Final Exam

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

Instructions

Please prepare reponses/solutions for the following questions. On the day of the exam, you will be given a new set of questions. You will use the solutions you've prepared for this exam during the exam.

During the exam, you will also be permitted to access the internet for publicly available content. You will not be 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 `final-exam`
2. Within the folder, create the script file `final-exam.rmd` with your solutions. Create a rendered report in `.pdf` output.
3. Add, commit, and push to your repo on github.com.
4. If you received an email from gradescope, upload your PDF to gradescope.

Questions

All questions, including extra credit are 5 pts.

1. In the examples below, please explain/define what is meant by the word "probability".

> I think there is an 0.8 probability that the defendent committed the crime.

This is an example of belief informed probability, where our prior information and knowledge allows us to estimate the probability. This is more like an opinion. Probability is the strength of our beliefs that the likelihood that the defendant committed the crime is 0.8.

> The probability of winning the powerball is less than 0.00001.

This could be an example of data informed probability where it is less of an opinion and more of a fact. This statement asserts that there is a very low chance of winning the powerball(0.00001).

2. The following table is based on a study of delivery method and postnatal depression (link). (You do not need to read the publication.) In a cohort of mothers, researchers collected delivery mode and depression scores (8 weeks postpartum). The data were collected in 1991 and 1992. While some planned vaginal deliveries did result in emergency caesarean section or assisted vaginal delivery, the variable of interest was the **planned** delivery mode.

Suppose that the cell probabilities were provided as a, b, c, and d as in the table below. Complete the rest of the table symbolically.

	Depression score < 13	Depression score ≥ 13	All
Planned vaginal delivery	a	b	a+b
row	a/(a+b)	b/(a+b)	
col	a/(a+c)	b/(b+d)	
Planned caesarean section delivery	c	d	c+d
row	c/(c+d)	d/(c+d)	
col	c/(a+c)	d/(b+d)	
All	a+c	b+d	a+b+c+d

3. Now suppose that rather than cell probabilities, conditional probabilities were collected. Define postnatal depression as a depression score ≥ 13, and let

$$e = P(\text{postnatal depression}|\text{Planned vaginal delivery})$$

$$f = P(\text{postnatal depression}|\text{Planned caesarean section delivery})$$

$$g = \text{incidence of planned caesarean section.}$$

Complete the table symbolically.

	Depression score < 13	Depression score ≥ 13	All
Planned vaginal delivery	1-g-e(1-g)	e(1-g)	1-g
row	(1-g-e(1-g))/(1-g)	e	
col	(1-g-e(1-g))/(1-g-e(1-g)+g-fg)	(e(1-g))/(e(1-g)+fg)	
Planned caesarean section delivery	g-fg	fg	g
row	(g-fg)/g	f	
col	(g-fg)/(1-g-e(1-g)+g-fg)	fg/(e(1-g)+fg)	
All	1-g-e(1-g)+g-fg	e(1-g)+fg	1

4. (Continuing from the previous problem.) If planned caesarean section is 30% of all deliveries, and the risk of postnatal depression is 0.1 in the planned vaginal delivery group and 0.15 in planned caesarean section delivery groups, what is

$$P(\text{Planned caesarean section delivery}|\text{Depression score } < 13)?$$

(g-f*g)/(1-g-e*(1-g)+g-f*g)=0.2881356

5. Suppose observational data were collected in which depression rates matched the proportions in question **4**. Would the data support the conclusion that caesarean section delivery leads to higher rates of depression? If not, why not? (Hint: Recall chapter 8 of the text "Understanding Uncertainty".)

Because this is observational rather than experiemental data, we cannot make a conclusion on the effectiveness of the two treatments. We did not control for other confounding factors. These observational data only imply correlation and can not determine causation.

6. The Monte Hall problem is a classic game show. Contestants on the show where shown three doors. Behind one randomly selected door was a sportscar; behind the other doors were goats.

At the start of the game, contestants would select a door, say door A. Then, the host would open either door B or C to reveal a goat. At that point in the game, the host would ask the contestant if she would like to change her door selection. Once a contestant decided to stay or change, the host would open the chosen door to reveal the game prize, either a goat or a car.

In this problem, consider a **modified** version of the Monte Hall problem in which the number of doors is **variable**. Rather than 3 doors, consider a game with 4 or 5 or 50 doors. In the modified version of the game, a contestant would select an initial door, say door A. Then, the host would open **one** of the remaining doors to reveal a goat. At that point in the game, the host would ask the contestant if she would like to change her door selection. Once a contestant decided to stay or change, the host would open the chosen door to reveal the game prize, either a goat or a car.

Consider two strategies:

1. Always stay with the first door selected.
2. Always switch to the unopened door.

The function `game` below plays a single game of Monte Hall. The function returns a vector of length two, the first element is the prize under strategy 1 and the second element is the prize under strategy 2. The function has a single input parameter, N, which is the number of doors in the game.

Use the `game` function to estimate the probability that strategy 1 results in a goat and strategy 2 results in a car. Let **N=5**.

```r
suppressPackageStartupMessages(require(magrittr))
suppressPackageStartupMessages(require(dplyr))

game <- function(N){
  if(N<3) stop("Must have at least 3 doors")
  prize <- sample(c(rep("goat",N-1),"car"), N)
  guess <- sample(1:N,1)
  game <- data.frame(door = 1:N, prize = prize, stringsAsFactors = FALSE) %>%
    mutate(first_guess = case_when(
      door == guess ~ 1
      , TRUE ~ 0
    )) %>%
    mutate(potential_reveal = case_when(
        first_guess == 1 ~ 0
      , prize == "car" ~ 0
      , TRUE ~ 1
    )) %>%
    mutate(reveal = 1*(rank(potential_reveal, ties.method = "random") == 3)) %>%
    mutate(potential_switch = case_when(
      first_guess == 1 ~ 0
      , reveal == 1 ~ 0
      , TRUE ~ 1
    )) %>%
    mutate(switch = 1*(rank(potential_switch, ties.method = "random") == 3))
  c(game$prize[game$first_guess == 1], game$prize[game$switch == 1])
}


n_simulations<-1000
d1<-replicate(n_simulations,game(5)) |> t()
head(d1)
```

```
##      [,1]   [,2]
## [1,] "goat" "car"
## [2,] "goat" "car"
## [3,] "goat" "goat"
## [4,] "goat" "goat"
## [5,] "goat" "car"
## [6,] "goat" "car"
```

```
count <- sum(d1[,1] == "goat" & d1[,2 ] == "car")
probability <- count / n_simulations

probability
```

