5.3 Random Variables

Random Variable: a variable whose value is a numerical outcome of a random phenomenon; we typically use capital letters such as $X, Y, Z, \ldots$ to label random variables (rv's)

Probability Distribution (Function): the probability distribution (function) of a random variable tells us which values are possible for this random variable and which probabilities have been assigned to each of these values

Discrete Random Variable: a discrete rv has isolated (i.e., separated) possible values; the range is "countable"; note that the book speaks of a "finite" number of possible values — so our definition is more general here (for example, the set $\{1, 2, 3, 4, \ldots\}$ is countable, but since there is no upper bound, it is not finite)

Continuous Random Variable: a continuous rv takes an entire interval (i.e., a continuous set) of numbers as possible values (for example, all numbers between 0 and 1 or all values on the real axis)

<u>Example:</u>

(i) X = outcome of flipping a coin: discrete/continuous

(ii) Y = outcome of tossing a die: discrete/continuous

(iii) Z = age in years of Stat 2000 students: discrete/continuous

(iv) A = body height of Stat 2000 students: discrete/continuous

(v) B = eye color of Stat 2000 students: discrete/continuous

(vi) C = GPA of Stat 2000 students: discrete/continuous

Discrete Random Variables

A discrete rv X has a finite (or countable) number of possible values. The probability distribution of X lists the values and their probabilities or it provides a mathematical formula how to determine the probability for each particular value:

Value of X	x_1	x_2	$\dots$	x_k	$\dots$	x_n
Probability	p_1	p_2	$\dots$	p_k	$\dots$	p_n

The probabilities p_i must satisfy that

(i) $0 \leq p_i \leq 1$ for every p_i

(ii) $\sum_{i=1}^{n} p_i = 1$ in the finite case or $\sum_{i=1}^{\infty} p_i = 1$ in the countable (but infinite) case

The probability of any event is calculated by summing up the probabilities p_i of the outcomes that make up the event.

Example:

Which of these are valid probability assignments?

(i) For a rv Z: $p_1 = 0.5, p_2 = 0.3, p_3 = -0.4, p_4 = 0.6$: valid/invalid

(ii) For a rv Y: $p_1 = 0.5, p_2 = 0.3$, everything else has probability 0: valid/invalid

(iii) For a rv X: $p_1 = 0.95, p_2 = 0.049, p_3 = 0.001$, everything else has probability 0: valid/invalid

Example:

Assume Z is a rv that represents the lifespan in days of a newly baught light bulb. Most of the bulbs from this series last for 300 days, some for 500 days, and some for 1000 days. But when the manufacturer ceased to test how long indivual bulbs would operate after 1000 days, some were still operational at that time. So we cannot be sure that there isn't a single light bulb (perhaps in the entire annual production of 50,000,000 bulbs) that lasts for 10,000 days. Obviously, the probability for this would be very small but it would still be > 0. This is an example for a discrete rv that has a countable (but not finite) number of possible values.

Formal Notation:

We use a cap letter, e.g., X, to denote a random variable and a small letter, e.g., x, to denote a possible value. Then, the following notation is used to denote the **probability that the random variable X takes the value x**:

$$p_x = p(x) = P(X = x)$$

All of these notations mean the same. In different books, you might encounter only one of these notations, but none of the other notations.

$P(X = x)$ is the **probability distribution (function)** for the rv X and may be given as a graph, in tabular form, via a formula, etc.

In terms of Chapter 4 of the lecture notes, a random variable X relates to a particular experiment or an observational study. Each possible value x_i of the probability distribution (function) that has a probability > 0 is a possible outcome of this experiment.

