HW1 (120 points)

STAT 131A

Due Date: 11:59 pm, February 10, 2021

Instructions

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. In addition, there are some probability problems.

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 answers should be inserted into the .Rmd file. If you should do code, we will have provided you with a chunk where you should put your code. If you should answer a question with your words, they should be written after the > symbol.

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 chunks

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, we will ask you to save your answer with as a particular variable. In order to report this number so that we can grade it you must print it.

Example For example, suppose we ask you to simulate data from a N(0,1), find the median, and then comment whether it is less than 4. We would give you a code chunk that looks like this (notice we set the seed for you so everyone gets the same answer, do not delete this line, nor should you change the seed):

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

You might input your code 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
```

And then you would give the answer in a comment following your code. You can also talk about the results of your R code in your answer using the variable you saved. This is useful if you change your code, then your answer will update (though of course not your conclusions!)

> My answer is that the median of x is 0.016 which is less than 4.

Look at the .Rmd, to see how I put the answer to the R code `round(medx,3)` in my answer. (I use the function `round` to round the data to 3 digits after the decimal to make it readable).

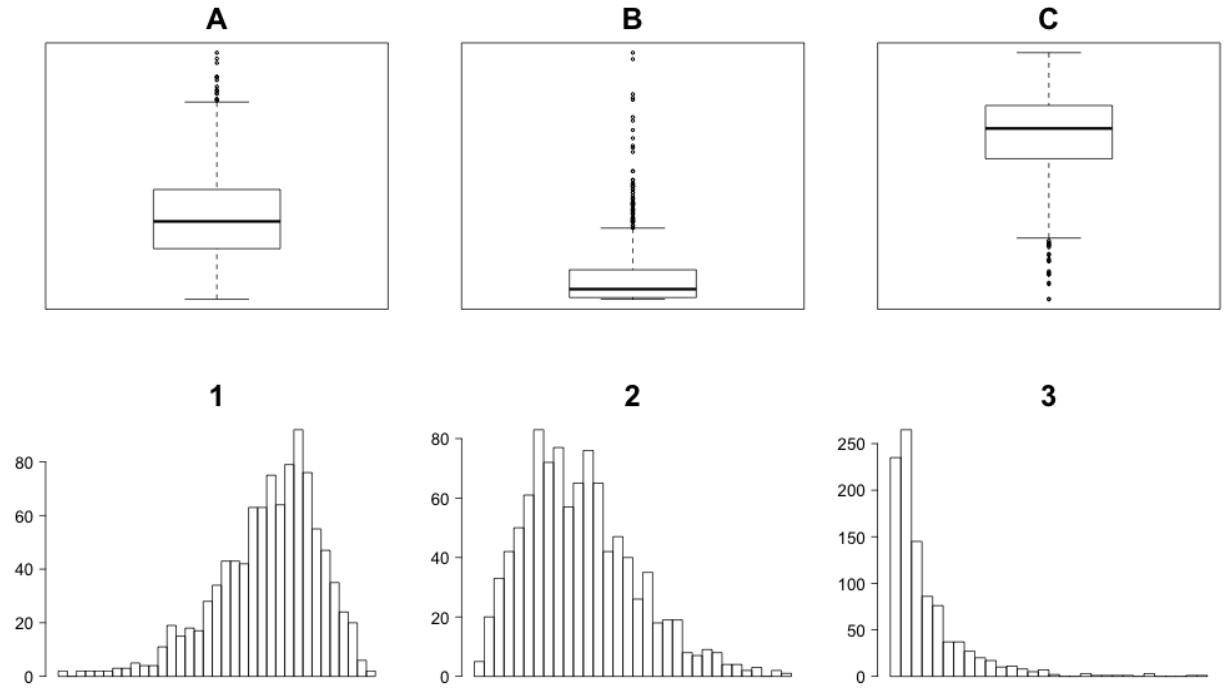
Also notice in the .Rmd how I used the `>` before my answer. This is how you put text in block quotes in Rmarkdown.

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 server, 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, just like for the labs.

Questions

Question 1 (6 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.



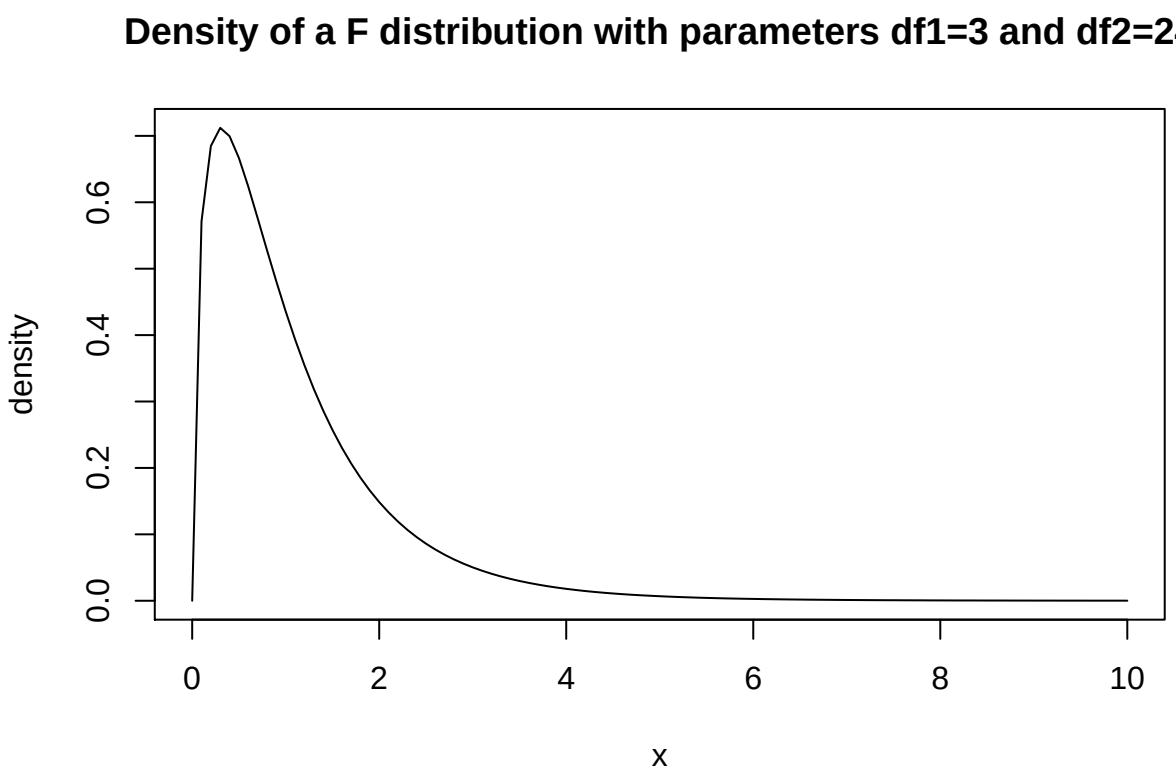
Answer is A-2, B-3, C-1. (B) The distance between the third quantile (75 percentile) andmedian is larger than the distance between the first quantile (25 percentile) and the median.The values of the outliers are large. Thus, the distribution may be right skewed. (A) Thedistribution may be right skewed as well, but the median of A is greater than B. We canconclude that A-2, B-3. (C) The distance between the third quantile (75 percentile) andmedian is less than the distance between the first quantile (25 percentile) and the median.And the values of the outliers are small Thus, the distribution may be left skewed. The onlyleft skewed distribution in the histograms is 1.

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.

An example of a common distribution is the F distribution. The functions for this distributions are `df`, `pf` and `rf`; 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 F distribution has two parameters called `df1` and `df2` by R (`df` stands for 'degrees of freedom'). 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 F distribution with parameters `df1=3` and `df2=24` using the density function. Describe how this distribution compares to a normal probability distribution.

