Final Project

Ingrid J. Lu 260773949; Grace Ma 260761707 ; Xinbei Wan 260777034; Wanqi Wang 261023391

compiled on December 15, 2021

Context

Your sister and her partner are expecting a child soon, and she just went to her obstetrician for her routine check. Because she in her second trimester, her obstetrician asks her to do a fetal cardiogram. The results will not get back to her until a week later. Your sister is a bit of a hypochondriac, so she is afraid that there is something wrong with her child.

So, her partner turns to you, a Masters student who is studying inferential statistics for health, hoping to find some results to calm your sister down.

Luckily, you have identified a publicly available dataset that includes thousands of fetal cardiogram results, and the classification of these babies' health status. You need to convince your sister that she and her child will be safe.

Note: the objective of this exercise is to consolidate all the important concepts covered in EPIB607. When answer each question, be sure to include any units and assumptions and define all parameters, when appropriate. The following questions are based on the publicly available dataset "Fetal Classification", please find all attribute information of the data from the hyperlink.

Use the following code to set-up your dataframe:

```r
# Set-up
df_fh <- readr::read_csv(here::here("fetal_health.csv")) %>%
  select(!starts_with("histogram"))
```

Question 1 Data Visualization and Summary Statistics

a)

Is this data ready for you to work with? If no, transform it into a ready to use form, if yes, explain why.

Solution: Yes, this data is ready to use because it is in a tidy format. Each column is a variable and each row is one observation. Each cell contains 1 value of the variable.

b)

Your sister is concerned that her baby does not move as much, which could be a sign of an unhealthy pregnancy.

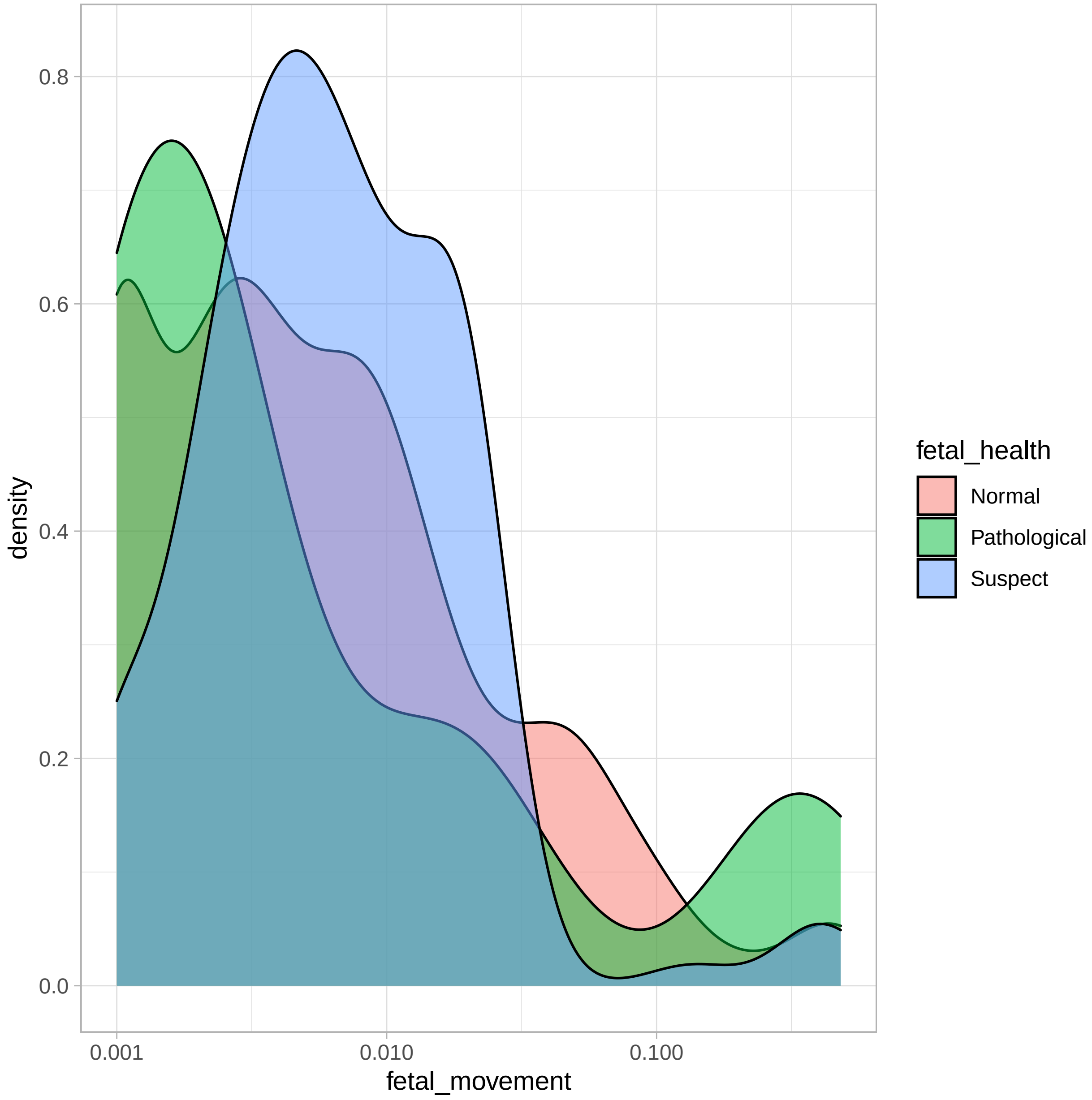
Use a density graph, show her the distribution of fetal movements according to different fetal health classifications. You do not have to interpret this graph.

