# Probability Review

Luke Sonnet

March 31, 2016

Much of the language and structure is borrowed from course notes for MIT's 17.800 course, developed by Teppei Yamamoto, Jens Hainmueller, and those that they borrowed rom.. This is not meant to be comprehensive but just to get you thinking in these terms. Review your 200A notes and check out the Aronow and Miller 2015 textbook.

## 1 Basic Probability Theory

We have a sample space $\Omega = \{A, B, \dots\}$. The sample space contains all possible events or outcomes of an experiment or data generating process. If we are considering a coin-toss, then $\Omega = \{H, T\}$ where the outcomes can either be heads or tails. We use probability, and specifically probability functions to tell us the probability of each event. Following Chad's notation, we can write $p(A)$ to denote the probability function that tells us the probability of event $A$.

I'm not going to cover basic set theory operations. Review your 200A notes. But given this set-up, we can formally define a probability function $p(\cdot)$ as a function that is defined over all subsets of a sample space $\Omega$ and that satisfies three probability axioms:

1. Nonnegativity - $p(A) \geq 0$ for all $A$ in the set of all events $\Omega$
2. Normalization - $p(\Omega) = 1$
3. Additivity - if $A_1, A_2, \dots$ are disjoint then $p(A_1 \cup A_2 \cup \cdots) = p(A_1) + p(A_2) + \dots$

Now I'm going to discuss random variables, leaving the discussion of conditional and marginal probabilities for that context.

## 2 Random Variables

Dealing with events is difficult when there become many possible events. For coin tosses, events can be easy, for today's temperature, the events may be infinite since temperature is continuous.

A random variable $X$ is a real-valued function that maps the sample space to real numbers, $X : \Omega \to \mathbb{R}$. In other words, a random variable is a function that returns a real number to each of the outcomes of the experiment, or events in the sample space.

Imagine a coin toss:

$$X = \begin{cases} 0 & \text{if we get tails} \\ 1 & \text{if we get heads} \end{cases}$$

The probability mass function (or probability discrete function for continuous random variables), assigns probability for each realized value $x \in X$. It is often written as $f_X(x)$; Chad uses $p(x)$. It will follow the probability axioms.

If our coin is fair, then

$$p(x) = \begin{cases} 1/2 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \end{cases}$$

This can be simplified as

$$p(x) = 0.5 \;\; \forall \, x \in X$$

A continuous random variable has a continuous range of possible realizations, or outcomes. If $X$ is a continuous random variable, then it is characterized by a probability density function that Chad would write as, again $p(x)$.

$$\Pr(a \leq X \leq b) = \int_a^b p(x)\, dx$$

For a uniform distribution that ranges from 0 to 4, we can write it as $X \sim Uniform(0, 4)$. The the probability density function of $X$ is

$$p(x) = \begin{cases} 1/4 & 0 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

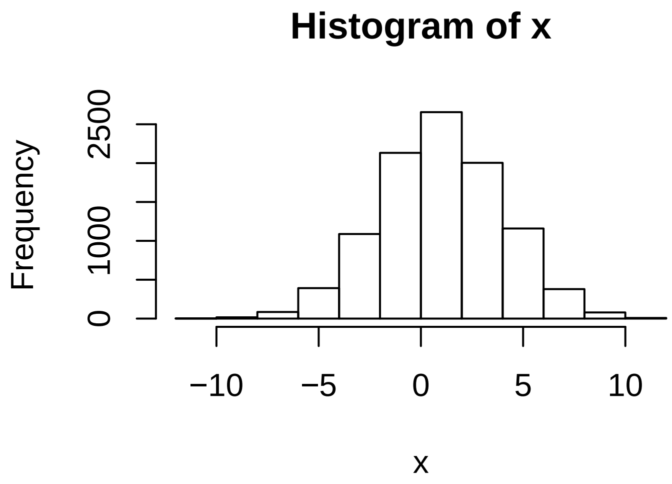
We commonly work with the normal distribution, which can be described by its mean and variance. We often write $X \sim N(\mu, \sigma^2)$. The PDF of $X$ is

