Lecture 12: The Central Limit Theorem

Relevant textbook passages:
Pitman [18]: Section 3.3, especially p. 196
Larsen–Marx [12]: Section 4.3, (Appendix 4.A.2 is optional)

12.1 Fun with CDFs

Recall that a random variable $X$ on the probability space $(\Omega, \mathcal{F}, P)$ has a cumulative distribution function (cdf) $F$ defined by

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

12.1.1 Quantiles

Assume $X$ is a random variable with a continuous, strictly increasing cumulative distribution function $F$.

- Then the equation $P(X \leq x) = F(x) = p$ has a unique solution $x_p = F^{-1}(p)$ for every $p$ with $0 < p < 1$, namely $x_p = F^{-1}(p)$.
- (If we allow $x$ to take on the values $\pm \infty$, then we can be sure that a solution exists for the cases $p = 0$ and $p = 1$.)

12.1.1 Definition When $X$ has a continuous, strictly increasing cumulative distribution function $F$, the value $x_p = F^{-1}(p)$ is called the $p^{th}$ quantile of the distribution $F$. The mapping $F^{-1}: p \mapsto x_p$ is called quantile function of the distribution $F$.

Some of the quantiles have special names.

- When $p$ has the form of a percent, $x_p$ is called a percentile.
- The $p = 1/2$ quantile is called the median.
- The $p = 1/4$ quantile is the first quartile, and the $p = 3/4$ is called the third quartile.
- The interval between the first and third quartiles is called the interquartile range.
- There are also quintiles, and deciles, and, I’m sure, others as well.

12.1.2 Proposition Let $F$ be the cdf of the random variable $X$. Assume that $F$ is continuous. Then $F \circ X$ is a random variable that is uniformly distributed on $[0, 1]$. In other words,

$$P(F(X) \leq p) = p, \quad (0 \leq p \leq 1).$$

Note that we are not assuming that $F$ is strictly increasing.

Proof: Since $F$ is continuous, its range has no gaps, that is, the range includes $(0, 1)$. Now fix $p \in [0, 1]$. There are three cases, $p = 0$, $0 < p < 1$, and $p = 1$. 
• The case $p = 1$ is trivial, since $F$ is bounded above by 1.

• In case $0 < p < 1$, define

$$x_p = \sup\{x \in \mathbb{R} : F(x) \leq p\}$$
$$= \sup\{x \in \mathbb{R} : F(x) = p\}$$

and note that $x_p$ is finite. In fact, if $F$ is strictly increasing, then $x_p$ is just the $p^{th}$ quantile defined above.

By continuity,

$$F(x_p) = p.$$

By the construction of $x_p$, for all $x \in \mathbb{R}$,

$$F(x) \leq p \iff x \leq x_p. \quad (1)$$

So, replacing $x$ by $X$ above,

$$F(X) \leq p \iff X \leq x_p$$

so

$$P(F(X) \leq p) = P(X \leq x_p) = F(x_p) = p.$$

• The above argument works for $p = 0$ if 0 is in the range of $F$. But if 0 is not in the range of $F$, then $F(X) > 0$ a.s., so $P(F(X) = 0) = 0.$

### 12.1.2 The quantile function in general

Even if $F$ is not continuous, so that the range of $F$ has gaps, there is still a “fake inverse” of $F$ that is very useful. Define

$$Q_F : (0, 1) \to \mathbb{R}$$

by

$$Q_F(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

When $F$ is clear from the context, we shall simply write $Q$.

When $F$ is strictly increasing and continuous, then $Q_F$ is just $F^{-1}$ on $(0, 1)$.

More generally, flat spots in $F$ correspond to jumps in $Q_F$ and vice-versa. The key property is that for any $p \in (0, 1)$ and $x \in \mathbb{R}$,

$$Q_F(p) \leq x \iff p \leq F(x), \quad (2)$$

or equivalently

$$F(x) \geq p \iff x \geq Q_F(p).$$

Compare this to equation (1).

### 12.1.3 Example

Refer to Figure 12.1 to get a feel for the relationship between flats and jumps. The cumulative distribution function $F$ has a flat spot from $x$ to $y$, which means $P(x < X < y) = 0$. Now $\{v : F(v) \geq p\} = [x, \infty)$, so $Q_F(p) = x$.

It also has jump from $p$ to $q$ at $y$, which means $P(y) = q - p$. For $p'$ in the gap $p < p' < q$, we have $\{v : F(v) \geq p'\} = [y, \infty)$, so $Q_F(p') = y$. This creates a flat spot in $Q_F$. 

\[\square\]
12.1.4 Proposition Let $F$ be a cumulative distribution function and let $U$ be a Uniform$[0, 1]$ random variable. That is, for every $p \in [0, 1]$, $P(U \leq p) = p$. Note that $P(U \in (0, 1)) = 1$ so with probability one $Q_F(U)$ is defined. Then $Q_F(U)$ is a random variable and the cumulative distribution function of the random variable $Q_F(U)$ is $F$.

Proof: From (2),

$$P(Q_F(U) \leq x) = P(U \leq F(x)) = F(x),$$

where the last equality comes from the fact that $U$ is uniformly distributed.

The usefulness of this result is this:

If you can generate a uniform random variable $U$, then you can create a random variable $X$ with any distribution $F$ via the transformation $X = Q_F(U)$.

12.2 * Stochastic dominance and increasing functions

Recall that a random variable $X$ stochastically dominates the random variable $Y$ if for every value $x$,

$$P(X > x) \geq P(Y > x).$$

This is equivalent to $F_X(x) \leq F_Y(x)$ for all $x$.

12.2.1 Proposition Let $X$ stochastically dominate $Y$, and let $g$ be a nondecreasing real function. Then

$$E g(X) \geq E g(Y).$$

If $g$ is strictly increasing, then the inequality is strict unless $F_X = F_Y$.

