MichaelY_HW2_Probability

Michael Y.

February 17, 2019

Homework - Chapter 2 - Probability -

Exercises 2.6, 2.8, 2.20, 2.30, 2.38, 2.44 (pp.116-126)

Datasets:

2.14 - health_coverage

2.17 - global_warming_pew (approximate from poll)

2.18 - health_coverage

2.19 - burger

2.20 - assortive_mating

2.30 - books

2.43 - cats

Exercise 2.6 - Dice rolls.

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

(a) getting a sum of 1?

This result is not possible, so the probability is zero.

(b) getting a sum of 5?

You can get a sum of 5 by rolling (1,4), (2,3), (3,2), or (4,1) on the pair of dice.
This is 4 permutations out of a possible 6*6 = 36 outcomes, so 4/36 = 1/9 = 0.1111111 .

(c) getting a sum of 12?

You can only get this by rolling (6,6), which is 1 chance out of 36, so the result is 1/36 = 0.02777778 .

Exercise 2.8 Poverty and language.

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, they are not disjoint, as 4.2% fall into both categories.

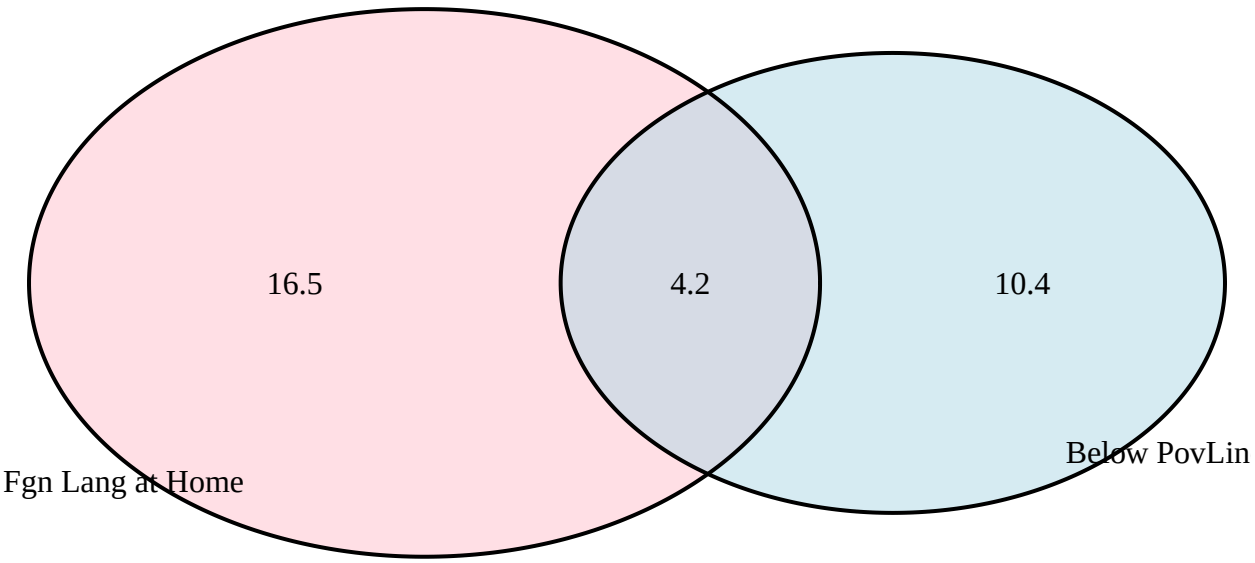
(b) Draw a Venn diagram summarizing the variables and their associated probabilities.

```
###install.packages('VennDiagram')
library(VennDiagram)
```

```
## Loading required package: grid
```

