

# Unit 20: Random Variables



## PREREQUISITES

Students should be familiar with the background on probability covered in Unit 18, Introduction to Probability, and Unit 19, Probability Models. They should also be familiar with the material on the normal distribution contained in Units 7 and 8, Normal Curves and Normal Calculations, respectively.

## ADDITIONAL TOPIC COVERAGE

Additional coverage of probability can be found in *The Basic Practice of Statistics*, Chapter 10, Introducing Probability. Unit 21, Binomial Distributions, continues the discussion on distributions of discrete random variables.

## ACTIVITY DESCRIPTION

This activity is based on data from the March Supplement Survey 2012, part of the Current Population Survey (CPS), which is sponsored jointly by the U.S. Census Bureau and by the U.S. Bureau of Labor Statistics. The series of questions that serve as the basis for this activity focuses on households. Keep in mind that a household can consist of more than one family. The nested sequence of survey questions is as follows:

Any children in household?

    If yes, any eat school lunch?

        If yes, any get lunch free?

            If yes, how many get lunch free?

The answer to each question can be viewed as a random phenomenon. Hence, in the activity, each question generates a random variable.

The first question is a yes or no question and hence, the random variable associated with it has two possible values. We assume that the random variable is a model for what is true in the

population of all U.S. households. Therefore, the mean and standard deviation of the random variable are considered to be population characteristics. In question 1, students use a column approach to calculate the mean and variance. This column approach is particularly useful in question 4 when students have to find the mean and variance of a random variable that has 8 possible values. In addition, this column approach to finding the mean and variance can be easily adapted to spreadsheets.

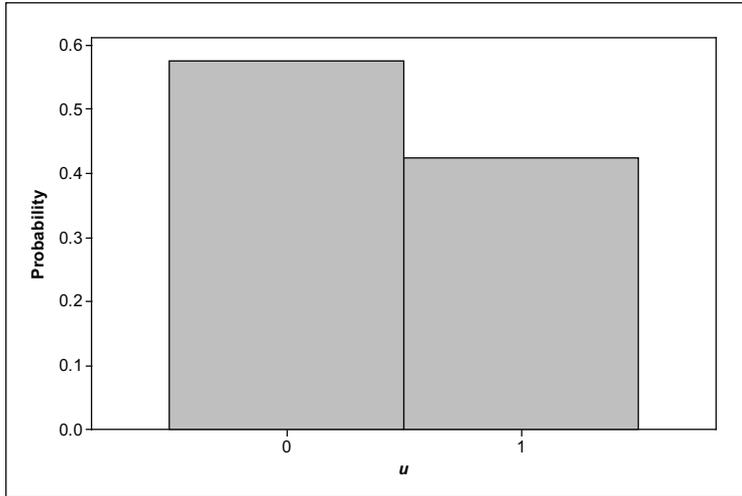
Question 2 provides the raw data from the 2012 March Supplement Survey and asks students to use those data to estimate the probabilities. In 2(b) students are asked to create an area model for the first two random variables. That model allows them to view the nesting of the questions. Check that students understand that to find a proportion of a proportion, you multiply proportions.

# THE VIDEO SOLUTIONS

1. Random variable  $x$  can take on the values 0, 1, 2, 3, or 4.
2. The probabilities add up to 1.
3. The most likely number of heads is 2.
4. Failure of at least one of the O-rings.
5. Multiplication Rule.
6.  $P(\text{at least one field joint failed}) = 1 - p(0)$ .

# UNIT ACTIVITY SOLUTIONS

1. a.



b. It is less than 0.5. The balance point would be at 0.5 if both bars in the probability histogram were equal in height. Since the bar over  $u = 0$  is taller than the bar over  $u = 1$ , the balance point must be less than where the two bars meet, at 0.5.

c.

$u$	$p(u)$	$(u)(p(u))$
0	0.586	0
1	0.432	0.432
Sum =		0.432

d.

$u$	$p(u)$	$(u - 0.432)^2$	$((u - 0.432)^2)(p(u))$
0	0.586	0.1866	0.1094
1	0.432	0.3226	0.1394
Sum =			0.2487

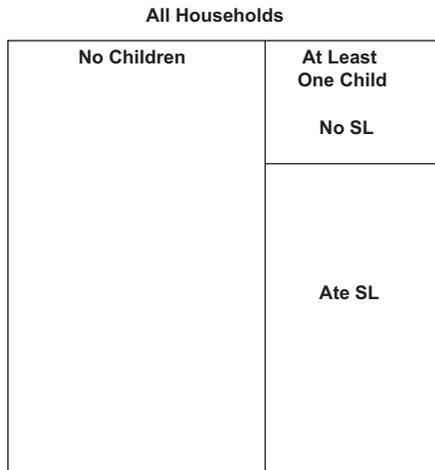
$$\sigma_u^2 \approx 0.2487 ; \sigma_u \approx 0.4987$$

2. a.

