DS 2006 Midterm 2

NAME: Makhdum Mourad Shah

NOTE: The `.rmd` version of the file is available here: (link)

Instructions

Please prepare reponses/solutions for the following questions.

Allowable resources

You may use the solutions you've prepared for the prep questions during the exam. You are welcome to use your homeworks, deliverables, or class notes. You are permitted to access the internet for publicly available content. You are not allowed to communicate with anyone via the internet or any other means during the exam. This includes, but is not limited to:

- No messaging, emailing, or using social media to contact others.
- No posting questions or seeking answers on forums, chat rooms, chat bots (including large language models like ChatGPT), or any collaborative platforms.
- No sharing or discussing exam content with peers through any online or electronic medium.

You may **NOT** discuss any aspect of the exam or prep questions with anyone other than the instructor or TA. You may **NOT** share code or documents.

Submission instructions

1. Within your course repo, create a folder called `Midterm2`
2. Within the folder, create the script file `exam.rmd` with your solutions. Create a rendered report in `.pdf` output.
3. Add, commit, and push to your repo on github.com.

Questions

Exam questions are organized into sections cooresponding to the learning outcomes of the course.

Section 1. Tools of the data scientists

> Learning objective: Use the tools of data scientists
>
> Learning objective: Implement best programming/coding practices

1.1 [5 pts] Write your name at the prompt above (line 6 in the script).

1.2 [5 pts] When you are done with the exam, please render this report as a pdf document.

1.3 [5 pts] The following is a schematic of a project folder, with subfolders and files.

```
project
|
|---code
|       script.rmd
|
|---data
|       survey-responses.csv
|
|---docs
```

Supposing the `code` subfolder is the designated working directory, write the command to be included in the `script.rmd` file which will read the `survey-responses.csv` data, avoiding absolute file paths?

```r
survey <- read.csv("../survey-responses.csv")
```

Section 2. Probability & Diagnostics

> Learning objective: Compare and contrast different definitions of probability, illustrating differences with simple examples

> Learning objective: Express the rules of probability verbally, mathematically, and computationally.

> Learning objective: apply cross table framework to the special case of binary outcomes

2.1 [5 pts] In a particular town, 1% of all individuals have a certain rare disease. There's a test for this disease that correctly identifies a sufferer 99% of the time (true positive rate) but also falsely identifies the disease in 2% of the healthy population (false positive rate).

Complete the following table of cell, row, and column probabilites based on the information about the prevalence, true positive rate, and false positive rate. You are welcome to use an excel spreadsheet (link) which will automatically create the table for any combination of prevalence, sensitivity, and specificity. (It might save time to insert a screen shot of the table rather than manually creating the table.)

	Disease +	Disease -	
Test +			
→ cell	0.0099	0.0198	0.0297
→ row	0.333	0.6666	
→ col	0.99	0.02	
Test -			
→ cell	0.0001	0.9702	0.9703
→ row	0.0001	0.9999	
→ col	0.01	0.98	
	0.01	0.99	1

2.2 [5 pts] Suppose that a new test is developed and approved 98% true positive rate and 1% false positive rate. As a consumer, which would you prefer? Be specific about your reasoning and the quantities that you are using to make a decision.

I would still prefer the first test, given that it is more serious if I have the disease and I test negative for it. Therefore, I would prefer a higher PPV as it is more important to me.

2.3 [10 pts] An audit of an email filtering system resulted in a dataset of 10000 emails, each manually verified as spam or not spam. In addition to the type of email, the dataset indicates if the filter sent the email to the inbox or the junk folder.

The following command reads the dataset into memory. From the data, generate an estimate of the positive predictive value and the negative predictive value of the spam filter.

PPV= P(Spam|Junk Folder) 0.1792/0.2214

NPV= P(Not Spam|Inbox Folder) 0.7859/0.7786

```
library(magrittr)
d1 <- readRDS(url("https://tgstewart.cloud/spam-data.RDS"))
cross_table <- table(d1$Type, d1$Folder)%>%proportions()%>%addmargins()
cross_table
```

