Chapter 3 - Probability

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?
(b) getting a sum of 5?
(c) getting a sum of 12?

Answer Dice Rolls:

All possible outcomes (Sample) in terms of sum of pair of fair dice;

```
row_1 <- c(2,3,4,5,6,7)
row_2 <- c(3,4,5,6,7,8)
row_3 <- c(4,5,6,7,8,9)
row_4 <- c(5,6,7,8,9,10)
row_5 <- c(6,7,8,9,10,11)
row_6 <- c(7,8,9,10,11,12)

dice.roll <- matrix(c(row_1, row_2, row_3, row_4, row_5, row_6), nrow=6, byrow = TRUE)
dice.roll
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    2    3    4    5    6    7
## [2,]    3    4    5    6    7    8
## [3,]    4    5    6    7    8    9
## [4,]    5    6    7    8    9   10
## [5,]    6    7    8    9   10   11
## [6,]    7    8    9   10   11   12
```

(a) Probability of getting a sum of 1 is;

$$0/36 = 0$$

(b) Probability of getting a sum of 5 is;

$$4/36 = 1/9$$

(c) 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?
(b) Draw a Venn diagram summarizing the variables and their associated probabilities.
(c) What percent of Americans live below the poverty line and only speak English at home?
(d) What percent 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?
(f) Is the event that someone lives below the poverty line independent of the event that the person speaks a foreign language at home?

Answer Poverty and Language

(a) Since 4.2% of the survey participants fall into the both categories, living below the poverty line and speaking a foreign language at home are not disjoint. The answer is no.

(b) Creating a venn diagram requested in r

```
install.packages('VennDiagram', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Anil Akyildirim/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

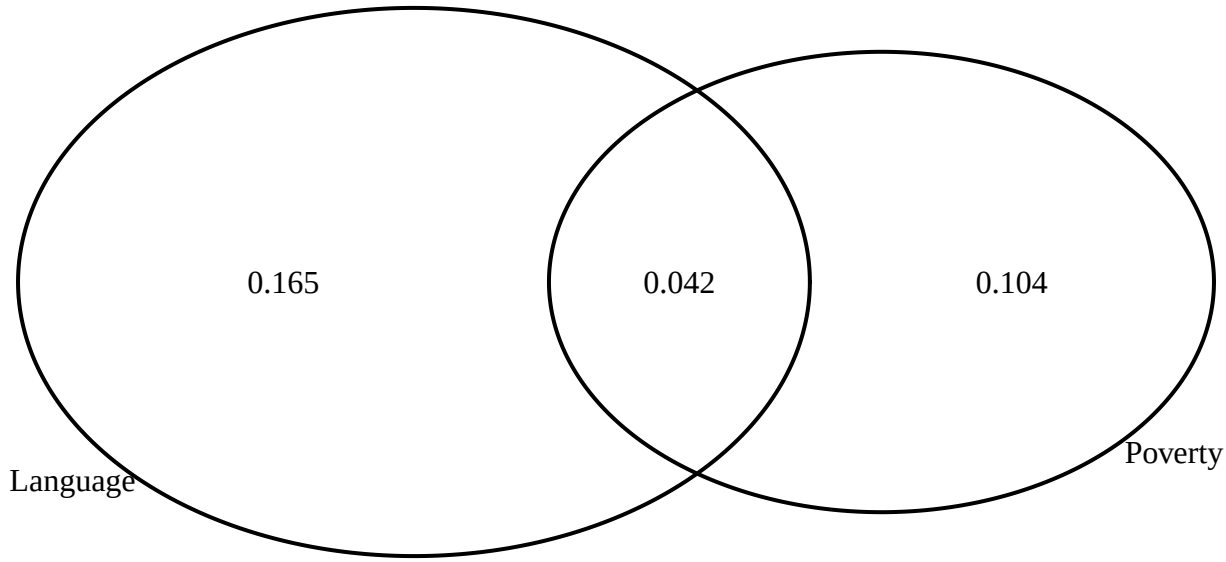
```
## package 'VennDiagram' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Anil Akyildirim\AppData\Local\Temp\RtmpEfHXsZ\downloaded_packages
```

```
library(VennDiagram)
```

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

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

```
grid.newpage()
venn.plot <- draw.pairwise.venn(area1 = 0.146,
                                area2 = 0.207,
                                cross.area = 0.042,
                                category = c("Poverty", "Language"))
```




```
venn.plot
```

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

(c) American's live below the poverty line and only speak English at home;

```
# Americans live below the powerty line
below_poverty <- 0.104+0.042

# Amerincas speak only English
speak_english_only <- 1-(0.165+0.042)

poverty_line_and_speak_english_only <- below_poverty * speak_english_only
poverty_line_and_speak_english_only
```