```r
# Insert code here for plotting the density function.
curve(df(x, df1 = 3, df2 = 24),
      ylab = "density",
      xlim =c(0, 10),
      main = "Density of a F distribution with parameters df1=3 and df2=24")
```



The F distribution is right skewed, while the normal distribution is centered at and symmetric around its mean. The F distribution is defined on $(0, \infty)$, while the normal distribution is defined on $(-\infty, \infty)$.

b. (5 points) Find the probability of an observation from this distribution being between 1 and 4, using R commands, and make sure the answer prints out so that it shows up in the pdf.

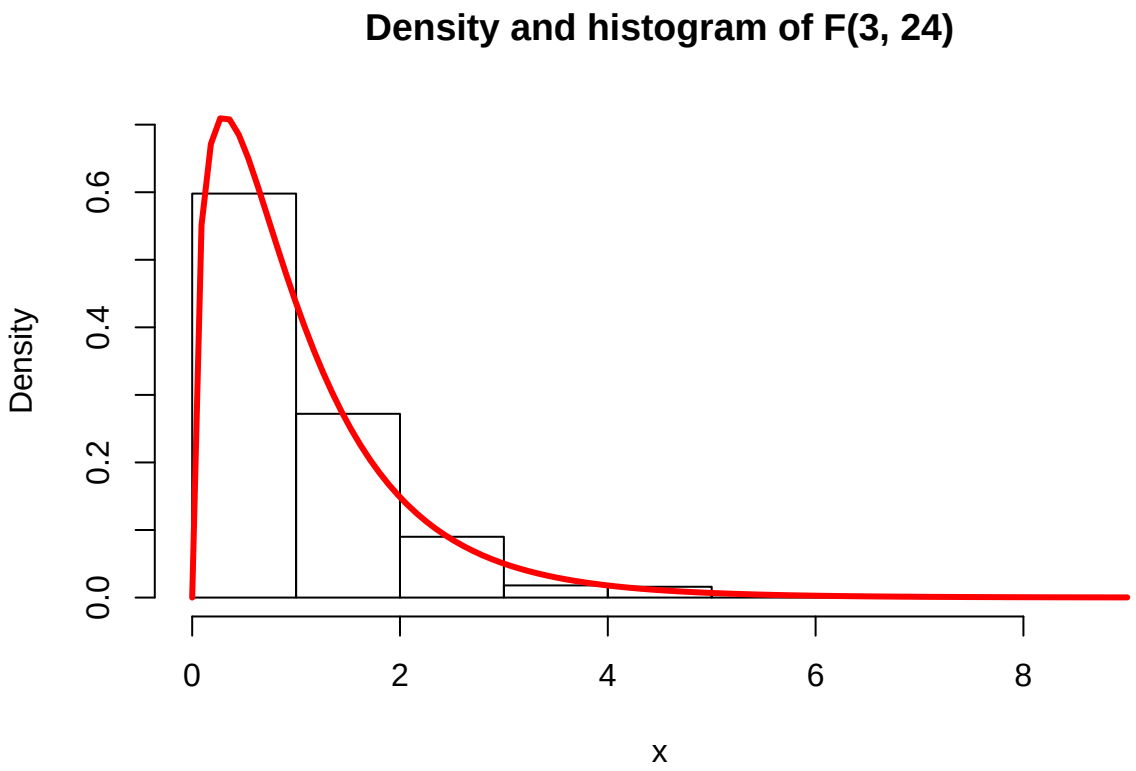
```r
# Insert code here for calculating the probability of being  in [1,4]
# Save the response as 'probF'
probF <-pf(4, df1 = 3, df2 = 24)- pf(1, df1 = 3, df2 = 24)
probF
```

```
## [1] 0.3906144
```

c. (15 points) Now simulate 500 observations from the above F 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.

