Foundations of Medical Statistics

Statistical Inference 7: Approximate hypothesis tests

Aims

The aim of this session is to introduce three approximate tests: the likelihood ratio, Wald and score tests.

Objectives

At the end of this session you should understand and be able to apply these tests in one-parameter contexts.

7.1 Approximate versus exact tests

In the previous session we described how to use the (log-)likelihood ratio to find the best test statistic to compare two simple hypotheses. Once we found the best statistic, we then had to 'hunt around' to obtain its sampling distribution, in order to quantify the evidence against the null hypothesis.

We restricted ourselves to simple situations: means and variances of samples from Normal distributions, when we could readily find the sampling distribution of the best test statistic.

Usually, however, finding this sampling distribution will be very difficult. One option is to use computer simulation to estimate the sampling distribution under various hypotheses. Another often simpler option is to use an approximate test. We describe three such tests in this session.

7.2 Likelihood ratio test

Recall that for testing simple hypotheses, H$_0$: $\theta = \theta_0$ vs H$_1$: $\theta = \theta_1$, the Neyman-Pearson lemma tells us that the most powerful test uses a statistic derived from the (log-) likelihood ratio:

$$l_{H_0} - l_{H_1} = l(\theta_0) - l(\theta_1)$$

When the hypotheses are composite this can be generalised in the following way. Suppose the hypotheses specify ranges, H$_0$: $\theta \in \omega_0$ vs H$_1$: $\theta \in \omega_1$, where ω_0, ω_1 specify the required subsets of the possible values of the parameter. The generalised test then compares the *maximum* value of the likelihood for each range:

$$\log \frac{\max\limits_{H_0}[L(\theta|\text{data})]}{\max\limits_{H_1}[L(\theta|\text{data})]} = \max_{H_0}[l(\theta)] - \max_{H_1}[l(\theta)] = \max_{\theta\in\omega_0}[l(\theta)] - \max_{\theta\in\omega_1}[l(\theta)]$$

Tests of this form, though not generally optimal, are usually nonoptimal in situations where no optimal test exists. They usually perform well, playing a central role in testing, as maximum likelihood estimates do in estimation.

Typically the null hypothesis is simple and the alternative is composite and two-sided:

$$H_0\colon \theta = \theta_0 \ \text{ vs } \ H_1\colon \theta \neq \theta_0$$

$$\Rightarrow \quad \max_{H_0}[l(\theta)] - \max_{H_1}[l(\theta)] = l(\theta_0) - l(\hat{\theta}) = llr(\theta_0)$$

Note that we are able to replace $\max_{H_1}[l(\theta)]$ with $l(\hat{\theta})$ since under the alternative hypothesis the maximum value of the (log-)likelihood is given at the MLE.

Now, crucially, if the null hypothesis is true, we know (Inference 4.4) the approximate distribution for -2 times this quantity:

$$H_0\colon \theta = \theta_0 \quad \Rightarrow \quad -2\, llr(\theta_0) \mathrel{\dot\sim} \chi_1^2$$

We can use this distribution to quantify evidence against the null hypothesis and in favour of the alternative: we reject $H_0\colon \theta = \theta_0$ with size α if

$$-2\, llr(\theta_0) > \chi_{1,(1-\alpha)}^2$$

e.g. $-2\, llr(\theta_0) > 3.84$ when $\alpha = 0.05$.

This test is known as the likelihood (or log-likelihood) ratio test.

Note:
Potentially this wastes information since 1) the distribution is approximate except for the mean of the Normal distribution with known variance; 2) the form of the test is not most powerful against any simple alternative hypothesis. Nevertheless for reasonable sample sizes its performance is usually good and the test has two great advantages: 1) it is simple; 2) the p-value does not depend on the parameter scale: i.e. the p-value is the same for all transformations of the parameter, provided the transformation is a strictly increasing or decreasing function of the parameter.

EXERCISE 7.2.1

Suppose k events are observed in n subjects, and a binomial model with parameter π is assumed. Obtain the statistic $-2llr(\pi_0)$ for testing $H_0\colon \pi = \pi_0$ vs $H_1\colon \pi \neq \pi_0$, where $\pi_0 = 1/2, k = 40, n = 100$. Do you reject H_0 at the 5% level?

