HW1

Graded

Student

Sangwon Ji

Total Points

119 / 119 pts

Question 1

Question 1

8 / 8 pts

✔ **– 0 pts** Correct: A-3, B-1, C-4,D-2

– 1 pt Incorrect: A matched

– 1 pt Incorrect: B matched

– 1 pt Incorrect: C matched

– 1 pt Incorrect D matched

– 4 pts No reasoning provided

– 2 pts Inadequate reasoning

Question 2

Question 2 **46** / 46 pts

2.1 **a** **10** / 10 pts

✔ **– 0 pts** Correct

– 1 pt Plot does not have x-axis/y-axis label, title

– 4 pts A wrong function was used. (instead of dgamma)

– 2 pts Incorrect shape/rate parameters

2.2 **b** **5** / 5 pts

✔ **– 0 pts** Correct

– 2 pts Incorrect calculation of area $P(X < .1)$

– 2 pts Incorrect calculation of area $P(X > 1.5)$

– 1 pt incorrect addition of pgamma0.1 and pgamma 1.5

2.3 **c** **15** / 15 pts

✔ **– 0 pts** Correct

– 7 pts Incorrect density curve

– 7 pts Incorrect histogram

– 1 pt Missing title

– 2 pts Incorrect histogram x-axis parameter.

– 4 pts Incorrect y-axis

2.4 **d** **8** / 8 pts

✔ **– 0 pts** Correct:

```
probGammaEst = mean(GammaObs <.1 | GammaObs >1.5)
probGammaEst
```

– 8 pts Uses density curve or gamma functions instead of the sampled data

– 2 pts Error in calculating the proportion

– 6 pts did not calculate using sampled data

– 4 pts Did not calculate the mean

2.5 **e** **8** / 8 pts

✔ **– 0 pts** The first and/or the second options will be better estimated.

– 8 pts Incorrect

– 4 pts picked one incorrect answer

Question 3

Question 3

10 / 10 pts

✔ **– 0 pts** Correct:

```
with(mydf,hist(Data,main="Histogram of 'Data'",las=1,breaks=100,freq = F))
f<-function(x){dnorm(x, mean = mean(mydf$Data), sd = sd(mydf$Data))}
curve(f,add = T,col = "red",lwd=3)
```

– 1.5 pts Incorrect mean

– 1.5 pts Incorrect SD used

– 2 pts Uses frequency histogram instead of density

– 2 pts x-axis and y-axis swapped

Question 4

Question 4

35 / 35 pts

4.1 a

5 / 5 pts

✔ **– 0 pts** Correct:

```
heart$cp =factor(heart$cp,levels=c(1, 2, 3, 4), labels =c("typical angina", "atypical angina", "non-ang
```

– 5 pts incorrect

– 2.5 pts missing labels

4.2 b

10 / 10 pts

✔ **– 0 pts** Correct:

```
table(heart$cp,heart$num)
```

– 5 pts Does not provide any comment

– 3 pts Inadequate/irrelevant comment

4.3 c

10 / 10 pts

✔ **– 0 pts** Correct:

```
hist(heart$trestbps, freq = F, breaks = 20,main = "Resting blood pressure",xlab = "mm Hg")
lines(density(heart$trestbps))
```

– 1 pt Missing x label, titlle

– 6 pts Uses dnorm to generate, instead of data

– 2 pts Inadequate/irrelevant comment

– 1 pt incorrect x-axis parameter

– 2 pts Made frequency instead of density histogram

4.4 d

10 / 10 pts

✔ **– 0 pts** Correct

– 5 pts No comment

– 3 pts Inadequate/irrelevant comment

Question 5

Question 5 **20** / 20 pts

5.1 **a** **5** / 5 pts

✔ **– 0 pts** Correct:
C = 3

– 2 pts Wrong limits used in the integral

– 2 pts Calculation error

– 5 pts Missing answer

5.2 **b** **5** / 5 pts

✔ **– 0 pts** Correct;
$F(x) = x^3$

– 2 pts Wrong limits in the intergral

– 2 pts Calculation error

– 5 pts missing answer

5.3 **c** **5** / 5 pts

✔ **– 0 pts** Correct;
F(1/3) = 1/27

– 2 pts Calculation error

– 5 pts missing answer

5.4 **d** **5** / 5 pts

✔ **– 0 pts** Correct
$P(0.1 \leq x \leq 0.5) = F(0.5) - F(0.1) = 0.124$

– 2 pts Incorrect use of $F(x)$ or $f(x)$

– 5 pts incorrect

– 5 pts Missing Answer

Question 6

Late Penalty **0** / 0 pts

✔ **– 0 pts** On time

– 11.9 pts 1 day late

– 23.8 pts 2 day late

– 35.7 pts 3 day late

No questions assigned to the following page.