As a special case, when $g(x) = x$, we have $E X \geq E Y$.

I will prove this in two ways. The first proof is kind of lame, but it was the sort of argument that was made when I learned about stochastic dominance.
Proof of a special case: The first proof is for the special where $X$ and $Y$ are strictly bounded in absolute value by $b$, and have densities $f_X$ and $f_Y$, and the function $g$ is continuous continuously differentiable. Then the expected value of $g(X)$ is obtained via the integral

$$\int_{-b}^{b} g(x) f_X(x) \, dx,$$

so integrating by parts we see this is equal to

$$g(t) F_X(t)]_{-b}^{b} - \int_{-b}^{b} F_X(x) g'(x) \, dx.$$

Likewise $E g(Y)$ is equal to

$$g(t) F_Y(t)]_{-b}^{b} - \int_{-b}^{b} F_Y(x) g'(x) \, dx.$$

Now by assumption $F_X(-b) = F_Y(-b) = 0$ and $F_X(b) = F_Y(b) = 1$, so the first term in each expectation is identical. Since $g$ is nondecreasing, $g' \geq 0$ everywhere. Since $F_X \leq F_Y$ everywhere, we conclude $E g(X) \geq E g(Y)$. \qed

A general proof: Proposition 12.1.4 states if $U$ is a Uniform[0,1] random variable, then $Q_F(U)$ is a random variable that has distribution $F$. Now observe that if $F_X \leq F_Y$, then $Q_X \geq Q_Y$. Therefore, since $g$ is nondecreasing,

$$g(Q_X(U)) \geq g(Q_Y(U))$$

Since expectation is a positive linear operator, we have

$$E g(Q_X(U)) \geq E g(Q_Y(U)).$$

But $g(Q_X(U))$ has the same distribution as $g(X)$ and $g(Q_Y(U))$ has the same distribution as $g(Y)$, so their expectations are the same. \qed

For yet a different sort of proof, based on convex analysis, see Border [3].

Aside: The second proof used the concept of a coupling. A coupling of the pair $(X, Y)$ is a pair $(X', Y')$ of random variables, perhaps defined on a different probability space, with the property that $X' \sim X$ and $Y' \sim Y$. Proposition 12.1.4 implies that if $X$ stochastically dominates $Y$ then there is a coupling $(X', Y') = (Q_X, Q_Y)$ satisfying $X' \geq Y'$. Here $Q_X, Q_Y$ are defined on the probability space $([0,1], \mathcal{B}, \lambda)$ where $\mathcal{B}$ is the Borel $\sigma$-algebra of the unit interval, and $\lambda$ is Lebesgue measure (length).

12.3 Convergence in distribution

When should we say that two random variables $X$ and $Y$ are “close” in terms of their distributions? Or rather, when should we say that a sequence of random variables $X_n$ converges to a random variable $X$ in a “distributional” sense? What do we mean by that?

We already have one example, the DeMoivre–Laplace Limit Theorem 10.5.1. It says that says that a Binomial$(n, p)$ random variable $X_n$ has standardization $X_n^*$ that is close to a standard normal random variable $Z$ in the sense that

$$\lim_{n \to \infty} P(a \leq X_n^* \leq b) = P(a \leq Z \leq b).$$

What we want is the “right” way to generalize this notion of convergence. We especially want a notion of convergence that expressed in terms of the cumulative distribution functions of the random variables so that we do not need to refer to the underlying probability space. We can rephrase the DeMoivre–Laplace result in a couple of ways.
• Ideally we would like to be able to say that in order for a sequence or random variables \( X_n \) to “converge distributionally” to \( X \) we would have that for every (Borel) function \( f \),

\[
\lim_{n \to \infty} E f(X_n) = E f(X).
\]

After all, \( E f(X) \) depends only on the distribution of \( X \).

• A slightly less ambitious idea, more in keeping with the DeMoivre–Laplace approach is this. For every \( a < b \),

\[
\lim_{n \to \infty} F_{X_n}(b) - F_{X_n}(a) = F_Z(b) - F_Z(a),
\]

which suggests that we might want to require that \( F_n \to F_X \) pointwise, where \( F_n \) is the cumulative distribution function of \( X_n^* \). But in terms of indicator functions, the condition becomes

\[
\lim_{n \to \infty} E 1_{[a,b]}(X_n^*) = E 1_{[a,b]}(Z).
\]

But it turns out that both these suggestions are too stringent to be very useful. To see why, let’s consider some very simple examples.

12.3.1 Example In the first example, our random variables will be degenerate. That is, they are essentially just numbers. Let \( X \) take the value 0 with probability one, and let each \( X_n \) take on the value \( 1/n \) with probability one. In other words,

\[
X = 0 \text{ a.s. and } X_n = \frac{1}{n} \text{ a.s.}
\]

Surely we wish to have a notion of “distributional convergence” that implies \( X_n \to X \). Can we capture this in terms of the cumulative distribution functions?

The cumulative distribution functions \( F \) of \( X \) and \( F_n \) of \( X_n \) are just

\[
F(x) = \begin{cases} 
0 & x < 0 \\
1 & x \geq 0
\end{cases} \quad \text{and} \quad F_n(x) = \begin{cases} 
0 & x < \frac{1}{n} \\
1 & x \geq \frac{1}{n}
\end{cases}.
\]

See Figure 12.2.

In some sense the cumulative distribution functions get close, but how precisely? Observe that \( \max_x |F_n(x) - F(x)| = 1 \neq 0 \), so uniform convergence of the cumulative distribution functions is too much to ask for if we want to consider these random variables to be close. What about mere pointwise convergence? That fails here too since \( 0 = F_n(0) \neq F(0) = 1 \).

