# MA 2B RECITATION 01/26/11

## 1. Announcements

The midterm will be handed out next week and will probably be due Monday, February 6. Official announcements and midterm review times will be announced shortly.

## 2. Probability Densities

We're now ready to study continuous distributions in depth. While there are a number of things about continuous distributions that you need to be careful about, the generalization from discrete to continuous is straightforward. Essentially, all you are doing is replacing sums with integrals.

Let's start with a warmup exercise.

**Example 1.** You forgot the 8-digit password (only numbers) necessary to log into your computer. If you try all possible passwords at random with uniform probability, discarding the unsuccessful ones, what is the expected number of attempts needed to find the correct password?

*Solution.* Let $n = 10^8$ be the number of all possible passwords and let $X$ be the number of attempts before the correct password is found (including the successful attempt). The probability that the correct password is found at the $k$th attempt is

$$P(\{X = k\}) = \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdots \frac{n-k+1}{n-k+2} \cdot \frac{1}{n-k+1} = \frac{1}{n}.$$

This is because $\frac{n-1}{n}$ is the probability that the password is not found at the first attempt, $\frac{n-2}{n-1}$ is the probability that it is not found at the third attempt, etc. Finally, $\frac{1}{n-k+1}$ is the probability that the password is found at the $k$th attempt. Therefore,

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} kP(\{X = k\}) = \sum_{k=0}^{n} \frac{k}{n} \\ &= \frac{n+1}{2} = \frac{100,000,001}{2} = 50,000,000.5. \end{aligned}$$

□

These are the kinds of problems that you should be able to do on your exam; namely, to apply the language of probability to analyze a specific scenario. The gambler's ruin problem from the last homework is also a good example.

*Date*: January 26, 2011.

### 2.1. **Highlights.**

Here are the most important concepts to absorb.

Let $X$ be a random variable with density function $f$. We define the **expectation** to be

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

We are also very interested in two quantities related to the expectation: the **variance**, given by

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2$$

and the **standard deviation** given by

$$SD(X) = \sqrt{\mathrm{Var}(X)}.$$

*Remark* 2.1. Since these concepts are so important, we often denote the standard deviation by $\sigma$ and the variance by $\sigma^2$.

How should you think about these terms? Expectation, like I said before, it what is sounds like. It is the value that your variable is "expected to take." In the continuous case, we're just replacing the sum with an integral. (Did you know that the integral symbol is just a long $S$? The notation was actually standardized by Leibniz, whose picked a the long $S$ [actually, a long "ſ" in German], to stand for "summa" or "sum.")

Similarly, the variation is just a measure of the amount of variation of the values of that variable. Historically, probabilists chose the most boring and obvious words to name their definitions and theorems, so you end up with things like "expectation", "variance", and "Central Limit Theorem." However, the world of probability is fascinating enough, so we can forgive them for uninspired nomenclature.

Things to keep in mind when working with continuous probability:

- The properties and rules about things like expectation that we showed for discrete distributions carry over to the continuous case. For instance, if $X$ and $Y$ are independent RVs, we still have $E(XY) = E(X)E(Y)$.
- There is a great summary of facts on p.248-249. In particular, you might often find it easier to try and use one of the tricks listed there than to blindly try to integrate to get the expected value.

While we might have the all the definitions above, it doesn't mean that we can always find them. Here's an example of a (discrete) random variable with undefined variance.

