Chapter 6: Welch's Uniform example re-visited.

6.1 Introduction.

In Chapter 5 we showed that there are some serious difficulties with the application of the restricted conditional principle; notably: frequent absence of an exact ancillary statistic, difficulty of choosing between competing ancillary statistics, and lack of continuity of inference. We can resolve these problems by adopting the unrestricted CP of Birnbaum, but in that case we will need to abandon frequentism altogether and adopt the LP. If we wish to remain within the frequentist framework, we may still take the view that Cox's conditional approach is superior to conventional frequentist inference because it excludes an irrelevant part of the sample space. In view of the unsatisfactory nature of many standard unconditional inferences (outlined in Chapter 3), it makes sense to pursue the conditional option and, in particular, to examine how far we can extend its scope while still remaining frequentist. By re-examining Welch's Uniform example in detail, we are able to identify serious flaws in both of the methods used to date; this insight points us in the direction of an alternative conditional approach, one that produces superior results. In later chapters we will find that the new conditional approach is applicable to many other, more realistic, scenarios.

Welch's example is valuable because it is a simple case where an ancillary statistic (the range, R) allows us to contrast the unconditional Neyman-Pearson approach with the conditional approach of Fisher; the advantages of this were recognised immediately. Welch was convinced of the superiority of Neyman's method and his view did not change as a result of studying this example though surely he did not study it very closely. Later commentators looked at the conditional features of the confidence intervals and tests and were inclined to support Fisher's point of view. However we believe that both sides of the debate have overlooked some important features of the example.

We will start by re-stating the connection between Neyman-Pearson critical regions and the likelihood ratio statistic, and considering the issue of randomisation.

6.2 Randomising: Adding arbitrary requirements to a test.

Consider a test of two simple hypotheses. If the critical region, $\Re$, is not based strictly on the likelihood ratio statistic (i.e. is not of the form: $\{\underset{\sim}{x} : LR(\underset{\sim}{x}) \leq k\}$, for some k), then the critical region is arbitrary with respect to the error probabilities α and β, that is, some *other* critical region(s) will yield the same error probabilities. In such a case, even if $\Re$ is most powerful, it is not uniquely so. In the context of two simple hypotheses, such a test is in breach of the sufficiency principle because the likelihood ratio statistic is sufficient on a binary parameter space, and if $\Re$ is not of the form given above, then $\exists \underset{\sim}{x}_1, \underset{\sim}{x}_2 : LR(\underset{\sim}{x}_1) = LR(\underset{\sim}{x}_2)$ and $\underset{\sim}{x}_1 \in \Re, \underset{\sim}{x}_2 \notin \Re$, that is, two observations giving the same value of a sufficient statistic nevertheless produce different inferences.

A test of this kind is sometimes used in order to obtain a conventional significance level, say 5%, when the likelihood ratio statistic (and usually also the natural statistic, $\underset{\sim}{X}$) is discrete[1].

Example 6.1

Suppose $X \sim Bin(5, p)$ and we want to test H: $p = \frac{1}{2}$ against K: $p = \frac{3}{4}$ at the 5% level. The null and alternative distributions of X, and the likelihood ratio of each value of x, are shown in table below. (Note that small values of the likelihood ratio (e.g. $\frac{32}{243}$ and $\frac{32}{81}$) are associated with large values of x (5 and 4).)

[1] Stuart *et al*., p.174.

Table 6.1

x	0	1	2	3	4	5
$P_H(x)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
$P_K(x)$	$\frac{1}{1024}$	$\frac{15}{1024}$	$\frac{90}{1024}$	$\frac{270}{1024}$	$\frac{405}{1024}$	$\frac{243}{1024}$
$LR(x)=\frac{P_H(x)}{P_K(x)}$	$\frac{32}{1}$	$\frac{32}{3}$	$\frac{32}{9}$	$\frac{32}{27}$	$\frac{32}{81}$	$\frac{32}{243}$

The significance level for the critical region, $\{x: LR(x) \le \frac{32}{243}\} \equiv \{x = 5\}$, is $\frac{1}{32} < 5\%$, while the significance level for the critical region, $\{x: LR(x) \le \frac{32}{81}\} \equiv \{x = 4 \text{ or } 5\}$, is $\frac{6}{32} > 5\%$. No value of k for a critical region of the form $\{\underset{\sim}{x}: LR(\underset{\sim}{x}) \le k\}$ produces a significance level of exactly 5%.

We can solve this problem by using a 'randomised test', which introduces a device capable of producing an event with a probability of 12%. Let V be any random variable, independent of X, such that $P(V = v^*) = 0.12$, for some v^*. If we use the rejection rule: 'reject H if $\{x = 5\}$ *or* $\{x = 4$ *and* $v = v^*\}$', the significance level of the test is:

$$P_H(X = 5) + P_H(X = 4) \cdot P(V = v^*) = \tfrac{1}{32} + (0.12 \times \tfrac{5}{32}) = 5\%.$$

There is something very unsatisfactory about this rejection rule. Although it produces the required long run failure rate, it does so by making the inference critically dependent on the observation of an irrelevant variable, V. This introduces an element into the definition of the rejection region over and above that which can be described in terms of the likelihood ratio of the data. We can see that this is in breach of the sufficiency principle (for *any* parameter space $\Theta \subseteq (0,1)$) because the two observations $\{x = 4, v = v'\}$ and $\{x = 4, v = v^*\}$ result in different inferences even though x is a sufficient statistic for p and is the same in each case.

This approach may make sense in some quality control environments but not when we are interested in evidence, since $\{x = 4, v = v'\}$ and $\{x = 4, v = v^*\}$ contain exactly the same evidence about the relative status of any two hypotheses about p. It was for this reason that Birnbaum regarded the frequentist 'confidence concept of statistical

evidence' as necessarily incorporating "the sufficiency concept, expressed in the general refusal to use randomised tests".[2]

6.3 Both analyses of the Uniform model are wrong.

Let us return to the Uniform example. There are two distinct problems with this example; one is the existence of an unrecognised and unintentional arbitrariness identical to that produced by the use of randomising devices, the second is that the choice of significance level is often inappropriate and introduces a bias into the procedure as well as reducing the success rate unnecessarily. These issues only become apparent when we look at tests of two simple hypotheses, however their implications flow through to tests of composite hypotheses and confidence intervals (as shown below). The problems occur regardless of which type of inference – conditional (on r) or unconditional – we use.

$X_1,\ldots,X_n$ are independent and identically distributed Uni$(\theta-\frac{1}{2},\theta+\frac{1}{2})$ random variables. An optimal hypothesis test, at significance level α, will have a rejection region that maximises the power of the test; in the process it produces conditional significance levels that vary with r, sometimes higher and sometimes lower than the nominal level α. By contrast Fisher's method uses a rejection region where all the conditional significance levels, α_r, are individually equal to α (and thus the overall level also equals α) but which, as a result, has slightly lower overall power than the Neyman-Pearson test. When $\alpha = 0$, the two approaches produce the same rejection region since each of the α_r values must also be equal to *zero*. This is also true for interval estimation; the Neyman-Pearson 100% confidence interval and Fisher's 100% interval are both equal to $(\theta \pm \frac{1}{2}(1-r))$, the conditional coverage being 100% for all r as well as overall. This is the shortest interval that is certain to contain θ.

[2] Birnbaum, A. (1970). Statistical methods in scientific inference, *Nature*, **225**, p. 1033, quoted in Giere, p. 9.

(All the technical details that follow are for the case $n = 2$ although the issues discussed arise for all n.) For a test of H: $\theta = \theta_1$ versus K: $\theta = \theta_2$ ($\theta_2 > \theta_1$, WLOG) where $n = 2$, the optimal, level α, Neyman-Pearson rejection region is

$$(m, r) : m > \begin{cases} \theta_1 + \frac{1}{2}(1-r), & r \geq \sqrt{\alpha} \\ \theta_1 + \frac{1}{2}(1-r) - (\sqrt{\alpha} - r), & r < \sqrt{\alpha}. \end{cases}$$

Fisher's rejection region, based on the distribution of M conditional upon the observed r is

$$m : m > [\theta_1 + \tfrac{1}{2}(1-r) - \alpha(1-r)].$$

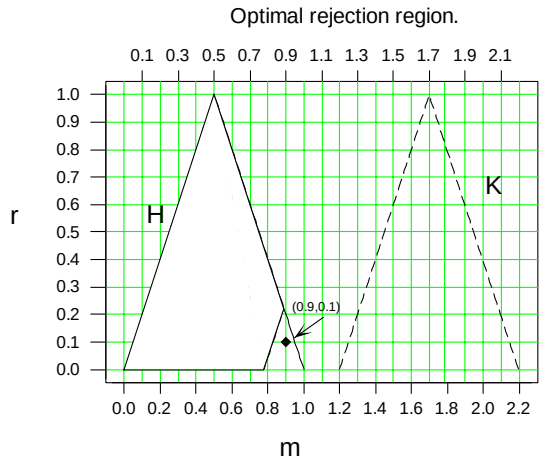
The characteristics of each test are somewhat dependent on how far apart the hypothesised values are, that is, on $\Delta = \theta_2 - \theta_1 > 0$. We assume that $\alpha > 0$ is a fixed value; the value of α used in our diagrams is 5%.

