Ch. 14 – 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.

Ex14.1) Toss a coin 3 times. Let X be the number of heads. What is the prob. dist'n?

Table 14X0

x	$P(X = x)$

Two noticeable characteristics for discrete probability distribution:

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

Ex14.2) Find the probabilities of the following events:

"no heads":

"at least one head":

"less than 2 heads":

Ex14.3) Refer back to Ex13.19. Suppose Bob and Mark play this game twice and that each play is independent of the other. Define X as the total number of hits on two independent plays. Define Y_i as the total number of hits on play i. Find the probability distribution of X.

Ex14.4) Suppose you roll two dice. If you roll 7 or 11, you win \$20. Otherwise, you win nothing. a) Let X be your winnings. Find the probability distribution of X. b) Suppose you pay \$10 to play the game. Let Y be your net profit. Find the probability distribution of Y.

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).

Ex14.5) Find the mean for Ex14.1).

Ex14.6) 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

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.

Ex14.7) Using Ex14.3), what is the expected total number of hits on two independent plays?

Ex14.8) Using Ex14.4), what are the expected winnings? The expected net profit? How much would you pay to play this game?

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

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

Ex14.9) Find the standard deviation for X in Ex14.1).

Ex14.10) Using Ex14.3), find the standard deviation for the total number of hits on two independent plays.

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.

Using probability symbols, point 1 is denoted by

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

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.

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

$$P(a) = 0 \qquad \text{and } P(b) = 0$$

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 $P(X \le a)$.

Ex14.11) Suppose we have a "uniform" distribution where obtaining each value between two 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 $P(X \leq 1.5)$?
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 <u>normal distribution</u> 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.

- 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.

Standard Normal Distribution

Def'n: The <u>standard normal distribution</u> 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).
$\mu \pm 2\sigma$ 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) = $P(Z < z) = P(Z \leq z)$ $\qquad Z \sim \mathrm{N}(0, 1)$

Helpful tips:
- 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 $P(-z \leq Z \leq z) = 1 - 2P(Z \leq -z)$

Ex14.12) *Examples with z-scores (finding prob.):*
a) $P(Z < -3.14) =$
b) $P(Z > 1.44) =$
OR $P(Z > 1.44) =$
c) $P(-3.14 \leq Z \leq 1.44) =$
d) $P(-2.00 \leq Z \leq 2.00) =$

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

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

Ex14.13) *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?

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

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

Ex14.14) *Examples with z-scores (finding values):*
Use the same X as in Ex14.13) to answer the following:

a) What value denotes the top 5%?

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

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

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

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

Ex14.15) Let X be the temperature in Edmonton on a random day in March and Y be the temperature on a random day in February. Suppose all days are independent and that

$E(X) = 4$, $V(X) = 3$ $\qquad$ $E(Y) = –3$, $V(Y) = 1$

a) Find the mean and standard deviation of $W = 4X – 3Y – \pi$.

b) Find the mean and standard deviation for the total of two random days in March.

c) Find the mean and standard deviation for the difference between the total of two random days in March and the total of three random days in February.

d) Find the mean and standard deviation of the average of two random days in March and one random day in February.