The *Law of Large Numbers* says roughly this:

> The probability distribution of practically any random variable can be determined to any desired degree of accuracy as nearly certainly as desired by sampling that random variable independently and often enough.

First this note applies Chebyshev's Inequality to justify the Law of Large Numbers in a typical special case. Next comes a description of the Central Limit Theorem, which is proved valid in a very special case. Sometimes this Theorem is confused with the Law of Large Numbers; both ideas are important for most practical applications of probability.

Consider a random variable x distributed over a given finite population of individuals i or sample space of "outcomes" i. Actually x is a function that takes the value x_i at i, which will occur or be chosen at random with probability prob(i). The simplest nontrivial random variable is just x_i := prob(i). Often prob(i) is the same for all individuals i but we shall not assume this; we do take for granted, as usual, that every prob(i) ≥ 0 and that $\sum_i$ prob(i) = 1. More generally, the probability that any preassigned subset **$** of the population will include a not-yet-specified individual chosen at random from the whole population is $\sum_{i \text{ in } \$}$ prob(i).

The *Mean, Average* or *Expected* value of x over the whole space is denoted in this note by

$$Æ(x) := \sum_i \text{prob}(i)\cdot x_i = \sum_X \text{Probability}(x = X)\cdot X .$$

The last sum is over all values $X = x_i$ that x takes in the given population. This note concerns the estimation of Æ(x), given function x, without knowing those probabilities in advance.

To know the probability distribution of x is to know Probability(x = X) for every number X or, more usefully, to know Probability(X ≤ x ≤ X+Δ) for every X and Δ ≥ 0. A way to estimate this latter probability, given X and Δ, is to define another random variable y thus:

$$y_i := 1 \text{ if } X \le x_i \le X+\Delta ; \text{ otherwise } y_i := 0 .$$

Then Æ(y) = Probability(X ≤ x ≤ X+Δ); can you see why? This is why we wish to know how to estimate Æ(…) in general, not merely for one random variable x. In other words, Æ(…) is a *functional*, a function whose explicit argument is a function (and whose implicit argument is a population or sample space); Æ maps functions (defined over populations) to numbers.

Æ is a *linear* functional in the following sense: If x and y are two random variables over a population, and if µ and ß are constants (each taking just one value over the population), then Æ(µ·x + ß·y) = µ·Æ(x) + ß·Æ(y); can you see why? More generally, however, for an arbitrary function $f(x, y)$ we almost always find that $Æ(f(x, y)) \ne f(Æ(x), Æ(y))$. This is why the estimation of Æ can be technically challenging.

Random variables x and y are called " Statistically Independent " (or just "Independent") if

$$\text{Probability}(x = X \text{ and } y = Y) = \text{Probability}(x = X)\cdot\text{Probability}(y = Y)$$

for all constants X and Y, in which case Æ(x·y) = Æ(x)·Æ(y); can you see why? But when x and y are not independent, Æ(x·y) – Æ(x)·Æ(y) equals something called the *covariance* of x and y, as we shall see later. Thus, multiplication of random variables is quite different from addition because Æ(x+y) = Æ(x)+Æ(y) regardless of independence. Whether random variables are independent is always important though sometimes difficult to ascertain.

Random Sampling

Suppose we plan to select (but not remove) an individual, to be called I, from the given population. This I will be a *Random Sample* if Probability$(I = i)$ = prob(i). Similarly for another random sample J. Then the two samples will be regarded as statistically independent if

$$\text{Probability}(I = i \text{ and } J = j) = \text{prob}(i)\cdot\text{prob}(j) .$$

The future tense is used here because the word "random" may be inappropriate to describe a sample J after it has been selected. Moreover, nobody knows how to choose samples that are *perfectly* random and independent although some pretty good approximations are known; the art of systematic random sampling is a topic discussed in Statistics courses, especially courses about the *Design of Experiments*. The generation of good *pseudo-random* numbers is treated in vol. II of D.E. Knuth's *The Art of Computer Programming* (Addison-Wesley).

For example, tossing a coin has two outcomes, heads and tails, that are sampled ostensibly at random every time the coin is tossed; however, the outcomes can be both biased and correlated if the tosser repeats too accurately his motions for each toss. A casting director chooses extras for a movie's crowd scene not by sampling them at random from whoever is available, but rather by correlating her selections to ensure that the crowd looks more nearly "representative" of the population intended by the script-writer. Japanese flower arrangements look random only if some artistry goes into their placement.

So, random sampling is hypothetical if not mythical. And to the extent that individuals can be sampled at random, so can a random variable x; we shall let $X := x_I$ denote the sample-value of x obtained from individual I sampled at random. We shall contemplate large numbers n of random samples $X_1, X_2, X_3, \ldots, X_n$ of random variable x corresponding respectively to individuals $I_1, I_2, I_3, \ldots, I_n$ to be selected (but not removed) at random and independently from a population. And then we shall compare several *statistics* :

$\bar{x} := Æ(x) =$ the *mean* of x over the population, with

$Æ(\overline{X})$ where $\overline{X} := (X_1 + X_2 + X_3 + \ldots + X_n)/n =$ the samples' mean; and

$\sigma^2 := Æ((x - \bar{x})^2) =$ the *variance* of x over the population, with

$Æ(S^2)$ where $S^2 := ((X_1-\overline{X})^2 + (X_2-\overline{X})^2 + \ldots + (X_n-\overline{X})^2)/n =$ the samples' variance.

