21 Testing Problems

Testing is certainly another fundamental problem of inference. It is interestingly different from estimation. The accuracy measures are fundamentally different and the appropriate asymptotic theory is also different. Much of the theory of testing has revolved somehow or other on the Neyman-Pearson lemma, which has led to a lot of ancillary developments in other problems in mathematical statistics. Testing has led to the useful idea of local alternatives, and like maximum likelihood in estimation, likelihood ratio, Wald, and Rao score tests have earned the status of default methods, with a neat and quite unified asymptotic theory. Because of all these reasons, a treatment of testing is essential. We discuss the asymptotic theory of likelihood ratio, Wald, and Rao score tests in this chapter. Principal references for this chapter are Bickel and Doksum (2001), Ferguson (1996), and Sen and Singer (1993). Many other specific references are given in the sections.

21.1 Likelihood Ratio Tests

The likelihood ratio test is a general omnibus test applicable, in principle, in most finite dimensional parametric problems. Thus, let \( X^{(n)} = (X_1, \ldots, X_n) \) be the observed data with joint distribution \( P^\theta \ll \mu_n, \theta \in \Theta \), and density \( f_\theta(x^{(n)}) = dP_\theta^n/d\mu_n \). Here, \( \mu_n \) is some appropriate \( \sigma \)-finite measure on \( X_n \), which we assume to be a subset of an Euclidean space. For testing \( H_0 : \theta \in \Theta_0 \) vs. \( H_1 : \theta \in \Theta - \Theta_0 \), the likelihood ratio test (LRT) rejects \( H_0 \) for small values of

\[
\Lambda_n = \frac{\sup_{\theta \in \Theta_0} f_\theta(x^{(n)})}{\sup_{\theta \in \Theta} f_\theta(x^{(n)})}
\]

The motivation for \( \Lambda_n \) comes from two sources:

(a) The case where \( H_0, H_1 \) are each simple, a most powerful (MP) test is found from \( \Lambda_n \) by the Neyman-Pearson lemma.

(b) The intuitive explanation that for small values of \( \Lambda_n \) we can better match the observed data with some value of \( \theta \) outside of \( \Theta_0 \).

LRTs are useful because they are omnibus tests and because otherwise optimal tests for a given sample size \( n \) are generally hard to find outside of the exponential family. However the LRT is not a universal test. There are important examples where the LRT simply cannot be used, because the null distribution of the LRT test statistic depends on nuisance parameters. Also, the exact distribution of the LRT
statistic is very difficult or impossible to find in many problems. Thus, asymptotics become really important. But the asymptotics of the LRT may be nonstandard under nonstandard conditions.

Although we only discuss the case of a parametric model with a fixed finite dimensional parameter space, LRTs and their useful modifications have been studied when the number of parameters grows with the sample size $n$, and in nonparametric problems. See, e.g., Portnoy(1988),Fan,Hung and Wong(2000), and Fan and Zhang(2001).

We start with a series of examples, which illustrate various important aspects of the likelihood ratio method.

21.2 Examples

Example 21.1. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ and consider testing

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu \neq 0.$$ 

Let $\theta = (\mu, \sigma^2)$. Then,

$$\Lambda_n = \sup_{\theta \in \Theta} (1/\sigma^n) \exp \left( -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right) = \left( \frac{\sum_i (X_i - \bar{X}_n)^2}{\sum_i X_i^2} \right)^{n/2}$$

by an elementary calculation of mles of $\theta$ under $H_0$ and in the general parameter space. By another elementary calculation, $\Lambda_n < c$ is seen to be equivalent to $t_n^2 > k$ where

$$t_n = \sqrt{n} \frac{\bar{X}_n}{\sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2}}$$

is the t-statistic. In other words, the t-test is the LRT. Also, observe that,

$$t_n^2 = \frac{n \bar{X}_n^2}{\frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2}$$

$$= \frac{\sum_i X_i^2 - \sum_i (X_i - \bar{X}_n)^2}{\frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2}$$

$$= \frac{(n - 1) \sum_i X_i^2}{\sum_i (X_i - \bar{X}_n)^2} - (n - 1)$$

$$= (n - 1) \frac{X_n^2}{\sum_i (X_i - \bar{X}_n)^2} - (n - 1)$$

320
This implies,
\[
\Lambda_n = \left( \frac{n - 1}{t^2_n + n - 1} \right)^{n/2}
\]
\[\Rightarrow \log \Lambda_n = \frac{n}{2} \log \frac{n - 1}{t^2_n + n - 1}\]
\[\Rightarrow -2 \log \Lambda_n = n \log \left( 1 + \frac{t^2_n}{n - 1} \right)\]
\[= n \left( \frac{t^2_n}{n - 1} + o_p \left( \frac{t^2_n}{n - 1} \right) \right)\]
\[\overset{\mathcal{L}}{\sim} \chi_1^2\]

under \(H_0\) since \(t_n \overset{\mathcal{L}}{\rightarrow} N(0, 1)\) under \(H_0\).

**Example 21.2.** Consider a multinomial distribution \(MN(n, p_1, \cdots, p_k)\). Consider testing

\(H_0 : p_1 = p_2 = \cdots = p_k\) vs. \(H_1 : H_0\) is not true.

Let \(n_1, \cdots, n_k\) denote the observed cell frequencies. Then by an elementary calculation,

\[\Lambda_n = \prod_{i=1}^{k} \left( \frac{n}{k n_i} \right)^{n_i}\]

\[\Rightarrow -\log \Lambda_n = n \left( \log \frac{k}{n} \right) + \sum_{i=1}^{k} n_i \log n_i\]

The exact distribution of this is a messy discrete object and so asymptotics will be useful. We illustrate the asymptotic distribution for \(k = 2\). In this case,

\[-\log \Lambda_n = n \log 2 - n \log n + n_1 \log n_1 + (n - n_1) \log (n - n_1)\]

Let \(Z_n = (n_1 - n/2)/\sqrt{n}/4\). Then,

\[-\log \Lambda_n = n \log 2 - n \log n\]
\[+ \frac{n + \sqrt{n} Z_n}{2} \log \frac{n + \sqrt{n} Z_n}{2} + \frac{n - \sqrt{n} Z_n}{2} \log \frac{n - \sqrt{n} Z_n}{2}\]
\[= -n \log n + \frac{n + \sqrt{n} Z_n}{2} \log (n + \sqrt{n} Z_n) + \frac{n - \sqrt{n} Z_n}{2} \log (n - \sqrt{n} Z_n)\]
\[= \frac{n + \sqrt{n} Z_n}{2} \log (1 + \frac{Z_n}{\sqrt{n}}) + \frac{n - \sqrt{n} Z_n}{2} \log (1 - \frac{Z_n}{\sqrt{n}})\]
\[= \frac{n + \sqrt{n} Z_n}{2} \left( \frac{Z_n}{\sqrt{n}} - \frac{Z_n^2}{2n} + o_p \left( \frac{Z_n^2}{n} \right) \right)\]
\[
\frac{n - \sqrt{n}Z_n}{2} \left( -\frac{Z_n}{\sqrt{n}} - \frac{Z_n^2}{2n} + o_p \left( \frac{Z_n^2}{n} \right) \right)
= \frac{Z_n^2}{2} + o_p(1).
\]

Hence \(-2 \log \Lambda_n \xrightarrow{\mathcal{L}} \chi^2_1\) under \(H_0\) as \(Z_n \xrightarrow{\mathcal{L}} N(0, 1)\) under \(H_0\).

