Ch. 16 – Random Variables

Def'n: A random variable is a numerical measurement of the outcome of a random phenomenon.

A discrete random variable is a random variable that assumes separate values.

→ # of people who think stats is dry

The probability distribution of a discrete random variable lists all possible values that the random variable can assume and their corresponding probabilities.

Notation: X = random variable; x = particular value;
$P(X = x)$ denotes probability that X equals the value x.

Ex16.1) Toss a coin 3 times. Let X be the number of heads.

8 possible outcomes: HHH, HHT, HTH, THH, HTT, THT, TTH, TTT

Table 16X0

x	$P(X = x)$
0	0.125
1	0.375
2	0.375
3	0.125

Two noticeable characteristics for discrete probability distribution:

1. $0 \le P(X = x) \le 1$ for each value of x
2. $\sum P(X = x) = 1$

Ex16.2) Find the probabilities of the following events:

"no heads": $P(X = 0) = 0.125$

"at least one head": $P(X \ge 1) = P(X = 1) + P(X = 2) + P(X = 3)$
$= 0.375 + 0.375 + 0.125 = 0.875$

"less than 2 heads": $P(X < 2) = P(X = 0) + P(X = 1) = 0.125 + 0.375 = 0.500$

The population mean μ of a discrete random variable is a measure of the center of its distribution. It can be seen as a long-run average under replication. More precisely,

$$\mu = \sum x_i P(X = x_i)$$

Sometimes referred to as $\mu = E(X)$ = the expected value of X.

Keep in mind that μ is not necessarily a "typical" value of X (it's not the mode).

Ex16.3) Using Table 16X0,

$$\mu = \sum x_i P(X = x_i) = (0)(0.125) + (1)(0.375) + (2)(0.375) + (3)(0.125) = 1.5$$

→ On average, the number of heads from 3 coin tosses is 1.5.

Ex16.4) Toss an unfair coin 3 times (hypothetical). Let X be as in previous example.

x	$P(X = x)$
0	0.10
1	0.05
2	0.20
3	0.65

$$\mu = \sum x_i P(X = x_i)$$
$$= (0)(0.10) + (1)(0.05) + (2)(0.20) + (3)(0.65)$$
$$= 2.4$$

As 2nd example shows, interpretation of μ as a measure of center of a distribution is more useful when the distribution is roughly symmetric, less useful when the distribution is highly skewed.

The population standard deviation σ of a discrete random variable is a measure of variability of its distribution. As before, the standard deviation is defined as the square root of the population variance σ^2, given by

$$\sigma^2 = \sum (x_i - \mu)^2 P(X = x_i) = \sum x_i^2 P(X = x_i) - \mu^2$$

Ex16.5) From Table 16X0,

$$\sigma^2 = \sum x_i^2 P(X = x_i) - \mu^2 = [0^2(1/8) + 1^2(3/8) + 2^2(3/8) + 3^2(1/8)] - 1.5^2 = \frac{24}{8} - \frac{9}{4} = \frac{6}{8} = \frac{3}{4}$$

<u>Ch. 17 – Probability Models</u>

Continuous Distributions

Def'n: A <u>continuous random variable</u> assumes any value contained in one or more intervals.
→ average alcohol intake by a student, average alcohol outtake by a student
The <u>probability distribution</u> of a continuous r.v. is specified by a curve.

Two noticeable characteristics for continuous probability distribution (sans calculus):
1. The probability that X assumes a value in any interval lies in the range 0 to 1.
2. The interval containing all possible values has probability equal to 1, so the total area under the curve equals 1.

(corresponding diagrams drawn in class)

Using probability symbols, point 1 is denoted by

$P(a \le X \le b)$ = Area under the curve from a to b

The probability that a continuous random variable X assumes a single value is always zero. This is because the area of a line, which represents a single point, is zero.

(diagram drawn in class)

In general, if a and b are two of the values that X can assume, then

$P(a) = 0$ and $P(b) = 0$

Hence, $P(a \le X \le b) = P(a < X < b)$. For a continuous probability distribution, the probability is always calculated for an interval, such as $P(X > b)$ or $P(X \le a)$.