Solution:

```
#----Question 1----------

## 1.b
df_fh$fetal_health <- ifelse(df_fh$fetal_health == 1,
                             "Normal",
                             ifelse(df_fh$fetal_health == 2,
                                    "Suspect",
                                    "Pathological"))

df_fh %>%
  ggplot(aes(x = fetal_movement, fill = fetal_health))+
  geom_density(alpha = 0.5) +
  scale_x_log10()
```



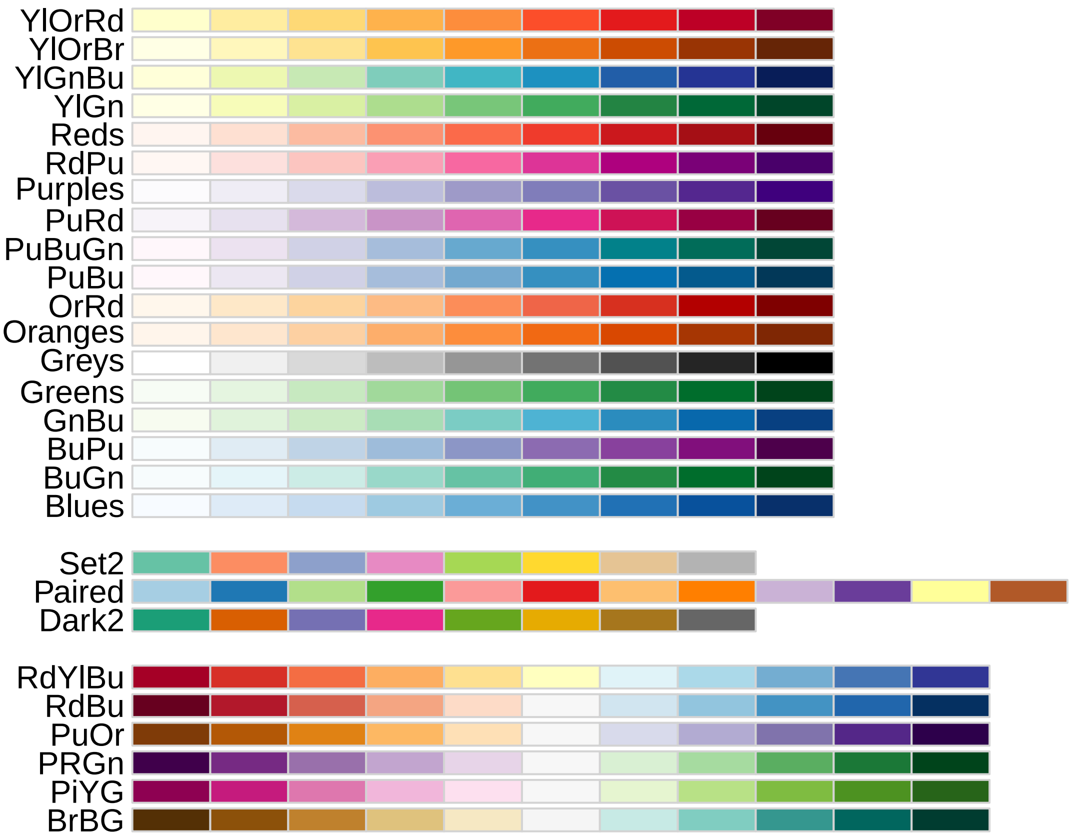
c)

Your sister's partner tells you that he is colourblind, please fix the graph produced above to provide him with a colourblind-friendly graph.

