Statistics 213 – 2.1: Random Variables

UNIVERSITY OF CALGARY

Textbook:

4.1, 4.2, 4.3

Objectives:

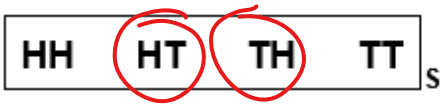
- Know the definition of a ***random variable***
- Know the definition of a ***discrete random variable***
- Be able to construct and use a ***probability distribution table***
- Be able to sketch and/or interpret a probability distribution graph
- Be able to calculate the ***expected value***, ***variance***, and ***standard deviation*** of a random variable as well of a linear transformation of a random variable

Motivation:

To best introduce the concept of random variables in statistics, let's briefly review how, up until this point, we've been approaching probability calculations in class. We'll go back to a basic example from the 1.1 notes.

Example

An experiment is performed in which two fair coins are tossed. The sample space of this experiment is listed here:



Let A be the event that we observe exactly one head. As we've learned in the previous units, to calculate the probability of observing exactly one head, we calculate $P(A)$ as follows:

$$P(A) = \frac{n(A)}{n(S)} = \frac{2}{4} = \frac{1}{2}$$

In this case, we are interested in a specific outcome. We defined this outcome of interest as event A and calculate the probability of A occurring in any given instance of the experiment.

Notice that we really don't take into account the other possible outcomes of the experiment. In other words, we don't define an event for observing exactly zero heads or an event for observing exactly two heads. We just lump these other outcomes into A^C, the complement of the event we're interested in.

But there is another way we could look at these outcomes. Suppose we were to focus on *the number of heads* that could result from the experiment described above. As we can see from the sample space, a given iteration of this experiment could result in any one of the following number of heads:

- Zero heads
- One head
- Two heads

In other words, the number of heads is ***variable*** - it varies from iteration to iteration and does so randomly (assuming the coins and the method for flipping the coins are all fair) - and can take on the values zero, one, or two.

We will now be focusing on quantities that can randomly take on one of multiple different values (quantities like the number of heads resulting from tossing two coins) and using these quantities to calculate probabilities. That is, rather than defining a specific event for each possible outcome we're interested in, we're going to be using quantities that can represent any of the possible values that an outcome of interest can attain.

Random Variables

In statistics, a ***random variable*** (often just called a variable) is a quantity whose values are the numerical values associated with the random outcomes of an experiment. Often times, we represent a random variable with a letter.

The best way to get an understanding of what random variables are is by looking at a few examples.

Example

An experiment is performed in which two fair coins are tossed. Let X be the number of heads resulting from an iteration of this experiment.

Here, X is a random variable that can take on values:

$$x = \{0, 1, 2\}$$

Example

A family is planning on having four children. Let Y be the number of boys they ultimately end up having.

Here, Y is a random variable that can take on values:

$$y = \{0, 1, 2, 3, 4\}$$

uppercase for what could happen

lowercase for what did happen

Example

A hockey game is played between rival teams. Let W be the number of goals scored by one of the teams in the game.

Here, W is a random variable that can take on values:

$$w = \{0, 1, 2, 3 \cdots\} = [0, \infty) \quad w \in \mathbb{Z}$$

Now we can think of probability in terms of these random variables instead of specifically defined events.

Examples

- What is the probability that $X = 2$? → two heads when tossing twice = $P(x=2) = \left(\frac{1}{2}\right)^2$
- What is the probability that $Y = 0$? → zero boys out of 4
- What is the probability that $W = 4$?

$P(Y = y)$ ← realizing a possibility; concept →

We will be discussing two types of random variables as the course progresses: discrete random variables and continuous random variables.

Discrete random variables are variables that can take on a countable number of values in a given range. Think of the values of a discrete variable as being "steps" - the possible values jump or step up from one number to the next. Often (but not always!), discrete random variables only take on integer values. categories that are mutually exclusive

Continuous random variables are variables that can take on an infinite (uncountable) number of values in a given range. Think of the values of a continuous variable as being a "ramp" - since all values are possible, there is no "jumping" or "stepping" from one value to the next. measurements continuum

For now, we will focus only on discrete random variables, so don't worry about continuous random variables just yet!

