Week 5 - Oct 5 - The Structure of Randomness

This week we take a slight detour into the world of Probability. This material is fairly abstract. In fact it is least as philosophical as mathematical. And historical too, thanks to The Drunkard’s Walk book.

As that book explains, it is a long and difficult intellectual process to understand how we can separate random chance from real, predictable processes. So this week our focus will be on conceptual understanding so that you really know what's going on when we get statistics like the famous "margin of error". In a few weeks in your assignment 2 you will have to use the right language to explain what those statistics mean. In the final exam, you'll have to explain the conceptual intellectual steps required to do so.

This week you should be learning:

- Why we need inferential statistics at all!
- That the social/political/economic events we want to understand are usually combinations of systematic and random causes.
- The structure of randomness is somewhat predictable. The predictability is based on the laws of probability.
- The most important structure of randomness is the Normal Distribution (bell curve).

You'll learn most of that here and in reading the sections of the Drunk book that I point you to.

In our first class we'll see how probabilities add up and in the second we'll try to learn some fundamental things about your shoes, which is just like learning about wars or voting turnout.

This week in the lab you won't be doing anything directly related to this material; you will continue to develop your data analysis skills in Stata.

This week's material is fairly abstract. But keep in mind that we always apply these statistical-philosophical ideas to real cases where we want to learn about how the world really works.

DO NOT EXPECT THINGS TO BE CRYSTAL CLEAR THIS WEEK. PART OF THE OBJECTIVE IS TO HAVE YOU DOUBT THINGS YOU NATURALLY THINK AND WONDER ABOUT HOW THE WORLD WORKS.

Why do Things Happen

- What an absurd title to kick off this week's material? It's a bit absurd, but since it's a very general question we need a very general answer.

There's an answer running through the Drunkard's Walk and through the history of ideas that it presents. The answer appears first on page 4 (have a look), but it's kind-of hidden there. He
writes: "The outline of our lives, like the candle's flame, is continuously coaxed in new directions by a variety of random events that, along with our responses to them, determine our fate. As a result, life is both hard to predict and hard to interpret."

When you first read this, in light of the intense story in the page before that, you probably don't really notice that this passage gives an answer to the "Why do things happen?" question. You probably see this passage as summarizing a lesson about randomness in our lives and not about the combination of systematic, predictable forces AND randomness. That's the real answer: Things happen (are caused by) two kinds of things:

1. systematic, predictable forces or actions
2. a collection of random forces (plural! lots of them!)

Take a moment and think of something that happened to you today that was affected by a 'systematic, predictable force or action' -- call that experience A. Now think of a second experience 'B' that was affected by some random forces. Was B also affected by some systematic/predictable forces? Was A also affected by random forces? Here's a question to keep in mind: how do we know which forces were random and which are systematic/predictable?

Notice that Mlodinow didn't write that life is impossible to predict or impossible to interpret; just hard to predict and hard to interpret.

The theme of the book is that prediction of the future and interpretation of the past is fundamentally about separating predictable forces from random forces. The answer to our question appears slightly more explicitly over the paragraph spanning pages 10-11 and the next one. And it will appear many times in various guises throughout the book.

Now, if you're really thinking about this, and reading the Drunkard's book with this framework in mind, you will start to realize that this is not a binary distinction: systematic versus random, even though we often contrast them. In life and in social science statistics, when we have something we want to know about, we usually find that there is a continuum from causes we know well to causes that we can't know or describe systematically.

Think about smoking and lung cancer for a moment. We know that smoking is a systematic cause of lung cancer. We know that does not mean that everyone who smokes will get lung cancer. And we know that not everyone who gets lung cancer was a smoker. Smoking is not a 'deterministic' cause of lung cancer. We know that lots of other genetic and environmental factors are responsible for whether individual smokers and non-smokers alike get lung cancer. I was tempted to write "lots of other, random, genetic and environmental factors" but that would give you the impression that we can't ever account for some of those factors and that I'm confident they are truly random.

