FIT2086 Lecture 4 Notes, Part I
Central Limit Theorem

Dr. Daniel F. Schmidt*

September 7, 2020

1 Central Limit Theorem

The central limit theorem (CLT) is often called the most important theorem, or result, in all of statistics. Why is that? The answer lies in the fact that the CLT tells us how many of the quantities we study as part of statistics and data science are asymptotically distributed; more specifically, it tells us that they are asymptotically distributed as per a normal (Gaussian) distribution (recall that asymptotically means "for large values"). This means that we can often approximate the exact distribution of many quantities we study by a normal distribution, which is much easier to work with.

There are a number of ways the central limit theorem can be framed. Let us start with a very simple statement of a special case of the central limit theorem that will largely suffice for our needs: let $Y_1, \ldots, Y_n$ be random variables (RVs) with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$. Then, for large n, the distribution of the sum $\sum_{i=1}^{n} Y_i$ is approximately normally distributed, with a mean of $n\mu$ and a variance of $n\sigma^2$; the closeness of the approximation improves with increasing n. Note the extreme weakness of our assumptions: we assume only that the random variables *have* a mean of μ and a variance of σ^2 – as long as they satisfy those properties, then regardless of their distribution, their sum for large n will be approximately normally distributed. This is obviously a *powerful* result. More formally:

Fact 1 (Central Limit Thereom)**.** Let $Y_1, \ldots, Y_n$ be random variables (RVs) with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$. Then

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2) \tag{1}$$

as $n \to \infty$, where "a $\xrightarrow{d}$ b" means that the quantity a converges in distribution to the quantity b in the limit.

In words, the CLT says that sums of many RVs, each with a finite mean and variance, are approximately normally distributed, with the approximation getting better and better the greater the number of RVs being added together. This has a number of interesting implications for data science practitioners.

1.1 Natural Phenomena are often Normally Distributed

One particularly interesting implication of the central limit theorem is that it helps to explain why so many natural phenomena appear to be normally distributed. As an example, consider the heights of adults in a homogenous population. These have been empircally shown to be well approximated by normal distributions in a large number of studies. At first this might seem surprising, but the central limit theorem offers an answer as to why this may be the case.

To see how the CLT offers an explanation to this phenemena, consider that a person's height is known to be determined by a sum of many different factors: there are genetic determinants, in which millions of genetic markers across the genome have been shown to be associated with height, and there are environmental factors, such as dietary choices and behavioural patterns. If we treat the contribution of each of these factors to an individuals height as the realisations of many random variables, we see that a individual's height can be thought of being determined by a sum of the effects of many RVs. Appealing to the CLT tells us that we expect this sum to be (approximately) normally distributed, which is largely borne out by real observational data.

1.2 Many Distributions are Normal in the Limit

A second interesting implication of the CLT is that many standard, parametric distributions essentially become equivalent to the normal distribution for large values of one (or more) of their parameters. As a consequence of this, these distributions can frequently be well approximated by the normal distribution, at least for large values of some of their parameters, which can make them substantially easier to work with. To get an idea of this phenomena, we will examine the behaviour of two of the basic distributions that were previously introduced in Lecture 2.

1.2.1 Binomial Distribution

Let us begin by examining one of the more surprising examples of the CLT in action. Recall the binomial distribution:

$$p(M = m \,|\, \theta) = \binom{n}{m} \theta^m (1-\theta)^{(n-m)}.$$

This distribution models the total number of successes, M, that occur in n Bernoulli trials, with each trial having a probability of success of θ. If $Y_1, \ldots, Y_n$ are the Bernoulli trials (binary RVs), then the number of successes can be written as the sum

$$M = \sum_{i=1}^{n} Y_i.$$

As M is the sum of n RVs, and as each RV Y_i has mean $\mathbb{E}[Y_i] = \theta$ and $\mathbb{V}[Y_i] = \theta(1-\theta)$ (by the properties of Bernoulli RVs, see Lecture 2), the CLT (1) tells us that for large values of n the number of successes is approximately normally distributed as

$$M \xrightarrow{d} N(n\theta, n\theta(1-\theta)).$$

Upon reflection, this result is actually quite astonishing: it says that if you tossed a coin n times and recorded the number of successes, then repeated this many times and produced a histogram of the number of successes, you would find that for large values of n that this distribution is approximately normal. That is, adding together zeros and ones in sufficient number produces an (essentially) normally distributed random variable!

