Chapter 9: Exhaustive ancillary statistics for more general cases.

9.1 When a testing scenario is not log-symmetric.

In the 'two machines' scenario, each machine has a given probability of being picked and this is constant under H and K. This feature is necessary if we are not to lose information in the process of conditioning upon the choice of machine. If a statistic possesses this feature, the argument for conditioning upon its observed value is compelling.[1]

In Chapter 8 we identified a number of cases where the statistic $A = |\ln Y|$ has the same distribution under H and K; this statistic is also attractive on the following grounds. If the likelihood ratio, Y, is a measure of the evidence for H relative to K, then $y = 3$ indicates the same evidence, for H relative to K, as $y = \frac{1}{3}$ does, for K relative to H. Thus we may say that $|\ln y|$ measures the weight of evidence favouring (either) one hypothesis over the other, i.e. it indicates the degree to which the data distinguishes between the two hypotheses. By conditioning upon $|\ln Y|$ we can take into account how informative *our data* is (to the question at issue), rather than using results that average over subsets of the sample space that were not observed and are more informative or less informative than our data. However, this option is only available when $|\ln Y|$ is ancillary, which is not always the case.

In general, if $A = |\ln Y|$, the density function of A (f_A) is related to the density function of Y (f_Y) by

$$f_A(\ln y) = y \cdot f_Y(y) + \tfrac{1}{y} \cdot f_Y(\tfrac{1}{y}).$$

[1] It sometimes seems appropriate to condition on a non-ancillary statistic on the basis that we narrow the sample space down to observations with (in some sense) the same level of reliability as our own. In order to make a case for this, it is necessary to argue that the conditioning process accesses more (extra) information than is lost by the conditioning; this is a much harder case to make.

To show that $|\ln Y|$ is not always ancillary, we need only consider the exponential model:

$$f_X(x;\theta) = \tfrac{1}{\theta} e^{-x/\theta},\ x > 0,\ \theta > 0,$$

with hypotheses H: $\theta = 1$ versus K: $\theta = 2$. The likelihood ratio takes values in the range $y \in (0, 2)$. Now consider the event $A = |\ln Y| = \ln 3$, which is equivalent to $Y \in \{\tfrac{1}{3}, 3\}$. In the exponential case, $f_Y(3) = 0$ since '3' is outside the domain of Y and hence $f_A(\ln 3) = \tfrac{1}{3} f_Y(\tfrac{1}{3})$. This value is not the same under the two hypotheses, since $f_Y(\tfrac{1}{3})$ has *one-third* of the value, under H, that it has under K, by definition of Y as the likelihood ratio. A is not ancillary and $|\ln Y| = \ln 3$ is a result that favours K over H; we would lose this information if we conditioned on the event.

Can we identify ancillary statistics that can be used as the basis of conditional inference in those cases where $|\ln Y|$ is not ancillary and therefore cannot be used? In particular, can we identify statistics that are *exhaustive* as well as ancillary, i.e. that partition the support of Y into sets containing only *two* elements?

In this chapter, we show that, whenever Y is a continuous variable, we can identify such a statistic.

9.2 The 'difference of distribution functions statistic'.

Theorem identifying an exhaustive ancillary statistic on a BPS.

Preliminaries.

Consider a parameter of interest, θ, defined on a parameter space Θ. For a given value of θ, the distribution of the random variable X is completely specified. We consider competing simple hypotheses of the form H: $\theta = \theta_1$ versus K: $\theta = \theta_2$ where

θ_1 and θ_2 are any distinct members of Θ, defining the binary parameter space (BPS): $\Theta_B = \{\theta_1, \theta_2\}$.

Define the likelihood ratio statistic for H versus K as:

$$Y = \frac{g(X;\theta_1)}{g(X;\theta_2)} = LR(X),$$

where $g(x;\theta)$ is the density of x given θ.

Define f_H and f_K as the density functions of Y under H and K respectively. Note that Y is its own likelihood ratio as well as the likelihood ratio of X, i.e. $Y = \frac{f_H(Y)}{f_K(Y)}$.

Since Y is the MSS of $\theta \in \Theta_B$, we can base our inference on the value of y, rather than x, with no loss of information.

Theorem.

Suppose that the likelihood ratio statistic, Y, is a continuous variable[2] under both H and K, and $(c,d) \subseteq (0,\infty)$ is the shortest interval containing both the support of Y under H and the support of Y under K. It follows that $c < 1 < d$.

Define F_H and F_K as the distribution functions of Y under H and K respectively, i.e. $F_{\square}(y) = \int_c^y f_{\square}(r)dr$. Note that $F_H(c) = F_K(c) = 0$ and $F_H(d) = F_K(d) = 1$.

Define the ***difference of distribution functions statistic*** **(DDF statistic)** as:

$$\boxed{D(Y) = F_K(Y) - F_H(Y).}$$

Then this statistic is an exhaustive ancillary statistic on Θ_B.

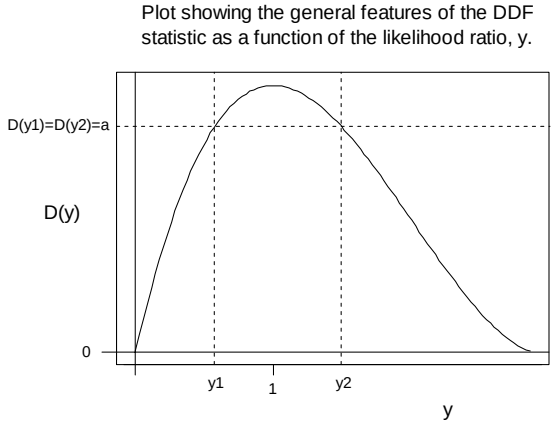
[2] Except that Y may have a positive probability mass, p, at $y = 1$.

Proof.

From the definition of Y as the likelihood ratio statistic, it is easy to show that $D(y)$ is a continuous function of y, taking non-negative values on (c,d), with a maximum turning point at $y=1$ and no other stationary points. Its maximum value is

$$D(1)=\int_c^1 F_K(r)dr<1.$$

Figure 9.1



Hence, for any $a\in(0,D(1))$, the equation $D(y)=a$ has exactly two distinct solutions in y, say y_1 and y_2, where (WLOG) $y_1<1<y_2$.

To show that $D(Y)$ is ancillary, we must show that it has the same distribution under H and K.

The value of the distribution function of $D(Y)$ at a is $P_{\square}(D(Y)<a)$.

$a=D(y_1)=D(y_2)$, hence $D(Y)<a$ if and only if either $Y<y_1$ or $Y>y_2$ (see **Figure 9.1** above), thus:

$$\begin{aligned}P_{\square}(D(Y)<a)&=P_{\square}(Y<y_1)+P_{\square}(Y>y_2)\\&=F_{\square}(y_1)+1-F_{\square}(y_2).\end{aligned}$$

It follows that

$$\begin{aligned}
&P_H(D(Y)<a)-P_K(D(Y)<a)\\
&=\{F_H(y_1)+1-F_H(y_2)\}-\{F_K(y_1)+1-F_K(y_2)\}\\
&=\{F_K(y_2)-F_H(y_2)\}-\{F_K(y_1)-F_H(y_1)\}\\
&=D(y_2)-D(y_1)\\
&=a-a\\
&=0,\quad(\forall a).
\end{aligned}$$

That is, for all a in the domain of $D(Y)$, the distribution function of $D(Y)$ at a is the same under the two hypotheses, hence $D(Y)$ has the same distribution under H and K. $D(Y)$ is a function of the MSS (i.e. Y) and is thus ancillary on Θ_B, in the restricted sense. Since it partitions the sample space of Y into sets containing exactly two values (except for the set {1}), it is an *exhaustive* ancillary statistic.
Q.E.D.

Range of application of the methodology given in this chapter.

The theorem developed above is always applicable when the following conditions are met.

Under both hypotheses, the LR statistic, $y=LR(x;\theta_1,\theta_2)=\frac{f_X(x;\theta_1)}{f_X(x;\theta_2)}$, is a continuous variable, except that it may have a positive probability mass at the point $y=1$.

