Many familiar distributions are special cases, including the exponential ($P = 1$) and chi-squared ($\lambda = \frac{1}{2}, P = n/2$). The **Erlang distribution** results if P is a positive integer. The mean is P/λ, and the variance is P/λ^2.

3.4.6. THE BETA DISTRIBUTION

Distributions for models are often chosen on the basis of the range within which the random variable is constrained to vary. The lognormal distribution, for example, is sometimes used to model a variable that is always nonnegative. For a variable constrained between 0 and $c > 0$, the **beta distribution** has proved useful. Its density is

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{x}{c}^{\alpha-1} 1 - \frac{x}{c}^{\beta-1} \frac{1}{c}. \tag{3-40}$$

This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if $\alpha = \beta$, asymmetric otherwise, and can be hump-shaped or U-shaped. The mean is $c\alpha/(\alpha + \beta)$, and the variance is $c^2\alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)^2]$. The beta distribution has been applied in the study of labor force participation rates.[7]

3.4.7. THE LOGISTIC DISTRIBUTION

The normal distribution is ubiquitous in econometrics. But researchers have found that for some microeconomic applications, there does not appear to be enough mass in the tails of the normal distribution; observations that a model based on normality would classify as "unusual" seem not to be very unusual at all. One approach has been to use thicker-tailed symmetric distributions. The **logistic distribution** is one candidate; the cdf for a logistic random variable is denoted

$$F(x) = \Lambda(x) = \frac{1}{1 + e^{-x}}.$$

The density is $f(x) = \Lambda(x)[1 - \Lambda(x)]$. The mean and variance of this random variable are zero and $\pi^2/3$.

3.4.8. DISCRETE RANDOM VARIABLES

Modeling in economics frequently involves random variables that take integer values. In these cases, the distributions listed thus far only provide approximations that are sometimes quite inappropriate. We can build up a class of models for discrete random variables from the **Bernoulli distribution** for a single binomial outcome (trial)

$$\text{Prob}(x = 1) = \alpha,$$
$$\text{Prob}(x = 0) = 1 - \alpha,$$

where $0 \leq \alpha \leq 1$. The modeling aspect of this specification would be the assumptions that the success probability α is constant from one trial to the next and that successive trials are independent. If so, then the distribution for x successes in n trials is the **binomial distribution,**

$$\text{Prob}(X = x) = \binom{n}{x} \alpha^x (1 - \alpha)^{n-x}, \quad x = 0, 1, \ldots, n.$$

[7] Heckman and Willis (1976).

The mean and variance of x are $n\alpha$ and $n\alpha(1 - \alpha)$, respectively. If the number of trials becomes large at the same time that the success probability becomes small so that the mean $n\alpha$ is stable, the limiting form of the binomial distribution is the **Poisson distribution,**

$$\text{Prob}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

The Poisson distribution has seen wide use in econometrics in, for example, modeling patents, crime, recreation demand, and demand for health services.

3.5. The Distribution of a Function of a Random Variable

We considered finding the expected value of a function of a random variable. It is fairly common to analyze the random variable itself, which results when we compute a function of some random variable. There are three types of transformation to consider. One discrete random variable may be transformed into another, a continuous variable may be transformed into a discrete one, and one continuous variable may be transformed into another.

The simplest case is the first one. The probabilities associated with the new variable are computed according to the laws of probability. If y is derived from x and the function is one to one, then the probability that $Y = y(x)$ equals the probability that $X = x$. If several values of x yield the same value of y, then $\text{Prob}(Y = y)$ is the sum of the corresponding probabilities for x.

The second type of transformation is illustrated by the way individual data on income are typically obtained in a survey. Income in the population can be expected to be distributed according to some skewed, continuous distribution such as the one shown in Figure 3.1.

