IS606 Homework 2

Daniel Dittenhafer

September 13, 2015

2.6 Dice rolls (p116)

If you roll a pair of fair dice, what is the probability of:

Assuming:

- Six sided dice
- Values 1 - 6 (no zero)

a. getting a sum of 1? The minimum sum from a pair of dice, given the assumptions above, would be 2. Since a sum of 1 is not part of the set of outcomes, the probability would be 0.

b. getting a sum of 5? How many ways can a sum of 5 be the result of 2 dice?

Roll	1	2	3	4
die 1	1	2	3	4
die 2	4	3	2	1

There are 4 outcomes which can result in a sum of 5, and 36 total outcomes possible (6 X 6), therefore the probability is $\frac{4}{36} = \frac{1}{9} \approx 0.1111111$

c. getting a sum of 12? There is only one outcome from 2 dice which sum to 12: a 6 and 6 (boxcars). As such, the probablity is:

$\frac{1}{36} \approx 0.0277778$

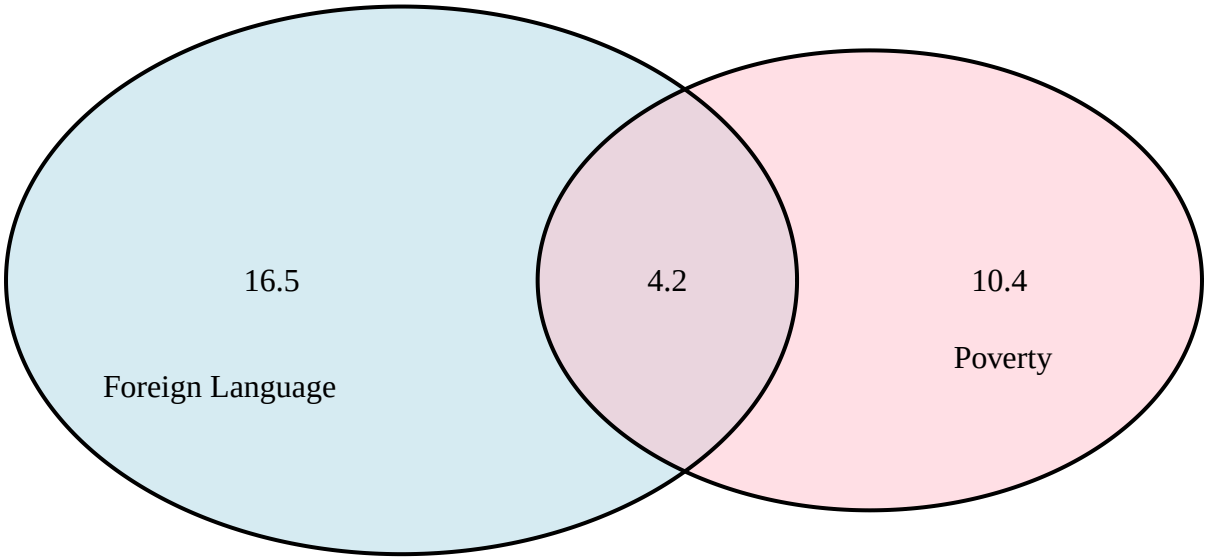
2.8 Poverty and language (p117)

a. Are living below the poverty line and speaking a foreign language at home disjoint? No. Specifically, one could be living below the poverty line and speaking a foreign language at home, or one could living below the poverty line only, or speaking a foreign language at home only. In the case described in the question, 4.2% fall into both categories.

b. Draw a Venn diagram summarizing the variables and their associated probabilities. The R code segment below shows using the `VennDiagram` package to create a Venn diagram for the variables and probabilities described in the exercise description:

```
library(VennDiagram)
pov <- 14.6
forLang <- 20.7
both <- 4.2
povOnly <- pov - both
forLangOnly <- forLang - both
```

```
venn.plot <- draw.pairwise.venn(pov,
                    forLang,
                    cross.area=both,
                    c("Poverty", "Foreign Language"),
                    fill=c("pink", "lightblue"),
                    cat.dist=-0.08,
                    ind=FALSE)
grid.draw(venn.plot)
```



c. What percent of Americans live below the poverty line and only speak English at home? Based on the results of the Venn diagram, 10.4% of American 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? Using the General Addition Rule from the text, where P(A or B) = P(A) + P(B) - P(A and B), then:

