Chapter 8: Exhaustive conditional inference for log-symmetric models.

8.1 Ancillary statistics for a binary parameter space.

We were able to clarify the situation in Welch's Uniform example by conditioning upon a statistic, A_{ij}, that is ancillary for a binary parameter space $\Theta_B \equiv \{\theta_i, \theta_j\}$. This allowed us to identify some critically important features of the example and to make inferences much more sensible than those produced by, either, unconditional ('optimal') theory or conditioning on a variable (R) that is ancillary for the larger parameter space, $\mathbb{R}$. A striking feature of this case is that the inference conditional on R, which is now standard, seems quite reasonable until we consider two simple hypotheses, at which point the flaws in the approach become apparent. Given a binary parameter space, conditioning on A_{ij} satisfies all the requirements of the *restricted* conditionality principle advocated by Cox, and even that advocated by Fisher. Such a principle does not lead to the likelihood principle when coupled with the sufficiency principle as Birnbaum's unrestricted CP does.

In this chapter and those that follow, we will look at what happens when we condition consistently on ancillary statistics of this type. The fact that we are using a restricted form of CP means that we can remain within the general framework of frequentist inference, replacing the measures α, β and p-value by their conditional counterparts which, by Cox's argument, are more relevant to the question at issue. Since our ancillary statistics must be functions of the MSS for Θ_B, some statistics that are functions of the MSS for larger parameter spaces are ruled out of contention[1]; on the other hand, the requirement that the statistic have the same distribution for all θ is

[1] Let Θ_B be a binary parameter space, which is a proper subset of the larger space Θ, and suppose that S is a MSS over Θ, then $Y = LR(X)$, which is the MSS over Θ_B, must be a function of S. Any (restricted) ancillary statistic over Θ must be a function of S, but this does not guarantee that it will also be a function of Y. If it is not, it will not be a candidate for the role of ancillary statistic over Θ_B.

less onerous for a binary parameter space than for a larger space; thus, some statistics that are not ancillary for larger spaces are ancillary for our purposes. We will find that we are often able to identify ancillary statistics for all the binary subsets of the natural parameter space even in cases where no such statistic exists for the natural parameter space *per se*; thus the scope of our conditional inference is much wider.

Fisher/Cox conditioning, based on large parameter spaces, usually produces results that are quite radically different from unconditional (optimal) inference; consider, for instance, Cox's two-stage example and the difference between the 5% conditional and unconditional rejection regions (§4.3). We might imagine that these new inferences might be (in some sense) *closer* to 'likelihood' inference, since the LP can be derived from the unrestricted CP, but this is not the case. When isolated sub-experiments, of the type identified by Cox, are analysed by the usual frequentist methods, the inference can be every bit as inconsistent with the LP, or law of likelihood (LL), as an unconditional inference; in particular, even when the conditional significance level is low, there is no general upper bound on the value of the critical likelihood ratio. This is true even when the restricted CP is extended to allow for conditioning on *approximately* ancillary statistics – a version that is less restrictive than Birnbaum's CP[2] regarding the distribution of the statistic. We might infer from this that the difference between the two conditionality principles (whether or not the ancillary statistic is a function of the MSS) is critical and prevents any convergence of the methods. However we will show that, when a *binary* parameter space is used, the restricted version of the CP produces results that are, in a meaningful way, more consistent with both the LP and the LL than any existing frequentist methods. Thus, our approach occupies a position between frequentist and likelihood methods; although it is a frequentist method, constraints in terms of the likelihood ratio arise naturally from it.

In this chapter we consider models with a particular type of symmetry property that enables us, not only to identify ancillary statistics for binary parameter spaces, but also to provide complete details of the inferential results that follow from conditioning upon them. Since such statistics exist for all binary subsets of the natural parameter

[2] Birnbaum's ancillary statistic has exactly the same distribution for all θ, whereas an approximate ancillary statistic has a distribution that may vary slightly over θ.

space, it is also possible to define conditional confidence intervals (CCI) for the parameter of interest. Inference on the mean of a Normal population where the variance is either known or can be estimated very reliably is among these cases; thus, inference about any θ based on its maximum likelihood estimator comes under this heading as long as the sample is reasonably large and the mean and variance are functionally independent (in Chapter 10 we consider cases were the latter requirement is not met).

Note that all p-values are one-sided since they are associated with a particular simple alternative hypothesis. If the observed value of the likelihood ratio statistic (i.e. the LR of the observed data) is y_0, then the conventional p-value of this data is $p(y_0) = P_H(Y \le y_0)$, which is the smallest value of α that could lead us to reject H on observing this data.

8.2 Conditional inference on the mean of a Normal population.

Weak and strong data identified by an ancillary statistic.

In ***Example 3.2***, we looked at a test of the hypotheses H: $\mu = 0$ versus K: $\mu = 3$ based on $T \sim N(\mu,1)$. We split the sample space, $\mathbb{R}$, into two regions:

$$\tau_1 \equiv [1.0, 2.0]$$
$$\text{and } \tau_2 \equiv \mathbb{R} \setminus \tau_1.$$

The data in τ_1 provides weaker evidence regarding the two hypotheses than data in τ_2. The optimal 5% test is defined by the rule *Reject H if t>1.645*. When we calculated the error probabilities of this test, conditional upon the data being weak ($T \in \tau_1$), we found them to be $\alpha_1 = 20.05\%$, and $\beta_1 = 47.78\%$ (compared with the unconditional values, $\alpha = 5\%$ and $\beta = 8.78\%$), verifying the view that data in this range does not produce reliable results.

It is appropriate to condition on the event $T \in \tau_1$ because it has the same probability under H as under K. Our ancillary statistic is $A*$, defined by:

$$A* = A*(T) = \begin{cases} 1, & T \in \tau_1 \\ 2, & T \in \tau_2. \end{cases}$$

$A*$ is ancillary (in the restricted sense) for the binary parameter space $\Theta_B \equiv \{0,3\}$ since it has the same distribution $\forall \mu \in \Theta_B$ and is a function of the MSS on Θ_B, as follows.

The MSS on any binary parameter space is the likelihood ratio statistic, $LR(T)$. In general, if $T \sim N(\mu, \sigma^2)$, then for testing H: $\mu = \mu_1$ against K: $\mu = \mu_2$, the likelihood ratio is:

$$\begin{aligned} LR(t) &= \frac{f_H(t)}{f_K(t)} \\ &= \exp\{\tfrac{1}{2\sigma^2}[(t-\mu_2)^2 - (t-\mu_1)^2]\} \\ &= \exp\{\tfrac{1}{\sigma^2}[(\mu_1-\mu_2)(t-\bar{\mu})]\}, \end{aligned}$$

$$\text{where } \bar{\mu} = \tfrac{(\mu_1+\mu_2)}{2}.$$

Thus any function of t that is defined symmetrically around the value $\bar{\mu}$ is a function of the likelihood ratio. In our example, $\bar{\mu} = 1.5$ and the sets τ_1 and τ_2, which define $A*$, are symmetric about this point, therefore $A*$ can be written in terms of the likelihood ratio (i.e. the MSS), as follows:

$$A* = \begin{cases} 1, & LR(t) \in [e^{-1.5}, e^{1.5}] \\ 2, & \text{otherwise.} \end{cases}$$

An equivalent two-stage experiment.

We can use $A*$ to construct a notional two-stage experiment that has the same probabilistic structure as the experiment observing t.

STAGE 1.

We take a single observation of the random variable, A^*, observing a to be either 1 or 2. This is an observation from the following distribution (that applies under both H and K).

Table 8.1

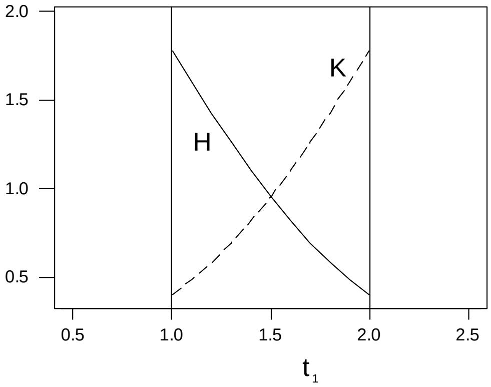
a	1	2
$P(A^* = a)$	0.1359	0.8641

(Note that this distribution shows how likely we are to get weak or strong data.)

STAGE 2.

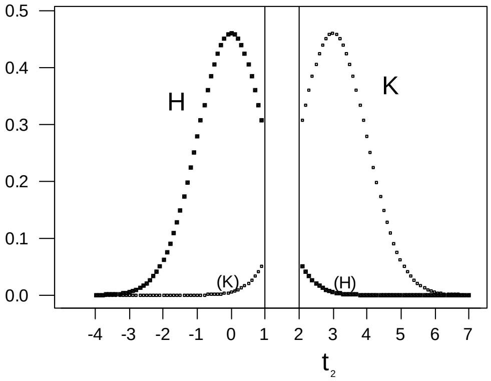
If $a = 1$ in stage 1, we observe the value of a random variable $T_1 \in [1.0, 2.0]$ that has one of the densities shown in the following plot, under H and K (respectively).

Figure 8.1



If $a = 2$ from stage 1, then we observe the value of a random variable $T_2 \in \mathbb{R} \setminus [1.0, 2.0]$ that has one of the densities shown in the following plot.

Figure 8.2



The conventional inference, based on T, combines the distributions of T_1 and T_2 using, as weights, the probabilities of the two possible values of a; this combination is then used regardless of what the result of stage 1 (observed value of a) actually was. As argued in previous chapters, the alternative approach is to use only the observed value and distributions (under H and K) of the statistic that is observed in the second stage of the experiment. This is more appropriate since the outcome of stage 1 is not directly informative about the test ($LR(a)=1, \forall a$), but T_2 is a more reliable statistic, for the test, than T_1 (so $A*$ is a precision index[3]), and the test result should reflect this fact.

8.3 Defining an exhaustive ancillary statistic.

The argument for conditioning on the strength of evidence is compelling, but we can derive different results from the same data if we partition the sample space of T differently; many such divisions will give rise to ancillary statistics. Further, we can get a better picture of the accuracy of the stage 2 variable if we divide the sample space into more than two subsets while still maintaining the ancillarity. The further away from $\bar{\mu}$ the data is, the stronger our information about which hypothesis is true.

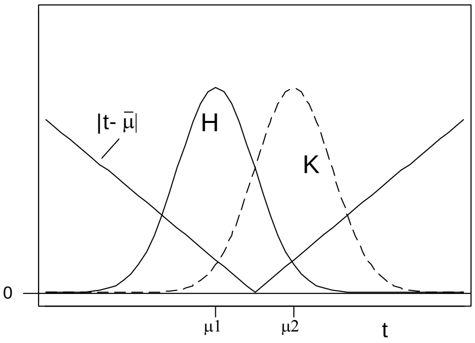
[3] Not necessarily the most discriminating precision index.

Therefore we define a new ancillary statistic:

$$A = G(T) = | T - \bar{\mu} | .$$

A is a continuous variable on $\mathbb{R}^+$ with a distribution that is the same under H and K, as can be deduced from the following plot[4].

Figure 8.3



A is a function of the MSS, $LR(T)$, as follows:

$$A = \tfrac{\sigma^2}{|\mu_1 - \mu_2|} \cdot | \ln LR(T) | .$$

The coefficient of $| \ln LR(T) |$ is a positive constant for any given σ^2 and Θ_B, and thus $| \ln LR(T) |$ is also an ancillary statistic. Since A and $| \ln LR(T) |$ are one-to-one functions of each other, they partition the sample space of T in exactly the same way and, if we condition on the observed value of either variable, we will obtain the same results. Any variable that is a function of A is also ancillary, but, unless it is a one-to-one function of A, it will be less informative. $A*$ is an example of this, since

$$A* = \begin{cases} 1, \text{ if } A \le \tfrac{1}{2} \\ 2, \text{ if } A > \tfrac{1}{2}. \end{cases}$$

[4] I.e. check that $P(| T - \bar{\mu} | < c)$ is the same under H and K.

If we want to find the most relevant error probabilities of any particular test criterion, we should condition on the observed value of A.

We call A *exhaustive* because it possesses the following feature.

> A statistic, $A = \Psi(Y)$, that is ancillary on a binary parameter space, is *exhaustive* if it partitions the support of the likelihood ratio statistic, Y, into subsets, all of which contain exactly *two* values of y, except for the subset $\{y = 1\}$.

This definition is motivated by the fact that any subsets produced by an ancillary partition of the likelihood ratio statistic must contain more than one value unless that value is $y = 1$. It follows that subsets containing two elements are the smallest that can be produced. Thus, an exhaustive ancillary statistic partitions the support of Y maximally with respect to the size of the subdivisions. Since ancillary statistics that are one-to-one functions of each other create the same partitions, it follows that any one-to-one function of an exhaustive ancillary statistic (EAS) is also an EAS. The statistic $\frac{\sigma^2}{|\mu_1-\mu_2|} \cdot |\ln Y|$ is exhaustive since $|\ln Y|$ is an EAS, as follows: $|\ln Y| = a > 0$ $\Leftrightarrow$ $y \in \{e^{-a}, e^{+a}\}$ and $|\ln Y| = 0$ $\Leftrightarrow$ $y = 1$. (From here on, we define A as $|\ln Y|$ for greater simplicity.)

8.4 Conditional probability in the limit.

A is a continuous variable and hence $P(A = a) = 0, \forall a$. In order to condition upon the observed value of A, we must define the relevant conditional probability in the limit.

If the observed value of Y is y and $A = \Psi(Y)$, then the observed value of A is $\Psi(y) = a$. Let[5]

$$\begin{aligned}\vec{P}(Y = y \mid A = a) &= \lim_{\varepsilon \to 0} P(Y \in (y-\varepsilon, y] \mid A \in (\Psi(y-\varepsilon), \Psi(y)]) \\ &= \lim_{\varepsilon \to 0} \left\{ \frac{P(Y \in (y-\varepsilon, y])}{P(A \in (\Psi(y-\varepsilon), \Psi(y)])} \right\}\end{aligned}$$

This definition is convenient, but it also makes sense for our purposes. For any fixed $\varepsilon > 0$, and i in some set I_ε, a discrete variable, say A^ε, defined by

$$A^\varepsilon = i, \text{ if } A \in (i-\varepsilon, i]$$

is ancillary (since it is a function of A) and we can gain information by conditioning upon it; this is unproblematic since A^ε is a discrete variable. By making ε smaller, we can get increasingly accurate information about the amount of evidence obtainable from the data; the greatest amount of information is in the limit as $\varepsilon \to 0$, which is to say, as $A^\varepsilon \to A$ and $I_\varepsilon \to \mathbb{R}^+$. Viewing A as the limiting case of A^ε, as $\varepsilon \to 0$, provides the above definition of probability conditional upon $A = a$. The symbol $\vec{P}$ reminds us that this ‘probability’ is defined in the limit, although it will be indistinguishable from a conventional conditional probability derived from a discrete A^ε with sufficiently small ε.

What features does a conventional z-test have, conditional on the observed value of A? In order to answer this question we need to derive the (limiting) conditional distribution of Y *given* $A = |\ln y_0|$.

[5] We assume that the function $\Psi(\cdot)$ is monotonic increasing on $(y-\varepsilon, y]$ for suitably small ε. This is true for all the ancillary functions of $Y = LR$ considered in this and later chapters of this work as long as $y < 1$. If $y > 1$ we will find that $\Psi(\cdot)$ is a monotonic decreasing function in $[y, y+\varepsilon)$ allowing us to use the formula:

$$\vec{P}(Y = y \mid A = a) = \frac{\lim_{\varepsilon \to 0} P(Y \in [y, y+\varepsilon))}{\lim_{\varepsilon \to 0} P(A \in (\Psi(y+\varepsilon), \Psi(y)])}$$

8.5 Conditioning in the log-symmetric case.

Definition of log-symmetry.

Let Y be the likelihood ratio of a random variable, T, with respect to simple hypotheses H and K. If $A = |\ln Y|$ has the same distribution under H and K, we say that the scenario is *log-symmetric* and it follows that A is ancillary, in the restricted (Cox) sense, and exhaustive on the binary parameter space defining H and K.

Conditional probabilities.

Such a structure uniquely defines the conditional distributions (under H and K) of Y given $A = |\ln Y|$, the details of which are derived below.

When we condition on the observed value of A ($a_0 = |\ln y_0|$), the conditional distribution of Y given $A = a_0$ is (in the limit) dichotomous, that is, Y may only take those two values (y_0 and y_0^{-1}) consistent with the observed a_0[6]. We need to derive the probabilities of these two values under both H and K.

First, we derive some helpful relationships.

The observed value of Y, y_0, may be greater than or less than *one*, i.e. in some cases the observed value will be the smaller of the two values consistent with a_0 and in other cases it will be the larger of the two values. We cannot assume one way or the other without loss of generality, however we may distinguish between the two

[6] That is, *if* $|\ln Y| = a_0$, *then* $\ln Y$ must be either $-a_0$ or $+a_0$ and Y must be either e^{-a_0} or e^{+a_0}. Since $a_0 = |\ln y_0|$, it follows that the only two values of Y consistent with this value of A are $\exp\{-|\ln y_0|\}$ and $\exp\{+|\ln y_0|\}$ which are y_0 and y_0^{-1} (*not* necessarily respectively).

elements of the conditional sample space, according to size, without making such an assumption; thus, let

$$\begin{aligned}&y_1 = \min\{y_0, y_0^{-1}\}\\ &\Rightarrow y_1^{-1} = \max\{y_0, y_0^{-1}\}\\ &\text{hence } y_1 < 1 < y_1^{-1}.\end{aligned}$$

The observed value, y_0, will be sometimes y_1 and sometimes y_1^{-1}.

