Chapter 4: Seminal moments in the history of the conditioning controversy.

4.1 Introduction.

Neyman and Pearson defined an optimal test by reference to its *power* and this was a major point of contention between them and Fisher ("I am a little sorry that you have been worrying yourself at all with that unnecessarily portentous approach to tests of significance represented by the Neyman and Pearson critical regions etc."[1]). The difference between the theories appeared to be largely philosophical since Fisher's test statistics (maximum likelihood estimators) generally produced most powerful tests. However, it eventually became apparent that, in cases where a Fisherian ancillary statistic exists, Fisher's conditioning requirement results in an inference that is substantially different from that of Neyman. In this chapter we look at three papers that have made a major contribution to the debate about conditioning. The earliest of the three is that of Welch from 1939, only six years after the publication of the Neyman-Pearson theorem. The other two papers were published twenty years later during the period 1958-1962. Each of these works generated further literature on the topic and, together, they highlight all the most important features of the debate as well as motivating the development of the approach described in later chapters of this work.

4.2 Welch (1939).

In 1939 B. L. Welch published "On confidence limits and sufficiency, with particular reference to parameters of location". The paper aimed to critically compare the theory of Fisher with that of Neyman and Pearson. One of the most important aspects of the paper is that it showed that the two approaches could produce very different

[1] Fisher, quoted in Lehmann (1993), p. 1245.

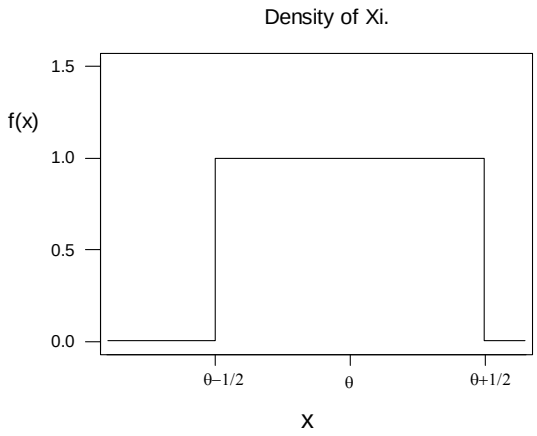
results even when applied to a straightforward example. The choice of example was an inspired one; it is very simple with only one parameter and a structure that is easy to understand. Despite this simplicity, the data give rise to a statistic that is ancillary according to Fisher's stringent definition. The fact that the ancillary statistic is embedded in the data instead of standing out as, say, the first part of a two-stage sample is also informative since it shows that such phenomena are not always easy to recognise (however, once recognised, it is easy to interpret). It was Fisher's position that inferences should be carried out conditional upon the observed value of such an ancillary statistic whereas Neyman and Pearson's approach aimed to maximise the overall power, and was thus unconditional.

The Uniform example – Part I.

This is the example used by Welch. Our model is a random sample of size n from a Uniform$[\theta-\frac{1}{2},\theta+\frac{1}{2}]$ population; $X_1,\ldots,X_n$ are independent and identically distributed with density:

$$f(x;\theta)=\begin{cases}1, & x\in[\theta-\frac{1}{2},\theta+\frac{1}{2}]\\ 0, & \text{otherwise.}\end{cases}$$

Figure 4.1



X is equally likely to lie within any interval of a given length in $[\theta-\frac{1}{2},\theta+\frac{1}{2}]$; such variables are sometimes described as lying 'randomly' in the interval. The probability that X lies outside $[\theta-\frac{1}{2},\theta+\frac{1}{2}]$ is *zero*. We know the width of the interval in which X must lie, it is *one*; we do not know where this interval lies on the real line. We could use any point on the interval as the unknown parameter and have defined θ to be the centre of the interval.

The minimal sufficient statistic.

For any $n \geq 2$, the minimal sufficient statistic for $\theta \in \mathbb{R}$ is given by the first and n^{th} order statistics, $(X_{(1)}, X_{(n)})$; that is, the largest and smallest values together contain all the information about $\theta \in \mathbb{R}$ available in $X_1,\ldots,X_n$. To see why this is so, observe that θ must lie in the interval $[X_{(n)}-\frac{1}{2}, X_{(1)}+\frac{1}{2}]$ since $\theta-\frac{1}{2} < X_{(1)} < X_{(n)} < \theta+\frac{1}{2}$. Once we know the left and right extremes of the data, knowing the positions between them occupied by other data-values does not add to our knowledge about θ. Any one-to-one function of $(X_{(1)}, X_{(n)})$ is also a minimal sufficient statistic for $\theta \in \mathbb{R}$.

The ancillary statistic, R.

Define $M = \dfrac{X_{(1)} + X_{(n)}}{2}$, the point half way between the two extreme order statistics (sometimes called the 'midrange' of the sample), and $R = X_{(n)} - X_{(1)}$, the sample 'range'. Then (M, R) is a one-to-one function of $(X_{(1)}, X_{(n)})$ and hence a minimal sufficient statistic for $\theta \in \mathbb{R}$. It is obvious that the behaviour of R will not be influenced by the interval's position on the real line, and it is easy to show that the distribution of R is the same for all values of θ. Since the likelihood of θ, $L(\underset{\sim}{x};\theta)$, equals *one* for all $\theta \in (x_{(n)}-\frac{1}{2}, x_{(1)}+\frac{1}{2})$ and *zero* elsewhere, any point on the interval $[X_{(n)}-\frac{1}{2}, X_{(1)}+\frac{1}{2}]$ qualifies as a maximum likelihood estimator of θ, including M, which is the only unbiased MLE. Thus the statistic, R, satisfies Fisher's definition

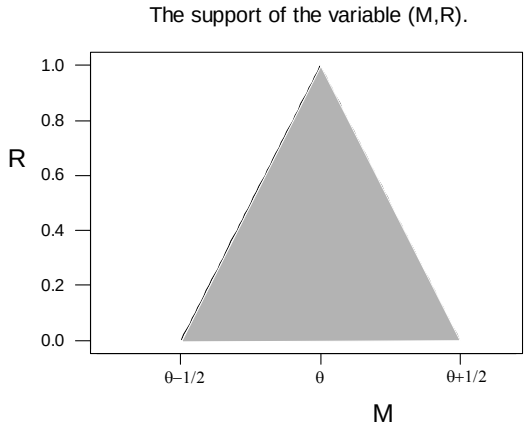
of an ancillary statistic and appears to be a good indicator of the reliability of M as an estimator of θ (in the same way that the random sample size N was a good indicator of the reliability of $\bar{X}$ as an estimator of μ (see Chapter 2)).

