18.02 LECTURE NOTES ON PROBABILITY

Continuous Probability

Discrete probability describes games of chance with a list of outcomes, such as whether a coin lands on heads or tails, or a die lands on one of the values 1 through 6. In contrast, *continuous probability* concerns quantities that can take on all possible values in a continuum. In the discrete case, the probability of an outcome or an average value is expressed as a sum, whereas in the continuous case these values are described using integrals.

The basic equation of probability theory is

$$\text{PROBABILITY} = \frac{\text{PART}}{\text{WHOLE}}$$

Note that the probability is a number between 0 and 1.

Example 1. We say that x is *uniformly distributed* on the interval $0 \le x \le 10$ if any value of x is as likely as any other. In this case, the probability that $1 < x < 7$ is

$$\frac{\text{PART}}{\text{WHOLE}} = \frac{7-1}{10} = \frac{6}{10} = 60\%$$

If $0 \le a < b \le 10$, then probability that $a < x < b$, is given by the formula

$$P(a < x < b) = \frac{1}{10}\int_a^b dx = \frac{b-a}{10}$$

The *probability density* of x is $1/10$ on $0 \le x \le 10$ and zero outside this interval. More generally, x can be distributed by a nonnegative function $g(x)$ so that

$$P(a < x < b) = \int_a^b g(x)dx$$

Because the total probability is 1, we need

$$\int_{-\infty}^{\infty} g(x)dx = 1$$

In our example, $g(x) = 1/10$ on $0 \le x \le 10$ and $g(x) = 0$ outside this interval. With continuous variables like x it does not matter whether we include the ends $x = a$ and $x = b$ or not. We interpret the events $x = a$ and $x = b$ as happening with zero probability. Thus, $P(a < x < b) = P(a \le x \le b)$.

Example 2. Consider a point (x, y) distributed according to the weighting or density $\delta(x, y) = x^2 + y^2$ on the unit disk D, $x^2 + y^2 < 1$. The probability that (x, y) is in a portion R of D is

$$P((x, y) \text{ in } R) = \frac{\text{PART}}{\text{WHOLE}} = \frac{\text{mass}(R)}{\text{mass}(D)} = \frac{1}{M}\iint_R \delta dA$$

where $M = \iint_D \delta dA$ is the total mass of D. We also write

$$P((x, y) \text{ in } R) = \iint_R \frac{\delta}{M} dA = \iint_R g(x, y)dA; \quad \text{where} \quad g(x, y) = \frac{\delta(x, y)}{M},$$

In other words, the probability density $g(x,y) = \delta(x,y)/M$ is normalized so that the total integral is 1.

$$\iint_D g(x,y)dxdy = 1.$$

Using polar coordinates,

$$M = \int\int_D \delta dA = \int_0^{2\pi}\int_0^1 r^2 r dr d\theta = \frac{\pi}{2}$$

If R is the ring $a < r < b$, then

$$P(a < r < b) = \int_0^{2\pi}\int_a^b \frac{1}{M} r^2 r dr d\theta = \int_a^b \frac{1}{M} r^3 dr \int_0^{2\pi} d\theta = \int_a^b 4r^3 dr$$

Thus, by integrating in the θ variable, we obtain the probability density in the remaining variable r. The probability density of r in this example is $g(r) = 4r^3$ for $0 \le r \le 1$ and $g(r) = 0$ outside this interval. As usual, the total probability

$$P(0 \le r \le 1) = \int_0^1 4r^3 dr = 1$$

Example 3. The normal distribution. The value of the constant $M = \displaystyle\int_{-\infty}^{\infty} e^{-x^2} dx$ is very important in probability theory. It gives us the normalizing factor to use when defining

$$G(x) = \frac{1}{M} e^{-x^2}$$

as a probability density, that is, M is the constant we need in order that

$$\int_{-\infty}^{\infty} G(x)dx = 1$$

The function $G(x)$ is the well-known bell curve or normal distribution.