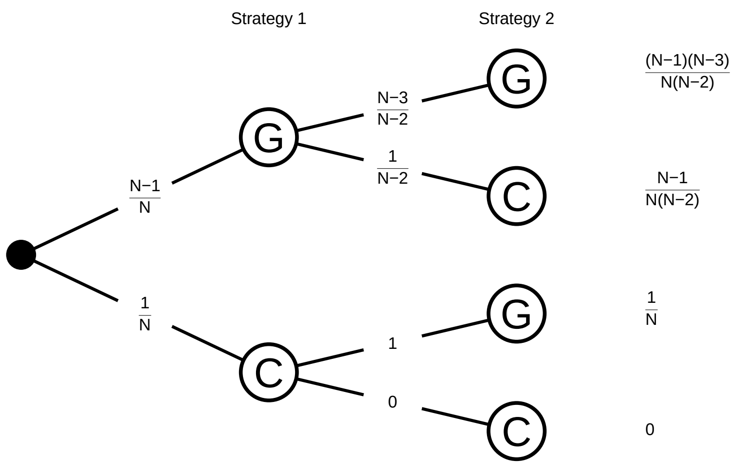
```
## [1] 0.27
```

7. Consider the following tree, a possible analytic solution proposed by your classmate for the Monte Hall game with N doors. Your classmate argues that at the start of the game, there is only $\frac{1}{N}$ chance of getting the car in the initial guess. Consequently, there is a $\frac{N-1}{N}$ of selecting a goat in the initial guess. The initial guess is the outcome of strategy 1.

If strategy 1 results in a goat, then the outcome of strategy 2 is either a goat or car. As the host as revealed a door with a goat behind it, there are N-2 doors to choose from, 1 of which hides a car and N-3 of which hide goats. (Or so your classmate argues.)

Likewise, your classmate argues that if strategy 1 results in a car, then the outcome of strategy 2 must be a goat.

Multiplying the probabilities along the pathway, your classmate argues, generates the probability of the path itself.



The joint distribution of the outcomes of strategy 1 and strategy 2 can be represented, then, with the following contingency table.

		Statragy 2	
		Car	Goat
Statragy 1	Car	0	$\frac{1}{N}$
	Goat	$\frac{N-1}{N(N-2)}$	$\frac{(N-1)(N-3)}{N(N-2)}$

Using simulation, check the solution of your classmate for N=5. Show the contingency table which results from simulation next to the proposed analytic solution proposed by your classmate. How well does the simulation solution match the proposed solution?

(N-1)/N(N-2) 4/(5*2)=4/10=0.4

```r
library(dplyr)
table(d1)
```

```
## d1
##  car goat
##  454 1546
```

8. Calculate the relative and absolute simulation error of your simulated probability in question **6**, supposing that the your classmate's solution in question **7** is correct.

```r
absolute_error<-abs(0.4-probability)
absolute_error
```

```
## [1] 0.13
```

```r
relative_error<-(abs(0.4-probability))/0.4
relative_error
```

```
## [1] 0.325
```

9. Consider a test for a rare genetic condition. Let T+ denote a test result that indicates the condition is present, while T- denotes absence. Let D+ and D- denote the true status of the disease.

Using the following information,

- P(T+|D+) = .85, and
- P(T-|D-) = .95, and
- P(D+) = 0.001

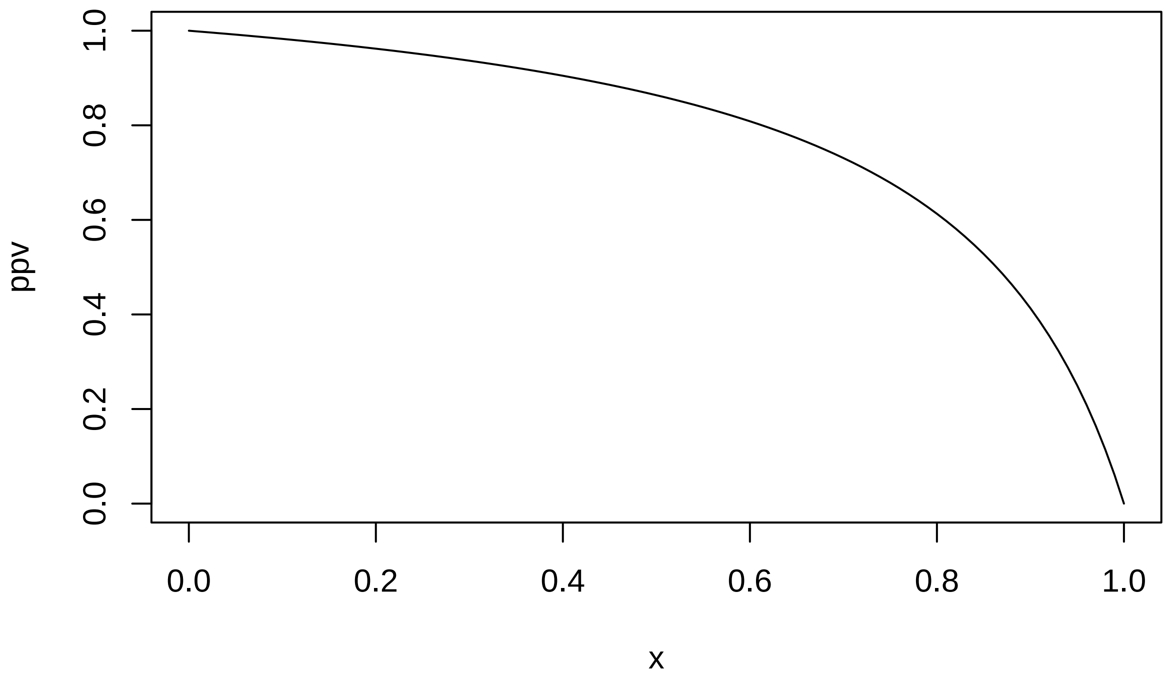
calculate the **negative** predictive value of the test, P(D-|T-).