HW1

STAT 131A

Spring 2023

Goals of Homework: In this homework, you should be reviewing the ideas of boxplots, histograms, and continuous distributions, and how to work with them in R. You are also demonstrating basic ability to work in R language and create R markdown files in RStudio.

General Instructions: This homework is given in the form of a R markdown file (HW1.Rmd) that you have learned about in lab, as well as a "compiled" pdf version for easy reading. To answer these questions, you should open the R markdown file in Rstudio, save it to a new file name (e.g. 'HW1_Purdom.Rmd'). Your text answer should start with `>` and after the question, and your code should be inserted into the chunks provided.

Remember to completely answer the question. For example, if you are asked to make a plot and comment on it, don't forget to add the comment! For this homework, we have put prompts like "My answer is …" to show you how this works; you may replace that with your text.

Instructions for code: For those questions that request you to edit or create R code, we have already put in R chunks for where your code should go. Depending on the instructions for the questions, inside the chunks you should either correct the existing code or insert the code needed to complete the assignment. **Do not change the names of the R chunks:** our tests as to whether your code works depends on these names.

If you are asked to use R to find a numeric number, this will not require text, but we will ask you to save your answer with as a particular variable. In order to see this number in your report (and therefore be able to discuss it in other problems) you should print it. For example, we might give you a code chunk that looks like this:

```
set.seed(4291)
# insert code here
# save the median of your simulated data as 'medx'
```

And you might complete it like so:

```
set.seed(4291)
# insert code here
# save the median of your simulated data as 'medx'
x<-rnorm(1000)
medx<-median(x)
# My answer: the median of x is:
print(medx)
```

```
## [1] 0.01612433
```

You can also talk about it in your answer using the value you saved. This is useful if you change your code, then your answer will update (though of course not your conclusions!). Notice how I can put in R code to round the data so as to not print out 8 digits.

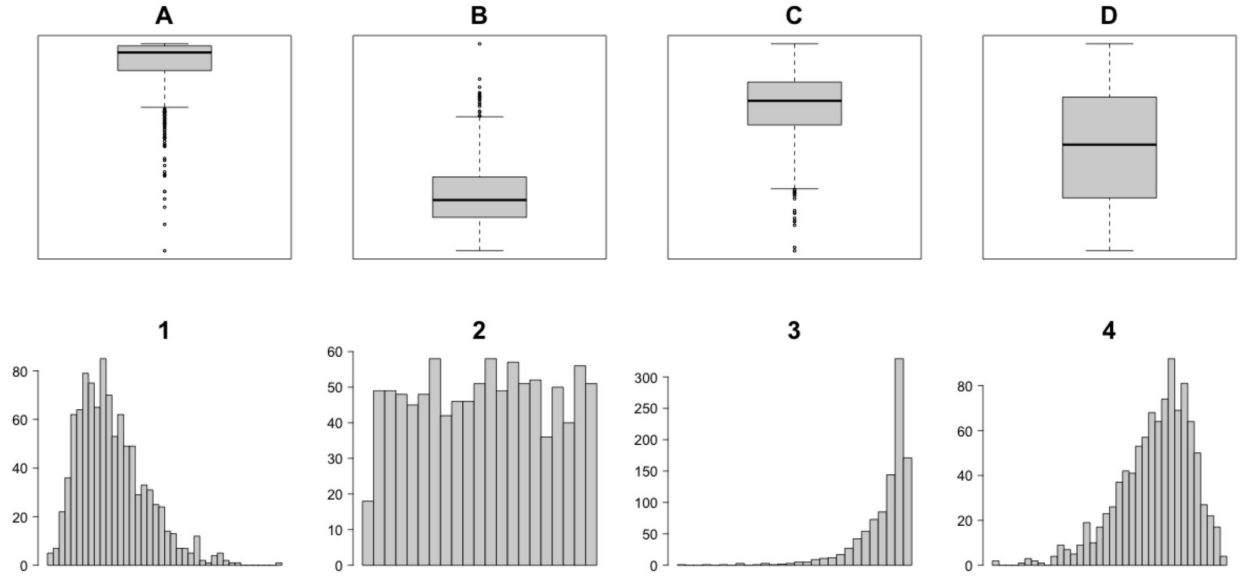
> My answer is that the median of x is 0.016

Question assigned to the following page: 1

Instructions for Submission: We have set up the rmarkdown file so that it should compile into a pdf (the default is into html). If you are not working on the servers, but on your own laptop, you should test that you can compile this file into pdf before changing it. You should then submit both the *.pdf and the *.Rmd file to Gradescope for grading.

Questions

Question 1 (8 points) Below are both the boxplots and frequency histograms of three different datasets. The axes are not labelled to give the actual values of the data, but all three datasets have the same median value. Identify which boxplot goes with which histogram, and explain why.