So back to the opening question: why do things happen. In this case, we ask: why do some people get lung cancer. We can break our answer into three categories: 1) There are some known systematic factors like smoking. 2) Some other (currently) unknown systematic factors. For instance maybe using air freshener increases the risk of lung cancer. It's possible, but we don't
know yet. 3) Truly random forces. To understand this last category, let's take an example where we know an outcome is random. If you toss a fair coin, we know the chances of getting heads or tails are the same. So the 'cause' of the outcome of any particular toss is randomness.

So we can think of non-systematic causes not as fundamentally random, but as things we don't know enough about or can't measure well enough to know if they themselves are systematic causes.

So how do we know that smoking is a significant cause of lung cancer? Or that shipments of automatic weapons is a significant cause of civil conflict in developing countries? Or that in the US, getting a raise is a significant cause of voting for Republican candidates?

Think of how you would answer this "how do we know" question before you go on to the next page.

- Individual Cases are Just Cases; Patterns Emerge when we look at many cases

How can smoking be a cause of lung cancer if, for instance, two women who live on my street smoked for over 60 years and didn't get lung cancer?

I'm sure you'll agree that smoking made these women more likely to get lung cancer. So they didn't get lung cancer, presumably, because of a whole set of causes we have to consider random because we don't know enough (yet) to know that they are causes of not getting lung cancer.

One of the key ideas discussed in the Drunkard's Walk is that our beliefs about the role of systematic causes and random processes can depend on whether we look at individual cases or a whole large set of cases.

For instance, when we talk about the record of a single hollywood studio executive there is a tendency to downplay the role of randomness because we think we know so much about specific causes. But if we take a step back, it becomes clear that randomness can play an important role in outcomes.

The author of The Drunkard's Walk is not very direct about this point because he likes a good personal story and because he wants to let the point unfold over the course of the book. But I'm warning you now: For us the book is not about individual stories and random twists and turns in life. And it's not really about how humans make mistakes as they interpret data, although we will talk about this a bit. The key idea for us is that the structure of randomness is somewhat predictable.
Another way to put this point is to look at the following website and understand how it drives home the point that most history is really interesting but **totally useless** for a certain kind of understanding that helps us predict the effects of (potential) causes: [http://www.cracked.com/article_17298_6-random-coincidences-that-created-modern-world.html](http://www.cracked.com/article_17298_6-random-coincidences-that-created-modern-world.html). (Don't blame me if you waste time clicking other links on the site).

For our purposes, the point of the *Drunkard's Walk* book is about how we see "the order in chaos", the title of Chapter 8. Read just the first three pages of that chapter now.

And then, to see how this really matters for politics, have a look at [http://themonkeycage.org/blog/2012/01/27/electoral-fraud-in-russia-report-from-the-russian-blogosphere/](http://themonkeycage.org/blog/2012/01/27/electoral-fraud-in-russia-report-from-the-russian-blogosphere/) and then remember that when you turn back to the *Drunkard's Walk* book again and read the first few pages of chapter 5!

- **Chaos first, Order second**

The big story of the book is that over many centuries, and thanks to some pioneering thinkers, we came to realize that since everything is chaos -- or affected by lots of random causes -- we need to know what the patterns produced by chaos -- randomness -- look like before we can recognize systematic, orderly departures from that overwhelming randomness.

The intellectual tool we need to understand randomness is obviously **PROBABILITY**. If some process that we repeat doesn't turn out the same way every time, we can only describe the pattern of ways that it does turn out -- the "distribution" of the ways it turns out, if you will. Flipping a coin and throwing six-sided dice are the most familiar examples, but they are actually confusing ones for understanding how probability can help us learn about social and political affairs because they seem like a physical process that we can fully describe. That's because we think of them as perfect 'probability machines' where we know that in the very very long run they will produce equal numbers of each outcome (1/6th of the time for each face of the die, for example). The other thing about them that we think about less, however, is that for any **one time** you flip a coin or throw a die, the result is determined by the underlying equal probability **AND** a whole bunch of physical factors like how you're holding it, how far it drops, how hard you throw it, and the surface it lands on.

