DATA 602 Assignment 1

Michael Ellsworth, ID 30101253

September 16, 2019

Question 1

A recent survey [1] of employed Canadians found that 40%, or 4-in-10, would find it difficult to meet their financial obligations if their paycheque was delayed by one-week. You are to randomly select two employeed Canadians. Compute the probability that:

a.

Both would find it difficult to meet their financial obligations if their paycheque was delayed by one-week.

$$
\begin{aligned}
P(\text{both find it difficult}) &= P(Diff_1 \cap Diff_2) \\
&= P(Diff_1) * P(Diff_2) \text{ (Both events are independent)} \\
&= 0.4 * 0.4 \text{ (The probability of each event is the same)} \\
&= 0.16
\end{aligned}
$$

```
#alternative
dbinom(2, 2, 0.4)
```

```
## [1] 0.16
```

b.

Neither would find it difficult to meet their financial obligations if their paycheque was delayed by one-week.

$$
\begin{aligned}
P(\text{neither find it difficult}) &= P(Diff_1^c \cap Diff_2^c) \\
&= P(Diff_1^c) * P(Diff_2^c) \text{ (both events are independent)} \\
&= (1 - P(1 - Diff_1)) * (1 - P(Diff_2^c)) \text{ (The complement } P(Diff_1^c) = 1 - P(\\
&= 0.6 * 0.6 \\
&= 0.36
\end{aligned}
$$

---

[1] (https://www.thestar.com/news/canada/2018/09/05/data-shows-fewer-canadians-are-living-paycheque-to-paycheque-but-more-are-overwhelmed-by-debt.html)

```r
#alternative
dbinom(0, 2, 0.4)

## [1] 0.36
```

c.

At least one of the two would find it difficult to meet their financial obligations if their paycheque was delayed by one-week.

$$
\begin{aligned}
P(\text{at least one find it difficult}) &= P(Diff_1 \cup Diff_2) \\
&= P(Diff_1) + P(Diff_2) - P(Diff_1 \cap Diff_2) \\
&= P(Diff_1) + P(Diff_2) - P(Diff_1) * P(Diff_2) \\
&= 0.4 + 0.4 - 0.4 * 0.4 \\
&= 0.64
\end{aligned}
$$

```r
#alternative
dbinom(1, 2, 0.4) + dbinom(2, 2, 0.4)

## [1] 0.64
```

d.

Suppose you are to randomly pick n-employed Canadians in such a way that the probability of ***at least one of them*** *will not be able to meet their financial obligations if their paycheque is delayed by one -week is 0.95. Compute the minimum number of employed Canadians you would have to randomly select. In other words, compute the* ***sample size*** *n. (Hint: $ln(a^b) = b * ln(a)$...)*

This problem can be solved with a while loop that tests the probability of n until the Probability exceeds 0.95.

```r
# Solve part d via while loop
# Start by testing the minimum sample size "n" at 1
n <- 1
P <- 0
# Set the condition of the while loop so that it runs until the probability
exceeds 0.95
while(P < 0.95){
  P <- 0
  # Run a for loop to calculate the cumulative probability
  for(i in 1:n){
    # Sums each binomial distribution from 1 to n
```

```
    # This does not include 0 as we need at least 1 person unable to meet
their financial obligations
    P <- P + dbinom(i, n, 0.4)
  }
  # After the probability is calculated, an if statement determines if we
should increase n
  if(P < 0.95){
    n <- n + 1
  }
  # If the probability exceeds 0.95, then we have our minimum "n"!
  else{
    n
  }
}
# Returns the probability of the calculated minimum "n""
P
```

```
## [1] 0.953344
```

```
# Returns the minimmum sample size "n"
n
```

```
## [1] 6
```

Question 2

For Question 2, you are asked to create the following simulation: Toss a fair-die 1000 times then compute the sum of the 1000 tosses. For example, ${\cal S} = \{Toss1, Toss2, \cdots, Toss1000 \}$. Then $\sum_{i=1}^{1000} T oss_i =$?

Step 1: Create a series of vectors to hold output

```
nsims = 1000
outcome = numeric(nsims)
```

Step 2: Run the Simulation

```
for(i in 1:nsims){
  outcome[i] = sample(c(1,2,3,4,5,6), 1, replace=FALSE)
  }
simresult = data.frame(outcome)
head(simresult,3)
```

```
##   outcome
## 1       2
## 2       6
## 3       5
```

```
tail(simresult,3)
```

```
##      outcome
## 998        1
## 999        3
## 1000       1
```

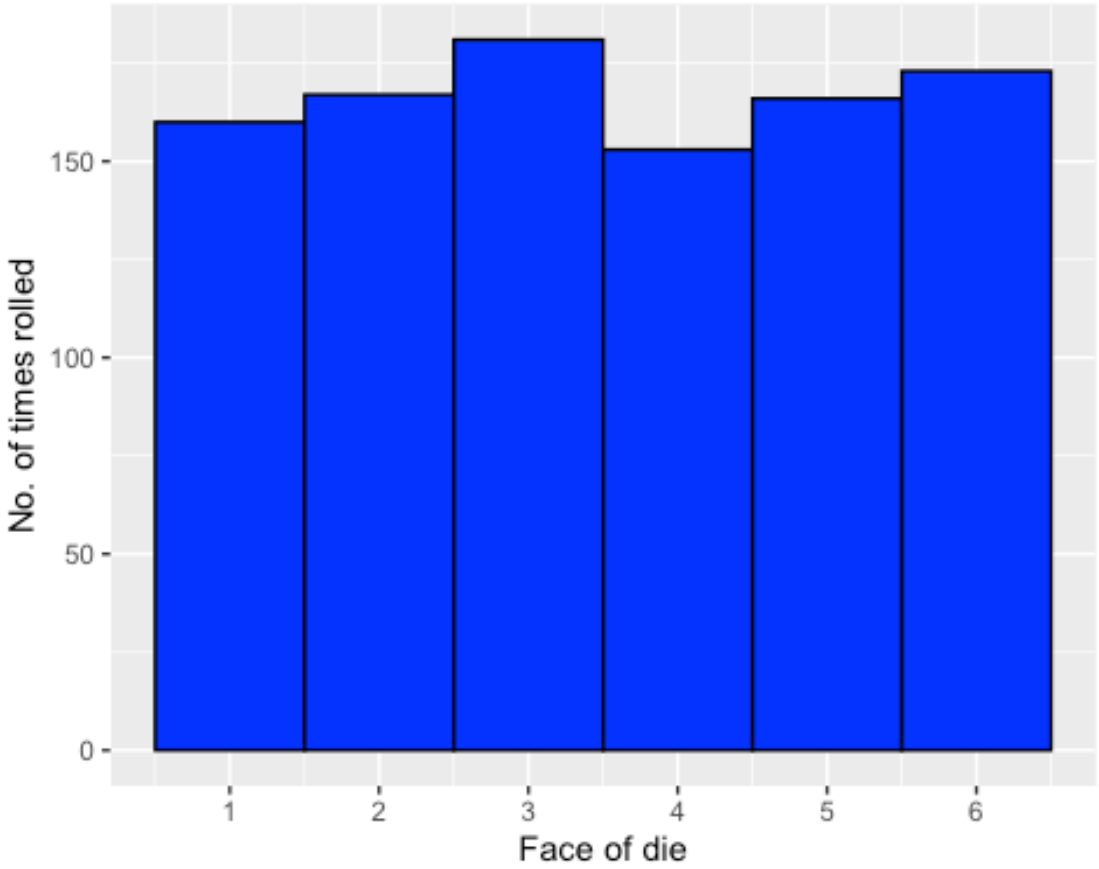
Step 3: Visualize the Simulation with ggplot2

```
# Load packages
library(ggplot2)
library(dplyr)

sum(simresult$outcome)

## [1] 3517

simresult %>%
  ggplot(aes(x = outcome)) +
  geom_histogram(binwidth=1, fill='blue', col='black') +
  scale_x_continuous(breaks = c(1, 2, 3, 4, 5, 6)) +
  ylab("No. of times rolled") +
  xlab("Face of die")
```



