Project Draft ETA

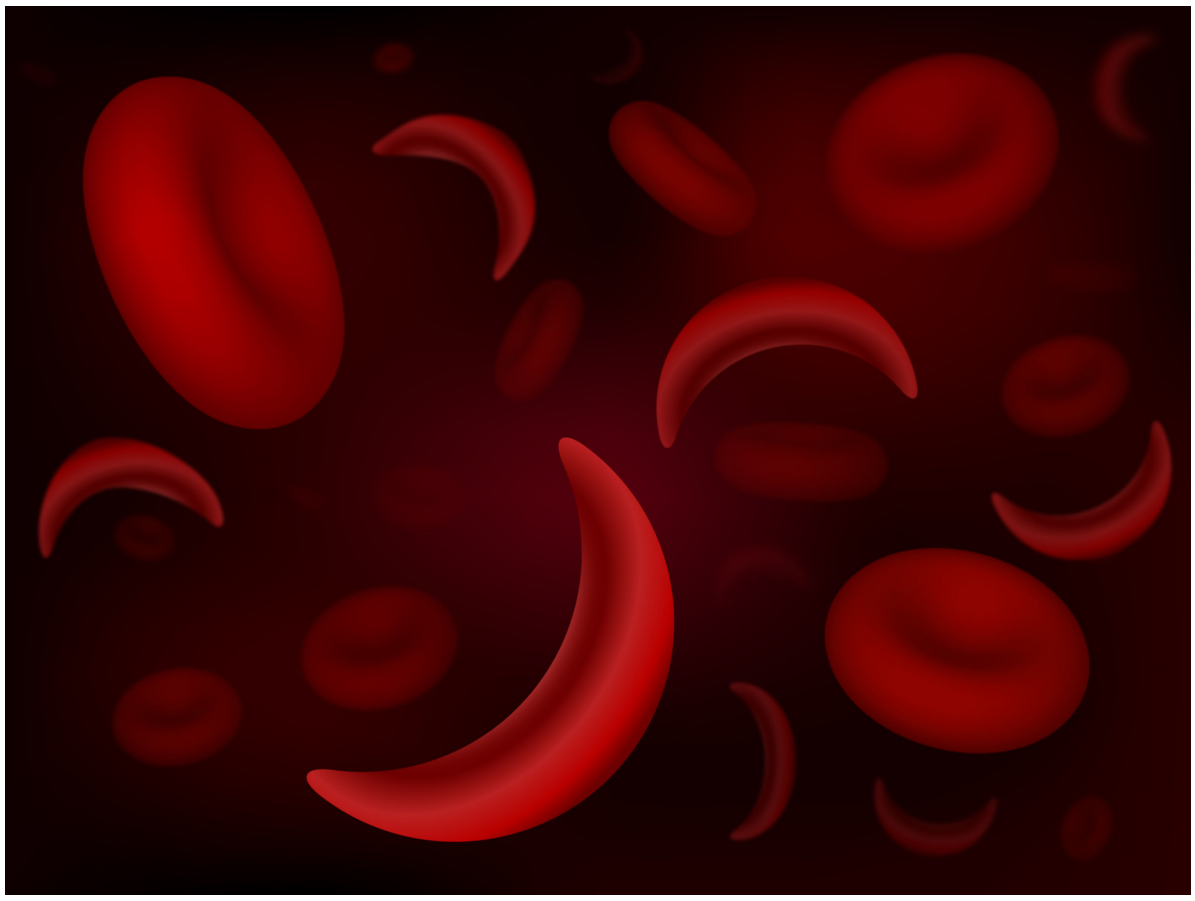
Pratima Niroula

Michel Ruiz-Fuentes

Parunjodhi Munisamy

December 2, 2021

Sickle Cell Disease Death Toll in terms of Age group

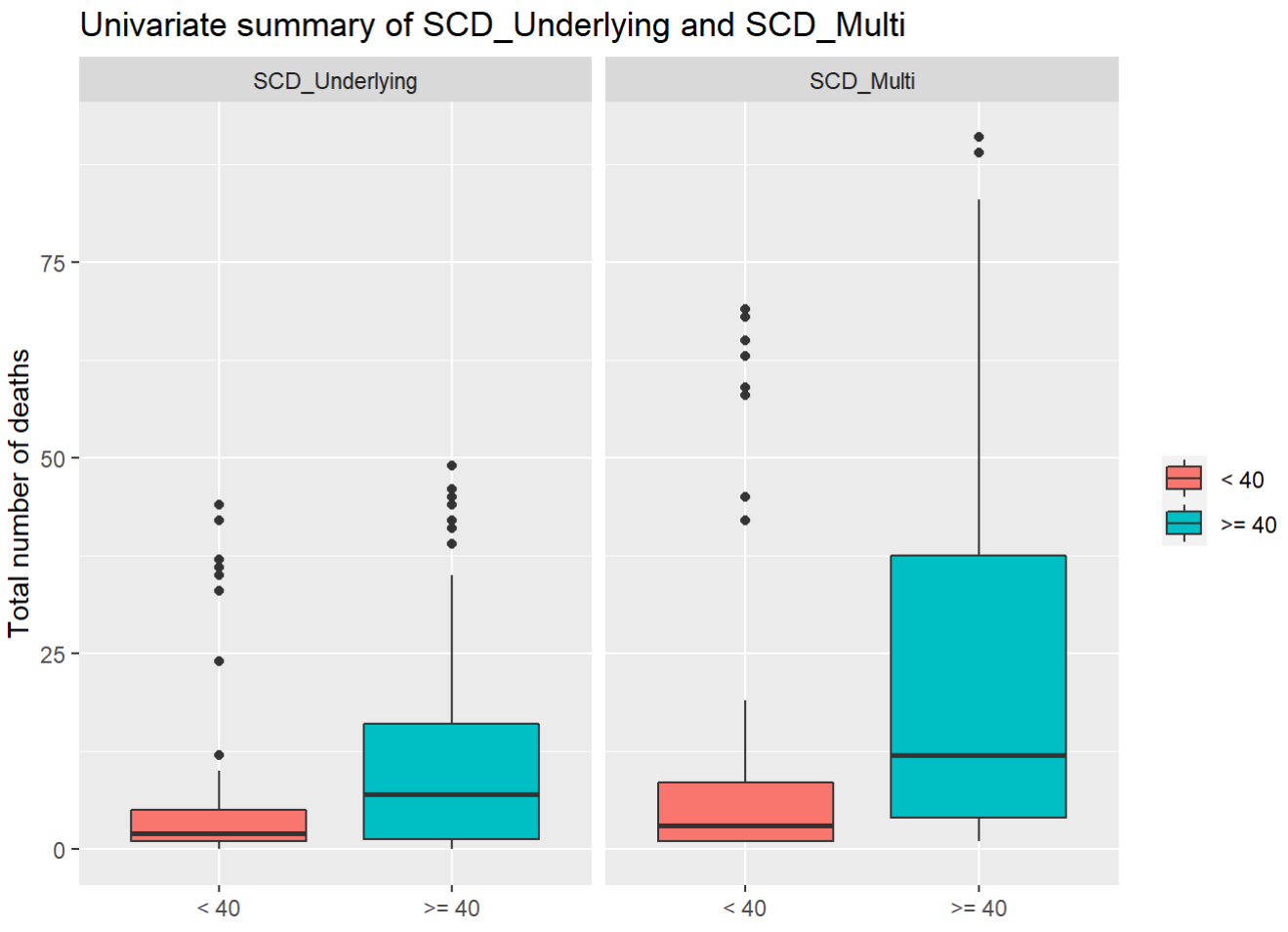


Sickle Cell Disease (SCD) is the most common genetic disorder in the world and affects about 100,000 people in the United States. As SCD is a genetic disease, symptoms start showing around 5 months old. However, as life expectancy for adults with SCD increases because of improvements in medical care, it becomes more important to analyze differences in SCD with age so as to focus on providing better care for all the complications that arise with aging with SCD. We chose to compare patients with SCD over 40 years of age and under 40 because prior studies have shown that hemoglobin, indirect bilirubin and platelet counts were significantly lower in patients over 40. Our ultimate goal would be to use this data and our conclusion to shed more light on the disproportionate impact on racial communities. For instance, several papers have shown that black communities are the most vulnerable to SCD.