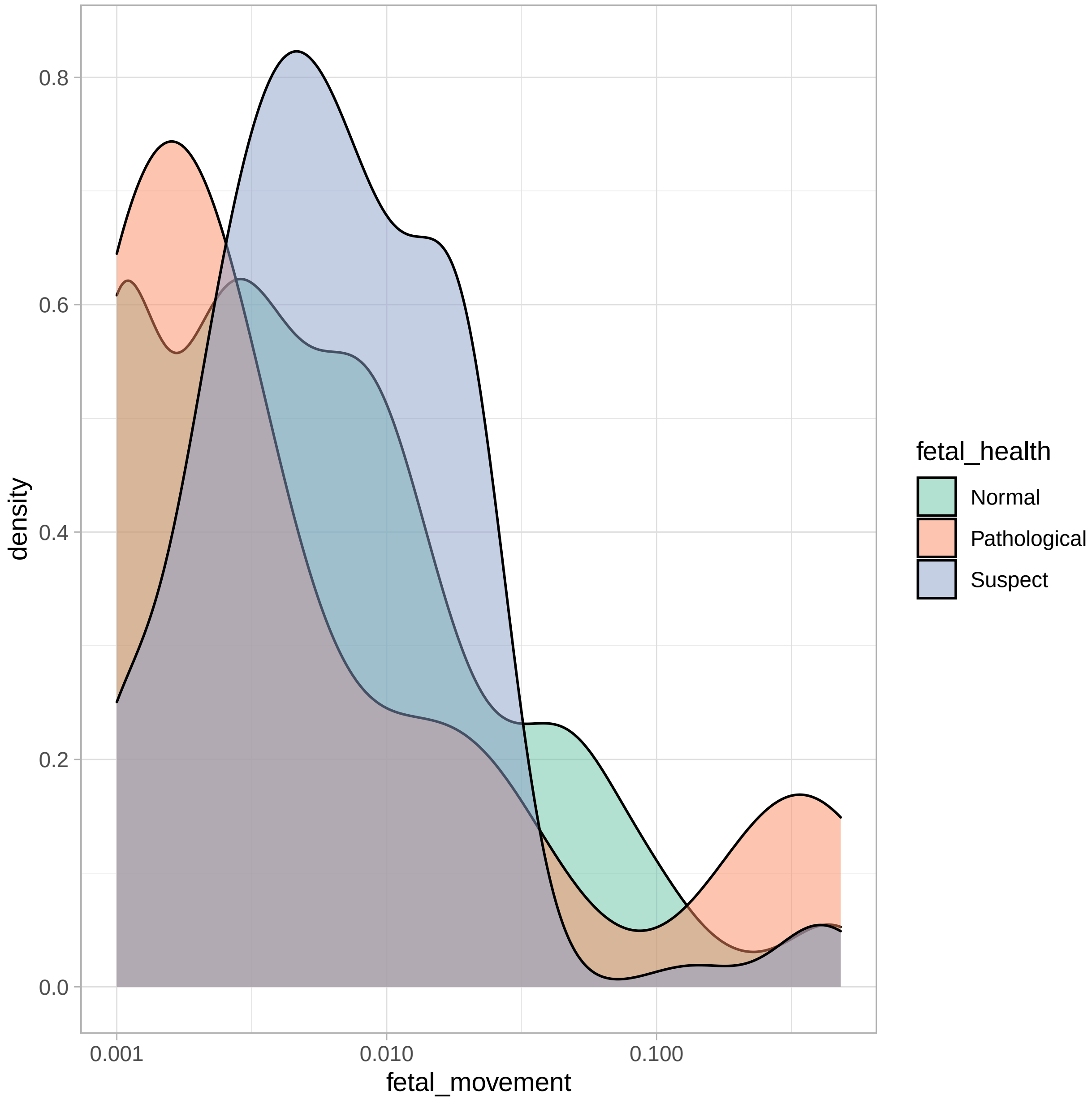
Solution: We can use the "RColorBrewer" package and look for a colourblind-friendly palette. Then apply the package to our ggplot.

```
# 1.c

display.brewer.all(colorblindFriendly = TRUE)
```



```
df_fh %>%
  ggplot(aes(x = fetal_movement, fill = fetal_health))+
  geom_density(alpha = 0.5) +
  scale_x_log10() +
  scale_fill_brewer(palette="Set2")
```



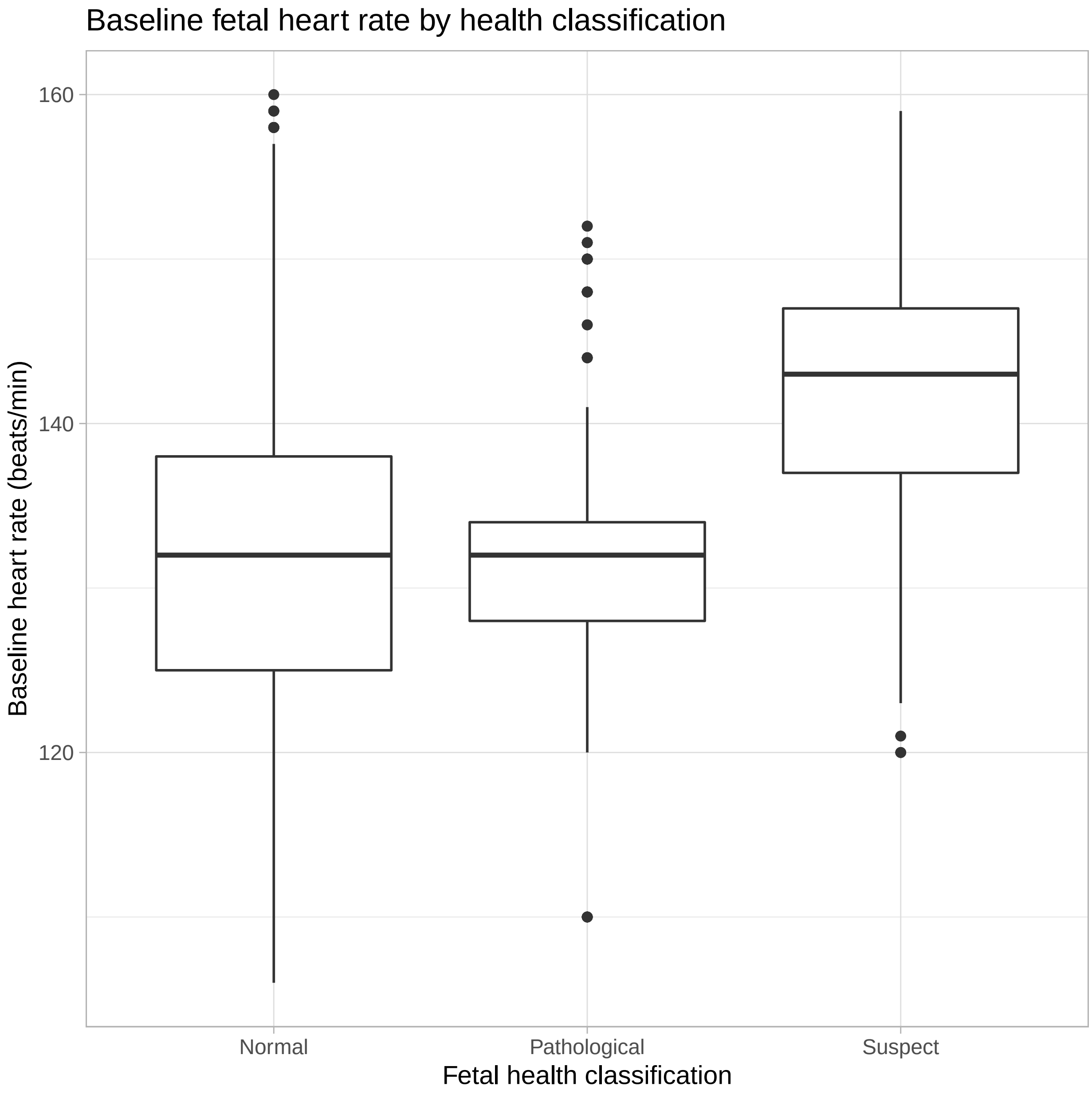
d)

Your sister also said she thinks her baby's heart rate is faster than normal, which could be a sign of unhealthy pregnancy.

