Chapter 7: An alternative measure of evidence.

In Neyman-Pearson hypothesis testing, the error probabilities are primary. The probability of Type I error (significance level) is fixed in advance of the experiment and determines the critical likelihood ratio (CLR). If a unique, optimal NP test exists for testing two simple hypotheses, it will have higher power than any other test with the same significance level. We can often design a test so that both error probabilities are low by choosing an appropriate sample size. There is no doubt that the error probabilities are an important feature of experimental design, however, NP inference goes much further by making them the basis of the inference. Although the most powerful test is based on the likelihood ratio statistic (which is the minimal sufficient statistic for a binary parameter space), the critical likelihood ratio is determined by α. If, on the other hand, the LR were primary, α would be determined by the choice of CLR. This distinction is important because a result due to Robbins (see below) shows that, while a low CLR guarantees a low α, the converse is not true; we have already cited a case (***Example 3.1***) where α is conventionally low and yet the CLR is high, with the result that we reject H when the data is far more consistent with H than with K. Thus, low error probabilities, alone, are not an adequate criterion for a good test; the test result must tell us about something that we are interested in, only then is the fact that it will be correct a high proportion of the time of value.

Cox distinguished[1] between "making statements by a rule with certain specified long-run properties" and finding out "what we can learn from the data that we have". The former describes an approach that only has good error probabilities, i.e. where the 'statements' have been designed with the sole purpose of producing low error probabilities and have no other intrinsic quality. Yet, when introductory textbooks focus on examples where the data *does* comprise strong evidence against H as well as having a low p-value, they create a perception that these things are the same, and the widespread use of composite hypotheses further muddies the water since our ability to intuitively assess the evidence is lost – evidence against H relative to what?

[1] Cox (1958), pp. 360, 361.

In this work, we are interested in evidential inference and, specifically, in finding a good answer to the question 'Does the data constitute strong evidence against H relative to K?' This is one reason why we discuss most issues in the context of a comparison of two simple hypotheses. It can be objected that this is not a particularly realistic scenario: many tests that are carried out for practical purposes involve one or more composite hypotheses; for example, $\mu = 0$ versus $\mu > 0$. Since many of the problems with conventional inference are most easily observed in the case of two simple hypotheses, it may be tempting to believe that the scope of these problems is similarly limited; however this leaves some important questions unanswered. Is it claimed that a methodology that does not work (with respect to assessing evidence) in the case of two simple hypotheses nevertheless does work when we use composite hypotheses? Why should this be so, and how could it be proved? How do we interpret a composite hypothesis? Does $\mu > 0$ mean the same as the disjunction of $\{\mu = r\}$ over all $r \in \mathbb{R}^+$, and, if so, why would we expect there to be a single meaningful answer to a large number of varied questions, or accept an approach that can be shown not to work for certain simple components of the composite? If, on the other hand, $\mu > 0$ does not amount to a disjunction of many statements, what exactly does it mean, and how can we hope to interpret the result evidentially (or in any other way) when we do not know what issue we are investigating? Without clear answers to these questions it is impossible to make the case that standard methodology works for this scenario. It is plausible that the problems with standard methodology are simply obscured (rather than absent) in the composite-hypothesis context because of the greater complexity and vagueness in such a case; these features make it almost impossible to answer the question 'Does the conventional inference *make sense* here?' Being unable to answer this question, or even understand its meaning, does not amount to answering it in the affirmative.

7.1 Using the likelihood ratio as a measure of evidence.

The Neyman-Pearson theorem shows that a critical region based on the likelihood ratio statistic produces the most powerful test (where such a thing exists); this critical region contains all the data with a *relatively* small likelihood ratio. It is widely believed that a small p-value or (equivalently) the rejection of H at a low significance level signifies data that constitutes strong evidence against H (see Efron quotation in Chapter 2). An alternative view is that the likelihood ratio value measures the evidence *in absolute terms* (a position adopted most recently by Royall).

In the later chapters of this work we will extend the manner and scope of conditioning within the frequentist framework. In this chapter we argue that the likelihood ratio is a good non-frequentist measure of the evidence favouring one (simple) hypothesis relative to another, and show how it is used. This will provide us with a concrete measure with which to compare the results of conditional inferences developed in the following chapters, though we will also, at times, assess them directly by common-sense reference to the null and alternative distributions, as we did with conventional inferences in Chapters 3 & 6.

The quantitative law of likelihood.