Note that, at least until the samples have been drawn, each of $X_1, X_2, X_3, \ldots, X_n$ is a random variable distributed the same way as x is. Consequently $\overline{X}$ and S^2 are random variables too, but over a population of n-tuples $(I_1, I_2, I_3, \ldots, I_n)$ composed from the n-fold Cartesian product of the given population with itself. Until the samples have been drawn,

$$\text{Probability}((I_1, I_2, I_3, \ldots, I_n)) = \text{prob}(I_1)\cdot\text{prob}(I_2)\cdot\text{prob}(I_3)\cdot\ldots\cdot\text{prob}(I_n)$$

because the samples are independent. It follows that each of $X_1, X_2, X_3, \ldots, X_n$ is independent of all others (why?) so $Æ(X_k\cdot X_m) = Æ(X_k)\cdot Æ(X_m) = Æ(x)\cdot Æ(x) = \bar{x}^2$ for $1 \le k < m \le n$. But every $Æ(X_k^2) = Æ(x^2)$ generally differs from $\bar{x}^2$, as we shall see. Moreover, random variables $\overline{X}$ and S^2 are not generally independent of each other nor of the samples-to-be X_k.

Let's digress for a moment to consider two random variables x and y that are not necessarily independent. Analogous to $\bar{x}$ and σ^2 are the statistics

$\bar{y} := Æ(y) =$ the mean of y over the given population, and
$\tau^2 := Æ((y - \bar{y})^2) =$ the variance of y over the population.
To these we add the statistic
$\gamma := Æ((x-\bar{x})\cdot(y-\bar{y})) =$ the *covariance* of x and y over the given population.
In the very special case that x and y are independent we find that
$\gamma = Æ(x\cdot y - \bar{x}\cdot y - x\cdot\bar{y} + \bar{x}\cdot\bar{y}) = \bar{x}\cdot\bar{y} - \bar{x}\cdot\bar{y} - \bar{x}\cdot\bar{y} + \bar{x}\cdot\bar{y} = 0$.
In general $\gamma \neq 0$ except by accident; and it is possible to find $\gamma = 0$ even though x and y are not independent. In all cases, $Æ(x\cdot y) = \bar{x}\cdot\bar{y} + \gamma$; can you see why? Moreover $\gamma^2 \le \sigma^2\cdot\tau^2$; to see why, consider the discriminant of a quadratic $Æ((\mu\cdot(y-\bar{y}) - (x-\bar{x}))^2) \ge 0$ for all real μ .

About terminology: *Variances* σ^2 and τ^2 are the squares respectively of *standard deviations* σ and τ which are nonnegative by convention. And here "independent" can mean "Pairwise Independent" because this note allows for ostensibly bizarre situations like when, of three random variables x, y and z , every two are independent though those two determine the third.

Now the expected values of a few sample statistics can be computed for comparison with population statistics:

$Æ(\bar{X}) = Æ(\sum_k X_k)/n = \sum_k Æ(X_k)/n = \sum_k Æ(x)/n = n\cdot\bar{x}/n = \bar{x}$, and

$Æ((\bar{X} - Æ(\bar{X}))^2) = Æ((\sum_k (X_k - \bar{x})/n)^2) = Æ(\sum_k \sum_m (X_k - \bar{x})\cdot(X_m - \bar{x}))/n^2$

$= (n\cdot\sigma^2 + 0)/n^2$ (since X_k and X_m are independent if $k \neq m$)

$= \sigma^2/n$.

Therefore, as a random variable, the samples' mean $\bar{X}$ has the same mean $\bar{x}$ as has x over the whole population. But $\bar{X}$ has a variance σ^2/n smaller than the population's variance σ^2 , which is *NOT* the same as the samples' variance S^2 though they are close enough to justify the Law of Large Numbers, as we shall see later. First we digress to …

Chebyshev's Inequality: If a random variable x has mean $\bar{x}$ and standard deviation σ , then

Probability($|x - \bar{x}| \ge \sigma/\lambda$) $\le \lambda^2$ for every positive $\lambda < 1$.

Proof: Let **$** be that subset of the population's individuals i for which $|x_i - \bar{x}| \ge \sigma/\lambda$. Then

$\sigma^2 = \sum_{\text{all } i} \text{prob}(i)\cdot(x_i - x)^2$

$\ge \sum_{i \text{ in } \$} \text{prob}(i)\cdot(x_i - \bar{x})^2 \ge \sum_{i \text{ in } \$} \text{prob}(i)\cdot(\sigma/\lambda)^2$

$= (\sigma^2/\lambda^2)\cdot$Probability($|x - \bar{x}| \ge \sigma/\lambda$) . Divide by σ^2/λ^2 to finish the proof.

Chebyshev's Inequality tends to be extremely pessimistic because Probability($|x - \bar{x}| \ge \sigma/\lambda$) is almost always very much tinier than λ^2 . Without additional information about x this λ^2 cannot be replaced by something smaller because there are random variables x that satisfy Probability($|x - \bar{x}| \ge \sigma/\lambda$) $= \lambda^2$ for at least one $\lambda > 0$. For example suppose x takes only three values, namely $x = \pm 1$ each with probability $\lambda^2/2$, and $x = 0$ with probability $1 - \lambda^2$; then $\bar{x} = 0$, $\sigma = \lambda$, and Probability($|x - \bar{x}| \ge \sigma/\lambda$) $= \lambda^2$ exactly. Later we shall see how pessimistic Chebyshev's Inequality is; for now it is adequate to prove …

The Law of Large Numbers: If a random variable x has mean $\bar{x}$ and standard deviation σ then, given any two tiny positive tolerances μ and $ß$, choosing a big $n > (\sigma/\mu)^2/ß$ will ensure that the samples' mean $\overline{X}$, of n independent random samples of x, differs from the population's mean $\bar{x}$ by less than μ except with probability smaller than $ß$.