To get an intuitive understanding of why this is the case one can think about the behaviour of the random variable M when $\theta = 0.5$ (i.e., a fair coin) as n grows. For large n, seeing M close to zero or n

is highly unlikely, as we expect to see more than a small number of heads or tails in any long sequence of coin tosses. Conversely, seeing a number of heads M near to $n/2$ (i.e., the expected value of M) is much more likely as we expect a sufficiently long sequence to have a roughly equal number of zeros and ones. More generally, the further M is from the expected value $n/2$, in either direction, the less likely we expect such a value of M to be due to the roughly equal distribution of heads and tails in a long sequence of coin tosses. Such a distribution of M – more probably to be near the mean, much less likely to be far away from the mean – is clearly a rough characterisation of the general shape of a normal distribution.

1.3 Poisson Distribution

The second distribution we will examine that "becomes normal" in the limit is the Poisson distribution:

$$p(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

This distribution models the number of events occuring in a fixed time period if the events occur at a rate determined by λ. For simplicity, let us assume the rate λ is an integer (this has no real effect on the result) and consider the RV $S \sim \text{Poi}(\lambda)$; from Lecture 2 we know that if $Y_1, \ldots, Y_\lambda \sim \text{Poi}(1)$ then

$$S = \sum_{i=1}^{\lambda} Y_i \sim \text{Poi}(\lambda).$$

This implies that any $\text{Poi}(\lambda)$ RV is equal to the sum of λ $\text{Poi}(1)$ RVs. From the properties of Poisson distributions we know that of these RVs has $\mathbb{E}[Y_i] = 1$ and $\mathbb{V}[Y_i] = 1$, and therefore by the CLT we have

$$S \xrightarrow{d} N(\lambda, \lambda)$$

as $\lambda \to \infty$. In words, if the rate at which events occur is very large, then the distribution of the number of events occurring in a time period will be approximately normally distributed.

A further interesting consequence of the normality of the Poisson for "large λ" also derives from the divisibility of the Poisson distribution. The divisibility property implies that if the rate of occurrence of events over some time period T is λ_T, then the rate of occurrence of events of a time period $k\,T$ is $k\,\lambda_T$. Therefore, while the counts of events over small time periods may not be normally distributed (if λ is small), they will *always* be approximately normally distributed if the time period over which the counts are taken is sufficiently long. For example, the number of babies born in an hour time period is unlikely to be normally distributed, but the total number of babies born *in a year* is much more likely to be normally distributed as the rate of babies being born in a year is 365×24 times greater than the rate of babies being born in one hour. Interestingly, this offers further support to the discussion of why natural phenomena are frequently normally distributed that was discussed in Section 1.1.

2 The CLT and the Distribution of the Sample Mean

In Lecture 3 we derived a very general result characterising the mean and variance of the sample mean under quite weak assumptions. In particular, if $Y_1, \ldots, Y_n$ are RVs with mean $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$ the sample mean

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{2}$$

was shown to satisfy

$$\mathbb{E}\left[\bar{Y}\right] = \mu, \text{ and } \mathbb{V}\left[\bar{Y}\right] = \frac{\sigma^2}{n}.$$

That is, the average value of the sample mean is equal to the mean of any one observation from our population, and the variance of our sample mean is equal to the variance of any one sample from our population, divided by the sample size n. These results are exact under our assumptions, and are higly useful for two reasons: (i) many estimators are equivalent to the sample mean (for example, maximum likelihood estimate of normal μ parameter, or ML estimation of Poisson rate parameter λ), and (ii) they give us some idea of the behaviour and variability of the sample mean when data is assumed to come from different populations. But what about the *distribution* of $\bar{Y}$ (i.e., the sampling distribution $p(\bar{Y})$)? As discussed in Lecture 3, this is in general much more difficult to derive and depends on the exact distribution that we assume for the population (remember, the population is the infinite large source of data from which we are draw our sample). However, as the sample mean is simply the sum of the RVs $Y_1, \ldots, Y_n$, we see that we can use the CLT to get an asymptotic (in the sample size n) distribution for our sample mean, once again under very weak assumptions.

Let us assume our RVs are independent and satisfy $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$; then from the CLT we know that

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2).$$

Using the fact that $\mathbb{V}[Y_i/n] = \mathbb{V}[Y_i]/n^2$ to conclude the following important result.

Fact 2. Let $Y_1, \ldots, Y_n$ be independently distributed random variables (RVs) with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$. Then

$$\bar{Y} \xrightarrow{d} N(\mu, \sigma^2/n)$$

as $n \to \infty$.

This tells us that under quite weak assumptions (finite mean and variance, independently distributed RVs) that for large sample sizes n we can obtain an approximate sampling distribution of the sample mean $\bar{Y}$, and more importantly that this approximate sampling distribution is normal with mean equal to the mean of a single observation from the population (i.e., $\mathbb{E}[Y_i]$) and variance equal to the variance of a single observation from the population divided by the sample size n (i.e., $\mathbb{V}[Y_i]/n$). Crucially, due to the CLT, this approximation improves in accuracy as n gets larger and larger. In the next two Chapters we will see how this distributional information can be used to answer several important types of data science questions.