Looking at the 3 different classification of fetal health status and each fetus' baseline heart rate, provide an appropriate graphic summarizing the distribution of each of baseline heart rates for each class. Be sure to provide the correct title and label for the plot.

Solution:

```r
# 1.d
df_fh %>%
  group_by(fetal_health) %>%
  ggplot(aes(x = fetal_health, y = `baseline value`)) +
  geom_boxplot() +
  labs(title = "Baseline fetal heart rate by health classification",
       x = "Fetal health classification",
       y = "Baseline heart rate (beats/min)")
```



e)

Comment on the boxplot, what are the characteristics of each category?

Solution:

The objective was to show the distribution of baseline heart rates in each health category. The graph above shows that the median fetal baseline heart rate was similar between Normal and Pathological fetuses; both are lower than the suspect group. The range of baseline heart rate was wider for the normal group, and the range for the pathological group is the narrowest. The baseline heart rates were rightly skewed for Normal and Pathological groups. There is considerable overlap between the boxplots of normal category and pathological category, it does not appear that the baseline heart rate significantly differs between these two groups.

Question 2

a)

Your sister thinks that more uterine contraction means that her fetus is unhealthy and she is more likely to give a pre-term birth.

Is her statement true? What is the mean uterine contraction for each class?

Solution:

```
#----Question 2--------------
# 2.a
df_fh_n <- df_fh %>% filter(fetal_health == "Normal")
mean(df_fh_n$uterine_contractions)
```

```
## [1] 0.004780665
```

```
df_fh_s <- df_fh %>% filter(fetal_health == "Suspect")
mean(df_fh_s$uterine_contractions)
```

```
## [1] 0.002389831
```

```
df_fh_p <- df_fh %>% filter(fetal_health == "Pathological")
mean(df_fh_p$uterine_contractions)
```

```
## [1] 0.003784091
```

Looking at the mean, this statement is not true. The Normal category has the highest number of mean uterine contractions. And the mean uterine contraction in the Pathological category is lower than the mean in the Normal category. The Suspect category has the lowest mean uterine contraction out of the three classes.

Nevertheless, further testing is required to confirm this observation.

b)

Since we have a small sample size for those who are suspected to be pathological and those who are determined to be pathological, what is one method that we can use to artificially create a pseudo-population and calculate the median of these two groups? State your assumptions.

Solution: Bootstrapping, the assumptions are simple random sampling and the samples are representative of the population.

```
# 2.b
B <- 1000
set.seed(3949)

R_s <- replicate(B, { # first argument, # of replicates
  df_fh_s %>%
    dplyr::slice_sample(n = nrow(df_fh_s), replace = T) %>%
    summarise(mean = mean(df_fh_s$uterine_contractions)) %>%
    pull(mean)
})

mean(R_s)
```

```
## [1] 0.002389831
```

```
R_p <- replicate(B, { # first argument, # of replicates
  df_fh_p %>%
    dplyr::slice_sample(n = nrow(df_fh_p), replace = T) %>%
    summarise(mean = mean(df_fh_p$uterine_contractions)) %>%
    pull(mean)
})

mean(R_p)
```

```
## [1] 0.003784091
```

c)

What is one weakness of using the bootstrapping method?

Solution: The bootstrapping method assumes that the sample is representative of the whole population. However, when we have a small sample size, that is not always the case. There could be sampling errors which can skew the distribution of the sample. And the bootstrapping method cannot correct this error as it can only sample within this small set of data. As the saying goes, "garbage in, garbage out". If the sample is not representative of the population, the results obtained from bootstrap will not be accurate and will result in an incorrect inference of the population parameter.

Question 3. p-value, power

For the purpose of this question only, we treat the 2126 individuals as **the entire target population of newborns**.

a)

Calculate the mean and standard deviation of the baseline fetal heart rate.

```
#----Question 3--------------
# 3.a
mean_hr <- mean(df_fh$`baseline value`)
mean_hr
```

```
## [1] 133.3039
```

```
sd_hr <- sqrt(var(df_fh$`baseline value`)*(length(df_fh$`baseline value`)-1)/length(df_fh$`baseline valu
sd_hr
```

```
## [1] 9.83853
```

b)

Your sister claimed that, she read on a magazine, that the baseline fetal heart rates of fetuses with “suspect” health status are above average. Take a simple random sample of 10 fetuses with “suspect” health status, and measure their heart rate to obtain a sample mean of 141.68. Heart rates are scaled to be normally distributed. Does the sample provide evidence to reject null hypothesis? State your null and alternative hypothesis.

```
# 3.b
pnorm(q = 141.68, mean = 133.30, sd = 9.84/sqrt(10), lower.tail = FALSE)
```

```
## [1] 0.003539786
```

$H_0 : \mu = 133.30$, $H_A : \mu > 133.30$

The p-value of one sided test is 0.0035. This sample provides evidence against the null hypothesis. The p-value tells us the probability of observing the sample size mean of 141.68 under the null hypothesis distribution is very unlikely.

c)

So your sister asks you now, what is the probability that you can detect the baseline fetal heart rates of fetuses with “suspect” health status are at least 8.38 heart beats higher than average, using a one-sided test and sample size 10 and a 0.05 level test?

