by Cox, it makes a critical difference whether or not we are prepared to condition on a statistic that is *not* part of the minimal sufficient statistic (MSS) for $\theta \in \Theta$; this is because Birnbaum's theorem uses the unrestricted version of the conditional principle. A conditional principle based on Cox's definition of ancillarity will not yield the likelihood principle when coupled with the sufficiency principle. This is not to say that it will not radically alter the way we make inferences by comparison with the unconditional approach, but the general approach underlying frequentist inference may remain intact. By contrast, if we base a conditional principle on the less restrictive notion of ancillarity and still retain the sufficiency principle, this will force us to accept the likelihood principle which is associated with alternative theories of inference and would require us to abandon frequentist inference altogether.

Another look at the proof of Birnbaum's theorem shows why this is true. In the two stopping rules example, the outcomes $E_1 \mapsto x = 14$ and $E_2 \mapsto y = 14$ have proportional likelihoods in the umbrella experiment E^*. Thus the likelihood principle is entailed by the sufficiency and conditional principles (as shown again below) and this dictates that the inference made about p should be the same in either case.

[16] Helland (1995b).

a) $\{(1,14)$ from $E^*\} \equiv_p \{(2,14)$ from $E^*\}$ by the SP,

b) $\{(1,14)$ from $E^*\} \equiv_p \{x=14$ from $E_1\}$ by the CP,

c) $\{(2,14)$ from $E^*\} \equiv_p \{y=14$ from $E_2\}$ by the CP.

Hence,

d) $\{x=14$ from $E_1\} \equiv_p \{y=14$ from $E_2\}$ as required by the LP.

The variable, A, tells us 'which sub-experiment' or 'which stopping rule' was performed in stage two. Birnbaum takes the view that we should condition on the observed value of A, leaving only the experiment that was actually performed in the picture since the nature and probability of the other sub-experiment is irrelevant; this is his conditional principle (CP) and it yields b) and c) above. However, the variable, A, is not a function of the minimal sufficient statistic[17] for $p \in (0,1)$, so that, if we were only prepared to condition on a statistic that is part of the MSS, statements b) and c) (and hence the LP) would not result.

In Cox's two-stage experiment with sub-experiments involving Normal distributions, the 'which machine' variable *is* part of the minimal sufficient statistic for the parameter of interest, μ, in the parameter space $\mathbb{R}$. Cox's example contained two sub-experiments, $E_i \ (i=1,2)$, that observe the value of $X_i \sim N(\mu, \sigma_i^2)$, a model consistent with two machines of different reliability or the random choice of two possible sample sizes (in which case X_i is the sample mean). Why is the 'which sub-experiment' variable part of the MSS in the Normal case but not the Bernoulli case? Simply because, in the Bernoulli case the two outcomes $x=14$ from E_1 and $y=14$ from E_2 have proportional likelihoods, so the MSS does not distinguish between them. A minimal sufficient statistic categorises the sample space so that any two observations will have the same value of the MSS if, and only if, they have likelihoods proportional to each other. Since A distinguishes between these two

[17] As follows: Let S be the MSS, then A is a function of S if and only if two outcomes that give the same value of S must also give the same value of A. As noted previously $s = S((a,u)) = \begin{cases} (a,u), & u \neq 14 \\ u, & u = 14 \end{cases}$, hence s takes the same value when $(a,u)=(1,14)$ as when $(a,u)=(2,14)$, but a is not the same in these two cases. Hence A is not a function of S.

outcomes (because it tells us which sub-experiment was performed) it is evident that A is not a function of the MSS. By comparison, in the case where the sub-experiments produce Normal variables, there are no values associated with the two sub-experiments that have proportional likelihoods for all values of $\mu \in \mathbb{R}$. It follows that A, which distinguishes between the two sub-experiments, must be a function of the MSS since any two outcomes from different sub-experiments will have different (i.e. non-proportional) likelihoods and hence produce different values of the MSS.

Thus the restricted conditional principle, in combination with the SP, does not produce the LP. This is because, for any two outcomes from a hypothetical umbrella experiment, we can, by conditioning on a variable that *is* part of the MSS and applying the sufficiency principle, produce either results b) and c) (if the two outcomes have different likelihoods) but not a) (for the same reason), or produce result a) (if the two outcomes have proportional likelihoods) but not b) or c). We can never produce results a), b) and c), and therefore are not bound by the likelihood principle. It follows that the RCP will be attractive to anyone who wishes to address the relevance issue (at least, to some degree) but remain a frequentist.

II. Arguments about conditional principles.

Leaving aside the implications of choosing one or another conditional principle, what are the direct arguments in favour of each?