![Figure 12.2. Are these cumulative distribution functions close?](image)

Now consider the indicator function \( f(x) = 1_{[0,1]}(x) \). Then for each \( n \),

\[
E f(X_n) = P(0 < X_n) = 1, \quad \text{but} \quad E f(X) = P(0 < X) = 0.
\]
Thus $E f(X_n) \not\rightarrow E f(X)$. In this example, it is plain to see that the problem is that indicator functions are not continuous. But even continuous functions can pose problems. □

12.3.2 Example Consider next the random variables $X$ and $X_n$, where

$$X = 0 \text{ a.s., } 
X_n = \begin{cases} 
0 & \text{with probability } 1 - \frac{1}{n} \\
\frac{1}{n^2} & \text{with probability } \frac{1}{n}.
\end{cases}$$

So

$$F(x) = \begin{cases} 
0 & x < 0 \\
1 & x \geq 0
\end{cases} \quad \text{and} \quad F_n(x) = \begin{cases} 
0 & x < 0 \\
1 - \frac{1}{n} & 0 \leq x < \frac{1}{n^2} \\
1 & x \geq \frac{1}{n^2}.
\end{cases}$$

Note that $F_n \xrightarrow{\text{uniformly}} F$, so we would like to be able to say that $X_n$ gets “closer” to $X$. But consider the unbounded continuous function

$$f(x) = x.$$ 

Then $E f(X) = 0$, but for each $n$, we have $E f(X_n) = n$, and again we have $E f(X_n) \not\rightarrow E f(X)$. □

---

Figure 12.3. Are these cumulative distribution functions close?

---

- But now consider a bounded continuous function $f$ and the same random variables as in Example 12.3.2. Let $B$ a bound on $f$, that is, $|f(x)| \leq B$ for all $x$. Then $E f(X) = f(0)$, and $f(X_n) = \frac{n-1}{n} f(0) + \frac{1}{n} f(n^2)$. Then

$$|E f(X_n) - E f(X)| = \frac{1}{n} |f(n^2)| \leq \frac{B}{n} \rightarrow 0.$$ 

- Next consider a bounded continuous function $f$ and the same random variables as in Example 12.3.1. Then $E f(X) = f(0)$, and $f(X_n) = f(1/n)$. Since $f$ is continuous, $f(1/n) \rightarrow f(0)$, so

$$E f(X_n) \rightarrow E f(X).$$

Not in Pitman. This suggests the following definition.
12.3.3 Definition The sequence $X_n$ of random variables converges in distribution to the random variable $X$, written $X_n \xrightarrow{D} X$, if for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$E f(X_n) \to E f(X).$$

Letting $F_n$ denote the cumulative distribution function of $X_n$ and $F$ denote the cumulative distribution function of $X$, we also say that $F_n$ converges in distribution to $F$, written

$$F_n \xrightarrow{D} F.$$

12.3.4 Remark This is not exactly what we had hoped for, since it is not obvious how to express this property easily in terms of the cumulative distribution functions. That is why the next theorem is useful.

There are several equivalent notions, based on the notion of a convergence determining class of functions.

12.3.5 Theorem Let $X_n$ be random variables with cumulative distribution functions $F_n$, and let $X$ be a random variable with cumulative distribution function $F$. The following are equivalent.

1. $F_n \xrightarrow{D} F$, that is, for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$E f(X_n) \to E f(X).$$

2. For every bounded uniformly continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$E f(X_n) \to E f(X).$$

3. For every continuous function $f : \mathbb{R} \to \mathbb{R}$ with bounded support,\(^1\)

$$E f(X_n) \to E f(X).$$

4. For every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ with bounded continuous \(n^{th}\) order derivatives and bounded support,

$$E f(X_n) \to E f(X).$$

5. $F_n(x) \to F(x)$ at every $x$ where $F$ is continuous.

For now, I shall not prove this theorem, but I may add a proof or references in the future. See, e.g., Breiman[4, § 8.3, pp. 163–164] and Billingsley [2, Chapter 1].

• Some authors use the last equivalent property, pointwise convergence at all points of continuity of $F$, as the definition of convergence in distribution, but it is hard to understand why that is an interesting concept. It is the theorem that makes the property interesting.

12.3.6 Remark Note that convergence in distribution depends only on the cumulative distribution function, so random variables can be defined on different probability spaces and still converge in distribution. This is not true of convergence in probability or almost-sure convergence.

\(^1\) The support of a function $f : \mathbb{R} \to \mathbb{R}$ is the closure of $\{x : f(x) \neq 0\}$. 

KC Border v. 2020.10.21::10.28
Aside: It is possible to quantify how close two distribution functions are using the Lévy metric, which is defined by

\[ d(F, G) = \inf \{ \varepsilon > 0 : (\forall x) \left[ F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon \right] \}. \] (3)

See, e.g., Billingsley [2, Problem 4, pp. 21–22]. Note that since \( F \) and \( G \) are bounded between 0 and 1, we must have \( d(F, G) \leq 1 \).

It may not seem that this expression is symmetric in \( F \) and \( G \), but it is. For suppose \( \varepsilon > d(F, G) \), as defined in (3). Pick an arbitrary \( \bar{x} \in \mathbb{R} \). Then applying (3) to \( x = \bar{x} \), we get

\[ F(\bar{x} - \varepsilon) - \varepsilon \leq G(\bar{x}) \leq F(\bar{x} + \varepsilon) + \varepsilon. \]

Applying (3b) to \( x = \bar{x} - \varepsilon \) gives

\[ G(\bar{x} - \varepsilon) \leq F(\bar{x}) + \varepsilon, \]

and applying (3a) to \( x = \bar{x} + \varepsilon \) gives

\[ F(\bar{x}) - \varepsilon \leq G(\bar{x} + \varepsilon). \]

Rearranging the last two inequalities gives

\[ G(\bar{x} - \varepsilon) - \varepsilon \leq F(\bar{x}) \leq G(\bar{x} + \varepsilon) + \varepsilon. \]

That is, \( \varepsilon > d(F, G) \implies \varepsilon > d(G, F) \). Symmetry implies the converse, so we see that \( d(F, G) = d(G, F) \).

