Data 606 - Homework 3 - Probability

Cameron Smith

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?

Assuming the dice each have 6 sides, labelled 1 through 6, then (a) sum of 1 is 0/36 (impossible) (b) sum of 5 is 4/36 (c) sum of 12 is 1/36

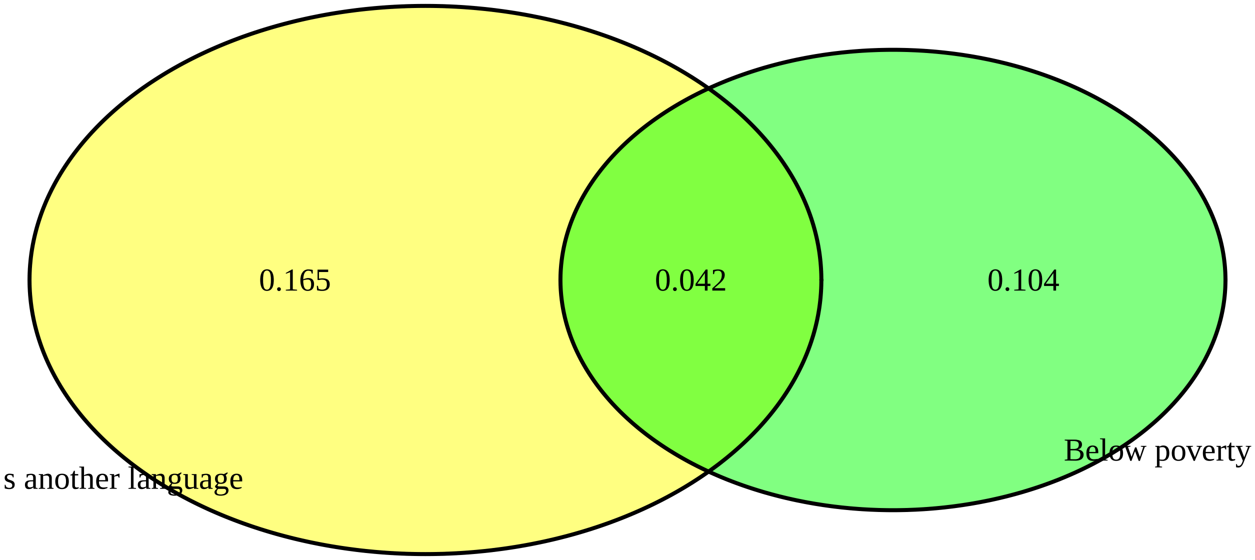
Poverty and language. (3.8, p. 93) The American Community Survey is an ongoing survey that provides data every year to give communities the current information they need to plan investments and services. The 2010 American Community Survey estimates that 14.6% of Americans live below the poverty line, 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?

```r
# Create variables based on above
below_poverty_line <- .146
lang_other_english <- .207
both <- .042
```

(a) No, they are not disjoint since they can occur at the same time
(b) See diagram below

```r
grid.newpage()
draw.pairwise.venn(area1 = below_poverty_line,
                   area2 = lang_other_english,
                   cross.area = both,
                   fill = c("green", "yellow"),
                   category = c("Below poverty line", "Speaks another language"),
                   scaled = TRUE)
```



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

(c) 10.4% live below poverty line and speak only English

```
below_poverty_line - both
```

```
## [1] 0.104
```

(d) 31.1% chance of living below poverty line or only speaking English

```
(below_poverty_line + lang_other_english) - (both)
```

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

(e) 68.9% live above poverty line and speak only English

```
1 - (below_poverty_line + lang_other_english - both)
```

```
## [1] 0.689
```

(f) No they are not independent.

```
# Testing rule for independent processes - P(A and B) = P(A) x P(B)
below_poverty_line * lang_other_english
```

```
## [1] 0.030222
```

```
isTRUE(both == (below_poverty_line * lang_other_english))
```

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

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

(a) 70.6% (78 + 19 + 11 + 23 + 13 people w/ blue eyes divided by 204 total people)
(b) 68.4% (78 people divided by 114 total people)
(c) 35.2% (19 people divided by 54 total people), and 30.5% (11 people by 36 total people)
(d) It appears that they are not independent, based on the fact that the probabilities of each of the combinations are different.

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.

(a) 18.5% (28 books out of 95 total multiplied by 59 books out of 94 books total)
(b) 22.4%

```
fiction1st <- (72/95)
hardcover2ndpossibility1 <- (28/94)
hardcover2ndpossibility2 <- (27/94)

fiction1st *
  (((hardcover2ndpossibility1 * fiction1st)) +
     ((hardcover2ndpossibility2 * (1 - fiction1st))))
```

```
## [1] 0.2238039
```

(c) 22/3% (72 / 95 multiplied by 28 / 95)
(d) The size of the population (# of books in each category) is large enough that one book does not make much of a difference.

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.

Average revenue per passenger: $15.70 Variance: $207.21 Standard deviation: $14.39

See code block below for more info:

```
bag_costs <-c(0, 25, 35)
bag_perc <- c(.54,.34,.12)

# Calc average revenue per passenger
one_bag_revenue <- bag_costs[2] * bag_perc[2]
two_bags_revenue <- (bag_costs[2] + bag_costs[3]) * bag_perc[3]

per_person_revenue <- one_bag_revenue + two_bags_revenue

## Calculation of standard deviation

# Start with variance.  Same as for ’normal’ statistics except that probability included as well for ea
variance <- ((bag_costs[1] - per_person_revenue)^2 * .54) +
  ((bag_costs[2] - per_person_revenue)^2 * .34) +
  ((bag_costs[3] - per_person_revenue)^2 * .12)

# Standard deviation is square root of variance
standard_deviation <- sqrt(variance)

# Printing of the calculated answers
per_person_revenue
```

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

```
variance
```

```
## [1] 207.21
```

```
standard_deviation
```

```
## [1] 14.39479
```

(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.