We will discuss a number of issues that arise when $n = 2$; these phenomena also occur for general n, only the numerical details vary. Instead of using the natural data $\underset{\sim}{x} = (x_1, x_2)$, we will use (m, r). We can do this even when there are more than two data values since it is a sufficient statistic and using this form allows us to look at the ancillary statistic explicitly.

Distributions – conditional & unconditional.

The statistic (M, R) has a (joint) density dependent on the parameter θ. Since it is a bivariate statistic, the density lies in three dimensions. The values that m can take are dependent on the value of r; for any given value of r in [0,1], m can only take values in the interval $[\theta \pm \frac{1}{2}(1-r)]$ as shown below.

Figure 4.2

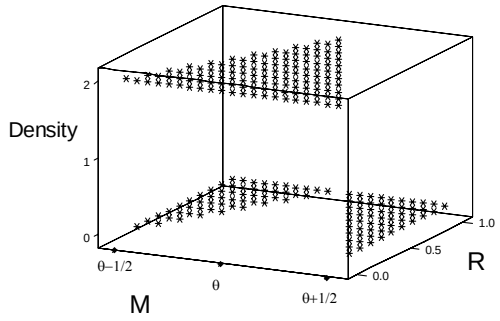


It is only when the two data points are on top of each other ($r = 0$) that the midrange, which is then equivalent to the two data points, can lie anywhere in the interval

$[\theta \pm \frac{1}{2}]$; when the range is large, the two data points are (relatively) far apart in the interval, and the point half way between them (m) must be fairly close to the centre, θ. When $r = 1$, the data points are one unit apart and must lie on the two bounds of the uniform distribution and $m = \theta$. Thus it is evident that the larger the range, the more information we have about θ. This is also apparent from the fact that the interval $[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}]$, which we know must contain the value θ (and is the shortest interval that must do so), can be written as $[M \pm \frac{1}{2}(1-R)]$ and has a width of $(1-R)$.

When $n = 2$, the density of (M, R) is a (3 dimensional) uniform over the triangular support (shown above), with a height of *two* (shown below).

Figure 4.3

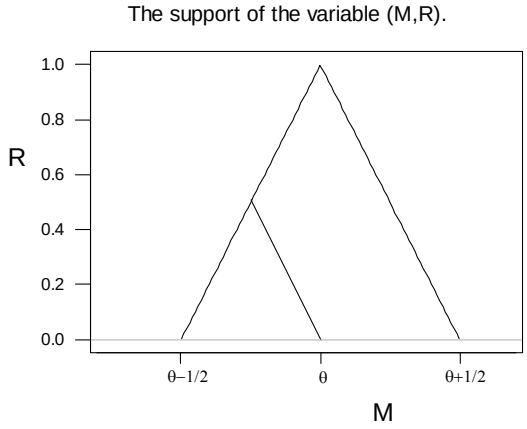


Since this distribution is uniform, we can avoid the usual integrations when calculating probabilities; the probability that (M, R) will fall within any particular region within the triangular support is simply the area of that region as a proportion of the total area of the support, or equivalently (since the area of the support is *one half*), twice the area of the region. For example, the region defined by:

$$\Lambda = \{(m, r) : (\theta - \tfrac{(1-r)}{2} < m < \theta - \tfrac{r}{2}) \,\&\, (0 < r < \tfrac{1}{2})\}$$

corresponds to the bottom left triangle below.

Figure 4.4



This area is 25% of the total area of the support and thus $P_\theta((M,R) \in \Lambda) = 0.25$.

The marginal distribution of R is independent of θ and (for the case $n = 2$) given by the density $f(r) = 2(1-r), \;\; 0 < r < 1$ (from which we can show that the average range from a sample of two observations is 1/3).

The conditional distribution of M given that $R = r \in [0,1]$ is also uniform, given by:

$$f_{M|R=r}(m) = \frac{1}{(1-r)}, \;\; (\theta - \tfrac{(1-r)}{2}) < m < (\theta + \tfrac{(1-r)}{2}).$$

For example, if $\theta = 1$, then the conditional distribution of M given that $R = 0.2$ is $f_{M|R=0.2}(m) = 1.25, \;\; 0.6 < m < 1.4$, i.e. given that $R = 0.2$, $M \sim \text{Uni}(0.6,1.4)$.

Neyman-Pearson theory uses the distribution of (M,R) to make an inference about θ, whereas Fisher's theory uses the conditional distribution of M given that $R = r$ (the post-experiment observed value of the range) and thereby takes into account the reliability of M for that particular value of R (similar to conditioning on the random sample size N).

Welch's paper did two things. The first is technical; he showed, that the different principles behind the two theories lead to different results, so it is not the case that they are really the same principle differently stated. This had not formerly been

apparent since standard tests had yielded the same numerical results from both approaches. Secondly, having established, and illuminated, the differences between the methods, Welch attempted to argue that the approach of Neyman and Pearson is superior to that of Fisher.

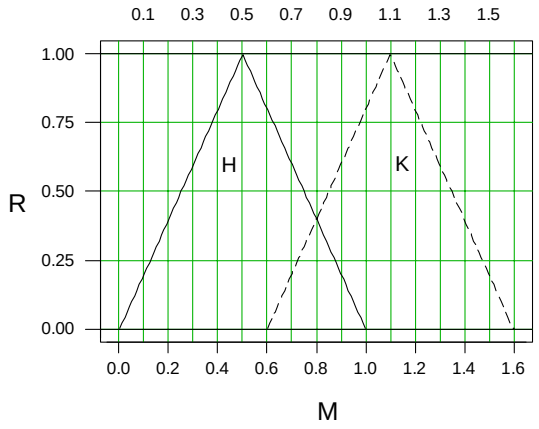
We will start by looking at the first issue. (Welch wrote about a two-sided confidence interval whereas we will look at a one-sided hypothesis test in order to highlight the connections with later parts of this work. All the interesting phenomena and arguments can be observed equally well in terms of either hypothesis tests or confidence intervals.)

