ECON 4130
Revised Sept. 2018

Lecture Notes to Rice Chapter 5

By H. Goldstein

1.1

Chapter 5 gives an introduction to probabilistic approximation methods, but is insufficient for the needs of an adequate study of econometrics. The common non-linear nature of economic models often requires approximation methods for a tractable empirical analysis. An excellent summary of asymptotic (approximation) techniques can, for example, be found in chapter 4 in W.H. Greene's book, "Econometric Analysis", Prentice Hall (any edition). With the tool kit in that book you can handle a large number of approximation problems common in econometrics. This course does not go all the way to Greene's summary, but should represent a good basis. The step from this course up to Greene's level should not be very large.

There are many probabilistic convergence concepts available, of which two, *convergence in probability* and *convergence in distribution* are discussed or implied in Rice.

Def. 1 Convergence in Probability.

> *Let* $Y_1, Y_2, \ldots, Y_n, \ldots$ *be a sequence of r.v.'s. Then* Y_n *converges in probability to a constant, c, as* $n \to \infty$ *(written shortly* $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$ *or* $\operatorname{plim}_{n\to\infty} Y_n = c$*), if , for any* $\varepsilon > 0$, $P(|Y_n - c| > \varepsilon) \to 0$ *as* $n \to \infty$.
>
> *Or equivalently:* $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$ *if, for any* $\varepsilon > 0$, $P(|Y_n - c| \le \varepsilon) \to 1$ *as* $n \to \infty$.

Example 1: If $X_1, X_2, \ldots$ are *iid* with $E(X_i) = \mu$ and $\operatorname{Var}(X_i) = \sigma^2$, then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \underset{n\to\infty}{\overset{P}{\to}} \mu$ (one of the laws of large numbers proven by Chebyshev's inequality).

Example 3 below shows that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \underset{n\to\infty}{\overset{P}{\to}} \sigma^2$, and also $S = \sqrt{S^2} \overset{P}{\to} \sigma$ (proven by the continuity properties of limits in probability described below). This shows that $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = S^2$, and $\hat{\sigma} = S$ are all *consistent* estimators as defined in the next paragraph..

Definition of consistency. In general, suppose that θ is an unknown parameter in a model and $\hat{\theta}$ an estimator for θ depending on *n* observations. If $\hat{\theta} \xrightarrow[n\to\infty]{P} \theta$, we say that $\hat{\theta}$ is a *consistent* estimator for θ.

This is a rather weak property, but is usually considered a minimum requirement for the behavior of an estimator when the number of observations grows large. Even if it is a weak property, it turns out to be a very useful (and much used) concept in econometric handling of approximation problems.

> [**Note on the law of large numbers.** In the lectures we gave a simple proof of the law of large numbers for sample means based on Chebyshev's inequality. That proof assumes that the variance, $\text{var}(X_i) = \sigma^2$, exists. It can be proven, however, that this assumption is not necessary. Thus: If $X_1, X_2, \ldots$ are *iid* with $E(X_i) = \mu$, then $\bar{X} \xrightarrow[n\to\infty]{P} \mu$ (without any assumptions on the variance). This is a classical result in probability theory which requires a somewhat deeper proof.]

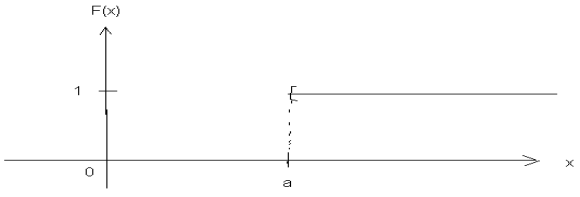
1.2 Trivial distributions.

It is sometimes convenient to interpret constants as special r.v.'s. Let a be any constant (a real number). We may interpret *a* as a random variable by introducing the r.v., *X*, by $P(X = a) = 1$. Hence *X* can only take one value, *a*. The probability mass function is then given by $p(a) = P(X = a) = 1$. By the definition of expectation and variance (that does not exclude this special case), we have (check formally!), $E(X) = a$ and $\text{var}(X) = 0$.

The cdf of *X* becomes

$$(1) \qquad F(x) = P(X \le x) = \begin{cases} 0 & \text{for } x < a \\ 1 & \text{for } x \ge a \end{cases} \qquad \text{(see figure 1)}$$

Figure 1



We may call this distribution *the trivial distribution at a*.
Note that $F(x)$ is continuous everywhere except at $x = a$.

(2) The moment generating function (mgf) for *X* with the trivial distribution at *a*, is $M(t) = e^{ta}$.

(i.e. $M(t) = Ee^{tX} = e^{ta}P(X = a) = e^{ta}$).

Let $a_1, a_2, \ldots, a_n, \ldots$ be a sequence of constants converging to *a* as $n \to \infty$ (see appendix 1 (A1) for the concept of a sequence). This means (slightly more precise than presented in Sydsæter I): F*or any fixed* $\varepsilon > 0$*, there is a number N such that* $|a_n - a| \le \varepsilon$ *for every* $n \ge N$. From this definition it follows that convergence of sequences in the usual sense can be considered as a special case of convergence in probability.

(3) *If* $a_n \underset{n\to\infty}{\to} a$*, then* $a_n \overset{P}{\underset{n\to\infty}{\to}} a$ *(where the* a_n*'s are interpreted as r.v.'s)*

Proof: Let $\varepsilon > 0$ be arbitrarily small. We need to show that $P(|a_n - a| \le \varepsilon) \underset{n\to\infty}{\to} 1$. But this probability must be either 0 or 1 according to if $|a_n - a| \le \varepsilon$ is false or true (since $a_n, a,$ and ε are constants and therefore fixed and not subject to random variation). Hence, choosing N such that $|a_n - a| \le \varepsilon$ for all $n \ge N$, we have

$$P(|a_n - a| \le \varepsilon) = \begin{cases} 1 & \text{if } |a_n - a| \le \varepsilon \text{ is true, which it is for all } n \ge N \\ 0 & \text{if } |a_n - a| \le \varepsilon \text{ is false} \end{cases}$$

This shows that $P(|a_n - a| \le \varepsilon) \underset{n\to\infty}{\to} 1$ since the probability is 1 for all *n* large enough. Q.E.D[1].

[1] Q.E.D. means "**end of proof**". It is short for the latin expression: *quod erat demonstrandum*.

