Chapter 3: Some problems with evidential frequentist inference.

3.1 Tests of two simple hypotheses.

Consider testing hypothesis H against hypothesis K, where H and K are both simple hypotheses that (together with the background assumptions) completely specify the distribution of the test statistic.

The trial interpretation.

There is a widespread view, encouraged by textbooks, that hypothesis tests are so constructed that H is only rejected when there is very strong evidence against it. Consistent with this view, hypothesis testing is sometimes described as analogous to a criminal trial where the null hypothesis says that the defendant is innocent and the alternative hypothesis says otherwise. In this construction, the null hypothesis is the default or starting position and can only be overthrown by very strong evidence against it. In a section entitled "Significance testing procedures: statistician as juror", Smith writes "In all significance testing procedures, H_0 is assumed to be true until the test statistic indicates otherwise, beyond reasonable doubt.[1]".

Similarly Efron states,

> **…because there is a vested interest in discrediting H, conservative statistical methods have been developed which demand a rather stiff level of evidence before H is declared invalid. The frequentist theory, which is dominant in hypothesis testing, accomplishes this by requiring that the probability of falsely rejecting H in favour of K, when H is true, be held below a certain small level, usually .05.[2]**

[1] Smith, p.516.
[2] Efron, p. 241.

Example 3.1

Suppose we are interested in the mean height (cm) of the males in a particular population where the heights of males can be assumed to be Normally distributed and the standard deviation is known to be 8cm. We have a random sample of $n = 16$ individuals whose heights we have measured producing an average of 167.9cm.

The previous generation who grew up with food shortages had a mean height of 163cm; we wish to test the hypothesis of no change against a claim that the population mean has increased dramatically, by 16cm, to 179cm.

We have independent and identically distributed random variables, $X_1, \ldots, X_{16}$, where $X_i \sim N(\mu, 8^2)$ and data $\overline{x} = 167.9$.

Suppose we carry out a test of H: $\mu = 163$ versus K: $\mu = 179$ at the 1% significance level, then we will reject H if $\overline{x} \geq 163 + z_{0.99}\frac{8}{\sqrt{16}} = 167.653$ (where $z_\gamma = \Phi^{-1}(\gamma)$). Since $\overline{x} = 167.9$, we will reject H in favour of K even at the 1% level. However, if we look at the graphs below, it does not appear that this data provides evidence beyond reasonable doubt that μ equals 179 rather than 163. On the contrary, the data (167.9) is clearly more consistent with H than with K.

Figure 3.1

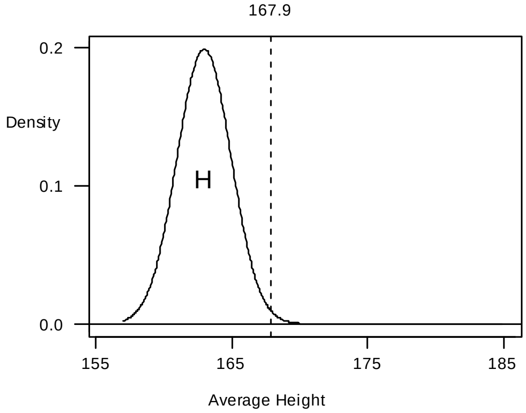
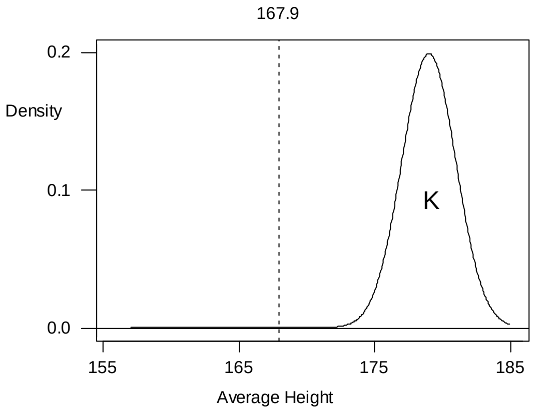


Figure 3.2



According to Kendall and Stuart, we should stick with H under these circumstances:

It is perfectly possible for a sample of observations to be a rather "unlikely" one if the original hypothesis were true; but it may be much more "unlikely" on another hypothesis. If the situation is such that we are forced to choose one hypothesis or another, we shall obviously choose the first, notwithstanding the "unlikeliness" of the observations. The problem of testing a hypothesis is essentially one of choice between it and some other or others.[3]

Bias in favour of K.

Here we have an example where, despite the small p-value (less than 1%) we do not have strong evidence against H, relative to K, and any test that tells us to reject H in favour of K under these circumstances is clearly biased[4] in favour of K.

The idea that hypothesis tests are automatically biased in favour of H so that H is like the presumption of innocence – only rejected when the evidence against it (relative to

[3] Kendall & Stuart, p. 164.