Under this interpretation, a given value of the likelihood ratio corresponds to a given level of evidence for H relative to K. This association between the likelihood ratio and the level of evidence does not depend on the model[2] or other context of the problem as it does in frequentist inference; in frequentist inference, the same critical likelihood ratio is associated with widely varying α -values, significant in some cases but not others, depending on the model and hypotheses. Royall's Quantitative Law of Likelihood describes the connection between likelihood ratio and evidence, as follows.

[2] Except in so far as the model affects the value of the likelihood ratio itself through the two likelihood values.

The Quantitative Law of Likelihood[3] (QLL):

If hypothesis H implies that the likelihood of a random variable X at x is $L_H(x)$, while hypothesis K implies that the likelihood is $L_K(x)$, then the observation $X = x$ is evidence supporting H over K if and only if $L_H(x) > L_K(x)$ [i.e. LR>1], and the likelihood ratio, $L_H(x)/L_K(x)$, measures the strength of that evidence.

This is a quantitative extension of the Law of Likelihood (LL), defined by Hacking[4]:

***d* [data] supports *h* better than *i* whenever the likelihood ratio of *h* to *i* given *d* exceeds 1.**

(Hacking's one-way implication ('whenever') is usually replaced by the two-way 'if and only if', which Royall also employs. [5])

Before we look at basing inferences on the QLL, we need to consider some criticisms of the LL. Fitelson (2007) describes a number of cases that seem to be counter-examples to the LL (and thus also to the QLL). Some depend on allocating certain *priors* to the hypotheses "so, it seems to me that Likelihoodists needn't be swayed by such examples"[6], but another is rather more serious. It is possible to identify cases where the likelihood ratio of the data is more than *one* despite the fact that the data entails K but not H[7]; in such circumstances, it seems unreasonable to interpret the data as support for H over K, as the LL requires. However, this does not pose a problem for the application of the LL (or QLL) *in this work*, for the following reason. As will be seen in Chapters 8 and 9, the methodology we develop involves only *binary* parameter spaces within which the two (simple) hypotheses are logical opposites, i.e. in all cases, $K \equiv\sim H$. It is easy to show that, under these circumstances, Fitelson's

[3] Royall, p. 3.
[4] Hacking, p. 71.
[5] See, for instance Fitelson, p. 3.
[6] Fitelson, p. 5.
[7] Fitelson, p.5.

paradoxical cases cannot arise (since the likelihood ratio must be *zero* if the data entails K). Fitelson also shows that, under our circumstances (i.e. where $K \equiv\sim H$) the LL is consistent with "*any* Bayesian relevance measure of degree of non-relational confirmation."[8] Thus, (in our case) the reader is free to accept the LL criterion for any one of a number of different reasons. The QLL says that the likelihood ratio measures the degree to which H is favoured over K, but this measure is also, in our case, equivalent to a number of others; for instance, $LR(E) = r > 1$ if and only if $\frac{l(H,E)}{l(K,E)} = r^2 > 1$ (where $l(H^*, E)$ is the Bayesian relevance measure of degree of non-relational confirmation defined by $l(H^*, E) \equiv \frac{P(E|H^*)}{P(E|\sim H^*)}$, see Fitelson, p. 7). In short, it is not necessary to accept that the QLL criterion is appropriate in all circumstances, in order to accept its use in the context of this work.

Features of the likelihood ratio as a measure of evidence.

Any inference method based solely on the likelihood ratio as a measure of evidence, in accordance with the QLL, has the following features:

a) The method satisfies the likelihood principle (LP) and, hence, the CP (both restricted and unrestricted) and the SP.

b) The inference will be the same for the same observations produced by different stopping rules (as long as they produce the same LR).

c) The inference will tend to be very sensitive to the exact specification of both hypotheses – not just the 'null' hypothesis.

d) It will not be possible to test a composite hypothesis other than in very exceptional circumstances; for instance we can only test simple H against composite K using data $\underset{\sim}{x}$ if $LR(\underset{\sim}{x}) = L_H(\underset{\sim}{x}) \big/ L_{K_i}(\underset{\sim}{x})$ is the *same* $\forall\ K_i \in \mathrm{K}$.

Note also that, by symmetry, a LR of *one* corresponds to data that is neutral regarding the two hypotheses.

[8] Fitelson, p. 11.