P(D-|T-)=0.999841972

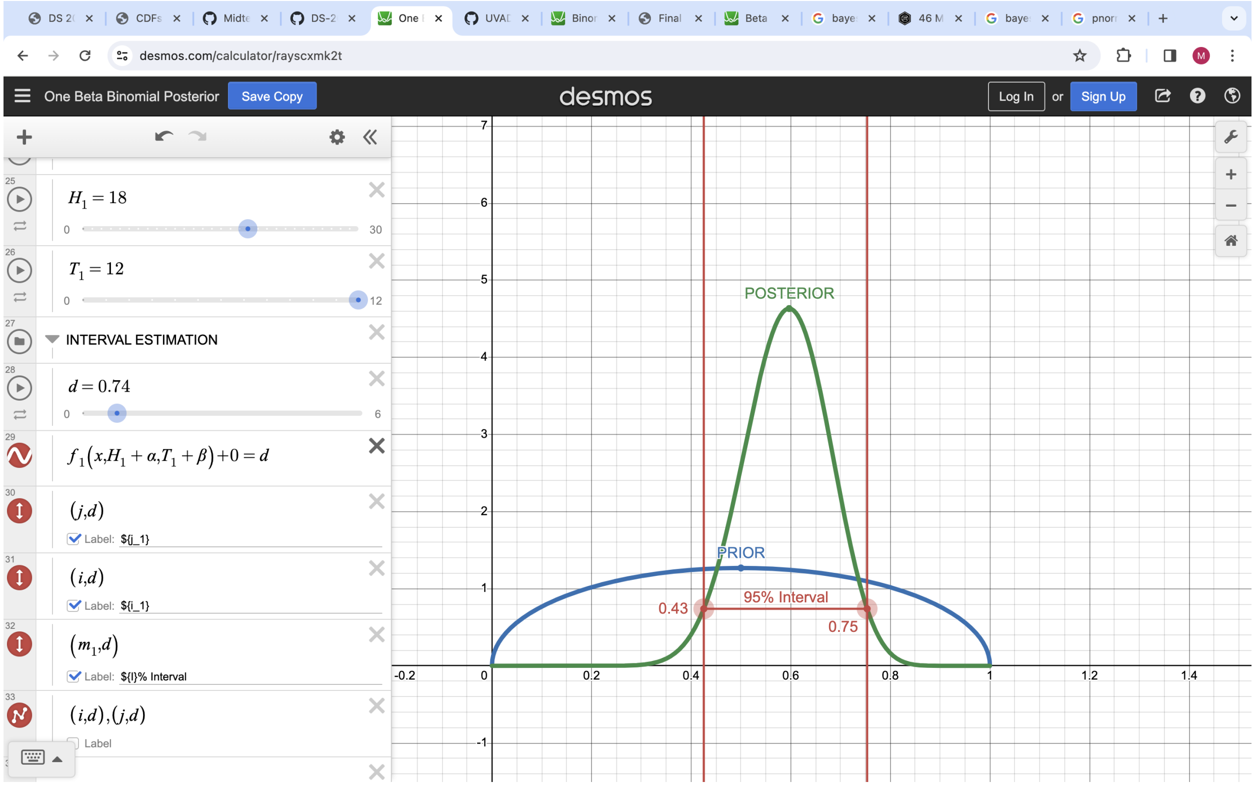
10. Create a plot that shows how the **positive** predictive value is a function of the prevalence of disease, P(D+). Keep the sensitivity and specificity the same as the previous question.

In the graph x is prevalence.

```r
y<-0:1
ppv<- function(y)(0.95*(1-y))/(0.95*(1-y)+(y-0.85*y))
plot(ppv)
```



11. Suppose an upcoming election for UVA student body president is between two candidates. In a survey of 30 students, 18 voiced support for candidate A. Use the Desmos calculator (link) to fit a probability model with Bayesian methods for the election, specifically the probability that candidate A is the prefered by the student body. Report the 95% credible interval. (Provide a screen shot of the calculator with your solution.)