[4] We use the term 'biased' because it is by far the most appropriate. Note that this is not the same notion of bias usually referred to in the term 'biased test' (although our definition also applies to tests); the conventional meaning is $\exists \theta_i \in \Theta_K : \beta(\theta_i) > 1 - \alpha$, which is less clear cut than the form of bias examined in this section.

K) is very strong – is clearly false. In most of the examples shown in introductory textbooks, the null and alternative hypotheses are fairly close together relative to the spread of the test statistic. In practice hypotheses are just as likely to be far apart. In the Normal case, the standard deviation of the test statistic, $\bar{X}$, is $\sigma/\sqrt{n}$ and the difference between the hypothesised means, $|\mu_1 - \mu_2|$, will be large in terms of $\sigma/\sqrt{n}$ whenever n is sufficiently large – with enough data, any two hypotheses can be 'far apart'. For a test of two hypotheses (H: $\mu = \mu_1$ versus K: $\mu = \mu_2$) with known σ, it seems reasonable to say that when the cut-off value (in terms of $\bar{x}$) is closer to μ_1 than to μ_2 then the test is biased in favour of K and that the test is only biased in favour of H when the cut-off value is closer to μ_2 (than μ_1). Since the cut-off value (for a right-sided test) is $\mu_1 + z_{1-\alpha}\,\sigma/\sqrt{n}$, the test will be biased in favour of K whenever

$$(\mu_1 + z_{1-\alpha}\,\sigma/\sqrt{n}) < (\mu_2 + \mu_1)/2\,.$$

For fixed values of μ_1, μ_2, σ and α, this occurs whenever n is sufficiently large,

$$\text{i.e. } n > \frac{4z_{1-\alpha}^2\sigma^2}{(\mu_2 - \mu_1)^2}\,.$$

In *Example 3.1*, the key to the existence of the bias in favour of K and the misleading nature of the test result lies in the choice of significance level. A significance level of 1% is fairly low, but the probability of Type II error (β) was much lower $(7 \times 10^{-7})\%$. We make α small in order that the null hypothesis will be wrongly rejected in only a small proportion of cases. Sometimes this means that β (=1-power) is quite high and $\alpha < \beta$ in which case the test is biased in favour of H[5]. However, keeping α small does not guarantee that the test will be biased in favour of H or even that the test will *not* be strongly biased in favour of K. Although β is a decreasing function of α (so that, all other things being equal, making α smaller will

[5] For the *Normal* case, the following three conditions are equivalent and can be interpreted as 'bias in favour of K':

- c is closer to μ_1 than μ_2, where c is the critical value for $\bar{x}$.
- $\alpha > \beta$,
- $LR(c) > 1$.

make β bigger), β is also a function of other factors such as σ and n, so it is perfectly possible for a small α to produce an even smaller β. When $\beta < \alpha$ we are more likely to wrongly **reject** H than to wrongly **accept** H indicating that the bias is now in favour of K. It is odd that this problem typically occurs when n is large. In such a case we have a lot of information as indicated by the fact that we are able to produce a test with both error rates so low. Here, if anywhere, we ought not to have a problem.

The difficulty arises because the theory seems to have been devised to deal with the challenges associated with limited data. In such a case, it would be easy to end up with a high significance level. The theory emphasises that we must not do this, that there may be serious consequences associated with wrongly rejecting H. (In fact, many books encourage us to identify the null hypothesis by asking which hypothesis we would be more afraid to wrongly reject; this is often further complicated by the view that placing the burden of evidence on any new claim should dominate this judgement. Thus the common identification of the null hypothesis with '*no* effect' is justified on the basis that we would not want a finding that (say) a new drug is more effective when it is not, although it is not clear that the ill effects of adopting a new drug, that is no more effective than the old, are really greater than the ill effects of failing to identify a more successful new drug.) For this reason we must always choose α to be appropriately small – at most 5%, but 1% or lower if the consequences of wrongly rejecting H are particularly bad. This emphasis ignores the fact that we will sometimes have a lot of data or measurements from an extremely accurate machine.