It is now not hard to show that if \( d(F, F_n) \to 0 \), then \( F_n \xrightarrow{D} F \): For suppose \( d(F, F_n) = \varepsilon_n \to 0 \) and that \( \bar{x} \) is a point of continuity of \( F \). Then \( F(\bar{x} - \varepsilon_n) - \varepsilon_n \leq F_n(\bar{x}) \leq F(\bar{x} + \varepsilon_n) + \varepsilon_n \). Since \( F \) is continuous at \( \bar{x} \), we have \( F(\bar{x} + \varepsilon_n) + \varepsilon_n \to F(x) \) and \( F(\bar{x} - \varepsilon_n) - \varepsilon_n \to F(x) \), so \( F_n(\bar{x}) \to F(x) \). Since \( \bar{x} \) is an arbitrary point of continuity of \( F \), we have \( F_n \xrightarrow{D} F \).

The converse is slightly more difficult, and I’ll leave it for an exercise or a more advanced course.

Aside: For bounded random variables with values in some bounded interval \([a, b]\), you can show that \( F_n \to F \) if and only if \( \int_a^b |F_n(x) - F(x)| \, dx \to 0 \), but for some reason, this is not used a lot.

Aside: The fact above does not imply that if \( X_n \xrightarrow{D} X \), then \( E[X_n] \to E[X] \), since the function \( f(x) = x \) is not bounded. Ditto for the variance. There is a more complicated result for unbounded functions that applies often enough to be useful. Here it is.

12.3.7 Theorem (Breiman [4, § 8.3, pp. 163–164, exercise 14]) Let \( X_n \xrightarrow{D} X \), and let \( g \) and \( h \) be continuous functions such that \( |h(x)| \to \infty \) as \( x \to \pm \infty \), and \( |g(x)/h(x)| \to 0 \) as \( x \to \pm \infty \). If \( \lim \sup_{n \to \infty} E[|h(X_n)|] < \infty \), then \( E\left(g(X_n)/h(X_n)\right) \to E\left(g(X)/h(X)\right) \).

As a consequence,

If \( X_n \xrightarrow{D} X \), and \( E|X_n^2| \) is bounded, then \( E[X_n] \to E[X] \) and \( \text{Var}(X_n) \to \text{Var}(X) \).

12.4 Central Limit Theorem

We saw in the Law of Large Numbers that if \( S_n \) is the sum of \( n \) independent and identically distributed random variable with finite mean \( \mu \) and standard deviation \( \sigma \), then the sample mean \( A_n = S_n/n \) converges to \( \mu \) as \( n \to \infty \). This is because the mean of \( A_n = S_n/n \) is \( \mu \) and the standard deviation is equal to \( \sigma/\sqrt{n} \), so the distribution is collapsing around \( \mu \).

What if we don’t want the distribution to collapse? Let’s standardize the average \( A_n \):

\[ A_n^* = \frac{A_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n}{\sigma/\sqrt{n}} - \mu = \frac{n}{\sigma/\sqrt{n}} \frac{S_n}{\sqrt{n} \sigma} = \frac{S_n}{\sqrt{n} \sigma}. \]
Or equivalently, by Proposition 7.2.2, let’s look at $S_n/\sqrt{n}$ rather than $S_n/n$. The mean of $S_n/\sqrt{n}$ is $\sqrt{n} \mu$ and the standard deviation is $\sigma$. The standardization is

$$\frac{S_n - n \mu}{\sqrt{n} \sigma} = S_n - \frac{n \mu}{\sqrt{n} \sigma},$$

which is the same as the standardization of $A_n$.

One version of the Central Limit Theorem tells what happens to the distribution of the standardization of $A_n$. The DeMoivre–Laplace Limit Theorem is a special case of the Central Limit Theorem that applies to the Binomial Distribution. Cf. Pitman [18, p. 196] and Larsen–Marx [12, Theorem 4.3.2, pp. 246–247].

12.4.1 Central Limit Theorem, v. 1 Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables. Let $\mu = E X_i$ and $0 < \text{Var} X_i = \sigma^2 < \infty$. Define $S_n$ by $S_n = \sum_{i=1}^{n} X_i$.

Then

$$\frac{S_n - n \mu}{\sqrt{n} \sigma} \xrightarrow{d} N(0, 1).$$

The proof of this result is beyond the scope of this course, but I have included a completely optional appendix to these notes sketching it.

- Note that we have to assume that the variance is positive, that is, the $X_i$’s are not degenerate. If all the random variables are degenerate, there is no way to standardize them, and the limiting distribution, if any, will be degenerate, not normal.

But armed with this fact, we are now in a position to explain one of the facts mentioned in Lecture 10.3, page 10–6. I will prove it for the case where $X$ has a finite variance $\sigma^2$, but it is true without that restriction.

12.4.2 Proposition If $X$ and $Y$ are independent with the same distribution having finite variance $\sigma^2$, and if $(X + Y)/\sqrt{2}$ has the same distribution as $X$, then $X$ has a normal $N(0, \sigma^2)$ distribution.

Proof: (Cf. Breiman [4, Proposition 9.2.1, p. 186]) Note first that $E X = 0$ since

$$E X = E(X + Y)/\sqrt{2} = \sqrt{2} E X.$$

Now let $X_1, X_2, X_3, \ldots$ be a sequence of independent random variables having the same distribution as $X$. Then

$$X \sim (X_1 + X_2)/\sqrt{2} \sim \frac{X_1 + X_3 + X_4}{\sqrt{4}} = \frac{X_1 + X_2 + X_3 + X_4}{\sqrt{4}}, \text{ etc.}$$

So

$$\frac{X}{\sigma} \sim \frac{X_1 + \cdots + X_n}{\sqrt{n} \sigma} \text{ whenever } n \text{ is a power of 2},$$

but the right-hand side converges in distribution to a $N(0, 1)$ distribution.