```r
set.seed(51920)
# Insert code here for simulating the data and making a histogram from the data.
# Save the simulated data as `simF`.
simF <- rf(500, df1 = 3, df2 = 24)
hist(simF, freq = FALSE, ylim =c(0, 0.7), xlab = "x",
     main = "Density and histogram of F(3, 24)")
```

```
curve(df(x, df1 = 3, df2 = 24), add = T, col = "red", lwd = 3)
```



d. (8 points) Imagine that this simulated data was actually observed data given to you and you didn't know the actual probability distribution of the data. You want to use this simulated data to estimate the probability distribution. What would be your *estimate* of being between between 1 and 4? (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?

```
# Insert code here for estimating the probability of being in [1,4]
# from simulated data created in previous chunk.
# Save the response as 'probFEst'
probFEst = mean(1 <= simF & simF <= 4)
probFEst
```

```
## [1] 0.38
```

It is close to the actual probability for the true distribution.

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 < 0.5
- Probability of an observation between 1 and 4
- Probability of an observation > 4

The second would be likely to be better estimated from the data, since it has the largest prob-ability. The third one would require larger sample size since the probability of an observation$>$ 4 is comparatively small.

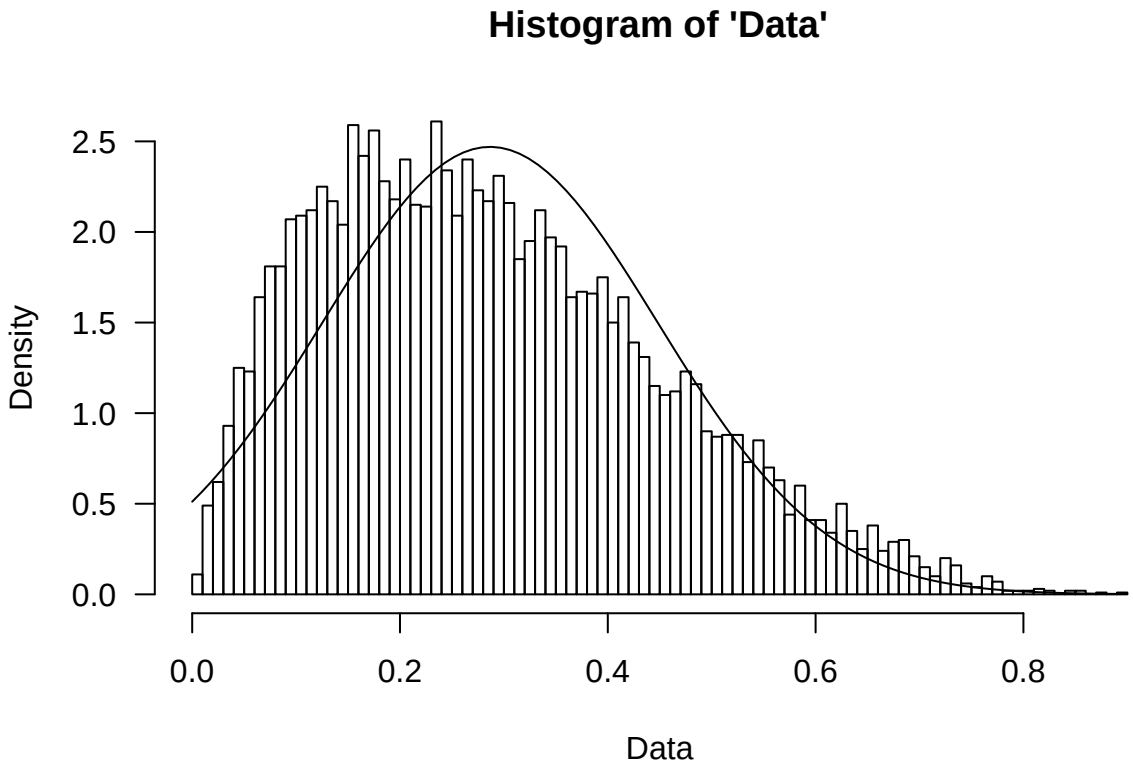
Question 3 (10 points) For the small set of toy data saved in 'histogramData.txt', we want to compare the distribution of this data to a normal density curve to see if it might be distributed reasonably closely to normal. Read in the data with the code below, making sure the data matches what you see 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
## 1 0.06802080
## 2 0.13054032
## 3 0.32468638
## 4 0.05886398
## 5 0.10394246
## 6 0.48470550
```