It is tempting to assume that an approach that works well with limited data will work even better with large amounts of data and this should be the case, the fact that it is not points to a flaw in the method. The problem lies with fixing α; α is not allowed to exceed 5%, but, to prevent the test from being biased against H, it is necessary, not only that α be small, but that it be at least as small as β. In the Normal distribution case, the requirement $\alpha \leq \beta$ is equivalent to the requirement that $LR(c) \leq 1$ (where c is the critical value for $\overline{x}$). (For models with skewed distributions these two criteria may not be exactly equivalent and 'critical likelihood ratio less than one' may be preferred over $\alpha < \beta$ as the definition of bias in favour of H. However this requirement can generally be satisfied by making α 'sufficiently much' less than β.) If we want there to be no bias in favour of K, we must make sure that α is extremely small whenever we have a very large amount of information. Yet we choose α in advance and textbooks do not suggest that a significance level of, say, 10^{-20} might be appropriate in certain circumstances, simply to prevent bias in the test. In the *heights* example, using the rejection region $[171,\infty)$, instead of $[167.653,\infty)$ would have given us $\alpha = \beta = 3.2\times10^{-5}$ instead of $\alpha = 10^{-2}$, $\beta = 7\times10^{-9}$. This would be a particularly reliable test, since both the error probabilities are so low, and surely this is what we should expect to happen when the hypotheses are so far apart and so easy to distinguish from each other. When n is very large (or for some other reason it is easy to choose between the two hypotheses) the significance level should automatically become very small, but α is chosen from a conventional range of values according to the seriousness of the Type I error and no other criteria. The optimality in Neyman-Pearson inference applies to β only, for a pre-determined, fixed value of α. There is no sense in which the value of α is optimised. In the *heights* example we see that using $\alpha = 1\%$ – usually regarded as a low value – forces us to reject H in favour of K although the data is much more consistent with H; we can easily create similar examples for any α (no matter how small) by making the two hypotheses far enough apart. The problem is not solved by using p-values instead of fixed α levels since it becomes necessary for us to rethink the idea that small p-values indicate evidence against H relative to K; in certain circumstances a p-value of 10^{-10} may not be at all significant in this sense.

In the 1967 edition of their classic text Kendall & Stuart discussed this problem and came to the following conclusion:

The [null] hypothesis tested will only be rejected … if we keep α fixed as n increases. There is no reason why we should do this: we can determine α in any way we please, and it is rational … to apply the gain in sensitivity arising from increased sample size to the reduction of α as well as of β. It is only the habit of fixing α at certain conventional levels which leads to the paradox. If we allow

α to decline as n increases, it is no longer certain that a very small [i.e., 'small' by comparison with the difference between the two hypotheses] departure from H_0 will cause H_0 to be rejected: this now depends on the rate at which α declines.[6]

The 1991 edition added the following sentence to this passage:

A reasonable, though arbitrary, solution is to make α equal to β at the smallest departure from H_0 that is of practical importance.[7]

(Thus, in the previous example, we would place c exactly midway between the two hypothesised values.)

It follows from this that a given 'significant' p-value implies evidence against H relative to K to a greater degree when the sample size is small; this is consistent with the view of Lindley & Scott. It is just as strong a counter-example to Fisherian methods as to those of Neyman and Pearson since it contradicts Fisher's belief (still widely accepted) that the p-value alone allows one to make evidential inferences. Despite the prominence of Kendall & Stuart's text, this particular piece of advice has fallen on deaf ears for forty years; no textbook advises students to draw their selection of significance level from anything other than the small range of conventional values usually considered appropriate, nor is there any indication that the value of n is relevant to choosing α.

Two further points need to be made. Clearly it is not enough to base α on n alone since other reliability factors are equally relevant. In the Normal case, the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$ and is thus equal to *one* regardless of whether we have $\sigma = 20$ and $n = 400$ or $\sigma = 2$ and $n = 4$. If $|\mu_1 - \mu_2| = 10$, then in both cases the hypotheses are *ten* standard deviations apart suggesting that α will need to be extremely small if the test is not to be heavily biased in favour of K, yet in the second case we have a sample size of *four* which would not normally be regarded as 'large'.

[6] Kendall & Stuart, p. 183.
[7] Stuart, *et al*., p. 193.

Also, for fixed values of n and σ, the value $|\mu_1 - \mu_2|$ is important in choosing an appropriate value of α. If we want all our tests to have bias of the same level and direction, we must choose α based on these three factors and this raises the possibility that we might not see a given value of α as having any particular significance or interpretation. Thus suppose we were to carry out two tests: one of H1: $\mu = \mu_1$ versus K1: $\mu = \mu_2$, and one of H2: $\mu = \mu_3$ versus K2: $\mu = \mu_4$, where n and σ are the same for both tests but $|\mu_1 - \mu_2| \neq |\mu_3 - \mu_4|$. If we want the bias in favour of H1, in test 1, to be of the same magnitude as the bias in favour of H2, in test 2, we will need to choose different values of α for the two tests. If we then (say) reject H1 in favour of K1 and reject H2 in favour of K2, it is tempting to attach exactly the same evidential significance (however defined) to the rejection of the null hypothesis in both cases; thus the significance level is no longer the significant measure even when the sample sizes are the same, and the precise nature of the alternative hypothesis has much more influence over the test result than is currently the case.