7.3 Wald test

Like the likelihood ratio test, the Wald test is appropriate for testing hypotheses of the form: $H_0\colon \theta = \theta_0$ vs $H_1\colon \theta \neq \theta_0$.

However it is based on the *quadratic approximation* to the log-likelihood ratio introduced in Inference 5, rather than the log-likelihood ratio itself.

Note that the likelihood ratio test of 7.2 uses the log-likelihood ratio itself, with the approximation (in contexts apart for the mean of the Normal distribution with known variance) residing in the approximate distribution of $-2llr$.

In contrast the Wald test in this section, and the score test in 7.4, not only rely on this approximate distribution, they also require the quadratic approximation to the real log-likelihood ratio.

Recall from Inference 5.2

$$llr(\theta) \approx -\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2 \quad \text{(asymptotically)}$$

where M is the MLE $\hat{\theta}$, and $S = \sqrt{-1/l''(\hat{\theta})}$. Note that $-l''(\hat{\theta})$ is known as the *observed Fisher information* (see section 7.5), and is obtained by simply substituting $\hat{\theta}$ for θ in minus the second derivative of the log-likelihood.

And we know that under the null hypothesis:

$$H_0\colon \theta = \theta_0 \quad \Rightarrow \quad -2\, llr(\theta_0) \mathrel{\dot{\sim}} \chi_1^2 \quad \text{(asymptotically)}$$

$$\Rightarrow \quad -2 \times -\frac{1}{2}\left(\frac{M-\theta_0}{S}\right)^2 = \left(\frac{M-\theta_0}{S}\right)^2 \mathrel{\dot{\sim}} \chi_1^2$$

$$\Rightarrow \quad \left(\frac{M-\theta_0}{S}\right) \mathrel{\dot{\sim}} N(0,1)$$

In Figure 7.3 below, the quadratic curve (dashed line) is chosen so that its peak, and its curvature at the peak, are the same as the real log-likelihood function (solid line). The difference between $\hat{\theta}$ and θ_0 shown indicates the evidence against the null value given by the Wald test. In the test this difference is calibrated against the curvature (i.e. standard error): the same $\hat{\theta} - \theta_0$ offers less evidence against the null value when the curvature is shallower (standard error larger), since this corresponds to a smaller log-likelihood ratio (vertical difference). The evidence of the likelihood ratio test is given by the vertical difference between $l(\theta_0)$ and $l(\hat{\theta})$.

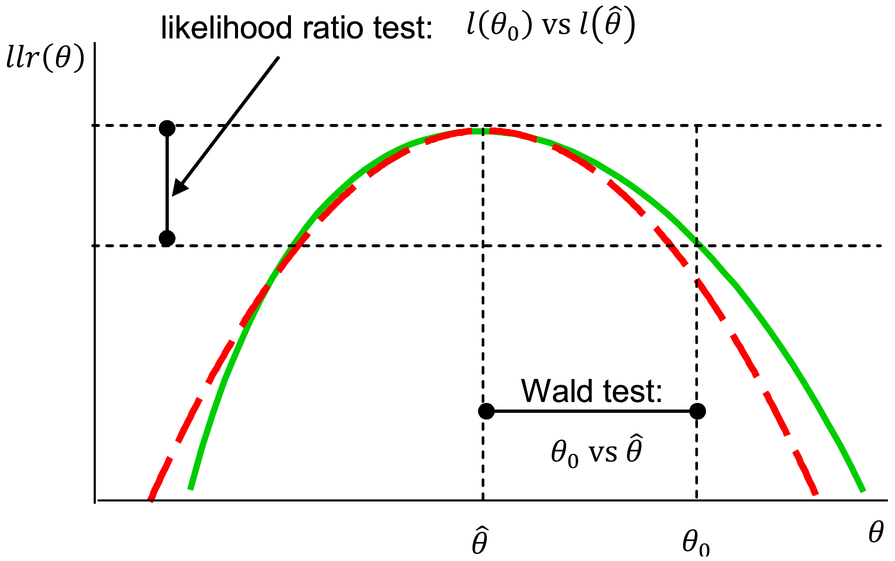
likelihood ratio test: $l(\theta_0)$ vs $l(\hat{\theta})$
$llr(\theta)$
Wald test:
θ_0 vs $\hat{\theta}$
$\hat{\theta}$
θ_0
θ