**Remark:** The popular test in this problem is the Pearson chi-square test which rejects \(H_0\) for large values of

\[
\frac{\sum_{i=1}^k (n_i - n/k)^2}{n/k}
\]

Interestingly this test statistic too has an asymptotic \(\chi^2_1\) distribution as does the LRT statistic.

**Example 21.3.** This example shows that the chi-square asymptotics of the LRT statistic fail when parameters have constraints or somehow there is a boundary phenomenon.

Let \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} N_2(\theta, I)\) where the parameter space is restricted to \(\Theta = \{\theta = (\theta_1, \theta_2) : \theta_1 \geq 0, \theta_2 \geq 0\}\). Consider testing

\(H_0 : \theta = 0 \text{ vs. } H_1 : H_0 \text{ is not true.}\)

We would write the \(n\) realizations as \(x_i = (x_{1i}, x_{2i})^T\) and the sample average would be denoted by \(\bar{x} = (\bar{x}_1, \bar{x}_2)\). The mle of \(\theta\) is given by,

\[
\hat{\theta} = \begin{pmatrix}
\bar{x}_1 \lor 0 \\
\bar{x}_2 \lor 0
\end{pmatrix}
\]

Therefore by an elementary calculation,

\[
\Lambda_n = \frac{\exp(-\sum_i x_{1i}^2/2 - \sum_i x_{2i}^2/2)}{\exp(-\sum_i (x_{1i} - \bar{x}_1 \lor 0)^2/2 - \sum_i (x_{2i} - \bar{x}_2 \lor 0)^2/2)}
\]

**Case 1:** \(\bar{x}_1 \leq 0, \bar{x}_2 \leq 0\). In this case \(\Lambda_n = 1\) and \(-2 \log \Lambda_n = 0\).

**Case 2:** \(\bar{x}_1 > 0, \bar{x}_2 \leq 0\). In this case \(-2 \log \Lambda_n = n\bar{x}_1^2\).

**Case 3:** \(\bar{x}_1 \leq 0, \bar{x}_2 > 0\). In this case \(-2 \log \Lambda_n = n\bar{x}_2^2\).
Case 4: $\pi_1 > 0, \pi_2 > 0$. In this case $-2 \log \Lambda_n = n\pi_1^2 + n\pi_2^2$.

Now, under $H_0$ each of the above four cases has a probability $1/4$ of occurrence. Therefore, under $H_0$,

$$-2 \log \Lambda_n \xrightarrow{c} \frac{1}{4} \delta_{[0]} + \frac{1}{2} 
 \chi_1^2 + \frac{1}{4} \chi_2^2,$$

in the sense of the mixture distribution being its weak limit.

**Example 21.4.** In bio-equivalence trials, a brand name drug and a generic are compared with regard to some important clinical variable, such as average drug concentration in blood over a 24 hour time period. By testing for a difference, the problem is reduced to a single variable, often assumed to be normal. Formally, one has $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ and the bio-equivalence hypothesis is:

$$H_0 : |\mu| \leq \epsilon \text{ for some specified } \epsilon > 0.$$

Inference is *always* significantly harder if known constraints on the parameters are enforced. Casella and Strawderman (1980) is a standard introduction to the normal mean problem with restrictions on the mean; Robertson, Wright and Dykstra (1988) is an almost encyclopedic exposition.

Coming back to our example, to derive the LRT, we need the restricted mle and the non-restricted mle. We will assume here that $\sigma^2$ is known. Then the MLE under $H_0$ is

$$\hat{\mu}_{H_0} = \bar{X} \text{ if } |\bar{X}| \leq \epsilon$$

$$\quad = \epsilon \text{sign}(\bar{X}) \text{ if } |\bar{X}| > \epsilon.$$

Consequently, the LRT statistic $\Lambda_n$ satisfies,

$$-2 \log \Lambda_n = 0 \text{ if } |\bar{X}| \leq \epsilon$$

$$\quad = \frac{n}{\sigma^2} (\bar{X} - \epsilon)^2 \text{ if } \bar{X} > \epsilon$$

$$\quad = \frac{n}{\sigma^2} (\bar{X} + \epsilon)^2 \text{ if } \bar{X} < -\epsilon.$$ 

Take a fixed $\mu$. Then,

$$P_\mu(|\bar{X}| \leq \epsilon) = \Phi \left( \frac{\sqrt{n}(\epsilon - \mu)}{\sigma} \right) + \Phi \left( \frac{\sqrt{n}(\epsilon + \mu)}{\sigma} \right) - 1$$

$$P_\mu(\bar{X} < -\epsilon) = \Phi \left( -\frac{\sqrt{n}(\epsilon + \mu)}{\sigma} \right)$$

323
and

\[ P_\mu(\bar{X} > \epsilon) = 1 - \Phi\left(\frac{\sqrt{n}(\epsilon - \mu)}{\sigma}\right) \]

From these expressions we get:

**Case 1:** If \(-\epsilon < \mu < \epsilon\) then \(P_\mu(|\bar{X}| \leq \epsilon) \to 1\) and so \(-2 \log \Lambda_n \xrightarrow{\mathcal{L}} \delta_{\{0\}}\).

**Case 2.a:** If \(\mu = -\epsilon\) then each of \(P_\mu(|\bar{X}| \leq \epsilon)\) and \(P_\mu(\bar{X} < -\epsilon)\) converges to \(1/2\). In this case, \(-2 \log \Lambda_n \xrightarrow{\mathcal{L}} \frac{1}{2} \delta_{\{0\}} + \frac{1}{2} \chi^2_1\), i.e., the mixture of a point mass and a chisquare distribution.

**Case 2.b:** Similarly if \(\mu = \epsilon\) then again, \(-2 \log \Lambda_n \xrightarrow{\mathcal{L}} \frac{1}{2} \delta_{\{0\}} + \frac{1}{2} \chi^2_1\).

**Case 3:** If \(\mu > \epsilon\) then \(P_\mu(|\bar{X}| \leq \epsilon) \to 0\) and \(P_\mu(\bar{X} > \epsilon) \to 1\). In this case, \(-2 \log \Lambda_n \xrightarrow{\mathcal{L}} NC\chi^2(1, (\mu - \epsilon)^2/\sigma^2)\), a noncentral chisquare distribution.

**Case 4:** Likewise, if \(\mu < -\epsilon\) then \(-2 \log \Lambda_n \xrightarrow{\mathcal{L}} NC\chi^2(1, (\mu + \epsilon)^2/\sigma^2)\).