However, 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))
f<-function(x){dnorm(x, mean =mean(mydf$Data), sd =sd(mydf$Data))}
curve(f,add=TRUE)
```



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

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

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal num
## 1  63   1  1      145  233   1       2     150     0     2.3     3  0    6   0
## 2  67   1  4      160  286   0       2     108     1     1.5     2  3    3   2
## 3  67   1  4      120  229   0       2     129     1     2.6     2  2    7   1
## 4  37   1  3      130  250   0       0     187     0     3.5     3  0    3   0
## 5  41   0  2      130  204   0       2     172     0     1.4     1  0    3   0
## 6  56   1  2      120  236   0       0     178     0     0.8     1  0    3   0
```

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.
- `restecg`: resting electrocardiographic results
- Value 0: normal
- Value 1: having ST-T wave abnormality (T wave inversions and/or ST levation or depression of > 0.05 mV)
- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- `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
- `chol` the serum cholestoral in mg/dl

Notice that `cp` and `restecg` are encoded in this data as numeric values, but are actually categorical (this is a quite common practice). 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$cp` and `heart$restecg` everywhere):

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

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

```
with(heart, table(restecg))
```

```
## restecg
##   0   1   2
## 147   4 146
```

a. (5 points) Change both of these variables to be `factor` variables *inside your* ***`heart`*** *data frame*. Give the different levels of the `cp` variable the labels described above. For `restecg`, use the labels "normal", "ST-T wave", "ventricular hypertrophy"

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

```
                      levels = c(0, 1, 2),
                      labels = c("normal", "ST-T wave", "ventricular hypertrophy"))
```

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","restecg")])
```

```
##                 cp                         restecg
##  typical angina  : 23   normal                 :147
##  atypical angina : 49   ST-T wave              :  4
##  non-anginal pain: 83   ventricular hypertrophy:146
##  asymptomatic    :142
```

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

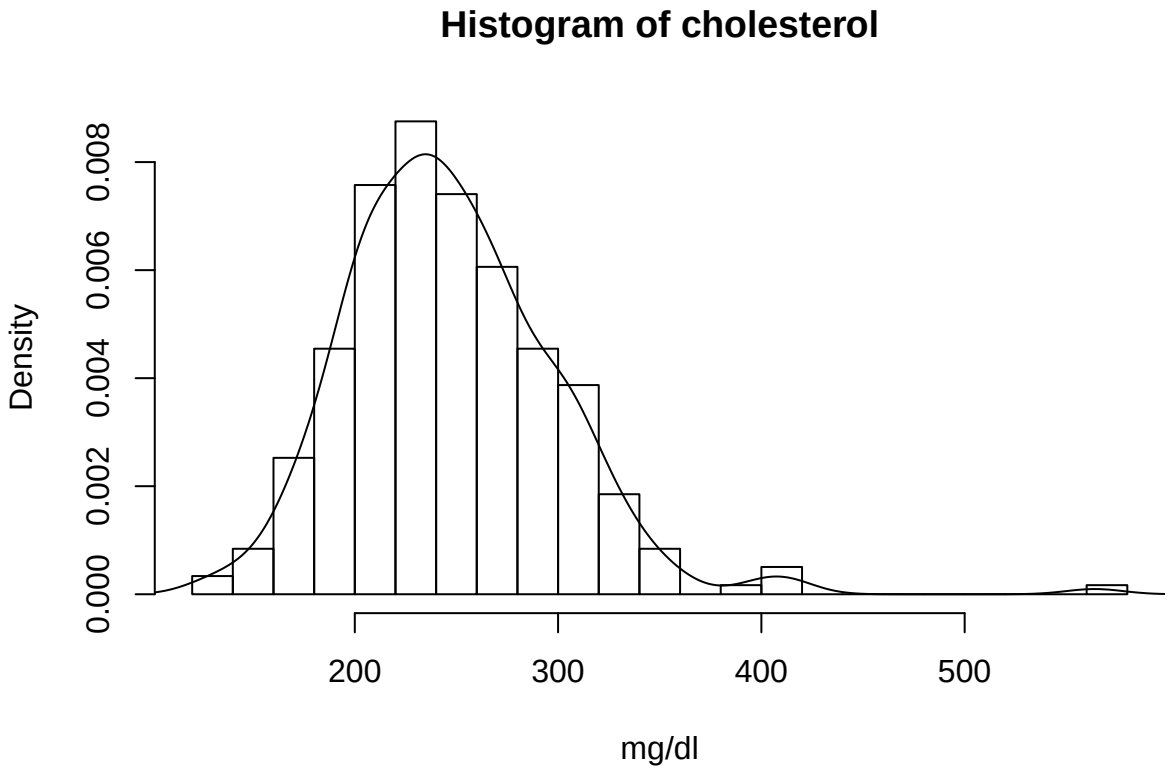
```
# Insert code here for contingency table
table(heart$cp, heart$num)
```