Figure 7.3. Likelihood ratio and Wald tests: solid line is log-likelihood ratio, dashed is quadratic approximation

***Example** 7.3.1 Binomial distribution*

Suppose k events are observed in n subjects, and a binomial model is assumed with parameter π. We will perform the Wald test of $H_0: \pi = \pi_0$ vs $H_1: \pi \neq \pi_0$ both in terms of the untransformed parameter, and using instead the log odds parameter.

1. Using parameter π, the values of M and S are $M = k/n = p,\ S = \sqrt{p(1-p)/n}$ (see 5.2).
   For the Wald test,

$$H_0: \pi = \pi_0 \quad \Rightarrow \quad W = \left(\frac{p - \pi_0}{\sqrt{p(1-p)/n}}\right) \dot{\sim} N(0,1)$$

   and we therefore refer W to a $N(0,1)$ distribution.

2. Using instead the transformation $\beta = \log[\pi/(1-\pi)]$, the value of M is now $\log[k/(n-k)]$, and it can be shown that now $S = \sqrt{\frac{1}{k} + \frac{1}{n-k}}$. So the Wald test becomes,

$$H_0: \beta = \beta_0 \quad \Rightarrow \quad W = \left(\frac{\log\left[\frac{k}{n-k}\right] - \log\left[\frac{\pi_0}{1-\pi_0}\right]}{\sqrt{\frac{1}{k} + \frac{1}{n-k}}}\right) \dot{\sim} N(0,1)$$

   and we again refer W to a $N(0,1)$ distribution.

EXERCISE 7.3.1

Evaluate the results of a Wald test based on π and $\log\left[\frac{\pi}{1-\pi}\right]$ for $H_0: \pi = 1/2$ vs $H_1: \pi \neq 1/2$, if $K \sim Bin(100, \pi)$ and we observe $k = 40$.

7.4 Score test

The score test is based on a different quadratic approximation to that underlying the Wald test. While the Wald test most accurately approximates the shape of the real log-likelihood at its maximum, when $\theta = \hat{\theta}$, the score test is based on a quadratic approximating the shape of the real log-likelihood at the null value, θ_0. Thus, for the score test, the gradient and curvature of the quadratic approximation are set to equal the gradient and curvature of the real log-likelihood at the null value. We define U to be the gradient of the log-likelihood at $\theta = \theta_0$, and V to be minus the expectation of the curvature of the log-likelihood at $\theta = \theta_0$, assuming the null hypothesis is true:

$$U = l'(\theta)|_{\theta=\theta_0} = l'(\theta_0)$$

$$V = -E[l''(\theta)]|_{\theta=\theta_0} = -E[l''(\theta_0)]$$

Note that $-E[l''(\theta)]$ is known as the *expected Fisher information*, and while U requires only substitution of the null value in the first derivative, V requires taking the expectation of the second derivative (by assuming the null value is true) as well as substitution.

Then consider the quadratic approximation:

$$q(\theta) = -\frac{V}{2}\left(\frac{U}{V} + \theta_0 - \theta\right)^2$$

Note that:

- $q(\theta)$ has a maximum of 0 when $\theta = \theta_0 + \frac{U}{V}$
- $q(\theta_0) = -\frac{U^2}{2V}$
- $q'(\theta_0) = \left[V\left(\frac{U}{V} + \theta_0 - \theta\right)\right]_{\theta=\theta_0} = V\left(\frac{U}{V}\right) = U = l'(\theta_0)$ as required
- $q''(\theta_0) = [-V]_{\theta=\theta_0} = -V = E[l''(\theta_0)]$ as required

Since, the quadratic $q(\theta)$ approximates $llr(\theta)$,

$$H_0\colon \theta = \theta_0 \quad \Rightarrow \quad -2\,q(\theta_0) = \frac{U^2}{V} \dot{\sim}\ \chi_1^2$$

so we refer the statistic $\frac{U^2}{V}$ to a χ_1^2 distribution. Alternatively, we can refer $\frac{U}{\sqrt{V}}$ to a $N(0,1)$ distribution. As with the likelihood ratio and Wald tests, the approximation improves with sample size.