First consider the case where the hypothesised values of θ are sufficiently far apart that the supports either do not overlap or overlap by a region with a probability of less than α. (The probability of any part of the intersection of the two supports is the same under both hypotheses.)

Large Δ, unconditional (NP) test.

In the following diagrams, the *acceptance region* has no grid markings; the remainder of the union of the two supports constitutes the rejection (critical) region.

Figure 6.1



This test has the following features:

i. The α-sized part of the rejection region, which is in the null support (i.e. the support of (M, R) under H), is arbitrarily chosen. If we replaced it with another area of the same size in the support under H, it would not affect either of the error probabilities; it follows that we could find a rejection region with the same optimal error probabilities to include (or exclude) *any* particular data point in the support under H.

ii. The probability of the Type II error, β, is *zero* (power of 100%) whereas $\alpha > 0$. In a symmetric case like this, it follows that the test is biased in favour of K.

iii. The power is 100%. Usually, reducing α will reduce the power as well, but not in this case. We could reduce α to *zero* and still have power of 100%; thus, no matter what level of (relative) importance we attach to α and β, it is clear that using a positive α is inefficient, that is, α is *unnecessarily* large. Reducing α to *zero* will remove the bias, remarked in ii, at no cost.

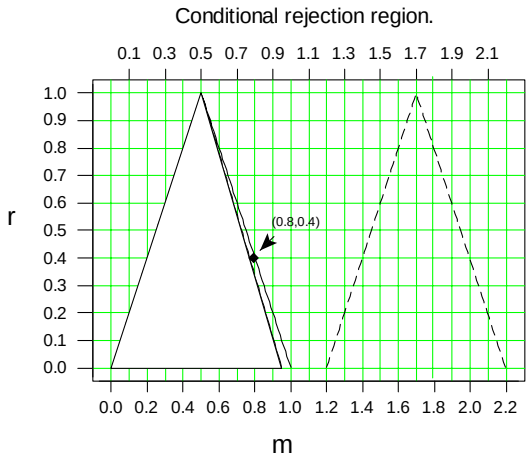
iv. The test rejects H in favour of K when (for instance) $(m, r) = (0.9, 0.1)$ (see **Figure 6.1**). This is worse than merely rejecting H when the data is more consistent with H than with K (as previously observed when $\beta < \alpha$); this data simply cannot

occur if K is true; observing this data, we know that H is certainly true yet we reject H because of it.

v. The supports shown in the diagram above do not overlap at all; if they overlap by an area with probability *less than* α, i.e. with probability of $\alpha - \delta$ $(\delta > 0)$, then the same points apply with the following modifications: (i) the δ-sized part of the rejection region is chosen arbitrarily, (ii) remains unchanged, (iii) we can reduce the significance level from α to $\alpha - \delta$ while retaining power of 100%, this will reduce the bias in favour of K, (iv) still true for data in the 'δ' part of the rejection region.

Large Δ, conditional (Fisher) test.

Figure 6.2

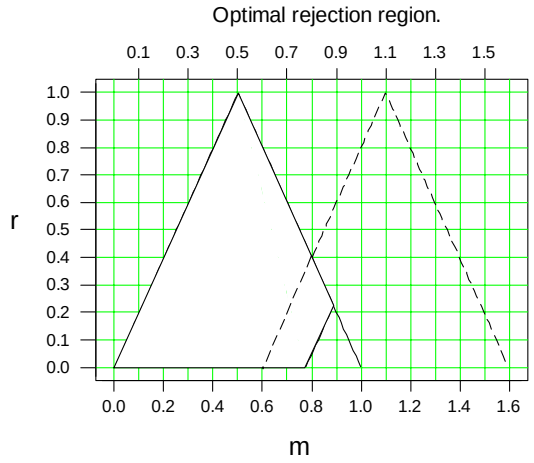


The test conditional upon r (see above) has the same unsatisfactory features detailed above for the Neyman-Pearson case; simply replace α, β and *power* with α_r, β_r and conditional power in the discussion. The data $(m, r) =$ (0.8,0.4) (shown above) can make the required point in (iv).

Small Δ, NP test.

Now consider the case where Δ is small, i.e. the hypothesised values of θ are relatively close together.

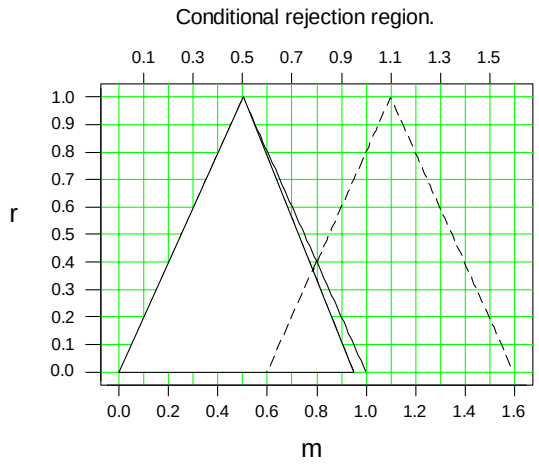
Figure 6.3



Again the critical region is arbitrarily defined; the α-sized part of the rejection region in the null support could equally have been sited anywhere in the overlap area, which is larger, without affecting either of the error probabilities. As in the previous case, the highest power test is not unique – we could choose from an infinite number of them. The rejection region usually quoted (shown above) may seem reasonable because we are used to having the rejection region for a right-sided test ($\theta_2 > \theta_1$) on the right side of the null support; this often produces a unique most powerful rejection region, but in this case it is only one of many. This happens because of features peculiar to the Uniform model: $x = 0.5$ is no *more* consistent with a Uni(0,1) than is $x = 0.98$, even though the first value is at the centre of the distribution and the latter is close to the edge; the uniform likelihood does not rise as we move closer to the centre as so many other densities do. It is important to emphasise how unacceptable this arbitrary element is; two analysts using exactly the same data (and same α) can quite validly make opposite inferences simply by choosing to use different rejection regions, and both of them can call their test 'most powerful' because no other test of the same significance level is more powerful.

Small Δ, conditional test.

Figure 6.4



These comments also apply to the conditional test when it comes to most of those values of r associated with the area where the supports overlap. If $r < (1-\Delta-\alpha)/(1-\alpha)$, then (under H) m can only take values in the interval $(\theta_1 \pm \frac{1}{2}(1-r))$, part of which is in the acceptance region and part in the rejection region. That proportion of the interval for which we reject H could be placed *anywhere* within the overlap interval, $(\theta_2 - \frac{1}{2}(1-r), \theta_1 + \frac{1}{2}(1-r))$, without changing α_r or β_r – it does not have to be on the extreme right of the interval as in the region shown above. For example, when $r = 0.1$, the rejection region for m is [0.905, 1.550], of which [0.905,0.950] is in the null support; this interval could be replaced by an interval of the same width anywhere within [0.65,0.95] without changing either of the error probabilities.

The conditional test also has a problem not present in the Neyman-Pearson test when Δ is small. Part of the rejection region associated with larger r lies completely outside the support under K, encouraging us to reject H when we see data proving that H is true (e.g. data $(m, r) = (0.69, 0.6)$ in the above plot); for the Neyman-Pearson test this only happens when $\Delta > 1 - \sqrt{\alpha}$, but for the conditional test it is a problem for

some r, no matter what the values of α and $\Delta(>0)$. It also follows that, for many values of r, we could reduce α_r to *zero* without increasing β_r; it would seem to be a requirement of any reasonable methodology that $\alpha_r = 0$ in such a case.

These problems have some unpleasant implications for interval estimates as well as tests. We will consider only the interval estimates conditional on r, now widely regarded as superior to the unconditional intervals.

Confidence intervals conditional on *r*.