Since A has the same distribution under both hypotheses, then, letting f_H and f_K be the density functions[7] of Y under the respective hypotheses, and utilising the relationship between the density of A and the density of Y, it follows that:

$$y_1^{-1} f_H(y_1^{-1}) + y_1 f_H(y_1) = y_1^{-1} f_K(y_1^{-1}) + y_1 f_K(y_1).$$

We can use this, in combination with the fact that $f_H(u) / f_K(u) = u \ (\forall u)$, to show that:

$$\text{(i) } f_H(y_1^{-1}) = y_1 f_H(y_1) \text{ and (ii) } f_K(y_1^{-1}) = y_1^3 f_K(y_1).$$

We want to find $\vec{P}(Y = y_1 \ given \ A = |\ln y_0|)$, and since $|\ln y_0| = |\ln y_1|$, this is equal to:

$$\begin{aligned}&\lim_{\varepsilon\to 0} P(\{y_1 - \varepsilon < Y < y_1\} \ given \ \{|\ln(y_1-\varepsilon)| < |\ln Y| < |\ln y_1|\})\\ &= \lim_{\varepsilon\to 0}\left\{\frac{P(\{y_1 - \varepsilon < Y < y_1\})}{P(\{|\ln(y_1-\varepsilon)| < |\ln Y| < |\ln y_1|\})}\right\}\\ &= \lim_{\varepsilon\to 0}\left\{\frac{P(\{y_1 - \varepsilon < Y < y_1\})}{P(\{y_1 - \varepsilon < Y < y_1\}) + P(\{y_1^{-1} < Y < (y_1-\varepsilon)^{-1}\})}\right\}\end{aligned}$$

We can divide the numerator and denominator by $\varepsilon > 0$ and hence write this expression as:

[7] This derivation is constructed on the basis that Y is a continuous variable, but note that the results are also valid for Y discrete or partly discrete.

$$\lim_{\varepsilon\to 0}\left\{\frac{P(\{y_1-\varepsilon<Y<y_1\})/\varepsilon}{P(\{y_1-\varepsilon<Y<y_1\})/\varepsilon+P(\{y_1^{-1}<Y<(y_1-\varepsilon)^{-1}\})/\varepsilon}\right\}$$
$$=\lim_{\varepsilon\to 0}\left\{\frac{\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon}{[\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon]+[\{F((y_1-\varepsilon)^{-1})-F(y_1^{-1})\}/\varepsilon]}\right\}$$
$$=\frac{\lim_{\varepsilon\to 0}[\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon]}{\lim_{\varepsilon\to 0}[\{F(y_1)-F(y_1-\varepsilon)\}/\varepsilon]+\lim_{\varepsilon\to 0}[\{F((y_1-\varepsilon)^{-1})-F(y_1^{-1})\}/\varepsilon]}$$
$$=\frac{f(y_1)}{f(y_1)+\lim_{\varepsilon\to 0}[\{F((y_1-\varepsilon)^{-1})-F(y_1^{-1})\}/\varepsilon]}$$

where F is the distribution function of Y, and the density function, f, is its derivative.

Let $H(u)=F(u^{-1})$, then

$$\lim_{\varepsilon\to 0}[\{F((y_1-\varepsilon)^{-1})-F(y_1^{-1})\}/\varepsilon]$$
$$=\lim_{\varepsilon\to 0}[\{H(y_1-\varepsilon)-H(y_1)\}/\varepsilon]$$
$$=-H'(y_1)$$
$$=f(y_1^{-1})\cdot y_1^{-2}.$$

Thus,

$$\vec{P}(Y=y_1\mid A=|\ln y_0|)$$
$$=\frac{f(y_1)}{f(y_1)+f(y_1^{-1})\cdot y_1^{-2}.}$$

Using (i) and (ii) above, it follows that:

$$\vec{P}_H(Y=y_1\mid A=|\ln y_0|)$$
$$=\frac{f_H(y_1)}{f_H(y_1)+f_H(y_1^{-1})\cdot y_1^{-2}}$$
$$=\frac{f_H(y_1)}{f_H(y_1)+y_1 f_H(y_1)\cdot y_1^{-2}}$$
$$=\frac{y_1}{(1+y_1)},$$

and

$$\begin{aligned}&\vec{P}_K(Y = y_1 \mid A =|\ln y_0|)\\ &= \frac{f_K(y_1)}{f_K(y_1) + f_K(y_1^{-1})\cdot y_1^{-2}}\\ &= \frac{f_K(y_1)}{f_K(y_1) + y_1^3 f_K(y_1)\cdot y_1^{-2}}\\ &= \frac{1}{(1+y_1).}\end{aligned}$$

Hence the distribution of Y given that $A =|\ln y_0|$ is that shown in **Table 8.2**.

Table 8.2

y	$y_1 = \min\{y_0, y_0^{-1}\}$	$y_1^{-1} = \max\{y_0, y_0^{-1}\}$
$\vec{P}_H(y)$	$\frac{y_1}{(1+y_1)}$	$\frac{1}{(1+y_1)}$
$\vec{P}_K(y)$	$\frac{1}{(1+y_1)}$	$\frac{y_1}{(1+y_1)}$

Features of the conditional distributions.

Some examples of the conditional distributions of Y are as follows.

If we observe data with a likelihood ratio of $\frac{1}{3}$ (i.e. $y_0 = \frac{1}{3}$) then we have also observed the variable A taking the value $\ln 3$. The distribution of Y conditional upon $A = \ln 3$ is:

Table 8.3

y	$\frac{1}{3}$	3
$\vec{P}_H(y)$	$\frac{1}{4}$	$\frac{3}{4}$
$\vec{P}_K(y)$	$\frac{3}{4}$	$\frac{1}{4}$

(Note that, as should be the case, Y is the likelihood (probability) ratio for the conditional distributions as it was for the unconditional distributions.)

In this case the conditional p-value of the observation $y_0 = \frac{1}{3}$ is

$\vec{P}_H(Y \le \frac{1}{3} \mid A = \ln 3) = \frac{1}{4}$.

On the other hand, if we observe data with a likelihood ratio of 10, the distribution of Y given than $A = \ln 10$ is:

Table 8.4

y	$\frac{1}{10}$	10
$\vec{P}_H(y)$	$\frac{1}{11}$	$\frac{10}{11}$
$\vec{P}_K(y)$	$\frac{10}{11}$	$\frac{1}{11}$

This gives a conditional p-value for the observation $y_0 = 10$ of

$\vec{P}_H(Y \le 10 \mid A = \ln 10) = 1$.

Note that, when a is larger, there is a greater difference between the (conditional) distributions of Y under H and under K and this makes it easier to distinguish between the two hypotheses on the basis of y. If, in the first case ($A = \ln 3$) we used a rejection rule *Reject H when* $y = \frac{1}{3}$, then the conditional significance level and probability of Type II error $(\alpha_a$ and $\beta_a)$ are both 25%, whereas, in the second case ($A = \ln 10$) the rule *Reject H when* $y = \frac{1}{10}$ produces conditional error probabilities of 9.09%. For any $a = |\ln y_0|$, and rule *Reject H when* $y = y_1$ the two conditional error probabilities are *the same* and equal $y_1/(1+y_1) = 1/(1+e^{+a})$, which is a one-to-one function of a. This shows that A is a good precision index[8].

[8] It can also be argued directly from a likelihood perspective that $|\ln Y|$ measures the ‘absolute (weight of) evidence’ in the data and is thus a ‘natural’ precision index.

It is clear that any value of y greater than *one* has a conditional p-value of 100%. From a likelihood point of view, such a value indicates that the data is more consistent with H than K.

We have identified the exhaustive ancillary statistic, $|\ln Y|$, and derived the distributions of the minimal sufficient statistic, Y, conditional upon it. In §8.6 we will use this to show that conventional inference on μ has disturbing conditional properties before developing (in §8.7) an approach with better conditional features.

8.6. Exhaustive conditional inference on the mean of a Normal population.

The conventional and conditional significance levels of a standard z-test.

We have shown that in Cox's two-stage example and Welch's Uniform example, the unconditional, 'optimal' test has conditional features (for instance, significance level) that vary depending on the value of the ancillary statistic. In this section we look at the conventional inference on the mean of a Normal population (variance known) (popularly called 'z-tests') and assess the significance level and power of these tests conditional on the given value of the exhaustive ancillary statistic $A = |\ln Y|$. We will find that the conditional value of the significance level varies greatly, even a test with a nominal level of 5% producing conditional significance levels as high as 100%, and that these unreasonable levels correspond to cases where the usual interpretation of the standard test is intuitively wrong.

Suppose $T \sim N(\mu, \sigma^2)$ (σ^2 known) and we test H: $\mu = \mu_1$ against K: $\mu = \mu_2$, then the optimal rejection rule for a significance level α is:

$$\text{Reject H when } \begin{cases} t \le \mu_1 - z_{1-\alpha} \cdot \sigma, & \mu_2 < \mu_1 \\ t \ge \mu_1 + z_{1-\alpha} \cdot \sigma, & \mu_2 > \mu_1. \end{cases}$$

However the *conditional* significance level of this test, which is a function of $a_0 = |\ln y_0| = |\ln LR(t_0)|$, will usually not equal α. In Cox's example, the result of a coin toss dictates which of two Normal populations (same mean, different variance) is sampled from. The unconditional test quotes a significance level that is the average of the two conditional significance levels associated with the different populations. The same is true here, except that our ancillary statistic, $A = |\ln Y|$, can take an infinite number of values rather than *two*. A population of Y, associated with each value of $a_0 = |\ln y_0|$, is described by the conditional distribution of Y given $A = a_0$, shown again below.

y	$y_1 = \min\{y_0, y_0^{-1}\}$	$y_1^{-1} = \max\{y_0, y_0^{-1}\}$
$\vec{P}_H(y)$	$\frac{y_1}{(1+y_1)}$	$\frac{1}{(1+y_1)}$
$\vec{P}_K(y)$	$\frac{1}{(1+y_1)}$	$\frac{y_1}{(1+y_1)}$

(Where $y_1 < 1 < y_1^{-1}$.)

This distribution can be written in terms of a_0 as:

Table 8.5

y	$y_1 = e^{-a_0}$	$y_1^{-1} = e^{+a_0}$
$\vec{P}_H(y)$	$\frac{e^{-a_0}}{(1+e^{-a_0})}$	$\frac{1}{(1+e^{-a_0})}$
$\vec{P}_K(y)$	$\frac{1}{(1+e^{-a_0})}$	$\frac{e^{-a_0}}{(1+e^{-a_0})}$

The conventional, α -level, rejection rule (above) can be written in terms of y, instead of t, as follows:

$$\text{Reject H when } y < \exp\{\tfrac{\delta}{2}(\delta - 2z_{1-\alpha})\} = CLR(\delta,\alpha), \text{ where } \delta = \tfrac{|\mu_1 - \mu_2|}{\sigma}.$$

The value $CLR(\delta,\alpha) = \exp\{\tfrac{\delta}{2}(\delta - 2z_{1-\alpha})\}$ is the *critical likelihood ratio* of the conventional α -level test and is dependent on μ_1, μ_2 and σ only through the value of δ; we will often abbreviated this to 'CLR' where its dependence on δ and α is understood. The CLR is the cut-off point for the rejection region in terms of the likelihood ratio statistic, Y, instead of the natural variable T, hence $\alpha = P_H(Y < CLR)$[9].

In order to calculate the conditional significance levels of the conventional test, we need to distinguish between three cases.

i. $CLR < y_1$ (i.e. $CLR < e^{-a_0}$)

ii. $y_1 \le CLR < y_1^{-1}$ (i.e. $e^{-a_0} \le CLR < e^{+a_0}$)

iii. $CLR \ge y_1^{-1}$ (i.e. $CLR \ge e^{+a_0}$).

The conditional significance level, α_{a_0}, is derived from the distribution of Y, conditional on a_0, by $\alpha_{a_0} = \vec{P}_H(Y \le CLR \mid A = a_0)$. From the conditional distribution of Y, we can see that the three cases, above, give rise to different conditional significance levels as follows.

i. $\alpha_{a_0} = 0$

ii. $\alpha_{a_0} = \dfrac{y_1}{(1+y_1)} = \dfrac{e^{-a_0}}{(1+e^{-a_0})}$

iii. $\alpha_{a_0} = 1$.

[9] All NP tests are inherently left-sided when written in terms of the likelihood ratio statistic, Y.

For any given test (fixed values of α and δ), the expression $\delta - 2z_{1-\alpha}$ is fixed and must be either *negative*, *zero*, or *positive*. In each of these instances, we can observe, at most, *two* of the above cases regarding the CLR since, for instance, if $\delta - 2z_{1-\alpha} < 0$ then $CLR < 1$ and this rules out case iii. We can derive the formulae for α_a, for all possible cases, as shown in the table below.

Table 8.6

A: $\delta < 2z_{1-\alpha} \Rightarrow CLR < 1$	B: $\delta = 2z_{1-\alpha} \Rightarrow CLR = 1$	C: $\delta > 2z_{1-\alpha} \Rightarrow CLR > 1$
$\alpha_a = \begin{cases} 0, & a < -\ln(CLR) \\ \dfrac{e^{-a}}{(1+e^{-a})}, & a \geq -\ln(CLR) \end{cases}$	$\alpha_a = \dfrac{e^{-a}}{(1+e^{-a})}, \ \forall a$	$\alpha_a = \begin{cases} 1, & a \leq \ln(CLR) \\ \dfrac{e^{-a}}{(1+e^{-a})}, & a > \ln(CLR) \end{cases}$

(Where $e^{-a} = y_1 = \min(y_0, y_0^{-1})$.)

The conventional significance level as an average.

In Cox's two-stage example it is easy to see that the nominal (i.e. unconditional) significance level of the optimal test is the average of the two conditional significance levels, which can also be thought of as the pre-experiment *expected value* of the conditional significance level, i.e. $E(\alpha_A)$. Our intuition is that the conditional significance level for the population that was actually observed is more relevant than the average value, which involves the population not observed. We will now show that the nominal level of the conventional z-test (α) is the average of all the conditional significance levels, weighted by the distribution of the ancillary statistic $A = |\ln Y|$, i.e. $\alpha = E(\alpha_A)$.

In order to find the expected value of α_a over all possible values of a, we need to find the density of the random variable A. Using the fact that $A = |\ln Y|$ it can easily be established that

$$F_A(a) = \Phi(\tfrac{\delta}{2} + \tfrac{a}{\delta}) - \Phi(\tfrac{\delta}{2} - \tfrac{a}{\delta}), \quad a > 0$$

and hence that

$$\begin{aligned} f_A(a) &= \tfrac{1}{\delta}[\phi(\tfrac{\delta}{2} + \tfrac{a}{\delta}) + \phi(\tfrac{\delta}{2} - \tfrac{a}{\delta})] \\ &= \tfrac{1}{\delta\sqrt{2\pi}}[\exp(-\tfrac{1}{2}(\tfrac{\delta}{2} + \tfrac{a}{\delta})^2) + \exp(-\tfrac{1}{2}(\tfrac{\delta}{2} - \tfrac{a}{\delta})^2)] \\ &= \tfrac{1}{\delta\sqrt{2\pi}}\exp(-\tfrac{1}{2}(\tfrac{\delta}{2} + \tfrac{a}{\delta})^2)[1 + \exp(a)]. \end{aligned}$$

The conventional significance level, α, is equal to the expected value of the random variable α_A, that is, $\alpha = E(\alpha_A) = \int_a \alpha_a \cdot f_A(a)da$. The proof of this is shown below for the case where $\delta > 2z_{1-\alpha}$. The proofs for the two other cases are similar.

Let $\delta > 2z_{1-\alpha}$, then

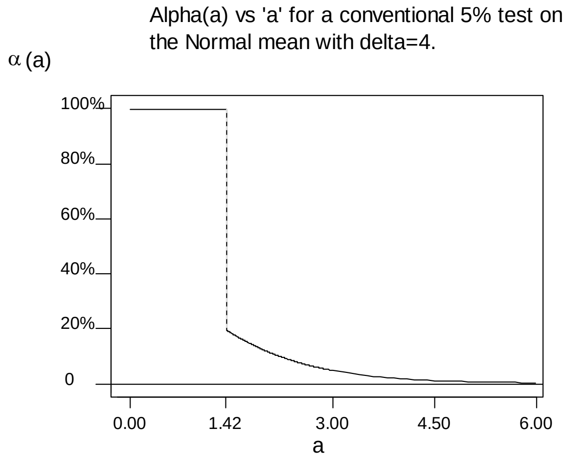
$$\begin{aligned} \int_a \alpha_a \cdot f_A(a)da &= \int_0^{[\ln(CLR)=]\delta(\delta-2z)/2} 1 \cdot f_A(a)da + \int_{\delta(\delta-2z)/2}^{\infty} \frac{e^{-a}}{(1+e^{-a})} f_A(a)da \\ &= [F_A(\delta(\delta-2z)/2)] + \int_{\delta(\delta-2z)/2}^{\infty} \frac{1}{(1+e^{a})} \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\delta}{2}+\frac{a}{\delta})^2}(1+e^{a})da \\ &= [\Phi(\delta - z) - \Phi(z)] + \int_{\delta(\delta-2z)/2}^{\infty} \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\delta}{2}+\frac{a}{\delta})^2} da \\ &= [\Phi(\delta - z) - \Phi(z)] + \int_{\delta-z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du \\ &= [\Phi(\delta - z) - \Phi(z)] + [1 - \Phi(\delta - z)] \\ &= 1 - \Phi(z_{1-\alpha}) \\ &= 1 - (1-\alpha) \\ &= \alpha. \end{aligned}$$

To illustrate this fact we will look at the conditional properties that apply to any standard 5% z-test with $\delta = 4$ (for example, $\mu_1 = 0,\ \mu_2 = 4,\ \sigma = 1$). The value of $\ln(CLR) = \frac{\delta}{2}(\delta - 2z_{1-\alpha}) = 2(4 - 2z_{0.95})$ is 1.42 and, hence, the conditional significance level is:

$$\alpha_a = \begin{cases} 1, & a \le 1.42 \\ \dfrac{e^{-a}}{(1+e^{-a})}, & a > 1.42. \end{cases}$$

The following plot shows α_a as a function of a.

Figure 8.4

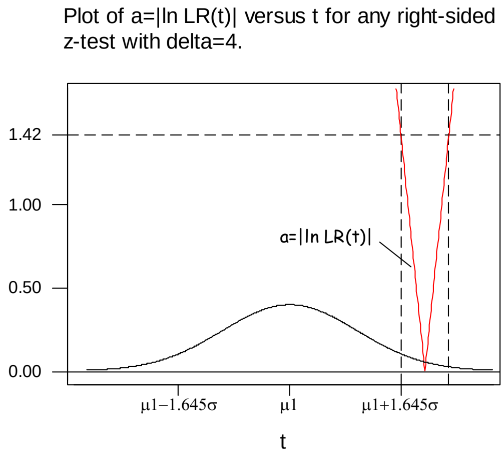


When a is large, the likelihood of the data under one hypothesis is much larger than that under the other hypothesis, and the conditional significance level becomes increasingly close to *zero* (while the conditional power becomes increasingly close to 100%) even though this is a 5% test; but when a is small the conditional significance level is much larger than the nominal 5%.