1.3 The continuity property of probability limits.

(See appendix 1 (A2) for some useful facts about continuous functions.)

Theorem 1

(4)

Let $X_n, Y_n, \quad n = 1,2,\ldots$ *be two sequences of r.v.'s such that* $X_n \xrightarrow[n\to\infty]{P} c$ *and* $Y_n \xrightarrow[n\to\infty]{P} d$ *. Let* $g(x)$ *be continuous at* $x = c$ *and* $h(x, y)$ *be continuous at* $x = c$ *and* $y = d$*. Then*

$$g(X_n) \xrightarrow[n\to\infty]{P} g(c) \quad \text{and} \quad h(X_n, Y_n) \xrightarrow[n\to\infty]{P} h(c, d)$$

(This is also true when h has more than two arguments.)

[A proof for those interested (optional reading) is given in **appendix 2**.]

Example 2. Suppose that $X_n \xrightarrow[n\to\infty]{P} c$. Then also $Z_n = X_n\left(1-\frac{1}{n}\right) \xrightarrow[n\to\infty]{P} c$. Here we use that $h(x, y) = xy$ is continuous (see appendix 1 (A2)), and that $Y_n = 1-\frac{1}{n} \xrightarrow[n\to\infty]{P} 1$ because of (3).

[I.e. $h(X_n, Y_n) = X_n Y_n \xrightarrow{P} h(c,1) = c\cdot 1 = c$ since $X_n \xrightarrow{P} c$ and $Y_n \xrightarrow{P} 1$.]

Example 3. Suppose that $X_1, X_2, \ldots$ are *iid* with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. Then $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \xrightarrow[n\to\infty]{P} \sigma^2$ (i.e. S^2 is consistent for σ^2).

Reason (short argument): It follows from theorem 1 that

$$S^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} - (\bar{X})^2\right] \xrightarrow[n\to\infty]{P} 1\cdot\left[\mu^2 + \sigma^2 - \mu^2\right] = \sigma^2$$

Explanation: By the law of large numbers (see further details below), $\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} \xrightarrow{P} E(X_i^{\,2}) = \mu^2 + \sigma^2$, and $\bar{X} \xrightarrow{P} \mu$. Then, since $h(x, y) = x - y^2$ is continuous, we get from theorem 1 that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} - (\bar{X})^2 = h\left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2},\ \bar{X}\right) \xrightarrow[n\to\infty]{P} h(E(X_i^{\,2}), \mu) = \mu^2 + \sigma^2 - \mu^2 = \sigma^2 .$$

Finally, multiplying by $\frac{n}{n-1}$ (which converges to 1 as $n \to \infty$) and the same argument as in example 2, we get $S^2 \xrightarrow{P} \sigma^2$.

From this we also obtain that $S = \sqrt{S^2} \xrightarrow{P} \sigma$ since $g(x) = \sqrt{x}$ is continuous (theorem 1 again).

[*Some more details*: Put $Z_i = X_i^2$. Since $X_1, X_2, \ldots$ are *iid*, then $Z_1, Z_2, \ldots$ are *iid* as well with $E(Z_i) = E(X_i^2) = \mu^2 + \sigma^2$. Then, by the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \bar{Z} \xrightarrow[n\to\infty]{P} E(Z_i) = \mu^2 + \sigma^2$. (Note that, because of the note to example 1, we do not have to bother about the variance of Z_i.)]

Exercise 1. Show that the sample correlation, $r = \frac{S_{XY}}{S_X S_Y}$ is a consistent estimator for the population correlation, $\rho = \text{corre}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$, based on a *iid* random sample of pairs, $(X_1,Y_1), (X_2,Y_2), \ldots, (X_n,Y_n)$ (meaning that the *n* pairs are independent and have all the same joint distribution).

Hint: To prove the consistency of the sample covariance, write $S_{XY} = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}\right]$. Note that $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i$ is a mean $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, where $Z_i = X_i Y_i, \quad i = 1, 2, \ldots$ are *iid* rv's. Hence, $\bar{Z}$ converges in probability to $E(Z_i) = E(X_i Y_i) = \text{cov}(X_i, Y_i) + \mu_X \mu_Y$. Then use the continuity of the function $g_1(x, y, z) = x - y \cdot z$, and finally that $g_2(x, y, z) = \frac{z}{\sqrt{x}\sqrt{y}}$ also is continous.

Note that *r* is *not* unbiased as an estimator of ρ. On the other hand, the fact that it is consistent, justifies its use for large *n*. Simulation studies and other investigations show in addition that it behaves reasonably well even in smaller samples, and is therefore the most common estimator of ρ.

1.4 Convergence in distribution

In the introductory statistics course, the following version of the central limit theorem (CLT) is presented:

Let $X_1, X_2, \ldots$ be *iid* with $\mathrm{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$ (implying that $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n}$). Then, for large *n* ($n \geq 30$ usually considered sufficient), we have

$$Z_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \overset{\text{approximately}}{\sim} N(0,1) \quad (\text{"}\sim\text{" means "is distributed as"})$$

This statement is somewhat un-precise. What we mean is that "Z_n converges in distribution to *Z*, where *Z* ~ N(0, 1), as $n \to \infty$ ". (We write this shortly, $Z_n \underset{n\to\infty}{\overset{D}{\to}} Z$, or simply $Z_n \overset{D}{\to} Z$). The formal mathematical definition, given in Rice, is:

Def. 2 (Convergence in distribution)

> *Let $Y_1, Y_2, \ldots$ be a sequence of r.v..'s with cdf's, $F_n(y) = P(Y_n \leq y)$, and Y a r.v. with cdf $F(y) = P(Y \leq y)$. We say that $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$ if $F_n(y) \underset{n\to\infty}{\to} F(y)$ for every y where the limit cdf, $F(y)$, is continuous.*
>
> (Then, for large *n*, $Y_n \overset{\text{approx.}}{\sim} F(y)$)

This means: If the limit cdf, $F(y)$, is continuous at $y = a$ and $y = b$, then
$P(a < Y_n \leq b) = F_n(b) - F_n(a) \underset{n\to\infty}{\to} F(b) - F(a) = P(a < Y \leq b)$
Hence, $P(a < Y_n \leq b) \approx P(a < Y \leq b)$ for large *n*.

The importance of this property follows from the fact that we quite often find ourselves in a situation where Y_n (being e.g. a complicated estimator) has a very complicated distribution while the limit distribution of *Y* is quite simple (often normal). Hence, for large *n*, we may be able to replace complicated probability statements about Y_n with simple probability statements about *Y*.

