Dis 5: Random Variables; Discrete Probability Distributions

Relevant textbook chapter: 7
Ch 7 handout and solution offered by Dr. Pac can be accessed here: Handout Solution
This handout incorporates reviews with all exercises from Dr. Pac's original handout.

1 Motivation

- Two weeks ago, we talked about probability, which formally discusses the likelihood of an event
- Relating this back to data:
  - Say that you have population data, and you'd like to present all the data points in some way
  - One way is to list off every single observation from the population (like an Excel spreadsheet; we will see how this is similarly done in Stata in the Stata Q&A session)
  - Now, with probability, an alternative way is to list off the unique values from the population, and then use probability to describe the frequency – or the likelihood – of hitting a specific value
    - This alternative way of describing data gives us a **distribution**
    - Knowing the distribution of the population data gives us information on how likely certain values will be included in a sample
      $\Rightarrow$ helps with our long-term goal of making inference
- Between this and the next discussion, we will examine two types of distribution: **discrete** and **continuous**
- But before getting there, we'll start off talking about **random variables**, which will simplify our discussion on distributions

2 Random Variables

- A random variable **assigns a number to each outcome** of an experiment.
  e.g. Let X be a random variable recording the outcome of rolling a fair, six-sided die, then

$$\begin{aligned} \text{Rolling a 1} &\quad\rightarrow\quad X = 1 \\ \text{Rolling a 2} &\quad\rightarrow\quad X = 2 \\ &\cdots \\ \text{Rolling a 6} &\quad\rightarrow\quad X = 6 \end{aligned}$$

- Notice that each outcome has been assigned a number:
  - When all possible outcomes are listed, this is equivalent to the **sample space** from our probability discussion
  - When all possible outcomes are listed, these describe the set of possible values that the **population** data can take on

- Now, if someone describes the associated probabilities at each of the values of a random variable (using a table / formula / graph / something else), then we have a **probability distribution**.

  e.g. Continue with the example of rolling a fair, six-sided die. The following describes a probability distribution:

  $$P(X = x) = \frac{1}{6} \quad \text{where } x \in \{1, 2, 3, 4, 5, 6\}$$

  – With a probability distribution, we have the probability described at any possible value in the **population** data. This means that we can calculate **parameters** using a probability distribution:

Parameter Name	**Notation**	**Formula**	**Shortcut**
Expected value (mean)	$E(X) = \mu = \mu_X$	$\sum_x xP(x)$	-
Variance	$V(X) = \sigma^2 = \sigma_X^2$	$\sum_x (x - \mu)^2 P(x)$	$E(X^2)$ $-[E(X)]^2$
Standard deviation	$\sigma = \sigma_X$	$\sqrt{V(X)}$	-

  – When there is more than one probability distribution (say, distributions for random variables X and Y), then **parameters** on the relationship between two random variables can be expressed:

Parameter Name	**Notation**	**Formula**	**Shortcut**
Covariance	$Cov(X, Y) = \sigma_{XY}$	$\sum_x \sum_y (x - \mu_X)(y - \mu_Y) P(x, y)$	$E(XY)$ $-E(X)E(Y)$
Correlation (of coefficient)	$Corr(X, Y) = \rho_{XY}$	$\frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$	-

  – Common parameter operations, when the random variable is transformed in some way:
  (Let X and Y be random variables, a, b, c, d be constants.)

    * Expected value (mean):
      · $E(c) = c$
      · $E(aX + b) = aE(X) + b$
      · $E(X + Y) = E(X) + E(Y)$
    * Variance:
      · $V(c) = 0$
      · $V(aX + b) = a^2 V(X)$
      · $V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X, Y)$
    * Covariance:
      · $Cov(a, b) = 0$
      · $Cov(X, X) = V(X)$
      · $Cov(aX + b, cY + d) = acCov(X, Y)$

- There are two types of probability distributions. Each associated with a specific type of random variable:

  1. **Discrete probability distribution**:

Generated by a **discrete random variable**, which means that numbers assigned to the random variable are countable.