25 $H_1 = 18$ (0 to 30)

26 $T_1 = 12$ (0 to 12)

27 INTERVAL ESTIMATION

28 $d = 0.74$ (0 to 6)

29 $f_1(x, H_1 + \alpha, T_1 + \beta) + 0 = d$

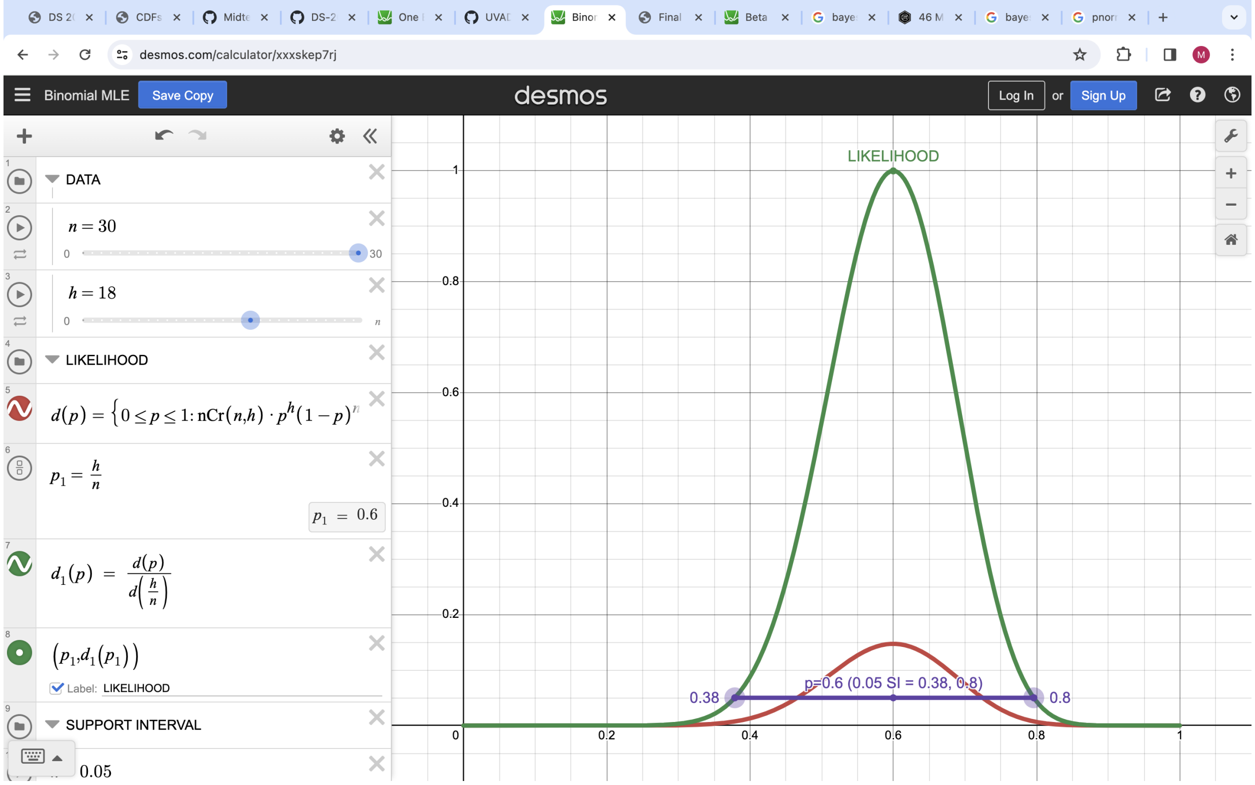
30 (j, d) Label: ${j_1}

31 (i, d) Label: ${i_1}

32 (m_1, d) Label: ${l}% Interval

33 $(i, d), (j, d)$ Label

12. Suppose an upcoming election for UVA student body president is between two candidates. In a survey of 30 students, 18 voiced support for candidate A. Use the Desmos calculator (link) to fit a probability model with Maximum Likelihood for the election, specifically the probability that candidate A is the prefered by the student body. Report the 1/20 support interval. (Provide a screen shot of the calculator with your solution.)



1 DATA

2 $n = 30$ (0 – 30)

3 $h = 18$ (0 – n)

