Econ4130, Statistics 2
Tore Schweder, 6 October 2003

Lecture note to Rice chapter 5: Limit theorems

Section 4.5 in Rice on moment generating functions is a technical prelude to the Central limit theorem: that the empirical mean $\overline{X}$ based on n independent observations is approximately normally distributed, and that the approximation is better the larger the sample size n, and also the closer the common distribution of the observations is to the normal distribution. The essential facts of the moment generating function are:

$$\begin{aligned} \text{Definition} &: \quad M_X(t) = E\left[e^{tX}\right] \\ \text{Uniqueness (Property A, p143)} &: \quad M_X(t) = M_Y(t) \text{ for } t \in (-a,a) \Rightarrow X \overset{D}{=} Y \\ \text{Moments (Property B, p144)} &: \quad M_X^{(r)}(0) = E\left[X^r\right] \\ \text{Continuity (Theorem A, p167)} &: \quad M_{X_n}(t) \to M_X(t) \Rightarrow X_n \overset{D}{\to} X \end{aligned}$$

The notation here is

$$\begin{aligned} \text{Definition} &: \quad X \overset{D}{=} Y \text{ when } X \text{ and } Y \text{ have the same distribution,} \\ \text{Definition} &: \quad X_n \overset{D}{\to} X \text{ when } F_{X_n}(x) \to F_X(x) \text{ for all continuity points } x \text{ of } F_X, \end{aligned}$$

and $X_n \overset{D}{\to} X$ is said to be convergence in distribution.

The following version of the Central limit theorem is saying the same as Theorem B, p169, but it is formulated in terms of the mean of standardized observations rather than a sum of unstandardized ones.

Theorem 1 (Central limit theorem) *Let $X_1, X_2, \cdots$ be independent and identically distributed (i.i.d.) with zero expectation and unit variance. The empirical mean based on a sample of size n, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ has approximately a normal distribution with zero mean and variance $1/n$. That is,*

$$\sqrt{n}\overline{X}_n \overset{D}{\to} Z$$

where Z has a standard normal distribution.

Proof. *Let $M(t) = E\left[e^{tX_i}\right]$ be the moment generating function, which is continuous and twice differentiable around 0 since the expectation and the variance exist. $M_X'(0) = EX_i = 0$ and $M_X''(0) = EX^2 = var(X_i) + (EX)^2 = 1 + 0 = 1$. Since $t\sqrt{n}\overline{X}_n = \frac{t}{\sqrt{n}}\sum_{i=1}^n X_i = \frac{t}{\sqrt{n}}S_n$, by definition*

$$M_{\sqrt{n}\overline{X}_n}(t) = M_{S_n}\left(\frac{t}{\sqrt{n}}\right).$$

By independence $M_{S_n}(r) = E\left[e^{rS_n}\right] = E\left[\prod_{i=1}^n e^{rX_i}\right] = \prod_{i=1}^n E\left[e^{rX_i}\right] = M(r)^n$. Thus,

$$M_{\sqrt{n}\overline{X}_n}(t) = M\left(\frac{t}{\sqrt{n}}\right)^n.$$

As $n \to \infty$, $\frac{t}{\sqrt{n}} \to 0$. We therefore have by second order Taylor expansion

$$M\left(\frac{t}{\sqrt{n}}\right) = M(0) + \frac{t}{\sqrt{n}}M^{'}(0) + \frac{1}{2}\frac{t^2}{n}M^{''}(0) + \varepsilon_n = 1 + \frac{1}{2}\frac{t^2}{n} + \varepsilon_n.$$

Since ε_n can be shown to tend faster to zero than $\frac{1}{2}\frac{t^2}{n}$, we have

$$M\left(\frac{t}{\sqrt{n}}\right)^n \to e^{\frac{1}{2}t^2} = M_Z(t).$$

By the continuity property and the uniqueness property of the moment generating function, we conclude that $\sqrt{n}\overline{X}_n \xrightarrow{D} Z$. ■

The proof given here is very much the proof in Rice. It is included since it is so beautiful and so important. It was Laplace who around 1800 played with these concepts and realized that $EX_i = 0 \Rightarrow M_X^{'}(0) = 0$, and therefore (i) that two terms are needed in the Taylor expansion, and (ii) it is $\sqrt{n}\overline{X}_n$ and not $\overline{X}_n$ that has moment generating function approximately $\left(1 + \frac{1}{2}\frac{t^2}{n}\right)^n$, which is needed to get the desired result.