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

```
grid.newpage()
draw.pairwise.venn(area1 = 14.6, area2 = 20.7, cross.area = 4.2,
                   fill = c("light blue", "pink"),
                   category = c("Below PovLine     ", "          Fgn Lang at Home"))
```



```
## (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?

$$\mathbb{P}\left(BelowPovertyLine \wedge EnglishAtHome\right) = \mathbb{P}\left(BelowPovertyLine \wedge !FgnLangAtHome\right) = 14.6\% - 4.2\% = 10.4\%$$

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

$$\mathbb{P}\left(BelowPovertyLine \vee FgnLangAtHome\right) = \mathbb{P}\left(BelowPovertyLine\right) + \mathbb{P}\left(FgnLangAtHome\right) - \mathbb{P}\left(BelowPovertyLine \wedge Fg$$

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

$$\mathbb{P}\left(AbovePovertyLine \wedge EnglishAtHome\right) = 100\% - \mathbb{P}\left(BelowPovertyLine \vee FgnLangAtHome\right) = 100\% - 31.1\% = 68.9\%$$

(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:

$$\mathbb{P}\left(BelowPovertyLine \wedge FgnLangAtHome\right) = 4.2\%$$
$$\mathbb{P}\left(BelowPovertyLine\right) * \mathbb{P}\left(FgnLangAtHome\right) = 14.6\% * 20.7\% = 3\%$$

As the above quantities are not equal, the events are not independent.

Exercise 2.20 Assortative mating.

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.

```
# load the data supplied by the text
library(openintro)
```

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

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

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

```
data("assortive.mating")
# compute a table of counts of pairs
mytable <- table(assortive.mating)
mytable
```

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

```
# confirm that I've got the same figures:
eyes=matrix(c(78,23,13,
              19,23,12,
              11,9,16),
            3,3,byrow = T,
            dimnames = list(c("Male-Blue","Male-Brown","Male-Green"),
                            c("Fem-Blue","Feb-Brown","Fem-Green")))
eyes
```

```
##            Fem-Blue Feb-Brown Fem-Green
## Male-Blue        78        23        13
## Male-Brown       19        23        12
## Male-Green       11         9        16
```

```
# compute the sums of the columns
with_colsums=rbind(eyes,COLSUMS=colSums(eyes))
# compute the sums of the rows
bigeyes = cbind(with_colsums,ROWSUMS=rowSums(with_colsums))
bigeyes
```

```
##            Fem-Blue Feb-Brown Fem-Green ROWSUMS
## Male-Blue        78        23        13     114
## Male-Brown       19        23        12      54
## Male-Green       11         9        16      36
## COLSUMS         108        55        41     204
```

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

The question asks whether either member of any randomly chosen couple has blue eyes.
This means either:
the **male** has blue eyes, or his **partner** has blue eyes, or **both** have blue eyes.

The probability of the union (either has blue eyes) is equal to:
the probability that the male has blue eyes,
PLUS the probability that the female has blue eyes,
MINUS the probability that BOTH have blue eyes (to prevent double-counting.)