$v$	Frequency	Probability
0	32,491	0.303
1	74,690	0.697
Total	107,181	1

b.  $(0.432)(0.697) \approx 0.301$

Sample diagram:

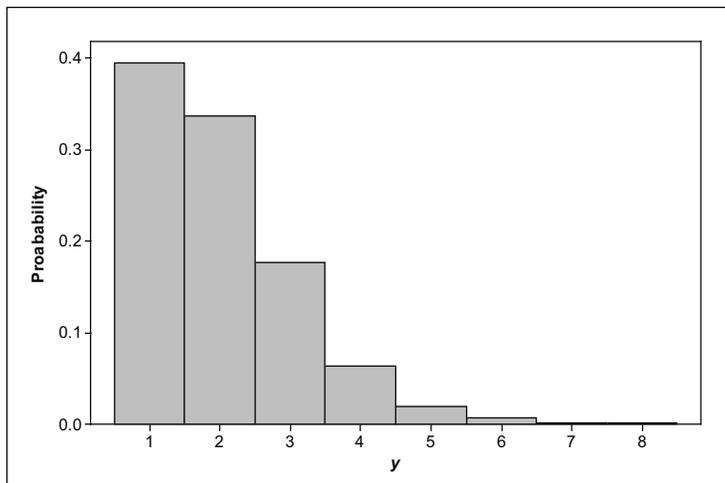


3. We would expect the proportion:

$(\text{number households participating in free lunch program}) / 74,690$

to be approximately 0.568. Hence, we get  $(0.568)(74,690) = 42,424$  households where at least one child participates in the free lunch program.

4. a.



b. Sample answer: Around 2.

c.  $\mu_y \approx 2.007$

y	p(y)	(y)(p(y))
1	0.394	0.394
2	0.337	0.674
3	0.176	0.528
4	0.063	0.252
5	0.019	0.095
6	0.007	0.042
7	0.002	0.014
8	0.001	0.008
Sum =		2.007

d.

y	p(y)	(y - 2.007)^2	((y - 2.007)^2)(p(y))
1	0.394	1.0140	0.3995
2	0.337	0.0000	0.0000
3	0.176	0.9860	0.1735
4	0.063	3.9720	0.2502
5	0.019	8.9580	0.1702
6	0.007	15.9440	0.1116
7	0.002	24.9300	0.0499
8	0.001	35.9160	0.0359
Sum =			1.1909

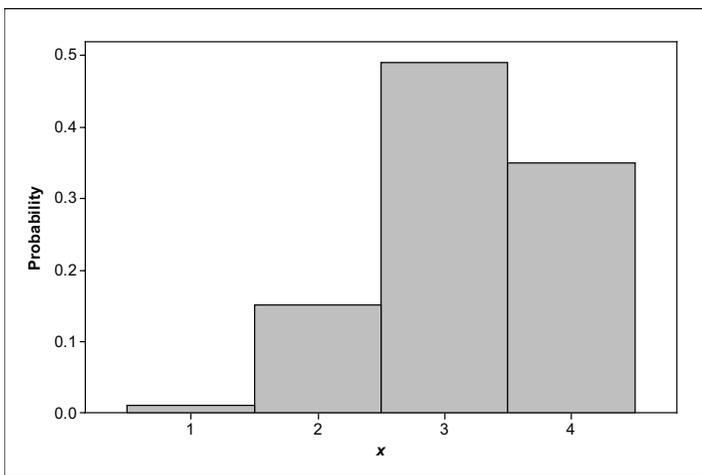
$\sigma_y^2 \approx 1.1909$  ;  $\sigma_y \approx 1.091$

# EXERCISE SOLUTIONS

1. a.  $P(x \geq 3) = P(x = 3 \text{ or } x = 4) = p(3) + p(4) = 0.49 + 0.35 = 0.84$  .

b.  $P(x < 3) = 1 - P(x \geq 3) = 1 - 0.84 = 0.16$  The events “below B” and “B or above” are complementary events. Therefore the two events sum to 1.