This, in turn, is satisfied by the following condition, in terms of the natural statistic, X. **Let the natural variable, X, be a continuous random variable on the support, $\varsigma_X(\theta)$ (possibly dependent on θ), and with density $f_X(x;\theta)$. Then, for $i=1,2,$ there should not be any interval, $\mathrm{I}\subseteq\varsigma_X(\theta_i)$, such that $y=LR(x;\theta_1,\theta_2)$ is a constant not equal to *one*, for all $x\in\mathrm{I}$.**

The following type of well-known structure satisfies these requirements. Suppose X has densities that are regular cases of the exponential class of continuous type so that the following conditions are met.

The natural parameter space, $\Theta \equiv (\gamma, \delta)$ is an interval, and

$$\forall\, \theta \in \Theta,\ f_X(x;\theta) = \exp[A(\theta)B(x) + C(\theta) + D(x)],\ a < x < b,$$

where

- neither a nor b depends on θ,
- $A(\theta)$ is a non-trivial continuous function of θ,
- both $B'(x) \not\equiv 0$ and $D(x)$ are continuous functions of x.

In this situation, our condition is equivalent to requiring that (for $\theta_1, \theta_2 \in \Theta$) there be no interval $\mathrm{I} \subseteq (a,b)$ such that $B(x)$ equals some constant, $c \neq \frac{C(\theta_2)-C(\theta_1)}{A(\theta_1)-A(\theta_2)}$, for all $x \in \mathrm{I}$. (A version of this condition can be extended to the regular exponential class where θ is a vector rather than *one*-dimensional.)

Examples of cases where our method is applicable.

Suppose $X \sim N(\mu, 4)$, and thus a member of the regular continuous exponential class with: $A(\mu) = \frac{\mu}{4}$, $B(x) = x$. Clearly, there is no interval, $\mathrm{I} \subseteq \mathbb{R}$, such that $B(x) = x$ is constant for all $x \in \mathrm{I}$, hence we can apply the theorem to all μ_1 and μ_2 for this model.

Let X have an Exponential distribution with a mean of θ, i.e. $f_X(x;\theta) = \theta^{-1} e^{-x/\theta}$. This density is a member of the regular continuous exponential class with $A(\theta) = -\frac{1}{\theta}$ and $B(x) = x$. Again it is clear that we can apply this theorem to all binary parameter spaces associated with the model.

A note on the Fisherian structure of the DDF statistic.

The DDF statistic is ancillary (on Θ_B) in the restricted sense but it also satisfies Fisher's more stringent notion of ancillarity.

A Fisherian ancillary statistic[3], $\mathcal{F}$, has the same distribution for all θ in the given parameter space, Θ, and also satisfies $\mathcal{S} \equiv (\mathcal{F}, \mathcal{M})$, where $\mathcal{S}$ is the MSS of $\theta \in \Theta$, and $\mathcal{M}$ is the maximum likelihood estimator of $\theta \in \Theta$.

To see that $D(Y)$ is a Fisherian ancillary statistic on Θ_B, recall that Y is the MSS. The maximum likelihood estimate of $\theta \in \Theta_B$, based on y, is whichever of θ_1 and θ_2 has the higher likelihood when $Y = y$. Thus the MLE of $\theta \in \Theta_B$ is:

$$\mathcal{M}(y) = \begin{cases} \theta_1, \text{ if } y > 1 \\ \theta_2, \text{ if } y < 1. \end{cases}$$

The MSS, Y, is equivalent to $(D(Y), \mathcal{M}(Y))$ since a specific $D(y)$ is associated with only two values of y (one less and one greater than *one*), and $\mathcal{M}(y)$ identifies whether y is greater than or less than *one*. Hence $D(y)$ and $\mathcal{M}(y)$, together, uniquely define y, and vice versa. Thus $D(Y)$ is a Fisherian ancillary statistic on the binary parameter space.

How good is the DDF statistic?

Comparisons of alternative ancillary statistics and attempts to find a non *ad hoc* method for choosing between them are major themes in conditional inference.

In our context – that of conditioning on statistics that are ancillary with respect to binary parameter spaces – the same issues arise. We have just defined a widely applicable algorithm for finding an ancillary statistic based on the distribution

[3] Fisher (1956), Basu (1964).

functions of the likelihood ratio. How good are such statistics; do they have any optimal properties?

From the proof of the theorem, we can see that such statistics are exactly ancillary on the BPS, i.e. they have exactly the same distribution under the two hypotheses. Also they are ancillary in the restricted sense favoured by Cox and Fisher, since they are all functions of the likelihood ratio statistic, Y, which is the MSS, for any BPS. (And they satisfy all of Fisher's requirements[4], which are more stringent than what we have termed 'restricted ancillarity'.) In addition, they are exhaustive, meaning that they partition the sample space of Y into the smallest[5] subsets that can still be ancillary.

Ancillary statistics that are '*maximal*'[6] are generally regarded as superior to those that are not, where:

An ancillary statistic, A, is *maximal* if the existence of an ancillary statistic, B, such that $A = g(B)$ implies that $B = h(A)$.

This is simply to say that, if there is another ancillary statistic, and A is a function of it, A must be a *one-to-one* function of it so that they are equivalent. This is important because any non-one-to-one function of an ancillary statistic is *less* informative, since it does not partition the sample space of Y so finely. If A is not a one-to-one function of B, it implies that B is superior to A because it separates, into different categories (subsets), all the values of y that are separated by A and more besides. (Note that, in the conventional context, there may be more than one maximal ancillary statistic.)

It is clear, however, that if an ancillary statistic is exhaustive, it must also be maximal. If A is an exhaustive ancillary statistic (EAS) and A is a non-one-to-one function of B, then certain values of b must be associated only with single values of y, other than *one*. This being so, B cannot be ancillary. Being exhaustive entails being maximal but not vice versa.

[4] But Fisher would have wanted these applied to a natural PS, not a BPS.
[5] 'Smallest' in the sense of 'containing the smallest number of elements'.
[6] Cox (1971).

9.3 The DDF statistics for some log-symmetric testing scenarios.

In §9.2, we noted that the Normal location test satisfies the conditions for the existence of an ancillary DDF statistic. We already know that $|\ln Y|$ is an EAS in this case; how does it compare with the DDF statistic?

In the Normal case, the formula for the DDF statistic is as follows.

$$D(y) = \left| \Phi\left(\frac{\ln y}{\delta} + \frac{\delta}{2} \right) - \Phi\left(\frac{\ln y}{\delta} - \frac{\delta}{2} \right) \right|,$$

where $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$, and Φ is the distribution function of the standard Normal variable, Z. Thus $D(y) = P\left(Z \in \left(\frac{\ln y}{\delta} \pm \frac{\delta}{2} \right) \right) = P(Z \in I_y)$. The width of I_y is δ, which is fixed. Thus, the closer to *zero* the centre of the interval is, the larger $D(y)$ is. The centre of the interval is $\frac{\ln y}{\delta}$, and hence $D(y)$ increases (monotonically) as $|\ln y|$ decreases. Thus $D(Y)$ is a one-to-one function of $|\ln Y|$; the two exhaustive ancillary statistics we have identified are *equivalent* – conditioning on them produces the same result.

The DDF statistic is also ancillary in the Cauchy location case. In this case it can also be shown (more laboriously) that $D(y_1) = D(y_2)$ if and only if $y_2 = y_1^{-1}$ and hence $D(Y)$ is a one-to-one function of $|\ln Y|$ and equivalent to it.

9.4 Conditioning on the DDF statistic: definitions and results.

The conditional distribution of the LR given the DDF statistic.

