Life Expectancy & Education Level

Richard Lin

2024-04-01

First, we set a random seed for out test. Since we're going to be using bootstrapping, this part of the process is really important to generate the random samples, which is why we do it first.

```
SEED <- 13013020;
```

Introduction and motivations.

Does the estimated life expectancy at birth vary depending on the education level of different countries?

Data cleaning.

Before performing the test the data was wrangled to have a more manageable and organized data set. This section describes the process of data wrangling the data went through previous to the testing.

To perform the test we will use two numerical variables: 'goal_4_dash' to measure quality education, and 'sowc_demographics___life-expectancy-at-birth-years_2021-0' to measure child labor.

Because of this, we'll extract the variable 'goal_4_score' and create a data set containing only this information and the country codes. We also renamed variables `Goal 4 Dash` and `Country Code ISO3`:

```
goal_4_information <- read_csv("sdr_fd5e4b5a.csv") %>%
  select(`Goal 4 Dash`, `Country Code ISO3`, country_label) %>%
  rename(goal_4_classifier = `Goal 4 Dash`,
         iso3_codes = `Country Code ISO3`)
```

```
## New names:
## Rows: 206 Columns: 59
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (36): Goal 1 Dash, Goal 1 Trend, Goal 2 Dash, Goal 2 Trend, Goal 3 Dash,... dbl
## (23): ...1, Goal 1 Score, Goal 2 Score, Goal 3 Score, Goal 4 Score, Goal...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
glimpse(goal_4_information)
```

```
## Rows: 206
## Columns: 3
## $ goal_4_classifier <chr> "SDG achieved", "Challenges remain", "Challenges rem~
## $ iso3_codes        <chr> "FIN", "SWE", "DNK", "DEU", "AUT", "FRA", "NOR", "CZ~
## $ country_label     <chr> "Finland", "Sweden", "Denmark", "Germany", "Austria"~
```

Next, we will extract the variable 'sowc_demographics___life-expectancy-at-birth-years_2021-0', as well as the ISO3 country codes. We also renamed the variables to 'life_expectancy_at_birth' and 'iso3_codes' respectively.

```
life_expectancy_information <- read_csv("country_indicators.csv") %>%
  select("sowc_demographics__life-expectancy-at-birth-years_2021-0", "iso3") %>%
  rename( life_expectancy_at_birth =
            "sowc_demographics__life-expectancy-at-birth-years_2021-0",
          iso3_codes="iso3")
```

```
## New names:
## Rows: 218 Columns: 1332
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (8): iso3, hdr_hdicode, hdr_region, wbi_income_group, wbi_lending_cat... dbl
## (1324): ...1, sowc_demographics__population-thousands-2021_total, sowc_d...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
glimpse(life_expectancy_information)
```

```
## Rows: 218
## Columns: 2
## $ life_expectancy_at_birth <dbl> 61.9824, 76.4626, 76.3767, 80.3684, 61.6434, ~
## $ iso3_codes               <chr> "AFG", "ALB", "DZA", "AND", "AGO", "AIA", "AT~
```

Now we're going to merge the two data sets into one using the 'iso3_codes' variable that they have in common.

```
data_research_question_3 <- merge(
  goal_4_information, life_expectancy_information, by="iso3_codes")

glimpse(data_research_question_3)
```

```
## Rows: 193
## Columns: 4
## $ iso3_codes               <chr> "AFG", "AGO", "ALB", "AND", "ARE", "ARG", "AR~
## $ goal_4_classifier        <chr> "Major challenges", "Major challenges", "Chal~
## $ country_label            <chr> "Afghanistan", "Angola", "Albania", "Andorra"~
## $ life_expectancy_at_birth <dbl> 61.9824, 61.6434, 76.4626, 80.3684, 78.7104, ~
```

Finally, we will split the countries into 4 groups each representing a different level of SDG Goal 4 achievement. These are, "SDG Achieved", "Challenges remain", "Major challenges remain" and "Significant challenges remain". Each of these data sets contains countries only in these groups and their respective 'life_expectancy_at_birth'.

```
achieved <- data_research_question_3 %>% filter(
  goal_4_classifier == "SDG achieved")

challenges_remain <- data_research_question_3 %>% filter(
  goal_4_classifier == "Challenges remain")

major_challenges_remain <- data_research_question_3 %>% filter(
  goal_4_classifier == "Major challenges")

significant_challenges_remain <- data_research_question_3 %>% filter(
  goal_4_classifier == "Significant challenges")
```

We can now use these data sets to perform the bootstrapping process. In the next section, we will construct a bootstrapping sample distribution for each education level group. This will enable us to find an estimate

of the median 'life_expectancy_at_birth' for each group to therefore make inferences.