```
##
##                     0  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
```

> Those with the most severe final heart disease were overwhelmingly asymptomatic withregards to chest pain. Those that are asymptomatic are much more likely to have severeoutcomes at any level

c. (10 points) Create a histogram of cholesterol (`chol`), 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$chol, freq = F, breaks = 20,
     main = "Histogram of cholesterol",xlab = "mg/dl")
lines(density(heart$chol))
```



My answer ...

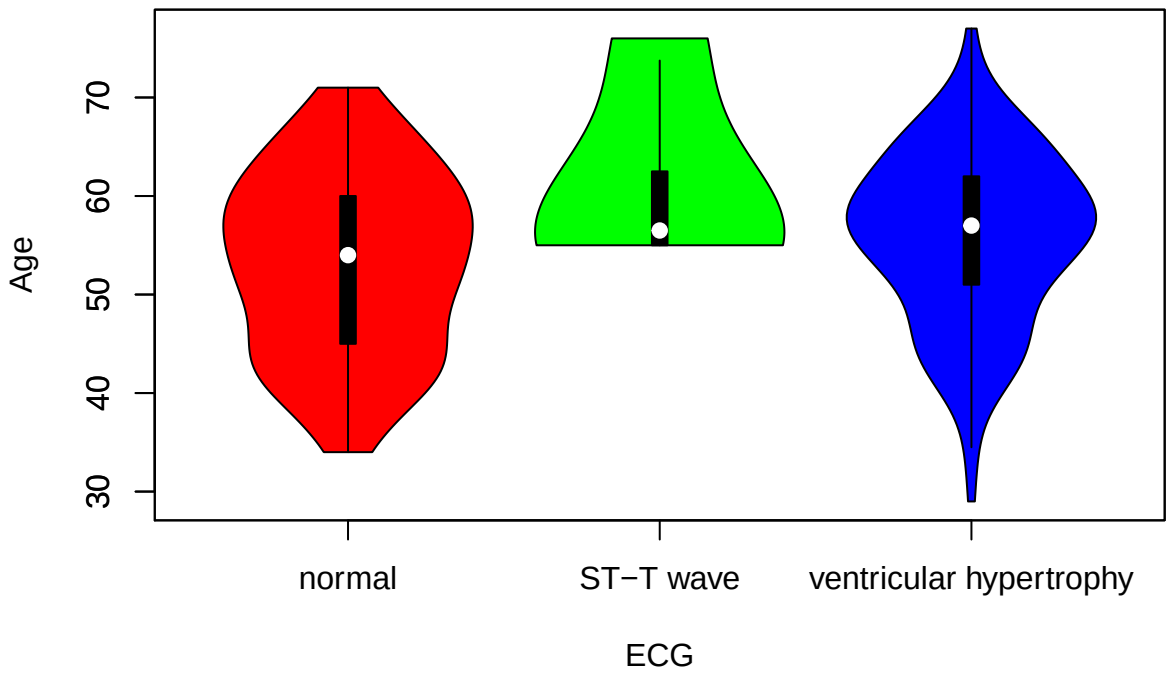
d. (10 points) Create a violin plot of the age of the patients (`age`), separated by their resting electrocardiographic results (`restecg`). In otherwords, on one plot, a violin plot of `age` for each of the 5 categories of the diagnosis [Use Professor Purdom's version `vioplot2` of the vioplot function. The `source` command given below accesses it from the web].