Describing Discrete Random Variables: Probability Distributions

whole numbers

When we're dealing with a discrete random variable, it is often useful for us to, if possible, list all the values that the random variable can assume as well as to list the probabilities associated with those values.

The ***probability distribution*** of a discrete random variable is a table, graph, or formula that specifies every possible value that a random variable can assume along with the probabilities associated with each of these values.

Suppose we have a discrete random variable X. The probability distribution for X must meet the following two requirements (we first saw these properties in the 1.1 notes):

1. The probability that X can take on a given value x must be between 0 and 1. Let x_i represent the i^{th} observation (instance) of X. Then:

$$0 \leq P(X = x) \leq 1$$

2. The sum of probabilities of all n possible values of X must sum to 1. Let x_i represent the i^{th} observation (instance) of X. Then:

$$\sum_{\text{all } x} P(X = x) = 1 \quad \approx \quad \sum_{i=1}^{n} P(X = x) = 1$$

Probability Distribution Tables

A ***probability distribution table*** is a useful way of displaying all the values a random variable can assume as well as the probabilities associated with those values. A probability distribution table is also useful for calculating probabilities involving a random variable.

Example 2.1.1

An experiment is performed in which two fair coins are tossed. Let X be the number of heads that result from this experiment. Fill in the following probability distribution table for this scenario.

	TT	TH, HT	HH
x	$x = 0$	$x = 1$	$x = 2$
P(X = x)	$\left(\frac{1}{2}\right)^2 = \frac{1}{4}$	$2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{2}{4}$	$\left(\frac{1}{2}\right)^2 = \frac{1}{4}$

pdf ≈ pmf Probability density function, Probability mass function

$$p(x) = \binom{2}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{2-x} \text{ where } x = 0, 1, 2$$

Example 2.1.2

During the 2007 baseball season, Alex Rodriguez played 74 games in which he had exactly four at-bats. The following partially-completed probability distribution table shows the distribution of Y, where Y is the number of hits Rodriguez got per game.

1. Complete the probability distribution table.

Y	0	1	2	3	4
Frequency	22	27	15	$74-22-27-15-2=8$	2
P(Y = y)	$\frac{22}{74} = 0.297$	$0.3649 = \frac{27}{74}$	$0.2027 = \frac{15}{74}$	$\frac{8}{74} = 0.102$	$0.027 = \frac{2}{74}$

→ this needs to = 74

→ this column needs to equal 1

$Y = 0 : 4$

prob

2. What is the probability that Rodriguez had fewer than two hits in a given game?

$P(Y < 2) = P(Y \leq 1) = P(Y=0) + P(Y=1)$

$= \frac{22}{74} + \frac{27}{74}$

$= \frac{49}{74}$

→ means $<$ not $\leq$

if its @ most then $\leq$

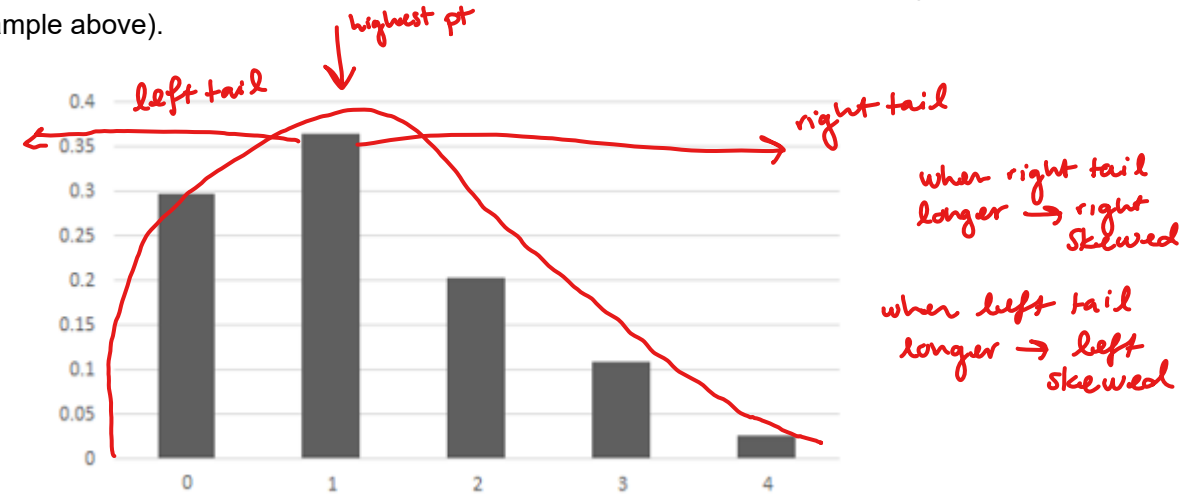
Probability Distribution Graphs

A probability distribution table can give us a lot of useful information, but sometimes it's easier to see the "tendency" or pattern of a random variable in a graph.