To compute $M = \displaystyle\int_{-\infty}^{\infty} e^{-x^2} dx$, rewrite M^2 in a clever way, as in lecture:

$$\begin{aligned} M^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right) = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right) e^{-y^2} dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2} e^{-y^2} dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2-y^2} dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r dr d\theta = \int_0^{2\pi} \frac{1}{2} d\theta = \pi \end{aligned}$$

Therefore, $M^2 = \pi$ and $M = \sqrt{\pi}$. In all, we have

$$G(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}; \quad \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-x^2} dx = 1$$

The importance of this function to probability was discovered by Abraham de Moivre around 1700. We have used the name $G(x)$ in honor of Karl Friedrich Gauss, who laid the foundations of probability theory (along with the method of least squares) in the process of

finding better ways to make use of measurements in astronomy. Any function of the form e^{-ax^2+bx+c} is called a Gaussian.

In order for you to recognize the normal distribution in the future, you will need to recognize all its scalings, related to the parameters a, b and c above. The scaling of the Gaussian will be discussed in the optional section at the end of these Notes. That section is not necessary for 18.02, but it will give you a brief look at some tools and terminology in the theory of probability.

Conditional probability

To choose $0 \le x \le 1$ and $0 \le y \le 1$ independently "at random" means

$$P((x, y) \text{ in } R) = \text{area}(R); \quad R \text{ in unit square}$$

Thus $P(x \ge 1/2) = 1/2$. But the probability changes when we add information:

$$P(x \ge 1/2 \mid xy = 1/1000) = ?$$

This notation means the probability that $x \ge 1/2$ given that we already know $xy = 1/1000$. It is known as a *conditional probability*.

Computing conditional probabilities of this kind is very closely related to computing integrals using a change of variable. The conditional probability density turns out to be the Jacobian factor, renormalized so that the total probability is one.

Recall that if $u = x$ and $v = xy$, then on Thursday we showed that

$$\int_0^1 \int_0^1 dxdy = \int_0^1 \int_v^1 \frac{du}{u} dv$$

Note that the interesting parts of this calculation are that the Jacobian $J = 1/u$ and that the range of u with v fixed is $v \le u \le 1$. Consider any fixed value $xy = v = v_0$, and consider the inner integral

$$\int_{v_0}^1 \frac{1}{u} du$$

The idea is that if $xy = v_0$, then $v_0 \le u \le 1$ is the full range for $u = x$ and that Jacobian factor $1/u$ is the probability density on that interval. We need to need to normalize by this total mass.

$$M = \int_{v_0}^1 \frac{1}{u} du = -\ln(v_0)$$

For $xy = v_0$ fixed, this means that $x > 1$ and $x < v_0$ never happen. In between, $v_0 \le a < b \le 1$,

$$P(a \le x \le b \mid xy = v_0) = \frac{\int_a^b \frac{du}{u}}{M} = \frac{\text{part}}{\text{whole}}$$

Therefore, with $a = 1/2$, $b = 1$, $v_0 = 1/2^{10} = 1/1024$ (close enough to $1/1000$)

$$P(x \ge 1/2 \mid xy = 1/2^{10}) = \frac{\int_{1/2}^1 \frac{du}{u}}{\int_{1/2^{10}}^1 \frac{du}{u}} = \frac{-\ln(1/2)}{-\ln(1/2^{10})} = \frac{\ln 2}{10 \ln 2} = \frac{1}{10}$$

Why this formula works. There is a difficulty with fixing $xy = v_0$. The condition confines us to a single curve, which has zero area and hence zero probability. So when we

divide the part by the whole, we are dividing zero by zero. To repair this difficulty, consider a small interval $v_0 \le v \le v_0 + \Delta v$, carry out the computation of the ratio, and take the limit as Δv tends to zero. This is the correct way of thinking about what it means to know that $xy = .001$ because any computation that reports the value .001 does it with a roundoff error Δv. For practical purposes, the limit of the ratio as Δv tends to zero is the same as the ratio at any fixed band of values of v corresponding to $\pm\Delta v = \pm 10^{-6}$ or smaller. (Matlab is accurate to 15 digits.)