Cox advocated the more restrictive definition of ancillarity, but his own justification of conditioning does not depend on the ancillary statistic being part of a MSS, rather it derives its force from the clear irrelevance of the non-performed sub-experiments. If the sub-experiment is chosen by an uninformative random mechanism, should it make any difference whether the sub-experiments are Bernoulli (as above) or involve Normal variates, as in Cox's paper? Should not the unperformed experiment be left out of consideration in either case?

If we condition only on variables that are part of the MSS (as Cox requires), there will be some unpleasant consequences. In Cox's example, a random mechanism chooses either of two machines (or, equivalently, the sample size) and this defines the (known) variance of a Normal statistic where the parameter of interest is the mean. The 'which sub-experiment' variable is part of the MSS as long as the parameter space for μ contains an interval, but, if the parameter space is (for example) the set $\Theta \equiv \{\mu_1, \mu_2, \mu_3, ...\mu_k\}$, then the stage one variable is not part of the MSS for $\mu \in \Theta$, and hence we should not condition upon its value, yet the issue of relevance is as potent as ever. Furthermore, we showed in §5.1 that, when $\Theta \equiv \{0, 5\}$ and the 'which sub-experiment' variable is not a function of the MSS, we can still define a winning betting strategy, based on the conditional success rates.

If we are only prepared to condition on statistics that are functions of the MSS, slight changes to the scenario can make a big difference to the inference drawn; Savage used this fact as the basis for one of the most effective arguments against so limiting the conditional principle. Durbin (1970) suggested that the conditional principle should be based on the definition of ancillarity used by Cox, pointing out that such a principle, together with the sufficiency principle, would not entail the likelihood principle. He argued, "although Birnbaum's sufficiency principle implies that as a function of the observations, evidential meaning depends only on the minimal sufficient statistic, … the domain of application of [the unrestricted] conditional principle is not restricted to statistics which are minimally sufficient."[18] In reply[19], Savage pointed out that conditioning only when the ancillary statistic is part of the MSS destroys the continuity of the inference process, as follows.

A problem caused by the restricted CP.

Consider a series of umbrella experiments, E_j^*, where the ancillary first stage chooses one of two sub-experiments, $E_{1,j}$ or E_2, with probabilities ½ and ½. Suppose, for convenience, that all the sub-experiments produce data from the same set $\mathfrak{X}$ and that, for all $\theta \in \Theta,\ j \in \mathbb{N}$ and $x \in \mathfrak{X},\ L_{1,j}(x;\theta) = c(x)L_2(x;\theta) + \varepsilon_j(x)$, where, for all

[18] Durbin, p. 395.
[19] Savage (1970).

x and j, $\varepsilon_j(x) \neq 0$, and $\varepsilon_j(x) \to 0$ as $j \to \infty$. Then as $j \to \infty$, the likelihood values of x from the two sub-experiments become arbitrarily close to being proportional. However, since, for any finite j, $L_{1,j}(x)$ and $L_2(x)$ are not *exactly* proportional, the first stage result, A, will be part of the MSS (as long as certain relations between $L_{1,j}(x_i)$ and $L_2(x_k)$ are ruled out) and we should make our inferences conditional on the observed value of A. Now consider another umbrella experiment, E_∞^*, where $L_{1,\infty}(x;\theta) = c(x)L_2(x;\theta)$ for all x, then A is not part of the MSS and we ought not to condition on a, instead carrying out an unconditional inference. Consider the two experiments E_j^* and E_∞^*, both resulting in the outcome $(A, X_A) = (2, x)$; by making j sufficiently large, we can reach a point where there is no identifiable difference between the two experimental models (as well as no difference between the outcomes), yet as long as we are aware that A is part of the MSS in the former experiment but not in the latter, we should feel obliged to make different – perhaps radically different – inferences in the two cases. Savage noted that such situations occur in practice; for instance, where a model is based on sampling without replacement, the likelihood will converge to that of a model based on sampling with replacement as the population size increases. Intuitively, as one model converges to another, we would like the inferences from a given result to converge as well; we cannot be comfortable with a theory that produces substantially different inferences in virtually identical circumstances.

Does the unrestricted CP conflict with the SP?

Durbin (above) seems to imply that the unrestricted conditional principle is in conflict with the sufficiency principle. To examine this question further, we introduce a new example. In the two stopping rules case, the stage one variable was not part of the MSS but the stage-two variable, alone, was sub-sufficient; the case where the stage-two variable is sufficient is more enlightening.

Consider the case where the two sub-experiments have a common support and proportional probabilities (or densities). Such an example is easiest to construct for a binary parameter space $\Theta_B \equiv \{\theta_1, \theta_2\}$; the following instance is due to Berger & Wolpert.[20]

i. Distribution of $X_1 \in \{10, 20, 30\}$ for sub-experiment 1 (E_1).