Proof: For any $n > (\sigma/\mu)^2/ß$ set $\lambda := \sigma/(\mu\cdot\sqrt{n}) < \sqrt{ß}$ to infer that Probability($|\overline{X}-\bar{x}| \geq \mu$) $< ß$ from Chebyshev's inequality because the standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$. End of proof.

The Law of Large Numbers is often misapplied. For example consider a large number n of fair tosses of a fair coin just as likely to come up Heads as Tails. The expected number of each is $n/2$, from which some people wrongly infer that the difference between the numbers of Heads and Tails is likely to be small, and more likely as n increases. If the coin has come up Heads rather more often then Tails for a while, these people would bet that Tails are more likely to appear in the next several tosses. Not so! Even if the tosses are perfectly fair, that difference can be proved almost certainly bigger than any big number chosen in advance, while the ratio of the numbers of Heads and Tails is almost certain to differ from 1 by less than any tiny positive number chosen in advance, provided the number n of tosses is chosen big enough in advance. Choosing n in advance is obligatory lest the Law of Large Numbers, as stated above, be violated. It is violated when n is chosen by drawing ever more samples until a tolerance is exceeded, and stopping then. No matter how unlikely this stopping event may be, unless it is impossible it will surely occur at least once if Fate is tempted often enough.

Any application of the foregoing Law of Large Numbers to estimate the mean $\bar{x}$ of x uses an estimate of the variance σ^2 of x to decide how big the sample size n should be; but if $\bar{x}$ is not yet known where can an estimate of σ^2 come from? From the samples' variance S^2? Not exactly. First, until the samples have been drawn, S^2 is a random variable. Second, it is likely to somewhat underestimate σ^2; in other words, S^2 is a *statistically biased estimator* of σ^2. More precisely, as shall be proved next,

$$Æ(S^2) = (1 - 1/n)\cdot\sigma^2 .$$

Lemma: If independent random variables y_j all have mean $Æ(y_j) = 0$ and respective variances $Æ(y_j^2) = \tau_j^2$, then $Æ((\sum_j y_j)^2) = \sum_j \tau_j^2$.

Proof: $Æ((\sum_j y_j)^2) = Æ(\sum_k \sum_j y_k\cdot y_j) = \sum_k \sum_j Æ(y_k\cdot y_j) = \sum_j Æ(y_j^2) + \sum_k \sum_{j\neq k} 0 = \sum_j \tau_j^2$.

Now set every $y_j := -(X_j-\bar{x})$ except $y_k := (n-1)\cdot(X_k-\bar{x})$ for any positive $k \leq n$ to find that

$$Æ((n\cdot X_k - n\cdot\overline{X})^2) = Æ((y_k + \sum_{j\neq k} y_j)^2) = (n-1)^2\cdot\sigma^2 + (n-1)\cdot\sigma^2 = n\cdot(n-1)\cdot\sigma^2 .$$

Consequently

$$n^3\cdot Æ(S^2) = Æ(\sum_k (n\cdot X_k - n\cdot\overline{X})^2) = \sum_k Æ((n\cdot X_k - n\cdot\overline{X})^2) = n^2\cdot(n-1)\cdot\sigma^2 .$$

Divide by n^3 to conclude that $Æ(S^2) = (1 - 1/n)\cdot\sigma^2$ as claimed above. End of proof.

An initial batch of n samples could be drawn to provide an estimate $S^2/(1 - 1/n)$ of σ^2 after which at least $(\sigma/\mu)^2/ß$ new samples would very likely estimate $\bar{x}$ adequately; but these are almost always far too many new samples because Chebyshev's inequality is so pessimistic.

The Central Limit Theorem Summarized

Its proof is difficult, not for everybody. It says something astonishing, roughly this:

> Practically regardless of how the random variable x is distributed, there is one universal *Normal Distribution* by which the probability distribution of the samples' mean $\overline{X}$ comes to be approximated ever better as the number n of samples increases.

This Normal Distribution is characterized by a function $\Phi(z)$ that increases smoothly from $\Phi(-\infty) := 0$ through $\Phi(0) = 1/2$ to $\Phi(+\infty) = 1$ with a derivative $\Phi'(z) := \exp(-z^2/2)/\sqrt{(2\pi)}$. The graph of $\Phi(z)$ is its own reflection in its midpoint; $\Phi(-z) + \Phi(z) = 1$. And as $z \to +\infty$, $\Phi(\pm z)$ approaches its limits $\Phi(\pm\infty)$ extremely rapidly; for every $z > 0$ it can be proved that

$$0 < \Phi'(z)/(z + 1/z) < \Phi(-z) = 1 - \Phi(z) < \Phi'(z)/z \ .$$

Tables of values of $\Phi(z)$ are available widely, especially in Statistics texts. Physicists more often use the *Error Function* $\mathrm{erf}(z) = 2\Phi(z\sqrt{2})-1$. Computer programs based upon continued fractions or other formulas can compute $\Phi(z)$ as accurately as need be though not so quickly as we would like; but it has been proved that no formula that invokes algebraic operations ($+, -, \cdot, /, \sqrt{}$) and elementary transcendental functions like exp, ln, tan, arctan, … only finitely often can compute $\Phi(z)$ exactly. $\Phi(z)$ and its derivative $\Phi'(z)$ are plotted on the next page.