Ex17.1) Suppose we have a "uniform" distribution where obtaining each value between 2 endpoints has equal probability. Suppose the endpoints are 0 and 2.
a) What is the probability of $P(X < 1.5)$?
b) What is the probability of $P(X \le 1.5)$?
c) What is the probability of $P(0.5 < X < 2.5)$?

The Normal Distribution

- most widely used and most important of all (continuous) probability distributions
- the normal distribution has 2 *parameters*: μ and σ
- the density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- the *normal (distribution) curve*, when plotted, gives a bell-shaped curve such that
  1. The total area under the curve is 1.0.
  2. The curve is symmetric about the mean (or bell-shaped).
  3. The two tails of the curve extend indefinitely.

  (diagram drawn in class)
- there is not just one normal curve, but a *family* of normal curves. Each different set of μ and σ gives a different curve. μ determines the center of the distribution and σ gives the spread of the curve. (diagrams drawn in class)

Standard Normal Distribution

Def'n: The standard normal distribution is the normal distribution with $\mu = 0$ and $\sigma = 1$. It is the distribution of normal z-scores.

Recall Empirical Rule. $\mu \pm \sigma$ gives middle 68.26% of the data. In terms of z-scores, this is the interval (-1.0, 1.0).

(diagram drawn in class)

$\mu \pm 2\sigma$ gives middle 95.44% of the data. z-score interval of (-2.0, 2.0).
$\mu \pm 3\sigma$ gives middle 99.74% of the data. z-score interval of (-3.0, 3.0).

Using Table of Standard Normal Curve Areas:

For any number z between -3.90 and 3.90 and rounded to 2 decimal places, Table Z gives

(area under curve to the left of z) = $P(Z < z) = P(Z \leq z)$ $\qquad Z \sim N(0, 1)$

Tips & tricks:

- diagrams are helpful
- Complement: $P(Z \geq z) = P(Z > z) = 1 - P(Z \leq z)$
- Symmetry: $P(Z \geq z) = P(Z \leq -z)$
- $P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$
- If $z > 0$, then $P(-z \leq Z \leq z) = 1 - 2P(Z \leq -z)$

Ex17.2) *Examples with z-scores (finding prob.):*

a) $P(Z < -3.14) = 0.0008$

b) $P(Z > 1.44) = 1 - P(Z < 1.44) = 1 - 0.9251 = 0.0749$

OR $P(Z > 1.44) = P(Z < -1.44) = 0.0749$

c) $P(-3.14 \leq Z \leq 1.44) = P(Z \leq 1.44) - P(Z \leq -3.14) = 0.9251 - 0.0008 = 0.9243$

d) $P(-2.00 \leq Z \leq 2.00) = 1 - 2P(Z \leq -2.00) = 1 - 2(0.0228) = 0.9544$

Standardizing a normal distribution:

$X \sim N(\mu, \sigma)$ and $Z \sim N(0, 1)$. What is Z? $Z = \dfrac{X-\mu}{\sigma}$, $z = \dfrac{x-\mu}{\sigma}$

$P(X \le x) \rightarrow P\left(\dfrac{X-\mu}{\sigma} \le \dfrac{x-\mu}{\sigma}\right) = P(Z \le z)$

Ex17.3) *Examples with z-scores (standardizing):*
Find the following probabilities for $X \sim N(75, 6.5)$.

a) What is the probability of getting a value greater than 94.5?

$$P(X > 94.5) \rightarrow P\left(\frac{X-\mu}{\sigma} > \frac{94.5-75}{6.5}\right) = P(Z > 3.00) = P(Z < -3.00) = 0.0013$$

b) What is the probability of getting a value between 71.75 and 84?

$$P(71.75 \le X \le 84) \rightarrow P\left(\frac{71.75-75}{6.5} \le \frac{X-\mu}{\sigma} \le \frac{84-75}{6.5}\right) = P(-0.50 \le Z \le 1.38)$$

$$= P(Z \le 1.38) - P(Z \le -0.50) = 0.9162 - 0.3085 = 0.6077$$

Identifying values:
Using the area under the curve, you can find appropriate z values; so, what are the corresponding x values?
$x = \mu + z\sigma$

Ex17.4) *Examples with z-scores (finding values):*
Use the same X as in Ex17.3) to answer the following:

a) What value denotes the top 5%?

