Dis 5: Random Variables; Discrete Probability Distributions

Related 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 the handout given by Dr. Pac.

1 Motivation

- Two weeks ago, we talked about probability, which formally discusses the likelihood of an event
- Tying this back to data:
  - Say that you have access to the population data
  - One way for us to present the population data is to list off every single observation from the population (like what you saw last week from running `browse` on a dataset in Stata)
  - 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 in a bit

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

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

- 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):
      - $E(c) =$
      - $E(aX + b) =$
      - $E(X + Y) =$
    - Variance:
      - $V(c) =$
      - $V(aX + b) =$
      - $V(X \pm Y) =$

* Covariance:
  * $Cov(a, b) =$

  * $Cov(X, X) =$

  * $Cov(aX + b, cY + d) =$

- 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
>
> 2. $X =$ amount of time it takes for a student to complete a 60-minute exam
>
> 3. $X =$ the number of rolls it takes to get a 6 from rolling a six-sided die

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?

   (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.

   (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?

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.

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

(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 $V(Z)$?

		Y	
		0.0	0.5
X	0.0	0.4	0.1
	0.5	0.1	0.4

(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?

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 $E(X_1)$?

(b) What is $E(X_1|X_2 = 1)$?

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.)

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

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

(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?

(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)?

TABLE 1 Binomial Probabilities

Tabulated values are $P(X \le k) = \sum_{x=0}^{k} p(x_i)$. (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	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	0.6723	0.4095	0.2262	0.0490