$H_0 : \mu = 133.30$, $H_A : \mu > 141.30$

```
# 3.c
# cutoff to reject the null
cutoff <- qnorm(p = 0.95, mean = 133.30, sd = 9.84/sqrt(10))
cutoff
```

```
## [1] 138.4183
```

```
# probability of observing this cutoff or greater under the alternative
pnorm(q = cutoff, mean = 141.68, sd = 9.84/sqrt(10), lower.tail = FALSE)
```

```
## [1] 0.8527324
```

d)

A sample size of 10 fetuses with "suspect" health status will have at least 85% power to detect a difference of 8.38 heart beats. Use a simulation based approach to reproduce the sample size calculation for the baseline fetal heart rates of fetuses with "suspect" health status and average.

```
# 3.d
set.seed(490)

power_distribution <- replicate(n = 1000, expr={
  sample.size <- 10

  suspect <- rnorm(sample.size, mean = 141.68, sd = 9.83)
  SEM <- sd(suspect)/sqrt(sample.size)

  pnorm(q = mean(suspect), mean = 133.30, sd =SEM, lower.tail = FALSE) < 0.05

} )

prop.table(table(power_distribution))
```

```
## power_distribution
## FALSE  TRUE
## 0.146 0.854
```

The percentage of samples that results in a p-value less than 0.05 is 85.4%, which shows the study is powered at 85% to detect the difference.

Question 4.1

Suppose this data represents the **population of newborns in one hospital** and you take a simple random sample of 100 babies from the population.

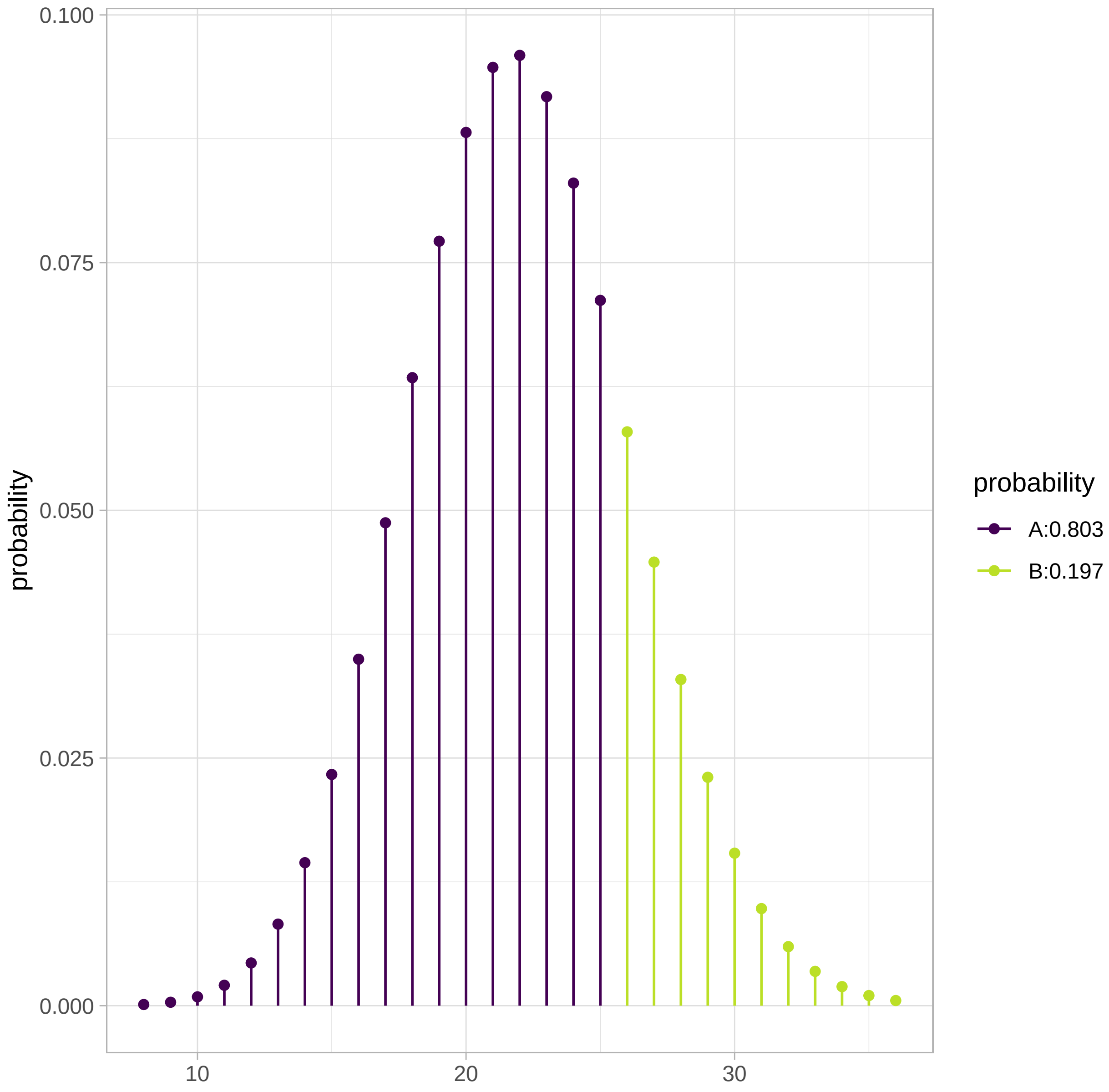
a)