In fact, the CLT (as it is known to its friends) is even more general.
12.5 ★ A CLT for non-identical distributions

What if the $X_i$’s are not identically distributed? The CLT can still hold. The following result is taken from Breiman [4, Theorem 9.2, pp. 186–188]. There are other variations. See also Feller [8, Theorem VIII.4.3, p. 262], Lindeberg [15], Loève [16, 21.B, p. 292]; Hall and Heyde [10, 3.2].

12.5.1 A more general CLT Let $X_1, X_2, \ldots$ be independent random variables (not necessarily identically distributed) with $E X_i = 0$ and $0 < \text{Var} X_i < \infty$, and $E |X_i^3| < \infty$. Define

$$S_n = \sum_{i=1}^{n} X_i.$$ 

Let

$$s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2 = \text{Var} S_n,$$

so that $s_n$ is the standard deviation of $S_n$. Assume that

$$\limsup_n \frac{1}{s_n^3} \sum_{i=1}^{n} E |X_i^3| = 0,$$

then

$$\frac{S_n}{s_n} \xrightarrow{d} N(0,1).$$

The importance of this result is that for a standardized sum of a large number of independent random variables, each of which has a negligible part of the total variance, the distribution is approximately normal.

This is the explanation of the observed fact that many characteristics of a population have a normal distribution. In applied statistics, it is used to justify the assumption that data are normally distributed.

12.6 ★ The Berry–Esseen Theorem

One of the nice things about the Weak Law of Large Numbers is that it gave us information on how close we probably were to the mean. The Berry–Esseen Theorem gives information on how close the cdf of the standardized sum is to the standard normal cdf. The statement and a proof may be found in Feller [8, Theorem XVI.5.1., pp. 542–544]. See also Bhattacharya and Ranga Rao [1, Theorem 12.4, p. 104].

12.6.1 Berry–Esseen Theorem Let $X_1, \ldots, X_n$ be independent and identically distributed with expectation 0, variance $\sigma^2 > 0$, $E(|X_i^3|) = \rho < \infty$. Let $F_n$ be the cumulative distribution function of $S_n/\sqrt{n}\sigma$. Then for all $x \in \mathbb{R}$,

$$|F_n(x) - \Phi(x)| \leq \frac{3\rho}{\sigma^3 \sqrt{n}}.$$

12.7 ★ The CLT and densities

Under the hypothesis that each $X_i$ has a density and finite third absolute moment, it can be shown that the density of $S_n/\sqrt{n}$ converges uniformly to the standard normal density at a rate of $1/\sqrt{n}$. See Feller [8, Section XVI.2, pp. 533ff].
12.8 ★ The “Second Limit Theorem”

Fréchet and Shohat [9] give an elementary proof of a generalization of another useful theorem on convergence in distribution, originally due to Andrey Andreyevich Markov (Андре́й Андре́еви́ч Марков), 2 which he called the Second Limit Theorem. More importantly, their proof is in English. The statement here is taken from van der Waerden [23, § 24.F, p. 103]. You can also find this result in Breiman [4, Theorem 8.48, pp. 181–182].

12.8.1 Markov–Fréchet–Shohat Second Limit Theorem Let $F_n$ be a sequence of cdfs where each $F_n$ has finite moments $\mu_{k,n}$ of all orders $k = 1, 2, \ldots$, and assume that for each $k$, $\lim_{n \to \infty} \mu_{k,n} = \mu_k$, where each $\mu_k$ is finite. Then there is a cdf $F$ such that the $k$th moment of $F$ is $\mu_k$. Moreover, if $F$ is uniquely determined by its moments, then

$$F_n \xrightarrow{D} F.$$ 

An important case is the standard Normal distribution, which is determined by its moments

$$\mu_k = \begin{cases} 0 & \text{if } k \text{ is odd} \\ \frac{(2m)!}{2^{m}m!} & \text{if } k = 2m. \end{cases}$$

Mann and Whitney [17] used this to derive the asymptotic distribution of their eponymous test statistic [23, p. 277], which we shall discuss in Section 26.4.

12.9 ★ Slutsky’s Theorem

My colleague Bob Sherman assures me that the next result, known as Slutsky’s Theorem [21], is incredibly useful in his work. (I also recently [Dec, 2019] spoke with statistical genetics post-doc who used this result in estimating heritability under assortative mating.) This version is taken from Cramér [6, § 20.6, pp. 254–255].

12.9.1 Theorem (Slutsky’s Theorem) Let $X_1, X_2, \ldots$, be a sequence of random variables with cdfs $F_1, F_2, \ldots$. Assume that $F_n \xrightarrow{D} F$. Let $Y_1, Y_2, \ldots$, satisfy $\text{plim } Y_n = c$, where $c$ is a real constant. Then, (slightly abusing notation),

$$(X_n + Y_n) \xrightarrow{D} F(x - c), \quad (X_nY_n) \xrightarrow{D} F(x/c) \ (c > 0), \quad (X_n/Y_n) \xrightarrow{D} F(cx) \ (c > 0).$$

One of the virtues of this theorem is that you do not need to know anything about the independence or dependence of the random variables $X_n$ and $Y_n$. The proof is elementary, and may be found in Cramér [6, § 20.6, pp. 254–255], Jacod and Protter [11, Theorem 18.8, p. 161], or van der Waerden [23, § 24.G, pp. 103–104].