For a derivation of $q(\theta)$, U and V from the Normal quadratic function see the (non-examinable) Appendix at the end of this document.

The score test, in effect, compares the gradient at the null value with the gradient ($= 0$) at the maximum: the greater the difference, the greater the evidence against the null.

Again, the curvature is used to calibrate the gradient, but in a less transparent way than for the Wald test (see end of Appendix).

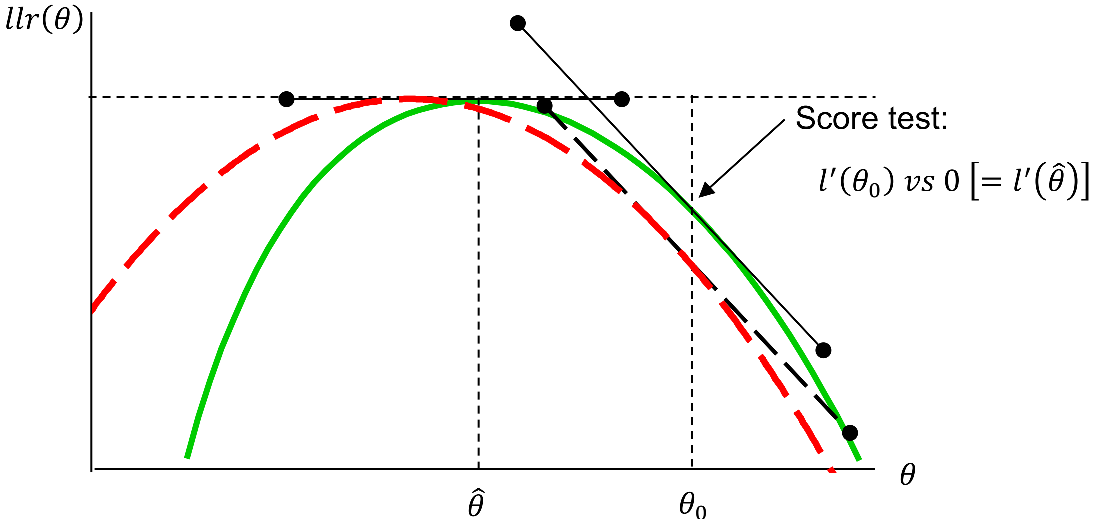

Figure 7.4. Score test: solid line is log-likelihood ratio, dashed is quadratic approximation

In Figure 7.4, although the value of the score quadratic approximation (dashed curve) and of the real log-likelihood (solid curve) are not equal at the null value θ_0 (the shapes are at different heights here), their gradient and curvature are the same here. The further the null value gradient is from zero, the less support the data give to the null hypothesis.

Example *7.4.1 Binomial distribution (continued)*
$K{\sim}Bin(100,\pi)$ and we observe $k = 80$. $H_0: \pi = \pi_0$ vs $H_1: \pi \neq \pi_0$, where $\pi_0 = 0.2$. To perform the score test we obtain U and V, and refer U^2/V to χ_1^2. To make the structure of the test statistic clearer, we will only substite in the numerical values at the end, and introduce $p = k/n$:

$$l(\pi|k) = k\log(\pi) + (n-k)\log(1-\pi)$$

$$l'(\pi) = \frac{k}{\pi} - \frac{n-k}{1-\pi} = \frac{k(1-\pi)-(n-k)\pi}{\pi(1-\pi)} = \frac{k-k\pi-n\pi+k\pi}{\pi(1-\pi)} = \frac{k-n\pi}{\pi(1-\pi)}$$
$$= \frac{k/n-\pi}{\pi(1-\pi)/n} = \frac{p-\pi}{\pi(1-\pi)/n}$$

