Module 5 – Intro to Probability Distributions

Reading: Sections 2.9, 3.1 – 3.5

Random Variables

- So far, we have used descriptions to identify the outcomes of a sample space (brands, condition, color)
- In this course, it will be helpful to assign a number to each outcome.
- We can call each type of outcome a variable (test result, actually getting an illness)
- Since the final result of that variable is not known, we call it a random variable.

- A random variable assigns a real number to each outcome in the sample space.
- We express random variables with uppercase letters.
- We express individual outcomes of the random variable with lowercase letters.

For example, the random variable X represents the volume of liquid evaporated during an experiment. After the experiment is conducted and we know that volume, we say that the volume is x.

Just like sample spaces, random variables can be discrete or continuous.

Examples of discrete random variables:

- no. of scratches on a surface
- number of defective parts
- number of errors received

Examples of continuous random variables:

- electrical current
- temperature
- time

Probability Distribution

The probability distribution of a random variable X is a description of the probabilities associated with all the possible values of X. For discrete random variables, the probability distribution is usually presented as a table.

Probability Mass Function

The probability mass function (PMF), $f(x)$ is a function that describes the probability distribution for a discrete random variable. The function must satisfy:

- $f(x) \geq 0$ for all x
- $\sum_{i=1}^{n} f(x_i) = 1$
- $f(x) = P(X = x)$

From these definitions, we see that the sum of all the probabilities in the probability distribution of a random variable must be equal to 1.

Example: Camera Flash Test

Suppose you perform three tests on a camera flash and used the data to develop the following probability distribution, where X represents the number of times the flash passed the test. What is the probability that if you pick one of the cameras at random, it's flash only passes 2 of the tests?

X	$f(x)$
0	0.008
1	0.096
2	0.384
3	0.512

- From the table, we see that the probability that $X = 2$ is $f(2) = P(X = 2) = 0.128$.
- It is good practice to make sure the table is correct by ensuring that the sum of probabilities add up to 1.

Cumulative Distribution Function

A cumulative distribution function, $F(x)$ is a function that shows the probability that a random variable is equal to or less than a given outcome. The cumulative distribution function (CDF) must satisfy the following:

- $F(x) = P(X \leq x)$
- $0 \leq F(x) \leq 1$
- If $x \leq y$ then $F(x) \leq F(y)$

Example: Camera Flash Test

Write a cumulative distribution function for the table in the previous example.

- To write the CDF, it is necessary to add all of the probabilities above the probability of a given outcome

X	$f(x)$	$F(x)$
0	0.008	0.008
1	0.096	0.104
2	0.384	0.488
3	0.512	1

- Note that the final value of any CDF is always 1.
- It is important that you learn how to do this function in Excel or another spreadsheet program.

Mean

The mean is a measure of the center of a probability distribution. It is sometimes called a weighted average.

$$\mu = E(X) = \sum_{i=1}^{n} x_i f(x_i)$$

Note that this mean is not the same as the arithmetic average that you are used to, unless your random variable has equally likely outcomes.

Variance

The variance is a measure of the dispersion or variability in a probability distribution. It is the expected squared deviation of a variable from its mean.

$$\sigma^2 = V(X) = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i) = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2$$

The last term in the equation above is sometimes expressed as $E(X^2) - [E(X)]^2$. The derivation is relatively easy if we understand how probability distributions work.

Standard Deviation

The standard deviation is the square root of the variance. It is the expected deviation of a variable from its mean.

$$\sigma = SD(X) = \sqrt{\sigma^2}$$

Example: Measures of a Discrete Probability Distribution

For the data in the previous examples, determine the expected value of tests that a camera flash passes and the expected deviation of the number of tests passed from the mean.

- The wording in the example is just jargon for "find the mean and the standard deviation"
- It is best to learn how to find these in Excel.
- It is recommended that you first write the equations in Excel manually but then to familiarize yourself with built-in functions.

- For the mean, let's first find $xf(x)$ for each value.