A ***probability distribution graph*** plots probability (0 to 1) on the y-axis and the values of the random variable on the x-axis. A graph like this can be made using software (like Excel or R) or can be sketched by hand.

Example

The following graph shows the probabilities of the number of hits per four at-bats for Alex Rodriguez in the 2007 season (from the example above).



From the graph, it's easy to see that Rodriguez has a greater tendency to get either no hits or just one hit in four at-bats.

The general shape of graphs like this can be described in terms of *symmetry* or *skew*.

A ***symmetric*** graph is one that looks the same on both sides of the "middle" of the graph. → no skew

A ***right-skewed*** graph is one that has a "tail" pointing off in the right direction (the mass of the graph is on the left-hand side).

A ***left-skewed*** graph is one that has a "tail" pointing off in the left direction (the mass of the graph is on the right-hand side).

large numbers are more likely

Describing Discrete Random Variables: Summary Values

We often want to be able to give more of a description of a random variable than just its probability distribution. One way to do so is to provide some "summary values" of the random variable - values that can be used to describe the tendency and shape of the probability distribution.

The ***mean*** or ***expected value*** of a discrete random variable X is the average value of X over a (theoretical) infinite number of repetitions of an experiment. It is a measure of "central tendency" and is calculated as follows:

This is a weighted average like calculating grades

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i p_i$$

where x_i is the i^{th} value that the random variable X can assume and p_i is the probability of that i^{th} value occurring.

$E[X] = 90(0.25) + 65(0.5) + 70(0.45)$

The ***variance*** of a discrete random variable X is a measure of the variability of X over an infinite number of repetitions of an experiment. It is a measure of "spread" and is calculated as follows:

$$VAR(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$$

second moment - first moment^2

The ***standard deviation*** of a discrete random variable is another measure of the variability of X and is simply the (positive) square root of the variance:

$$SD(X) = \sqrt{VAR(X)}$$

variance

Example 2.1.3

An experiment is performed in which two fair coins are tossed. Let X be the number of heads that result. The following is a probability distribution table for X.

x	0	1	2
P(X = x)	0.25	0.50	0.25

$E[X] = \quad = 0(0.25) + 1(0.5) + 2(0.25)$

1. Find the expected value of X.

x	0	1	2
P(x)	0.25	0.5	0.25

$$\sum x_i p(x_i) = 0(0.25) + 1(0.5) + 2(0.25)$$
$$= 0.5 + 0.5$$
$$= 1$$

2. Find the variance of X.

$$Var(x) = E\left[(x - E(x))^2\right] = E(x^2) - E(x)^2$$

Step 1

$$E(x) = \sum x_i p(x_i) = 1$$

Step 2

$$E(x^2) = \sum x_i^2 p(x_i) = 0^2(0.25) + 1^2(0.5) + 2^2(0.25)$$
$$= 0 + 0.5 + 1$$
$$= 1.5$$

Step 3

$$V(x) = E(x^2) - E(x)^2$$
$$= 1.5 - (1)^2$$
$$= 0.5$$

$$SD(x) = \sqrt{V(x)} = \sqrt{0.5} = 0.7071068$$

Range $1 - SD$ to $1 + SD$

$1 - 0.7071068$ to $1 + 0.7071068$

Example 2.1.4

A lottery game is played as follows: you select three integers (without replacement) from zero to nine (zero and nine included). At the end of the week, the lottery operator randomly selects three integers (without replacement) from zero to nine (zero and nine included). If you match the lottery operator on all three numbers, you win $100. If you match on exactly two numbers, you win $10. Matching exactly one number means you win $1. If you do not match any of the numbers, you have to pay $5.