Two different approaches.

Example 4.1

Consider a test of two hypotheses H: $\theta = 0.5$ versus K: $\theta = 1.1$, where X_1, X_2 are independent and both have the uniform distribution on $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. We will translate the data from (x_1, x_2) to (m, r). Under both hypotheses, (M, R) is uniformly distributed, with a height of *two*. The support of (M, R) depends on the hypothesis; both supports are shown below.

Figure 4.5



The rejection region for a Neyman-Pearson test is an area in the (m,r)-plane. Welch showed that an optimal 5% test can be obtained using the following area.

Figure 4.6

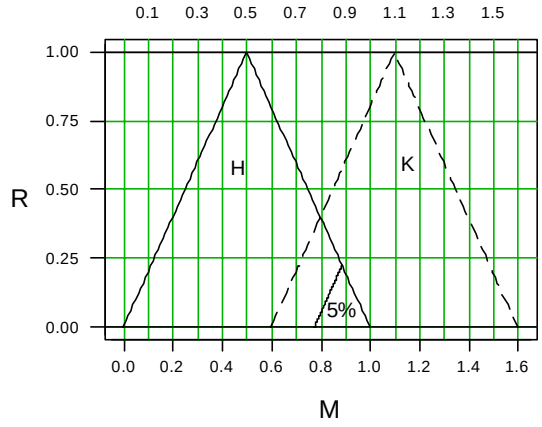
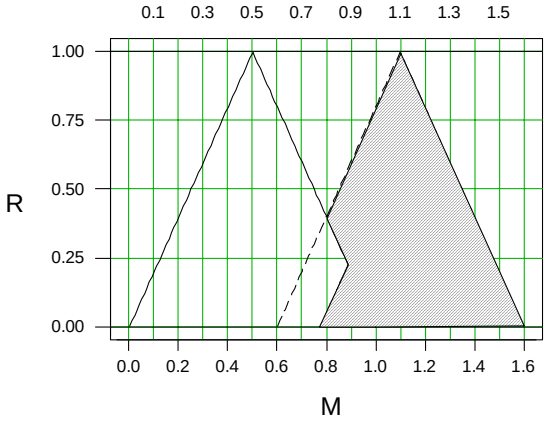


Figure 4.7

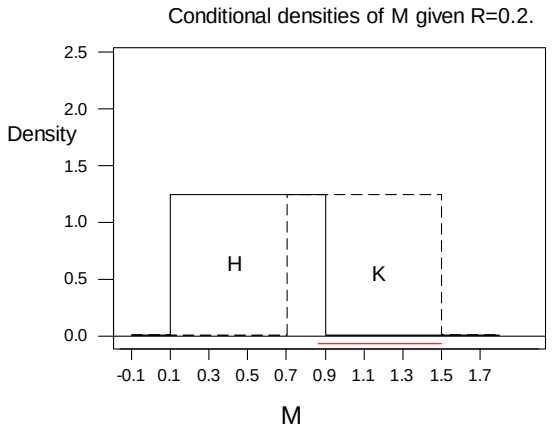


The shaded area above is the critical (rejection) region. If the data produces a value of (m,r) in this region, we will reject H in favour of K at the 5% level. The top graph shows that the probability of this region, under H, is 5%. (To check that this is true, note that each of the grid rectangles has an area of 0.025 units and is thus one twentieth of the total triangle area, and the small triangle labelled 5% has the same area as one of the rectangles.) Of course, that part of the support under K that does not overlap with the support under H is also part of the rejection region but does not

contribute to the value of α (5%) since it has a probability, under H, of *zero*. The power of the test is the probability of the shaded region under K, and is obviously fairly high. No other rejection region with a probability, under H, of 5% has a higher power than this one, so it is 'optimal' in the Neyman-Pearson sense.

Now let us look at a Fisherian test. Given that we observe data with a range of r, we will base our inference on the *conditional* distribution of M *given* $R = r$. For example, if $R = 0.2$, the conditional distribution of M is Uniform on $[\theta \pm 0.4]$. The conditional distributions under H and K are shown below; the height of the densities is 1.25.

Figure 4.8

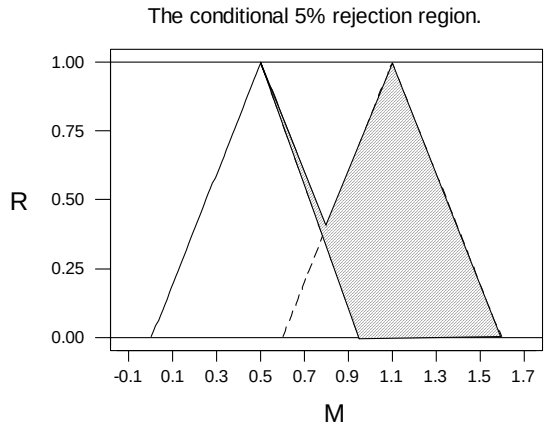


The (conditional) 5% rejection region for m is [0.86,1.50] (highlighted in the diagram above) since $P_H(0.86 < M < 1.50 \mid R = 0.2) = (0.90 - 0.86) \times 1.25 = 5\%$.
Welch pointed out that this type of test is conditionally most powerful, in the sense that there is no other rejection region with a *conditional* significance level of 5% that has higher *conditional* power than this test (the conditional power is $P_K(0.86 < M < 1.50 \mid R = 0.2)$).

In order to make a valid comparison between the two approaches we need to look at how they work for different values of r. Fisher's approach requires us to consider only the conditional distribution associated with the value of R that occurs in the experiment. We can find a Fisherian rejection region in the (m, r)-plane by combining the rejection regions (in m) for all the different possible values of r.

