Chapter 3 - Probability

```
library(VennDiagram)
```

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

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

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?

The probability of getting a sum of 1 is 0, or 0/36, since it is not possible to obtain a sum of 1 if a pair of fair dice are rolled.

(b) getting a sum of 5?

The probability of getting a sum of 5 is 4/36, or 1/9.

(c) getting a sum of 12?

The probability of getting a sum of 12 is 1/36.

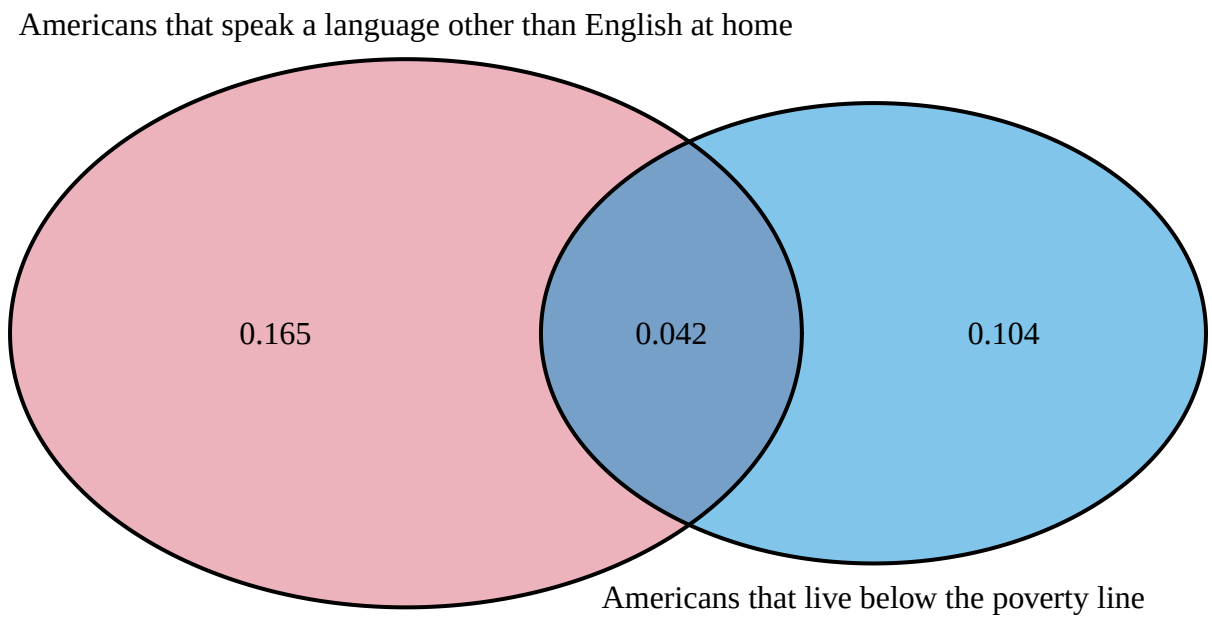
Poverty and language. (3.8, p. 93) The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 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 at home are not disjoint, since it is possible to fall into both categories. There are people that both live below the poverty line and speak a language other than English at home.

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

```
venn.plot <- draw.pairwise.venn(
  area1 = 0.146,
  area2 = 0.207,
  cross.area = 0.042,
  category = c("Americans that live below the poverty line",
               "Americans that speak a language other than English at home"),
  fill = c("#008DD5", "#DA667B"),
  cat.pos = c(0, 180)
)
```



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

$PL = Americans\ living\ below\ poverty\ line$
$FL = Americans\ who\ speak\ a\ foreign\ language\ at\ home$

$P(PL\ or\ FL) = P(PL) - P(PL\ and\ FL)$
$P(PL\ and\ notFL) = P(0.146) - P(0.042)$
$P(PL\ and\ notFL) = \mathbf{0.104}$

The percent of Americans that live below the poverty line and only speak English at home is 10.4%.

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

By using the General Addition Rule: the percent of Americans that live below the poverty line or speak a foreign language at home is:

$PL = Americans\ living\ below\ poverty\ line$
$FL = Americans\ who\ speak\ a\ foreign\ language\ at\ home$

$P(PL\ or\ FL) = P(PL) + P(FL) - P(PL\ and\ FL)$

$P(PL\ or\ FL) = P(0.146) + P(0.207) - P(0.042)$

$P(PL\ or\ FL) = \mathbf{0.311}$

About 31.1% of Americans live below the poverty line or speak a foreign language at home.

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

The percent of Americans that live above the poverty line and only speak English at home is:

$P((PL\ or\ FL)^c) = 1 - 0.311$
$P((PL\ or\ FL)^c) = \mathbf{0.689}$

About 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?

