CHAPTER 2
Basic Statistical Concepts

2.1 Introduction

One chapter cannot possibly review what one learned in one or two pre-requisite courses in statistics. This is an econometrics book, and it is imperative that the student have taken at least one solid course in statistics. The concepts of a random variable, whether discrete or continuous, and the associated probability function or *probability density function* (p.d.f.) are assumed known. Similarly, the reader should know the following statistical terms: Cumulative distribution function, marginal, conditional and joint p.d.f.'s. The reader should be comfortable with computing mathematical expectations, and familiar with the concepts of independence, Bayes Theorem and several continuous and discrete probability distributions. These distributions include: the Bernoulli, Binomial, Poisson, Geometric, Uniform, Normal, Gamma, Chi-squared (χ^2), Exponential, Beta, t and F distributions.

Section 2.2 reviews two methods of estimation, while section 2.3 reviews the properties of the resulting estimators. Section 2.4 gives a brief review of test of hypotheses, while section 2.5 discusses the meaning of confidence intervals. These sections are fundamental background for this book, and the reader should make sure that he or she is familiar with these concepts. Also, be sure to solve the exercises at the end of this chapter.

2.2 Methods of Estimation

Consider a Normal distribution with mean μ and variance σ^2. This is the important "Gaussian" distribution which is symmetric and bell-shaped and completely determined by its measure of centrality, its mean μ and its measure of dispersion, its variance σ^2. μ and σ^2 are called the population parameters. Draw a random sample $X_1, \ldots, X_n$ independent and identically distributed (IID) from this population. We usually estimate μ by $\widehat{\mu} = \bar{X}$ and σ^2 by

$$s^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1).$$

For example, μ = mean income of a household in Houston. $\bar{X}$ = sample average of incomes of 100 households randomly interviewed in Houston.

This estimator of μ could have been obtained by either of the following two methods of estimation:

(i) Method of Moments

Simply stated, this method of estimation uses the following rule: Keep equating population moments to their sample counterpart until you have estimated all the population parameters.

B.H. Baltagi, *Econometrics*, Springer Texts in Business and Economics, DOI 10.1007/978-3-642-20059-5_2,

© Springer-Verlag Berlin Heidelberg 2011

Population	Sample
$E(X) = \mu$	$\sum_{i=1}^{n} X_i/n = \bar{X}$
$E(X^2) = \mu^2 + \sigma^2$	$\sum_{i=1}^{n} X_i^2/n$
$\vdots$	$\vdots$
$E(X^r)$	$\sum_{i=1}^{n} X_i^r/n$

The normal density is completely identified by μ and σ^2, hence only the first 2 equations are needed

$$\widehat{\mu} = \bar{X} \qquad \text{and} \qquad \widehat{\mu}^2 + \widehat{\sigma}^2 = \sum_{i=1}^{n} X_i^2/n$$

Substituting the first equation in the second one obtains

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} X_i^2/n - \bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$$

(ii) Maximum Likelihood Estimation (MLE)

For a random sample of size n from the Normal distribution $X_i \sim N(\mu, \sigma^2)$, we have

$$f_i(X_i; \mu, \sigma^2) = (1/\sigma\sqrt{2\pi}) \exp\left\{-(X_i - \mu)^2/2\sigma^2\right\} \quad -\infty < X_i < +\infty$$

Since $X_1, \ldots, X_n$ are independent and identically distributed, the joint probability density function is given as the product of the marginal probability density functions:

$$f(X_1, \ldots, X_n; \mu, \sigma^2) = \prod_{i=1}^{n} f_i(X_i; \mu, \sigma^2) = (1/2\pi\sigma^2)^{n/2} \exp\left\{-\sum_{i=1}^{n}(X_i - \mu)^2/2\sigma^2\right\} \quad (2.1)$$

Usually, we observe only one sample of n households which could have been generated by any pair of (μ, σ^2) with $-\infty < \mu < +\infty$ and $\sigma^2 > 0$. For each pair, say (μ_0, σ_0^2), $f(X_1, \ldots, X_n; \mu_0, \sigma_0^2)$ denotes the probability (or likelihood) of obtaining that sample. By varying (μ, σ^2) we get different probabilities of obtaining this sample. Intuitively, we choose the values of μ and σ^2 that maximize the probability of obtaining this sample. Mathematically, we treat $f(X_1, \ldots, X_n; \mu, \sigma^2)$ as $L(\mu, \sigma^2)$ and we call it the likelihood function. Maximizing $L(\mu, \sigma^2)$ with respect to μ and σ^2, one gets the first-order conditions of maximization:

$$(\partial L/\partial \mu) = 0 \qquad \text{and} \qquad (\partial L/\partial \sigma^2) = 0$$

Equivalently, we can maximize $\log L(\mu, \sigma^2)$ rather than $L(\mu, \sigma^2)$ and still get the same answer. Usually, the latter monotonic transformation of the likelihood is easier to maximize and the first-order conditions become

$$(\partial \log L/\partial \mu) = 0 \qquad \text{and} \qquad (\partial \log L/\partial \sigma^2) = 0$$

For the Normal distribution example, we get

$$\log L(\mu; \sigma^2) = -(n/2)\log \sigma^2 - (n/2)\log 2\pi - (1/2\sigma^2)\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\partial \log L(\mu; \sigma^2)/\partial \mu = (1/\sigma^2)\sum_{i=1}^{n}(X_i - \mu) = 0 \Rightarrow \widehat{\mu}_{MLE} = \bar{X}$$

$$\partial \log L(\mu; \sigma^2)/\partial \sigma^2 = -(n/2)(1/\sigma^2) + \sum_{i=1}^{n}(X_i - \mu)^2/2\sigma^4 = 0$$

$$\Rightarrow \widehat{\sigma}^2_{MLE} = \sum_{i=1}^{n}(X_i - \widehat{\mu}_{MLE})^2/n = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$$

Note that the moments estimators and the maximum likelihood estimators are the same for the Normal distribution example. In general, the two methods need not necessarily give the same estimators. Also, note that the moments estimators will always have the same estimating equations, for example, the first two equations are always

$$E(X) = \mu \equiv \sum_{i=1}^{n} X_i/n = \bar{X} \qquad \text{and} \qquad E(X^2) = \mu^2 + \sigma^2 \equiv \sum_{i=1}^{n} X_i^2/n.$$

For a specific distribution, we need only substitute the relationship between the population moments and the parameters of that distribution. Again, the number of equations needed depends upon the number of parameters of the underlying distribution. For e.g., the exponential distribution has one parameter and needs only one equation whereas the gamma distribution has two parameters and needs two equations. Finally, note that the maximum likelihood technique is heavily reliant on the form of the underlying distribution, but it has desirable properties when it exists. These properties will be discussed in the next section.

So far we have dealt with the Normal distribution to illustrate the two methods of estimation. We now apply these methods to the Bernoulli distribution and leave other distributions applications to the exercises. We urge the student to practice on these exercises.

Bernoulli Example: In various cases in real life the outcome of an event is binary, a worker may join the labor force or may not. A criminal may return to crime after parole or may not. A television off the assembly line may be defective or not. A coin tossed comes up head or tail, and so on. In this case $\theta = \Pr[\text{Head}]$ and $1 - \theta = \Pr[\text{Tail}]$ with $0 < \theta < 1$ and this can be represented by the discrete probability function

$$\begin{array}{lll} f(X;\theta) & = \theta^X(1-\theta)^{1-X} & X = 0, 1 \\ & = 0 & \text{elsewhere} \end{array}$$

The Normal distribution is a continuous distribution since it takes values for all X over the real line. The Bernoulli distribution is discrete, because it is defined only at integer values for X. Note that $P[X = 1] = f(1;\theta) = \theta$ and $P[X = 0] = f(0;\theta) = 1 - \theta$ for all values of $0 < \theta < 1$. A random sample of size n drawn from this distribution will have a joint probability function

$$L(\theta) = f(X_1, \ldots, X_n; \theta) = \theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n - \sum_{i=1}^{n} X_i}$$

with $X_i = 0, 1$ for $i = 1, \ldots, n$. Therefore,

$$\begin{array}{rcl} \log L(\theta) & = & (\sum_{i=1}^{n} X_i)\log\theta + (n - \sum_{i=1}^{n} X_i)\log(1-\theta) \\ \dfrac{\partial \log L(\theta)}{\partial \theta} & = & \dfrac{\sum_{i=1}^{n} X_i}{\theta} - \dfrac{(n - \sum_{i=1}^{n} X_i)}{(1-\theta)} \end{array}$$

Solving this first-order condition for θ, one gets

$$(\sum_{i=1}^{n} X_i)(1-\theta) - \theta(n - \sum_{i=1}^{n} X_i) = 0$$

which reduces to

$$\widehat{\theta}_{MLE} = \sum_{i=1}^{n} X_i/n = \bar{X}.$$

This is the frequency of heads in n tosses of a coin.

For the method of moments, we need

$$E(X) = \sum_{X=0}^{1} Xf(X,\theta) = 1.f(1,\theta) + 0.f(0,\theta) = f(1,\theta) = \theta$$

and this is equated to $\bar{X}$ to get $\widehat{\theta} = \bar{X}$. Once again, the MLE and the method of moments yield the same estimator. Note that only one parameter θ characterizes this Bernoulli distribution and one does not need to equate second or higher population moments to their sample values.

2.3 Properties of Estimators

(i) Unbiasedness

$\widehat{\mu}$ is said to be unbiased for μ if and only if $E(\widehat{\mu}) = \mu$

For $\widehat{\mu} = \bar{X}$, we have $E(\bar{X}) = \sum_{i=1}^{n} E(X_i)/n = \mu$ and $\bar{X}$ is unbiased for μ. No distributional assumption is needed as long as the X_i's are distributed with the same mean μ. Unbiasedness means that "on the average" our estimator is on target. Let us explain this last statement. If we repeat our drawing of a random sample of 100 households, say 200 times, then we get 200 $\bar{X}$'s. Some of these $\bar{X}$'s will be above μ some below μ, but their average should be very close to μ. Since in real life situations, we observe only one random sample, there is little consolation if our observed $\bar{X}$ is far from μ. But the larger n is the smaller is the dispersion of this $\bar{X}$, since $\text{var}(\bar{X}) = \sigma^2/n$ and the lesser is the likelihood of this $\bar{X}$ to be very far from μ. This leads us to the concept of efficiency.

(ii) Efficiency

For two unbiased estimators, we compare their efficiencies by the ratio of their variances. We say that the one with lower variance is more efficient. For example, taking $\widehat{\mu}_1 = X_1$ versus $\widehat{\mu}_2 = \bar{X}$, both estimators are unbiased but $\text{var}(\widehat{\mu}_1) = \sigma^2$ whereas, $\text{var}(\widehat{\mu}_2) = \sigma^2/n$ and {the relative efficiency of $\widehat{\mu}_1$ with respect to $\widehat{\mu}_2$} $= \text{var}(\widehat{\mu}_2)/\text{var}(\widehat{\mu}_1) = 1/n$, see Figure 2.1. To compare all unbiased estimators, we find the one with minimum variance. Such an estimator if it exists is called the MVU (minimum variance unbiased estimator). A lower bound for the variance of any unbiased estimator $\widehat{\mu}$ of μ, is known in the statistical literature as the Cramér-Rao lower bound, and is given by

$$\text{var}(\widehat{\mu}) \geq 1/n\{E(\partial \log f(X;\mu)/\partial\mu)\}^2 = -1/\{nE(\partial^2 \log f(X;\mu)/\partial\mu^2)\} \tag{2.2}$$

where we use either representation of the bound on the right hand side of (2.2) depending on which one is the simplest to derive.

Example 1: Consider the normal density

$$\log f(X_i;\mu) = (-1/2)\log\sigma^2 - (1/2)\log 2\pi - (1/2)(X_i-\mu)^2/\sigma^2$$

$$\partial \log f(X_i;\mu)/\partial\mu = (X_i-\mu)/\sigma^2$$

$$\partial^2 \log f(X_i;\mu)/\partial\mu^2 = -(1/\sigma^2)$$

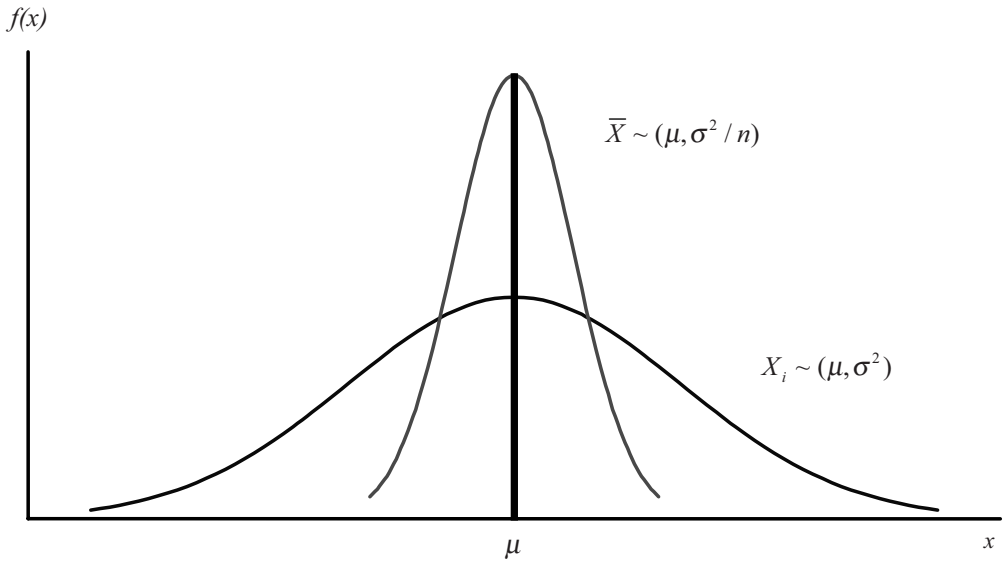


Figure 2.1 Efficiency Comparisons

with $E\{\partial^2 \log f(X_i;\mu)/\partial\mu^2\} = -(1/\sigma^2)$. Therefore, the variance of any unbiased estimator of μ, say $\widehat{\mu}$ satisfies the property that $\text{var}(\widehat{\mu}) \geq \sigma^2/n$.

Turning to σ^2; let $\theta = \sigma^2$, then

$$\log f(X_i;\theta) = -(1/2)\log\theta - (1/2)\log 2\pi - (1/2)(X_i-\mu)^2/\theta$$

$$\partial \log f(X_i;\theta)/\partial\theta = -1/2\theta + (X_i-\mu)^2/2\theta^2 = \{(X_i-\mu)^2 - \theta\}/2\theta^2$$

$$\partial^2 \log f(X_i;\theta)/\partial\theta^2 = 1/2\theta^2 - (X_i-\mu)^2/\theta^3 = \{\theta - 2(X_i-\mu)^2\}/2\theta^3$$

$E[\partial^2 \log f(X_i;\theta)/\partial\theta^2] = -(1/2\theta^2)$, since $\text{E}(X_i-\mu)^2 = \theta$. Hence, for any unbiased estimator of θ, say $\widehat{\theta}$, its variance satisfies the following property $\text{var}(\widehat{\theta}) \geq 2\theta^2/n$, or $\text{var}(\widehat{\sigma}^2) \geq 2\sigma^4/n$.