Conditional on $R = r$, $M \sim \text{Uni}(\theta \pm \frac{1}{2}(1-r))$. Based on this, we can find the conditional 100% confidence interval for θ, $\mathbb{C}_{100}(m) \equiv (m \pm \frac{1}{2}(1-r))$, which has a width of $(1-r)$ (and is the same as the 100% *unconditional* confidence interval for θ). Conventionally (and in accordance with the conditional rejection region for positive α examined above)[3], the conditional 90% confidence interval is defined as $(m \pm 0.45(1-r))$, i.e. it is the central *nine-tenths* of $\mathbb{C}_{100}(m)$. However *any* interval of the same width, lying entirely within $\mathbb{C}_{100}(m)$ – not necessarily in the centre – will also have coverage of 90%. More generally, any interval within $\mathbb{C}_{100}(m)$ of width $(1-\alpha)(1-r)$ has coverage of $100(1-\alpha)\%$, and any union of disjoint intervals all within $\mathbb{C}_{100}(m)$ having a combined width of $(1-\alpha)(1-r)$ also has total coverage of $100(1-\alpha)\%$. It is impossible to choose between these options on the basis of either width or coverage. If all these intervals (and even unions of intervals) are considered valid – and there is no statistical basis for preferring any one to another – it follows that *any* point in $\mathbb{C}_{100}(m)$ can be included in or excluded from a valid confidence interval of any coverage less than 100% purely on the whim of the analyst. Only when $\alpha = 0$ is there a unique shortest[4] interval with the specified coverage, namely $\mathbb{C}_{100}(m)$.

[3] See, for instance, Welch (1939), uncontested in the subsequent literature.
[4] Usually intervals are required to be 'shortest on average'; since we are conditioning on r it is not appropriate to average over data with differing ranges, averaging over m (for a given range) is redundant since all these intervals have the same width.

Views on randomisation vary, but conditional inference has, from the time of Fisher onwards, always been discussed in the context of evidential inference; the aim is to extract all the relevant information from the data, not merely achieve predetermined, low error rates (see Fisher, Cox, Pratt, and Birnbaum, for example); this rules out methods with random components. The proponents of conditional inference would not have approved of this inference had they been aware of its arbitrary attributes.

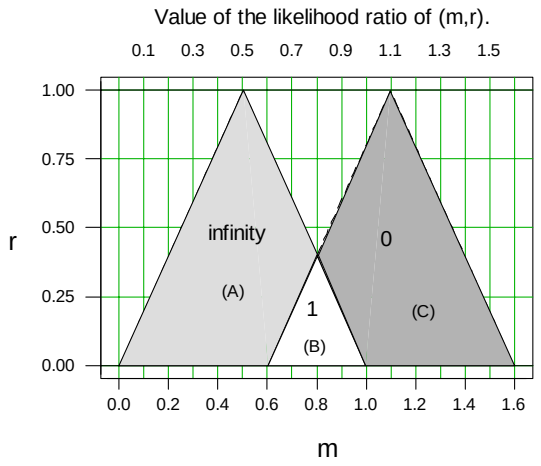
6.4 Using the likelihood ratio statistic in the Uniform example.

To understand what is going wrong, we need to look at the likelihood ratio statistic, $LR(\underset{\sim}{X})$. For each x_i $(i = 1,...,n)$, the density, $f(x_i;\theta)$, is either *one* or *zero* under the uniform model, so the likelihood (joint density) of $(x_1,\ldots,x_n)$, $L(\underset{\sim}{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$, can also only take the values *one* or *zero*. The likelihood is *one* only when $\theta - \frac{1}{2} < x_1,\ldots,x_n < \theta + \frac{1}{2}$ or, equivalently, when $\theta - \frac{1}{2}(1-r) < m < \theta + \frac{1}{2}(1-r)$. Since, for any θ, the likelihood of $\underset{\sim}{x}$ can only take these two values, there are at most *four* possibilities for the likelihood ratio[5] of any $\underset{\sim}{x}$: $\frac{0}{0}$, $\frac{0}{1} = 0$, $\frac{1}{0} = \infty$, and $\frac{1}{1} = 1$. The first of these indicates that the data we have observed cannot occur under either hypothesis; assuming that either H or K is true (or that we do not test H against K when it is clear that neither is true), we may rule this out. It follows that the likelihood ratio statistic is a discrete variable taking only three values even though the original X variables are continuous. This explains why a seemingly reasonable critical region was arbitrary; when the likelihood ratio statistic is discrete, only a limited number of α-values can be used without resorting to the equivalent of a randomising device, however unconsciously.

[5] $\frac{0}{0}$ and $\frac{1}{0}$ are both formally undefined. The first we can exclude for the reason given below, the second we interpret as ∞ i.e. $\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}$, where $\varepsilon \to 0$ from above, since likelihoods are non-negative. The actual value assigned to the likelihood ratio is important (for our purposes) only up to their ordering; clearly, for any $0 < \varepsilon < 1$, the value $\frac{1}{\varepsilon} > 1 > 0$.

We can consider the likelihood ratio as a function of (m,r), instead of as a function of $\underset{\sim}{x}$, since (m,r) is a sufficient statistic. The diagram below shows the likelihood ratio as a function of (m,r) for the hypotheses H: $\theta = 0.5$ versus K: $\theta = 1.1$.

Figure 6.5



The areas in the (m,r)-plane corresponding to the three different likelihood ratio values are labelled A, B and C. When $\alpha > 0$, the conditional rejection region always contains part, but not all, of A, and it often contains part, but not all, of B as well. (B will not exist if Δ is large enough.) The unconditional rejection region $(\alpha > 0)$ contains either part, but not all, of B (when Δ is small) or part, but not all, of A (when Δ is large). (Only in the very specific case $\Delta = 1 - \sqrt{\alpha}$ can the unconditional rejection region be described in terms of values of the likelihood ratio statistic, this is a special case of option 'a', discussed below.) A rejection region of the form $\{(m,r) : LR(m,r) \le k\}$ would include (or exclude) *all* of any area of constant likelihood ratio. When we use the conventional critical regions, described above, we reject or accept H partly on the basis of features of the data over and above the likelihood ratio. The part of B that is included in the rejection region is chosen on a basis that is arbitrary with respect to the likelihood ratio, and therefore also with respect to the error probabilities; this is why we could change it without changing the error probabilities. One of the results of this is that, although no other test at the 5%

(conditional or unconditional) level has higher power (conditional or unconditional) than these tests, the tests are not *uniquely* most powerful. Any part of A that is included in the rejection region is also arbitrarily chosen and adds to the value of α without adding anything to the power.

We can get rid of this arbitrary quality by basing our rejection region strictly on the value of $LR(m,r)$. Since the likelihood ratio can only take three distinct values, our choice of rejection rules (and α or α_r) is very limited; we should reject H when the likelihood ratio is 'small' and it only remains to decide how small, i.e. which value to use as the cut-off value, k. The likelihood ratio statistic can only take the values $0,1$ or ∞; we can rule out using ∞ as the cut-off value, not (according to frequentist reasoning) because it is so large (although it is), but because it would produce a significance level of 100%. There are two other options.

Rule (a): Reject H whenever $LR(m,r) \le 1$, i.e. when $(m,r) \in \mathrm{B} \cup \mathrm{C}$.

In this test, we reject H whenever it is *possible* that K is true. The power of the test is 100% ($\beta = 0$) and the test is unique in achieving this power for the given significance level; it is therefore uniquely optimal in the Neyman-Pearson sense. However, the significance level, α, depends on Δ, and is given by:

$$\begin{cases} (1-\Delta)^2, & 0 < \Delta \le 1 \\ 0, & \Delta > 1. \end{cases}$$

This will be greater than 5% whenever $\Delta < 1 - \sqrt{0.05} \approx 0.776$; for instance, if $\Delta = 0.5$, the significance level is 25%.

The conditional significance level, α_r, is

$$\begin{cases} \dfrac{(1-\Delta)-r}{(1-r)}, & (r,\Delta) \in \{[0,1-\Delta)\times(0,1)\} \\ 0, & (r,\Delta) \in \{(1-\Delta,1)\times(0,1)\} \cup \{[0,1]\times[1,\infty)\}. \end{cases}$$

Since this varies with r, the test is not Fisherian. We can still calculate the conditional probabilities if we believe them to be more relevant; the conditional power is 100% for all r. The conditional significance level will be greater than 5% whenever $\Delta < 1$ and $r < 1-(\Delta/0.95)$. For instance if $\Delta = 0.5$ and $r = 0.3$ then $\alpha_r = 40\%$.

Clearly there is a problem with using this rejection region; whether or not we condition on r, our significance level will often be unacceptably high. Since the power is always *one* $(\beta = 0)$ but α (or α_r) is frequently positive, we can see that this test is often biased in favour of K and never in favour of H. We should consider using the alternative cut-off value for the likelihood ratio.

Rule (b): Reject H whenever $LR(m,r) = 0$, i.e. when $(m,r) \in \mathrm{C}$.