7.2 Royall's canonical experiment.[9]

If the likelihood ratio measures the evidence in favour of H relative to K, with *one* corresponding to neutrality, when is the value large enough or small enough to constitute strong or significant evidence in favour of one or other hypothesis? There is no probabilistic interpretation of the LR since it lies in the interval $(0,\infty)$ rather than $(0,1)$. The best way to establish the meaning of the value is to look at a simple example and consider how strong the evidence needs to be before we regard it as significant. In order to do this, we need to stipulate prior probabilities, on our two hypotheses, of ½ and ½ so that our judgement is due to the data only, not to any initial preference between the hypotheses.

Suppose we have a coin before us and are interested in two hypotheses regarding $p = \mathrm{P}(head)$, namely:

H: the coin is fair, i.e. $p = \frac{1}{2}$
K: the coin is double headed, i.e. $p = 1$.

To establish the appropriate priors, suppose that we possess two coins one of which is indeed fair and the other double-headed and that one of these two coins has been randomly selected, with a probability of *one half*, and placed before us. This prevents us from being influenced by the view that K is intrinsically less plausible than H.

Clearly, we can dismiss K as soon as we obtain even one *tail*, but let us consider the case where all tosses of the coin result in a *head*; how many heads do we need to throw in order to feel that we have significant evidence that the coin is double-headed rather than fair? The likelihood ratios of H relative to K of some possible outcomes are shown below.

Table 7.1

Outcomes	*h*	*hh*	*hhh*	*hhhh*	*hhhhh*	*hhhhhh*	*hhhhhhh*
Likelihood ratio	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$

[9] Royall, p. 11.

If you feel that tossing the coin three times and getting a head on each occasion constitutes significant evidence that the coin is double-headed rather than fair, then you regard a likelihood ratio of $\frac{1}{8}$ or of 8 as being 'significant'. If two heads out of two tosses leaves you very doubtful, then you do not regard a likelihood ratio of $\frac{1}{4}$ (or 4) as significant. Royall takes the view that a likelihood ratio of 8 is "fairly strong evidence" and a likelihood ratio of 32, corresponding to five heads out of five tosses, is "quite strong evidence".[10] (My own intuition is that 3 *heads* is *not* significant evidence.) In any given case, the strength of evidence represented by the likelihood ratio of the observed data can be judged by reference to this simple example; for instance, if your data, $\underset{\sim}{x}$, has a likelihood ratio of 500 (or 1/500), the evidence favours one of your hypotheses nearly as strongly as *nine heads out of nine tosses* favours K, while, if your data has a likelihood ratio of 1/6.8 (a figure that arises in testing hypotheses about the mean of a Normal variate), the evidence is not as strong as that from the outcome *hhh*.

7.3 Basing a dichotomous inference on the likelihood ratio.

The likelihood ratio is a continuous measure and there is no need to use it only in a yes-or-no format. However, in the interest of comparing it with standard hypothesis testing techniques we will consider the case where we want to specify a critical likelihood ratio (CLR) and use it as the basis for accepting H as opposed to K (when $LR(\underset{\sim}{x}) > \text{CLR}$) or rejecting H in favour of K (when $LR(\underset{\sim}{x}) \leq \text{CLR}$). Since we are principally interested in rejecting H when there is strong evidence against H relative to K, we will use a CLR of $1/\lambda$ where $\lambda \gg 1$. The two error probabilities of the test, $\alpha(\lambda)$ and $\beta(\lambda)$, depend on λ; by the Neyman-Pearson theorem, the power $\kappa(\lambda) = 1 - \beta(\lambda)$ will be the highest that can be produced by any dichotomous test of level $\alpha(\lambda)$, since the test is based on the likelihood ratio statistic.

[10] Royall, p. 26.

Bias.

We saw in Chapter 3 that trying to control the bias between the hypotheses can lead us towards a LR-based inferential rule. The bias (between the two hypotheses) is readily identifiable, for any dichotomous rule, from the CLR itself. (This is also true in frequentist inference; but in that case the actual value of the CLR is almost never calculated. The cut-off value used in the test is the critical value for the natural statistic – not the critical value of the likelihood ratio statistic. Also, when either of the hypotheses is composite, there is no single CLR; the CLR of the test varies depending on which component of the composite hypothesis we consider.) The rule will be unbiased only if the CLR is *one*; we can construct a test that is biased in favour of H and only rejects H when there is strong evidence in favour of K (relative to H) by making the CLR sufficiently far below *one*, for example by rejecting H only when $LR(\underset{\sim}{x}) \leq \frac{1}{32}$; any test where the CLR>1 is biased in favour of K and the larger the CLR, the stronger the bias.

Contrasts and criticisms of frequentist inference via the likelihood ratio measure of evidence.

How significant is statistical significance?

