1 The Strong Law of Large Numbers (SLLN) and the Central Limit Theorem

Contents

One way of interpreting the probability, $P(A)$, of an event A is as the (long-run) proportion of times that the event occurred when sequentially repeating the experiment in question. For example, when we say that "the probability that a fair coin lands Heads (H) equals 1/2", we can interpret that as follows:

If we flip a coin over and over again, then the long-run proportion of times that it landed H equals 1/2.

If we let $X_i = 1$ if the i^{th} flip lands H and $X_i = 0$ if it lands T, then we can express this as

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \frac{1}{2}. \tag{1}$$

A precise and general mathematical statement of this notion is called the *Strong Law of Large Numbers (SLLN)*:

Theorem 1.1 (Strong Law of Large Numbers(SLLN)) *If $\{X_i : i \geq 1\}$ is any iid sequence of random variables with $E|X| < \infty$, then with probability one (wp1) it holds that*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = E(X), \tag{2}$$

by which we mean that

$$P\Big(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = E(X)\Big) = 1.$$

In words: "With probability one, the empirical average of the $\{X_i\}$ equals the expected value."

Note that for any iid sequence $\{X_i\}$ and any real-valued function $g = g(x)$, the sequence $\{g(X_i)\}$ too forms an iid sequence and hence if $E|g(X)| < \infty$, then wp1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(X_i) = E(g(X)).$$

For example, if for a fixed subset $A \subset \mathbb{R}$, one defines $g(X_i) = I\{X_i \in A\}$, the indicator, then we get

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I\{X_i \in A\} = P(X \in A);$$

"The proportion of times that X_i falls in A is equal to $P(X \in A)$."

An immediate application is to rolling a die over and over, with X_i the outcome of the i^{th} roll $X_i \in \{1, 2, 3, 4, 5, 6\}$. Then wp1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I\{X_i = k\} = P(X = k) = \frac{1}{6},\ k \in \{1, 2, 3, 4, 5, 6\}.$$

1.1 Proving the SLLN

We will be satisfied to prove a weaker version of SLLN, known as the *Weak Law of Large Numbers (WLLN)* which gives ample intuition as to why the SLLN must hold. For notation, we let $\mu = E(X)$, $\sigma^2 = Var(X)$ and define

$$\begin{aligned} X(n) &\stackrel{\text{def}}{=} \sum_{i=1}^{n} X_i, \ n \geq 1 & (3) \\ \overline{X}(n) &\stackrel{\text{def}}{=} \frac{1}{n} X(n) = \frac{1}{n} \sum_{i=1}^{n} X_i, \ n \geq 1. & (4) \end{aligned}$$

Proposition 1.1 (WLLN) *If $\{X_i : i \geq 1\}$ is any iid sequence of random variables with finite variance $\sigma^2 = Var(X) < \infty$, then for all $\epsilon > 0$, no matter how small,*

$$P(|\overline{X}(n) - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0, \ \textit{as } n \to \infty.$$

Equivalently,

$$P(|\overline{X}(n) - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2} \to 1, \ \textit{as } n \to \infty.$$

Proof : Recall Markov's inequality $P(|X| > x) \leq \frac{E(X)}{x}$, $x > 0$, specifically the version $P(|X| > x) \leq \frac{E(X^2)}{x^2}$, $x > 0$. Note that by linearity of expectation, $E(\overline{X}(n)) = \frac{1}{n}E(X(n)) = \frac{1}{n}(n\mu) = \mu$, and by independence of the X_i (and recalling $Var(cX) = c^2 Var(X)$ for any constance c), we have

$$\begin{aligned} E|\overline{X}(n) - \mu|^2 &= Var(\overline{X}(n)) \\ &= \frac{1}{n^2} Var(X(n)) \\ &= \frac{1}{n^2} Var(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2} n Var(X) \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Finally, from Markov's inequality,

$$P(|\overline{X}(n) - \mu| > \epsilon) \leq \frac{E|\overline{X}(n) - \mu|^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

■

In words, this result says that for any $\epsilon > 0$, the probability that $\overline{X}(n)$ differs from μ by more than ϵ tends to 0 as the "sample size" n increases.

1.2 Application to Monte Carlo simulation

Suppose we wish to compute an integral

$$\alpha = \int_0^1 g(x)dx,$$

for a function g that has no known antiderivative.

Observe that if $U \sim unif(0,1)$, then the expected value of the random variable $g(U)$ is equal to the desired integral[1]:

$$E(g(U)) = \int_0^1 g(x)dx = \alpha.$$

Thus if we take an iid sequence of the uniforms $\{U_i : i \geq 1\}$ and define $X_i = g(U_i)$, the SLLN asserts that wp1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = E(X) = \alpha. \tag{5}$$

This immediately yields an approximation for α: Choose a "large" value of n, and use

$$\alpha \approx \overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{6}$$

The implementation would be carried out by your computer, which can sequentially "generate/simulate" the U_i. This is known as *Monte Carlo simulation.*

As a nice example for illustrating the method, suppose you wish to estimate the value of π. We know that π is the area of the unit disk in $\mathbb{R}^2$. So it can be obtained as

$$\pi = 4\int_0^1 \sqrt{1-x^2}dx.$$

In other words, we can implement Monte Carlo simulation by using the function $g(x) = 4\sqrt{1-x^2},\ x \in (0,1)$:

$$\pi \approx \frac{4}{n}\sum_{i=1}^{n} \sqrt{1-U_i^2}.$$

Sample code in Python:

```
summ = 0
n = 1000 #sample size
for i in range(n):
    U = numpy.random.uniform(0,1)
    summ += 4*math.sqrt(1-math.pow(U,2))
summ/n
```

Using $n = 1000$ as an example yielded (one run of it) $\pi \approx 3.1622$, increasing to $n = 10,000$ yielded $\pi \approx 3.1415$ which is exact to 4 decimal places.