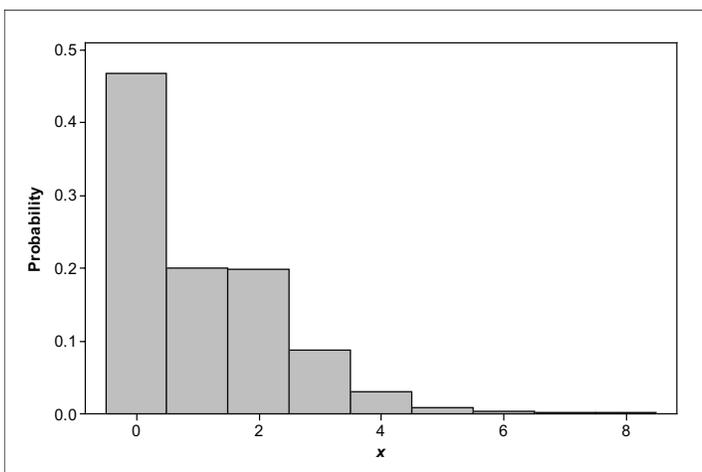
c. The histogram below shows that if a 12th grade student is randomly selected, the most likely outcome is  $x = 3$ , and the next most likely is  $x = 4$ . Not surprisingly, the least likely outcome is  $x = 1$ ; students with D averages probably drop out of high school before completing the 12th grade.



2.a.  $P(x \geq 1) = 1 - P(x < 1) = 1 - p(0) = 0.532$

b.  $P(2 \leq x \leq 4) = P(x = 2 \text{ or } x = 3 \text{ or } x = 4) = p(2) + p(3) + p(4)$   
 $= 0.199 + 0.087 + 0.031 = 0.317$

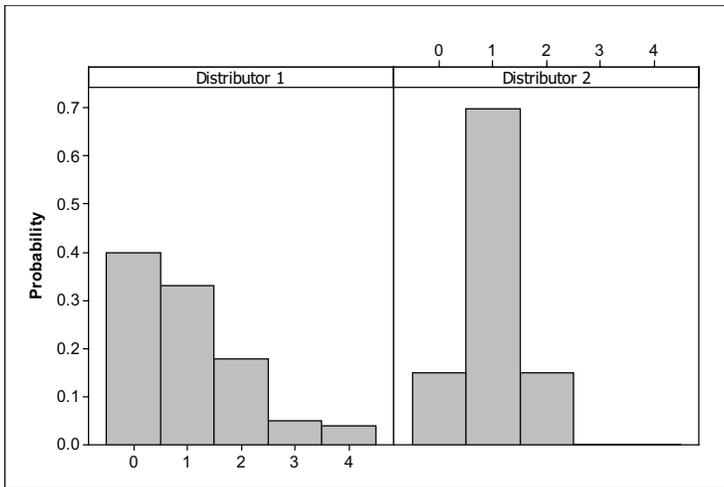
c. The histogram is skewed to the right.



d.  $\mu = 1.068$  children under 15 per household. The calculations follow:

$$0 \times 0.468 + 1 \times 0.200 + 2 \times 0.199 + 3 \times 0.087 + 4 \times 0.031 + 5 \times 0.009 + 6 \times 0.003 + 7 \times 0.002 + 8 \times 0.001 = 1.068 \text{ child per household.}$$

3. a. The histogram for Distributor 1 is skewed to the right and the histogram for Distributor 2 is symmetric.



b. Using the approach outlined in the unit activity, we get  $\mu_x = 1$  and  $\mu_y = 1$ .

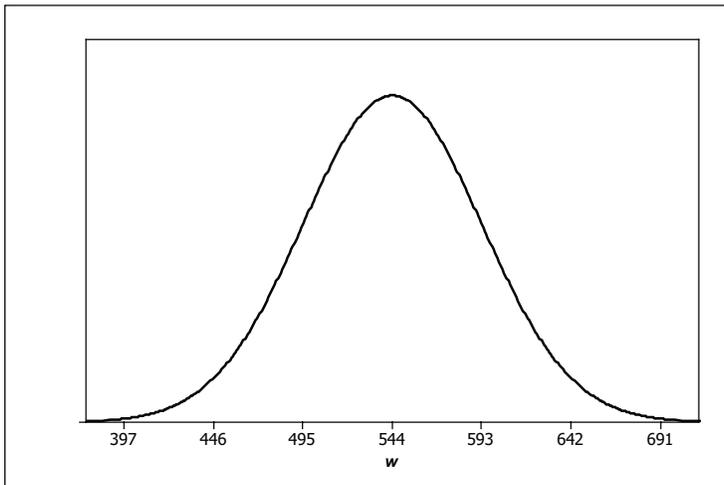
x or y	$p(x)$	$p(y)$	$(x)(p(x))$	$(y)(p(y))$
0	0.40	0.15	0	0
1	0.33	0.70	0.33	0.7
2	0.18	0.15	0.36	0.3
3	0.05	0	0.15	0
4	0.04	0	0.16	0
Sum =			1	1

c. Using the approach outlined in the unit activity, we get  $\sigma_x^2 = 1.14$  and  $\sigma_y^2 = 0.3$ . Therefore,  $\sigma_x \approx 1.07$  and  $\sigma_y \approx 0.55$ . (Calculations follow.)