Boostrapping.

Assuming that our random sample is representative of the full population, we will generate random samples by randomly sampling with replacement from the observed sample (the data under the 'life_expectancy_at_birth' variable).

We decided to find an estimate for the median, as the mean might be very sensitive to possible outliers.

To generate resamples, we used a for loop and the sample() function, and we sampled with replacement, so replace=TRUE. For each of the four groups created, SDG4 achieved, challenges remain in achieving SDG4, major challenges remain in achieving SDG4 and significant challenges remain in achieving SDG4 we performed a simulation of 10000 repetitions, with sample size 26, 56, 48 and 57 respectively.

Then, the test statistics from the sample were stores in tibbles, which enabled us to plot easily the sample distribution for the estimated median 'life_expectancy_at_birth' for each of the groups in separate histograms.

Then, we estimated a potential range of values for the true median 'life_expectancy_at_birth' using a 95% confidence interval. This allowed us to be 95% sure that the true median 'life_expectancy_at_birth' for a given group is in that range (between quantiles 2.5 and 97.5 of the distribution).

We will proceed with the bootstrapping:

Education level group: **SDG Achieved**

```
n1 <- 26
repetitions <- 10000
sim1 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim1 <- sample(na.omit(achieved$life_expectancy_at_birth),
                     size = n1, replace=TRUE)
  sim_median1 <- median(new_sim1)
  sim1[i] <- sim_median1
}
sim1 <- tibble(median = sim1)
```

Next, we calculated a 90% confidence interval:

```
quantile(sim1$median, c(0.05, 0.95))
```

```
##       5%      95%
## 74.13025 78.71290
```

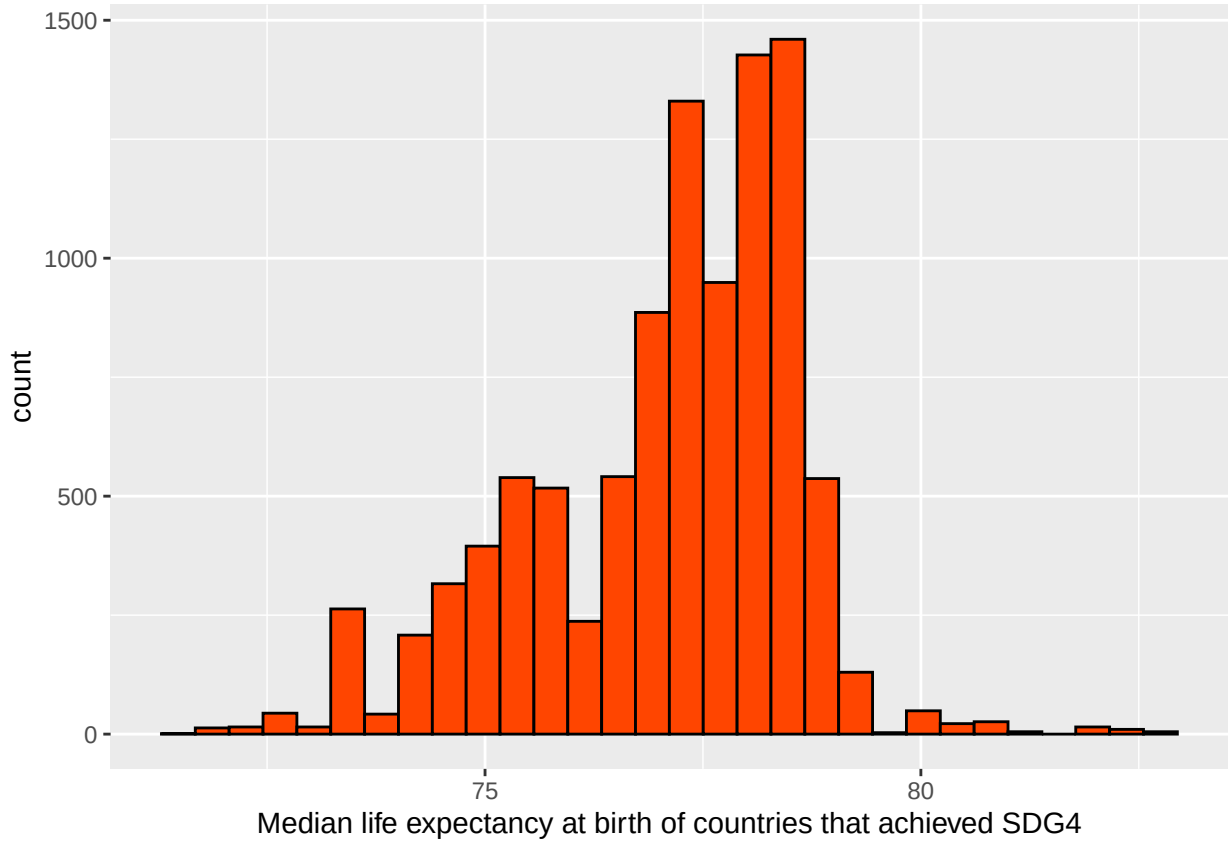
From this we know that the true median life expectancy at birth for countries who have already achieved SDG 4 is between 73.62 and 78.72.