When analyzing our dataset revolving around SCD, we concluded that our population of interest is people with sickle cell disease in the United States. This dataset also has 129 observations and 7 variables, but our investigations will focus on three numerical variables: `SCD_Multi` , `SCD_Underlying` , `SCD_COVID_19` , and one categorical variable: Age Group.

The variable Age Group would be ordinal because there is a natural ordering in those data points and the years of this sample collection range from 2019 to early 2021. Moving onto numerical variables,there are three: `SCD_Underlying` , `SCD_multi` and `SCD_COVID_19` . These are all discrete variables because they are counts of the number of deaths. `SCD_Underlying` represents the number of deaths with Sickle Cell Disease listed as underlying cause of death, `SCD_Multi` represents the number of deaths with Sickle Cell Disease listed as underlying or contributing cause of death and `SCD_COVID_19` represents the number of deaths with Sickle Cell Disease and COVID-19. We chose to primarily focus on comparing `SCD_Multi` and `SCD_Underlying` because we were observing the deaths between 2019 and early 2020, and we did not have 2019 data for individuals that died of COVID-19 and SCD.Additionally, the range of `SCD_Underlying` variables is 0-49, `SCD_Multi` ranges from 1-91 and `SCD_COVID_19` ranges from 0-12.

Furthermore, as mentioned above based on studies we read about Sickle Cell disease, most of the other laboratories divided the age groups equal to and below 40 years (<) and equal to and above 40 years (≥). Not to mention the probability of SCD occurring in older aged people is higher than in younger people.



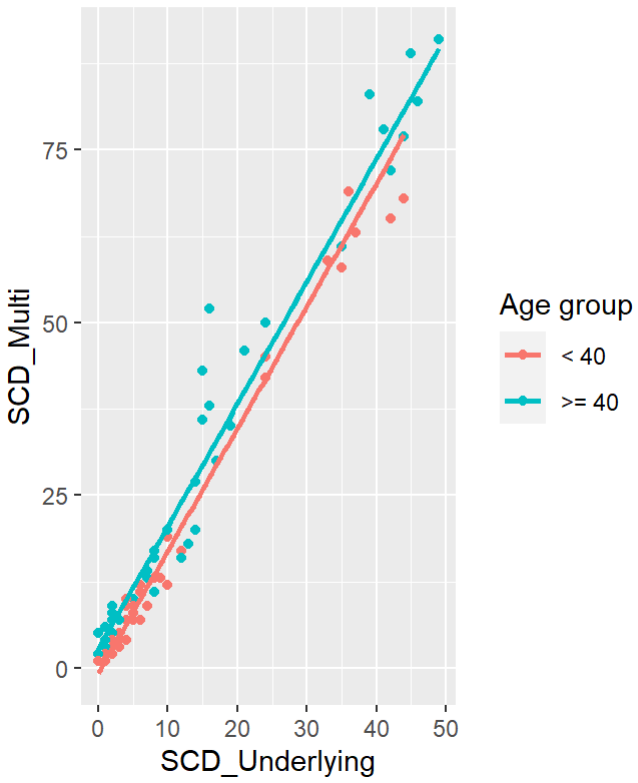
Comparing the two sets of box plots, we see that regardless of the type of SCD, the age group produces significant differences in total of deaths between its categories (at least in this dataset). This is in line with the papers we read and based our grouping on.