For all values of $a < D(1)$, there are only two values of y that satisfy $D(y) = a$ (see **Figure 9.1**). Thus the conditional distribution of $Y \mid D(Y) = a$ is dichotomous (as was

the conditional distribution of Y given $|\ln Y| = a$) and we need to find the probabilities associated with the two values. We define the conditional probability in the limit, as before; let $D(y) = a$, then

$$\vec{P}(Y = y \mid D(Y) = a) = \lim_{\varepsilon \to 0} P(Y \in (y-\varepsilon, y] \mid D(Y) \in (D(y-\varepsilon), D(y)])$$
$$= \lim_{\varepsilon \to 0} \left\{ \frac{P(Y \in (y-\varepsilon, y])}{P(D(Y) \in (D(y-\varepsilon), D(y)])} \right\}.$$

Recall that $D(y)$ is an increasing function for $y \in (0,1)$ and a decreasing function for $y \in (1,\infty)$. We use the following notation: if $D(y_1) = D(y_2) = a$ $(y_1 < 1 < y_2)$, then $y_1 = D_1^{-1}(a)$ and $y_2 = D_2^{-1}(a)$. That is, $\{y_1, y_2\}$ is the set in the partition of the support of Y that corresponds to the observation $D(Y) = a$ and y_1 is the smaller of the two values, and y_2 the larger; either y_1 or y_2 was the observed value of y.

Let $y_1 < 1$, we want to find $\vec{P}(Y = y_1 | D(Y) = a)$, where $D(y_1) = a$. (The observed value of y was either $y_1 = D_1^{-1}(a)$ or $y_2 = D_2^{-1}(a)$, and hence the observed value of $D(y)$ was $a = D(y_1)$.)

$\vec{P}(Y = y_1 \mid D(Y) = a)$ is equal to:

$$\lim_{\varepsilon \to 0} P[\{y_1 - \varepsilon < Y < y_1\} \mid \{D(y_1 - \varepsilon) < D(Y) < D(y_1)\}]$$
$$= \lim_{\varepsilon \to 0} \left\{ \frac{P(\{y_1 - \varepsilon < Y < y_1\})}{P(\{D(y_1 - \varepsilon) < D(Y) < D(y_1)\})} \right\}$$
$$= \lim_{\varepsilon \to 0} \left\{ \frac{P(\{y_1 - \varepsilon < Y < y_1\})}{P(\{D_1^{-1}(D(y_1 - \varepsilon)) < Y < D_1^{-1}(D(y_1))\}) + P(\{D_2^{-1}(D(y_1)) < Y < D_2^{-1}(D(y_1 - \varepsilon))\})} \right\}$$

Dividing the numerator and denominator by $\varepsilon > 0$ gives:

$$\lim_{\varepsilon\to 0}\left\{\frac{P(\{y_1-\varepsilon<Y<y_1\})/\varepsilon}{P(D_1^{-1}(D(y_1-\varepsilon))<Y<D_1^{-1}(D(y_1))\})/\varepsilon+P(\{D_2^{-1}(D(y_1))<Y<D_2^{-1}(D(y_1-\varepsilon))\})/\varepsilon}\right\}$$

$$=\lim_{\varepsilon\to 0}\left\{\frac{\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon}{[\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon]+[\{F(D_2^{-1}(D(y_1-\varepsilon)))-F(D_2^{-1}(D(y_1)))\}/\varepsilon]}\right\}$$

$$=\frac{\lim_{\varepsilon\to 0}[\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon]}{\lim_{\varepsilon\to 0}[\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon]+\lim_{\varepsilon\to 0}[\{F(D_2^{-1}(D(y_1-\varepsilon)))-F(D_2^{-1}(D(y_1)))\}/\varepsilon]}$$

$$=\frac{f(y_1)}{f(y_1)+\lim_{\varepsilon\to 0}[\{F(D_2^{-1}(D(y_1-\varepsilon)))-F(D_2^{-1}(D(y_1)))\}/\varepsilon]}$$

$$=\frac{f(y_1)}{f(y_1)+\lim_{\varepsilon\to 0}[\{G(y_1-\varepsilon)-G(y_1)\}/\varepsilon]}$$

$$=\frac{f(y_1)}{f(y_1)-G'(y_1)},$$

where F is the distribution function of Y, the density function, f, is its derivative, and $G\equiv F\circ D_2^{-1}\circ D$.

Using the fact that $D_2^{-1}(D(y_1))=y_2$, and $D'(y)=f_K(y)-f_H(y)$, we find that:

$$G'(y_1)=\frac{f(y_2)[f_K(y_1)-f_H(y_1)]}{[f_K(y_2)-f_H(y_2)]}.$$

Hence,

$$\vec{P}[Y=y_1\mid D(Y)=a]=\frac{f(y_1)[f_K(y_2)-f_H(y_2)]}{f(y_1)[f_K(y_2)-f_H(y_2)]-f(y_2)[f_K(y_1)-f_H(y_1)]}$$

and

$$\vec{P}_H[Y = y_1 \mid D(Y) = a] = \frac{f_H(y_1)[f_K(y_2) - f_H(y_2)]}{f_H(y_1)[f_K(y_2) - f_H(y_2)] - f_H(y_2)[f_K(y_1) - f_H(y_1)]}$$

$$= \left[1 - \frac{y_2 f_K(y_2)[f_K(y_1) - y_1 f_K(y_1)]}{y_1 f_K(y_1)[f_K(y_2) - y_2 f_K(y_2)]}\right]^{-1}$$

$$= \left[1 - \frac{y_2[1 - y_1]}{y_1[1 - y_2]}\right]^{-1}$$

$$= \frac{y_1(y_2 - 1)}{(y_2 - y_1)}.$$

Similarly,

$$\vec{P}_K[Y = y_1 \mid D(Y) = a] = \frac{(y_2 - 1)}{(y_2 - y_1)}.$$

Thus the conditional distribution of Y *given* $D(Y) = a$ is:

Table 9.1

y	$y_1 = D_1^{-1}(a) < 1$	$y_2 = D_2^{-1}(a) > 1$
$\vec{P}_H[Y = y \mid D(Y) = D(y_1)]$	$\frac{y_1(y_2 - 1)}{(y_2 - y_1)}$	$\frac{y_2(1 - y_1)}{(y_2 - y_1)}$
$\vec{P}_K[Y = y \mid D(Y) = D(y_1)]$	$\frac{(y_2 - 1)}{(y_2 - y_1)}$	$\frac{(1 - y_1)}{(y_2 - y_1)}$

In the log-symmetric case, $D(y)$ is equivalent to $|\ln y|$ and $y_2 = y_1^{-1}$. Substituting this into the above expressions gives the conditional formulae from Chapter 8, $\vec{P}_H(Y = y_1 \mid A = |\ln y_1|) = \frac{y_1}{(1+y_1)}$, etc.

9.5 The pairing function.

In the above notation, y_1 and y_2 are related by the equation $D(y_1) = D(y_2)$ and are the only values of y where D takes a particular value. We now define a function that describes their relationship directly.

Let $A = \Psi(Y)$ be any exhaustive ancillary statistic (EAS); we define the *pairing function of* A, $\pi_A(y) : \mathbb{R}^+ \to \mathbb{R}^+$, as the unique function with the property:

$$\boxed{\Psi(y) = \Psi(\pi_A(y)),\ \forall y.}$$

Thus the two values, y and $\pi_A(y)$, are associated with the same value of a, or, equivalently, the partition on the support of Y created by A, produces subsets, all of the form $\{y, \pi_A(y)\}$.

The pairing function of any exhaustive ancillary statistic has the following properties:

i. π is a one-to-one function.
ii. π is its own inverse, i.e. $\pi \equiv \pi^{-1}$.
iii. $\pi(1) = 1$.
iv. If A_1 and A_2 are equivalent[7] EA statistics, then $\pi_{A_1} \equiv \pi_{A_2}$.

Thus the equivalence of two (or more) EAS is indicated by their common pairing function.

From the general structure of the function $D(\cdot)$, it follows that the pairing function of any DDF statistic is *monotone decreasing*, in addition to having the above properties.

The conditional probabilities associated with any DDF statistic can be written in terms of the pairing function, as follows.

