Summary of Lecture 3

Reading:

- Sections 2.1-2.4, 3.3 and 3.4 of OpenIntro Statistics
- Strogatz article on Bayes rule

The main concept we introduced in Lecture 3 was that of a discrete random variable (rv). We use rv as a model for data collected in the world or to talk about future outcomes. For example, we could use rv to describe data collected from a manufacturing process or we could talk about the probability that a new mortgage in a certain zip-code defaults.

A **random variable** (our model) is specified by two things:

1. A list of possible outcomes (things that can happen)
2. Probability for each outcome

Only one outcome can occur. The probability is the fraction of times we see an outcome in the long run. Sum of probabilities for all outcomes must be 1.

A Bernoulli random variable is a dummy variable which specifies probability for outcome "0" and outcome "1." These outcomes can denote different things in the real world (for example, sex of a customer, whether market goes up or down tomorrow, whether a tossed coin lands heads or tails). We write $X \sim \text{Bernoulli}(p)$ to denote a Bernoulli rv with probability of the outcome "1" equal to p.

Discrete Uniform distribution is used to model a finite number of values that occur with the same probability. That is, if we have N outcomes, then probability of each one of the outcome is $/1N$. For example, we can model rolling a 6-sided die with a discrete Uniform distribution. Before rolling a die, we do not know what the outcome of a roll will be. However, we know that if a die is fair that it is equally probable to get any number between 1 and 6, hence, we can use a discrete Uniform distribution as a model.

To compute probability of a subset of outcomes (that cannot happen simultaneously), we sum probabilities of each outcome in the subset. For example,

$$P(X = a \text{ OR } X = b) = P(X = a) + P(X = b)$$

or

$$P(a < X < b) = \sum_{a<x<b} P(X = x).$$

Probability that we do not see an outcome a, can be computed as one minus probability that we do see the outcome a, that is, $P(X \neq a) = 1 - P(X = a)$.

Expectation of a discrete random variable represents a typical value under our model. The **expected value** of X is computed as

$$E(X) = \sum_{j=1}^{m} P(X = x_j) \cdot x_j.$$

The expected value of a function $g(X)$ is defined as:

$$E\left(g(X)\right) = \sum_{j=1}^{m} g(x_j) \cdot P(X = x_j).$$

The variance of a discrete random variable X captures how outcomes vary around the expected value of X. We compute the variance as

$$\mathrm{Var}(X) = \sum_{j=1}^{m} P(X = x_j) \cdot [x_j - E(X)]^2 = E(X^2) - (E(X))^2.$$

Variance is the expected squared distance of the random variable X from its mean.

A more intuitive way to understand the spread of a distribution is to look at the **standard deviation**:

$$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$$

If $X \sim \mathrm{Bernoulli}(p)$, then $E(X) = p$ and $\mathrm{Var}(X) = p \cdot (1 - p)$.

The **joint model** specifies relationships between outcomes of two (or more) random variables. We can specify the joint model in the same way as before. We list all possible outcomes and their probabilities. In the case of two random variables, specifying the joint model is not too hard. However, when we need to specify a joint model between many random variables, it becomes difficult to list all possible outcomes. However, we will be able to use simple building blocks learned in this lecture to specify more complex models later on.

We defined the joint model via simpler building blocks: **marginal distribution** and **conditional distribution**. In our class example we had a marginal model for whether economy will go up or down the following quarter. This model was given by a marginal distribution ($E \sim \mathrm{Bernoulli}(p)$). We also had two models for next quarter sales. One model captured uncertainty in sales when economy was up, while the other captured uncertainty in sales when economy was down. These two models were given by conditional distributions ($P(S \mid E = 1)$ and $P(S \mid E = 0)$). Given the marginal distribution and conditional distributions, we can obtain the joint distribution as $p(s, e) = p(s \mid e) \cdot p(e)$.

Given a joint distribution of X and Y, we can obtain marginal distributions and conditional distributions. The marginal distribution of X is obtained from the joint distribution and it tells us what we know about X, while ignoring any information about Y. The conditional distribution of Y given $X = x$ tells us what we think about Y given that we observed a particular realization of X.