Despite the prominent place give to the likelihood ratio statistic in the Neyman-Pearson theorem, there is an immense difference between methods such as that of Royall (or any method consistent with the LP) and that of Neyman & Pearson or the Fisher-Neyman-Pearson hybrids. This is because, in the likelihood methods, the likelihood ratio is interpreted in absolute rather than relative terms. The Neyman-Pearson theorem justifies rejecting H in favour of K when the likelihood ratio is less than a constant k, which is chosen so that (under H) most of the likelihood ratios observed from repetitions of the experiment will be larger than k; however this does not amount to rejecting H whenever the likelihood ratio is small because k may not be small. By contrast Royall's interpretation of the likelihood ratio is constant. A likelihood ratio of $\frac{1}{2}$ constitutes very weak evidence against H relative to K and this is

just as true for experiments that produce likelihood ratios in the range $[\frac{1}{2},3]$, so that $\frac{1}{2}$ is the strongest evidence against H that can ever be observed, as for those where the likelihood ratio lies in $[\frac{1}{1000},3]$, so that much stronger evidence against H is observable.

Under Royall's paradigm, a given LR is interpreted the same way in all contexts[11], and only likelihood ratios less than *one* are regarded as evidence against H (relative to K) to any degree whatsoever. By contrast, in frequentist inference, there is no *general* connection between the CLR and α, or between $LR(\underset{\sim}{x})$ and the p-value of $\underset{\sim}{x}$. In frequentist inference, the CLR (k_α) depends on the context, but it is usually considered that α does not (i.e. that a given value of α corresponds to a given rigor, regardless of the context); this gives rise to the problem that occurs when the sample size is very large and so is the CLR, with the result that H is rejected when the data is far more consistent with H than with K. There is not, in general, an upper bound on the CLR for any given value of α; even when α is very small, we can find a model and hypotheses to make the CLR arbitrarily large. Tests on the mean of a Normal population illustrate this point, as follows.

The LR of the critical (cut-off) point for a one-sided, α-level z-test is:

$$\text{CLR=LR}(c_\alpha) = \exp\{\tfrac{\delta}{2}(\delta - 2z_{1-\alpha})\},$$

where $\delta = \frac{|\mu_1-\mu_2|}{\sigma}$ and $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. For $\delta > z_{1-\alpha}$ this is an increasing function of δ, indicating that, for any α, σ and μ_1, it is possible to make the CLR arbitrarily large by choosing a value of μ_2 that is far enough away from μ_1. Some instances are shown below for the case $\sigma = 1,\ \mu_1 = 0,\ \alpha = 5\%,\ c_\alpha = 1.645$.

[11] This is not to say that we must use the same CLR in all contexts; we can vary the level of rigor (i.e. bias) depending on our aims and circumstances. However, a LR of $\frac{1}{32}$, in one context, always amounts to *stronger* evidence against H than a LR of $\frac{1}{10}$, in another context, regardless of how much the contexts vary.

Table 7.2

μ_2	1.645	2.000	3.290	5.000	10.000
CLR=LR(1.645)	$0.258 \approx \frac{1}{3.9}$	$0.275 \approx \frac{1}{3.6}$	1.00	71.933	3.726×10^{14}

When testing $\mu = 0$ versus $\mu = 10$, we reject H in favour of K at the 5% level even when the likelihood ratio of the data is of the order of 10^{14} and therefore favours H over K to about the same degree as 48 consecutive *heads* favours the double-headed hypothesis in the canonical example. (Intuitively, this rule makes no sense; an observation such as (say) $x = 1.8$ (in the rejection region) is far more consistent with $\mu = 0$ than $\mu = 10$.) Even when μ_2 and μ_1 are closer together, we do not require very strong evidence against H in order to reject it. The value of μ_2 that produces the smallest CLR is $\mu_2 = c_\alpha$, but even in this case the CLR is $0.258 \approx \frac{1}{3.9}$, indicating that we may reject H in favour of K when the evidence favours K to a *lesser* degree than the outcome *hh* (LR=$\frac{1}{4}$) favours the double-headed hypothesis. For the Normal location case, the 5% criterion never requires strong evidence against H.

This answers the question that we posed in Chapter 3, 'Does rejecting H in favour of a composite hypothesis necessarily imply than that for *some component* of the composite (i.e. some value of μ_2) the evidence strongly favours μ_2 over μ_H?' The answer is *no*, since we see above that we may reject H: $\mu = 0$ in favour of K: $\mu > 0$ at the 5% level and yet, for no value of $\mu_2 > 0$ is the CLR less than $\frac{1}{3.9}$; no data in the rejection region favours any hypothesised value of μ to a much greater degree than $\mu = 0$. In the Normal case, if we make α small enough, we can reach a point where rejecting a composite hypothesis does mean that the CLR is significantly small for *some* component of the alternative hypothesis, but it is necessary to use significance levels that are a lot smaller than the conventional 5%.