Hence, this furthers our interest in learning about how life expectancy and age affects people who have different types of sickle cell disease. Exploring this avenue in more statistical depth will give us meaningful conclusions about the role in aging in the likelihood of passing from SCD.

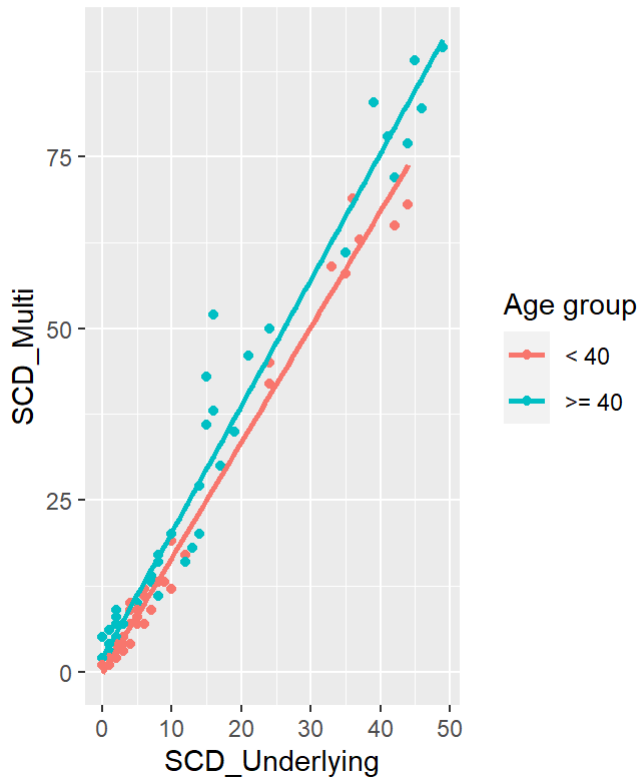
However, before we conduct our hypothesis testing, we have to look at which variables we should choose to focus on so that we can get meaningful results. As mentioned above, we can potentially infer trends among `SCD_Underlying` deaths, `SCD_Multi` deaths and our two different age groups. The best way to go about this is to create linear models who look at whether there is a relationship between these three variables.

So, first we are looking at a linear model where we assume age group does not have an effect on the relationship between `SCD_Underlying` and `SCD_Multi` deaths. This is otherwise known as the parallel slopes model. Our second model will look at the relationship between `SCD_Underlying` and `SCD_Multi` deaths but this time, we assume that the age group affects this relationship. This model is known as the interactions model.

Parallel Slopes Model between SCD_Underlying and SCD_Multi affected by age group



Interactions Model between SCD_Underlying and SCD_Multi affected by age group



Comparing the parallel slopes model with the interaction model, we can see that there is not much of a difference. The trend between `SCD_Underlying` and `SCD_Multi` deaths remains a positive and linear association. That is as `SCD_Underlying` deaths increase, `SCD_Multi` deaths increase also by the same amount on average.

Subsequently, after examining these linear trends, we decided to create our hypothesis testing around age group. Going from this, we then also group the deaths from `SCD_Underlying`, `SCD_Multi` and `SCD_COVID_19` into one variable: total_deaths.

We are now ready to carry out our hypothesis testing on our data directly instead of our linear models since the results from the latter were not very conclusive. As good practice, before doing anything else, we set our significance level to be 0.05.

Next up, we write out our hypotheses. Our null hypothesis is the one where we maintain the status quo, that is in this case, the proportion of total deaths from sickle cell disease for people aged 40 and above is 0.5. On the other hand, our alternate hypothesis, which is the one where we go against the status quo, is that the proportion of total deaths from sickle cell disease for people aged 40 and above is greater than 0.5. These hypotheses might seem banal since as people age, they are more likely to get sick. However, according to this article (https://www.medscape.com/answers/205926-15332/at-what-age-do-the-symptoms-of-sickle-cell-disease-scd-typically-develop), “Sickle cell disease (SCD) usually manifests early in childhood”. Moreover, there is evidence (https://ashpublications.org/blood/article/132/17/1750/39513/How-I-treat-the-older-adult-with-sickle-cell) that the cutoff at 40 years old makes sense statistically.