Two important things that you need to know:

- Given a joint distribution, compute the marginal and conditional distributions;

$$\begin{aligned} P(X = x) &= \sum_y P(X = x, Y = y) \\ P(Y = y) &= \sum_x P(X = x, Y = y) \\ P(Y = y \mid X = x) &= \frac{P(Y = y, X = x)}{P(X = x)} \end{aligned}$$

- Given marginal and conditional distributions, compute the joint distribution.

$$\begin{aligned} P(Y = y, X = x) &= P(X = x) \cdot P(Y = y \mid X = x) \\ &= P(Y = y) \cdot P(X = x \mid Y = y) \end{aligned}$$

Bayes theorem allows us to compute conditional probability $p(x \mid y)$ given the marginal distribution $P(X = x)$ and the conditional distribution $p(y \mid x)$. In particular, we have

$$P(X = x \mid Y = y) = \frac{P(Y = y, X = x)}{P(Y = y)} = \frac{P(Y = y, X = x)}{\sum_x P(Y = y, X = x)} = \frac{P(X = x) P(Y = y \mid X = x)}{\sum_x P(X = x) P(Y = y \mid X = x)}.$$

For any sequence of n random variables $Y_1, \ldots, Y_n$ we can write the joint model in terms of the marginal distribution and conditional distributions as

$$p(y_1, y_2, \ldots, y_n) = p(y_n \mid y_{n-1}, y_{n-2}, \ldots, y_2, y_1) \cdot \ldots \cdot p(y_3 \mid y_2, y_1) \cdot p(y_2 \mid y_1) \cdot p(y_1)$$

Even though we can use the above representation to specify any joint model, it becomes difficult to specify all the conditional models in the above equation. However, by assuming simplifying relationships between r.v.s we only need to specify a marginal model or a joint model between a pair of r.v.s. We saw how i.i.d model and Markov model assumed different things about the world in order to simplify the joint model. These assumptions are usually only approximately correct, however, they are still extremely useful.

Two random variables are independent if knowledge about one of them does not change what we know about the other one. Given a joint distribution, you will need to be able to check if the two r.v.s are dependent or independent. There are three ways to do so. To show that they are dependent, you only need to find an outcome (x,y) for which $P(X=x,Y=y)4 is not equal to $P(X = x) \cdot P(Y = y)$. To show that they are independent, you need to do more work. One way is to show that for all outcomes (x,y) it holds that $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$. Another way to show independence is to demonstrate that for all (x,y) it holds that $P(Y = y \mid X = x) = Pr(Y = y)$.

We often want to summarize a sample of n data points using random variables. We would like to be able to talk about probability that stock market goes down 5 days in a row, for example. As mentioned above, it is difficult to make a joint model for 5 random variables. However, we can assume that these 5 random variables follow an **i.i.d. model**. In that case, we only need to specify the marginal distribution of one random variable. Computing the joint probability for (say) 5 i.i.d random variables can be done using

$$\begin{aligned} P(X_1 = x_1, & X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5) \\ & = P(X_1 = x_1) \cdot P(X_2 = x_2) \cdot P(X_3 = x_3) \cdot P(X_4 = x_4) \cdot P(X_5 = x_5), \end{aligned}$$

since all the random variables are independent of each other (the first i. in i.i.d.). Furthermore, we have that random variables are identically distributed (the i.d. in i.i.d.), so we can obtain $P(X_i = x_i)$ from the marginal distribution.

One way to specify a r.v. is through mathematical formulas. Let $X_1, X_2, \ldots, X_n$ denote n i.i.d. Bernoulli(p) random variables.
The **binomial distribution** is the probability distribution for the total number of successes:

$$Y = X_1 + X_2 + \ldots X_n = \sum_{i=1}^{n} X_i.$$

We write $Y \sim \text{binomial}(n, p)$. This distribution can be used to model situations when we try something n times, each time we succeed with probability p, and all the trials are independent of each other. The random variable Y models the number of successes in these n trials. We have that

$$\begin{aligned} E[Y] &= np \\ \text{Var}[Y] &= np(1-p) \\ P(Y = y) &= \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \text{ for } y = 0, 1, 2, \ldots, n. \end{aligned}$$

Sometimes an i.i.d. model is not appropriate. For example, the stock price $P_t4 on day t, is not independent of the price on day before $t - 1$. However, we can often assume that differences in prices $E_t = P_t - P_{t-1}$ are approximately i.i.d. for $t = 1, 2, \ldots$. In this case, we are assuming that the price P_t depends only on P_{t-1} and not on the whole history of prices $P_0, P_1, \ldots, P_{t-2}, P_{t-1}$. This assumption is called the Marko assumption or the Markov model. The Markov model is commonly used for modeling sequences. The Markov model assumes that the probability of X_t depends only on X_{t-1}. Under the Markov model, we need to specify the marginal distribution for the first r.v. and the conditional distribution of X_t given X_{t-1}. It is assumed that this conditional distribution does not change over time. That is, the conditional distribution $P(P_t = p_t \mid P_{t-1} = p_{t-1})$ does not change with t. With these assumptions, we can compute the probability of a sequence of (say) 5 observations as

$$\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4, X_5 = x_5) \\
&\quad = P(X_1 = x_1) \cdot P(X_2 = x_2 \mid X_1 = x_1) \cdot P(X_3 = x_3 \mid X_2 = x_2) \cdot P(X_4 = x_4 \mid X_3 = x_3) \cdot P(X_5 = x_5 \mid X_4 = x_4).
\end{aligned}$$