$PL = Americans\ living\ below\ poverty\ line$
$FL = Americans\ who\ speak\ a\ foreign\ language\ at\ home$

$P(PL)xP(FL) = (0.146)x(0.207)$
$P(PL)xP(FL) = \mathbf{0.030}$

This does not equal $P(PL\ and\ FL)$, which is 0.042. Therefore, the events are dependent (not independent from one another).

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

I also created a table of proportions of the data above to help with the following questions:

		Partner (female)			
		Blue	Brown	Green	Total
Self (male)	Blue	0.382	0.113	0.064	0.559
	Brown	0.093	0.113	0.059	0.265
	Green	0.054	0.044	0.078	0.176
	Total	0.529	0.270	0.201	1.000

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

First, we need to calculate the proportions:

$108/204 = 0.529$ **(proportion of males with blue eyes)**
$114/204 = 0.559$ **(proportion of females with blue eyes)**
$78/204 = 0.382$ **(proportion of both partners with blue eyes)**

((0.529) + (0.559)) - (0.382) = 0.706

The probability is roughly 71%, since we had to sum the proportions of respondents who independently have blue eyes, but then subtract the proportion of instances where both partners had blue eyes (to not count them twice).

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

$P(female = blue \mid male = blue) = \frac{78}{114} =$ **0.684**

The conditional probability here is roughly 68% that a randomly chosen male respondent with blue eyes has a partner with blue eyes.

(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?

$P(female = blue \mid male = brown) = \frac{19}{54} =$ **0.352**

$P(female = blue \mid male = green) = \frac{11}{36} = \mathbf{0.305}$

The conditional probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes is roughly 35%.

The conditional probability that a randomly chosen male respondent with green eyes has a partner with blue eyes is roughly 30%.

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

To check whether or not eye colors of male respondents and their partners are independent, we can take a look at the conditional probabilities from above and use the Multiplication Rule for independece processes:

$P(female = blue)xP(male = blue) = P(0.529)xP(0.559) =$ **0.296 (assuming independence)**

However,

$P(female = blue\ and\ male = blue) = \frac{78}{204} = \mathbf{0.382}$

These two values are not equal, and therefore the eye colors of male respondents and their partners are dependent (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.

The probability of drawing a hardcover book first then a paperback fiction book second when drawing without replacement is:

```
calc <- (28 / 95) * (59 / 94)
calc
```

```
## [1] 0.1849944
```

Approximately 18.5%.

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

This answer depends on whether or not the first fiction book selected was hardcover or paperback. See below for both answers.

The probability of drawing a paperback fiction book first then a hardcover book second when drawing without replacement is:

```
calc <- (72 / 95) * (28 / 94)
calc
```

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

Approximately 22.6%.

However, if a hardcover fiction book was selected first then a hardcover book second when drawing without replacement is:

```
calc <- (72 / 95) * (27 / 94)
calc
```

```
## [1] 0.2176932
```

Approximately 21.8%.

(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.

The probability of drawing a paperback fiction book first then a hardcover book second when drawing with replacement is:

```
calc <- (72 / 95) * (28 / 95)
calc
```

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

Approximately 22.3%.

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

The reason parts (b) and (c) are very similar is due to the sample size of only taking one book out at a time - which is under 10% of the population of books on the bookshelf. Therefore, taking a book off the bookshelf and not replacing it back after the first observation makes the observation nearly independent in this case.

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.

I've filled in the table below with computations for this model:

i	1	2	3	Total
x_i	0 USD	25 USD	60 USD	
$P(X = x_i)$	0.54	0.34	0.12	
$x_i \times P(X = x_i)$	0	8.50	7.20	15.70
$x_i - u$	-15.70	9.30	44.30	
$(x_i - u)^2$	246.49	86.49	1962.49	
$(x_i - u)^2 \times P(X = x_i)$	133.10	29.41	235.50	398.01

ANSWER: The average revenue per passenger is about $15.70. The corresponding standard deviation is $19.95.

Calculations for the above table:

```
25 * 0.34
```

```
## [1] 8.5
```

```
60 * 0.12
```

```
## [1] 7.2
```

```
expected_value <- 0 + 8.5 + 7.2
expected_value
```

```
## [1] 15.7
```

```
0 - 15.70
```

```
## [1] -15.7
```

```
25 - 15.70
```

```
## [1] 9.3
```

```
60 - 15.70
```

```
## [1] 44.3
```

```r
(-15.70)^2
```

```
## [1] 246.49
```

```r
(9.30)^2
```

```
## [1] 86.49
```

```r
(44.30)^2
```

```
## [1] 1962.49
```

```r
val1 <- 246.49 * 0.54
val2 <- 86.49 * 0.34
val3 <- 1962.49 * 0.12

variance <- sum(val1, val2, val3)

standard_deviation <- sqrt(variance)
standard_deviation
```

```
## [1] 19.95019
```

(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.

```r
120 * 15.70
```

```
## [1] 1884
```

The airline should expect a revenue of approximately $1,884 for a flight of 120 passengers. The standard deviation will be about $19.95. However, there may be some sampling variability so the actual amount may differ slightly from our approximation.