```
male_blue = sum(mytable["blue",])
male_blue
```

```
## [1] 114
```

```
female_blue = sum(mytable[,"blue"])
female_blue
```

```
## [1] 108
```

```
both_blue = mytable["blue","blue"]
both_blue
```

```
## [1] 78
```

```
total = sum(mytable)
total
```

```
## [1] 204
```

```
pct_at_least_one_blue = (male_blue + female_blue - both_blue ) / total
pct_at_least_one_blue
```

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

$$\mathbb{P}(MaleBlue \vee FemBlue) = \mathbb{P}(MaleBlue) + \mathbb{P}(FemBlue) - \mathbb{P}(MaleBlue \wedge FemBlue) = \frac{114}{204} + \frac{108}{204} - \frac{78}{204} = \frac{144}{204} = 0.7059$$

Thus there are 144 pairs, out of the 204 total, where one or more of the members has blue eyes; the ratio is 0.7058824.

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

Conditional probability, via Bayes' Rule:

There are 114 male respondents with **blue** eyes, all are represented across the **top** row of the grid.
Randomly choose one of those 114 blue-eyed males; the number of those with a **blue-eyed partner** is 78 .

The percentage is 0.6842105 .

$$\mathbb{P}(FemBlue|MaleBlue) * \mathbb{P}(MaleBlue) = \mathbb{P}(MaleBlue \wedge FemBlue)$$

$$\mathbb{P}(FemBlue|MaleBlue) = \frac{\mathbb{P}(MaleBlue \wedge FemBlue)}{\mathbb{P}(MaleBlue)} = \frac{\left(\frac{78}{204}\right)}{\left(\frac{114}{204}\right)} = \frac{78}{114} = .6842$$

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

There are 54 male respondents with **brown** eyes, all are represented across the **second** row of the grid.
Randomly choose one of those 54 brown-eyed males; the number of those with a blue-eyed partner is 19 .
The probability is 0.3518519 .

$$\mathbb{P}(FemBlue|MaleBrown) * \mathbb{P}(MaleBrown) = \mathbb{P}(MaleBrown \wedge FemBlue)$$

$$\mathbb{P}(FemBlue|MaleBrown) = \frac{\mathbb{P}(MaleBrown \wedge FemBlue)}{\mathbb{P}(MaleBrown)} = \frac{\left(\frac{19}{204}\right)}{\left(\frac{54}{204}\right)} = \frac{19}{54} = .3519$$

What about the probability of a randomly chosen male respondent with green eyes having a partner with blue eyes?

There are 36 male respondents with **green** eyes, all are represented across the **third** row of the grid.
Randomly choose one of those 36 green-eyed males; the number of those with a blue-eyed partner is 11.

The probability is 0.3055556 .

$$\mathbb{P}(FemBlue|MaleGreen) * \mathbb{P}(MaleGreen) = \mathbb{P}(MaleGreen \wedge FemBlue)$$

$$\mathbb{P}\left(FemBlue|MaleGreen\right) = \frac{\mathbb{P}\left(MaleGreen \wedge FemBlue\right)}{\mathbb{P}\left(MaleGreen\right)} = \frac{\left(\frac{11}{204}\right)}{\left(\frac{36}{204}\right)} = \frac{11}{36} = .3056$$

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

No, the eye colors of male respondents and their partners are not independent.

Explain your reasoning.

If **independent**, then the probabilities would multiply, i.e.,

$$\mathbb{P}\left(MaleBlue\right) * \mathbb{P}\left(FemBlue\right) = \mathbb{P}\left(MaleBlue \wedge FemBlue\right)$$

But, we do not have this:

$$\left(\frac{114}{204}\right) * \left(\frac{108}{204}\right) \neq \left(\frac{78}{204}\right).5588 * .5294 \neq .3824.2958 \neq .3824$$

What we do have is:

$$\mathbb{P}\left(MaleBlue\right) * \mathbb{P}\left(FemBlue|MaleBlue\right) = \mathbb{P}\left(MaleBlue \wedge FemBlue\right) \left(\frac{114}{204}\right) * \left(\frac{78}{114}\right) = \left(\frac{78}{204}\right).5588 * .6842 = .3824.3824 =$$

This illustrates the conditional dependence.

In order to have independence, the values of the initial matrix would have to be rearranged as follows (keeping the totals in each row and column the same, and ignoring the fact that the results are not integers, thus implying "fractional people") :

```
ROWSUMS=rowSums(eyes)
COLSUMS=colSums(eyes)
TOTAL=sum(eyes)
INDEPeyes=(ROWSUMS %o% COLSUMS) / sum(eyes)
### These are the values which would be required for independence:
INDEPeyes
```

```
##            Fem-Blue Feb-Brown Fem-Green
## Male-Blue  60.35294 30.735294 22.911765
## Male-Brown 28.58824 14.558824 10.852941
## Male-Green 19.05882  9.705882  7.235294
```

```
INDEPfoo=rbind(INDEPeyes,COLSUMS=colSums(INDEPeyes))
bigINDEPeyes = cbind(INDEPfoo,ROWSUMS=rowSums(INDEPfoo))
### This demonstrates that the row and column sums are the same as the initial grid
bigINDEPeyes
```