$$\mathbf{H_o \rightarrow p = 0.5}$$

$$\mathbf{H\alpha \rightarrow p > 0.5}$$

The next step in our hypothesis testing is to calculate our own test statistic, that is from our original dataset, how many of the total deaths are accounted for by people aged 40 and above.

```
## # A tibble: 2 x 2
##   agegroup  prop
##   <chr>    <dbl>
## 1 < 40     0.394
## 2 >= 40    0.606
```

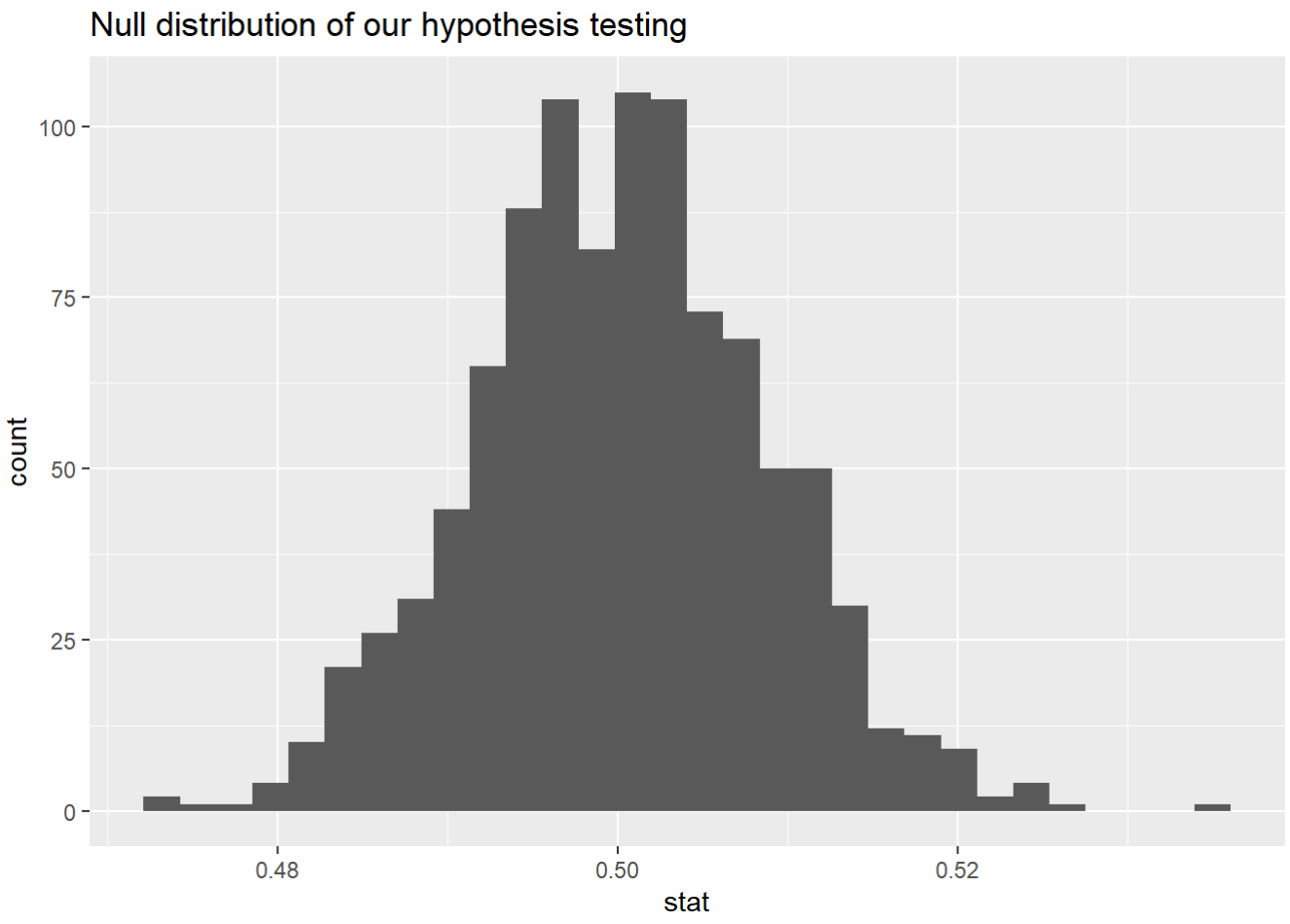
From the above calculations, the proportion of total deaths of people aged 40 and above is 0.61 to 2d.p. Therefore, our test statistics is 0.61.