```
## [1] 0.115778
```

(d) Percent Americans live below the poverty line or speak a foreign language is;

```
0.165 + 0.042 + 0.104
```

```
## [1] 0.311
```

31.1

(e) Percent of Americans live above the poverty line and only speak English at home is;

```
above_poverty <- 1- below_poverty
above_poverty_and_speak_english_only <- above_poverty * speak_english_only
above_poverty_and_speak_english_only
```

```
## [1] 0.677222
```

67.7

(f)If they are independent it should give us

$$P(A) * P(B) = P(AandB)$$

```
PAB <- 14.6 *20.7

if (PAB==4.2){
  print("Lives Below Poverty Line and Speaks a Foreign language at home is independent")
} else{
  print("Lives Below Poverty Line and Speaks a Foreign language at home is dependent")
}
```

```
## [1] "Lives Below Poverty Line and Speaks a Foreign language at home is dependent"
```

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?
(b) What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes?
(c) Whatistheprobabilitythatarandomlychosenmalerespondentwithbrowneyeshasapartner with blue eyes? What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?
(d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning.

Answer Assortative mating
(a) Proabability of Male Blue Eyes or Female Blue Eyes

$$P(AorB) = P(A) + P(B) - P(AandB)$$

Total Outcome (sample) is 204 Total Male Outcome with Blue eyes is 114 Total Female outcome with Blue eyes is 108 The Male Blue eyes AND Female Blue Eyes is 78

```
r total <- 204 mb <- 114 fb <- 108 mb_and_fb <- 78 mb_or_fb <- mb + fb - mb_and_fb
mb_or_fb
## [1] 144
r probability_mb_or_fb <- mb_or_fb / total probability_mb_or_fb
## [1] 0.7058824
```

The answer is 0.705
(b) Probability Male Blue Eyes has a Partner with Blue eyes (Female Blue Eyes)

```
r total_blue_male <- 114 mb_and_fb / total_blue_male
## [1] 0.6842105
```

The answer is 0.68
(c) i) Probability Random Male with Brown Eyes has a partner with blue eyes
“‘r total_brown_male <- 54 mbr_and_fb <- 19
mbr_and_fb / total_brown_male “‘

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

The answer is 0.35
ii) Probability male respondent with green eyes partner with blue eyes
“‘r total_green_male <- 36 mg_and_fb <- 11
mg_and_fb / total_green_male “‘

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

The answer is 0.30
(d) It is not independent. Basic example is that in any combination

$$P(A) * P(B) = P(AandB)$$

does not stand.

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.
(b) Determine the probability of drawing a fiction book first and then a hardcover book second, when drawing without replacement.
(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.
(d) The final answers to parts (b) and (c) are very similar. Explain why this is the case.

Answer Book on a bookshelf

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

t: Total books = 95 h: Hard cover book = 28 pf: Paperback Fiction books= 59

```
h <-28/95
pf <-59/(95-1) # we need to take one out because of without replacement
h_pf<- h*pf
h_pf
```

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

The answer is 0.18

(b) Probability of drawing a fiction book first then a hardcover second without replacement.

t: Total books = 95 f: Fiction book = 72 h: Hard cover book =28

```
f <- 72/95
h2 <- 28/(95-1) # we need to take out 1 because of without replacement
f_h2 <- f*h2
f_h2
```

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

The answer is 0.225

(c) same scenario as b but this time we are placing back the book prior to second draw

t: Total books = 95 f: Fiction book = 72 h: Hard cover book =28

```r
f <- 72/95
h3 <- 28/(95) # we are not taking out the 1 as we placed back the book
f_h3 <- f*h3
f_h3
```

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

The answer is 0.223

(d) Explanation of why part b and c is similar. The total book amount is 95 and we are only taking out a book one by one. In b, even though we didnt replace the book, the book we draw (which is 1) is not significant enough to make a difference. This is seen in finding c. The sample size is large enough to keep the finding b and c similar.

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.
(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.

Answer Baggage fees

fee_1 : 25$ fee_2 : 35$ fee_3 : 0$ (no luggage no fee)

no_luggage: 54% one_luggage: 34% two_luggage: 12%

more than two luggage is negligible.

Creating a data frame to display this

```
fee_1 <- 0
fee_2 <- 25
fee_3 <- 35

no_luggage <- 0.54
one_luggage <- 0.34
two_luggage <- 0.12

pas_1 <- "0 bags"
pas_2 <- "1 bag"
pas_3 <- "2 bags"


passengers <- c(pas_1, pas_2, pas_3)
luggage <- c(no_luggage, one_luggage, two_luggage)
fee <- c(fee_1, fee_2, fee_3)

df <- data.frame(passengers, luggage, fee)
df
```

```
##   passengers luggage fee
## 1     0 bags    0.54   0
## 2      1 bag    0.34  25
## 3     2 bags    0.12  35
```

Finding Average Revenue