```
##             Fem-Blue Feb-Brown Fem-Green ROWSUMS
## Male-Blue   60.35294 30.735294 22.911765     114
## Male-Brown  28.58824 14.558824 10.852941      54
## Male-Green  19.05882  9.705882  7.235294      36
## COLSUMS    108.00000 55.000000 41.000000     204
```

```
### Let's try rounding the results to get integer "people", as numerically close to independence as poss
roundINDEPeyes = round(INDEPeyes,0)
roundINDEPfoo=rbind(roundINDEPeyes,COLSUMS=colSums(roundINDEPeyes))
bigroundINDEPeyes = cbind(roundINDEPfoo,ROWSUMS=rowSums(roundINDEPfoo))
###The rounding has caused the sums in the middle row and middle column to increase by 1 vs. the origina
bigroundINDEPeyes
```

```
##            Fem-Blue Feb-Brown Fem-Green ROWSUMS
## Male-Blue        60        31        23     114
## Male-Brown       29        15        11      55
## Male-Green       19        10         7      36
## COLSUMS         108        56        41     205
```

```
### adjust the rounding to get back to the same overall totals
adjustment = -0.06
roundINDEPeyes = round(INDEPeyes+adjustment,0)
roundINDEPfoo=rbind(roundINDEPeyes,COLSUMS=colSums(roundINDEPeyes))
bigroundINDEPeyes = cbind(roundINDEPfoo,ROWSUMS=rowSums(roundINDEPfoo))
### This is the closest integer result to perfect independence
bigroundINDEPeyes
```

```
##            Fem-Blue Feb-Brown Fem-Green ROWSUMS
## Male-Blue        60        31        23     114
## Male-Brown       29        14        11      54
## Male-Green       19        10         7      36
## COLSUMS         108        55        41     204
```

```
### These are the number by which each original number needs to be adjusted to yield independence
changes = bigroundINDEPeyes-bigeyes
changes
```

```
##            Fem-Blue Feb-Brown Fem-Green ROWSUMS
## Male-Blue       -18         8        10       0
## Male-Brown       10        -9        -1       0
## Male-Green        8         1        -9       0
## COLSUMS           0         0         0       0
```

Looking just at the case of Blue eyes, under the above adjustments, we would (approximately) have:

$$\mathbb{P}\left(FemBlue\right) * \mathbb{P}\left(MaleBlue\right) \approx \mathbb{P}\left(MaleBlue \wedge FemBlue\right)$$

$$\left(\frac{108}{204}\right) * \left(\frac{114}{204}\right) \approx \left(\frac{60}{204}\right) .5294 * .5588 \approx .2941 .2958 \approx .2941$$

This illustrates that, up to the necessary rounding to integer "people", independence of eye-color selection would exist under the above combinations. (It would have to be demonstrated that the above approximate equality prevailed for every pair, but the above example has been constructed in such a way to force this.)

Exercise 2.30 - Books on a bookshelf.

The table below shows the distribution of books on a bookcase based on whether they are nonfiction or fiction and hardcover or paperback.

```
books=matrix(c(13,59, 15, 8),2,2,byrow = T,
             dimnames = list(c("Fiction","Nonfiction"),
```

```
                              c("Hardcover","Paperback")))
books
```

```
##            Hardcover Paperback
## Fiction           13        59
## Nonfiction        15         8
```

```
with_colsums=rbind(books,COLSUMS=colSums(books))
bigbooks = cbind(with_colsums,ROWSUMS=rowSums(with_colsums))
bigbooks
```

```
##            Hardcover Paperback ROWSUMS
## Fiction           13        59      72
## Nonfiction        15         8      23
## COLSUMS           28        67      95
```

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

The probability of drawing a **hardcover** book first is 0.2947368

$$\mathbb{P}\left(HardcoverFirst\right) = \left(\frac{28}{95}\right) = .2947$$

We don't know whether the first book drawn was a hardcover **fiction** book or a hardcover **non-fiction** book, but it turns out that this does not matter.
If the first book drawn was **hardcover fiction**, then the updated shelf is:

