Lec 4*: Random Variables; Population Distributions

1 Motivation

- Last lecture, 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
  - 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 end goal of making inference
- In this lecture, 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

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

- 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:

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

*Some exercise questions are taken from or slightly modified based on Dr. Gregory Pac's Econ 310 discussion handout.

– 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

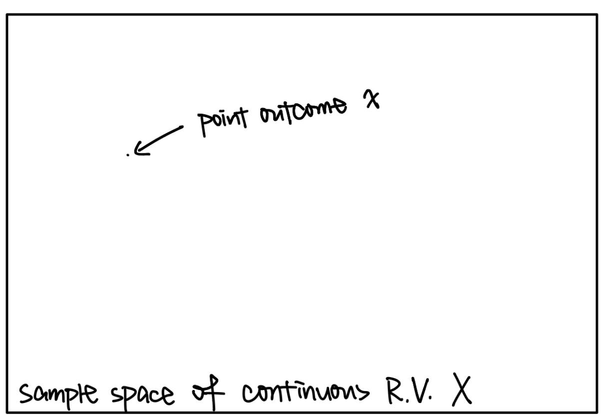
- A discrete probability distribution needs to describe the probability $P(x)$ at all possible x values that a discrete random variable X can take on.
- What are some conditions that a discrete probability distribution needs to satisfy?
  1. $0 \leq P(x) \leq 1$ for all x
  2. $\sum_x P(x) = 1$

- One important example of discrete probability distribution: **Binomial distribution**
  - **Binomial distribution** is a 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 $x = 0, 1, 2, \ldots, n$
      - $E(X) = np$
      - $V(X) = np(1-p)$

4 Continuous Probability Distributions

4.1 Density Functions

- When we examined discrete probability distributions, we said that a discrete probability distribution should describe point probability $P(x)$ at all possible x outcome values.
- Ideally, we would like to provide a similar definition for continuous probability distributions.
- But there's a problem: when random variable X is continuous, **point probability equals to 0 at every single point** ($P(X = x) = 0$ for all x).
  - Reason 1: A continuous random variable has uncountable amount of values, so if each outcome value has probability $\varepsilon > 0$, then the sum of all probabilities would equal to ∞ instead of 1.
  - Reason 2: Say that the rectangle below represents the sample space for a continuous random variable. The probability of hitting a point within the rectangle is the area of the point divided by the area of the rectangle. However, a point has area $= 0$, so the point probability $= 0$.



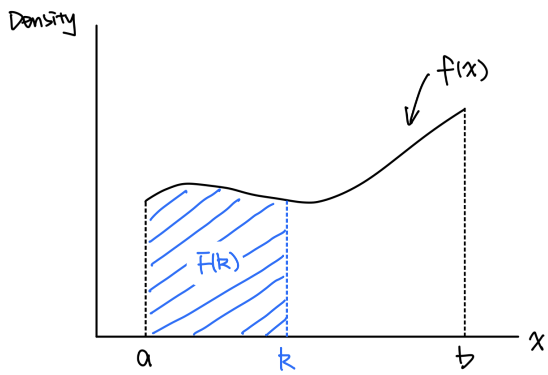
- So our definition for a continuous probability distribution needs to be modified (slightly).
- Solution: describe **density** of x instead of probability.

Definition 1 (Probability density function (PDF)). A function $f(x)$ is called a probability density function (PDF) over $a \leq x \leq b$ if it satisfies the following two criteria:

1. $f(x) \geq 0$ for all x between a and b, and
2. Total area under the curve of $f(x)$ between a and b is 1.

Definition 2 (Cumulative density function (CDF)). A cumulative density function (CDF) describes probability up to a point x. That is, CDF $F(x) = P(X \leq x)$ for random variable X.

Example.



- In this graph, $f(x)$ is the density function for random variable X, and area under the curve describes the cumulative density. For example, the cumulative density up until point k is $F(k)$, which is the total area up until point k.
- The cumulative density describes probability up to a point. Therefore,

$$P(X \leq k) = F(k) = \text{Area under } f(x) \text{ up until point } k$$

- To make sure the density function can properly describe all outcomes, we need to make sure of the following two things:
  1. it's possible to calculate probability $\Rightarrow$ it's possible to calculate area under the density function $\Rightarrow f(x) \geq 0$
  2. total probability should sum up to 1 $\Rightarrow F(b) = 1$

- With the help of density functions, we can finally define a continuous probability distribution:

Definition 3 (Continuous probability distribution). A continuous probability distribution describes a valid PDF $f(x)$ at all possible x values for a continuous random variable X.