[7] That is, they are one-to-one functions of each other.

$$\boxed{\begin{aligned}\vec{P}_H[Y = y \mid D(Y) = D(y)] &= \frac{y(\pi_D(y)-1)}{(\pi_D(y)-y)},\\ \vec{P}_K[Y = y \mid D(Y) = D(y)] &= \frac{(\pi_D(y)-1)}{(\pi_D(y)-y)}.\end{aligned}}$$

Only rarely can we solve the equation $D(y) = D(\pi(y))$ analytically, to find the general form of $\pi(\cdot)$. For any explicit D, we can always find the relevant conditional values (cp-value etc.) by numerically solving the equation $D(y_0) = D(\pi(y_0))$ (for $\pi(y_0)$), where y_0 is the particular value obtained from the experiment. This allows us to find the conditional results in any particular case, but our ability to talk in general about (for example) the relationship between the conditional and unconditional p-values is limited. There is one important exception to this, a property common to all conditional tests derived from a DDF statistic, and this will be discussed in §9.7. (For log-symmetric scenarios, $\pi(y) = y^{-1}$ is known, for all y, allowing us to discuss the general nature of the conditional test results, as in the previous chapter.)

One result that can be shown to apply[8], in general, for the pairing function of any DDF statistic, is:

$$\lim_{y\to 1}\{\tfrac{d}{dy}\pi(y)\} = \pi'(1) = -1.$$

This result is useful because it helps to prove a general result about the cp-value, namely,

$$\boxed{\lim_{y\to 1} cp(y) = 50\%}$$

(where $y \to 1$ from *below*).

Note that the pairing function for testing H against K, and the pairing function (denoted π^*) for testing K (as null) against H, are related by: $\pi^*(y) = \{\pi(\tfrac{1}{y})\}^{-1}$. When the test is log-symmetric, $\pi(y) = y^{-1}$ and the two functions are the same; thus the reverse test is also log-symmetric.

[8] Using the facts that $\pi'(y) < 0,\ \forall y$ and $D(y) - D(\pi(y))$ is constant and the formula for $D(y)$.

9.6 The most relevant error probabilities.

We can use the conditional distribution of Y to find the error probabilities of any test criterion, conditional on the observed value of the DDF statistic. Since the DDF statistic is always exhaustive, these have a good claim to being the *most relevant* error probabilities of that test.

If a test is not to have unnecessarily low power, it must have a critical region of the form given by the Neyman-Pearson theorem, i.e. $y = LR(\underset{\sim}{x}) \in (0, y_c]$. When $D(y) = a$, any such test has conditional significance level and power as follows.

Suppose we observe data with a likelihood ratio, y_0, such that $D(y_0) = a$, then (letting $y_1 = D_1^{-1}(a) < 1$ and $\pi(y_1) = D_2^{-1}(a) > 1$), the relevant, conditional significance level (α_a), power (κ_a) and probability of Type II error (β_a), of any test of the form *Reject H when* $y \le y_c$, depend on the value of y_c and are as shown below.

Table 9.2

	α_a	κ_a	β_a
When $y_c < D_1^{-1}(a)$.	0	0	1
When $D_1^{-1}(a) \le y_c < D_2^{-1}(a)$.	$\frac{y_1(\pi(y_1)-1)}{(\pi(y_1)-y_1)}$	$\frac{(\pi(y_1)-1)}{(\pi(y_1)-y_1)}$	$\frac{(1-y_1)}{(\pi(y_1)-y_1)}$
When $y_c \ge D_2^{-1}(a)$.	1	1	0

Note that if the critical likelihood ratio, y_c, is greater than *one*, then $\exists a: \alpha_a = 1$, $\kappa_a = 1$ and $\beta_a = 0$, that is, for some data in the rejection region, the most relevant significance level of the test is 100%.

9.7 A general requirement for sensible inferences.

The conditional distributions of Y, given the DDF statistic, are shown again below (where $y_2 = \pi(y_1)$).

y	y_1	y_2
$\vec{P}_H[Y = y \mid D(Y) = D(y_1)]$	$\frac{y_1(y_2 - 1)}{(y_2 - y_1)}$	$\frac{y_2(1 - y_1)}{(y_2 - y_1)}$
$\vec{P}_K[Y = y \mid D(Y) = D(y_1)]$	$\frac{(y_2 - 1)}{(y_2 - y_1)}$	$\frac{(1 - y_1)}{(y_2 - y_1)}$

($y_1 < 1 < y_2$.)

Because the conditional distribution always has this form, it follows that $\forall y_2,\ cp(y_2) = 100\%$. Thus, *in general*, (as in the log-symmetric case):

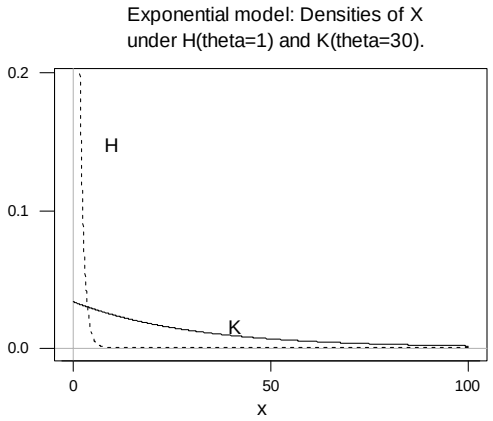
$$\boxed{cp(y) = 100\% \text{ whenever } y > 1.}$$

When we use an exhaustive conditional test based on the DDF ancillary statistic, no data with a likelihood ratio of more than *one* can ever be significant for rejecting H in favour of K. As we noted earlier, this contrasts hugely with conventional methods where, for a high power test, H may be rejected even though $y = 10^{14}$ and, in fact, there is no limit to how large the LR can be while still having a small p-value. However, any rejection region, based on a critical likelihood ratio greater than *one*, will inevitably have a conditional significance level of 100%, for some value of $a = D(y)$. To see this in practice, consider the following example. (The exponential model is not log-symmetric for any hypotheses.)

Example 9.1

Suppose $X \sim Expo(\theta)$ and we want to test H: $\theta = 1$ against K: $\theta = 30$. The distributions associated with these hypotheses are very different so it should be a simple matter to identify data that constitutes strong evidence against H relative to K (the likelihood ratio can be arbitrarily close to *zero*), however the conventional critical region contains many values that do not constitute such evidence.

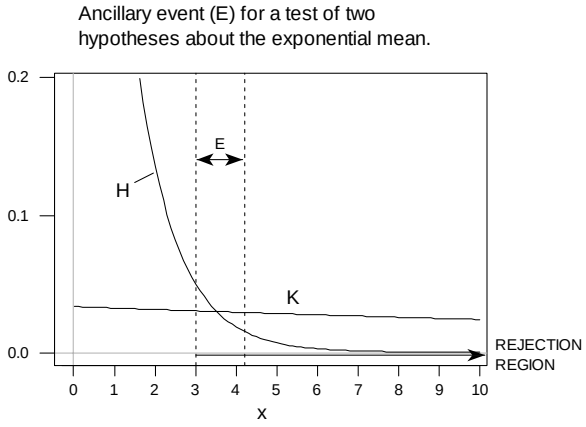
Figure 9.2



The standard 5% test rejects H whenever $x \geq \ln 20 \approx 3$, that is, when $y = LR(x) \leq 1.66$. Since the critical likelihood ratio is greater than *one*, it must be the case that the conditional significance level of this test is 100% for some values of $a = D(y)$. We can show this quite simply without going so far as to condition on the exact value of a.

Consider the event $E \equiv \{3.00 < X < 4.16\}$; this event has the same probability under H and K, as can be seen from the plot below. Thus, when E occurs, we should condition on this fact in order to derive the more relevant error probabilities. (The cases 'E occurs' and 'E does not occur' are analogous to 'machine A is used' and 'machine B is used'.)

Figure 9.3



The event, E, lies entirely within the rejection region, $[3,\infty)$, and, thus, the probability that we will reject H, given E, is 100% (under either hypothesis). The relevant significance level of the test is 100%, as is the relevant power; thus we can read nothing into the fact that we have rejected H.

Since the cp-value of any data with a likelihood ratio greater than *one* is *always*[9] 100%, we can evaluate the significance of such data without going to the trouble of deriving $D(y)$, $\pi(y)$ or any ancillary events. Thus, in ***Example 9.1***, the observation $x = 3.2$ lies in the rejection region (p-value<5%). We could use the fact that ancillary event E has occurred to show that we can read nothing into this, but there is no need; simply by calculating the likelihood ratio of the data:

$$y = LR(3.2) = 30\exp\{\tfrac{-29\times 3.2}{30}\} = 1.36 > 1,$$

we can show that the cp-value is 100% and deduce that the observation does not constitute strong evidence against H.

