Probability

Kenneth A. Ribet

Math 10A
Election Day, 2016

Yesterday's pop-in lunch!

Breakfast this Thursday (November 10) at 9AM.

Breakfast next Monday (November 17) at 9AM.

Friday is a UC holiday.

Today's class is, in part, a review of the fundamental concepts that we've seen so far. We'll move on to new stuff after a few minutes. I've tried to make up detailed slides for these reasons:

- I'm going to a somewhat fancy lunch today at noon and may not have time to change into my chalk-friendly jeans.
- Some of you are poll watching.
- Some of you are freaked out by a chemistry exam.
- Some of you are worried about tomorrow's 10A quiz.

Review

A probability space is a set Ω with a probability function P on subsets of Ω. For example, Ω could be the set of candidates for US President and $P(\omega)$ for $\omega \in \Omega$ could be your estimate for the probability that ω will be elected.

A random variable is a numerical function X on Ω. We write

$$X : \Omega \to \mathbf{R}.$$

For example, $X(\omega)$ could be the dollar amount that you win if ω is elected. If $X(\omega)$ is negative, this means that you lose money if ω is elected.

The online notes (and maybe also the book) says that a random variable is *discrete* if the set of its values is finite (or a "discrete") infinite set like the set of integers.

Meanwhile a random variable X is *continuous* if there's a PDF $f(x)$ such that

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

for real numbers a and b with $a \leq b$.

Generally mathematicians think of "discrete" and "continuous" as antonyms, but it should be easy to concoct Xs that are neither discrete or continuous. Nonetheless, it's basically the case that all Math 10 random variables are either discrete or continuous.

Further, it's supposed to be true that:

- Discrete math is done when we spring forward;
- Continuous math is done when we fall back.

However, Math 10A is full of discrete examples that motivate the continuous definitions. (See online notes and the Schreiber book.)

Review

The CDF of a random variable X is the function

$$F(x) = P(X \leq x).$$

In the continuous case, this means:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

We write "$f(t)\,dt$" instead of "$f(x)\,dx$" simply to avoid having x play two roles in the same formula.

In probability theory, a CDF can be any decent function $F(x)$ that increases as x increases and runs between 0 (at $x = -\infty$) and 1 (at $x = +\infty$). For a continuous X, the CDF is the antiderivative of $f(x)$ that takes the value 0 "at" $-\infty$. For a discrete X, the CDF is a "step function."

Note that the variable x stands for a real number while X is a real-valued function.

The *mean value* or *expected value* of a random variable X is the quantity

$$E[X] = \int_{-\infty}^{\infty} x\, dF(x).$$

In the continuous case, which is our focus today, $dF(x) = F'(x)\, dx = f(x)\, dx$, where $f(x)$ is the PDF. In that case:

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

.

For quite a few cases that we care about, the value of $E[X]$ is easy to figure out for symmetry reasons. For example, if $f(x)$ is even, then $E[X] = 0$, as we saw on Thursday. More generally, if the graph of $f(x)$ is symmetric about the line $x = \mu$, then μ is the expected value of X.

Finally, if a is a real number, the expected value of $X - a$ is $E[X] - a$.

Variance

The *variance* of X is the quantity

$$\text{Var}[X] = E[(X - \mu)^2],$$

where $\mu = E[X]$. Thus Var[X] tries to measure how far X strays from its mean, μ, on average.

The *standard deviation* of X is the square root of Var[X]. Because Var involves squaring, it makes sense to some people to take a square root at the end.

If X is discrete with values $\{x_i\}$, then

$$\mathrm{Var}[X] = \sum_i (x_i - \mu)^2 P(X = x_i).$$

The corresponding formula for a continuous X with PDF $f(x)$ is

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx.$$

As Schreiber explains, you can derive the continuous formula from the discrete formula by taking limits: the sums for the approximating discrete random variables become Riemann sums that converge to the desired integral.

A really important case is that where X is continuous and has PDF equal to the standard Gaussian

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The mean for this Gaussian is 0, so the standard deviation is just

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\, dx.$$

We will calculate the integral (using integration by parts) and see that its value is $\sqrt{2\pi}$.

Thus the standard Gaussian has mean 0 and standard deviation 1.

Because Gaussians are (regarded as) so important, you should be able to compute the mean and standard deviation associated to the more general PDF

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

The mean is μ and the standard deviation is σ. (We take $\sigma > 0$.)

Chebyshev

Suppose that X has mean μ and standard deviation σ. Then

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

for integers $k \geq 1$. For $k = 1$, this just says that the probability is non-negative—that's no new information. For $k = 2$, it says that the probability of being no more than two standard deviations out from the mean is at least $\frac{3}{4}$.

For example:

$$\frac{1}{\sqrt{2\pi}} \int_{-2}^{2} e^{-x^2/2}\, dx \geq \frac{3}{4}.$$

The numerical value of the integral is ≈ 0.9545 according to Wikipedia. For $k = 3$, the numerical value is ≈ 0.9973.

Proof of Chebyshev

The proof for continuous random variables is given on page 553 of the Schreiber text. I'm going to explain the proof on the chalkboard to make sure that everyone in the room is on board with it.

Hope that's OK.