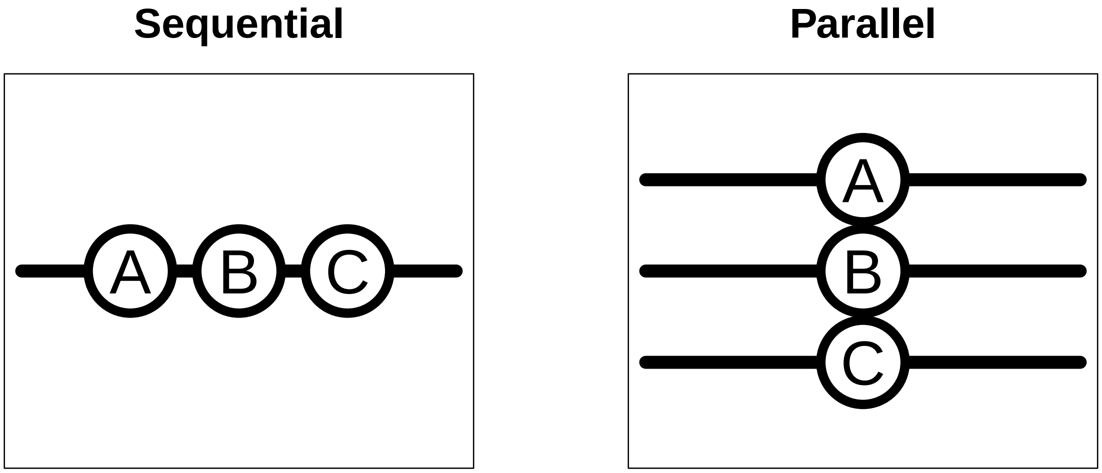
```
##
##               Inbox   Junk    Sum
##   Not spam 0.7589 0.0422 0.8011
##   Spam     0.0197 0.1792 0.1989
##   Sum      0.7786 0.2214 1.0000
```

Section 3. Simulation

> Learning objective: Use probability models to build simulations of complex real world processes to answer research questions

3.1 [5 pts] Consider two systems of three components (say A, B, and C). In the first, a failure of any component leads to a failure of the entire system. In the second, the components are redundent, and a failure only occurs if all three components fail.

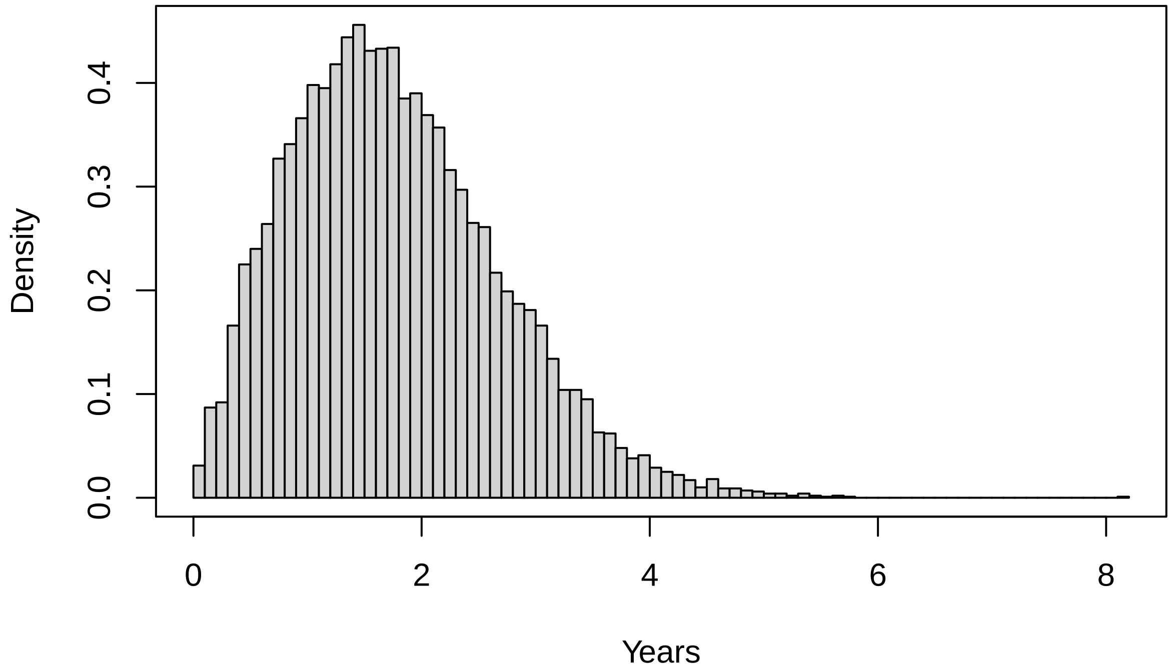
```
## pdf
##   2
```



Sequential

Parallel

Suppose the failure time of an individual component is a random variable with the following distribution.



The following function `sysfail` generates replicates of the time to system failure (years) for the sequential and parallel systems. The input parameter `R` is the number of replicates that the fuction will return.

```r
sysfail <- function(R){
    A <- array(rweibull(R*3,2,2),dim = c(R,3))
    data.frame(sequential = apply(A,1,min), parallel = apply(A,1,max))
}
```

The following provides an estimate of how much longer the parallel system will last compared to the sequential system by simulating 25 different system failures.