---

2 Not to be confused with his son, Andrey Andreyevich Markov, Jr. (Андре́й Андре́еви́ч Марков).
A proof of a CLT

This section was written by my colleague Bob Sherman. I have edited Bob’s text a bit, to make it stylistically more like the rest of these notes, but also because his keyboard does not have a Shift key. The argument is originally due to Aleksandr Mikhailovich Lyapunov (Алекса́ндр Миха́йлович Ляпуно́в) (formerly known as Alexandre Liapounoff) [13]. The exposition is based in part on Pollard [19, § 3.4, pp. 50–52]. There is a similar argument on Terence Tao’s blog [22].

Recall the following Fun Facts:

**FF1** If $E f(X_n) \to E f(X)$ for all bounded uniformly continuous real-valued $f$, then $X_n \xrightarrow{d} X$.

**FF2** If $E f(X_n) \to E f(X)$ for all bounded real-valued $f$ having bounded continuous derivatives of all orders, then $X_n \xrightarrow{d} X$.

**FF3** Suppose $X$, $Y$, and $W$ are random variables where $X$ is independent of $(Y, W)$, $E Y = EW$, $E Y^2 = EW^2$, and $E Y^3 < \infty$ and $E W^3 < \infty$. Finally, suppose $f$ is a real-valued function of a real variable having bounded continuous derivatives up to order 3 (i.e., $f \in C^3(R)$). then there exists a constant $C$:

$$|E f(X + Y) - E f(X + W)| \leq C E |Y|^3 + C E |W|^3$$

**FF3** is easily established by doing Taylor expansions of $f(X + Y)$ and $f(X + W)$ about $X$, then taking differences, and finally taking expectations.

We need one more fun fact. To develop it, let

$$Z = X_1 + \cdots + X_K,$$

where the $X_j$’s are independent random variables with finite third moments. Write $\sigma_j^2$ for $E X_j^2$. Standardize so that $E X_j = 0$ and $\sigma_j^2 + \cdots + \sigma_K^2 = 1$. We want to establish conditions under which $Z \xrightarrow{d} N(0, 1)$ as $K \to \infty$.

To this end, independently of the $X_j$’s, choose independent $N(0, \sigma_j^2)$ random variables $Z_j$, $j = 1, \ldots, K$. Note that $Z_1 + \cdots + Z_K$ is distributed $N(0, 1)$.

**Aside:** Why does this argument only work if the $Z_j$’s are Normal? It is because the normal distribution is the only zero-mean, sum-stable distribution with finite variance. There are lots of sum-stable distributions, continuous and discrete alike. Examples: Discrete: Poisson, Binomial, see Casella and Berger[5, p. 216ff.]. Note: Poisson is not mean zero and (I think) if you center 2 Poissons to zero mean, then the sum is no longer Poisson. Continuous: Cauchy.

For $j = 1, \ldots, K$, define the gap-toothed hybrid sum

$$S_j = X_1 + \cdots + X_{j-1} + Z_{j+1} + \cdots + Z_K.$$

Yes, $S_j$ is missing a tooth—its $j$th tooth! Note that

- $S_K + X_K = X_1 + \cdots + X_K = Z$ and
- $S_1 + Z_1 = Z_1 + \cdots + Z_K \sim N(0, 1)$.
- Furthermore, for $j = 1, \ldots, K - 1$,

$$S_j + X_j = X_1 + \cdots + X_{j-1} + Z_j + \cdots + Z_K + X_j$$

$$= X_1 + \cdots + X_j + Z_{j+2} + \cdots + Z_K + Z_{j+1} = S_{j+1} + Z_{j+1}.$$
Choose $f \in C^3(R)$. To prove that $Z \xrightarrow{d} N(0,1)$ as $K \to \infty$ it’s enough to show that $E f(Z) \to E f(N(0,1))$ as $K \to \infty$. (Why?)

In light of the above remarks regarding $S_j$, $X_j$, and $Z_j$, we get the telescoping result that

$$E f(Z) - E f(N(0,1)) = E f(S_K + X_K) - E f(S_1 + Z_1)$$

$$= E f(S_1 + X_1) - E f(S_1 + Z_1)$$

$$+ E f(S_2 + X_2) - E f(S_2 + Z_2)$$

$$\vdots$$

$$+ E f(S_{K-2} + X_{K-2}) - E f(S_{K-2} + Z_{K-2})$$

$$+ E f(S_{K-1} + X_{K-1}) - E f(S_{K-1} + Z_{K-1})$$

$$+ E f(S_K + X_K) - E f(S_K + Z_K).$$

Then apply the triangle inequality followed by an application of FF3 with $X = S_j$, $Y = X_j$, and $W = Z_j$ to the telescoped sum above get that

$$|E f(Z) - E f(N(0,1)| \leq \sum_{j=1}^{K} |E f(S_j + X_j) - E f(S_j + Z_j)|$$

$$\leq C \sum_{j=1}^{K} E |X_j|^3 + C \sum_{j=1}^{K} E |Z_j|^3$$

Let’s call this last fact fun fact four, or FF4, for short.

We are now ready to prove the following result:

12.10.1 Lyapunov CLT For each $n$, let $Z_n = X_{n1} + \cdots + X_{n,K(n)}$ where the $X_{nj}$’s are independent with zero means and variances that sum to unity. If the Lyapunov condition, namely,

$$\sum_{j=1}^{K(n)} E |X_{nj}|^3 \to 0 \text{ as } n \to \infty$$

is satisfied, then $Z_n \xrightarrow{d} N(0,1)$.

