A Brief Introduction to Continuous Probability

Up to now we have focused exclusively on *discrete* probability spaces Ω, where the number of sample points $\omega \in \Omega$ is either finite or countably infinite (such as the integers). As a consequence we have only been able to talk about *discrete* random variables, which take on only a finite (or countably infinite) number of values.

But in real life many quantities that we wish to model probabilistically are *real-valued*; examples include the position of a particle in a box, the time at which an certain incident happens, or the direction of travel of a meteorite. In this lecture, we discuss how to extend the concepts we've seen in the discrete setting to this continuous setting. As we shall see, everything translates in a natural way once we have set up the right framework. The framework involves some elementary calculus but (at this level of discussion) nothing too scary.

Continuous uniform probability spaces

Suppose we spin a "wheel of fortune" and record the position of the pointer on the outer circumference of the wheel. Assuming that the circumference is of length ℓ and that the wheel is unbiased, the position is presumably equally likely to take on any value in the real interval $[0,\ell]$. How do we model this experiment using a probability space?

Consider for a moment the (almost) analogous discrete setting, where the pointer can stop only at a finite number m of positions distributed evenly around the wheel. (If m is very large, then presumably this is in some sense similar to the continuous setting.) Then we would model this situation using the discrete sample space $\Omega = \{0, \frac{\ell}{m}, \frac{2\ell}{m}, \ldots, \frac{(m-1)\ell}{m}\}$, with uniform probabilities $\Pr[\omega] = \frac{1}{m}$ for each $\omega \in \Omega$. In the continuous world, however, we get into trouble if we try the same approach. If we let ω range over all real numbers in $[0,\ell]$, what value should we assign to each $\Pr[\omega]$? By uniformity this probability should be the same for all ω, but then if we assign to it any positive value, the sum of all probabilities $\Pr[\omega]$ for $\omega \in \Omega$ will be ∞! Thus $\Pr[\omega]$ must be zero for all $\omega \in \Omega$. But if all of our sample points have probability zero, then we are unable to assign meaningful probabilities to any events!

To rescue this situation, consider instead any non-empty *interval* $[a,b] \subseteq [0,\ell]$. Can we assign a non-zero probability value to this interval? Since the total probability assigned to $[0,\ell]$ must be 1, and since we want our probability to be uniform, the logical value for the probability of interval $[a,b]$ is

$$\frac{\text{length of } [a,b]}{\text{length of } [0,\ell]} = \frac{b-a}{\ell}.$$

In other words, the probability of an interval is proportional to its length.

Note that intervals are subsets of the sample space Ω and are therefore *events*. So in continuous probability, we are assigning probabilities to certain basic events, in contrast to discrete probability, where we assigned probability to *points* in the sample space. But what about probabilities of other events? Actually, by specifying the probability of intervals we have also specified the probability of any event E which can be written as the disjoint union of (a finite or countably infinite number of) intervals, $E = \cup_i E_i$. For then we can write $\Pr[E] = \sum_i \Pr[E_i]$, in analogous fashion to the discrete case. Thus for example the probability that the pointer ends up in the first or third quadrants of the wheel is $\frac{\ell/4}{\ell} + \frac{\ell/4}{\ell} = \frac{1}{2}$. For all practical purposes, such events are all we really need.[1]

Continuous random variables

Recall that in the discrete setting we typically work with *random variables* and their distributions, rather than directly with probability spaces and events. The simplest example of a continuous random variable is the position X of the pointer in the wheel of fortune, as discussed above. This random variable has the *uniform* distribution on $[0,\ell]$. How, precisely, should we define the distribution of a continuous random variable? In the discrete case the distribution of a r.v. X is described by specifying, for each possible value a, the probability $\Pr[X = a]$. But for the r.v. X corresponding to the position of the pointer, we have $\Pr[X = a] = 0$ for every a, so we run into the same problem as we encountered above in defining the probability space.

The resolution is essentially the same: instead of specifying $\Pr[X = a]$, we instead specify $\Pr[a \le X \le b]$ for all intervals $[a,b]$[2]. To do this formally, we need to introduce the concept of a *probability density function* (sometimes referred to just as a "density", or a "pdf").