$$\Rightarrow \quad U = l'(\pi_0) = \frac{p-\pi_0}{\pi_0(1-\pi_0)/n}$$

Since we need to take the expectation to derive V, we will make the random variable, K, explicit (all other terms are fixed):

$$l''(\pi|K) = -\frac{K}{\pi^2} - \frac{n-K}{(1-\pi)^2}$$

Then we use that $E[K] = n\pi$:

$$E[l''(\pi|K)] = -\frac{E[K]}{\pi^2} - \frac{n-E[K]}{(1-\pi)^2} = -\frac{n\pi}{\pi^2} - \frac{n-n\pi}{(1-\pi)^2} = -\frac{n}{\pi} - \frac{n(1-\pi)}{(1-\pi)^2}$$
$$= -\frac{n}{\pi} - \frac{n}{1-\pi} = -\frac{n(1-\pi)+n\pi}{\pi(1-\pi)} = -\frac{n-n\pi+n\pi}{\pi(1-\pi)} = -\frac{n}{\pi(1-\pi)}$$

$$-E[l''(\pi|K)] = \frac{n}{\pi(1-\pi)}$$

$$\Rightarrow \quad V = -E[l''(\pi_0)] = \frac{n}{\pi_0(1-\pi_0)}$$

This gives

$$\frac{U^2}{V} = \frac{\dfrac{(p-\pi_0)^2}{(\pi_0(1-\pi_0)/n)^2}}{\dfrac{n}{\pi_0(1-\pi_0)}} = \frac{(p-\pi_0)^2}{(\pi_0(1-\pi_0)/n)^2}\frac{\pi_0(1-\pi_0)}{n} = \frac{(p-\pi_0)^2}{\pi_0(1-\pi_0)/n} \mathrel{\dot\sim} \chi_1^2$$

or

$$\frac{U}{\sqrt{V}} = \frac{(p-\pi_0)}{\sqrt{\pi_0(1-\pi_0)/n}} \mathrel{\dot\sim} N(0,1)$$

under the null hypothesis.

We see that in this binomial context the score statistic has the same numerator as the Wald test statistic, but the standard error is derived under the null hypothesis.

Now substituting in our values for p, π_0, n:

$$\frac{U^2}{V} = \frac{(80/100-0.2)^2}{0.2\,(1-0.2)/100} = 225$$

giving very strong evidence at less than the 1% level to reject H_0, implying that $\pi > 0.2$.

EXERCISE 7.4.1
Evaluate the results of the score test when $K{\sim}Bin(100,\pi)$, $k=40$ and $\pi_0=0.5$.

7.5 Comparison between LLR, Wald and Score tests

1. The likelihood ratio test measures the distance between the log-likelihoods at the null value and the MLE (Figure 7.3). The greater this difference (whose distribution is approximate in non-Normal contexts), the less support the data (represented by the MLE) give to the null hypothesis.

   The Wald test measures the difference between the MLE and the null value, using the standard error (obtained from the quadratic approximation) to calibrate this difference (with an approximate distribution in non-Normal contexts). Again, the larger this difference, the less support for the null hypothesis.

   The score test measures the slope of the log-likelihood at the null value (Figure 7.4), calibrated by the curvature of the quadratic approximation (which is constant for a quadratic, unlike the real log-likelihood ratio in most non-Normal contexts). The steeper this slope (again with an approximate distribution in non-Normal contexts), the 'further' from the slope at the MLE, and the less support for the null hypothesis. In certain contexts the score test has the practical advantage that it does not require the computation of the MLE (since we know the gradient at the MLE to be zero), and depends only on the null value specified.

2. The score test uses $V = -E[l''(\theta_0)]$, the expected (Fisher) information under the null hypothesis $\theta = \theta_0$; but in contexts where the expectation cannot be taken, the statistic $V = -l''(\theta_0)$ can be used (and in some contexts the expectation evaluates to this in any case).