Note that, if one finds an unbiased estimator whose variance attains the Cramér-Rao lower bound, then this is the MVU estimator. It is important to remember that this is only a lower bound and sometimes it is not necessarily attained. If the X_i's are normal, $\bar{X} \sim N(\mu, \sigma^2/n)$. Hence, $\bar{X}$ is unbiased for μ with variance σ^2/n equal to the Cramér-Rao lower bound. Therefore, $\bar{X}$ is MVU for μ. On the other hand,

$$\widehat{\sigma}^2_{MLE} = \textstyle\sum_{i=1}^n (X_i - \bar{X})^2/n,$$

and it can be shown that $(n\widehat{\sigma}^2_{MLE})/(n-1) = s^2$ is unbiased for σ^2. In fact, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ and the expected value of a Chi-squared variable with $(n-1)$ degrees of freedom is exactly its degrees of freedom. Using this fact,

$$E\{(n-1)s^2/\sigma^2\} = E(\chi^2_{n-1}) = n-1.$$

Therefore, $E(s^2) = \sigma^2$.[1] Also, the variance of a Chi-squared variable with $(n-1)$ degrees of freedom is twice these degrees of freedom. Using this fact,

$$\text{var}\{(n-1)s^2/\sigma^2\} = \text{var}(\chi^2_{n-1}) = 2(n-1)$$

or

$$\{(n-1)^2/\sigma^4\}\text{var}(s^2) = 2(n-1).$$

Hence, the $\text{var}(s^2) = 2\sigma^4/(n-1)$ and this does not attain the Cramér-Rao lower bound. In fact, it is larger than $(2\sigma^4/n)$. Note also that $\text{var}(\widehat{\sigma}^2_{MLE}) = \{(n-1)^2/n^2\}\text{var}(s^2) = \{2(n-1)\}\sigma^4/n^2$. This is smaller than $(2\sigma^4/n)$! How can that be? Remember that $\widehat{\sigma}^2_{MLE}$ is a biased estimator of σ^2 and hence, $\text{var}(\widehat{\sigma}^2_{MLE})$ should not be compared with the Cramér-Rao lower bound. This lower bound pertains only to unbiased estimators.

Warning: Attaining the Cramér-Rao lower bound is only a sufficient condition for efficiency. Failing to satisfy this condition does not necessarily imply that the estimator is not efficient.

Example 2: For the Bernoulli case

$$\log f(X_i;\theta) = X_i \log\theta + (1-X_i)\log(1-\theta)$$

$$\partial \log f(X_i,\theta)/\partial\theta = (X_i/\theta) - (1-X_i)/(1-\theta)$$

$$\partial^2 \log f(X_i;\theta)/\partial\theta^2 = (-X_i/\theta^2) - (1-X_i)/(1-\theta)^2$$

and $E[\partial^2 \log f(X_i;\theta)/\partial\theta^2] = (-1/\theta) - 1/(1-\theta) = -1/[\theta(1-\theta)]$. Therefore, for any unbiased estimator of θ, say $\widehat{\theta}$, its variance satisfies the following property:

$$\text{var}(\widehat{\theta}) \geq \theta(1-\theta)/n.$$

For the Bernoulli random sample, we proved that $\mu = E(X_i) = \theta$. Similarly, it can be easily verified that $\sigma^2 = \text{var}(X_i) = \theta(1-\theta)$. Hence, $\bar{X}$ has mean $\mu = \theta$ and $\text{var}(\bar{X}) = \sigma^2/n = \theta(1-\theta)/n$. This means that $\bar{X}$ is unbiased for θ and it attains the Cramér-Rao lower bound. Therefore, $\bar{X}$ is MVU for θ.

Unbiasedness and efficiency are finite sample properties (in other words, true for any finite sample size n). Once we let n tend to ∞ then we are in the realm of *asymptotic properties*.

Example 3: For a random sample from any distribution with mean μ it is clear that $\widetilde{\mu} = (\bar{X} + 1/n)$ is not an unbiased estimator of μ since $E(\widetilde{\mu}) = E(\bar{X} + 1/n) = \mu + 1/n$. However, as $n \to \infty$ the lim $E(\widetilde{\mu})$ is equal to μ. We say, that $\widetilde{\mu}$ is *asymptotically unbiased* for μ.

Example 4: For the Normal case

$$\widehat{\sigma}^2_{MLE} = (n-1)s^2/n \quad \text{and} \quad E(\widehat{\sigma}^2_{MLE}) = (n-1)\sigma^2/n.$$

But as $n \to \infty$, lim $E(\widehat{\sigma}^2_{MLE}) = \sigma^2$. Hence, $\widehat{\sigma}^2_{MLE}$ is *asymptotically unbiased* for σ^2.

Similarly, an estimator which attains the Cramér-Rao lower bound in the limit is *asymptotically efficient*. Note that $\text{var}(\bar{X}) = \sigma^2/n$, and this tends to zero as $n \to \infty$. Hence, we consider$\sqrt{n}\bar{X}$ which has finite variance since $\text{var}(\sqrt{n}\bar{X}) = n\ \text{var}(\bar{X}) = \sigma^2$. We say that the asymptotic variance of $\bar{X}$ denoted by $\text{asymp.var}(\bar{X}) = \sigma^2/n$ and that it attains the Cramér-Rao lower bound in the limit. $\bar{X}$ is therefore asymptotically efficient. Similarly,

$$\text{var}(\sqrt{n}\widehat{\sigma}^2_{MLE}) = n\ \text{var}(\widehat{\sigma}^2_{MLE}) = 2(n-1)\sigma^4/n$$

which tends to $2\sigma^4$ as $n \to \infty$. This means that $\text{asymp.var}(\widehat{\sigma}^2_{MLE}) = 2\sigma^4/n$ and that it attains the Cramér-Rao lower bound in the limit. Therefore, $\widehat{\sigma}^2_{MLE}$ is asymptotically efficient.

(iii) Consistency

Another asymptotic property is consistency. This says that as $n \to \infty$ $\lim \Pr[|\bar{X} - \mu| > c] = 0$ for any arbitrary positive constant c. In other words, $\bar{X}$ will not differ from μ as $n \to \infty$.

Proving this property uses the Chebyshev's inequality which states in this context that

$$\Pr[|\bar{X} - \mu| > k\sigma_{\bar{X}}] \leq 1/k^2.$$

If we let $c = k\sigma_{\bar{X}}$ then $1/k^2 = \sigma^2_{\bar{X}}/c^2 = \sigma^2/nc^2$ and this tends to 0 as $n \to \infty$, since σ^2 and c are finite positive constants. A sufficient condition for an estimator to be consistent is that it is asymptotically unbiased and that its variance tends to zero as $n \to \infty$.[2]

Example 1: For a random sample from *any* distribution with mean μ and variance σ^2, $E(\bar{X}) = \mu$ and $\text{var}(\bar{X}) = \sigma^2/n \to 0$ as $n \to \infty$, hence $\bar{X}$ is consistent for μ.

Example 2: For the Normal case, we have shown that $E(s^2) = \sigma^2$ and $\text{var}(s^2) = 2\sigma^4/(n-1) \to 0$ as $n \to \infty$, hence s^2 is consistent for σ^2.

Example 3: For the Bernoulli case, we know that $E(\bar{X}) = \theta$ and $\text{var}(\bar{X}) = \theta(1-\theta)/n \to 0$ as $n \to \infty$, hence $\bar{X}$ is consistent for θ.

Warning: This is only a sufficient condition for consistency. Failing to satisfy this condition does not necessarily imply that the estimator is inconsistent.

(iv) Sufficiency

$\bar{X}$ is sufficient for μ, if $\bar{X}$ contains all the information in the sample pertaining to μ. In other words, $f(X_1, \ldots, X_n/\bar{X})$ is independent of μ. To prove this fact one uses the factorization theorem due to Fisher and Neyman. In this context, $\bar{X}$ is sufficient for μ, if and only if one can factorize the joint p.d.f.

$$f(X_1, \ldots, X_n; \mu) = h(\bar{X}; \mu) \cdot g(X_1, \ldots, X_n)$$

where h and g are any two functions with the latter being only a function of the X's and independent of μ in form and in the domain of the X's.

Example 1: For the Normal case, it is clear from equation (2.1) that by subtracting and adding $\bar{X}$ in the summation we can write after some algebra

$$f(X_1, .., X_n; \mu, \sigma^2) = (1/2\pi\sigma^2)^{n/2} e^{-\{(1/2\sigma^2)\sum_{i=1}^{n}(X_i - \bar{X})^2\}} e^{-\{(n/2\sigma^2)(\bar{X}-\mu)^2\}}$$

Hence, $h(\bar{X}; \mu) = e^{-(n/2\sigma^2)(\bar{X}-\mu^2)}$ and $g(X_1, \ldots, X_n)$ is the remainder term which is independent of μ in form. Also $-\infty < X_i < \infty$ and hence independent of μ in the domain. Therefore, $\bar{X}$ is sufficient for μ.

Example 2: For the Bernoulli case,

$$f(X_1, \ldots, X_n; \theta) = \theta^{n\bar{X}}(1-\theta)^{n(1-\bar{X})} \quad X_i = 0, 1 \quad \text{for } i = 1, \ldots, n.$$

Therefore, $h(\bar{X}, \theta) = \theta^{n\bar{X}}(1-\theta)^{n(1-\bar{X})}$ and $g(X_1, \ldots, X_n) = 1$ which is independent of θ in form and domain. Hence, $\bar{X}$ is sufficient for θ.

Under certain regularity conditions on the distributions we are sampling from, one can show that the MVU of any parameter θ is an unbiased function of a sufficient statistic for θ.[3] Advantages of the maximum likelihood estimators is that (i) they are sufficient estimators when they exist. (ii) They are asymptotically efficient. (iii) If the distribution of the MLE satisfies certain regularity conditions, then making the MLE unbiased results in a unique MVU estimator. A prime example of this is s^2 which was shown to be an unbiased estimator of σ^2 for a random sample drawn from the Normal distribution. It can be shown that s^2 is sufficient for σ^2 and that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. Hence, s^2 is an unbiased sufficient statistic for σ^2 and therefore it is MVU for σ^2, even though it does not attain the Cramér-Rao lower bound. (iv) Maximum likelihood estimates are invariant with respect to continuous transformations. To explain the last property, consider the estimator of e^{μ}. Given $\widehat{\mu}_{MLE} = \bar{X}$, an obvious estimator is $e^{\widehat{\mu}_{MLE}} = e^{\bar{X}}$. This is in fact the MLE of e^{μ}. In general, if $g(\mu)$ is a continuous function of μ, then $g(\widehat{\mu}_{MLE})$ is the MLE of $g(\mu)$. Note that $E(e^{\widehat{\mu}_{MLE}}) \neq e^{E(\widehat{\mu}_{MLE})} = e^{\mu}$, in other words, expectations are not invariant to all continuous transformations, especially nonlinear ones and hence the resulting MLE estimator may not be unbiased. $e^{\bar{X}}$ is not unbiased for e^{μ} even though $\bar{X}$ is unbiased for μ.

In summary, there are two routes for finding the MVU estimator. One is systematically following the derivation of a sufficient statistic, proving that its distribution satisfies certain regularity conditions, and then making it unbiased for the parameter in question. Of course, MLE provides us with sufficient statistics, for example,

$$X_1, \ldots, X_n \sim \text{IIN}(\mu, \sigma^2) \Rightarrow \widehat{\mu}_{MLE} = \bar{X} \quad \text{and} \quad \widehat{\sigma}^2_{MLE} = \textstyle\sum_{i=1}^{n}(X_i - \bar{X})^2/n$$

are both sufficient for μ and σ^2, respectively. $\bar{X}$ is unbiased for μ and $\bar{X} \sim N(\mu, \sigma^2/n)$. The Normal distribution satisfies the regularity conditions needed for $\bar{X}$ to be MVU for μ. $\widehat{\sigma}^2_{MLE}$ is biased for σ^2, but $s^2 = n\widehat{\sigma}^2_{MLE}/(n-1)$ is unbiased for σ^2 and $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ which also satisfies the regularity conditions for s^2 to be a MVU estimator for σ^2.

Alternatively, one finds the Cramér-Rao lower bound and checks whether the usual estimator (obtained from say the method of moments or the maximum likelihood method) achieves this lower bound. If it does, this estimator is efficient, and there is no need to search further. If it does not, the former strategy leads us to the MVU estimator. In fact, in the previous example $\bar{X}$ attains the Cramér-Rao lower bound, whereas s^2 does not. However, both are MVU for μ and σ^2 respectively.

(v) Comparing Biased and Unbiased Estimators

Suppose we are given two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of θ where the first is unbiased and has a large variance and the second is biased but with a small variance. The question is which one of these two estimators is preferable? $\widehat{\theta}_1$ is unbiased whereas $\widehat{\theta}_2$ is biased. This means that if we repeat the sampling procedure many times then we expect $\widehat{\theta}_1$ to be on the average correct, whereas $\widehat{\theta}_2$ would be on the average different from θ. However, in real life, we observe only one sample. With a large variance for $\widehat{\theta}_1$, there is a great likelihood that the sample drawn could result in a $\widehat{\theta}_1$ far away from θ. However, with a small variance for $\widehat{\theta}_2$, there is a better chance of getting a $\widehat{\theta}_2$ close to θ. If our loss function is $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$ then our risk is

$$\begin{aligned} R(\widehat{\theta}, \theta) &= E[L(\widehat{\theta}, \theta)] = E(\widehat{\theta} - \theta)^2 = MSE(\widehat{\theta}) \\ &= E[\widehat{\theta} - E(\widehat{\theta}) + E(\widehat{\theta}) - \theta]^2 = \text{var}(\widehat{\theta}) + (\text{Bias}(\widehat{\theta}))^2. \end{aligned}$$

Minimizing the risk when the loss function is quadratic is equivalent to minimizing the Mean Square Error (MSE). From its definition the MSE shows the trade-off between bias and variance. MVU theory, sets the bias equal to zero and minimizes var$(\widehat{\theta})$. In other words, it minimizes the above risk function but only over $\widehat{\theta}$'s that are unbiased. If we do not restrict ourselves to unbiased estimators of θ, minimizing MSE may result in a biased estimator such as $\widehat{\theta}_2$ which beats $\widehat{\theta}_1$ because the gain from its smaller variance outweighs the loss from its small bias, see Figure 2.2.

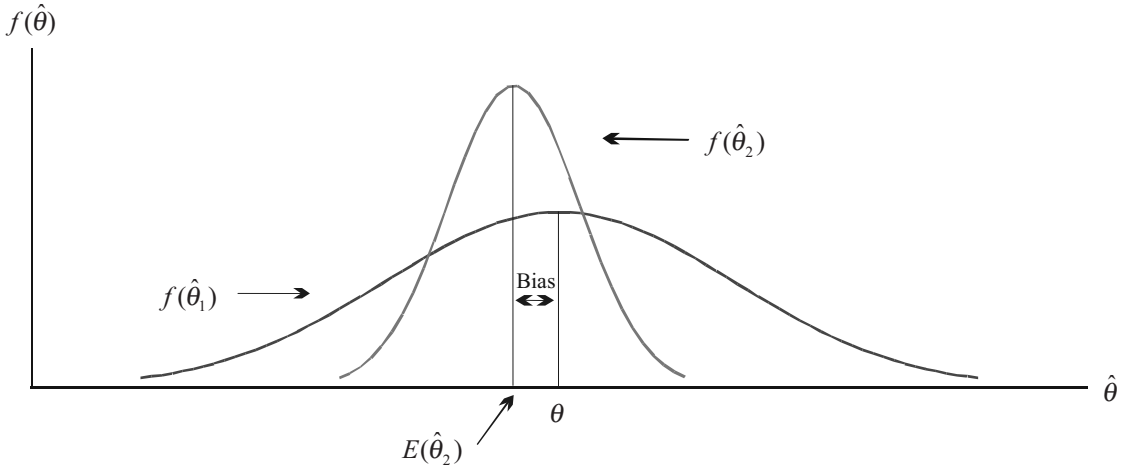


Figure 2.2 Bias Versus Variance

2.4 Hypothesis Testing

The best way to proceed is with an example.