Table 5.6

x	*10*	*20*	*30*
$P_{\theta_1}(X_1 = x)$	0.090	0.055	0.855
$P_{\theta_2}(X_1 = x)$	0.900	0.050	0.050
$LR_1(x) = \dfrac{P_{\theta_1}(X_1 = x)}{P_{\theta_2}(X_1 = x)}$	0.1	1.1	17.1

ii. Distribution of $X_2 \in \{10, 20, 30\}$ for sub-experiment 2 (E_2).

Table 5.7

x	*10*	*20*	*30*
$P_{\theta_1}(X_2 = x)$	0.026	0.803	0.171
$P_{\theta_2}(X_2 = x)$	0.260	0.730	0.010
$LR_2(x) = \dfrac{P_{\theta_1}(X_2 = x)}{P_{\theta_2}(X_2 = x)}$	0.1	1.1	17.1

The first stage of the experiment consists of randomly selecting one of the two sub-experiments according to the fixed probabilities $P(E_1) = q$ and $P(E_2) = 1 - q$ ($0 < q < 1$, known and independent of $\theta \in \{\theta_1, \theta_2\}$), and the second stage consists of carrying out the chosen sub-experiment. The result of the umbrella experiment can be written (a, u), where u is x_1 if $a = 1$ and x_2 if $a = 2$. It is tempting to think that any

[20] Berger & Wolpert, p.21.

data should be interpreted by reference to the sub-experiment that actually produced it, i.e. conditional on the value of a.

Clearly (A,U) is a sufficient statistic for θ, however, in this case, it is easy to show that the minimal sufficient statistic is U alone, suggesting that all the information about $\theta \in \Theta_B$ contained in (a,u) is present in u while a contributes nothing. If we condition on the observed value of A, we are allowing a to influence our interpretation of the data. Does this mean that the (unrestricted) CP conflicts with what we know about the sufficient statistic, i.e. does it conflict with the SP?

Let us consider the outcome $u = 10$. For a test of H: $\theta = \theta_1$ versus K: $\theta = \theta_2$, the outcomes $(a,u) = (2,10)$ and $(a,u) = (1,10)$ have the following conditional and unconditional p-values:

- p-value of (2,10) *conditional upon* $a = 2$ is $P_{\theta_1}(X_2 = 10)$ $(= 2.6\%)$,
- p-value of $(1,10)$ *conditional upon* $a = 1$ is $P_{\theta_1}(X_1 = 10)$ $(= 9\%)$,
- unconditional p-value of (2,10) is $qP_{\theta_1}(X_1 = 10) + (1-q)P_{\theta_1}(X_2 = 10)$ (dependent on q),
- unconditional p-value of (1,10) is also $qP_{\theta_1}(X_1 = 10) + (1-q)P_{\theta_1}(X_2 = 10)$.

When we apply the unrestricted CP in a frequentist framework, we get the conditional p-values (above) and these are different when $a = 1$ and $a = 2$, contradicting the SP.[21] By contrast, we can see that the unconditional p-values are consistent with the sufficiency principle since they produce the same inference from $u = 10$ regardless of the value of a; however, relevance is still a problem; if we use unconditional methods, the p-value of any outcome (a,u) depends critically on the value of q, which is surely irrelevant. This puts us in a difficult position because the concept of sufficiency is apparently at odds with our common sense, however the sufficiency principle it not necessarily the cause of the problem.

[21] That is, since U is a sufficient statistic for $\theta \in \Theta_B$, any observations that produce the same value of u (like (1,1) and (2,1)) should produce the same inference, for example p-value.

Many discussions have focussed on whether we ‘ought’ or ‘ought not’ to condition on a statistic that is not part of the MSS, but it is more natural to ask why it would make any difference. Clearly it does make a difference when we use p-values, yet this is counter to our understanding of minimal sufficiency. A MSS is usually interpreted as maximally removing irrelevant or junk information from the raw data, so a statistic that is not part of the MSS is presumably junk. For example, we may observe the *colour of the experimenter’s socks* along with the outcome of the experiment but it will not be a function of the MSS. We can regard the sock variable as ancillary since the probability of any particular colour is independent of the parameter of interest, then we can try to decide whether or not to condition upon it since it is not part of the MSS. In fact, our results will be the same regardless of whether we condition on sock colour or not, since the sample space and associated probabilities for the rest of the experiment will be the same no matter what the colour of the socks actually is. It is neither necessary nor wrong to condition on this variable, it simply makes no difference, and surely this is what we would expect from information that is only junk.