Consider the area of the whole region,

$$\text{area}(v_0 < xy < v_0 + \Delta v) = \int_{v_0}^{v_0+\Delta v} \int_v^1 \frac{du}{u} dv$$

On a very short interval $v_0 \le v \le v_0 + \Delta v$, the inner integral is nearly constant:

$$\int_v^1 \frac{du}{u} \approx \int_{v_0}^1 \frac{du}{u}$$

Therefore,

$$\int_{v_0}^{v_0+\Delta v} \int_v^1 \frac{du}{u} dv \approx \int_{v_0}^{v_0+\Delta v} \int_{v_0}^1 \frac{du}{u} dv = \Delta v \int_{v_0}^1 \frac{du}{u}$$

Similarly, for $v_0 + \Delta v \le a < b \le 1$, the area of the part is

$$\text{area}(a < x < b \text{ and } v_0 < xy < v_0 + \Delta v) = \int_{v_0}^{v_0+\Delta v} \int_a^b \frac{du}{u} dv = \Delta v \int_a^b \frac{du}{u}$$

So the factor Δv cancels in the ratio of the part to the whole,

$$P(a < x < b \mid v_0 \le v \le v_0 + \Delta v) \approx \frac{\Delta v \int_a^b \frac{du}{u}}{\Delta v \int_{v_0}^1 \frac{du}{u}} = \frac{\int_a^b \frac{du}{u}}{\int_{v_0}^1 \frac{du}{u}}$$

In the limit as $\Delta v \to 0$,

$$P(a < x < b \mid v = v_0) = \frac{\int_a^b \frac{du}{u}}{\int_{v_0}^1 \frac{du}{u}}$$

There are pitfalls in dealing with zero area sets like $xy = v_0$. For example, one may be tempted to use the ratio of the arclength of the curve on the portion $a < x < b$ to the total arclength of $xy = v_0$. This gives the wrong answer. The reason is that the band $v_0 \le xy \le v_0 + \Delta v$ does not have uniform thickness. If it did, then the weighting or density would be equivalent to arclength. Put another way, if one tries to partition the square into subsets of uniform thickness around curves of the form $xy = v_0$ this cannot be done without overlapping bands or bands that miss substantial sections. By contrast, the bands $v_0 \le xy \le v_0 + \Delta v$ are compatible with partitioning without gaps or overlaps: use pieces of the form $k\Delta v \le xy \le (k+1)\Delta v$, $k = 0, 1, \ldots$.

Expected value, variance, and standard deviation: Rescaling the normal distribution.

This section is optional reading. It introduces a few standard notions and terminology from probability theory. Recall that if a variable x is distributed with a density $g(x)$, then

$$P(a < x < b) = \int_a^b g(x)dx, \quad g(x) \geq 0$$

Since the total integral is 1, the average value or *mean* of x is

$$\mu = \int_{-\infty}^{\infty} xg(x)dx$$

(μ for mean). In probability, the upper case letter X denotes something called a random variable, which can be viewed as a quantity that will vary depending on each sampling of the variable.[1] The *mean* or *expected value* of X, $E(X)$, is a theoretical value for what one would expect to get if one averaged over several samples of the variable X. This quantity is the same as the average value or mean value of x,

$$E(X) = \int_{-\infty}^{\infty} xg(x)dx = \mu$$

One can take the expected value of any function of $f(X)$. The formula is

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

Again, this is a weighted average of f. The expected value of X is just the special case $f(X) = X$. It is also interesting to evaluate the expectation of functions like $f(X) = X^k$, called the kth moments of X, and $f(X) = e^{tX}$.