Here we come close to the boundary between competing theories of inference. Suppose that we adjust α in the way suggested by Kendall and Stuart (or a more sophisticated version); is this an innocuous alteration to the theory, or does it cut at the heart of the assumptions behind using frequentist inference? On the face of it we may adjust the significance level to take into account anything, including our ability to discriminate between the hypotheses, without conflicting with the theory of Neyman and Pearson – we understand that the choice of α is ours to make even if we do not take advantage of this freedom often. However, if we want to calculate the value of α so that there is no bias in favour of K, we start to wander into the territory of alternative inferential theories. In a case where the usual values of α all produce tests biased (to some degree) in favour of K, we may decide to make α still smaller – for instance, just small enough so that there is no bias in the test. This will amount to choosing the critical value, c, according to the formula $LR(c) = 1$; the value of c so derived will then dictate the values of α and β. Insofar as we are making the critical value of the likelihood ratio (in this case *one*) dominate α and β, this is coming close to being a *likelihood method* rather than a frequentist method. The modification has substantially changed the emphasis of the theory: we are now making α extremely small, *not* because the consequences of wrongly rejecting H are very serious (the justification allowed by Neyman-Pearson theory), but rather because using a conventional value of α will cause the test to give us ‘misleading’ results. These results are not misleading in the sense that they miscalculate the error probabilities (which are at the core of Neyman-Pearson inference), nor are the probabilities unreasonably large; it is simply that the conventional choice of α gives intuitively silly results. If we decide that data lying in the 1% (optimal) rejection region does not necessarily justify rejecting H but that data with a low likelihood ratio does, we have abandoned frequentist inference entirely.

Insensitivity to K.

In this discussion, we talked in terms of the evidence for either one of the hypotheses relative to the other because this is what those who carry out hypothesis tests are primarily interested in. While most of us recognise that the basis of this type of inference is controlling the error probabilities so that the test results will not be too often wrong, we tend to believe that we can use this mechanism to find out what the data has to say about the hypotheses. The following feature of tests throws doubt on this belief.

Suppose that $T \sim N(\mu, 1)$ and we want to test H: $\mu = 0$ against various alternative hypotheses. For data $t = 2$, the results are as follows.

i. Reject H: $\mu = 0$ in favour of K1: $\mu = \frac{1}{10}$ at the 5% level.
ii. Reject H: $\mu = 0$ in favour of K2: $\mu = 3$ at the 5% level.
iii. Reject H: $\mu = 0$ in favour of K3: $\mu = 10$ at the 5% level.
iv. Reject H: $\mu = 0$ in favour of K4: $\mu = 10^{20}$ at the 5% level.
v. Accept H: $\mu = 0$ as opposed to K5: $\mu = -\frac{1}{4}$ at the 5% level (or any level less than 98%).

Clearly our observation, $t = 2$, favours $\mu = 0$ over $\mu = -\frac{1}{4}$, but it favours $\mu = 0$ over $\mu = 10^{20}$ even more strongly, so why do we accept H in the former case but reject it in the latter? The formal answer, of course, is that all these inferences are consistent with maintaining a (long-run) Type I error rate of 5%. However, the power and also (as discussed above) the bias between the two hypotheses, varies with K.

In Neyman-Pearson theory, we reject H in favour of K if the likelihood ratio of the observed data is *relatively* small (in the sense that it is in the smallest α -proportion of values that we would observe, in the long-run, under H), rather than if it is *actually* small; for instance, we reject H if the likelihood ratio it is in the 'smallest 5%' of likelihood ratios produced under H. When we change the alternative hypothesis but keep the null hypothesis the same, the likelihood ratio of each value of the test statistic also changes. However, in a case like that above, as long as the value specified by K remains on the *same side* of the null value, the likelihood ratios remain in the same order. That is, for any values of the test statistic, say t_1 and t_2, if

$$LR(t_1; \mu_H, \mu_{K1}) < LR(t_2; \mu_H, \mu_{K1})$$

then

$$LR(t_1; \mu_H, \mu_{K2}) < LR(t_2; \mu_H, \mu_{K2}) .$$

Thus the critical value of the test statistic will remain the same although the critical value of the likelihood ratio will change, while still being the value on the boundary of the smallest 5%.

If, in the above case we switch sides so that we test H: $\mu = 0$ against a negative alternative, the order of the likelihood ratios (in terms of the test statistic) will be reversed so that

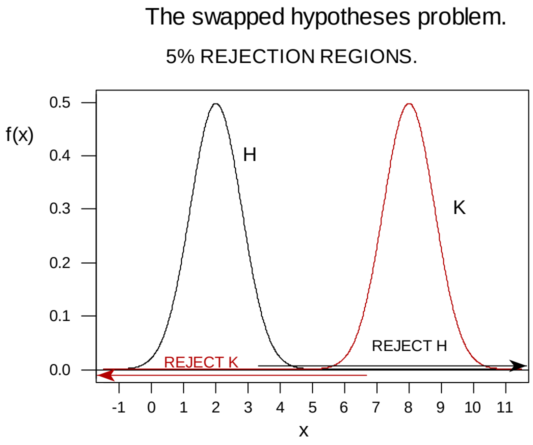
$$\begin{aligned} &LR(t_1; \mu_H, \mu_{pos}) < LR(t_2; \mu_H, \mu_{pos}) \\ &\Leftrightarrow LR(t_1; \mu_H, \mu_{neg}) > LR(t_2; \mu_H, \mu_{neg}). \end{aligned}$$

The critical value of the test statistic will change because the values that were formally associated with small values of the likelihood ratio will now be associated with large values of the new likelihood ratio, and vice versa. Thus the result of a test (whether described in accept/reject terms for a fixed α or given in the form of the p-value of the data) is influenced by the precise details of K only insofar as this affects the *order* of the likelihood ratios (as a function of the test statistic) not insofar as it affects the *value* of the likelihood ratios. This leads to some counter-intuitive results including the issue of bias discussed above.