Table 7.3

Smallest CLR	One-sided α
$\frac{1}{8}$	2.07%
$\frac{1}{16}$	0.92%
$\frac{1}{32}$	0.42%

Even a one-sided test at the 2.5% level (or a 5% two-sided test) does not produce a significantly small value of the CLR for any component of the alternative; the smallest value at this α-level is $\frac{1}{6.8}$. The one-sided p-value must be well under $\frac{1}{2}$% before the test requires 'quite strong' evidence ($\text{LR} \leq \frac{1}{32}$) in favour of *any* of the alternative values.

In the Normal location case, the likelihood ratio statistic, for any two simple hypotheses, takes values in the interval $(0, \infty)$; thus, it is always theoretically possible to observe data with a genuinely small LR. However, there are models and hypotheses where no data has a LR than is sufficiently small to constitute strong evidence against H relative to K, and yet it is still possible to get small p-values.

Example 7.1.

Consider a test on the mean of an Exponential population based on a single observation. The density of the variable X is given by:

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

For the hypotheses H: $\theta = 2$ versus K: $\theta = 1$, the likelihood ratio is:

$$y = LR(x) = f(x;2) / f(x;1) = \tfrac{1}{2} e^{x/2}.$$

Thus, the $LR(x)$ only takes values in the interval $(\frac{1}{2}, \infty)$; a likelihood ratio of $\frac{1}{2}$ is the strongest evidence against H relative to K that sampling x can ever produce. This

evidence is not very strong, only the same as the evidence for the double-headed coin contained in the outcome from a single coin toss (*h*). Yet we can easily reject H using conventional methods since the p-value – a *relative* measure – is small whenever the likelihood ratio is among the smallest *that we can possibly get* in the given context. Thus, in this case, p-value$(x) = 1 - e^{-x/2} \to 0$ as $x \to 0$; for instance, the p-value of $x = 0.002$ is $\frac{1}{10}\%$, but the LR of this value is still greater than $\frac{1}{2}$. It is in the nature of this experiment that it can never produce even moderately strong evidence in favour of K relative to H, yet we can still get highly 'statistically significant' results.

Confidence intervals and likelihood intervals.

In addition to performing tests, we can find interval estimates for θ based on the level of evidence in the data. These are called 'likelihood intervals' (LI) and have the same relationship to likelihood test results as confidence intervals have to hypothesis test results. Recall that a (two-sided) $100(1-\alpha)\%$ CI, based on data $\underset{\sim}{x}$, contains all and only those values of θ that would not be rejected in favour of *any* alternative value by a conventional hypothesis test conducted at level $\alpha/2$. (If the tests involved are Neyman-Pearson optimal, then so are the confidence intervals.) Similarly, for any $\lambda > 1$, the $1/\lambda$ LI contains all those values of θ that (when specified as the null hypothesis, H) would not be rejected in favour of any alternative (K) by a test using the criterion:

Reject H in favour of K whenever $LR(\underset{\sim}{x}) = f_H(\underset{\sim}{x}) / f_K(\underset{\sim}{x}) \le 1/\lambda$.

Thus, the value $\theta' \in 1/\lambda$ LI if and only if $\forall \theta \in \Theta,\ f_{\theta'}(\underset{\sim}{x}) > \frac{1}{\lambda} f_{\theta}(\underset{\sim}{x})$. If $\hat{\theta}$ is a maximum likelihood estimate of θ, based on $\underset{\sim}{x}$, then $f_{\hat{\theta}}(\underset{\sim}{x}) \ge f_{\theta}(\underset{\sim}{x}),\ \forall \theta$, and a necessary and sufficient condition for θ' to be in the LI is: $f_{\theta'}(\underset{\sim}{x}) > \frac{1}{\lambda} f_{\hat{\theta}}(\underset{\sim}{x})$. For example, θ' is in the $\frac{1}{8}$LI for θ, if and only if $f_{\theta'}(\underset{\sim}{x})$ is at least one-eighth of the maximum value

reached by the likelihood of $\underset{\sim}{x}$ as a function of θ ($\underset{\sim}{x}$ fixed). It is a fact that likelihood intervals associated with evidentially significant levels, e.g. 1/8 or less, are usually wider than the 95% confidence interval based on the same data, sometimes much wider. This can give the impression that likelihood inference is less informative, but the phenomenon is due to the fact that the criteria for excluding a value from a LI are more reasonable and more stringent than those for excluding a value from a CI. The following example, again involving the exponential model, illustrates this point.

