CS 70 Spring 2007 — Discrete Mathematics for CS — Luca Trevisan — Lectures 24-25

Revised May 3

The Normal Distribution and the Central Limit Theorem

Suppose we have a random variable X (it is helpful to think of the example of the random variable X that is $+1$ with probability $1/2$ and -1 with probability $1/2$) of finite expectation and finite variance, that describes the outcome of a certain process. Suppose we repeat the process several times independently, and so have random variables $X_1, \ldots, X_n, \ldots$ that are mutually independent and have each the same distribution as X. What can we say about $A_n = \frac{1}{n}(X_1 + \ldots + X_n)$? We know that $\mathrm{E}(A_n) = \mathrm{E}(X)$, that $\mathrm{Var}(A_n) = \mathrm{Var}(X)/n$, and that, by the law of large number, the probability that A_n deviates from its expectation by any fixed error margin goes to zero when n goes to infinity.

In this lecture we will see that much more can be said about A_n: for large n, if we plot the probabilities of the various possible values of A_n, the plot looks like a "bell-shaped curve" that spikes around the expectation.

This "bell-shaped curve" is the so-called *Normal distribution*,[1] and it is the first (and only) non-discrete distribution we will meet in this course.

Continuous Random Variables

For random variables that take on continuous real values, it no longer makes sense to talk about $\Pr[X = a]$. As an example, consider a r.v. X that has the uniform distribution on the continuous interval $[0,1]$. Then for any single point $0 \le a \le 1$, we have $\Pr[X = a] = 0$. However, clearly it is the case that, for example, $\Pr[\frac{1}{4} \le X \le \frac{3}{4}] = \frac{1}{2}$. So in place of point probabilities $\Pr[X = a]$, we need a different notion of "distribution" for continuous random variables.

Definition 24.1 (density function): For a real-valued r.v. X, a real-valued function $f(x)$ is called a *(probability) density function* for X if

$$\Pr[X \le a] = \int_{-\infty}^{a} f(x)dx.$$

Thus we can think of $f(x)$ as defining a curve, such that the area under the curve between points $x = a$ and $x = b$ is precisely $\Pr[a \le X \le b]$. Note that we must always have $\int_{-\infty}^{\infty} f(x)dx = 1$. (Why?) As an example, for the uniform distribution on $[0,1]$ the density would be

$$f(x) = \begin{cases} 0 & \text{for } x < 0; \\ 1 & \text{for } 0 \le x \le 1; \\ 0 & \text{for } x > 1. \end{cases}$$

[1] As the name suggest, this distribution arises a lot in practice.

[Check you agree with this. What would be the density for the uniform distribution on $[-1,1]$?]

Expectations of continuous r.v.'s are computed in an analogous way to those for discrete r.v.'s. Thus

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

And also

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2, \quad \text{where } \mathrm{E}(X^2) = \textstyle\int_{-\infty}^{\infty} x^2 f(x) dx.$$

To get some familiarity with these notions, let X be the uniform distribution over $[0,1]$, let $f(x)$ be its density function, and let us compute the expectation and the variance of X.

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x dx = \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2}$$

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2 dx = \left.\frac{x^3}{3}\right|_0^1 = \frac{1}{3}$$

and so

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$

As another example, consider the random variable Y that is uniformly distributed in the interval $[a,b]$, where $a < b$ are two real numbers. The density function of Y is going to be zero for $x < a$ and for $x > b$, is going to be a constant in $[a,b]$, and its integral has to be 1. We can see that the density function $f(x)$ of Y must be

$$f(x) = \begin{cases} 0 & \text{for } x < a; \\ \frac{1}{b-a} & \text{for } a \le x \le b; \\ 0 & \text{for } x > b. \end{cases}$$

The expectation of Y is then

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} x f(x) dx = \int_a^b \frac{1}{b-a} x dx = \frac{1}{b-a} \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2}\frac{1}{b-a}(b^2 - a^2) = \frac{b+a}{2}$$

Computing the variance of Y would now be a tedious exercise in calculus. We can, however, do the calculation in a more direct way.