Data are normally reported categorically, as shown in the lower part of the figure. Thus, the random variable corresponding to observed income is a discrete transformation of the actual underlying continuous random variable. Suppose, for example, that the transformed variable y is the mean income in the respective interval. Then

$$\begin{aligned}
\text{Prob}(Y = \mu_1) &= P(-\infty < X \leq a),\\
\text{Prob}(Y = \mu_2) &= P(a < X \leq b),\\
\text{Prob}(Y = \mu_3) &= P(b < X \leq c),
\end{aligned}$$

and so on, which illustrates the general procedure.

If x is a continuous random variable with pdf $f_x(x)$ and if $y = g(x)$ is a continuous monotonic function of x, then the density of y is obtained by using the change of variable technique to find the cdf of y:

$$\text{Prob}(yb) = \int_{-\infty}^{b} f_x(g^{-1}(y)) \mid g^{-1\prime}(y) \mid dy.$$

This equation can now be written as

$$\text{Prob}(yb) = \int_{-\infty}^{b} f_y(y)\, dy.$$

Thus, in our examples, γ_n^2 is $O(1/n)$, $\text{Var}[x_{(1),n}]$ is $O(1/n^2)$ and $o(1/n)$, S_n is $O(n^2)$ ($\delta = -2$ in this case), $\log L(\theta)$ is $O(n)$ ($\delta = -1$), and c_n is $O(1)$ ($\delta = 0$). Important particular cases that we will encounter repeatedly in our work are sequences for which $\delta = 1$ or -1.

The notion of order of a sequence is often of interest in econometrics in the context of the variance of an estimator. Thus, we see in Section 4.4.3 that an important element of our strategy for forming an asymptotic distribution is that the variance of the limiting distribution of $\sqrt{n}(\bar{x}_n - \mu)/\sigma$ is $O(1)$. In Example 4.17, the variance of m_2 is the sum of three terms that are $O(1/n)$, $O(1/n^2)$, and $O(1/n^3)$. The sum is $O(1/n)$, because $n\text{Var}[m_2]$ converges to $\mu_4 - \sigma^4$, the numerator of the first, or *leading term*, whereas the second and third terms converge to zero. This term is also the *dominant term* of the sequence. Finally, consider the two divergent examples in the preceding list. S_n is simply a deterministic function of n that explodes. However, $\log L(\theta) = n \log \theta - \theta\Sigma_i x_i$ is the sum of a constant that is $O(n)$ and a random variable with variance equal to n/θ. The random variable "diverges" in the sense that its variance grows without bound as n increases.

4.5. Efficient Estimation: Maximum Likelihood

The principle of **maximum likelihood** provides a means of choosing an asymptotically efficient estimator for a parameter or a set of parameters. The logic of the technique is best illustrated in the setting of a discrete distribution. Consider a random sample of 10 observations from a Poisson distribution: 5, 0, 1, 1, 0, 3, 2, 3, 4, and 1. The density for each observation is

$$f(x_i, \theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}.$$

Since the observations are independent, their joint density, which was identified in Section 4.3.2 as the **likelihood** for the sample, is

$$\begin{aligned} f(x_1, x_2, \ldots, x_{10} | \theta) &= \prod_{i=1}^{10} f(x_i, \theta) \\ &= \frac{e^{-10\theta}\,\theta^{\Sigma_{i=1}^{10} x_i}}{\prod_{i=1}^{10} x_i!} \\ &= \frac{e^{-10\theta}\,\theta^{20}}{207,360}. \end{aligned}$$

The last line gives the probability of observing *this particular sample,* assuming that a Poisson distribution with as yet unknown parameter θ generated the data. What value of θ would make this sample most probable? Figure 4.6 plots this function for various values of θ. It has a single mode at $\theta = 2$, which would be the maximum likelihood estimate, or MLE, of θ.