Example 1: The Economics Departments instituted a new program to teach micro-principles. We would like to test the null hypothesis that 80% of economics undergraduate students will pass the micro-principles course versus the alternative hypothesis that only 50% will pass. We draw a random sample of size 20 from the large undergraduate micro-principles class and as a simple rule we accept the null if x, the number of passing students is larger or equal to 13, otherwise the alternative hypothesis will be accepted. Note that the distribution we are drawing from is Bernoulli with the probability of success θ, and we have chosen only two states of the world H_0; $\theta_0 = 0.80$ and $H_1; \theta_1 = 0.5$. This situation is known as testing a *simple* hypothesis versus another *simple* hypothesis because the distribution is completely specified under the null or alternative hypothesis. One would expect $(E(x) = n\theta_0)$ 16 students under H_0 and $(n\theta_1)$ 10 students under H_1 to pass the micro-principles exams. It seems then logical to take $x \geq 13$ as the cut-off point distinguishing H_0 from H_1. No theoretical justification is given at this stage to this arbitrary choice except to say that it is the mid-point of $[10, 16]$. Figure 2.3 shows that one can make two types of errors. The first is rejecting H_0 when in fact it is true, this is known as *type I error* and the probability of committing this error is denoted by α. The second is accepting H_0 when it is false. This is known as *type II error* and the corresponding probability is denoted by β. For this example

$$\begin{aligned}\alpha &= \Pr[\text{rejecting } H_0/H_0 \text{ is true}] = \Pr[x < 13/\theta = 0.8] \\ &= b(n=20; x=0; \theta=0.8) + .. + b(n=20; x=12; \theta=0.8) \\ &= b(n=20; x=20; \theta=0.2) + .. + b(n=20; x=8; \theta=0.2) \\ &= 0 + .. + 0 + 0.0001 + 0.0005 + 0.0020 + 0.0074 + 0.0222 = 0.0322\end{aligned}$$

where we have used the fact that $b(n; x; \theta) = b(n; n-x; 1-\theta)$ and $b(n; x; \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ denotes the binomial distribution for $x = 0, 1, \ldots, n$, see problem 4.

		True World	
		$\theta_0 = 0.80$	$\theta_1 = 0.50$
Decision	θ_0	No error	Type II error
	θ_1	Type I error	No Error

Figure 2.3 Type I and II Error

$$\begin{aligned}\beta &= \Pr[\text{accepting } H_0/H_0 \text{ is false}] = \Pr[x \geq 13/\theta = 0.5] \\ &= b(n=20; x=13; \theta=0.5) + .. + b(n=20; x=20; \theta=0.5) \\ &= 0.0739 + 0.0370 + 0.0148 + 0.0046 + 0.0011 + 0.0002 + 0 + 0 = 0.1316\end{aligned}$$

The rejection region for $H_0, x < 13$, is known as the *critical region* of the test and $\alpha = \Pr[\text{Falling in the critical region}/H_0 \text{ is true}]$ is also known as the *size* of the critical region. A good test is one which minimizes both types of errors α and β. For the above example, α is low but β is high with more than a 13% chance of happening. This β can be reduced by changing the critical region from $x < 13$ to $x < 14$, so that H_0 is accepted only if $x \geq 14$. In this case, one can easily verify that

$$\begin{aligned}\alpha &= \Pr[x < 14/\theta = 0.8] = b(n=20;\ x=0; \theta=0.8) + .. + b(n=20, x=13, \theta=0.8) \\ &= 0.0322 + b(n=20; x=13;\ \theta=0.8) = 0.0322 + 0.0545 = 0.0867\end{aligned}$$

and

$$\begin{aligned}\beta &= \Pr[x \geq 14/\theta = 0.5] = b(n=20;\ x=14; \theta=0.5) + .. + b(n=20; x=20; \theta=0.5) \\ &= 0.1316 - b(n=20; x=13;\ \theta=0.5) = 0.0577\end{aligned}$$

By becoming more conservative on accepting H_0 and more liberal on accepting H_1, one reduces β from 0.1316 to 0.0577 but the price paid is the increase in α from 0.0322 to 0.0867. The only way to reduce both α and β is by increasing n. For a fixed n, there is a tradeoff between α and β as we change the critical region. To understand this clearly, consider the real life situation of trial by jury for which the defendant can be innocent or guilty. The decision of incarceration or release implies two types of errors. One can make $\alpha = \Pr[\text{incarcerating/innocence}] = 0$ and $\beta =$ its maximum, by releasing *every* defendant. Or one can make $\beta = \Pr[\text{release/guilty}] = 0$ and $\alpha =$ its maximum, by incarcerating *every* defendant. These are extreme cases but hopefully they demonstrate the trade-off between α and β.

The Neyman-Pearson Theory

The classical theory of hypothesis testing, known as the Neyman-Pearson theory, fixes $\alpha = \Pr(\text{type I error}) \leq$ a constant and minimizes β or maximizes $(1-\beta)$. The latter is known as the *Power* of the test under the alternative.

The Neyman-Pearson Lemma: If C is a critical region of size α and k is a constant such that

$$(L_0/L_1) \leq k \text{ inside } C$$

and

$$(L_0/L_1) \geq k \text{ outside } C$$

then C is a most powerful critical region of size α for testing $H_0; \theta = \theta_0$, against $H_1; \theta = \theta_1$.

Note that the likelihood has to be completely specified under the null and alternative. Hence, this lemma applies only to testing a simple versus another simple hypothesis. The proof of this lemma is given in Freund (1992). Intuitively, L_0 is the likelihood function under the null H_0 and L_1 is the corresponding likelihood function under H_1. Therefore, (L_0/L_1) should be small for points inside the critical region C and large for points outside the critical region C. The proof of the theorem shows that any other critical region, say D, of size α cannot have a smaller probability of type II error than C. Therefore, C is the best or most powerful critical region of size α. Its power $(1-\beta)$ is maximum at H_1. Let us demonstrate this lemma with an example.

Example 2: Given a random sample of size n from $N(\mu, \sigma^2 = 4)$, use the Neyman-Pearson lemma to find the most powerful critical region of size $\alpha = 0.05$ for testing $H_0; \mu_0 = 2$ against the alternative $H_1; \mu_1 = 4$.

Note that this is a simple versus simple hypothesis as required by the lemma, since $\sigma^2 = 4$ is known and μ is specified by H_0 and H_1. The likelihood function for the $N(\mu, 4)$ density is given by

$$L(\mu) = f(x_1, \ldots, x_n; \mu, 4) = (1/2\sqrt{2\pi})^n \exp\left\{-\sum_{i=1}^n (x_i - \mu)^2/8\right\}$$

so that

$$L_0 = L(\mu_0) = (1/2\sqrt{2\pi})^n \exp\left\{-\sum_{i=1}^n (x_i - 2)^2/8\right\}$$

and

$$L_1 = L(\mu_1) = (1/2\sqrt{2\pi})^n \exp\left\{-\sum_{i=1}^n (x_i - 4)^2/8\right\}$$

Therefore

$$L_0/L_1 = \exp\left\{-\left[\sum_{i=1}^n (x_i - 2)^2 - \sum_{i=1}^n (x_i - 4)^2\right]/8\right\} = \exp\left\{-\sum_{i=1}^n x_i/2 + 3n/2\right\}$$

and the critical region is defined by

$$\exp\left\{-\sum_{i=1}^n x_i/2 + 3n/2\right\} \leq k \quad \text{inside } C$$

Taking logarithms of both sides, subtracting $(3/2)n$ and dividing by $(-1/2)n$ one gets

$$\bar{x} \geq K \quad \text{inside } C$$

In practice, one need not keep track of K as long as one keeps track of the direction of the inequality. K can be determined by making the size of $C = \alpha = 0.05$. In this case

$$\alpha = \Pr[\bar{x} \geq K/\mu = 2] = \Pr[z \geq (K-2)/(2/\sqrt{n})]$$

where $z = (\bar{x} - 2)/(2/\sqrt{n})$ is distributed $N(0,1)$ under H_0. From the $N(0,1)$ tables, we have

$$\frac{K-2}{(2/\sqrt{n})} = 1.645$$

Hence,

$$K = 2 + 1.645(2/\sqrt{n})$$

and $\bar{x} \geq 2 + 1.645(2/\sqrt{n})$ defines the most powerful critical region of size $\alpha = 0.05$ for testing $H_0; \mu_0 = 2$ versus $H_1; \mu_1 = 4$. Note that, in this case

$$\begin{aligned} \beta &= \Pr[\bar{x} < 2 + 1.645(2/\sqrt{n})/\mu = 4] \\ &= \Pr[z < [-2 + 1.645(2/\sqrt{n})]/(2/\sqrt{n})] = \Pr[z < 1.645 - \sqrt{n}] \end{aligned}$$

For $n = 4$; $\beta = \Pr[z < -0.355] = 0.3613$ shown by the shaded region in Figure 2.4. For $n = 9$; $\beta = \Pr[z < -1.355] = 0.0877$, and for $n = 16$; $\beta = \Pr[z < -2.355] = 0.00925$.

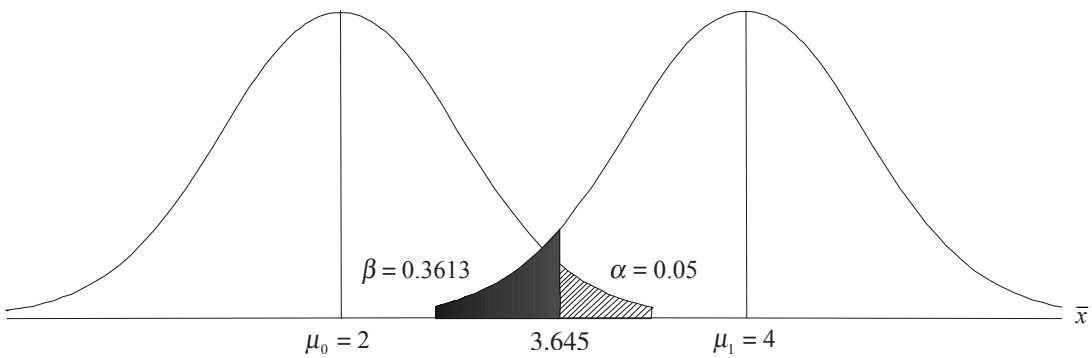


Figure 2.4 Critical Region for Testing $\mu_0 = 2$ against $\mu_1 = 4$ for $n = 4$

This gives us an idea of how, for a fixed $\alpha = 0.05$, the minimum β decreases with larger sample size n. As n increases from 4 to 9 to 16, the $\text{var}(\bar{x}) = \sigma^2/n$ decreases and the two distributions shown in Figure 2.4 shrink in dispersion still centered around $\mu_0 = 2$ and $\mu_1 = 4$, respectively. This allows better decision making (based on larger sample size) as reflected by the critical region shrinking from $\overline{x} \geq 3.65$ for $n = 4$ to $\overline{x} \geq 2.8225$ for $n = 16$, and the power $(1 - \beta)$ rising from 0.6387 to 0.9908, respectively, for a fixed $\alpha \leq 0.05$. The power function is the probability of rejecting H_0. It is equal to α under H_0 and $1 - \beta$ under H_1. The ideal power function is zero at H_0 and one at H_1. The Neyman-Pearson lemma allows us to fix α, say at 0.05, and find the test with the best power at H_1.

In example 2, both the null and alternative hypotheses are simple. In real life, one is more likely to be faced with testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. Under the alternative hypothesis, the distribution is not completely specified, since the mean μ is not known, and this is referred to as a *composite* hypothesis. In this case, one cannot compute the probability of type II error

since the distribution is not known under the alternative. Also, the Neyman-Pearson lemma cannot be applied. However, a simple generalization allows us to compute a Likelihood Ratio test which has satisfactory properties but is no longer uniformly most powerful of size α. In this case, one replaces L_1, which is not known since H_1 is a composite hypothesis, by the maximum value of the likelihood, i.e.,

$$\lambda = \frac{\max L_0}{\max L}$$

Since max L_0 is the maximum value of the likelihood under the null while max$L$ is the maximum value of the likelihood over the whole parameter space, it follows that max$L_0 \leq$ max$L$ and $\lambda \leq 1$. Hence, if H_0 is true, λ is close to 1, otherwise it is smaller than 1. Therefore, $\lambda \leq k$ defines the critical region for the Likelihood Ratio test, and k is determined such that the size of this test is α.

Example 3: For a random sample $x_1, \ldots, x_n$ drawn from a Normal distribution with mean μ and variance $\sigma^2 = 4$, derive the Likelihood Ratio test for $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. In this case

$$\max L_0 = (1/2\sqrt{2\pi})^n \exp\left\{-\textstyle\sum_{i=1}^n (x_i - 2)^2/8\right\} = L_0$$

and

$$\max L = (1/2\sqrt{2\pi})^n \exp\left\{-\textstyle\sum_{i=1}^n (x_i - \bar{x})^2/8\right\} = L(\widehat{\mu}_{MLE})$$

where use is made of the fact that $\widehat{\mu}_{MLE} = \bar{x}$. Therefore,

$$\lambda = \exp\left\{\left[-\textstyle\sum_{i=1}^n (x_i - 2)^2 + \sum_{i=1}^n (x_i - \bar{x})^2\right]/8\right\} = \exp\left\{-n(\bar{x} - 2)^2/8\right\}$$

Hence, the region for which $\lambda \leq k$, is equivalent after some simple algebra to the following region

$$(\bar{x} - 2)^2 \geq K \quad \text{or} \quad |\bar{x} - 2| \geq K^{1/2}$$

where K is determined such that

$$\Pr[|\bar{x} - 2| \geq K^{1/2}/\mu = 2] = \alpha$$

We know that $\bar{x} \sim N(2, 4/n)$ under H_0. Hence, $z = (\bar{x} - 2)/(2/\sqrt{n})$ is $N(0, 1)$ under H_0, and the critical region of size α will be based upon $|z| \geq z_{\alpha/2}$ where $z_{\alpha/2}$ is given in Figure 2.5 and is the value of a $N(0, 1)$ random variable such that the probability of exceeding it is $\alpha/2$. For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, and for $\alpha = 0.10$, $z_{\alpha/2} = 1.645$. This is a two-tailed test with rejection of H_0 obtained in case $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

Note that in this case

$$LR \equiv -2\log\lambda = (\bar{x} - 2)^2/(4/n) = z^2$$

which is distributed as χ_1^2 under H_0. This is because it is the square of a $N(0, 1)$ random variable under H_0. This is a finite sample result holding for any n. In general, other examples may lead to more complicated λ statistics for which it is difficult to find the corresponding distributions and hence the corresponding critical values. For these cases, we have an asymptotic result

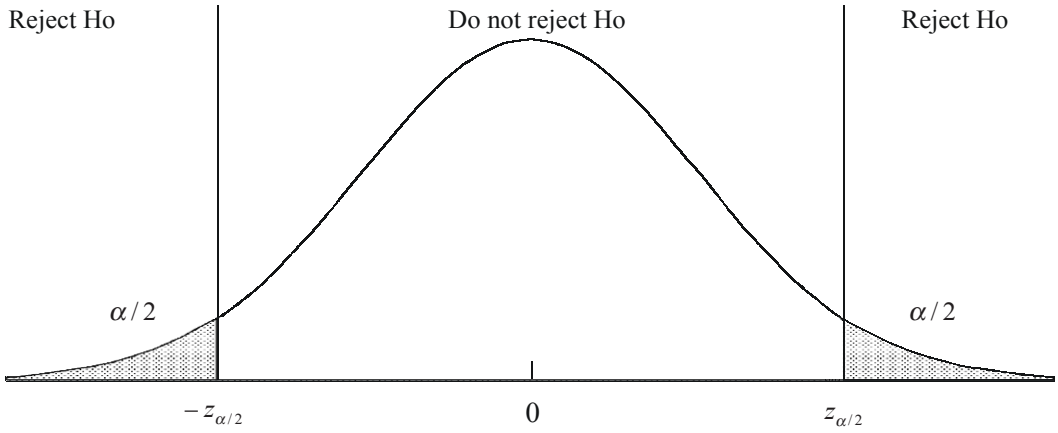


Figure 2.5 Critical Values

which states that, for large n, $LR = -2\log\lambda$ will be asymptotically distributed as χ^2_ν where ν denotes the number of restrictions that are tested by H_0. For example 2, $\nu = 1$ and hence, LR is asymptotically distributed as χ^2_1. Note that we did not need this result as we found LR is exactly distributed as χ^2_1 for any n. If one is testing H_0; $\mu = 2$ and $\sigma^2 = 4$ against the alternative that H_1; $\mu \neq 2$ or $\sigma^2 \neq 4$, then the corresponding LR will be asymptotically distributed as χ^2_2, see problem 5, part (f).