This test tells us to reject H whenever H cannot possibly be true. The significance level of the test (conditional on r or unconditional) is *zero* $\forall\Delta$. This is a better test than option 'a' because the significance levels (of all kinds) are *zero* (thus we call it the *zero-level* test); the test is never biased against H (from either a conditional or unconditional point of view). The rule, *Reject H whenever* $(m,r) \in C$, satisfies the requirements of both the (unconditional) optimal Neyman-Pearson test and the (conditional) Fisherian test; the former, because it has the (unique) highest power of any $\alpha = 0$ level test, and the latter, because the α_r values are all the same (*zero*), and the conditional power values are optimal for each r. However, the test still has both conditional and unconditional features, and we may ask which are more relevant. To answer this, we need a context. Clearly, there are circumstances (for instance, some

quality control situations) where the individual values of β_r may not be important, but the average, β, is. However, for the purposes of this discussion, we assume that we are primarily interested in identifying those measures that help us answer the question 'What does the data-set say about the evidence for the hypothesis H relative to K?' Below we give the formulae for the conditional and unconditional probability of Type II error; by looking at some specific cases, we may be able to assess which of these measures is more informative for our purposes.

Comparison of conditional and unconditional error probabilities for the *zero-level* test.

The unconditional probability of Type II error, β, is given by:

$$\beta = \begin{cases} (1-\Delta)^2, & 0 < \Delta < 1 \\ 0, & \Delta \geq 1. \end{cases}$$

The Type II error probability conditional upon r, β_r, is:

$$\beta_r = \begin{cases} \dfrac{(1-\Delta)-r}{(1-r)}, & (r,\Delta) \in \{[0,1-\Delta)\times(0,1)\} \\ 0, & (r,\Delta) \in \{(1-\Delta,1)\times(0,1)\} \cup \{[0,1]\times[1,\infty)\}. \end{cases}$$

We can check that β is the average (over r) of the β_r values, as follows.

$$\begin{aligned} E(\beta_R) &= \int_0^{1-\Delta} \beta_r \cdot f_R(r)dr \\ &= \int_0^{1-\Delta} \frac{(1-\Delta)-r}{(1-r)} \cdot 2(1-r)dr \\ &= (1-\Delta)^2. \end{aligned}$$

This test is never biased in favour of K; it is sometimes biased in favour of H and sometimes unbiased. Which tests we think are biased depends on whether we take the conditional or unconditional point of view. From the unconditional point of view, the

test is biased in favour of H when $\beta > \alpha = 0$. This happens whenever the two supports overlap, i.e. when $B \not\equiv \varnothing$. Otherwise the test is unbiased and $\alpha = \beta = 0$ so we can tell with certainty which of the two hypotheses is true; you would expect to be able to do this when the two supports do not overlap.

From a 'conditional on r' point of view, the test is biased in favour of H only when $\beta_r > \alpha_r = 0$, where r is the value of R observed in the experiment. This happens whenever the two supports overlap at the r level, i.e. when a horizontal line at height r passes through B. Otherwise the test is unbiased and $\alpha_r = \beta_r = 0$. This requirement is not as strong as $B \equiv \varnothing$, yet, since we can tell with certainty which hypothesis is true when our data is of this form, it seems clear that the conditional version of power is the superior measure. To illustrate this, consider the test of $H: \theta = 0.5$ versus $K: \theta = 1.1$ $(\Delta = 0.6)$. Suppose that we observe $r = 0.75$ (see **Figure 6.5** above); any value of $(m, 0.75)$ that we could possibly observe (under either hypothesis) completely rules out either H or K; thus the 'relevant' error probabilities are surely *zero*. The conditional values ($\alpha_{0.75}$ and $\beta_{0.75}$) are both *zero*, whereas the unconditional β is $P_K(B) = 16\%$ – a fact that is clearly irrelevant. Does the conditional power always work as well as this? Consider the data point $(m, r) = (0.8, 0.25)$; this point is in area B and, thus, not in the rejection region so we accept H; the conditional power when $r = 0.25$ is 80% which is high enough for us to consider that this data tends to support H over K to some degree. However, examination of the diagram suggests that the data is equally consistent with the two hypotheses; we will return to this point later.

The confidence interval.

The confidence interval that corresponds to the test that rejects H if and only if $LR(m, r) = 0$ is the 100% interval for θ: $(m \pm \frac{(1-r)}{2})$. The uniform case is unusual in having a meaningful 100% confidence interval and this interval is produced by both the Neyman-Pearson and Fisher approaches, that is, they agree about the $100(1-\alpha)\%$ interval *when* $\alpha = 0$. These intervals have the shortest average overall length of any

intervals with a 100% success rate, and this is equally true when we confine ourselves to looking at (long-run) data for any particular $R = r$. Note that there is no disagreement between the formal coverage or confidence level of this interval and what we know about its properties (as there was with the 90% unconditional intervals); it contains 100% of the possible values of θ and is indeed called a '100% confidence interval'.

6.5 Using a 'better' ancillary statistic for the Uniform case.

We have illuminated a number of issues by going directly to the likelihood ratio statistic; is there any other approach that is similarly edifying? In this section we show that the appropriate inference becomes obvious if we condition on an alternative ancillary statistic that has the Fisherian structure for the binary parameter space $\Theta_B = \{\theta_1, \theta_2\}$.

A Fisherian ancillary statistic on the binary parameter space.

A Fisherian ancillary statistic[6] is any statistic A such that $S \equiv (A, \hat{\theta})$ where A has the same distribution $\forall \theta \in \Theta$, S is minimal sufficient for $\theta \in \Theta$, and $\hat{\theta}$ is a maximum likelihood estimator (MLE) of $\theta \in \Theta$.

When we are dealing with a binary parameter space, the likelihood ratio statistic is the minimal sufficient statistic for θ in that space. In the Uniform case, the likelihood ratio statistic is a discrete statistic taking only three values (see **Figure 6.5** where it is shown as a function of (m, r)); (M, R) is not minimal sufficient for this parameter space since it discriminates unnecessarily between data points with the same likelihood ratio value. Since $LR(\underset{\sim}{X})$ is minimal sufficient, so is its natural logarithm: $\ln\{LR(\underset{\sim}{X})\}$. The maximum likelihood estimate, a function of the data $\underset{\sim}{x}$, is also dependent on the nature of the parameter space since it is that value of θ *in the*

[6] Fisher (1956), Basu (1963).

parameter space for which the likelihood $L(\underset{\sim}{x};\theta)$ is maximum. When we are considering the parameter space $\mathbb{R}$, the MLE is not unique but M is the unique unbiased MLE. When we are considering the parameter space Θ_B, the maximum likelihood estimator, $\hat{\theta}(\underset{\sim}{X})$, is whichever one of the parameter values has the higher likelihood given $\underset{\sim}{X}$. Since $LR(\underset{\sim}{X}) = L(\underset{\sim}{X};\theta_1)/L(\underset{\sim}{X};\theta_2)$, it follows that the maximum likelihood statistic can be written as:

$$\hat{\theta}(\underset{\sim}{X}) = \begin{cases} \theta_1, & \text{if } LR(\underset{\sim}{X}) > 1 \\ \theta_2, & \text{if } LR(\underset{\sim}{X}) < 1. \end{cases}$$

If $LR(\underset{\sim}{X}) = 1$, the likelihood is constant over Θ_B and it is pointless to talk of a maximum or even maxima.

Consider the statistic $(A(\underset{\sim}{X}), \hat{\theta}(\underset{\sim}{X}))$ where $A(\underset{\sim}{X}) = |\ln\{LR(\underset{\sim}{X})\}|$. Since $LR(\underset{\sim}{x})$ can only take the values 0, 1, and ∞, $\ln\{LR(\underset{\sim}{x})\}$ can only take the values $-\infty$, 0 and ∞ and $A(\underset{\sim}{x})$ can only take the values ∞ and 0.

Table 6.2

$(A(\underset{\sim}{x}), \hat{\theta}(\underset{\sim}{x}))$	$\ln\{LR(\underset{\sim}{x})\}$
$(0, *)$	0
(∞, θ_1)	∞
(∞, θ_2)	$-\infty$

There are three distinct values of both $(A(\underset{\sim}{x}), \hat{\theta}(\underset{\sim}{x}))$ and $\ln\{LR(\underset{\sim}{x})\}$, in a one-to-one correspondence with each other, hence $(A(\underset{\sim}{x}), \hat{\theta}(\underset{\sim}{x}))$ is a one-to-one function of $\ln\{LR(\underset{\sim}{x})\}$ and is thus a minimal sufficient statistic for $\theta \in \Theta_B$.