First note that, if X is a random variable, and c, d are constants, then we know that $\mathrm{E}(cX + d) = c\mathrm{E}(X) + d$. We have also proved that $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$, and it is easy to see from the definition that $\mathrm{Var}(X + d) = \mathrm{Var}(X)$, so that, combining the two rules, we have

$$\mathrm{Var}(cX + d) = c^2\mathrm{Var}(X)$$

Back to the random variable Y that is uniformly distributed in $[a,b]$, we can think of it as $a+(b-a)X$ where X is a random variable uniformly distributed in $[0,1]$. Using this observation, we get

$$\mathrm{E}(Y) = a + (b-a)\mathrm{E}(X) = a + \frac{b-a}{2} = \frac{b+a}{2}$$

which is a different way to get the result that we computed by solving an integral, and

$$\mathrm{Var}(Y) = (b-a)^2\mathrm{E}(X) = \frac{(b-a)^2}{12}$$

The Normal Distribution

Now we are in a position to define the Normal distribution. We start from a special case.

Definition 24.2 (Standard Normal distribution): The *standard Normal distribution* is the distribution with density function

$$f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}.$$

The reason for the factor $\sqrt{2\pi}$ is that this is the value of the integral of the exponential function $\mathrm{e}^{-(x^2/2)}$. So we have to normalize by this value to get $\int_{-\infty}^{\infty} f(x)dx = 1$.

Since we have $f(x) = f(-x)$, we see that if Y has a standard normal distribution, then

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} x f(x)dx = 0$$

Computing the variance is a bit harder, and we just give the final result

$$\mathrm{Var}(Y) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}dx = 1$$

Now we can finally go back to our original question of approximating the probability distribution of a sum of independent random variables.

Say we have ± 1 random variables $X_1,\ldots,X_n,\ldots$ independent and identically distributed such that $\Pr[X_i = -1] = \Pr[X_i = 1] = 1/2$. Then define the random variable

$$S_n = \frac{1}{\sqrt{n}} \cdot (X_1 + \ldots + X_n)$$

Since $(X_1 + \ldots + X_n)$ has variance n and standard deviation $\sqrt{n}$, the random variable S_n measures how much larger or smaller our sum is compared to its standard deviation.

We can now state the punchline of this lecture: for large n, the random variable S_n "tends to" a standard Normal random variable.

Theorem 24.1: **[Central Limit Theorem – Baby Version]** *Let S_n be defined above, Y be a standard Normal random variable, and $a < b$ be any two constants, then*

$$\lim_{n\to\infty} \Pr[a \le S_n \le b] = \Pr[a \le Y \le b] = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

If, for large n, S_n is well approximated by a standard Normal random variable Y, then the sum $(X_1 + \ldots + X_n)$ is well approximated by the random variable $\sqrt{n} \cdot Y$. Similarly (check it) the number of heads in n coin tosses is well approximated by the random variable $\frac{n}{2} + \frac{\sqrt{n}}{2} \cdot Y$, where Y is standard Normal. In general, random variables of the form $cY + d$ where Y is standard Normal are very useful. In order to effectively use such random variables as approximations of discrete random variables, we need to figure out what their density functions are.

Here is a helpful notion. If Y is a random variable of density function f_Y, the *cumulative distribution function* pf Y is the function

$$F_Y(a) = \Pr[Y \le a]$$

Just like the density function does, the cumulative distribution function tells us all that there is to know about a random variable. The two functions are related by the fact that

$$F_Y(a) = \int_{-\infty}^{a} f_y(x)dx$$

and so we have that the density function is the derivative of the cumulative distribution function:

$$f_y(x) = \frac{dF_Y(x)}{dx}$$

Now suppose that Y is standard Normal, and that we have another random variable Z defined as $Z = cY + d$. Then there is a simple relation between the cumulative distribution function of Z and the one of Y:

$$F_Z(x) = \Pr[cY + d \le x] = \Pr[Y \le (x-d)/c] = F_Y((x-d)/c)$$

and this translates into a relation between their density functions:

$$f_Z(x) = \frac{dF_Z(x)}{dx} = \frac{dF_Y((x-d)/c)}{dx} = \frac{1}{c} \cdot \frac{dF_Y((x-d)/c)}{d((x-d)/c} = \frac{1}{c} f_Y((x-d)/c)$$

At this point, we have a pretty good understanding of a random variable of the form $cY + d$ where Y is normal. It's density function is $\frac{1}{c\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-d)^2}{c^2}}$ and, by previous calculations, we see that $\mathrm{E}(cY + d) = d$ and $\mathrm{Var}(cX + d) = c^2$. Distributions of this type are called *normal distributions*.