Comment on any differences between ages for the different levels of heart disease.

```
#loads Prof. Purdom's function:
source("https://www.stat.berkeley.edu/~epurdom/RcodeForClasses/myvioplot.R")
# Insert code here for violin plots:
palette(rainbow(3))
vioplot2(heart$age, heart$restecg, col =palette(), ylab = "Age", xlab = "ECG")
```

```
## Loading required package: vioplot

## Warning: package 'vioplot' was built under R version 3.6.2

## Loading required package: sm

## Warning in fun(libname, pkgname): couldn't connect to display ":0"

## Package 'sm', version 2.2-5.6: type help(sm) for summary information

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.2

##
## Attaching package: 'zoo'

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

```
##     as.Date, as.Date.numeric
```



Not many people have ST-T wave, 4 to be precise, so the plot appears truncated. Thedistributions of age for the normal and ventricular hypertrophy groups appear similar.

Probability problems

Question 5 (3 points) Say we have a box with 2 golden rings and 3 silver rings in it. I reach in and pull out one ring at a time, at random without replacement, and stop when I get a golden ring. Let X be the number of tries that it takes, up to and including the first time I pull out a golden ring. What are the possible values of X and their probabilities?

Solution:
The possible values of X are 1, 2, 3, 4:

$$P(X=1) = \frac{2}{5} = 0.4$$
$$P(X=2) = \frac{3}{5} \times \frac{2}{4} = 0.3$$
$$P(X=3) = \frac{3}{5} \times \frac{2}{4} \times \frac{2}{3} = 0.2$$
$$P(X=4) = \frac{3}{5} \times \frac{2}{4} \times \frac{1}{3} \times \frac{2}{2} = 0.1$$

Question 6 A box of 100 tickets contains 45 red tickets, 50 blue tickets, and 5 yellow tickets. 20 tickets will be drawn from this box at random,*without* replacement.

a. (1 point) What is the expected number of **red** tickets in the sample?

Solution:
Let X be the color a ticket in the sample: 1 if it is red and 0 if it is not.
X is a Bernoulli trial, with $p = 0.45$ and we have 20 tickets in the sample.
Let us define Y as the sum of the 20 Bernoulli trials. We note that Y follows a Binomial(n = 20, p = 0.45) distribution.