We say a random variable u is *Distributed Normally* with mean $\bar{u}$ and variance v^2 just when

$$\text{Probability}(\, u \le U \,) = \Phi((U-\bar{u})/v) \quad \text{for all real } U \text{ or, equivalently,}$$
$$\text{Probability}(\, \ddot{U} < u \le \hat{U} \,) = \Phi((\hat{U}-\bar{u})/v) - \Phi((\ddot{U}-\bar{u})/v) \quad \text{whenever } \ddot{U} < \hat{U} \ .$$

It turns out that $\overline{X}$ is distributed approximately Normally with mean $\bar{x}$ and variance σ^2/n ;

$$\text{Probability}(\, \overline{X} \le U \,) \approx \Phi((U-\bar{x})/(\sigma/\sqrt{n})) \ ,$$

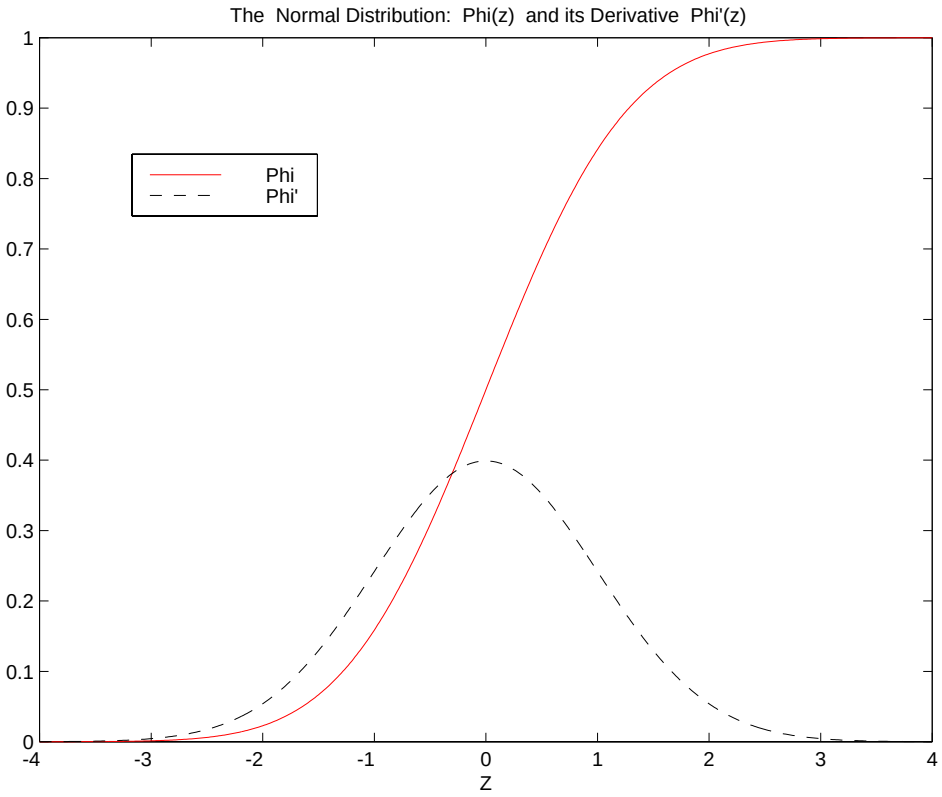
and this approximation improves as n increases. However, out on the *tails* of the distribution where $|\overline{X}-\bar{x}|/(\sigma/\sqrt{n})$ exceeds 3 or 4 , the approximation improves so slowly that its use to estimate tiny probabilities of extreme departures from the mean is imprudent. An appropriate use for the Central Limit Theorem is to estimate where the values of $\overline{X}$ are most likely to be found; this estimate depends upon n, $\bar{x}$ and σ but is otherwise affected little by the way x is distributed.

For example let us estimate $p_3 := \text{Probability}(\, |\overline{X}-\bar{x}| < 3\cdot\sigma/\sqrt{n} \,)$. Chebyshev's inequality implies $p_3 > 1 - 1/3^2 = 0.8888\ldots$, but this underestimates p_3 substantially when n is big, in which case the Central Limit Theorem implies that $p_3 \approx \Phi(3) - \Phi(-3) = 0.9973\ldots$.

More generally, for $k = 1, 2, 3, \ldots$ let y_k be *mutually* independent random variables (the probability of each is unaffected by whatever may be known about all others) with respective means $\bar{y}_k$ and standard deviations τ_k , and let $y := y_1 + y_2 + y_3 + \ldots + y_n$ for some large n . This y has mean $\bar{y} := \bar{y}_1 + \bar{y}_2 + \bar{y}_3 + \ldots + \bar{y}_n$ and variance $\tau^2 := \tau_1^2 + \tau_2^2 + \tau_3^2 + \ldots + \tau_n^2$ according to the Lemma above. Provided $\max_{1\le k\le n} \tau_k^2/\tau^2 \to 0$ as $n \to +\infty$, the Central Limit Theorem says that $(y-\bar{y})/\tau$ is distributed ever more nearly Normally with mean 0 and variance 1 as n increases. Again, the approximation is best for central tendencies but remains relatively inaccurate out on the tails.