Note that each time you run a simulation you will obtain slightly different answers; for example if you used $n = 1000$ again (using different iid uniforms), you might get $\pi \approx 3.1422$, and so on.

[1] More generally, if X has a density $f(x)$ on $(0,1)$ then $E(g(X)) = \int_0^1 g(x)f(x)dx$. The density of U is $f(x) = 1$

In our next section, we will address the general issue of *how accurate is the SLLN approximation*

$$E(X) \approx \overline{X}(n).$$

This involves what is called the *Central Limit Theorem* which in turn involves the *normal probability distribution.*

1.2.1 Normal distribution with mean μ and variance σ^2: $N(\mu, \sigma^2)$

We start with a rv Z which has a *normal distribution with mean* 0 *and variance* 1. It is denoted by $N(0,1)$ and has probability density function denoted by $\phi(x)$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \ x \in \mathbb{R}.$$

This function $\phi(x)$ is symmetric about 0, $\phi(-x) = \phi(x)$, and its graph yields the famous "bell curve". Such a random variable Z with this distribution is called a *standard* or *unit* normal, and we denote this by writing "$Z \sim N(0,1)$". This distribution is one of the most fundamental and important ones in all of probability theory and statistics, with numerous applications. The CDF is not explicit (i.e., we can't integrate it out explicitly), we denote it by $\Phi(x)$;

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy, \ x \in \mathbb{R}.$$

For each x, $\Phi(x)$ is the area under the curve ϕ to the left of x. Numerical methods are used to compute $\Phi(x)$ to any desired level of accuracy, and Tables are available in the back of statistics textbooks with useful values of $\Phi(x)$. For example,
$P(Z \leq 1) = \Phi(1) = 0.8413, \ P(Z \leq 2) = \Phi(2) = 0.9772, \ P(Z \leq 3) = \Phi(3) = 0.99987.$

By symmetry $P(Z \leq -x) = P(Z > x)$ for any $x \geq 0$, and $P(Z \leq 0) = P(Z \geq 0) = 0.5$. We thus can obtain from the above values that

$$P(-1 \leq Z \leq 1) \approx 0.68, \ P(-2 \leq Z \leq 2) \approx 0.95, \ P(-3 \leq Z \leq 3) \approx 0.99, \tag{7}$$

or equivalently that

$$P(|Z| \leq 1) \approx 0.68, \ P(|Z| \leq 2) \approx 0.95, \ P(|Z| \leq 3) \approx 0.99, \tag{8}$$

This is very revealing: in particular it tells us that essentially all the mass of Z lies within the interval $[-3, 3]$; i.e., it is rare that a rv Z would take on a value above 3 or below -3: $P(|Z| > 3) < 0.01$. Since the standard deviation of Z is $\sigma = 1$, we can say in words that

99% *of the mass of Z lies within 3 standard deviations of its mean.*

That $\phi(x)$ integrates to 1, hence indeed defines a probability density function, is proved by using polar coordinates: Letting

$$S = \int_{-\infty}^{\infty} \phi(x) dx,$$

observe that

$$\begin{aligned} S^2 &= \Big(\int_{-\infty}^{\infty}\phi(x)dx\Big)\Big(\int_{-\infty}^{\infty}\phi(y)dy\Big) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-(x+y)^2}{2}}dxdy \\ &= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} e^{-r^2/2}rdrd\theta. \\ &= \int_{0}^{\infty} e^{-r^2/2}rdr \\ &= \int_{0}^{\infty} e^{-u}du \\ &= 1, \end{aligned}$$

where the last integral was derived by a simple change of variables $u = r^2/2,\ du = rdr$. Thus $S = 1$.

It is immediate that $E(Z) = 0$, since

$$\int_{-\infty}^{0} x\phi(x)dx = -\int_{0}^{\infty} x\phi(x), dx$$

and ($u = x^2/2$ change of variables),

$$\int_{0}^{\infty} xe^{-x^2/2}dx = \int_{0}^{\infty} e^{-u}du = 1 < \infty.$$

Thus $Var(Z) = E(Z^2)$, and by using integration by parts it follows that $1 = E(Z^2) = \int_{-\infty}^{\infty} x^2\phi(x)dx$.

The more general $N(\mu, \sigma^2)$ dstribution

For any $\mu \in \mathbb{R}$ and any $\sigma > 0$, we can define a new rv

$$X = \sigma Z + \mu, \tag{9}$$

the distribution of which is called the *normal distribution with mean μ and variance σ^2*. We denote this by writing, "$X \sim N(\mu, \sigma^2)$". Because $E(Z) = 0$ and $Var(Z) = 1$, we see from (9) (and recalling the properties of expected value and variance) that indeed $E(X) = \mu$, and $Var(X) = \sigma^2$. All we have done is shift the mean of Z to be μ and modified its variance to be σ^2. Since $F(x) = P(X \leq x) = P(Z \leq (x - \mu)/\sigma) = \Phi((x - \mu)/\sigma)$, we can compute $F(x)$ by using the values for the unit normal CDF $\Phi(x) = P(Z \leq x)$. Moreover, we immediately conclude using a change of variables ($u = \sigma y + \mu,\ du = \sigma dy$) that the CDF is given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy,\ x \in \mathbb{R},$$

and by taking the derivative $f(x) = F'(x)$ the density is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathbb{R}.$$

$f(x)$ is symmetric about its mean μ and again has a bell-shaped curve. Moreover, by using (9), (7) generalizes to

$$\begin{aligned} P(\mu-\sigma \le X \le \mu+\sigma) &\approx 0.68 \qquad (10)\\ P(\mu-2\sigma \le X \le \mu+2\sigma) &\approx 0.95 \qquad (11)\\ P(\mu-3\sigma \le X \le \mu+3\sigma) &\approx 0.99. \qquad (12)\end{aligned}$$

Just as we saw for Z, this says that essentially 99% of the mass of any $X \sim N(\mu, \sigma^2)$ lies within 3 standard deviations of its mean. Note that the larger the value of the variance σ^2 is, the larger are the intervals containing the mass. That is because a larger variance makes the value of X less predictable; its average distance from the mean μ becomes larger.