Note that, if the limit distribution is *N*(0,1) (which is often the case), then the limit cdf (usually written $\Phi(x) = P(Z \leq x)$, where $Z \sim N(0,1)$) is continuous for all *x*.

Another useful technical comment is that *convergence in probability* can be interpreted as a special case of *convergence in distribution* by the following lemma:

(5)

$Y_n \xrightarrow[n\to\infty]{P} c$ is equivalent to $Y_n \xrightarrow[n\to\infty]{D} Y$ where Y is the trivial r.v. at c (i.e. $P(Y=c)=1$) with the trivial cdf as in (1)*. (The last statement we may simply write $Y_n \xrightarrow[n\to\infty]{D} c$.)*

[For those interested, a proof is written out in appendix 2.]

1.5 Determination of limit distributions

It turns out difficult (usually) to use the definition of limit in distribution directly to derive a limit distribution. Therefore, a number of techniques and tools have been developed for this purpose in the literature. One important tool is by means of moment generating functions (mgf's) formulated as theorem A in Rice, chapter 5, and cited below in theorem 2. (An even more important tool is by means of so-called *characteristic functions,* (see Rice at the end of section 4.5), which requires complex analysis and is omitted here.)

Theorem 2 (Theorem A in Rice, chapter 5)

Let $Y_n, \quad n=1,2,\ldots$ be a sequence of r.v.'s with cdf's, $Y_n \sim F_n(y)=P(Y_n \le y)$. Suppose that the mgf's, $M_n(t)=Ee^{tY_n}$, exist for all n. Let Y be a r.v. with cdf, $F(y)$ and mgf $M(t)=Ee^{tY}$, and assume that $M_n(t) \xrightarrow[n\to\infty]{} M(t)$ for all t in an open interval that contains 0. Then $Y_n \xrightarrow[n\to\infty]{D} Y$ (i.e. $F_n(y) \xrightarrow[n\to\infty]{} F(y)$ for all y where $F(y)$ is continuous).

Note that if $F(y)$ is the cdf of a normal distribution (which is most often the case), then $F(y)$ is continuous for all y. So, in that case, $F_n(y) \xrightarrow[n\to\infty]{} F(y)$ for all *y*, and $P(a<Y_n \le b) \approx P(a<Y \le b)$ for all *a* and *b* when *n* is large.

Example 4 (example A in Rice, section 5.3)

We simplify the argument in Rice by using l'Hôpital's rule instead of his series argument.
Let $X_n \sim \text{pois}(\lambda_n),\ n = 1,2,\ldots$ where $\lambda_1, \lambda_2, \ldots$ is a sequence of numbers (see appendix 1 (A1)) such that $\lambda_n \underset{n\to\infty}{\to} \infty$. Then, since X_n is poisson distributed, we have $\mathrm{E}(X_n) = \mathrm{var}(X_n) = \lambda_n$. We will show that the standardized

$$Z_n = \frac{X_n - \mathrm{E}(X_n)}{\sqrt{\mathrm{var}(X_n)}} = \frac{X_n - \lambda_n}{\sqrt{\lambda_n}} = \frac{1}{\sqrt{\lambda_n}} X_n - \sqrt{\lambda_n}$$

converges in distribution to $Z \sim N(0,1)$, which follows if we can show that the *mgf* of Z_n converges to the *mgf* of $Z \sim N(0, 1)$, i.e. $M_Z(t) = e^{t^2/2}$. The mgf of X_n is (see Rice, section 4.5, example A):

$$M_{X_n}(t) = e^{\lambda_n (e^t - 1)}$$

We have from before that, if *X* and *Y* are r.v.'s such that $Y = a + bX$, the *mgf* of Y is, $M_Y(t) = e^{at} M_X(bt)$. Hence

$$M_{Z_n}(t) = e^{-t\sqrt{\lambda_n}} M_{X_n}\left(\frac{1}{\sqrt{\lambda_n}} t\right) = e^{-t\sqrt{\lambda_n}} \cdot e^{\lambda_n (e^{t/\sqrt{\lambda_n}} - 1)} \qquad \text{or}$$

$$\ln(M_{Z_n}(t)) = -t\sqrt{\lambda_n} + \lambda_n \left(e^{t/\sqrt{\lambda_n}} - 1\right)$$

Put $x = \frac{1}{\sqrt{\lambda_n}}$. Since $\lambda_n \underset{n\to\infty}{\to} \infty$, we have $x \underset{n\to\infty}{\to} 0$. From l'Hôpital's rule we get

$$\ln(M_{Z_n}(t)) = -\frac{t}{x} + \frac{1}{x^2}\left(e^{xt} - 1\right) = \frac{e^{xt} - 1 - xt}{x^2} \underset{x\to 0}{\to} \lim_{x\to 0} \frac{te^{xt} - t}{2x} = \lim_{x\to 0} \frac{t^2 e^{tx}}{2} = \frac{t^2}{2}$$

Since e^x is a continuous function of *x*, $M_{Z_n}(t) \underset{n\to\infty}{\to} e^{t^2/2}$, implying $Z_n \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$.
(*End of example.*)

We will now repeat Rice's proof of the central limit theorem (CLT) supplied with some details. Note that this proof can be taken as **optional reading,** which means that a proper understanding of the proof is not required for exam purposes. However, the result itself including the more practical version given in the corollary in (6) below, must be understood properly.

Theorem 3 (CLT, theorem B in Rice, section 5.3)

Let $X_1, X_2, \ldots$ *be a sequence of iid r.v.'s with* $\mathrm{E}(X_i) = 0$ *and* $\mathrm{var}(X_i) = \sigma^2$.

Let $S_n = \sum_{i=1}^{n} X_i$ *(implying* $\mathrm{E}(S_n) = 0$ *and* $\mathrm{var}(S_n) = n\sigma^2$*). Then*

$$\frac{S_n}{\sqrt{\mathrm{var}(S_n)}} = \frac{S_n}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{D} Z \sim N(0,1) \quad \textit{(or} \quad P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) \underset{n\to\infty}{\to} \Phi(x) \textit{ for all } x \textit{ since}$$

$\Phi(x) = P(Z \le x)$ *is continuous everywhere).*

[**Note.** The proof is only given here for the special case that the *mgf* of X_j, $M(t) = \mathrm{E}(e^{tX_j})$, exists in an open interval containing 0, which is not always the case (see the note to (A5) in appendix 1). The proof for the general case is almost identical to the given one, but based instead on characteristic functions (defined by $g(t) = \mathrm{E}(e^{itX_j})$ where *i* is the complex number, $\sqrt{-1}$). Characteristic functions exist for every probability distribution. Such a proof, however, requires some knowledge of complex analysis, and is omitted here.]