For proofs see W. Feller's *An Introduction to Probability Theory and its Applications* vol. II 2d ed. (1971, Wiley); it is heavy reading. A comparatively elementary treatment of a special case appears in the following pages.

Graph of the Normal Distribution: $\Phi(z)$ and its Derivative $\Phi'(z)$.



The Demoivre-Laplace Limit Theorem

Now we shall demonstrate the Central Limit Theorem's validity in the special case of a huge number n of fair and independent tosses of a fair coin. This is definitely not for everybody!

This case was discussed in 1718 by Abraham DeMoivre, a Huguenot who had fled to England from France because Louis XIV revoked in 1685 the religious tolerance promulgated by Henri IV's Edict of Nantes in 1598. In 1812 DeMoivre's discussion was refined by Pierre Simon Laplace, whose name too is now attached to this demonstration. Many texts and notes exhibit flawed versions of this demonstration; I hope this one isn't flawed too.

Let random variable h_n count how many heads will appear after n independent and fair tosses of a fair coin as likely (probability = 1/2) to come up heads as tails. We already know that

$$\text{Probability}(\, h_n = k \,) \;=\; {}^nC_k/2^n$$

(where combinatorial coefficient ${}^nC_k = n!/(k!\cdot(n-k)!)$) for all integers k provided we accept the convention that ${}^nC_k = 0$ whenever $k < 0$ or $k > n$. Also known is that h_n has mean $\bar{h}_n := Æ(h_n) = n/2$ and standard deviation $v_n := \sqrt{Æ((h_n-\bar{h}_n)^2)} = \sqrt{n}/2$; see our text's Example 22 on pp. 281-2. To confirm the Central Limit Theorem we must prove that, as n tends

towards $+\infty$, the random variable $u_n := (h_n - \bar{h}_n)/v_n$ becomes distributed ever more nearly like a normal random variable with mean 0 and variance 1, which means the approximation

$$\text{Probability}(\ \ddot{U} < u_n \le \hat{U}\ ) \approx \Phi(\hat{U}) - \Phi(\ddot{U}) \quad \text{whenever } \ddot{U} < \hat{U}$$

becomes ever more accurate as n increases *provided* $\ddot{U}$ and $\hat{U}$ are fixed first.

The function

$$\begin{aligned} F_n(U) :=\ & \text{Probability}(\ u_n \le U\ ) = \text{Probability}(\ h_n \le \bar{h}_n + v_n\cdot U\ ) \quad \dots \text{ but } h_n \text{ is an integer} \\ =\ & \text{Probability}(\ h_n \le \lfloor \bar{h}_n + v_n\cdot U \rfloor\ ) = \text{Probability}(\ u_n \le (\lfloor \bar{h}_n + v_n\cdot U \rfloor - \bar{h}_n)/v_n\ ) \\ =\ & F_n(\ (\lfloor \bar{h}_n + v_n\cdot U \rfloor - \bar{h}_n)/v_n\ )\ . \end{aligned}$$

In other words, $F_n(U)$ is for each $n > 0$ a nondecreasing step-function of U determined by its values at regularly spaced discrete arguments $U = U_{n,k} := (k - \bar{h}_n)/v_n$ for all integers k. Since $F_n(U) = 0$ for all $U < -\sqrt{n}$ (do you see why?), the set of all differences

$$\begin{aligned} F_n(U) - F_n(U - 1/v_n) =\ & \text{Probability}(\ U - 1/v_n < u_n \le U\ ) \\ =\ & \text{Probability}(\ \lfloor \bar{h}_n + v_n\cdot U \rfloor - 1 < h_n \le \lfloor \bar{h}_n + v_n\cdot U \rfloor\ ) \\ =\ & \text{Probability}(\ h_n = \lfloor \bar{h}_n + v_n\cdot U \rfloor\ ) \\ =\ & {}^nC_k/2^n \quad \text{where integer } k = \lfloor \bar{h}_n + v_n\cdot U \rfloor \end{aligned}$$

determines F_n too by a telescoping sum: $F_n(U) = \sum_{j\ge 0}\ (\ F_n(U - j/v_n) - F_n(U - (j+1)/v_n)\ )$.

Thus our strategy is to deduce the approximation $F_n(U) \approx \Phi(U)$ from a proof that the differenced approximation $F_n(U) - F_n(U - 1/v_n) \approx \Phi(U) - \Phi(U - 1/v_n)$ has high *relative* accuracy if n is big enough. But as $n \to +\infty$ these differences tend to zero since $1/v_n = 2/\sqrt{n} \to 0$; to remedy that we divide by $1/v_n$ and find that $(\ \Phi(U) - \Phi(U - 1/v_n)\ )/(1/v_n) \to \Phi'(U) > 0$ as $n \to +\infty$. This simplifies our strategy, reducing our task to the proof that also, for any fixed U,

$$(\ F_n(U) - F_n(U - 1/v_n)\ )/(1/v_n) \to \Phi'(U) \quad \text{as } n \to +\infty\ .$$