To properly use dice and coins to understand probability you only learn the first of those two things: How each flip is 50/50, how the chances of getting three heads in a row is 1/8, how the chances of getting four of the same side in a row is also 1/8, and so on. That's the kind of thing covered, for example, in chapter 4 of the *Drunkard's Walk*.

But to properly use probability to help you with social science, you need to think of the second of those things: that a given coin flip is a product of the underlying process (the 50/50-ness of the coin) and a whole bunch of randomness. So the really huge new idea for you in this course is
that once you start thinking of individual people like Sherry Lansing (movies) and Roger Maris (baseball) and Barack Obama merely as cases of some bigger phenomena you care about (box office returns, home runs, re-election of an incumbent President), then you are forced to think about your phenomenon as being produced by a set of forces just like a set of physical forces produces a head or a tail on a coin flip. Just different forces, with different probabilities. Some are systematic and some are random.

In other words, underlying any political outcome we want to analyze -- civil war, environmental performance of countries, re-election, income distribution -- is a process that generates outcomes \textit{probabilistically}. For example, at any given time there either is or is not militarized civil conflict (MCC) in a set of countries. That's like a coin flip in the sense that it is a binary variable: conflict or not-conflict. And in any given year, a developing country has a certain probability of having a civil conflict. Imagine it's just like a coin: 50/50. Then if all the other causes of civil conflict are really unknowable and thus we consider them random, we should observe half the countries in any given year with a civil conflict. Of course, there would be a lot of randomness that's adding up to the 50/50 chance of a civil conflict and sometimes we would see more than half of the countries with conflict and other years we'd see less than half. But if we somehow knew that the systematic processes that generate these conflicts add up to 50/50 we would then be in a position to look for certain features of countries that cause them to have less or more conflict. If we observed that only one in four democracies had these conflicts while three in four non-democracies did, we'd start to think that democracy is a systematic cause that prevents civil conflict. See, we had to know about randomness (or, chaos) before we could recognize systematicness (or, order). (I made up that word).

Just like you said you were doing when you told yourself how you know that smoking causes lung cancer.

But here's the problem that makes it even more important that we understand probability and how to apply it: We don't know for sure that countries are 50/50 to have a conflict. We would only \textbf{INFER} that from some data we collect. Countries aren't coins, after all. And as you'll see this week, shoes aren't coins. And shoes are more like countries. You'll see what I mean if you come to class on Wednesday.

\textbf{Now you should read chapters 3, 4, and 5 in the Drunkard's Walk. BUT READ THE LAST PARAGRAPH OF CHAPTER 5 FIRST}. The point is that in chapters 3, 4, 5 you learn the conceptual foundation of probability when we know the underlying probabilities (like a coin flip). It's a different thing to take this knowledge of probability and use it when we don't know the underlying probabilities. A challenge of \textit{inference}.

In our second class this week we'll experience a very strange example of this challenge.

- \textbf{Systematic, Underlying, Orderly Causes (probabilities and means)}

\textbf{Enabled:}
Statistics Tracking
I'm afraid what we mean by underlying causes here is a little different than the way we've been thinking about causes so far. We have been thinking of them the natural way: as in A causes B. But here we're going to think of systematic causes as underlying probabilities of outcomes of a given 'data generating process'; while all other things that create the outcome are randomness. Remember, we're only dealing with one variable at a time for the moment, so we can't think of causes as represented by other variables.

Flipping a coin involves an underlying probability of .5 for both outcomes. The results of actual coin flips essentially start at .5 and then randomness gives you a particular flip of heads or tails. Simple enough. But we'll use this same way of thinking for real political data. So if the probability of a trade dispute between any two countries is .5, then the underlying probability is .5 and the actual result -- whether or not there's a trade dispute -- is that starting point (.5), plus randomness. (Forget, for the moment, that it's not all randomness and we know there are other causes). If the probability of a full scale war between any two countries is .02, then the actual result -- war or peace -- is .02 plus randomness (all other causes). If the probability of an American voter voting for Obama is .57, then any actual vote is a product of the .57 underlying probability plus randomness (other causes).