My answer is A-3, B-1, C-4, D-2.First, for boxplot A matches number number 3, becuase it's similar to number 4 but the difference is that the number is less evenly distributed compared to number 4, and therefore corresponds to the one that is mostly concentrated Also, looking at the outliers, I was able to distinguish from all others. Boxplot B is matched with number 1, becuase it's the only histogram and boxplot that is skewed right-skewed and therefore corresponds with each other. Boxplot C is matched with number 4, becuase it's a left skewed, but it's evenly distributed compared to number 3, therefore corresponds with each other. Boxplot D is matched with number 2 because it is among the most evenly distributed histogram and corresponds to boxplot D which is evenly distributed as well.

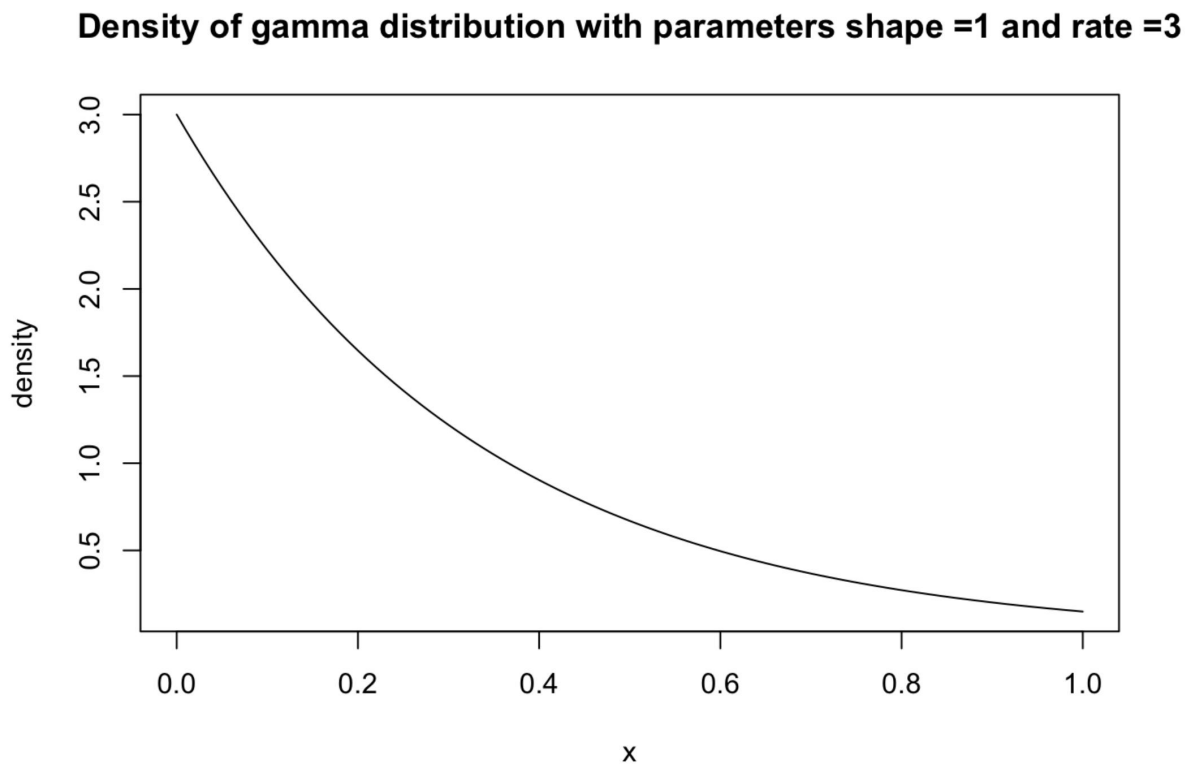
Question 2: Simulation from parametric distributions in R There are many other standard continuous distributions other than the normal distribution that are important in statistics. Just like the normal distribution, R provides functions to plot their density curves, calculate probabilities, and simulate data.

Questions assigned to the following page: 2.1 and 2.2

An example of a common distribution is the gamma distribution. The functions for these distributions are `dgamma`, `pgamma` and `rgamma`; these correspond to the same functions `dnorm`, `pnorm`, and `rnorm` that you have seen for the normal distribution. However, different distributions have different parameters. While the normal distribution has the mean (`mean`) and variance/standard deviation (`sd`), other distributions have other parameters. The gamma distribution has two parameters called `shape` and `rate` by R. We are not going to worry too much right now about what those parameters mean, other than to note that they change the probability distribution.

a. (10 points) Plot the density of a gamma distribution with parameters `shape=1` and `rate=3`` using the density function. Describe how this distribution compares to a normal probability distribution.

```
# Insert code here for plotting the density function.
curve(dgamma(x, shape =1, rate = 3),
ylab = "density",
main = "Density of gamma distribution with parameters shape =1 and rate =3")
```



My answer is that compared to the normal probability distribution, the gamma distribution is right skewed while the normal probability distribution is tend to be symmetry and bell shaped.

b. (5 points) Find the probability of an observation from this distribution being less than .1 or greater than 1.5, using R commands, and make sure the answer prints out so that it shows up in the pdf.