Example: "Fair" Die
random variable Y = outcome of tossing a die
We can indicate the probability as

$$p_y = p(y) = P(Y = y) = \begin{cases} 1/6, & \text{for } y = 1, 2, 3, 4, 5, 6 \\ 0, & \text{otherwise} \end{cases}$$

Alternatively, we can write

y	$p(y)$	$F(y)$
1	1/6	1/6
2	1/6	
3	1/6	
4	1/6	
5	1/6	
6	1/6	

In this notation, we introduce the **cumulative probability function** (or **cumulative distribution function (cdf)**) of Y which is defined as

$$F(y) = P(Y \leq y) = \sum_{z \leq y} p(z)$$

To fill in the missing values into the table above, we have to calculate

$$\begin{aligned} F(1) &= P(Y \leq 1) = \sum_{z \leq 1} p(z) = p(1) = 1/6 \\ F(2) &= P(Y \leq 2) = \sum_{z \leq 2} p(z) = \ldots \\ F(3) &= P(Y \leq 3) = \sum_{z \leq 3} p(z) = \ldots \\ \ldots \end{aligned}$$

Common Graphical Representations for Probability Functions:

It is common/useful to summarize a probability function (of a discrete rv) through either one of the following three graphical representations:

- **probability histogram** (in case of equally spaced possible values)

- **spike graph**

- **graph of** $F(y)$

It is possible to obtain any of the other representations once we know one of these representations. In particular, we can determine the probability distribution function from a cumulative distribution function and vice versa.

Continuous Random Variables

A continuous rv X can take all values in an interval of numbers. This interval may be

- open on one side, e.g.,
- open on both sides, e.g.,
- closed on both side, e.g.,

In any of these cases, the number of possible outcomes is no longer finite and it is not even countable, i.e., there are inifnite many numbers in any interval.

The **probability distribution** of X is described by a density curve. The probability of any event is the area under the density curve related to that event. Recall the following facts from Section 2.4 (page 15) from the Lecture Notes:

A **probability density function (pdf)** for a continuous random variable X is a function $f(x)$ with the properties

(i) $f(x) \geq 0$ for all values of x

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$, i.e., the area between the graph of $f(x)$ and the horizontal axis equals 1

(iii) $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$, i.e., the probability that X will take a value within the interval $[a, b]$ equals the area under the graph of $f(x)$ between a and b.

Note:

Recall that $P(X = c) = 0$ for any value c for a continuous random variable, i.e., the probability that a continuous rv X will take any specific value c equals 0.

Note:

There is a similarity between **probability distributions** for **discrete** random variables and **pdf's** for **continuous** random variables:

$$p(x) \geq 0 \text{ and } f(x) \geq 0$$

and also

$$\sum_i p(x_i) = 1 \text{ and } \int_{-\infty}^{\infty} f(x)dx = 1$$

The **cumulative distribution function (cdf)**, sometimes also called **cumulative probability function**, for a continuous random variable X is the function $F(x)$, where

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Example:

Which example of a well–known continuous rv X have we seen earlier in this course?
How does its pdf and its cdf look like?

Note:

A cdf $F(x)$ for a continuous rv X has no jumps. It is a continuous curve (i.e., it can be drawn without ever raising the pen and moving it to another location) that remains constant or increases (but never decreases) when x is being increased. A cdf for a continuous rv never jumps like the cumulative probability function for a discrete rv Y.

Example: Well–balanced spinner

This is an example of a **Uniform** distribution:

Reasonable probabilities are

$$\begin{aligned} P(\frac{1}{4} \leq X \leq \frac{3}{4}) &= \\ P(\frac{4}{8} \leq X \leq \frac{5}{8}) &= \\ \text{i.e., } P(a \leq X \leq b) &= \qquad\qquad \text{for } 0 \leq a \leq b \leq 1 \end{aligned}$$

How can we describe this behavior with a pdf and a cdf? Use

$$f(x) =$$

Note the following behavior of $F(x) = P(X \leq x)$:

$$\begin{aligned}
P(X \leq -1) &= \\
P(X \leq 0) &= \\
P(X \leq \frac{1}{4}) &= \\
P(X \leq \frac{1}{2}) &= \\
P(X \leq \frac{3}{4}) &= \\
P(X \leq 1) &= \\
P(X \leq 10) &=
\end{aligned}$$

Therefore, we can conclude that

$$F(x) =$$

Normal Distribution

The most famous continuous probability distribution is the **Normal distribution**, first introduced in Section 2.4 of the Lecture Notes. The pdf of a random variable X that relates to a Normal distribution with **mean** μ and **variance** σ^2, i.e., $X \sim N(\mu, \sigma^2)$, is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \text{for } -\infty < x < \infty, \sigma^2 > 0,$$

where $\exp(x) = e^x$, $\pi \approx 3.1416$, and $e \approx 2.7183$.