Those are all binary variables. The same logic applies to multi-category and continuous variables. If the mean income in Canada in 2006 was $35,067.69, then any given person in your census dataset has that income plus randomness (which can be positive or negative). Now, the weird thing is that you know you know that each person has their own reasons to have lower or higher incomes than that. But when we are thinking of that variable (income) on its own, and we're really only concerned with its distribution, we think of each case abstractly. Together, all Canadians' incomes are just a distribution with a mean and a standard deviation. When we look at one Canadian in our census dataset, we can think of his/her income as being drawn randomly from the distribution of all Canadians' incomes. The "expected value" of that one person's income is the average income; any deviation from the average is a result of the fact that we randomly drew one Canadian to look at.

Go and look at the income of the first person in your dataset. That income, like all others, is drawn from the distribution of incomes that has a mean of $35,067.69 and a standard deviation of $55,697.02.

This way of thinking is the foundation for all of the ideas in the rest of the course on sampling and expressing how certain we are about our conclusions from samples.

- Averages (of political phenomena) are things (events) added up!

The statistics we're concerned with here are all *aggregations*. That is, we have multiple cases and we report an average or other statistics based on averages.
Recall that each case that goes into the average is a product of an underlying mean and some randomness. If we have a certain number of cases in our average we know that in the long run the randomness in each case will cancel out when we're adding them all up.

We can think of the randomness itself, in the very long run, as having a mean value of zero. Because it's randomness to us, we might as well give it an average value of zero. After all, if it's randomness it has no bias; if it has a bias then it will end up being part of the mean in the first place!

So as we get more and more cases in our data, the random element in our average is going to get closer and closer to zero: we'll "zero-in" on the true value we're trying to estimate.

But more important is the fact that if we know the underlying probabilities or means and standard deviations of a distribution, then probability theory can tell us how much variation we will get in our averages. It simply does this by taking the probability of each occurrence and then adding up these probabilities as many times as we have cases.

The simplest example is with binary outcomes that add up to what is called the "binomial distribution". We will explore this. But you MUST think of this not as a way to learn the binomial distribution but instead to learn the CONCEPTUAL underpinnings of sampling theory. The binomial distribution is just an example. It's the example you encounter in the section of the Drunk book starting on page 95. In class and in the lab we'll do it ourselves.

But you might as well go through a preparation exercise now. Go to this applet: [http://www.rossmanchance.com/applets/CoinTossing/CoinToss.html](http://www.rossmanchance.com/applets/CoinTossing/CoinToss.html) and do the following things:

- Set both the number of tosses and repetitions to 1 and flip the coin a few times. Each time you simply get either a heads or a tails.
  - Do you get half heads and half tails? If you do it twice? Ten times?
- Then set the number of tosses to 10 and keep repetitions at 1. You know immediately that each flip is separate and so we could get any result from 0 to 10 heads. But you don't think 10 heads is as likely as 5, do you? Why is that? (See above).
- Now leave tosses at 10 and set repetitions to 100. This means you have tossed the coin ten times and the proportion of heads you got is recorded on the histogram, just like before, but then the applet does this how-many-heads-out-of-ten-flips thing 100 times.
  - What is the most common outcome? How often do you see zero heads? If the underlying probability is .5, why is 0 or 10 heads so uncommon?
- Finally, leave things as they are but press ‘toss the coin’ a number of times…. Like a lot! And then compare the shape to the normal distribution depicted here - [http://demonstrations.wolfram.com/TheNormalDistribution/](http://demonstrations.wolfram.com/TheNormalDistribution/).