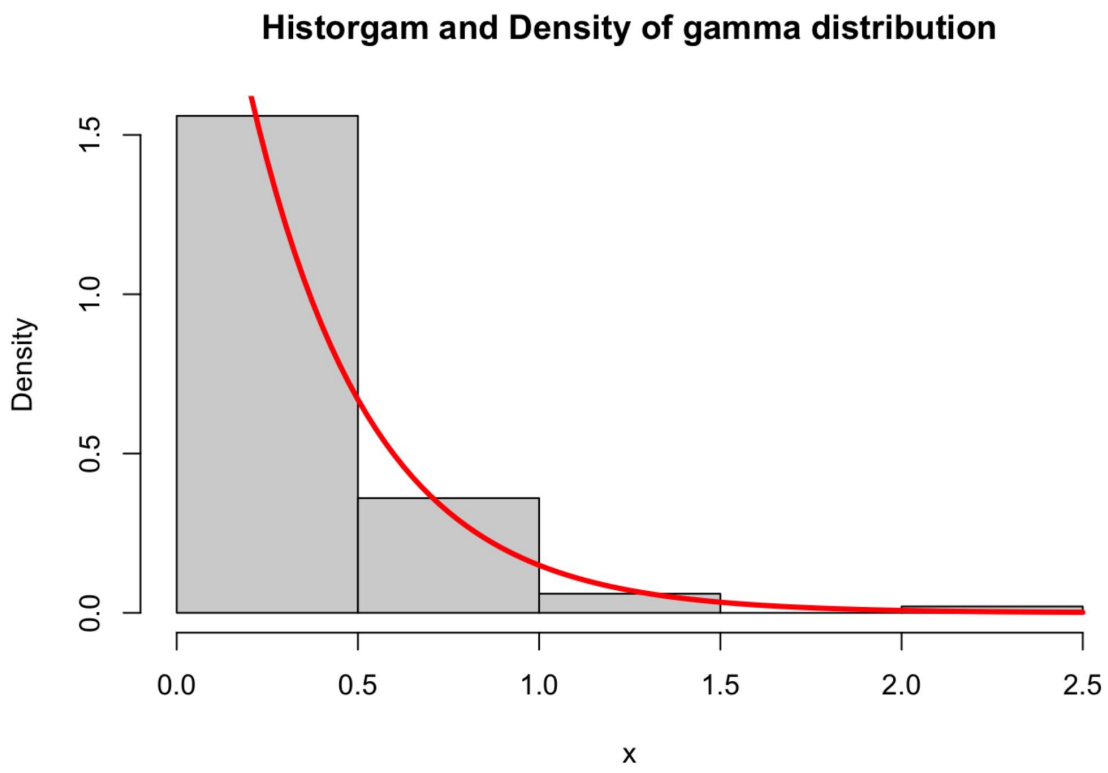
Questions assigned to the following page: 2.3, 2.2, and 2.4

```
# Insert code here for calculating the probability of being  <.1 or >1.5
# Save the response as 'probGamma' and print it
or_Assign <- pgamma(1.5, shape = 1, rate =3) - pgamma(.1, shape =1, rate = 3)
probGamma <- 1 - or_Assign
probGamma
```

```
## [1] 0.2702908
```

c. (15 points) Now simulate 100 observations from the above gamma distribution and plot a histogram of the simulated data. Overlay the true probability density that you plotted above on top of the histogram. Color the true density curve red and make it thicker than the default so that it stands out.

```
set.seed(51920)
# Insert code here for simulating the data and making a histogram from the data.
simGamma <- rgamma(100, shape =1, rate = 3)
hist(simGamma , freq = FALSE,xlab = "x",
     main = "Historgam and Density of gamma distribution")
curve(dgamma(x, shape =1, rate = 3), col ="red", lwd = "3", add=TRUE)
```



d. (8 points) Imagine that this simulated data was actually observed data and you didn't know the actual probability distribution of the data and you want to use this simulated data to estimate the probability distribution. What would be your estimate of being less than .1 or greater than 1.5? (calculate it in R using the simulated data and make sure the answer prints out so that it shows up in the pdf). How does it compare to what we know is the actual probability that you calculated in (b) above?

Questions assigned to the following page: 3, 2.4, and 2.5

```
# Insert code here for estimating the probability of being <.1 or >1.5 from simulated da
ta created in previous chunk.
# Save the response as 'probGammaEst'
probGammaEst <- (sum(simGamma <.1) + sum (simGamma > 1.5))/length(simGamma)
probGammaEst
```

```
## [1] 0.29
```

My answer is Not much, but there is some difference, but I would say it came out to be close to the estimation that was made earlier at 2b.

e. (8 points) Comparatively, which of the following probabilities would be likely to be better estimated from this simulated data, and which would need more data (explain your reasoning)

- Probability of an observation < 0.1
- Probability of an observation between 0.5 and 1.0
- Probability of an observation > 1

My answer is probability of an observation between 0.5 and 1.0. I think for this case, because it seems to have the largest probability of observation is the biggest among all. For probability of an obervation >1, I think we will need more data on that.