Recall that $F_n(U) - F_n(U - 1/v_n) = {}^nC_k/2^n$ where integer $k = \lfloor \bar{h}_n + v_n\cdot U \rfloor$. For any fixed U this integer $k = \lfloor \bar{h}_n + v_n\cdot U \rfloor = \lfloor (n + U\cdot\sqrt{n})/2 \rfloor$ increases somewhat irregularly as n increases. To attenuate that irregularity we define $u := (k - \bar{h}_n)/v_n = 2(\lfloor (n + U\cdot\sqrt{n})/2 \rfloor - n/2)/\sqrt{n}$, which is designed to satisfy $k = (n + u\cdot\sqrt{n})/2 = \lfloor (n + U\cdot\sqrt{n})/2 \rfloor$ with $U - 2/\sqrt{n} < u \le U$. Clearly $u \to U$ as $n \to +\infty$, and now

$$\begin{aligned} (\ F_n(U) - F_n(U - 1/v_n)\ )/(1/v_n) =\ & {}^nC_k\cdot\sqrt{n}/2^{n+1} = n!\cdot\sqrt{n}/(k!\cdot(n-k)!\cdot 2^{n+1}) \\ =\ & n!\cdot\sqrt{n}/(((n + u\cdot\sqrt{n})/2)!\cdot((n - u\cdot\sqrt{n})/2)!\cdot 2^{n+1})\ . \end{aligned}$$

Now is the time to invoke Stirling's Approximation $n! \approx \sqrt{2\pi\cdot n}\cdot(n/e)^n$ (proved in the class notes on *Some Inequalities*) at three places; after a lot of algebraic simplification we find

$$(\ F_n(U) - F_n(U - 1/v_n)\ )/(1/v_n) \approx (1 - u/\sqrt{n})^{u\sqrt{n}/2}/(\sqrt{2\pi}\cdot(1 - u^2/n)^{(n+1)/2}\cdot(1 + u/\sqrt{n})^{u\sqrt{n}/2})\ .$$

Calculus classes teach that if $t \to T$ as $K \to \pm\infty$ then $(1 + t/K)^K \to \exp(T) = e^T$, which implies here that, as $n \to +\infty$,

$(1 - u/\sqrt{n})^{u\sqrt{n}/2} \to \exp(-U^2/2)$, $(1 - u^2/n)^{(n+1)/2} \to \exp(-U^2/2)$, $(1 + u/\sqrt{n})^{u\sqrt{n}/2} \to \exp(U^2/2)$, and consequently, as claimed above, for any U fixed in advance, …

$$\left(F_n(U) - F_n(U - 1/v_n) \right)/(1/v_n) \;=\; {}^nC_k\cdot\sqrt{n}/2^{n+1} \;\to\; \exp(-U^2/2)/\sqrt{2\pi} \;=\; \Phi'(U)\,.$$

Thus the differenced approximation $F_n(U) - F_n(U - 1/v_n) \approx \Phi(U) - \Phi(U - 1/v_n)$ has now been proved for each U to have arbitrarily high relative accuracy if n is big enough; in other words,

$$\left(F_n(U) - F_n(U - 1/v_n)\right)/\left(\Phi(U) - \Phi(U - 1/v_n)\right) \;\to\; 1 \quad \text{as} \quad n \to +\infty$$

even though numerator and denominator of the ratio $(\ldots)/(\ldots)$ both approach 0. Convergence to 1 has been proved on the assumption that U is fixed *before* $n \to +\infty$; the proof and its conclusion are invalid if U is allowed to vary too wildly with n. Still, theorems about *Uniform Convergence* taught in *Real Analysis* classes say that, given any *finite* interval $Ü \le U \le Û$ in advance, we can keep $\left|\left(F_n(U) - F_n(U - 1/v_n)\right)/\left(\Phi(U) - \Phi(U - 1/v_n)\right) - 1\right|$ as tiny as we like for all values U in that interval simultaneously by taking n big enough. This fact and one more will be needed to complete the proof of the validity of the Central Limit Theorem for fair coin tosses.

Because $F_n(U)$ is a nondecreasing step-function of U, we find whenever $0 < Û - Ü < 1/v_n$ that

$$0 \le F_n(Û) - F_n(Ü) \le F_n(Û) - F_n(Û - 1/v_n) \le \max_k {}^nC_k/2^n \approx \Phi'(0)/v_n \to 0 \quad \text{as} \quad n \to +\infty$$

and similarly $0 < \Phi(Û) - \Phi(Ü) < \Phi'(0)/v_n \to 0$.

Now choose any finite interval $Ü \le U \le Û$, and set integer $J := \lfloor (Û - Ü)\cdot v_n \rfloor = \lfloor (Û - Ü)\cdot\sqrt{n}/2 \rfloor$. This J grows with $\sqrt{n}$, but always $0 \le Û - J/v_n - Ü < 1/v_n$, so $F_n(Û - J/v_n) - F_n(Ü) \to 0$ and $\Phi(Û - J/v_n) - \Phi(Ü) \to 0$ as $n \to +\infty$. Meanwhile, because the relative error in every sum of positive terms is no worse than the worst relative error in any term, the approximation

$$\begin{aligned} F_n(Û) - F_n(Û - J/v_n) \;&=\; \textstyle\sum_{0\le j<J} \left(F_n(Û - j/v_n) - F_n(Û - (j+1)/v_n) \right) \\ &\approx\; \textstyle\sum_{0\le j<J} \left(\Phi(Û - j/v_n) - \Phi(Û - (j+1)/v_n) \right) \;=\; \Phi(Û) - \Phi(Û - J/v_n) \end{aligned}$$

can be made as accurate as we please by taking n big enough. Add to both sides their respective vanishing differences $F_n(Û - J/v_n) - F_n(Ü)$ and $\Phi(Û - J/v_n) - \Phi(Ü)$ to finish the proof that

$$F_n(Û) - F_n(Ü) \to \Phi(Û) - \Phi(Ü) \quad \text{as} \quad n \to +\infty\,.$$

Appendix: To estimate how quickly $\Phi(u)$ decays as $u \to -\infty$, we compute