$E(Y) = np = 20(0.45) = 9$

b. (2 points) What is the probability that there are no **yellow** tickets among the 20 tickets drawn?

Solution:
Let A_i be the event that the i-th ticket drawn is not yellow.
The probability we are calculating is $P(A_1 \cup A_2 \cup ... \cup A_{20})$.
$P(A_1 \cup A_2 \cup ... \cup A_{20}) = \frac{95}{100} \times \frac{94}{99} \times ... \times \frac{76}{81} = 0.3193$.
The probability that there are no yellow tickets among the 20 tickets drawn is 0.3193.

c. (4 points) Now we will put these 20 tickets back in the box, shuffle the tickets, and draw 20 tickets again, at random without replacement. If there is at least one yellow ticket in our sample, we will stop. If not, we will repeat the procedure of drawing 20 tickets at random without replacement and checking for at least one yellow ticket. What is the probability that we will see a sample with at least one yellow ticket for the first time on the third try?

Solution:
Let W be the number of attempts in drawing 20 tickets until the seeing a yellow ticket inthat attempt.
We can see that W follows a geometric distribution with probability of success $p = 1-0.3193 = 0.6807$.
The probability of failure is calculated and defined in the previous part b.
$P(W = 3) = (1-p)^2 \times p = 0.3193^2 \times 0.6807 = 0.0694$

Question 7 Consider the following density function defined on $(0, 1) : f(x) = cx^2$.

a. (3 points) Find the value of c.

$\int_0^1 f(x)dx = 1 = \int_0^1 cx^2dx = c\int_0^1 x^2dx = 1$ So we have $c = \frac{1}{\int_0^1 x^2dx} = \frac{1}{\frac{1}{3}(1)^2 - \frac{1}{3}(0)^2} = 3$

b. (2 points) Find the cdf.

$F(x) = \int_0^x f(t)dt = \int_0^x 3t^2dt = (x)^3 - (0)^3 = x^3$

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

Plugging in $\frac{1}{3}$, we have $F(\frac{1}{3}) = (\frac{1}{3})^3 = \frac{1}{27}$

d. (1 points) What is $P(0.1 \leq X \leq 0.5)$?

$P(0.1 \leq X \leq 0.5) = \int_{0.1}^{0.5} f(t)dt = (0.5)^3 - (0.1)^3 = 0.124$

Question 8 (6 points) Sketch the pdf and cdf of a random variable that is uniform on [-1,1].

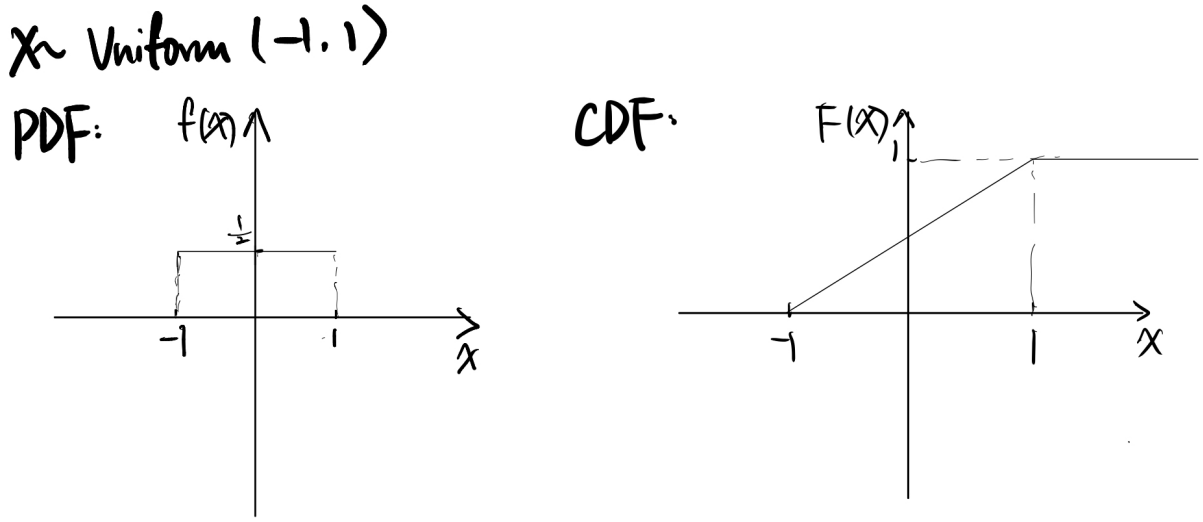


Figure 1: Alt