Visualizations:

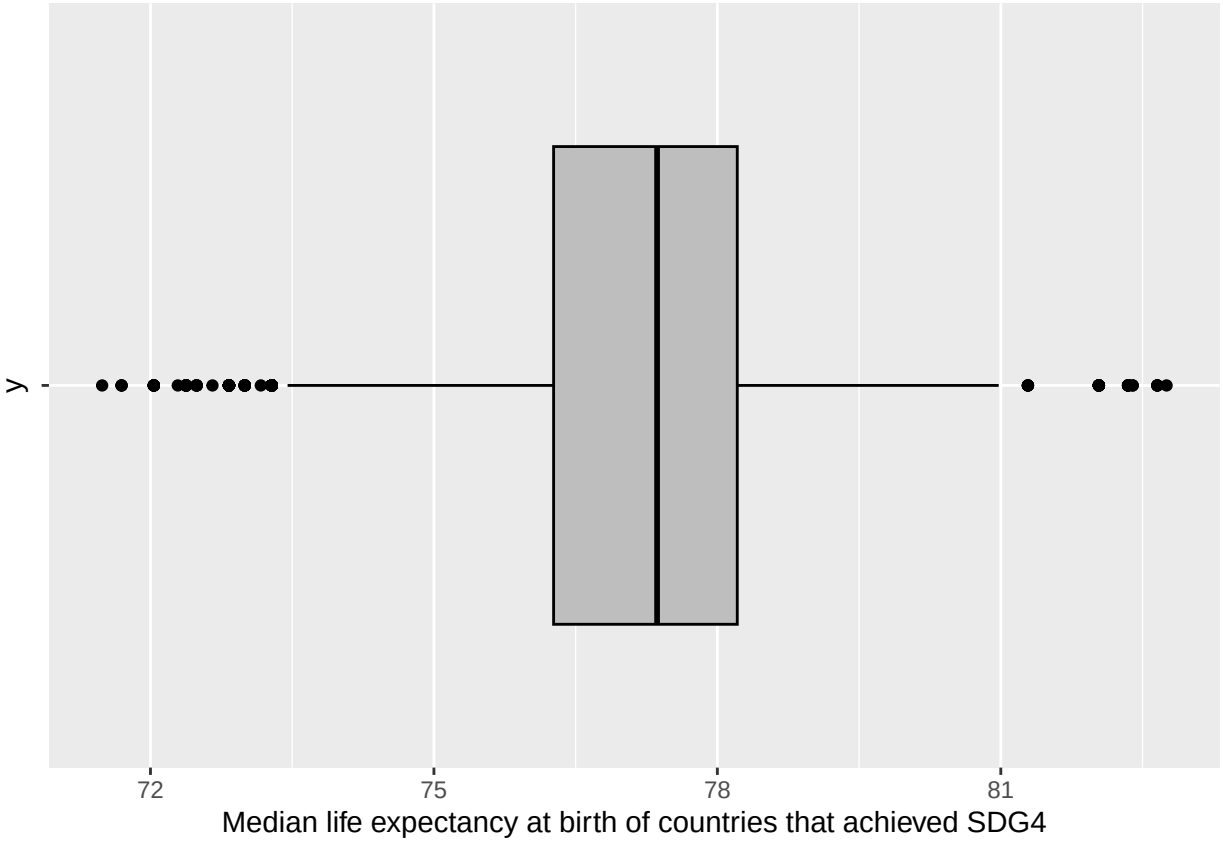
```
hist1 <- ggplot(data = sim1, aes(x = median))+
            geom_histogram(colour = "black",
                 fill = "orangered1",
                 bins = 30) +
            labs(x = "Median life expectancy at birth of countries that achieved SDG4")

boxplot1 <- ggplot(data = sim1, aes(x = median, y = "")) +
```

```
            geom_boxplot(colour="black", fill="gray") +
            labs(x = "Median life expectancy at birth of countries that achieved SDG4")
```

```
hist1
```




```
boxplot1
```