2.1 Example 1: CLT and sample mean for normal populations

If we assume our population is distributed as per a $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, i.e., it is normally distributed, then using the CLT we find

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\mu, n\sigma^2)$$

so that the sample mean satisfies

$$\bar{Y} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

as $n \to \infty$. In fact, from Lecture 3, we know that if our population is normally distributed then the sample mean is *exactly* normally distributed for all sample sizes n, so in this case the CLT is not necessary.

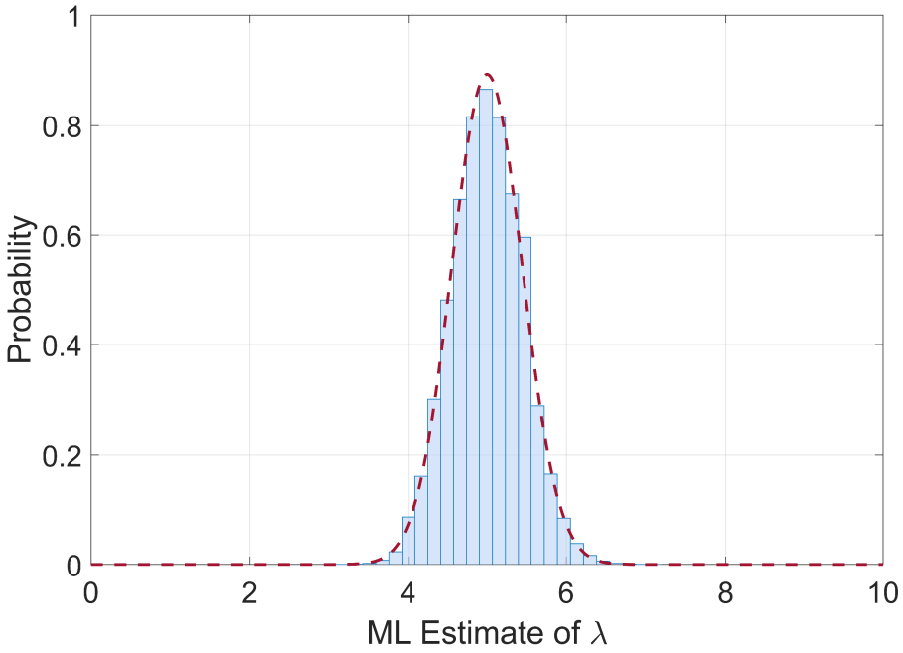

Figure 1: Histogram of $\hat{\lambda}_{\text{ML}}$ from $1,000,000$ data samples, each of size $n = 25$ and generated from a Poi(5) distribution. Also plotted is the normal $N(5, 0.2)$ approximation to the sampling distribution.

2.2 Example 2: CLT and sample mean for Poisson populations

A second example in which the CLT is much more useful is when the population is distributed as a per a Poisson distribution with rate λ, i.e., $Y_1, \ldots, Y_n \sim \text{Poi}(\lambda)$. Recall that the maximum likelihood estimator for λ is:

$$\hat{\lambda}(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

which is equivalent to the sample mean (2). Under our assumed population we know that $\mathbb{E}[Y_i] = \lambda$ and $\mathbb{V}[Y_i] = \lambda$ (see Lecture 2), and from the CLT (1) we know that

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N(n\lambda, n\lambda)$$

as $n \to \infty$; therefore,

$$\hat{\lambda} \xrightarrow{d} N\left(\lambda, \frac{\lambda}{n}\right)$$

as $n \to \infty$. This gives us an approximate distribution of the estimate of the Poisson rate parameter which gets better and better as n gets larger. In fact, for $\lambda > 2$ the approximation is very good even for sample sizes as small as $n = 10$. Figure 1 shows the distribution of $\hat{\lambda}$ as calculated by simulation and plotted against the normal approximation. As an be seen for $n = 25$ it is virtually indistinguishable from the normal approximation.

3 The CLT and other estimators

The CLT can be applied to many other estimators to derive approximate sampling distributions. For example, recall the ML estimator of the variance parameter σ^2 of a normal distribution:

$$\hat{\sigma}^2(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

This estimator is not exactly equivalent to the sample mean (2) so how do we apply the CLT? If we define the RVs

$$E_i = (Y_i - \bar{Y})^2$$

we can write the above estimator as

$$\hat{\sigma}^2(Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^{n} E_i,$$

which is clearly a sample mean of these new RVs $E_1, \ldots, E_n$, and if they have a finite mean and variance (which they do, if we assume the population RVs Y_i have finite mean and variance) we can apply the CLT, and $\hat{\sigma}^2$ will be approximately normally distributed for large n. In fact, this result holds for many estimators that on the surface don't appear to be equivalent to the sample mean, although a detailed study of these estimators is beyond the scope of the subject.