Consider maximizing the function directly. Since the log function is monotonically increasing and easier to work with, we usually maximize $\ln L(\theta)$ instead:

$$\begin{aligned} \ln L(\theta) &= -10\theta + 20 \ln \theta - 12.242, \\ \frac{d \ln L(\theta)}{d\theta} &= -10 + \frac{20}{\theta} = 0 \Rightarrow \hat{\theta} = 2, \end{aligned}$$

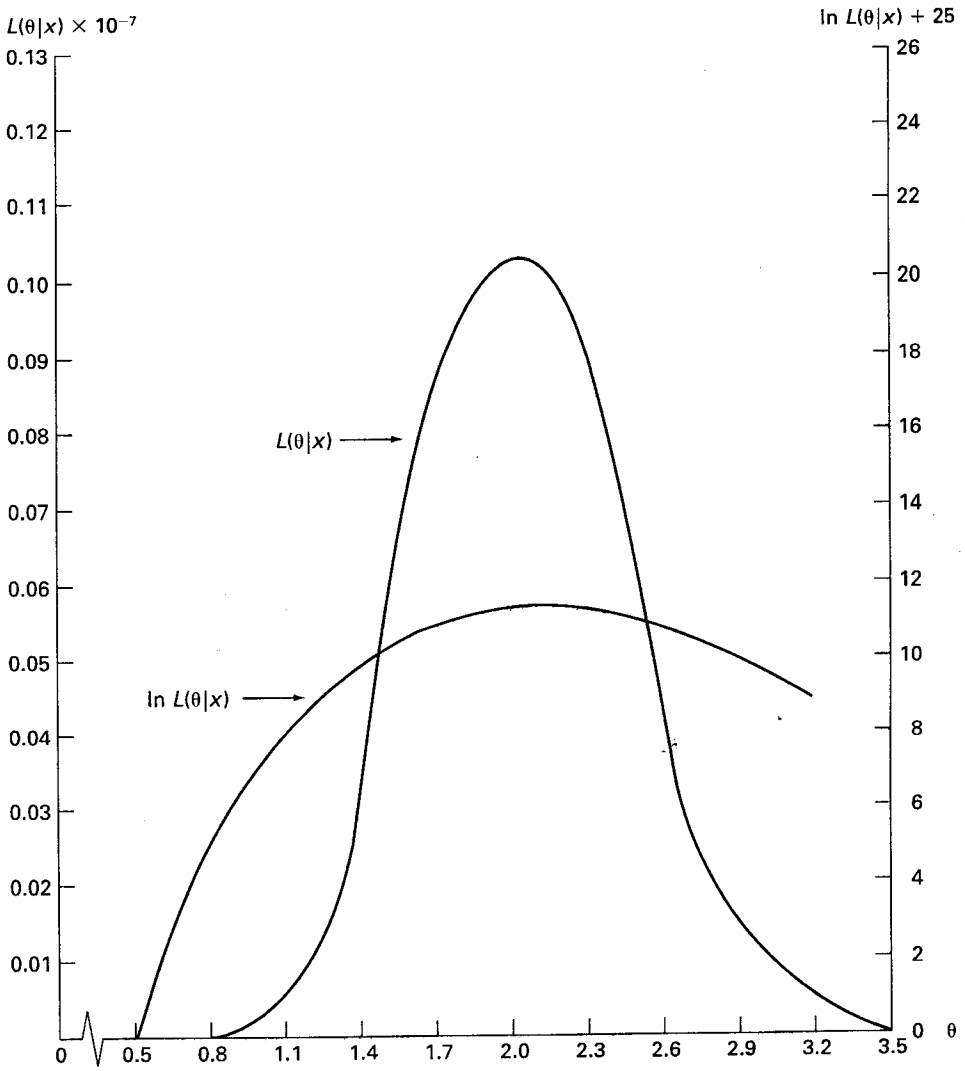


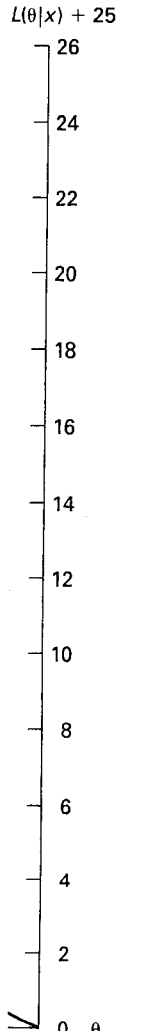
FIGURE 4.6 Likelihood and Log-likelihood Functions for a Poisson Distribution