2. **Continuous probability distribution**:

   Generated by a **continuous random variable**, which means that numbers assigned to the random variable are NOT countable.

Aside: what does it mean to be countable?

(A pretty loose definition:) as long as you can sequentially count all the numbers assigned – even though it might take forever – then such series of numbers is considered as countable.

Exercise. Is the following random variable discrete or continuous?

1. $X =$ whether the result from a fair coin flip is head or not

   X is a discrete random variable. When the outcome is head, $X = 1$; when the outcome is not head (i.e. tail), $X = 0$. These two numbers are certainly countable.

2. $X =$ amount of time it takes for a student to complete a 60-minute exam

   X is a continuous random variable. The outcome is assigned to anything between 0 minute and 60 minutes, but if someone tries to count all possible time between 0 and 60 minutes, they won't be able to.

   More explicitly, say that someone wants to count from 0 to 60 minutes with 0.01 minute increment. However, there are time between 0.00 minute and 0.01 minute (such as 0.003 minute) that will be left uncounted. If this person then lower the increment to 0.001, the same argument would apply since some even smaller time would be left uncounted.

   The point being that, regardless of how small the time increment is made for counting, there always exists some even smaller time increment. Thus, one can never sequentially count all potential outcomes, which is why X is considered as a continuous random variable.

3. $X =$ the number of rolls it takes to get a 6 from rolling a six-sided die

   X is a discrete random variable. The outcome is $1, 2, 3, 4, \ldots$. While it might be the case that it will take you infinite number of rolls to get a 6, as long as you keep counting, then you can sequentially count all outcomes. Thus, numbers assigned to X are countable, which means X is discrete.

   Notice the difference between this and the second exercise. In this exercise, the increment between all potential outcomes is always 1, so as long as you keep counting, all outcomes could be counted. In the second exercise, even if you have infinite amount of time, the time increment used for counting can always be made smaller, meaning that there is just no way for you to sequentially count all the time between 0 and 60 minutes to begin with.

3 Discrete Probability Distributions

- A discrete probability distribution needs to describe the probability $P(x)$ at all possible x values that a discrete random variable X can take on.
- What are some conditions that a discrete probability distribution needs to satisfy?

1. $0 \leq P(x) \leq 1$ for all x
2. $\sum_x P(x) = 1$

Aside: What if, instead of looking among all possible outcomes, we want to narrow our focus to only certain outcomes for discrete random variable X?

For example, say that we already know for random variable Y, it has taken on the value y, meaning that $Y = y$ is the space that we want to restrict ourselves to. How should we describe the probability distribution and the conditions that such distribution must satisfy?

Solution: use **discrete conditional probability distribution**

Conditions that a discrete conditional probability distribution needs to satisfy:

1. $0 \leq P(x|y) \leq 1$ for all x
2. $\sum_x P(x|y) = 1$

- Two examples of discrete probability distribution:
  1. **Binomial distribution**: distribution of success among n trials
     - Random variable $X \sim \text{Binomial}(n, p)$ if the following holds:
       * There are fixed number (n) of trials.
       * Each trial has two outcomes: success, and failure.
       * $P(\text{success}) = p$ is constant across all trials.
       * Trials are independent.
     - Once $X \sim \text{Binomial}(n, p)$ is established, then
       * $P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$
       * $E(X) = np$
       * $V(X) = np(1-p)$
  2. **Poisson distribution**: distribution of success within a fixed time period / fixed interval, with success arriving at rate $\mu > 0$
     - Random variable $X \sim \text{Poisson}(\mu)$ if the following holds:
       * Number of success in any interval is independent of the number of success in any other interval.
       * Probability of a success in any equal-size interval is constant.
       * Probability of a success is proportional to the size of the interval.
       * Probability of more than one success in an interval approaches 0 as the interval becomes smaller.
     - Once $X \sim \text{Poisson}(\mu)$ is established, then
       * $P(x) = \frac{e^{-\mu} \mu^x}{x!}$ for $x = 0, 1, 2, \ldots$
       * $E(X) = \mu$
       * $V(X) = \mu$