Step 4: Using Simulation to Compute a Probability

*In this step, you are going to take the simulation of a die toss done **3000** times and use the results to compute the probability that the outcome of the die toss is either a 5 or a 6.*

```
nsims_step4 = 3000
outcome_step4 = numeric(nsims_step4)
fivesix = numeric(nsims_step4)
for(i in 1:nsims_step4){
  outcome_step4[i] = sample(c(1,2,3,4,5,6), 1, replace=FALSE)
  fivesix[i] = if (outcome_step4[i] == 5 || outcome_step4[i] == 6) 1 else 0
  }
simresult_step4 = data.frame(outcome_step4)
head(simresult_step4, 3)
```

```
##   outcome_step4
## 1             6
## 2             6
## 3             6
```

```
tail(simresult_step4, 3)
```

```
##      outcome_step4
## 2998             1
## 2999             2
## 3000             3
```

```
simresult_fivesix = data.frame(fivesix)
head(simresult_fivesix, 3)
```

```
##   fivesix
## 1       1
## 2       1
## 3       1
```

```
tail(simresult_fivesix, 3)
```

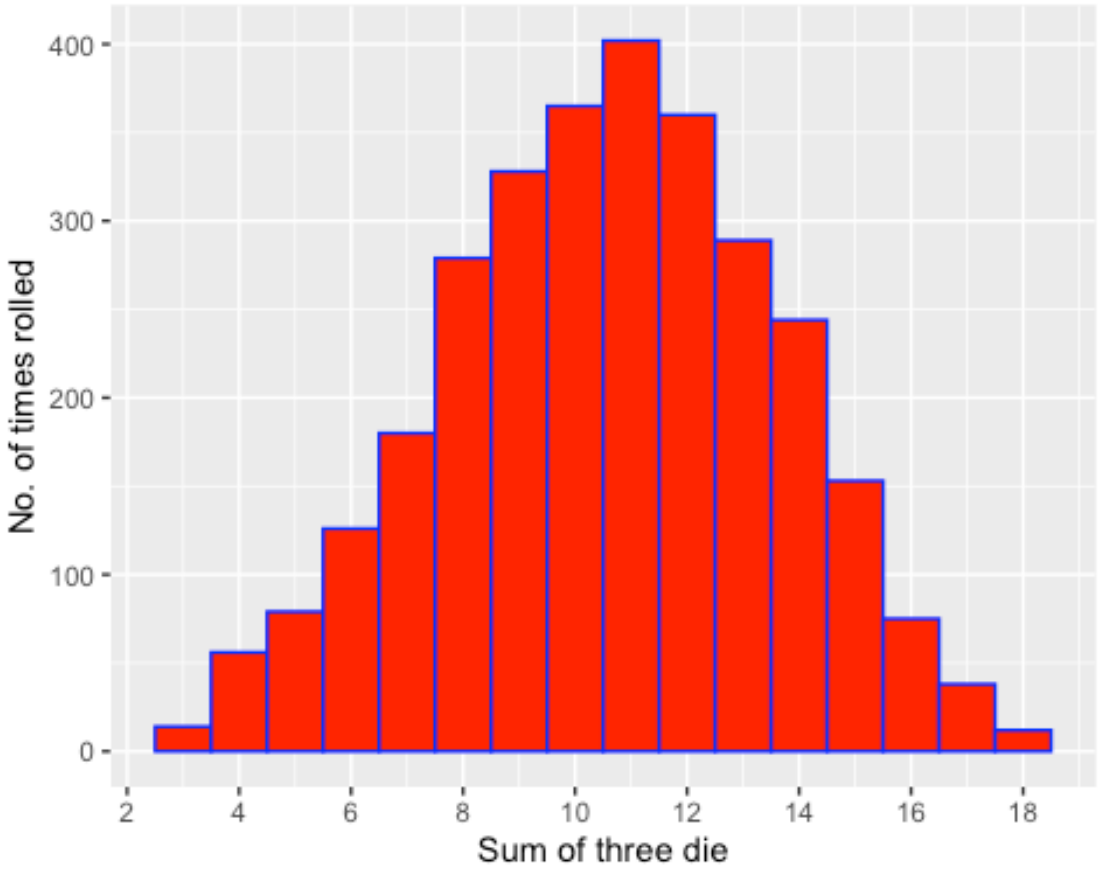
```
##      fivesix
## 2998       0
## 2999       0
## 3000       0
```

```
sum(simresult_fivesix$fivesix)/nsims_step4
```

```
## [1] 0.319
```

Suppose a trial consisted of the three die tosses. An element in the sample space $o_i = (toss1, toss2, toss3)$*. On each trial, you wish to observe if the sum of the three tosses is 14 or more. For example, a* $(5,6,3)$ *outcome sums to 14 and satisifed the condition sum* ≥ 14*. You wish to estimate the probability of observing a sum of 14 or more when three fair die are tossed. Run 3000 simulations.* Compute $P(Sum \geq 14)$.

```r
nsims_q2 = 3000
outcome_q2 = numeric(nsims_q2)
sum14more = numeric(nsims_q2)
for(i in 1:nsims_q2){
  outcome_q2[i] = sum(sample(c(1,2,3,4,5,6), 3, replace=TRUE))
  sum14more[i] = if (outcome_q2[i] >= 14) 1 else 0
  }
simresult_q2 = data.frame(outcome_q2)
simresult_sum14more = data.frame(sum14more)
simresult_q2 %>%
  ggplot(aes(x = outcome_q2)) +
  geom_histogram(binwidth = 1, fill='red', col='blue') +
  scale_x_continuous(breaks = c(seq(2, 20, 2))) +
  ylab("No. of times rolled") +
  xlab("Sum of three die")
```




```r
sum(simresult_sum14more$sum14more)/nsims_q2
```

```
## [1] 0.174
```

Question 3

An abbreviated deck of 20 cards consists of four suits (♡, ♢, ♠, ♣) and the following denominations (10, Jack, Queen, King, Ace). You pick at random five cards, or a `hand', without replacement.

a.

Compute the probability that your hand will have neither ♠s nor ♣s. There are a total of 20 cards of which 5 cards are to be chosen. The number of ways this can occur is:

$$\binom{20}{5}$$

There are a total of 5 spades and 5 clubs in this deck of 20 cards. We need to determine the probability that 0 of those 10 cards are selected and that 5 of the remaining cards are selected:

$$\binom{10}{5}\binom{10}{0}$$

$$\begin{aligned} P(X = 0) = &\frac{{10 \choose 0}{10 \choose 5 - 0}}{{20 \choose 5}} \hspace{0.5in} \\ \\ = & \frac{1 * 252}{15504} \\ \\ \approx &\ 0.0163 \end{aligned}$$

```
## [1] 0.016
```

b.

Compute the probability your hand will consist of a 10, Jack, Queen, King, and Ace of the same suit. For example, $(10♡, J♡, Q♡, K♡, Ace♡)$.

Since their are only four suits, their are only four possible combinations for the 5 cards to consist of a 10, Jack, Queen, King and Ace of the same suit.