When we do this we find that the conditional approach produces the following rejection region, which is substantially different from the Neyman-Pearson critical region.

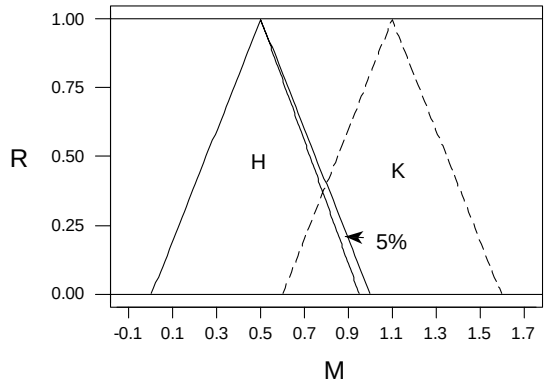
Figure 4.9



(If we were to draw a horizontal line across this plot at the point $R = 0.2$, we would hit the left side of the rejection area at the point $m = 0.86$ and the right side at the point $m = 1.5$.)

That part of the rejection region in the support of (M, R) under H is highlighted below.

Figure 4.10



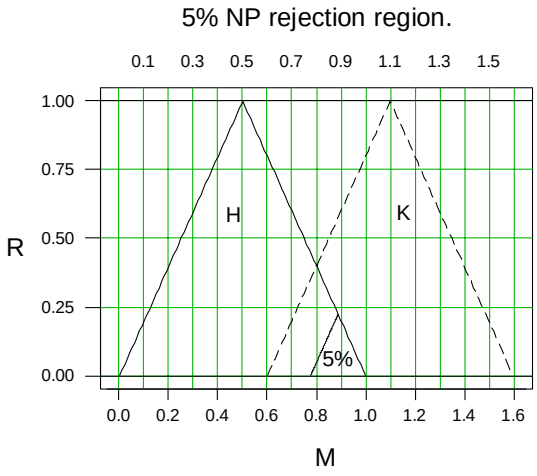
Since, the probability (under H) associated with the conditional rejection region is 5% for each value of r, it is also 5% when averaged over all values of r as indicated above. Note that this rejection region (call it $\mathfrak{C}$ for 'conditional') has two characteristics:

i. $P_H((M,R) \in \mathfrak{C}) = 5\%$

ii. $P_H((M,R) \in \mathfrak{C} \mid R = r) = 5\%$, *for all* r.

The unconditional (Neyman-Pearson) rejection region (call it $\mathfrak{N}$) satisfies (i) but does not satisfy (ii).

Figure 4.11



In fact, since (see above) the height of the '5% rejection region triangle' in the support of H is $1/\sqrt{20} \approx 0.224$, we can see that, for all $r > 0.224$, $P_H((M,R) \in \mathfrak{N} \mid R = r) = 0$, and, by contrast, if r is close to *zero*, $P_H((M,R) \in \mathfrak{N} \mid R = r)$ is close to 22.4%. (All the triangles in this diagram have *base = height*.) Thus, if our value of r happens to be small (the two data points are very close together), we know that the (conditional) probability of Type I error associated with the optimal rejection region $\mathfrak{N}$ is more than 20%, even though the nominal level of the test is 5%.

A critical question ignored.

At the heart of the issue is this question: is it the conditional properties (significance level, power) or the unconditional properties that really matter? We cannot answer this question solely by reference to the theories of Neyman and Pearson or Fisher. From their theories we know their views, but their theories result from their views, they do not wholly explain or justify them. The Neyman-Pearson theorem tells us how to find the test with highest unconditional power for a given unconditional significance level, but does not explain why this should be more important than the conditional power and significance level. If the rival theories are no help, what can we look at? We could look for a higher order principle that seems compelling and ask which of the two approaches is more consistent with it. Alternatively, we could take a more pragmatic approach; for a given set of data and model, we could look at the different inferences produced by the two approaches and ask ourselves which inference seems to make more sense, intuitively, in the light of the data. According to Jaynes, this approach is rarely adopted ("Let me make what, I fear, will seem to some a radical, shocking suggestion: *the merits of any statistical method are not determined by the ideology which led to it"*[2]). Welch's paper is a prime example of this shortcoming; the uniform example could have been used to assess the performance of each method from a practical point of view; instead, he chose to assess them only by reference to the conflicting theories, or rather, by reference to one of the theories to which he was deeply attached. Since both the competing tests have the same unconditional significance level, the theory of Neyman and Pearson states that they should be compared on the basis of their unconditional power and whichever one has the higher power is superior. The Neyman-Pearson test is (of course) the one with the higher unconditional power. Welch went to the trouble of showing this by calculating the power of the two tests even though he could have deduced the result directly from the Neyman-Pearson theorem. On the basis of this, he concluded that the Neyman-Pearson test is the better of the two tests. His argument for using this particular criterion was simply that the unconditional power is "the real power"[3] of the test.

[2] Jaynes (1983), p. 154.
[3] Welch, p. 63.

The Uniform example – Part II.

At the time, Welch's paper was accepted on its own terms despite the defects in its approach. Thus, Bartlett (1939) was concerned that Welch's paper cast doubt on the value of 'quasi-sufficiency' (a concept related to ancillary statistics and conditional inference):

Confining our attention, however, to problems which are primarily problems in one unknown only, we require to examine relations of the [conditional] type further, in view of some recent comments by Welch on the extent to which any conditional statistic … can claim to be sufficient[4].

Hotelling (1940) also thought the paper cast doubt on Fisher's method:

Criticisms of these applications of fiducial probability have been made by M. S. Bartlett [1936] and B. L. Welch [1939], and the field of applicability of such methods is still in need of elucidation[5].

Neyman, not surprisingly, also regarded the paper as lending support to his own methods:

In this paper various general claims of Fisher are analysed, essentially from the point of view of [Neyman-Pearson] confidence intervals, and tested on appropriate examples. Among other things it is found that the fears of inconsistencies in the theory of confidence intervals are unfounded[6].