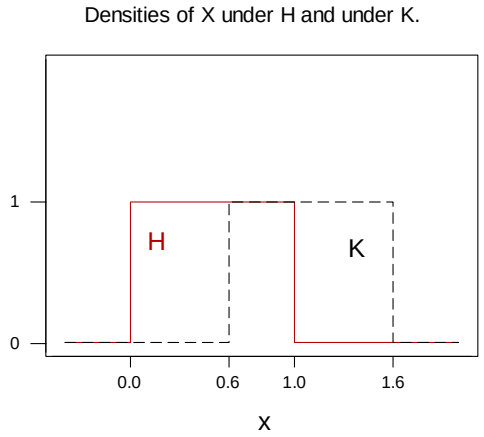
It is easy to show that A is an ancillary statistic on Θ_B. Note that (M, R) is still sufficient for $\theta \in \Theta_B$, even though not minimal sufficient, hence $LR(\underset{\sim}{X}) = LR(M, R)$ and $P(A = 0) = P\{LR(\underset{\sim}{X}) = 1\} = P\{LR(M, R) = 1\} = P\{(M, R) \in \mathrm{B}\}$. This probability

is the same under both hypotheses (i.e. for both $\theta \in \Theta_B$) – see **Figure 6.5**. Since A is dichotomous, it follows that it is ancillary over Θ_B and, since A and $\hat{\theta}$ together comprise the minimal sufficient statistic, it follows that A is a Fisherian ancillary statistic on Θ_B (as well as satisfying Cox's requirements). Thus, according to Fisher's theory, we should condition on the observed value of A when carrying out tests on $\theta \in \Theta_B$.

Interpreting A as a precision index.

The statistic A has a simple intuitive interpretation. Consider the distribution of the raw data-values X_i. We are sampling randomly from a distribution which is either $\text{Uni}(\theta_1 \pm \frac{1}{2})$ or $\text{Uni}(\theta_2 \pm \frac{1}{2})$. If the two distributions do not overlap, we should be able to tell (from any data), with absolute certainty, which is the true hypothesis. Neither maximising the power nor conditioning on r made this automatic; in both cases the conventional choice of positive α seemed reasonable but made it difficult to interpret the evidence correctly. The situation should be quite straightforward even when the distributions do overlap. The X_i variables have one density or another as shown in the diagram below.

Figure 6.6



In such a case, any data vector will either be *completely informative* regarding the two hypotheses or *completely uninformative*. For example, consider again testing H: $\theta = 0.5$ versus K: $\theta = 1.1$. Under H, $X_i \sim$ Uni(0,1) whereas under K, $X_i \sim$ Uni(0.6,1.6). If all the data falls in the overlap interval, [0.6,1.0], the two hypotheses are equally consistent with the data and we are none the wiser for observing it. On the other hand, if *not* all the data is in the overlap region, we know with certainty which of the two hypotheses is true[7]. The statistic, A, distinguishes between these two cases: $A(\underset{\sim}{x}) = 0$ indicates that all the x's are in the overlap region and the data is completely uninformative, while $A(\underset{\sim}{x}) = \infty$ indicates that they are not all on the overlap interval and so the data is completely informative. For the example above, $A(\underset{\sim}{x}) = 0$ if and only if all the x's are in [0.6,1.0]. (It is now easier to see that A has the same distribution under the two hypotheses.) It would seem that A is the ultimate 'precision index' for the test, even more so than R.

(The statistic A is similar to that which we used as a precision index in the Binomial case. In that case, such a statistic existed for only a limited range of hypothesis pairs, in contrast to the present case. In both cases, the *absolute value of the natural logarithm of the likelihood ratio statistic* is an ancillary statistic on Θ_B.)

Conditioning on $A = |\ln LR(\underset{\sim}{X})|$.

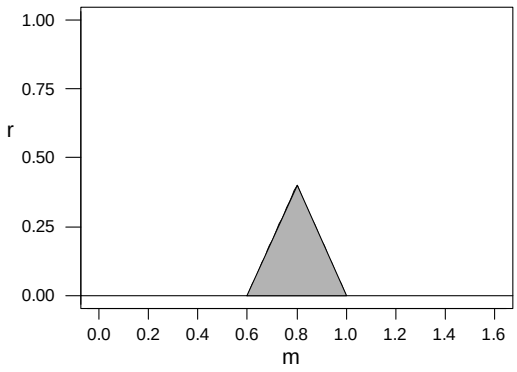
What happens when we condition on the observed value of A? Again we will use the sufficient statistic (M, R) in preference to the data vector $\underset{\sim}{X}$. Having observed $A = a$ we should base our tests on the conditional distributions (under H and K) of (M, R) *given* $A = a$.

[7] We are not considering the further possibilities that the data shows that (i) neither hypothesis can be true, or (ii) the underlying model assumptions are incorrect. Were the first situation to arise, we could discontinue the test of those particular hypotheses. The assessment of an inference procedure is usually made on the assumption that the specified model is correct.

(i) When $A = 0$ …

When we observe that all our observations are lying in the overlap region, we have observed $A = 0$. Conditional upon this being so, (M, R) must lie in the area we called 'B'.

Figure 6.7



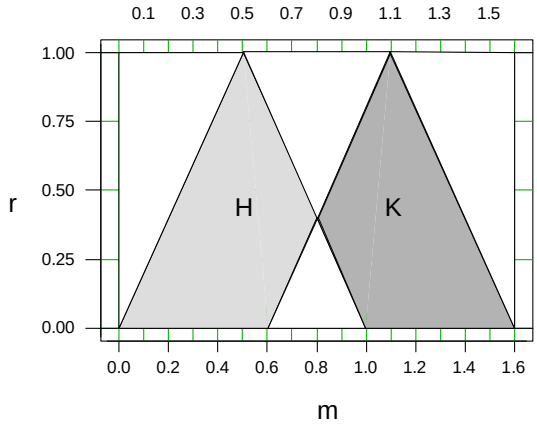
The support of (M, R) (under H *or* K), given that $A = 0$, is shown above. The density is the same under both hypotheses and varies with n; when $n = 2$ the density is constant, $f_H(m,r) = f_K(m,r) = 12.5$ over this area. Since (M, R) has exactly the same conditional distribution under H and K, it is not informative regarding these two hypotheses. Although this is obvious from a common sense point of view, neither of the previous approaches that we examined revealed this fact. If we insist upon defining a rejection rule, there are only two options open to us: either we reject H when we observe any (m,r) in this region, in which case the conditional significance level, $\alpha_{A=0}$ =100% as is the conditional power ($\beta_{A=0} = 0$), or we accept H for all (m,r) in this region, in which case $\alpha_{A=0} = 0$ as is the conditional power (and $\beta_{A=0} = 100\%$). In other words, since there is no basis whatsoever for distinguishing between different parts of this area we must either accept or reject H throughout the area and one of the error probabilities will be 100% while the other is *zero*. We might choose to let $\alpha_{A=0}$ be *zero* in the interest of keeping the test biased in favour of H rather than in favour of K; this means that we will always accept H when

(m,r) is in this region, but the fact that the conditional power is *zero* means that we cannot read anything into our failure to reject H. This brings us, rather laboriously, back to the fact that was obvious, namely that we can infer nothing useful from this type of result vis-à-vis H and K. Note however that the sheer irrationality of using any value of α other than *zero* (or possibly 100%) is obvious once we have conditioned on $A=0$; this was not the case when we did not conditional nor when we conditioned on R. The same is true when we observe $A=\infty$.

(ii) When $A=\infty$ …

The supports of (M,R) under H and under K conditional upon $A=\infty$ are shown below.

Figure 6.8



The white areas of the diagram are not part of either support – the supports do not overlap. Since this is the case we can tell with absolute certainty which is the true hypothesis by observing whether (m,r) lies in the support under H or the support under K. Thus we reject H if and only if (m,r) lies in the support under K and hence not in the support under H; for this test the conditional error probabilities are $\alpha_{A=\infty}=\beta_{A=\infty}=0$, consistent with the fact that, in these circumstances, the result of the test cannot be wrong. Again it is instantly obvious that the only appropriate significance level is *zero*.

If we want to define an accept/reject rule for data based on a fixed sample size (rather than optional stopping), the obvious course is to set both the conditional significance levels, α_a, equal to *zero*. The resulting test is the *zero-level* test discussed in §6.4 and the analogous confidence intervals are the 100% intervals.

Which power is most relevant?

For this test, all the significance levels – conditional (whether upon R or A) or unconditional – are *zero*. However the various versions of power are not all the same; the power conditional upon the observed value of A is the one that gives us the best idea about the evidence contained in the data.