If $a < 1.42$, the conditional significance level is 100%. We can examine the reason for this in terms of the original Normal random variable, T.

Consider the right-sided test of H: $\mu = \mu_1$ versus K: $\mu = \mu_1 + 4\sigma$. The standard 5% rejection region for t is $[\mu_1 + 1.645\sigma, \infty)$. Consider the event $A \leq 1.42$, this is equivalent to $\mu_1 + 1.645\sigma \leq T \leq \mu_1 + 2.355\sigma$. If this ancillary event occurs, it is evident that the probability, under H, of rejecting H (the conditional significance level) is *one* since t must be in the rejection region $[\mu_1 + 1.645\sigma, \infty)$. Similarly, the conditional power of this test is *one*. Since this is true when we condition upon $A \leq 1.42$, it must also be true if we condition upon A being equal to any exact value in $(0,1.42]$ (since a significance level can never exceed *one*). These details can be seen in the following plot. Note that whenever $|\ln LR(t)|$ is less than 1.42, t is in the rejection region, i.e. t is greater than $\mu_1 + 1.645\sigma$.

Figure 8.5



Plot of a=|ln LR(t)| versus t for any right-sided z-test with delta=4.

This is similar to some of the anomalies that we encountered in the Uniform case. When the event $A \in (0,1.42]$ occurs, it tells us nothing about the question at issue because it has the same probability under both hypotheses, but, when it occurs, we will always reject H. The fact that the (conditional) significance level is *one* shows that we can read nothing into this; on the other hand, the unconditional significance level of 5% tempts us to interpret this as evidence against H relative to K. The unconditional significance level is an average value and, in this case, it is low only because of the contributions from those conditional significance levels associated with the unobserved cases where $A > 1.42$ (i.e. where the data is more informative than it is here). This can be further illustrated by considering the distribution of α_A.

Since α_A is a function of A, it is itself a random variable. The unconditional significance level of the conventional test remains α no matter what data we observe, whereas the conditional significance level is obtained by conditioning on a, which is a function of the data.

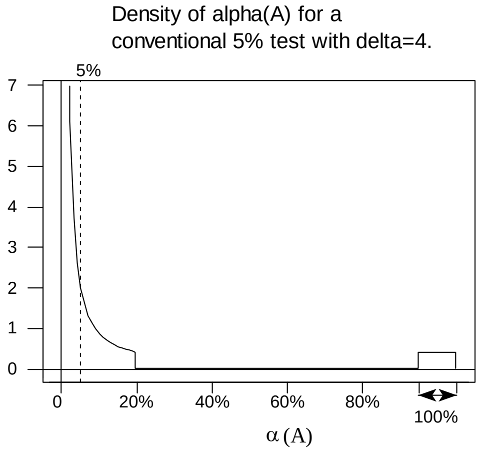
For the case $\delta > 2z_{1-\alpha}$, $P(\alpha_A = 1) = P(A < \frac{\delta}{2}(\delta - 2z)) = \Phi(\delta - z) - (1-\alpha) > 0$, thus the random variable α_A is partly discrete having a probability mass at the point ‘1’ but is continuous elsewhere with density:

$$f_{\alpha(A)}(u) = \frac{1}{u^2(1-u)\delta}\cdot\phi\left(\frac{\delta}{2}+\frac{\ln(u^{-1}-1)}{\delta}\right)$$
$$= \frac{1}{u^2(1-u)\delta\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\left(\frac{\delta}{2}+\frac{\ln(u^{-1}-1)}{\delta}\right)^2\right),\ u\in\left(0,\frac{1}{(1+CLR)}\right).$$

(NB: The density tends towards infinity as $\alpha_a \to 0$.)

In the plot below, we show the distribution of α_A for the case $\alpha = 5\%$, $\delta = 4$. (The probability mass at $\alpha_a = 100\%$ is shown by an area of that size so that the total area is *one*.)

Figure 8.6



The plot shows that the expected value of the random variable, α_A, is 5% – the nominal significance level of the test. This value is influenced by both the outlying probability mass at $\alpha_A = 100\%$ and the very high densities associated with values of α_A close to zero; in fact the probability that α_A is close to 5% is quite low. For any observed data, only one value of a will have occurred, yet the conditional significance levels associated with *every possible value* of a contribute to the conventional significance level. Since a contains no information about the question at issue, but indicates how informative our data is, it is wrong to allow the inference to be affected by values of a that did not occur. The test result should reflect the information available to us.

The conditional power of the conventional z-test.

Since the conditional distribution of Y given $A = a_0 = |\ln y_0|$ is as below (**Table 8.2** shown again):

y	$y_1 = \min\{y_0, y_0^{-1}\}$	$y_1^{-1} = \max\{y_0, y_0^{-1}\}$
$\vec{P}_H(y)$	$\frac{y_1}{(1+y_1)}$	$\frac{1}{(1+y_1)}$
$\vec{P}_K(y)$	$\frac{1}{(1+y_1)}$	$\frac{y_1}{(1+y_1)}$

and the conventional, α-level, rejection rule is:

$$\text{Reject H when } y < \exp\{\tfrac{\delta}{2}(\delta - 2z_{1-\alpha})\} = CLR(\alpha,\delta), \text{ where } \delta = \tfrac{|\mu_1-\mu_2|}{\sigma},$$

we need to distinguish between the same three cases (below) in order to calculate the conditional power.

i. $CLR < y_1$ (i.e. $CLR < e^{-a_0}$)

ii. $y_1 \le CLR < y_1^{-1}$ (i.e. $e^{-a_0} \le CLR < e^{+a_0}$)

iii. $CLR \ge y_1^{-1}$ (i.e. $CLR \ge e^{+a_0}$).

The conditional power, $\kappa_{a_0} = 1 - \beta_{a_0}$ (where β_a is the probability of Type II error conditional upon $A = a$), is derived from the distribution of Y, conditional on a_0, by $\kappa_{a_0} = \vec{P}_K(Y \le CLR \mid A = a_0)$. From the conditional distribution of Y, we can see that the three cases, above, give rise to different conditional power levels as follows.

i. $\kappa_{a_0} = 0$

ii. $\kappa_{a_0} = \dfrac{1}{\left(1+y_1\right)} = \dfrac{1}{(1+e^{-a_0})}$

iii. $\kappa_{a_0} = 1$.

Equivalently,

i. $\beta_{a_0} = 1$

ii. $\beta_{a_0} = \dfrac{y_1}{\left(1+y_1\right)} = \dfrac{e^{-a_0}}{(1+e^{-a_0})}$

iii. $\beta_{a_0} = 0$.

Thus, if the data we observe is such that $y_1 \le CLR < y_1^{-1}$ (i.e. $e^{-a_0} \le CLR < e^{+a_0}$), it follows that the conditional probability of Type II error is the same as the conditional probability of Type I error even when the unconditional error probabilities were not the same.

The following table shows the conditional power results for any given test (fixed values of α and δ).

Table 8.7

A: $\delta < 2z_{1-\alpha} \Rightarrow CLR < 1$	B: $\delta = 2z_{1-\alpha} \Rightarrow CLR = 1$	C: $\delta > 2z_{1-\alpha} \Rightarrow CLR > 1$
$\kappa_a = \begin{cases} 0, & a < -\ln(CLR) \\ \dfrac{1}{(1+e^{-a})}, & a \ge -\ln(CLR) \end{cases}$	$\kappa_a = \dfrac{1}{(1+e^{-a})}, \ \forall a$	$\kappa_a = \begin{cases} 1, & a \le \ln(CLR) \\ \dfrac{1}{(1+e^{-a})}, & a > \ln(CLR) \end{cases}$

These results are summarised in the table below, which gives the conditional error probabilities of both kinds for any fixed α-level z-test.

Table 8.8

	α_a	β_a
(i) $\delta < 2z_{1-\alpha}$ & $a < -\ln(CLR)$ [Equivalent to: CLR<1 & $a <\| \ln(CLR) \|$.]	0	1
(ii) $\delta > 2z_{1-\alpha}$ & $a < \ln(CLR)$ [Equivalent to: CLR>1 & $a <\| \ln(CLR) \|$.]	1	0
(iii) Otherwise [i.e. $a >\| \ln(CLR) \|$].	$\frac{e^{-a}}{1+e^{-a}}$	$\frac{e^{-a}}{1+e^{-a}}$

The error probabilities are non-trivial, and the test meaningful, only when $a > | \ln CLR |$. Since large values of a indicate that we have more informative data, this makes sense. Also note that $\frac{e^{-a}}{1+e^{-a}} = \frac{1}{1+e^{a}}$ is a decreasing function of a, and hence the conditional error probabilities become smaller and the test more reliable as a increases. The two extreme cases, (i) and (ii), occur when $a = | \ln y | < | \ln CLR |$. In terms of the Normal variate, t, this is equivalent to the requirement: $| t - \bar{\mu} | < d = z_{1-\alpha} \cdot \sigma$; thus, t is 'close' to $\bar{\mu}$ (the point half-way between the two values) indicating 'weak' data, and the ancillary set $\{t : | t - \bar{\mu} | < d\}$ is *either* wholly within the α -level rejection region (case (ii) when the hypotheses are relatively far apart so that all the weakest observations are in the rejection region) *or* wholly outside the rejection region (case (i) when the hypotheses are close together so that all the weakest data is outside the rejection region). The conditional error probabilities give a more accurate account of the reliability of the test result than the conventional error probabilities.

Conditional p-values and conventional p-values.

In this section, we show that for tests on the Normal mean (σ known or accurately estimable) the p-value, conditional upon the observed value of the ancillary statistic $A = | \ln Y |$, is greater than the conventional p-value no matter what the data, hypotheses, or value of σ.

Let $T \sim N(\mu, \sigma^2)$ where σ^2 is known. Let the data be $t \in \mathbb{R}$ and the binary parameter space $\Theta_B \equiv \{\mu_1, \mu_2\} \in \mathbb{R}^2$ $(\mu_1 \neq \mu_2)$, and $\sigma \in \mathbb{R}^+$; H states that $\mu = \mu_1$ while K states that $\mu = \mu_2$.

Then, $y = LR(t) = \exp\{\frac{1}{2\sigma^2}[(t-\mu_2)^2 - (t-\mu_1)^2]\} \in \mathbb{R}^+$.

The conventional p-value is:

$$\text{p-value}(t) = \begin{cases} \Phi(\frac{t-\mu_1}{\sigma}), & \mu_2 < \mu_1 \\ 1-\Phi(\frac{t-\mu_1}{\sigma}), & \mu_2 > \mu_1. \end{cases}$$

This can be written in terms of y as:

$$p_\delta(y) = \Phi(\ln y \cdot \delta^{-1} - \delta / 2),$$

dependent on $\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$.

Denote by '$cp(y)$' the p-value of y conditional upon $A = a = |\ln y|$, i.e. $cp(y) = \vec{P}_H(Y \leq y \mid A = |\ln y|)$. From the conditional distribution of Y given $A = a = |\ln y|$ (see **Table 8.2**), it follows that:

$$cp(y) = \begin{cases} 100\%, & y \geq 1 \\ \dfrac{y}{(1+y)}, & y < 1. \end{cases}$$

When $y \geq 1$[10] (i.e. t is closer to μ_1 than to μ_2), $cp(y) = 100\%$ since the observed value of Y is the larger of the two possible values consistent with observing $A = a$. On the other hand, ($\forall \delta$) the conventional p-value, $p_\delta(y)$, goes to 100% only as $y \to \infty$, hence $cp(y) > p_\delta(y)$ whenever $y > 1$ (finite). In order to show that this is also true when $y < 1$, we need to take into account the different possible values of δ.

[10] When $y = 1$, $y^{-1} = y$ thus the conditional distribution contains only the observed value and (conditionally) $\vec{P}(Y \leq y) = \vec{P}(Y = y) = 1$. Where Y is a continuous variable (unconditionally), there is no need to distinguish between $y < 1$ and $y \leq 1$ since $P(Y = 1) = 0$.

First, fix y as any arbitrary value less than *one* and regard $p_\delta(y)$ as $g(\delta)$: a function of $\delta \in \mathbb{R}^+$.

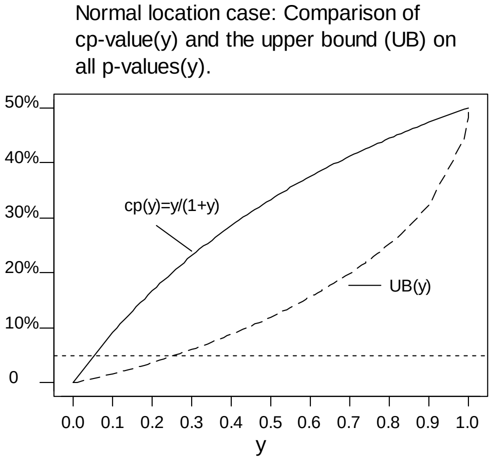
$$\frac{d}{d\delta} g(\delta) = -\phi(\ln y \cdot \delta^{-1} - \delta/2) \times (\ln y \cdot \delta^{-2} + \tfrac{1}{2})$$
$$= 0, \text{ if and only if } \delta = \delta_0 = \sqrt{-2\ln y}.$$

It can easily be shown that $\frac{d^2}{d\delta^2} g(\delta)|_{\delta=\delta_0} = \frac{-\phi(-\delta_0)}{\delta_0} < 0$ and hence g has a unique maximum at $\delta = \delta_0$, i.e. $\max_\delta g(\delta) = g(\sqrt{-2\ln y}) = \Phi(-\sqrt{-2\ln y})$.

Thus, for any $y < 1$, $p_\delta(y) \le \Phi(-\sqrt{-2\ln y})$ ($\forall \delta$); we may call $\Phi(-\sqrt{-2\ln y})$ $'\mathrm{UB}(y)'$ to indicate that it is, for all δ, an *upper bound* on the conventional p-value, $p(y)$.

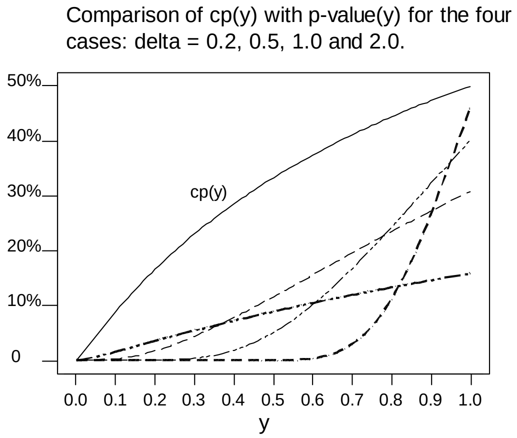
When $y < 1$, the conditional p-value is $cp(y) = y/(1+y)$. From the following plot it is clear that $\mathrm{UB}(y) < cp(y), \forall y < 1$, and hence it follows that $p_\delta(y) < cp(y), \forall y, \delta$.

Figure 8.7



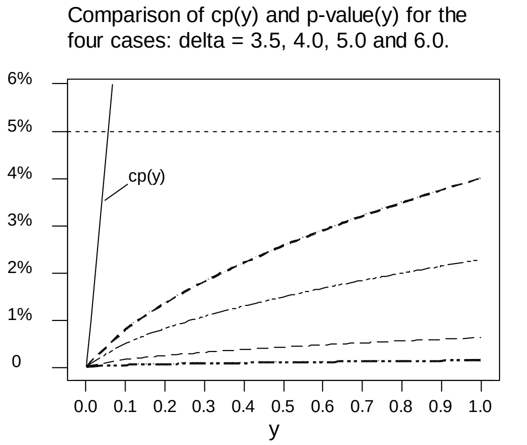
Note that $cp(y) \le 5\%$ only when $y \le \frac{1}{19} \approx 0.053$ whereas the upper bound has values under 5% for a wider range of y-values. Since it is an *upper* bound, the p-values themselves are generally lower and thus significant for even more values of y. Unlike the cp-value, the p-value of y varies with the value of δ; the following plot shows $p_\delta(y)$ for several values of δ in the range 0.2 to 2.0.

Figure 8.8



We noted earlier that there is a particular tendency for conventional tests to reject H unreasonably when the two hypotheses are far apart, i.e. when δ is large. The following plot illustrates this point.

Figure 9.9



Since $p_\delta(y) \to \Phi(\frac{-\delta}{2})$ as $y \to 1$, and $\Phi(\frac{-\delta}{2}) \to 0$ as $\delta \to \infty$, it follows that y can be close to *one* and still have a small p-value when δ is large; in the plot above, the p-values of $y = 0.99999$ are less than 5% in all four cases. From a likelihood point of view, such data is almost exactly neutral regarding the two hypotheses; the *conditional* p-value ($\frac{0.99999}{1.99999} = 49.99975\%$) is consistent with this view unlike the conventional p-values. When we condition on $A = |\ln(0.99999)|$, we reduce the sample space to one containing only two outcomes, $y = 0.99999$ and $y = 0.99999^{-1}$, both of which are in almost equal agreement with the two hypotheses; in such a situation, the failure rate of any rule allowing us to reject H is high and the conditional p-value reflects this fact.

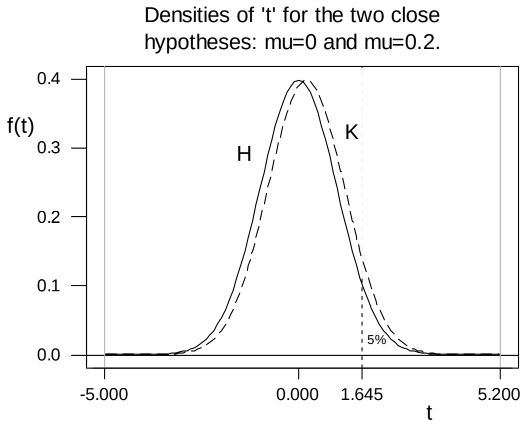
Close hypotheses.