1.3 Chebyshev's inequality

Using a variation of Markov's inequality, we obtain for any random variable X with finite variance σ^2 that it holds for integers $k \ge 1$ that

$$P(|X-\mu| > k\sigma) \le \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2},$$

known as *Chebyshev's inequality.* Equivalently,

$$P(|X-\mu| \le k\sigma) \ge 1 - \frac{1}{k^2}. \qquad (13)$$

This is an upper bound on how much of the distribution of X lies within k standard deviations from its mean μ. Using $k = 3$ we thus get

$$P(|X-\mu| \le k\sigma) \ge 1 - \frac{1}{3^2} = \frac{8}{9} = 0.89.$$

As we just saw in the previous section, for the special case when $X \sim N(\mu, \sigma^2)$, the true probability for $k = 3$ is even higher, 0.99. But it is nice to know that regardless of the distribution, it must be at least 0.89.

2 The Central Limit Theorem

While the SLLN asserts for an iid sequence $\{X_i\}$, that wp1

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu = E(X),$$

as $n \to \infty$, it does not address the issue of how good an approximation to μ the empirical average $\overline{X}(n)$ is for a given n. Recalling that $E(\overline{X}(n)) = \mu = E(X)$ and $Var(\overline{X}(n)) = \frac{\sigma^2}{n}$, where $\sigma^2 = Var(X)$, the Central Limit Theorem (CLT) fills in this gap. It asserts that

For n large, the probability distribution of $\overline{X}(n)$ is approximately $N(\mu, \frac{\sigma^2}{n})$.

Since we can represent a rv having the $N(\mu, \frac{\sigma^2}{n})$ distribution as $\frac{\sigma}{\sqrt{n}}Z + \mu$, with $Z \sim N(0,1)$, we then have for n large that $\overline{X}(n) \approx \frac{\sigma}{\sqrt{n}}Z + \mu$, hence subtracting μ and dividing by $\frac{\sigma}{\sqrt{n}}$ implies that

$$Z_n \stackrel{\text{def}}{=} \frac{\overline{X}(n) - \mu}{\frac{\sigma}{\sqrt{n}}} \approx Z \sim N(0,1)$$

Note that by multiplying the numerator and denominator of Z_n by n, we equivalently can re-write Z_n as

$$Z_n = \frac{X(n) - n\mu}{\sigma\sqrt{n}},$$

and so we can also express the result as

$$\frac{X(n) - n\mu}{\sigma\sqrt{n}} \approx Z \sim N(0,1), \quad \text{for } n \text{ large.}$$

Here is the rigorous statement:

Theorem 2.1 (Central Limit Theorem) *For any iid sequence of rvs $\{X_i\}$ with finite and non-zero variance, $0 < \sigma^2 < \infty$, it holds that as $n \to \infty$ the distribution of $Z_n = \frac{\overline{X}(n)-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{X(n)-n\mu}{\sigma\sqrt{n}}$ converges to $N(0,1)$:*

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z) = P(Z \leq z), \; z \in \mathbb{R}.$$

A proof of this result (in such generality) is beyond the scope of these notes; but for some special cases we will give some proofs later (Section 4).

The importance of the CTL can not be understated: For **any** probability distribution $F(x) = P(X \leq x)$ (for the X_i) with finite non-zero variance $0 < \sigma^2 < \infty$, Z_n becomes exactly $N(0,1)$ as $n \to \infty$.

We can use this to compute how close $\overline{X}(n)$ is to μ: Recalling Equation (8), we know that $P(|Z| \leq 2) \approx 0.95$.

Thus for n large, $P(|Z_n| \leq 2) \approx 0.95$. Re-writing this we get

$$P(|\overline{X}(n) - \mu| \leq 2\frac{\sigma}{\sqrt{n}}) \approx 0.95$$

For example, choosing $n = 100$, and using (say) $\sigma = 1$, we get

$$P(|\overline{X}(n) - \mu| \leq \frac{1}{5}) \approx 0.95,$$

which we can re-write further as

$$P(\overline{X}(n) - 0.20 \leq \mu \leq \overline{X}(n) + 0.20) \approx 0.95,$$

which means that with probability 0.95, the true (unknown) mean μ lies within the interval

$$[\overline{X}(n) - 0.20, \; \overline{X}(n) + 0.20]. \tag{14}$$

This is an example of what is known as a 95% *confidence interval for estimating* μ, which in general is given by $\overline{X}(n) \pm 2\frac{\sigma}{\sqrt{n}}$. By using a larger sample size n, not only will $\overline{X}(n)$ change to a more accurate estimate for μ, but will yield an even smaller interval with the same confidence level. For example, if $n = 10,000$, with $\sigma = 1$, then $2\frac{\sigma}{\sqrt{n}} = 0.02$, ten times smaller; (14) becomes $\overline{X}(n) \pm 0.02$. We could also use $P(|Z| \leq 3) = 0.99$ instead of $P(|Z| \leq 2) = 0.95$, so as to get a higher probability 0.99, for our estimate interval which would change to $\overline{X}(n) \pm 3\frac{\sigma}{\sqrt{n}}$; a 99% confidence interval. This would enlarge the interval size, however; interval (14) becomes

$$[\overline{X}(n) - 0.30,\ \overline{X}(n) + 0.30], \tag{15}$$

but again by increasing the sample size we can bring it down again. Using $n = 10,000$ yields

$$[\overline{X}(n) - 0.03,\ \overline{X}(n) + 0.03], \tag{16}$$

and it is now a 99% confidence interval.