4 Exercises

1. Suppose you are offered an opportunity to play a game where you are paid based on the outcome of a series of flips of a (fair) coin. You win two dollars if the first head comes up on toss one, four dollars if the first head comes up on toss two, eight dollars if the first head comes up on toss three, sixteen dollars if the first head comes up on toss four, and so on.

   (a) Without doing any math, how much would you be willing to pay to play this game?
   This answer is completely subjective, but most people are willing to pay less than $5.

   (b) Define the payoff to this game as a random variable X. Is this random variable discrete or continuous? Double check that the probability distribution for this random variable is well-defined.
   The probability distribution for the random payoff variable is:

x	$P(x)$
$2^1 = 2$	$\left(\frac{1}{2}\right)^1 = \frac{1}{2}$
$2^2 = 4$	$\left(\frac{1}{2}\right)^2 = \frac{1}{4}$
$2^3 = 8$	$\left(\frac{1}{2}\right)^3 = \frac{1}{8}$
$2^4 = 16$	$\left(\frac{1}{2}\right)^4 = \frac{1}{16}$
$\vdots$	$\vdots$

   This is a discrete random variable, since it takes on a countable number of possible outcomes.
   The probability distribution is well-defined, since

   i. $P(X = x) = \left(\frac{1}{2}\right)^x$ is between 0 and 1, and
   ii. The probabilities sum to one

   To see that the probabilities sum to one, recall that a geometric series looks like the following:

$$r, r^2, r^3, \ldots \quad \text{where } |r| < 1$$

   Now, the sum of a geometric series takes the following form:

$$r + r^2 + r^3 + \ldots = \frac{r}{1-r}$$

   In this case, we have a geometric series with $r = \frac{1}{2}$, which means the sum of all probabilities is

$$\left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \ldots = \frac{\frac{1}{2}}{1-\frac{1}{2}} = 1$$

   (c) What is the expected value of the payoff from this game? Given this, if you've taken Intermediate Micro, do you recall how much a risk neutral individual should be willing to pay?
   The expected value of the payoff from this game is

$$E(X) = \sum_x xP(x) = 2 \times \frac{1}{2} + 4 \times \frac{1}{4} + 8 \times \frac{1}{8} + 16 \times \frac{1}{16} + \ldots = \infty$$

Recall from Intermediate Micro: a risk neutral individual is willing to pay the expected value of the payoff. Thus, with the payoff being infinite, he or she should be willing to play the game no matter what the cost.

Now, recall your answer from part (a). While the game's expected payoff is infinite, most people are only willing to pay a very little amount of money to play the game. This phenomenon is known as the St. Petersburg Paradox.

2. The table below contains the joint distribution of the random variables X and Y representing the percentage return for Xenon Incorporated and Yellow Company:

		Y	
		0.0	0.5
X	0.0	0.4	0.1
	0.5	0.1	0.4

(a) Find the population mean, variance, and standard deviation of X and Y.

We can calculate the marginal probabilities for X using the joint probability table:

$$P(X = 0.0) = P(X = 0.0, Y = 0.0) + P(X = 0.0, Y = 0.5) = 0.4 + 0.1 = 0.5$$
$$P(X = 0.5) = P(X = 0.5, Y = 0.0) + P(X = 0.5, Y = 0.5) = 0.1 + 0.4 = 0.5$$

Then we use these to calculate the population mean:

$$\begin{aligned} E(X) &= \sum_x xP(x) \\ &= 0.0 \times P(X = 0.0) + 0.5 \times P(X = 0.5) \\ &= 0.0 \times 0.5 + 0.5 \times 0.5 = 0.25 \end{aligned}$$