So, unfortunately, even the null asymptotic distribution of \(- \log \Lambda_n\) depends on the exact value of \(\mu\).

**Example 21.5.** Consider the Behrens-Fisher problem with

\[
X_1, X_2, \ldots, X_m \overset{iid}{\sim} N(\mu_1, \sigma_1^2) \\
Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu_2, \sigma_2^2)
\]

and all \(m + n\) observations are independent. We want to test

\[ H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2. \]

Let \(\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2\) denote the restricted mle of \(\mu, \sigma_1^2\) and \(\sigma_2^2\) respectively, under \(H_0\). They satisfy the equations,

\[
\hat{\sigma}_1^2 = (\bar{X} - \hat{\mu})^2 + \frac{1}{m} \sum_i (X_i - \bar{X})^2 \\
\hat{\sigma}_2^2 = (\bar{Y} - \hat{\mu})^2 + \frac{1}{n} \sum_j (Y_j - \bar{Y})^2 \\
0 = \frac{m(\bar{X} - \hat{\mu})}{\hat{\sigma}_1^2} + \frac{n(\bar{Y} - \hat{\mu})}{\hat{\sigma}_2^2}
\]
This gives,

$$-2 \log \Lambda_n = \left( \frac{1}{m} \sum (X_i - \bar{X})^2 + \left( \frac{1}{m} \sum (X_i - \bar{X})^2 \right) \right)^\frac{m}{2} \left( \frac{1}{n} \sum (Y_j - \bar{Y})^2 + \left( \frac{1}{n} \sum (Y_j - \bar{Y})^2 \right) \right)^\frac{n}{2}$$

However, the LRT fails as a matter of practical use because its distribution depends on the nuisance parameter $\sigma_1^2/\sigma_2^2$, which is unknown.

**Example 21.6.** In point estimation, the mles in nonregular problems behave fundamentally differently with respect to their asymptotic distributions. For example, limit distributions of mles are not normal (see Chapter 16). It is interesting, therefore, that limit distributions of $-2 \log \Lambda_n$ in nonregular cases still follow the chi-square recipe, but with different degrees of freedom.

As an example, suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} U[\mu - \sigma, \mu + \sigma]$ and suppose we want to test $H_0: \mu = 0$. Let $W = X_{(n)} - X_{(1)}$ and $U = \max(X_{(n)}, -X_{(1)})$. Under $H_0$, $U$ is complete and sufficient and $W/U$ is an ancillary statistic. Therefore, by Basu’s theorem (Basu (1955)), under $H_0$, $U$ and $W/U$ are independent. This will be useful shortly to us. Now, by a straightforward calculation,

$$\Lambda_n = \left( \frac{W}{2U} \right)^n$$

$$\Rightarrow -2 \log \Lambda_n = 2n(\log 2U - \log W)$$

Therefore, under $H_0$,

$$E e^{it(-2 \log \Lambda_n)} = E e^{2nit(\log 2U - \log W)} = \frac{E e^{2nit(- \log W)}}{E e^{2nit(- \log 2U)}}$$

by the independence of $U$ and $W/U$. This works out, on doing the calculation, to

$$E e^{it(-2 \log \Lambda_n)} = \frac{n - 1}{n(1 - 2it) - 1} \rightarrow \frac{1}{1 - 2it}$$

Since $(1 - 2it)^{-1}$ is the characteristic function of the $\chi_2^2$ distribution, it follows that

$$-2 \log \Lambda_n \overset{\mathcal{L}}{\rightarrow} \chi_2^2$$

Thus inspite of the nonregularity, $-2 \log \Lambda_n$ is asymptotically chi-square, but we gain a degree of freedom! This phenomenon holds more generally in nonregular cases.
Example 21.7. Consider the problem of testing for equality of two Poisson means. Thus let \( X_1, X_2, \ldots, X_m \overset{iid}{\sim} \text{Poi}(\mu_1) \) and \( Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{Poi}(\mu_2) \), where \( X_i \)'s and \( Y_j \)'s are independent. Suppose we wish to test \( H_0 : \mu_1 = \mu_2 \). We assume \( m, n \to \infty \) in such a way that
\[
\frac{m}{m+n} \to \lambda \text{ with } 0 < \lambda < 1
\]
The restricted mle for \( \mu = \mu_1 = \mu_2 \) is
\[
\frac{\sum_i X_i + \sum_j Y_j}{m+n} = \frac{m\bar{X} + n\bar{Y}}{m+n}
\]
Therefore, on an easy calculation,
\[
\Lambda_{m,n} = \frac{(m\bar{X} + n\bar{Y})/(m+n))^{m\bar{X}+n\bar{Y}}}{\bar{X}^{m\bar{X}}\bar{Y}^{n\bar{Y}}}
\]
\[
\Rightarrow -\log \Lambda_{m,n} = m\bar{X} \log \bar{X} + n\bar{Y} \log \bar{Y}
\]
\[
-(m\bar{X} + n\bar{Y}) \log \left( \frac{m}{m+n} \bar{X} + \frac{n}{m+n} \bar{Y} \right)
\]
The asymptotic distribution of \(-\log \Lambda_{m,n}\) can be found by a two-term Taylor expansion, by using the multivariate delta theorem. This would be a more direct derivation, but we will instead present a derivation based on an asymptotic technique that is useful in many problems. Define,
\[
Z_1 = Z_{1,m} = \frac{\sqrt{m}(\bar{X} - \mu)}{\sqrt{\mu}}, \quad Z_2 = Z_{2,m} = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sqrt{\mu}}
\]
where \( \mu = \mu_1 = \mu_2 \) is the common value of \( \mu_1 \) and \( \mu_2 \) under \( H_0 \). Substituting in the expression for \( \log \Lambda_{m,n} \) we have,
\[
-\log \Lambda_{m,n} = (m\mu + \sqrt{m}\mu Z_1) \left[ \log \left( 1 + \frac{Z_1}{\sqrt{m\mu}} \right) + \log \mu \right]
\]
\[
+ (n\mu + \sqrt{n}\mu Z_2) \left[ \log \left( 1 + \frac{Z_2}{\sqrt{n\mu}} \right) + \log \mu \right]
\]
\[
- ((m+n)\mu + \sqrt{m}\mu Z_1 + \sqrt{n}\mu Z_2) \left[ \log \left( 1 + \frac{m}{m+n} \frac{Z_1}{\sqrt{m\mu}} + \frac{n}{m+n} \frac{Z_2}{\sqrt{n\mu}} \right) + \log \mu \right]
\]
\[
= (m\mu + \sqrt{m}\mu Z_1) \left( \frac{Z_1}{\sqrt{m\mu}} - \frac{Z_1^2}{2m\mu} + O_p(m^{-3/2}) + \log \mu \right)
\]
\[
+ (n\mu + \sqrt{n}\mu Z_2) \left( \frac{Z_2}{\sqrt{n\mu}} - \frac{Z_2^2}{2n\mu} + O_p(n^{-3/2}) + \log \mu \right)
\]
326
\[-((m + n)\mu + \sqrt{m\mu}Z_1 + \sqrt{n\mu}Z_2)\left(\frac{m}{m+n}\frac{Z_1}{\sqrt{m\mu}} + \frac{n}{m+n}\frac{Z_2}{\sqrt{n\mu}}\right)\]

\[-\frac{m}{(m+n)^2}Z_1^2 + \frac{n}{(m+n)^2}Z_2^2\]

\[-\frac{\sqrt{mn}}{m+n}Z_1Z_2 + O_p(\min(m, n)^{-3/2}) + \log \mu\]