Definition 18.1 (Density): A *probability density function* for a random variable X is a function $f : \mathbf{R} \to \mathbf{R}$ satisfying

$$\Pr[a \le X \le b] = \int_a^b f(x)dx \qquad \text{for all } a \le b.$$

Let's examine this definition. Note that the definite integral is just the area under the curve f between the values a and b. Thus f plays a similar role to the "histogram" we sometimes draw to picture the distribution of a discrete random variable.

In order for the definition to make sense, f must obey certain properties. Some of these are technical in nature, which basically just ensure that the integral is always well defined; we shall not dwell on this issue here since all the densities that we will meet will be well behaved. What about some more basic properties of f? First, it must be the case that f is a non-negative function; for if f took on negative values we could find an interval in which the integral is negative, so we would have a negative probability for some event! Second, since the r.v. X must take on some value everywhere in the space, we must have

$$\int_{-\infty}^{\infty} f(x)dx = \Pr[-\infty < X < \infty] = 1. \tag{1}$$

I.e., the total area under the curve f must be 1.

A caveat is in order here. Following the "histogram" analogy above, it is tempting to think of $f(x)$ as a "probability." However, $f(x)$ doesn't itself correspond to the probability of anything! For one thing, there is no requirement that $f(x)$ be bounded by 1 (and indeed, we shall see examples of densities in which $f(x)$ is greater than 1 for some x). To connect $f(x)$ with probabilities, we need to look at a very small interval $[x, x+\delta]$ close to x; then we have

$$\Pr[x \le X \le x+\delta] = \int_x^{x+\delta} f(z)dz \approx \delta f(x). \tag{2}$$

[1] A formal treatment of which events can be assigned a well-defined probability requires a discussion of *measure theory*, which is beyond the scope of this course.

[2] Note that it does not matter whether or not we include the endpoints a,b; since $\Pr[X = a] = \Pr[X = b] = 0$, we have $\Pr[a < X < b] = \Pr[a \le X \le b]$.

Thus we can interpret $f(x)$ as the "probability per unit length" in the vicinity of x.

Now let's go back and put our wheel-of-fortune r.v. X into this framework. What should be the density of X? Well, we want X to have non-zero probability only on the interval $[0,\ell]$, so we should certainly have $f(x) = 0$ for $x < 0$ and for $x > \ell$. Within the interval $[0,\ell]$ we want the distribution of X to be uniform, which means we should take $f(x) = c$ for $0 \le x \le \ell$. What should be the value of c? This is determined by the requirement (1) that the total area under f is 1. The area under the above curve is $\int_{-\infty}^{\infty} f(x)dx = \int_0^\ell c dx = c\ell$, so we must take $c = \frac{1}{\ell}$. Summarizing, then, the density of the uniform distribution on $[0,\ell]$ is given by

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

Expectation and variance of a continuous random variable

By analogy with the discrete case, we define the expectation of a continuous r.v. as follows:

Definition 18.2 (Expectation): The expectation of a continuous random variable X with probability density function f is

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

Note that the integral plays the role of the summation in the discrete formula $\mathrm{E}(X) = \sum_a a \Pr[X = a]$.

Example: Let X be a uniform r.v. on the interval $[0,\ell]$. Then

$$\mathrm{E}(X) = \int_0^\ell x \frac{1}{\ell} dx = \left[\frac{x^2}{2\ell}\right]_0^\ell = \frac{\ell}{2}.$$

This is certainly what we would expect!

We will see more examples of expectations of continuous r.v.'s in the next section.

Since variance is really just another expectation, we can immediately port its definition to the continuous setting as well:

Definition 18.3 (Variance): The variance of a continuous random variable X with probability density function f is

$$\mathrm{Var}(X) = \mathrm{E}((X - \mathrm{E}(X))^2) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left(\int_{-\infty}^{\infty} x f(x) dx\right)^2.$$

Example: Let's calculate the variance of the uniform r.v. X on the interval $[0,\ell]$. From the above definition, and plugging in our previous value for $\mathrm{E}(X)$, we get

$$\mathrm{Var}(X) = \int_0^\ell x^2 \frac{1}{\ell} dx - \mathrm{E}(X)^2 = \left[\frac{x^3}{3\ell}\right]_0^\ell - \left(\frac{\ell}{2}\right)^2 = \frac{\ell^2}{3} - \frac{\ell^2}{4} = \frac{\ell^2}{12}.$$

The factor of $\frac{1}{12}$ here is not particularly intuitive, but the fact that the variance is proportional to ℓ^2 should come as no surprise. Like its discrete counterpart, this distribution has large variance.

