

The Central Limit Theorem (CLT) and the Law of Large Numbers (LLN)

The Law of Large Numbers

Let

$$X_1, X_2, \dots$$

be an infinite sequence of *iid* observations with expected value μ and variance σ^2 . Let

$$A_k = (X_1 + \dots + X_k)/k$$

be the k^{th} partial average. We know that

$$EA_k = \mu$$

and

$$\text{var}A_k = \sigma^2/k.$$

Thus when k gets large, A_k is a random variable with mean μ and very small variance. It follows (with some additional definitions and mathematics) that the sequence of A_k values always converges to μ . This is the “law of large numbers.”

A related fact is that if $B_k = A_k - \mu$, then B_k always converges to zero.

The Central Limit Theorem

Since $B_k = A_k - \mu$ converges to zero, we can multiply B_k term by term with a sequence going to infinity and try to get things to balance out (i.e. not converge to zero or infinity). We will construct a sequence

$$Z_k = C_k \cdot B_k$$

where the C_k are constants (not random). The goal is that the C_k are big enough to counteract the tendency of B_k to converge to zero, but not so big that the sequence blows up.

Since the variance of B_k is σ^2/k , and

$$\text{var}(C_k B_k) = C_k^2 \text{var}(B_k),$$

if we want the variance to stay bounded, then C_k should be \sqrt{k} :

$$Z_k = \sqrt{k}(A_k - \mu).$$

For every value of k , the expected value of Z_k is zero and the variance of Z_k is σ^2 .

What can we say about the distribution of Z_k ? For any particular value of k , the distribution of Z_k can be complex. But as k grows large, the distribution of Z_k becomes approximately normal with mean zero and variance σ^2 . Surprisingly, this is true regardless of the distribution of A_k (as long as some technical conditions are satisfied – in particular, A_k must have a finite variance).

- Program 1

```
## Generate a 20 x 10000 array of Bernoulli trials with success
## probability 0.2.
X <- array(runif(20*10000), c(20, 10000))
X <- (X < 0.2)

## Get the proportion of successes in each column.
Y <- colMeans(X)

## This should be approximately normal with expected value 0.
Z <- sqrt(20) * (Y - 0.2)

## Are the results of this command consistent with the CLT?
summary(Z)

## How can you explain where the result of this command comes from?
var(Z)
```

After running this program, type the commands `hist(Z)` and `qqnorm(Z)` to get a histogram and a normal quantile-quantile (“QQ”) plot. Are these consistent with the CLT?

Now modify the program to use different distributions for the raw data in X , and different sample sizes (i.e. the number of rows of X). Evaluate your results in the context of the CLT.