Proof (optional reading):

Assume that the common *mgf* of $X_1, X_2, \ldots$, $M(t) = \mathrm{E}(e^{tX_i})$, exists in an open interval, (a,b), where $a < 0 < b$. Then, according to (A7) in appendix 1, $M(t)$, has continuous derivatives of all orders in (a,b).

Since $X_1, X_2, \ldots$ are independent and identically distributed, the *mgf* of S_n is

$$M_{S_n}(t) = \mathrm{E}\left(e^{t\sum_{i=1}^{n} X_i}\right) = \mathrm{E}\left(e^{tX_1}e^{tX_2}\cdots e^{tX_n}\right) = \mathrm{E}\left(e^{tX_1}\right)\mathrm{E}\left(e^{tX_2}\right)\cdots\mathrm{E}\left(e^{tX_n}\right) = M(t)^n$$

Putting $Z_n = \dfrac{S_n}{\sigma\sqrt{n}}$, we obtain the *mgf*, $M_{Z_n}(t) = M\left(\dfrac{t}{\sigma\sqrt{n}}\right)^n$

Applying Taylor's formula (see (A4) in appendix 1) to $M(t)$, we have

$M(t) = M(0) + tM'(0) + \frac{t^2}{2}M''(0) + \frac{t^3}{3!}M'''(c)$ where c is somewhere between 0 and t. We have $M(0) = \mathrm{E}(e^{0X_i}) = 1$, $M'(0) = \mathrm{E}(X_i) = 0$, and $M''(0) = \mathrm{E}(X_i^2) = \sigma^2$. Hence

$$M(t) = 1 + \frac{t^2}{2}\sigma^2 + \frac{t^3}{6}M'''(c)$$

Substituting into $M_{Z_n}(t)$, we obtain

$$M_{Z_n}(t) = M\left(\frac{t}{\sigma\sqrt{n}}\right)^n = \left[1 + \frac{\left(\frac{t}{\sigma\sqrt{n}}\right)^2}{2}\sigma^2 + \frac{\left(\frac{t}{\sigma\sqrt{n}}\right)^3}{6}M'''(c_n)\right]^n$$

or

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + R_n\right]^n \quad \text{where} \quad R_n = \frac{t^3}{6\sigma^3 n^{\frac{3}{2}}}M'''(c_n), \text{ and } c_n \text{ lies between}$$

0 and $\frac{t}{\sigma\sqrt{n}}$.

We will now prove that $n \cdot R_n \underset{n\to\infty}{\to} 0$, i.e. $n \cdot R_n = \frac{t^3}{6\sigma^3\sqrt{n}}M'''(c_n) \underset{n\to\infty}{\to} 0$

Since c_n lies between 0 and $\frac{t}{\sigma\sqrt{n}}$, and $\frac{t}{\sigma\sqrt{n}} \underset{n\to\infty}{\to} 0$, we must have that $c_n \to 0$. Therefore, $M'''(c_n) \to M'''(0)$ since $M'''(t)$ is continuous in 0 (see (A7) in appendix 1). Hence, $M'''(c_n)$ is bounded, and $M'''(c_n)/\sqrt{n} \to 0$, which proves that $n \cdot R_n \underset{n\to\infty}{\to} 0$.

We finally get $M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + R_n\right]^n = \left[1 + \frac{a_n}{n}\right]^n$ where $a_n = \frac{t^2}{2} + n \cdot R_n \to \frac{t^2}{2}$

Thus, using (A6) in appendix 1, we get $M_{Z_n}(t) \to e^{t^2/2}$, which is the *mgf* of $N(0,1)$. Property A in Rice, section 4.5, tells us that the *mgf* uniquely determines the probability distribution. Hence, $Z_n \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$. Q.E.D.

In practice the following reformulation of the CLT is the most common or practical:

Corollary (CLT)

(6)

If $X_1, X_2, \ldots \sim iid$, *with* $\mathrm{E}(X_i) = \mu$ *and* $\mathrm{var}(X_i) = \sigma^2$, *then*
$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$, *which means that* $\bar{X} \overset{\text{approximately}}{\sim} N(\mu, \frac{\sigma^2}{n})$
for large n.

Proof. We show how the result follows from Rice's CLT in theorem 3: Put $Y_i = X_i - \mu$. Then, $Y_1, Y_2, \ldots \sim iid$, $\mathrm{E}(Y_i) = 0$ and $\mathrm{var}(Y_i) = \sigma^2$. We can then use theorem 3:

$$\frac{\sum_{i=1}^{n} Y_i}{\sigma\sqrt{n}} = \frac{n\bar{X} - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \underset{n\to\infty}{\overset{D}{\to}} Z \sim N(0,1)$$

$$\Rightarrow \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \overset{\text{approx.}}{\sim} N(0,1) \text{ for large } n \Rightarrow \sqrt{n}(\bar{X} - \mu) \overset{\text{approx.}}{\sim} N(0, \sigma^2) \text{ for large } n$$

$$\Rightarrow \bar{X} - \mu \overset{\text{approx.}}{\sim} N(0, \frac{\sigma^2}{n}) \text{ for large } n \Rightarrow \bar{X} \overset{\text{approx.}}{\sim} N(\mu, \frac{\sigma^2}{n}) \text{ for large } n. \quad \text{Q.E.D.}$$

[Note that we in the proof several times have used the well known property of the normal distribution: If $X \sim N(\eta, \tau^2)$, then $a + bX \sim N(a + b\eta, b^2\tau^2)$ where *a,b* are constants.]