and

$$\frac{d^2 \ln L(\theta)}{d\theta^2} = \frac{-20}{\theta^2} < 0 \Rightarrow \text{this is a maximum.}$$

The solution is the same as before. Figure 4.6 also plots the log of L to illustrate the result.

In a continuous distribution, the analogy to the probability of observing the given sample is not exact, since a particular sample has probability zero. The principle is the same for either, however. The joint density of the n observations, which may be univariate (x_i) or multivariate ($\mathbf{x}_i$), is the product of the individual densities. This joint



density is the **likelihood function,** defined as a function of the unknown parameter vector, $\boldsymbol{\theta}$:

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{x}_i, \boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{X}), \tag{4-43}$$

where $\mathbf{X}$ is used to indicate the sample data. It is usually simpler to work with the log of the likelihood function:

$$\ln L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i, \boldsymbol{\theta}). \tag{4-44}$$

The values of the parameters that maximize this function are the maximum likelihood estimates, generally denoted $\hat{\boldsymbol{\theta}}$. Since the logarithm is a monotonic function, the values that maximize L are the same as those that maximize $\ln L$. The likelihood function and its logarithm, evaluated at $\boldsymbol{\theta}$, are usually denoted L and $\ln L$, respectively. The necessary condition for maximizing $\ln L(\boldsymbol{\theta})$ is

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{4-45}$$

This is called the **likelihood equation.**

EXAMPLE 4.18 Poisson Likelihood Function

In sampling from a Poisson population,

$$\ln L(\theta) = -n\theta + (\ln \theta) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!),$$

$$\frac{\partial \ln L(\theta)}{\partial \theta} = -n + \frac{1}{\theta} \sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\theta}_{\text{ML}} = \bar{x}_n.$$

EXAMPLE 4.19 Likelihood for the Normal Distribution

In sampling from a normal distribution with mean μ and variance σ^2, the log-likelihood function and the likelihood equations for μ and σ^2 are

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^{n} \left[\frac{(x_i - \mu)^2}{\sigma^2} \right], \tag{4-46}$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0, \tag{4-47}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0. \tag{4-48}$$

To solve the likelihood equations, multiply (4-47) by σ^2 and solve for $\hat{\mu}$, then insert this solution in (4-48) and solve for σ^2. The solutions are

$$\hat{\mu}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$$

llustrate the

ng the given
nciple is the
may be uni-
s. This joint

H_0 is rejected if θ_0 exceeds the upper limit or is less than the lower limit. Equivalently, H_0 is rejected if

$$\left|\frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}\right| > C_{1-\alpha/2}.$$

In words, the hypothesis is rejected if the estimate is too far from θ_0, where the distance is measured in standard error units. The critical value is taken from the t or standard normal distribution, whichever is appropriate.

EXAMPLE 4.33 Testing a Hypothesis About a Mean with a Confidence Interval

For the results in Example 4.29, test $H_0: \mu = 1.98$ versus $H_1: \mu \neq 1.98$, assuming sampling from a normal distribution:

$$t = \left|\frac{\bar{x} - 1.98}{s/\sqrt{n}}\right| = \left|\frac{1.63 - 1.98}{0.102}\right| = 3.43.$$

The 95 percent critical value for $t(24)$ is 2.064. Therefore, reject H_0. If the critical value for the standard normal table of 1.96 is used instead, then the same result is obtained.

