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?

$$\frac{0}{36} = 0.000$$

(b) getting a sum of 5?

$$\frac{4}{36} = 0.083$$

(c) getting a sum of 12?

$$\frac{1}{36} = 0.027$$

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. A person may live in poverty and speak a foriegn language. These are not mutually exclusive events.

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

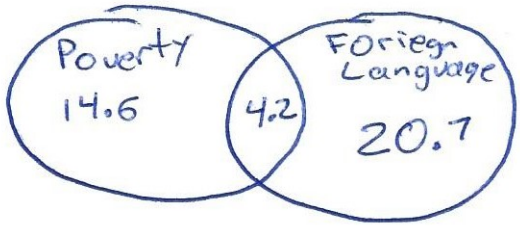

Figure 1: Live Below Poverty Rate and/or Speak a Foriegn Language

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

4.2%

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

Calculate a Non-Mutually Exclusive OR scenario :

$$P(A \text{ or } B) = \frac{P(A) + P(B)}{P(A * B)} = \frac{.146 + .207}{.042} = .311$$

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

95.8%

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

No. By definition, they are dependent. Knowing that an observation lives below the poverty line increases the probability that a foriegn language is spoken at home.

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?

108 couples in which the women have blue eyes.

$$P(A) = \frac{108}{204} = 0.52943$$

114 couples in which the men have blue eyes.

$$P(B) = \frac{114}{204} = 0.5588$$

78 couples in which both men and women have blue eyes.

$$P(A \text{ and } B) = \frac{78}{204} = 0.3823$$

$$P(A \text{ or } B) = P((A) + P(B)) - P(A * B) = (.53 + .55 - .38 = .70$$

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

108 couples in which the women have blue eyes. $P(A) = \frac{108}{204} = 0.5294$

(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?

There are 54 men with brown eyes. 19 of which have partners with blue eyes.

$$\frac{19}{54} = 0.35$$

There are 36 men with green eyes. 11 of which have partners with blue eyes.

$$\frac{11}{36} = 0.30$$

Note: *If the question was what are the odds of selecting a couple in which the male has green eyes and the female has blue eyes then the answer becomes:*

$$\frac{11}{204} = 0.05$$

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

Yes. If I look at the 36 men with green eyes, I noticed that 44% of them have partners with green eyes. The overall percentage of Scandinavian women is 20% so it appears that having green eyes increases the probability of having a partner with green eyes.

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.

Draw hardcover first :

$$\frac{28}{95} = 0.29$$

Draw paperback fiction book second :

$$\frac{67}{94} = 0.71$$

Probability of both:

$$0.29 * 0.71 = .20$$

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

Draw fiction first :

$$\frac{72}{95} = 0.76$$

Draw hardcover second :

$$\frac{28}{94} = 0.30$$

Probability of both:

$$0.76 * 0.3 = .23$$

(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.

Draw fiction first :

$$\frac{72}{95} = 0.76$$

Draw hardcover second (restore the fiction book) :

$$\frac{28}{95} = 0.29$$

Probability of both:

$$0.76 * 0.29 = .22$$

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

Reducing the population from 95 to 94 is not that dramatic. If you compare with replacement vs without replacement on a larger scale, it would be more noticable.

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.

```
# create a data set in which 54 observations generate $0 in revenue,
#          34 generate $25, and 12 generate $60
column_revenue <- c(rep(0, 54), rep(25,34), rep(60,12))

#calculate the expected revenuep per person
expected_revenue<-sum(column_revenue)/length(column_revenue)

# same as (( 0 * .54 ) + ( 25 * .34 ) + ( 60 * .12 ))
# variance ? (0^2 * .54) + (25^2 * .34) + (60^2 * .12)

sprintf("The expected revenue ( or mean ) per person is %.3f", expected_revenue)
```

```
## [1] "The expected revenue ( or mean ) per person is 15.700"
```

```
var_recalc = sum( (((column_revenue - expected_revenue)^2)/length(column_revenue))  )

val_count = length(column_revenue)

var_recalc=0

for (n in 1:val_count)
{
  md = column_revenue[n]-expected_revenue
  md_sq<-md^2

     # be careful you divide first, then add
  var_recalc=var_recalc+(md_sq/(val_count-1))
}

sd_recalc=sqrt(var_recalc)


sprintf("The standard deviation, calculated manually, is %.3f", sd_recalc)
```

```
## [1] "The standard deviation, calculated manually, is 20.051"
```

```
sprintf("The standard deviation, using sd(), is %.3f", sd(column_revenue))
```

```
## [1] "The standard deviation, using sd(), is 20.051"
```

(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.

I tried a few different methods. I settled on the brute force method of prorating the original probabilites X 120.

```r
# original variance was 402

expected_revenue_120=expected_revenue * 120
sprintf("The total revenue expected for 120 passengers is $%.2f", expected_revenue_120)
```

```
## [1] "The total revenue expected for 120 passengers is $1884.00"
```

```r
prorated_0bag = 120 * .54 * 100
prorated_1bag = 120 * .34 * 100
prorated_2bag = 120 * .12 * 100


# calculate variance from the linear equation ? doesnt seem to work
# var_recalc<-(25^2 * prorated_1bag) + (60^2 * prorated_2bag )


# given 120 passengers
# 64.8 will pay 0
# 40.8 will pay 25
# 14.4 will pay 60


prorated_to_120 <- c(rep(0, prorated_0bag), rep(25,prorated_1bag), rep(60,prorated_2bag))

var(prorated_to_120) # s.b something a little smaller than 402
```

```
## [1] 397.9128
```

```r
sprintf("The standard deviation, roughly prorating the quantities to 120, is %.3f", sd(prorated_to_120))
```

```
## [1] "The standard deviation, roughly prorating the quantities to 120, is 19.948"
```

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.

Normal although the \$100,000 or more category might be hiding a long tail.

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

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.

I have some understanding that women make less than men, but since that information is not provided, I will assume gender and salary are independent.

$$0.41 * 0.622 = .25$$

(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.

The percentage of the total population that is female and makes less than 50,000 is

$$0.41 * 0.718 = .29$$