Education level group: **SDG Challenges remain**

```
n2 <- 56
repetitions <- 10000
sim2 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim2 <- sample(na.omit(challenges_remain$life_expectancy_at_birth) ,size = n2, replace=TRUE)
  sim_median2 <- median(new_sim2)
  sim2[i] <- sim_median2
}
sim2 <- tibble(median = sim2)
```

Next, we calculated a 90% confidence interval:

```
quantile(sim2$median, c(0.05, 0.95))
```

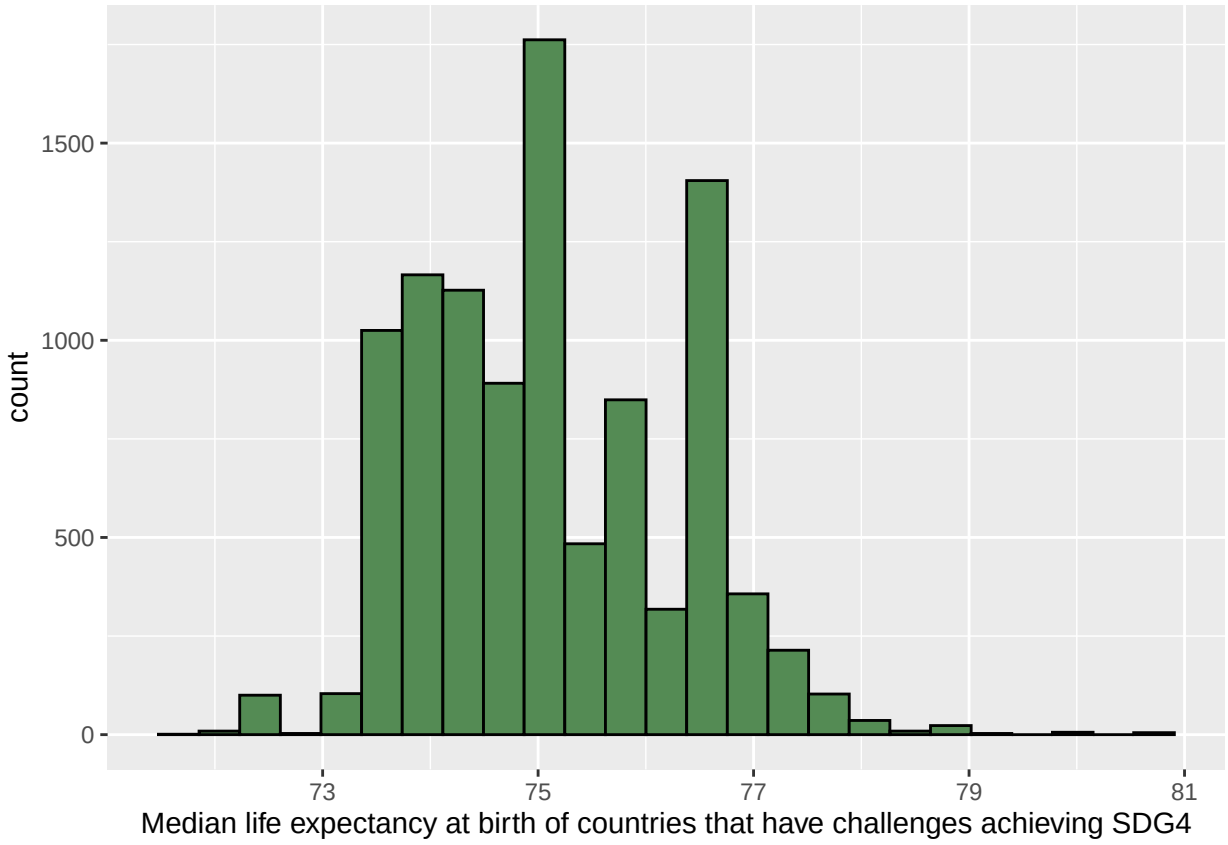
```
##       5%      95%
## 73.67645 77.06710
```