If the test is one-sided, as in

$$H_0: \theta \geq \theta_0,$$
$$H_1: \theta < \theta_0,$$

then the critical region must be adjusted. Thus, for this test, H_0 will be rejected if a point estimate of θ falls sufficiently below θ_0. (Tests can usually be set up by departing from the decision criterion, "What sample results are inconsistent with the hypothesis?")

EXAMPLE 4.34 One-Sided Test About a Mean

A sample of 25 from a normal distribution yields $\bar{x} = 1.63$ and $s = 0.51$. Test

$$H_0: \mu \leq 1.5,$$
$$H_1: \mu > 1.5.$$

Clearly, no observed $\bar{x}$ less than or equal to 1.5 will lead to rejection of H_0. Using the borderline value of 1.5 for μ, we obtain

$$\text{Prob}\left(\frac{\sqrt{n}(\bar{x} - 1.5)}{s} > \frac{5(1.63 - 1.5)}{0.51}\right) = \text{Prob}(t_{24} > 1.27).$$

This is approximately 0.11. This value is not unlikely by the usual standards. Hence, at a significant level of 0.11, we would not reject the hypothesis.

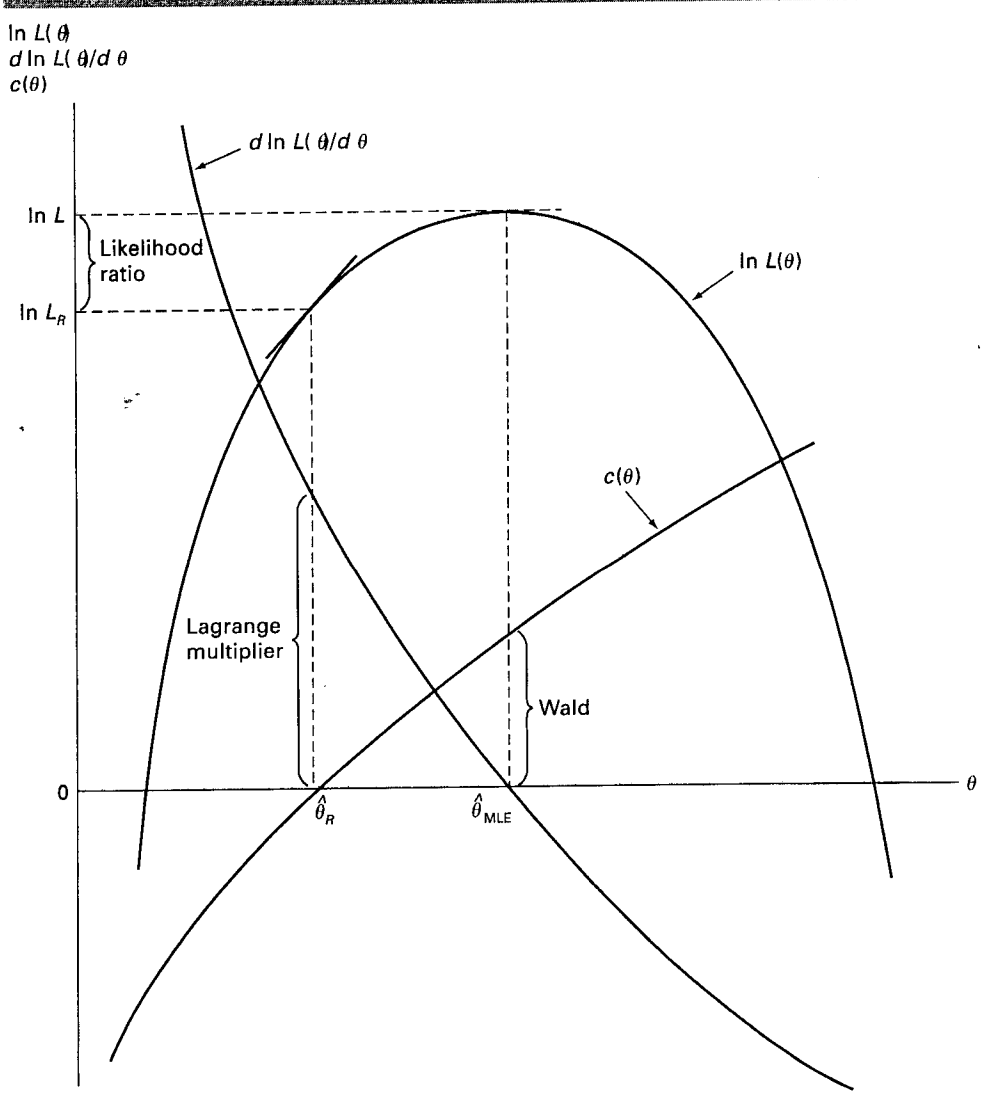
4.9.3. THREE ASYMPTOTICALLY EQUIVALENT TEST PROCEDURES

The next several sections will discuss the most commonly used test procedures: the likelihood ratio, Wald, and Lagrange multiplier tests. [Extensive discussion of these procedures is given in Godfrey (1988).] We consider maximum likelihood estimation

of a parameter θ and a test of the hypothesis H_0:$c(\theta) = 0$. The logic of the tests can be seen in Figure 4.8.[21] The figure plots the log-likelihood function ln $L(\theta)$, its derivative with respect to θ, d ln $L(\theta)/d\theta$, and the constraint $c(\theta)$. There are three approaches to testing the hypothesis suggested in the figure:

- **Likelihood ratio test.** If the restriction $c(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on

FIGURE 4.8 Three Bases for Hypothesis Tests.



[21] See Buse (1982). Note that the scale of the vertical axis would be different for each curve. As such, the points of intersection have no significance.

the difference, $\ln L - \ln L_R$, where L is the value of the likelihood function at the unconstrained value of θ and L_R is the value of the likelihood function at the restricted estimate.