Similarly, we can find $E(X^2)$:

$$\begin{aligned} E(X^2) &= \sum_x x^2 P(x) \\ &= 0.0^2 \times P(X = 0.0) + 0.5^2 \times P(X = 0.5) \\ &= 0.0^2 \times 0.5 + 0.5^2 \times 0.5 = 0.125 \end{aligned}$$

By the shortcut method, the variance of X is

$$V(X) = E(X^2) - [E(X)]^2 = 0.125 - (0.25)^2 = 0.0625$$

Taking the squared root of $V(X)$ gives us the standard deviation of X as $\sigma_X = \sqrt{V(X)} = \sqrt{0.0625} = 0.25$

Since the marginal distribution of Y is the same as for X, it has the same mean, variance, and standard deviation as X.

(b) Find the population covariance and correlation coefficient between X and Y

To calculate the covariance via the shortcut method, we first need $E(XY)$:

$$\begin{aligned}
E(XY) &= \sum_x \sum_y xyP(x,y) \\
&= 0.0 \times 0.0 \times P(0.0, 0.0) + 0.0 \times 0.5 \times P(0.0, 0.5) \\
&\quad + 0.5 \times 0.0 \times P(0.5, 0.0) + 0.5 \times 0.5 \times P(0.5, 0.5) \\
&= 0.0 \times 0.0 \times 0.4 + 0.0 \times 0.5 \times 0.1 \\
&\quad + 0.5 \times 0.0 \times 0.1 + 0.5 \times 0.5 \times 0.4 \\
&= 0.1
\end{aligned}$$

By the shortcut method, the covariance of X and Y is

$$\sigma_{XY} = Cov(X,Y) = E(XY) - E(X)E(Y) = 0.1 - 0.25 \times 0.25 = 0.0375$$

From (a), we know that $\sigma_X = \sigma_Y = 0.25$. Thus, the correlation of coefficient is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0.0375}{0.25 \times 0.25} = 0.6$$

(c) Let $Z = \frac{1}{2}X + \frac{1}{2}Y$ represent the return on a 50/50 mix of the two assets. What are $E(Z)$ and $V(Z)$?

Using the properties of expectation:

$$\begin{aligned}
E(Z) &= E\left(\frac{1}{2}X + \frac{1}{2}Y\right) \\
&= \frac{1}{2}E(X) + \frac{1}{2}E(Y) = \frac{1}{2} \times 0.25 + \frac{1}{2} \times 0.25 = 0.25
\end{aligned}$$

And by properties of variance:

$$\begin{aligned}
V(Z) &= V\left(\frac{1}{2}X + \frac{1}{2}Y\right) \\
&= V\left(\frac{1}{2}X\right) + V\left(\frac{1}{2}Y\right) + 2Cov\left(\frac{1}{2}X, \frac{1}{2}Y\right) \\
&= \frac{1}{4}V(X) + \frac{1}{4}V(Y) + 2 \times \frac{1}{2} \times \frac{1}{2}Cov(X,Y) \\
&= \frac{1}{4} \times 0.0625 + \frac{1}{4} \times 0.0625 + 2 \times \frac{1}{2} \times \frac{1}{2} \times 0.0375 = 0.05
\end{aligned}$$

(d) Suppose a shift in the market changes the return for Xenon Incorporated, the new return is $X^* = 3X + 0.10$. What is the mean and variance of Xenon Incorporated's return after this market shift?