Therefore $n(a) = 4$.

The total number of combinations is consistent with part a, and can be displayed as:

$$\binom{20}{5}$$

.

$$\begin{aligned} P(\text{straight flush}) = & \frac{n(a)}{{20 \choose 5}} \hspace{0.5in} \\ \\ = & \frac{4}{15504} \\ \\ \approx &\ 0.0003 \end{aligned}$$

```
## [1] 3e-04
```

c.

Compute the probability that you get a three-of-a-kind. For example: (10♡, 10 ◊, 10♠, *J*♣, *K*♡).

For each individual card, their are four possible cards to choose from and we want to choose three of them to make a three-of-a-kind:

$$\binom{4}{3}$$

From the other 15 cards in the deck (excluding one from the deck that we don't want), we want to choose 2 of them:

$$\binom{15}{2}$$

Considering there are 5 types of cards in the deck, we need to multiply these combinations by 5.

$$\begin{aligned} P(\text{three of a kind}) = &\ 5 * \frac{{4 \choose 3} * {16 \choose 2}}{{20 \choose 5}} \hspace{0.5in} \\ \\ = &\ \frac{2400}{15504} \\ \\ \approx &\ 0.1548 \end{aligned}$$

```
## [1] 0.1548
```

d.

Compute probability that one observes two Aces and two diamonds.

Since the Ace of diamonds exists, we need to split the number of combinations in two: the number of combinations of two aces and two diamonds **without** the ace of diamonds, and the number of combinations of two aces and two diamonds **with** the ace of diamonds.

For the **without** case, there are 19 total cards to choose from:

- 3 cards to select from to select two aces $\binom{3}{2}$
- 4 cards to select from to select two diamonds $\binom{4}{2}$
- 12 remaining cards to select the other card in our 5 card hand $\binom{12}{1}$.

For the **with** case, there are 20 total cards to choose from:

- 1 card to select from to select the ace of diamonds $\binom{1}{1}$
- 3 cards to select from to select the other ace we need $\binom{3}{1}$
- 4 cards to select from to select the other diamond we need $\binom{4}{1}$

- 12 remaining cards to select the other 2 cards in our 5 card hand $\binom{12}{2}$.

We will need to add the combinations in the **without** case and the **with** case in order to determine the total number of combinations that are possible.

Without

$$\begin{aligned} n(\text{Pair of Aces w/o Diamonds}) &= \binom{3}{2} \\ n(\text{Pair of Diamonds w/o Ace}) &= \binom{4}{2} \\ n(\text{Other cards we need}) &= \binom{12}{1} \end{aligned}$$

With

$$\begin{aligned} n(\text{Ace of Diamonds}) &= \binom{1}{1} \\ n(\text{Other Ace}) &= \binom{3}{1} \\ n(\text{Other Diamond}) &= \binom{4}{1} \\ n(\text{Other cards we need}) &= \binom{12}{2} \end{aligned}$$

With and **without**

$$\begin{aligned} n(\text{With}) + n(\text{With}) &= \binom{3}{2}\binom{4}{2}\binom{12}{1} + \binom{1}{1}\binom{3}{1}\binom{4}{1}\binom{12}{2} \\ &= 1008 \end{aligned}$$

The probability is now calculated by:

$$\begin{aligned} &= \frac{1008}{15504} \\ &\approx 0.065 \end{aligned}$$

```
## [1] 0.065
```

Question 4

An oil and gas executive needs to fly from Calgary, Alberta (airport code YYC) to Washington-Dulles (airport code IAD) to attend a meeting with lobbyists about the building of a certain pipeline. Because there is no direct flight from YYC to IAD, this traveler has to fly from YYC to a different city, then connect with a flight to IAD. The traveler has airline options. Airline AA will connect through Dallas, Airline UA will connect through Chicago, or Airline D which

connects through Minneapolis-St.Paul. Taking into their past experiences with flying with the three airlines in question, this executive hints that the probability of flying with Airline AA is 0.15. The probability they will fly with Airline D is three times more than the probability of flying with Airline UA. Historical data has shown that 15% of passengers who fly with Airline AA miss their connecting flights in Dallas. Similarly, 10% of Airline D passengers and 30% of Airline UA passengers miss their connecting flights.

a.

The executive has called the office of the lobby-group to say they have missed their connecting flight. Compute the probability that the executive called from Chicago (or is flying Airline UA).

First, with the given information, we can calculate the individual probabilities of this executive taking each airline, assuming there are no other airlines available to connect to IAD. We are given the probability that the executive will fly with Airline AA $P(AA) = 0.15$ and also, the probability that the executive will fly with Airline D $P(D) = 3 * P(UA)$. Now we can calculate both $P(D)$ and $P(UA)$:

$$\begin{aligned} P(AA) + P(UA) + P(D) &= 1 \\ 0.15 + P(UA) + 3 * P(UA) &= 1 \\ P(UA) &\approx 0.21 \\ P(D) &\approx 0.21 * 3 \\ P(D) &\approx 0.64 \end{aligned}$$

Given that we know that the executive missed their connecting flight, we need to determine the probability of the executive calling from Chicago: $P(\text{UA} \mid \text{Missed})$

$$P(\text{UA} \mid \text{Missed}) = \frac{P(\text{UA} \cap \text{Missed})}{P(\text{Missed})}$$

$$P(\text{UA} \mid \text{Missed}) = \frac{P(\text{UA} \cap \text{Missed})}{P(\text{UA} \cap \text{Missed}) + P(\text{AA} \cap \text{Missed}) + P(\text{D} \cap \text{Missed})}$$

$$P(\text{UA} \mid \text{Missed}) = \frac{0.1 * 0.2125}{0.1 * 0.2125 + 0.6375 * 0.3 + 0.15 * 0.15}$$

$$P(\text{UA} \mid \text{Missed}) = \frac{0.02125}{0.235}$$

$$P(\text{UA} \mid \text{Missed}) \approx 0.09$$

b.

The executive has not missed a connecting flight and made it to IAD. Compute the probability that they flew on a Delta flight.

Given that we know that the executive did not miss their connecting flight, we need to determine the probability of the executive flew through Delta: $P(\text{D} \mid \text{Not Missed})$. Much of the information we calculated in part a can be used in part b.

$$P(\text{D} \mid \text{Not Missed}) = \frac{P(\text{D} \cap \text{Not Missed})}{P(\text{Not Missed})}$$

$$P(\text{D} \mid \text{Not Missed}) = \frac{P(\text{D} \cap \text{Not Missed})}{1 - P(\text{Missed})}$$

$$P(\text{D} \mid \text{Not Missed}) = \frac{0.6375 * (1 - 0.3)}{1 - 0.235}$$

$$P(\text{D} \mid \text{Not Missed}) = \frac{0.44625}{0.765}$$

$$P(\text{D} \mid \text{Not Missed}) \approx 0.583$$

c.

Provide a statement that interprets the meaning of the probability computed in (b) in the ***context of these data****.*

If the executive makes it to Washington without missing a connecting flight, there is a greater than 50% chance that they flew Delta. This means that the majority of the flights taken by this executive when arriving in Washington without missing a connecting flight are with Delta.

Even though the probability of missing their connecting flight with Delta is the highest of the given airlines, the executive's preference of flying Delta $P(D) \approx 0.64$ dominates the probability of flying Delta without missing a connecting flight.