From this we know that the true median life expectancy at birth for countries who still have challenges regarding SDG 4 is between 73.67 and 77.20.

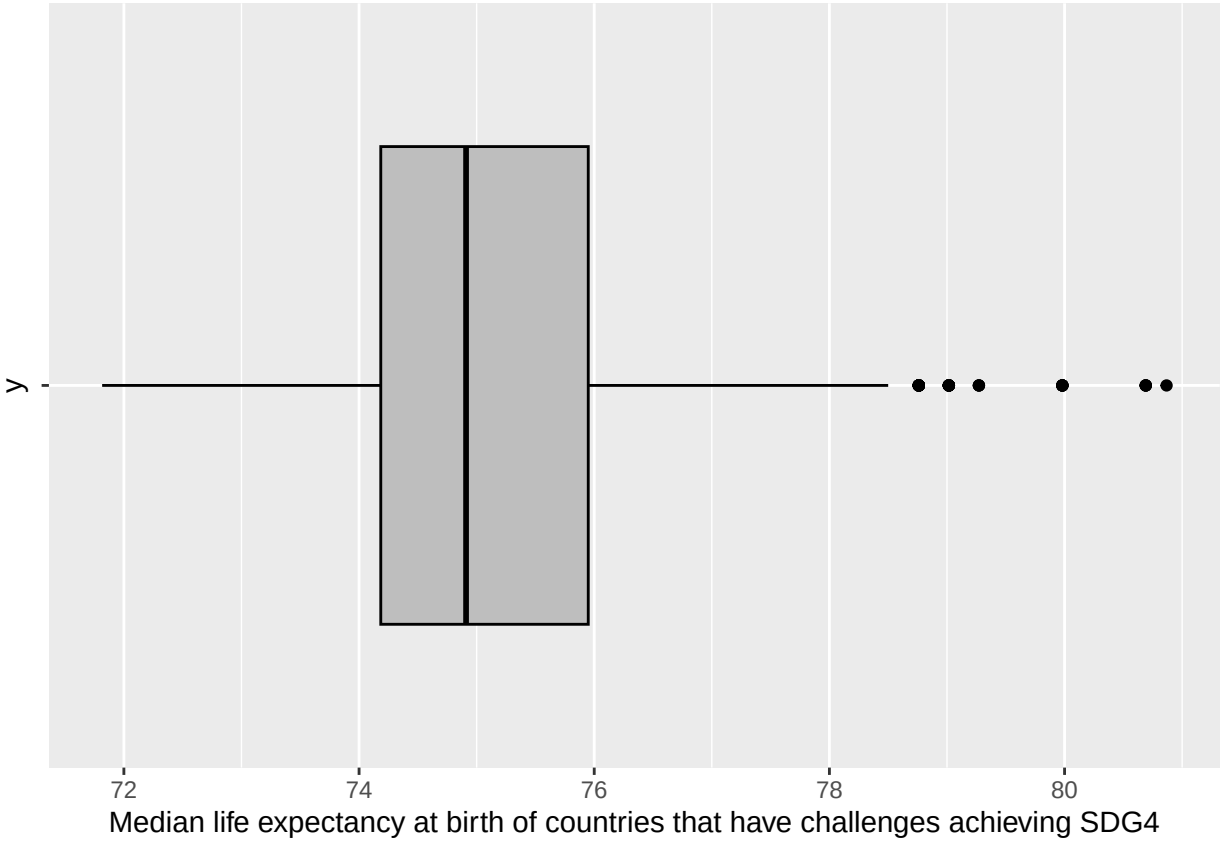
Visualizations

```
hist2 <- ggplot(data = sim2, aes(x = median))+
            geom_histogram(colour = "black",
                 fill = "palegreen4",
```

```
                bins = 25) +
            labs(x = "Median life expectancy at birth of countries that have challenges achieving SDG4")

boxplot2 <- ggplot(data = sim2, aes(x = median, y = "")) +
              geom_boxplot(colour="black", fill="gray") +
              labs(x = "Median life expectancy at birth of countries that have challenges achieving SDG4

hist2
```




```
boxplot2
```