By the properties of expectation:

$$E(X^*) = E(3X + 0.10) = 3E(X) + 0.10 = 3 \times 0.25 + 0.10 = 0.85$$

And by the properties of variance:

$$V(X^*) = V(3X + 0.10) = 9V(X) = 9 \times 0.0625 = 0.5625$$

3. The table below contains the joint distribution of the random variables X_1 and X_2, which represent pass/fail grades on two quizzes:

		X_2	
		0	1
X_1	0	0.30	0.40
	1	0.10	0.20

(a) What is $E(X_1)$?
First, we calculate the marginal probabilities of X_1 using the joint probability table:

$$P(X_1 = 0) = P(X_1 = 0, X_2 = 0) + P(X_1 = 0, X_2 = 1) = 0.30 + 0.40 = 0.7$$
$$P(X_1 = 1) = P(X_1 = 1, X_2 = 0) + P(X_1 = 1, X_2 = 1) = 0.10 + 0.20 = 0.3$$

Then we use these to calculate the expected value:

$$\begin{aligned} E(X_1) &= \sum_{x_1} x_1 P(x_1) = 0 \times P(X_1 = 0) + 1 \times P(X_1 = 1) \\ &= 0 \times 0.7 + 1 \times 0.3 = 0.3 \end{aligned}$$

(b) What is $E(X_1|X_2 = 1)$?
To calculate the conditional expectation, we next need to know the probability distribution conditional on $X_2 = 1$. We get this by dividing the joint probabilities in the second column of the table by the marginal probability of $X_2 = 1$:

$$\begin{aligned} P(X_1 = 0|X_2 = 1) &= \frac{P(X_1 = 0, X_2 = 1)}{P(X_2 = 1)} = \frac{P(X_1 = 0, X_2 = 1)}{P(X_1 = 0, X_2 = 1) + P(X_1 = 1, X_2 = 1)} \\ &= \frac{0.4}{0.4 + 0.2} = \frac{2}{3} \\ P(X_1 = 1|X_2 = 1) &= \frac{P(X_1 = 1, X_2 = 1)}{P(X_2 = 1)} = \frac{P(X_1 = 1, X_2 = 1)}{P(X_1 = 0, X_2 = 1) + P(X_1 = 1, X_2 = 1)} \\ &= \frac{0.2}{0.4 + 0.2} = \frac{1}{3} \end{aligned}$$

We use these to calculate the conditional expectation:

$$\begin{aligned} E(X_1|X_2 = 1) &= \sum_{x_1} x_1 P(x_1|X_2 = 1) \\ &= 0 \times P(X_1 = 0|X_2 = 1) + 1 \times P(X_1 = 1|X_2 = 1) \\ &= 0 \times \frac{2}{3} + 1 \times \frac{1}{3} = \frac{1}{3} \end{aligned}$$

4. To help market their latest blockbuster sports franchise, Croquet 2015, EA Sports decides to distribute a free demo of their video game. Suppose each customer who plays the demo buys the full game with probability 0.8. For this problem, you may find it helpful to reference the probability tables on the last page.

   (a) Let X_n be the total number of sales after n customers have played the demo. What distribution does X_n have? Does it seem likely that the conditions of this distribution are satisfied? (Regardless of your answer, for the rest of the problem you may assume the conditions are satisfied.)

   Because it's the number of successes in n trials, this has the basic setup of a binomial random variable. But in order for the binomial distribution to yield valid answers, we must check whether the conditions for this distribution are satisfied.

   - First, are there fixed amount of trials?
     Yes, the number of trials $= n$ always.
   - Second, does each trial results in a success or a failure?
     Yes, in this case a success occurs when a demo-playing customer purchases the full game.
   - Third, is the probability of success constant across all trials?
     Yes, in this case the probability of success is $p = 0.8$, and this probability remains the same for all trials.
   - Finally, is each trial's outcome independent of the outcome of every other trial?
     This one is debatable. Notice that while probability of success stays at 0.8, it doesn't mean that trials need to be independent. Say that two customers before you played the demo, one has purchased the game, and the other didn't. If your decision on purchasing the game is influenced by both consumers, and their degree of impact on you is the same, then you'll still end up purchasing the game with a probability of 0.8, but your decision is no longer independent of other trials.

   (b) What are the expected value and variance of X_1?