If Y_i are i.i.d. with expectation μ and standard deviation σ, the standardized variables $X_i = (Y_i - \mu)/\sigma$ satisfies the Central limit theorem. Therefore $\sqrt{n}\frac{\overline{Y}_n - \mu}{\sigma} \xrightarrow{D} Z$, which means that $\overline{Y}_n$ is approximately normally distributed with expectation μ and variance σ^2/n.

Rice does not mention the simpler concept convergence in probability to a constant c, denoted $\xrightarrow{P} c$. The basic properties are

Definition:	$X_n \xrightarrow{P} c \Leftrightarrow P(\lvert X_n - c\rvert > \varepsilon) \to 0$ for all $\varepsilon > 0$
Continuity:	$X_n \xrightarrow{P} c \Rightarrow g(X_n) \xrightarrow{P} g(c)$ when g is continuous at c.

Try to prove the continuity property of convergence in probability.

An important question for an estimator is whether it is consistent, that is, whether it converge in probability to the correct value whatever that value is. Let $\widehat{\theta}_n$ be an estimator for the parameter θ based on a sample of size n.

Definition: $\widehat{\theta}_n$ is consistent whenever $\widehat{\theta}_n \xrightarrow{P} \theta$ for all possible values of θ.

Theorem 2 (Law of large numbers) *Let $X_1, X_2, \cdots$ be i.i.d. with expectation μ. Then $\overline{X}_n \xrightarrow{P} \mu$.*

Proof. *Parallel to the proof of the Central limit theorem, $M_{\overline{X}_n}(t) = M\left(\frac{t}{n}\right)^n = \left(M(0) + \frac{t}{n}M^{'}(0) + \varepsilon_n\right)^n = \left(1 + \frac{t}{n}\mu + \varepsilon_n\right)^n \to e^{t\mu} = M_Y(t)$ where Y is a degenerate random variable with all its mass at μ, $P(Y = \mu) = 1$. By continuity and uniqueness of moment generating function, we have proved the theorem.* ■

Rice could also have included Slutsky's lemma. This lemma is extensively used in statistics and econometrics in conjunction with the Law of large numbers, convergence in probability and the Central limit theorem, to prove that certain statistics are approximately normally distributed in large samples. The example at the end of this note illustrates its use.

Lemma 3 (Slutsky) *Let $\{A_i\}$, $\{B_i\}$ and $\{X_i\}$ be three sequences of random variables (or constants). If $A_n \xrightarrow{P} a$, $B_n \xrightarrow{P} b$ and $X_n \xrightarrow{D} X$, then $Y_n = A_n + B_n X_n \xrightarrow{D} a + bX$*

The proof of this lemma requires an ε and δ argument that is tedious, but not terribly exciting. It is excluded here. That A_n and B_n converge in probability means that their probability mass pile up closer and closer to a and b respectively. The distribution of Y_n must therefore nearly be the same as that of $a + bX_n$, and since X_n converge in distribution, Slutsky's conclusion is most reasonable.