An application: Buffon's needle

Here is a simple application of continuous random variables to the analysis of a classical procedure for estimating the value of π known as *Buffon's needle*, after its 18th century inventor Georges-Louis Leclerc, Comte de Buffon.

Here we are given a needle of length ℓ, and a board ruled with horizontal lines at distance ℓ apart. The experiment consists of throwing the needle randomly onto the board and observing whether or not it crosses one of the lines. We shall see below that (assuming a perfectly random throw) the probability of this event is exactly $2/\pi$. This means that, if we perform the experiment many times and record the *proportion* of throws on which the needle crosses a line, then the Law of Large Numbers (Lecture 16) tells us that we will get a good estimate of the quantity $2/\pi$, and therefore also of π.

To analyze the experiment, we first need to specify the probability space. Let's go straight to random variables. Note that the position where the needle lands is completely specified by two random variables: the vertical distance Y between the midpoint of the needle and the closest horizontal line, and the angle Θ between the needle and the vertical. The r.v. Y ranges between 0 and $\ell/2$, while Θ ranges between $-\pi/2$ and $\pi/2$. Since we assume a perfectly random throw, we may assume that their *joint distribution* has density $f(y,\theta)$ that is uniform over the rectangle $[0,\ell/2]\times[-\pi/2,\pi/2]$. Since this rectangle has area $\frac{\pi\ell}{2}$, the density should be

$$f(y,\theta) = \begin{cases} 2/\pi\ell & \text{for } (y,\theta)\in[0,\ell/2]\times[-\pi/2,\pi/2]; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

As a sanity check, let's verify that the integral of this density over all possible values is indeed 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(y,\theta)dyd\theta = \int_{-\pi/2}^{\pi/2}\int_{0}^{\ell/2} \frac{2}{\pi\ell}dyd\theta = \int_{-\pi/2}^{\pi/2}\left[\frac{2y}{\pi\ell}\right]_0^{\ell/2} d\theta = \int_{-\pi/2}^{\pi/2}\frac{1}{\pi}d\theta = \left[\frac{1\theta}{\pi}\right]_{\pi/2}^{\pi/2} = 1.$$

This is an analog of equation (1) for our joint distribution; rather than the area under the simple curve $f(x)$, we are now computing the area under the "surface" $f(y,\theta)$. But of course the result should again be the total probability mass, which is 1.

Now let E denote the event that the needle crosses a line. How can we express this event in terms of the values of Y and Θ? Well, by elementary geometry the vertical distance of the endpoint of the needle from its midpoint is $\frac{\ell}{2}\cos\Theta$, so the needle will cross the line if and only if $Y\le\frac{\ell}{2}\cos\Theta$. Therefore we have

$$\Pr[E] = \Pr[Y\le\tfrac{\ell}{2}\cos\Theta] = \int_{-\pi/2}^{\pi/2}\int_{0}^{(\ell/2)\cos\theta} f(y,\theta)dyd\theta.$$

Substituting the density $f(y,\theta)$ from equation (3) and performing the integration we get

$$\begin{aligned}
\Pr[E] &= \int_{-\pi/2}^{\pi/2} \int_0^{(\ell/2)\cos\theta} \frac{2}{\pi\ell} dy d\theta \\
&= \int_{-\pi/2}^{\pi/2} \left[\frac{2y}{\pi\ell}\right]_0^{(\ell/2)\cos\theta} d\theta \\
&= \int_{-\pi/2}^{\pi/2} \frac{2}{\pi\ell}\frac{\ell}{2} \cos\theta d\theta \\
&= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \cos\theta d\theta \\
&= \frac{1}{\pi}[\sin\theta]_{-\pi/2}^{\pi/2} \\
&= \frac{2}{\pi}.
\end{aligned}$$

This is exactly what we claimed at the beginning of the section!