2.1 Practical considerations

Sample size n

In practice, a large sample size n for constructing a confidence interval might not be available; it might involve for example, rare medical data in which very few people have a certain disease ($n = 10$ (say)), or some other application in which samples are very expensive or difficult to acquire. As a rule of thumb one needs about $n \geq 36$, so as to ensure that the CLT is kicking in. In Monte Carlo simulation, however, the samples are generated by your computer and typically can be very large if needed.

Confidence interval size relative to what is being measured

One also needs to be careful not to get a useless confidence interval. For example if we are trying to estimate the average weight of a certain animal, then we would not want an interval like 5.3 ± 500 pounds, which is useless. One would want the interval length to be a small fraction of what is being measured, such as 5.3 ± 0.1 pounds.

More precise Z values for constructing confidence intervals

When constructing 95% and 99% confidence intervals using (8), we used the values 2 and 3 only for simplicity so as to offer a simple 'rule of thumb' that is easy to remember. In practice we would use the precise values 1.96 and 2.58:

$$P(|Z| \leq 1.96) = 0.95,\ P(|Z| \leq 2.58) = 0.99. \tag{17}$$

This then leads to

95% and 99% confidence intervals

$$\overline{X}(n) \pm 1.96\frac{\sigma}{\sqrt{n}},\ \overline{X}(n) \pm 2.58\frac{\sigma}{\sqrt{n}}. \tag{18}$$

In general a $100(1-\alpha)\%$ confidence interval (where $\alpha \in (0,1)$) is obtained by using the value of z, denoted by z_α, such that $P(|Z| \leq z_\alpha) = 1 - \alpha$ yielding the interval

$$\overline{X}(n) \pm z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Using $\alpha = 0.05$ and 0.01 yields the desired values $z_{0.05} = 1.96$ and $z_{0.01} = 2.58$ for 95% and 99% confidence intervals respectively. In general the various z_α are called *z-scores.*

2.1.1 One-sided (upper/lower) confidence intervals

We have been constructing *two-sided* intervals $\overline{X}(n)\pm$, but one can also construct *one-sided* intervals (regions) such as the *upper* $\mu \le \overline{X}(n) + 1.645\frac{\sigma}{\sqrt{n}}$, where $\Phi(1.646) = P(Z \le 1.646) = 0.95$.

This says that we are 95% confident that the true value of μ lies *below* $\overline{X}(n) + 1.645\frac{\sigma}{\sqrt{n}}$. The *lower* one: we are 95% confident that the true value of μ lies *above* $\overline{X}(n) - 1.645\frac{\sigma}{\sqrt{n}}$. $\Phi(2.33) = P(Z \le 2.33) = 0.99$ for obtaining the analogous one-sided 99% confidence intervals.

Such one-sided intervals are particularly useful in *hypothesis testing* situations (you will learn about this when taking a course in statistics). For example if you wish to test the hypothesis that the public voting support of a particular proposal is *at least* 51%, or you wish to test the hypothesis that a new beverage contains *at most* 40 gms of sugar per bottle.

2.1.2 Sample variance $s^2(n)$

If we are trying to estimate an unknown mean μ, then in general we also would not know the variance σ^2, and hence we would need a good estimate of it so as to explicitly construct a confidence interval such as $\overline{X}(n) \pm 1.96\frac{\sigma}{\sqrt{n}}$. This is provided by what is called the *sample variance*

$$s^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}(n))^2, \tag{19}$$

which uses exactly the same n data points used for the sample mean. It can be proved that for iid data $\{X_i : i \ge 1\}$,

$$\sigma^2 = \lim_{n\to\infty} s^2(n),\ wp1 \quad (s^2(n) \text{ is a } \textit{consistent} \text{ estimator for } \sigma^2) \tag{20}$$

$$\sigma^2 = E(s^2(n)),\ n \ge 2 \quad (s^2(n) \text{ is an } \textit{unbiased} \text{ estimator for } \sigma^2). \tag{21}$$

The definition of $s^2(n)$ has the division by $n-1$ instead of by n because (it turns out) that is (mathematically) required to ensure that $s^2(n)$ is an unbiased estimator.

It turns out that the CLT still holds if we replace σ^2 by $s^2(n)$ (and use $s(n) \overset{\text{def}}{=} \sqrt{s^2(n)}$).

Recalling (18), the confidence intervals for estimating μ are then

95% and 99% confidence intervals: using the sample variance

$$\overline{X}(n) \pm 1.96\frac{s(n)}{\sqrt{n}},\ \overline{X}(n) \pm 2.58\frac{s(n)}{\sqrt{n}}. \tag{22}$$

2.2 Using the CLT to approximate sums of iid random variables

Given iid rvs $\{X_i : 1 \le i \le n\}$ with $0 < \sigma^2 < \infty$, we know from the CLT that for n large, $Z_n \overset{\text{def}}{=} \frac{\overline{X}(n)-\mu}{\frac{\sigma}{\sqrt{n}}} \approx Z \sim N(0,1)$, and thus the sum

$$X(n) \overset{\text{def}}{=} \sum_{i=1}^{n} X_i \approx \sigma\sqrt{n}Z + n\mu \sim N(n\mu, n\sigma^2).$$

