Chapter 3 - Probability

Samuel I Kigamba

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?

```
impossible, the least we can get is a 2.
```

(b) getting a sum of 5?

```
there are 4 possible rolls that produce a sum of 5. thus the probability is 4/36 or 11.11%
```

(c) getting a sum of 12?

```
There is only one possible roll that produce a sum of 12. The probability is 1/36 or 2.78%
```

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.

poverty<- .146 lang_other<-.207 both<- .042

```
#install.packages("VennDiagram")
#install.packages("formattable")
library(VennDiagram)
```

```
## Warning: package 'VennDiagram' was built under R version 3.5.3
```

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

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

```
## Warning: package 'futile.logger' was built under R version 3.5.3
```

```
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.5.3
```

```
library(formattable)
```

```
## Warning: package 'formattable' was built under R version 3.5.3
```

```
##
## Attaching package: 'formattable'
```

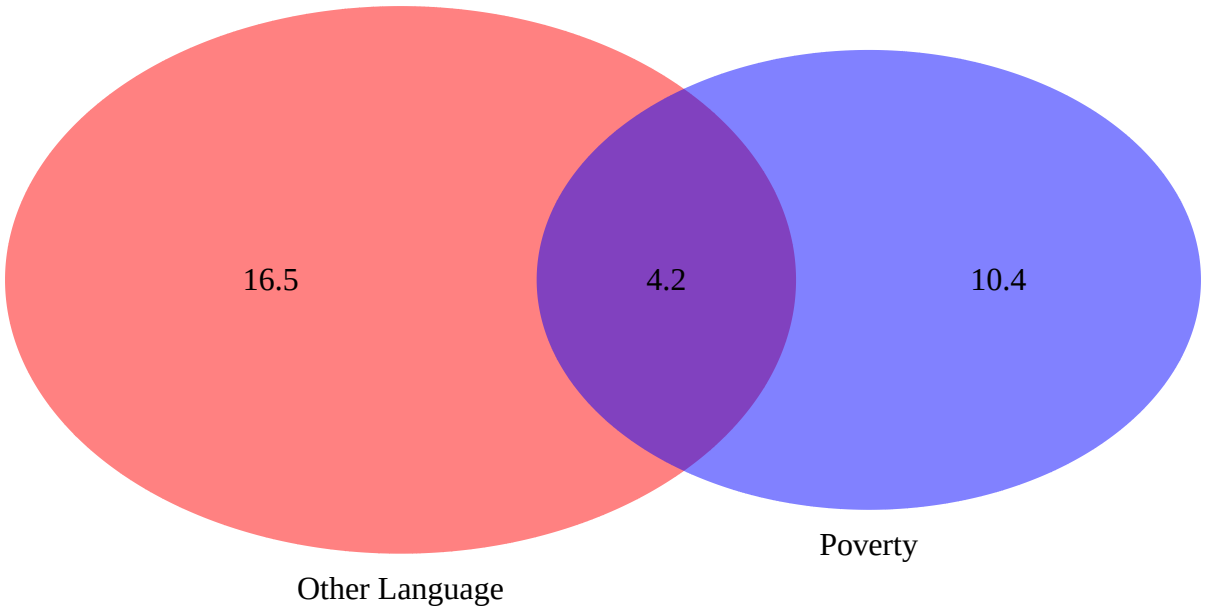
```
## The following object is masked from 'package:xtable':
##
##     digits
```

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

```
No. There are those who speak a foreign language at home and live below poverty line.
```

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

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




```
## (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?

```
P(Poverty) - P(Poverty and Foreign) = 14.6 - 4.2 = 10.4%
```

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

```
P(Poverty) + P(Foreign) - P(Poverty and Foreign) = 14.6 + 20.7 - 4.2 = 31.1%
```

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

```
P(English and NoPoverty)= 100 - P(Povery) + P(Other) - P(Poverty and Other) = 100 - (20.7 + 14.6 -
```

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

```
P(Poverty|Engish) = 10.4 /(69.9+10.4) = 0.1295143
P(Poverty|NonEnglish) = 4.2 / (16.5 + 4.2) = 0.2028986

There is a higher probability of being in Poverty given NonEnglish @ (20.3%) versus Poverty given E
```

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

```
f_blue <- 108 / 204
f_green <- 41 / 204
f_brown <- 55 / 204

m_blue <- 114 / 204
m_green <- 36 / 204
m_brown <- 54 / 204

both_blue <- 78 / 204
both_green <- 16 / 204
both_brown <- 23 / 204
```

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

```
either_blue <- f_blue + m_blue - both_blue
either_blue
```

```
## [1] 0.7058824
```

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

```
m_blue_both_blue <- both_blue/m_blue
m_blue_both_blue
```

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

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