$\boldsymbol{X}$	$\boldsymbol{f(x)}$	$\boldsymbol{xf(x)}$
0	0.008	0
1	0.096	0.096
2	0.384	0.768
3	0.512	1.536

- Then, we add those values up to obtain the mean: $\mu = 2.4$
- For the standard deviation, we must first determine the variance. First, let's find $(x - \mu)^2 f(x)$ for each value.

$\boldsymbol{X}$	$\boldsymbol{f(x)}$	$\boldsymbol{xf(x)}$	$\boldsymbol{(x-\mu)^2 f(x)}$
0	0.008	0	0.04608
1	0.096	0.096	0.18816
2	0.384	0.768	0.06144
3	0.512	1.536	0.18432

- Then, we add those values to obtain the variance: $\sigma^2 = 0.48$
- The standard deviation is the square root of the variance $\sigma = \sqrt{\sigma^2} \approx 0.693$

Discrete Uniform Distributions

If each outcome in a random variable is equally likely to occur, then its probability distribution is uniform. It's probability mass function is then:

$$f(x) = \frac{1}{n}$$

There are special cases of discrete uniform distributions where the mean and variance can be calculated using simplified formulas. You can read them in Section 3.4 and complete Examples 3.10, 3.11, and 3.12 for practice.

Bernoulli Trial

- Jacob Bernoulli was a mathematician and member of the prominent Bernoulli family
- He developed the Bernoulli process, a sequence of binary random variables with consistent probability distributions.

First, some definitions:

- A binary variable is one that only has two possible outcomes.
- A Bernoulli trial is a binary variable where there are only two outcomes: success or failure
- The outcomes are typically expressed in the random variable as 1 (success) and 0 (failure)
- The probability of success is denoted as p. The probability of failure is then $1-p$.
- Examples:
    - Flipping a coin. 2 outcomes: Heads or Tails
    - Asking a yes/no question in a survey. 2 outcomes: Yes or No
    - Testing if a screen works. 2 outcomes: Works or Does Not Work
    - Observing a car drive through a highway without incident. 2 outcomes: No Incident or Incident
- It is assumed that Bernoulli trials are independent of each other, that means that the probabilities of success and failure remain constant throughout the experiment.

Binomial Distribution

If an experiment consists of n Bernoulli trials, then its random variable is called a binomial random variable and has a binomial distribution, whose PMF is defined as:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Recall that the first term in the equation is the Binomial coefficient and is defined by $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

Binomial distributions are another special case where the mean and variance can be calculated using simplified formulas. You can read them in Section 3.5 and complete Examples 3.13, 3.14, 3.15, and 3.16 for practice.

Math Review: Binomial Expansion

The Binomial coefficient receives its name because it is the coefficient found in the binomial expansion. For two constants a and b, the binomial expansion is:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Note that if $a = p$ and $b = 1 - p$ then the Binomial expansion is 1. Practically, this is because the Binomial expansion shows the sum of probabilities of a Binomial distribution, and we know that the sum of probabilities of a probability distribution is 1.

Recommendation: Probability Distributions and their Measures

To help complete homework and prepares to quizzes and exams, I recommend that you fill out this table throughout the semester as we cover more probability distributions. I have filled it out with this week's material.

Distribution	PMF/PDF	μ	σ^2
Discrete	$f(x)$	$\sum x_i f(x_i)$	$\sum (x_i - \mu)^2 f(x_i)$
Discrete Uniform	$\frac{1}{n}$	$\frac{\sum x_i}{n}$	$\frac{\sum (x_i - \mu)^2}{n}$
Discrete Uniform Distribution with consecutive integers $a, a+1, a+2 \ldots b$.	$\frac{1}{n}$	$\frac{b+a}{2}$	$\frac{(b-a+1)^2 - 1}{12}$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	np	$np(1-p)$

Homework: 2.9.1, 3.1.7, 3.2.5, 3.3.3, 3.4.3, 3.5.3

Project Part 3: Derive the expression $V(X) = E(X^2) - [E(X)]^2$ starting from the general definition of variance $\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i)$. Show and explain all your steps.

References

Montgomery, Runger, *Applied Statistics and Probability for Engineers*