The next result that we present, is an extremely useful result for statistical practice:

Theorem 4 (Slutsky's lemma)[2]

Let A_n, B_n, X_n *be r.v.'s such that* $A_n \underset{n\to\infty}{\overset{P}{\to}} a$ *(constant),* $B_n \underset{n\to\infty}{\overset{P}{\to}} b$ *(constant),*
and $X_n \underset{n\to\infty}{\overset{D}{\to}} X$. *Then* $A_n X_n + B_n \underset{n\to\infty}{\overset{D}{\to}} aX + b$
In particular, if $A_n \underset{n\to\infty}{\overset{P}{\to}} 0$, *then* $A_n X_n + B_n \underset{n\to\infty}{\overset{P}{\to}} b$ *(because of (5) above).*

[2] Note that in econometric literature (see e.g. in Greene's book, "Econometric Analysis"), it is usually theorem 1 on the continuity property of *p*lim that is referred to by "Slutsky's theorem", while, in the statistical literature it is usually this theorem 4 that is meant. It appears that Slutsky proved several simpler versions of both these two and other limit results in a paper in 1925. *In this course we will refer to theorem 4 as Slutsky's lemma (or theorem)*, since this result is one of the most important results of the course, and since it takes a little bit of training in exercises to learn to use it properly.

The proof is a straightforward, but somewhat lengthy, ε, δ - argument along the lines illustrated in appendix 2, and is omitted here.

Here we illustrate the result by making some arguments for confidence intervals presented in the introductory statistics course more precise.

Example 5. (Confidence intervals)

(Note: *It is recommended that you study this example thoroughly and learn the argument used. In particular note how Slutsky's lemma is used in the argument. The example also gives an example of why the concept of consistency is useful*.)

Suppose $X_1, X_2, \ldots$ are *iid*, with $\mathrm{E}(X_i) = \mu$ (unknown) and $\mathrm{var}(X_i) = \sigma^2$. We want a confidence interval (CI) with degree of confidence, $1-\alpha$, for the unknown μ. Even if the common distribution, $F(x)$, for the X_i's, is unknown, the distribution of $\bar{X}$ is approximately known for large *n* ($n \geq 30$ usually considered sufficient) because of the CLT, which we utilize as follows:

For large *n*, $Z_n = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{approx.}}{\sim} N(0,1)$. Hence, $P(-z_{\frac{\alpha}{2}} \leq Z_n \leq z_{\frac{\alpha}{2}}) \approx 1-\alpha$ where $z_{\frac{\alpha}{2}}$ is the upper $\alpha/2$-point in $N(0,1)$. Manipulating the probability (do it!), we get (as in the basic statistics course)

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) \approx 1-\alpha$$

Thus, if σ is known, then an approximately $1-\alpha$ CI for μ is given by

(7) $$\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

In practice σ is usually unknown, but according to Slutsky's lemma, σ can be replaced by a consistent estimator, as the following argument shows:

Put $U_n = \dfrac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$ where $\hat{\sigma} = \sqrt{S^2} = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ is consistent for σ (see example 3). We then have

$$U_n = \frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} = \frac{\sigma}{\hat{\sigma}} \cdot \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\sigma}{\hat{\sigma}} \cdot Z_n$$

Since $\frac{\sigma}{\hat{\sigma}} \underset{n\to\infty}{\overset{P}{\to}} \frac{\sigma}{\sigma} = 1$ (see theorem 1 and example 3), and $Z_n \underset{n\to\infty}{\overset{D}{\to}} Z$, we have from Slutsky's lemma $U_n \underset{n\to\infty}{\overset{D}{\to}} 1 \cdot Z = Z \sim N(0,1)$. Hence, for large *n*, $P(-z_{\frac{\alpha}{2}} \le U_n \le z_{\frac{\alpha}{2}}) \approx 1-\alpha$.
Manipulating this (do it!), we get

$$P\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \bar{X} + z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right) \approx 1-\alpha$$

which gives the approximate $1-\alpha$ CI for μ: $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}$. Simulation studies show that the approximation is usually satisfactory for $n \ge 30$.

Notice that the CI is the same as the CI in (7) where we have replaced the unknown σ with a consistent estimator $\hat{\sigma}$, and that *it is Slutsky's lemma that allows us to do that*.

We have a similar state of affairs for poisson- and binomial models:

The poisson case: Suppose that the number, *X*, of working accidents during *t* time units in a large firm, is ~ pois(λt), where λ is the unknown expected (i.e. long run average) accident rate per time unit in the firm. Then $\mathrm{E}(X) = \lambda t = \mathrm{var}(X)$, which implies that $\hat{\lambda} = \frac{X}{t}$ is an unbiased estimator of λ. Since $\mathrm{var}(\hat{\lambda}) = \frac{\lambda}{t} \underset{t\to\infty}{\to} 0$, it follows from Chebyshev's inequality (check!) that $\hat{\lambda}$ is consistent for λ as well as $t \to \infty$ (i.e., $\hat{\lambda} \underset{t\to\infty}{\overset{P}{\to}} \lambda$). From example 4 we get that

$$Z_t = \frac{X - t\lambda}{\sqrt{t\lambda}} = \frac{t\hat{\lambda} - t\lambda}{\sqrt{t}\sqrt{\lambda}} = \sqrt{t}\,\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda}} \underset{t\to\infty}{\overset{D}{\to}} Z \sim N(0,1) \quad \text{since } t\lambda \to \infty \text{ as } t \to \infty.$$

Slutzky's lemma shows that we can replace λ by $\hat{\lambda}$ in the denominator of Z_t without destroying the approximation substantially, i.e.,

$$U_t = \sqrt{t}\,\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}}} = \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}}\cdot Z_t \underset{t\to\infty}{\overset{D}{\to}} 1\cdot Z = Z \sim N(0,1) \quad \text{since} \quad \frac{\sqrt{\lambda}}{\sqrt{\hat{\lambda}}} \overset{P}{\to} 1 \text{ as } t\to\infty,$$ using

that the function, $g(x) = \sqrt{\lambda}/\sqrt{x}$ is continuous in *x*. We then get for large *t* (the criterion $t\lambda \geq 10$ is usually considered sufficient), the following approximation

$$P\left(\hat{\lambda} - z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{\lambda}}}{\sqrt{t}} \leq \lambda \leq \hat{\lambda} + z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{\lambda}}}{\sqrt{t}}\right) \approx 1-\alpha$$

then gives an approximate $1-\alpha$ CI for λ: $\hat{\lambda} \pm z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{\lambda}}}{\sqrt{t}}$.

Discuss the binomial case yourself.