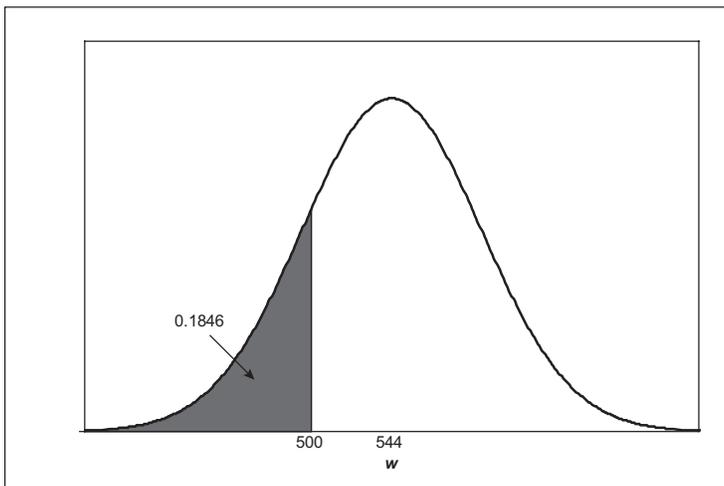
x or y	$p(x)$	$p(y)$	$((x-1)^2)(p(x))$	$((y-1)^2)(p(y))$
0	0.40	0.15	0.40	0.15
1	0.33	0.70	0	0
2	0.18	0.15	0.18	0.15
3	0.05	0	0.20	0
4	0.04	0	0.36	0
Sum =			1.14	0.3

d. For both distributors, the mean number of defects in lots of four is 1. Hence, on average, over many, many lots, there will be one defect out of four, regardless of which distributor is used. However, the standard deviations show that for Distributor 2 the number of defects is more concentrated about the mean than it is for Distributor 1. Hence, when purchasing from Distributor 1, there will be more variability in a long sequence of lots than when purchasing from Distributor 2. Given the expected number of defects out of four is the same for both distributors, it makes sense to go with the distributor that has the more consistent lots.