After calculating our test statistic, we have to create our null world. This is where we use bootstrapping since we have only one sample which is our data set. We will sample with replacement where each resample is of the same size as our original dataset/sample that is 3187. However, to create the null world, we first have to create a dataframe that resembles it, that is where the proportions of total deaths from sickle cell disease for people aged 40 and above is 0.5. And by de facto, this means that our new data frame will have half of the total deaths accounted for by people aged below 40 and the other half by people aged 40 and above.

```
data_boostrap <- tibble(
  agegroup = c(rep(">= 40", 1593), rep("< 40", 1594))
)
```

```
null_dist <- data_boostrap %>%
  specify(response = agegroup, success = ">= 40") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop")
```

Now that we have our null bootstrapped distribution dataframe, we can plot it.



We then move on to calculating our p-value. The p-value in this case is the probability of getting our test statistic or a greater value given the null hypothesis is true. Since the model does not include any values after 0.53, we can calculate the p-value to be 0.

Since our p-value is less than our significance level (0.05), we therefore reject the null hypothesis. This does mean that the alternate hypothesis is favored but not accepted regardless. In other words, we are certain that the proportion of total deaths from sickle cell disease for people aged 40 and above is not equal 0.5.

Thus, this raises the question of what should we do with this conclusion now. Statistics is not only about producing statistically significant results. It is also about using those results and applying them in a ethical and unbiased way to improve the current policies and/or circumstances. We know that most people who have SCD in the US are black. In the future we would like to explore SCD death with respect to race and ethnicity as well as age group. We would also explore disparity in SCD in the US and globally. One action we hope could be implemented is raising more awareness about SCD especially among older people.

Nevertheless, a limitation of our data is that there is no information about where the sample was gotten from and thus we couldn't account for other confounding variables such. It is also hard to decide how to compare different numerical variables given that they were death counts but with slight variations, such as `SCD_Multi` , which accounted for sickle cell disease as well as multiple other diseases in the patient and `SCD_COVID_19` which accounted for sickle cell disease and COVID in a patient. Our focus was on comparing SCD-Underlying and SCD_Multi and future work could incorporate comparing `SCD_COVID_19` with the other variables. It would especially be interesting to explore `SCD_COVID_19` against race, seeing as how both sickle cell disease and COVID have been known to disproportionately affect marginalized communities.

References:

1) Manipulate image size in R Markdown (https://bookdown.org/yihui/rmarkdown-cookbook/figure-size.html)
2) Sickle Cell Disease and its Toll Compared in Different Age Groups in Study (https://sicklecellanemianews.com/2017/05/09/sickle-cell-disease-and-its-toll-compared-in-age-groups-in-pisces-study/)
3) How Does Sickle Cell Disease Affect People Differently (https://sickle-cell.com/demographics)
4) The Older Sickle Cell Patient (https://pubmed.ncbi.nlm.nih.gov/15164373/)
5) Sickle Cell Disease FAQs (https://www.sicklecelldisease.org/sickle-cell-health-and-disease/faqs/)
6) https://www.nidirect.gov.uk/conditions/sickle-cell-disease-sickle-cell-anaemia (https://www.nidirect.gov.uk/conditions/sickle-cell-disease-sickle-cell-anaemia)

…