Definition 24.3 (Normal distribution): The *Normal distribution with mean μ and variance σ^2* is the distribution with density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}.$$

If you plot the above density function $f(x)$, you will see that it is a symmetrical bell-shaped curve centered around the mean μ. Its height at the mean is about 0.4, and its width is determined by the variance σ^2, as follows: about 68% of the mass is contained in the interval of width σ either side of the mean, 95% in the interval of width 3σ, and 99.7% in the interval of width 3σ either side of the mean. (Note that, to get the correct scale, deviations are on the order of σ, the standard deviation, rather than σ^2, the variance.)

Now we are in a position to state the Central Limit Theorem. Because our treatment of continuous distributions has been rather sketchy, we shall be content with a rather imprecise statement. This can be made completely rigorous without too much extra effort.

Theorem 24.2: **[Central Limit Theorem]** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common expectation $\mu = \mathrm{E}(X_i)$ and variance $\sigma^2 = \mathrm{Var}(X_i)$ (both assumed to be $< \infty$). Define $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then as $n \to \infty$, the distribution of A_n approaches the Normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.*

Note that the variance is $\frac{\sigma^2}{n}$ (as we would expect) so the width of the bell-shaped curve decreases by a factor of $\sqrt{n}$ as n increases.

The Central Limit Theorem is actually a very striking fact. What it says is the following. If we take an average of n observations of absolutely any r.v. X, then the distribution of that average will be a bell-shaped curve centered at $\mu = \mathrm{E}(X)$. Thus all trace of the distribution of X disappears as n gets large: all distributions, no matter how complex,[2] look like the Normal distribution when they are averaged. The only effect of the original distribution is through the variance σ^2, which determines the width of the curve for a given value of n, and hence the rate at which the curve shrinks to a spike.

The Central Limit Theorem and Confidence Intervals

Let us return to the question of how many people we need to poll in order to have, with high confidence, a good approximation of the opinions of the general population.

Suppose a p fraction of people are Democrat, we pick a random sample n of people, we ask them their party affiliation, and we let A_n be the random variable denoting what fraction of polled people are Democrat. (That is, the number of Democrat respondents divided by n.) As we computed, $\mathrm{E}(A_n) = p$, making it an *unbiased estimator*, and its variance depends on p, but is always at most $1/4\sqrt{n}$. Suppose we want to make sure that, with probability at least 95%, our result A_n is within $\pm 4\%$ of the actual value p; how many people do we need to poll? That is, how large should n be so that

$$\Pr[|A_n - p| > .04] \leq .05 \ ?$$

Using Chebyshev's inequality, we found that $n = 3,125$ was sufficient. It turns out that this was quite an overestimation, and that we can get a better approximation via the Central Limit Theorem. Let A'_n be the Normal distribution with expectation p and variance $(p - p^2)/n$. We know that there is a probability at most 5% that a Normal distribution deviates from its expectation by more than 2 standard deviations, so

$$\Pr\left[|A'_n - p| > 2\sqrt{(p - p^2)/n}\right] \leq .05$$

So if we find a n such that the two standard deviations of A'_n are less than 4%, then we have solved our problem. As usual we don't know $(p - p^2)$, but we know it's at most $1/4$, so

[2] We do need to assume that the mean and variance of X are finite.

$$2\sqrt{(p-p^2)/n} \leq \frac{1}{\sqrt{n}} \leq .04 \text{ provided that } n \geq 625$$

And so we have that, for $n \geq 625$,

$$\Pr[|A'_n - p| > .04] \leq \Pr[|A'n - p| > 2\sqrt{(p-p^2)n}] \leq .05$$

The difference between this calculation and the one we did using Chebyshev's inequality is that here, if choose $n = 625$, we have to rely on the assumption that A_n is close to A'_n for $n = 625$, while the Central Limit Theorem is only telling us that the two distribution are close in the limit. Chebyshev's inequality, instead, bounds A_n directly and gives a valid result for every n.