Joint distribution and independence for continuous random variables

In analyzing the Buffon's needle problem, we used the notion of *joint density* of two continuous random variables without formally defining the concept. But its definition should be obvious at this point:

Definition 18.4 (Joint Density): A *joint density function* for two random variable X and Y is a function $f: \mathbf{R}^2 \to \mathbf{R}$ satisfying

$$\Pr[a \le X \le b, c \le Y \le d] = \int_c^d \int_a^b f(x,y) dx dy \qquad \text{for all } a \le b \text{ and } c \le d.$$

So in the previous example, a uniform joint density on the rectangle $[0,\ell/2]\times[-\pi/2,\pi/2]$ simply means that the probability of landing in any small rectangle in that rectangle is proportional to the area of the small rectangle.

In analogy with (2), we can connect $f(x,y)$ with probabilities by looking at a very small square $[x,x+\delta]\times[y,y+\delta]$ close to (x,y); then we have

$$\Pr[x \le X \le x+\delta, y \le Y \le y+\delta] = \int_y^{y+\delta}\int_x^{x+\delta} f(u,v) du dv \approx \delta^2 f(x,y). \tag{4}$$

Thus we can interpret $f(x,y)$ as the "probability per unit area" in the vicinity of (x,y).

Recall that in discrete probability, two r.v.'s X and Y are said to be *independent* if the events $X = a$ and $Y = c$ are independent for every a, c. What about for continuous r.v.'s?

Definition 18.5 (Independence for Continuous R.V.'s): Two continuous r.v.'s X,Y are *independent* if the events $a \le X \le b$ and $c \le Y \le d$ are independent for all a,b,c,d.

What does this definition say about the joint density of independent r.v.'s X and Y? Applying (4) to connect the joint density with probabilities, we get, for small δ:

$$\begin{aligned}
\delta^2 f(x,y) &\approx \Pr[x \le X \le x+\delta, y \le Y \le y+\delta] \\
&= \Pr[x \le X \le x+\delta]\Pr[y \le Y \le y+\delta] \\
&\approx \delta f_1(x) \times \delta f_2(y) = \delta^2 f_1(x) f_2(y),
\end{aligned}$$

where f_1 and f_2 are the (marginal) densities of X and Y respectively. So for independent r.v.'s, the joint density is the product of the marginal densities (cf. the discrete case, where joint distributions are the product of the marginals).

In the Buffon's needle problem, it is easy to check that Y and Θ are independent r.v.'s , each of which is uniformly distributed in its respective range.

Two more important continuous distributions

We have already seen one important continuous distribution, namely the uniform distribution. In this section we will see two more: the *exponential* distribution and the *normal* (or *Gaussian*) distribution. These three distributions cover the vast majority of continuous random variables arising in applications.

Exponential distribution: The exponential distribution is a continuous version of the geometric distribution, which we have already met. Recall that the geometric distribution describes the number of tosses of a coin until the first Head appears; the distribution has a single parameter p, which is the bias (Heads probability) of the coin. Of course, in real life applications we are usually not waiting for a coin to come up Heads but rather waiting for a system to fail, a clock to ring, an experiment to succeed etc.

In such applications we are frequently not dealing with discrete events or discrete time, but rather with *continuous* time: for example, if we are waiting for an apple to fall off a tree, it can do so at any time at all, not necessarily on the tick of a discrete clock. This situation is naturally modeled by the exponential distribution, defined as follows:

Definition 18.6 (Exponential distribution): For any $\lambda > 0$, a continuous random variable X with pdf f given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

is called an *exponential* random variable with parameter λ.

Like the geometric, the exponential distribution has a single parameter λ, which characterizes the *rate* at which events happen. We shall illuminate the connection between the geometric and exponential distributions in a moment.

First, let's do some basic computations with the exponential distribution. We should check first that it is a valid distribution, i.e., that it satisfies (1):

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^{\infty} = 1,$$

as required. Next, what is its expectation? We have

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} \lambda x e^{-\lambda x} dx = \left[-xe^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx = 0 + \left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^{\infty} = \frac{1}{\lambda},$$

where for the first integral we used integration by parts.