In our experiment, the distributions of X_i from the two sub-experiments are different, but the fact that the MSS does not distinguish between the sub-experiments tells us that these differences are irrelevant (according to the usual interpretation of *sufficiency*) and ought not to influence our result; this, in turn, suggests that there is something wrong with our mode of analysis. Any two outcomes that produce the same value of a sufficient statistic will also have the same likelihood ratio (LR) value (for any two values of θ in the parameter space). If we were to base our inference solely on the observed value of the likelihood ratio instead of using p-values, our inference from any u would be the same regardless of which sub-experiment was involved, but would no longer depend on q (see likelihood ratios in above table). If we use methods of inference that are consistent with the likelihood principle, the two conditional principles are effectively the same, because, whenever A is not a function of the MSS, conditioning upon it has no effect. In contrast, frequentist inference uses measures, such as p-values, that cannot agree with *both* the form of relevance addressed by the sufficiency principle *and* that addressed by the (unrestricted) conditional principle; thus, *in the context of frequentist inference*, sufficiency and conditional relevance are inherently contradictory; of the three ‘principles’ – frequentism, sufficiency and (conditional) relevance – only two can co-exist; the problem with the conditional p-values (above) may be attributed to the fact that they are p-values (i.e. a frequentist measure) rather than to the fact that they are conditional. Is the decision to condition on a statistic even when it is not a function of the MSS more counter to the *spirit* of the sufficiency principle than the decision to accept a mode of inference that is altered by such conditioning? The differences of opinion expressed by writers on this topic stem from no technical controversy but simply reflect which principles they regard as non-negotiable and which as optional. The following statement by Gordon is typical: “it is possible to construct examples in which conditioning on an ancillary statistic which is not a function of the MSS seems the appropriate procedure to adopt… accordingly, the ‘Sufficiency Principle’ is violated”[22], but this should not be taken too literally since the *accordingly* only applies in the context of frequentism. This context is so often assumed, by writers of this persuasion, that it often vanishes from sight leaving the reader with the impression that there is a real and inherent conflict between the unrestricted-conditional principle and the sufficiency principle. The issue of ‘relevance’ as described by Cox and the unrestricted conditional principle that comes naturally from it pose a problem for frequentists who do not have a convincing alternative principle with which to oppose it and are therefore reduced to using its supposed conflict with the sufficiency principle as a reason for rejecting it. Yet it is only when we use a frequentist procedure that conditioning on a statistic that is not a function of the MSS makes any difference and creates a conflict with the SP.

In many ways it is strange that the notion of sufficiency holds such sway with frequentists when the idea behind it (which is essentially the same as that behind the likelihood principle) is not connected to that theory conceptually. This peculiarity caused Good to see Birnbaum’s paper as “primarily a contribution to the sociology of statistics rather than to its logic”.[23] Since Birnbaum’s theorem is redundant from the point of view of Bayesians or any who are already convinced of the likelihood principle, it is only important to non-likelihoodists who nevertheless believe in the sufficiency principle: “Birnbaum’s result can win over only those statisticians who

[22] Gordon, p. 273.
[23] Good, I. J. in Birnbaum (1962) discussion, p. 312.

find [the SP] compelling for some other [i.e. non-Bayesian] reason, such as by being compelled by the weight of Fisher's authority".[24] Surprisingly, there are a great many of these; even Joshi, who is extremely hostile to the likelihood principle and hence to Birnbaum's theorem, never went so far as to reject the sufficiency principle altogether. In "Fallacy in the proof of Birnbaum's theorem" (1990), he argued that the sufficiency principle ought not to be applied to the scenario described by Birnbaum, but he did not repudiate the principle, preferring to re-interpret it in such a way as to bar it from co-existing with the conditional principle; this cuts to the heart of the problem for frequentists since individually neither principle is at odds with frequentism.

Using the restricted CP allows the frequentist framework to remain intact but at the cost of severe lack of continuity in the inference process; the unrestricted CP in concert with the SP requires us to abandon frequentism altogether in favour of some approach consistent with the LP, but it is clear that in this context there is no conflict between either the logic or the spirit of the two principles.

More problems with restricted conditional theory.

In 1958 Cox wrote of choosing a restricted ancillary statistic to partition the sample space of the MSS, taking "the finest such division, assumed here to be unique subject to regularity conditions"[25]. Unfortunately, it soon became apparent that the assumption of uniqueness was not always warranted and this is a major problem for those pursuing Cox-type relevance by conditioning on an ancillary statistic that is a function of the MSS. Suppose that two statistics A and B are both functions of X and are both ancillary (in Cox's sense) for the parameter space of interest; which statistic should we condition on? If the statistic (A,B) is itself ancillary, there is no problem as we can condition on both; this situation covers, but is not limited to, the case where one of the statistics is simply a finer version of the other, and it will apply whenever the two statistics are stochastically independent. However in some cases it happens that a partition involving both statistics is not ancillary because the

[24] Good, I. J. in Birnbaum (1962) discussion, p. 312, 313.