Example 4 *Consider a situation where unemployed have a constant rate of being employed. This is perhaps not very realistic because of individual differences between the unemployed. Some are more attractive than others, and some are less active in their job search. Disregarding this individual heterogeneity, and assuming stability in the labour marked, the assumption of a constant rate θ is at least a starting point. Let X be the waiting time until employment for a person that just lost employment. That the employment rate is constant means that the person has the same conditional probability of being employed in the waiting time interval $< x, x+t]$ given that the person was still unemployed after having waited x units of time, regardless of x. For small t, this conditional probability is θt to make the employment rate $\theta : P(X \le x + t | X > x)/t \to \theta$. You can show that X has this property if it is exponentially distributed with density*

$$f(x) = \theta e^{-\theta x} \quad x > 0.$$

We know that $E(X) = 1/\theta$ and $var(X) = 1/\theta^2$. The moment generating function of X is $M(t) = \int_0^\infty e^{tx}\theta e^{-\theta x} dx = \theta/(\theta - t)$ for $t < \theta$. Thus, $M^{'}(0) = 1/\theta$ and $M^{''}(0) = 2/\theta^2 = E(X^2)$ making $var(X) = 1/\theta^2$. Assume now that you shall observe a large sample of unemployment spells, $X_1, X_2, \cdots, X_n$ for the purpose of estimating $\theta = 1/E(X)$. From this last equality, $\widehat{\theta} = 1/\overline{X}$ is a natural estimator. What are the statistical properties of $\widehat{\theta}$? By the Law of large numbers, $\overline{X}_n \xrightarrow{P} 1/\theta$. Since the function $g(x) = 1/x$ is continuous at any $x > 0$, $\widehat{\theta}_n = 1/\overline{X}_n \xrightarrow{P} g(1/\theta) = \theta$ whatever $\theta > 0$ is. The estimator is consistent. Is it also approximately normally distributed? And what is its approximate expectation and variance for large n? From the Central limit theorem, $\sqrt{n}(\theta \overline{X}_n - 1) \xrightarrow{D} Z$, the standard normal. Indeed, $\theta X_i - 1$ have expectation 0 and variance 1. Therefore, by Slutsky's lemma,

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) = -\frac{1}{\overline{X}_n}\sqrt{n}(\theta\overline{X}_n - 1) \xrightarrow{D} -\theta Z.$$

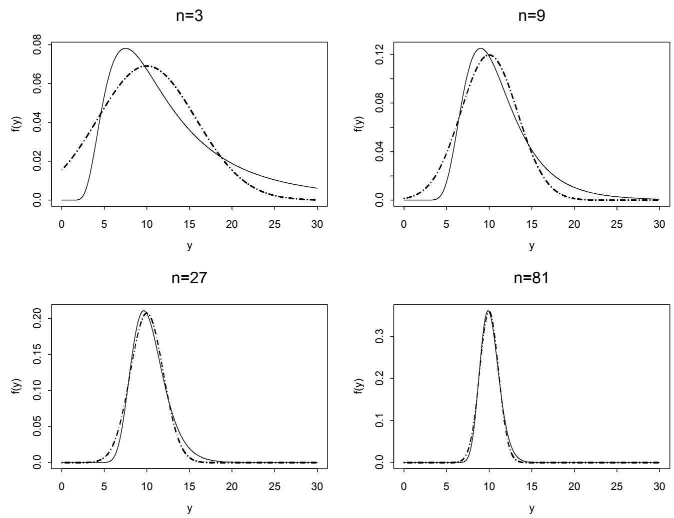

Figure 1: Density of $Y = \widehat{\theta}_n$ for $\theta = 10$ and for various values of n, and approximating normal densities (dashed lines).

The conclusion is thus that $\widehat{\theta}_n$ is approximately normally distributed with expectation θ and with standard deviation $\theta/\sqrt{n}$. The moment generating function tells us that

$$M_{\overline{X}_n}(t) = \left(\frac{n\theta}{n\theta - t}\right)^n.$$

Why? This is the moment generating function of the gamma distribution with shape parameter $\alpha = n$ and scale parameter $\lambda = n\theta$. The probability density of $\widehat{\theta}_n = Y$ is thus

$$f(y) = \frac{(n\theta)^n}{\Gamma(n)} y^{-n-2} e^{-\frac{n\theta}{y}} \quad y > 0.$$

This density and the density of the approximating normal distribution is plotted for $\theta = 10$ and n=3,9,27 and 81 in the figure. Note the change in the vertical scale in the four diagrams.