In all cases where the DDF statistic is ancillary, it follows that:

No data justifies the rejection of H in favour of K if it has a likelihood ratio of more than *one.* This is true even when the data lies in the optimal rejection region defined by a conventionally small value of α.

[9] Assuming only that the DDF statistic is ancillary for which the continuity of Y is a sufficient condition.

To prove this claim, we generalise the argument used in ***Example 9.1***.

Let $\Re$ be a (Neyman-Pearson) *best critical region* in the support of a random variable, X, for rejecting H in favour of K and let x_0 be the observed value of this variable; suppose that x_0 lies in the rejection region and has a likelihood ratio of more than *one*. That is:

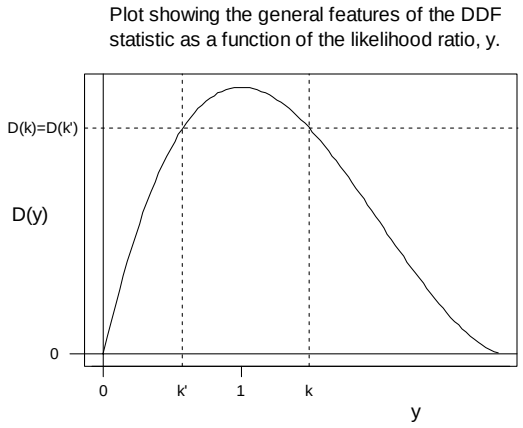
a) $\Re \equiv \{x : y = LR(x) \leq k\}$.

b) $x_0 \in \Re$.

c) $y_0 = LR(x_0) > 1$.

From the above, it follows that $k > 1$ (this is not inconsistent with $\alpha = P_H(X \in \Re)$ being small). Since $k > 1$, we can write it as $k = D_2^{-1}(D(k))$, where $D(Y)$ is the DDF statistic. From the earlier theory on $D(y)$, it follows that there exists a value, $k' = D_1^{-1}(D(k)) < 1$, (see below).

Figure 9.4



Consider the event $E \equiv \{k' \leq Y \leq k\}$. E is an ancillary event[10] since $k' \leq Y \leq k$ is equivalent to $D(Y) \geq D(k)$ (see plot above) and this depends only on the ancillary statistic, $D(Y)$.

[10] An event having the same probability under each hypothesis.

Note the following two results:

i. The event, E, has occurred if $X = x_0$,

ii. E occurs $\Rightarrow \{X \in \Re\}$.

The first claim is true because $(X = x_0) \Rightarrow (Y > 1)$ and, since $x_0 \in \Re$, $(X = x_0) \Rightarrow (Y \leq k)$, thus $(X = x_0) \Rightarrow (1 < Y \leq k) \Rightarrow (k' \leq Y \leq k)$. The second claim is true because $X \in \Re$ if and only if $Y \leq k$ and this is implied by $k' \leq Y \leq k$.

When we observe data, x_0, the event E has occurred and, since it is ancillary, we should calculate the error probabilities conditional upon E. Since $P(X \in \Re \mid E) = 100\%$ under both hypotheses, the conditional significance level and conditional power are 100%. When E occurs, we always reject H, no matter which hypothesis is true, thus we can read nothing into our result. The relevant error probabilities show that we cannot sensibly reject H on the basis of observing x_0.

The fact that we cannot reject H unless (as a minimum requirement) the likelihood ratio of the data is less than *one* brings us closer to the *law of likelihood* because it is consistent with the view that only a likelihood ratio less than *one* represents any evidence in favour of K (relative to H). In this, our results differ markedly, not only from those of unconditional tests, but also from those derived from the type of conditional tests represented by Cox's example. Conditioning on the non-exhaustive ancillary statistic in Cox's example can produce a small conditional p-value from data with a likelihood ratio of *any size*, leading to the rejection of H.

An aside on the definition of p-value.

Conditional upon the observed value of the DDF statistic, Y has a discrete distribution. Throughout this work, we use the Fisherian definition of p-value. Suppose that X is any random variable and that Y is a one-to-one function of X, then the Fisherian p-value of x_0 for a left-sided test is $P_H(X \leq x_0)$ and for a right-

sided test is $P_H(X \geq x_0)$. An alternative definition that is sometimes used in discrete cases (where, for instance, X takes integer values) gives the left-sided p-value as $P_H(X \leq x_0 - 1) + \frac{1}{2} P_H(X = x_0)$. The advantage of this version is that the left-sided and right-sided p-values of x_0 (i.e. $P_H(X \leq x_0 - 1) + \frac{1}{2} P_H(X = x_0)$ and $\frac{1}{2} P_H(X = x_0) + P_H(X \geq x_0 + 1)$) sum to *one* instead of summing to $1 + P_H(X = x_0)$. This is seen as a desirable feature for two-sided p-value functions[11] and is a feature of the Fisherian p-value when X is *continuous*. We retain the Fisherian definition for two main reasons. The first is that, in the context of binary parameter spaces, all tests are one-sided (in terms of Y) and further, we have argued (see Chapter 3) that two-sided tests are nonsensical. The second is that we wish to retain the usual relationship between a significance level, α, and the p-value of x; namely that the two statements 'x *is in the* α*-level rejection region*' and '*p-value*$(x) \leq \alpha$' are equivalent[12]. Note however that using the modified definition would not alter the results of our conditional inference in any practical way. All values of $y < 1$ would still have the same cp-value, while values of $y > 1$ would have varying cp-values, all greater than 50%, rather than the cp-value of 100% that we have derived. Since no p-value of more than 50% will be regarded as significant, the interpretation of the data would not be changed.

Swapping hypotheses.

Exhaustive conditional inference, based on the DDF statistic, produces consistent results for swapped hypotheses. The conditional error probabilities, α and β, for testing H versus K are equal to the conditional error probabilities, respectively, β^* and α^*, for testing K versus H.

E. C. inference has a feature that we have already noticed in conditional tests that use $|\ln Y|$ as the ancillary statistic. A minimum requirement for rejecting H in favour of

[11] I.e. the two-sided p-value of the fixed value x_0, as a function of θ.

[12] For the discrete case we must assume that $\exists x' : \text{p-value}(x') = \alpha$.

K is that the likelihood ratio[13] is less than *one* and a minimum requirement for rejecting K in favour of H must be that the (same) likelihood ratio is greater than *one*. Thus we can never find the situation that often arises in conventional inference, where the same data would lead us to reject H in favour of K and also to reject K in favour of H, were the hypotheses reversed. Conditioning on the DDF statistic, and using any bound ($0 < \bar{\alpha} < 1$) on the conditional significance level of both tests, partitions the support of Y into three intervals. The first contains 'small' values of y that cause us to reject H as the null hypothesis in favour of K, the third contains 'large' values of y that cause us to reject K as null hypothesis in favour of H, and the second contains 'medium sized' values that do not lead us to reject either hypothesis in favour of the other. In fact, if we let the observed value of $y = \frac{f_{X,H}(x)}{f_{X,K}(x)}$ be y_0 and $D(y_0) = a_0$, then, it can easily be shown that the second interval is:

$$\left(\frac{\bar{\alpha} D_2^{-1}(a_0)}{[D_2^{-1}(a_0) - (1-\bar{\alpha})]}, \frac{[1-(1-\bar{\alpha})D_1^{-1}(a_0)]}{\bar{\alpha}} \right),$$

and that this interval contains the value *one*.