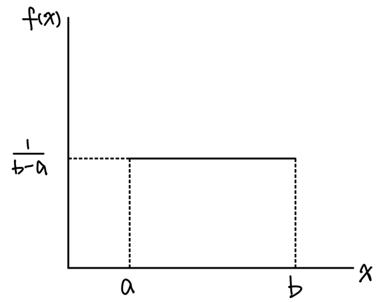
How does this compare with the discrete case?

	Discrete Prob Dist	Continuous Prob Dist
Describes ... at all valid x	$P(x)$	$f(x)$
Range of measure for all valid x	$0 \leq P(x) \leq 1$	$f(x) \geq 0$
How to make sure all valid x are covered	$\sum_x P(x) = 1$	$F(b) = \int_a^b f(x)dx = 1$

4.2 Examples of Continuous Probability Distribution

4.2.1 Uniform Distribution

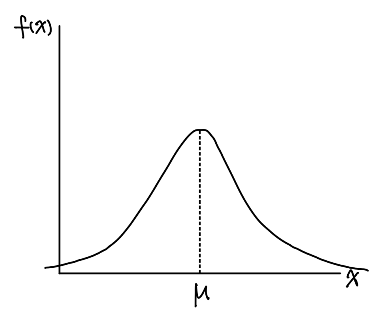
- If X is uniformly distributed between point a and b, then $X \sim \text{Uniform}(a, b)$
- PDF: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$



- $E(X) = \frac{a+b}{2}$
- $V(X) = \frac{(b-a)^2}{12}$

4.2.2 Normal Distribution

- If X is normally distributed with expected value μ and variance σ^2, then $X \sim N(\mu, \sigma^2)$
- PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ for $-\infty < x < \infty$



- Usually, for random variable X that follows a normal distribution, it is helpful to standardize the variable so that we are examining a transformed variable with **standard normal distribution** instead.
  - A random variable Z that follows standard normal distribution is denoted as $Z \sim N(0, 1)$
  - How to transform X to be standard normal? $\Rightarrow$ Since $X \sim N(\mu, \sigma^2)$, $\frac{X-\mu}{\sigma} = Z \sim N(0, 1)$

5 Exercises

1. 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

		Y	
		0.0	0.5
X	0.0	0.4	0.1
	0.5	0.1	0.4

(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 $V(Z)$?

(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?

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

		$X_2 = 0$	$X_2 = 1$
X_1	0	0.30	0.40
	1	0.10	0.20

(a) What is $E(X_1)$?

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

3. 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)?

4. The weekly output of a steel mill is a uniformly distributed random variable that lies between 110 and 175 metric tons

   (a) What is the probability the steel mill will produce more than 150 metric tons next week?

(b) What is the probability the mill will produce between 120 and 160 metric tons next week?

(c) What is the expected value and variance of the mill's weekly output?

5. An analysis of the amount of interest paid monthly by Visa cardholders is normally distributed with a mean of \$27 and a standard deviation of \$6. (Note: A probability table for a standard normal can be found on the last page of the handout.)

   (a) What proportion of Visa cardholders pay less than \$30 in interest?

(b) What proportion pay more than \$42 in interest?

(c) What proportion pay less than \$15 in interest?

6. Answer the following questions.

(a) Let $Z \sim N(0, 1)$. If $P(Z \leq A) = 0.75$, then what is A?

(b) Let $X \sim N(3, 49)$. If $P(X > D) = 0.25$, then what is D?

(c) Let $Y \sim N(\mu, 49)$. If $P(Y < 3) = 0.75$, then what is μ?

(d) Let $M \sim N(3, \sigma^2)$. If $P(M > 4) = 0.4$, then what is σ?

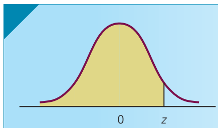
TABLE 1 Binomial Probabilities

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	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

TABLE 3 *(Continued)*



$P(-\infty < Z < z)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990