When the hypotheses are far apart the conventional method gives strange results, rejecting H when the LR is very large and the data is far more consistent with H than with K. This is especially disturbing since these tests have very high power and so appear to be reliable. However, there are also problems when the hypotheses are close together. In such a case there is very little difference between the distribution of T under the two hypotheses and, thus, observing t is not a good basis for choosing between the hypotheses – no data amounts to strong evidence one way or the other. For some models, the likelihood ratio is strictly bounded in such circumstances. For the Normal location model, the likelihood ratio theoretically takes values in the range $(0, \infty)$ in all cases, however if we put realistic bounds on t, these will flow through to the likelihood ratio.

Example 8.1.

Consider $T \sim N(\mu, 1)$ and suppose we want to test H: $\mu = 0$ against K: $\mu = 0.2$. The hypothesised values are one-fifth of a standard deviation apart.

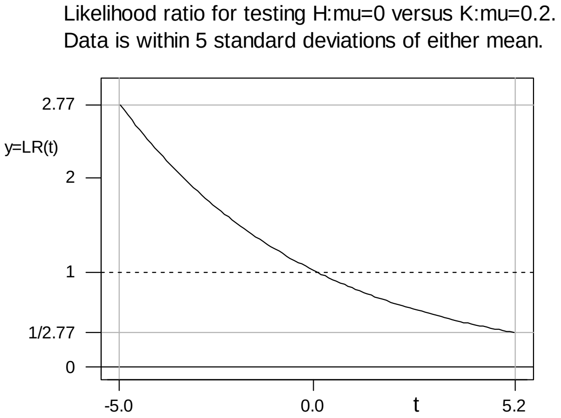
Figure 8.10



Consider all possible data lying within *five* standard deviations of either of the hypothesised means, i.e. $t \in [-5.0, +5.2]$. Clearly the two distributions are very alike; T behaves in much the same way under either hypothesis. The poor design of this test is reflected in the low power; however, this does not prevent us from observing data (in this range) that has a very small p-value. Any $t > 1.645$ has a p-value less than 5%. Thus we may easily reject H in favour of K.

The likelihood ratio bears out our intuition that none of this data is convincing in either direction.

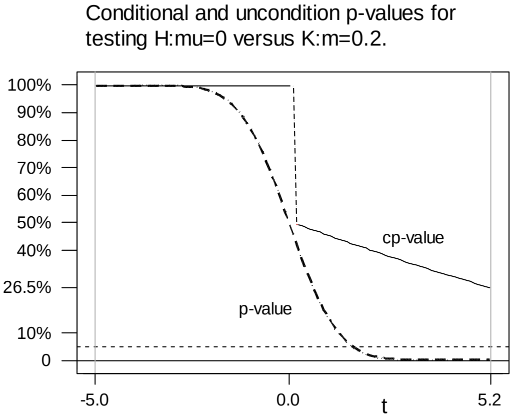
Figure 8.11



The likelihood ratio lies within the range $[\frac{1}{2.77}, 2.77]$ – none of the data in this range favours one hypothesis over the other to the extent that the outcome *hh* favours the double-headed hypothesis over the fair coin hypothesis. It seems wrong that we can observe p-values that are very close to *zero* in such a case.

Compare the p-value with the cp-value. When $y < 1$ (i.e. $t > 0.1$) the cp-value is $\frac{y}{(1+y)}$. Since y is never less than $\frac{1}{2.77}$, the cp-value is similarly bounded below and is greater than 26.5% for all data in this range (see below).

Figure 8.12



The conventional test is quite inadequate at responding to the challenges of this situation. The low power indicates that the experiment is not very likely to produce data that rejects H in favour of K, even if K is true. It follows that we can infer nothing from a failure to reject H. However this fact becomes irrelevant (except as a criticism of the design of the experiment) if we do, in fact, reject H, and we are able to do so surprisingly easily, based on data that is close to neutral for these hypotheses ($LR \leq \frac{1}{1.362}$, for p-value$\leq 5\%$).

An examination of standard tests on composite hypotheses.

In Chapter 3, we looked at tests of a simple null hypothesis and a composite alternative hypothesis, and asked whether rejection of the null hypothesis necessarily implies that the data is much more consistent with *some* component of the composite alternative than with the null hypothesis. According to likelihood theory, the answer is *no* (see Chapter 7). What answer does our conditional approach produce? Consider the following case (modified from ***Example 3.1***).

Example 8.2.

$X_1,\ldots,X_{16}$ are independent and identically distributed $N(\mu,8^2)$ variables. We want to test H: $\mu=163$ versus K: $\mu>163$, with data $\overline{x}=166.3$. Since $\overline{X}\sim N(\mu,2^2)$, this data has a conventional p-value of $1-\Phi(\frac{166.3-163}{2})\approx 4.95\%$, and we can reject H at the 5% level.

Although this is significant, it is obvious that the data is more consistent with the hypothesis H than with (say) the hypothesis $\mu=180$, which is a component of the composite hypothesis K. Clearly the strength of evidence against H relative to any given component of K is very varied. Is there some component of K that the evidence in the data strongly favours relative to H?

Since this is a Normal location test, we can use the conditional results for log-symmetric models. The values of $\mu_1=163$ and $\sigma=8$ are fixed, as is $\overline{x}=166.3$ (called t in the general theory). When we consider different components of K, we are allowing μ_2 to vary within the domain $(163,\infty)$. Since $cp(y)=\frac{y}{(1+y)}$ is a monotone increasing function of the likelihood ratio (y), it is smallest and most ‘significant’ when the LR is smallest. The LR varies depending on the value of μ_2 and it is easy to show that it takes the smallest value when $\mu_2=\overline{x}=166.3$. When $\mu_2=166.3$, the LR value of the data is $y=0.25634$. This value is greater than ¼ so Royall would not regard it as significant – the evidence is less strong than that from data *hh* in the paradigm coin tossing case. The cp-value is $\frac{0.25634}{1.25634} = 20.4\%$ which is far from significant. A test of H against any other component of K will produce a larger cp-value (and LR) than the one calculated here. On the basis of either the LR or the cp-value, we can say that the evidence from the data $\bar{x} = 166.3$ does not justify rejecting H: $\mu = 163$ in favour of *any* hypothesis nominating a larger value of μ; in particular, all such tests will produce a cp-value of at least 20.4%.

If the data were further away from μ_1, the result would be different. How far away must it be in order that the cp-value, relative to *some* component of K, is significant?

Let $T \sim N(\mu, \sigma^2)$, and we observe data t. Whenever t is on the side of μ_1 *other* than that defined by K, it follows that, for all components of K, $y = LR(t) > 1$ and $cp(y) = 100\%$; only when this is not the case is there any possibility of obtaining conditionally significant results.

In view of this, let H: $\mu = \mu_1$ and let K be defined as follows:

$$\text{K:}\begin{cases} \mu > \mu_1, \text{ if } t > \mu_1 \\ \mu < \mu_1, \text{ if } t < \mu_1. \end{cases}$$

Over the components of K, the minimum LR occurs at $\mu_2 = t$, hence

$$y_{\min} = \min_{K_i} LR(t) = \frac{f(t;\mu_1)}{f(t;t)} = \exp\left\{\frac{-(t-\mu_1)^2}{2\sigma^2}\right\}.$$

For a test of H: $\mu = \mu_1$ against K$_t$: $\mu = t$, the cp-value is smallest and is equal to

$$\min_{K_i} \text{ cp-value}(t) = cp(y_{\min}) = \left[\exp\left\{\frac{(t-\mu_1)^2}{2\sigma^2}\right\}+1\right]^{-1}.$$

The cp-value is only significant for any component of K if it is significant for the particular component K$_t$ and for this to be true we need to have

$$t: \left[\exp\left\{\frac{(t-\mu_1)^2}{2\sigma^2}\right\}+1\right]^{-1} \leq \gamma,$$

where γ is some appropriately small value such as 5%, defining a significantly small cp-value.

This inequality holds only when

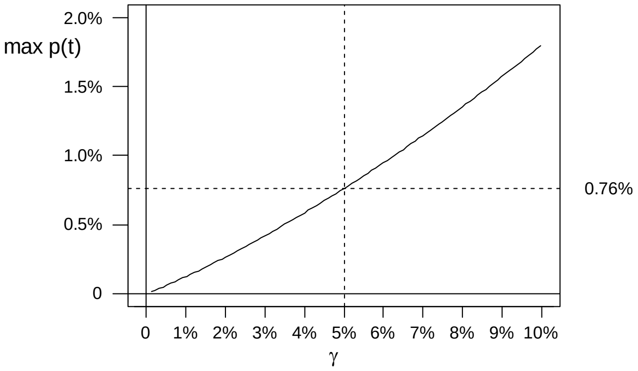
$$\frac{|t-\mu_1|}{\sigma} \geq \sqrt{2\ln(\tfrac{1-\gamma}{\gamma})}$$

(subject to t being on the side of μ_1 specified by K). The conventional p-value of t can be written as: $\text{p-value}(t) = \Phi\left(\frac{-|t-\mu_1|}{\sigma}\right)$, and hence the above inequation can be written as:

$$-\Phi^{-1}(\text{p-value}(t)) \geq \sqrt{2\ln(\tfrac{1-\gamma}{\gamma})}$$
$$\Rightarrow \text{p-value}(t) \leq \Phi\left(-\sqrt{2\ln(\tfrac{1-\gamma}{\gamma})}\right).$$

The following plot shows this relationship; '$\max p(t)$' denotes the right hand side of the inequality.

Figure 8.13

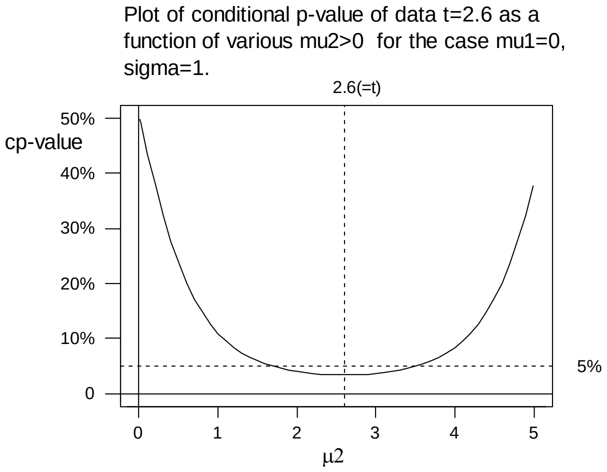


Only when the conventional p-value is less than 0.76%, is the cp-value under 5% for the test of H versus K$_t$; it follows that whenever $\text{p-value}(t) > 0.76\%$, there is no component of K for which the cp-value of the data is less than 5%.

This fact also gives us a general result for tests of two simple hypotheses on the Normal mean. Whenever the conventional p-value of an observation is greater than 0.76%, it follows that no conditional test of the null hypothesis against *any* simple alternative hypothesis will deliver a cp-value less than 5%. (More generally, whenever the conventional p-value of an observation is greater than $\Phi\left(-\sqrt{2\ln(\frac{1-\gamma}{\gamma})}\right)$, it follows that no conditional test of the null hypothesis against any simple alternative hypothesis will deliver a cp-value less than γ.)

Even when the p-value is less than 0.76%, it is still misleading to quote the result of a test with a composite alternative as, for instance, '*Reject H:* $\mu = 0$ *in favour of K:* $\mu > 0$' since this does not identify which of the components of K produce significant results and which do not. If $T \sim N(\mu,1)$ and we want to test H: $\mu = 0$ against K: $\mu > 0$, the data $t = 2.6$ has a conventional p-value of 0.466%. Since this is less than 0.76%, we know that the cp-value will be less than 5% for some components of K. The plot below shows the cp-values for alternatives in the range (0,5].

Figure 8.14



Against the alternative $\mu_2 = 2.6$, the data has a cp-value of 3.3%, and for any alternative value of μ in the range (1.6, 3.6) the cp-value is 5% or less, but for some values outside this range the cp-value is very high. Any reasonable summary of the significance of the observation $t = 2.6$ to the null hypothesis should distinguish between these cases.

8.7 Creating hypothesis tests with good conditional properties.

We have examined the conventional z-test from a conditional point of view, using the statistic $A = |\ln Y|$ which is ancillary, in the restricted sense, on the general binary parameter space (μ_1, μ_2). Although the z-test is widely used, and generally considered unproblematic, our conditional analyses reveal that the conditional error probabilities of (say) a 5% test can be appallingly high, and these cases correspond to those where tests of simple hypotheses produce intuitively unpalatable results (see Chapter 3). It is clear that A is an effective precision index and can be used to discriminate between more and less informative data.

In this section, we use A to construct a new test that has acceptable conditional properties; this approach is applicable to all log-symmetric scenarios. We then compare the structure of this test, and its associated confidence intervals, with that of the standard z-test and z-intervals. We find that the new test possesses some striking features more usually associated with a non-frequentist approach and produces results that are much more intuitively reasonable than the standard results.

A bound on the conditional significance level.

For any fixed test criterion (rejection region), the conditional significance level (and power) of the test depends on the value of $A = |\ln Y|$ observed, thus the relevant conditional significance level of the test can not be known until after the experiment has been performed, in contrast to the unconditional (i.e. average) error probabilities. Since the conditional distributions are discrete, we cannot make all the conditional significance levels equal without introducing a randomising variable and breaching the *sufficiency principle*. However, we can, if we wish, define a rejection rule that puts an upper bound, $\bar{\alpha} < 100\%$, on all the possible values of α_a.

Consider a rule of the form: *reject H in favour of K when* $y \le k$. First note that if $k > 1$, there will be some observable a such that $\alpha_a = 100\%$ because both values of y (y_1 and y_1^{-1}) associated with a are in the rejection region (since $1 < y_1^{-1} < k$); thus we must choose $k < 1$. If neither value of y, associated with a, is in the rejection region, then $\alpha_a = 0 < \bar{\alpha}$ and this is acceptable. There remain the cases where the smaller value of y ($y_1 = e^{-a} < 1$) is in the rejection region and the larger value is not. In such a case the conditional significance level is $\alpha_a = \frac{y_1}{(1+y_1)}$ and this is less than or equal to $\bar{\alpha}$ if and only if $y_1 \le \frac{\bar{\alpha}}{1-\bar{\alpha}}$.

Hence, the rule

$$\boxed{\text{Reject H in favour of K when } y \le k = \tfrac{\bar{\alpha}}{1-\bar{\alpha}}}$$

has conditional significance levels (α_a) that are less than or equal to $\bar{\alpha}$ for all a.

If we want to ensure that the relevant (i.e. conditional) significance level cannot exceed (say) 5%, then the rule *reject H in favour of K when* $y \le \frac{1}{19}$ is appropriate. Since the rule rejects H when the likelihood ratio is smaller than a certain value, the test produces the highest conditional power that can be associated with the observed conditional significance level $(\forall a)$, i.e. Neyman-Pearson optimality applies conditionally (as Welch pointed out in 1939).

The relevant significance level.

Even though we may define a rule determined by an upper bound on the conditional significance levels, it is important to recognise that it is the conditional significance level *for the observed value of* a (say, a_0) that is the relevant Type I error probability for the test. This is because α_{a_0} is the error probability for, as it were, 'the machine that was actually used' in the experiment. The value, $\bar{\alpha}$, allows us to identify significant values of y before the experiment is performed, but, for making an after-experiment inference, it is the value α_a that is important. For example, if the experiment produces data, $\underset{\sim}{x}_0 : y_0 = LR(\underset{\sim}{x}_0) = \frac{1}{32}$, it is immediately obvious that we can reject H using $\bar{\alpha} = 5\%$, since $\frac{1}{32} < \frac{1}{19}$, and so $\alpha_{a_0} < 5\%$. From the conditional distribution of Y *given* $A = |\ln y_0|$ (shown below), it is clear that the relevant significance level of the test is $\alpha_{a_0} = \frac{1}{33} = 3.03\%$ where $a_0 = |\ln \frac{1}{32}|$ was the observed value of A.

Distribution of Y given that $A = |\ln \frac{1}{32}|$.

Table 8.9

y	$\frac{1}{32}$	32
$\vec{P}_H(Y = y)$	$\frac{1}{33}$	$\frac{32}{33}$

$$\Rightarrow \alpha_{a_0} = \vec{P}_H(Y \le \tfrac{1}{19} \mid A = a_0) = \vec{P}_H(Y = \tfrac{1}{32} \mid A = |\ln \tfrac{1}{32}|) = \tfrac{1}{33}.$$

(The distinction between α_{a_0} and $\bar{\alpha}$ ought not to be identified with the distinction between the p-value(x) and α, in conventional tests. In this case α_{a_0} is a genuine significance level, as discussed below.)

The relevant significance level and the cp-value.

What is the relationship between the relevant significance level and the cp-value of the data? Like the cp-value, the relevant significance level, α_{a_0}, is a function of the data; in this it differs from the unconditional α which depends only on the rejection rule. Unlike the cp-value, it depends on the data only through the ancillary statistic, A, that is, α_{a_0} depends on the rejection rule and on which ‘machine’ was used in the sub-experiment (the latter outcome being part of the experimental result). However, when $A = |\ln Y|$, the following relation holds:

Whenever we use a test of the form 'Reject H when $y \le \frac{\bar{\alpha}}{1-\bar{\alpha}}$'
and the data is such that $cp(y_0) \le \bar{\alpha} < 1$, then $cp(y_0) = \alpha_{a_0}$.

In other words, if the observed value, y_0, is significant, relative to some rule that precludes conditional significance levels of 100%, then the distinction between the conditional p-value of the data and the conditional (relevant[11]) significance level of the test vanishes – they are the same. This relation holds because of the exhaustive nature of the ancillary statistic, $A = |\ln Y|$, specifically, the fact that A partitions the sample space of Y into sets each containing only two values of Y ($y_1(a)$ and $[y_1(a)]^{-1}$) either side of *one*. Cox's conditional example[12] (for instance) does not possess this property.

Using the relevant significance level to rank the strength of evidence.