To compute the variance, we need to evaluate

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} \lambda x^2 e^{-\lambda x} dx = \left[-x^2 e^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x} dx = 0 + \frac{2}{\lambda}\mathrm{E}(X) = \frac{2}{\lambda^2},$$

where again we used integration by parts. The variance is therefore

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Let us now explore the connection with the geometric distribution. Note first that the exponential distribution satisfies, for any $t \geq 0$,

$$\Pr[X > t] = \int_t^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_t^\infty = e^{-\lambda t}. \tag{5}$$

In other words, the probability that we have to wait more than time t for our event to happen is $e^{-\lambda t}$, which is a geometric decay with rate λ.

Now consider a discrete-time setting in which we perform one trial every δ seconds (where δ is very small—in fact, we will take $\delta \to 0$ to make time "continuous"), and where our success probability is $p = \lambda\delta$. Making the success probability proportional to δ makes sense, as it corresponds to the natural assumption that there is a fixed *rate* of success *per unit time*, which we denote by $\lambda = p/\delta$. The number of trials until we get a success has the geometric distribution with parameter p, so if we let the r.v. Y denote the time (in seconds) until we get a success we have

$$\Pr[Y > k\delta] = (1-p)^k = (1-\lambda\delta)^k \qquad \text{for any } k \geq 0.$$

Hence, for any $t > 0$, we have

$$\Pr[Y > t] = \Pr[Y > (\tfrac{t}{\delta})\delta] = (1-\lambda\delta)^{t/\delta} \approx e^{-\lambda t},$$

where this final approximation holds in the limit as $\delta \to 0$ with $\lambda = p/\delta$ fixed. (We are ignoring the detail of rounding $\frac{t}{\delta}$ to an integer since we are taking an approximation anyway.)

Comparing this expression with (5) we see that this distribution has the same form as the exponential distribution with parameter λ, where λ (the success rate per unit time) plays an analogous role to p (the probability of success on each trial)—though note that λ is not constrained to be ≤ 1. Thus we may view the exponential distribution as a continuous time analog of the geometric distribution.

Normal Distribution: The last continuous distribution we will look at, and by far the most prevalent in applications, is called the *normal* or *Gaussian* distribution. It has two parameters, μ and σ.

Definition 18.7 (Normal distribution): For any μ and $\sigma > 0$, a continuous random variable X with pdf f given by

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

is called a *normal* random variable with parameters μ and σ. In the special case $\mu = 0$ and $\sigma = 1$, X is said to have the *standard normal* distribution.

A plot of the pdf f reveals a classical "bell-shaped" curve, centered at (and symmetric around) $x = \mu$, and with "width" determined by σ. (The precise meaning of this latter statement will become clear when we discuss the variance below.)

Let's run through the usual calculations for this distribution. First, let's check equation (1):

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} = 1. \tag{6}$$

The fact that this integral evaluates to 1 is a routine exercise in integral calculus, and is left as an exercise (or feel free to look it up in any standard book on probability or on the internet).

What are the expectation and variance of a normal r.v. X? Let's consider first the standard normal. By definition, its expectation is

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{0} xe^{-x^2/2}dx + \int_{0}^{\infty} xe^{-x^2/2}dx \right) = 0.$$

The last step follows from the fact that the function $e^{-x^2/2}$ is symmetrical about $x = 0$, so the two integrals are the same except for the sign. For the variance, we have

$$\begin{aligned}\mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \left[-xe^{-x^2/2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = 1.\end{aligned}$$

In the first line here we used the fact that $\mathrm{E}(X) = 0$; in the second line we used integration by parts; and in the last line we used (6) in the special case $\mu = 0$, $\sigma = 1$. So the standard normal distribution has expectation $\mathrm{E}(X) = 0 = \mu$ and variance $\mathrm{Var}(X) = 1 = \sigma^2$.

Now suppose X has normal distribution with general parameters μ, σ. We claim that the r.v. $Y = \frac{X-\mu}{\sigma}$ has the standard normal distribution. To see this, note that

$$\Pr[a \leq Y \leq b] = \Pr[\sigma a + \mu \leq X \leq \sigma b + \mu] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\sigma a+\mu}^{\sigma b+\mu} e^{-(x-\mu)^2/2\sigma^2} dx = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-y^2/2} dy,$$

by a simple change of variable in the integral. Hence Y is indeed standard normal. Note that Y is obtained from X just by shifting the origin to μ and scaling by σ. (And we shall see in a moment that μ is the mean and σ the standard deviation, so this operation is very natural.)