12.10.2 Remark Before we prove this, note that the Standard CLT corresponds to $K(n) = n$ and $X_{nj} = \frac{X_j - \mu}{\sqrt{\sigma^2}}$ where the $X_j$’s are iid $(\mu, \sigma^2)$ random variables. This is a triangular array set-up where $\text{Var} X_{nj} = \frac{1}{n}$ for each $j$. Also, as an example of involving a more general “ragged” array, consider, for $j = 1, \ldots, K(n)$,

$$X_{nj} \sim \text{Binomial}(K(n), p) - K(n)p \left(\frac{\sum_{j=1}^{K(n)} K(n)p(1-p)}{K(n)p(1-p)}\right)^{1/2}$$

where, for each $n$, the $X_{nj}$ are independent and $\text{Var} X_{nj} = \frac{1}{K(n)}$. Note that in this case, the $n^{th}$ and the $m^{th}$ rows of random variables could be completely different, whereas in the standard set-up there is overlap: For example, in the standard set-up, row 5 uses $X_1, \ldots, X_5$ and row 6 uses $X_1, \ldots, X_5, X_6$. It is also interesting to consider the “ragged” binomial case where $p$ is replaced by $p_n = \frac{\lambda}{K(n)}$ for fixed $\lambda$. In this case, $p_n \to 0$ as $n \to \infty$ and there is a nonnormal limit. Do you know what it is?

Proof of Theorem 12.10.1: Fix $f \in C^3(R)$. We must show that as $n \to \infty$,

$$E f(z_n) \to E f(N(0,1)).$$
The replacement normal random variables are denoted $Z_{n1}, \ldots, Z_{nK(n)}$. The sum $Z_{n1} + \cdots + Z_{nK(n)}$ has a $N(0,1)$ distribution. Write $\sigma_{nj}^2$ for $\text{Var} \ X_{nj}$. Apply FF4 to get

$$|E f(z_n) - E f(N(0,1))| \leq C \sum_{j=1}^{K(n)} E |X_{nj}|^3 + C \sum_{j=1}^{K(n)} \sigma_{nj}^3 E |N(0,1)|^3.$$  

Note that

$$\sigma_{nj}^3 = (E X_{nj}^2)^{3/2} \leq E |X_{nj}|^3 \text{ why?}$$

(Jensen: The function $H(x) = x^{3/2}$, $x \geq 0$, is convex. Take $y = |X_{nj}|^2$. Then $H(E y) \leq E H(y)$. Cf. $(E y)^2 \leq E y^2$: Variance is nonnegative.)

It follows that the LHS is bounded by a multiple of $\sum_{j=1}^{K(n)} E |X_{nj}|^3$. Apply the Lyapunov condition to get the result.

Here are some questions to ponder regarding the result and its proof.

- Could $K(n) = 5$ for all $n$ and the Lyapunov condition both hold? No. Since $\sigma_{n1}^2 + \cdots + \sigma_{n5}^2 = 1$ at least one of the $\sigma_{nj}^2$ must be bounded away from zero. But then the corresponding $E |X_{nj}|^3$ will also be bounded away from zero, making it impossible for the LC to hold. Just apply Jensen to see that $(E |X_{nj}|^3)^{2/3} \geq E |X_{nj}|^2 = \sigma_{nj} > c > 0$. I.e., take $Y = |X_{nj}|^3$, $H(x) = x^{2/3}$, $x \geq 0$, concave. Then $H(E Y) \geq E H(Y)$.

In other words, must $K(n) \to \infty$ as $n \to \infty$ for the Lyapunov condition to hold? Yes.

- Must $K(n) \to \infty$ as $n \to \infty$ for the normality result to hold? No, take $X_{nj} = Z_{nj}$, $j = 1, 2, 3, 4, 5$.

- Must the $X_{nj}$’s be identically distributed? No.

- Must the $X_{nj}$’s be independent?

12.10.3 Exercise Verify the Lyapunov and Lindeberg conditions for the standard set-up $X_{nj} = \frac{X_j - \mu}{\sqrt{n} \sigma}$.

- Lyapunov: $\sum_{j=1}^{n} E |X_{nj}|^3 = \frac{1}{\sqrt{n}} E |X_1|^3 \to 0$.

- Lindeberg: $\sum_{j=1}^{n} E \left( |X_{nj}|^2 1_{(|X_{nj}| > \varepsilon)} \right) = E \left( |X_1|^2 1_{(|X_1| > \sqrt{n} \varepsilon)} \right) \to 0$ by the Dominated Convergence Theorem 6.12.2.

For simplicity, standardize: Take $\mu = 0$ and $\sigma^2 = 1$, so that $X_{nj} = \frac{X_j}{\sqrt{n}}$. Also, $K(n) = n$. □

12.10.4 Remark Lyapunov [14] weakens his third moment condition to $2 + \delta$ moments, $\delta > 0$.

The next result characterizes convergence in distribution of a sequence of random vectors as convergence in distribution of an arbitrary linear combination of the terms in the sequence. It is due to Harald Cramér and Herman Wold [7].

12.10.5 Theorem (Cramér–Wold device) Let $X_n$ and $X$ be random $K \times 1$ vectors. Then the Cramér–Wold device says that as $n \to \infty$,

$$X_n \xrightarrow{D} X$$

if and only if for every fixed $K \times 1$ vector $T$,

$$T'X_n \xrightarrow{D} T'X.$$
We prove this shortly using characteristic functions. First we look at a very important application of this general result.

For \( i = 1, \ldots, n \), consider independent and identically distributed random \( K \)-dimensional vectors \( U_i = (U_{i1}, \ldots, U_{iK})' \) with

\[
\begin{align*}
\mathbf{E} U_i &= (\mu_1, \ldots, \mu_K)' = \mu \\
\text{Var} U_i &= \mathbf{E}(U_i - \mu)(U_i - \mu)' = \Sigma.
\end{align*}
\]

Also, define

\[
X_n = (X_{n1}, \ldots, X_{nK})' = \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_{i1} - \mu_1), \ldots, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_{iK} - \mu_K) \right)'
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i - \mu).
\]

This implies that

\[
\begin{align*}
\mathbf{E} X_n &= 0 \\
V(X_n) &= \mathbf{E} X_n X_n' = \Sigma.
\end{align*}
\]

Now for a fixed, arbitrary \( T \in \mathbb{R}^K \),

\[
T'X_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} T'(U_i - \mu)
\]

\( \xrightarrow{D} N(0, T'S\Sigma T) \) (why?)