However, twenty years later the attitude had changed as authors began to look more closely at the practical implications of having a conditional significance level that varies in a known way as r varies, but is ignored. In 1959, Lehmann had used the Uniform example in a discussion about sequential analysis. Observing that a large

[4] Bartlett, p. 391.
[5] Hotelling, p. 275.
[6] Neyman (1941), p. 129.

range provides more information about θ, he advocated a stopping rule that continues the sampling process until the range reaches a certain value.

Consider, for example, observations from the uniform distribution over the interval $(\theta-\frac{1}{2},\theta+\frac{1}{2})$ and the problem of estimating θ. …[A] sample of size n can practically pinpoint θ if the range is sufficiently close to 1, or it can give essentially no more information than a single observation if the range is close to 0. Again, the required sample size should be determined sequentially.[7]

In his 1961 review of Lehmann's book, Pratt argued that Lehmann had not grasped the wider implications of recognising that the amount of information available varies with r. In the previous section we showed that, for larger values of r, the conditional significance level of the optimal Neyman-Pearson test is *zero*, it follows that an analogous confidence interval will contain all possible values of θ consistent with the data (i.e. the confidence interval will be $[\theta\pm\frac{1}{2}(1-r)]$). For example, if $r>1/\sqrt{20}$, the 90% confidence interval[8] will contain all possible values of θ; in fact it is quite possible for even a 50% confidence interval to contain all possible values of θ, and, at the other end of the spectrum, for (say) an 80% confidence interval to be the empty set.[9]

Thus it appears that there are certainly occasions when inferences should be conditional, but optimal [i.e. Neyman-Pearson] decision procedures are not. … a confidence interval may include all or no possible values of the parameter, so that the confidence level measures no one's confidence … Satisfactory criteria have never been given for choosing a "good" *inference* procedure in the Neyman-Pearson formulation.[10]

[7] Lehmann (1959), p. 7.
[8] Throughout this work, the term 'confidence interval' (or CI), when unqualified, refers to the 'optimal' Neyman-Pearson (i.e. 'uniformly most accurate') confidence interval.
[9] See Pratt (1961) for more details.
[10] Pratt (1961), p. 166.

The same year, Fraser also criticised the Neyman-Pearson approach in the Uniform case. Having observed that the Neyman-Pearson confidence interval is more 'accurate'[11] than the Fisher interval, he continued:

There is however another side to the comparison. When the permissible range for θ is small the [95%] confidence interval embraces not 95% but the *full* range of possible values for θ. It is not hard to see what is happening: when the range of permissible values is short the confidence interval takes the *full* range on the grounds that in probability there may be another occasion when the range of permissible values is larger and *less* than 95% can be chosen and still maintain the long run 95% average. Is the long run average more important than the specialized knowledge of the particular situation? … My preference weighs heavily in favour of the fiducial [i.e. Fisher's] interval for Welch's example.[12]

Pierce, in 1973, cited Welch's Uniform example as a case where the optimal Neyman-Pearson method gives results that are fundamentally misleading. He thought that the use of fiducial methods only in cases where no optimal Neyman-Pearson method exists had obscured the fact that even Neyman-Pearson *optimal* methods have unfortunate conditional properties.

By 1975, Robinson was referring to Welch's example as "possibly the best known counterexample for Neyman's version of confidence interval theory"[13], and Keifer (1982)[14] in the entry on 'conditional inference' in the *Encyclopaedia of Statistical Sciences* quoted Welch's case as the paradigmatic example favouring conditioning.

Welch's example had undergone an amazing transformation from a case lending support to Neyman's theory to evidence of a major flaw in it. It had also helped to motivate a substantial modification of the theory (but for a description of the patchy acceptance, or even awareness, of this modification, see §5.4). Why did this happen? Once practitioners started to look in detail at the results (for example, confidence

[11] "Accuracy" in confidence intervals is analogous to power in hypothesis tests. The confidence interval analogue to a *most powerful unbiased test* is a *most accurate unbiased interval*; the Neyman-Pearson confidence interval in the Uniform case has this optimality property.
[12] Fraser (1961), p. 671.
[13] Robinson (1975), p. 155.
[14] Keifer (1982), p. 105.

intervals) generated from particular data by the optimal methods, their defects were to some degree obvious; Welch had failed to do this, instead concentrating on features of the *process* (like power), but the paper discussed in the following section contributed substantially to this change of heart.

4.3 Cox (1958).

1958 saw the publication of "Some problems connected with statistical inference" by D. R. Cox[15]. In part of this paper, Cox revisited the issue of whether or not an inference should be carried out conditional upon the observed value of an ancillary statistic and came down firmly on the side of conditioning. The compelling nature of this paper came in part from the use of a simple example using Normal distributions and a two-stage experimental process.

Cox's two-stage example.

The parameter of interest is a population mean, μ.

The first stage of the experiment utilises a device that produces one of two possible outcomes, each with a probability of ½ (the outcome of a toss of a fair coin, for instance). Call this an observation of the random variable, A, where $a \in \{1,2\}$.

In the second stage we observe the value of a random variable, X_a, where $X_a \sim N(\mu, \sigma_a^2)$ and:

$$\sigma_a^2 = \begin{cases} \sigma_1^2, & a = 1 \\ \sigma_2^2, & a = 2. \end{cases}$$

Suppose also that σ_1^2 and σ_2^2 are known, distinct values. Essentially this example is the same as that in which the sample size is randomly chosen and the sample mean is

[15] Cox (1958).

used as the test statistic (or estimator) for making inferences about the population mean. Here the sample size is fixed at *one* but the variance of X is still dependent on the outcome of the first stage.