Question 5

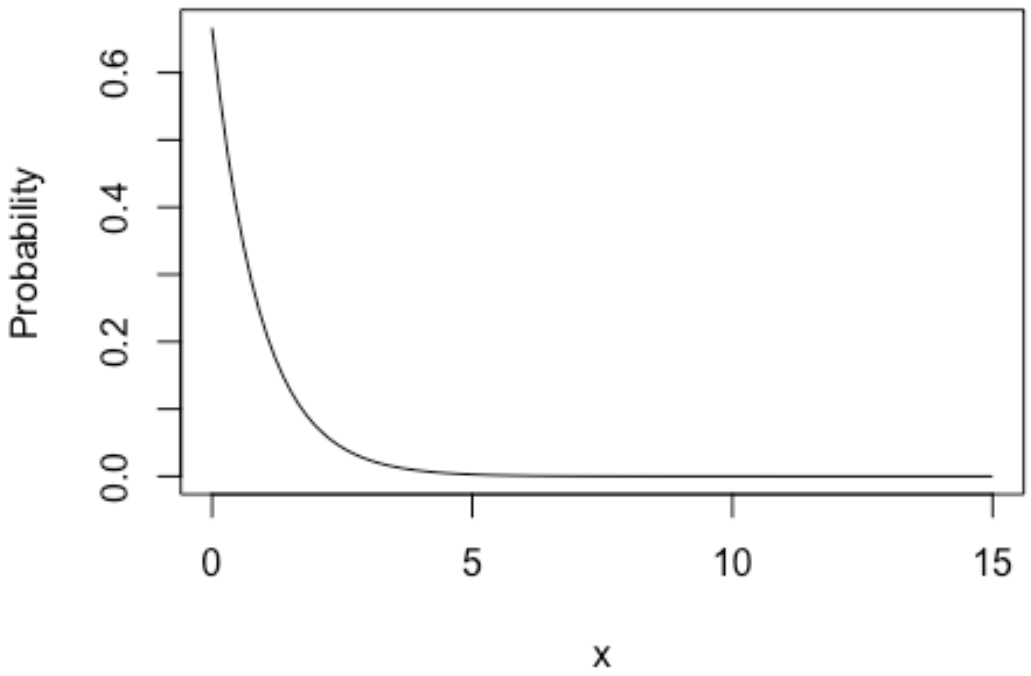
A random variable X has the following probability distribution function

$$P(X = x) = \frac{2}{3^{x + 1}} \hspace{0.5in} x = 0, 1, 2, \cdots$$

a.

Using R Studio, create a display that shows the probability distribution of this particular random variable X. Refer to the various code provided for examples appearing in both Probability Module 4 and Review Exercise 5 from Thursday, September 5th. For values of x, use ***xvalues = 0:15****.*

```r
pdf_5a <- function(x){
  2 / (3**(x + 1))
}
curve(pdf_5a, from = 0, to = 15, n = 100, ylab = "Probability")
```



b.

Compute $P(X > 3)$.

```r
prob_5b <- integrate(pdf_5a, lower = 3, upper = 15)$value
round(prob_5b, digits = 4)
```

```
## [1] 0.0225
```

c.

Compute the mean or expected value of X, $E(X)$ *or* μ_X. *(Hint: In computing* $E(X)$, *change the upper limit on xvalues from 15 to 50...)*

```r
pdf_mean_5c <- function(x){
  x * 2 / (3**(x + 1))
}
mean_5c <- integrate(pdf_mean_5c, lower = 0, upper = Inf)$value
round(mean_5c, digits = 4)
```

```
## [1] 0.5524
```

d.

Compute the standard deviation of X, $SD(X)$ or σ_X.

```
pdf_sd_5d <- function(x){
  x**2 * 2 / (3**(x + 1))
}
sd_5d <- (integrate(pdf_sd_5d, lower = 0, upper = Inf)$value -
mean_5c**2)**0.5
round(sd_5d, digits = 4)

## [1] 0.8369
```

e.

Consider the interval $(\mu_X - \sigma_X, \mu_X + \sigma_X)$. Compute $P(\mu_X - \sigma_X < X < \mu_X + \sigma_X)$.

```
prob_5e <- integrate(pdf_5a, lower = (mean_5c - sd_5d), upper = (mean_5c +
sd_5d))$value
round(prob_5e, digits = 4)

## [1] 0.6977
```

Question 6

Ipsos-Reid released the results of a poll indicating that 60% of Canadians disagree with internet companies handing over private information to authorities (such as the police). You are to randomly pick $n = 40$ Canadians. Compute the probability that:

a.

Exactly 30 will disagree with internet companies handing over private information to authorities.

```
prob_6a <- choose(40, 30) * 0.6**30 * (1 - 0.6)**(40 - 30)
round(prob_6a, digits = 4)

## [1] 0.0196

#alternative
round(dbinom(30, 40, 0.6), digits = 4)

## [1] 0.0196
```

b.

Between 28 and 35 (inclusive) will disagree with internet companies handing over private information to authorities.

```r
prob_6b <- 0
for(i in 28:35){
  prob_6b <- prob_6b + choose(40, i) * 0.6**i * (1 - 0.6)**(40 - i)
}
round(prob_6b, digits = 4)
```

```
## [1] 0.1285
```

```r
#alternative
round(pbinom(35, 40, 0.6) - pbinom(27, 40, 0.6), digits = 4)
```

```
## [1] 0.1285
```

c.

Suppose you are to randomly inspect n-Canadians on this issue until you find the 15-th to disagree with internet companies handing over private information. Compute the probability that $n = 25$.

```r
prob_5c <- choose(25, 15) * 0.6**15 * (1 - 0.6)**(25-15)
round(prob_5c, digits = 4)
```

```
## [1] 0.1612
```

```r
#alternative
round(dbinom(15, 25, 0.6), digits = 4)
```

```
## [1] 0.1612
```

Question 7

*You and four friends decide to play "Odd Person Out". In this game, the five of you each toss a fair coin. The person who throws the **odd outcome** has to pay for the next round of drinks/coffee/kombutcha/whatever-you-all-fancy. For example, if one person flips a head while the other four flips tails, then the person who flipped the head has to pay for all five, and vice versa. Should such an outcome not occur, everyone flips again **until** the "odd person out" occurs. Presuming all five toss a fair coin, the random variable X that counts the number of tosses needed to observe "odd person out" is given by*

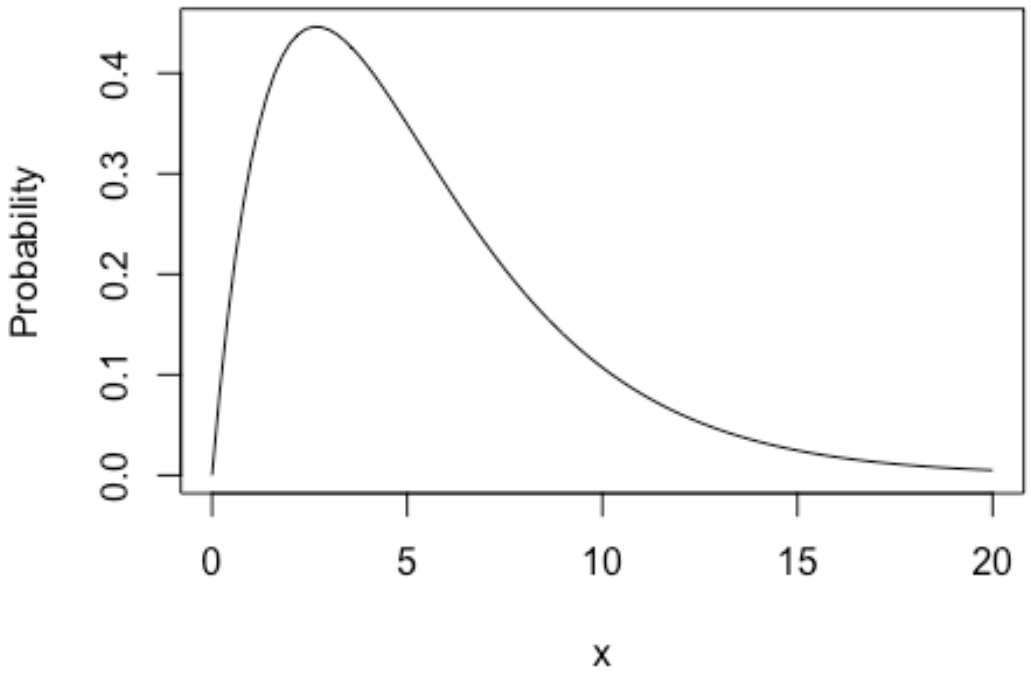
$$ P(X = x) = \left(0.6875\right)^{x -1}\left(0.3125 \right) \hspace{0.5in} x = 1, 2, 3, 4, \cdots $$