What is the probability that your sample contains more than 25 babies has abnormal health status (1 = health, 2 = suspect, 3 = Pathological)?

```
## ---- Question4.1-----------------------------------------------------------
# 4.1.a
df_fh %>%
  select(fetal_health)%>%
  filter(fetal_health != 1)%>%
  nrow()
```

```
## [1] 2126
```

```
#there are 471 abnormal in this population
#the probability of having abnormal is 471/2126 = 0.22
1 - mosaic::xpbinom(q = 25, size = 100, prob = 0.22)
```



```
## [1] 0.1972269
```

The probability of having more than 30 abnormal is 0.197.

b)

Turns out that your sample actually contains 20 babies with abnormal health status. What is the 95% confidence interval of this proportion? Can you use a normal approximation for this sample? Why or why not?

```
# 4.1.a
mosaic::binom.test(x = 20, n = 100, ci.method = "Clopper-Pearson")
```

```
##
##
##
## data:  20 out of 100
## number of successes = 20, number of trials = 100, p-value = 1.116e-09
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1266556 0.2918427
## sample estimates:
## probability of success
##                    0.2
```

```
mosaic::binom.test(x = 20, n = 100, ci.method = "Wald")
```

```
##
##  Exact binomial test (Wald CI)
##
## data:  20 out of 100
## number of successes = 20, number of trials = 100, p-value = 1.116e-09
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1216014 0.2783986
## sample estimates:
## probability of success
##                    0.2
```

The 95% CI using Clopper-Pearson method: [0.127,0.292] The 95% CI using exact method(ie: normal approximation): [0.122,0.278] The two methods give similar 95% CIs. Normal approximation can be used here since the sample size is large enough to generate a binomial distribution approximating the normal distribution and for CLT to kick in.

c)

Another sample taken have the same proportion of event but the sample size is now only 10 and the count of abnormal is 2. Calculate the 95% CI using this sample and compare it with the one you have in b). Describe their difference and the reason why.

```
# 4.1.c
mosaic::binom.test(x = 2, n = 10, ci.method = "Clopper-Pearson")
```

```
##
##
##
## data:  2 out of 10
## number of successes = 2, number of trials = 10, p-value = 0.1094
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.02521073 0.55609546
## sample estimates:
## probability of success
##                    0.2
```

The 95% CI for this sample is: [0.0252, 0.556]. 95% CI in c) with sample size = 10 is wider than the 95% CI in b) with sample size = 100. The sample size is different so the standard error is different. The 95% CI is calculated using the formula:

$$\bar{y} - 1.96(\frac{\sigma}{\sqrt{n}}), \bar{y} + 1.96(\frac{\sigma}{\sqrt{n}})$$

If the sample size n is larger, the standard error(sigma/sqrt n) is smaller. The value 1.96*standard error is also smaller, and results in a narrower confident interval.

Question 4.2

You continue to work with the simple random sample. This time you take 100 babies from the a different hospital as the sample from the population.

a)

Your sample contains 30 babies with abnormal health status. What is the rate and 95% CI of health abnormality? Interpret your result.

```
# 4.2.a
stats::poisson.test(x = 30, T = 100)
```

```
##
##  Exact Poisson test
##
## data:  30 time base: 100
## number of events = 30, time base = 100, p-value = 4.154e-16
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##  0.2024087 0.4282687
## sample estimates:
## event rate
##        0.3
```

Rate: 0.3 95% CI: [0.20,0.43] There is a 95% chance that the confidence interval [0.20,0.43] captures the true rate.

b)

According to the data of the population with 2126 babies, 471 babies have abnormal health status. The expectation of the baby having abnormal health status of the original hospital is only 0.22. Does your sample suggest that the babies coming from two hospitals have significantly different rate of abnormal health? Calculate the 95% CI for the rate ratio both by hand and using a one-step canned function.

```
# 4.1.b
stats::poisson.test(x = 30, T = 100)
```

```
##
##  Exact Poisson test
##
## data:  30 time base: 100
## number of events = 30, time base = 100, p-value = 4.154e-16
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##  0.2024087 0.4282687
## sample estimates:
## event rate
##        0.3
```

```
0.2024087/0.22
```

```
## [1] 0.9200395
```

```
0.4282687/0.22
```

```
## [1] 1.946676
```

```
stats::poisson.test(x = 30, T = 471/2126*100)
```

```
##
##  Exact Poisson test
##
## data:  30 time base: 471/2126 * 100
## number of events = 30, time base = 22.154, p-value = 0.1092
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##  0.9136327 1.9331192
## sample estimates:
## event rate
##    1.35414
```

95% CI by hand: [0.92.1.95] 95% CI by canned function: [0.91,0.933] According to the Poisson test and by hand calculation, two 95% CI for the rate ratio both contains the null value 1. The 95% CI and the p-value suggests that the difference between two hospitals is not statistically significant.

Question 5

a)

Your sister's partner said, their obstetrician told them that the Fetal Heart Rate could be a reflection of a lower value of short term variability. Can you conduct a linear regression to test it? Is it significant?

Solution:

```
#---Question 5----------------
# 5.a
reg1 <- lm(mean_value_of_long_term_variability ~ `baseline value`, data = df_fh)
summary(reg1)
```

```
##
## Call:
## lm(formula = mean_value_of_long_term_variability ~ ‘baseline value‘,
##     data = df_fh)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.615 -3.526 -0.756  2.647 42.525
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.63425    1.65791   6.414 1.74e-10 ***
## ‘baseline value‘ -0.01835    0.01240  -1.480    0.139
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.627 on 2124 degrees of freedom
## Multiple R-squared:  0.00103,    Adjusted R-squared:  0.0005595
## F-statistic:  2.19 on 1 and 2124 DF,  p-value: 0.1391
```

According to the linear regression, there is no significant association between baseline heart rate and value of short term variability.

b)

Now you are getting curious with this dataset, and you want to use a logistic regression and predict the classification of fetal health diagnosis (Pathological vs non-pathological) using fetal movement as a binary determinant (normal vs. abnormal rate of movement). State the regression equation, Describe how will you process these three variables to fit a logistic regression.