[25] Cox (1958), p. 361.

distribution of (A, B) varies with θ. In such a case we must choose between conditioning on A or on B. In 1964, Basu showed that ancillary statistics are not necessarily unique or compatible; this also applies to Cox's definition and in 1971 Cox attempted to address the problem by developing a theoretical framework on which to base the selection of an 'optimal' ancillary statistic, namely, that which "separates the expected information into components that differ as widely as possible".[26] His aim was to ensure that the possible samples be "sorted as selectively as possible into sets, all the samples in any one set having the same amount of [Fisher] information".[27] Though the approach is reasonable enough, it relies heavily on Fisher's general theory and Cox himself thought it "admittedly rather arbitrary".[28] Efron (1978) observed,

> **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 … clearly structured. 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.[29]**

Buehler (1982) discussed the historical development and use of different notions of ancillarity. He was particularly interested in using ancillary statistics as precision indices in certain contexts and noted, "non-uniqueness of ancillary statistics has been considered a weakness of theories of conditional inference"[30]. In an overview published in 1988, Cox conceded that lack of uniqueness still posed a problem for the theory, "one difficulty… is that the decomposition $S = (T, A)$ may not be unique … although reasonably plausible criteria of choice between competing ancillaries are available"[31], and in the same year Berger & Wolpert stated, " the choice of a relevant subset or an ancillary statistic or a partition $\{x_s : s \in \varsigma\}$ can be very uncertain…

[26] Cox (1971), p. 251.
[27] Cox (1971), p. 254.
[28] Cox (1971), p. 254.
[29] Efron, p. 240.
[30] Buehler (1982), p. 586.
[31] Cox (1988), p. 318.

Researchers working with ancillarity attempt to define "good" ancillary statistics to condition upon, but, … there appear to be no completely satisfactory definitions".[32]

The other major problem that plagues conditional frequentism is that often there are no exactly ancillary statistics for $\theta \in \Theta$. It may seem that this is not a problem since it leaves us free to perform an unconditional inference, however the intuition that some data sets are more informative than others may still be present. Moreover the situation, described by Savage, where two models that are very similar produce quite different inferences (because an exact ancillary statistic exists in one case but not the other) remains a problem. For this reason there has evolved a practice of (sometimes) conditioning on statistics that are approximately ancillary - this has done nothing to simplify the theory.

The question of approximate ancillarity (distribution weakly dependent on θ) was raised by Cox and Hinkley (1974, p. 34), and recently has been the subject of substantial work… Ancillarity is to the conditional principle as approximate ancillarity is to what? A new principle seems to be needed.[33]

A more serious aspect [than lack of uniqueness] is that such a decomposition $[S = (T, A)]$ **often does not exist. In the latter case, and in general when we are driven to the use of approximate methods, considerations of continuity suggest that approximate arguments should be as similar to 'exact' answers as possible. More specifically, if exact theory indicates conditioning on statistics with a certain property, then approximate theory should condition on statistics that have that same property approximately,** ***in some sense to be defined*** [my emphasis]**. Failure to do this would lead to the position where a minor perturbation of a model made a radical difference to the mode of inference.**[34]

[32] Berger & Wolpert, p. 16.
[33] Buehler (1982), p. 586.
[34] Cox (1988), p. 318.

Restricting the CP allows one to remain a frequentist but creates a great many other problems. The question of which of the two conditional principles to support remains a confused one. In the *Encyclopedia of Statistical Sciences* (1982), Keifer[35] states that "currently in the literature" we take an ancillary statistic to be any statistic with a distribution independent of θ without requiring that it also be a function of the MSS, and Lehmann & Sholtz (1992), summarising the research up to this date, use the same convention, but Stuart et al. (1999) use Cox's definition adding "it should be noted that some authors do not require (T_r, A) to be minimally sufficient … this increases the problems associated with the non-uniqueness of the ancillary statistic and we shall restrict attention to the minimally sufficient case".[36]

The last point is an interesting one. On the face of it, the less restrictive our definition of 'ancillarity', the more problems we will have with lack of uniqueness and competing inconsistent inferences, but this is another claim that is only true when we assume a frequentist framework. Birnbaum's theorem tells us that any inference method consistent with the likelihood principle will inevitably satisfy the (unrestricted) conditional principle; that is, it will satisfy it with respect to each and every such ancillary statistic, simultaneously. As long as we are prepared to abandon frequentist inference, as we are virtually obliged to do if we want to apply the unrestricted CP, there need not be a problem and the more demanding of the two conditional principles will automatically be satisfied (as will the restricted version).