4 LIKELIHOOD

5 $d(p) = \{0 \le p \le 1 : \mathrm{nCr}(n,h) \cdot p^h (1-p)^n$

6 $p_1 = \frac{h}{n}$ $p_1 = 0.6$

7 $d_1(p) = \frac{d(p)}{d\left(\frac{h}{n}\right)}$

8 $(p_1, d_1(p_1))$

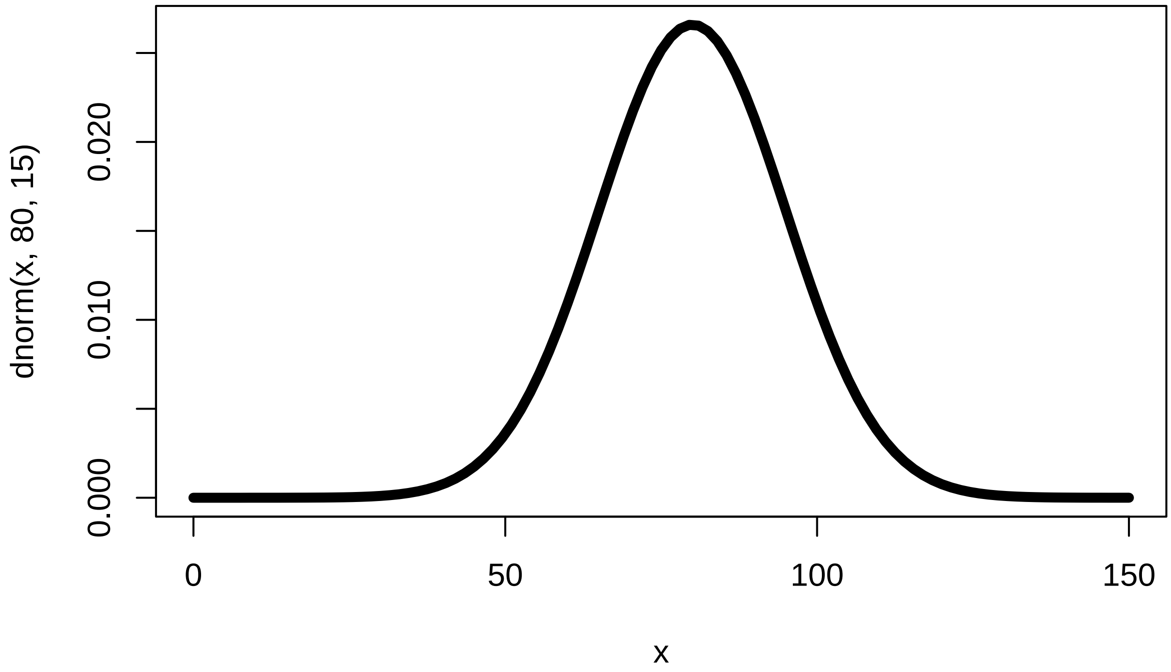
☑ Label: LIKELIHOOD

9 SUPPORT INTERVAL

0.05

13. Suppose diastolic blood pressure (DBP) follows a normal distribution with mean 80 mmHg and SD 15 mmHg. What is the probability that a randomly sampled person's DBP exceeds 104 mmHg?

```
curve(dnorm(x,80,15),0,150,lwd=5)
```



```
p<-1-pnorm(104,80,15)
p
```

```
## [1] 0.05479929
```

14. Suppose a laptop manufacturer sourced batteries from two different vendors. In testing the batteries, the manufacturer collected the following data on time to battery depletion.

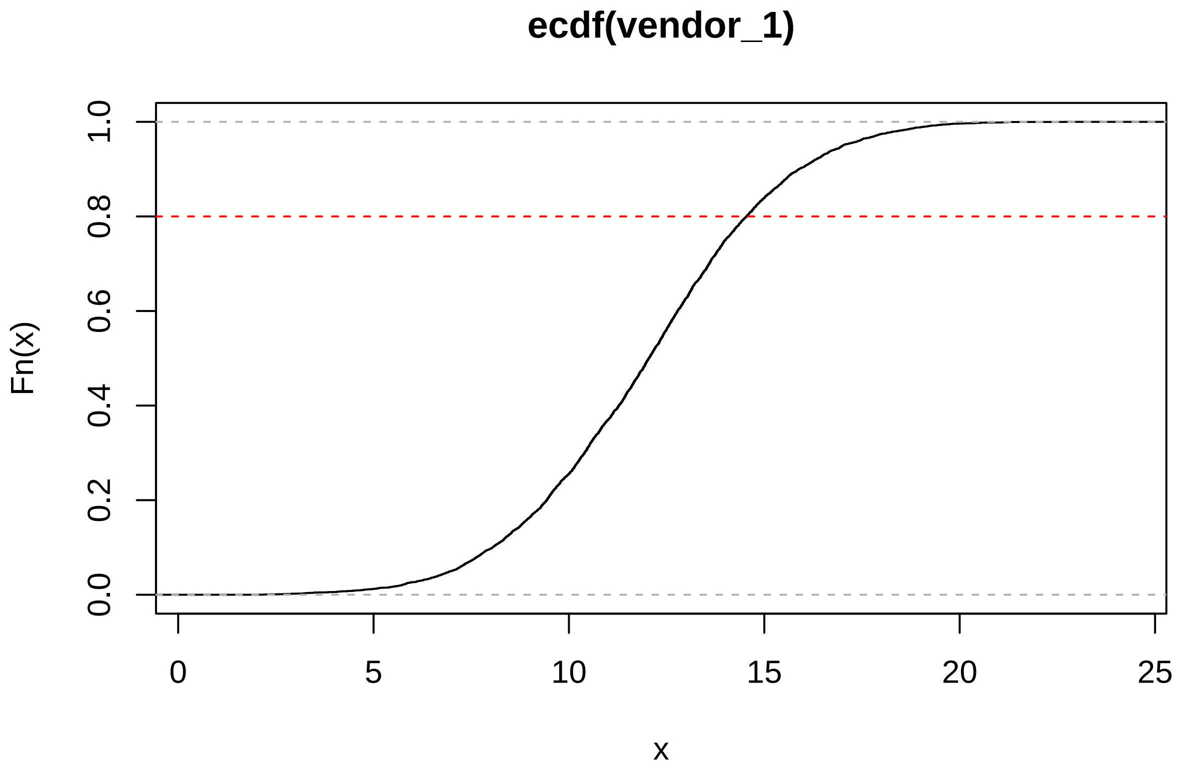
```
d1 <- readRDS(url("https://tgstewart.cloud/battery-data.RDS"))

head(d1)
```

