Final Project

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

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:

```
# 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.

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.

0.1 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.

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.

e)

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

1 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?

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.

c)

What is one weakness of using the bootstrapping method?

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.

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.

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?

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.

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)?

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?

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.

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.

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.

Question 5

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.

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.

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.

2 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)

b)

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

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?