Appendix 1 (mathematical prerequisites for Rice, chapter 5)

First some review of sequences and continuous functions:

(A1) Sequences (see also Sydsæter I, section 6.4 on sequences ("tallfølger"))

An example of a sequence is:

(i) $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, \frac{1}{n}, \ldots$ (or described more shortly simply as $1, \frac{1}{2}, \frac{1}{3}, \ldots$)

which is a sequence of numbers continued indefinitely. It consists of infinitely many numbers, one for each integer, *n*. Abstractly we can describe a sequence (of numbers), $a_1, a_2, a_3, \ldots, a_n, \ldots$ simply as *a function*, a_n , defined for each natural number *n*. Thus, the sequence (i) can also be described as

$$a_n = \frac{1}{n} \quad \text{for } n = 1,2,3,\ldots.$$

We see that this particular sequence converges to 0 as *n* increases, i.e.,

$\lim_{n\to\infty} a_n = \lim_{n\to\infty}\frac{1}{n} = 0$. However, in general a sequence does not have to converge. For example, the sequence (ii) does not converge

(ii) $-1,\ 1,\ -1,\ 1, \ldots\ldots, (-1)^n, \ldots$ (or $a_n = (-1)^n$ for $n = 1,2,3,\ldots$)

Most number sequences we meet in this course, however, converge. For example

(iii) $2, \frac{3}{2}, \frac{4}{3}, \ldots\ldots, \frac{n}{n-1}, \ldots$ (or $a_n = \frac{n}{n-1}$ for $n = 2,3,4,\ldots$)

which converges to 1. (Note that $\frac{n}{n-1} = \frac{n-1+1}{n-1} = 1 + \frac{1}{n-1} \to 1$ as $n \to \infty$).

A famous sequence is the following:

(iv) $\left(1 + \frac{1}{n}\right)^n$ for $n = 1,2,3,\ldots$

which converges to $e = 2,718281828\ldots$ as $n \to \infty$ (proven in **A6** below).

We also talk about a sequence of random variables, $X_1, X_2, X_3, \ldots\ldots, X_n, \ldots$, which just means that the r.v. X_n is well defined for any natural number. When we say that the infinite sequence of r.v.'s, $X_1, X_2, X_3, \ldots$, is an *iid* sequence, we mean that all the (infinite number of) variables are observed under identical conditions (i.e., they have the same distribution) and observed independently of each other. From this sequence we can define other sequences, for example the sequence of means $(\bar{X})$: $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots$, where

$\bar{X}_1 = X_1$,

$\bar{X}_2 = \frac{1}{2}(X_1 + X_2)$

…….

$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ and so on.

(Note that we have here put an index, *n*, on the mean $\bar{X}_n$ to underline its dependence on *n*, i.e., the number of observations used. Usually we drop that index from the notation if the number of observations is fixed and understood from the context.)

(A2) Continuous functions (review)

Definition (see Sydsæter I, section 6.9): A function $y = f(x)$ is continuous in *x* if for *any* sequence, $x_1, x_2, x_3, \ldots$ that belongs to the area of definition of *f* and converges to *x*, ($\lim_{n\to\infty} x_n = x$), then also $f(x_n) \to f(x)$ as $n \to \infty$[3].

[3] Similarly for several arguments in *f*: $z = f(x, y)$ is continuous in (x, y) if, for every sequence, $x_1, x_2, \ldots \to x$ and $y_1, y_2, \ldots \to y$ we have that $f(x_n, y_n) \to f(x, y)$ as $n \to \infty$.

In this course we often need to check that a function is continuous in order to use various results from the theory. There are some simple rules you should know that makes it easy in most cases to check that a function is continuous simply by looking at the expression for the function:

(i) All elementary functions used in this course are continuous. Those include e.g.:
linear functions, $y = ax + b$ where *a*,*b* are constants,
power functions, $y = x^n$ where *n* is an integer, or
$y = x^r$ for any real *r* when $x > 0$,
exponentials, $y = \exp(x) = e^x$ or $y = a^x$ for any $a > 0$,
log functions, $y = \log(x)$ when $x > 0$,
the gamma function, $y = \Gamma(x)$ when $x > 0$.

(ii) If $y = f(x)$ and $y = g(x)$ are both continuous, then all the following functions are continuous as well:

(a) $y = c \cdot f(x)$ where *c* is a constant,
(b) $y = f(x) + g(x)$
(c) $y = f(x) \cdot g(x)$
(d) $y = f(x)/g(x)$ when $g(x) \neq 0$
(e) $y = f(g(x))$ - a function of a function.

(iii) The rules under (ii) are still valid if *f* and *g* depend on more than one variable. For example, if $f(x, y)$ and $g(x, y)$ are both continuous in *x* and *y*, then $h(x, y) = f(x, y)/g(x, y)$ is continuous in *x* and *y* (when $g(x, y) \neq 0$) and so on.

Examples:
Using (i) and (ii)(a and b), we see that $y = 2x^5 - 3x + 4$ is continuous, and, more generally, any polynomial in *x*.

$h(x, y) = x^2 - xy$ is continuous since both x^2 and xy are continuous (the last one because of (iii)).

The pdf of the N(μ, σ^2) distribution, $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, we immediately see is continuous since:

$x - \mu$ is continuous $\Rightarrow -\frac{1}{2\sigma^2}(x-\mu)^2$ is continuous (using (i) and (ii)e and noting that $-\frac{1}{2\sigma^2}$ is just a constant),

$\Rightarrow \quad e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ is continuous (using (i) and (ii)e),

$\Rightarrow \quad f(x)$ is continuous (using (ii)a).

The following results are much used in probability theory (a motivation can be found in Sydsæter I, section 7.6).

(A3)

(i) For any real a, e^a can be expressed as an infinite series

$$e^a = \sum_{i=0}^{\infty} \frac{a^i}{i!} = 1 + a + \frac{a^2}{2!} + \cdots + \frac{a^n}{n!} + \cdots$$

(ii) If c is a common factor, it can be taken outside the sum,

$$\sum_{i=0}^{\infty} c\frac{a^i}{i!} = c\sum_{i=0}^{\infty} \frac{a^i}{i!} = ce^a$$

[**Note.** The theory of infinite series is not treated in the mathematics curriculum, except geometric series, so we will not go into this here. We only mention that the precise mathematical meaning of the infinite sum is as a limit of a corresponding sequence of numbers (see A1 above), $s_n = 1 + a + \frac{a^2}{2!} + \cdots + \frac{a^n}{n!}, \qquad n = 1, 2, 3, \ldots.$

Then it can be shown that

$$e^a = \lim_{n\to\infty} s_n = \lim_{n\to\infty}\left(1 + a + \frac{a^2}{2!} + \cdots + \frac{a^n}{n!}\right),$$

is well defined and true for every *a* – which is the precise meaning of the statement in (*i*) (we say that the series is *convergent* if the corresponding sequence converges).

