Data606 Homework 2 Jagdish Chhabria

Jagdish Chhabria

February 6, 2019

Graded: Graded: 2.6, 2.8, 2.20, 2.30, 2.38, 2.44

Chapter 2 - Introduction to Data, Problem 2.6

Dice rolls. If you roll a pair of fair dice, what is the probability of

(a) getting a sum of 1? 0

(b) getting a sum of 5? 4/36

(c) getting a sum of 12? 1/36

```
d1<-c(1,2,3,4,5,6)
d2<-c(1,2,3,4,5,6)
#length(d1)
d1d2<-NULL
for(i in seq(d1[1], length(d1))) {
  for(j in seq(d2[1], length(d2))) {
  d1d2<-c(d1d2,d1[i]+d2[j])
  }
}
#print(d1d2)
#d1d2<-as.factor(d1d2)
p1=sum(d1d2=='1')/length(d1d2)
p5 = sum(d1d2=='5')/length(d1d2)
p12 = sum(d1d2=='12')/length(d1d2)
cat("The probability of getting a sum of 1 is:", round(p1,2), "\n")
```

```
## The probability of getting a sum of 1 is: 0
```

```
cat("The probability of getting a sum of 5 is:", round(p5,4), "\n")
```

```
## The probability of getting a sum of 5 is: 0.1111
```

```
cat("The probability of getting a sum of 12 is:", round(p12,4), "\n")
```

```
## The probability of getting a sum of 12 is: 0.0278
```

Chapter 2 - Introduction to Data, Problem 2.8

(a) Are living below the poverty line and speaking a foreign language at home disjoint? No. Because there are people who are both living below the poverty line and speak a foreign langauge at home.

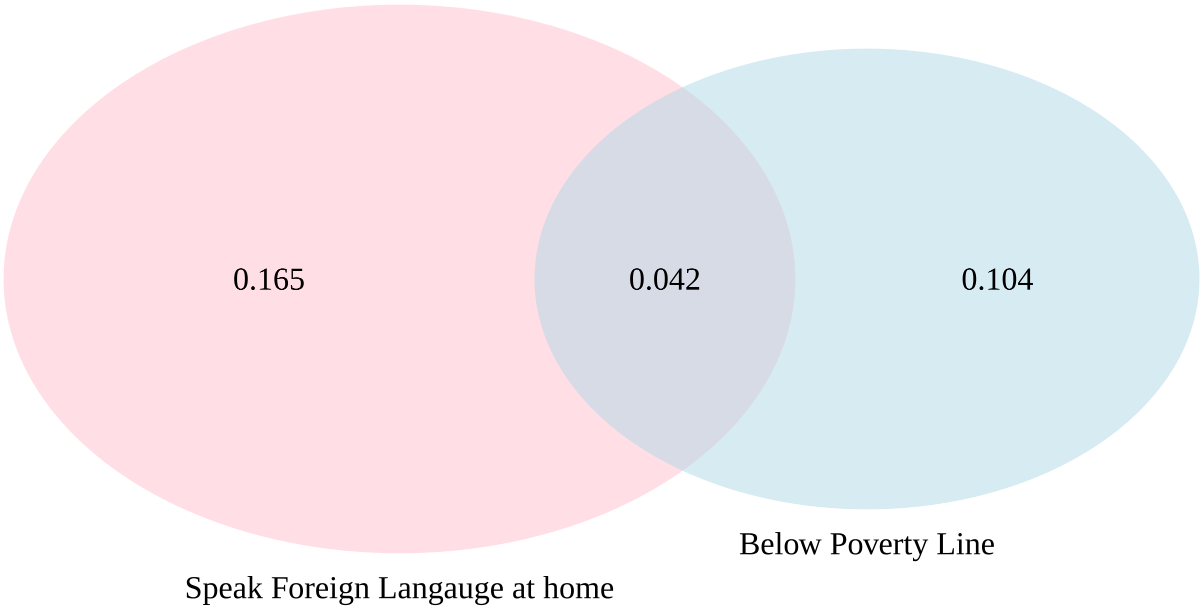
(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

P(below poverty line) = 0.146 P(foreign language) = 0.207 P(below poverty line and speak foreign language) = 0.042

```
library(VennDiagram)
```