   One could solve this problem exactly as we've been doing up until now, using the definition of expected value and variance:

$$E(X_1) = 0 \times 0.2 + 1 \times 0.8 = 0.8$$
$$V(X_1) = (0 - 0.8)^2 \times 0.2 + (1 - 0.8)^2 \times 0.8 = 0.16$$

   But it's much easier to make use of the fact that this is a binomial distribution with $n = 1$ and $p = 0.8$:

$$E(X_1) = np = 1 \times 0.8 = 0.8$$
$$V(X_1) = np(1 - p) = 1 \times 0.8 \times (1 - .8) = 0.16$$

   (c) What is the expected value and variance of X_5?

   Noting that it's a binomial random variable with $n = 5$ and $p = 0.8$, we have:

$$E(X_5) = np = 5 \times 0.8 = 4$$
$$V(X_5) = np(1 - p) = 5 \times 0.8 \times 0.2 = 0.8$$

(d) Assuming $n = 5$, what is the probability at least 1 customer buys the full game? What is the probability exactly 5 customers buy the full game?

For the probability at least 1 customer buys the full game, the easiest way to approach this question is to begin with the complement rule, and notice that the only outcome that satisfies $X_5 \leq 0$ is when $X_5 = 0$:

$$P(X_5 > 0) = 1 - P(X_5 \leq 0)$$

Now, if you look up the value of $P(X_5 \leq 0)$ from the binomial table, you'll find that $P(X_5 \leq 0) = 0.0003$. So we have

$$P(X_5 > 0) = 1 - 0.0003 = 0.9997$$

Alternatively, recall that the binomial probability distribution for a binomial random variable is:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n$$

this means that

$$P(X_5 \leq 0) = P(X_5 = 0) = \frac{5!}{0!(5-0)!}(0.8)^0(1-0.8)^{5-0} = 0.00032$$

So we obtain

$$P(X_5 > 0) = 1 - 0.00032 = 0.99968$$

For the probability exactly five customers buy the full game, we can get this directly using the binomial formula:

$$P(X_5 = 5) = \frac{5!}{5!(5-5)!}(0.8)^5(1-0.8)^{5-5} = 0.32768$$

Or, if you prefer to use the binomial table, we get:

$$P(X_5 = 5) = 1 - P(X_5 \leq 4) = 1 - 0.6723 = 0.3277$$

(You might notice that the probability you get from using the binomial formula is a little bit different from the one from using the binomial table. This is mainly the result of rounding error, so it's okay.)

(e) Assuming $n = 5$, what is the probability that the number of customers who buy the full game lies between 1 and 4 (including both 1 and 4)?

We've already done most of the work, since by the complement rule:

$$P(1 \leq X_5 \leq 4) = P(X_5 > 0) - P(X_5 = 5) = 0.672$$

Alternatively, notice that

$$P(1 \leq X_5 \leq 4) = P(X_5 \leq 4) - P(X_5 \leq 0)$$

Using the binomial table, we find that $P(X_5 \leq 4) = 0.6723$, and that $P(X_5 \leq 0) = 0.0003$, so this yields the same result.

TABLE 1 Binomial Probabilities

Tabulated values are $P(X \leq k) = \sum_{x=0}^{k} p(x_i)$. (Values are rounded to four decimal places.)

***n* = 5**

	p														
k	**0.01**	**0.05**	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.75**	**0.80**	**0.90**	**0.95**	**0.99**
0	0.9510	0.7738	0.5905	0.3277	0.2373	0.1681	0.0778	0.0313	0.0102	0.0024	0.0010	$\boxed{0.0003}$	0.0000	0.0000	0.0000
1	0.9990	0.9774	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0156	0.0067	0.0005	0.0000	0.0000
2	1.0000	0.9988	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.1035	0.0579	0.0086	0.0012	0.0000
3	1.0000	1.0000	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.3672	0.2627	0.0815	0.0226	0.0010
4	1.0000	1.0000	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.7627	$\boxed{0.6723}$	0.4095	0.2262	0.0490