On further algebra, from above, by breaking down the terms, and on cancellation,

\[-\log \Lambda_{m,n} = Z_1^2 \left(\frac{1}{2} - \frac{m}{2(m+n)}\right) + Z_2^2 \left(\frac{1}{2} - \frac{n}{2(m+n)}\right)\]

Assuming that \(m/(m+n) \to \lambda\), it follows that,

\[-\log \Lambda_{m,n} = \frac{1}{2} \left(\sqrt{\frac{n}{m+n}}Z_1 - \sqrt{\frac{m}{m+n}}Z_2\right)^2 + o_p(1)\]

where \(N_1, N_2\) are independent \(N(0, 1)\). Therefore, \(-2\log \Lambda_{m,n} \overset{\mathcal{L}}{\longrightarrow} \chi^2_1\), since \(\sqrt{1 - \lambda N_1 - \sqrt{\lambda}N_2} \sim N(0, 1)\).

**Example 21.8.** This example gives a general formula for the LRT statistic for testing about the natural parameter vector in \(q\)-dimensional multiparameter exponential family. What makes the example special is a link of the LRT to the Kullback-Leibler divergence measure in the exponential family.

Suppose \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} p(x|\theta) = e^{\theta T(x)}c(\theta)h(x)(d\mu)\). Suppose we want to test \(\theta \in \Theta_0\) vs. \(H_1: \theta \in \Theta - \Theta_0\) where \(\Theta\) is the full natural parameter space. Then, by definition,

\[-2\log \Lambda_n = -2 \log \sup_{\theta \in \Theta_0} c(\theta)^n \exp(\theta' \sum_i T(x_i))\sup_{\theta \in \Theta} c(\theta)^n \exp(\theta' \sum_i T(x_i))\]

\[= -2 \sup_{\theta \in \Theta} [n \log c(\theta_0) + n\theta_0' \mathcal{T}] + 2 \sup_{\theta \in \Theta} [n \log c(\theta) + n\theta' \mathcal{T}]\]

\[= 2n \sup_{\theta \in \Theta} \inf_{\theta_0 \in \Theta_0} [(\theta - \theta_0)' \mathcal{T} + \log c(\theta) - \log c(\theta_0)]\]

on a simple rearrangement of the terms. Recall now (see Chapter 2) that the Kullback-Leibler divergence, \(K(f, g)\) is defined to be,
\[ K(f, g) = E_f \log \frac{f}{g} = \int \log \frac{f(x)}{g(x)} f(x) \mu(dx) \]

It follows that,

\[ K(p_\theta, p_{\theta_0}) = (\theta - \theta_0)' E_{\theta} T(X) + \log c(\theta) - \log c(\theta_0) \]

for any \( \theta, \theta_0 \in \Theta \). We will write \( K(\theta, \theta_0) \) for \( K(p_\theta, p_{\theta_0}) \). Recall also that in the general exponential family, the unrestricted mle \( \hat{\theta} \) of \( \theta \) exists for all large \( n \) and furthermore, \( \hat{\theta} \) is a moment estimate given by \( E_{\theta = \hat{\theta}}(T) = \bar{T} \). Consequently,

\[ -2 \log \Lambda_n = 2n \sup_{\theta \in \Theta} \inf_{\theta_0 \in \Theta_0} [(\hat{\theta} - \theta_0)' E_{\hat{\theta}}(T) + \log c(\theta) - \log c(\theta_0) +
\]

\[ (\theta - \hat{\theta})' E_{\hat{\theta}}(T) + \log c(\theta) - \log c(\hat{\theta})] \]

\[ = 2n \sup_{\theta \in \Theta} \inf_{\theta_0 \in \Theta_0} [K(\hat{\theta}, \theta_0) + (\theta - \hat{\theta})' E_{\hat{\theta}}(T) + \log c(\theta) - \log c(\hat{\theta})] \]

\[ = 2n \inf_{\theta_0 \in \Theta_0} K(\hat{\theta}, \theta_0) + 2n \sup_{\theta \in \Theta} [(\theta - \hat{\theta})' \bar{T}_n + \log c(\theta) - \log c(\hat{\theta})] \]

The second term in the above vanishes as the the supremum is attained at \( \theta = \hat{\theta} \). Thus we ultimately have the identity,

\[ -2 \log \Lambda_n = 2n \inf_{\theta_0 \in \Theta_0} K(\hat{\theta}, \theta_0) \]

The connection is beautiful. Kullback-Leibler divergence being a measure of distance, \( \inf_{\theta_0 \in \Theta_0} K(\hat{\theta}, \theta_0) \) quantifies the disagreement between the null and the estimated value of the true parameter \( \theta \), when estimated by the mle. The formula says that when the disagreement is small one should accept \( H_0 \).

The link to the K-L divergence is special for the exponential family; for example, if

\[ p_{\theta}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}, \]

then, for any \( \theta, \theta_0 \), \( K(\theta, \theta_0) = (\theta - \theta_0)^2 / 2 \) by a direct calculation. Therefore,

\[ -2 \log \Lambda_n = 2n \inf_{\theta_0 \in \Theta_0} \frac{1}{2}(\hat{\theta} - \theta)^2 = n \inf_{\theta_0 \in \Theta_0} (X - \theta_0)^2 \]

If \( H_0 \) is simple, i.e., \( H_0 : \theta = \theta_0 \), then the above expression simplifies to \( -2 \log \Lambda_n = n(X - \theta_0)^2 \), just the familiar chisquare statistic.
21.3 Asymptotic Theory of Likelihood Ratio Test Statistics

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$, densities with respect to some dominating measure $\mu$. We present the asymptotic theory of the LRT statistic for two types of null hypothesis.

Suppose $\theta \in \Theta \subseteq \mathbb{R}^q$ for some $0 < q < \infty$, and suppose as an affine subspace of $\mathbb{R}^q$, $\dim \Theta = q$. One type of null hypothesis we consider is $H_0 : \theta = \theta_0$ (specified). A second type of null hypothesis is that for some $0 < r < q$, and functions $h_1(\theta), \ldots, h_r(\theta)$, linearly independent,

$$h_1(\theta) = h_2(\theta) = \cdots = h_r(\theta) = 0.$$

For each case, under regularity conditions to be stated below, the LRT statistic is asymptotically a central chi-square under $H_0$ with a fixed degree of freedom, including the case where $H_0$ is composite.