Our rejection region is that labelled 'K' and filled with the darker colour in **Figure 6.8**. The unconditional approach described the power of this test as $84\% = P_K[(M,R) \in \text{ rejection region}]$. The power conditional upon R depends on the observed value of r, for instance if $r = 0.25$ then the conditional power equals $P_K[(M,R) \in \text{ rejection region } | R = 0.25] = 80\%$, whereas if $r = 0.5$ then the conditional power is 100%. The power conditional upon A depends on the observed value of a, if $a = 0$ and the data is uninformative, then the conditional power is *zero*, whereas when $a = \infty$ and the data is completely informative, the conditional power is 100%. Let us consider a particular observation.

The observation $(m,r) = (0.8, 0.25)$ is in the overlap area, which means that its occurrence does not help us to choose between H and K. It is not in the rejection region, so we will not reject H; how much can we read into this fact regarding support for the hypothesis H? The unconditional power and the power conditional upon $r = 0.25$ are (respectively) 84% and 80%. We remarked earlier that this gives the counter-intuitive impression that the data favours H somewhat (since with high power we would expect to reject H if it was false). Let us consider this question from the 'conditional upon a' point of view. We have observed the event $A = 0$ but this was equally likely to occur under either hypothesis so we can infer nothing from it; the

power conditional upon $A = 0$ is *zero*, which means that we should infer nothing from our failure to reject H. The alternative values of 84% and 80% are utterly misleading. In fact the unconditional power and the power conditional upon r are each the average of two values, one of which is completely irrelevant. Consider the value 80% which is the power conditional upon r being equal to 0.25. When $r = 0.25$, we can distinguish between two cases with respect to A: either (m, r) is in the overlap area or it is not. The power conditional upon *both* (i) $R = 0.25$ *and* (ii) data in the overlap area (i.e. $A = 0$) is:

$$P_K[(M, R) \in \text{ rejection region } | (R = 0.25 \,\&\, A = 0)] = 0 .$$

By contrast the power conditional upon both (i) $R = 0.25$ and (ii) data not in the overlap area ($A = \infty$) is:

$$P_K[(M, R) \in \text{ rejection region } | (R = 0.25 \,\&\, A = \infty)] = 100\% .$$

Now, when $r = 0.25$, the probabilities of A being *zero* or *infinity* are 0.2 and 0.8 respectively[8], and the power conditional upon R being 0.25 can be seen to be the average of the two conditional[9] powers weighted according to their probabilities, thus

$$\begin{aligned}
&P_K[(M, R) \in \text{ rejection region } | R = 0.25] \\
&= P_K[(M, R) \in \text{ rejection region } | (R = 0.25 \,\&\, A = 0)] \times P_K(A = 0 \,|\, R = 0.25) + \\
&P_K[(M, R) \in \text{ rejection region } | (R = 0.25 \,\&\, A = \infty)] \times P_K(A = \infty \,|\, R = 0.25) \\
&= (0\% \times 0.2) + (100\% \times 0.8) \\
&= 80\%.
\end{aligned}$$

If we condition on R but not on A, the power will be the average over both the possible values of A, yet in any given case A must be either *zero* or infinity. If it is *zero*, why should a failure-rate for the '$A = \infty$' case play any part in our assessment, and if it is infinity, why should we be influenced by a value relevant only when $A = 0$?

[8] A and R are not independent.

[9] That is, conditional upon A in addition to R.

Should we condition on both R and A? We have just shown that conditioning on R alone is inadequate – when we condition on A, as well, it makes a striking difference and brings our inference into line with common sense. Can we condition on A alone or do we also need R? In the numerical example above, the power conditional on A alone was either *zero* or 100%. This was also true of the power conditional upon A and $R = 0.25$; is R always redundant once we have conditioned on A? The answer is *yes* and this can be shown in general as follows.

Since the significance level is *zero* (overall and also for any ancillary subset based on R and/or A), we need only consider how the power of the test is interpreted. The power conditional on both R and A is given by

$$\text{power}(r,a) = P_K[(M,R) \in \text{ rejection region } | (R = r \,\&\, A = a)]$$

Now,

$$\begin{aligned} \text{power}(r,\infty) &= P_K[(M,R) \in \text{ rejection region } | (R = r \,\&\, A = \infty)] \\ &= 100\% \\ &= \text{power}(a = \infty), \end{aligned}$$

since, when $A = \infty$ and K is true, (M,R) must be in the rejection region no matter what the value of r.

The event $A = 0$ can only occur if $\Delta < 1$ and $r < 1 - \Delta$, in which case

$$\begin{aligned} \text{power}(r,0) &= P_K[(M,R) \in \text{ rejection region } | (R = r \,\&\, A = 0)] \\ &= 0 \\ &= \text{power}(a = 0), \end{aligned}$$

since $(M,R) \in$ rejection region implies that $A = \infty$.

Thus, for all a and r the power conditional upon $A = a$ and $R = r$ is the same as the power conditional upon $A = a$ alone; once we have conditioned upon A, conditioning upon R is redundant.

We can see that, while the traditional approach to conditioning using the ancillary statistic R and a positive value of α is misleading, conditioning can still play a part in helping us to understand the problem. *Although R is ancillary, conditioning upon it does not help us to understand the important features of this model; it does not reveal the problems inherent in using a positive α and, even when $\alpha = 0$, the power conditional upon r is still misleading.* By contrast, if we condition on $A_{ij} = |\ln LR(\underset{\sim}{X};\theta_i,\theta_j)|$, whenever we want to test θ_i against θ_j, the problems and solutions immediately become apparent. In one way, this is a more complex procedure since, unlike R, A_{ij} has to be re-defined for every pair (θ_i,θ_j) (order not important). Nevertheless, this approach illuminates the issue of confidence intervals, as well as tests, since it shows that the 100% confidence interval is the only appropriate interval to use.

Using A_{ij} to understand the Uniform confidence interval.

This does not mean that the value of R has no significance; it is true (as Lehmann said) that, when r is large enough, you can virtually pinpoint the value of θ, but this can be understood in terms of the relationship between R and the statistics A_{ij}. The 100% confidence interval for $\theta \in \mathbb{R}$ is $\mathbb{C}_{100}(m) = (m \pm \frac{(1-r)}{2})$, which has a width of $(1-r)$. Any value of θ outside this interval cannot possibly have produced the observed data; any value within is consistent with the data, but between any two values of θ (say, θ_i and θ_j) both lying within the interval we can make no meaningful judgement on the basis of the data. We cannot narrow the interval to (say) a 90% interval in order to see if one of the values drops out because such an interval would be completely arbitrary and we could equally well justify using a different interval which reverses the preference between the two θ's. This is consistent with the fact that $A_{ij}(m,r) = 0$ for any two values (θ_i & θ_j) both within the 100% interval. The larger r is, the narrower $\mathbb{C}_{100}(m)$ is, and this has implications for the 'number' of hypothesis pairs for which the data is decisive, i.e. able to distinguish the true hypothesis. For example, suppose that $r = \frac{1}{2}$, then $\mathbb{C}_{100}(m)$ is $m \pm \frac{1}{4}$ and

whenever two hypothesised values are more than half a unit apart, (at least) one of them must be completely inconsistent with the data; if r is larger, say $\frac{4}{5}$, then $\mathbb{C}_{100}(m)$ is $m \pm \frac{1}{10}$ and whenever two values are at least one-fifth of a unit apart, we will be able to make a confident judgement between them. A large value of r narrows down the range of possible θ values, but it is going too far to say that we will get more information about θ, relevant to *any* context, when r is larger. For choosing between given θ_i and θ_j, even data with a very small range can be perfectly informative if one or more of the observations lies outside the overlap area; alternatively, if Δ is small and there is a large overlap between the two supports, even data with a fairly large range may fail to give us any information that will help to pick the correct hypothesis. As an ancillary statistic, A_{ij} is superior to R because, in any given context (i.e. for any pair of hypotheses), it perfectly distinguishes between informative and uninformative data.

Once θ_i and θ_j are fixed, we are faced with only three situations:

i. We can dismiss both hypotheses because both values are outside $\mathbb{C}_{100}(m)$.

ii. Both hypotheses are plausible and *equally so* since both values are inside $\mathbb{C}_{100}(m)$.

iii. One hypothesis can be rejected in favour of the other with *zero* error probabilities because one of the values is inside $\mathbb{C}_{100}(m)$ and the other is outside.

We can partition $\mathbb{R}$ into two subsets: $\mathbb{R} \setminus \mathbb{C}_{100}(m)$ and $\mathbb{C}_{100}(m)$; the first contains all the values that cannot possibly be θ, given the data, the second contains all the values (of θ) that could *possibly* have given rise to the data; this interval is narrower when r is larger, but given the uniform model, there is no rational basis for preferring any one of these values over another to even the slightest degree.