**Example 2.** Let $X$ be a RV with distribution

$$P(\{X = \frac{3^k}{2^k}\}) = \frac{1}{2^k}$$

for $k = 1, 2, \ldots$. This discrete distribution is well-defined since

$$\sum_{k=1}^{\infty} \frac{1}{2^k} = \frac{1/2}{1-\frac{1}{2}} = 1.$$

The expectation is given by

$$E(X) = \sum_{k=1}^{\infty} \frac{3^k}{2^k} \cdot \frac{1}{2^k} = \sum_{k=1}^{\infty} \left(\frac{3}{4}\right)^k = \frac{3/4}{1-\frac{3}{4}} = 3$$

which converges since $\frac{3}{4} < 1$. However, if we try and compute the expectation of $X^2$, we get

$$E(X^2) = \sum_{k=1}^{\infty} \left(\frac{3^k}{2^k}\right)^2 \frac{1}{2^k} = \sum_{k=1}^{\infty} \left(\frac{9}{8}\right)^k$$

which diverges since $\frac{9}{8} > 1$. Therefore, the expectation of $X^2$ is undefined and so the variance of $X$ is also undefined.

2.2. **Why the Normal Distribution is Awesome.** The reason why the normal distribution is so amazing is because of the following result. The rigorous definition is in the book. Here's how you should think about it.

**Theorem 2.2.** ***(Central Limit Theorem)*** *If you have $n$ independent identically distributed (IID) random variables with finite variance, then as $n \to \infty$, the sum (or average) of the $n$ variables "converges" to the normal distribution.*

*Remark* 2.3. The convergence here is a little special, it's called *convergence in distribution*, and so it's saying that the cumulative distribution function converges pointwise to the normal distribution.

Maybe the awesomeness didn't sink in, so let's take a look at it again. What this says is that if you have IID random variables, *regardless of distribution*, their sum as $n \to \infty$ *always* has a normal distribution. The random variables could have a Poisson, Binomial, uniform, or some crazy distribution that's not even common enough to have a name, but as long as they are IID, their sum will *always be normal*.

This is a magical mathematical fact, and is the reason why the normal distribution has such a peculiar definition. (Why do we divide by $1/\sqrt{2\pi}$? Why is the exponential of some squared value? Why is knowing just the mean and variance enough?) It's because the normal distribution is just the right thing that these sums converge to. So we know the how. But do we the why? Why does the world behave this way? Answering that question leads to deep philosophical questions about the nature of the universe or discussions about the unreasonable effectiveness of mathematics, among other things. But that is far beyond the scope of this class.

Instead, let's try and apply this theorem!

**Example 3.** Suppose you have some process that spits out random numbers, like a random number generator, a numerical simulation, a machine that's measuring some phenomenon, etc. Suppose that you find out somehow (or just know) that the mean of the numbers is 10 and the standard deviation is 5. Suppose that you run the process 100 times and record the results. What's the probability that after getting 100 numbers out of my generator, their sum is over 1100?

*Solution.* We can use the Central Limit Theorem. While we don't know the distribution, we do know that the sum of the random numbers approaches a normal distribution.

What we need to do now is the scale the problem by the standard deviation, and then use the information we have about the standard normal distribution. In probability language, we want to find

$$P(X_1 + \cdots + X_{100} \geq 1100).$$

The Central Limit Theorem says that this sum has the same distribution as a random variable $Y$ that is normal with mean $\mu = 10 \cdot 100 = 1000$ and standard deviation $\sigma = 5 \cdot \sqrt{100} = 50$. Therefore,

$$P(Y \geq 1100) = 1 - \Phi\left(\frac{1100 - 1000}{50}\right) = 1 - \Phi(2) = 0.02275.$$

□

## 3. Change of Variable

Given a random variable $X$ and a function $g(X)$, we're often interested in the behavior of $g(X)$. What can we learn about $g(X)$ if we only know the distribution of $X$? It turns out that in a particular situation, we can learn quite a lot.

**Theorem 3.1.** *Let $X$ be a random variable with density $f_X(x)$ on the range $(a, b)$. Let $Y = g(X)$, where $g$ is* ***strictly increasing or decreasing*** *on $(a, b)$. The range of $T$ is then an interval $(g(a), g(b))$, and the density of $Y$ on this interval is given by*

$$f_Y(y) = f_X(x) \frac{1}{|\frac{dy}{dx}|}$$

*where $y = g(x)$.*

There are techniques to analyze the distribution of $g(X)$ without the assumption that I bolded above, but they are usually studied in more advanced courses in probability and will probably not be touched upon in this class.