In unconditional inference, the p-value of the data is sometimes preferred to the accept/reject result (based on the pre-determined significance level) because it interprets the data in a more refined way. For example, when $\sigma = 1$, a 5% right-sided z-test of H: $\mu = 0$, will reject H whenever $t > 1.645$. The results $t = 3$ and $t = 5$ both lead us to reject H at this level, however, the latter result seems like stronger evidence against H than the former – a fact reflected by their p-values (and, for this reason, the p-value is sometimes vaguely interpreted as a measure of evidence of some kind). On the other hand, the p-value cannot be interpreted as a long-run success rate associated (purely) with the method, since it depends on the data that was observed in a particular performance of the experiment, unlike the figure '5%'. Contrast this with the relevant significance level. Consider the case H: $\mu = 0$ versus K: $\mu = 5$ ($\sigma = 1$) where $t = 5$ appears to be stronger evidence against H than $t = 3$. If we let $\bar{\alpha} = 10\%$, then we will reject H when the likelihood ratio, y, is less than 0.111. The likelihood ratios of the observations are: $y_0 = 0.082$, when $t_0 = 3$, and $y_0 = 3.73 \times 10^{-6}$, when $t_0 = 5$. Thus both observations would lead us to reject H based on $\bar{\alpha} = 10\%$. However, the two observations are associated with different values of $a_0 = |\ln y_0|$ (i.e.

[11] 'Relevant' meaning that the value of A conditioned upon is that actually observed in the experiment.
[12] Cox (1958).

different ‘machines’) and hence the relevant significance level of the test is different when $t = 3$ than when $t = 5$. Since both values are in the rejection region we know that the respective significance levels are equivalent to the cp-values where $cp(y_0) = \frac{y_0}{(1+y_0)}$, hence when $t_0 = 3$ ($a_0 = 2.5$), $\alpha_{a_0} = 7.59\%$, and when $t_0 = 5$ ($a_0 = 12.5$), $\alpha_{a_0} = (3.7 \times 10^{-4})\%$. The value, α_{a_0}, indicates the long-run error rate for rejecting H, when H is true, assuming that we use the same rejection rule and the value of A is the same as that observed here. This can be thought of as the Type I error probability associated with the machine we actually used and it depends on the data only in so far as we count the choice of machine as part of the data. The two values $t = 3$ and $t = 5$ are associated with different values of α_{a_0} because they were produced by different machines; specifically, $t = 5$ was produced by a machine that is better at distinguishing between H and K and produces smaller (non-trivial) error probabilities than the machine that produced $t = 3$. This allows us to interpret the difference between the significance of $t = 3$ and $t = 5$ by reference to genuine error probabilities and we can see that $t = 5$ is indeed more significant, at least, in the sense of being a significant outcome associated with a more rigorous test.

So far we are in agreement with the conventional view that $t = 5$ is the more significant of the two results[13], however note that the values quoted for the relevant significance levels are substantially different from the conventional p-values.

Table 8.10

t_0	p-value (t_0)	α_{a_0}
3	0.10%	7.59%
5	$(2.87 \times 10^{-5})\%$	$(3.7 \times 10^{-4})\%$

Although the conventional p-value ranks the data in the right order (in terms of strength of evidence) it does not accurately assess what that strength is. This is because the p-value is the mean of a number of values, most of which are irrelevant *and smaller* than the cp-value (see §8.11). When we make proper allowance for the

[13] This is also consistent with a likelihood interpretation since $LR(5) < LR(3)$.

ancillary set in which the data, $t = 3$, lies, it is apparent that the rejection rule we are using has a failure rate (H being true) of over 7%, whereas (for all $\mu_2 > 0$) the conventional p-value of the data is 0.1%. The job of ranking the data, for which the p-value is valued, is much better performed by the relevant conditional significance level.

The cp-value and the likelihood ratio.

In earlier chapters we noted that in conventional tests, including z-tests, a small significance level does not imply that the critical likelihood ratio (CLR) of the test is small, nor does a small p-value imply that the likelihood ratio of the data is small. A conventional significance level, such as 5%, may produce a CLR anywhere in $\mathbb{R}^+$, depending on the model and hypotheses. Similarly, an observation may have a large likelihood ratio, even much greater than *one*, and still have a significantly small p-value. The only general connection between the likelihood ratio and significance measures is the following (where H and K are simple hypotheses).

$$\boxed{\text{With respect to any null hypothesis, H, p-value}(x) < LR(x) = \tfrac{p_H(x)}{P_K(x)},\ \forall\text{K.}}$$

This ensures that, as long as the LR is sufficiently small (for some K), the conventional p-value will be significantly small, but not vice versa. A small p-value is a necessary but not a sufficient condition for the existence of a small likelihood ratio (for some alternative hypothesis).

However, when we look at the conditional approach using the ancillary statistic $|\ln Y|$, the relationship changes to the following.

$$\boxed{\text{For all H, K and } x\text{, cp-value}(x) = \begin{cases} \dfrac{LR(x)}{(1+LR(x))}, & LR(x) < 1 \\ 100\%, & LR(x) \geq 1. \end{cases}}$$

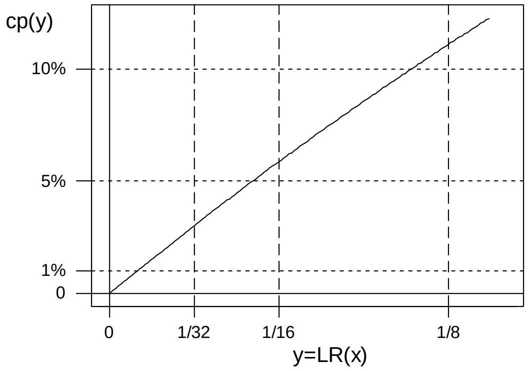
Thus, the cp-value is a function of the likelihood ratio of the data and this function is the same across all model/hypothesis combinations that are log-symmetric. In particular it holds for all tests on the Normal mean (σ^2 known) regardless of the hypotheses or variance. Moreover, this relationship is one-to-one over those values where the $LR(x) \in [0,1)$. Within the log-symmetric class of models/hypotheses, any two observations with the same likelihood ratio will also have the same cp-value. This is a localised version of the likelihood principle[14]. If an experiment, E_1, produces data x_1 for testing H_1 versus K_1, and, an experiment, E_2, produces data x_2 for testing H_2 versus K_2, and $LR_1(x_1) = LR_2(x_2)$, then the LP states that x_1 is evidence for H_1 relative to K_1 to exactly the same degree that x_2 is evidence for H_2 relative to K_2. We have shown that, in such a case, as long as $(E_1, \mathrm{H}_1, \mathrm{K}_1)$ and $(E_2, \mathrm{H}_2, \mathrm{K}_2)$ are both members of the log-symmetric class, the $\text{cp-value}_1(x_1)$ and $\text{cp-value}_2(x_2)$ are the same and the conditional interpretation of the two results with be the same. The converse is also true as long as the $LR_i(x_i) < 1$ $(i = 1,2)$, i.e. identical cp-values imply identical likelihood ratios.

Any reasonable (i.e. $\bar{\alpha} < 1$) construction of a rejection region in terms of the cp-value(x) can be re-written in terms of $LR(x)$, and this translation is the same in all log-symmetric cases. For instance, the regions $\{\underset{\sim}{x} : \text{ cp-value}(\underset{\sim}{x}) \le \bar{\alpha}\}$ and $\{\underset{\sim}{x} : LR(\underset{\sim}{x}) \le \frac{\bar{\alpha}}{1-\bar{\alpha}}\}$ are equivalent to each other. Thus, in the log-symmetric scenario, any $\bar{\alpha}$ bounded conditional test ($\bar{\alpha} < 1$) is equivalent to the dichotomous likelihood test where $\frac{1}{\lambda} = \frac{\bar{\alpha}}{1-\bar{\alpha}}$. Furthermore, cp-values in the significant range roughly correspond to likelihood ratios also in the significant range (see plot below).
There is no conflict between the interpretation of y directly (as a likelihood ratio) and through $cp(y)$.

[14] The real LP is not restricted to any locale – that is a big part of its force. However, the result given here still represents a big change by comparison with conventional z-tests, which are only likelihood-consistent across a class of cases where $\frac{|\mu_1 - \mu_2|}{\sigma}$ is constant. In contrast, our tests are consistent across all Normal location scenarios and even beyond.

Figure 8.15

Log-symmetric case. A plot showing the relationship between the conditional p-value and the likelihood ratio statistic.



The fact that only data with a likelihood ratio of less than *one* can possibly produce a significant result for rejecting H in favour of K, and that only data with a likelihood ratio[15] of more than *one* can possibly cause us to reject K (as null hypothesis) in favour of H (as alternative), can be seen as agreeing with that part of the Law of Likelihood which states that, when the likelihood ratio is *one*, the evidence is neutral between the two hypotheses. This contrasts greatly with unconditional inference where data with a LR of *one* may be seen as providing strong evidence against one hypothesis relative to another if it has a small p-value.

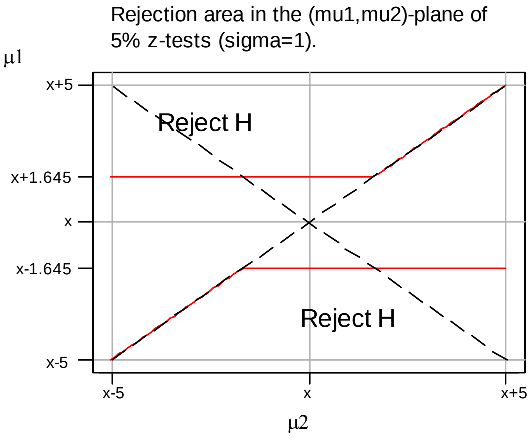
If all model/hypothesis-pair scenarios in existence were log-symmetric (which they are not), it would follow that we could satisfy both the LL and the LP by carrying out frequentist inferences conditional upon $A = |\ln Y|$. Even the case where $y > 1$ and the cp-value is not a one-to-one function of the LR, could be dealt with by testing K (as null hypothesis) versus H.

[15] Still meaning $\frac{p_H}{p_K}$.

Rejection areas in the (H,K)-plane.

We can illuminate the differences between the conventional and conditional approaches to the Normal location case by observing the areas in the (μ_1, μ_2) -plane where H: $\mu = \mu_1$ is rejected in favour of K: $\mu = \mu_2$, based on the data x (we use $\sigma = 1$). We look first at the conventional 5% test.

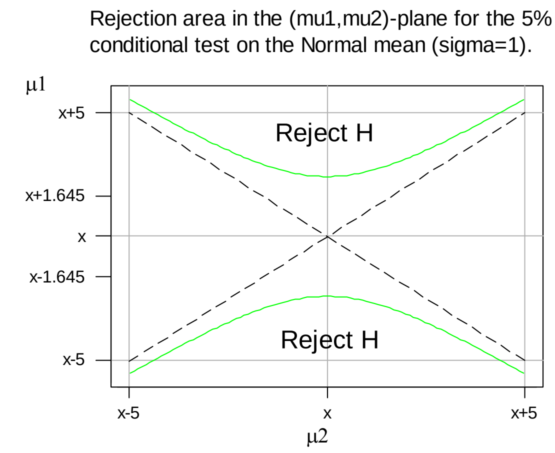
Figure 8.16



The main diagonal shows where $\mu_1 = \mu_2$; we are only interested in values of (μ_1, μ_2) not on this line. The areas labelled 'Reject H' indicate the values of (μ_1, μ_2) for which the hypothesis $\mu = \mu_1$ (being the *null* hypothesis) would be rejected in favour of the alternative μ_2, at the 5% level, based on the data x. The $LR(x)$ *equals one* on both of the diagonals and the rejection area contains some, but not all, of these points. This shows that this approach breaches the *likelihood principle* (but not the SP, since this applies only within a given test and each value of (μ_1, μ_2) corresponds to a different test).

The following plot shows the rejection areas for the conditional test using $\bar{\alpha} = 5\%$.

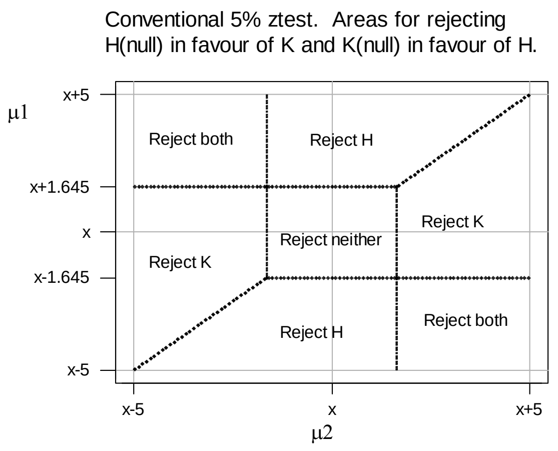
Figure 8.17


Rejection area in the (mu1,mu2)-plane for the 5% conditional test on the Normal mean (sigma=1).

For the conditional test, the boundaries of the rejection areas are likelihood ratio contours, in this case $\{(\mu_1, \mu_2) : LR(x) = \frac{1}{19}\}$.

We can also use plots to examine what happens when we reverse the order of the hypotheses, i.e. when we are interested, both, in testing μ_1 (as the null value) against the alternative, μ_2, and in testing μ_2 (as the null value) against μ_1. The following plot shows the results for conventional 5% tests. (For convenience, a hypothesis is only 'rejected' when it is acting as null hypothesis.)

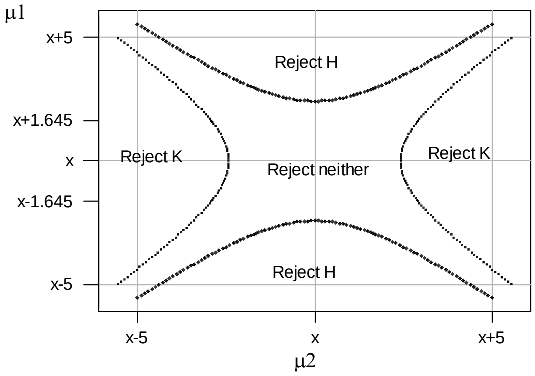
Figure 8.18


Conventional 5% ztest. Areas for rejecting H(null) in favour of K and K(null) in favour of H.

In the areas labelled ‘Reject K’, the hypothesis H (when null) is not rejected but the hypothesis K (when null) is rejected. Note that there are some (μ_1, μ_2) for which 5% tests will, both, reject H in favour of K and reject K in favour of H, in both cases on the basis of the same data x. This feature shows that the rejection of one hypothesis in favour of another, even at a low significance level, cannot be interpreted as indicating that the data constitutes strong evidence for the one hypothesis relative to the other. Any reasonable conception of ‘evidence’ excludes the possibility that any data can be, at one and the same time, strong evidence against H relative to K and strong evidence against K relative to H.

The following plot shows the results when both tests are performed conditionally with $\bar{\alpha} = 5\%$.

Figure 8.19



The area where we reject H does not overlap with that where we reject K, thus rejection of either hypothesis can be interpreted as telling us something about the evidence. This is not surprising since the conditional test is equivalent to a test based on the value of the likelihood ratio and this value can be interpreted directly as a measure of evidence. Note that the area in this plot where the evidence is weak (labelled ‘Reject neither’) is distributed among all four[16] inferences by the conventional approach (**Figure 8.18**).

[16] ‘Reject H’, ‘Reject K’, ‘Reject both’, ‘Reject neither’.

8.8 Interval estimates.

Where there is an appropriate ancillary statistic for each and every binary parameter space in the natural parameter space, we can use the conditional approach to produce a conditional confidence interval (CCI) for θ based on the conditional tests. In the Normal location case, there is an ancillary statistic of the form $A = |\ln Y|$ for any σ and any binary parameter space in $\mathbb{R}^2$ hence we can find a CCI for μ. We do this by using the following relationship between test results and intervals that applies in both conventional frequentist inference and also in likelihood inference.

An interval at 'level' $l(\gamma)$ will contain a value of θ, θ_0, if and only if there does not exist any value $\theta' \in \Theta$ (the natural parameter space) such that a test of H:$\theta = \theta_0$ versus K:$\theta = \theta'$ rejects H at the γ level.

In likelihood inference, the level stated for a likelihood interval (LI) is that of the corresponding dichotomous tests, i.e. a $\frac{1}{\lambda}$ LI can be derived from the results of $\frac{1}{\lambda}$-level tests (thus $l(\gamma) = \gamma$).

Conventional confidence intervals are able to play two roles simultaneously. Firstly, these intervals have the property that, in the long run, they include θ a given proportion of the time (the *coverage* of the interval); and, secondly, they summarise the accept/reject results of all possible tests of two simple hypotheses carried out at a given significance level. These two features are connected by the relation: $coverage = 1 - 2 \times significance\ level$, i.e. $l(\gamma) = 100(1-2\gamma)\%$. The coverage property is uncontroversial but its usefulness has been called into question, partly, because it is a property of the *method* with no implications for any observed confidence interval derived from data[17], and partly because the existence of any ancillary statistic for the natural parameter space can lead to the situation where we know that the success rates of certain, identifiable, sub-classes of intervals are different from each other and from the nominal (i.e. average) coverage. (For many years, the Welch example has been

[17] See especially Pratt (1961), p.165 for a concise and witty discussion of why this is such a serious shortcoming.

regarded as the classic example of this phenomenon. We have shown that, while it constitutes an interesting case, it does not provide a good argument in favour of (Fisherian) conditioning, as is usually thought. A better instance of conditioning solving the 'variable coverage rates problem' is given in our 'Gradient Model' discussed in Chapter 10.)

The more important role of intervals may, therefore, be as a summary of the results of all possible tests. A simple formula for the confidence interval circumvents the need to carry out many separate hypothesis tests. It follows that the interval is only as good as the tests to which it is equivalent. The common expectation of intervals is that they should contain all the 'plausible' values of θ or 'all values of θ that are reasonably consistent with the data'. In the absence of posterior probabilities, this has no easy interpretation. However, we can certainly base an interval method on our conditional tests. In that case the interval comprises just those values that are consistent with the data in the very definite sense that '*the data does not provide much more evidence for any other hypothesis than for this one*'. We know that a conventional confidence interval does not satisfy this requirement because standard tests often reject a hypothesised value based on data that does not constitute strong evidence against it in favour of any alternative. It follows that values are excluded unnecessary from conventional intervals, which are, as a result, shorter than likelihood intervals or conditional confidence intervals. The shortness of the conventional interval tends to be regarded as a point in its favour because it *homes in* on θ better than (say) a likelihood interval. But this assumes an evidential interpretation that is not justified by the unconditional methodology. Certainly intervals should not be wider than necessary but it seems that our requirement that they contain all values consistent with the data (as defined above) does make the extra length of the LI and CCI a necessity.

Conditional confidence intervals for the log-symmetric case.