Question 3: (10 points) For the small set of toy data saved in 'histogramData.txt', we want to compare the distribution of this data to a normal density curve to see if it might be distributed reasonably closely to normal. Read in the data with the code below, making sure the data matches what you see in the pdf.

```
# Make sure this code works for you and creates output exactly like that seen on pdf.
mydf<-read.table("histogramData.txt",header=TRUE)
head(mydf)
```

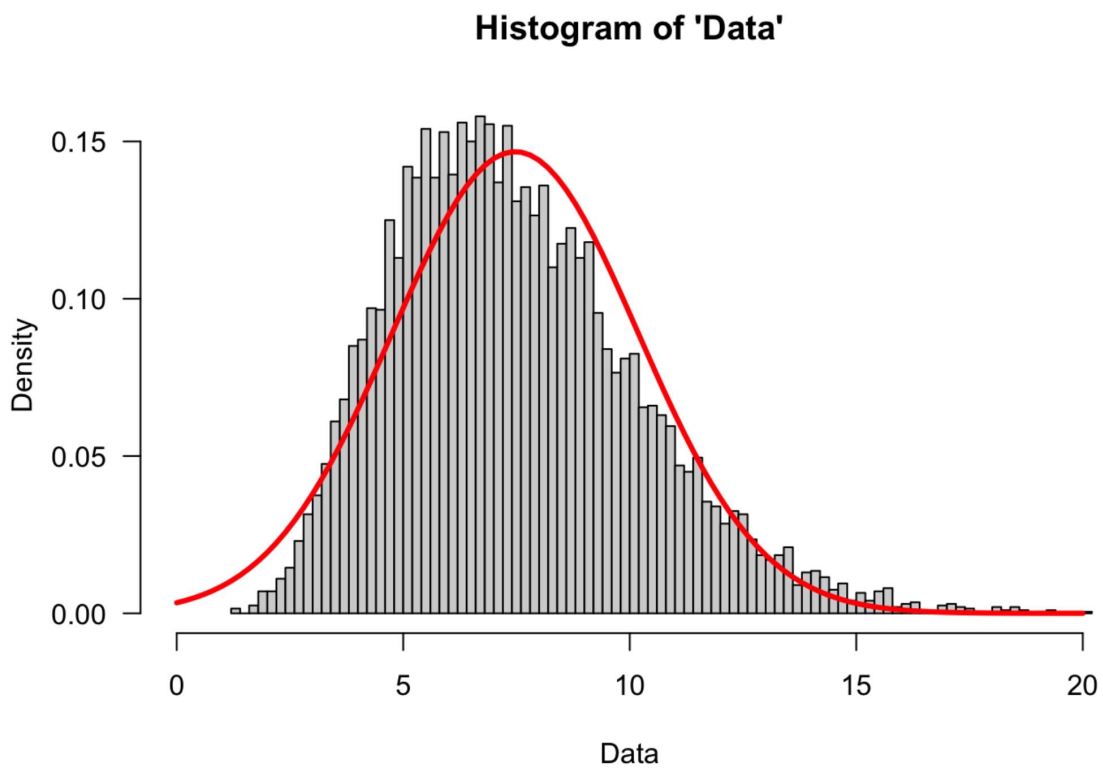
	Data <dbl>
1	9.478014
2	6.462729
3	5.919282
4	8.523029
5	7.094029
6	14.777504

6 rows

Question assigned to the following page: 3

However, the code below for drawing a normal density curves using `dnorm` function (below) does not result in the normal density curve showing up on top of the histogram. Correct the code so that the normal curve shows up on the plot and overlays on top of the data in a reasonably way for a comparison. [Hint: there may be multiple problems with the code].

```
# Correct this code:
with(mydf,hist(Data,main="Histogram of 'Data'",las=1,breaks=100, freq = FALSE, xlim= c
(0, 20)))
f<-function(x){dnorm(x, mean =mean(mydf$Data), sd= sd(mydf$Data))}
curve(f,add=TRUE, col="red",lwd = 3)
```



Question 4: We will consider a dataset consisting of data collected on patients under-going angiography in the 1980's to determine a diagnosis of coronary artery disease at the Cleveland Clinc in Cleveland, Ohio. Angiography is an invasive procedure requiring involving injecting an agent into the blood vessel and imaging using X-ray based techniques. In addition to the final diagnosis, 13 less invasive (and expensive) measurements were taken of each patient, such as blood pressure and heart rate under exercise. The goal was to determine how accurately some combination of these less invasive measures could accurately predict heart disease.