The swapped hypotheses problem.

Another feature of conventional tests that casts doubt on any evidential interpretation is that the results from swapped hypotheses, in cases where the power is high, are contradictory. Thus suppose we have a typical high power test, such as $H:\mu=2$ versus K: $\mu=8$, based on $T \sim N(\mu,1)$. Then the 5% BCR for rejecting H (as null hypothesis) in favour of K overlaps with the 5% BCR for rejecting K (as null hypothesis) in favour of H.

Figure 3.3



Interpreted in the typical evidential fashion, we would have to say that the data, $x=5.5$, represents *both* strong evidence against H relative to K *and* strong evidence against K relative to H. This claim is not consistent with any reasonable conception of ‘evidence’. Again, it is particularly disturbing that this occurs when the sample is large, since, in such cases, we are particularly confident of our results.

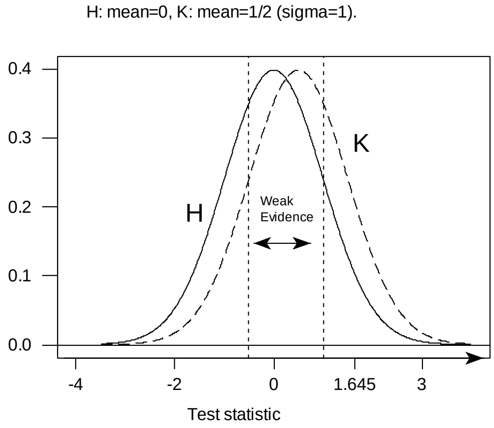
Weak evidence.

It is widely recognised that a test may give misleading results (or be misinterpreted so as to be misleading) if the power is low. Specifically, if the power of the test is low and the test result is not 'significant' (H is not rejected), then interpreting this as evidence in favour of H is inappropriate since the test is unlikely to reject H even when K is true. This has lead to calls for the compulsory reporting of power in journal articles. For many standard tests, the power declines towards α (which is small) as the hypothesised values converge and will thus be low whenever the values are sufficiently close together.

In ***Example 3.1*** the observed data was more consistent with H than K, but not very consistent with either hypothesis. We can define as giving 'weak evidence' any data which is not much more consistent with one hypothesis than with the other, i.e. data which is not much help in deciding which hypothesis is true. If a test statistic has the same support for all hypotheses (e.g. for all $\theta \in \Theta$), then, for any two hypotheses, there is some data that is weak, in this sense.

When the two hypotheses are close together, and the power of any test with a conventional significance level is low, all the weakest data will fall outside the rejection region and, when observed, will cause us to 'accept H'. For example, if our test statistic is $N(\mu,1)$, and we test H: $\mu = 0$ versus K: $\mu = \frac{1}{2}$ at the 5% level, we will reject H if the test statistic is greater than 1.645, giving the test a power of about 13%. The data that is close to $\frac{1}{4}$ (the point half way between μ_1 and μ_2) provides only weak evidence about which hypothesis is true. We could define as 'weak' all data in the range, say, $[-\frac{1}{2},1]$, and we note that this data always leads us to accept H, as shown below.

Figure 3.4

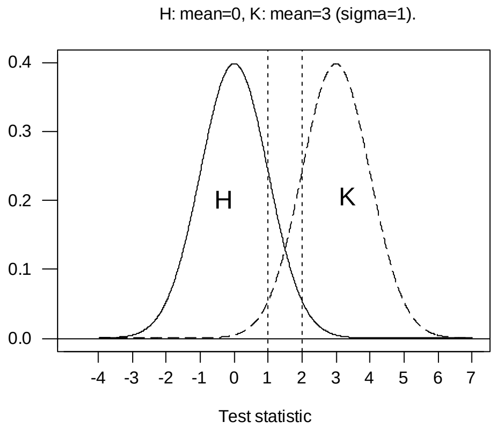


This problem may not be too serious if the analyst is conscious of the low power of the test and declines to read anything significant into the failure to reject H (although is this desirable if we observe $t = -4$?), but what about the case where the power is high?