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

```
grid.newpage()
draw.pairwise.venn(0.146, 0.207, 0.042, category = c("Below Poverty Line", "Speak Foreign Langauge at ho
    2), fill = c("light blue", "pink"), alpha = rep(0.5, 2), cat.pos = c(0,
    0), cat.dist = rep(0.025, 2))
```



```
## (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? 10.4%

(d) What percent of Americans live below the poverty line or speak a foreign language at home? 31.1%

(e) What percent of Americans live above the poverty line and only speak English at home? 68.8%

(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. If the 2 events were independent, then P(A and B) = P(A) x P(B) i.e. 0.042 = 0.146 x 0.207 But 0.042 <> 0.030 Therefore the 2 events are not independent

Chapter 2 - Introduction to Data, Problem 2.20

```
library(openintro)
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following objects are masked from 'package:datasets':
```

```
##
##     cars, trees
```

```
#d1<-data("assortive.mating")
#str(assortive.mating)
table(assortive.mating)
```

```
##          partner_female
## self_male blue brown green
##     blue    78    23    13
##     brown   19    23    12
##     green   11     9    16
```

```
(d2<-prop.table(table(assortive.mating)))
```

```
##          partner_female
## self_male       blue      brown      green
##     blue  0.38235294 0.11274510 0.06372549
##     brown 0.09313725 0.11274510 0.05882353
##     green 0.05392157 0.04411765 0.07843137
```

```
#str(d2)
```

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

```
#library(plyr)
#count(assortive.mating, 'assortive.mating$self_male')
#(prob.m_blue <- sum(assortive.mating$self_male=="blue")/nrow(assortive.mating))
(mblue=d2[1,1]+d2[1,2]+d2[1,3])
```

```
## [1] 0.5588235
```

```
(mbrown=d2[2,1]+d2[2,2]+d2[2,3])
```

```
## [1] 0.2647059
```

```
(mgreen=d2[3,1]+d2[3,2]+d2[3,3])
```

```
## [1] 0.1764706
```

```
(fblue=d2[1,1]+d2[2,1]+d2[3,1])
```

```
## [1] 0.5294118
```

```
(fbrown=d2[1,2]+d2[2,2]+d2[3,2])
```

```
## [1] 0.2696078
```

```
(fgreen=d2[1,3]+d2[2,3]+d2[3,3])
```

```
## [1] 0.2009804
```

```
cat("The probability of randomly chosen male respondent or his partner having blue eyes is:", round((pr
```

```
## The probability of randomly chosen male respondent or his partner having blue eyes is: 0.7059
```

```
#(prob.m_blue_or_f_blue = mblue+fblue-d2[1,1])
```

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

```
#(mblue)
#colnames(d2)
cat("The probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes
```

```
## The probability that a randomly chosen male respondent with blue eyes has a partner with blue eyes i
```

(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?

```
cat("The probability that a randomly chosen male respondent with brown eyes has a partner with blue eye
```

```
## The probability that a randomly chosen male respondent with brown eyes has a partner with blue eyes
```

```
cat("The probability that a randomly chosen male respondent with green eyes has a partner with blue eye
```

```
## The probability that a randomly chosen male respondent with green eyes has a partner with blue eyes
```

(d) Does it appear that the eye colors of male respondents and their partners are independent? Explain your reasoning. No, it does not appear to be independent because the probablity of a female partner with blue eyes is much lower for males with green and brown eyes as compared to males with blue eyes. If it was independent, these 3 probabilities would be roughly the same. Also, if the eye colors were independent, then probability of a male with blue eyes having a female partner with blue eyes would be the same as the probability of a male with blue eyes multiplied by the probability of a female with blue eyes i.e.