In the dataset `heartDisease.csv` you will find a (comma-deliminated) dataset with the 14 variables (the 13 non-invasive measurements and the final diagnosis). Below we give you the command to read in this data, as well as the command to print out the first few rows of the dataset. Make sure that you can do this correctly and that it matches the result in the pdf version of the homework

Question assigned to the following page: 4.1

```
heart<-read.csv("heartDisease.csv",header=TRUE)
head(heart)
```

	age <dbl>	sex <dbl>	cp <dbl>	trestbps <dbl>	chol <dbl>	fbs <dbl>	restecg <dbl>	thalach <dbl>	exang <dbl>
1	63	1	1	145	233	1	2	150	0
2	67	1	4	160	286	0	2	108	1
3	67	1	4	120	229	0	2	129	1
4	37	1	3	130	250	0	0	187	0
5	41	0	2	130	204	0	2	172	0
6	56	1	2	120	236	0	0	178	0

6 rows | 1-10 of 15 columns

We will concentrate on four variables (the full description is in the `heartREADME.md` file for this data):

- `num` the final diagnosis on a integer scale of 0-4, with 0 being absence of heart disease and 4 the most severe.
- `cp` the type of chest pain the patient was suffering: (1) typical angina, (2) atypical angina, (3) non-anginal pain, (4) asymptomatic
- `age` the age of the patient
- `trestbps` resting blood pressure (in mm Hg) on admission to the hospital

Notice that `cp` is encoded in this data as numeric values, but is actually categorical (this is a quite common practice). Similarly, though `num` is on a numeric scale, it only takes on 5 values, not the continuous range. We can see this even without the above variable guide, by using `table` (notice how I can use `with` to avoid having to type `heart$cp` and `heart$num` everywhere):

```
with(heart, table(cp))
```

```
## cp
##   1   2   3   4
##  23  49  83 142
```

```
with(heart, table(num))
```

```
## num
##   0   1   2   3   4
## 160  54  35  35  13
```

a. (5 points) Change these variables to be `factor` variables in your `heart` data frame. Give the different levels of `cp` the labels described above. For `num` , you can keep the labels '1',…,'4', but for the value '0', give the label 'Absence'. [Hint, to check that you've done this conversion correctly, rerun the table command above and make sure you get the same values from before you changed anything.]

Questions assigned to the following page: 4.3, 4.1, and 4.2

```
# Insert code here for factor conversion
heart$cp = factor(heart$cp, levels = c(1,2,3,4), labels = c("typical angina", "atypical
angina", "non anginal pain", "asymptomatic"))
heart$num = factor(heart$num, levels = c(0, 1, 2, 3, 4), labels = c("Absence", "1", "2",
"3", "4"))
```

Once you have done that correctly, `summary` applied to the `heart` data frame (in the code chunk below) should show the table of their categories, rather than the numerical summary it shows now.

```
# Leave this code in place
summary(heart[,c("cp","num")])
```

```
##                 cp           num
##  typical angina  : 23   Absence:160
##  atypical angina : 49   1      : 54
##  non anginal pain: 83   2      : 35
##  asymptomatic    :142   3      : 35
##                         4      : 13
```

b. (10 points) Create a contingency table between the type of chest pain (`cp`) and the final diagnosis (`num`). Comment on the results.

```
# Insert code here for contingency table
table (heart$cp, heart$num)
```

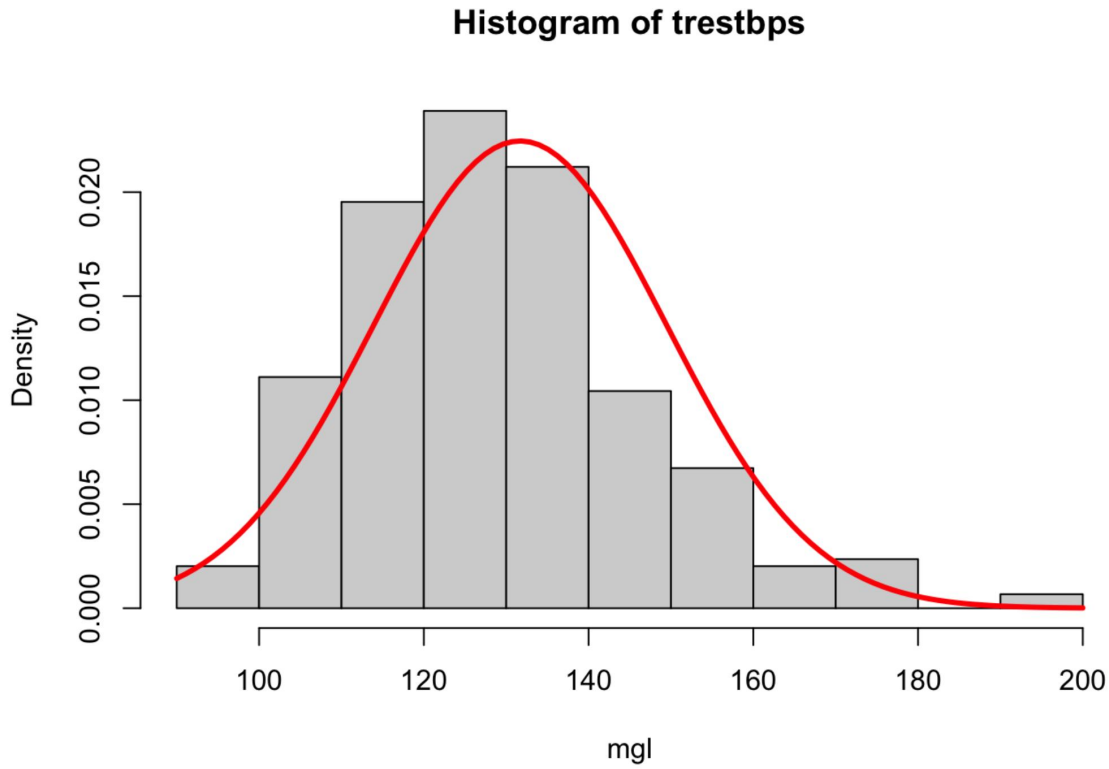
```
##
##                    Absence  1  2  3  4
##   typical angina        16  5  1  0  1
##   atypical angina       40  6  1  2  0
##   non anginal pain      65  9  4  4  1
##   asymptomatic          39 34 29 29 11
```