1. Let X represent your profit. Create a probability distribution table for X.

let Y be # out of 3 you choose that $Y = \{0, 1, 2, 3\}$

$X = x$	-5	1	10	100
$Y = y$	$y = 0$	$y = 1$	$y = 2$	$y = 3$
$P(Y = y)$	$\frac{\binom{3}{0}\binom{7}{3}}{\binom{10}{3}} = \frac{35}{120}$	$\frac{\binom{3}{1}\binom{7}{2}}{\binom{10}{3}} = \frac{63}{120}$	$\frac{\binom{3}{2}\binom{7}{1}}{\binom{10}{3}} = \frac{21}{100}$	$\frac{\binom{3}{3}\binom{7}{0}}{\binom{10}{3}} = \frac{1}{120}$

$P(y) = \frac{\binom{3}{y}\binom{7}{3-y}}{\binom{10}{3}}$ when $y = 0, 1, 2, 3$

2. What is the expected profit for this game?

(sum × prob)

1.65

3. What is the standard deviation of the profit for this game?

Linear Transformations of Random Variables

Sometimes a random variable - let's call it X — must undergo a linear transformation. A ***linear transformation*** of X can be achieved by doing one or more of the following:

- Adding or subtracting a constant to X
- Multiplying or dividing X by a constant

Note: Operations like X^2 or $\frac{1}{X}$ are not linear transformations!

When we perform a linear transformation of a random variable X, we can consider the result as a new random variable: let's call it Y. Assuming we know $E(X)$, we can easily calculate $E(Y)$ by knowing a few rules. The same can be said about the variance and standard deviations—if we know $VAR(X)$ (and thus $SD(X)$), we can calculate $VAR(Y)$ (and $SD(Y)$).

$E[x] = \sum x_i p(x_i) = \mu$

$E[x^2] = \sum x_i^2 p(x_i)$

$E(x)^2 \neq E(x^2)$

$V(x) = E(x^2) - E(x)^2 = \sigma^2$ (sigma)

$SD(x) = \sqrt{E(x^2) - E(x)^2} = \sigma$

Rules for Expected Values and Variances of Linear Transformations

Let X be a random variable with expected value $E(X)$ and variance $VAR(X)$ and let a be a constant.

- $E(aX) = E(a)E(X) = aE(X)$
- $E(X + a) = E(X) + E(a) = E(X) + a$ (same process holds for subtraction of *a*)
- $VAR(aX) = a^2VAR(X)$
- $VAR(X + a) = VAR(X) + VAR(a) = VAR(X) + 0$ (same process holds for subtraction of *a*)

In general, note that:

- $E(a) = a$ [the expected value of a constant is the constant itself]
- $VAR(a) = 0$ [the variance of a constant is 0]

The above rules can be combined for more complicated linear transformations.

Example 2.1.5

Let Y be a discrete random variable with $E(Y) = 4$ and $VAR(Y) = 25$.

1. Define $U = 6Y$. Find $E(U)$.

$E[6Y] = 6E[Y] = 6(4) = 24$

2. Define $V = Y - 5$. Find $E(V)$ and $VAR(V)$.

$E[Y - 5] = E[Y] - 5 = 4 - 5 = -1$

$VAR[Y - 5] = VAR[Y] + VAR[5] = 25 + 0 = 25$

3. Define $W = -3Y + 8$. Find $SD(W)$.

$E[-3Y + 8] = -3E[Y] + 8 = -3(4) + 8 = -12 + 8 = -4$

$V[-3Y + 8] = (-3)^2 V[Y] + V[8]$

$= 225$

$\sqrt{V(-3Y+8)} = \sqrt{225} = 15 \leftarrow SD(W)$

In some situations, linear transformations of a given random variable are “hidden” within a word problem.

Example 2.1.6

A salesman at a mall kiosk interacts with, at most, three customers per day. Let W represent the number of customers to which the salesman successfully sells his product in a single day. The following probability distribution table shows the probabilities for the values of W.

W	0	1	2	3
P(W = w)	0.54	0.3	0.12	0.04

Suppose that the salesman earns $100 for each successful sale. However, he must also pay a fixed out-of-pocket expense of $25 per day in order to run the kiosk. What is the salesman’s expected daily profit?