Solution:

Regression equation:

$$\frac{\pi}{1-\pi} = \frac{\pi_0}{1-\pi_0} * \theta^{mov}$$

$$log(\frac{\pi}{1-\pi}) = log(\frac{\pi_0}{1-\pi_0}) + mov * \theta \begin{cases} mov = 1 & \text{if movement is abnormal} \\ mov = 0 & \text{if movement is normal} \end{cases}$$

π is the probability of receiving a pathological diagnosis

π_0 is the probability of receiving a pathological diagnosis, when fetal movement is lower than 0.01

θ_1 is the odds ratio of receiving a pathological diagnosis, when fetal movement is equal to or higher than 0.01, and it is the **parameter of interest**.

mov is the risk parameter.

1. In a logistic regression, the outcome variable is binary variable. In this example, the fetal health outcome can be classified as Pathological diagnosis, and non-pathological diagnosis (including normal and suspect).
2. We can determine the abnormal threshold for the number of fetal movements per second based on previous literature and hypothesis and categorize this variable into a binary determinant.

c)

Suppose the normal number of fetal movements per second is below 0.01. Fit the logistic regression model and provide and interpret the 95% CI for you parameter of interest.

```
# 5.c
df_fh$f_mov <- ifelse(df_fh$fetal_movement>0.01, 1, 0)
df_fh$f_patho <- ifelse(df_fh$fetal_health=="Pathological", 1 , 0)

reg2 <- glm(f_patho~ f_mov, family = binomial (link = "logit"), data = df_fh)
summary(reg2)
```

```
##
## Call:
## glm(formula = f_patho ~ f_mov, family = binomial(link = "logit"),
##     data = df_fh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4975  -0.4048  -0.4048  -0.4048   2.2550
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.46056    0.08535 -28.829   <2e-16 ***
## f_mov        0.43339    0.22181   1.954   0.0507 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1214.0  on 2125  degrees of freedom
## Residual deviance: 1210.5  on 2124  degrees of freedom
## AIC: 1214.5
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(reg2)[2])
```

```
##    f_mov
## 1.542478
```

```
exp(confint(reg2))
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.07194141 0.1005501
## f_mov       0.98016928 2.3459592
```

θ is 1.542. This means that, the odds of receiving a pathological diagnosis in pregnancies with an abnormal fetal movement is 1.54 times the odds of receiving a pathological diagnosis in pregnancies with an normal fetal movement.

The confidence interval for the parameter of interest θ is 0.98, 2.35. We are 95% confident that the true odds ratio falls within this range.

d)

Can you calculate the intercept by hand? If yes, show your work, if no, explain what other values do you need to perform this calculation.

Solution: Using the equation $Intercept = log(\frac{\pi_0}{1-\pi_0})$, and since π_0 is the probability of receiving a pathological diagnosis, when fetal movement is lower than 0.01, we can first calculate the odds ratio of π_0 and then calculate the log this OR.

```
# 5.d
count <- df_fh %>%
  count(f_mov, f_patho)
prob <- count[2,3] / (count[2,3] + count[1,3])
odds <- prob/ (1-prob)
log(odds)
```

```
##           n
## 1 -2.460564
```

Question 6

a)

You saw on another paper, that the pathological diagnosis is dependent on fetal movements and abnormal long term variability, provide a regression equation for this model. Remember to define all parameters.

For the purpose of this question, code the presence and absence of abnormal long term variability as 0,1, using the percentage of time with abnormal long term variability. (0% of variability means no abnormal variability, any number higher than 0% suggests there is long term variability)

Solution:

Regression equation:

$$\frac{\pi}{1-\pi} = \frac{\pi_0}{1-\pi_0} * \theta_1^{mov} * \theta_2^{var} \begin{cases} mov = 1 & \text{if movement is equal to or higher than 0.01} \\ mov = 0 & \text{if movement is lower than 0.01} \\ var = 1 & \text{if there is abnormal variability} \\ var = 0 & \text{if there is no abnormal variability} \end{cases}$$

$$log(\frac{\pi}{1-\pi}) = log(\frac{\pi_0}{1-\pi_0}) + mov * log(\theta_1) + var * log(\theta_2)$$

π is the probability of receiving a pathological diagnosis

π_0 is the probability of receiving a pathological diagnosis, when fetal movement is lower than 0.01 and no abnormal long-term variability

θ_1 is the odds ratio of receiving a pathological diagnosis, when fetal movement is equal to or higher than 0.01, adjusting for long-term variability.

θ_2 is the odds ratio of receiving a pathological diagnosis, when there is abnormal long-term variability, adjusting for fetal movement.

mov and *var* are the risk parameters.

b)

Fit the regression equation, and compare the intercept value, is it different than the fitted value in Question 5.c? Why?