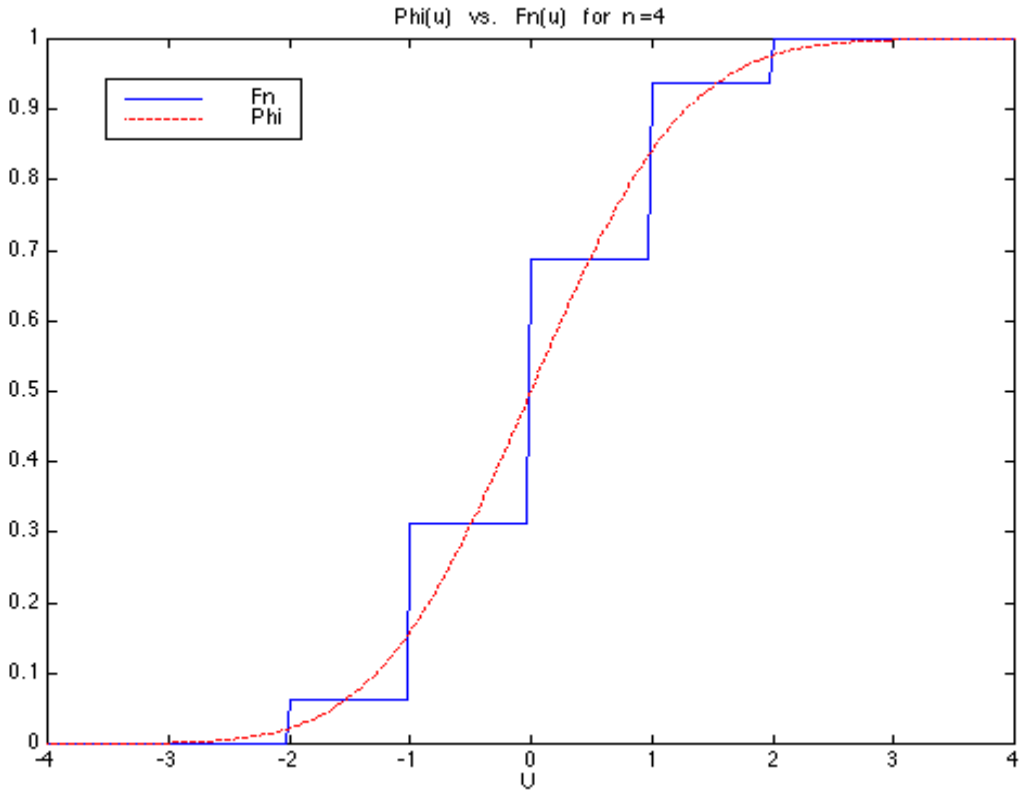
$$\begin{aligned} 0 < \textstyle\int_{-\infty}^{u} \Phi(v)\cdot dv \;&=\; \textstyle\int_{-\infty}^{u} \int_{-\infty}^{v} \Phi'(t)\cdot dt\cdot dv \;=\; \int_{-\infty}^{u} \int_{t}^{u} \Phi'(t)\cdot dv\cdot dt \;=\; \int_{-\infty}^{u} (u - t)\cdot\Phi'(t)\cdot dt \\ &=\; u\cdot\Phi(u) + \Phi'(u)\,) \qquad (\text{ because } \; d\Phi'(t)/dt = -t\cdot\Phi'(t) \;\text{ and }\; \Phi'(-\infty) = \Phi(-\infty) = 0\,), \end{aligned}$$

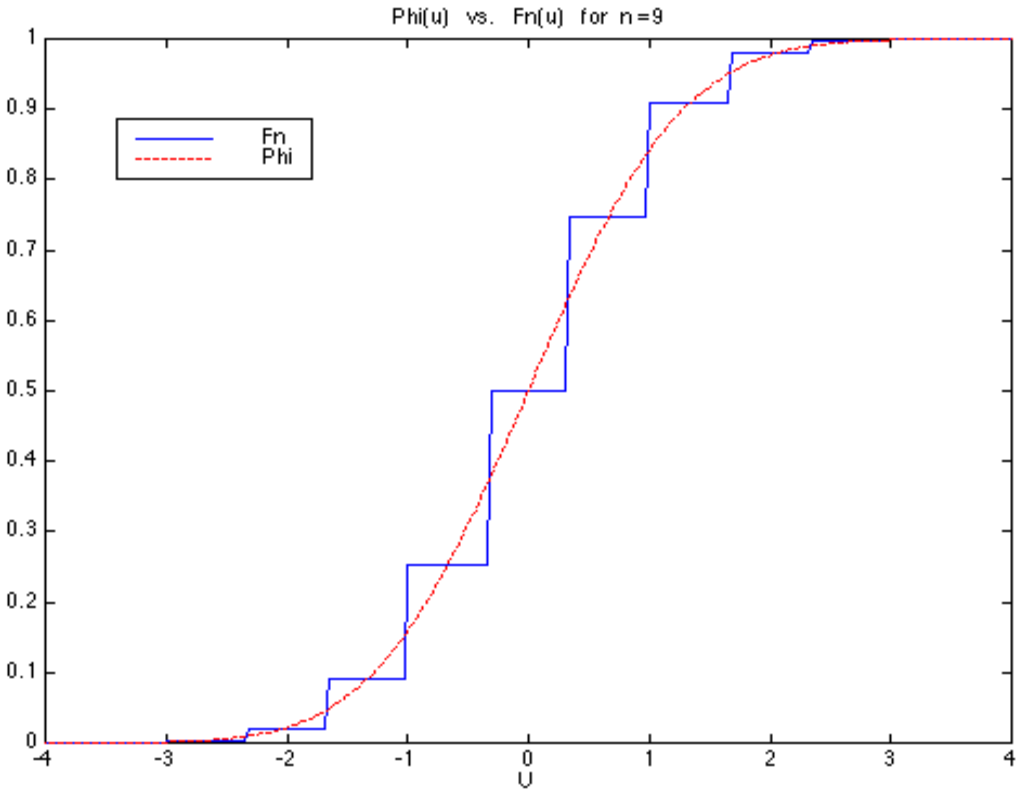
and infer that $0 < \Phi(u) < \Phi'(u)/(-u) = \exp(-u^2/2)/(-u\cdot\sqrt{2\pi})$ so long as $-u > 0$. An analogous estimate can be obtained for $F_n(U)$ by observing that ${}^{n-1}C_k - {}^{n-1}C_{k-1} = (1 - 2k/n)\cdot{}^nC_k$ and then $0 \le 2\sum_{i<k}\sum_{0\le j\le i} {}^nC_j \;=\; (2k - n)\cdot\sum_{j<k} {}^nC_j + k\cdot{}^nC_k$, whence $\sum_{j<k} {}^nC_j \le (k/(n - 2k))\cdot{}^nC_k$ so long as $k < n/2$. (Can you carry out the algebra?) Consequently, so long as $\sqrt{n} >> -u > 0$,