For any data, t, we can find an optimal conditional test (see §8.7) on any $\Theta_B \equiv (\mu_1, \mu_2)$ and this produces a relevant significance level of $\alpha_a = \alpha_a(\mu_1, \mu_2)$. Let[18]

$$\bar{\alpha} = \max_{(\mu_1,\mu_2)\in\mathbb{R}^2} \alpha_a(\mu_1, \mu_2),$$

i.e. $\bar{\alpha}$ is the achievable least upper bound on the relevant significance level of the tests on all the various hypothesis pairs.

The CCI for the Normal mean, μ, based on the observed value of $T \sim N(\mu, \sigma^2)$, and corresponding to tests with an upper bound on the conditional significance level of $\bar{\alpha}$, is given by:

$$\boxed{t_0 \pm \sigma\sqrt{2\ln(\tfrac{1-\bar{\alpha}}{\bar{\alpha}})}.}$$

Like the CI, and the likelihood interval (for this model), the CCI is symmetric around the observation t_0. Thus any such symmetric interval can be interpreted as either a CCI, CI or LI. Depending on the interpretation preferred, the *rigor* of the interval will then be determined by either $\bar{\alpha}$, α or λ (respectively).

The interpretations of six different intervals as likelihood intervals, conventional confidence intervals (CI), or conditional confidence intervals (CCI) are shown below in **Table 8.11**. For example (looking at the left-most column), the CCI using $\bar{\alpha} = 5\%$ is identical to the CI based on (one-sided) tests at $\alpha = 0.76\%$ (usually called a 98.48% CI), and to the $\frac{1}{19}$LI for μ. Note the two intervals at the right end of the table. The likelihood and conditional interpretations of the 90% and 95% conventional intervals suggest they are inadequate since $\frac{1}{\lambda}$ and $\bar{\alpha}$ are too big.

[18] More generally, the upper bound on the conditional significance level is defined by $\bar{\alpha} = \max_{(\theta_i,\theta_j)\in\Theta^2} \alpha_a(\theta_i, \theta_j)$.

Table 8.11

CCI: $\bar{\alpha}$	***5.00%***	***2.50%***	***5.88%***	***3.03%***	***20.54%***	***12.78%***
CI : α	***0.76%***	***0.34%***	***0.93%***	***0.42%***	***5.00%***	***2.50%***
LI: $\frac{1}{\lambda}$	***1/19***	***1/39***	***1/16***	***1/32***	***1/3.9***	***1/6.8***

(Note: the $\frac{1}{\lambda}$ LI for μ is: $\boxed{t_0 \pm \sigma\sqrt{2\ln\lambda}}$.)

The CCI and LI are in broad agreement (as are the analogous tests), since the CCI with significance level bound, $\bar{\alpha}$, is the same as the $\frac{1}{\lambda}$ LI with $\frac{1}{\lambda} = \frac{1-\bar{\alpha}}{\bar{\alpha}}$. Thus $\bar{\alpha} = \frac{1}{m}$ implies that $\frac{1}{\lambda} = \frac{1}{(m-1)}$, this means that conventionally significant values of $\bar{\alpha}$ (say, less than $\frac{1}{10}$) are associated with significantly small[19] values of $\frac{1}{\lambda}$.

8.9 The Cauchy location model.

For tests about the mean of a Normal population, $\mu = E(T)$, we find that, for any binary parameter space, $\{\mu_1, \mu_2\}$, the statistic

$$A = |\ln Y| = |\ln LR(T; \mu_1, \mu_2)| = |\ln\{\tfrac{f_T(T;\mu_1)}{f_T(T;\mu_2)}\}|$$

is ancillary (in the restricted sense). We have called this property log-symmetry, and it is not confined to the Normal location model.

If the density function of a variable X is given by

$$f_X(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}, \ x \in \mathbb{R}, \ \theta \in \mathbb{R},$$

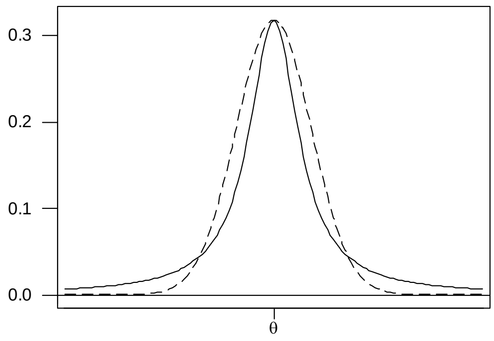
we say that X has a *Cauchy* distribution with location parameter $\theta = \text{median}(X)$ (and scale parameter *one*). By comparison with the Normal distribution, the Cauchy has very heavy tails, as shown below[20]. (When $\theta = 0$ this distribution is also called the *T* distribution with *one* degree of freedom.)

[19] As judged by Royall, for example.

[20] The densities shown are Cauchy(θ) and $N(\theta, \frac{\pi}{2})$.

Figure 8.20

Comparison of Cauchy (solid line) and Normal densities.



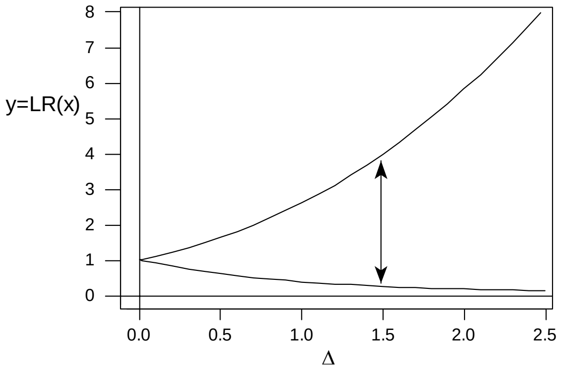
As in the Normal case (for fixed σ), the shape and spread of the Cauchy density is unaffected by changes to θ and the density is symmetric. Let x be a single observation from a Cauchy population, then, for any distinct values θ_1 and θ_2, the likelihood ratio of x is given by:

$$y = LR(x) = \frac{f_X(x;\theta_1)}{f_X(x;\theta_2)} = \frac{1+(x-\theta_2)^2}{1+(x-\theta_1)^2}.$$

For any fixed $\Theta_B \equiv \{\theta_1, \theta_2\}$, define $\Delta = |\theta_1 - \theta_2| > 0$, then $\forall x \in \mathbb{R}$, $y = LR(x)$ lies in the interval $(1/k, k)$ where $k = \frac{1}{2}(2 + \Delta^2 + \Delta\sqrt{4+\Delta^2})$ and is greater than *one*. These bounds are shown below as a function of Δ.

Figure 8.21

Bounds on y (as a function of delta).



The Cauchy location model is log-symmetric.

It is clear that the statistic $D = | X - \overline{\theta} |$ (where $\overline{\theta} = \frac{\theta_1 + \theta_2}{2}$) has the same distribution under each of the hypotheses defined by Θ_B; this is true by symmetry, as in the Normal case.

This case differs from the Normal case in that D is not a function of Y – the MSS – it is thus not ancillary in the restricted sense. However, if $A = | \ln Y |$ can be shown to be a function of D, then A must be ancillary in the restricted sense, since it is clearly a function of Y.

A is a function of D if and only if $D(x_1) = D(x_2) \Rightarrow A(x_1) = A(x_2), \ \forall x_1, x_2$. Suppose $D(x_1) = D(x_2) = d$, then either $x_1 = x_2$ in which case $A(x_1)$ obviously equals $A(x_2)$, or (WLOG) $x_1 = (\overline{\theta} - d)$ and $x_2 = (\overline{\theta} + d)$. If $LR(x_2) = [LR(x_1)]^{-1}$ then $A(x_1) = A(x_2)$.

Let $\delta = \overline{\theta} - \theta_1 = \theta_2 - \overline{\theta}, \quad (\delta \in \mathbb{R}),$ then:

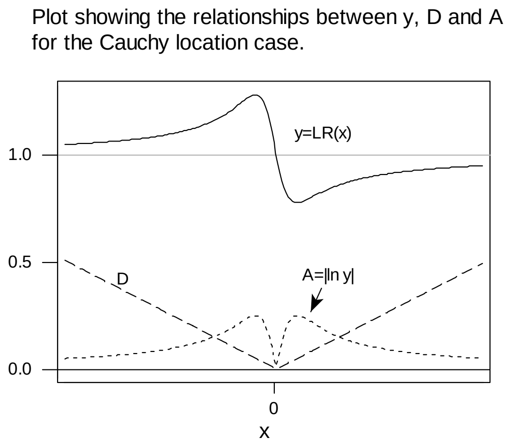
$$\begin{aligned} LR(x_1) &= \frac{1 + (\overline{\theta} - d - \theta_2)^2}{1 + (\overline{\theta} - d - \theta_1)^2} \\ &= \frac{1 + (-\delta - d)^2}{1 + (\delta - d)^2} \\ &= \frac{1 + (\delta + d)^2}{1 + (d - \delta)^2}, \end{aligned}$$

and

$$\begin{aligned} LR(x_2) &= \frac{1 + (\overline{\theta} + d - \theta_2)^2}{1 + (\overline{\theta} + d - \theta_1)^2} \\ &= \frac{1 + (d - \delta)^2}{1 + (\delta + d)^2} \\ &= \frac{1}{LR(x_1)}. \end{aligned}$$

Hence $A(x_1) = A(x_2)$. The following plot shows the relationships between these variables, i.e. that A is a function of Y and A is also a function of D.

Figure 8.22



Since A is a function of D, it must – like D – have the same distribution under both hypotheses, thus the Cauchy random variable is log-symmetric. It follows that the conditional probabilities of $Y \mid A$ are those already derived for the log-symmetric case and, hence, for $y < 1$, and all $\{\theta_1, \theta_2\}$, the p-value of y conditional upon $A = a = |\ln y|$ is $cp(y) = \dfrac{y}{(1+y)}$, as in the Normal case, while, for $y > 1$, $cp(y) = 100\%$.

Proof that the p-value is less than the cp-value $\forall y \in (\tfrac{1}{k}, k),\ \forall \theta_1, \theta_2$.

The conventional p-value is $p(y) = F_H(y)$, where F_H is the distribution function of the likelihood ratio statistic, Y, under the hypothesis (H) $\theta = \theta_1$. For $y > 1$, $cp(y) = 100\%$, but (since Y is continuous with a positive density on the interval $(\tfrac{1}{k}, k)$) $p(y) < 100\%$ for all $y < k$, hence for $1 < y < k$, $p(y) < cp(y)$.

To show that $p(y) < cp(y)$ when $y < 1$ we need to show that $F_H(y) < y/(1+y),\ \forall \theta_1, \theta_2$.

In order to prove this result, we treat $y<1$ as fixed and allow $\Delta =| \theta_1 - \theta_2 |$ to vary within the bounds consistent with the observation of y, i.e. $\Delta^2 \geq (1-y)^2 / y$. The scenario is complicated by the fact that Y is not a one-to-one function of the natural variable, as it was in the Normal case, instead there are two values of x (call them $x_1(y)$ and $x_2(y)$) associated with any given value of y.

$p_\Delta(y) = F_H(y;\Delta) = \frac{1}{\pi}\{\tan^{-1}(x_2(y)-\theta_1) - \tan^{-1}(x_1(y)-\theta_1)\}$, where $LR(x_1(y);\theta_1,\theta_2) = LR(x_2(y);\theta_1,\theta_2) = y$.

The latter equation can be solved to give:

$$x_1(y)-\theta_1 = \frac{(\theta_2-\theta_1)}{(1-y)} - \left\{\frac{y\Delta^2}{(1-y)^2} - 1\right\}^{1/2}$$

and

$$x_2(y)-\theta_1 = \frac{(\theta_2-\theta_1)}{(1-y)} + \left\{\frac{y\Delta^2}{(1-y)^2} - 1\right\}^{1/2} .$$

For fixed $y<1$, let[21]

$$\begin{aligned} g(\Delta) &= p_\Delta(y) \\ &= \frac{1}{\pi}\{\tan^{-1}[\frac{\Delta}{(1-y)} + \frac{\{y\Delta^2-(1-y)^2\}^{1/2}}{(1-y)}] - \tan^{-1}[\frac{\Delta}{(1-y)} - \frac{\{y\Delta^2-(1-y)^2\}^{1/2}}{(1-y)}]\}. \end{aligned}$$

Thus, for any given y, g is a function of Δ on the domain $\Im_y \equiv [\frac{|1-y|}{\sqrt{y}}, \infty)$. If, for all $y<1$, $\max_{\Delta\in\Im_y} g(\Delta) < y/(1+y)$, then it follows that $p(y) < cp(y),\ \forall y<1, \theta_1, \theta_2$. It is straightforward to show that the function g has maxima at two turning points occurring at $\Delta = \pm\sqrt{\frac{2(1-y)}{y}}$ (and not on the bound of the domain); both turning points produce the same value of $g(\Delta)$, i.e.

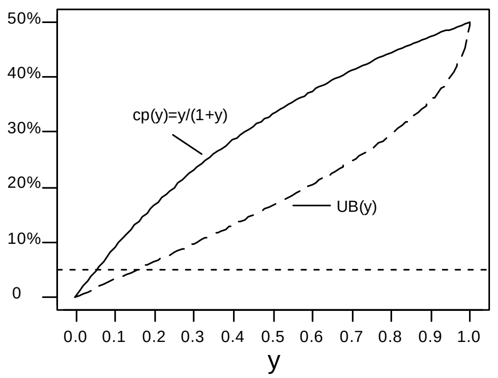
[21] This expression is appropriate regardless of whether θ_1 is greater or less than θ_2 since $\tan^{-1}(-u) = -\tan^{-1}(u)$.

$$\max(g) = \frac{1}{\pi}\{\tan^{-1}[\tfrac{1}{\sqrt{1-y}}(\sqrt{\tfrac{2}{y}} + \sqrt{1+y})] - \tan^{-1}[\tfrac{1}{\sqrt{1-y}}(\sqrt{\tfrac{2}{y}} - \sqrt{1+y})]\}.$$

Let $UB(y) = \max(g)$. This is an *upper bound* on $p(y)$ in the sense that, for all Δ consistent with y, and for all $y < 1$, $UB(y) \geq p_\Delta(y)$. Thus, if $y/(1+y) > UB(y)$ $(\forall y < 1)$, it follows that $cp(y) > p_\Delta(y)$, as claimed. The plot, below, shows that this is true.

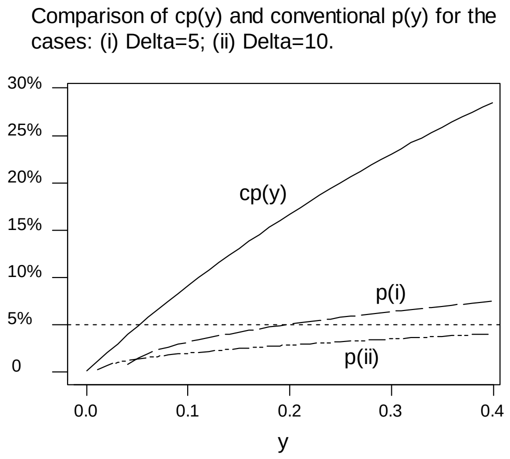
Figure 8.23

Comparison of cp-value(y) and the upper bound (UB) for all p-value(y) for the Cauchy location case.



The gap between $p_\Delta(y)$ and $cp(y)$ is generally larger than the above plot suggests since, for any given Δ (i.e. hypotheses), $p_\Delta(y)$ is less than $UB(y)$ for all observable values of y other than $y = 2/(2+\Delta^2)$. The conventional p-values for the cases $\Delta = 5$ and $\Delta = 10$ are shown in the plot below. Note that the conventional p-values are significant (less than 5%) for many values of y for which the cp-value is not. A likelihood ratio of $0.4 = \frac{1}{2.5}$ (viewed from a likelihood perspective) provides evidence against H that is not much stronger than that which tossing a single *head* provides for the two-headed hypothesis relative to the fair coin hypothesis. The cp-value of this observation is 28.6% (for all Δ), which is consistent with the likelihood interpretation, whereas the conventional p-values are less than 10% when $\Delta = 5$ and less than 5% when $\Delta = 10$.

Figure 8.24



(Note that, for any given value of Δ, y is at least $\frac{1}{2}\{2+\Delta^2-\Delta\sqrt{4+\Delta^2}\}>0$ which equals 0.037 for $\Delta=5$ and 0.010 for $\Delta=10$. For the sake of simplicity, $cp(y)$ and $UB(y)$ are shown as a functions of $y>0$ in the above plots.)

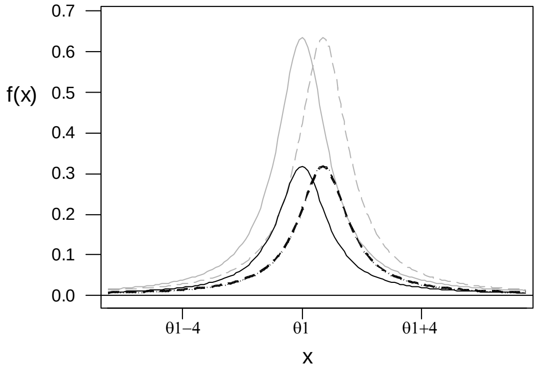
A limited range of likelihood ratios.

Example 8.3.

Consider a test of the form H: $\theta=\theta_1$ versus K: $\theta=\theta_1+\frac{1}{\sqrt{2}}$ (for any $\theta_1\in\mathbb{R}$), thus $\Delta=\frac{1}{\sqrt{2}}$ and $y\in[\frac{1}{2},2]$. In this case, y takes a very narrow range of values; no data supports either hypothesis over the other to any greater extent than the result *head* favours the double-headed hypothesis over the fair coin hypothesis.

This can be seen in the following plot. The densities of X under H (solid line) and under K (dotted line) are shown, together with twice the densities (in grey). For no value of x, is either of the densities ever more than twice the other density.

Figure 8.25



A single observation on the random variable X is not a good basis for choosing between these particular hypotheses because X behaves in much the same way regardless of whether it is H or K that is true. This flaw in the design of the experiment is reflected in the power of the conventional test (see below[22]), but not in the range of the p-value.

Table 8.12

Sig. Level, α	CLR*	Power, κ
1%	0.500370	1.999%
5%	0.509233	9.939%

[* Critical likelihood ratio.]