$P(Z > z) = 0.05$ $\rightarrow$ look for 0.95 in table; in the middle of 1.64 and 1.65, so $z = 1.645$
$x = \mu + z\sigma = 75 + 1.645(6.5) = 85.69$

To be in the top 5%, the corresponding 'x value' needs to be greater than 85.69.

b) What values bound the middle 70% of the data?

Look for 0.15 and 0.85 in table; $z = -1.04$ and $z = 1.04$, respectively.

$x_1 = \mu + z\sigma = 75 + (-1.04)(6.5) = 68.24$
$x_2 = \mu + z\sigma = 75 + 1.04(6.5) = 81.76$

The middle 70% of the data is represented by the interval (68.24, 81.76).

Combinations and Functions of Random Variables
For any constants a and b,

Means:
1. $E(a) = a$
2. $E(aX) = aE(X)$
3. $E(aX + b) = aE(X) + b$
4. $E(aX \pm bY) = aE(X) \pm bE(Y)$

Variances:
1. $V(a) = 0$
2. $V(aX) = a^2V(X)$
3. $V(aX + b) = a^2V(X)$
4. $V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm$ ~~$2ab\text{cov}(X, Y)$~~

Rule 4 for variance eliminates the last component only if X and Y are independent.

Ex17.5) If X_1, X_2, and Y are independent random variables such that

$E(X) = E(X_1) = E(X_2) = 4$ $\quad$ $E(Y) = -3$

$V(X) = V(X_1) = V(X_2) = 3$ $\quad$ $V(Y) = 1$

a) Find the mean and standard deviation of $W = X_1 + X_2$. (Note that $W \neq 2X$)

$E(W) = E(X_1 + X_2) = E(X_1) + E(X_2) = 4 + 4 = 2(4) = 8$ $\quad$ $E(2X) = 2E(X) = 2(4) = 8$

$V(W) = V(X_1 + X_2) = V(X_1) + V(X_2) = 3 + 3 = 2(3) = 6$ $\quad$ $V(2X) = 2^2V(X) = 4(3) = 12$

$\sigma = \sqrt{V(X_1 + X_2)} = \sqrt{6} = 2.449$

b) Find the mean and standard deviation of $T = 4X_1 - 3Y$.

$E(T) = E(4X_1 - 3Y) = 4E(X_1) + (-3)E(Y) = 4(4) + (-3)(-3) = 25$

$V(T) = V(4X_1 - 3Y) = 4^2V(X_1) + (-3)^2V(Y) = 16(3) + (9)(1) = 57$

$\sigma = \sqrt{V(T)} = \sqrt{57} = 7.550$

c) Find the mean and standard deviation of $K = \frac{X_1 + X_2 + Y}{3} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}Y$.

$$E(K) = E\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}Y\right) = E\left(\frac{1}{3}X_1\right) + E\left(\frac{1}{3}X_2\right) + E\left(\frac{1}{3}Y\right) = \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(Y)$$

$$= \frac{1}{3}(4) + \frac{1}{3}(4) + \frac{1}{3}(-3) = \frac{5}{3}$$

$$V(K) = V\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}Y\right) = V\left(\frac{1}{3}X_1\right) + V\left(\frac{1}{3}X_2\right) + V\left(\frac{1}{3}Y\right) = \left(\frac{1}{3}\right)^2 V(X_1) + \left(\frac{1}{3}\right)^2 V(X_2) + \left(\frac{1}{3}\right)^2 V(Y)$$

$$= \frac{1}{9}(3) + \frac{1}{9}(3) + \frac{1}{9}(1) = \frac{7}{9} = 0.777$$

$$\sigma = \sqrt{V(K)} = \sqrt{7/9} = 0.882$$