**Example 21.9.** Suppose $X_{11}, \ldots, X_{1n_1} \overset{iid}{\sim} \text{Poisson (}\theta_1\text{)}, X_{21}, \ldots, X_{2n_2} \overset{iid}{\sim} \text{Poisson (}\theta_2\text{)}$ and $X_{31}, \ldots, X_{3n_3} \overset{iid}{\sim} \text{Poisson (}\theta_3\text{)}$ where $0 < \theta_1, \theta_2, \theta_3 < \infty$, and all observations are independent. An example of $H_0$ of the first type is $H_0 : \theta_1 = 1, \theta_2 = 2, \theta_3 = 1$. An example of $H_0$ of the second type is $H_0 : \theta_1 = \theta_2 = \theta_3$. The functions $h_1, h_2$ may be chosen as $h_1(\theta) = \theta_1 - \theta_2, h_2(\theta) = \theta_2 - \theta_3$.

In the first case, $-2 \log \Lambda_n$ is asymptotically $\chi^2_q$ under $H_0$ and in the second case it is asymptotically $\chi^2_2$ under each $\mu$. This is a special example of a theorem originally stated by Wilks (1938); also see Lawley (1956). The degree of freedom of the asymptotic chi-square distribution under $H_0$ is the number of independent constraints specified by $H_0$; it is useful to remember this as a general rule.

**Theorem 21.1.** Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta(x) = f(x|\theta) = \frac{dP_\theta}{d\mu}$ for some dominating measure $\mu$. Suppose $\theta \in \Theta$, an affine subspace of $\mathbb{R}^q$ of dimension $q$. Suppose that the conditions required for asymptotic normality of any strongly consistent sequence of roots $\hat{\theta}_n$ of the likelihood equation hold (see Chapter 16). Consider the problem $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Let

$$\Lambda_n = \frac{\prod_{i=1}^n f(x_i|\theta_0)}{l(\hat{\theta}_n)},$$

where $l(\cdot)$ denotes the likelihood function. Then $-2 \log \Lambda_n \overset{L}{\to} \chi^2_q$ under $H_0$.

The proof can be seen in Ferguson (1996), Bickel and Doksum (2001), or Sen and...
Remark: As we have seen in our illustrative examples earlier in this chapter, the theorem can fail if the regularity conditions fail to hold. For instance, if the null value is a boundary point in a constrained parameter space, then the result will fail.

Remark: This theorem is used in the following way. To use the LRT with an exact level \( \alpha \), we need to find \( c_{n,\alpha} \) such that \( P_{H_0}(\Lambda_n < c_{n,\alpha}) = \alpha \). Generally, \( \Lambda_n \) is so complicated as a statistic that one cannot find \( c_{n,\alpha} \) exactly. Instead we use the test that rejects \( H_0 \), when

\[
-2 \log \Lambda_n > \chi^2_{\alpha, q}.
\]

The theorem above implies that \( P(-2 \log \Lambda_n > \chi^2_{\alpha, q}) \to \alpha \) and so \( P(-2 \log \Lambda_n > \chi^2_{\alpha, q}) - P(\Lambda_n < c_{n,\alpha}) \to 0 \) as \( n \to \infty \).

Under the second type of null hypothesis the same result holds with the degree of freedom being \( r \). Precisely, the following holds:

**Theorem 21.2.** Assume all the regularity conditions in the previous theorem. Define \( H_{q \times r} = H(\theta) = \left( \frac{\partial}{\partial \theta_j} h_j(\theta) \right)_{1 \leq j \leq q} ; \) it is assumed that the required partial derivatives exist. Suppose for each \( \theta \in \Theta_0 \), \( H \) has full column rank. Define

\[
\Lambda_n = \sup_{\theta: h_j(\theta) = 0, 1 \leq j \leq r} \frac{l(\theta)}{l(\hat{\theta}_n)}.
\]

Then, \( -2 \log \Lambda_n \xrightarrow{\mathcal{L}} \chi^2_r \) under each \( \theta \) in \( H_0 \).

A proof can be seen in Sen and Singer(1993).

### 21.4 Distribution under Alternatives

To find the power of the test that rejects \( H_0 \) when \( -2 \log \Lambda_n > \chi^2_{\alpha,d} \) for some \( d \), one would need to know the distribution of \( -2 \log \Lambda_n \) at the particular \( \theta = \theta_1 \) value where we want to know the power. But the distribution under \( \theta_1 \) of \( -2 \log \Lambda_n \) for a fixed \( n \) is also generally impossible to find. So we may appeal to asymptotics.

However, there cannot be a nondegenerate limit distribution on \([0, \infty)\) for \( -2 \log \Lambda_n \) under a fixed \( \theta_1 \) in the alternative. The following simple example illustrates this difficulty.
Example 21.10. Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu, \sigma^2$ both unknown, and we wish to test $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$. We saw earlier that

$$
\Lambda_n = \left( \frac{\sum (X_i - \bar{X})^2}{\sum X_i^2} \right)^{n/2} = \left( \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - X)^2 + n\bar{X}^2} \right)^{n/2}.
$$

$$
\Rightarrow -2 \log \Lambda_n = n \log \left( 1 + \frac{n\bar{X}^2}{\sum (X_i - X)^2} \right) = n \log \left( 1 + \frac{\bar{X}^2}{\frac{1}{n} \sum (X_i - \bar{x})^2} \right)
$$

Consider now a value $\mu \neq 0$. Then $\bar{X}^2 \overset{a.s}{\to} \mu^2(>0)$ and $\frac{1}{n} \sum (X_i - \bar{X})^2 \overset{a.s}{\to} \sigma^2$. Therefore, clearly $-2 \log \Lambda_n \overset{a.s}{\to} \infty$ under each fixed $\mu \neq 0$. There cannot be a bona fide limit distribution for $-2 \log \Lambda_n$ under a fixed alternative $\mu$.

However, if we let $\mu$ depend on $n$ and take $\mu_n = \frac{\Delta}{\sqrt{n}}$ for some fixed but arbitrary $\Delta$, $0 < \Delta < \infty$, then $-2 \log \Lambda_n$ still has a bona fide limit distribution under the sequence of alternatives $\mu_n$.

Sometimes alternatives of the form $\mu_n = \frac{\Delta}{\sqrt{n}}$ are motivated by arguing that one wants the test to be powerful at values of $\mu$ close to the null value. Such a property would correspond to a sensitive test. These alternatives are called Pitman alternatives.

The following result holds. We present the case $h_i(\theta) = \theta_i$ for notational simplicity.

**Theorem 21.3.** Assume the same regularity conditions as in the previous theorems. Let $\theta_{q \times 1} = (\theta_1, \eta)'$ where $\theta_1$ is $r \times 1$ and $\eta$ is $(q - r) \times 1$, a vector of nuisance parameters. Let $H_0 : \theta_1 = \theta_{10}$ (specified) and let $\theta_n = (\theta_{10} + \frac{\Delta}{\sqrt{n}}, \eta)$. Let $I(\theta)$ be the Fisher Information matrix, with $I_{ij}(\theta) = -E \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X|\theta) \right)$. Denote $V(\theta) = I^{-1}(\theta)$ and $V_r(\theta)$ the upper $(r \times r)$ principal submatrix in $V(\theta)$. Then $-2 \log \Lambda_n \overset{\mathcal{L}}{\to} NC \chi^2(r, \Delta V_r^{-1}(\theta_0, \eta) \Delta)$ where $NC \chi^2(r, \delta)$ denotes the noncentral $\chi^2$ distribution with $r$ degrees of freedom and noncentrality parameter $\delta$.