Likelihood Ratio, Wald and Lagrange Multiplier Tests

Before we go into the derivations of these three tests we start by giving an intuitive graphical explanation that will hopefully emphasize the differences among these tests. This intuitive explanation is based on the article by Buse (1982).

Consider a quadratic log-likelihood function in a parameter of interest, say μ. Figure 2.6 shows this log-likelihood $\log L(\mu)$, with a maximum at $\widehat{\mu}$. The Likelihood Ratio test, tests the null hypothesis $H_0; \mu = \mu_0$ by looking at the ratio of the likelihoods $\lambda = L(\mu_0)/L(\widehat{\mu})$ where

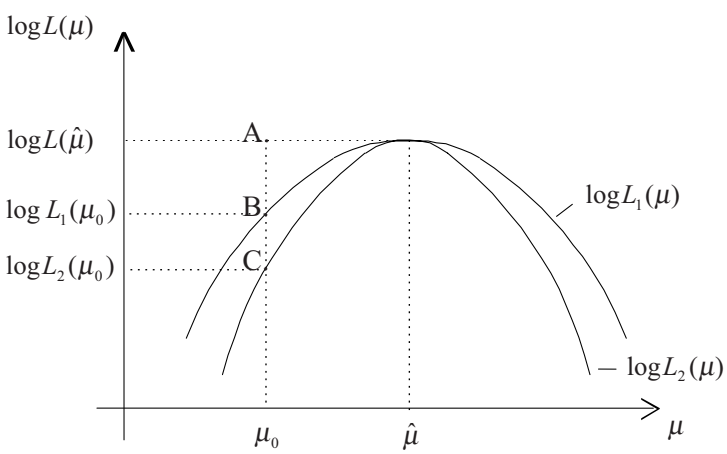


Figure 2.6 Wald Test

$-2\log\lambda$, twice the difference in log-likelihood, is distributed asymptotically as χ^2_1 under H_0. This test differentiates between the top of the hill and a preassigned point on the hill by evaluating the height at both points. Therefore, it needs both the restricted and unrestricted maximum of the likelihood. This ratio is dependent on the distance of μ_0 from $\widehat{\mu}$ and the curvature of the log-likelihood, $C(\mu) = |\partial^2 \log L(\mu)/\partial\mu^2|$, at $\widehat{\mu}$. In fact, for a fixed $(\widehat{\mu} - \mu_0)$, the larger $C(\widehat{\mu})$, the larger is the difference between the two heights. Also, for a given curvature at $\widehat{\mu}$, the larger $(\widehat{\mu} - \mu_0)$ the larger is the difference between the heights. The Wald test works from the top of the hill, i.e., it needs only the unrestricted maximum likelihood. It tries to establish the distance to μ_0, by looking at the horizontal distance $(\widehat{\mu} - \mu_0)$, and the curvature at $\widehat{\mu}$. In fact the Wald statistic is $W = (\widehat{\mu} - \mu_0)^2 C(\widehat{\mu})$ and this is asymptotically distributed as χ^2_1 under H_0. The usual form of W has $I(\mu) = -E[\partial^2 \log L(\mu)/\partial\mu^2]$ the information matrix evaluated at $\widehat{\mu}$, rather than $C(\widehat{\mu})$, but the latter is a consistent estimator of $I(\mu)$. The information matrix will be studied in details in Chapter 7. It will be shown, under fairly general conditions, that $\widehat{\mu}$ the MLE of μ, has $\text{var}(\widehat{\mu}) = I^{-1}(\widehat{\mu})$. Hence $W = (\widehat{\mu} - \mu_0)^2/\text{var}(\widehat{\mu})$ all evaluated at the unrestricted MLE. The Lagrange-Multiplier test (LM), on the other hand, goes to the preassigned point μ_0, i.e., it only needs the restricted maximum likelihood, and tries to determine how far it is from the top of the hill by considering the slope of the tangent to the likelihood $S(\mu) = \partial \log L(\mu)/\partial\mu$ at μ_0, and the rate at which this slope is changing, i.e., the curvature at μ_0. As Figure 2.7 shows, for two log-likelihoods with the same $S(\mu_0)$, the one that is closer to the top of the hill is the one with the larger curvature at μ_0.

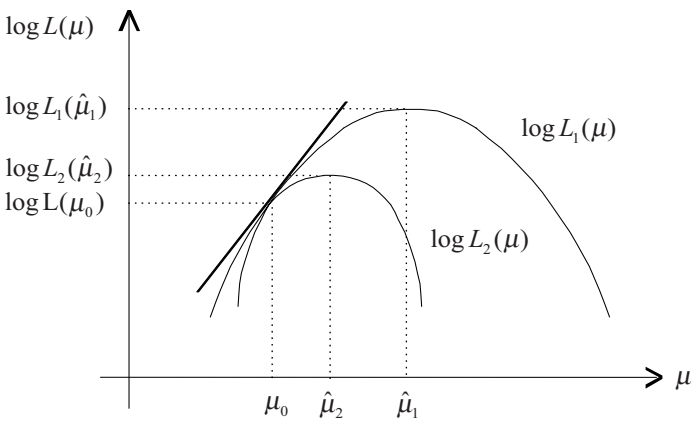


Figure 2.7 LM Test

This suggests the following statistic: $LM = S^2(\mu_0)\{C(\mu_0)\}^{-1}$ where the curvature appears in inverse form. In the Appendix to this chapter, we show that the $E[S(\mu)] = 0$ and $\text{var}[S(\mu)] = I(\mu)$. Hence $LM = S^2(\mu_0)I^{-1}(\mu_0) = S^2(\mu_0)/\text{var}[S(\mu_0)]$ all evaluated at the restricted MLE. Another interpretation of the LM test is that it is a measure of failure of the restricted estimator, in this case μ_0, to satisfy the first-order conditions of maximization of the unrestricted likelihood. We know that $S(\widehat{\mu}) = 0$. The question is: to what extent does $S(\mu_0)$ differ from zero? $S(\mu)$ is known in the statistics literature as the *score*, and the LM test is also referred to as the score test. For a more formal treatment of these tests, let us reconsider example 3 of a random sample $x_1, \ldots, x_n$ from a $N(\mu, 4)$ where we are interested in testing $H_0; \mu_0 = 2$ versus $H_1; \mu \neq 2$. The likelihood function $L(\mu)$ as well as $LR = -2\log\lambda = n(\bar{x} - 2)^2/4$ were given in example 3. In

fact, the score function is given by

$$S(\mu) = \frac{\partial \log L(\mu)}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i - \mu)}{4} = \frac{n(\bar{x} - \mu)}{4}$$

and under H_0

$$\begin{aligned} S(\mu_0) &= S(2) = \frac{n(\bar{x} - 2)}{4} \\ C(\mu) &= |\frac{\partial^2 \log L(\mu)}{\partial \mu^2}| = |-\frac{n}{4}| = \frac{n}{4} \end{aligned}$$

and $I(\mu) = -E\left[\dfrac{\partial^2 \log L(\mu)}{\partial \mu^2}\right] = \dfrac{n}{4} = C(\mu)$.

The Wald statistic is based on

$$W = (\widehat{\mu}_{MLE} - 2)^2 I(\widehat{\mu}_{MLE}) = (\bar{x} - 2)^2 \cdot \left(\frac{n}{4}\right)$$

The LM statistic is based on

$$LM = S^2(\mu_0) I^{-1}(\mu_0) = \frac{n^2(\bar{x} - 2)^2}{16} \cdot \frac{4}{n} = \frac{n(\bar{x} - 2)^2}{4}$$

Therefore, $W = LM = LR$ for this example with known variance $\sigma^2 = 4$. These tests are all based upon the $|\bar{x} - 2| \geq k$ critical region, where k is determined such that the size of the test is α. In general, these test statistics are not always equal, as is shown in the next example.

Example 4: For a random sample $x_1, \ldots, x_n$ drawn from a $N(\mu, \sigma^2)$ with *unknown* σ^2, test the hypothesis $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. Problem 5, part (c), asks the reader to verify that

$$LR = n \log\left[\frac{\sum_{i=1}^{n}(x_i - 2)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \text{ whereas } W = \frac{n^2(\bar{x} - 2)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and } LM = \frac{n^2(\bar{x} - 2)^2}{\sum_{i=1}^{n}(x_i - 2)^2}.$$

One can easily show that $LM/n = (W/n)/[1+(W/n)]$ and $LR/n = \log[1+(W/n)]$. Let $y = W/n$, then using the inequality $y \geq \log(1 + y) \geq y/(1 + y)$, one can conclude that $W \geq LR \geq LM$. This inequality was derived by Berndt and Savin (1977), and will be considered again when we study test of hypotheses in the general linear model. Note, however that all three test statistics are based upon $|\bar{x} - 2| \geq k$ and for finite n, the same exact critical value could be obtained from the Normally distributed $\bar{x}$. This section introduced the W, LR and LM test statistics, all of which have the same asymptotic distribution. In addition, we showed that using the normal distribution, when σ^2 is known, $W = LR = LM$ for testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. However, when σ^2 is unknown, we showed that $W \geq LR \geq LM$ for the same hypothesis.

Example 5: For a random sample $x_1, \ldots, x_n$ drawn from a Bernoulli distribution with parameter θ, test the hypothesis $H_0; \theta = \theta_0$ versus $H_1; \theta \neq \theta_0$, where θ_0 is a known positive fraction. This example is based on Engle (1984). Problem 4, part (i), asks the reader to derive LR, W and LM for H_0; $\theta = 0.2$ versus H_1; $\theta \neq 0.2$. The likelihood $L(\theta)$ and the Score $S(\theta)$ were derived in section 2.2. One can easily verify that

$$C(\theta) = |\frac{\partial^2 \log L(\theta)}{\partial \theta^2}| = \frac{\sum_{i=1}^{n} x_i}{\theta^2} + \frac{n - \sum_{i=1}^{n} x_i}{(1 - \theta)^2}$$

and

$$I(\theta) = -E\left[\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right] = \frac{n}{\theta(1-\theta)}$$

The Wald statistic is based on

$$W = (\widehat{\theta}_{MLE} - \theta_0)^2 I(\widehat{\theta}_{MLE}) = (\bar{x} - \theta_0)^2 \cdot \frac{n}{\bar{x}(1-\bar{x})} = \frac{(\bar{x} - \theta_0)^2}{\bar{x}(1-\bar{x})/n}$$

using the fact that $\widehat{\theta}_{MLE} = \bar{x}$. The LM statistic is based on

$$LM = S^2(\theta_0) I^{-1}(\theta_0) = \frac{(\bar{x} - \theta_0)^2}{[\theta_0(1-\theta_0)/n]^2} \cdot \frac{\theta_0(1-\theta_0)}{n} = \frac{(\bar{x} - \theta_0)^2}{\theta_0(1-\theta_0)/n}$$

Note that the numerator of the W and LM are the same. It is the denominator which is the $\text{var}(\bar{x}) = \theta(1-\theta)/n$ that is different. For Wald, this $\text{var}(\bar{x})$ is evaluated at $\widehat{\theta}_{MLE}$, whereas for LM, this is evaluated at θ_0.

The LR statistic is based on

$$\log L(\widehat{\theta}_{MLE}) = \textstyle\sum_{i=1}^n x_i \log \bar{x} + (n - \sum_{i=1}^n x_i)\log(1-\bar{x})$$

and

$$\log L(\theta_0) = \textstyle\sum_{i=1}^n x_i \log \theta_0 + (n - \sum_{i=1}^n x_i)\log(1-\theta_0)$$

so that

$$\begin{aligned} LR &= -2\log L(\theta_0) + 2\log L(\widehat{\theta}_{MLE}) = -2[\textstyle\sum_{i=1}^n x_i(\log\theta_0 - \log\bar{x}) \\ &\quad + (n - \sum_{i=1}^n x_i)(\log(1-\theta_0) - \log(1-\bar{x}))] \end{aligned}$$

For this example, LR looks different from W and LM. However, a second-order Taylor-Series expansion of LR around $\theta_0 = \bar{x}$ yields the same statistic. Also, for $n \to \infty$, plim $\bar{x} = \theta$ and if H_0 is true, then all three statistics are asymptotically equivalent. Note also that all three test statistics are based upon $|\bar{x} - \theta_0| \geq k$ and for finite n, the same exact critical value could be obtained from the binomial distribution. See problem 19 for more examples of the conflict in test of hypotheses using the W, LR and LM test statistics.

Bera and Permaratne (2001, p. 58) tell the following amusing story that can bring home the interrelationship among the three tests: "Once around 1946 Ronald Fisher invited Jerzy Neyman, Abraham Wald, and C.R. Rao to his lodge for afternoon tea. During their conversation, Fisher mentioned the problem of deciding whether his dog, who had been going to an "obedience school" for some time, was disciplined enough. Neyman quickly came up with an idea: leave the dog free for some time and then put him on his leash. If there is not much difference in his behavior, the dog can be thought of as having completed the course successfully. Wald, who lost his family in the concentration camps, was adverse to any restrictions and simply suggested leaving the dog free and seeing whether it behaved properly. Rao, who had observed the nuisances of stray dogs in Calcutta streets did not like the idea of letting the dog roam freely and suggested keeping the dog on a leash at all times and observing how hard it pulls on the leash. If it pulled too much, it needed more training. That night when Rao was back in his Cambridge dormitory after tending Fisher's mice at the genetics laboratory, he suddenly realized the connection of Neyman and Wald's recommendations to the Neyman-Pearson LR and Wald tests. He got an idea and the rest is history."

2.5 Confidence Intervals

Estimation methods considered in section 2.2 give us a point estimate of a parameter, say μ, and that is the best bet, given the data and the estimation method, of what μ might be. But it is always good policy to give the client an interval, rather than a point estimate, where with some degree of confidence, usually 95% confidence, we expect μ to lie. We have seen in Figure 2.5 that for a $N(0,1)$ random variable z, we have

$$\Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] = 1 - \alpha$$

and for $\alpha = 5\%$, this probability is 0.95, giving the required 95% confidence. In fact, $z_{\alpha/2} = 1.96$ and

$$\Pr[-1.96 \leq z \leq 1.96] = 0.95$$

This says that if we draw 100 random numbers from a $N(0,1)$ density, (using a normal random number generator) we expect 95 out of these 100 numbers to lie in the $[-1.96, 1.96]$ interval. Now, let us get back to the problem of estimating μ from a random sample $x_1, \ldots, x_n$ drawn from a $N(\mu, \sigma^2)$ distribution. We found out that $\widehat{\mu}_{MLE} = \bar{x}$ and $\bar{x} \sim N(\mu, \sigma^2/n)$. Hence, $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ is $N(0,1)$. The point estimate for μ is $\bar{x}$ observed from the sample, and the 95% confidence interval for μ is obtained by replacing z by its value in the above probability statement:

$$\Pr[-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}] = 1 - \alpha$$

Assuming σ is known for the moment, one can rewrite this probability statement after some simple algebraic manipulations as

$$\Pr[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}) \leq \mu \leq \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})] = 1 - \alpha$$

Note that this probability statement has random variables on both ends and the probability that these random variables sandwich the unknown parameter μ is $1 - \alpha$. With the same confidence of drawing 100 random $N(0,1)$ numbers and finding 95 of them falling in the $(-1.96, 1.96)$ range we are confident that if we drew a 100 samples and computed a 100 $\bar{x}$'s, and a 100 intervals $(\bar{x} \pm 1.96\ \sigma/\sqrt{n})$, μ will lie in these intervals in 95 out of 100 times.