Now we can read off the expectation and variance of X from those of Y. For the expectation, using linearity, we have

$$0 = \mathrm{E}(Y) = \mathrm{E}\left(\frac{X-\mu}{\sigma}\right) = \frac{\mathrm{E}(X) - \mu}{\sigma},$$

and hence $\mathrm{E}(X) = \mu$. For the variance we have

$$1 = \mathrm{Var}(Y) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{\mathrm{Var}(X)}{\sigma^2},$$

and hence $\mathrm{Var}(X) = \sigma^2$.

The bottom line, then, is that the normal distribution has expectation μ and variance σ^2. (This explains the notation for the parameters μ, σ.)

The fact that the variance is σ^2 (so that the standard deviation is σ) explains our earlier comment that σ determines the "width" of the normal distribution. Namely, by Chebyshev's inequality, a constant fraction of the distribution lies within distance (say) 2σ of the expectation μ.

Note: The above analysis shows that, by means of a simple origin shift and scaling, we can relate any normal distribution to the standard normal. This means that, when doing computations with normal distributions, it's enough to do them for the standard normal. For this reason, books and online sources of mathematical formulas usually contain tables describing the density of the standard normal. From this, one can read off the corresponding information for any normal r.v. X with parameters μ, σ^2, from the formula

$$\Pr[X \leq a] = \Pr[Y \leq \tfrac{a-\mu}{\sigma}],$$

where Y is standard normal.

The normal distribution is ubiquitous throughout the sciences and the social sciences, because it is the standard model for any aggregate data that results from a large number of independent observations of the same random variable (such as the heights of females in the US population, or the observational error in a physical experiment). Such data, as is well known, tends to cluster around its mean in a "bell-shaped" curve, with the correspondence becoming more accurate as the number of observations increases. A theoretical explanation of this phenomenon is the Central Limit Theorem, which we next discuss.

The Central Limit Theorem

Recall from Lecture Note 17 the Law of Large Numbers for i.i.d. r.v.'s X_i's: it says that the probability of *any* deviation α of the sample average $A_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ from the mean, however small, tends to zero as the number of observations n in our average tends to infinity. Thus by taking n large enough, we can make the probability of any given deviation as small as we like.

Actually we can say something much stronger than the Law of Large Numbers: namely, the distribution of the sample average A_n, for large enough n, looks like a *normal distribution* with mean μ and variance $\frac{\sigma^2}{n}$. (Of course, we already know that these are the mean and variance of A_n; the point is that the distribution becomes normal.) The fact that the standard deviation decreases with n (specifically, as $\frac{\sigma}{\sqrt{n}}$) means that the distribution approaches a sharp spike at μ.

Recall from last section that the density of the normal distribution is a symmetrical bell-shaped curve centered around the mean μ. Its height and width are determined by the standard deviation σ as follows: the height at the mean is about $0.4/\sigma$; 50% of the mass is contained in the interval of width 0.67σ either side of the mean, 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 rather than σ^2.)

To state the Central Limit Theorem precisely (so that the limiting distribution is a constant rather than something that depends on n), we shift the mean of A_n to 0 and scale it so that its variance is 1, i.e., we replace A_n by

$$A'_n = \frac{(A_n - \mu)\sqrt{n}}{\sigma} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

The Central Limit Theorem then says that the distribution of A'_n converges to the *standard normal* distribution.

Theorem 18.1: **[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 $A'_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$. Then as $n \to \infty$, the distribution of A'_n approaches the standard normal distribution in the sense that, for any real α,*

$$\Pr[A'_n \le \alpha] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-x^2/2} dx \quad \text{as } n \to \infty.$$

The Central Limit Theorem is 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,[3] 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.

In class, we saw experimentally how the distribution of A_n behaves for increasing values of n, when the X_i have the geometric distribution with parameter $\frac{1}{6}$.

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