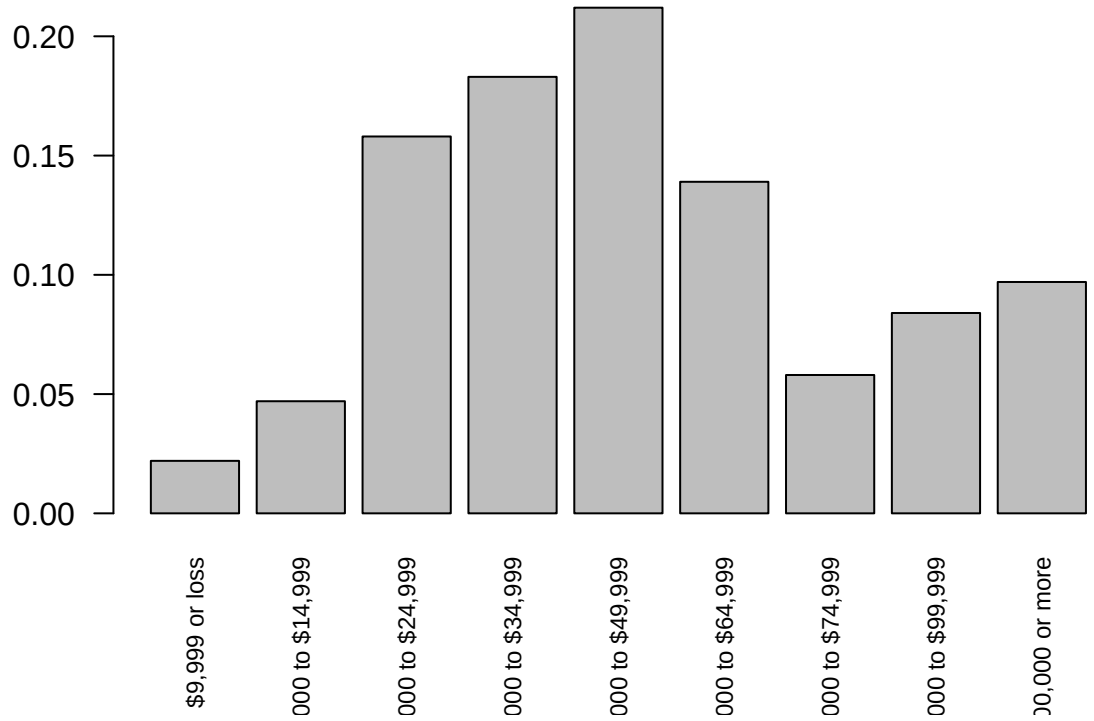
Income and gender. (3.38, p. 128) The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised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

```
percentages <- c(2.2, 4.7, 15.8, 18.3, 21.2, 13.9, 5.8, 8.4, 9.7)
income_brackets <- c("$1 to $9,999 or loss", "$10,000 to $14,999", "$15,000 to $24,999", "$25,000 to $3

table <- data.frame(income_brackets, percentages)
barplot(prop.table(table$percentages), names.arg = table$income_brackets, cex.names = .75, las = 2)
```



After plotting this in R, I can see that the distribution is a somewhat normal distribution, with a majority of incomes from \$35,000 to \$49,000. It is unimodal.

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

The probability that a randomly chosen US resident makes less than $50,000 per year is:

```
prob_under_50k <- 0.022 + 0.047 + 0.158 + 0.183 + 0.212
prob_under_50k
```

```
## [1] 0.622
```

The probability is 62.2%.

(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.

The probability that a randomly chose US resident makes less than $50,000 per year and is a female can be calculated by the doing the following:

```
0.622 * 0.41
```

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

$F = female$
$Below\ 50K = Income\ below\ 50,000\ dollars$

$P(female\ and\ Below\ 50K) = P(female) \times P(Below\ 50K) = 0.622 \times 0.410 = 0.255$

The probability is approximately 25.5%. Given that there are many different income brackets in this table and distribution, we are assuming that the number of women are split pretty evenly across all income brackets.

(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.

Given that my calculation of 25.5% is well below the conclusion made that 71.8% of females make less than $50,000 per year, my assumption in part (c) is invalid. This shows that there is an unequal distribution of females across the different income brackets in this dataset, and that most females seem to fall into lower income brackets, while men tend to fall into higher income brackets.