$20.7 + 14.6 - 4.2 = 31.1\%$

e. What percent of Americans live above the poverty line and only speak English at home? 79.3% English-speaking Americans - 10.4 English speaking Americans in Povery = 68.9% Americans living above 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? Using the test of the Multiplication Rule for independent events (p86,87), do poverty and language satisfy the rule:

P(A and B) = P(A) x P(B)

$0.146 \times 0.207 = 0.030222$

This is not equal to 0.042, therefore it fails the Multiplication Rule for independent events test - the events are not independent and therefore knowing information about one of the events does provide information about the outcome of the other event.

2.20 Assortative mating (p121)

a. What is the probability that a randomly chosen male respondent or his partner has blue eyes? This is not a disjoint event, because there is a scenario when both the male and partner have blue eyes. There are 108 females in the study with blue eyes, and another 114 males with blue eyes, with 78 where male and partner have blue eyes, out of a total of 204 couples in the study.

P(Male Blue or Partner Blue) = $\frac{114}{204} + \frac{108}{204} - \frac{78}{204} = 0.7058824$

b. What is the probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes? My initial interpretation of this question was as P(Male Blue and Partner Blue):

P(Blue and Blue) = $\frac{78}{204} = 0.3823529$

Revisiting this answer, when interpreted as P(Partner Blue given Male Blue):

P(Partner Blue given Male Blue) = $\frac{78}{114} = 0.6842105$

c. What is the probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes? Green eyes and blue eyes? This sounds like the question is P(Partner Blue given Male Brown), and P(Partner Blue given Male Green). Thinking in this way:

P(Partner Blue given Male Brown) = $\frac{19}{54} = 0.3518519$

P(Partner Blue given Male Green) = $\frac{11}{36} = 0.3055556$

d. Does it appear that the eye color of male respondents and their partners are independent? Explain your reasoning. Looking at proportions by male eye color, there is definitely an affinity to selecting a partner with the same eye color. Given this analysis, the eye color of male respondents and their partners does not appear to be independent.

```
fBlue <- c(78,19,11)
fBrown <- c(23,23,9)
fGreen <- c(13,12,16)
df <- data.frame(fBlue, fBrown, fGreen)
row.names(df) <- c("mBlue", "mBrown", "mGreen")
df$sum <- c(sum(df["mBlue",]), sum(df["mBrown",]), sum(df["mGreen",]))

dfProp <- df / df$sum
dfProp
```

```
##            fBlue    fBrown    fGreen sum
## mBlue  0.6842105 0.2017544 0.1140351   1
## mBrown 0.3518519 0.4259259 0.2222222   1
## mGreen 0.3055556 0.2500000 0.4444444   1
```

2.30 Books on a bookshelf (p123)

The table below shows the distribution of books on a bookcase:

Type/Format	Hardcover	Paperback	Total
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. First we identify the marginal probability for Hardcover:

P(Hardcover) = $\frac{28}{95} = 0.2947368$

Then the joint probability of paperback fiction (w/o replacement):

P(Paperback Fiction) = $\frac{59}{94} = 0.6276596$

P(Hardcover and Paperback Fiction) = $0.2947368 \times 0.6276596 = 0.1849944$

b. Determine the probability of drawing a fiction book first and then a hardcover book second when drawing without replacement. First we identify the marginal probability for Fiction:

P(Fiction) = $\frac{72}{95} = 0.7578947$

Then the marginal probability of hardcover fiction (w/o replacement):

P(Hardcover) = $\frac{28}{94} = 0.2978723$

P(Fiction and Hardcover) = $0.7578947 \times 0.2978723 = 0.2257559$

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. We know the marginal probability for Fiction:

P(Fiction) = $\frac{72}{95} = 0.7578947$

But now the marginal probability of hardcover fiction is based on replacement:

P(Hardcover) = $\frac{28}{95} = 0.2947368$

P(Fiction and Hardcover) = $0.7578947 \times 0.2947368 = 0.2233795$

d. The final ansewrs to parts (b) and (c) are very similar. Explain why this is the case. The probabilities are very similar because the marginal probabilities that go into the result are also very similar. The only difference is the replacement which simply changes the denominator of the second book selection by 1. Given the number of books on the bookcase (94 vs 95), replacement has very little effect initally. On the other hand, after removing 90 books from the self, replacement (or not) would have a significant effect.