Example 3.2

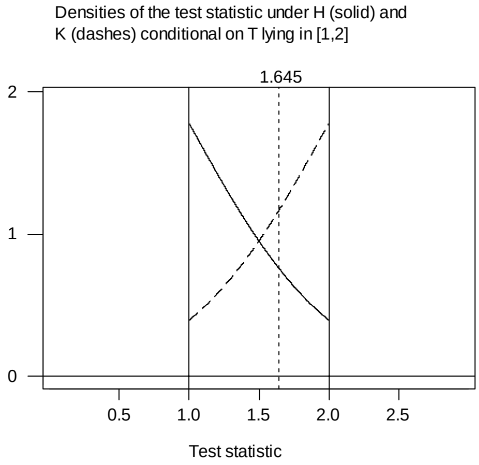
Below we show the densities relevant to the two hypotheses H: $\mu = 0$ and K: $\mu = 3$ where $\sigma = 1$. If we test these hypotheses at the 5% level, the cut-off value is again 1.645 but now the power of the test is 91.23%. Nevertheless it is clear that some data is not very helpful when it comes to choosing between H and K. When the test statistic lies in the interval $[1.0, 2.0]$, it is hard to feel that we have much evidence of which hypothesis is true. Some of the weak data in this interval lies in the acceptance region and some in the rejection region. Yet the fact that both the error probabilities are low ($\alpha = 5\%, \beta = 8.78\%$) would lead many people to regard the result of the test – accept H or reject H – with confidence; this confidence is unjustified when the data is weak and we can easily tell whether the data is weak just by examining it.

Figure 3.5



We can gain some insight into whether or not our intuition is correct by looking at the weak data in isolation. Note that, by symmetry, the data is just as likely to be weak when H is true as when K is true. We can calculate the error probabilities for this test *given that* the data is weak.

Figure 3.6



Given that the data is weak, the probabilities of Type I error, α_W, and Type II error, β_W, are:

$$\alpha_W = P_H(T > 1.645 \mid 1 < T < 2) = 20.05\%,$$
$$\text{and } \beta_W = P_K(T < 1.645 \mid 1 < T < 2) = 47.78\%.$$

These values are consistent with our intuition that when the data is in the region [1,2] we cannot infer much about which hypothesis is the true one. It seems that making allowance for low power is not enough to prevent us from misinterpreting test results. The belief that we can place great confidence in a test result whenever α and β are both small, seems to imply that, in such a case, all data is bound to be highly informative, yet this is not true. When the power is high, the usual interpretations of significant as well as non-significant results may be misleading.

In the low-power example given earlier $(\mu = 0$ versus $\mu = \tfrac{1}{2})$, data in the rejection region does constitute genuine evidence against H relative to K. As long as we refuse to be influenced by non-significant results (as is sometimes advised), we will not be misled (the weak data is all confined to the acceptance region). However, even when the power is low, this is not always the case; in the following example the power is low but data in the rejection region cannot reasonably be interpreted as evidence against H.

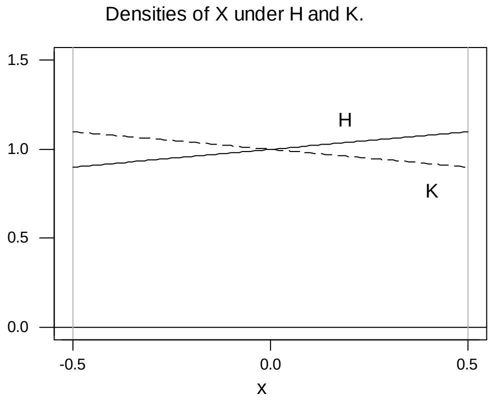
Example 3.3

Suppose that the distribution of a random variable X is dependent on a parameter, θ, through the model

$$f(x;\theta) = 1 + \theta x,\ x \in (-\tfrac{1}{2}, \tfrac{1}{2}),\ \theta \in (-2, 2).$$

The densities of X under H: $\theta = 0.2$ and K: $\theta = -0.2$ are shown below.

Figure 3.7



The two distributions are very alike so that a single observation on X is a poor basis on which to test these hypotheses; in this case all the data is weak. The 5% (optimal) rejection region is $(-0.50, -0.44)$ and the test has a power of 6.04% – extremely low, as we would expect. The low power of the test indicates that we can read nothing into a failure to reject H. However, comprehending this point will not necessarily prevent us from misinterpreting the data for, if $x < -0.44$, we can reject H at the 5% level and yet such data in no way indicates a reasonable level of evidence against H relative to K. Whenever the likelihood ratio statistic for a test is a continuous variable, it is possible to define a BCR associated with a low significance level. This is true even when no data constitutes strong evidence against H relative to K. In such a case, there is weak data in both the rejection and acceptance regions and neither result is reliable. This is true notwithstanding the fact that the test has a genuinely low significance level, indeed, this phenomenon can occur for arbitrarily small significance levels.