It has taken 10 rounds to observe "odd person out", or $X = 10$. Did it take more trials than expected to observe "odd person out" or less? Ensure you incorporate course content in your explaination.

In this problem, we want to calculate the expected number of trials it would take to observe "odd person out" in order to compare it to the number of trials given in the question (10). We can calculate the expected number of trials to observe "odd person out" through:

$$\mu_X = E(X) = \sum_{all\ x} x\,P(X = x)$$

```
pdf_mean_7 <- function(x){
  x * 0.6875**(x - 1) * 0.3125
}
curve(pdf_mean_7, from = 0, to = 20, ylab = "Probability")
```




```
round(integrate(pdf_mean_7, lower = 0, upper = Inf)$value, digits = 4)
```

```
## [1] 3.2376
```

Since the expected amount of trials to observe the "odd person out" is approximately 3, 10 rounds exceeds the expected amount of trials. This is clearly visible on the Probability Distribution Function curve in which a right skew of the curve occurs well before 10 trials.

Question 8

In a certain beverage manufacturer's factory, an automated soft-drink filling machine is to fill 2-litre bottles with product, the amount of soft-drink slightly varying from one 2-litre bottle to the next in according with a Normal probability model with a mean of $\mu = 1.89$ *litres and a standard deviation of* $\sigma = 0.05$ *litres.*

a.

You are to randomly pick a 2-litre bottle off the production line and measure its contents. Compute the probability that the amount of soft-drink dispensed into this bottle is between 1.83 and 1.91 litres.

```
mean_8 <- 1.89
sd_8 <- 0.05
ndf <- function(x){
  1 / (2 * pi * sd_8**2)**0.5 * exp(-(x - mean_8)**2 / (2 * sd_8 **2))
}
prob_8a <- integrate(ndf, lower = 1.83, upper = 1.91)$value
round(prob_8a, digits = 4)
```

```
## [1] 0.5404
```

```
#alternative
round((pnorm(1.91, mean = mean_8, sd = sd_8) - pnorm(1.83, mean = mean_8, sd
= sd_8)), digits = 4)
```

```
## [1] 0.5404
```

b.

Find the 90th-percentile and interpret is meaning in the context of these data.

```
round(qnorm(0.9, mean_8, sd_8), digits = 4)
```

```
## [1] 1.9541
```

This value represents 90% of all 2-litre bottles meaning 90% of all 2-litre bottles will contain ~1.95 litres of soft-drink or less.

$$P(X < 1.954078) = 0.9$$

c.

*What proportion of **all** 2-litre bottles will be filled to overflow?*

```
prob_8c <- integrate(ndf, lower = 2, upper = Inf)$value
round(prob_8c, digits = 4)
```

```
## [1] 0.0139
```

```
#alternative
round((1 - pnorm(2, mean_8, sd_8)), digits = 4)
```

```
## [1] 0.0139
```

$$P(X > 2) \approx 0.014$$

d.

You are to randomly pick 50 2-litre bottles for inspection, measuring the amount of product dispensed into each of the bottles. Compute the probability that between 5 and 10 of these bottles will have less than 1.85 litres of soft-drink.

```
#Compute the probability that any selected bottle will have less than 1.85
litres
prob_8d1 <- pnorm(1.85, mean = mean_8, sd = sd_8)

#Use the binomial distribution to calculate the probability of selecting
#between 5 and 10 bottles of 1.85 litres or less out of 50
prob_8d <- 0
for(i in 5:10){
  prob_8d <- prob_8d + choose(50, i) * prob_8d1**i * (1 - prob_8d1)**(50 - i)
}
round(prob_8d, digits = 4)

## [1] 0.4892

#alternative
round((pbinom(10, 50, prob_8d1) - pbinom(4, 50, prob_8d1)), digits = 4)

## [1] 0.4892
```

$$P(5 < X < 10) \approx 0.49$$

Question 9

The data file GSS2002.csv consists of data resulting from the General Social Suvery (GSS) that tracks various demographic, characteristics, and views on social and political issues since the early 1970s. This file can be imported into R with the following command:

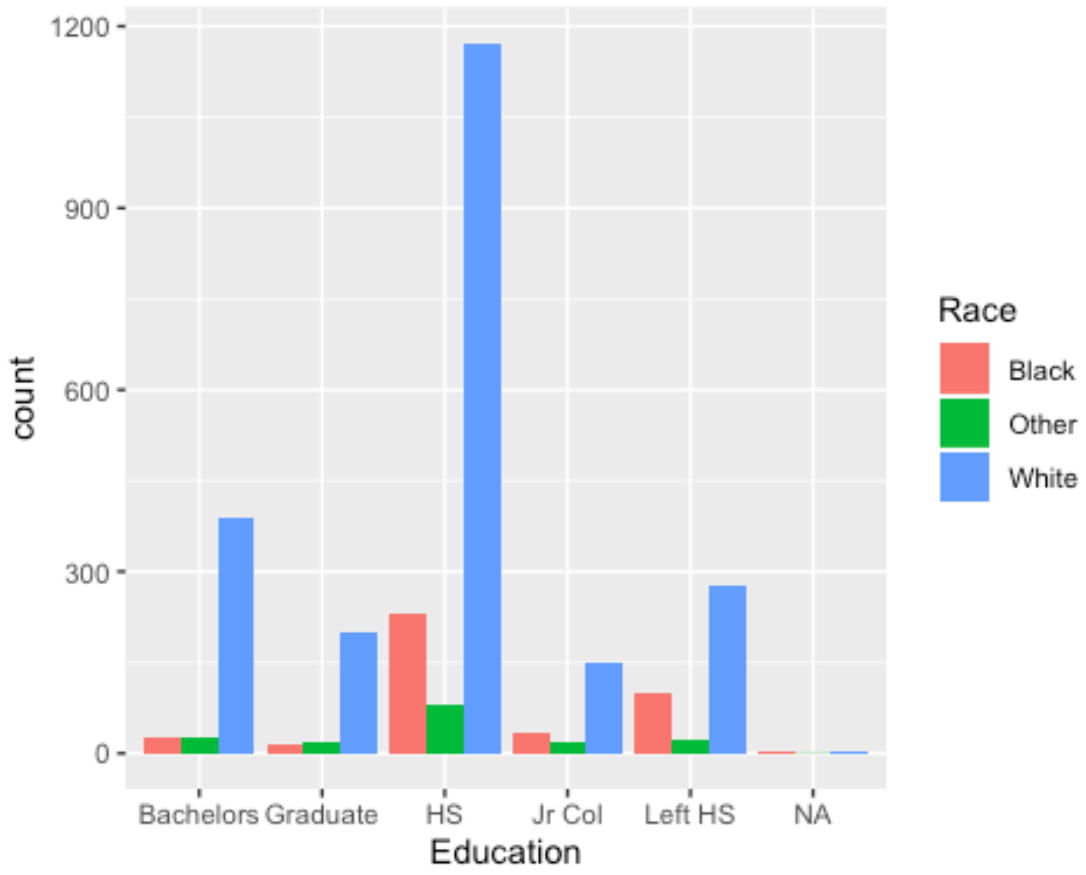
```
gss = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/GSS2002.csv")
```

*This data set consists of a various categorical variables. You can run the **head(gss, # of rows to see)** command to inspect the different variable names. Or, the command **columnnames(gss)** will return the names of the different columns/variables. To determine the different values that are possible on a certain categorical variable, the command **levels(dataframename$variablename)** will return the different values.*

a.