Solution:

```r
#----Question 6-----------------
# 6.b
df_fh$ab_var <- ifelse(
  df_fh$percentage_of_time_with_abnormal_long_term_variability == 0, 0, 1)

reg3 <- glm(f_patho ~ f_mov + ab_var, family = binomial (link = "logit"), data = df_fh)
summary(reg3)
```

```
##
## Call:
## glm(formula = f_patho ~ f_mov + ab_var, family = binomial(link = "logit"),
##     data = df_fh)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5081  -0.4182  -0.4182  -0.3866   2.2938
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3926     0.1074 -22.284   <2e-16 ***
## f_mov         0.4106     0.2230   1.841   0.0656 .
## ab_var       -0.1635     0.1632  -1.002   0.3165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1214.0  on 2125  degrees of freedom
## Residual deviance: 1209.5  on 2123  degrees of freedom
## AIC: 1215.5
##
## Number of Fisher Scoring iterations: 5
```

The intercept for this model is -2.3926, which is different from the model above, because it is calculating the log of the odds of receiving a pathological diagnosis, without abnormal long-term variability nor fetal movement. Meanwhile, the intercept in the previous model computes the log of the odds of receiving a pathological diagnosis, only without abnormal fetal movement.

c)

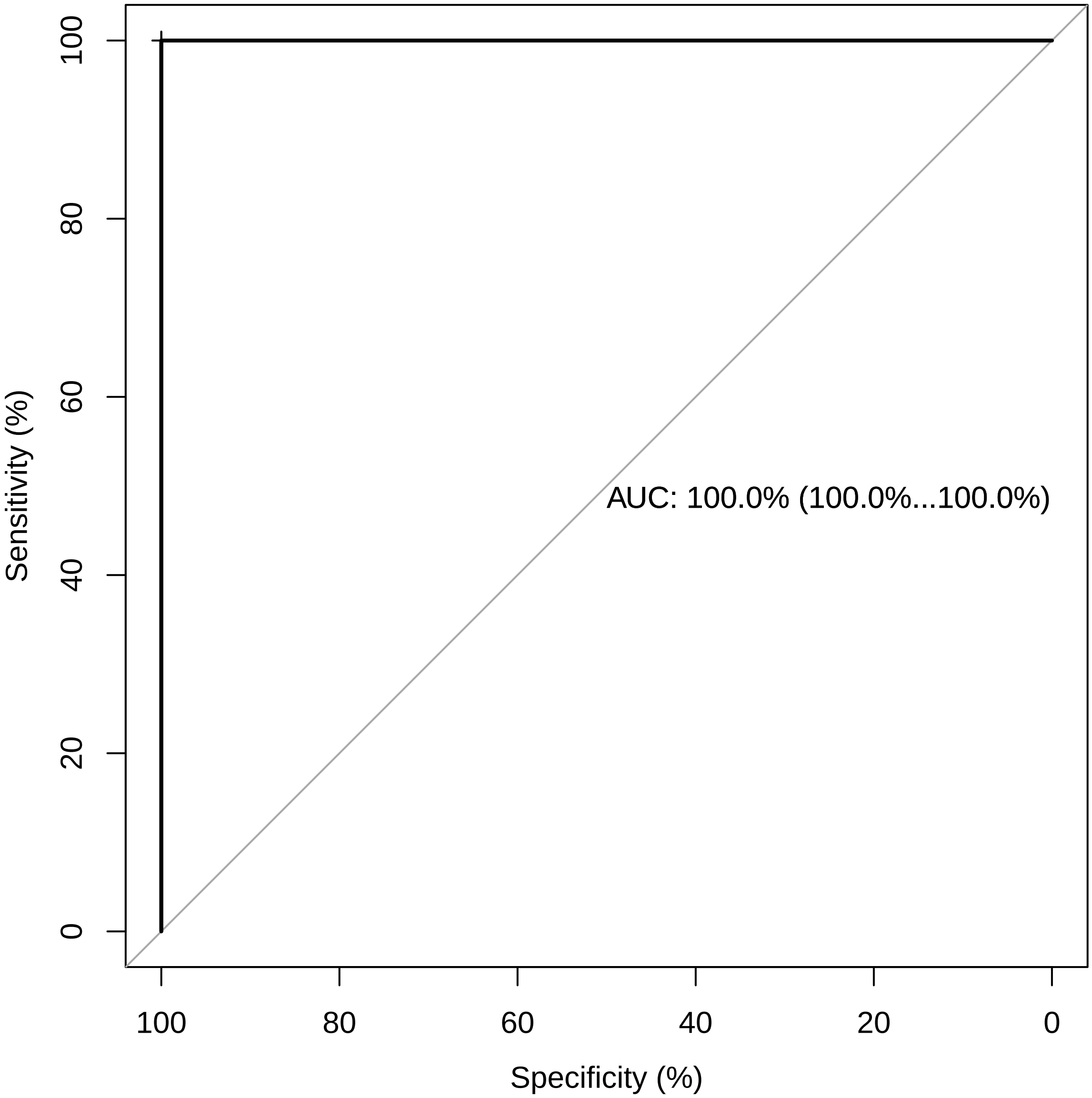
You decided to draw a ROC curve and see how well your model works out. Use the pROC package to draw the curve, can the model produce accurate predictions?

Solution: According to the ROC curve, the model can predict the diagnosis very well. The AUROC is 100%, suggesting that our model can predict the diagnosis with 100% specificity and 100% sensitivity.

```
# 6.c
rocobj1 <- plot.roc(df_fh$f_mov, fitted(reg3),
                    percent=TRUE,
                    ci=TRUE,
                    print.auc=TRUE)

ciobj <- ci.se(rocobj1,
               specificities=seq(0, 100, 5))

plot(ciobj, type="shape", col="#1c61b6AA")
plot(ci(rocobj1, of="thresholds", thresholds="best"))
```


AUC: 100.0% (100.0%...100.0%)
Sensitivity (%)
0
20
40
60
80
100
Specificity (%)
100
80
60
40
20
0