Therefore we can explicitly approximate the CDF of the sum $X(n)$ in terms of the $N(0,1)$ CDF $\Phi(z) = P(Z \le z)$:

$$\begin{aligned} P(X(n) \le x) &\approx P(\sigma\sqrt{n}Z + n\mu \le x) \\ &= P(Z \le \frac{x - n\mu}{\sigma\sqrt{n}}) \\ &= \Phi\Big(\frac{x - n\mu}{\sigma\sqrt{n}}\Big), \ x \in \mathbb{R}. \qquad (23) \end{aligned}$$

Similarly, letting $\overline{\Phi}(z) = 1 - \Phi(z) = P(Z > z)$, we have the tail approximation,

$$P(X(n) > x) \approx \overline{\Phi}\Big(\frac{x - n\mu}{\sigma\sqrt{n}}\Big), \ x \in \mathbb{R}. \qquad (24)$$

Using these approximations can be very useful since in general such probabilities might be very difficult to compute directly.

2.2.1 The normal approximation to the binomial distribution

A famous example is to approximate a binomial (n, p) distribution, valid because such a random variable has the representation as a sum of iid rvs;

$$X(n) = \sum_{i=1}^{n} X_i,$$

where the $\{X_i\}$ are iid Bernoulli (p) rvs. $\mu = E(X_i) = p, \ \sigma^2 = Var(X_i) = p(1-p)$. We thus conclude, for example, that (for n sufficiently large),

$$P(X(n) > x) \approx \overline{\Phi}\Big(\frac{x - np}{\sqrt{np(1-p)}}\Big), \ x \in \mathbb{R}. \qquad (25)$$

For example, suppose that $p = 0.4$ and $n = 500$, and we wish to estimate $P(X(500) > 215)$, the probability that out of 500 trials, there are > 215 successes.

Then we can use the approximation

$$P(X(500) > 215) \approx \overline{\Phi}\Big(\frac{215 - (500)(0.4)}{\sqrt{500(0.4)(0.6)}}\Big) = \overline{\Phi}(1.37) = P(Z > 1.37). \qquad (26)$$

Using a Z table or Z calculator we obtain $P(Z > 1.37) = 0.085$.

2.2.2 Continuity correction for the binomial distribution

We can refine the binomial approximation in Equation (25) by using what is called a *continuity correction* based on the fact that since a binomial is *discrete*, we have, for example that $P(X(n) > 215) = P(X(n) > 215.5)$.

This is particularly important when using the approximation for the probability mass function $P(X(n) = k)$, for then otherwise we always would get 0, since Z is continuous; $P(Z = x) = 0$ for any x. So instead we use

$$\begin{aligned} P(X(n) = k) &= P(k - 0.5 < X(n) < k + 0.5) \qquad (27) \\ &\approx P\Big(\frac{k - 0.5 - np}{\sqrt{np(1-p)}} < Z < \frac{k + 0.5 - np}{\sqrt{np(1-p)}}\Big). \qquad (28) \end{aligned}$$

$$P(X(n) > k) = P(X(n) > k+0.5) \approx P\Big(Z > \frac{k+0.5-np}{\sqrt{np(1-p)}}\Big). \tag{29}$$

$$P(X(n) < k) = P(X(n) < k-0.5) \approx P\Big(Z < \frac{k-0.5-np}{\sqrt{np(1-p)}}\Big). \tag{30}$$

For example, we can use it to refine approximation (26),

$$P(X(500) > 215) = P(X(500) > 215.5) \approx P(Z > 1.41) = 0.079,$$

and

$$P(X(500) = 215) = P(214.5 < X(500) < 215.5) \approx P(1.32 < Z < 1.41) = 0.0142.$$

As the reader can check (using computational software for the binomial distribution), the true answers are 0.0789 and 0.0142 respectively, in particular this continuity correction yields a more accurate answer than before for $P(X(500) > 215)$.

One can also use $\leq$, to obtain a common (equivalent) variation of continuity correction:

$$P(X(n) \geq k) = P(X(n) \geq k-0.5) \approx P\Big(Z > \frac{k-0.5-np}{\sqrt{np(1-p)}}\Big). \tag{31}$$

$$P(X(n) \leq k) = P(X(n) \leq k+0.5) \approx P\Big(Z < \frac{k+0.5-np}{\sqrt{np(1-p)}}\Big). \tag{32}$$

This is equivalent to what we already presented because $P(X(n) > k) = P(X(n) \geq k+1)$ and $P(X(n) \leq k) = P(X(n) < k+1)$.

What about the Poisson approximation to the binomial?

Recall that we also have another approximation to the binomial; the Poisson distribution with mean $\alpha = np$, but that is only useful *when n is large and p is small.* The normal approximation only requires that n be large; any p is ok as long as n is large enough. **As a rule of thumb: to use the normal approximation for the binomial distribution make sure that** $np(1-p) \geq 10$. For example if $p = 0.5$, then n should be ≥ 40. If $p = .01$, then n should be ≥ 1000.

When the distribution is a continuous one, as a general rule of thumb it is suggested to use $n \geq 36$.

For example: suppose we wish to approximate $P(X(50) > 55.8)$ when the X_i are iid exponentials at rate $\lambda = 1$: $P(X \leq x) = 1 - e^{-x},\ x \geq 0$. $\mu = E(X) = 1$ and $\sigma^2 = 1$.

$$P(X(50) > 55.8) \approx P(Z > \frac{55.8 - n\mu}{\sigma\sqrt{n}}) = P(Z > \frac{5.8}{\sqrt{50}}) = P(Z > 0.820) = 0.206.$$

3 Simulating from the normal distribution

As we well know, given a $Z \sim N(0,1)$ we can transform it into an $X \sim N(\mu, \sigma^2)$ via setting $X = \sigma Z + \mu$. Thus it suffices to have a simulation algorithm for generating iid copies of $Z \sim N(0,1)$.

