Chapter 3 - Probability

Zachary Palmore

Dice rolls. (3.6, p. 92) If you roll a pair of fair dice, what is the probability of

(a) getting a sum of 1?

Zero, since they are independent events and the lowest value possible to roll per die is 1. The lowest sum you could roll is 2 with both die equal to 1.

(b) getting a sum of 5?

There are a few options to roll a sum of 5 using two dice. They are:

- 1+4 = 5
- 4+1 = 5
- 2+3 = 5
- 3+2 = 5

Since all of these possibilities equal the sum of 5 we can say there are 4 possible outcomes where the sum of the two separate die will equal 5.

```
sumsof5 <- 4
```

To express the chances of any of these possible outcomes, we first find the total number of all possible outcomes.

For example, one die can roll any whole number 1 - 6, meaning it has 6 possible outcomes. They are 1, 2, 3... and so on to 6. The other die has the same number of possible outcomes. Our total number of possible outcomes with both die being rolled at the same time is found through the product of each die's total possible outcomes.

6 * 6 = 36 outcomes

```
# Outcomes of 1 die
die_outcomes <- 6
#Total outcomes
totaloutcomes <- die_outcomes*die_outcomes
```

To find the probability of rolling the two die such that the sum of its values equals 5 we can express it as a fraction 4/36.

```
# This is equal to 4/36
prob_sumof5 <- sumsof5 / totaloutcomes
```

Expressed as a percentage, there is an 11.11% chance of rolling two die such that the sum of its values equals 5.

```
percent_probsumof5 <- prob_sumof5*100
```

(c) getting a sum of 12?

The only outcome possible when rolling two die such that the sum of their values equals 12 is by rolling 6 on both die at the same time.

- 6+6 = 12

Borrowing from the total outcomes in the sum of 5 probability where the total outcomes possible is 36 we can express the result as a fraction 1/36 because there is only one possible outcome of the total outcomes where the sum of the die equals 12 on a pair of fair, six-sided die.

```
sumof12 <- 1
prob_sumof12 <- sumof12 / totaloutcomes
prob_sumof12
```

```
## [1] 0.02777778
```

Expressed as a percentage, there is about a 2.78% chance of the sum of the die being equal to 12.

```
percent_probsumof12 <- prob_sumof12*100
percent_probsumof12
```

```
## [1] 2.777778
```

Poverty and language. (3.8, p. 93) The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 20.7% speak a language other than English (foreign language) at home, and 4.2% fall into both categories.

(a) Are living below the poverty line and speaking a foreign language at home disjoint?

No, living below the poverty line and speaking a foreign language are not mutally exclusive. There can be (and are) people who speak a foreign language and are in poverty and there are also people who speak a foreign language who are not below the poverty line.

(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

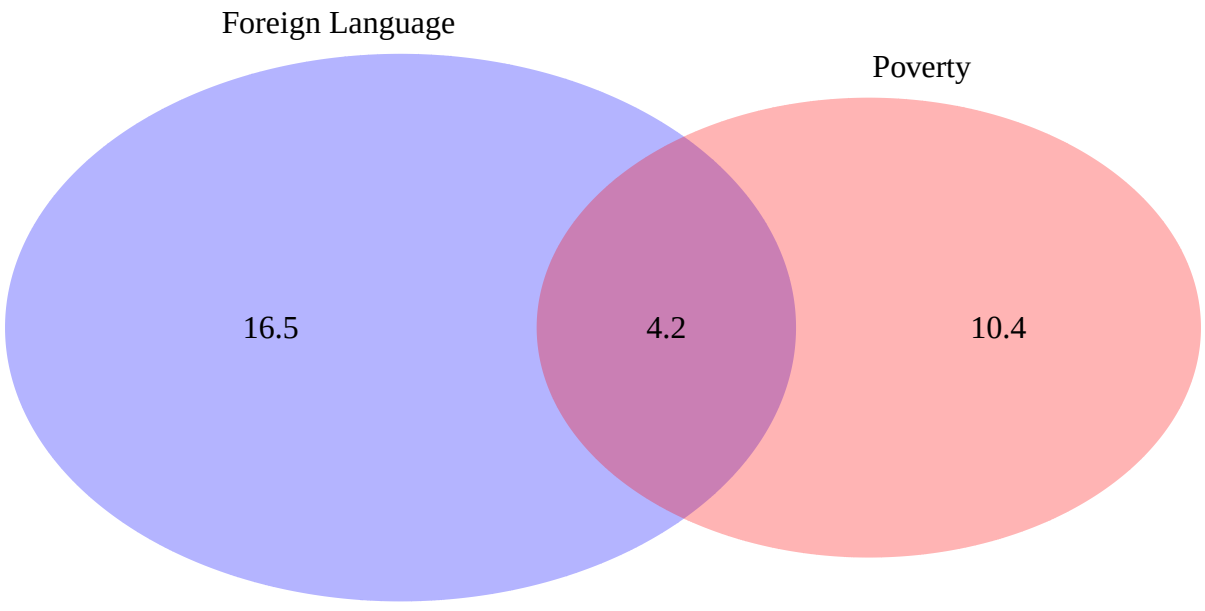
Poverty at 14.6% in the circle on the right, 4.2% in the overlapping portion of the circles, and 20.7% in the foreign language circle.

```
library(VennDiagram)
```

```
## Loading required package: grid
```

```
## Loading required package: futile.logger
```

```
grid.newpage()
draw.pairwise.venn(20.7, 14.6, 4.2, category = c("Foreign Language", "Poverty"), lty = rep("blank", 2),
```



```
## (polygon[GRID.polygon.1], polygon[GRID.polygon.2], polygon[GRID.polygon.3], polygon[GRID.polygon.4],
```