Education level group: **SDG Major Challenges remain**

```
n3 <- 48
repetitions <- 10000
sim3 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim3 <- sample(na.omit(major_challenges_remain$life_expectancy_at_birth) ,size = n3, replace=TRUE)
  sim_median3 <- median(new_sim3)
  sim3[i] <- sim_median3
}
sim3 <- tibble(median = sim3)
```

Next, we calculated a 90% confidence interval:

```
quantile(sim3$median, c(0.05, 0.95))
```

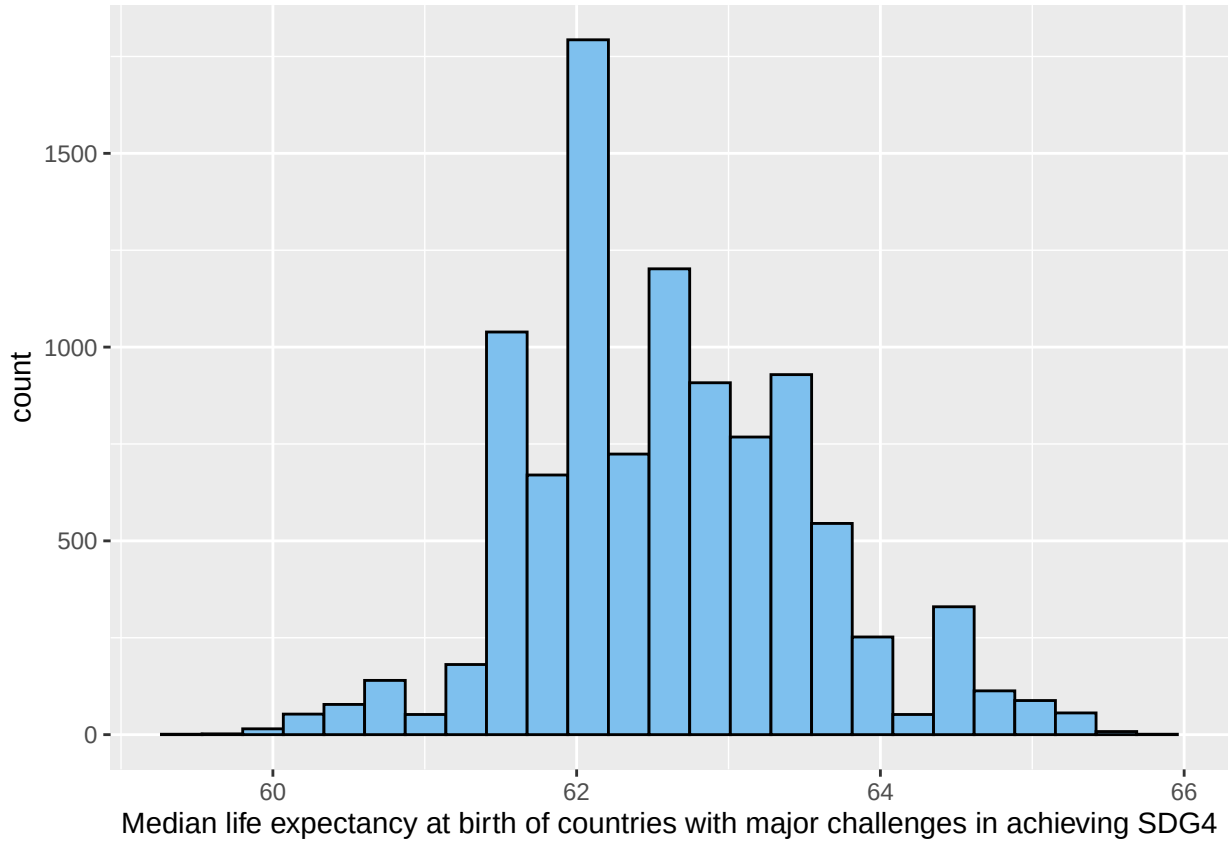
```
##       5%      95%
## 61.20495 64.36360
```