**Remark:** The theorem is essentially proved in Sen and Singer (1993). The theorem can be restated in an obvious way for the testing problem $H_0 : g(\theta) = (g_1(\theta), \ldots, g_r(\theta)) = 0$.

**Example 21.11.** Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. Let $\theta = (\mu, \sigma^2) = (\theta_1, \eta)$, where $\theta_1 = \mu$ and $\eta = \sigma^2$. Thus $q = 2$ and $r = 1$. Suppose we want to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. Consider alternatives $\theta_n = (\mu_0 + \frac{\Delta}{\sqrt{n}}, \eta)$. By a familiar calculation,

$$
I(\theta) = \begin{pmatrix} \frac{1}{\eta} & 0 \\ 0 & \frac{1}{2\eta^2} \end{pmatrix} \Rightarrow V(\theta) = \begin{pmatrix} \eta & 0 \\ 0 & 2\eta^2 \end{pmatrix}
$$

and $V_r(\theta) = \eta$. 331
Therefore, by the above theorem $-2\log \Lambda_n \xrightarrow{p_{\theta_0}} NC\chi^2(1, \frac{\Delta^2}{\eta}) = NC\chi^2(1, \frac{\Delta^2}{\eta^2})$

**Remark:** One practical use of this theorem is in approximating the power of a test that rejects $H_0$ if $-2\log \Lambda_n > c$.

Suppose we are interested in approximating the power of the test when $\theta_1 = \theta_{10} + \epsilon$. Formally, set $\epsilon = \frac{\Delta}{\sqrt{n}} \Rightarrow \Delta = \epsilon \sqrt{n}$. Use $NC\chi^2(1, \frac{nc^2}{\operatorname{Var}(\theta_{10}, \eta)})$ as an approximation to the distribution of $-2\log \Lambda_n$ at $\theta = (\theta_{10} + \epsilon, \eta)$. Thus, the power will be approximated as $P(NC\chi^2(1, \frac{nc^2}{\operatorname{Var}(\theta_{10}, \eta)}) > c)$.

### 21.5 Bartlett Correction

We saw that under regularity conditions $-2\log \Lambda_n$ has asymptotically a $\chi^2(r)$ distribution under $H_0$ for some $r$. Certainly, $-2\log \Lambda_n$ is not exactly distributed as $\chi^2(r)$. In fact, even the expectation is not $r$. It turns out that $E(-2\log \Lambda_n)$ under $H_0$ admits an expansion of the form $r(1 + \frac{a}{n} + o(n^{-1}))$. Consequently, $E\left(\frac{-2\log \Lambda_n}{1 + \frac{a}{n} + R_n}\right) = r$ where $R_n$ is the remainder term in $E(-2\log \Lambda_n)$. Moreover, $E\left(\frac{-2\log \Lambda_n}{1 + \frac{a}{n}}\right) = r + O(n^{-2})$. Thus, we gain higher order accuracy in the mean by rescaling $-2\log \Lambda_n$ to $-2\log \Lambda_n - \frac{a}{n}$. For the absolutely continuous case, the higher order accuracy even carries over to the $\chi^2$ approximation itself, i.e.

$$P\left(\frac{-2\log \Lambda_n}{1 + \frac{a}{n}} \leq c\right) = P(\chi^2(r) \leq c) + O(n^{-2});$$

Due to the rescaling, the $\frac{1}{n}$ term that would otherwise be present has vanished. This type of rescaling is known as the Bartlett correction. See Bartlett(1937), Wilks(1938), McCullagh and Cox(1986), and Barndorff-Nielsen and Hall(1988) for detailed treatment of Bartlett corrections in general circumstances.

### 21.6 The Wald and Rao Score Tests

Competitors to the LRT are available in the literature. They are general and can be applied to a wide selection of problems. Typically, the three procedures are asymptotically first order equivalent. See Wald(1943), and Rao(1948) for the first introduction of these procedures.

We define the Wald and the score statistics first. We have not stated these definitions under the weakest possible conditions. Also note that establishing the asymptotics
of these statistics will require more conditions than are needed for just defining them.

**Definition 21.1.** Let \( X_1, X_2, \ldots, X_n \) be iid \( f(x|\theta) = \frac{dP_n}{d\mu}, \theta \in \Theta \subseteq \mathbb{R}^q, \mu \) a \( \sigma \)-finite measure. Suppose the Fisher information matrix \( I(\theta) \) exists at all \( \theta \). Consider \( H_0 : \theta = \theta_0 \). Let \( \hat{\theta}_n \) be the MLE of \( \theta \). Define \( Q_W = n(\hat{\theta}_n - \theta_0)'I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \). This is the Wald test statistic for \( H_0 : \theta = \theta_0 \).

**Definition 21.2.** Let \( X_1, X_2, \ldots, X_n \) be iid \( f(x|\theta) = \frac{dP_n}{d\mu}, \theta \in \Theta \subseteq \mathbb{R}^q, \mu \) a \( \sigma \)-finite measure. Consider testing \( H_0 : \theta = \theta_0 \). Assume that \( f(x|\theta) \) is partially differentiable with respect to each coordinate of \( \theta \) for every \( x, \theta \), and the Fisher information matrix exists and is invertible at \( \theta = \theta_0 \). Let \( U_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(x_i|\theta) \). Define \( Q_S = \frac{1}{n} U_n'(\theta_0)I^{-1}(\theta_0)U_n(\theta_0) \); \( Q_S \) is the Rao score test statistic for \( H_0 : \theta = \theta_0 \).

**Remark:** As we have seen in Chapter 16, the MLE \( \hat{\theta}_n \) is asymptotically multivariate normal under the Cramér-Rao regularity conditions. Therefore, \( n(\hat{\theta}_n - \theta_0)'I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \) is asymptotically a central chi-square, provided \( I(\theta) \) is smooth at \( \theta_0 \). The Wald test rejects \( H_0 \) when \( Q_W = n(\hat{\theta}_n - \theta_0)'I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \) is larger than a chi-square percentile.

On the other hand, under simple moment conditions on the score function \( \frac{\partial}{\partial \theta} \log f(x_i|\theta) \), by the CLT \( \frac{U_n(\theta)}{\sqrt{n}} \) is asymptotically multivariate normal, and therefore the Rao score statistic \( Q_S = \frac{1}{n} U_n'(\theta_0)I^{-1}(\theta_0)U_n(\theta_0) \) is asymptotically a chi-square. The score test rejects \( H_0 \) when \( Q_S \) is larger than a chi-square percentile. Notice that in this case of a simple \( H_0 \), the \( \chi^2 \) approximation of \( Q_S \) holds under less assumptions than would be required for \( Q_W \) or \(-2 \log \Lambda_n\). Note also that to evaluate \( Q_S \), computation of \( \hat{\theta}_n \) is not necessary, which can be a major advantage in some applications. Here are the asymptotic chi-square results for these two statistics. A version of the two parts of the theorem below is proved in Serfling (1980) and in Sen and Singer (1993). Also see van der Vaart (1998).