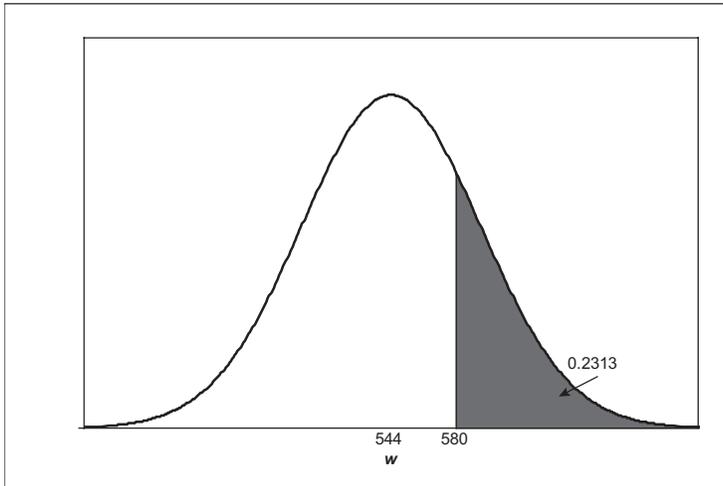
4. a



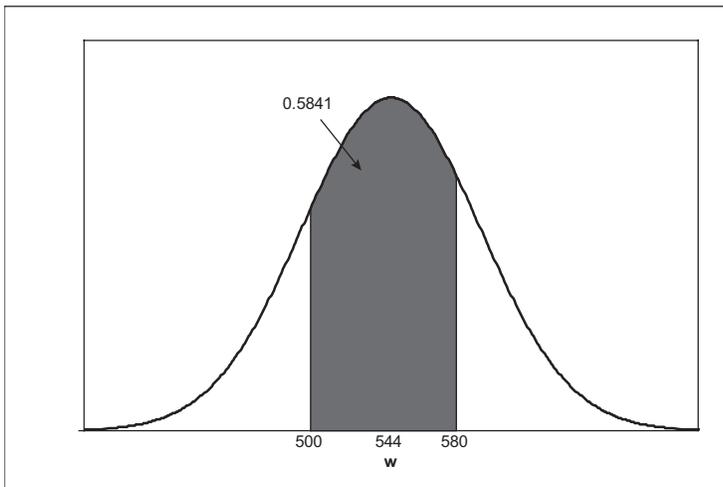
b.  $P(w < 500) = 0.1846$



c.  $P(w \geq 580) = 0.2313$



d.  $P(500 \leq w \leq 580) = 1 - (0.2313 + 0.1846) = 0.5841$



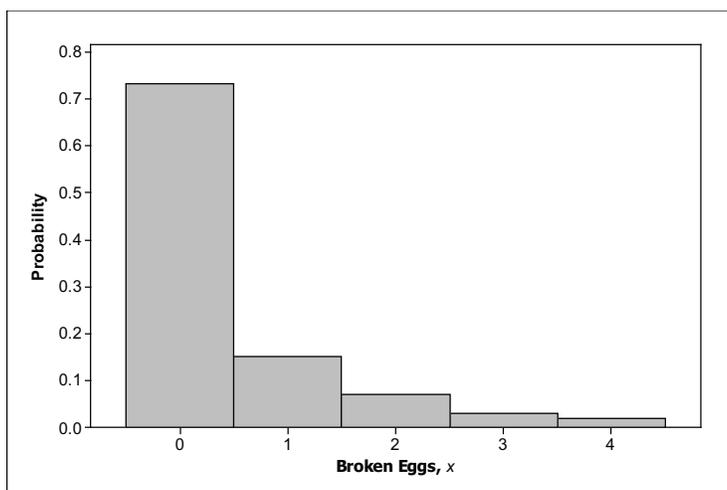
# REVIEW QUESTIONS SOLUTIONS

1. a.  $P(x < 4) = p(0) + p(1) + p(2) + p(3) = 0.73 + 0.15 + 0.07 + 0.03 = 0.98$ .

$p(4) = 1 - 0.98 = 0.02$ . This means that out of many, many egg cartons, roughly 2% will have four broken eggs.

b.  $P(x \geq 2) = p(2) + p(3) + p(4) = 0.07 + 0.03 + 0.02 = 0.12$

c.



d.  $\mu = (0)(0.73) + (1)(0.15) + (2)(0.07) + (3)(0.03) + (4)(0.02) = 0.46$ . This means that in the long run, after opening many, many cartons, the average number of broken eggs is 0.46. It is also the balance point for the probability histogram.

2. a. This is an example of a discrete random variable. There will be a finite number of Cheerios in a box. The weight limit of the box adds a cap to the number of Cheerios that will fit in a box.

b. This is an example of a continuous random variable. Time to finish takes values in an interval.

c. This is discrete. Possible amounts are separated by pennies. You could put the possible amounts in a list.

Comment: In some situations, money is treated as a continuous random variable. For example, this is true for determining the amount of money in a savings account that is accruing interest. Of course, the bank always rounds down to the nearest penny when paying out money.

d. Length is an example of a continuous random variable that can take values in an interval.

3. a.  $S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}$

b.

$x$	0	1	2	3
$p(x)$	0.125	0.375	0.375	0.125

$$\mu = (0)(0.125) + (1)(0.375) + (2)(0.375) + (3)(0.125) = 1.5$$

$$\sigma^2 = (0 - 1.5)^2(0.125) + (1 - 1.5)^2(0.375) + (2 - 1.5)^2(0.375) + (3 - 1.5)^2(0.125) = 0.75$$

$$\sigma = \sqrt{0.75} \approx 0.866$$

c.

$y$	1	3
$p(y)$	0.75	0.25

$$\mu = (1)(0.75) + (3)(0.25) = 1.5$$

$$\sigma^2 = (1 - 1.5)^2(0.75) + (3 - 1.5)^2(0.25) = 0.75 ; \sigma = \sqrt{0.75} \approx 0.866$$

d. In this case, the value of  $w$  is always 3.

$w$	3
$p(w)$	1

$$\mu = (3)(1) = 3$$

$$\sigma^2 = (3 - 3)^2(1) = 0 ; \sigma = 0$$

4. a.  $P(w < 68) \approx 0.2525$

b.  $P(w \geq 75) \approx 0.0478$

c.  $P(68 < w < 75) \approx 0.6997$