The fact that the power is very small – only about twice the significance level – is a flaw in the design of the experiment but is usually regarded as affecting the inference only by limiting our ability to infer anything useful from a *failure* to reject H. Note that, if we were to observe a value of $y \le 0.50037$, we would reject H at the 1% level;

[22] These values are found by numerically solving the equation:

$$F_{Y,\theta}(y) = \tfrac{1}{\pi}\{\tan^{-1}(\theta_1 - \theta + \tfrac{1}{(1-y)\sqrt{2}} + \sqrt{[\tfrac{y}{2(1-y)^2} - 1]}) - \tan^{-1}(\theta_1 - \theta + \tfrac{1}{(1-y)\sqrt{2}} - \sqrt{[\tfrac{y}{2(1-y)^2} - 1]})\}$$

to find

a) $CLR : F_{Y,\theta_1}(CLR) = \alpha,$ and

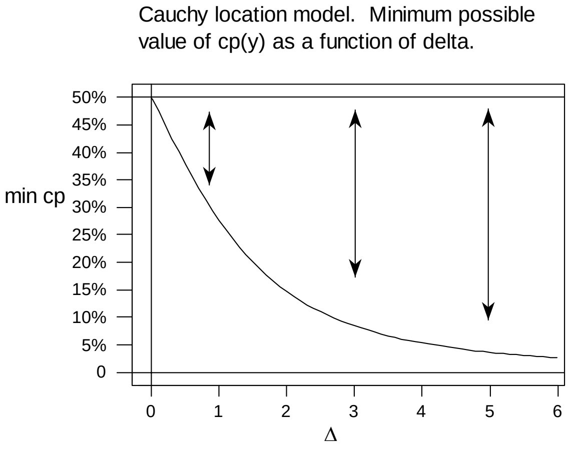
b) $\kappa = F_{Y,\theta_1+\frac{1}{\sqrt{2}}}(y_\alpha)$.

this result would appear to be truly significant since the probability of Type I error is small (1%), unlike the probability of Type II error.

This type of problem can occur in any case where the likelihood ratio statistic, Y, is a (non-degenerate) continuous statistic. No matter how restricted the domain of Y, there will be some observable values of y that produce statistically significant results since $p(y) = F_{Y,H}(y) \to 0$ as $y \to y_L$ (its lower bound). This is not true of cp-values. In any log-symmetric case, $cp(y) = \frac{y}{(1+y)}$ $(\forall y < 1)$ and thus if y is bounded below by some y_L it follows that $cp(y)$ is itself bounded below by the value $\frac{y_L}{(1+y_L)}$. In the present example $y_L = \frac{1}{2}$ and thus $cp(y)$ can never be less than 33%; no data that can be obtained from this poorly designed experiment will be interpreted as significant evidence for rejecting H in favour of K, if the inference is made conditional upon A.

The cp-value reflects the facts about this scenario – the limited range of likelihood ratios – much more accurately than does the p-value. The following plot shows the minimum possible value (over y) of $cp(y)$ as a function of $\Delta = |\theta_1 - \theta_2|$, in the Cauchy case. For any given value of Δ and $y < 1$, $cp(y)$ must lie above the curved line. The smaller Δ is, the more restricted are the possible values of y (i.e. the more similar the distributions of X under H and K), and the more limited the range of cp-values; it is not possible find a cp-value of 5% (or less) when $\Delta \le 4$.

Figure 8.26

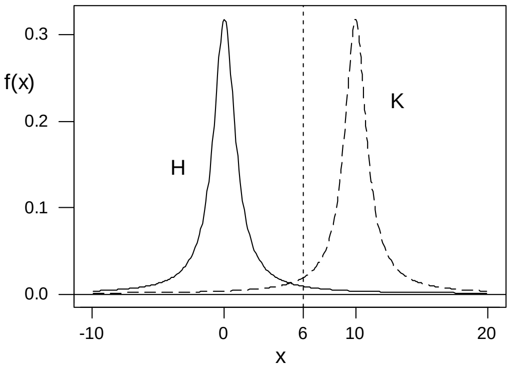


However, disparities between the likelihood ratio and the p-value do not *only* occur when the hypotheses are close together.

Example 8.4.

Consider a test of H: $\theta = 0$ versus K: $\theta = 10$. In this case, y can be observed anywhere in the interval $[\frac{1}{102}, 102]$, so it is possible to observe data that provides strong evidence against H relative to K (or *vice versa*). Does this mean we can confidently interpret a small p-value as such evidence? Suppose we observe $x = 6$ (shown below).

Figure 8.27



It is clear that $x = 6$ is only slightly more consistent with K than with H. We can confirm this by calculating the likelihood ratio $y_0 = 0.4595 = \frac{1}{2.18}$. Thus, again, the evidence is of about the same weight as that from a single *head* in the paradigm coin-tossing example. The cp-value is not significant: $cp(y) = 31.5\%$. However, the conventional p-value of this observation is 4.32%. Even though y can take values that are much smaller than the observed $y_0 = 0.4595$, the (unconditional) probability of these values is so small that $P_H(Y \le 0.4595) = 4.32\%$. Thus the observed value y_0 is *relatively* small (probabilistically speaking) and so produces a significant result even though it is not actually small and there are observations with much smaller likelihood ratios. When we condition on $|\ln Y|$, this effect is greatly mitigated.

Averaging over unobserved values.

As always, the conventional p-value (or significance level) is the mean, over a, of the conditional probabilities (or levels). Since only one value of a is observed in any given experiment, it is inappropriate to quote the average over all possible values. Consider again the above example where $\Delta = 10$. If we observe the data $y_0 = 0.4595$, then the unconditional p-value is significant at 4.32%. The various conditional probabilities that go to make up this unconditional probability are:

$$\rho(a) = P_H(Y \le 0.4595 \mid A = a), \quad a \in (0, \ln 102].$$

If f_A is the density function of A, then the conventional p-value can be written as:

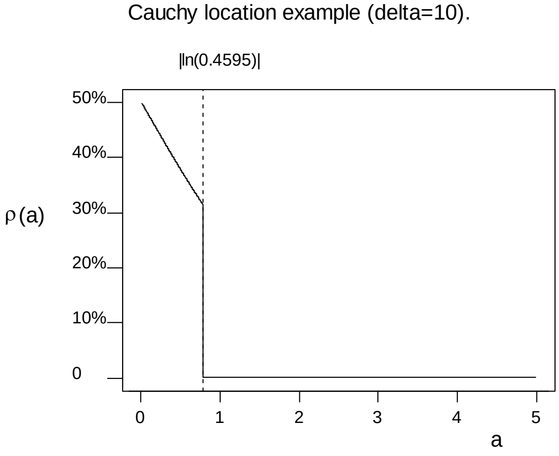
$$p(0.4595) = \int_0^{\ln 102} \rho(a) \cdot f_A(a)da = E(\rho(A)).$$

We can define the conditional probabilities $\rho(a)$ for every value of a, but note that only one of these probabilities is a conditional p-value – namely, $\rho(a_0)$ where $a_0 = |\ln 0.4595|$. (To find a cp-value, we condition on the value of a that actually occurred.)

Why is the p-value so small?

We can see why conditioning makes such a difference when we will look at the values of $\rho(a)$ associated with different values of a.

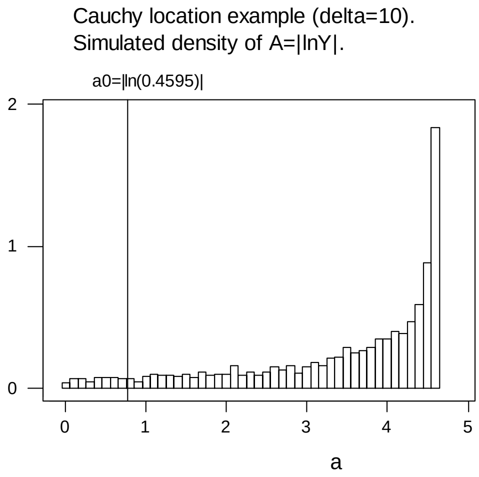
Figure 8.28



$\rho(a) = 0$ unless $a \le a_0 = |\ln 0.4595| \approx 0.78$, since is it possible to observe $y \le 0.4595$ only when a is in this range. Thus many of the values in our average are *zero*, although the value that we observed, $\rho(a_0) = cp(y_0)$, is not.

The other influence on the overall average is the distribution of the ancillary statistic A. The following histogram shows the distribution of A simulated from a sample of 3000 instances.

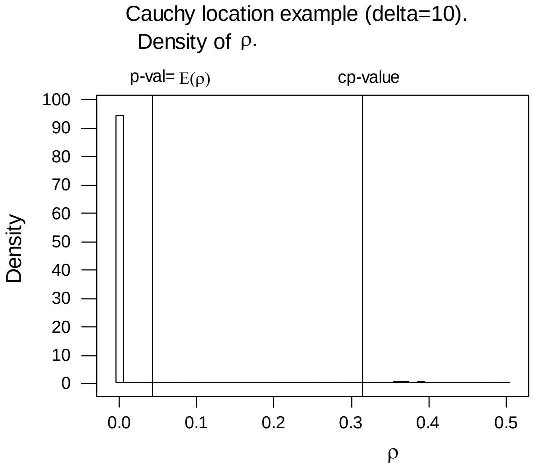
Figure 8.29



The values of a for which $\rho(a) = 0$ have a very high probability and this affects the conventional p-value, which is the expected value of $\rho(A)$.

We can identify $E(\rho(A))$ from the (simulated) distribution of the random variable, $\rho(A)$.

Figure 8.30



There is a large probability mass associated with the value $\rho = 0$, this is so great that the densities associated with the other possible values of ρ in [0.315,0.5) barely show on the histogram. Clearly the high probability that $\rho = 0$ has a huge effect on the mean, $E(\rho) =$ p-value, and drags it down towards *zero*. Before the experiment, the probability that we would observe a case where $\rho = 0$ was very high; this fact dominates the conventional p-value, even though, in our particular case, ρ turned out to be 31.5%. Only by using the cp-value can we remove the influence of the unobserved values of A, and, hence, of $\rho(A)$[23]. Before the experiment, there was a high probability that $A =| \ln Y |$ would be large, i.e. that $Y = LR(X)$ would be very large or very small. This reflects the fact that the two hypotheses are far apart and the bulk of the observable x-values are much more likely under one hypothesis than the other. In such a case, we have a high expectation that the data we observe will provide definite evidence one way or the other. However, if we are unlucky, the experiment will produce data that is not much more likely under one hypothesis than the other. In such a case, we should accept that this has happened and not attempt to modify the result by including the more definite evidence that we might have, but did

[23] Since $\rho(A)$ is a function only of A, it is also ancillary in our sense and we can think in terms of conditioning on the observed value of ρ if we wish (the results are the same). Note that this only applies when we are interested specifically in the p-value of the data $y = 0.4595$, since ρ was defined for this purpose. That ρ can not be used as an equivalent general purpose ancillary statistic is evident from the fact that it is not a one-to-one function of A.

not, observe. Using the average p-value instead of the conditional p-value amounts to quoting the strength of evidence that we *deserved* to get (in view of our experimental design), rather than that actually produced by the experiment.

Generalisation.

We can make the Cauchy model more general by allowing a scale parameter other than *one*. If X is Cauchy with a known scale parameter, σ, and median θ which is the parameter of interest, then

$$f_X(x;\theta,(\sigma)) = \frac{\sigma}{\pi(1+(\frac{x-\theta}{\sigma})^2)},\ x\in\mathbb{R},\ \theta\in\mathbb{R},\ (\sigma>0)\,.$$

Our discussion generalises to this case.

8.10 Other log-symmetric scenarios.

We will briefly mention some other models that are log-symmetric so that the same relationship exists between $y = LR(x)$ and cp-value(x) as in the Normal and Cauchy location cases.

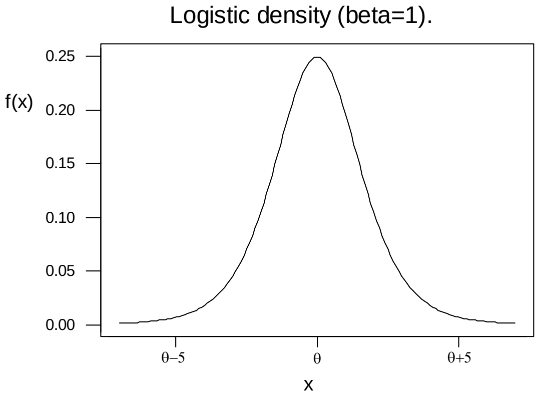
The Logistic (location) model.

Consider the continuous random variable, X, with density dependent on the location parameter, θ, and a fixed, known scale parameter β according to:

$$f_X(x;\theta,(\beta)) = \frac{e^{-(x-\theta)/\beta}}{\beta\{1+e^{-(x-\theta)/\beta}\}^2},\ x\in\mathbb{R},\ \theta\in\mathbb{R},\ \beta\in\mathbb{R}^+.$$

Then X has a *logistic* distribution and its density is symmetric around θ, as shown below.

Figure 8.31

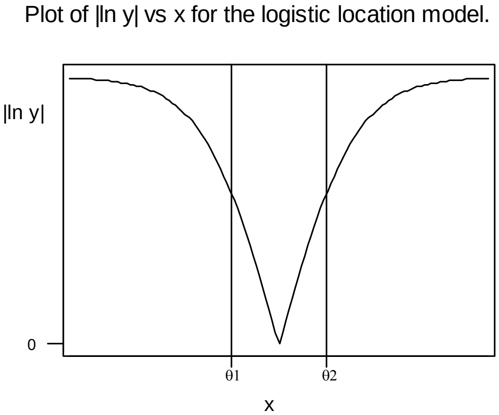


For a test of θ_1 versus θ_2, the likelihood ratio of a single observation, x, is given by:

$$y = LR(x) = \frac{e^{(\theta_1-\theta_2)/\beta}\{1+e^{-(x-\theta_2)/\beta}\}^2}{\{1+e^{-(x-\theta_1)/\beta}\}^2}.$$

Y is a continuous variable on the support $(\frac{-|\theta_1-\theta_2|}{\beta}, \frac{|\theta_1-\theta_2|}{\beta})$ and, since $|\ln y|$ is symmetric around $\frac{(\theta_1+\theta_2)}{2}$ (shown below), is follows that this model is log-symmetric.

Figure 8.32



Thus, for all $(\theta_1, \theta_2) \in \mathbb{R}^2$ and $\beta > 0$ (fixed), $cp(y)$ is the same function of y as in the Normal and Cauchy cases.

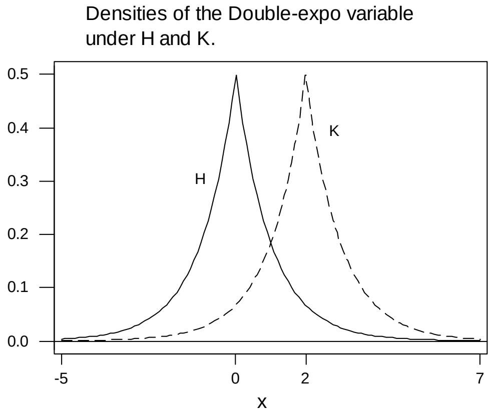
The Double-exponential (location) model.

Consider the continuous random variable, X , with density dependent on the location parameter, θ, according to:

$$f_X(x;\theta) = \tfrac{1}{2}\exp\{-|x-\theta|\},\ x \in \mathbb{R},\ \theta \in \mathbb{R}.$$

The plot below shows the densities of X under H: $\theta = 0$ and K: $\theta = 2$.

Figure 8.33

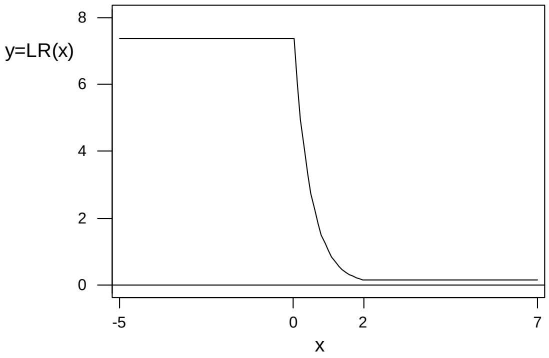


Although X is a continuous variable $(\forall\theta)$, $Y = LR(X)$ is not, since (for instance) when $\theta_1 < \theta_2$

$$y = \frac{\tfrac{1}{2}\exp\{-|x-\theta_1|\}}{\tfrac{1}{2}\exp\{-|x-\theta_2|\}} = \begin{cases} \exp\{+|\theta_1-\theta_2|\},\ x < \theta_1 \\ \exp\{-|\theta_1-\theta_2|\},\ x > \theta_2. \end{cases}$$

This is shown in the following plot.

Figure 8.34



$$y = \begin{cases} e^2 \approx 7.4, & x < 0 \\ e^{-2} \approx 0.14, & x > 2. \end{cases}$$

Thus $P_H(Y = e^{-2}) = P_H(X > 2) > 0$, i.e. Y has positive probability mass at e^2 and e^{-2}. Y is partly continuous and partly discrete. However, by symmetry, it is easy to show that $|\ln Y|$ (which is also partly discrete) still has the same distribution under H and K and hence this model is log-symmetric for all $\{\theta_1, \theta_2\} \in \mathbb{R}^2$. It follows from this that the same relationship between cp-value(x) and $LR(x)$ applies here as in the Normal, Cauchy and Logistic cases.

The Bernoulli model with symmetric hypotheses & stopping rules.

We now look at two models where Y is discrete.

Consider a series of independent Bernoulli trials each resulting in either *success* or *failure* with constant probabilities p and $(1-p)$ respectively; the parameter of interest is p. The trials continue until brought to a halt by the *stopping rule*, R. We consider only that subset of this class of experiments satisfying both the following conditions:

- The binary parameter space for p is of the form $(\theta, 1-\theta)$, where $\theta \in (0,1)$.
- The stopping rule, R, is 'symmetric' in the following sense: If we were to define a second stopping rule by swapping over the roles played by *success* and *failure* in the definition of R, the two rules would have the same meaning.

Suppose a model/hypothesis-pair satisfies these requirements, then that case is log-symmetric, i.e. $|\ln Y|$ is ancillary, in the restricted sense, on the binary parameter space. We prove this below.