While $f(x)$ for a Normal distribution looks difficult, it is even impossible to indicate the cdf $F(x)$ for a Normal distribution. There exists no closed mathematical expression for this cdf.

Therefore, we have to make frequent use of the **Standard Normal distribution** $N(0,1)$ which is tabulated. Indeed, Table A in the textbook contains numerical values of the cdf $F(z)$ for given values of z for the Standard Normal distribution. We obtain values for

$$\Phi(z) = F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt,$$

i.e., the cdf $F(z)$ for tabulated values of z for a random variable $Z \sim N(0,1)$, from Table A.

5.4 Mean and Variance of a Random Variable

It is useful to summarize features of a discrete distribution in terms of the same kind of measures of center and variability we used for data sets. Similar definitions could be given for continuous distributions as well, but this goes beyond the scope of this course since we need to know more about integration.

Mean (or Expected Value) of a discrete random variable X:

$$\mu = E(X) = \sum_i x_i p(x_i)$$

Variance of a discrete random variable X:

$$\begin{aligned}
\sigma^2 &= Var(X) \\
&= \sum_i (x_i - \mu)^2 p(x_i) \\
&= \left(\sum_i x_i^2 p(x_i) \right) - \mu^2 \\
&= \left(\sum_i x_i^2 p(x_i) \right) - \left(\sum_i x_i p(x_i) \right)^2
\end{aligned}$$

Note:
The last two lines for the calculation of the variance of a discrete rv represent a shortcut that is similar to the shortcut used for the calculation of the variance for data sets.

Standard Deviation of a discrete random variable X:

$$\sigma = \sqrt{\sigma^2}$$

Law of Large Numbers:
If we repeat a random phenomenon many times and record the value of X each time, the average $\overline{X}$ will get closer and closer to the true mean μ as we make more and more repetitions. This fact is called the **Law of Large Numbers**.

Empirical Rule:

For a bell–shaped random variable X, we can indicate an empirical rule that is similar to the one we introduced for data sets:

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &\approx 0.68 \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 0.95 \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 0.997 \end{aligned}$$

Example: "Fair" Die:

i	y_i	$p(y_i)$	$y_i p(y_i)$	$y_i^2 p(y_i)$
1	1	1/6		
2	2	1/6		
3	3	1/6		
4	4	1/6		
5	5	1/6		
6	6	1/6		

Now calculate

$$\mu = \sum_i y_i p(y_i) =$$

$$\sum_i y_i^2 p(y_i) =$$

$$\sigma^2 =$$

$$\sigma =$$

So, for a random variable X that represents a fair die, the mean μ is ..... , the variance σ^2 is ..... , and the standard deviation σ is ..... .

Example:

In a previous Stat 2000 class, the following point gains and point losses between Quiz 1 and Quiz 2 have been reported:

# Students	Point Gain/Loss
2	-5
3	-1
3	0
4	+1
2	+2
4	+3
2	+5

Let the random variable Y represent the point gain/loss.

Questions:

- What is the probability $p(y)$ to gain (or lose) y points?
- One student missed Quiz 2. What is the expected point gain/loss for this student?
- Assuming a bell–shaped curve, in which interval (centered around the mean) would 95% of future observations from the same distribution fall (e.g., students from another session of the same course, that did not take Quiz 2 yet)?

i	y_i	$p(y_i)$	$y_i p(y_i)$	$y_i^2 p(y_i)$
1	-5			
2	-4			
3	-3			
4	-2			
5	-1			
6	0			
7	1			
8	2			
9	3			
10	4			
11	5			

Now calculate

$$\mu = \sum_i y_i p(y_i) =$$

$$\sum_i y_i^2 p(y_i) =$$

$$\sigma^2 =$$

$$\sigma =$$

From the empirical rule, we get

Therefore, 95% of future observations are expected in the interval ... .