The $N(0,1)$ density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}},\ x \in \mathbb{R}, \tag{33}$$

and the CDF is

$$\Phi(x) = P(Z \le x) = \int_{-\infty}^{x} \phi(y)dy \; x \in \mathbb{R}.$$

The inverse transform method can't be used to simulate a Z because we do not have an explicit functional form of the CDF $\Phi(x)$ let alone its inverse $\Phi^{-1}(y)$. One might first try to approximate $\Phi^{-1}(y)$ by an explicit tractable function so as to use the inverse transform method to obtain approximate copies of Z, and that is an approach sometimes used in practice. However, we can actually *exactly* simulate copies of Z using a clever different approach called the *polar method.* What is interesting about this method is that it requires the use of 2 iid $Unif(0,1)$ rvs and in return hands you back 2 iid copies of Z, $X = Z_1, Y = Z_2$.

Polar Method

Suppose that X and X are iid copies of $N(0,1)$. If we graph the vector (X, Y) in the Cartesian $x - y$ plane and then transform it into polar coordinates,

$$R^2 = X^2 + Y^2 \in \mathbb{R}_+ \tag{34}$$

$$\Theta = \arctan(Y/X) \in [0, 2\pi), \tag{35}$$

then from classical multi-dimensional calculus, we can compute the joint density of (R^2, Θ) by using the *Jacobian matrix/determinant* of the invertible polar coordinates transformation h given by

$$(x, y) \longrightarrow h(x, y) = (h_1(x, y), h_2(x, y)) = (x^2 + y^2, \arctan(y/x)).$$

When this is done (proof below), we conclude that the joint density of (R^2, Θ) denoted by $g(u, \theta)$ (i.e., $u = r^2 = x^2 + y^2$), is given by a product of an exponential density at rate $1/2$ (mean 2), and a Uniform$(0, 2\pi)$ density :

$$g(u, \theta) = \frac{1}{2} e^{-u/2} \cdot \frac{1}{2\pi}, \; u > 0, \; \theta \in (0, 2\pi). \tag{36}$$

Summarizing:

1. R^2 has an exponential distribution at rate $1/2$ (mean 2).
2. Θ has a continuous uniform distribution over the interval $[0, 2\pi)$.
3. R^2 and Θ are independent random variables.

Proof of Equation (36)

Proof : The joint density of $f(x, y)$ of (X, Y) is given by the product of two $N(0,1)$ densities (Equation (33)):

$$f(x, y) = \phi(x)\phi(y) = \frac{1}{2\pi} e^{\frac{-(x^2+y^2)}{2}} = \frac{1}{2\pi} e^{u/2}.$$

From 2-dimensional calculus involving inverse transformations, the density of $g(u, \theta)$ is therefore given by

$$g(u, \theta) = \frac{1}{2\pi} e^{u/2} \times |J|^{-1},$$

where $|J|$ denotes the determinant of the Jacobian matrix J of partial derivatives

$$J = \begin{bmatrix} \frac{\partial h_1}{\partial x} & \frac{\partial h_1}{\partial y} \\ \frac{\partial h_2}{\partial x} & \frac{\partial h_2}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x & 2y \\ \frac{-y/x^2}{1+(y/x)^2} & \frac{1/x}{1+(y/x)^2} \end{bmatrix} = \begin{bmatrix} 2x & 2y \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix}. \tag{37}$$

It is immediately verified that $|J| = 2$, and hence $|J|^{-1} = 1/2$ yielding the joint density in (36). ■

Using the above facts in reverse we conclude that if R^2 has an exponential distribution with mean 2, and independently Θ has a continuous uniform distribution over the interval $(0, 2\pi)$, then (converting back into Cartesian coordinates), with radius $R = \sqrt{R^2}$, we have that the following 2 rvs X, Y are iid $N(0, 1)$:

$$\begin{aligned} X &= R\cos\Theta \\ Y &= R\sin\Theta \end{aligned}$$

Letting U_1, U_2 be iid $Unif(0, 1)$, we can generate our exponential via $R^2 = -2\ln(U_1)$ and our uniform via $\Theta = 2\pi U_2$ leading to

Polar Algorithm

1. Generate two iid $Unif(0, 1)$ rvs, U_1, U_2
2. Set $R^2 = -2\ln(U_1)$, $\Theta = 2\pi U_2$ and set $R = \sqrt{R^2}$.
3. Set
$$\begin{aligned} X &= R\cos\Theta \\ Y &= R\sin\Theta. \end{aligned}$$
4. Stop. Output X, Y.

Remark 3.1 Using the Jacobian matrix J with the reciprocal of its determinant $|J|^{-1}$ (via $g(u, \theta) = f(x, y)|J|^{-1}$) is simply a multi-dimensional version of what we would do in one dimension. For example, consider $X \sim exp(\lambda)$. Let us use the transformation $h(x) = x^2,\ x > 0$ and then $Y = h(X) = X^2$. What is the density of X^2? We can first derive it directly by computing its CDF and then differentiating it:
$F_Y(y) = P(Y \le y) = P(X \le \sqrt{y}) = 1 - e^{-\lambda\sqrt{y}}$.
$f_Y(y) = F'(y) = \frac{\lambda}{2\sqrt{y}}e^{-\lambda\sqrt{y}}$.

