Lecture 15: One distribution to rule them all. Part 2. Central Limit Theorem–Normal Approximation to Everything

CSCI2244-Randomness and Computation

April 22, 2019

1 Central Limit Theorem

The last lecture illustrated the fact that the sum of independent identically distributed Bernoulli random variables is approximately normally distributed. This is an instance of a *much* more general phenomenon—nearly every random variable has this property.

To be more precise, let X be a random variable for which $\mu = E(X)$ and $\sigma^2 = Var(X)$ are defined. Let $X_1, \ldots, X_n$ be mutually independent random variables, each with the same distribution as X. Think of this as making n independent repetitions of an experiment whose outcome is modeled by the random variable X. Our claim is that the sum of the X_i is approximately normally distributed. Again we adjust the mean and standard deviation to be 0 and 1; then the precise statement is

$$\lim_{n\to\infty} P(a < \frac{X_1 + \cdots X_n - n\mu}{\sigma\sqrt{n}} < b) = \Phi(b) - \Phi(a).$$

This is called the *Central Limit Theorem*. Earlier we saw, with the Law of Large Numbers, that the deviation of the average of n independent identical random variables from its mean approaches 0 as n grows larger. The Central Limit Theorem says more: it tells us how that deviation is distributed.

2 Examples.

Example. The posted iPython notebook contains graphical illustrations of several instances of the Central Limit Theorem, with where X is (a) the outcome of a single die, (b) a very asymmetric discrete random variable, (c) the exponential distribution, which is a very asymmetric continuous random variable. In every case you can see the convergence to the normal distribution.

Example. Roll a die a large number N of times, and let A_N be the average roll. What is the probability that A_N is between 3 and 4? That is, estimate

$$P(3 < A_N < 4).$$

We have $A_N = S_N/N$, where S_N is the sum of the outcomes of N rolls. The roll of a single die has mean $\mu = 3.5$ and variance $35/12 = 2.91667$, so the standard deviation is about 1.708. By the Central Limit Theorem

$$\begin{aligned} P(3 < A_N < 4) &= P(3N < S_N < N) \\ &= P(\frac{-0.5N}{1.708\sqrt{N}} < \frac{S_N - 3.5N}{1.708\sqrt{N}} < \frac{0.5N}{1.708\sqrt{N}}) \\ &\approx \Phi(\frac{0.5N}{1.708\sqrt{N}}) - \Phi(-\frac{0.5N}{1.708\sqrt{N}}) \\ &= 1 - 2\Phi(\frac{0.5N}{1.708\sqrt{N}}) \\ &= 1 - 2\Phi(-0.292\sqrt{N}) \end{aligned}$$

(The next-to-last last line follows from the symmetry of the normal density.)

If we're using the `scipy` library, we can use `norm.cdf` to compute Φ. This gives 0.8557 if $N = 25$. One way to think of the problem is that we are asked the probability that a normal random variable is within $0.292\sqrt{N}$ standard deviations of the mean.

Suppose instead that we are asked the inverse problem: How many rolls do we need to make to ensure that the average roll is between 3 and 4 with probability 0.98? We then have to solve

$$1 - 2\Phi(-0.292\sqrt{N}) = 0.98,$$

or, equivalently

$$\Phi(-0.292\sqrt{N}) = 0.01,$$

so

$$N = (\Phi^{-1}(0.01)/-0.292)^2 \approx 63.5.$$

In `scipy` we can use `norm.ppf` to compute the inverse cdf Φ^{-1}.

These calculations beg the question of how big N has to be for the normal approximation to be accurate. After all, $N = 25$ does not *seem* terribly large; is our estimate of 0.8557 really accurate?

The textbook gives a bound on the error in the normal distribution. In more careful work, we really should check this. Here we will just do a little reality check and compute the probability fo $N = 25$ *exactly*, by using the convolution function (see the posted examples). Here we find

$$P(3N < S_N < 4N) = 0.839$$

and

$$P(3N \leq S_N \leq 4N) = 0.871.$$

The agreement is very good. As with the approximation to the binomial distribution, we get a better result if we go between the integer values of the discrete distribution.

3 Normal distribution in nature.

(For details here, see the Grinstead and Snell book.)

The Central Limit Theorem holds under much weaker hypotheses: It is not necessary for the random variables X_i to be identically distributed, only to obey some modest requirements on the sequence of values of X_i and their variances. We get the same conclusion about the normal approximation.

This is sometimes advanced as an explanation of a striking phenomenon: Many measurements obtained from natural sources appear to follow a normal distribution. A frequently cited example is people's heights. The idea here (for what it's worth—I don't find it enormously convincing) is that within an idealized population of adults, all the variation in height is due to genetic differences, and thus the difference from the population mean can be viewed as a sum of an assortment of mutually independent random variables, each one representing the contribution to overall height of a specific gene. Thus the stronger version of the Central Limit Theorem just cited applies, and the heights should be approximately normally distributed.

4 Is it normal?

We illustrate these ideas with numbers based on real human height data. (See the posted iPython notebook demo, which contains the figures described below.) If you plot a histogram of the heights, the data certainly appear to follow a kind of bell shape–but is it really normal?

One possible visual check is to estimate the mean μ and variance σ^2 from the data. So if the data set is the sequence of values

$$(x_1, \ldots, x_N)$$

we set

$$\mu = \frac{1}{N} \sum_{j=1}^{N} x_j,$$

and

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{N} x_j^2 - \mu^2.$$

We then superimpose the graph of the normal density on the histogram, adjusting the vertical scaling so that the two plots have the same maximum height.

Another visual check is to use something called a *quantile-quantile plot* (qq-plot). If you want to compare experimental data

$$(x_1, \ldots, x_N)$$

to a known distribution with cdf F, we proceed as follows: Let us sort the original data as

$$x'_1 \leq x'_2 \leq \cdots \leq x'_N.$$

For each i, we look at the rank i/N of the value x'_i, and then set

$$y_i = F^{-1}(\frac{i}{N}).$$

The qq-plot is the scatter plot of the points (x'_i, y_i).

Let's see what this gives us if the data follows an approximately normal distribution with mean μ and variance σ^2, and we compare it against the standard normal distribution. In this case the values $\frac{x_i - \mu}{\sigma}$ approximately follow the standard normal distribution, and thus

$$\frac{x'_i - \mu}{\sigma} \approx \Phi^{-1}(\frac{i}{N}) = y_i,$$

so

$$x_i' \approx \sigma y_i + \mu,$$

and thus the points of the scatter plot lie on a straight line. The notebook illustrates how to generate qq-plots in `matplotlib`, as well as the result of the experiment with the height data.