```
average_revenue <-(df$luggage*df$fee)/(sum(df$luggage))
average_revenue
```

```
## [1] 0.0 8.5 4.2
```

Average Revenue for o bags per passenger is 0 Average Revenue for 1 bag per passenger is $8.5 Average Revenue for 2 bag per passenger is $4.2

Average Revenue is;

```
sum(average_revenue)
```

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

Finding Standard Deviation with below formula:

```
# calculate variance
variance <- (df$fee-average_revenue)^2
variance
```

```
## [1]   0.00 272.25 948.64
```

```
sd <- sqrt(variance)
sd
```

```
## [1]  0.0 16.5 30.8
```

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

Answer Income and Gender

Creating the data frame outlined Income vs Total

```
income <- c("$1 to $9,999 or loss", "$10,000 to $14,999", "$15,000 to $24,999", "$25,000 to $34,999", "$
total <- c(0.022, 0.047, 0.158, 0.183, 0.212, 0.139, 0.058, 0.084, 0.097)

df_2 <- data.frame(income, total)
df_2
```

```
##                 income total
## 1 $1 to $9,999 or loss 0.022
## 2   $10,000 to $14,999 0.047
## 3   $15,000 to $24,999 0.158
## 4   $25,000 to $34,999 0.183
## 5   $35,000 to $49,000 0.212
## 6   $50,000 to $64,999 0.139
## 7   $65,000 to $74,999 0.058
## 8   $75,000 to $99,999 0.084
## 9     $100,000 or more 0.097
```

(a) Describe the distribution of total personal income.

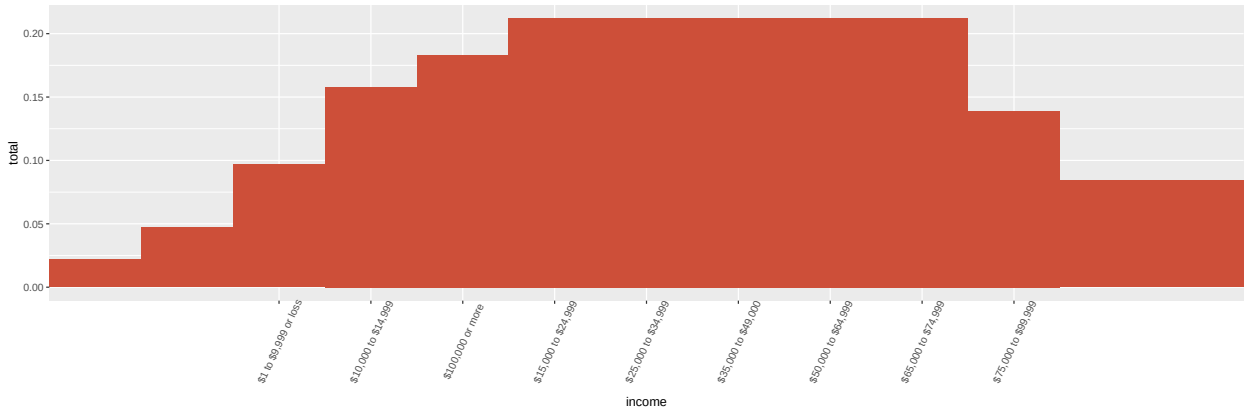
Creating a bar plot to outline the distribution.

```
install.packages('ggplot2', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Anil Akyildirim/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Anil Akyildirim\AppData\Local\Temp\RtmpEfHXsZ\downloaded_packages
```

```
## Warning: position_stack requires non-overlapping x intervals
```



(b) Probability Random Chosen US resident makes less than $50K per year.

```
p_less_10<-0.022
p_bet_10_15<-0.047
p_bet_25_15<-0.158
p_bet_35_25<-0.183
p_bet_50_35<-0.212

p_less_50<-p_less_10 + p_bet_10_15 + p_bet_25_15 + p_bet_35_25 + p_bet_50_35
p_less_50
```

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

The answer is 0.622

(c) Probability random chosen US resident that makes less than 50K and female

Assuming the gender and income is independent from each other.

p_female : Probability of random chosen female :0.41

p_less_50_female = p_less_50 * p_female

```r
p_female <- 0.41
p_less_50_female <- p_less_50 * p_female
p_less_50_female
```

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

The answer is 0.25

(d) My assumption was not correct. If gender and income was independent, the probability of a random chosen US resident that makes less than 50K would be ~25% however it is 71.8% . This means that gender and income are dependent variables.