The majority of the discussions of the last forty years (outlined in Chapters 4 & 5) support the view that conditioning is often desirable because it is the only way in which a (frequentist) inference can take into account the fact that some data is more informative than other data. What affect has this had on the every day application of statistics?

[35] Keifer (1982), p. 105.
[36] Stuart *et al*., p. 433.

5.3 Conditional inference today.

There is limited recognition of the issue.

Many comments in papers addressing the foundations of inference give the impression that one or other form of the CP has now been accepted by frequentists generally, at least for use in evidential contexts; for instance:

Robinson (1975)

Today it is widely accepted by adherents of confidence interval theory that they should perform their analyses conditional on the value of ancillary statistics.[37]

Keifer (1982)

The classical interval is what one would use if its optimum properties were criteria of chief concern, but many practitioners will not find those unconditional properties as important as conditional assessment of precision based on the value of [an ancillary statistic].[38]

Lehmann (1993)

…we might say that we prefer the [conditional] test in a scientific situation where [unconditional] long-run considerations are irrelevant and only the circumstances at hand [i.e. the observed value of the ancillary statistic] matter.[39]

However, there is still a confused attitude towards the criticism of conventional Neyman-Pearson theory inherent in both the conditional principles. This is not simply a matter of conflicting points of view; it is noticeable that the debate has been confined to the private sphere of specialists in foundational inference and is not covered in general textbooks or courses, either as an interesting controversy or a dogmatic point of view of either kind. Textbooks aimed at senior undergraduates discuss the Neyman-Pearson theorem and the highest power criterion with no mention of the issue of relevance addressed by the conditional principle and the fact that this

[37] Robinson (1975), p. 155.
[38] Keifer (1982), p. 105.
[39] Lehmann (1993), p. 1246.

conflicts with the power criterion. What makes this more striking is that these issues are easy to describe and well within the comprehension of undergraduate students. (But one of my colleagues argued that this topic should be avoided as it would “worry the students” and perhaps this view is more widespread than we like to think.)

Apart from its theoretical implications, the CP is applicable in a number of realistic and simple cases taught at the introductory level; for instance, regression analyses. The model

$$\underset{\sim}{Y}=\alpha\underset{\sim}{1}+\beta\underset{\sim}{X}+\underset{\sim}{\varepsilon},$$

where the independent variable, $\underset{\sim}{X}$, is random and the conventional independences apply, is a clear case for conditioning since $\underset{\sim}{X}$ is ancillary to the parameters of interest, α,β and σ_{ε}^{2}. Despite this, the issue is not addressed in these terms in the (very large number of) multivariate analysis textbooks aimed at students of statistically dependent disciplines. One of the few to discuss the random-independent-variable model is Hair *et al.* (1998), which contains the following passage:

A random independent variable is one in which the levels are selected at random. When using a random independent variable, the interest is not just in the levels examined but rather in the larger population of possible independent variable levels from which we selected a sample.
Most regression models based on survey data are random effects models. …
The estimation procedures for models using both types of independent variables [random and fixed] are the same except for the error terms. In the random effects models, a portion of the random error comes from the sampling of the independent variables.[40]

There is no mention in this text of any of the issues associated with the conditioning debate. The justification for making different inferences depending on whether the independent variable is fixed or random (i.e. for not conditioning) is contained in the assertion: *the interest is not just in the levels examined but rather in the larger population of possible independent variable levels from which we selected a sample.*

[40] Hair *et al.*, p. 166.

Clearly this is not true as regards estimating, or performing tests on, α, β and σ_{ε}^{2}, rather it appears to be a version of Welch's old argument, that the random case must be treated differently because we *might have* observed a different value of $\underset{\sim}{x}$; Cox's counter-example cast doubt upon this view, as long ago as 1958.

Ignorance of this issue within the statistics-using community has also been apparent in research; in the mid-late 1980's, Greenland *et al.* wrote several papers in the *American Journal of Epidemiology* attacking the widespread use of standardised measures, in part, via a regression example:

The preceding objection to standardized coefficients and correlation applies if all the variables are continuous. In particular, it applies even if all the variables are jointly normally distributed. To briefly illustrate this, suppose that we observe two study populations, A and B, and that for each individual i in these populations, the value of a continuous outcome y is causally determined by an antecedent variable x according to a mechanism such that $y_i = \alpha + \beta x_i + \varepsilon_i$, where α and β are unknown parameters that do not change across individuals or populations, each ε_i is a normal random variable with mean zero and variance 16, and the variables ε_i are independent across individuals and independent of x. In particular, note that the biologic effect of a change in x will be the same no matter which population an individual comes from. Suppose further that x is normally distributed in both populations, with mean zero and variance 9 in population A, and mean zero and variance 4 in population B. If $\beta = 1$, within each population we will observe the following relationships between x and y:

Population	**Linear regression coefficient**	**Standardized linear coefficient**	**Proportion of y variance explained by x**	**Correlation of x and y**
A	**1.00**	**0.60**	**0.36**	**0.60**
B	**1.00**	**0.45**	**0.20**	**0.45**

… only the ordinary regression coefficient properly reflects the unrelatedness of effect to study population: despite the constancy of effect, the populations differ in their standardized regression coefficients, proportions of variance explained, and correlations.
… The standard deviation of an exposure variable is a function of the marginal distribution of the variable. Consequently, standardized coefficients give a distorted measure of biologic effect because they depend not only on the magnitude of the biologic effect but also on the distribution of the risk factor … In other words, a standardized coefficient confounds the effect of a factor on risk with the background frequencies of both the factor and the outcome.[41]

The example given is one where conditioning is appropriate, since x is ancillary with respect to α and β; conditioning on x would remove the differences between the values attributed to the two populations A and B since the *distribution* (as distinct from the *value*) of x would no longer play a part in the analysis[42] and the authors' comments clearly tend towards a question of 'relevance'; yet, nearly thirty years after Cox's paper, the connection between their problem and the issue of conditioning is not made, even though the use of statistics in epidemiology is more sophisticated than that in many other areas.

While there has been much discussion of this topic in that part of the statistical research literature that deals with foundational issues, very little of it seems to have filtered down to those using statistics at a practical level. This may be because theoretical statisticians are still somewhat conflicted about the issue.

The account of conditioning in Stuart et. al. (1999).

Some such conflict seems to influence the account of conditioning given in the most recent edition of *Kendall's advanced theory of statistics.*[43] This work (in its various editions dating back to 1943) has been prized for its comprehensive coverage of

[41] Greenland *et al*. (1986), pp. 204-206.
[42] We do not suggest that it would resolve all the issues associated with standardization.
[43] Stuart *et al*.

frequentist theory and methodology and has generally given a balanced account of the conflicting schools of thought (for example, between Neyman-Pearson and Fisher). However the most recent edition fails to give a fair, or even clear, account of the issues surrounding conditioning.

Thus, Stuart is unclear about the appropriateness of conditioning in regression, referring to it, in passing, in the section on conditional inference[44], but then arguing as follows in the regression section:[45]

> **…in experimental statistics, it is often the case that the regressor variables are set to pre-assigned levels, so that even the idea of random variation among such variables is unreasonable. In such circumstances, it is preferable to carry out the analysis *conditionally*, given the values of the regressor variables. …The analysis will proceed *conditionally* upon the *x*-values.**

If $\underset{\sim}{x}$ is truly fixed, rather than being the observed value of a random $\underset{\sim}{X}$, this terminology makes very little sense; we do not 'condition' on the value of constants, the issue of conditioning only arises when we must choose between using the observed value, $\underset{\sim}{x}$, or the (non-trivial) distribution of the random $\underset{\sim}{X}$. Note how the main justification for conditioning (removing irrelevancies) is missing from this account.

Other aspects of Stuart's discussion of the conditional principle also indicate a reluctance to discuss the 'relevance' or 'two machines' argument for conditioning. The easiest way to convey this argument is to present Cox's Normal sub-experiments[46] example; Stuart does, but in a manner so cryptic as to render the example almost useless:[47]

[44] Stuart *et al*., p. 434.
[45] Stuart *et al*., p. 538.
[46] Recall, we toss a coin to choose the sample size. The conditional approach utilises only the sample size actually used.
[47] Stuart *et al*., p. 434.

Example 26.6 (Conditional versus unconditional test procedures (Cox and Hinkley, 1974))

An experiment involves drawing a sample from a $N(\theta, \sigma^2)$ population where σ^2 is known. The sample size, A, is either n or kn (k an integer, $k > 1$) and is selected by spinning a fair coin. The conditional information is n / σ^2, given $A = n$, and kn / σ^2, given $A = kn$; and $\hat{\theta} = \bar{X}$ in both cases. The UMP test, given A, is of the usual form, but power considerations based upon the unconditional experiment lead to a different and less intuitively appealing test procedure.

Almost everything of interest and note in Cox's example is missing from or obscured by this account. The central argument about the irrelevance of 'the sample size that was not used' is absent and, because of this, it is not necessary for the authors to justify limiting the definition of ancillarity to statistics that are functions of the MSS – a restriction that does not follow from the relevance argument. The 'less intuitively appealing' test procedure (which is, in fact, UMP in the traditional sense, unlike the conditional test) is not shown so it is impossible for the reader to judge how unappealing it is, nor are we shown how great a difference conditioning can make to the inference. Describing the conditional inference as being 'of the usual form' is misleading since it is the test that would be performed if the sub-experiment was the whole experiment but that is not the scenario being discussed here.