9.8 Wald's sequential probability-ratio stopping rule and exhaustive conditional inference.

In testing two simple hypotheses, any method consistent with the LP will make the same inference from given data arising from experiments with different stopping rules, as long as the stopping rules produce the same likelihood ratio function (as they often do). This is counter to what happens in conventional frequentist inference where the interpretation of any given data is very sensitive to the stopping rule that produced it. Thus 14 *heads* out of 20 coin tosses will be interpreted differently depending on whether the experiment was designed to terminate after 20 tosses or terminate after 14 heads. Because E. C. inference is not consistent with the LP, it will also be sensitive to the stopping rule but it is not as sensitive as conventional

[13] Defined as f_H / f_K.

inference. We have already discussed the case of symmetric stopping rules for Bernoulli trials on certain binary parameter spaces (see §8.10). This showed that E. C. inference could produce a common interpretation of data from experiments based on different stopping rules in cases where the unconditional inferences would have been different. (In §9.9 we will see that the converse cannot happen.)

In this section we show that, if we use any sampling regime and it gives rise to data with a likelihood ratio reasonably far from *one*, then the same data produced by a particular version of Wald's sequential probability ratio (SPR) sampling regime is associated with almost identical E. C. inference results. This is striking because data is usually interpreted quite differently if it comes from a 'fixed sample size' regime (for example), rather than from an SPR regime.

Wald's sequential probability-ratio test uses a specific stopping rule designed to create a sample space containing only observations that clearly favour one hypothesis over the other (the kind of data we might call 'strong'). The following exposition is from Kendall and Stuart[14] re-worded to match our terminology and notations.

Suppose we take m values in succession from a population $f(x;\theta)$. At any stage the ratio of the probabilities of the sample on hypotheses H ($\theta = \theta_1$) and K ($\theta = \theta_2$) is

$$y_m = \frac{\prod_{i=1}^{m} f(x_i;\theta_1)}{\prod_{i=1}^{m} f(x_i;\theta_2)}.$$

We select two numbers L and U, related to the desired type I and type II error probabilities (α and β), and set up a sequential test as follows: so long as $L < y_m < U$ we continue sampling; at the first occasion when $y_m \le L$ we accept K; [or] at the first occasion when $y_m \ge U$ we accept H. (This experiment terminates with a probability of *one*.)

[14] Kendall & Stuart, pp. 599-602.

The values of L and U necessary to produce error probabilities of approximately α and β are $L = \frac{\alpha}{1-\beta}$ and $U = \frac{1-\alpha}{\beta}$. The approximation is due to the end-effects (i.e. the fact that the final value of y_m will overshoot the bound to some degree) but it can be shown[15] that the approximation is very accurate when L and U are derived from conventionally small values of α and β, i.e. when L and U are reasonably far away from *one* ($L < 1 < U$) and we will assume this to be the case. Then 'accepting K' amounts to observing a genuinely small likelihood ratio constituting strong evidence against H relative to K and 'accepting H' amounts to observing a genuinely large likelihood ratio, constituting strong evidence against K relative to H. This sampling regime ensures that the sample space contains no values of y between L and U and, for practical purposes, no values of y that are much less than L or much more that U (since there is not much over-shooting). Hence the sample space contains only values that are clustered close to L (on the lower side) and close to U (on the upper side). Assuming that L and U were chosen appropriately, the unconditional type I and II error probabilities can be derived to a high degree of accuracy as $\alpha = P_H(\text{'Accept K'}) = \frac{L(U-1)}{(U-L)}$ and $\beta = P_K(\text{'Accept H'}) = \frac{(1-L)}{(U-L)}$. Because the unconditional sample space for this experiment is already virtually reduced to two values of y – L and U – the DDF statistic[16] is practically degenerate and conditioning on it will not significantly change the error probabilities, thus, α and β (above) can also be regarded as the exhaustive conditional error probabilities for this test.

---

[15] Kendall & Stuart, p. 601.

[16] We have defined the DDF statistic in terms of the distribution functions and these are conventionally continuous from the left; we have also stated that Y must be continuous (except possibly at $y = 1$) in order that the DDF statistic be ancillary. However, it is possible to widen the definition of DDF statistic so that we can find an exhaustive ancillary statistic in certain cases where Y is discrete but there is a high level of symmetry in its distributions. The modification takes the following form:

$$D^*(y) = \begin{cases} F_K(y) - F_H(y), & y \le 1 \\ P_K(Y < y) - P_H(Y < y), & y > 1. \end{cases}$$

Thus the distribution functions are defined to be continuous from the right when $y > 1$. When Y is continuous there is no distinction between $D(y)$ and $D^*(y)$ and we have not thought it worthwhile to add this extra complication in order to cover a small number of unusual cases. $D^*(y)$ is equivalent to $|\ln Y|$ in the Welch and Double-exponential cases, where Y is discrete or partly so (see Chapters 6 & 8), and also applies to Wald's model if we think of Y as being effectively discrete on $\{L, U\}$.

Now suppose that we perform an experiment sampling x-values from the same population as above and that we are interested in the same hypotheses about θ (H and K), but we do not use the same sampling regime. Nevertheless, at the end of the day, this experiment produces the same data, and hence the same likelihood ratio, as the SPR experiment. We call the observed likelihood ratio y_0, and $\pi(y_0)$ is its pair – the other root of the equation $D(y) = D(y_0)$, where $D(Y)$ is the DDF statistic for this scenario (which includes the sampling regime). As usual, let $y_1 = \min\{y_0, \pi(y_0)\} < 1$ and $y_2 = \max\{y_0, \pi(y_0)\} > 1$. Then the rule *Reject H in favour of K if* $y = y_1$ produces a test with a conditional type I error probability of $\alpha = \frac{y_1(y_2-1)}{(y_2-y_1)}$ and a conditional type II error probability of $\beta = \frac{(1-y_1)}{(y_2-y_1)}$ (see §9.4).

Since the same data was observed in both experiments, it follows that either $y_0 \approx L$ or $y_0 \approx U$; if, in addition, the SPR stopping rule was defined so that L and U are connected by the relation $D(L) = D(U)$, then $L \approx y_1$, $U \approx y_2$ and the conditional error probabilities are the same for both sampling regimes.

Since our conditional error probabilities have the same general structure as Wald's SPR error probabilities, it follows that, no matter what sampling regime (stopping rule) we use, if the data is reasonably informative, an SPR sampling regime can be found that can produce the same data and has the same conditional error probabilities. In such a case, exhaustive conditional inference overrides the differences between the SPR sampling regime and any other regime to produce the same results.

9.9 Inference Classes.

We end this Chapter by introducing a concept that is useful for comparing conventional frequentist inference with exhaustive conditional inference based on the ancillary DDF statistic, where 'inference' refers to the very specific issue of identifying data that constitutes strong evidence against one simple hypothesis relative to another. In conventional inference the p-value of the data is the basis for

distinguishing between data that constitutes strong[17] evidence against one hypothesis relative to the other and data that does not. In exhaustive conditional inference the cp-value plays the same role.

Let $\mathcal{M}$, be a model connecting a natural statistic, X, with a parameter of interest, θ, via some probability density $f_{\mathcal{M}}(x;\theta)$. Then we define, a *scenario*, $\mathcal{S} \equiv (\mathcal{M}, H, K)$, as a combination of the model with two distinct, ordered hypotheses specifying the value of θ.

In conventional inference, we reject H in favour of K when the p-value is small, and, in exhaustive conditional inference, when the cp-value is small. For any particular scenario, $\mathcal{S}$, both the p-value and the cp-value can be written as functions of the likelihood ratio statistic, $y = f_{H,X}(x) / f_{K,X}(x)$, which is the MSS of θ; we have called these functions $p(\cdot)$ and $cp(\cdot)$. In general these functions vary between scenarios; thus, when $\mathcal{S}_A$ and $\mathcal{S}_B$ are different scenarios, $p_A(\cdot)$ and $p_B(\cdot)$ may be different functions, as may $cp_A(\cdot)$ and $cp_B(\cdot)$.

We define an *inference-class* as any class of scenarios all associated with the same p-value function, $p(\cdot)$. In other words, all the scenarios in a particular inference-class give rise to the same value of $p(y)$, for all y. An *E.C. inference-class* does the same for exhaustive conditional inference, that is, all the scenarios in a particular E. C. inference-class give rise to the same value of $cp(y)$, for all y. Thus two scenarios in the same inference-class (or E. C. inference-class) result in the same evidential interpretation of any particular likelihood ratio.