Let \( X \sim N(0, \Sigma) \). Then \( T'X \sim N(0, T'S\Sigma T) \). It follows that as \( n \to \infty \),

\[
T'X_n \xrightarrow{D} N(0, T'S\Sigma T) \sim T'X.
\]

Deduce from the Cramér–Wold device that \( X_n \xrightarrow{D} X \) as \( n \to \infty \).
12.11 ★ Other Proofs

Chapter 2 of Ross and Peköz [20] provides a very nice elementary, but long, proof based on constructing a new sequence of random variables (the Chen–Stein construction) with the same distribution as the sequence $X_1 + \cdots + X_n / \sqrt{n\sigma}$ in such a way that we can easily compute their distance from a standard normal.

The Fourier transform approach to the CLT is algebraic and does not give a lot of insight, but the following appendix gives an outline.

Appended

You are not responsible for what follows.

12.12 ★ The characteristic function of a distribution

Some of you may have taken ACM 95, in which case you have come across Fourier transforms. If you haven’t, you may still find this comment weakly illuminating. You may also want to look at Appendix 4.A.2 in Larsen–Marx [12].

You may know that the exponential function extends nicely to the complex numbers, and that for a real number $u$,

$$e^{iu} = \cos u + i \sin u.$$  

where of course $i = \sqrt{-1}$.

The characteristic function $\varphi_X : \mathbb{R} \to \mathbb{C}$ of the random variable $X$ is a complex-valued function defined on the real line by

$$\varphi_X(u) = E e^{iuX}.$$  

This expectation always exists as a complex number (integrate real and imaginary parts separately) since $|e^{ix}| = 1$ for all $u, x$.

12.12.1 Fact If $X$ and $Y$ have the same characteristic function, they have the same distribution.

12.12.2 Fact (Fourier Inversion Formula) If $F$ has a density $f$ and characteristic function $\varphi$, then

$$\varphi(u) = \int_{-\infty}^{\infty} e^{iux} f(x) \, dx$$  

and we have the inversion formula

$$f(x) = \int_{-\infty}^{\infty} e^{-iux} \varphi(u) \, du.$$
12.12.1 Characteristic Function of an Independent Sum

12.12.3 Fact Let $X, Y$ be independent random variables on $(\Omega, \mathcal{F}, P)$. Then
\[
\varphi_{X+Y}(u) = E(e^{iu(X+Y)}) = E(e^{iuX}e^{iuY}) = E(e^{iuX})E(e^{iuY}) \quad \text{by independence} = \varphi_X(u)\varphi_Y(u).
\]

12.12.2 Characteristic Function of a Normal Random Variable

12.12.4 Fact Let $X$ be a standard normal random variable. Then
\[
\varphi_X(u) = E(e^{iuX}) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{iuX} e^{-\frac{1}{2}x^2} dx = e^{-\frac{u^2}{2}}.
\]
(This is not obvious.)

12.13 ⋆ The characteristic function and convergence in distribution

12.13.1 Theorem

\[
F_n \xrightarrow{D} F \iff \varphi_n(u) \to \varphi(u) \quad \text{for all } u.
\]


12.14 ⋆ The characteristic function and the CLT

Sketch of proof of the CLT: We consider the case where $E X_i = 0$. (Otherwise subtract the mean everywhere.)

Compute the characteristic function $\varphi_n$ of $\frac{X_1 + \cdots + X_n}{\sqrt{n}\sigma}$.

\[
\varphi_n(u) = E(e^{iu\frac{X_1 + \cdots + X_n}{\sqrt{n}\sigma}}) = E\left(\prod_{k=1}^{n} e^{iu\frac{X_k}{\sqrt{n}\sigma}}\right) = \left[E(e^{iu\frac{X_1}{\sqrt{n}\sigma}})^n\right] \quad \text{(identical distribution)}
\]

\[
= \left[E\left(1 + iu\frac{X_1}{\sqrt{n}\sigma} + \frac{1}{2} \left(iu\frac{X_1}{\sqrt{n}\sigma}\right)^2 + o\left(\frac{(iuX_1)^2}{\sqrt{n}\sigma}\right)\right)^n\right] \quad \text{(Taylor’s Theorem)}
\]

\[
= \left[1 + 0 - \frac{u^2\sigma^2}{2n\sigma^2} + o\left(\frac{-u^2}{2n}\right)\right]^n \quad \text{(} E X = 0, \ E X^2 = \sigma^2) \]

Now use the well-known fact: $\lim_{x \to 0}(1 + ax + o(bx))^\frac{1}{x} \to e^a$, so as $n \to \infty$, $\varphi_n(u) \to e^{-\frac{u^2}{2}}$, which is the characteristic function of $N(0, 1)$.

KC Border  v. 2020.10.21::10.28
Proof of the well-known fact:

\[
\ln \left( \left(1 + ax + o(bx) \right)^{\frac{1}{x}} \right) = \frac{1}{x} \ln \left(1 + ax + o(bx) \right)
= \frac{1}{x} \left[ \ln (1 + ax \ln'(1) + o(x)) \right]
= \frac{1}{x} \left[ ax + o(x) \right]
= a + \frac{o(x)}{x} \to a \text{ as } x \to 0.
\]

Therefore \( e^{(1 + ax + \gamma(bx))^{\frac{1}{x}}} \to e^a \text{ as } x \to 0. \)

Bibliography