Example 7.2.

Consider the exponential model: $f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, x > 0, \theta > 0,$ where $\theta = E(X)$.

The conventional two-sided $100(1-\alpha)\%$ confidence interval for θ, based on a single observation, x, is $(x\{\ln(\frac{2}{\alpha})\}^{-1}, x\{\ln(\frac{2}{2-\alpha})\}^{-1})$. This interval excludes all (and only) those values that, if specified in the null hypothesis, would be rejected in favour of some alternative at the $\frac{\alpha}{2}$ significance level.

For a fixed value of x, the value of $\theta \in \mathbb{R}^+$ that maximises the likelihood (density) is $\theta = x$, thus the maximum value taken by the likelihood function is $\frac{1}{x} e^{-x/x} = \frac{1}{x} e^{-1} = m$ (for 'maximum'). The $\frac{1}{\lambda}$ likelihood interval for θ is the interval that excludes all (and only) those values of θ such that $f(x;\theta) \le m/\lambda$, i.e. excluding all those values that would be rejected in favour of some alternative by a dichotomous likelihood test using an evidential level of $\frac{1}{\lambda}, (\lambda > 1)$.

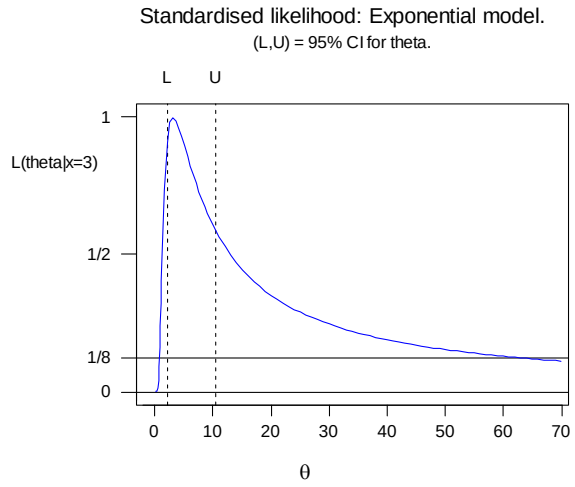
For this model, we have calculated confidence intervals with coverage of 99%, 98% and 95% and likelihood intervals with evidential levels $\frac{1}{32}, \frac{1}{16},$ and $\frac{1}{8}$, based on the data $x = 3$; the likelihood intervals are very much wider than the confidence intervals. (Note that the confidence intervals are based on tests using the levels ½ %, 1% and 2½% respectively.)

Table 7.4

$100(1-\alpha)\%$	Confidence Interval.	$\frac{1}{\lambda}$	Likelihood Interval.
99%	(1.001, 58.487)	1/32	(0.476, 285.714)
98%	(1.303, 28.474)	1/16	(0.548, 133.333)
95%	(2.164, 10.428)	1/8	(0.651, 60.606)

The lowest 'significant' evidential level is 1/8, hence the narrowest significant interval is the 1/8 interval; this is about the same width as the 99% confidence interval. The standard 95% confidence interval is very much narrower.

Figure 7.1



The plot shows the 95% confidence interval lower and upper bounds (L and U) on a plot of the standardised likelihood function. (The 1/8 likelihood interval, (0.65, 60.6), can be deduced from the points at which the horizontal line at 1/8 crosses the likelihood curve.) The maximum value of the function occurs at $\theta = x = 3$; the lower bound of the CI is very close to this point and has a standardised likelihood of over 0.9, so that the CI excludes (on the left side) many values that are almost as likely as the maximum likelihood estimate. Even the right bound has a standardised likelihood of more than 0.5. In fact, the 95% confidence interval is completely contained in the $\frac{1}{2}$- likelihood interval, which excludes all values of θ that are inconsistent with the data to at least the same degree that the hypothesis 'fair coin' is inconsistent with a

single coin toss resulting in a single head (relative to the hypothesis 'double-headed coin'). Thus the 95% CI excludes values of θ against which the evidence is weaker than that associated with a likelihood ratio of $\frac{1}{2}$. The 95% confidence procedure will indeed produce intervals that contain the true value of θ *ninety-five percent* of the time in the long run of samples, but it excludes a great many values of θ that are plausible according to this data, nor is the judgement consistent – some of the values excluded on the left side of the interval are more likely than some of the values (greater than 3) included on the right side.