Expected revenue would be $1,884 (average revenue * # of passengers) Variance would be $24,865.20 (variance * # of passengers) Standard deviation would be $157.69 (square root of variance calculated above)

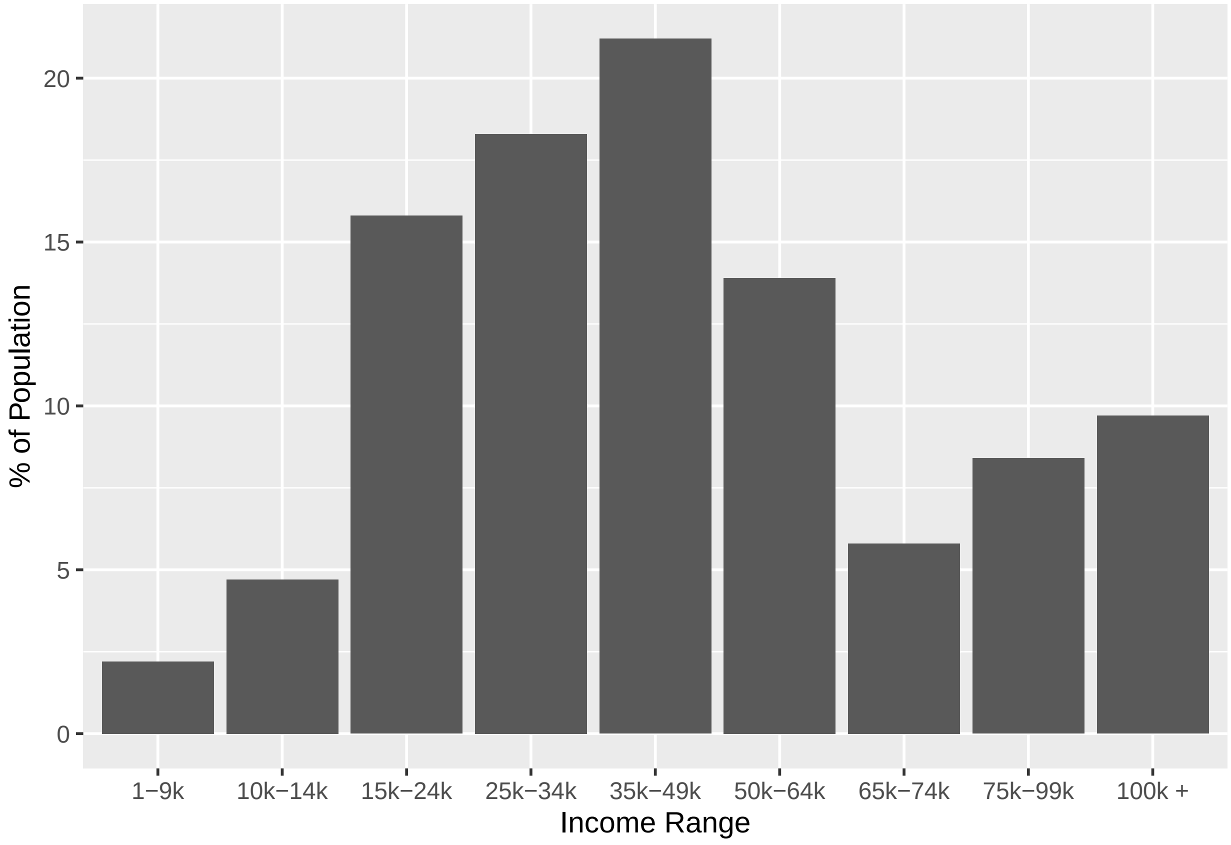
Income and gender. (3.38, p. 128) The relative frequency table below displays the distribution of annual total personal income (in 2009 inflation-adjusted dollars) for a representative sample of 96,420,486 Americans. These data come from the American Community Survey for 2005-2009. This sample is comprised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

A unimodal distribution peaking at the 35-49k range with a slight tail to the right.

```
income <- data.frame(c("1-9k", "10k-14k", "15k-24k", "25k-34k", "35k-49k", "50k-64k", "65k-74k", "75k-9
colnames(income) <- c("income", "perc_of_population")
ggplot(data=income, aes(x=factor(income, level=income), y=perc_of_population)) +
  geom_bar(stat="identity") +
  xlab("Income Range") + ylab("% of Population")
```



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

62.2%, which is the sum of the percentages for all ranges under 50k.

```
sum(income$perc_of_population[1:5])
```

```
## [1] 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 is 25.5%, calculated by the figure in (b) above multiplied by the % of females in the samples population (41%). It is assumed that these are independent of one another.

```
.622 * .41
```

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

(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 assumption is invalid as this infers that the variables are not independent. If they were independent then the female figure would equal the total figure (inclusive of females and males).

```
# Testing for independence: Does P(A and B) = P(A) x P(B)?

a_and_b <- .718
a <- .41
b <- .622

a*b
```

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

```
isTRUE(a_and_b == (a*b))
```

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

-End of Assignment-