Let $U = \#successes$, $V = \#\,failures$ (thus $\#trials = U + V$), then (U,V) is a sufficient statistic for p. Then,

$$P[(U,V)=(u,v)] = C_R(u,v)\,p^u(1-p)^v, \text{ where } (u,v) \in \varsigma_R{}^2,\ \varsigma_R \subseteq \{0,1,2,\ldots\}.$$

$C_R(u,v)$ is a combinatoric term, (based on the rule, R) which counts the number of ways in which the experiment can end with the result 'u successes and v failures'.

Without loss of generality let H: $p = \theta$ and K: $p = 1-\theta$ then

$$y = LR(u,v) = \tfrac{\theta^u (1-\theta)^v}{(1-\theta)^u \theta^v} = \exp\{(u-v)\cdot\ln(\tfrac{\theta}{(1-\theta)})\}.$$

Hence, $|\ln Y| = |(U-V)\cdot\ln(\tfrac{\theta}{1-\theta})|$.

Since θ is a constant, this is a one-to-one function of $|U-V|$ which is an equivalent ancillary statistic. To show that this case is log-symmetric, we need to show that $|U-V|$ has the same distribution under both hypotheses.

First note that, because the rule, R, is symmetric, $C_R(i,j)=C_R(j,i),\ \forall i,j$. Hence,

$$\begin{aligned}
&P(|U-V|=a)=P[(U-V)=a]+P[(V-U)=a]\\
&=\sum_{v=l(a)}^{m(a)} C_R(a+v,v)\,p^{a+v}(1-p)^v+\sum_{u=l(a)}^{m(a)} C_R(u,a+u)\,p^u(1-p)^{a+u}\\
&=\sum_{v=l(a)}^{m(a)} C_R(a+v,v)\,p^{a+v}(1-p)^v+\sum_{u=l(a)}^{m(a)} C_R(a+u,u)\,p^u(1-p)^{a+u}\\
&=\sum_{i=l(a)}^{m(a)} C_R(a+i,i)[\,p^{a+i}(1-p)^i+p^i(1-p)^{a+i}].
\end{aligned}$$

(The bounds $l(a)$ and $m(a)$ depending on R and a.)

$$\begin{aligned}
&P_H(|U-V|=a)\\
&=\sum_{i=l(a)}^{m(a)} C_R(a+i,i)[\theta^{a+i}(1-\theta)^i+\theta^i(1-\theta)^{a+i}],
\end{aligned}$$

and

$$\begin{aligned}
&P_K(|U-V|=a)\\
&=\sum_{i=l(a)}^{m(a)} C_R(a+i,i)[(1-\theta)^{a+i}\theta^i+(1-\theta)^i\theta^{a+i}]\\
&=P_H(|U-V|=a).
\end{aligned}$$

Thus, this is a log-symmetric scenario of which Birnbaum’s Binomial example with $p_1+p_2=1$ (discussed in Chapter 5) is a special case.

While the best-known symmetric stopping rule is the rule (fixed n) that gives rise to the Binomial distribution, it is easy to define many other such rules; for example, ‘sample until there are at least twice as many outcomes of one kind as the other’ or ‘sample until there are equal numbers of successes and failures or $n=100$’. Clearly the range of rules is very varied as are the sample spaces to which they give rise; it follows that a given experimental result will often produce very different (conventional) p-values, depending on which rule was used. However, if we condition on the observed value of $|\ln Y|$, the cp-value of any given data will be the same regardless of which symmetric stopping rule was used. This is immediate from the fact that any given data uniquely defines (u,v), which, in turn, defines y (shown above), and hence $cp(y)$ for the log-symmetric case. Thus, conditioning removes the effect that the stopping rule[24] has on the interpretation of the data.

The Uniform location (Welch) model.

Finally, we note that the Welch example, $\underset{\sim}{X} = (X_1, \ldots, X_n)$ where $X_i \sim Uni(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, is log-symmetric for all $\{\theta_1, \theta_2\} \in \mathbb{R}^2$, as we showed in Chapter 6. This is still true if we let $X_i \sim Uni(\theta - c, \theta + c)$ for any fixed, known c.

Summary.

The following are all log-symmetric test scenarios:

a) Normal location test with any variance and hypotheses.[25]

b) Cauchy location test with any scale parameter and hypotheses.

c) Logistic location test with any scale parameter and hypotheses.

d) Double-exponential location test with any scale parameter and hypotheses.

e) Test on the Bernoulli probability based on independent Bernoulli trials with a symmetric stopping rule and hypotheses such that $p_1 + p_2 = 1$.

f) Test on the location of a Uniform mean with any known spread and any hypotheses.

Any observed likelihood ratio is interpreted the same way by the cp-value, based on the ancillary statistic $|\ln Y|$, in all of the above cases.

[24] Among this class of rules.

[25] *Simple* hypotheses.

8.11 Why the cp-value is larger than the p-value, in all log-symmetric cases.

When we condition on the ancillary statistic, $A = |\ln Y|$, the cp-value of any data is always greater than the conventional p-value. This is surprising on two counts:

i. It contrasts with the conditional *significance levels*. These vary around the unconditional significance level (their mean), so that they are sometimes greater and sometimes less than the nominal level.

ii. In Cox's example of two Normal populations, it is clear that the unconditional p-value of any observation is the mean of the two conditional p-values. So, in that case, the cp-value is sometimes greater than the p-value and sometimes less, depending on which of the two populations we observed.

Why does the log-symmetric case produce such different results?

To answer this question we need to consider a fairly general version of the conditioning scenario. Suppose that we want to perform conditional inferences based on some statistic that has the same distribution under both hypotheses in the binary parameter space. All data can be transformed into the form (y, a) where y is the likelihood ratio of the data and a is the observed value of the 'ancillary' statistic being used[26]. If A is ancillary in the restricted sense, it will be a function of the MSS, Y, and the second part of this expression will be redundant, however, we will keep the more general formulation since, when Cox's example is applied to a binary parameter space, A is not a function of Y (and is not ancillary in the restricted sense).

Suppose that we observed data corresponding to (y_0, a_0); the conditional p-value of this data is $cp(y_0, a_0) = P_H(Y \le y_0 \mid A = a_0)$ (which may need to be defined in the limit).

[26] Note that (Y, A) is sufficient since Y is minimal sufficient. A is possibly ancillary in the unrestricted sense of having (only) the same distribution for all θ in the parameter space.

The unconditional p-value of the data is:

$$p(y_0,a_0)=P_H(Y\le y_0)=\int_a P_H(Y\le y_0\mid A=a)\cdot f_A(a)da.$$

(This value will be the same for *any* observable value of a_0 consistent with y_0.) Thus $p(y_0,a_0)$ is the mean, over a, of the conditional probabilities $P_H(Y\le y_0\mid A=a)$. It follows that these probabilities must vary around $p(y_0,a_0)$, but these probabilities are not conditional p-values of y_0 since a cp-value involves only values y_0 and a that correspond to some observable data.

For a fixed y_0 and any given value of a we define

$$y_0(a)=\max_{y\le y_0}\{y:(y,a)\text{ corresponds to some observable data}\}.$$

Thus $y_0(a)$ is the largest value of y, not exceeding y_0, with which we may observe the ancillary statistic taking the value a.

Then, when $y_0(a)$ exists, $P_H(Y\le y_0\mid A=a)=P_H(Y\le y_0(a)\mid A=a)=cp(y_0(a),a)$, i.e. it is the conditional p-value of the observation corresponding to $(y_0(a),a)$.

If, for some a, no such $y_0(a)$ exists, then $P_H(Y\le y_0\mid A=a)=0$ and is not the cp-value of any possible data.

Removing the expressions that are equal to *zero*, we can re-write $p(y_0,a_0)$ as:

$$p(y_0,a_0)=\int_{a:y_0(a)\text{ exists.}} cp(y_0(a),a)\cdot f_A(a)da.$$

This is not necessarily a mean over a since it may not cover all possible values of a (i.e. $\int_{a:y_0(a)\text{ exists.}} f_A(a)da$ may be less than *one*).

Cox's case.

A special case occurs when the conditional distributions of Y given the various values of A all have the same support, i.e. every possible value of Y is observable with each and every different value of A. (This can not happen if A is a function of Y [27].) In such a case, for all y_0 and a, $y_0(a)$ exists and is equal to y_0. This is true in Cox's case where the two values of a correspond to Normal populations with different variances. Any two hypotheses about the mean of a Normal population will produce likelihood ratios everywhere in the interval $(0,\infty)$ regardless of the variance; thus $y \in \mathbb{R}^+$ for both $a=1$ and $a=2$, and hence

$$y_0(a) = \max_{y \le y_0}\{y : (y,a) \text{ corresponds to some observable data}\} = y_0,\ \forall a .$$

As a result,

$$p(y_0, a_0) = \int_a cp(y_0, a) \cdot f_A(a)da, \quad \forall y_0, a_0.$$

The p-value of y_0 is thus the weighted mean of the conditional p-values of y_0 so that the conditional p-values of y_0 for all (i.e. both) the possible values of a vary around the unconditional p-value, one being larger and one smaller than the unconditional value.

In general, however, all we can say is that

$$p(y_0, a_0) = \int_{a:y_0(a)=y_0} cp(y_0, a) \cdot f_A(a)da + \int_{a:y_0(a)\neq y_0} cp(y_0(a), a) \cdot f_A(a)da .$$

The first part of the sum integrates over those values of a for which $y_0(a) = y_0$, i.e. values of a that are consistent with the observation of y_0; the second part integrates over those values of a that are not consistent with y_0 but are consistent with some value of y less than y_0 (and hence $y_0(a)$ exists); the cp-values in this component are not cp-values *of* y_0. The weights, $f_A(a)$, in these expressions will not necessarily sum to *one* across the two integrals since there may be values of a (with non-zero

[27] If A is a function of Y, there is no overlap between the supports of the conditional distributions of Y; in other words, A partitions the unconditional support of Y.

density or probability) for which no $y_0(a)$ exists. All the cp-values of y_0 are included in the first integral. Given the varied nature of the rest of the expression and of the weights in both parts, we cannot make any general statement about the relationship between the conventional p-value, $p(y_0, a_0)$, and the conditional p-values, $cp(y_0, a_0)$; in particular, it is not necessarily the case that the cp-values vary around the p-value.

We have already shown that the cp-values do not vary around the p-value in the Normal case where $A = |\ln Y|$; in this instance, A partitions the (unconditional) support of Y so that any value of y is consistent with only *one* value of a and $y_0(a) = y_0$ only if $a = |\ln y_0|$. (In fact, for $y_0 < 1$, $y_0(a) = e^{-a}$ where $a \leq |\ln y_0|$ and is non-existent otherwise.)

Contrast with significance levels.

When it comes to significance levels the situation is a little different. A test that rejects H in favour of K whenever $Y \leq y_c$, has a significance level (conditional or unconditional) that is calculated by reference to the fixed critical likelihood ratio value, y_c, regardless of the data observed. Thus the unconditional significance level is the mean of all the possible conditional significance levels, since

$$\alpha = P_H(Y \leq y_c) = \int_a P_H(Y \leq y_c \mid A = a) \cdot f_A(a) da = \int_a \alpha_a \cdot f_A(a) da .$$

The difference is due to the fact that $P_H(Y \leq y_0 \mid A = a)$ is a cp-value of y_0 only if it is possible to observe (y_0, a), whereas $P_H(Y \leq y_c \mid A = a)$ can be regarded as the significance level of the test, conditional upon $A = a$, even when $A = a$ is incompatible with $Y = y_c$ or, indeed, $Y \leq y_c$. This approach is correct in terms of identifying the failure rates associated with different values of A.

We have shown that for the Normal and Cauchy location models, the cp-value is always greater than the conventional p-value. We now show that this is true for any scenario that is log-symmetric.

Proof that the cp-value is greater than the p-value in all log-symmetric cases.

Let y_U be the least upper bound on the support of the unconditional distribution of Y, i.e. it is the smallest value of y such that $P_H(Y \le y) = 1$[28]. (In the Normal case $y_U = \infty$.)

Claim.
In any case where $A = |\ln Y|$ has the same distribution under hypotheses H and K, it follows that $\forall y < y_U$ the p-value conditional upon $A = |\ln y|$ will be greater than the unconditional p-value of y, i.e. $cp(y) > p(y)$.

Proof.
Assuming that Y is not degenerate at *one*, $y_U > 1$ because, otherwise, $f_H(y) \le f_K(y)$ everywhere, and they cannot both integrate to one.

In any log-symmetric case,

$$cp(y) = \begin{cases} 100\%, & y \ge 1 \\ \frac{y}{(1+y)}, & y < 1. \end{cases}$$

First note that if $y = y_U$ then $cp(y) = p(y) = 100\%$ so, even in this case, $cp(y) \nless p(y)$. If $1 \le y < y_U$, then, $cp(y) = 100\%$ while $p(y) < 100\%$ (since $y < y_U = \min_y \{y : p(y) = 100\%\}$) and hence $cp(y) > p(y)$.

Now consider the case where $y < 1$. (The following is written in terms appropriate for a continuous A; for the discrete case replace the integral and density with a sum and probability.)

[28] In the next chapter we will show that the cdf of Y under K is always greater than the cdf of Y under H, thus $P_H(Y \le y) = 1 \Rightarrow P_K(Y \le y) = 1$.

$$\begin{aligned} p(y_0) &= P_H(Y \le y_0) \\ &= \int_a \vec{P}_H(Y \le y_0 \mid A = a) \cdot f_A(a)da. \end{aligned}$$

If $y < 1$, then $a = |\ln y| = -\ln y$. Thus, if $y_0 < 1$, it follows that $y \le y_0$ if and only if $a = -\ln y \ge -\ln y_0$. Hence it follows that:

$$y_0(a) = \begin{cases} e^{-a}, & a \ge |\ln y_0| = -\ln y_0 \\ \text{non-existent, otherwise.} \end{cases}$$

Thus

$$\begin{aligned} p(y_0) &= \int_{a:y_0(a)\text{ exists.}} cp(y_0(a)) \cdot f_A(a)da \\ &= \int_{a \ge -\ln y_0} cp(e^{-a}) \cdot f_A(a)da \\ &= \int_{a \ge -\ln y_0} \frac{e^{-a}}{(1+e^{-a})} \cdot f_A(a)da \\ &\le \frac{\displaystyle\int_{a \ge -\ln y_0} \frac{e^{-a}}{(1+e^{-a})} \cdot f_A(a)da}{\displaystyle\int_{a \ge -\ln y_0} f_A(a)da} \\ &= E\left(\frac{e^{-A}}{(1+e^{-A})} \mid A \ge -\ln y_0 \right) \\ &= E(h(A) \mid A \ge -\ln y_0) \end{aligned}$$

Now note that $h(a) = \frac{e^{-a}}{(1+e^{-a})}$ is a decreasing function of a and thus, when a is in the range $[-\ln y_0, \infty)$, the maximum value of $h(a)$ is $h(-\ln y_0) = \frac{y_0}{(1+y_0)}$. Since this is the maximum value, it follows that the mean must be less than it, i.e. $E(h(A) \mid A \ge -\ln y_0) < \frac{y_0}{(1+y_0)}$. Hence $p(y_0) < \frac{y_0}{(1+y_0)} = cp(y_0)$.

Thus, in all log-symmetric cases, the conventional p-value of any observation overstates the significance of the data, in that, it is smaller than the relevant p-value of the data.

8.12 Comments and summary.

In this chapter we have described a number of scenarios where the statistic $A = |\ln Y| = |\ln\{LR(\underset{\sim}{X};\theta_1,\theta_2\}|$ is ancillary, in the restricted sense, on the binary parameter space $\Theta_B \equiv \{\theta_1,\theta_2\}$, that is, it has the same distribution $\forall\,\theta \in \Theta_B$ and is a function of the MSS for $\theta \in \Theta_B$ (i.e. Y). In all but the Bernoulli case this is true for all possible binary subsets of the natural (or conventional) parameter space and thus it is possible to define the conditional confidence interval based on A (§8.8) as well as conditional hypothesis tests. Of the cases examined here, the Normal location case is the most general and useful, because it applies to random samples of all sizes through the relation $X_i \sim N(\mu,\sigma^2) \Rightarrow T = \overline{X} \sim N(\mu,\frac{\sigma^2}{n})$.

In all the scenarios covered in this chapter, a single relationship connects the p-value of the data, *conditional upon* a, to the likelihood ratio of the data, in stark contrast to the lack of association between the conventional p-value and likelihood ratio. It follows that, in each case, any given value of the likelihood ratio is interpreted (via the cp-value) *the same way*, despite the differences between the various model/Θ_B combinations.

The conventional significance level of any test can be understood as the before-experiment expected value, over a, of the conditional significance levels associated with the various possible values of a. We have illustrated this relationship in the Normal (§8.6) and Cauchy (§8.9) cases. The p-value is also the expected value (over a) of certain conditional probabilities, which include the cp-value.

Whenever data, $\underset{\sim}{x}$, is in the α-level rejection region of a test but does not constitute strong evidence against H relative to K (i.e. $LR(\underset{\sim}{x}) \lll 1$), it follows that the relevant significance level of the test, α_a (where $a = |\ln LR(\underset{\sim}{x})|$), is unreasonably large, even though α is not (see ***Example 8.1***). Thus the conditional significance level of a test is consistent with our intuitions about the quality of that test.

An important feature of many ancillary statistics is that they function as ‘precision indices’, distinguishing between categories of more and less informative data (see Buehler (1982) for a formal definition of this concept). In this connection we remark that, if Y is a measure of the weight of evidence in favour of H relative to K, then $A = |\ln Y|$ is a measure of the absolute weight of evidence between the hypotheses (one way or the other) – each value of a being associated with a set of data points $\{\underset{\sim}{x} : A = a\}$ that is homogeneous in terms of the weight of evidence between the two hypotheses. A also indicates the size of the conditional error probabilities. For any a, if $y_1 < 1$ is a solution of the equation $|\ln y| = a$, then the test that rejects H if and only if $y = y_1$ is the only non-trivial conditional test of H versus K (i.e. the only test where neither of the conditional error probabilities is 100%). Such a test has equal conditional error probabilities and these are decreasing in a; when a is larger we can test both H (as null) versus K and K versus H and achieve lower error probabilities than when a is smaller – thus we might describe A as an ‘error index’ and this can also be seen as a measure of precision.