**Theorem 21.4.**

(a) If \( f(x|\theta) \) can be differentiated twice under the integral sign, \( I(\theta_0) \) exists and is invertible, and \( \{x : f(x|\theta) > 0\} \) is independent of \( \theta \), then under \( H_0 : \theta = \theta_0 \), \( Q_S \xrightarrow{\mathcal{L}} \chi^2_q \), where \( q = \text{dimension of } \Theta \).

(b) Assume the Cramér-Rao regularity conditions for the asymptotic multivariate normality of the MLE, and assume that the Fisher information matrix is con-
continuous in a neighborhood of $\theta_0$. Then, under $H_0: \theta = \theta_0$, $Q_W \xrightarrow{\mathcal{L}} \chi^2_q$, with $q$ as above.

**Remark:** $Q_W$ and $Q_S$ can be defined for composite nulls of the form $H_0: g(\theta) = (g_1(\theta), \ldots, g_r(\theta)) = 0$ and once again, under appropriate conditions, $Q_W$ and $Q_S$ are asymptotically $\chi^2_q$ for any $\theta \in H_0$. However, now the definition of $Q_S$ requires calculation of the restricted MLE $\hat{\theta}_{n,H_0}$ under $H_0$.

Again, consult Sen and Singer (1993) for a proof.

Comparisons between the LRT, the Wald test and Rao’s score test have been made using higher order asymptotics. See Mukerjee and Reid (2001), in particular.

### 21.7 Likelihood Ratio Confidence Intervals

The usual duality between testing and confidence intervals says that the acceptance region of a test with size $\alpha$ can be inverted to give a confidence set of coverage probability at least $(1 - \alpha)$. In other words, suppose $A(\theta_0)$ is the acceptance region of a size $\alpha$ test for $H_0: \theta = \theta_0$ and define $S(x) = \{\theta_0 : x \in A(\theta_0)\}$. Then $P_{\theta_0}(S(x) \ni \theta_0) \geq 1 - \alpha$ and hence $S(x)$ is a $100(1 - \alpha)%$ confidence set for $\theta$.

This method is called the inversion of a test. In particular, each of the LRT, the Wald test, and the Rao score test can be inverted to construct confidence sets that have asymptotically a $100(1 - \alpha)%$ coverage probability.

The confidence sets constructed from the LRT, the Wald test, and the score test are respectively called the likelihood ratio, Wald and score confidence sets. Of these, the Wald and the score confidence sets are ellipsoids because of how the corresponding test statistics are defined. The likelihood ratio confidence set is typically more complicated. Here is an example.

**Example 21.12.** Suppose $X_i \overset{iid}{\sim} \text{Bin}(1, p), 1 \leq i \leq n$. For testing $H_0: p = p_0$ vs. $H_1: p \neq p_0$, the LRT statistic is

$$\Lambda_n = \frac{p_0^x(1 - p_0)^{n-x}}{\sup_p p^x(1 - p)^{n-x}} \quad \text{where} \quad x = \sum_{i=1}^{n} X_i$$

$$= \frac{p_0^x(1 - p_0)^{n-x}}{(\frac{x}{n})^x(1 - (\frac{x}{n}))^{n-x}} = \frac{p_0^x(1 - p_0)^{n-x}n^n}{x^x(n - x)^{n-x}}$$

Thus the LR confidence set is of the form $S_{LR}(x) := \{p_0 : \Lambda_n \geq k\} = \{p_0 : p_0^x(1 - p_0)^{n-x} \geq k^*\} = \{p_0 : x \log p_0 + (n - x) \log(1 - p_0) \geq \log k^*\}$. 

334
The function $x \log p_0 + (n - x) \log(1 - p_0)$ is concave in $p_0$ and therefore $S_{LR}(x)$ is an interval. The interval is of the form $[0, u]$ or $[l, 1]$ or $[l, u]$ for $0 < l, u < 1$. However, $l, u$ cannot be written in closed form; see Brown, Cai, DasGupta (2001) for asymptotic expansions for them.

Next, $Q_W = n(\hat{p} - p_0)^2 I(\hat{p})$, where $\hat{p} = \frac{x}{n}$ is the MLE of $p$ and $I(p) = \frac{1}{p(1-p)}$ is the Fisher information function. Therefore

$Q_W = n(\frac{x}{n} - p_0)^2 \frac{\hat{p}(1-\hat{p})}{\hat{p}(1-\hat{p})}$.

The Wald confidence interval is $S_W = \{ p_0 : \frac{n(\hat{p} - p_0)^2}{\hat{p}(1-\hat{p})} \leq \chi_\alpha^2 \} = \{ p_0 : (\hat{p} - p_0)^2 \leq \frac{\hat{p}(1-\hat{p})}{n} \chi_\alpha^2 \}$ = $\{ p_0 : |\hat{p} - p_0| \leq \frac{\chi_\alpha}{\sqrt{n}} \sqrt{\hat{p}(1 - \hat{p})} \} = [\hat{p} - \frac{\chi_\alpha}{\sqrt{n}} \sqrt{\hat{p}(1 - \hat{p})}, \hat{p} + \frac{\chi_\alpha}{\sqrt{n}} \sqrt{\hat{p}(1 - \hat{p})}]$

This is the textbook confidence interval for $p$.

For the score test statistic, we need $U_n(p) = \sum_{i=1}^n \frac{\partial}{\partial p} \log f(x_i|p) = \sum_{i=1}^n \frac{\partial}{\partial p} [x_i \log p + (1 - x_i) \log(1 - p)] = \frac{x - np}{p(1-p)}$. Therefore, $Q_S = \frac{1}{n} \frac{(x - np_0)^2}{p_0(1-p_0)}$ and the score confidence interval is $S_S = \{ p_0 : \frac{(x - np_0)^2}{n p_0(1-p_0)} \leq \chi_\alpha^2 \} = \{ p_0 : (x - np_0)^2 \leq n p_0(1-p_0) \} = \{ p_0 : p_0^2(n^2 + n \chi_\alpha^2) - p_0(2nx + n \chi_\alpha^2) + x^2 \leq 0 \} = [l_S, u_S]$, where $l_S, u_S$ are the roots of the quadratic equation $p_0^2(n_\alpha^2 + n \chi_\alpha^2) - p_0(2nx + n \chi_\alpha^2) + x^2 = 0$.