2.38 Baggage fees (p124)

An airline changes the following baggage fees: \$25 for the first bag and \$35 for the second. Suppose 54% of passengers have no checked baggage, 34% have one piece of checked luggage, and 12% have two pieces. We suppose a negligible portion of people check more than two bags.

```r
prob <- c(0.54, 0.34, 0.12)
bags <- c(0, 1, 2)
fees <- c(0, 25, 25 + 35)
df38 <- data.frame(prob, bags, fees)
df38$weightRev <- df38$prob * df38$fees
df38
```

a. Build a probability model, compute the average revenue per passenger, and compute the corresponding standard deviation.

```
##   prob bags fees weightRev
## 1 0.54    0    0       0.0
## 2 0.34    1   25       8.5
## 3 0.12    2   60       7.2
```

```r
# Compute the average revenue per passenger
avgRevPerPax <- sum(df38$weightRev)
avgRevPerPax
```

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

```r
# Compute Variance
df38$DiffMean <- df38$weightRev - avgRevPerPax
df38$DiffMeanSqrd <- df38$DiffMean ^ 2
df38$DiffMeanSqrdTimesProb <- df38$DiffMeanSqrd * df38$prob
df38
```

```
##   prob bags fees weightRev DiffMean DiffMeanSqrd DiffMeanSqrdTimesProb
## 1 0.54    0    0       0.0    -15.7       246.49              133.1046
## 2 0.34    1   25       8.5     -7.2        51.84               17.6256
## 3 0.12    2   60       7.2     -8.5        72.25                8.6700
```

```r
# Compute standard deviation
varRevPerPax <- sum(df38$DiffMeanSqrdTimesProb)
sdRevPerPax <- sqrt(varRevPerPax)
sdRevPerPax
```

```
## [1] 12.62538
```

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. The following R code computes the revenue for a flight of 120 passengers:

```r
pax <- 120
avgFlightRev <- avgRevPerPax * pax
avgFlightRev
```

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

```r
# Standard Deviation
varFlightRev <- (pax ^ 2) * varRevPerPax
sdFlightRev <- sqrt(varFlightRev)
sdFlightRev
```

```
## [1] 1515.046
```

The standard deviation of the average revenue for the flight is valid only if the average revenue per passenger is independent of other random variables. Given that this is the only random variable, independence is not an issue.