## 4. Continuous Joint Distributions

These are defined as you might expect, as an analog of the discrete case. Namely, if $f$ is the joint distribution of random variables $X$ and $Y$, then

$$P((X, Y) \in B) = \iint_B f(x, y)\, dx\, dy.$$

Given $f$, we can find the marginal density, say, of $X$:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

**Example 4.** Suppose that $X$ and $Y$ are uniform random variables on $[0, 1]$. What is the probability that $X > Y$?

*Solution.* We define $B = \{(x, y) \mid x > y\}$. Then our desired probability is

$$\iint_B 1\, dx\, dy = \int_0^1 \int_y^1 1\, dx\, dy = \int_0^1 (1 - y)\, dy = 1 - \frac{1}{2} = \frac{1}{2}.$$

An alternate way to do this is that because $X$ and $Y$ are uniform on $[0, 1]$, we can visualize them as a square, and this is just the area of the triangle above the triangle but below the line $y = 1$. □

## 5. Independent Normal RVs

If you have independent normal variables, don't integrate! Any linear combination of normal variables is normal, look at the formula on page 363!

What if you multiple a $(0,1)$ normal distribution by a factor $c$? As expected, you get a $(0,c^2)$ distribution. What about in general? We can use the change of variables formula above. We are interested in the distribution of $cX$ where $X$ is a $(\mu,\sigma^2)$ normal random variable. Here $y(x)=cx$, so $x(y)=\frac{1}{c}y$. Therefore, the distribution function is

$$\begin{aligned} f_{cX}(y) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{1}{c}y-\mu)^2/\sigma^2}\frac{1}{c} \\ &= \frac{1}{\sqrt{2\pi}c\sigma} e^{-\frac{1}{2}\frac{1}{c^2}(y-c\mu)^2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}c\sigma} e^{-\frac{1}{2}(y-c\mu)^2/(c\sigma)^2}, \end{aligned}$$

which we see is density of the $(c\mu,(c\sigma)^2)$ normal distribution.

*Remark* 5.1. While we usually only care about the mean and the standard deviation when talking about the normal distribution, when we write down a $(\mu,\sigma^2)$ normal distribution, we are writing down its mean and variance.

**Adding is NOT multiplying.** Note that if we add $n$ normal random variables, the distribution of the sum has variance multipled by $n$, so the standard deviation is multipled by $\sqrt{n}$.

However, if we multiply a single random variable by $n$, the variance gets multiplied by $n^2$, as we saw.

This makes sense, since it's easier to make a single variable larger than to make the sum larger, where you must have all the summands large. This is an important point, and obvious once you "get it," but can be tricky if you're not used to thinking this way. *Make sure you understand!*

**Example 5.** Pick 2 numbers from a standard normal distribution and find their difference. What is the probability that their difference is greater than 1?

*Solution.* From page 363, we know that the distribution of $X+Y$ for standard normal independent $X$ and $Y$ is a normal (0,2) distribution. However, we want to $X-Y$. But the distribution of $-Y$ is the same as $Y$ because the normal distribution is symmetric the mean 0. Therefore, the distribution of $X-Y$ is also a normal (0,2) distribution.

Therefore, to find the probability that (the absolute value of) this is bigger than 1 is just

$$\Phi\left(\frac{-1}{\sqrt{2}}\right)+\left(1-\Phi\left(\frac{1}{\sqrt{2}}\right)\right)\approx 0.4795.$$

□

## 6. Operations

However, when we have just two general random variables $X$ and $Y$ with densities $f_X$ and $f_Y$, how do you find the density of $X+Y$? We just convolution. We saw this in the discrete case, and it is the same idea for the continuous case:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx.$$

**Example 6.** What the distribution of the sum of two uniform random variables $X$ and $Y$ on $(0,1)$? We know that $f_X(x) = f_Y(x) = \chi_{[0,1]}$. (Here, $\chi_{[0,1]}$ is the characteristic function on $[0,1]$, which is 1 at $[0,1]$ and 0 everywhere else. We saw this last quarter as a "staircase function" often used in situation where we applied the Laplace transform.)