If we were to condition on the value of A (which of itself tells us nothing about μ), we would simply perform the usual z-test or construct the usual z-interval based on the particular variance we 'observed'. Thus to perform a 5% test of H: $\mu = 0$ against any right-sided alternative, we would use the rule *Reject H if* $x > 1.645\sigma_a$, substituting the appropriate value of σ_a, based on the outcome of the first stage of the experiment, into this formula. The argument for doing this is very compelling and Cox described the central issue as one of *relevance*. If the result of stage 1 is that $A = 1$ and hence X has a variance of σ_1^2, then, although it is true that X *might have had* a variance of σ_2^2 (had stage 1 produced a different result), it is simply not relevant. We can make this scenario more concrete by thinking of it as a case where we are randomly assigned one of two different measuring devices or machines. x is the measurement made by our machine (for example, the weight of an object, where the true weight is μ). Both of the machines are unbiased, $E(X_a) = \mu$, but neither of them is perfectly accurate. One is less accurate than the other (possibly, much less) as indicated by the variance of X_a, and we know exactly how accurate each machine is. The central question is: given that we know which machine we have used, do we take into account only the accuracy of this machine when performing an inference on the experimental result, or do we allow ourselves to be influenced by the reliability of the other, unused, machine?

Disconcertingly, it turns out that the optimal Neyman-Pearson result is influenced by the accuracy of the unused machine (or machines, if there is more than one) and also by the probability of getting each machine. Cox showed that, when σ_1 is much greater than σ_2, the optimal 5% test of H: $\mu = 0$ against any right-sided alternative hypothesis uses (approximately) the rejection rule:

$$\text{Reject H if } x > \begin{cases} 1.28\sigma_1, & a = 1 \\ 5\sigma_2, & a = 2. \end{cases}$$

Thus when $A = 1$, the conditional significance level $(P_H(X_1 > 1.28\sigma_1))$ is close to 10%, while when $A = 2$, it is close to *zero*. The unconditional significance level is the average of these and is indeed 5%, since we were equally likely to observe $A = 1$ or 2, but we have the same varying *conditional* significance levels that we noticed in the optimal test in Welch's Uniform example.

To confirm that this approach is, nevertheless, Neyman-Pearson optimal, we will look more closely at an example of this type and cite the exact values rather than approximates.

Example 4.2

Suppose X_a is Normally distributed and we want to test H: $\mu = 0$ against K: $\mu = 5$. The variance of X_a is either $\sigma_1^2 = 4$ (probability $\frac{1}{2}$) or $\sigma_2^2 = 1$ (probability $\frac{1}{2}$) depending on the outcome of a toss of a fair coin (A). A test with *conditional* significance levels of 5% (for each a) will Reject H whenever $x > 1.64485\sigma_a$. Since the conditional significance levels (α_1 and α_2) of such a test are both equal to 5%, the overall (unconditional) significance level ($\alpha = \frac{1}{2}\alpha_1 + \frac{1}{2}\alpha_2$) is also 5%, so this is a '5% test' in the unconditional sense. The conditional power values are $\kappa_a = P_K(X_a > 1.64485\sigma_a \mid A = a)$ $(a \in \{1, 2\})$, which are $\kappa_1 = 80.3765\%$ and $\kappa_2 = 99.9603\%$, hence the power of the test ($\kappa = \frac{1}{2}\kappa_1 + \frac{1}{2}\kappa_2$) is 90.1684%.

On the other hand, the most powerful 5% test uses the rejection rule *Reject H whenever*:

$$x > \begin{cases} 2.62906 = 1.31453\sigma_1, & a = 1 \\ 2.53227 = 2.53227\sigma_2, & a = 2. \end{cases}$$

For this test $\alpha_1 = 9.4334\%$ and $\alpha_2 = 0.5666\%$ giving an overall significance level of 5% as required. The conditional power values are $\kappa_1 = 88.2084\%$ and $\kappa_2 = 99.3202\%$ giving an overall power of 93.7643%; this is higher than the power of the conditional test for the same (unconditional) significance level and therefore superior from a Neyman-Pearson point of view. These results are summarised below.

Table 4.1

Test	Conditional	Unconditional (NP)
α_1 (%)	5.0000	9.4334
α_2 (%)	5.0000	0.5666
α (%)	5.0000	5.0000
κ_1 (%)	80.3765	88.2084
κ_2 (%)	99.9603	99.3202
κ (%)	90.1684	93.7643

Were we to replicate this experiment infinitely many times (including the coin tossing part of the experiment), we would reject H 5% of the time when H is true using either test procedure, but, when K is true we would reject H 90.17% of the time, if we use the conditional test, but 93.76% of the time, if we use the optimal test. On the other hand, were we to replicate only the second stage of the test infinitely many times (for the observed value of a), then, when $a = 1$ and H is true, we will reject H 5% of the time for the conditional test, but 9.43% of the time for the optimal test, and when $a = 2$ and K is true, we will reject H 99.96% of the time for the conditional test but 99.32% of the time for the optimal test.

Welch and Cox hold exactly opposite positions on the central question: for Welch the overall power is the *real* power precisely because, although we happened to observe $A = 1$, we might have observed $A = 2$ and it would be wrong for us to ignore this; for Cox, once we have observed $A = 1$, the fact that we might have observed $A = 2$ becomes a complete irrelevance – we might have observed it, but we did not. (We find an echo of Welch, even today, in Mayo's frequent critical references to Bayesians

using features of the data that "just happened to occur"[16].) Since the optimal test has higher power over the unconditional sample space than the conditional approach, it follows that we must choose between the pursuit of relevance and the pursuit of (unconditional) power. In a passage that became famous, Cox identified the fundamental difference between the two approaches:

Now if the object of the analysis is to make statements by a rule with certain specified long-run properties [e.g. significance level, power], the unconditional test just given is in order, although it may be doubted whether the specification of desired properties is in this case very sensible. If, however, our object is to say 'what we can learn from the data that we have', the unconditional test is surely no good.[17]

Ancillary statistics and notional two-stage experiments.

The two-stage structure of this example means that it is easy to identify the ancillary statistic and probably also makes it easier for us to decide which parts of the sample space are relevant and which not. It is possible to think of all experimental designs that produce ancillary statistics as having a 'built-in' two-stage structure.

