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? -zero, because the lowest value on a die is 1, so the minimum sum can be 2
(b) getting a sum of 5?

```
#ways to get 5
#1-4, 4-1
#2-3,3-2

p1<-1/6
p4when1<-p1*1/6
p1when4<-p4when1
p2when3<-p4when1
p3when2<-p2when3

psum5<-p1when4+p4when1+p2when3+p3when2
psum5
```

```
## [1] 0.1111111
```

(c) getting a sum of 12? You must roll 2 sixes to get a twelve so,

```
p6<-1/6
psum12<-p6*p6
psum12
```

```
## [1] 0.02777778
```

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.

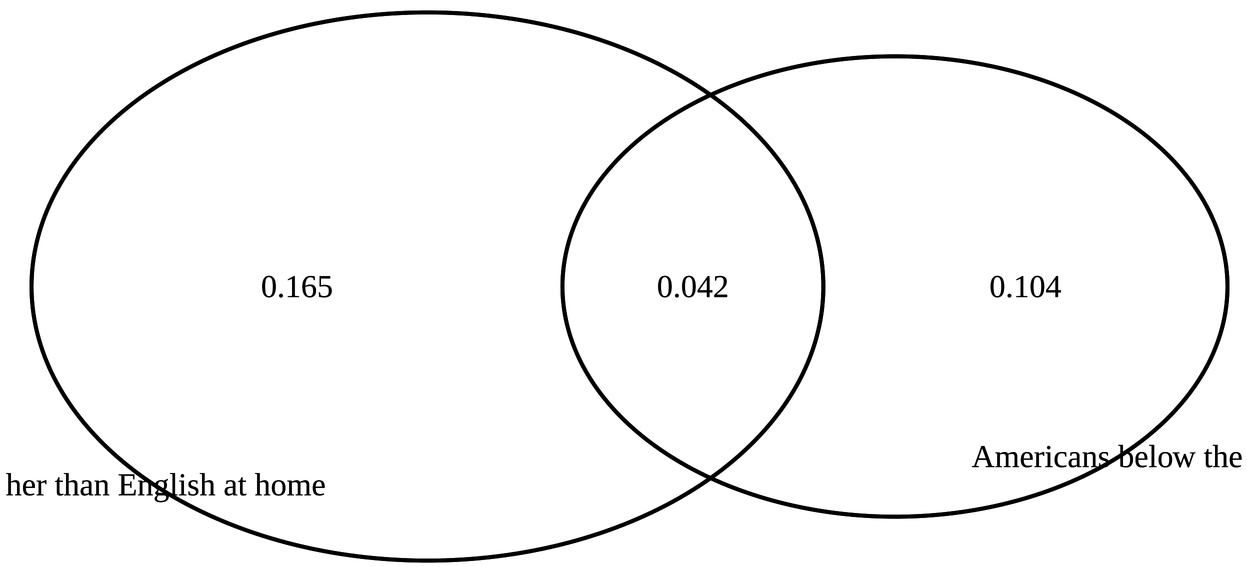
(a) Are living below the poverty line and speaking a foreign language at home disjoint?

No, because there is an intersection between the categories

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

```
venn<-draw.pairwise.venn(area1 = .146,area2 =.207,cross.area = .042, category = c("Americans below the p


grid.draw(venn)
```



(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?

.146+.207-.042=.311

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

P(above and only english)=1-.311=.689

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

P(below| FL)=P(below and FL)/P(FL)

.042/.207=.203

if they were independent than P(below-FL)=P(below) and .203 does not equal .311

so they are 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?

```
maleBlue<-114/204
femaleBlue<-108/204
bothBlue<-78/204

p<-maleBlue+femaleBlue-bothBlue

p
```

```
## [1] 0.7058824
```

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

```
pfBmB<-78/114

pfBmB
```

```
## [1] 0.6842105
```

(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) .35
(e) .31

```
pmBrFBl<-19/54
pmBrFBl
```

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

```
pmGFBl<-11/36
pmGFBl
```

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

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

```r
r1<-c(78,23,13)
r2<-c(19,23,12)
r3<-c(11,9,16)

dat<-rbind(r1,r2,r3)
dat<-prop.table(dat)
dat<-cbind(dat,Total=rowSums(dat))
dat<-rbind(dat,Total=colSums(dat))
dat
```

```
##                                             Total
## r1    0.38235294 0.11274510 0.06372549 0.5588235
## r2    0.09313725 0.11274510 0.05882353 0.2647059
## r3    0.05392157 0.04411765 0.07843137 0.1764706
## Total 0.52941176 0.26960784 0.20098039 1.0000000
```

Since the joint distributions of eye color for partner are not equal across a variable, there is a relationship between eye color and choice of mate's eye color.

Looking at the first column, we can see that 70 percent of blue eyed women selected blue eyed mates, although this corrolation seems to be strongest in blue eyed people and weakest in green eyed people.

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.

```
28/95* 59/94
```

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

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

```
72/95 * 28/94
```

```
## [1] 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.

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

(e)

```
72/95*28/95
```

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

This number is very similar because the sample size is large, and the lowering of the population by 1 doesn't have much of an effect.

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.