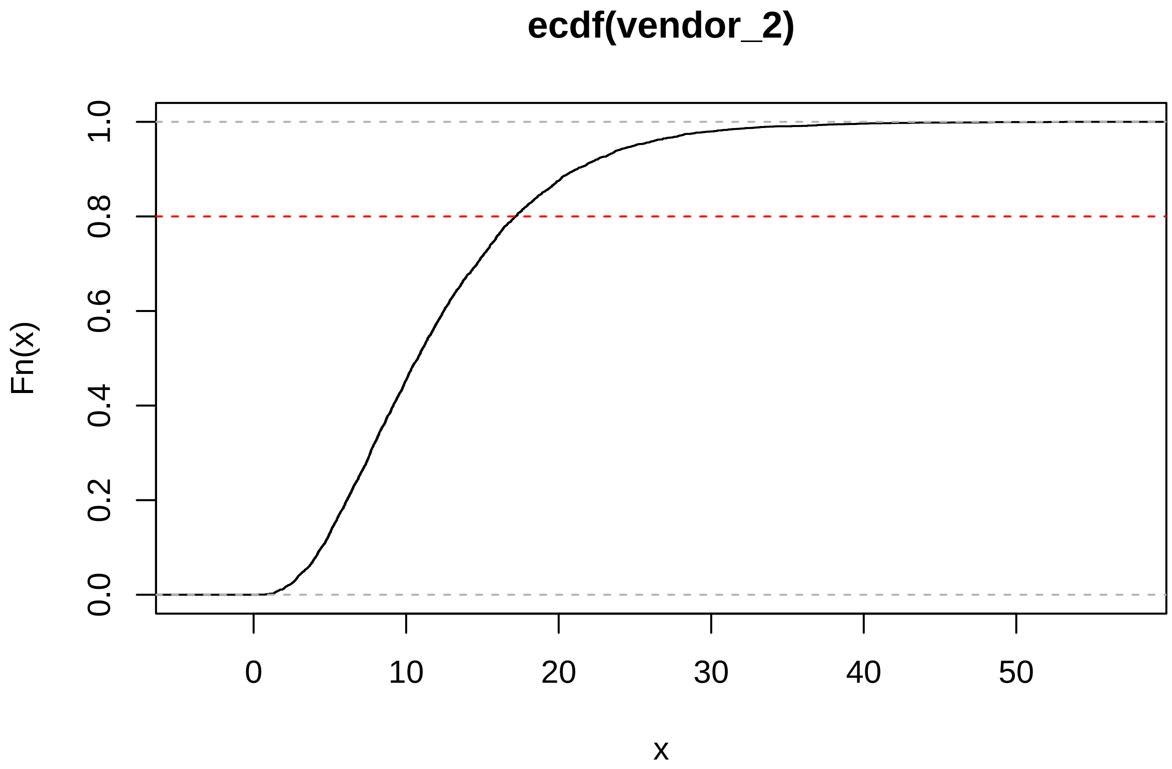
```
##   source      time
## 1      0 12.957865
## 2      0 12.972893
## 3      0 12.025714
## 4      1 18.317820
## 5      0 14.463882
## 6      1  8.110458
```

```
vendor_1 <- d1$time[d1$source == 0]
vendor_2 <- d1$time[d1$source == 1]

ecdf(vendor_1) |> plot()
abline(h=0.8, col="red",lty=2)
```



```
ecdf(vendor_2) |> plot()
abline(h=0.8, col="red",lty=2)
```



Using the data, generate a plot of the empirical CDF for time to battery depletion for each vendor. (Generate both eCDFs on the same plot, if possible.)

15. Based on the data, what is the 80th percentile for battery life (time to battery depletion) for each vendor?

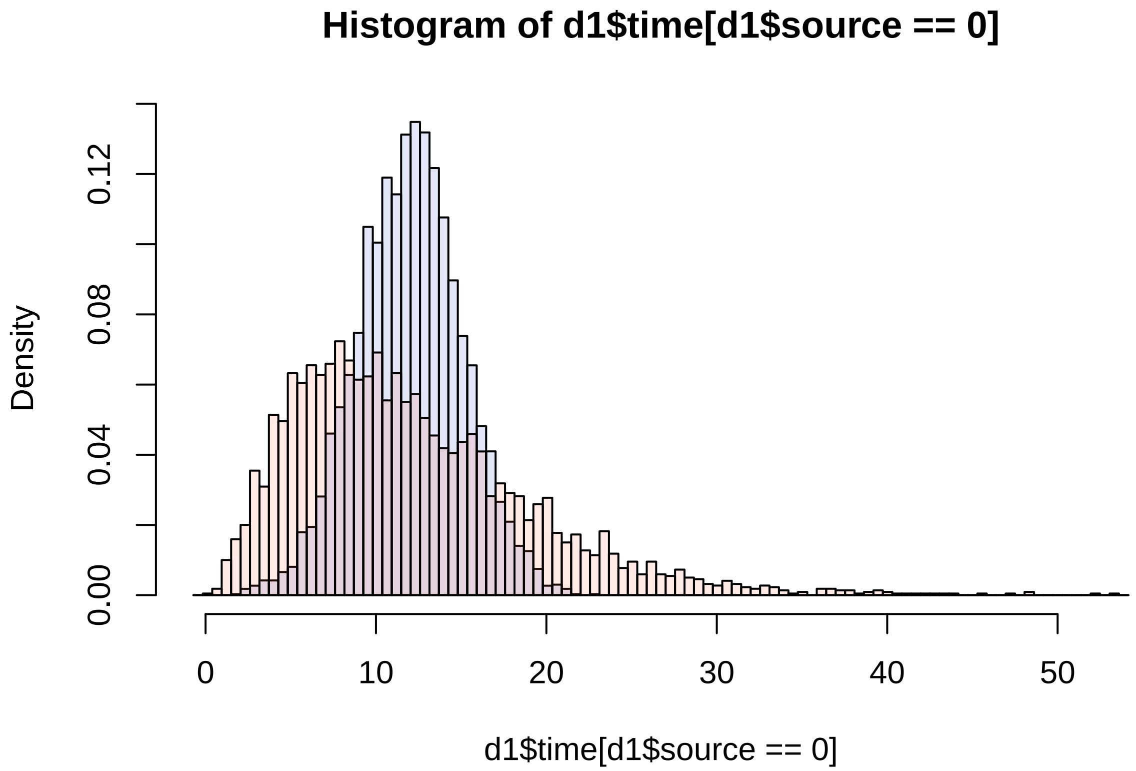
Shown in Plots above.