My answer is that With the 4th level of severness of the heart disease showed most with the asymptomatic and it was mostly shown regardless of the severness among all of the levels. It was to have the greatest outcome at any level. For Absence and most of the levels showed most in absence and going into numeric order for the rest. There were most absence, 1, 2, 3 , 4 most in orders among the symptoms.

c. (10 points) Create a histogram of resting blood pressure (`trestbps`), and overlay a density estimation curve on top of the histogram. Comment on the shape of the distribution.

```
# Insert code here for histogram/density curve
hist(heart$trestbps, freq = FALSE, breaks = 10, main = "Histogram of trestbps", xlab =
"mgl")
curve(dnorm(x, mean = mean(heart$trestbps), sd = sd(heart$trestbps)), add = TRUE, col =
"red", lwd =3)
```

Questions assigned to the following page: 4.3 and 4.4

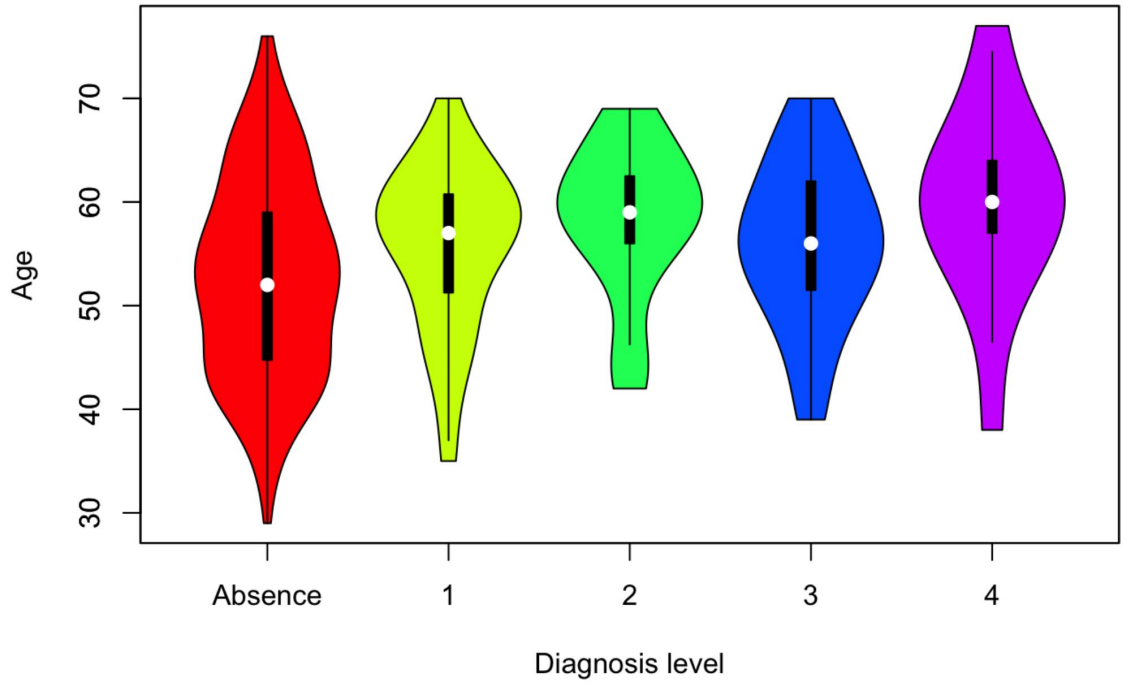


My answer is that the distribution is somewhat closer to the normal distribution, which is symmmetry, bell shaped. However its skwed to the right.