- **Wald test.** If the restriction is valid, then $c(\hat{\theta}_{\text{MLE}})$ should be close to zero since the MLE is consistent. Therefore, the test is based on $c(\hat{\theta}_{\text{MLE}})$. We reject the hypothesis if this is significantly different from zero.
- **Lagrange multiplier test.** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

These three tests are asymptotically equivalent under the null hypothesis, but they can behave rather differently in a small sample. Unfortunately, their small-sample properties are unknown, except in a few special cases. As a consequence, the choice among them is typically made on the basis of ease of computation. The likelihood ratio test requires calculation of both restricted and unrestricted estimators. If both are simple to compute, then this way to proceed is convenient. The Wald test requires only the unrestricted estimator, and the Lagrange multiplier test requires only the restricted estimator. In some problems, one of these estimators may be much easier to compute than the other. For example, a linear model is simple to estimate but becomes nonlinear and cumbersome if a nonlinear constraint is imposed. In this case, the Wald statistic might be preferable. Alternatively, restrictions sometimes amount to the removal of nonlinearities, which would make the Lagrange multiplier test the simpler procedure.

4.9.3.a. The Likelihood Ratio Test

Let $\boldsymbol{\theta}$ be a vector of parameters to be estimated, and let H_0 specify some sort of restriction on these parameters. Let $\hat{\boldsymbol{\theta}}_U$ be the maximum likelihood estimate of $\boldsymbol{\theta}$ obtained without regard to the constraints, and let $\hat{\boldsymbol{\theta}}_R$ be the constrained maximum likelihood estimator. If $\hat{L}_U$ and $\hat{L}_R$ are the likelihood functions evaluated at these two estimates, then the **likelihood ratio** is

$$\lambda = \frac{\hat{L}_R}{\hat{L}_U}. \tag{4-60}$$

This function must be between 0 and 1. Both likelihoods are positive, and $\hat{L}_R$ cannot be larger than $\hat{L}_U$. (A restricted optimum is never superior to an unrestricted one.) If λ is too small, then doubt is cast on the restrictions.