If σ is not known, and is replaced by s, then problem 12 shows that this is equivalent to dividing a $N(0,1)$ random variable by an independent χ^2_{n-1} random variable divided by its degrees of freedom, leading to a t-distribution with $(n-1)$ degrees of freedom. Hence, using the t-tables for $(n-1)$ degrees of freedom

$$\Pr[-t_{\alpha/2;n-1} \leq t_{n-1} \leq t_{\alpha/2;n-1}] = 1 - \alpha$$

and replacing t_{n-1} by $(\bar{x} - \mu)/(s/\sqrt{n})$ one gets

$$\Pr[\bar{x} - t_{\alpha/2;n-1}(s/\sqrt{n}) \leq \mu \leq \bar{x} + t_{\alpha/2;n-1}(s/\sqrt{n})] = 1 - \alpha$$

Note that the degrees of freedom $(n-1)$ for the t-distribution come from s and the corresponding critical value $t_{n-1;\alpha/2}$ is therefore sample specific, unlike the corresponding case for the normal density where $z_{\alpha/2}$ does not depend on n. For small n, the $t_{\alpha/2}$ values differ drastically from

Table 2.1 Descriptive Statistics for the Earnings Data

Sample: 1 595								
	LWAGE	WKS	ED	EX	MS	FEM	BLK	UNION
Mean	6.9507	46.4520	12.8450	22.8540	0.8050	0.1126	0.0723	0.3664
Median	6.9847	48.0000	12.0000	21.0000	1.0000	0.0000	0.0000	0.0000
Maximum	8.5370	52.0000	17.0000	51.0000	1.0000	1.0000	1.0000	1.0000
Minimum	5.6768	5.0000	4.0000	7.0000	0.0000	0.0000	0.0000	0.0000
Std. Dev.	0.4384	5.1850	2.7900	10.7900	0.3965	0.3164	0.2592	0.4822
Skewness	–0.1140	–2.7309	–0.2581	0.4208	–1.5400	2.4510	3.3038	0.5546
Kurtosis	3.3937	13.7780	2.7127	2.0086	3.3715	7.0075	11.9150	1.3076
Jarque-Bera	5.13	3619.40	8.65	41.93	238.59	993.90	3052.80	101.51
Probability	0.0769	0.0000	0.0132	0.0000	0.0000	0.0000	0.0000	0.0000
Observations	595	595	595	595	595	595	595	595

$z_{\alpha/2}$, emphasizing the importance of using the t-density in small samples. When n is large the difference between $z_{\alpha/2}$ and $t_{\alpha/2}$ diminishes as the t-density becomes more like a normal density. For $n = 20$, and $\alpha = 0.05, t_{\alpha/2;n-1} = 2.093$ as compared with $z_{\alpha/2} = 1.96$. Therefore,

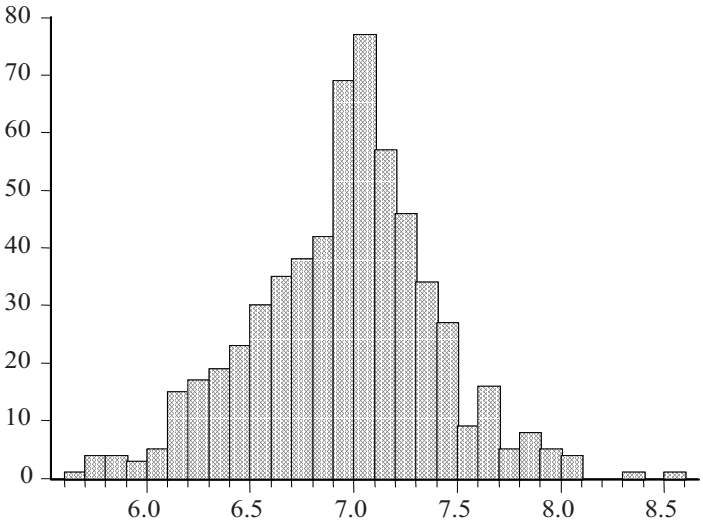
$$\Pr[-2.093 \leq t_{n-1} \leq 2.093] = 0.95$$

and μ lies in $\bar{x} \pm 2.093(s/\sqrt{n})$ with 95% confidence.

More examples of confidence intervals can be constructed, but the idea should be clear. Note that these confidence intervals are the other side of the coin for tests of hypotheses. For example, in testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$ for a known σ, we discovered that the Likelihood Ratio test is based on the same probability statement that generated the confidence interval for μ. In classical tests of hypothesis, we choose the level of confidence $\alpha = 5\%$ and compute $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$. This can be done since σ is known and $\mu = 2$ under the null hypothesis H_0. Next, we do not reject H_0 if z lies in the $(-z_{\alpha/2}, z_{\alpha/2})$ interval and reject H_0 otherwise. For confidence intervals, on the other hand, we do not know μ, and armed with a level of confidence $(1 - \alpha)\%$ we construct the interval that should contain μ with that level of confidence. Having done that, if $\mu = 2$ lies in that 95% confidence interval, then we cannot reject $H_0; \mu = 2$ at the 5% level. Otherwise, we reject H_0. This highlights the fact that any value of μ that lies in this 95% confidence interval (assuming it was our null hypothesis) cannot be rejected at the 5% level by this sample. This is why we do not say "accept H_0", but rather we say "do not reject H_0".

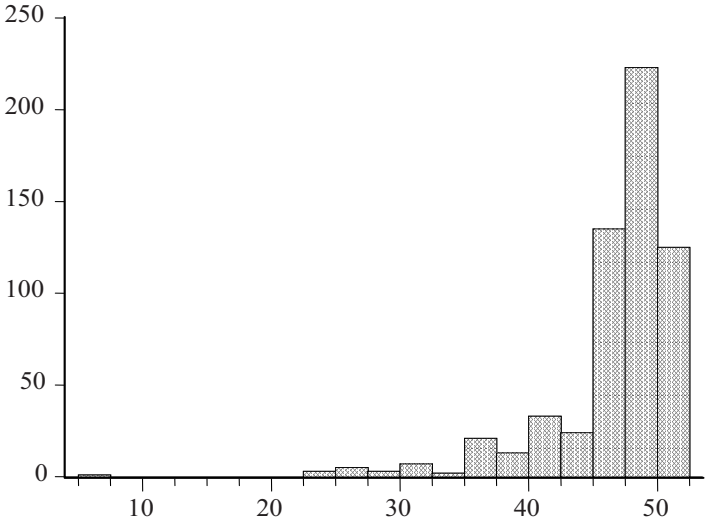
2.6 Descriptive Statistics

In Chapter 4, we will consider the estimation of a simple wage equation based on 595 individuals drawn from the Panel Study of Income Dynamics for 1982. This data is available on the Springer web site as EARN.ASC. Table 2.1 gives the descriptive statistics using EViews for a subset of the variables in this data set.



Series: LWAGE	
Sample 1 595	
Observations 595	
Mean	6.950745
Median	6.984720
Maximum	8.537000
Minimum	5.676750
Std. Dev.	0.438403
Skewness	–0.114001
Kurtosis	3.393651
Jarque-Bera	5.130525
Probability	0.076899

Figure 2.8 Log (Wage) Histogram



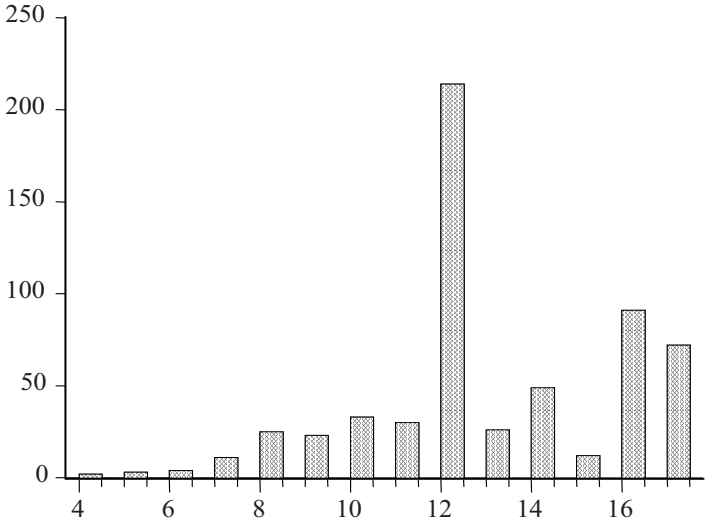
Series: WKS	
Sample 1595	
Observations 595	
Mean	46.45210
Median	48.00000
Maximum	52.00000
Minimum	5.000000
Std. Dev.	5.185025
Skewness	–2.730880
Kurtosis	13.77787
Jarque-Bera	3619.416
Probability	0.000000

Figure 2.9 Weeks Worked Histogram

The average log wage is \$6.95 for this sample with a minimum of \$5.68 and a maximum of \$8.54. The standard deviation of log wage is 0.44. A plot of the log wage histogram is given in Figure 2.8. Weeks worked vary between 5 and 52 with an average of 46.5 and a standard deviation of 5.2. This variable is highly skewed as evidenced by the histogram in Figure 2.9. Years of education vary between 4 and 17 with an average of 12.8 and a standard deviation of 2.79. There is the usual bunching up at 12 years, which is also the median, as is clear from Figure 2.10.

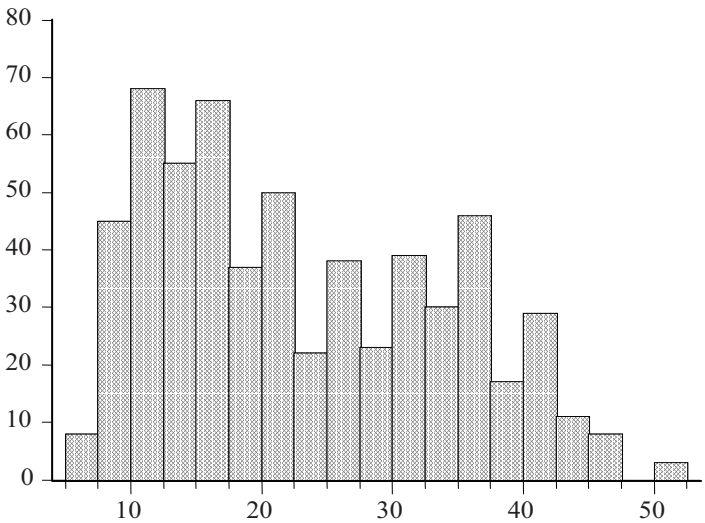
Experience varies between 7 and 51 with an average of 22.9 and a standard deviation of 10.79. The distribution of this variable is skewed to the left, as shown in Figure 2.11.

Marital status is a qualitative variable indicating whether the individual is married or not. This information is recoded as a numeric $(1, 0)$ variable, one if the individual is married and zero



Series: ED	
Sample 1595	
Observation 595	
Mean	12.84538
Median	12.00000
Maximum	17.00000
Minimum	4.000000
Std. Dev.	2.790006
Skewness	–0.258116
Kurtosis	2.712730
Jarque-Bera	8.652780
Probability	0.013215

Figure 2.10 Years of Education Histogram



Series: EX	
Sample 1 595	
Observation 595	
Mean	22.85378
Median	21.00000
Maximum	51.00000
Minimum	7.000000
Std. Dev.	10.79018
Skewness	0.420826
Kurtosis	2.008578
Jarque-Bera	41.93007
Probability	0.000000

Figure 2.11 Years of Experience Histogram

otherwise. This recoded variable is also known as a dummy variable. It is basically a switch turning on when the individual is married and off when he or she is not. Female is another dummy variable taking the value one when the individual is a female and zero otherwise. Black is a dummy variable taking the value one when the individual is black and zero otherwise. Union is a dummy variable taking the value one if the individual's wage is set by a union contract and zero otherwise. The minimum and maximum values for these dummy variables are obvious. But if they were not zero and one, respectively, you know that something is wrong. The average is a meaningful statistic indicating the percentage of married individuals, females, blacks and union contracted wages in the sample. These are 80.5, 11.3, 7.2 and 30.6%, respectively. We would like to investigate the following claims: (i) women are paid less than men; (ii) blacks are paid less than non-blacks; (iii) married individuals earn more than non-married individuals; and (iv) union contracted wages are higher than non-union wages.

Table 2.2 Test for the Difference in Means

	Average log wage	Difference
Male	$7,004	–0.474
Female	$6,530	(–8.86)
Non-Black	$6,978	–0.377
Black	$6,601	(–5.57)
Not Married	$6,664	0.356
Married	$7,020	(8.28)
Non-Union	$6,945	0.017
Union	$6,962	(0.45)

Table 2.3 Correlation Matrix

	LWAGE	WKS	ED	EX	MS	FEM	BLK	UNION
LWAGE	1.0000	0.0403	0.4566	0.0873	0.3218	–0.3419	–0.2229	0.0183
WKS	0.0403	1.0000	0.0002	–0.1061	0.0782	–0.0875	–0.0594	–0.1721
ED	0.4566	0.0002	1.0000	–0.2219	0.0184	–0.0012	–0.1196	–0.2719
EX	0.0873	–0.1061	–0.2219	1.0000	0.1570	–0.0938	0.0411	0.0689
MS	0.3218	0.0782	0.0184	0.1570	1.0000	–0.7104	–0.2231	0.1189
FEM	–0.3419	–0.0875	–0.0012	–0.0938	–0.7104	1.0000	0.2086	–0.1274
BLK	–0.2229	–0.0594	–0.1196	0.0411	–0.2231	0.2086	1.0000	0.0302
UNION	0.0183	–0.1721	–0.2719	0.0689	0.1189	–0.1274	0.0302	1.0000

A simple first check could be based on computing the average log wage for each of these categories and testing whether the difference in means is significantly different from zero. This can be done using a t-test, see Table 2.2. The average log wage for males and females is given along with their difference and the corresponding t-statistic for the significance of this difference. Other rows of Table 2.2 give similar statistics for other groupings. In Chapter 4, we will show that this t-test can be obtained from a simple regression of log wage on the categorical dummy variable distinguishing the two groups. In this case, the Female dummy variable. From Table 2.2, it is clear that only the difference between union and non-union contracted wages are insignificant.

One can also plot log wage versus experience, see Figure 2.12, log wage versus education, see Figure 2.13, and log wage versus weeks, see Figure 2.14.

The data shows that, in general, log wage increases with education level, weeks worked, but that it exhibits a rising and then a declining pattern with more years of experience. Note that the t-tests based on the difference in log wage across two groupings of individuals, by sex, race or marital status, or the figures plotting log wage versus education, log wage versus weeks worked are based on pairs of variables in each case. A nice summary statistic based also on pairwise comparisons of these variables is the correlation matrix across the data. This is given in Table 2.3.