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

So this function is fully described by $\mu$ and $\sigma$. Let $X \sim N(1, 9)$, then our PDF is

$$p(x) = \frac{1}{3\sqrt{2\pi}} e^{\frac{(x-1)^2}{18}}$$

If we sample $10,000$ times from this distribution, we can plot the data using the following distribution:



## 3 Expectation

One way to describe a random variable is by its expectation, which can also be considered as it's center of gravity. For a discrete random variable $X$, it is the sum of all of the possible realizations of $X$, $x \in X$ multiplied by the probability of those realizations.

$$\mathbb{E}[X] = \sum_{x \in X} xp(x)$$

If we want to consider what happens when we take the expectation of $X$ times some constant $a$, we just have to consider how it changes the right hand side of that equation. Because $a$ is constant, the probabilities will not change, simply the realizations $x \in X$ will change.

$$\begin{aligned}\mathbb{E}[aX] &= \sum_{x \in X} axp(x)\\ &= a\sum_{x \in X} xp(x)\\ &= a\mathbb{E}[X]\end{aligned}$$

Similarly, if $b$ is a constant as well,

$$\begin{aligned}\mathbb{E}[aX + b] &= \sum_{x \in X} (ax + b)p(x)\\ &= \sum_{x \in X} axp(x) + bp(x)\\ &= \sum_{x \in X} axp(x) + \sum_{x \in X} bp(x)\\ &= a\sum_{x \in X} xp(x) + b\sum_{x \in X} p(x)\\ &= a\mathbb{E}[X] + b\end{aligned}$$

Here we use the fact that $\sum_{x \in X} p(x) = 1$, meaning the sum of the probability of all possible realizations of $X$ equals 1. This all is very similar to continuous distribution, although instead of summing over possible realizations of $X$, we integrate over the range of possible $x$ values.

## 4 Joint Distributions

Two random variables $X$ and $Y$ have the joint probability mass function $p(x, y)$. This follows the same rules as a regular distribution function, in that the sum of all possible outcomes results must be 1 and no values are negative. In other words $\sum_{y \in Y} \sum_{x \in X} p(x, y) = 1$.

Sometimes we want to obtain $p(y)$ from this joint distribution. This is called marginalizing probability function.

$$p(y) = \sum_{x \in X} p(x, y) = \sum_{x \in X} Pr(X = x, Y = y)$$

Looks similar for continuous random variables. They must sum to 1 and can be marginalized. If $X, Y \in \mathbb{R}$, then $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y)\, dy\, dx = 1$. Marginalization looks like

$$p(y) = \int_{-\infty}^{\infty} p(x, y)\, dx$$

## 5 Conditional Probability

Working just with continuous distributions, let's think about the conditional PDF of $Y$ given $X$. We write it as $p(y|x)$. Using the multiplicative rule of probability, we can rewrite this as

$$p(y|x) = \frac{p(x, y)}{p(x)}$$

We can also summarize this conditional distribution with the expectation function, but I leave that for a future problem set.

## 6 Independence

Note that if two distributions are independent, then we can write $p(x, y) = p(x)p(y)$ for all $x$ and $y$. This implies that

$$\begin{aligned} p(y|x) &= \frac{p(x, y)}{p(x)} \\ &= \frac{p(x)p(y)}{p(x)} \\ &= p(y) \end{aligned}$$

Furthermore, this has implications for $E[XY]$. Imagine $X$ and $Y$ are continuous random variables.

$$\begin{aligned} \mathbb{E}[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyp(x, y)\, dy\, dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyp(x)p(y)\, dy\, dx \\ &= \int_{-\infty}^{\infty} xp(x)\, dx \int_{-\infty}^{\infty} yp(y)\, dy \\ &= \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

When $X$ and $Y$ are independent conditional on $Z$, we call this conditional independence and we can write it as $p(x, y|z) = p(x|z)p(y|z)$. This intuitively just means that for some realization of $Z$, $X$ and $Y$ are unrelated.