The only additional result from the theory of infinite series we need is the last statement that a common factor can be taken outside the sum. This particular series is mainly used to derive the *mgf* for a poisson r.v. (see Example A in Rice, section 4.5): $X \sim \text{pois}(\lambda)$ implies that the *mgf* is $M(t) = \text{E}(e^{tX}) = e^{\lambda(e^t - 1)}$]

Much of approximation theory in mathematics and probability theory is based on the famous **Taylor's formula** (see Sydsæter I, section 7.6):

(A4)

Let $f(x)$ be $n+1$ times differentiable in an interval that contains 0 and x. Then, $f(x)$ can be approximated by a polynomial as follows

$$f(x) = f(0) + \frac{x}{1!} f'(0) + \frac{x^2}{2!} f''(0) + \cdots + \frac{x^n}{n!} f^{(n)}(0) + R_{n+1}(x)$$

where the error term, $R_{n+1}(x)$, is $R_{n+1}(x) = \frac{x^{n+1}}{(n+1)!} f^{(n+1)}(c)$, *where c is a number lying somewhere between 0 and x.*

[**Note:** In **(A4)** we say that $f(x)$ is expanded around $x = 0$. From **(A4)** it follows that we can expand $f(x)$ around any other value, $x = \mu$, where *f* is differentiable: Write $f(x) = f(\mu + x - \mu)$ and define $g(h) = f(\mu + h)$ where $h = x - \mu$. Then $g(0) = f(\mu)$ and $g^{(n)}(0) = f^{(n)}(\mu)$. Applying **(A4)** to $g(h)$, we obtain an expansion of $f(x)$ around $x = \mu$:

(A5)

$$f(x) = g(x-\mu) = f(\mu) + \frac{x-\mu}{1!} f'(\mu) + \cdots + \frac{(x-\mu)^n}{n!} f^{(n)}(\mu) + \frac{(x-\mu)^{n+1}}{(n+1)!} f^{(n+1)}(c)$$

where c is a number lying somewhere between μ and *x*.]

Example 6. Rice section 4.6 gives examples of finding approximate expressions of expectations and variances. Let *X* be a r.v. with $\mathrm{E}(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$. Suppose we want the expectation and variance of a transformed r.v., $Y = g(X)$. If *g* is complicated it is often hard to find $E(Y)$ and var(*Y*) exactly. If $g(x)$ is differentiable around $x = \mu$, however, we can easily obtain approximate values by using Taylor expansion around μ. Ignoring the error term, we have from **(A5)** with *n* =1:

$$g(X) \approx g(\mu) + g'(\mu)(X - \mu)$$

By taking expected value and variance on both sides, we get (note that $g(\mu)$ and $g'(\mu)$ are constants)

$$E(g(X)) \approx g(\mu) \quad \text{and} \quad \mathrm{var}(g(X)) \approx [g'(\mu)]^2 \sigma^2$$

By including an extra term in the expansion, we may obtain a (hopefully – it depends on the error term) better approximation to the expectation:

$$g(X) \approx g(\mu) + g'(\mu)(X-\mu) + \frac{g''(\mu)}{2}(X-\mu)^2$$

gives

$$E(g(X)) \approx g(\mu) + \frac{\sigma^2}{2} g''(\mu)$$ (read example B in Rice, sec. 4.6)

Note that it is usually not a good idea in this context to include many terms in the Taylor approximation since terms like $(X-\mu)^r$ for larger *r* are often statistically quite unstable, which may destroy the approximation. (*End of example.*)

From **(A4)** we can now derive the following much used result (also used in the proof of the CLT):

(A6)

If $a_n, \quad n = 1,2,\ldots$ *is a sequence of numbers(see Sydsæter I, section 6.4) converging to a number, a (i.e.* $a_n \underset{n\to\infty}{\to} a$*), then*

$$\left(1+\frac{a_n}{n}\right)^n \underset{n\to\infty}{\to} e^a$$

Proof (optional reading): Taking log on both sides, the result follows if we can show that $n \cdot \ln\left(1+\frac{a_n}{\mathrm{n}}\right) \underset{n\to\infty}{\to} a$ (since e^x is a continuous function[4]). Put $x_n = \frac{a_n}{n}$. Then $n \cdot x_n = a_n \underset{n\to\infty}{\to} a$. Applying **(A4)** to the function, $f(x) = n \cdot \ln(1+\mathrm{x})$, with only one term plus error, we get $f(x) = f(0) + f'(c)x = \frac{n}{1+c}x$, where *c* is between 0 and *x*. Note that $f(0) = 0$. Therefore, $f(x_n) = n \cdot \ln(1+x_n) = \frac{n \cdot x_n}{1+c_n} \underset{n\to\infty}{\to} \frac{a}{1} = a$, using that $n \cdot x_n \underset{n\to\infty}{\to} a$ and that $c_n \underset{n\to\infty}{\to} 0$. The last statement follows since c_n always lies between 0 and x_n (implying $0 \le |c_n| \le |x_n|$), and $x_n = \frac{a_n}{n} \underset{n\to\infty}{\to} 0$ since the sequence, $a_n, \quad n = 1,2,\ldots$ converges to *a*, and therefore must be bounded (i.e., there is a number *C* such that $|a_n| \le C$ for all *n*). Q.E.D.

---

[4] By definition the continuity of e^x means that for any sequence, $x_1, x_2, x_3, \ldots$ converging to *x*, ($\lim_{n\to\infty} x_n = x$), then also $e^{x_n} \to e^x$ as $n \to \infty$. Put $x_n = n \cdot \ln\left(1+\frac{a_n}{n}\right)$.

Note: The rest of appendix 1 and 2 is optional reading.

In order to make the proof of the CLT (theorem 3, page 8) completely rigorous we need one more mathematical fact.