```
f_blue <- 108 / 204
f_green <- 41 / 204
f_brown <- 55 / 204

m_blue <- 114 / 204
m_green <- 36 / 204
m_brown <- 54 / 204
```

```
both_blue <- 78 / 204
both_green <- 16 / 204
both_brown <- 23 / 204

m_brown_f_blue <- 19 / 204

P_m_brown_f_blue <- m_brown_f_blue/m_brown
P_m_brown_f_blue
```

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

What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

```
f_blue <- 108 / 204
f_green <- 41 / 204
f_brown <- 55 / 204

m_blue <- 114 / 204
m_green <- 36 / 204
m_brown <- 54 / 204

both_blue <- 78 / 204
both_green <- 16 / 204
both_brown <- 23 / 204

m_green_f_blue <- 11 / 204

p_m_green_f_blue <- m_green_f_blue/m_green
p_m_green_f_blue
```

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

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

if the two events were independent we would expect to see similar proportions between the eye colors of both partners but this is not the case. We thus conclude that 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.

```
total <- 95
prob_Hc <- 28 / total
no_rep_total <- total - 1 # Without replacment, we removed 1 book
prob_pb_fict <- 59 / no_rep_total
final <- prob_Hc * prob_pb_fict
final
```

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

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

```
total <- 95
prob_fict <- 72 / total
no_rep_total <- total - 1 # Without replacment, we removed 1 book
prob_Hc <- 28 / no_rep_total
final <- prob_fict * prob_Hc
final
```

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

(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.

```
total <- 95
prob_fict <- 72 / total
no_rep_total <- total # With replacment, we put back the first book
prob_Hc <- 28 / no_rep_total
final <- prob_fict * prob_Hc
final
```

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

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

```
we have a sample size of 95 which is large enough. the loss of 1 book means that only 1% is lost wh
```

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.

```
fees <- c(0, 25, 35)
portions <- c(0.54, 0.34, 0.12)
model <- fees * portions

mean(model)
```

```
## [1] 4.233333
```

```
sd(model)
```

```
## [1] 4.250098
```

(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.

```
sum(model*120)
```

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

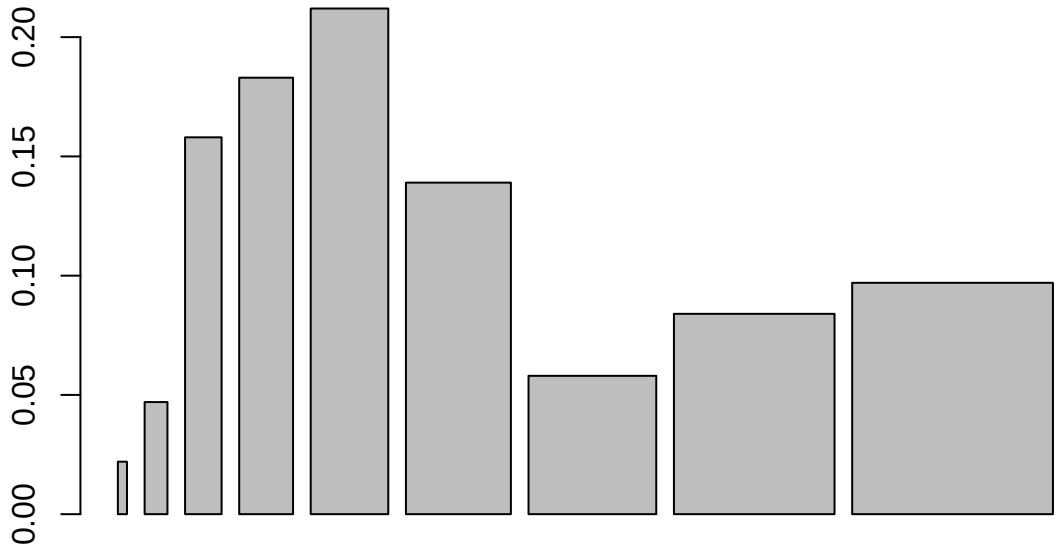
Income and gender. (3.38, p. 128) The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised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

```
inc <- c(5, 12.5, 20, 29.5, 42.5, 57.5, 70, 88, 110)
tot <- c(0.022, 0.047, 0.158, 0.183, 0.212, 0.139, 0.058, 0.084, 0.097)

barplot(tot, inc)
```



The data is right skewed with a long tail to the right, it is unimodal with only one prominent peak

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

```
sum(tot)       # Should add up to 100% (minor differences might be due to rounding off errors)
```

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

```
p_earn_less_than_50 <- sum(tot[0:5])  # sum the first 5 probabilities to get the prob of making less th
p_earn_less_than_50
```

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

(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.

```
p_earn_less_than_50_F <- sum(tot[0:5]) * 0.41
p_earn_less_than_50_F
```

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

my assumption is gender and income are independent. Though this is definitely not the case.

(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.

71.8% is significantly higher than 25%. this suggests a strong correlation between income and gende