```
first_bag<-25
second_bag<-35
two_bag<-first_bag+second_bag

#X*P(X)

E1<-.34*first_bag
E2<-.12*two_bag

EX<-E1+E2

EX
```

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

```
#Var

var<-(0-15.7)^2*.54+(25-15.7)^2*.34+(60-15.7)^2*.12




sd<-sqrt(var)

c("variance"=var,"standard deviation"=sd)
```

```
##           variance standard deviation
##          398.01000           19.95019
```

(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 E(X)is expected value per passenger so you can multiply it by the number of passengers. The SD will be the same

```
c("Expected Value"=120*EX, "sd"=sd)
```

```
## Expected Value             sd
##     1884.00000       19.95019
```

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?

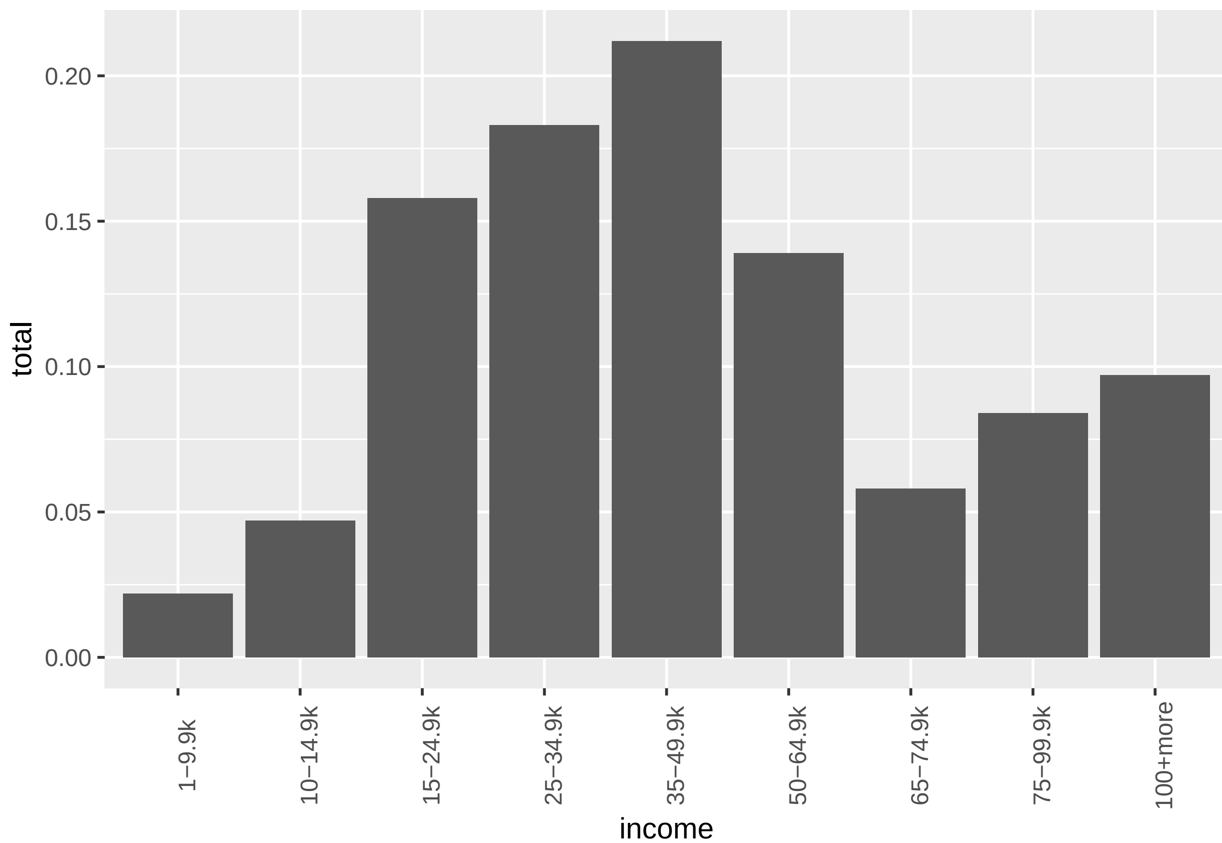
```r
income<-c("1-9.9k","10-14.9k","15-24.9k","25-34.9k","35-49.9k","50-64.9k","65-74.9k","75-99.9k","100+mo


total<-c(.022,.047,.158,.183,.212,.139,.058,.084,.097)

dat<-data.frame("income"=income,"total"=total)
dat$income<-factor(dat$income,levels=c("1-9.9k","10-14.9k","15-24.9k","25-34.9k","35-49.9k","50-64.9k",



ggplot(dat, aes(income, total))+geom_bar(stat="identity")+theme(axis.text.x = element_text(angle = 90))
```



The graph looks right skewed because of the larger proportion of very high salaries.

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

```
inc<-c(.022,.047,.158,.183,.212)

under50<-sum(inc)

under50*.41
```

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

If I assume men and women are distributed evenly, then I would make a tree diagram and multiply the proportion that makes less than 50k and the proportion of women.

(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.

If this is true, then 71.8 percent of this proportion would be female.

```
female_under_50<-under50*.718
female_under_50
```

```
## [1] 0.446596
```

(c) If .41 percent of females participated, and .622 of people make under 50, then, according to the sample a proportion of .255 of the women make under 50k.

If in the real world a .718 proportion exists and .467 of the US labor force is female, then a proportion of .33 of the working population should make less than 50k.

This is a very large discrepancy, and I would either think the data was erroneous or the sample was not large enough to capture this reality.