16. Using the data, generate a histogram for time to battery depletion for each vendor. (Generate both histograms on the same plot, if possible.)

```r
# Hints for plot
b1 <- d1$time %>% range %>% `+`(c(-1,1))
b2 <- seq(b1[1], b1[2], length=100)

# Source 1
hist(d1$time[d1$source == 0], breaks = b2, freq = FALSE, col = "#1338BE20", xlim = b1)

# Source 2
hist(d1$time[d1$source == 1], breaks = b2, add=TRUE, col = "#FF573320", freq = FALSE)
```



17. The function `rbatlife` was created to mimic the distribution of battery life from the previous problem. It will generate `N` draws from the distribution. Using `rbatlife`, what is the mean battery life for each vendor?

```
rbatlife <- function(N){
    g <- rbinom(N,1,.4)
    o <- rgamma(N,3,scale=4)*g + rnorm(N,12,3)*(1-g)
    data.frame(source = g, time = o)
}

d<-rbatlife(1000)
m1<-mean(d$time[d$source == 0])
m1
```

```
## [1] 11.92158
```

```
m2<-mean(d$time[d$source == 1])
m2
```

```
## [1] 11.84148
```

18. The following code will load the first 500 rows of the NHANES data, a large national survey about nutrition.

```r
library("Hmisc")
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
suppressPackageStartupMessages(require(dplyr))
Hmisc::getHdata(nhgh)
d1 <- nhgh[1:500,]
head(d1)
```

```
##     seqn    sex      age                 re         income tx dx    wt    ht
## 1  51624   male 34.16667 Non-Hispanic White [25000,35000)  0  0  87.4 164.7
## 3  51626   male 16.83333 Non-Hispanic Black [45000,55000)  0  0  72.3 181.3
## 5  51628 female 60.16667 Non-Hispanic Black [10000,15000)  1  1 116.8 166.0
## 6  51629   male 26.08333   Mexican American [25000,35000)  0  0  97.6 173.0
## 7  51630 female 49.66667 Non-Hispanic White [35000,45000)  0  0  86.7 168.4
## 10 51633   male 80.00000 Non-Hispanic White [15000,20000)  0  1  79.1 174.3
##      bmi  leg arml armc waist  tri  sub  gh albumin bun  SCr
## 1  32.22 41.5 40.0 36.4 100.4 16.4 24.9 5.2     4.8   6 0.94
## 3  22.00 42.0 39.5 26.6  74.7 10.2 10.5 5.7     4.6   9 0.89
## 5  42.39 35.3 39.0 42.2 118.2 29.6 35.6 6.0     3.9  10 1.11
## 6  32.61 41.7 38.7 37.0 103.7 19.0 23.2 5.1     4.2   8 0.80
## 7  30.57 37.5 36.1 33.3 107.8 30.3 28.0 5.3     4.3  13 0.79
## 10 26.04 42.8 40.0 30.2  91.1  8.6 15.2 5.4     4.3  16 0.83
```

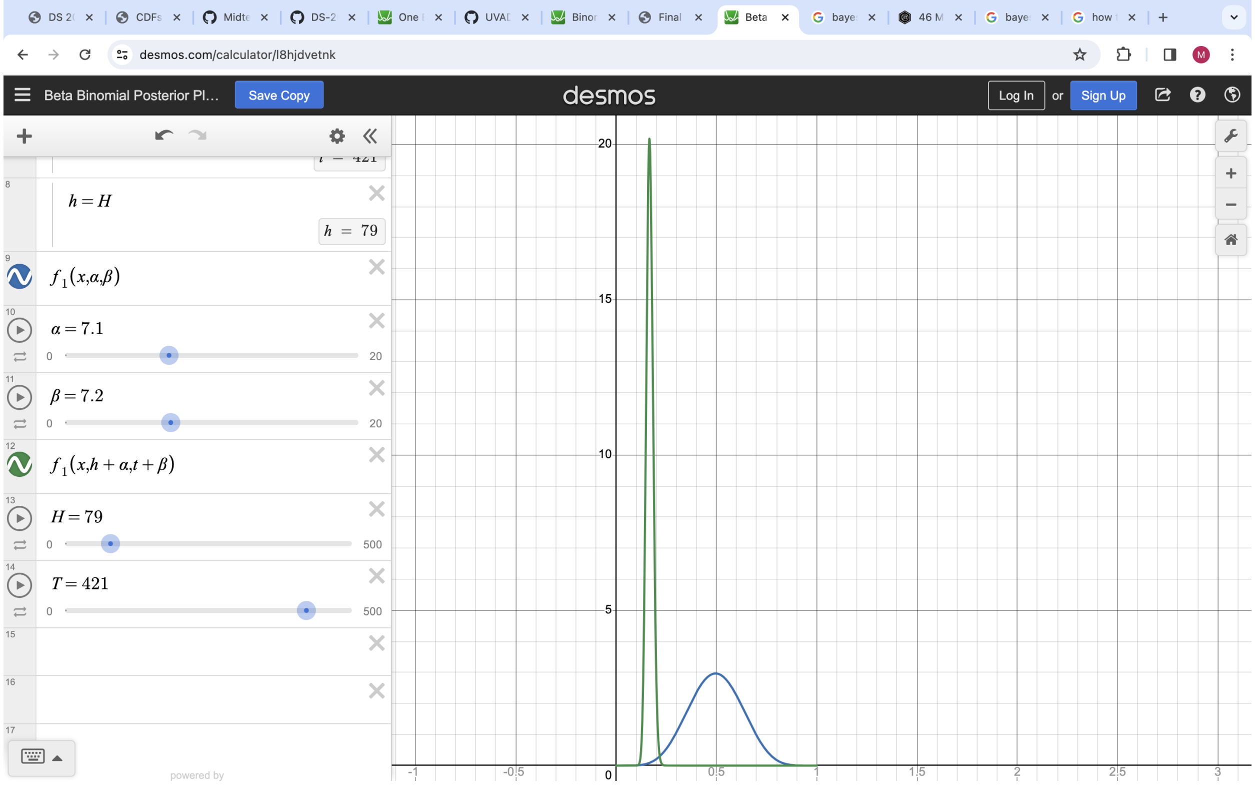
```r
positive_diabetes<-sum(d1$dx==1)
positive_diabetes
```

```
## [1] 79
```

```r
nrow(d1)
```

```
## [1] 500
```

Estimate the prevalence of diabetes (dx) for all respondents using Bayesian updating with a binomail likelihood and beta prior. Use the following Desmos calculator (link). Change α and β to control the prior. Use H (heads) and T (tails) to plug in the data. Take a screenshot of the posterior distribution and the prior.