6.6 Gambling on the Uniform example with a better ancillary statistic.

In Chapter 5 we used the ancillary statistic, R, to find a betting strategy that would beat the Neyman-Pearson 'optimal' 90% confidence interval for θ in the Uniform case. This strategy was highly effective and could equally have been used in tests of simple hypotheses where it would have worked under both H and K. We have argued that, when we are testing two simple hypotheses, the dichotomous statistic, A, is the appropriate ancillary statistic on which to condition. We can test this assertion by finding out if it is possible to use A to create a betting strategy that will beat the hypothesis test conditional upon R. We look at two cases: first, the hypothesis test using $\alpha_r = 5\%$ (for all r) with the rejection region on the right side of the null support (this approach is now widely regarded as the most appropriate for evaluating evidence – see *§4.1 (The uniform example – Part II)*); second, we will use $\alpha = 0$ (since it seems that this is the only appropriate value) and use the betting scenario to confirm that the power conditional upon a is the most relevant error probability for this test.

As before, we look at the case $n = 2$ and our two hypotheses are H: $\theta = 0.5$ and K: $\theta = 1.1$, hence $\Delta = 0.6 < 1$ and the two supports of (M, R) overlap. A is a function of (M, R) taking the value *zero* for $(m, r) \in \mathrm{B}$ (the overlap area) and the value *infinity* for $(m, r) \in \mathrm{A} \cup \mathrm{C}$. $\mathrm{B} \equiv \varnothing$ when $r > 0.4$; in such a case, A can only take the value infinity, nevertheless there is still a betting strategy that will work whenever $\alpha_r > 0$ (see below); when $\alpha_r = 0$ the error probabilities conditional on A are in agreement with those conditional on R. However, the probability that R is less than 0.4 is 64%; we will look, first, at the two cases $r = \frac{1}{3}$ and $r = \frac{1}{4}$.

Test 1: $r = \frac{1}{3},\ \alpha_r = 5\%$**.**

The conditional distribution of M given $R = \frac{1}{3}$ is $\text{Uni}(\theta \pm \frac{1}{2}(1-\frac{1}{3}))$, hence $\text{Uni}(\frac{1}{6}, \frac{5}{6})$ under H and $\text{Uni}(\frac{23}{30}, \frac{43}{30})$ under K. The conventional 5% critical region, conditional on r, has us reject H whenever $m > \theta_1 + (\frac{1}{2} - \alpha)(1-r) = \frac{1}{2} + (\frac{1}{2} - \frac{1}{20})(1-\frac{1}{3}) = \frac{4}{5}$, then when H is true (and $r = \frac{1}{3}$) we will reject H only 5% of the time. The conditional power of this test, $\kappa_{r=1/3}$, is $P_K(M > \frac{4}{5} \mid R = \frac{1}{3}) = 0.95$, hence $\beta_{r=1/3} = 5\%$.

If we believe that the success rates conditional upon r are the relevant values for interpreting the result of this test, then we should be prepared to bet that the test result is right, risking 0.95 dollars to win 0.05 dollars, or equally that the test result is wrong, risking 0.05 dollars to win 0.95 dollars (these odds work under both H and K since $\alpha_r = \beta_r = 5\%$ for $r = \frac{1}{3}$).

On the basis that the test result is more reliable (in fact, totally reliable) when $A = \infty$ and less reliable when $A = 0$, we adopt the strategy of betting that the result of the test is right whenever we observe $A = \infty$ and betting that the test result is wrong whenever we observe $A = 0$. (It is more enlightening to talk about betting in favour of or against the result of the test since this highlights the fact that some data is more or less informative, but note that this is equivalent to betting about the state of nature since we can deduce from the data what the test result is, i.e. betting that the test result is wrong amounts to betting that H is true when the data is in the rejection region, and amounts to betting that K is true when the data is in the acceptance region.) We randomly generated 1000 samples under both H and K and the cumulative results of this betting strategy are shown in the plots below.

Figure 6.9

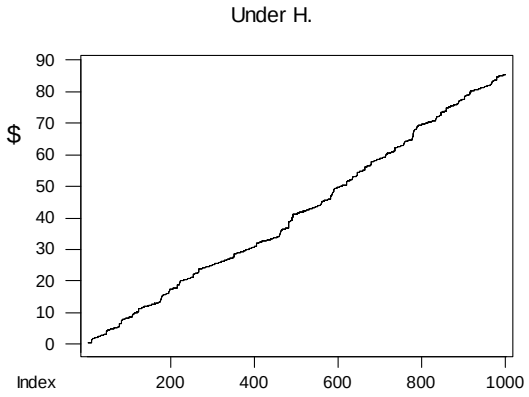
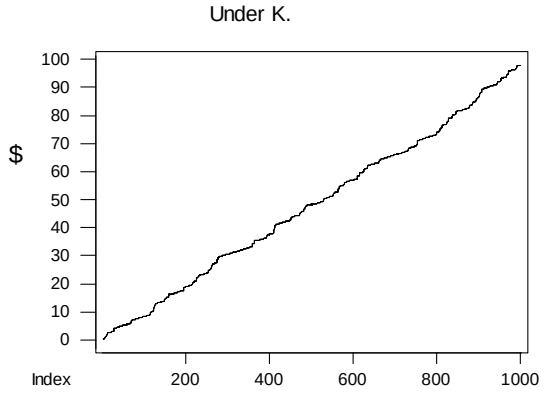


Figure 6.10



The strategy works exceedingly well regardless of which of the hypotheses is true and will also work if H and K appear in any proportions p and $1-p$ $(0 \le p \le 1)$. We can see the details of how the strategy worked by looking at the 1000 samples cross-classified according to the test result (inference) and the bet (value of a).

a) $R = \frac{1}{3}$**, H is true.**

Some of the probabilities associated with this scenario are as follows (r is equal to $\frac{1}{3}$, 'P_r' denotes *probability conditional on* $R = r$):

- $\alpha_r = 5\%$
- $P_r(A = 0) = 10\%$
- $P_r(\text{Result is wrong} \mid A = 0) = 50\%$
- $P_r(\text{Result is wrong} \mid A = \infty) = 0$.

Table 6.3

	Result right (Accept H)	**Result wrong (Reject H)**	**Total**
Bet 'wrong' $(A = 0)$	49 [-0.05]	45 [+0.95]	**94**
Bet 'right' $(A = \infty)$	906 [+0.05]	0 [-0.95]	**906**
Total	**955**	**45**	**1000**

The four probabilities given above are reflected in the respective relative frequencies: 45/1000 = 4.5%, 94/1000 = 9.4%, 45/94 = 47.9% and 0/96 = 0; the dollar profit made on each of the four types of result/bet combination is displayed in square brackets in the table.

b) $R = \frac{1}{3}$**, K is true.**

- $\beta_r = 5\%$
- $P_r(A = 0) = 10\%$
- $P_r(\text{Result is wrong} \mid A = 0) = 50\%$
- $P_r(\text{Result is wrong} \mid A = \infty) = 0$.

Table 6.4

	Result wrong (Accept H)	**Result right (Reject H)**	**Total**
Bet 'wrong' $(A = 0)$	60 [+0.95]	59 [-0.05]	**119**
Bet 'right' $(A = \infty)$	0 [-0.95]	881 [+0.05]	**881**
Total	**60**	**940**	**1000**

The four probabilities are again reflected in the observed relative frequencies: 60/1000 = 6%, 119/1000 = 11.9%, 60/119 = 50.4% and 0/881 = 0.

Test 2: $r = \frac{1}{4},\ \alpha_r = 5\%$.

For the case $r = \frac{1}{4}$, we used the same betting strategy based on the observed value of a. We still used a conditional (on r) significance level of 5% leading to the rule *Reject H when* $m > \frac{67}{80}$; in this case the power of the test is 85% so, under K, if we bet that the result is right we will either lose 0.85 dollars or win 0.15, and if we bet that the result is wrong, we either lose 0.15 dollars or win 0.85. The cumulative profits over 1000 bets are shown below for H and K.