Data, v, may be the product of a single stage experiment, but, if it has the structure $v = (a, x)$ where a is ancillary, then we can think of it as follows. Conceive of another experiment that involves two stages where the first randomly generates a from a distribution, f_A (independent of the parameter of interest) and the second generates x from $f_X(\cdot; a, \theta)$ (which depends upon both a and θ). All the features produced by this experiment, and considered relevant for any version of inference (sample space, probability mass functions for different θ etc.) are the same as those produced by the initial experiment yielding v directly. For example, in the Welch case $(n = 2)$, the distribution of (M, R) is equally consistent with either of the following scenarios:

[16] Mayo, p. 350, for example.
[17] Cox (1958), pp. 360, 361.

i. Sampling r from the 'triangular' population ($f_R(r) = 2(1-r)$) and then sampling m from the Uniform $(\theta \pm \frac{(1-r)}{2})$, or

ii. Sampling x_1 and x_2 from the Uniform $(\theta \pm \frac{1}{2})$ and then calculating m and r from them.

Should we make our inferences the same way in either case? It is difficult to argue that the physical features of the experiment should be allowed to dictate the form of the analysis even when they do not affect the mathematical structure; such an approach would require a new extra-mathematical theory. Unless you are prepared to contemplate this possibility, your position on the two-stage case will need to carry over to differently structured experiments with ancillary statistics embedded in the observations[18].

Cox's paper affected attitudes towards the Welch example, which he cited, without going into details, as a similar example to his own. After 1958, almost[19] all the references to the Uniform example cite it in criticism of unconditional Neyman-Pearson inference. In the same paper in which he discussed the Welch example, Fraser wrote "… fiducial [conditional] probability gives an answer to the question D. R. Cox in his 1958 paper felt that statistical inference should answer: 'What do the data tell us about θ?'"[20]. An interesting question, that we will not attempt to answer, is why it was that Cox was so much more successful in arguing the case for conditioning than Fisher had been. Perhaps it was simply that Fisher, who had contempt for Neyman-Pearson theory, never bothered to apply his argument to such a clear-cut case; his notoriously difficult personality may also have been a factor.

[18] An argument analogous to the above is sometimes given to justify the sufficiency principle by observing that any data can be thought of as the product of a two-stage experiment in which the first stage yields the value of a sufficient statistic, s, (distribution dependent on θ) and the second stage yields the 'rest' of the information from a distribution ($X \mid S = s$) independent of θ; since the second stage contains no information about θ, we can make our inference based solely on the outcome of the first stage or, in general, based solely on the value (and distribution) of a sufficient statistic.

[19] See Chapter 5 for a prominent exception.

[20] Fraser (1961), p. 670.

4.4 Birnbaum (1962).

Three Principles.

Recall the two principles defined in Chapter 2.

The Sufficiency Principle (SP):
For $\theta \in \Theta$, suppose an experiment results in data $\underset{\sim}{x} \in \mathfrak{X}$ and $s(\underset{\sim}{x}_1) = s(\underset{\sim}{x}_2)$, where $s(\underset{\sim}{X})$ is sufficient for $\theta \in \Theta$, then our inference about $\theta \in \Theta$ from $\underset{\sim}{x}_1$ should be the same as that from $\underset{\sim}{x}_2$.

This principle can be seen as asserting *the irrelevance of observations not part of a sufficient statistic.*

The Likelihood Principle (LP):
For $\theta \in \Theta$, suppose an experiment, E_1, can result in data, $x \in \mathfrak{X}$, with likelihood $L_1(\underset{\sim}{x};\theta)$, and an experiment, E_2, can result in data, $y \in \mathfrak{Y}$, with likelihood $L_2(\underset{\sim}{y};\theta)$. If, for some $\underset{\sim}{x}_0$ and $\underset{\sim}{y}_0$, it is the case that $L_1(\underset{\sim}{x}_0;\theta) = k \times L_2(\underset{\sim}{y}_0;\theta) \quad \forall \theta \in \Theta$ (where k is independent of θ), then our inference about $\theta \in \Theta$ from $\underset{\sim}{x}_0$ should be the same as that from $\underset{\sim}{y}_0$.

This principle can be seen as asserting *the irrelevance of outcomes not actually observed*.

Obviously (see Chapter 2), the LP entails the SP, which is simply a version of the LP with scope limited to a single experiment. We have also shown that Neyman-Pearson inference satisfies the SP (if we disallow randomising variables) but not the LP, which is breached as a result of the way in which tail areas are used to make inferences.

The unrestricted conditional principle.

Now consider a third principle, which we will call the conditional principle[21] (CP). This principle encapsulates the following view. If (to put it in concrete terms) we are allocated a certain 'machine' for performing our experiment, then the nature and probabilities of any other machines that might have been used, but were not, should not affect the inference we make from the outcome of our experiment, as long as those probabilities where independent of the value of θ. Equivalently, it can be described as asserting *the irrelevance of (component) experiments not actually performed.*

The Conditionality Principle (CP):

For $\theta \in \Theta$, suppose that in some experiment, E_1, we observe the value of a random variable $\underset{\sim}{X}_1 \in \mathfrak{X}_1$ with likelihood (density) $L_1(\underset{\sim}{x}_1;\theta)$ and in another experiment, E_2, we observe the value of a random variable $\underset{\sim}{X}_2 \in \mathfrak{X}_2$ with likelihood $L_2(\underset{\sim}{x}_2;\theta)$. Suppose also that we observe the value of a random variable $A \in \{1,2\}$ where $P(A=1) = P(A=2) = \frac{1}{2}$, independent of $\theta, \underset{\sim}{X}_1$ or $\underset{\sim}{X}_2$. Consider the two-stage experiment, E^*, where we first observe the value of a and then carry out E_1 if $a=1$ or E_2 if $a=2$, hence we observe the value of $(A, \underset{\sim}{X}_A)$. Then, $\forall a$, we should infer the same about $\theta \in \Theta$ from $(a, \underset{\sim}{x}_a)$ (derived from E^*) as from $\underset{\sim}{x}_a$ (derived from E_a).