Clearly $f_{X+Y}(z) = 0$ if $z \notin [0,2]$, so we will assume $z$ is in that interval

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \\ &= \int_{\max(z-1,0)}^{\min(1,z)} f_X(x)\, f_Y(z-x)\, dx \\ &= \int_{\max(z-1,0)}^{\min(1,z)} dx \\ &= \min(1,z) - \max(z-1,0). \end{aligned}$$

The graph of this is just z from 0 to 1 and $2-z$ from 1 to 2.

## 7. An interesting example

**Example 7.** Three candidates $A$, $B$, and $C$ are running for President. In order to predict the outcome of the election, a number of people, each selected at random and independently of the others, are asked their choice. The predicted percentage $p_X$ of votes for each candidate $X \in \{A,B,C\}$ is worked out by dividing the number of people in the sample who are going to vote for $X$ by the total number of people in the sample. How many people need to be in the sample to substantiate the following claim: (the wording is important) *For each candidate $X$, the probability that the predicted percentage $p_X$ is correct to within 1% is at least 0.95?*

(Note that I am saying "for each candidate, the probability that" instead of "the probability that for each candidate." The latter is a stronger statement, and more difficult. It also requires a larger sample.)

What sample size will make it possible to make the same claim if there are just 2 candidates? What about $n$ candidates? (Exclude the case $n = 1$, even though that is a common case in certain political systems.)

*Solution.* We need to estimate the smallest number $n$ of people in the sample such that $P(\{|p_X - r_X| \leq 0.01\}) \geq 0.95$ for any $r_X \in [0,1]$ for each candidate $X \in \{A,B,C\}$.

Let $S_X$ be the number of people in the sample who are going to vote for candidate $X$. Assuming that each person's preference is independent of the others, we can consider $S_X$ to be the sum of $n$ independent random variables, each taking value 1 with probability $r_X$ if a person will vote for $X$ or 0 otherwise.

The variance of each of these independent random variables is thus $r_X(1-r_X)$. Thus, by the central limit theorem,

$$P(\{|p_X - r_X| \leq 0.01\}) = P\left(\left\{\frac{|S_X - nr_X|}{\sqrt{nr_X(1-r_X)}} \leq 0.01\sqrt{\frac{n}{r_X(1-r_X)}}\right\}\right)$$
$$\approx \frac{1}{\sqrt{2\pi}}\int_{-0.01\sqrt{\frac{n}{r_X(1-r_X)}}}^{0.01\sqrt{\frac{n}{r_X(1-r_X)}}} e^{t^2/2}\,dt \geq 0.95.$$

This will be satisfied if

$$0.01\sqrt{\frac{n}{r_X(1-r_X)}} \geq 2.$$

In fact, we could use about 1.96 instead of 2.

Because for $r_X \in [0,1]$, the largest value of $r_X(1-r_X)$ is $\frac{1}{4}$, we find that this estimate is satisfied for all $r_X \in [0,1]$ if

$$n \geq 10,000.$$

The same value will be obtained for each candidate $X \in \{A, B, C\}$.

Therefore, a sample of about 10,000 people is sufficient to claim that the predicted percentage of votes is accurate to within 1% with probability at least 0.95 for each candidate. The answer the same for any number of candidates greater than 1. □

Given this, why don't we know the outcomes of elections? Or why don't we just save money by picking 10,000 people at random to decide the next President? The wording was an issue, of course. But the bigger thing is that peoples' preferences are usually far from independent.