But this is the same as starting with the density of X, $f_X(x) = \lambda e^{-\lambda x} = \lambda e^{-\lambda\sqrt{y}}$, and multiplying it by $(h'(x))^{-1} = \frac{1}{h'(x)}$: $h'(x) = 2x = 2\sqrt{y}$, and so $(h'(x))^{-1} = \frac{1}{2\sqrt{y}}$; indeed the product yields $f_Y(y) = \frac{\lambda}{2\sqrt{y}}e^{-\lambda\sqrt{y}}$.

4 Some proofs of the CLT in special cases

A classic method of proof for the CLT is to show that the moment generating function (MGF) of Z_n converges to that of $Z \sim N(0, 1)$, as $n \to \infty$, that is, that as $n \to \infty$,

$$M_{Z_n}(s) = E(e^{sZ_n}) \to E(e^{sZ}) = M_Z(s), s \in \mathbb{R}.$$

By MGF theory,[2] such convergence implies that Z_n converges to Z in distribution.

We first will derive $M_Z(s)$, and then use the above approach to prove the CLT in some special cases, such as when the iid X_i are Poisson distributed.

[2] For some probability distributions, a MGF does not exists (it is infinite for all $s \neq 0$), an example being the Weibull distribution $P(X > x) = e^{\lambda\sqrt{x}},\ x \ge 0$; if $Y \sim exp(\lambda)$, then $X = Y^2$ has such a distribution. Thus more generally, we use the *characteristic function*, $\phi(s) = E(e^{isX}),\ s \in \mathbb{R}$, where $i = \sqrt{-1}$; it is complex valued and always exists: $e^{ix} \overset{\text{def}}{=} \cos(x) + i\sin(x)$; $|e^{ix}| = 1$; $|E(e^{isX})| \le E(|e^{isX}|) = 1$.

4.1 MGF of the normal distribution

Proposition 4.1 *For $Z \sim N(0,1)$,*

$$M_Z(s) = e^{\frac{s^2}{2}},\ s \in \mathbb{R}.$$

More generally if $X \sim N(\mu, \sigma^2)$, then

$$M_X(s) = e^{\frac{s^2\sigma^2}{2} + s\mu},\ s \in \mathbb{R}.$$

Proof :

$$\begin{aligned}
M_Z(s) &= E(e^{sZ}) &(38)\\
&= \int_{-\infty}^{\infty} e^{sx}\phi(x)dx &(39)\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{sx} e^{\frac{-x^2}{2}} dx. &(40)\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2+2sx}{2}} dx &(41)\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-s)^2+s^2}{2}} dx &(42)\\
&= e^{\frac{s^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-s)^2}{2}} dx &(43)\\
&= e^{\frac{s^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du \ \ \text{(change of variables } u = x - s,\ du = dx) &(44)\\
&= e^{\frac{s^2}{2}} \int_{-\infty}^{\infty} \phi(u)du &(45)\\
&= e^{\frac{s^2}{2}} \times 1 &(46)\\
&= e^{\frac{s^2}{2}}, &(47)
\end{aligned}$$

where the second to last line follows since ϕ is a density hence integrates to 1.

We can express $X \sim N(\mu, \sigma^2)$ as $\sigma Z + \mu$ and thus

$$\begin{aligned}
M_X(s) &= E(e^{s\sigma Z + s\mu}) &(48)\\
&= e^{s\mu} e^{s\sigma Z} &(49)\\
&= e^{s\mu} M_Z(s\sigma) &(50)\\
&= e^{s\mu} e^{\frac{s^2\sigma^2}{2}} &(51)\\
&= e^{\frac{s^2\sigma^2}{2} + s\mu}. &(52)
\end{aligned}$$

■

From the above, we see that a MGF proof of the CLT is to prove that

$$\lim_{n\to\infty} E(e^{sZ_n}) = e^{\frac{s^2}{2}},\ s \in \mathbb{R}.$$

4.2 MGF of independent sums

If X_1, X_2 are independent rvs, then so are e^{sX_1} and e^{sX_2}; thus

$$\begin{aligned} M_{X_1+X_2}(s) &= E(e^{s(X_1+X_2)}) & (53)\\ &= E(e^{sX_1}e^{sX_2}) & (54)\\ &= E(e^{sX_1})E(e^{sX_2}) & (55)\\ &= M_{X_1}(s)M_{X_2}(s). & (56) \end{aligned}$$

More generally:

Proposition 4.2 *If $\{X_i : 1 \le i \le n\}$ are independent, then*

$$M_{\sum_{i=1}^n X_i(s)} = \prod_{i=1}^n M_{X_i}(s);$$

The MGF of an independent sum equals the product of the n individual MGFs. In particular, if the n rvs are independent and identically distributed (iid) copies of X, then

$$M_{\sum_{i=1}^n X_i}(s) = (M_X(s))^n.$$

The above provides a quick proof that the sum of independent normals is normal:

Proposition 4.3 *If $\{X_i : 1 \le i \le n\}$ are independent $N(\mu_i, \sigma_i^2)$ rvs, then the sum is normal: $\sum_{i=1}^n X_i \sim N(\mu, \sigma^2)$, where $\mu = \sum_{i=1}^n \mu_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$. In particular, If the $\{X_i\}$ are iid, then for each $n \ge 1$, $Z_n \sim N(0,1)$: There is no limit required, as $n \to \infty$, for applying the CLT, Z_n is already exactly normal for each $n \ge 1$.*

Proof : The proof is immediate by applying Proposition 4.2. For the first part: The MGF of the sum is the MGF of a $N(\mu, \sigma^2)$, with $\mu = \sum_{i=1}^n \mu_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$; uniqueness of MGMs implies that the sum is indeed distributed as $N(\mu, \sigma^2)$. Now applying this to the iid case (all are distributed as $N(\mu, \sigma^2)$ for a given fixed μ and σ^2), then $X(n) \sim N(n\mu, n\sigma^2)$. Thus we can write $X(n) = \sigma\sqrt{n}Z + n\mu$ and hence $Z_n = \frac{X(n)-n\mu}{\sigma\sqrt{n}} = Z \sim N(0,1)$. ■