This principle is consistent with Cox's view (and Fisher's) regarding the issue of relevance not adequately addressed by optimal Neyman-Pearson theory (or the concept of sufficiency), although Cox (like Fisher) advocated a principle that is weaker than that stated above, in order to be able to remain within the frequentist framework. (Arguments for different conditionality principles will be discussed in Chapter 5.) Note that repeated applications of the above principle will cover cases where the ancillary statistic, A, takes more than two values and is not necessarily uniformly distributed. Also note that none of these principles specify what our

[21] The term 'conditional principle', unqualified, always refers to this principle, which might also be called 'unrestricted' in contrast to the 'restricted CP' defined in Chapter 5.

inferences should be; they only specify that certain perceived irrelevancies not be allowed to influence the interpretation of the data.

Of these three principles, the SP is accepted within a wide range of different statistical theories; in a frequentist context it is due to Fisher, but the fact that Neyman-Pearson inference can be made consistent with it (despite being ultimately based on error probabilities) has probably reinforced its status.

The LP is consistent with Bayesian inference and it is possible to make non-Bayesian inferences that are also consistent with it, but all forms of frequentist theory are at odds with it and some of its implications (such as the irrelevance of the stopping rule) are regarded with distaste by many frequentists.

As we have seen, the CP (or the concept behind it) was first advocated by Fisher but dismissed by Neyman and Pearson, however, after Cox's paper and the later commentaries on Welch's example, it came into some favour among frequentists generally.

In 1962, A. Birnbaum presented to an eminent audience a long and intricate paper entitled "On the foundations of statistical inference". Birnbaum was "unusual among statisticians in that he actively sought contact with philosophers as well as with methodologists in various sciences"[22]. He was a frequentist, who supported the control of error probabilities but believed that frequentist inference needed to have a valid evidential interpretation to be useful in science; he described his paper as being about "informative inference" and "experimental evidence"[23]. In it he proved the theorem that bears his name and makes an astounding claim, namely, that the SP and CP together entail the LP (and vice versa) so that the LP is logically equivalent to the conjunction of the SP and CP; hence, that "two principles widely held by non-Bayesian statisticians … jointly imply an important consequence of Bayesian statistics".[24]

[22] Giere, p. 5.
[23] Birnbaum (1962), p. 269.
[24] Giere, p. 6.

Birnbaum's theorem: $\boxed{SP \ \& \ CP \Leftrightarrow LP.}$

It follows from this that anyone wishing to uphold *both* the SP *and* the CP must also uphold the LP, which, in turn, bars the use of frequentist methodologies.

Reactions to Birnbaum's theorem.

Birnbaum's theorem has provoked very strong reactions from the day it was first presented. The following responses are from its initial audience.

L. J. Savage.

… it seems to me that this is really a historic occasion. This paper is a landmark in statistics because it seems to me improbable that many people will be able to read this paper or to have heard it tonight without coming away with considerable respect for the likelihood principle. … not to take the principle seriously no longer seems possible … I, myself, came to take … Bayesian statistics … seriously only through recognition of the likelihood principle …[25]

Jerome Cornfield.

… I haven't quite recovered from the shock of seeing that two principles I had thought reasonable and one which I had thought doubtful imply each other. It is clear that I must either believe all three or disbelieve at least one of the two reasonable ones. What is not clear is on what basis this choice should be made. One basis for this choice is provided by consideration of a consequence of the likelihood principle – the irrelevance of the stopping rule[26]**.**

Irwin Bross.

[After calling the previous speakers "a mutual admiration society of Bayesians"]

… the scientific value of this recommendation is dubious … if this recommendation is examined from a practical standpoint, it is very bad advice. It would probably be very little short of disastrous to a scientist who followed it

[25] Savage, L. J. in Birnbaum (1962) discussion, p. 307.
[26] Cornfield, J. in Birnbaum (1962) discussion, p. 309.

… Finally I would like to point out that the basic themes of this paper were well-known to Fisher, Neyman, Egon Pearson and others, well back in the 1920's. But these men realised, as the author doesn't, that the concepts cannot be used directly for scientific *reporting*. So, they went on to develop confidence intervals in the 1930's, and these proved to be very useful. The author here proposes to push the clock back 45 years, but at least this puts him ahead of the Bayesians, who would like to turn the clock back 150 years[27].

George E. P. Box.
[Taking issue with Bross]
… although I pride myself on being a practical person … it would be very difficult to persuade an intelligent physicist that current statistical practice was sensible, but there would be much less difficulty with an approach via likelihood and Bayes' theorem[28].

Most of the responses reflected the ideological positions of their makers. Even those who were less dogmatic recognised that Birnbaum's theorem has very great implications and presents frequentist adherents of the CP with a real dilemma; the first edition of Kendall & Stuart to come out after Birnbaum's paper contains the following comment:

However the real question is whether we should [use the CP] … This question has far-reaching implications, since A. Birnbaum (1962) has shown that the Conditionality Principle implies (as well as being obviously implied by) the *Likelihood Principle*, which states that *only* the LF [likelihood function at x] need be regarded in making any statistical inference from observations. In particular, this has the consequence that the details of the sampling procedure which produced the observations (and the LF) are strictly irrelevant to subsequent statistical inference. Many, perhaps most, statisticians will find it intuitively unacceptable to eliminate the sample space from consideration in

[27] Bross, I. in Birnbaum (1962) discussion, p. 309, 310.
[28] Box, G. E. P. in Birnbaum (1962) discussion, p. 311.

making inferences from observations. If so, they must reject the Likelihood Principle, and the Conditionality Principle must automatically go with it[29].

Despite this ambivalence, Birnbaum's theorem has probably caused a number of people to reconsider the LP because of its equivalence to two such respectable axioms; E. T. Jaynes described the theorem as "The first proof of the 'likelihood principle' to be accepted by anti-Bayesians."[30]

"On the foundations of statistical inference" is a long paper; Birnbaum discusses many interesting points in addition to presenting the famous theorem. One of these points is that, for certain binary parameter spaces for the Bernoulli parameter, p, there exists an ancillary variable, with the result that the long-standing inferential methods used in this case are in breach of the conditionality principle. This example will be considered in detail in the following chapter.

[29] Stuart & Kendall (1967), p. 217.
[30] Jaynes (2003), p. 684.