(c) What percent of Americans live below the poverty line and only speak English at home?

This can be calculated using the percent of Americans that live below the poverty line and subtracting from it those who live below the poverty that also speak a foreign language at home.

```
Poverty <- 14.6
ForeignLan <- 20.7
Both <- 4.2
# We assume here that if they are not speaking a language other than
# English at home, they are only speaking English at home.
Eng_Poverty <- (Poverty - Both)
Eng_Poverty
```

```
## [1] 10.4
```

This indicates that of the 14.6% of Americans living below the poverty line, 10.4% of them speak only English at home.

(d) What percent of Americans live below the poverty line or speak a foreign language at home?

The percentage of Americans that live below the poverty line is given at 14.6%. The percentage of Americans that speak a foreign language at home is also given at 20.7% and those that fit into both categories is 4.2%.

To find the percentage of Americans that live below the poverty line or speak a foreign language at home (but not both) we sum the percentage below the poverty line with the percentage that speak a foreign language and subtract those who fall in both categories.

```
Poverty_OR_ForeignLang <- (sum(Poverty, ForeignLan) - Both)
Poverty_OR_ForeignLang
```

```
## [1] 31.1
```

The percentage of American that live below the poverty line or speak a foreign language at home is 31.1%.

(e) What percent of Americans live above the poverty line and only speak English at home?

Using the rule that probabilities must total to 1 (or 100%) for probability distributions, we can find the remaining English speakers above the poverty line from the calculated number of English speakers who live below the poverty line.

Given that 14.6% of people live below the poverty line, we subtract it from 100.

```
100 - 14.6
```

```
## [1] 85.4
```

This indicates that 85.4% of Americans are living above the poverty line. Now we find the percentage of people that are above the poverty line that also speak another language. We are given the percentage of Americans that speak a language other than English at home as 20.7%. We are also given those who live below the poverty line and speak a language other than English at home as 4.2% of Americans. To find those who are not in poverty but speak a foreign language at home, we calculate the difference.

```r
20.7 - 4.2
```

```
## [1] 16.5
```

This tells us 16.5% of Americans are not in poverty but speak another language. In the last step, we subtract this from the total number of people above the poverty line to get the people above the poverty line that only speak English at home.

```r
85.4 - 16.5
```

```
## [1] 68.9
```

Therefore, 68.9% of Americans live above the poverty line and only speak English at home.

(f) Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

Based on the multiplication rule they are not independent. This can be shown where the product of the probability of being below the poverty line and the probability of speaking a foreign language at home, do not equal the probability of the two together. In this case, the probability of the two together was given as 4.2%.

```r
Multiplication_Test <- (Eng_Poverty * (ForeignLan - 4.2)/100)
Multiplication_Test
```

```
## [1] 1.716
```

Since the given probability of .042 does not equal 1.716 the two cannot be independent events.

Assortative mating. (3.18, p. 111) Assortative mating is a nonrandom mating pattern where individuals with similar genotypes and/or phenotypes mate with one another more frequently than what would be expected under a random mating pattern. Researchers studying this topic collected data on eye colors of 204 Scandinavian men and their female partners. The table below summarizes the results. For simplicity, we only include heterosexual relationships in this exercise.