```
updatedbooks_aHF = bigbooks + matrix(c(-1,0,-1,0,0,0,-1,0,-1),3,3,byrow = T)
updatedbooks_aHF
```

```
##            Hardcover Paperback ROWSUMS
## Fiction           12        59      71
## Nonfiction        15         8      23
## COLSUMS           27        67      94
```

If the first book drawn was **hardcover nonfiction**, then the updated shelf is:

```
updatedbooks_aHN = bigbooks + matrix(c(0,0,0,-1,0,-1,-1,0,-1),3,3,byrow = T)
updatedbooks_aHN
```

```
##            Hardcover Paperback ROWSUMS
## Fiction           13        59      72
## Nonfiction        14         8      22
## COLSUMS           27        67      94
```

Regardless of which of the two cases we are in, the probability of drawing a **paperback fiction** book second (***without*** replacement) is 0.6276596

$$\mathbb{P}\left(PaperbackFictionSecond|HardcoverFirst\right) = \left(\frac{59}{94}\right) = .6277$$

The probability of **both** of the events happening is 0.1849944

$$\mathbb{P}\left(HardcoverFirst\right)*\left(PaperbackFictionSecond|HardcoverFirst\right) = \mathbb{P}\left(HardcoverFirst \wedge PaperbackFictionSecond\right) = \left(\frac{2}{9}\right.$$

```
bigbooks[3,1]/bigbooks[3,3]
```

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

```
bigbooks[1,2]/(bigbooks[3,3]-1)
```

```
## [1] 0.6276596
```

```
bigbooks[3,1]/bigbooks[3,3]*bigbooks[1,2]/(bigbooks[3,3]-1)
```

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

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

This is a more difficult problem, because we don't know whether the first book drawn was a **paperback fiction** or a **hardcover fiction**.

This presents two cases.

$$\mathbb{P}\left(FictionFirst \wedge HardcoverSecond\right) = \mathbb{P}\left(PaperbackFictionFirst \wedge HardcoverSecond\right) + \mathbb{P}\left(HardcoverFictionFirst \wedge Ha$$

Case 1:

If the first book drawn was **paperback fiction** :
The probability of this happening is 0.6210526

$$\mathbb{P}\left(PaperbackFictionFirst\right) = \left(\frac{59}{95}\right) = .6211$$

Then the remaining books on the shelf are:

```
updatedbooks_bPF = bigbooks + matrix(c(0,-1,-1,0,0,0,0,-1,-1),3,3,byrow = T)
updatedbooks_bPF
```

```
##            Hardcover Paperback ROWSUMS
## Fiction           13        58      71
## Nonfiction        15         8      23
## COLSUMS           28        66      94
```

```
updatedbooks_bPF[3,1]/updatedbooks_bPF[3,3]
```

```
## [1] 0.2978723
```

In this case, drawing a **hardcover** book second (***without*** replacement) has probability 0.2978723

$$\mathbb{P}\left(HardcoverSecond | PaperbackFictionFirst\right) = \left(\frac{28}{94}\right) = .2979$$

so the probability of both of these events happening is 0.1849944

$$\mathbb{P}\left(PaperbackFictionFirst\right) * \mathbb{P}\left(HardcoverSecond | PaperbackFictionFirst\right) = \left(\frac{59}{95}\right) * \left(\frac{28}{94}\right) = .6211 * .2979 = .1850$$

Case 2:

If the first book drawn was **hardcover fiction** :
The probability of this happening is 0.1368421

$$\mathbb{P}\left(HardcoverFictionFirst\right) = \left(\frac{13}{95}\right) = .1368$$

Then the remaining books on the shelf are:

```
updatedbooks_bHF = bigbooks + matrix(c(-1,0,-1,0,0,0,-1,0,-1),3,3,byrow = T)
updatedbooks_bHF
```

```
##            Hardcover Paperback ROWSUMS
## Fiction           12        59      71
## Nonfiction        15         8      23
## COLSUMS           27        67      94
```