Create a bar graph that demonstrates the distribution of race within each level of education. What can you infer from this bar graph?

```
gss %>%
  ggplot(aes(x = Education, fill = Race)) +
  geom_bar(position = "dodge")
```

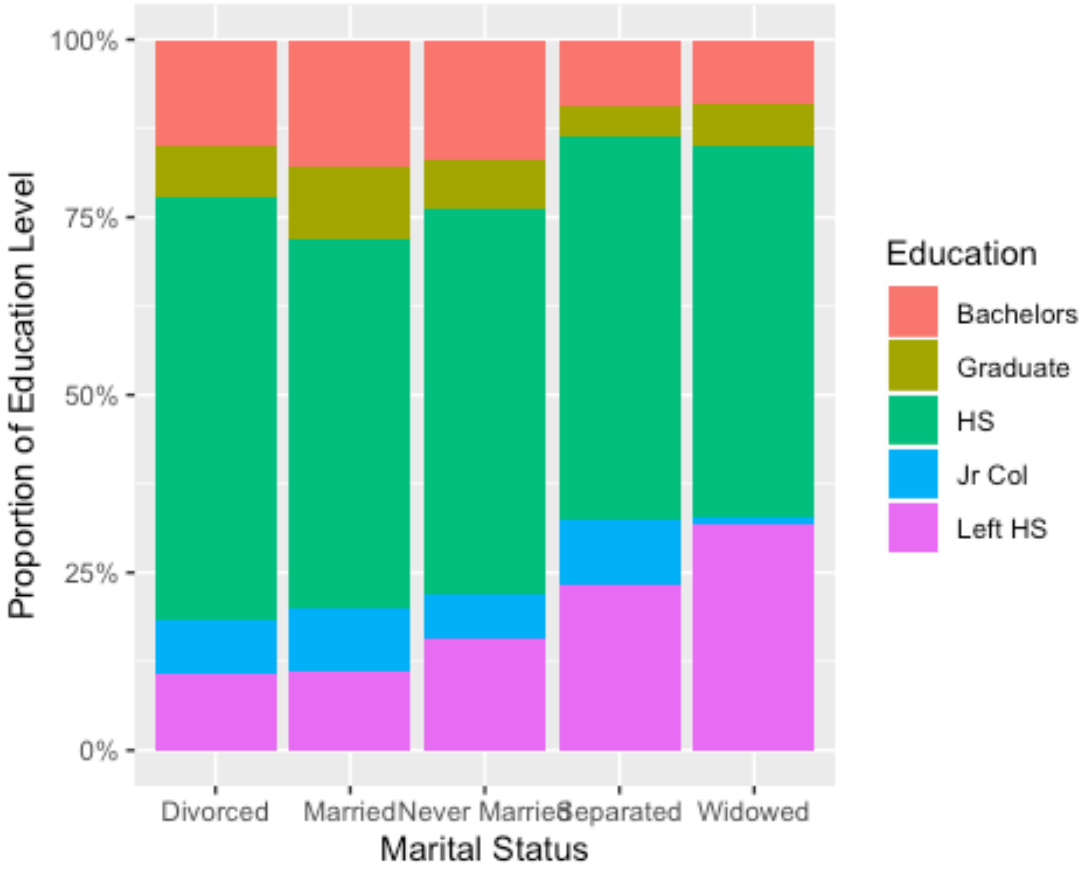


Based on the bar graph above, it can be inferred that at the Bachelors and Graduate levels of education, the representation of "White" is significantly more than at other levels of education.

b.

Create a data visualization that can be used to demonstrate if there is a relationship between one's marital status (Marital) and their education level.

```
gss %>%
  filter(!is.na(Education)) %>%
  group_by(Marital, Education) %>%
  count() %>%
  group_by(Marital) %>%
  mutate(percentage = n/sum(n)) %>%
  ggplot(aes(x = Marital, y = percentage, fill = Education)) +
  geom_bar(position = "stack", stat = "identity", aes(fill = Education))  +
  scale_y_continuous(labels = scales::percent) +
  ylab("Proportion of Education Level") +
  xlab("Marital Status")
```

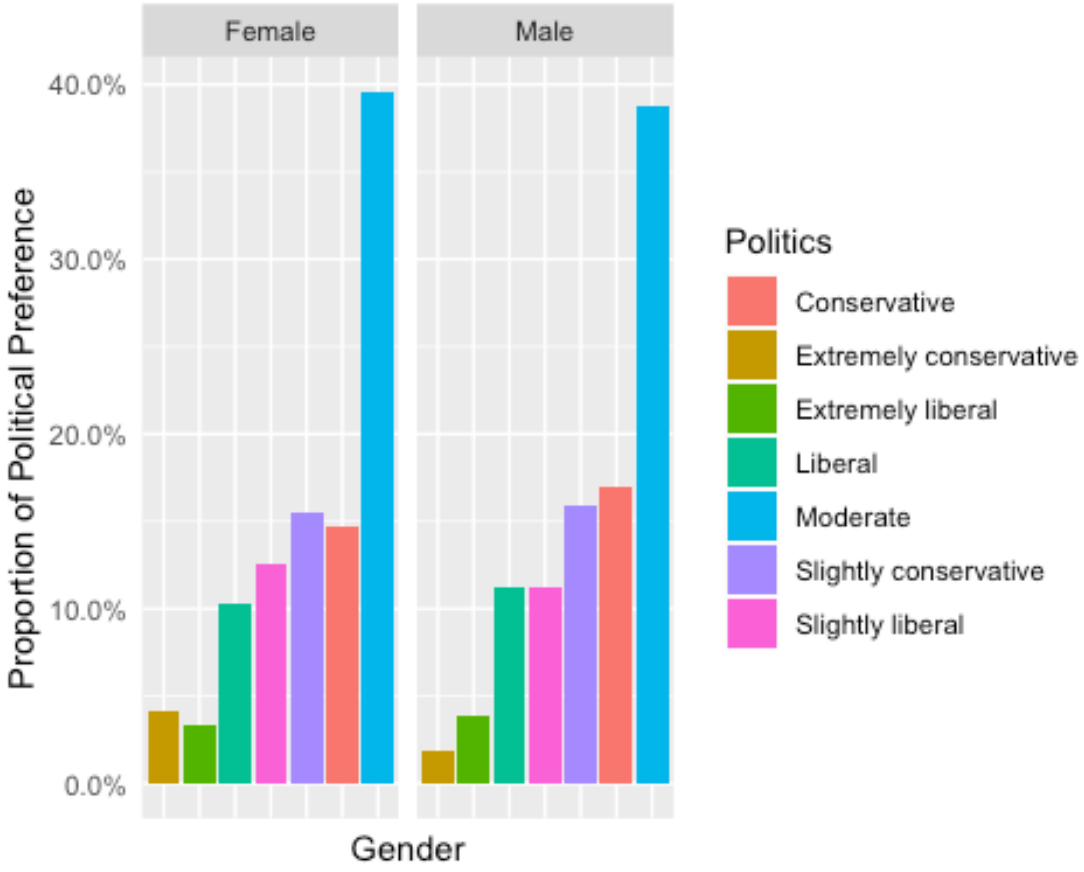


From the above bar graph, there does appear to be some relationship between those who have left highschool and the proportion of people separated and widowed.

c.