		Partner (female)			
		Blue	Brown	Green	Total
Self (male)	Blue	78	23	13	114
	Brown	19	23	12	54
	Green	11	9	16	36
	Total	108	55	41	204

(a) What is the probability that a randomly chosen male respondent or his partner has blue eyes?

The probability that a randomly chosen male respondent or his partner will have blue eyes is 0.54.

```
# male respondents with blue eyes is 114
male_blue <- 114
# female partners with blue eyes is 78 + 19 + 11
female_blue <- sum(78,19,11)
# Total (male) respondents
total_responses <- 204*2
P_blueeyes <- (female_blue + male_blue) / total_responses
P_blueeyes
```

```
## [1] 0.5441176
```

This means there is a 54.4% chance that a random respondent or his partner will have blue eyes.

(b) What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?

There is a 0.684 probability that a randomly selected male respondent with blue eyes has a partner with blue eyes.

```
# P male with blue eyes has a partner with blue eyes
# female partners with blue eyes is 78
# total male respondents with blue eyes is 114?
78/114
```

```
## [1] 0.6842105
```

In other words, there is a 68.4% chance that a blue eyed male has a blue eyed partner.

(c) What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

The probability of a randomly selected male with brown eyes having a blue eyed partner is 0.352 or a 35.2% chance.

```
# P male with brown eyes having a partner with blue eyes
# female partners with blue eyes is 19
# total male respondents with brown eyes is 54?
19/54
```

```
## [1] 0.3518519
```

The probability of a randomly selected male with brown eyes having a blue eyed partner is 0.306 or a 30.6% chance.

```
# P male with green eyes having a partner with blue eyes
# female partners with green eyes is 11
# total male respondents with green eyes is 36?
11/36
```

```
## [1] 0.3055556
```

(d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

Using the multiplication rule, we can set the variable A as male eye color and B as female eye color.

```
(114/204) * (108/204)
```

```
## [1] 0.2958478
```

```
78/204
```

```
## [1] 0.3823529
```

For another color, green:

```
(41/204)*(36/204)
```

```
## [1] 0.03546713
```

```
16/204
```

```
## [1] 0.07843137
```

For another color, brown:

```
(55/204) * (54/204)
```

```
## [1] 0.07136678
```

```
(23/204)
```

```
## [1] 0.1127451
```

It appears these events are not independent.

Books on a bookshelf. (3.26, p. 114) The table below shows the distribution of books on a bookcase based on whether they are nonfiction or fiction and hardcover or paperback.

		Format		
		Hardcover	Paperback	Total
Type	Fiction	13	59	72
	Nonfiction	15	8	23
	Total	28	67	95

(a) Find the probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacement.

Given that the total number of books is 95 and the total number of hardcover books is 28, the probability of drawing a hardcover book is 0.295.

```
28/95
```

```
## [1] 0.2947368
```

Without replacing the hardcover book, the probability of drawing a paperback fiction book is 59 (total fiction books) over the remaining total number of books at 94 (95 - 1).

```
59/94
```

```
## [1] 0.6276596
```

The probability of drawing a paperback fiction without replacement is 0.628.

Both events happening together creates a new probability that we can determine using the general multiplication rule.

```
(28/95)*(59/95)
```

```
## [1] 0.1830471
```

The probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacemen is 0.183.

(b) Determine the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement.

The total number of fiction books is (13+59) 72. The total number of books is 95.

```
72/95
```

```
## [1] 0.7578947
```

The probability of drawing a fiction book is 0.758. Without replacing the book, we find the odds of drawing a hardcover book.

The new total number of books is 94. Assuming that the previous fiction book was a paperback, the odds of pulling a hardcover next is 28/94.

```
28/94
```

```
## [1] 0.2978723
```

That gives us a 0.298 probability of drawing a hardcover book. Now, if the previous book was a hardcover book, then the odds change to 27/94 because one hardcover book has already been removed without replacement.

```
27/94
```

```
## [1] 0.287234
```

That gives us a different probability of 0.287 depending on the first draw.

Using the first possibility, that a paperback is drawn instead of a hardcover, then the total probability is the product of the two probabilities as stated in the general multiplication rule.

```
(72/95)*(28/94)
```

```
## [1] 0.2257559
```

Thus, the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement is 0.226.