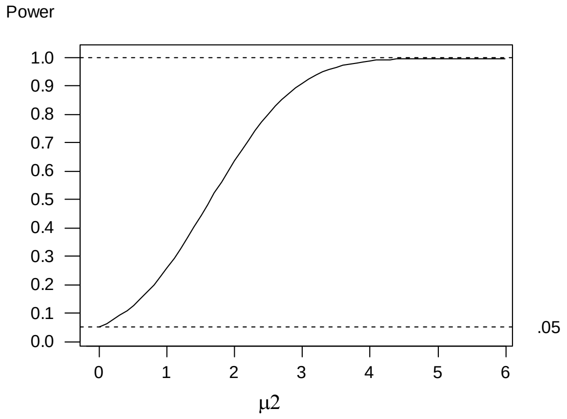
3.2 Tests against a composite alternative hypothesis.

One-sided tests.

Tests involving a composite alternative hypothesis are the most commonly performed, freeing the scientist from the need to be precise about what effect size is of interest. In the heights example, we could have tested $\mu = 163$ against the alternative $\mu > 163$ in order to determine whether the mean height had increased rather than trying to determine whether it had increased by a specific amount.

Suppose again that $T \sim N(\mu, 1)$ and we are testing H: $\mu = \mu_1$ against the one-sided alternative K: $\mu > \mu_1$. We can consider the various components that go to make up K. Since all the components of K produce the same α -level rejection region when used as simple alternative hypotheses, the same region is used for an α -level test of H against the composite K. The power of this test will vary depending on which component of K we look at and so we now treat it as a function of μ ($\mu \in \Theta_K$). Denote the power function by $\kappa(\mu) = 1 - \beta(\mu)$. The power will be low for alternative values close to the null value, i.e. $\kappa(\mu) \to \alpha$ as $\mu \to \mu_1$, and higher for values further away, i.e. $\kappa(\mu) \to 1$ as $|\mu - \mu_1| \to \infty$. Shown below is a plot of the power function for the test H:$\mu = 0$ versus K:$\mu > 0$ when α is 5%.

Figure 3.8



This test is uniformly most powerful; nevertheless the power is obviously very low for some values of μ (those close to the null value) so that accepting H is usually not regarded as evidence against these values.

We noted in the previous section, that the traditional results (based on any data) make more sense for some components of K than for others. When we observe the data $t = 2$, we will reject H in favour of K at the 5% level. When we think of *zero* versus the particular component *one-tenth*, the data is more consistent with the latter but surely not by very much (nothing like 'beyond reasonable doubt'); *zero* versus *two* is where it is most convincing to reject H since (among other reasons) *two* is the maximum likelihood estimate of μ from this data; for *zero* versus *ten*, the rejection seems wrong and even more so for more extreme alternatives. We could describe the test as having different biases for the different components of K just as it has different power values. For elements of K close to *zero*, the bias favours H, this bias weakens as the alternative gets further away from *zero* until it eventually turns into a bias in favour of the component of K – a bias which gets steadily stronger for larger and larger alternative values of μ. We could therefore argue that, since the bias in favour of K is strong for values of μ far away from the null value, it is wrong to regard the rejection of H as evidence in favour of these values relative to the null value.

If there was no bias in the tests of simple hypotheses, or if we insisted on having exactly the same bias for all tests, then the rejection regions would differ (over the components of K) and this would make testing a composite hypothesis far more complex. Thus the fact that many different alternative hypotheses all give the same result is very convenient although it does not make a lot of sense.

If we have doubts about the method or interpretation used in the testing of two simple hypotheses, these doubts must extend into the testing of a one-sided composite alternative. Aside from this, there is a separate question about the way in which we convert tests of simple hypotheses to a test of a composite alternative. Suppose we accept, for the moment, that the results of the simple hypothesis tests are appropriate, do we then approve of the mechanism for testing composite hypotheses, and if so, why? With all the simple alternative hypotheses on the 'same side' of the null hypothesis yielding the same result, our rejection rule for testing a one-sided composite alternative is equally consistent with two very different approaches:

a) We test $\mu = \mu_1$ against $\mu = \mu_2$ (at level α), for each $\mu_2 \in \Theta_K$, and if they *all* lead to the rejection of H, we reject H in favour of K at the appropriate level.

b) We test $\mu = \mu_1$ against $\mu = \mu_2$ (at level α), for each $\mu_2 \in \Theta_K$, and if *any one of them* leads to the rejection of H, we reject H in favour of K at the appropriate level.

For method (a), the rejection region for testing the composite hypothesis is the intersection of all the rejection regions for the individual tests, so the significance level will be some value $\le \alpha$. For method (b), the rejection region is the union of all the rejection regions for the individual tests, so the significance level will be some value $\ge \alpha$. In the case of one-sided alternatives the individual rejection regions are all the same and the intersection is the same as the union and hence the significance level is α.