When we compare conventional frequentist tests with tests based on observing the strength of evidence through the LR, and compare confidence intervals with likelihood intervals, we see that it is not the case that events which occur only rarely (under H) necessarily constitute strong evidence against H relative to a given K, or even, relative to *any* K, for a given model. When we compare H with a specific K, it is not even true that an event with an *arbitrarily* small probability will necessarily constitute evidence against H relative to K.

Why a small p-value is not enough.

For a test of two simple hypotheses, the conventional p-value of the data with a likelihood ratio of y is always less than y, as follows.
(Note that $\text{p-value}(\underset{\sim}{x}) = P_H(LR(\underset{\sim}{X}) \leq LR(\underset{\sim}{x})$.)

Let f and F be the density and distribution functions of the likelihood ratio statistic, Y, respectively, then $\forall y$, $\frac{f_H(y)}{f_K(y)} = y$. Let $y < 1$ (the proof is trivial for $y > 1$), then:

$$\begin{aligned} p(y) &= F_H(y) \\ &= \int_0^y f_H(r)dr \\ &= \int_0^y r \cdot f_K(r)dr \\ &< y \cdot \int_0^y f_K(r)dr \\ &= y \cdot F_K(y) \\ &< y. \end{aligned}$$

It follows that, if y is small, the p-value must also be small but not vice versa. Thus a small p-value is a necessary, *but not sufficient*, condition for the data to constitute strong evidence against H relative to K. The criterion for rejecting H in favour of K is not rigorous enough in frequentist inference.

Robbins' result.

Although data that occurs only rarely (under H) does not necessarily constitute reasonably strong evidence against H (relative to some K), the converse is true, i.e. data that constitutes strong evidence against H (relative to some K) occurs only rarely when H is true. Suppose that a random vector $\underset{\sim}{X}$ has density $f_H(\cdot)$ if H is true, and $f_K(\cdot)$ if K is true. We will reject H if we observe $\underset{\sim}{x}$ such that $LR(\underset{\sim}{x}) = f_H(\underset{\sim}{x}) / f_K(\underset{\sim}{x}) \leq 1/\lambda$. What is the probability that this will happen if H is actually true? The exact probability depends on f_H and f_K, but we can find a general upper bound.

For all $\underset{\sim}{x}$ such that $LR(\underset{\sim}{x}) \leq 1/\lambda$, $f_H(\underset{\sim}{x}) \leq f_K(\underset{\sim}{x}) / \lambda$, therefore:

$$\begin{aligned} P_H(\underset{\sim}{X} : LR(\underset{\sim}{X}) \leq 1/\lambda) &= \int_{\underset{\sim}{x}:LR(\underset{\sim}{x}) \leq 1/\lambda} f_H(\underset{\sim}{x})d\underset{\sim}{x} \\ &\leq \int_{\underset{\sim}{x}:LR(\underset{\sim}{x}) \leq 1/\lambda} (f_K(\underset{\sim}{x}) / \lambda)d\underset{\sim}{x} \\ &= \frac{1}{\lambda} \int_{\underset{\sim}{x}:LR(\underset{\sim}{x}) \leq 1/\lambda} f_K(\underset{\sim}{x})d\underset{\sim}{x} \\ &\leq \frac{1}{\lambda}. \end{aligned}$$

An even stronger result is due to Robbins (1970), following work by Ville and Wald; the version that most concerns us is given by Royall[12] in a form close to the following. Suppose that we use the stopping rule '*Continue sampling until* $LR(\underset{\sim}{x}) \leq 1/\lambda$'. This rule is designed to elicit evidence that favours K against H to the λ-degree; despite this, the probability of finishing the experiment in any *finite* time with the desired result is low whenever H is true, as follows.

Let $\underset{\sim}{X} = (X_1, \ldots, X_n)^T$, where $X_1, \ldots, X_n$ are independent and identically distributed random variables with density $f_H(\cdot)$ and $f_K(\cdot)$ under hypotheses H and K, and thus $LR(\underset{\sim}{X}) = \prod_{i=1}^{n} \{ f_H(X_i) / f_K(X_i) \}$. Then, $P_H\{LR(\underset{\sim}{X}) \leq 1/\lambda$ for some finite $n\} \leq 1/\lambda$.