4.3 MGF of the Poisson distribution

We next derive the MGF of the Poisson distribution with mean α;

$$P(X = k) = e^{-\alpha}\frac{\alpha^k}{k!},\ k \ge 0.$$

Proposition 4.4

$$M_X(s) = e^{\alpha(e^s-1)},\ s \ge 0. \tag{57}$$

Proof :

$$\begin{aligned} M_X(s) &= E(e^{sX}) \\ &= e^{-\alpha}\sum_{k=0}^{\infty} e^{sk}\frac{\alpha^k}{k!} \\ &= e^{-\alpha}\sum_{k=0}^{\infty}\frac{(\alpha e^s)^k}{k!} \\ &= e^{-\alpha}e^{\alpha e^s} \\ &= e^{\alpha e^s-\alpha} \\ &= e^{\alpha(e^s-1)}. \end{aligned}$$

■

Now we use Proposition 4.2 on the Poisson distribution to prove that the independent sum of Poisson rvs is Poisson:

Corollary 4.1 *If $\{X_i : 1 \le i \le 1\}$ are independent Poisson (α_i) rvs, then $Y = \sum_{i=1}^n X_i$ is Poisson (α) where $\alpha = \sum_{i=1}^n \alpha_i$. In particular, if the X_i are iid with the same mean α, then Y has a Poisson distribution with mean $n\alpha$.*

Proof : Using Proposition 4.2 with Equation (57) we have

$$\begin{aligned} M_Y(s) &= \prod_{i=1}^n M_{X_i}(s) && (58) \\ &= \prod_{i=1}^n e^{\alpha_i(e^s-1)} && (59) \\ &= e^{\sum_{i=1}^n \alpha_i(e^s-1)} && (60) \\ &= e^{\alpha(e^s-1)}. && (61) \end{aligned}$$

Thus Y has the MGF of a Poisson at rate $\alpha = \sum_{i=1}^n$, and so by uniqueness of MGFs, the result follows. ■

4.4 Proof of the CLT when the X_i are iid Poisson distributed

We now prove the CLT when the X_i are iid with a Poisson distribution. It is proved in a more general framework, but that is explained right after the proof.

Proposition 4.5 (CLT for the Poisson distribution) *Let $X^{(\beta)}$ denote a rv with a Poisson distribution with mean β. Then as $\beta \to \infty$,*

$$\frac{X^{(\beta)}-\beta}{\sqrt{\beta}} \Longrightarrow N(0,1) \text{ in distribution.}$$

Proof : Using Proposition 4.4,

$$\begin{aligned} E(e^{s(\frac{X^{(\beta)}-\beta}{\sqrt{\beta}})}) &= e^{-s\sqrt{\beta}}E(e^{\frac{s}{\sqrt{\beta}}X^{(\beta)}}) && (62) \\ &= e^{-s\sqrt{\beta}}e^{\beta(e^{\frac{s}{\sqrt{\beta}}}-1)}; && (63) \end{aligned}$$

we wish to prove that the above converges to $e^{\frac{s^2}{2}}$ as $\beta \to \infty$. Taking natural logarithms this is equivalent to showing that

$$-s\sqrt{\beta} + \beta(e^{\frac{s}{\sqrt{\beta}}} - 1) \to \frac{s^2}{2}.$$

Replacing β by β^2 equivalently we need to show that

$$-s\beta + \beta^2(e^{\frac{s}{\beta}} - 1) \to \frac{s^2}{2}.$$

Using the Taylor's series expansion, $e^{\frac{s}{\beta}} = 1 + \frac{s}{\beta} + \frac{s^2}{2\beta^2} + \sum_{k=3}^{\infty} \frac{s^k}{k!\beta^k}$, we have

$$\begin{aligned} -s\beta + \beta^2(e^{\frac{s}{\beta}} - 1) &= \frac{s^2}{2} + \beta^2 \sum_{k=3}^{\infty} \frac{s^k}{k!\beta^k} && (64) \\ &= \frac{s^2}{2} + \frac{\beta^2 s^3}{\beta^3} \sum_{k=3}^{\infty} \frac{s^{k-3}}{k!\beta^{k-3}} && (65) \\ &= \frac{s^2}{2} + \frac{s^3}{\beta} \sum_{k=3}^{\infty} \frac{s^{k-3}}{k!\beta^{k-3}}. && (66) \end{aligned}$$

But the error term tends to 0 for each s:

$$\begin{aligned} \frac{s^3}{\beta} \sum_{k=3}^{\infty} \frac{s^{k-3}}{k!\beta^{k-3}} &\le \frac{s^3}{\beta} \sum_{k=0}^{\infty} \frac{(\frac{s}{\beta})^k}{k!} && (67) \\ &= \frac{s^3}{\beta} e^{\frac{s}{\beta}} && (68) \\ &\to 0, \quad \text{as } \beta \to \infty. && (69) \end{aligned}$$

The result follows.

■

Note that if the $\{X_i\}$ are iid Poisson (α), then $X(n) \sim Poisson(n\alpha)$ by Corollary 4.1; $E(X(n)) = n\alpha$ and $Var(X(n)) = n\alpha$. So $Z_n = \frac{X(n) - n\alpha}{\sqrt{n\alpha}}$. Since $n\alpha \to \infty$ as $n \to \infty$, the above theorem (using $\beta = n\alpha$) proves that the CLT holds when the iid rvs are Poisson with any given mean α.