Of these two methods for dealing with composite hypotheses, the first is by far the more convincing; it is reasonable to suppose that we can lump together any options which all lead to the same conclusion. For instance, if incomplete information about an unknown animal suggests that it is more likely to be *any* given species of reptile than to be a duck, then we can safely say that the evidence points to a reptile rather than a duck. But if the evidence is more consistent with a crocodile than a duck and yet more consistent with a duck than with a blue-tongued lizard, we cannot make the same statement with any confidence. Method (b) seems like a poor approach to testing composite hypotheses, a point that becomes relevant when we try to test two-sided alternatives.

Two-sided tests.

The rejection region for testing H: $\mu = 0$ against the two-sided alternative K: $\mu \neq 0$, is the union of the rejection regions for the tests:

i. $\mu = 0$ versus $\mu > 0$, and
ii. $\mu = 0$ versus $\mu < 0$.

These in turn are each based on the (common) rejection region for simple alternatives discussed previously.

For example, when testing a Normal mean, we may reject H in favour of K if $t < -1.645$ or $t > 1.645$. The significance level is 10% rather than the 5% that would apply separately to the two one-sided tests. The test is optimal in the sense that it is uniformly most powerful unbiased (and there is no uniformly most powerful test for this K).

This approach is based on method (b). It cannot be based on (a) because, for conventional α, no data which causes us to reject H in favour of a left-sided alternative will also cause us to reject H in favour of a right-sided alternative, therefore the premise of (a) can never occur for a two-sided K. The particular version of (b) used for two-sided tests is as follows:

c) We test $\mu = 0$ against $\mu > 0$ and we test $\mu = 0$ against $\mu < 0$ (both at level $\alpha / 2$) and if *either one of them* leads to the rejection of H, we reject H in favour of K at the α level.

This approach is counter to common sense, since, when we reject H, we always do so in favour of K ($\mu \neq 0$) rather than in favour of whichever component of K caused the rejection – information readily available. (As always in Neyman-Pearson inference, the error probabilities are correct as statements of the long-run failure rates in repeated applications of the method; we dispute their evidential interpretation, not their design significance.)

What can we infer from rejecting H in favour of a composite alternative?

Despite these difficulties, tests of composite hypotheses might still be helpful as a first step towards investigating whether or not H is true, if rejecting H in favour of a composite K implies that the data is much more consistent with *some* component of K than with H (where H is still assumed to be a simple hypothesis), but is this true? To answer this, we need to break K up into its component simple hypotheses, since it is only when we look at two simple hypotheses that we have a natural intuitive sense of what constitutes strong evidence. If we reject H in favour of either a one-sided or two-sided K, then we would have rejected H in favour of any one of the values that lie to the right (left) of the null value, at either the α or $\alpha / 2$ level. Because of the varying degrees of bias, this does not mean that the evidence supports all these values better than H; does it necessarily support any value more than H? Suppose the parameter of interest is μ where $T \sim N(\mu, \sigma^2)$ and σ is known. The component of K that is most consistent with the data is that which corresponds to t, the observed value of the test statistic (for instance the sample mean). We reject H: $\mu = \mu_1$ in favour of (say) the right-sided[8] composite alternative ($\mu > \mu_1$) if $t > \mu_1 + z_{1-\alpha}\sigma$, in which case, we would also have rejected H in favour of K*: $\mu = t$ at the same significance level. Does this indicate that there is quite strong evidence that μ equals t rather than μ_1 ? Unfortunately not, since the test may be biased against H in favour of K*. This will be the case whenever $\alpha > \beta$, i.e. when $t > \mu_1 + 2z_{1-\alpha}\sigma$, in which case we cannot read anything into the fact that we have rejected H in favour of K*. This is not to say that the evidence never strongly supports K* over H, merely that we cannot deduce this from the test result. When $(\mu_1 + z_{1-\alpha}\sigma) < t < (\mu_1 + 2z_{1-\alpha}\sigma)$, the test rejects H even though it is not biased in favour of K*; if we had a way of measuring bias which would allow us to say (for instance) “We rejected H in favour of K* despite the fact that the test was strongly biased in favour of H”, we might choose to interpret this as evidence strongly favouring K* over H. In the absence of such a measure all we can say is that an unbiased test would have favoured K* over H, which is to say that the data was more consistent with K*, but not necessarily by very much. In Chapter 7 we will consider using the likelihood ratio as a measure of evidence and find that, by that standard, the rejection of H in favour of a composite alternative does not necessarily imply that there exists any hypothesised value that is much more consistent with the data than is the null hypothesis.

The results of orthodox ‘optimal’ hypothesis tests are usually interpreted as showing (for instance) that the set of data constitutes strong evidence against H relative to K. In this chapter we have shown that such evidential interpretations are not justified even when the power of the test is high. We will next examine the topic of *conditioning inference* before returning to consider some alternative approaches to finding evidence.

[8] Using a two-sided alternative would produce the same argument except with $\alpha / 2$ in place of α .