The *variance* $V(X)$ is a measure of the likelihood that X is far from its mean. It is the average of the square of the distance from X to its mean, $(x - \mu)^2$,

$$V(X) = E((X - \mu)^2) = \int_{-\infty}^{\infty} (x - \mu)^2 g(x)dx$$

Because the variance involves the square of distance, it is natural to take a square root (as in the Pythagorean theorem). The *standard deviation* is defined by

$$\sigma(X) = \sqrt{V(X)}$$

We can now explain the scaling of the normal distribution. There is a probability density for each $\sigma > 0$,

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

It has three properties:

$$\int_{-\infty}^{\infty} g_\sigma(x)dx = 1 \quad \text{(total integral 1)}, \tag{1}$$

[1] For instance, in Matlab the command rand(3,4) produces a 3×4 matrix with each element chosen uniformly distributed between 0 and 1. The entries are also independent of each other. Independence is a key concept in probability. But it would take us too far afield to discuss it in any detail.

$$\int_{-\infty}^{\infty} x g_\sigma(x)dx = 0 \quad \text{(mean 0)}, \tag{2}$$

$$\int_{-\infty}^{\infty} x^2 g_\sigma(x)dx = \sigma^2 \quad \text{(variance } \sigma^2\text{)}. \tag{3}$$

To shift the mean to μ, take

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

The standard deviation σ or the variance σ^2 measures how far the distribution is from its mean value. For larger σ, $g_\sigma(x)$ is flatter and has a smaller maximum value $1/\sqrt{2\pi}\sigma$. In other words, it is more weighted towards larger values, and X^2 is more likely to take on larger values. The formula for the variance above is an exact, quantitative expression of this qualitative comparison between the shapes of the graphs of the densities g_σ for different values of σ.

The function $G(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$ considered in Example 3 above equals g_σ for $\sigma = 1/\sqrt{2}$. Thus G is the normal distribution with mean 0 and standard deviation $1/\sqrt{2}$ (variance 1/2).

To confirm (1), (2), and (3), recall that we already showed that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

Change variables by $x = az$, $dx = adz$, to get

$$\int_{-\infty}^{\infty} e^{-a^2 z^2} a dz = \sqrt{\pi}$$

Putting $a^2 = \frac{1}{2\sigma^2}$,

$$\int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} \frac{1}{\sqrt{2}\sigma} dz = \sqrt{\pi} \tag{4}$$

and dividing by $\sqrt{\pi}$ gives (1).

Next, multiply (4) by $\sqrt{2}\sigma$ to obtain

$$\int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} dz = \sqrt{2\pi}\sigma \tag{5}$$

Differentiating the left side of (5) with respect to σ gives

$$\frac{d}{d\sigma} \int_{-\infty}^{\infty} e^{-z^2/2\sigma^2} dz = \int_{-\infty}^{\infty} \frac{\partial}{\partial\sigma} e^{-z^2/2\sigma^2} dz = \int_{-\infty}^{\infty} \frac{z^2}{\sigma^3} e^{-z^2/2\sigma^2} dz$$

On the other hand the derivative of right hand side of (5) with respect to σ is $\sqrt{2\pi}$. Hence,

$$\int_{-\infty}^{\infty} \frac{z^2}{\sigma^3} e^{-z^2/2\sigma^2} dz = \sqrt{2\pi}$$

Dividing by $\sqrt{2\pi}$ and multiplying by σ^2 yields (3).

Finally, (2) is obvious because the integrand is odd. To confirm that the mean is μ for the density $g_\sigma(x-\mu)$, change variables to $z = x - \mu$. Then, using (2) and (1),

$$\int_{-\infty}^{\infty} x g_\sigma(x-\mu)dx = \int_{-\infty}^{\infty} (z+\mu) g_\sigma(z)dz = \int_{-\infty}^{\infty} z g_\sigma(z)dz + \mu \int_{-\infty}^{\infty} g_\sigma(z)dz = 0 + \mu = \mu$$