(A7)

> *If the mgf,* $M(t) = \mathrm{E}(e^{tX})$ *of a r.v., X, exists for all t in an open interval containing 0 (i.e. for all* $t \in (a,b)$ *where* $a < 0 < b$*), then the n-th derivative,* $M^{(n)}(t)$*, exists for all* $n = 1,2,\ldots$ *in this interval. This implies, in particular that* $M^{(n)}(t)$ *is continuous in* (a,b) *for all n.*

[**Note.** This result is not hard to prove, but requires results from more advanced integration theory, and is therefore omitted here. Note also that **(A7)** shows that the assumption that $M(t)$ exists in an open interval around 0, is a quite strong assumption on the distribution of *X*. It implies that moments, $\mathrm{E}(X^r)$, of all orders $r = 1,2,\ldots$ exist. This follows since, $\mathrm{E}(X^r) = M^{(r)}(0)$ then exists for all *r*. The assumption is valid for most of the common distributions met in this course, but there are notable exceptions. For example it is not true for *t*-distributions, since, if *X* is *t*-distributed with ν degrees of freedom, then it can be shown that $\mathrm{E}(X^r)$ exists only for $r < \nu$.]

Appendix 2 (some proofs)

Proof of (4) (optional reading)
We will prove the *h*(*x*,*y*)-case. Try to write out a proof for the simpler *g*(*x*)-case yourself (in case you don't realize that the *g*-case follows directly from the *h*-case).

Suppose $X_n \underset{n\to\infty}{\overset{P}{\to}} c$ and $Y_n \underset{n\to\infty}{\overset{P}{\to}} d$ and that $h(x,y)$ is continuous for $x = c$, $y = d$. Choose an $\varepsilon > 0$ arbitrarily small. We need to prove that $P(|h(X_n,Y_n) - g(c,d)| \le \varepsilon) \underset{n\to\infty}{\to} 1$.

According to the meaning of continuity (see e.g. Sydsæter I, sec. 6.9), there is a $\delta > 0$ such that, whenever $|x - c| \le \delta$ and $|y - d| \le \delta$, then $|h(x,y) - h(c,d)| \le \varepsilon$.

Define events, A_n, B_n, C_n by $A_n = (|X_n - c| \le \delta)$, $B_n = (|Y_n - d| \le \delta)$, and $C_n = (|h(X_n,Y_n) - h(c,d)| \le \varepsilon)$.

We then have $A_n \cap B_n \Rightarrow C_n$ which implies that $P(C_n) \ge P(A_n \cap B_n)$. (Note that if *A*, *B* are events such that $A \Rightarrow B$, or $A \subset B$ interpreted as sets, then $P(A) \le P(B)$).
According to the definition of probability limit, $P(A_n) \underset{n\to\infty}{\to} 1$ and $P(B_n) \underset{n\to\infty}{\to} 1$.
This implies that $P(A_n \cap B_n) \underset{n\to\infty}{\to} 1$ since

$P(A_n \cap B_n) = P(A_n) + P(B_n) - P(A_n \cup B_n) \to 1+1-1=1$ as $n \to \infty$ (Note that $P(A_n \cup B_n) \geq P(A_n) \to 1$ implies that $P(A_n \cup B_n) \to 1$). Hence, since $P(C_n) \geq P(A_n \cap B_n)$, also $P(C_n) \to 1$ as $n \to \infty$. Q.E.D.

Proof of (5) (optional reading)

i) Suppose that $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$. We need to prove that $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$ where $P(Y=c)=1$. Let the cdf of Y_n be $F_n(y)$ and the cdf of Y be $F(y)$, i.e. the trivial cdf at c (see **1.2**)

$$F(y) = P(Y \leq y) = \begin{cases} 0 & \text{for } y < c \\ 1 & \text{for } y \geq c \end{cases}$$ Thus $F(y)$ is continuous for all $y \neq `c$.

Hence, according to the definition of convergence in distribution, we need to show that $F_n(y) \underset{n\to\infty}{\to} F(y)$ for all $y \neq `c$, or $F_n(y) \underset{n\to\infty}{\to} 0$ for $y < c$ and $F_n(y) \underset{n\to\infty}{\to} 1$ for $y > c$.

Again we use that if $A \Rightarrow B$, then $P(A) \leq P(B)$. Suppose $y > c$ (or $y - c > 0$). Then the following events satisfy

$$(|Y_n - c| \leq y - c) \Leftrightarrow (-(y-c) \leq Y_n - c \leq y - c) \Leftrightarrow (c-(y-c) \leq Y_n \leq c + y - c)$$
$$\Leftrightarrow (2c - y \leq Y_n \leq y) \Rightarrow (Y_n \leq y)$$

Hence $F_n(y) = P(Y_n \leq y) \geq P(|Y_n - c| \leq y - c) \underset{n\to\infty}{\to} 1$ since $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$. Therefore, we must have that $F_n(y) \underset{n\to\infty}{\to} 1$.

Now, suppose $y < c$ (i.e. $c - y > 0$). We have

$$(Y_n \leq y) \Leftrightarrow (-Y_n \geq -y) \Leftrightarrow (c - Y_n \geq c - y) \Rightarrow (|Y_n - c| \geq c - y) \Rightarrow (|Y_n - c| > \frac{c-y}{2})$$

Thus, $F_n(y) = P(Y_n \leq y) \leq P\left(|Y_n - c| > \frac{c-y}{2}\right) \underset{n\to\infty}{\to} 0$, which implies that $F_n(y) \underset{n\to\infty}{\to} 0$, and we have proven that $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$.

ii) Now, conversely, suppose that $Y_n \underset{n\to\infty}{\overset{D}{\to}} Y$ where $P(Y=c)=1$. Then $F_n(y) \underset{n\to\infty}{\to} F(y)$ for all $y \neq `c$. Let $\varepsilon > 0$ be arbitrary small. We have

$$P(|Y_n - c| \leq \varepsilon) = P(c - \varepsilon \leq Y_n \leq c + \varepsilon) \geq P(c - \varepsilon < Y_n \leq c + \varepsilon) = F_n(c+\varepsilon) - F_n(c-\varepsilon)$$

Since $F(y)$ is continuous for $y = c - \varepsilon$ and $y = c + \varepsilon$, the last expression converges to $F(c+\varepsilon) - F(c-\varepsilon) = 1 - 0 = 1$ as $n \to \infty$. Hence $P(|Y_n - c| \le \varepsilon) \underset{n\to\infty}{\to} 1$, and we have proven that $Y_n \underset{n\to\infty}{\overset{P}{\to}} c$. Q.E.D.