The key thing that should be obvious now is that we're adding up randomness. And adding up randomness has a particular shape -- The Normal Distribution -- that we can describe mathematically in terms of probabilities.
It's worth seeing the individual randomness getting added up here: http://www.math.psu.edu/dlittle/java/probability/plinko/index.html.

OK, you've seen it with the binomial distribution (zeros and ones, added up). Remember, it's the same logic with continuous distributions: Randomness around a systematic mean will be shaped like a bell curve.

- **Randomness Added Up = The Bell Curve (Normal Distribution)**

So when a whole bunch of separate truly random things are added up they always form a bell curve, or "The Normal Distribution". This fact appears in The Drunkard's Walk as the main point of chapter 7. But that chapter is called "Measurement and the Law of Errors". We're not back to measurement, are we? Well, not really; Mlodinow just finds it easier to illustrate the idea of multiple random things added up with the example of measurement error. But there's really no conceptual difference between all the random errors in a measurement, on one hand, and all the causes we don't know about, on the other. When you draw your one Canadian at random and look at his/her income, it's kind of like you're measuring the average income in an unbiased way, but there are lots of things that are knocking that person's income off the average. You could call those "measurement errors" or "unknown influences", it doesn't much matter!

You should read that chapter now and then come back here.

The point is that whatever is in that "error" or "unmeasured causes", there are LOTS of them, so they all add up to pushing your one case (or your set of 10 or 100 or 1000 cases) off of the underlying average income. They probably push a little, but it's very unlikely that they all end up pushing the same way. In fact, talking about one case is pretty silly here, since we know one case could be anywhere in the distribution. It matters a lot more that when we average all these individual outcomes in samples, THEN the errors or unmeasured causes are pushing off the underlying average but they're really unlikely to push all the same way for all of the cases. So the more cases we have the less and less likely we are to have an average very far from the underlying average. We'll learn the formal calculation for this next week.

For now, we need to know that there's some magic to the Normal Distribution that really helps us say how certain or uncertain we are about things we learn from random samples of data. The magic is in the Drunk book on pages 138-141 and in the textbook (also in Chapter 7) in section 7.2. Don't bother with the textbook now, we'll use it next week. You can wait til then to figure out this 68-95-99 rule that's mentioned at the top of page 140 in The Drunkard's Walk. Just make sure you've got that far in that book and that you're starting to see how probability theory is going to help us to know about the patterns of randomness so we can say how much we think we know about the systematic part of things.

- **So we only know about randomness. What about solid facts?**
All of this stuff so far about randomness is important. We know that a whole bunch of different randomness in measurements will add up to a normal distribution. That's called the "Central Limit Theorem" or CLT.

Although the CLT is proven mathematically the irony is that it's only knowledge about randomness. It would tell us for example that IF the real distribution of support for Obama vs. Romney is 57-43 in the whole US population, then any randomness around those numbers will be shaped like a bell curve. That's the same as starting with the assumption that IF a coin is perfectly fair (50/50 heads/tails) then we'll get a normal distribution of the number of heads out of a certain number of flips.

But in social science we don't want to make assumptions about things like this, we want to go out and collect data and use it to learn about the underlying processes (probabilities) that produce the phenomena we observe in the world. All those probability calculations with coins and stuff depend on assuming a certain underlying probability of outcomes. But we want to use real data to estimate bigger facts about the things we care about. Wars and re-election and representation and policy innovation, not coins.

If we care about the incidence of domestic violence in Canadian households, we want to collect some data and find out how bad is the situation. We can collect SOME data and we want to use that limited data to help us estimate how much domestic violence there is in the whole country. Or, we want to use current data on the prevalence of military coups to estimate the likely prevalence of military coups into the future. We don't know the true underlying tendencies of these things, we will only have the information we can collect.

This is put directly in the last paragraph of chapter 5 in the Drunkard's Walk that I had you read before you read all of chapters 3, 4, 5.

"the question of how to infer, from the data [actually] produced, the underlying probability of events -- the answer would not come for several decades more"

For us, it'll just be about five days!