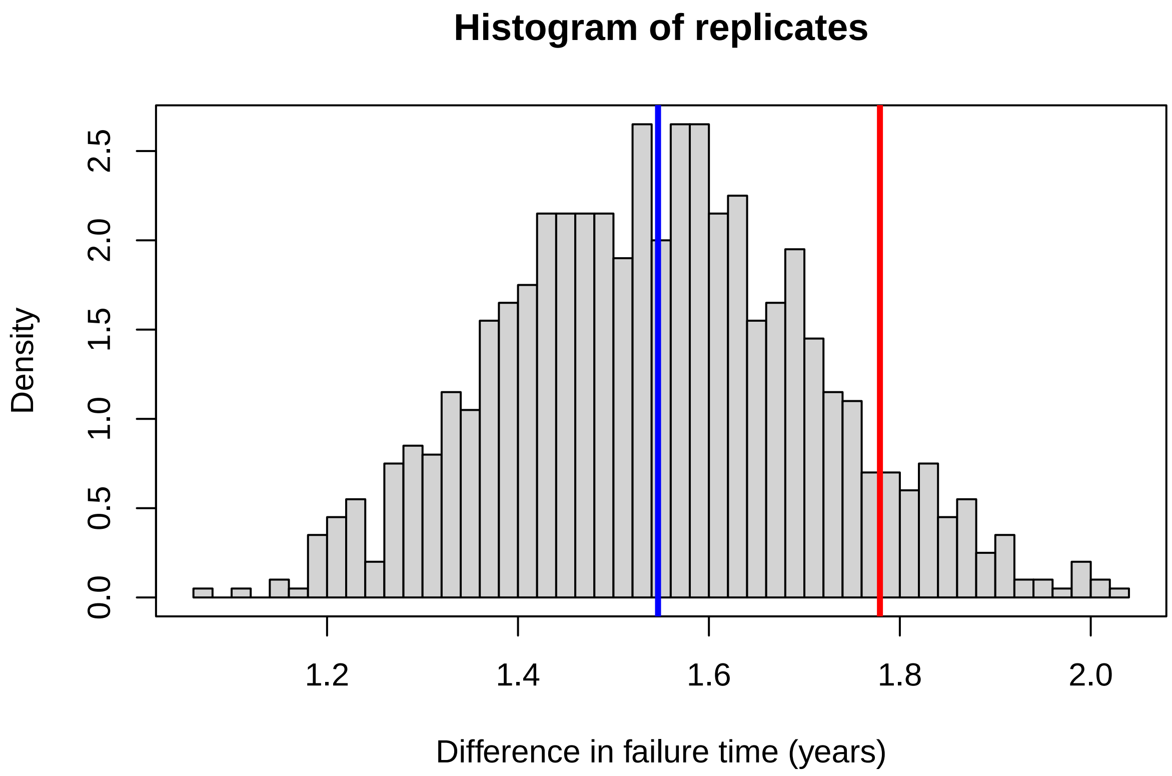
```r
set.seed(230583)
a1 <- sysfail(25)
a2 <- colMeans(a1)
diff(a2)
```

```
## parallel
## 1.779168
```

The calculated difference is based on pseudo-random data. The process can be repeated many times. The following code creates 1000 estimates. The redline refers to the single estimate generated above.

```r
R <- 25
replicates <- replicate(1000, sysfail(R) |> colMeans() |> diff())
hist(replicates, breaks = 50, freq = FALSE, xlab = "Difference in failure time (years)"); box()
```

```r
abline(v=diff(a2), lwd = 3, col = "red")
x<-mean(replicates)
abline(v=x, lwd = 3, col = "blue")
```



Add to the figure blue reference line for the mean of the 1000 estimates. (Simply edit the code chunk above. You do not need to create a second copy.)

3.2 [5 pts] What is the range (min and max) of the 1000 values you generated for the improved failure time estimate?

```
min_estimate <- min(replicates)
max_estimate <- max(replicates)


range<-max_estimate-min_estimate
range
```

```
## [1] 0.9543089
```

```
true_value <- x

# Calculate absolute errors for each estimate
absolute_errors <- abs(diff(a2) - x)

# Calculate the average absolute error
average_absolute_error <- mean(absolute_errors)

average_absolute_error
```

```
## [1] 0.2324734
```

```
relative_errors<-(diff(a2)/x)
average_relative_errors<-mean(relative_errors)
average_relative_errors
```

```
## [1] 1.150303
```

The range is 1.099354

3.3 [5 pts] What is the average absolute error of the 1000 estimates? Use the mean calculated in 3.1 as the “true” value.

The Avg abs error is 0.2177563

3.4 [5 pts] What is the average relative error of the 1000 estimates? Use the mean calculated in 3.1 as the “true” value.

1.139461

3.5 [5 pts] If you wanted to reduced the error by half, how many replicates (R) would you need to use?