Under H, the probability that we will observe data that provides evidence for K against H to the degree specified by λ is no greater than $1/\lambda$. This is an upper bound on the probability the exact value of which varies with the model and hypotheses; it shows that we can ensure that the probability (under H) of rejecting H in favour of K is no greater than (say) 5%, by using $\lambda = 20$ in our CLR; the probability of Type I error is then bounded above by 5% no matter what the model and hypotheses may be and no matter which stopping rule we use. The fact that this result is independent of the stopping rule means that, as long as the rejection criterion is reasonably stringent $(\lambda \gg 1)$, the result of the experiment cannot be rigged (except with a small probability of success) to reject a true hypothesis in favour of any specific alternative by using a biased stopping rule.

It is also the case[13] that, whenever we test a specific hypothesis against any other, there exists a finite sample size, $n = n(\lambda, \varepsilon)$, large enough that the evidence will favour whichever of the two hypotheses is true by a factor of λ with probability $1 - \varepsilon$ $(\varepsilon > 0)$.

[12] Royall, p. 7.
[13] Royall, pp. 7, 8.

Birnbaum noted that adherence to the likelihood principle is, in general, inconsistent with the controlling of error probabilities[14], and his supposed rejection of the likelihood principle[15] – which gave some comfort to frequentists[16] – was based on this point. However it is clear that the conflict only arises when we use composite hypotheses; for the case of two simple hypotheses it is quite possible to have your cake and eat it too (as pointed out by Giere[17], almost in an aside). In such a case we can retain the before-experiment error probabilities as an essential feature of the design while using appropriate values of the likelihood ratio of the data for the after-experiment evaluation of evidence.

Weak evidence.

In Chapter 3, we showed that observed data may be 'weak' in the sense of providing little information about which of two specified hypotheses is true, and that this can happen even when an optimal test has very low values of α and β. One of the reasons for performing conditional inferences has been to distinguish between data that is more or less informative about the question at issue.[18] Standard tests do not distinguish between weak and strong data; they put weak data into one or other or both (accept/reject) regions. High power tests put weak data in the rejection region where it counts towards the power of the test (often interpreted as 'the test's ability to detect that K is true') despite the fact that it is not good evidence for the truth of K rather than H. When the power is less high, some of the weak data lies in the acceptance region where it contributes to $1-\alpha$ ($= P(\text{Accept H}|\text{ H true})$) despite the fact that it is not good evidence for H. For fixed sample sizes, Royall has shown[19] that, in order to achieve a high probability that the data will be neither misleading (i.e. providing strong evidence against the true hypothesis) nor weak (giving no definite evidence either way), it is necessary to use sample sizes much larger that those which

[14] Unpublished MS, quoted in Giere.
[15] Birnbaum (1977), p. 24. "A concept of statistical evidence is not plausible unless it finds 'strong evidence for K as against H' with small probability (α) when H is true, and with much larger probability $(1-\beta)$ when K is true." This paper was published after Birnbaum's death in 1976.
[16] See Stuart, Ord & Arnold, p.440.
[17] Giere, p. 10.
[18] See Buehler (1982), Cox (1988).
[19] Royall, pp. 90-107.

produce high power for conventional significance levels. This is not surprising since it is a more ambitious aim; by making the power high and significance level low, we control the probability of misleading evidence, but not weak evidence.

7.4 Where to for frequentist inference?

The contrast afforded by likelihood inference highlights again the shortcomings of conventional frequentist inference when it comes to answering the question 'Does this data constitute strong evidence against H relative to K?' Does this mean that we must either live with the consequences of these problems, completely unalleviated, or abandon frequentism altogether in favour of a likelihood method? We have seen that restricted conditioning, of the forms championed by Cox or Fisher, does nothing to mitigate the problems that arise when we try to answer this question using the frequentist approach; the characteristics of frequentist inferences carried out on the sample spaces of sub-experiments (as usually defined) are no better (in any systematic sense) than those that arise from applications to umbrella experiments. Nevertheless, since unrestricted conditioning leads to the likelihood principle and methods consistent with the likelihood principle are free of these problems, it is plausible that some conditioning approach, short of unrestricted conditioning, may improve matters, while still allowing us to retain the frequentist framework. In the rest of this work, we show that it is possible to use Fisher/Cox conditioning in a way that is more fruitful than its traditional application. The resulting *exhaustive conditional inference* is wholly based on the frequentist approach and yet it is altogether free of some of the failings associated with frequentism and greatly mitigates others. This approach produces results that are radically different from conventional frequentist inferences and, in important respects, much closer to likelihood inferences.