   The Wald test uses the observed information $-l''(\hat{\theta})$. The expectation of this observed information under $\theta = \hat{\theta}$, $-\mathrm{E}[l''(\hat{\theta})]$, is generally not used: partly because in many common contexts this just evaluates to $-l''(\hat{\theta})$; and partly because it has been shown that, where they differ, the observed information is preferable to its expectation. These issues are discussed in Pawitan Y. *In All Likelihood: Statistical Modelling and Inference using Likelihood*, Oxford, 2001, Section 9.6.

3. If the log-likelihood is exactly quadratic (which is the case, for example, for the mean of a Normal distribution with known variance), the three tests will yield *exactly* the same answers. If the log-likelihood is close the quadratic, then the three tests will yield similar answers: the three tests are asymptotically equivalent.

4. The likelihood ratio test is the only one that gives the same result if the parameter is transformed. The Wald and score test approximations improve if a parameter transformation is chosen that improves the quadratic nature of the log-likelihood. This corresponds exactly to parameter transformation for improving confidence intervals (see 5.8); the same grounds for choice of transformation apply to the hypothesis-testing context.

5. If the MLE and null value are a long way apart, then quadratic approximation to the log-likelihood curve may not approximate well over the range covered by both these values (see Figures 7.3, 7.4). In this case the likelihood ratio test is preferable (though unless the sample size is small all three tests are likely to reject with small p-values).

6. If the results of the tests differ noticeably, even after choosing an appropriate scale for the parameter, then the sample size may be ‘too small’. So-called ‘exact’ p-values have been advocated, but a better solution is to use the real log-likelihood directly to measure the support for different values of the parameter (see Clayton and Hills Chapter 12).

7. Most parametric tests in wide use are either likelihood ratio tests, Wald tests or score tests, even if they were developed before likelihood based methods were introduced. For example, *Z* tests, *t*-tests and *F*-tests are likelihood ratio tests.

   In epidemiology many tests in use (developed before computers) are score tests. As well as being easier to compute than Wald tests, score tests also have the convenient property, for epidemiology, that unstratified versions are easily adaptable to stratified versions; e.g. Mantel-Haenszel test, log rank test.

   Wald tests are widely used in generalised linear models such as logistic, Poisson, Cox regressions to test individual coefficients.

We conclude that, when available, the likelhood ratio test is always preferable to the Wald or score tests for comparing $H_0\colon \theta = \theta_0$ vs $H_1\colon \theta \neq \theta_0$.

Appendix (*non-examinable*)

Quadratic approximation for the Score test

Consider a Normal log-likelihood quadratic $q(\theta) = -\frac{1}{2}\left(\frac{M-\theta}{S}\right)^2 = -\frac{1}{2S^2}(M-\theta)^2$. For the score test we require that:

$$q'(\theta_0) = U = l'(\theta_0) \tag{1}$$
$$q''(\theta_0) = -V = E[l''(\theta_0)] \tag{2}$$

Now

$$q'(\theta) = \frac{1}{S^2}(M-\theta)$$
$$q''(\theta) = -\frac{1}{S^2}$$

Equations (1) and (2) mean that we want:

$$q'(\theta_0) = \frac{1}{S^2}(M-\theta_0) = U \tag{3}$$
$$q''(\theta_0) = -\frac{1}{S^2} = -V \tag{4}$$

From equation (4):

$$V = \frac{1}{S^2}$$

and substituting this into equation (3):

$$V(M-\theta_0) = U$$
$$\Rightarrow \quad M = \frac{U}{V} + \theta_0 \tag{5}$$

Substituting for M (equation (5)) and S (equation (4)) in $q(\theta)$ gives:

$$q(\theta) = -\frac{V}{2}\left(\frac{U}{V} + \theta_0 - \theta\right)^2$$

as given in 7.4.

Having V in the denominator of the score test statistic might suggest that the larger V is, the smaller the test statistic (which would be surprising given that the larger V is, the smaller the standard error term S). However, U also depends on V. Note that:

$$U = V(M-\theta_0)$$

$$\Rightarrow \quad \frac{U^2}{V} = V(M-\theta_0)^2,$$

hence the larger V is, the larger the score test statistic.