```
d2["blue","blue"] == mblue * fblue
```

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

Given that the above is FALSE, it can be concluded that the eye colors of male respondents and their partners are not independent.

Chapter 2 - Introduction to Data, Problem 2.30

```
library(openintro)
table(books)
```

```
##              format
## type          hardcover paperback
##   fiction            13        59
##   nonfiction         15         8
```

```
(books1<-prop.table(table(books)))
```

```
##              format
## type           hardcover  paperback
##   fiction     0.13684211 0.62105263
##   nonfiction  0.15789474 0.08421053
```

```
(pfict=books1[1,1]+books1[1,2])
```

```
## [1] 0.7578947
```

```
(pnonfict=books1[2,1]+books1[2,2])
```

```
## [1] 0.2421053
```

```
(phardcov=books1[1,1]+books1[2,1])
```

```
## [1] 0.2947368
```

```
(ppaperbk=books1[1,2]+books1[2,2])
```

```
## [1] 0.7052632
```

```
#str(books1)
```

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

```
cat("The probability of drawing a hardcover book first then a paperback fiction book second
when drawing without replacement is:", 28/95 * 59/94)
```

```
## The probability of drawing a hardcover book first then a paperback fiction book second
## when drawing without replacement is: 0.1849944
```

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

The answer is the sum of P(hardcover fiction first and hardcover second) + P(paperback fiction first and hardcover second)

```
cat("The probability of drawing a fiction book first then a hardcover book second
when drawing without replacement is:", (13/95*27/94)+(59/95*28/94), "\n")
```

```
## The probability of drawing a fiction book first then a hardcover book second
## when drawing without replacement is: 0.2243001
```

(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.

```
cat("The probability of drawing a fiction book first then a hardcover book second
when drawing with replacement is:", (72/95*28/95), "\n")
```

```
## The probability of drawing a fiction book first then a hardcover book second
## when drawing with replacement is: 0.2233795
```

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

This is the case because of the low probability of the first fiction book being a hardcover as compared to it being a paperback. The probabilty of drawing a hardcover book second with replacement to that without replacement is a small difference between 28/95 versus 27/94.

Chapter 2 - Introduction to Data, Problem 2.38

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

```
prob.baggage<-c(0.54,0.34,0.12)
fees.baggage<-c(0,25,60)
#(prob.baggage*fees.baggage)
average_fee=weighted.mean(fees.baggage,prob.baggage)
stddev.fees=sqrt(sum(prob.baggage*(fees.baggage - average_fee)^2))

#p.no_bag=0.54
#p.one_bag=0.34
```

```
#p.two_bags=0.12
cat("Average revenue per passenger is $",average_fee, "\n")
```

```
## Average revenue per passenger is $ 15.7
```

```
cat("Standard deviation of revenue per passenger is $", round(stddev.fees,4), "\n")
```

```
## Standard deviation of revenue per passenger is $ 19.9502
```

(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.

```
npass=120
exp.bag.fees = npass*average_fee
stddev.bag.fees = sqrt(npass^2*stddev.fees^2)

cat("Average revenue for the flight is $",exp.bag.fees, "\n")
```

```
## Average revenue for the flight is $ 1884
```

```
cat("Standard deviation of average revenue for the flight is $", round(stddev.bag.fees,4), "\n")
```

```
## Standard deviation of average revenue for the flight is $ 2394.023
```

Chapter 2 - Introduction to Data, Problem 2.44

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:openintro':
##
##     diamonds
```

```
income_bands<-c("$1 to $9,999 or less", "$10,000 to $14,999", "$15,000 to $24,999", "$25,000 to $34,999
freq<-c(0.022, 0.047, 0.158, 0.183, 0.212, 0.139, 0.058, 0.084, 0.097)
(income.df<-data.frame(income_bands,freq))
```

```
##          income_bands  freq
## 1 $1 to $9,999 or less 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 $49,999 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
```

```
#ggplot(data=income.df)+geom_bar(mapping=aes(x=income_bands,y=freq))
#summary(income.df)
```

(a) Describe the distribution of total personal income. The distribution of total personal income is a little right-skewed. The median seems to be in the $35,000 to $49,999 income band.

(b) What is the probability that a randomly chosen US resident makes less than $50,000 per year? This probability is = 0.022+0.047+0.158+0.183+0.212 = 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. This probability is = p(<50k) x p(female) = 0.622 * 0.41 = 0.25502 The underlying assumption here is that these 2 outcomes are independent i.e. the probability of making less than $50K is not dependent on the gender of the survey respondent.

(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.

Given the above data, the the probability of making less than $50K and being female is 0.718 * 0.41 = 0.294380 Given the large sample size, this is a significant difference from the expected probability had these 2 been independent events (as calculated in c above). So the assumption made in (c) above does not seem to be valid.