```
updatedbooks_bHF[3,1]/updatedbooks_bHF[3,3]
```

```
## [1] 0.287234
```

In this case, drawing a **hardcover** book second (***without*** replacement) has probability 0.287234

$$\mathbb{P}\left(HardcoverSecond|HardcoverFictionFirst\right) = \left(\frac{27}{94}\right) = .2872$$

so the probability of both of these events happening is 0.0393057

$$\mathbb{P}\left(HardcoverFictionFirst\right) * \mathbb{P}\left(HardcoverSecond|HardcoverFictionFirst\right) = \left(\frac{13}{95}\right) * \left(\frac{27}{94}\right) = .1368 * .2872 = .0393$$

As indicated above, the final result is 0.2243001

```
bigbooks[1,2]/bigbooks[3,3]*updatedbooks_bPF[3,1]/updatedbooks_bPF[3,3] +bigbooks[1,1]/bigbooks[3,3]*up
```

```
## [1] 0.2243001
```

$$\mathbb{P}\left(FictionFirst \wedge HardcoverSecond\right) = \mathbb{P}\left(PaperbackFictionFirst \wedge HardcoverSecond\right) + \mathbb{P}\left(HardcoverFictionFirst \wedge Ha$$

(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.

This scenario is easier because the two events are independent.

The probability of drawing a **fiction** book first is 0.7578947

$$\mathbb{P}\left(FictionFirst\right) = \left(\frac{72}{95}\right) = .7579$$

Because the book is replaced on the shelf, the second draw is independent of the result of the first.

The probability of drawing a **hardcover** book second is 0.2947368

$$\mathbb{P}\left(HardcoverSecond\right) = \left(\frac{28}{95}\right) = .2947$$

Because of the independence of the events, the desired result is 0.2233795

$$\mathbb{P}\left(FictionFirst\right) * \mathbb{P}\left(HardcoverSecond\right) = \left(\frac{72}{95}\right) * \left(\frac{28}{95}\right) = .7579 * .2947 = .2234$$

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

For part (b), without replacement, the result is

$$\left(\frac{59}{95}\right) * \left(\frac{28}{94}\right) + \left(\frac{13}{95}\right) * \left(\frac{27}{94}\right) = \left(\frac{59 * 28 + 13 * 27}{8930}\right) = \left(\frac{2003}{8930}\right) = .2243$$

For part (c), with replacement, the result is

$$\left(\frac{72}{95}\right) * \left(\frac{28}{95}\right) = \left(\frac{2016}{9025}\right) = .2234$$

The difference between sampling *with* replacement vs. *without* replacement becomes more significant if the size of the population is quite small, and if the total number of items in the sample is more than five percent of the population. Here there are 95 books in the population but we are only taking 2 books, which is about 2 percent of the population. In the case where a larger sample is drawn without replacement, the difference would become more apparent. In such circumstances, a technique known as "**Finite population correction factor**" is recommended.

Exercise 2.38 – Baggage fees.

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.

```
baggagefee=c(0,25,60)
passengerpct=c(0.54, 0.34, 0.12)
revenue_a= baggagefee * passengerpct
revenue_a
```

```
## [1] 0.0 8.5 7.2
```

```
revenue_mu = sum(revenue_a)
revenue_mu
```

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

```
baggagefee_mu = baggagefee - revenue_mu
baggagefee_mu
```

```
## [1] -15.7   9.3  44.3
```

```
baggagefee_mu2 = baggagefee_mu^2
weighted_baggagefee_mu2 = baggagefee_mu2  * passengerpct
variance_bagfee = sum(weighted_baggagefee_mu2)
stdev_bagfee = sqrt(variance_bagfee)
stdev_bagfee
```

```
## [1] 19.95019
```

Average revenue per passenger is 15.7 . Standard Deviation is 19.950188 .

(b) About how much revenue should the airline expect for a flight of 120 passengers?

Revenue for a flight of 120 passengers should be about $1884 .

With what standard deviation?

Standard deviation of the above revenue estimate for 120 passengers would be $218.5433595 .

Note any assumptions you make and if you think they are justified.

The above calculations do not round the results to an integer number of passengers in each baggage group, so they allow for "fractional passengers."

Exercise 2.44 Income and gender.

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.