The signs of this correlation matrix give the direction of linear relationship between the corresponding two variables, while the magnitude gives the strength of this correlation. In Chapter 3, we will see that these simple correlations when squared give the percentage of

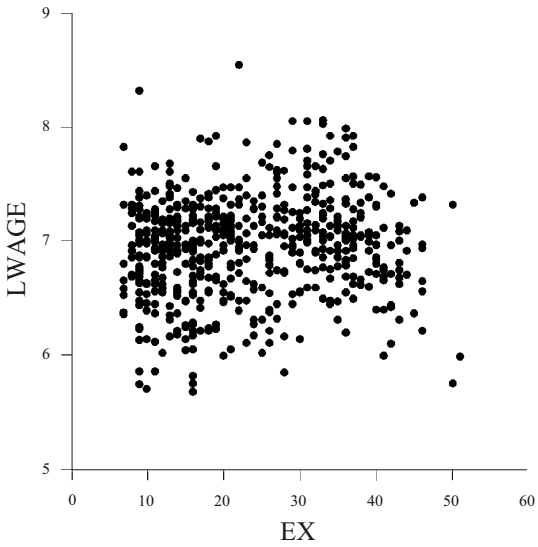


Figure 2.12 Log (Wage) Versus Experience

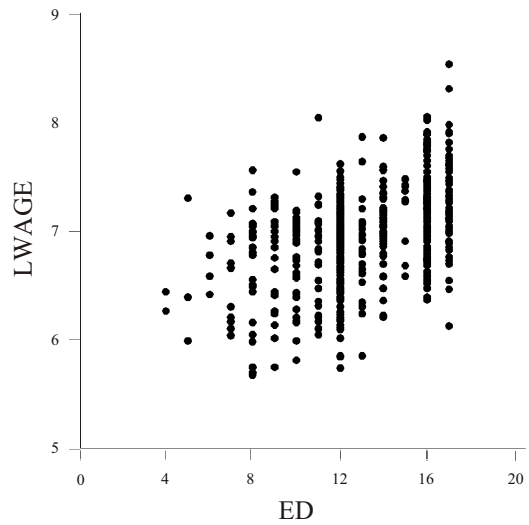


Figure 2.13 Log (Wage) Versus Education

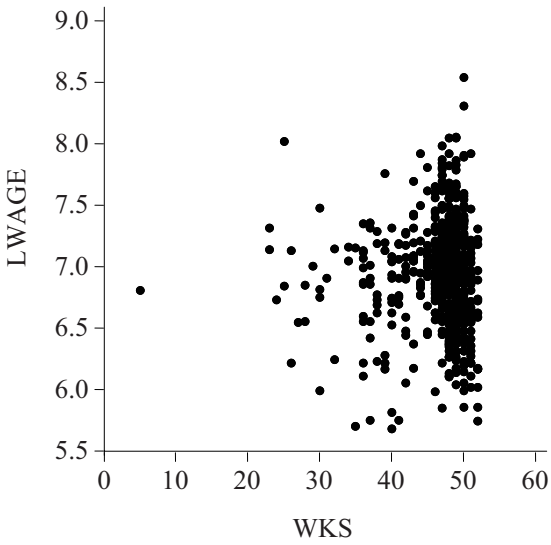


Figure 2.14 Log (Wage) Versus Weeks

variation that one of these variables explain in the other. For example, the simple correlation coefficient between log wage and marital status is 0.32. This means that marital status explains $(0.32)^2$ or 10% of the variation in log wage.

One cannot emphasize enough how important it is to check one's data. It is important to compute the descriptive statistics, simple plots of the data and simple correlations. A wrong minimum or maximum could indicate some possible data entry errors. Troughs or peaks in these plots may indicate important events for time series data, like wars or recessions, or influential observations. More on this in Chapter 8. Simple correlation coefficients that equal one indicate perfectly collinear variables and warn of the failure of a linear regression that has both variables included among the regressors, see Chapter 4.

Notes

1. Actually $E(s^2) = \sigma^2$ does not need the normality assumption. This fact along with the proof of $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$, under Normality, can be easily shown using matrix algebra and is deferred to Chapter 7.
2. This can be proven using the Chebyshev's inequality, see Hogg and Craig (1995).
3. See Hogg and Craig (1995) for the type of regularity conditions needed for these distributions.

Problems

1. *Variance and Covariance of Linear Combinations of Random Variables.* Let a, b, c, d, e and f be arbitrary constants and let X and Y be two random variables.
   (a) Show that $\text{var}(a + bX) = b^2 \text{ var}(X)$.
   (b) $\text{var}(a + bX + cY) = b^2\text{var}(X) + c^2 \text{ var}(Y) + 2bc \text{ cov}(X, Y)$.
   (c) $\text{cov}[(a + bX + cY), (d + eX + fY)] = be \text{ var}(X) + cf \text{ var}(Y) + (bf + ce) \text{ cov}(X, Y)$.
2. *Independence and Simple Correlation.*
   (a) Show that if X and Y are independent, then $E(XY) = E(X)E(Y) = \mu_x\mu_y$ where $\mu_x = E(X)$ and $\mu_y = E(Y)$. Therefore, $\text{cov}(X, Y) = E(X - \mu_x)(Y - \mu_y) = 0$.
   (b) Show that if $Y = a + bX$, where a and b are arbitrary constants, then $\rho_{xy} = 1$ if $b > 0$ and -1 if $b < 0$.
3. *Zero Covariance Does Not Necessarily Imply Independence.* Let $X = -2, -1, 0, 1, 2$ with $\Pr[X = x] = 1/5$. Assume a perfect quadratic relationship between Y and X, namely $Y = X^2$. Show that $\text{cov}(X, Y) = E(X^3) = 0$. Deduce that $\rho_{XY} = \text{correlation } (X, Y) = 0$. The simple correlation coefficient ρ_{XY} measures the strength of the *linear* relationship between X and Y. For this example, it is zero even though there is a perfect *nonlinear* relationship between X and Y. This is also an example of the fact that if $\rho_{XY} = 0$, then X and Y are not necessarily independent. $\rho_{xy} = 0$ is a necessary but not sufficient condition for X and Y to be independent. The converse, however, is true, i.e., if X and Y are independent, then $\rho_{XY} = 0$, see problem 2.
4. The *Binomial Distribution* is defined as the number of successes in n independent Bernoulli trials with probability of success θ. This discrete probability function is given by

$$f(X;\theta) = \binom{n}{X}\theta^X(1-\theta)^{n-X} \quad X = 0, 1, \ldots, n$$

   and zero elsewhere, with $\binom{n}{X} = n!/[X!(n-X)!]$.
   (a) Out of 20 candidates for a job with a probability of hiring of 0.1. Compute the probabilities of getting $X = 5$ or 6 people hired?
   (b) Show that $\binom{n}{X} = \binom{n}{n-X}$ and use that to conclude that $\text{b}(n, X, \theta) = \text{b}(n, n - X, 1 - \theta)$.
   (c) Verify that $E(X) = n\theta$ and $\text{var}(X) = n\theta(1 - \theta)$.
   (d) For a random sample of size n drawn from the Bernoulli distribution with parameter θ, show that $\bar{X}$ is the MLE of θ.
   (e) Show that $\bar{X}$, in part (d), is unbiased and consistent for θ.
   (f) Show that $\bar{X}$, in part (d), is sufficient for θ.

(g) Derive the Cramér-Rao lower bound for any unbiased estimator of θ. Is $\bar{X}$, in part (d), MVU for θ?

(h) For $n = 20$, derive the uniformly most powerful critical region of size $\alpha \leq 0.05$ for testing $H_0; \theta = 0.2$ versus $H_1; \theta = 0.6$. What is the probability of type II error for this test criteria?

(i) Form the Likelihood Ratio test for testing $H_0; \theta = 0.2$ versus $H_1; \theta \neq 0.2$. Derive the Wald and LM test statistics for testing H_0 versus H_1. When is the Wald statistic greater than the LM statistic?

5. For a random sample of size n drawn from the *Normal distribution* with mean μ and variance σ^2.

(a) Show that s^2 is a sufficient statistic for σ^2.

(b) Using the fact that $(n-1)s^2/\sigma^2$ is χ^2_{n-1} (without proof), verify that $E(s^2) = \sigma^2$ and that $\text{var}(s^2) = 2\sigma^4/(n-1)$ as shown in the text.

(c) Given that σ^2 is unknown, form the Likelihood Ratio test statistic for testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. Derive the Wald and Lagrange Multiplier statistics for testing H_0 versus H_1. Verify that they are given by the expressions in example 4.

(d) Another derivation of the $W \geq LR \geq LM$ inequality for the null hypothesis given in part (c) can be obtained as follows: Let $\tilde{\mu}$, $\tilde{\sigma}^2$ be the restricted maximum likelihood estimators under $H_0; \mu = \mu_0$. Let $\hat{\mu}$, $\hat{\sigma}^2$ be the corresponding unrestricted maximum likelihood estimators under the alternative $H_1; \mu \neq \mu_0$. Show that $W = -2\log[L(\tilde{\mu}, \hat{\sigma}^2)/L(\hat{\mu}, \hat{\sigma}^2)]$ and $LM = -2\log[L(\tilde{\mu}, \tilde{\sigma}^2)/L(\hat{\mu}, \tilde{\sigma}^2)]$ where $L(\mu, \sigma^2)$ denotes the likelihood function. Conclude that $W \geq LR \geq LM$, see Breusch (1979). This is based on Baltagi (1994).

(e) Given that μ is unknown, form the Likelihood Ratio test statistic for testing $H_0; \sigma = 3$ versus $H_1; \sigma \neq 3$.

(f) Form the Likelihood Ratio test statistic for testing $H_0; \mu = 2, \sigma^2 = 4$ against the alternative that $H_1; \mu \neq 2$ or $\sigma^2 \neq 4$.

(g) For $n = 20, s^2 = 9$ construct a 95% confidence interval for σ^2.

6. The *Poisson* distribution can be defined as the limit of a Binomial distribution as $n \to \infty$ and $\theta \to 0$ such that $n\theta = \lambda$ is a positive constant. For example, this could be the probability of a rare disease and we are random sampling a large number of inhabitants, or it could be the rare probability of finding oil and n is the large number of drilling sights. This discrete probability function is given by

$$f(X; \lambda) = \frac{e^{-\lambda}\lambda^X}{X!} \quad X = 0, 1, 2, \ldots$$

For a random sample from this Poisson distribution

(a) Show that $E(X) = \lambda$ and $\text{var}(X) = \lambda$.

(b) Show that the MLE of λ is $\widehat{\lambda}_{MLE} = \bar{X}$.

(c) Show that the method of moments estimator of λ is also $\bar{X}$.

(d) Show that $\bar{X}$ is unbiased and consistent for λ.

(e) Show that $\bar{X}$ is sufficient for λ.

(f) Derive the Cramér-Rao lower bound for any unbiased estimator of λ. Show that $\bar{X}$ attains that bound.

(g) For $n = 9$, derive the Uniformly Most Powerful critical region of size $\alpha \leq 0.05$ for testing $H_0; \lambda = 2$ versus $H_1; \lambda = 4$.

(h) Form the Likelihood Ratio test for testing H_0; $\lambda = 2$ versus H_1; $\lambda \neq 2$. Derive the Wald and LM statistics for testing H_0 versus H_1. When is the Wald test statistic greater than the LM statistic?

7. The *Geometric* distribution is known as the probability of waiting for the first success in independent repeated trials of a Bernoulli process. This could occur on the 1st, 2nd, 3rd,.. trials.

$$g(X;\theta) = \theta(1-\theta)^{X-1} \text{ for } X = 1, 2, 3, \ldots$$

(a) Show that $E(X) = 1/\theta$ and $\text{var}(X) = (1-\theta)/\theta^2$.

(b) Given a random sample from this Geometric distribution of size n, find the MLE of θ and the method of moments estimator of θ.

(c) Show that $\bar{X}$ is unbiased and consistent for $1/\theta$.

(d) For $n = 20$, derive the Uniformly Most Powerful critical region of size $\alpha \leq 0.05$ for testing H_0; $\theta = 0.5$ versus H_1; $\theta = 0.3$.

(e) Form the Likelihood Ratio test for testing H_0; $\theta = 0.5$ versus H_1; $\theta \neq 0.5$. Derive the Wald and LM statistics for testing H_0 versus H_1. When is the Wald statistic greater than the LM statistic?

8. The *Uniform* density, defined over the unit interval $[0, 1]$, assigns a unit probability for all values of X in that interval. It is like a roulette wheel that has an equal chance of stopping anywhere between 0 and 1.

$$\begin{aligned} f(X) &= 1 \qquad 0 \leq X \leq 1 \\ &= 0 \qquad \text{elsewhere} \end{aligned}$$

Computers are equipped with a Uniform (0,1) random number generator so it is important to understand these distributions.

(a) Show that $E(X) = 1/2$ and $\text{var}(X) = 1/12$.

(b) What is the $\Pr[0.1 < X < 0.3]$? Does it matter if we ask for the $\Pr[0.1 \leq X \leq 0.3]$?

9. The *Exponential* distribution is given by

$$f(X;\theta) = \frac{1}{\theta} e^{-X/\theta} \quad X > 0 \text{ and } \theta > 0$$

This is a skewed and continuous distribution defined only over the positive quadrant.

(a) Show that $E(X) = \theta$ and $\text{var}(X) = \theta^2$.

(b) Show that $\widehat{\theta}_{MLE} = \bar{X}$.

(c) Show that the method of moments estimator of θ is also $\bar{X}$.

(d) Show that $\bar{X}$ is an unbiased and consistent estimator of θ.

(e) Show that $\bar{X}$ is sufficient for θ.

(f) Derive the Cramér-Rao lower bound for any unbiased estimator of θ? Is $\bar{X}$ MVU for θ?

(g) For $n = 20$, derive the Uniformly Most Powerful critical region of size $\alpha \leq 0.05$ for testing $H_0; \theta = 1$ versus $H_1; \theta = 2$.

(h) Form the Likelihood Ratio test for testing $H_0; \theta = 1$ versus $H_1; \theta \neq 1$. Derive the Wald and LM statistics for testing H_0 versus H_1. When is the Wald statistic greater than the LM statistic?

10. The *Gamma* distribution is given by

$$\begin{aligned} f(X;\alpha,\beta) &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} X^{\alpha-1} e^{-X/\beta} && \text{for } X > 0 \\ &= 0 && \text{elsewhere} \end{aligned}$$

where α and $\beta > 0$ and $\Gamma(\alpha) = (\alpha - 1)!$ This is a skewed and continuous distribution.

(a) Show that $E(X) = \alpha\beta$ and $\text{var}(X) = \alpha\beta^2$.

(b) For a random sample drawn from this Gamma density, what are the method of moments estimators of α and β?

(c) Verify that for $\alpha = 1$ and $\beta = \theta$, the Gamma probability density function reverts to the Exponential p.d.f. considered in problem 9.

(d) We state without proof that for $\alpha = r/2$ and $\beta = 2$, this Gamma density reduces to a χ^2 distribution with r degrees of freedom, denoted by χ^2_r. Show that $E(\chi^2_r) = r$ and $\text{var}(\chi^2_r) = 2r$.

(e) For a random sample from the χ^2_r distribution, show that $(X_1 X_2, \ldots, X_n)$ is a sufficient statistic for r.

(f) One can show that the square of a $N(0,1)$ random variable is a χ^2 random variable with 1 degree of freedom, see the Appendix to the chapter. Also, one can show that the sum of independent χ^2's is a χ^2 random variable with degrees of freedom equal the sum of the corresponding degrees of freedom of the individual χ^2's, see problem 15. This will prove useful for testing later on. Using these results, verify that the sum of squares of m independent $N(0,1)$ random variables is a χ^2 with m degrees of freedom.

11. The *Beta* distribution is defined by

$$\begin{aligned} f(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} X^{\alpha-1}(1-X)^{\beta-1} && \text{for } 0 < X < 1 \\ &= 0 && \text{elsewhere} \end{aligned}$$

where $\alpha > 0$ and $\beta > 0$. This is a skewed continuous distribution.

(a) For $\alpha = \beta = 1$ this reverts back to the Uniform $(0,1)$ probability density function. Show that $E(X) = (\alpha/\alpha + \beta)$ and $\text{var}(X) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$.

(b) Suppose that $\alpha = 1$, find the estimators of β using the method of moments and the method of maximum likelihood.

12. The *t-distribution* with r degrees of freedom can be defined as the ratio of two independent random variables. The numerator being a $N(0,1)$ random variable and the denominator being the square-root of a χ^2_r random variable divided by its degrees of freedom. The t-distribution is a symmetric distribution like the Normal distribution but with fatter tails. As $r \to \infty$, the t-distribution approaches the Normal distribution.

(a) Verify that if $X_1, \ldots, X_n$ are a random sample drawn from a $N(\mu, \sigma^2)$ distribution, then $z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is $N(0,1)$.

(b) Use the fact that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ to show that $t = z/\sqrt{s^2/\sigma^2} = (\bar{X} - \mu)/(s/\sqrt{n})$ has a t-distribution with $(n-1)$ degrees of freedom. We use the fact that s^2 is independent of $\bar{X}$ without proving it.