According to the likelihood principle (LP), all scenarios regarding the same parameter of interest (θ) should be in the same inference class. Thus, for any given parameter of interest there should be only one, universal inference class. (Hacking's law of likelihood (LL) can be interpreted as implying that, across all parameters of interest, there should be only one single inference class. This would seem to follow from the claim that the likelihood ratio, y, is the (sole) measure of the evidence in x for H

[17] The level of strength is specified externally.

relative to K.) Thus conventional inference, based on the p-value, and E. C. inference, based on the cp-value, both contravene the LP. However, we would argue that E. C. inference is, in a sense, *closer* to likelihood inference because of the following fact.

Any two scenarios that lie in the same inference-class also lie in the same E. C. inference class, but the converse does not hold.

Changing from conventional inference to E. C. inference increases the size of some inference classes and decreases the size of none. In this sense E. C. inference brings us closer to the ideal of a single inference class. We prove the above claim as follows.

Proof.

Let $\mathcal{S}_A$ and $\mathcal{S}_B$ be two scenarios in the same (conventional) inference-class and let Y_A and Y_B be the likelihood ratio statistics associated with the two scenarios. Let the two hypotheses associated with the scenarios be $\theta = \theta_1$ (H) and $\theta = \theta_2$ (K) for scenario $\mathcal{S}_A$, and $\eta = \eta_1$ (H*) and $\eta = \eta_2$ (K*) for scenario $\mathcal{S}_B$.[18]

Y_A and Y_B must have the same distributions under (respectively) $\theta = \theta_1$ and $\eta = \eta_1$, since they have the same $p(y)$ for all y and $p(y)$ is the cumulative distribution function of the likelihood ratio statistic under the null hypothesis. Since the distribution functions are identical (i.e. $F^A_{\theta_1}(y) = F^B_{\eta_1}(y),\ \forall y$), the densities are also identical, i.e. $f^A_{\theta_1}(y) = f^B_{\eta_1}(y), \forall y$.

We also know that:

$$f^A_{\theta_1}(y) = y \cdot f^A_{\theta_2}(y),\ \forall y \text{ and}$$
$$f^B_{\eta_1}(y) = y \cdot f^B_{\eta_2}(y),\ \forall y.$$

[18] In the interest of generality, we do not assume that the two scenarios necessarily involve the same parameter, since it seems unnecessary to do so.

Hence it follows that $f^A_{\theta_2}(y) = f^B_{\eta_2}(y)\ (\forall y)$, and thus $F^A_{\theta_2}(y) = F^B_{\eta_2}(y)\ (\forall y)$, and

$$\begin{aligned} D_A(y) &= F^A_{\theta_2}(y) - F^A_{\theta_1}(y) \\ &= F^B_{\eta_2}(y) - F^B_{\eta_1}(y) \\ &= D_B(y)\ (\forall y). \end{aligned}$$

Hence $\pi_A(y) = \pi_B(y)$ and $cp_A(y) = cp_B(y),\ \forall y,$ that is, scenarios A and B are in the same E. C. inference-class.

We know from many examples that the converse does not hold. For example, all scenarios associated with the Normal location model are in the same E. C. inference-class (the log-symmetric class), but only those with the same value of $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$ are in the same conventional inference-class, i.e. $p_{\delta_1}(y) = p_{\delta_2}(y)$, for all y, if and only if $\delta_1 = \delta_2$.

Combining classes under E. C. inference.

When do two models, involving a particular parameter, produce the same E. C. inference even though the unconditional inferences are different, and why?

In terms of the two-stage experimental structure, $D(Y)$ describes the outcome of stage-one; this tells us nothing about the question at issue, but sets up the conditions under which stage-two of the experiment is performed.

Let ς_1 be the support of the likelihood ratio statistic (Y_1) under model 1 and ς_2 be the support of the LR statistic (Y_2) under model 2. The E. C. inference is the same for both models, if, and only if, $\pi_1(y) = \pi_2(y)$ on the intersection of the two supports[19] (i.e. $y \in \varsigma_1 \cap \varsigma_2$). This is the case only if, for any $y \in \varsigma_1 \cap \varsigma_2$, the conditional distribution of Y_1 given $D_1(Y_1) = D_1(y)$ is the same as the conditional distribution of

[19] And the two models are in the same *E. C. inference class* if, in addition, $\varsigma_1 \equiv \varsigma_2$.

Y_2 given $D_2(Y_2) = D_2(y)$, and, hence, the DDF statistics D_1 and D_2 are one-to-one functions of each other *on this set*.

When are the *unconditional* distributions of Y_1 and Y_2 different, despite this? The unconditional density of Y_j at the point y is the product of the density of the variable $D_j(Y_j)$ at the point $D_j(y)$ and the conditional density (at y) of Y_j *given that* $D_j(Y_j) = D_j(y)$, i.e. it is the product of the densities of the stage-one variable and the (conditional) stage-two variable. Since the second (conditional) density is the same in both cases, the products can only differ if the distribution of the variables D_1 and D_2 are not the same at the points $D_1(y)$ and $D_2(y)$, respectively.

We can simplify the discussion if we replace $D_2(y)$ by an ancillary statistic ($\tilde{D}_2(y)$), having the feature that $D_1(y) = \tilde{D}_2(y)\ \forall y \in \varsigma_1 \cap \varsigma_2$, as follows:

$$\text{let } \tilde{D}_2(y) = \begin{cases} D_1(y), & y \in \varsigma_1 \cap \varsigma_2 \\ D_2(y)+1, & y \in \varsigma_1' \cap \varsigma_2. \end{cases}$$

This transformation forces D_1 and $\tilde{D}_2$ to *take the same values* on their common domain while ensuring that $\tilde{D}_2$ is a one-to-one function of D_2 that can be used to produce the same conditional inference[20] for all $y \in \varsigma_2$. The unconditional distributions of Y_1 and Y_2 at y (common to both models) are different if and only if the ancillary statistics, $D_1(Y_1)$ and $\tilde{D}_2(Y_2)$, are differently distributed on their common support. This difference in distributions can amount *only* to a difference between the probabilities (densities) assigned to each value since the values themselves are the same. The stage-one distribution, i.e. the distribution of $D_1(Y_1)$ (or $\tilde{D}_2(Y_2)$), is uninformative about the question at issue since it is the same under H as under K; nevertheless, any difference between the distributions of $D_1(Y_1)$ and of $\tilde{D}_2(Y_2)$ is enough to ensure that the conventional, unconditional inferences are not the same. If we observe $D_1(Y_1) = a$ (say), the conditions for stage-two of the experiment

[20] Since $0 \le D(y) < 1,\ \forall y$. This is true for any DDF statistic.

are the same as when we observe $\tilde{D}_2(Y_2) = a$. Suppose, now, that in both cases these identical conditions give rise to the *same data*. The E. C. inferences will be the same, but, if $P(D_1(Y_1) = a) \neq P(\tilde{D}_2(Y_2) = a)$, then the conventional inferences will differ.

If both models are structured as umbrella experiments and have a sub-experiment[21] in common, then any given result from the sub-experiment will produce the same E. C. inference regardless of which umbrella model was used. However, if the sub-experiment does not have the same probability (density) under both (umbrella) scenarios then, even though we perform that particular sub-experiment, *and get the same result* in both cases, the conventional p-values will not be the same, and the two models will be in different conventional inference classes.

Which ancillary statistics are covered by E. C. inference?

Any statistic that is ancillary, in (at least) the weak sense of having the same distribution for all θ in the parameter space, defines notional sub-experiments within the main experiment. Each sub-experiment is associated with a unique value of the ancillary statistic and thus has a fixed probability of occurring no matter what the value of θ. Any outcome from the main experiment constitutes, in essence: (i) a choice of sub-experiment, and (ii) the outcome from that sub-experiment. Each sub-experiment is associated with a particular (sub) sample-space, which is a subset of the sample space of the whole (umbrella) experiment, i.e. the ancillary statistic defines a partition of the original sample space. The unrestricted conditionality principle states that we should make the same inference from any outcome of a given experiment regardless of whether that experiment stands alone or is a sub-experiment with respect to an ancillary statistic.