8 $h = H$

$h = 79$

9 $f_1(x,\alpha,\beta)$

10 $\alpha = 7.1$ 0 20

11 $\beta = 7.2$ 0 20

12 $f_1(x,h+\alpha,t+\beta)$

13 $H = 79$ 0 500

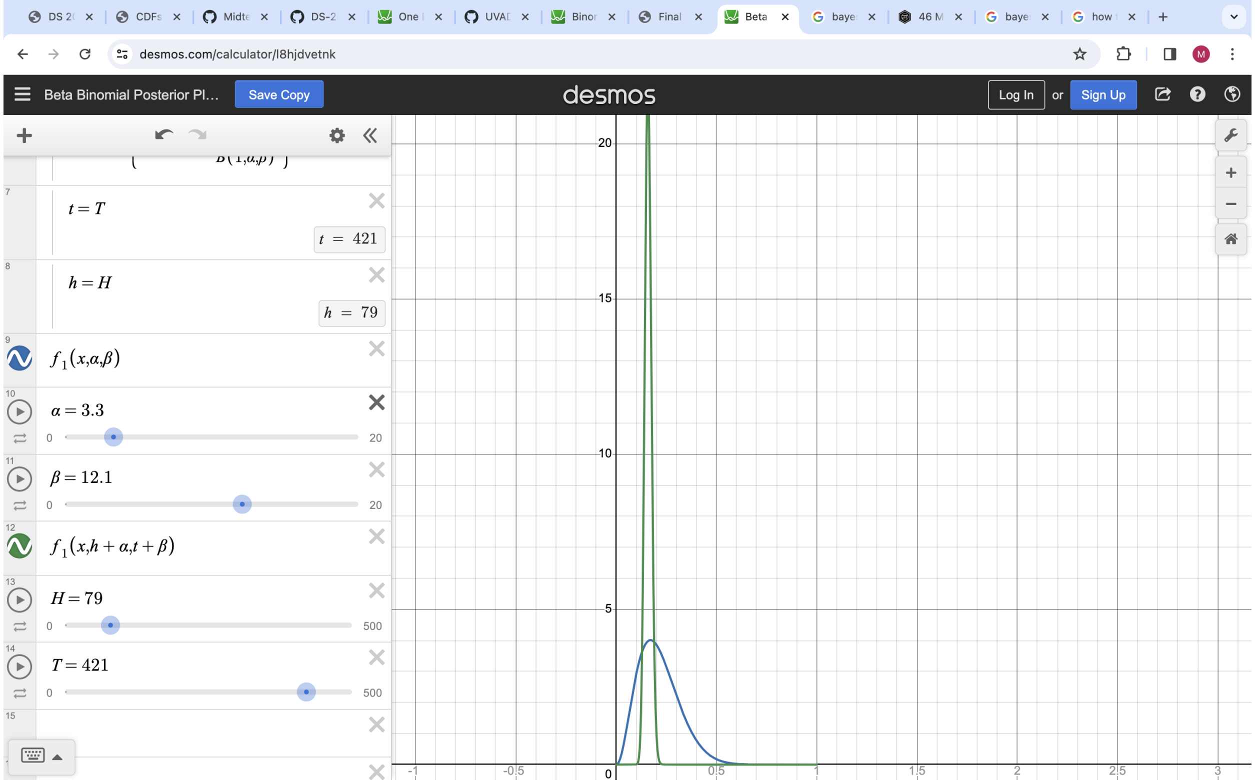
14 $T = 421$ 0 500

15

16

17

19. Reestimate the prevalence of diabetes (dx) with a more informative prior. Take a screenshot of the resulting posterior distribution with the new prior. Explain why the new prior is more informative.



$B(1,\alpha,\beta)$

7 $t = T$

$t = 421$

8 $h = H$

$h = 79$

9 $f_1(x,\alpha,\beta)$

10 $\alpha = 3.3$ 0 20

11 $\beta = 12.1$ 0 20

12 $f_1(x,h+\alpha,t+\beta)$

13 $H = 79$ 0 500

14 $T = 421$ 0 500

15

Shown in pdf/screenshot. The new prior is more informative because it is closer to the posterior and the data that we have.

20. Suppose the posterior distribution of the mean birthweight of infants whose mothers did not smoke was a normal distribution with mean $= 3100$ and standard deviation $= \sqrt{10}$. The symmetric density credible interval is calculated by identifying the 0.025 and 0.975 quantiles from the posterior. Calculate the interval.

```
lq <- qnorm(0.025, mean = 3100, sd = sqrt(10))
uq<- qnorm(0.975, mean = 3100, sd = sqrt(10))
print("The credible interval is between")
```

```
## [1] "The credible interval is between"
```

```
lq
```

```
## [1] 3093.802
```

```
print("and")
```

```
## [1] "and"
```

```
uq
```

```
## [1] 3106.198
```

EC1. A creative writting essay was submitted without the author's name. After informing the class of the unnamed essay, two students claimed to be the author. Student A is known to use exclamation points in 10% of sentences. Student B is known to use exlamation points in 5% of sentences. A review of the unnamed essay revealed that 5 of 60 sentences used an exlamation point. With this information, calculate

$$P(\text{Student A authored the essay } |5 \text{ of 60 sentences used an exlamation point}).$$

Using Bayes rule: (1/10*1/2)/1=1/20

EC2. Continuing the previous question, create a plot with number of exclamation points on the x-axis and the probability that student A authored the essay on the y-axis. exclamations<-1:100

EC3. (Continuing problems 14 - 17) If a battery lasted for 10 or fewer hours, what is the probability it was from source 1?

P(battery lasting 10 or fewer|source 1)