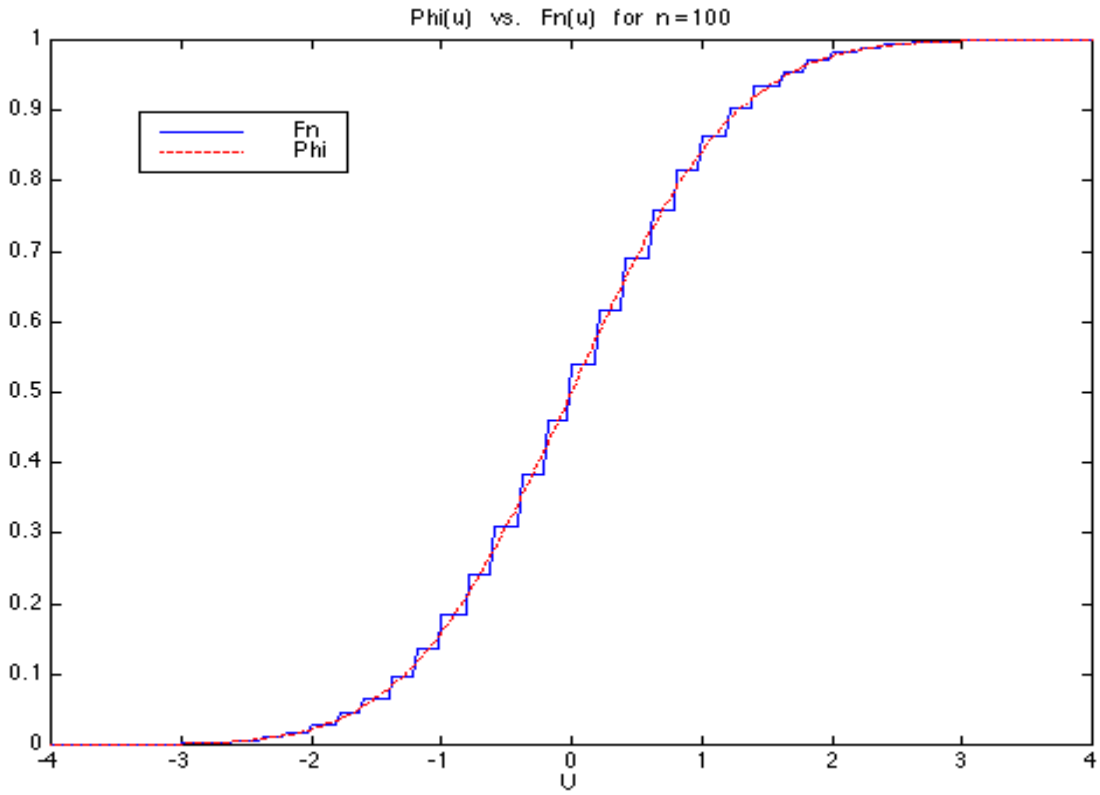
$$F_n(u - 1/v_n) \;\le\; \left((1 + u/\sqrt{n})/(-u)\right)\cdot\left(F_n(u) - F_n(u - 1/v_n) \right)/(1/v_n) \;\approx\; (1 + u/\sqrt{n})\cdot\Phi(u)/(-u)\,.$$

This means that $F_n(u)$ decays on its tail faster than $\Phi(u)$ does until n becomes big compared with $(-u)^2$. Still, $\Phi(u)$ decays rapidly enough as $u \to -\infty$ that a rather large number n of samples are generally needed before the Central Limit Theorem can approximate the tail of the Normal Distribution with a tolerably tiny relative error. It is a topic treated only in advanced texts on Probability and/or Statistics.

Graphs of $\Phi(u)$ vs. $F_n(u)$ for $n = 4,\ 9$ and 100.






Phi(u) vs. Fn(u) for n=100
Fn
Phi
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
U