(c) Calculate the probability of the scenario in part (b), except this time complete the calculations under the scenario where the first book is placed back on the bookcase before randomly drawing the second book.

Using the probabilities from part b,

```
(28/95)*(72/95)
```

```
## [1] 0.2233795
```

there is a 0.223 probability that both events occur with replacement.

(d) The final answers to parts (b) and (c) are very similar. Explain why this is the case.

The final answers to parts (b) and (c) are very similar because the sample size is large enough that a reduction in the sample size by 1, did not have a large impact. The larger the sample size, the less variation in probability when measured with or without replacement.

Baggage fees. (3.34, p. 124) An airline charges the following baggage fees: $25 for the first bag and $35 for the second. Suppose 54% of passengers have no checked luggage, 34% have one piece of checked luggage and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

(a) Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

```
carryon_cost <- 0
check1_cost <- 25
check2_cost <- 35
checked_fees <- c(carryon_cost, check1_cost, check2_cost)
# convert the percents to decimals prior to using them in the calculation
passenger_percent <- c(0.54, 0.34, 0.12)
avg_per_person <- sum(checked_fees * passenger_percent)
avg_per_person
```

```
## [1] 12.7
```

Now to find the standard deviation.

```
# Takes the square root of the variance function
# Uses average revenue per person as the mean
std_deviation <- sqrt(0.54*(0-avg_per_person)^2 + 0.34*(25-avg_per_person)^2 + 0.12*(35-avg_per_person)
```

The standard deviation is 14.08.

(b) About how much revenue should the airline expect for a flight of 120 passengers? With what standard deviation? Note any assumptions you make and if you think they are justified.

Given those same percentages of passengers that will and will not check their bags, the average revenue per person the airline should expect is $1,524 (120*12.7).

```
avg_revnue_120 <- 120*12.7
```

To see this another way, we can find the number of passengers in each category.

```
no_check <- 120*.54
one_check <- 120*.34
two_check  <- 120*.12
no_check
```

```
## [1] 64.8
```

```
one_check
```

```
## [1] 40.8
```

```
two_check
```

```
## [1] 14.4
```

There would be 64 passengers who did not check any bags, 40 who checked one bag, and 14 who checked two bags.

```
revenue_no_check <- (no_check * carryon_cost)
revenue_one_check <- (one_check * check1_cost)
revenue_two_check <- (two_check * check2_cost)
total_revenue_at120 <- sum(revenue_no_check, revenue_one_check, revenue_two_check)
total_revenue_at120
```

```
## [1] 1524
```

Here again, the total revenue is 1524 and the standard deviation is calculated by first finding the variance.

```
variance <- (std_deviation^2)
variance
```

```
## [1] 198.21
```

The variance is 198.21. With this, we can multiply the number of passengers by it then take the square root to find the standard deviation with 120 passengers.

```
sqrt(variance*120)
```

```
## [1] 154.2245
```

The standard deviation with 120 passengers is 154.22.

Income and gender. (3.38, p. 128) The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised of 59% males and 41% females.

Income	*Total*
$1 to $9,999 or loss	2.2%
$10,000 to $14,999	4.7%
$15,000 to $24,999	15.8%
$25,000 to $34,999	18.3%
$35,000 to $49,999	21.2%
$50,000 to $64,999	13.9%
$65,000 to $74,999	5.8%
$75,000 to $99,999	8.4%
$100,000 or more	9.7%

(a) Describe the distribution of total personal income.