The following claim is also liable to be misinterpreted.[48]

The results of 21.30-32 enable us to see that if the distribution of the sufficient statistics (A, T) is of the exponential form (21.73) [i.e. $f(\underset{\sim}{x} \mid \theta, \underset{\sim}{\psi}) = C(\theta, \underset{\sim}{\psi}) h(\underset{\sim}{x}) \exp\{\theta s(\underset{\sim}{x}) + \sum_{i=1}^{r} \psi_i t_i(\underset{\sim}{x})\}$], the use of the conditional distribution of T for given A will give UMPU tests.

This tends to give the impression that, when a variable of the exponential family (EF) type is involved, conditioning on an ancillary statistic will make no difference, that is, it will produce the usual (UMPU) test. If this were true it would greatly reduce the

[48] Stuart et al., p. 436.

importance of the conditioning issue in practice since many natural test statistics are of the EF type. However, an umbrella experiment will *not* usually involve an EF type variable just because the sub-experiments do (Cox's example is a good illustration of this[49]) and, in this case, conditioning *will* make a difference. Furthermore, even when the experiment produces an EF variate, it may still be the case that the decomposition $m=(a,t)$ (a being ancillary) does not have the form (21.73) and so conditioning on a will produce results that are different from the UMPU test. The Normal example discussed in Chapter 8 of this work is an instance where the maximal decomposition does not take this form even though the test statistic is Normal and therefore of the EF type; in this case conditioning makes an enormous difference.

Stuart (1999) [50] cites Welch's example of 1939 as a *counter-example* to the CP (as Welch himself regarded it), instead of, as is now usual, a counter-example to *unconditional* inference:

> **These examples [Cox and a regression case] suggest that conditioning is desirable, but the argument is not one-sided. For example, Welch (1939) gave an example which showed that the conditional test … may be uniformly less powerful than an alternative unconditional test.**

This passage implies that the examples of Cox and of Welch are contradictory and support opposite sides of the debate. But Cox, himself, referred to Welch's example as 'similar' to his own and they have the same essential qualities important to the question, notably, that the conditional test has lower overall power (this was obscured in Stuart's account of Cox) but seems to make a lot more sense than the unconditional test. Stuart also reiterates Welch's circular argument (see Chapter 4): inevitably, whenever a conditional test differs from the unconditional optimal Neyman-Pearson test, it must have uniformly lower power, but the real question is whether this power, which is the overall or average power rather than the power associated with that part of the sample space relevant to the question at issue, is the appropriate criterion to use

[49] That is, if $T=\begin{cases}X,\ a=1\\ Y,\ a=2\end{cases}$ then it may be the case that X and Y are both of EF type (for instance Normal) but the variable (A,T) is not of EF type.

[50] Stuart *et al*., p. 434.

(and the same goes for significance level). Like Welch before them, Stuart, Ord & Arnold seem to be assuming (but not arguing) that it is and then proceeding in a circle as though the last forty years of the debate (of which the following quote from Lehmann (1993) is a good example) had not occurred.[51]

> **… the principles of conditioning on the one hand and of maximizing the unconditional power on the other may be in conflict … This conflict disappears when it is realized that in such cases priority must be given to deciding on the appropriate frame of reference; that is, the real or hypothetical sequence of events that determine the meaning of any probability statement. Only after this has been settled do probabilistic concepts such as level and power acquire meaning, and it is only then that the problem of maximizing power comes into play.**

Reading the account of conditioning in *Stuart, Ord & Arnold*, we are left with an overall impression of confusion, surprising in a work of this eminence. Dislike of the likelihood principle (LP) may play a role in this, since they note that: "*The unrestrained use of the LP lacks appeal for many statisticians who find it intuitively unacceptable to ignore the sample space when making inferences, but a result of Birnbaum (1962) makes serious consideration of the LP necessary*".[52] Technically, this problem can be solved by switching to the *restricted* CP, which Stuart *et al.* favour (though not in the Welch case, where they would prefer not to condition at all). A problem with Cox's argument (that the unperformed sub-experiments are irrelevant) is that it justifies the *unrestricted* CP at least as well as the restricted version, and this must explain their reluctance to come to grips with this intuitively appealling argument, which best encapsulates the reasons for conditioning. It is surely as a result of this failure that they further confuse the two principles by wrongly assigning Birnbaum's description of the LP (as stating 'the irrelevance of unobserved outcomes') to the CP in place of his Cox-inspired description of the CP as stating 'the irrelevance of experiments not actually performed'[53].

[51] Lehmann (1993), p. 1247.
[52] Stuart *et al*., pp. 438-439.
[53] Birnbaum (1962), p. 271.