In order to reduce the error by half, the number of replicates must be quadrupled! the formula which explain the relationship berween error that we used in deliverable 2 is E=2ˆalpha/sqrt(4R)

3.6 [5 pts] Generate a plot of overlapping histograms to show the difference between R=25 and your R from the previous problem.

```
R2 <- 100 # Change this
replicates2 <- replicate(2000, sysfail(R2) |> colMeans() |> diff())

b1 <- seq(min(c(replicates,replicates2))-.01, max(c(replicates,replicates2))-0.01,by=0.02)
hist(replicates2, breaks = b1, col = "#ffabab50", freq = FALSE, main = "", xlim = range(replicates), xla
legend("topleft", legend = c("R=25", paste0("R=",R2)), col = c("#ffabab50","#6488ea59"),bty = "n", pch =
hist(replicates, breaks = b1, add=TRUE, col = "#6488ea59", freq = FALSE)
box()
```

3.7 [5 pts] Calculate the average absolute error of the 1000 estimates generate with the new choice of R? Did it change as you expected it to?

The average absolute error halfed.

Section 4. Confounding vs Causal Pathway

> Learning objective: define/describe confounding variables, Simpson's paradox, DAGs, and the causal pathway

4.1 [10 pts] The following function generates data from a cohort of individuals who agreed to be studied about heart disease. In the dataset, there is exercise level at age 20 (below average, above average), blood pressure at age 25 (low, normal, high), and heart disease at age 30 (present, absent).

Generate 10000 draws from the function and create the cross table for exercise level and heart disease. Calculate a summary of the effect of exercise by calculating the risk ratio:

$$RR = \frac{P(\text{heart disease present}|\text{below average exercise})}{P(\text{heart disease present}|\text{above average exercise})}$$

```r
library(magrittr)
heart_data <- function(R){
    ex <- rbinom(R,1,.5)
    bp <- rnorm(R,-ex+1/2,1)
    bp <- cut(bp,c(-Inf,-1,1,Inf), labels = FALSE)
    hd <- 1*(rnorm(R,bp-3,1.8)>0)
    data.frame(
        exercise = factor(ex,0:1,c("below average","above average"))
      , blood_pressure = factor(bp,1:3,c("low","normal","high"))
      , heart_disease = factor(hd, 0:1, c("absent","present"))
    )
}


set.seed(20240329); d1 <- heart_data(1000000)


cross_table <- table(d1$exercise, d1$heart_disease)%>%proportions()%>%addmargins()
cross_table
```

```
##
##                    absent  present      Sum
##   below average 0.328351 0.171560 0.499911
##   above average 0.372857 0.127232 0.500089
##   Sum           0.701208 0.298792 1.000000
```

combined= 0.343/0.2544

4.2 [10 pts] Stratify the table by blood pressure. As in the previous problem, calculate the same treatment effect in each strata.

```
d2<-d1[d1$blood_pressure=="low", ]
table(d2$exercise, d2$heart_disease)%>%proportions()%>%addmargins()
```

```
##
##                     absent   present       Sum
##   below average 0.1545983 0.0238541 0.1784524
##   above average 0.7121986 0.1093491 0.8215476
##   Sum           0.8667968 0.1332032 1.0000000
```

```
d3<-d1[d1$blood_pressure=="normal", ]
table(d3$exercise, d3$heart_disease)%>%proportions()%>%addmargins()
```

```
##
##                     absent   present       Sum
##   below average 0.3556224 0.1439744 0.4995968
##   above average 0.3561696 0.1442336 0.5004032
##   Sum           0.7117919 0.2882081 1.0000000
```

```
d4<-d1[d1$blood_pressure=="high", ]
table(d4$exercise, d4$heart_disease)%>%proportions()%>%addmargins()
```

```
##
##                      absent    present        Sum
##   below average 0.41115809 0.41114210 0.82230019
##   above average 0.08924982 0.08844999 0.17769981
##   Sum           0.50040791 0.49959209 1.00000000
```

4.3 [5 pts] Based on the summary of the treatment effect that you observed in the combined and stratified tables, is exercise associated with lower rates of heart disease?

low= 1.004

med= 1.004

high=1.004

The risk ratio is positive and remains the same across all tables, this suggests that mpre than average exercise is associated with lower rates of heart disease.

4.4 [5 pts] Which measure of treatment effect is most persuasive? The combined estimate or the stratified estimates? Which estimate(s) should you rely on? Explain why, creating a DAG to represent relationship between the variables.

The combined estimate is the most persuasive, because it is only looking at the relationship between heart disease and exercise. This is what we need. Blood pressure is irrelevant to us as it is unknown at the time of exercise. We know this because stratifiying the tables does not change the risk ratio. Therefore, it is not a confounding variable.

DAG=

Exercise → Heart disease

```
  ↓                  ↑

      Blodd Pressure
```