An example from a discrete distribution helps to fix these ideas. In estimating from a sample of 10 from a Poisson distribution at the beginning of Section 4.5, we found the MLE of the parameter θ to be 2. At this value, the likelihood, which is the probability of observing the sample we did, is 0.104×10^{-8}. Are these data consistent with $H_0: \theta = 1.8$? $L_R = 0.936 \times 10^{-9}$, which is, as expected, smaller. This particular sample is somewhat less probable under the hypothesis.

The formal test procedure is based on the following result.

THEOREM 4.20: Distribution of the Likelihood Ratio Test Statistic. *Under regularity, the large sample distribution of $-2 \ln \lambda$ is chi-squared, with degrees of freedom equal to the number of restrictions imposed.*

The null hypothesis is rejected if this value exceeds the appropriate critical value from the chi-squared tables. Thus, for the Poisson example,

$$-2 \ln \lambda = -2 \ln\left(\frac{0.0936}{0.104}\right) = 0.21072.$$

This chi-squared statistic with one degree of freedom is not significant at any conventional level, so we would not reject the hypothesis that $\theta = 1.8$ on the basis of this test.[22]

It is tempting to use the likelihood ratio test to test a simple null hypothesis against a simple alternative. For example, we might be interested in the Poisson setting in testing $H_0: \theta = 1.8$ against $H_1: \theta = 2.2$. But the test cannot be used in this fashion. The degrees of freedom of the chi-squared statistic for the likelihood ratio test equals the reduction in the number of dimensions in the parameter space that results from imposing the restrictions. In testing a simple null hypothesis against a simple alternative, this value is zero.[23] Second, one sometimes encounters an attempt to test one distributional assumption against another with a likelihood ratio test; for example, a certain model will be estimated assuming a normal distribution and then assuming a t distribution. The ratio of the two likelihoods is then compared to determine which distribution is preferred. This comparison is also inappropriate. The parameter spaces, and hence the likelihood functions of the two cases, are unrelated.

4.9.3.b. The Wald Test

A practical shortcoming of the likelihood ratio test is that it usually requires estimation of both the restricted and unrestricted parameter vectors. In complex models, one or the other of these estimates may be very difficult to compute. Fortunately, there are two alternative testing procedures, the Wald test and the Lagrange multiplier test, that circumvent this problem. Both tests are based on an estimator that is asymptotically normally distributed.

These two tests are based on the distribution of the full rank quadratic form considered at the end of Section 3.10.5. Specifically,

$$\text{If } \mathbf{x} \sim N_J[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \text{ then } (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \text{chi-squared}[J]. \qquad \textbf{(4-61)}$$

In the setting of a hypothesis test, under the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$, the quadratic form has the chi-squared distribution. If the hypothesis that $E(\mathbf{x}) = \boldsymbol{\mu}$ is false, however, then the quadratic form just given will, on average, be larger than it would be if the hypothesis were true.[24] This condition forms the basis for the test statistics discussed in this and the next section.

Let $\hat{\boldsymbol{\theta}}$ be the vector of parameter estimates obtained without restrictions. We hypothesize a set of restrictions

$$H_0: \mathbf{c}(\boldsymbol{\theta}) = \mathbf{q}.$$

---

[22] Of course, our use of the large-sample result in a sample of 10 might be questionable.

[23] Note that because both likelihoods are restricted in this instance, there is nothing to prevent $-2 \ln \lambda$ from being negative.

[24] If the mean is not $\boldsymbol{\mu}$, then the statistic in (4-61) will have a **noncentral chi-squared distribution.** This distribution has the same basic shape as the central chi-squared distribution, with the same degrees of freedom, but lies to the right of it. Thus, a random draw from the noncentral distribution will tend, on average, to be larger than a random observation from the central distribution.