Sometimes one ancillary statistic (say, A) 'covers' another ancillary statistic (B) in the following sense: conditioning upon A ensures that any outcome is interpreted the same way regardless of whether we locate it within the large sample space or within

[21] Where, as usual, the probability (density) of ending up in that sub-experiment is the same under H and K so that this fact is uninformative.

the appropriate sub-space, with respect to B. Thus any inference that satisfies the conditional principle with respect to A, automatically also satisfies it with respect to B. The most obvious case where this occurs is when A is a more refined version of B, i.e. B is a non-one-to-one function of A and, hence, all the sub-spaces defined by B are unions of one or more of the sub-spaces defined by A. However, these are not the only circumstances in which conditioning on one ancillary statistic can cover the effect of conditioning on another. In the Welch example (see Chapter 6), we noted that conditioning on $A = |\ln Y|$ covers the effect of conditioning on the range, R, even though R is not a function of A.

The DDF statistic satisfies a strong notion of ancillarity with respect to the binary parameter space (i.e. it is a function of the MSS), but its exhaustiveness means that conditioning upon it has far-reaching effects. What other ancillary statistics are covered by the DDF statistic?

E. C. inference covers the effect of conditioning on another ancillary statistic[22], A, if and only if the sub-experiments defined by A *all belong to the same E. C. inference class*. The proof of this is as follows. If all the sub-experiments belong in the same E. C. inference class, then they must give rise to exactly the same set of likelihood ratio values. For the sub-experiment defined by $A = a$, the DDF statistic at the value, y, is defined by $D_a(y) = F_{K,a}(y) - F_{H,a}(y)$. For the umbrella experiment covering the sub-experiments produced by every value of a, the distribution functions (under hypothesis i) are given by[23]:

$$\begin{aligned} F_i(y) &= P_i(Y \le y) \\ &= \int_a P_i(Y \le y \mid A = a) \cdot f_A(a)da \\ &= \int_a F_{i,a}(y) \cdot f_A(a)da. \end{aligned}$$

Hence, the DDF statistic for the 'umbrella' experiment (over a) is:

[22] A must be ancillary on a parameter space containing, or identical to, the BPS used for the E. C. inference.

[23] Or $\sum_a \{F_{i,a}(y) \cdot P(A = a)\}$ if A is discrete.

$$\begin{aligned} D(y) &= F_K(y) - F_H(y) \\ &= \int_a F_{K,a}(y) \cdot f_A(a)da - \int_a F_{H,a}(y) \cdot f_A(a)da \\ &= \int_a D_a(y) \cdot f_A(a)da. \end{aligned}$$

If each sub-experiment belongs to the same inference class, it follows that there exists a pairing function, $\pi : D_a(y) = D_a(\pi(y)) \ \forall y \ \forall a$.

Hence

$$\begin{aligned} & D(\pi(y)) \\ &= \int_a D_a(\pi(y)) \cdot f_A(a)da \\ &= \int_a D_a(y) \cdot f_A(a)da \\ &= D(y). \end{aligned}$$

Thus $\pi(\cdot)$ is also the pairing function for the umbrella experiment, which, therefore, belongs to the same inference class as all the sub-experiments; it follows that the exhaustive inference will interpret any outcome from such a sub-experiment the same way, regardless of whether or not it stands alone. Clearly, this is also a *necessary* condition for covering the effect of statistic A, since we need to ensure that:

$$\forall a \ \forall y : D(y) = D(\pi(y)) \Leftrightarrow D_a(y) = D_a(\pi(y)),$$

and this requires that $\forall a \ \forall y : \pi_a(y) = \pi(y)$.

When an ancillary statistic, A, defines sub-experiments that do not all belong to the same E. C. inference class, conditioning on the DDF for the umbrella experiment will not have the effect of ensuring that those sub-experiments are interpreted (as it were) in isolation. On the other hand, the DDF statistic partitions the support of the likelihood ratio statistic much more finely than most conventional ancillary statistics, and this may be considered more than adequate. We know that we cannot achieve the effect of conditioning on *all* 'weakly ancillary'[24] statistics *and* adhere to the sufficiency principle while remaining in the frequentist framework.

[24] The statistic has a distribution independent of θ but is not a function of the MSS.

What does this mean for Cox's example?

In Cox's example[25], the result (A) of a coin toss determined which of two Normal populations (same unknown mean, different known variances) was sampled from. Although Cox advocated conditioning on the observed value of A – because this provides more relevant results – he also insisted that an ancillary statistic should be a function of the MSS. The statistic A satisfies this requirement when $\mu \in \mathbb{R}$ but not when the parameter space does not contain an interval and, hence, not when $\Theta \equiv \{\mu_1, \mu_2\}$. When a conditional frequentist inference is based on an ancillary statistic that is not a function of the MSS, it necessarily breaches the SP. In ***Example 4.2***, we looked at an instance of Cox's scenario in which a hypothesis test was carried out by conditioning on A despite the fact that we were testing two simple hypotheses so that the parameter space was binary (i.e. $\{0,5\}$). Can we confirm that this approach breaches the SP?

For a binary parameter space, the LR is a minimal sufficient statistic and, in this case, the outcomes $(a=1, x_1 = 2.62906)$ and $(a=2, x_2 = 2.53227)$ both have the same likelihood ratio[26], i.e. $y = 0.851$. They therefore produce the same unconditional p-value (5%), but their p-values, *conditional on the observed value of* A, are (respectively) 9.4334% and 0.5666%. When we condition on A, we make different inferences based on two outcomes that produce the same value of a sufficient statistic. This is in breach of the SP and is undesirable since the usual interpretation of sufficiency is that the two outcomes contain exactly the same information about the question at issue. We seem to be left with an unpleasant choice between not conditioning on A, not satisfying the SP, or not carrying out a frequentist inference. Typically, these are the only options, but in Cox's case, we can get the effect of conditioning on A (and much more) by conditioning on the DDF statistic and this does not breach the sufficiency principle. Purely because Cox chose to use Normal models in his sub-experiments, the effect of his 'which population' ancillary statistic is covered by the DDF statistic for the umbrella experiment. Since both sub-

[25] Cox (1958).
[26] The (common) likelihood ratio of these two outcomes was the CLR for the 5% unconditional test – see *Example 4.2*.

experiments are for a Normal location model, they fall within the log-symmetric class of inference and so does the umbrella experiment (by the argument above). Thus the DDF statistic for the umbrella experiment produces the pairing function $\pi(y) = y^{-1}$ (as does each of the sub-experiments), and the E. C. inference gives $cp(y = 0.851) = \frac{0.851}{1.851} = 45.975\%$ as the (common) cp-value of either of the given outcomes. Since we get the same result for both outcomes, the method is not in breach of the SP and, since we get the same result from (for instance) $(a = 1, x_1 = 2.62906)$, observed from the umbrella experiment, as from $x = 2.62906$, observed from 'experiment 1' (now regarded as the whole experiment), we are satisfying the CP with respect to the ancillary statistic A.

The strength of the DDF statistic, as a basis for conditioning, means that it inevitably covers the effects of conditioning on some other statistics. Conditioning on the DDF statistic, instead of A (or nothing) produces results that are consistent with both the SP and CP (with respect to A) but the result is radically different. Neither the conventional p-value (5%) nor Cox's conditional values are consistent with the likelihood ratio of 0.85 ($\gg \frac{1}{2}$), whereas the cp-value of 46%, obtained by conditioning on the DDF statistic, is quite consistent with the LR.

Also note that, whenever our data has a likelihood ratio greater than or equal to *one*, it will be interpreted the same way by E. C. inference (a cp-value of 100%) regardless of whether or not the experiment from which is comes is embedded in a larger experiment; this is true even when the sub-experiments do not all come from the same class. Hence conditioning on the DDF will cover the effect of *any* other ancillary statistic whenever $y > 1$, without breaking the SP.

When we use E. C. inference, there is much more agreement about the interpretation to be placed on a given likelihood ratio (regardless of model differences) than when we use conventional inference.