Create a data visualization that can be used to demonstrate if there is a relationship between one's Gender and their Politics.

```
gss %>%
  filter(!is.na(Gender) & !is.na(Politics)) %>%
  group_by(Gender, Politics) %>%
  count() %>%
  group_by(Gender) %>%
  mutate(percentage = n/sum(n)) %>%
  ggplot(aes(x = reorder(Politics, percentage), y = percentage)) +
  geom_bar(position = "dodge", stat = "identity", aes(fill = Politics)) +
  facet_wrap(~Gender) +
  scale_y_continuous(labels = scales::percent) +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank()) +
  ylab("Proportion of Political Preference") +
  xlab("Gender")
```



From the above bar graph, there does not seem to be any relationship between gender and political preference

Question 10

*Refer to the **Default** data set in the* ISLR* package. This data set consists of 10000 cases. There are four different variables in this data set. "default" is a categorical variable that indicates if a person has defaulted on their credit card debt (Yes) or has not (No); the variable "student" flags a respondent as a student (Yes) or not (No); the third variable is the person's credit card balancing they are carrying, and the last variable "income" is the person's annual income.*

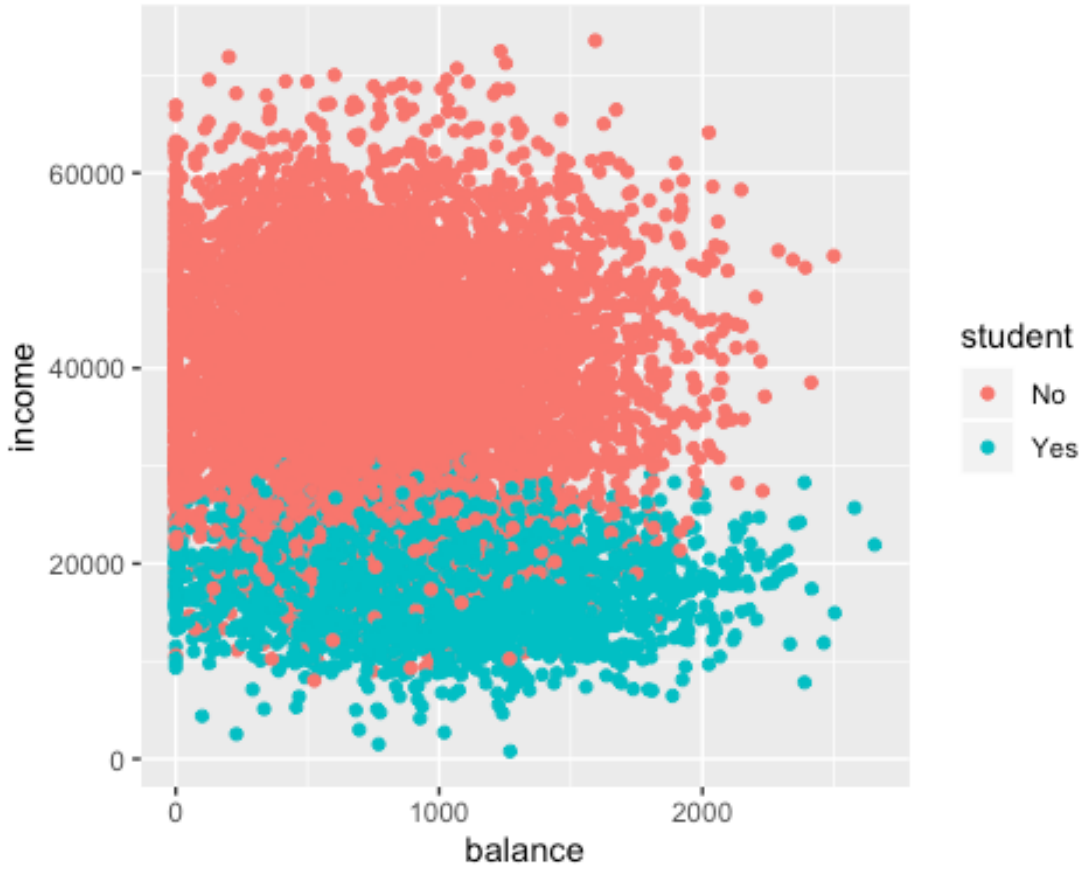
```
library(ISLR)
head(Default, 4)
```

```
##   default student   balance   income
## 1      No      No  729.5265 44361.63
## 2      No     Yes  817.1804 12106.13
## 3      No      No 1073.5492 31767.14
## 4      No      No  529.2506 35704.49
```

a.

Create a scatterplot that demonstrates the relationship between a person's income and their monthly balance they carry on their credit cards. Place the "income" variable as the y-axis and the "balance" variable as the x-axis. Within this visualization, differentiate between those who are students and those who are not.

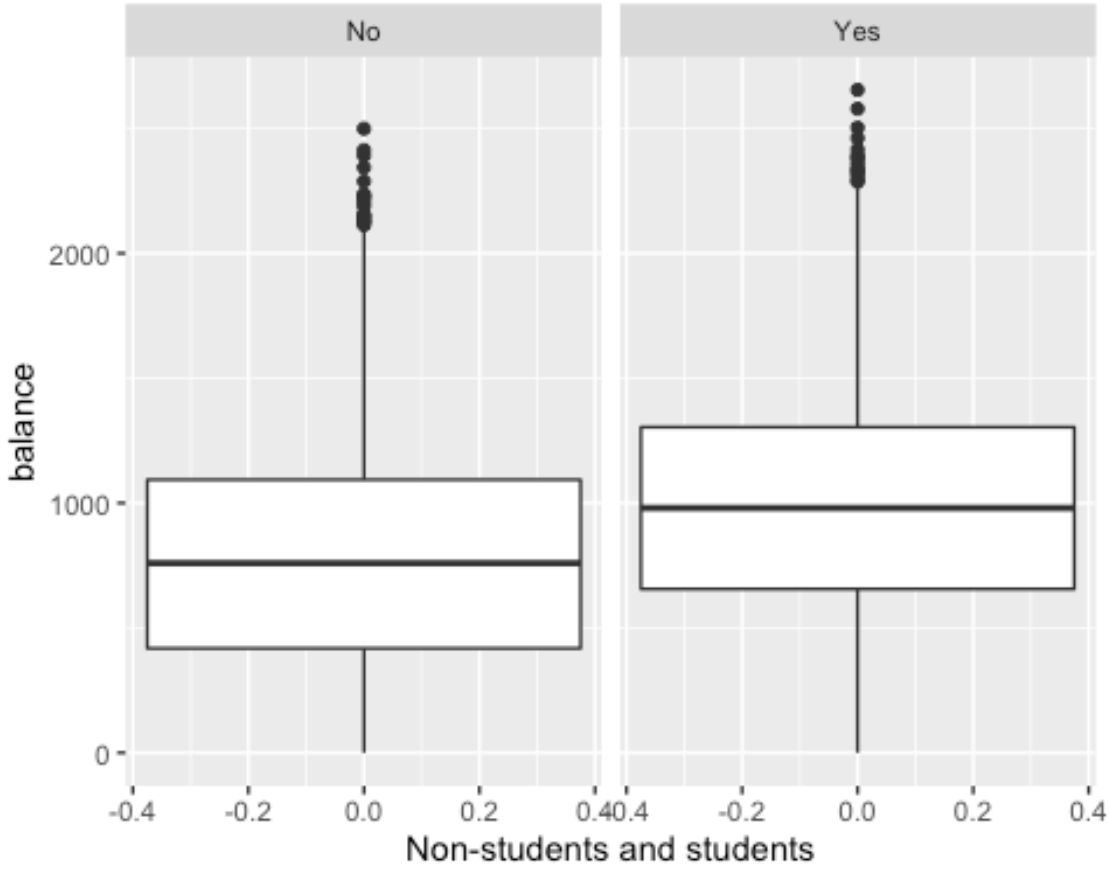
```
Default %>%
  ggplot(aes(x = balance, y = income, color = student)) +
  geom_point()
```



b.