d. (10 points) Create a violin plot of the age of the patients (`age`), separated by the final diagnosis (`num`) (i.e. on one plot, a violin plot of `age` for each of the 5 categories of the diagnosis). Comment on any differences between ages for the different levels of heart disease. [Use Professor Purdom's version `vioplot2` of the vioplot function. The `source` command given below accesses it from the web]

```
#loads Prof. Purdom's function:
source("https://www.stat.berkeley.edu/~epurdom/RcodeForClasses/myvioplot.R")
# Insert code here for violin plots:
palette(rainbow(5))
vioplot2(heart$age, heart$num, col = palette(), ylab = "Age", xlab = "Diagnosis level")
```

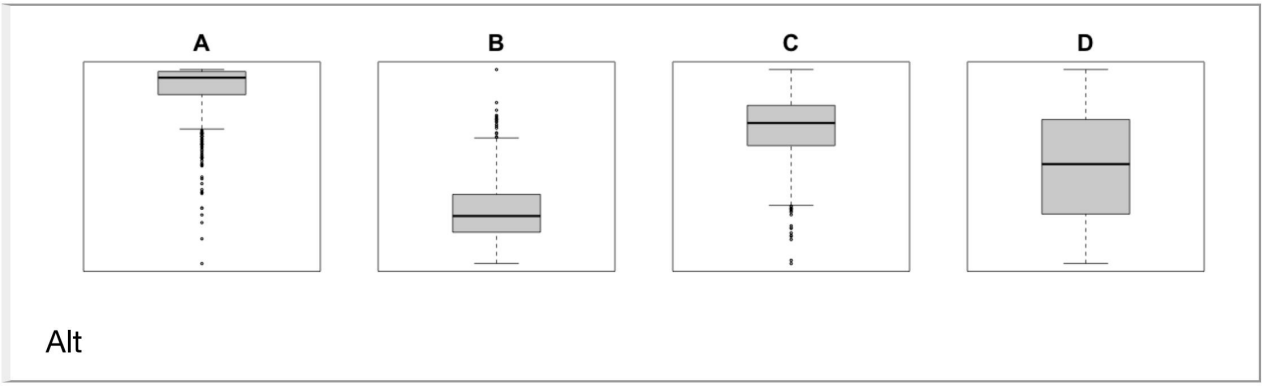
Question assigned to the following page: 4.4



My answer Only the level Absence showed it's reach to the 30's while else had started from at least of the mid-30s. However, not only reaching 30's but absence was in various fields up to 70's. Other 70's was only for the level 4. Others, 1,2, and 3 were showing somewhat similar upperlimit. Including 4 to the 1, 2, and 3, they had the similar distributions. For level 2, it has started from mid 40s. which explains the cutting at the bottom.

Calculations For the next question, if you know how to type your answers with latex code, you can write your answer by hand and save your work on a jpeg, one jpeg for each part. Save them in the same folder as this `HW1.Rmd` file. You can include it in the markdown using the same code I did for the boxplots above – i.e. below the question insert the following code

No questions assigned to the following page.



but replace `mysteryBoxplot.png` with your file name.

Alternatively, you are welcome to use other tools to combine/scan together the pdf for questions 1-4 and a file with your answers to question 5.

Question 5 Consider the following density function defined on (0,1) : $f(x) = cx^2$.

a. (5 points) Find the value of c.

Question assigned to the following page: 5.1

a) Find the value of C.

$$\int_0^1 f(x)\,dx = 1 = \int_0^1 cx^2\,dx = c\int_0^1 x^2\,dx = 1.$$

$$C = \frac{1}{\int_0^1 x^2\,dx} = \frac{1}{\frac{1}{3}(1)^2 - \frac{1}{3}(0)^2} = \underline{3}.$$

Alt

b. (5 points) Find the cdf.

Question assigned to the following page: 5.2

b) Find the cdf.

$$F(x) = \int_0^x f(t)dt = \int_0^3 3t^2 dt =$$

$$(x)^3 - (0)^3 = x^3.$$

or

$$F(x) = P(X \leq x) = P(0 < X < x)$$

Alt

c. (2 points) What is the cumulative distribution function F(x) evaluated at x = 1/3?

Question assigned to the following page: 5.3

C) What is the cumulative distribution function $F(x)$ evaluated at $X = \frac{1}{3}$?

This would be $F(\frac{1}{3})$.

$F(\frac{1}{3}) = (\frac{1}{3})^3 = \frac{1}{27}$

Alt

d. (3 points) What is $P(0.1 \leq X \leq 0.5)$?

Question assigned to the following page: 5.4

d) what is $P(0.1 \le x \le 0.5)$?

$$P(0.1 \le x \le 0.5) = \int_{0.1}^{0.5} f(t)\,dt = (0.5)^3 - (0.1)^3$$

$$= 0.125 - 0.001 = \underline{0.124}$$

Alt