```
percents = c(2.2, 4.7, 15.8, 18.3, 21.2, 13.9, 5.8,8.4, 9.7)
cuts = c(9999,14999,24999,34999,49999,64999,74999,99999,999999)
```

(a) Describe the distribution of total personal income.

It is a right-skewed distribution with a peak in the 35,000 - 50,000 bracket

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

```
low_percents <- percents[cuts<50000]
low_percents
```

```
## [1]  2.2  4.7 15.8 18.3 21.2
```

```
cumsum(low_percents)
```

```
## [1]  2.2  6.9 22.7 41.0 62.2
```

```
prob_below_50k <- sum(low_percents)
```

The probability that a randomly chosen US resident makes less than $50,000 per year is 62.2 percent .

$$\mathbb{P}\left(LowIncome\right) = .622$$

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

$$\mathbb{P}\left(Female\right) = .41$$

Assuming independence of income level vs. gender, the probablility of the above would be 62.2% * 41% = 25.502%

$$\mathbb{P}\left(LowIncome\right) * \mathbb{P}\left(Female\right) = \mathbb{P}\left(LowIncome \wedge Female\right) \ = .622 * .41 = .25502$$

Note any assumptions you make.

This assumes independence of income and gender, i.e.,

$$\mathbb{P}\left(LowIncome|Female\right) = \mathbb{P}\left(LowIncome\right) = .622$$

(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 above assumed that the 62.2% low-income applied equally to both females and males.

This means that the above assumption of independence is not valid, as

the probability of (low income and female) is not equal to (the probability of low income) * (probability of female.)

$$\mathbb{P}\left(Female\right) * \mathbb{P}\left(LowIncome|Female\right) = \mathbb{P}\left(LowIncome \wedge Female\right) \ = .41 * .718 = .29438$$

$$\mathbb{P}\left(LowIncome\right) * \mathbb{P}\left(Female\right) = \mathbb{P}\left(LowIncome \wedge Female\right) .622 * .41 = .25502 \neq .29438$$

Since the probabilities do not multiply, the variables (income and gender) are not independent.

The non-independent scenario described is consistent with the following table, based upon 100,000 people:

```
income=matrix(c(26238,11562,
                32762,29438),
              2,2,byrow = T,
              dimnames = list(c("Hi_Income","Low_Income"),
                              c("Male","Female")))
SumByGender=rbind(income,ByGender=colSums(income))
income_table = cbind(SumByGender,ByIncome=rowSums(SumByGender))
income_table
```

```
##             Male Female ByIncome
## Hi_Income  26238  11562    37800
## Low_Income 32762  29438    62200
## ByGender   59000  41000   100000
```

Here, the percentage of low-income females is 29438 divided by 41000 = 0.718,

while the percentage of low-income males is 32762 divided by 59000 = 0.5552881

If the income level were independent of gender, then the distribution would have to be as follows:

```
incomeindep=matrix(c(22302,15498,
                     36698,25502),
                     2,2,byrow = T,
            dimnames = list(c("Hi_Income","Low_Income"),
                            c("Male","Female")))
IndepSumByGender=rbind(incomeindep,ByGender=colSums(incomeindep))
incomeindep_table = cbind(IndepSumByGender,ByIncome=rowSums(IndepSumByGender))
incomeindep_table
```

```
##             Male Female ByIncome
## Hi_Income  22302  15498    37800
## Low_Income 36698  25502    62200
## ByGender   59000  41000   100000
```

Here, the percentage of low-income females would be 25502 divided by 41000 = 0.622,

which matches the percentage of low-income males: 36698 divided by 59000 = 0.622

Achieving such equalization would require the shifting of nearly 4000 individuals of each gender between income groupings (based upon 100,000 people):

```
incomeindep - income
```

```
##             Male Female
## Hi_Income  -3936   3936
## Low_Income  3936  -3936
```