These intervals all have the property

$$\lim_{n \to \infty} P_{p_0}(S \ni p_0) = 1 - \alpha.$$

See Brown, Cai, DasGupta (2001) for a comparison of their exact coverage properties; they show that the performance of the Wald confidence interval is extremely poor, and the score and the likelihood ratio intervals perform much better.
21.8 Exercises

For each of the following problems 1 to 18, derive the LRT statistic and find its limiting distribution under the null, if there is a meaningful one:

Exercise 21.1. \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} Ber(p), H_0 : p = \frac{1}{2} \text{ vs. } H_1 : p \neq \frac{1}{2}\). Is the LRT statistic a monotone function of \(|X - \frac{n}{2}|\)?

Exercise 21.2. \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 1), H_0 : a \leq \mu \leq b, H_1 : \text{not } H_0\).

Exercise 21.3. *\(X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 1), H_0 : \mu \text{ is an integer, } H_1 : \mu \text{ is not an integer.}\)

Exercise 21.4. *\(X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 1), H_0 : \mu \text{ is rational, } H_1 : \mu \text{ is irrational.}\)

Exercise 21.5. \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} Ber(p_1), Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} Ber(p_2), H_0 : p_1 = p_2, H_1 : p_1 \neq p_2; \text{ assume all } m + n \text{ observations are independent.}\)

Exercise 21.6. \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} Ber(p_1), Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} Ber(p_2); H_0 : p_1 = p_2 = \frac{1}{2}, H_1 : \text{not } H_0; \text{ all observations are independent.}\)

Exercise 21.7. \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} Pois(\mu), Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} Pois(\lambda), H_0 : \mu = \lambda = 1, H_1 : \text{not } H_0; \text{ all observations are independent.}\)

Exercise 21.8. *\(X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2), Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} N(\mu_2, \sigma_2^2), H_0 : \mu_1 = \mu_2, \sigma_1 = \sigma_2, H_1 : \text{not } H_0 \) (i.e., test that two normal distributions are identical); again, all observations are independent.

Exercise 21.9. *\(X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2), H_0 : \mu = 0, \sigma = 1, H_1 : \text{not } H_0.\)

Exercise 21.10. *\(X_1, X_2, \ldots, X_n \overset{iid}{\sim} BVN(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho), H_0 : \rho = 0, H_1 : \rho \neq 0.\)

Exercise 21.11. *\(X_1, X_2, \ldots, X_n \overset{iid}{\sim} MVN(\mu, I), H_0 : \mu = 0, H_1 : \mu \neq 0.\)

Exercise 21.12. *\(X_1, X_2, \ldots, X_n \overset{iid}{\sim} MVN(\mu, \Sigma), H_0 : \mu = 0, H_1 : \mu \neq 0.\)

Exercise 21.13. \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} MVN(\mu, \Sigma), H_0 : \Sigma = I, H_1 : \Sigma \neq I.\)

Exercise 21.14. *\(X_{ij} \overset{\text{indep}}{\sim} U[0, \theta_i], j = 1, 2, \ldots, n, i = 1, 2, \ldots, k; H_0 : \theta_i \text{ are equal, } H_1 : \text{not } H_0.\)

Remark: Notice the EXACT chisquare distribution.

336
Exercise 21.15. \( X_{ij} \overset{\text{indep}}{\sim} N(\mu_i, \sigma^2), H_0 : \mu_i \text{ are equal}, H_1 : \text{not } H_0. \)

Remark: Notice that you get the usual F test for ANOVA.

Exercise 21.16. \( * X_1, X_2, \ldots, X_n \overset{iid}{\sim} c(\alpha) Exp[-|x|^\alpha], \) where \( c(\alpha) \) is the normalizing constant, \( H_0 : \alpha = 2, H_1 : \alpha \neq 2. \)

Exercise 21.17. \( (X_i, Y_i), i = 1, 2, \ldots, n \overset{iid}{\sim} BVN(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho), H_0 : \mu_1 = \mu_2, H_1 : \mu_1 \neq \mu_2. \)

Remark: This is the paired t test.

Exercise 21.18. \( X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2), Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} N(\mu_2, \sigma_2^2), H_0 : \sigma_1 = \sigma_2, H_1 : \sigma_1 \neq \sigma_2; \) assume all observations are independent.

Exercise 21.19. \( * \) Suppose \( X_1, X_2, \ldots, X_n \) are iid from a two component mixture \( pN(0, 1) + (1-p)N(\mu, 1), \) where \( p, \mu \) are unknown.

Consider testing \( H_0 : \mu = 0 \text{ vs.} H_1 : \mu \neq 0. \) What is the LRT statistic? Simulate the value of the LRT by gradually increasing \( n \) and observe its slight decreasing trend as \( n \) increases.

Remark: It has been shown by Jon Hartigan that in this case \(-2\log \Lambda_n \to \infty.\)

Exercise 21.20. \( * \) Suppose \( X_1, X_2, \ldots, X_n \) are iid from a two component mixture \( pN(\mu_1, 1) + (1-p)N(\mu_2, 1), \) and suppose we know that \( |\mu_i| \leq 1, i = 1, 2. \) Consider testing \( H_0 : \mu_1 = \mu_2 \text{ vs.} H_1 : \mu_1 \neq \mu_2, \) when \( p, 0 < p < 1 \) is unknown.

What is the LRT statistic? Simulate the distribution of the LRT under the null by gradually increasing \( n. \)

Remark: The asymptotic null distribution is nonstandard, but known.

Exercise 21.21. \( * \) For each of Exercises 1, 6, 7, 9, 10, 11, simulate the exact distribution of \(-2\log \Lambda_n \) for \( n = 10, 20, 40, 100 \) under the null, and compare its visual match to the limiting null distribution.

Exercise 21.22. \( * \) For each of Exercises 1, 5, 6, 7, 9, simulate the exact power of the LRT at selected alternatives, and compare it to the approximation to the power obtained from the limiting nonnull distribution, as indicated in text.

Exercise 21.23. \( * \) For each of Exercises 1, 6, 7, 9, compute the constant \( a \) of the Bartlett correction of the likelihood ratio, as indicated in text.
Exercise 21.24. * Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Plot the Wald and the Rao score confidence sets for $\theta = (\mu, \sigma)$ for a simulated data set. Observe the visual similarity of the two sets by gradually increasing $n$.

Exercise 21.25. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Pois(\mu)$. Compute the likelihood ratio, Wald, and Rao score confidence intervals for $\mu$ for a simulated data set. Observe the similarity of the limits of the intervals for large $n$.

Exercise 21.26. * Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} Pois(\mu)$. Derive a two-term expansion for the expected lengths of the Wald and the Rao score intervals for $\mu$. Plot them for selected $n$ and check if either one is usually shorter than the other interval.

Exercise 21.27. Consider the linear regression model $E(Y_i) = \beta_0 + \beta_1 x_i, i = 1, 2, \ldots, n$, with $Y_i$ being mutually independent, distributed as $N(\beta_0 + \beta_1 x_i, \sigma^2), \sigma^2$ unknown. Find the likelihood ratio confidence set for $(\beta_0, \beta_1)$.

21.9 References


Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large, Trans. Amer. Math. Soc., 5, 426-482.