(c) For $n = 16$, $\bar{x} = 20$ and $s^2 = 4$, construct a 95% confidence interval for μ.

13. The *F-distribution* can be defined as the ratio of two independent χ^2 random variables each divided by its corresponding degrees of freedom. It is commonly used to test the equality of variances. Let s_1^2 be the sample variance from a random sample of size n_1 drawn from $N(\mu_1, \sigma_1^2)$ and let s_2^2 be the sample variance from another random sample of size n_2 drawn from $N(\mu_2, \sigma_2^2)$. We know that $(n_1-1)s_1^2/\sigma_1^2$ is $\chi^2_{(n_1-1)}$ and $(n_2-1)s_2^2/\sigma_2^2$ is $\chi^2_{(n_2-1)}$. Taking the ratio of those two independent χ^2 random variables divided by their appropriate degrees of freedom yields

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

which under the null hypothesis $H_0; \sigma_1^2 = \sigma_2^2$ gives $F = s_1^2/s_2^2$ and is distributed as F with (n_1-1) and (n_2-1) degrees of freedom. Both s_1^2 and s_2^2 are observable, so F can be computed and compared to critical values for the F-distribution with the appropriate degrees of freedom. Two inspectors drawing two random samples of size 25 and 31 from two shifts of a factory producing steel rods, find that the sampling variance of the lengths of these rods are 15.6 and 18.9 inches squared. Test whether the variances of the two shifts are the same.

14. *Moment Generating Function (MGF).*

    (a) Derive the MGF of the Binomial distribution defined in problem 4. Show that it is equal to $[(1-\theta)+\theta e^t]^n$.

    (b) Derive the MGF of the Normal distribution defined in problem 5. Show that it is $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$.

    (c) Derive the MGF of the Poisson distribution defined in problem 6. Show that it is $e^{\lambda(e^t-1)}$.

    (d) Derive the MGF of the Geometric distribution defined in problem 7. Show that it is $\theta e^t/[1-(1-\theta)e^t]$.

    (e) Derive the MGF of the Exponential distribution defined in problem 9. Show that it is $1/(1-\theta t)$.

    (f) Derive the MGF of the Gamma distribution defined in problem 10. Show that it is $(1-\beta t)^{-\alpha}$. Conclude that the MGF of a χ_r^2 is $(1-2t)^{-\frac{r}{2}}$.

    (g) Obtain the mean and variance of each distribution by differentiating the corresponding MGF derived in parts (a) through (f).

15. *Moment Generating Function Method.*

    (a) Show that if $X_1, \ldots, X_n$ are independent Poisson distributed with parameters (λ_i) respectively, then $Y = \sum_{i=1}^n X_i$ is Poisson with parameter $\sum_{i=1}^n \lambda_i$.

    (b) Show that if $X_1, \ldots, X_n$ are independent Normally distributed with parameters (μ_i, σ_i^2), then $Y = \sum_{i=1}^n X_i$ is Normal with mean $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$.

    (c) Deduce from part (b) that if $X_1, \ldots, X_n$ are IIN(μ, σ^2), then $\bar{X} \sim N(\mu, \sigma^2/n)$.

    (d) Show that if $X_1, \ldots, X_n$ are independent χ^2 distributed with parameters (r_i) respectively, then $Y = \sum_{i=1}^n X_i$ is χ^2 distributed with parameter $\sum_{i=1}^n r_i$.

16. *Best Linear Prediction.* (Problems 16 and 17 are based on Amemiya (1994)). Let X and Y be two random variables with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively. Suppose that

$$\rho = \text{correlation}(X, Y) = \sigma_{XY}/\sigma_X\sigma_Y$$

where $\sigma_{XY} = \text{cov}(X, Y)$. Consider the *linear* relationship $Y = \alpha + \beta X$ where α and β are scalars:

    (a) Show that the *best linear predictor* of Y based on X, where *best* in this case means the minimum mean squared error predictor which minimizes $E(Y-\alpha-\beta X)^2$ with respect to α and β is given by $\widehat{Y} = \widehat{\alpha} + \widehat{\beta} X$ where $\widehat{\alpha} = \mu_Y - \widehat{\beta}\mu_X$ and $\widehat{\beta} = \sigma_{XY}/\sigma_X^2 = \rho\sigma_Y/\sigma_X$.

(b) Show that the $\text{var}(\widehat{Y}) = \rho^2\sigma_Y^2$ and that $\widehat{u} = Y - \widehat{Y}$, the prediction error, has mean zero and variance equal to $(1-\rho^2)\sigma_Y^2$. Therefore, ρ^2 can be interpreted as the proportion of σ_Y^2 that is explained by the *best linear* predictor $\widehat{Y}$.

(c) Show that $\text{cov}(\widehat{Y}, \widehat{u}) = 0$.

17. *The Best Predictor.* Let X and Y be the two random variables considered in problem 16. Now consider predicting Y by a general, possibly non-linear, function of X denoted by $h(X)$.

(a) Show that the *best predictor* of Y based on X, where *best* in this case means the minimum mean squared error predictor that minimizes $E[Y - h(X)]^2$ is given by $h(X) = E(Y/X)$. **Hint:** Write $E[Y - h(X)]^2$ as $E\{[Y - E(Y/X)] + [E(Y/X) - h(X)]\}^2$. Expand the square and show that the cross-product term has zero expectation. Conclude that this mean squared error is minimized at $h(X) = E(Y/X)$.

(b) If X and Y are bivariate Normal, show that the *best predictor* of Y based on X is identical to the *best linear predictor* of Y based on X.

18. *Descriptive Statistics.* Using the data used in section 2.6 based on 595 individuals drawn from the Panel Study of Income Dynamics for 1982 and available on the Springer web site as EARN.ASC, replicate the tables and graphs given in that section. More specifically

(a) replicate Table 2.1 which gives the descriptive statistics for a subset of the variables in this data set.

(b) Replicate Figures 2.6–2.11 which plot the histograms for log wage, weeks worked, education and experience.

(c) Replicate Table 2.2 which gives the average log wage for various groups and test the difference between these averages using a t-test.

(d) Replicate Figure 2.12 which plots log wage versus experience. Figure 2.13 which plots log wage versus education and Figure 2.14 which plots log wage versus weeks worked.

(e) Replicate Table 2.3 which gives the correlation matrix among a subset of these variables.

19. *Conflict Among Criteria for Testing Hypotheses: Examples from Non-normal Distributions.* This is based on Baltagi (2000). Berndt and Savin (1977) showed that $W \geq LR \geq LM$ for the case of a multivariate regression model with normal distrubances. Ullah and Zinde-Walsh (1984) showed that this inequality is not robust to non-normality of the disturbances. In the spirit of the latter article, this problem considers simple examples from non-normal distributions and illustrates how this conflict among criteria is affected.

(a) Consider a random sample $x_1, x_2, \ldots, x_n$ from a Poisson distribution with parameter λ. Show that for testing $\lambda = 3$ versus $\lambda \neq 3$ yields $W \geq LM$ for $\bar{x} \leq 3$ and $W \leq LM$ for $\bar{x} \geq 3$.

(b) Consider a random sample $x_1, x_2, \ldots, x_n$ from an Exponential distribution with parameter θ. Show that for testing $\theta = 3$ versus $\theta \neq 3$ yields $W \geq LM$ for $0 < \bar{x} \leq 3$ and $W \leq LM$ for $\bar{x} \geq 3$.

(c) Consider a random sample $x_1, x_2, \ldots, x_n$ from a Bernoulli distribution with parameter θ. Show that for testing $\theta = 0.5$ versus $\theta \neq 0.5$, we will always get $W \geq LM$. Show also, that for testing $\theta = (2/3)$ versus $\theta \neq (2/3)$ we get $W \leq LM$ for $(1/3) \leq \bar{x} \leq (2/3)$ and $W \geq LM$ for $(2/3) \leq \bar{x} \leq 1$ or $0 < \bar{x} \leq (1/3)$.

References

More detailed treatment of the material in this chapter may be found in:

Amemiya, T. (1994), *Introduction to Statistics and Econometrics* (Harvard University Press: Cambridge).

Baltagi, B.H. (1994), "The Wald, LR, and LM Inequality," *Econometric Theory*, Problem 94.1.2, 10: 223–224.

Baltagi, B.H. (2000), "Conflict Among Critera for Testing Hypotheses: Examples from Non-Normal Distributions," *Econometric Theory*, Problem 00.2.4, 16: 288.

Bera A.K. and G. Permaratne (2001), "General Hypothesis Testing," Chapter 2 in Baltagi, B.H. (ed.), *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Berndt, E.R. and N.E. Savin (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrica*, 45: 1263–1278.

Breusch, T.S. (1979), "Conflict Among Criteria for Testing Hypotheses: Extensions and Comments," *Econometrica*, 47: 203–207.

Buse, A. (1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician,* 36 :153–157.

DeGroot, M.H. (1986), *Probability and Statistics* (Addison-Wesley: Mass.).

Freedman, D., R. Pisani, R. Purves and A. Adhikari (1991), *Statistics* (Norton: New York).

Freund, J.E. (1992), *Mathematical Statistics* (Prentice-Hall: New Jersey).

Hogg, R.V. and A.T. Craig (1995), *Introduction to Mathematical Statistics* (Prentice Hall: New Jersey).

Jollife, I.T. (1995), "Sample Sizes and the Central Limit Theorem: The Poisson Distribution as an Illustration," *The American Statistician*, 49: 269.

Kennedy, P. (1992), *A Guide to Econometrics* (MIT Press: Cambridge).

Mood, A.M., F.A. Graybill and D.C. Boes (1974), *Introduction to the Theory of Statistics* (McGraw-Hill: New York).

Spanos, A. (1986), *Statistical Foundations of Econometric Modelling* (Cambridge University Press: Cambridge).

Ullah, A. and V. Zinde-Walsh (1984), "On the Robustness of LM, LR and W Tests in Regression Models," *Econometrica*, 52: 1055–1065.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics* (Wiley: New York).

Appendix

Score and Information Matrix: The likelihood function of a sample $X_1, \ldots, X_n$ drawn from $f(X_i, \theta)$ is really the joint probability density function written as a function of θ:

$$L(\theta) = f(X_1, \ldots, X_n; \theta)$$

This probability density function has the property that $\int L(\theta) d\mathbf{x} = 1$ where the integral is over all $X_1, \ldots, X_n$ written compactly as one integral over $\mathbf{x}$. Differentiating this multiple integral with respect

to θ, one gets

$$\int \frac{\partial L}{\partial \theta} d\mathbf{x} = 0$$

Multiplying and dividing by L, one gets

$$\int \left(\frac{1}{L}\frac{\partial L}{\partial \theta}\right) \; L \; d\mathbf{x} = \int \left(\frac{\partial \log L}{\partial \theta}\right) L \; d\mathbf{x} = 0$$

But the *score* is by definition $S(\theta) = \partial \log L/\partial\theta$. Hence $E[S(\theta)] = 0$. Differentiating again with respect to θ, one gets

$$\int \left[\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) \; L + \int \left(\frac{\partial \log L}{\partial \theta}\right)\left(\frac{\partial L}{\partial \theta}\right)\right] d\mathbf{x} = 0$$

Multiplying and dividing the second term by L one gets

$$E\left[\frac{\partial^2 \log L}{\partial \theta^2} + \left(\frac{\partial \log L}{\partial \theta}\right)^2\right] = 0$$

or

$$E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right] \quad = E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right] = E[S(\theta)]^2$$

But $\text{var}[S(\theta)] = E[S(\theta)]^2$ since $E[S(\theta)] = 0$. Hence $\text{I}(\theta) = \text{var}[S(\theta)]$.

Moment Generating Function (MGF): For the random variable X, the expected value of a special function of X, namely e^{Xt} is denoted by

$$M_X(t) = E(e^{Xt}) = E(1 + Xt + X^2\frac{t^2}{2!} + X^3\frac{t^3}{3!} + ..)$$

where the second equality follows from the Taylor series expansion of e^{Xt} around zero. Therefore,

$$M_X(t) = 1 + E(X)t + E(X^2)\frac{t^2}{2!} + E(X^3)\frac{t^3}{3!} + ..$$

This function of t generates the moments of X as coefficients of an infinite polynomial in t. For example, $\mu = E(X) =$ coefficient of t, and $E(X^2)/2$ is the coefficient of t^2, etc. Alternatively, one can differentiate this MGF with respect to t and obtain $\mu = E(X) = M'_X(0)$, i.e., the first derivative of $M_X(t)$ with respect to t evaluated at $t = 0$. Similarly, $E(X^r) = M^r_X(0)$ which is the r-th derivative of $M_X(t)$ with respect to t evaluated at $t = 0$. For example, for the Bernoulli distribution;

$$M_X(t) = E(e^{Xt}) = \textstyle\sum_{X=0}^{1} e^{Xt}\theta^X(1-\theta)^{1-X} = \theta e^t + (1-\theta)$$

so that $M'_X(t) = \theta e^t$ and $M'_X(0) = \theta = E(X)$ and $M''_X(t) = \theta e^t$ which means that $E(X^2) = M''_X(0) = \theta$.

Hence,

$$\text{var}(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1-\theta).$$

For the Normal distribution, see problem 14, it is easy to show that if $X \sim N(\mu, \sigma^2)$, then $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ and $M'_X(0) = E(X) = \mu$ and $M''_X(0) = E(X^2) = \sigma^2 + \mu^2$.

There is a one-to-one correspondence between the MGF when it exists and the corresponding p.d.f. This means that if Y has a MGF given by e^{2t+4t^2} then Y is normally distributed with mean 2 and variance 8. Similarly, if Z has a MGF given by $(e^t + 1)/2$, then Z is Bernoulli distributed with $\theta = 1/2$.

Change of Variable: If $X \sim N(0,1)$, then one can find the distribution function of $Y = |X|$ by using the *Distribution Function* method. By definition the distribution function of y is defined as

$$\begin{aligned} G(y) &= \Pr[Y \leq y] = \Pr[|X| \leq y] = \Pr[-y \leq X \leq y] \\ &= \Pr[X \leq y] - \Pr[X \leq -y] = F(y) - F(-y) \end{aligned}$$

so that the distribution function of $Y, G(y)$, can be obtained from the distribution function of X, $F(x)$. Since the $N(0,1)$ distribution is symmetric around zero, then $F(-y) = 1 - F(y)$ and substituting that in $G(y)$ we get $G(y) = 2F(y) - 1$. Recall, that the p.d.f. of Y is given by $g(y) = G'(y)$. Hence, $g(y) = f(y) + f(-y)$ and this reduces to $2f(y)$ if the distribution is symmetric around zero. So that if $f(x) = e^{-x^2/2}/\sqrt{2\pi}$ for $-\infty < x < +\infty$ then $g(y) = 2f(y) = 2e^{-y^2/2}/\sqrt{2\pi}$ for $y \geq 0$.

Let us now find the distribution of $Z = X^2$, the square of a $N(0,1)$ random variable. Note that $dZ/dX = 2X$ which is positive when $X > 0$ and negative when $X < 0$. The *change of variable* method cannot be applied since $Z = X^2$ is not a monotonic transformation over the entire domain of X. However, using $Y = |X|$, we get $Z = Y^2 = (|X|)^2$ and $dZ/dY = 2Y$ which is always non-negative since Y is non-negative. In this case, the *change of variable* method states that the p.d.f. of Z is obtained from that of Y by substituting the inverse transformation $Y = \sqrt{Z}$ into the p.d.f. of Y and multiplying it by the absolute value of the derivative of the inverse transformation:

$$h(z) = g(\sqrt{z}) \cdot |\frac{dY}{dZ}| = \frac{2}{\sqrt{2\pi}} e^{-z/2} |\frac{1}{2\sqrt{z}}| = \frac{1}{\sqrt{2\pi}} z^{-1/2} e^{-z/2} \text{ for } z \geq 0$$

It is clear why this transformation will not work for X since $Z = X^2$ has two solutions for the inverse transformation, $X = \pm\sqrt{Z}$, whereas, there is one unique solution for $Y = \sqrt{Z}$ since it is non-negative. Using the results of problem 10, one can deduce that Z has a gamma distribution with $\alpha = 1/2$ and $\beta = 2$. This special Gamma density function is a χ^2 distribution with 1 degree of freedom. Hence, we have shown that the square of a $N(0,1)$ random variable has a χ^2_1 distribution.