2.44 Income and gender (p126)

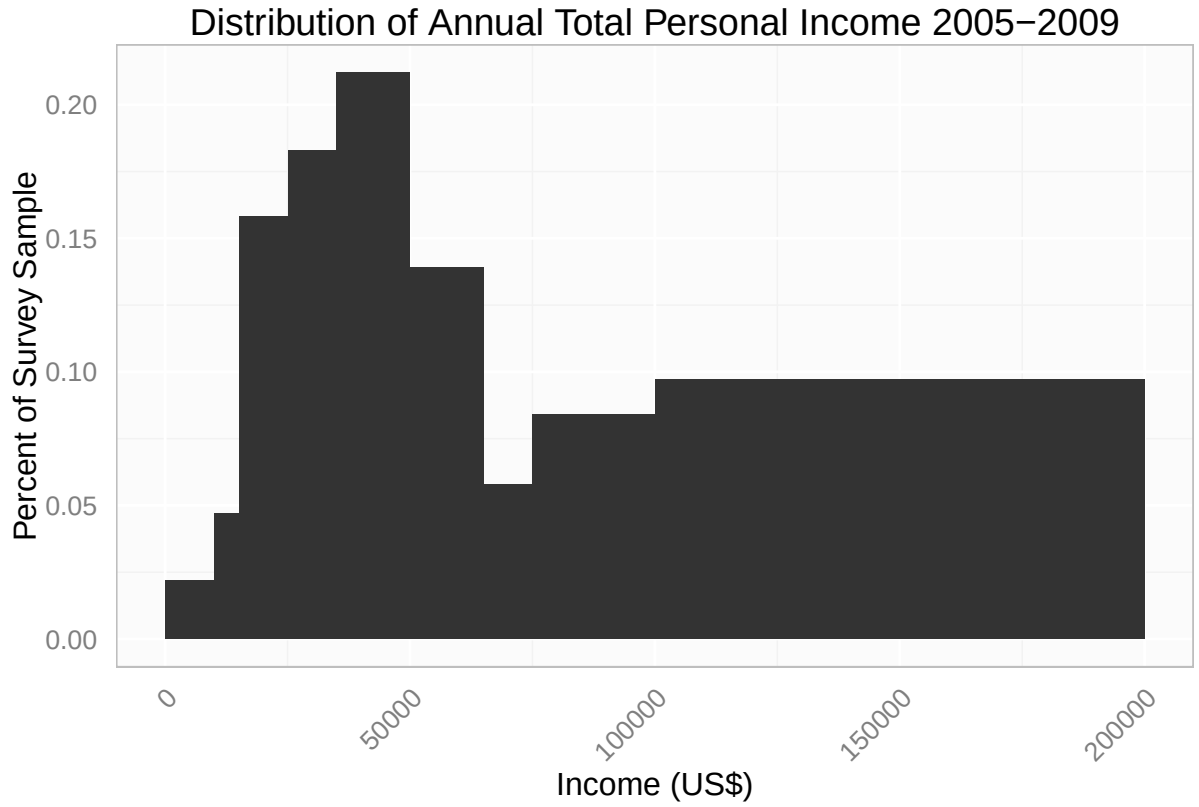
59% males, 41% females

```r
income <- c("$1 - $9,999 or loss",
            "$10,000 to $14,999",
            "$15,000 to $24,999",
            "$25,000 to $34,999",
            "$35,000 to $49,999",
            "$50,000 to $64,000",
            "$65,000 to $74,999",
            "$75,000 to $99,999",
            "$100,000 or more")
bounds <- c(1, 10000, 15000, 25000, 35000, 50000, 65000, 75000, 100000)
size <- c(9999, 4999, 9999, 9999, 14999, 14999, 9999, 24999, 99999)
center <- bounds + (size / 2)
total <- c(0.022, 0.047, 0.158, 0.183, 0.212, 0.139, 0.058, 0.084, 0.097)

df44 <- data.frame(income, center, total)
df44
```

```
##                income   center total
## 1 $1 - $9,999 or loss   5000.5 0.022
## 2  $10,000 to $14,999  12499.5 0.047
## 3  $15,000 to $24,999  19999.5 0.158
## 4  $25,000 to $34,999  29999.5 0.183
## 5  $35,000 to $49,999  42499.5 0.212
## 6  $50,000 to $64,000  57499.5 0.139
## 7  $65,000 to $74,999  69999.5 0.058
## 8  $75,000 to $99,999  87499.5 0.084
## 9    $100,000 or more 149999.5 0.097
```

a. Describe the distribution of total personal income. Plotting the points in the bar chart below, the visualization shows a concept of the distribution. The distribution is bimodel with peaks in the $35K - $50K range, and $100K+, and somewhat skewed to right given the upper range has an undefined upper bound.



b. What is the probability that a randomly chosen US resident makes less than $50,000 per year? By summing the percentages from the data for the disjoint, mutually exclusive outcomes of each income level, we can determine the total probability:

```
pr50K <- sum(df44[1:5,]$total)
pr50K
```

```
## [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. We don't know the relationship between the probability of an income of less than $50,000 and being female. Assuming they are independent events then P(A and B) = P(A) x P(B).

```
P_female <- 0.41
P_50K <- pr50K
# Compute
P_female_50K <- P_female * P_50K
# Show the result
P_female_50K
```

```
## [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. I'm interested in what percent of population is female and makes less than $50K... What is 71.8% of 41%?

```
# What percent of population is female and makes less than $50K?
actualFemale50K <- 0.718 * P_female
actualFemale50K
```

```
## [1] 0.29438
```

While 0.29438 is close to 0.25502, they are not equal. Therefore I conclude that making less than $50K and being female are not independent events.