Create side-by-side boxplots that will compare the distributions of balance owing between students and non-students.

```
Default %>%
  ggplot(aes(y = balance)) +
  geom_boxplot() +
  facet_wrap(~student) +
  xlab("Non-students and students")
```



c.

Compute the means, medians, standard deviations, x_5, x_{95} (the 5th and 95th percentiles, respectively) for the data you visually summarized in part (b).

```
student_stats <- Default %>%
  filter(student == "Yes") %>%
  summarise(mean = mean(balance),
            median = median(balance),
            sd = sd(balance),
            fifth_quant = quantile(balance, 0.05),
            ninetyfifth_quant = quantile(balance, 0.95))

non_student_stats <- Default %>%
  filter(student == "No") %>%
  summarise(mean = mean(balance),
            median = median(balance),
            sd = sd(balance),
            fifth_quant = quantile(balance, 0.05),
            ninetyfifth_quant = quantile(balance, 0.95))

student_stats
```

```
##       mean   median       sd fifth_quant ninetyfifth_quant
## 1 987.8182 979.9894 482.9097    172.9326          1811.757
```

```
non_student_stats
```

```
##       mean   median       sd fifth_quant ninetyfifth_quant
## 1 771.7704 759.1891 469.6749           0          1581.915
```

Question 11

A local courier service advertises that the amount of time they take to deliver a package can be modeled by the Normal distribution with a mean delivery time of 5.0 hours and a standard deviation of 1.5 hours. A random sample of $n = 12$ deliveries was taken, and the number of hours it took each to be delivered was recorded. The data appears in csv file.

a.

Read the data in this file into a data frame. Create both a density plot and a boxplot of these data.

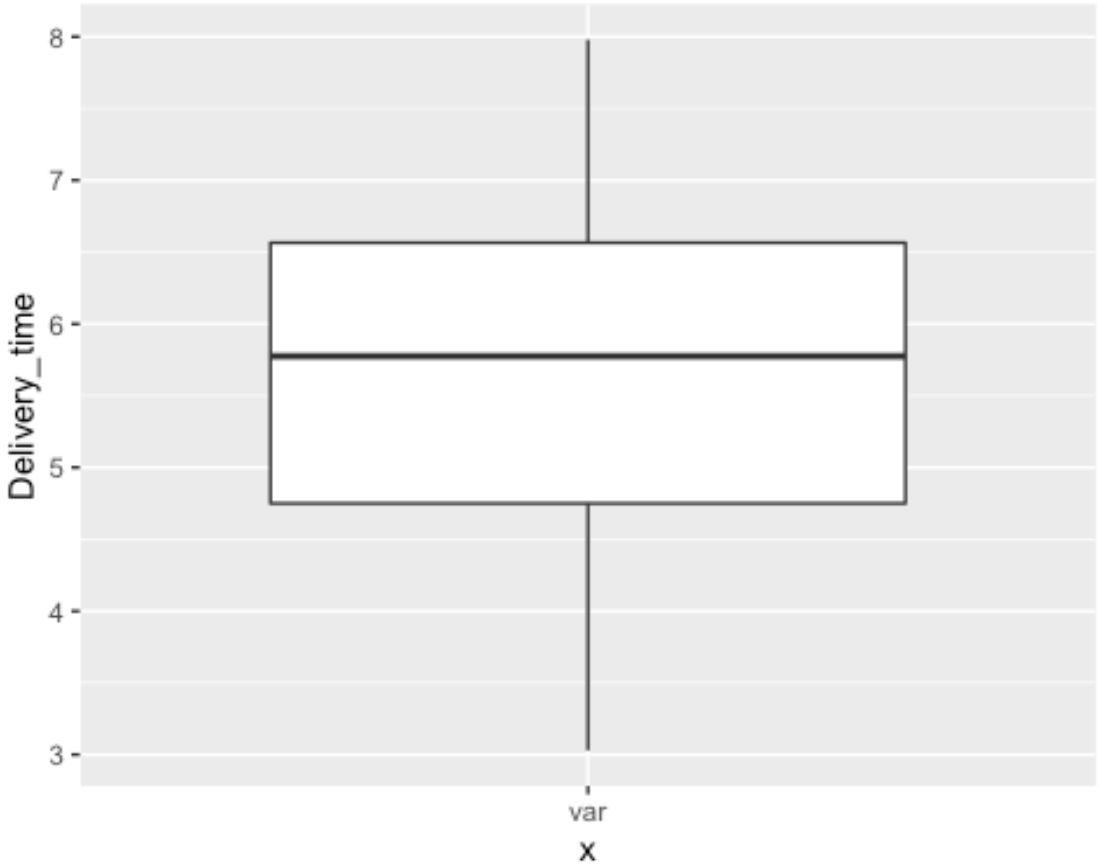
```
df <-
read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/Data602Assignment1Question11.csv")

head(df)
```

```
##   Delivery_time
## 1          3.03
## 2          6.33
## 3          6.50
## 4          5.22
## 5          3.56
## 6          6.76
```

```
df %>%
  ggplot(aes(x = Delivery_time)) +
  geom_density()
```


```
df %>%
  ggplot(aes(x = "var", y = Delivery_time)) +
  geom_boxplot()
```



b.

From this data, compute the sample mean, the sample median, the sample standard deviation, the first and third quartiles, and the 99th percentile.

```
# Sample mean
mean(df$Delivery_time)
```

```
## [1] 5.6875
```

```
# Sample median
median(df$Delivery_time)
```

```
## [1] 5.775
```

```
# Sample standard deviation
sd(df$Delivery_time)
```

```
## [1] 1.580369
```

```
# First quartile
quantile(df$Delivery_time)[2]
```

```
##  25%
## 4.75
```

```r
# Third quartile
quantile(df$Delivery_time)[4]
```

```
##   75%
## 6.565
```

```r
# 99th percentile
quantile(df$Delivery_time, 0.99)
```

```
##    99%
## 7.9778
```

```r
#alternative
favstats(df$Delivery_time)
```

```
##   min   Q1 median    Q3  max   mean       sd  n missing
##  3.03 4.75  5.775 6.565 7.98 5.6875 1.580369 12       0
```

c.

Suppose you were part of a marketing campaign to promote the efficiency of delivery times, as a part of the campaign there was a promise of delivery within a certain number of hours, beyond which there would be a refund for 1% of all deliveries. Provide the point of refund.

```r
round(qnorm(0.99, mean = 5.6875, sd = 1.580369), 4)
```

```
## [1] 9.364
```

The point of refund would be occur at approximately 9.4 hours.