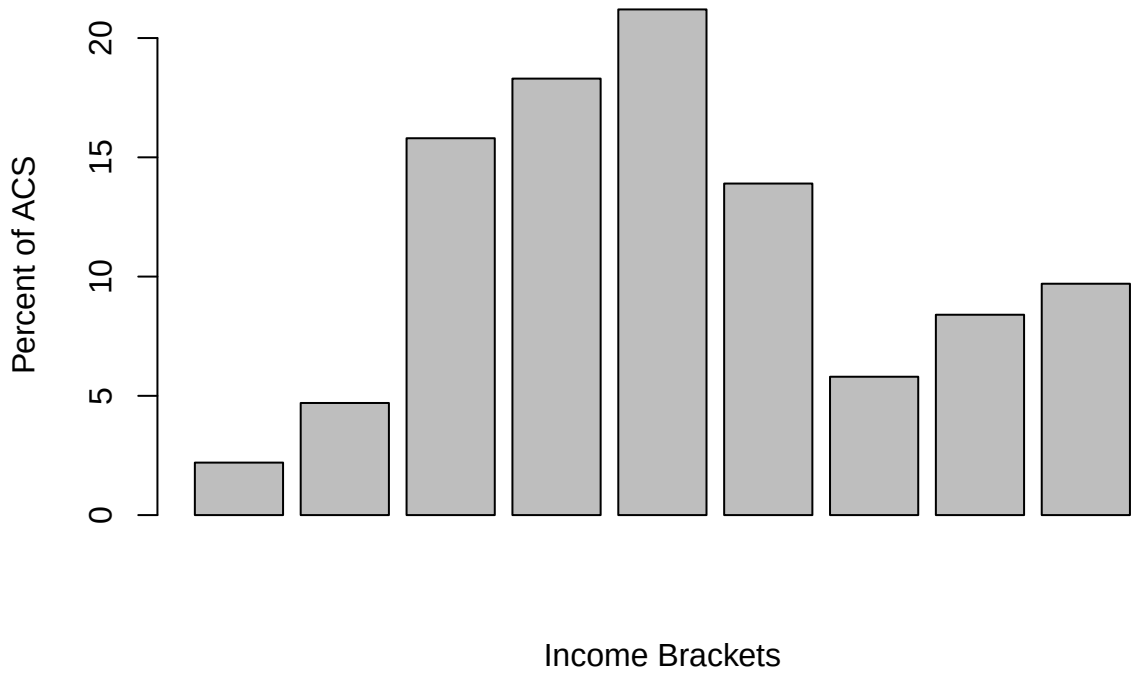
To visualize the distribution, I will create a histogram with the frequency values of each income level.

```
frequencies <- c(2.2,4.7,15.8,18.3,21.2,13.9,5.8,8.4,9.7)
barplot(frequencies, xlab = "Income Brackets", ylab = "Percent of ACS", title = "U.S. Income")
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): "title" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "title"
## is not a graphical parameter

## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): "title" is not a
## graphical parameter
```



This is a normal distribution with slightly higher frequencies in the highest two income ranges.

(b) What is the probability that a randomly chosen US resident makes less than $50,000 per year?

Let's first create a data frame to organize the data.

```
income_cat <- c("$1 to $9,999","$10,000 to $14,999","$15,000 to $24,999","$25,000 to $34,999","$35,000
USincome <- data.frame(income_cat, frequencies)
USincome
```

```
##           income_cat frequencies
## 1       $1 to $9,999         2.2
## 2 $10,000 to $14,999         4.7
## 3 $15,000 to $24,999        15.8
## 4 $25,000 to $34,999        18.3
## 5 $35,000 to $49,999        21.2
## 6 $50,000 to $64,999        13.9
## 7 $65,000 to $74,999         5.8
## 8 $75,000 to $99,999         8.4
## 9   $100,000 or more         9.7
```

Then sum the first 3 rows of the second column to find the total frequency of people making up to 50,000 per year.

```
lessthan50k <- sum(USincome[1:5, 2])
lessthan50k
```

```
## [1] 62.2
```

This indicates there is about a 62.2% chance that a randomly chosen US resident will makes less than $50,000 per year. In decimal form, that is a 0.622 probability.

(c) What is the probability that a randomly chosen US resident makes less than $50,000 per year and is female? Note any assumptions you make.

Given that the sample is 41% female we can use it to estimate the probability of females making less than 50,000 per year knowing the probability that any resident will make less than 50,000 per year.

```
female_lessthan50k <- (lessthan50k/100) * (41/100)
female_lessthan50k
```

```
## [1] 0.25502
```

Here, an assumption was made that the likelihood of males and females each individually making 50,000 per year are independent events. In other words, that they each have the same probability to make 50,000 per year.

(d) The same data source indicates that 71.8% of females make less than $50,000 per year. Use this value to determine whether or not the assumption you made in part (c) is valid.

This assumption must not be valid because the probability of females making less than 50,000 per year is greater than the probability of males and females having the same probability.

Using the general rule of multiplication and the assumption we already made (the probability of females making less than 50,000 per year) we can see why.

```
female_lessthan50k
```

```
## [1] 0.25502
```

```
real_female_lessthan50k <- (71.8/100)
real_female_lessthan50k
```

```
## [1] 0.718
```

Since these probabilities are not equal, they are not independent events.