Figure 6.11

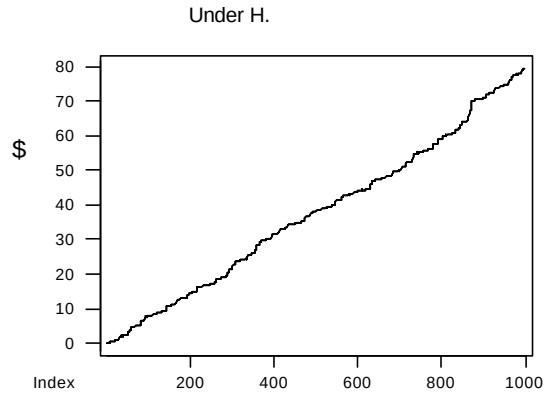
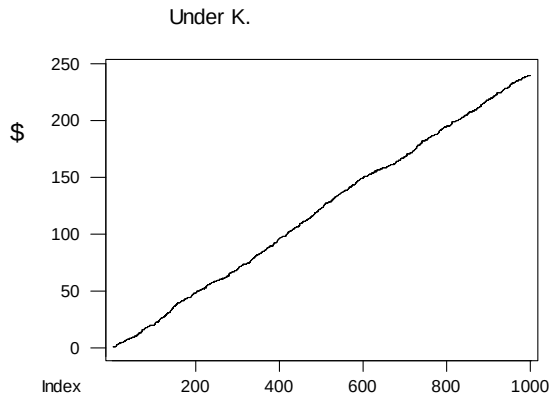


Figure 6.12



Again we can examine the 1000 samples to see that the strategy worked because of the conditional (on a) features of the inference procedure.

c) $R = \frac{1}{4}$**, H is true.**

- $\alpha_r = 5\%$
- $P_r(A = 0) = 20\%$
- $P_r(\text{Result is wrong} \mid A = 0) = 25\%$
- $P_r(\text{Result is wrong} \mid A = \infty) = 0$.

Table 6.5

	Result right (Accept H)	**Result wrong (Reject H)**	**Total**
Bet 'wrong' $(A = 0)$	147 [-0.05]	49 [+0.95]	**196**
Bet 'right' $(A = \infty)$	804 [+0.05]	0 [-0.95]	**804**
Total	**951**	**49**	**1000**

The relative frequencies corresponding to the four probabilities given above are respectively: 49/1000 = 4.9%, 196/1000 = 19.6%, 49/196 = 25% and 0/804 = 0.

d) $R = \frac{1}{4}$**, K is true.**

The probabilities are as follows:

- $\beta_r = 15\%$
- $P_r(A = 0) = 20\%$
- $P_r(\text{Result is wrong} \mid A = 0) = 75\%$
- $P_r(\text{Result is wrong} \mid A = \infty) = 0$.

Table 6.6

	Result wrong (Accept H)	**Result right (Reject H)**	**Total**
Bet 'wrong' $(A = 0)$	150 [+0.85]	48 [-0.15]	**198**
Bet 'right' $(A = \infty)$	0 [-0.85]	802 [+0.15]	**802**
Total	**150**	**850**	**1000**

These probabilities are reflected in the relative frequencies: 150/1000 = 15%, 198/1000 = 19.8%, 150/198 = 75.8% and 0/802 = 0.

This successful betting strategy, based on a, can be used for any value of $r < 0.4$ (or, generally, any value of $r < 1 - \Delta$ where $\Delta = |\theta_1 - \theta_2| < 1$). When $r > 0.4$, $a = \infty$ (always) so we cannot bet based on the value of a. However in this case the power of the test is 100% and since any positive value of α is inefficient, there is a strategy that will win against (say) a 5% test whenever H is true (and therefore overall as long as H is sometimes true); simply use the rejection rule for the significance level *zero*, and bet for or against H based on this rule, since the power of this rule is also 100%, you will break even in the long run when K is true and do better than even when H is true.

If we use $\alpha = 0$, as seems appropriate for this model, then all the conditional significance levels will also be *zero*, but the issue of power remains open. If we reject H whenever $(m, r) \in C$, should we judge the power of the test as being the power conditional on the observed value of r or conditional on the observed value of a? (We have already established that once we have conditioned on a, it makes no difference if we condition on r as well.) If we use the power conditional on r (κ_r) as the value on which to base the odds of the test being 'right' when K is true, will this be vulnerable to a betting strategy based on a? This question is only meaningful when $r < 1 - \Delta$ and $\Delta = |\theta_1 - \theta_2| < 1$, since otherwise κ_r and κ_a are the same.

Test 3: $r = \frac{1}{4},\ \alpha_r = 0$.

Consider the case where we observe data with a range of $\frac{1}{4}$, we have the same hypotheses H and K as before. Conditional upon $R = \frac{1}{4}$, $M \sim \text{Uni}(\frac{1}{8}, \frac{7}{8})$ under H and $M \sim \text{Uni}(\frac{29}{40}, \frac{59}{40})$ under K. We now reject H whenever $m > \frac{7}{8}$ in order to have a significance level of *zero*.

Whenever (in a certain state of nature) there is a probability of η that the test result is wrong and a probability of $1 - \eta$ that the test result is right, then the following payout scheme is 'fair' based on these probabilities whenever this state of nature occurs:

i. The punter bets that the result is right, and it is right and he wins η.

ii. The punter bets that the result is right and it is wrong and he loses $1 - \eta$.

iii. The punter bets that the result is wrong and it is wrong and he wins $1 - \eta$.

iv. The punter bets that the result is wrong and it is right and he loses η.

If $\eta = 0$ it follows that the result is always right and the second and third of these scenarios cannot occur; leaving us with either of two possibilities:

i. The punter bets that the result is right and 'wins' $\eta = 0$, or

ii. The punter bets that the result is wrong and 'loses' $\eta = 0$.

In other words, when the test result is a certainty (certainly right or certainly wrong) there are no non-*zero* payouts appropriate to the long run success rate. For our test with $\alpha = 0$ this is the case when H is true, so the punter will break even (in any number of bets) regardless of strategy.

When K is true, $\eta = \beta$ and $1-\eta = \kappa$ (the power). The power conditional upon r is $\kappa_{r=1/4} = P(M > \frac{7}{8} \mid R = \frac{1}{4}) = 80\%$, whereas the power conditional upon a is $\kappa_a = P(M > \frac{7}{8} \mid A = a)$ where

$$\kappa_a = \begin{cases} 0, & a = 0 \\ 100\%, & a = \infty. \end{cases}$$

As noted earlier, when $R = \frac{1}{4}$, the probability of A being *zero* is 20%, thus we can see that the value $\kappa_{r=1/4}$ is the weighted mean of the power values conditional on the two values of a, i.e. $\kappa_r = (0.2 \times \kappa_0) + (0.8 \times \kappa_\infty)$.

Based on κ_a, our strategy is to bet that the result is right when we observe $a = \infty$ and wrong when $a = 0$.

Under K, the result/bet combinations (cells) have the following joint probabilities; the appropriate payoffs, based on κ_r, are shown in brackets.

Table 6.7

	Result wrong (Accept H)	**Result right (Reject H)**	**Total**
Bet 'wrong' $(A = 0)$	20% [+0.8]	0% [-0.2]	**20%**
Bet 'right' $(A = \infty)$	0% [-0.8]	80% [+0.2]	**80%**
Total	**20%**	**80%**	**100%**

Thus the expected profit from a bet on a single random sample is:

$$(20\% \times 0.8) + 0 + 0 + (80\% \times 0.2) = 0.32 \text{ dollars}.$$

In fact this strategy ensures that the punter will not lose money on any single bet. Since this same strategy will not lose money even when H is true, it is workable as long as $\theta \in \Theta_B$. From this we can see that the relevant power for the test is κ_a rather than κ_r; when $A = 0$ the payouts should be based on κ_0 and $1-\kappa_0$ and when $A = \infty$ the payouts should be based on κ_∞ and $1-\kappa_\infty$, otherwise the odds will be beaten by this strategy.

In this chapter we have found that, in the Uniform (location) case, we can define ancillary statistics on the *binary* parameter spaces (instead of the usual *natural* parameter space, $\mathbb{R}$) that satisfy the requirements of the restricted conditional principle (and even Fisher's definition). Binary parameter spaces may hold the key to extending the scope of conditional inference, including in those common cases where no ancillary statistic exists on the natural parameter space. In the Uniform case there is an ancillary on the *natural parameter space* and yet those produced by the binary parameter spaces deliver much better results. Conditioning on A_{ij} removes and illuminates the problems that arise when we use either unconditional inference or inference based on the traditional ancillary statistic R; we can also argue that A_{ij} is more truly a precision index than is R. This raises the possibility that binary parameter spaces may produce better results generally; we have already noted (see Chapter 3) that using composite hypotheses (equivalent to using a parameter space with more than two elements) means that the question at issue is ill-defined and this may make it harder to identify *precision*. In the Uniform case it is evidently easier to identify a clear-cut precision index when we clarify the context of the test by specifying two simple hypotheses than when we have a large number of hypotheses in mind; it remains to be seen whether this advantage generalises to a greater range of cases.