Finally, if $X_1, \ldots, X_n$ are independently distributed then the distribution function of $Y = \sum_{i=1}^n X_i$ can be obtained from that of the X_i's using the *Moment Generating Function* (MGF) method:

$$\begin{aligned} M_Y(t) &= E(e^{Yt}) = E[e^{(\sum_{i=1}^n X_i)t}] = E(e^{X_1 t})E(e^{X_2 t})..E(e^{X_n t}) \\ &= M_{X_1}(t)M_{X_2}(t)..M_{X_n}(t) \end{aligned}$$

If in addition these X_i's are identically distributed, then $M_{X_i}(t) = M_X(t)$ for $i = 1, \ldots, n$ and

$$M_Y(t) = [M_X(t)]^n$$

For example, if $X_1, \ldots, X_n$ are IID Bernoulli (θ), then $M_{X_i}(t) = M_X(t) = \theta e^t + (1-\theta)$ for $i = 1, \ldots, n$. Hence the MGF of $Y = \sum_{i=1}^n X_i$ is given by

$$M_Y(t) = [M_X(t)]^n = [\theta e^t + (1-\theta)]^n$$

This can be easily shown to be the MGF of the Binomial distribution given in problem 14. This proves that the sum of n independent and identically distributed Bernoulli random variables with parameter θ is a Binomial random variable with same parameter θ.

Central Limit Theorem: If $X_1, \ldots, X_n$ are IID(μ, σ^2) from an unknown distribution, then $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is asymptotically distributed as $N(0,1)$.

Proof: We assume that the MGF of the X_i's exist and derive the MGF of Z. Next, we show that lim $M_Z(t)$ as $n \to \infty$ is $e^{1/2t^2}$ which is the MGF of $N(0,1)$ distribution. First, note that

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{Y - n\mu}{\sigma\sqrt{n}}$$

where $Y = \sum_{i=1}^{n} X_i$ with $M_Y(t) = [M_X(t)]^n$. Therefore,

$$\begin{aligned} M_Z(t) &= E(e^{Zt}) = E\left(e^{(Yt-n\mu t)/\sigma\sqrt{n}}\right) = e^{-n\mu t/\sigma\sqrt{n}} E\left(e^{Yt/\sigma\sqrt{n}}\right) \\ &= e^{-n\mu t/\sigma\sqrt{n}} M_Y(t/\sigma\sqrt{n}) = e^{-n\mu t/\sigma\sqrt{n}} [M_X(t/\sigma\sqrt{n})]^n \end{aligned}$$

Taking log of both sides we get

$$\log M_Z(t) = \frac{-n\mu t}{\sigma\sqrt{n}} + n\log[1 + \frac{t}{\sigma\sqrt{n}}E(X) + \frac{t^2}{2\sigma^2 n}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..]$$

Using the Taylor series expansion $\log(1+s) = s - \frac{s^2}{2} + \frac{s^3}{3} - ..$ we get

$$\begin{aligned} \log M_Z(t) &= -\frac{\sqrt{n}\mu}{\sigma}t + n\left\{\left[\mu\frac{t}{\sigma\sqrt{n}} + \frac{t^2}{2\sigma^2 n}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..\right]\right. \\ &\quad -\frac{1}{2}\left[\mu\frac{t}{\sigma\sqrt{n}} + \frac{t^2}{2\sigma^2 n}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..\right]^2 \\ &\quad \left.+\frac{1}{3}\left[\mu\frac{t}{\sigma\sqrt{n}} + \frac{t^2}{2\sigma^2\sqrt{n}}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..\right]^3 - ..\right\} \end{aligned}$$

Collecting powers of t, we get

$$\begin{aligned} \log M_Z(t) &= \left(-\frac{\sqrt{n}\mu}{\sigma} + \frac{\sqrt{n}\mu}{\sigma}\right)t + \left(\frac{(E(X^2)}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)t^2 \\ &\quad + \left(\frac{E(X^3)}{6\sigma^3\sqrt{n}} - \frac{1}{2}\cdot\frac{2\mu E(X^2)}{2\sigma^3\sqrt{n}} + \frac{1}{3}\frac{\mu^3}{\sigma^3\sqrt{n}}\right)t^3 + .. \end{aligned}$$

Therefore

$$\log M_Z(t) = \frac{1}{2}t^2 + \left(\frac{E(X^3)}{6} - \frac{\mu E(X^2)}{2} + \frac{\mu^3}{3}\right)\frac{t^3}{\sigma^3\sqrt{n}} + ..$$

note that the coefficient of t^3 is $1/\sqrt{n}$ times a constant. Therefore, this coefficient goes to zero as $n \to \infty$. Similarly, it can be shown that the coefficient of t^r is $1/\sqrt{n^{r-2}}$ times a constant for $r \geq 3$. Hence,

$$\lim_{n\to\infty} \log M_Z(t) = \frac{1}{2}t^2 \quad \text{and} \quad \lim_{n\to\infty} M_Z(t) = e^{\frac{1}{2}t^2}$$

which is the MGF of a standard normal distribution.

The Central Limit Theorem is a powerful tool for asymptotic inference. In real life we do not know what distribution we are sampling from, but as long as the sample drawn is random and we average (or sum) and standardize then as $n \to \infty$, the resulting standardized statistic has an asymptotic $N(0,1)$ distribution that can be used for inference.

Using a random number generator from say the uniform distribution on the computer, one can generate samples of size $n = 20, 30, 50$ from this distribution and show how the sampling distribution of the sum (or average) when it is standardized closely approximates the $N(0,1)$ distribution.

The real question for the applied researcher is how large n should be to invoke the Central Limit Theorem. This depends on the distribution we are drawing from. For a Bernoulli distribution, a larger n is needed the more asymmetric this distribution is i.e., if $\theta = 0.1$ rather than 0.5.

In fact, Figure 2.15 shows the Poisson distribution with mean = 15. This looks like a good approximation for a Normal distribution even though it is a discrete probability function. Problem 15 shows that the sum of n independent identically distributed Poisson random variables with parameter λ is a Poisson random variable with parameter $(n\lambda)$. This means that if $\lambda = 0.15$, an n of 100 will lead to the distribution of the sum being Poisson $(n\lambda = 15)$ and the Central Limit Theorem seems well approximated.

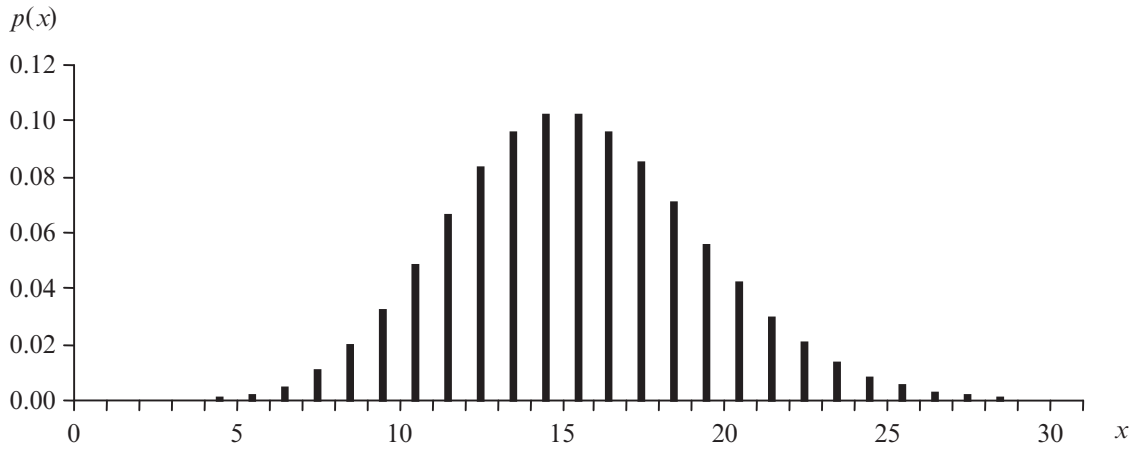


Figure 2.15 Poisson Probability Distribution, Mean = 15

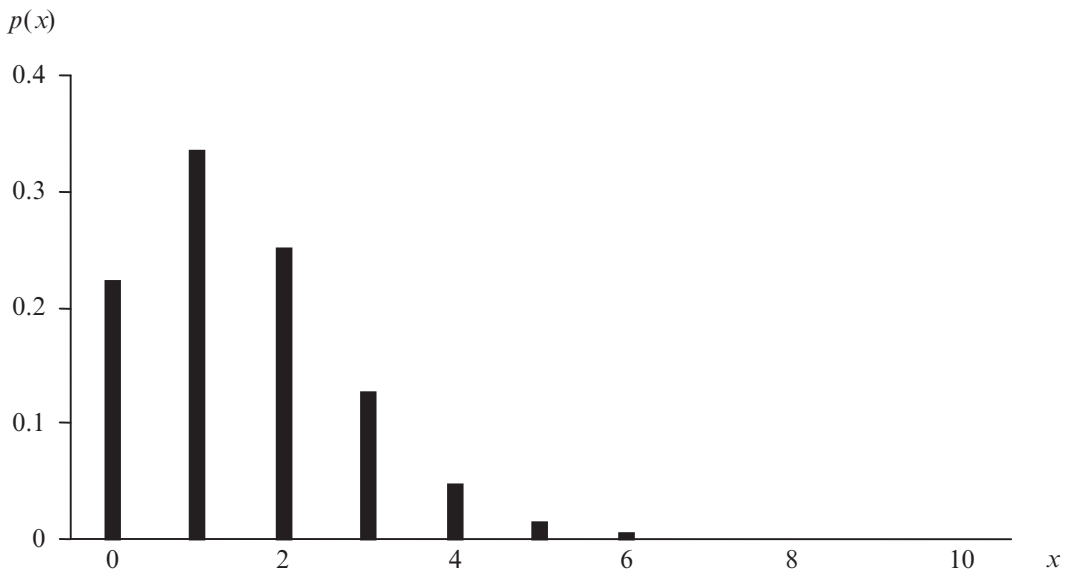


Figure 2.16 Poisson Probability Distribution, Mean = 1.5

However, if $\lambda = 0.015$, an n of 100 will lead to the distribution of the sum being Poisson ($n\lambda = 1.5$) which is given in Figure 2.16. This Poisson probability function is skewed and discrete and does not approximate well a normal density. This shows that one has to be careful in concluding that $n = 100$ is a large enough sample for the Central Limit Theorem to apply. We showed in this simple example that this depends on the distribution we are sampling from. This is true for Poisson ($\lambda = 0.15$) but not Poisson ($\lambda = 0.015$), see Joliffe (1995). The same idea can be illustrated with a skewed Bernoulli distribution.

Conditional Mean and Variance: Two random variables X and Y are bivariate Normal if they have the following joint distribution:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]\right\}$$

where $-\infty < x < +\infty$, $-\infty < y < +\infty$, $E(X) = \mu_X$, $E(Y) = \mu_Y$, $\text{var}(X) = \sigma_X^2$, $\text{var}(Y) = \sigma_Y^2$ and $\rho =$ correlation $(X, Y) = \text{cov}(X, Y)/\sigma_X \sigma_Y$. This joint density can be rewritten as

$$f(x,y) = \frac{1}{\sqrt{2\pi}\sigma_{_Y}\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y - \mu_Y - \rho\frac{\sigma_{_Y}}{\sigma_{_X}}(x-\mu_X)\right]^2\right\}$$
$$\cdot\frac{1}{\sqrt{2\pi}\sigma_X} \exp\left\{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2\right\} = f(y/x)f_1(x)$$

where $f_1(x)$ is the *marginal density* of X and $f(y/x)$ is the *conditional density* of Y given X. In this case, $X \sim N(\mu_X, \sigma_X^2)$ and Y/X is Normal with mean $E(Y/X) = \mu_Y + \rho\frac{\sigma_{_Y}}{\sigma_{_X}}(x - \mu_x)$ and variance given by $\text{var}(Y/X) = \sigma_Y^2(1-\rho^2)$.

By symmetry, the roles of X and Y can be interchanged and one can write $f(x,y) = f(x/y)\ f_2(y)$ where $f_2(y)$ is the marginal density of Y. In this case, $Y \sim N(\mu_Y, \sigma_Y^2)$ and X/Y is Normal with mean $E(X/Y) = \mu_X + \rho\frac{\sigma_X}{\sigma_{_Y}}(y - \mu_Y)$ and variance given by $\text{var}(X/Y) = \sigma_X^2(1-\rho^2)$. If $\rho = 0$, then $f(y/x) = f_2(y)$ and $f(x,y) = f_1(x)f_2(y)$ proving that X and Y are independent. Therefore, if $\text{cov}(X,Y) = 0$ *and* X and Y are bivariate Normal, then X and Y are independent. In general, $\text{cov}(X,Y) = 0$ alone does not necessarily imply independence, see problem 3.

One important and useful property is the *law of iterated expectations*. This says that the expectation of any function of X and Y say $h(X,Y)$ can be obtained as follows:

$$E[h(X,Y)] = E_X E_{Y/X}[h(X,Y)]$$

where the subscript Y/X on E means the conditional expectation of Y given that X is treated as a constant. The next expectation E_X treats X as a random variable. The proof is simple.

$$E[h(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(x,y)f(x,y)dxdy$$

where $f(x,y)$ is the joint density of X and Y. But $f(x,y)$ can be written as $f(y/x)f_1(x)$, hence $E[h(X,Y)] = \int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} h(x,y)f(y/x)dy\right] f_1(x)dx = E_X E_{Y/X}[h(X,Y)]$.

Example: This law of iterated expectation can be used to show that for the bivariate Normal density, the parameter ρ is indeed the correlation coefficient of X and Y. In fact, let $h(X,Y) = XY$, then

$$\begin{aligned} E(XY) &= E_X E_{Y/X}(XY/X) = E_X X E(Y/X) = E_X X[\mu_Y + \rho\frac{\sigma_{_Y}}{\sigma_{_X}}(X - \mu_X)] \\ &= \mu_X\mu_Y + \rho\frac{\sigma_{_Y}}{\sigma_{_X}}\sigma_X^2 = \mu_X\mu_Y + \rho\sigma_Y\sigma_X \end{aligned}$$

Rearranging terms, one gets $\rho = [E(XY) - \mu_X\mu_Y]/\sigma_X\sigma_Y = \sigma_{XY}/\sigma_X\sigma_Y$ as required.

Another useful result pertains to the unconditional variance of $h(X,Y)$ being the sum of the mean of the conditional variance and the variance of the conditional mean:

$$\text{var}(h(X,Y)) = E_X\text{var}_{Y/X}[h(X,Y)] + \text{var}_X E_{Y/X}[h(X,Y)]$$

Proof: We will write $h(X,Y)$ as h to simplify the presentation

$$\text{var}_{Y/X}(h) = E_{Y/X}(h^2) - [E_{Y/X}(h)]^2$$

and taking expectations with respect to X yields $E_X\text{var}_{Y/X}(h) = E_X E_{Y/X}(h^2) - E_X[E_{Y/X}(h)]^2$
$= E(h^2) - E_X[E_{Y/X}(h)]^2$.

Also, $\text{var}_X E_{Y/X}(h) = E_X[E_{Y/X}(h)]^2 - (E_X[E_{Y/X}(h)])^2 = E_X[E_{Y/X}(h)]^2 - [E(h)]^2$ adding these two terms yields

$$E(h^2) - [E(h)]^2 = \text{var}(h).$$