From this we know that the true median life expectancy at birth for countries who still have major challenges regarding SDG 4 is between 60.74 and 64.49.

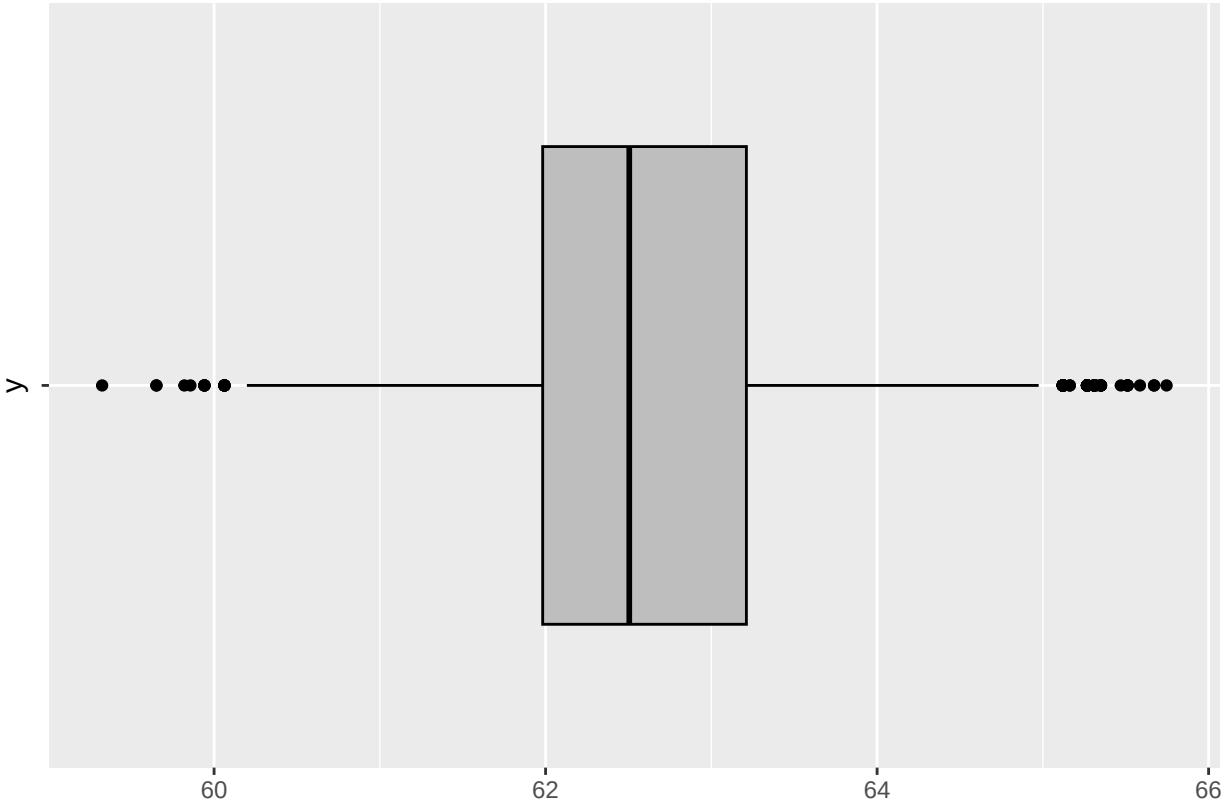
Visualizations

```
hist3 <- ggplot(data = sim3, aes(x = median))+
            geom_histogram(colour = "black",
                 fill = "skyblue2",
```

```
                bins = 25) +
            labs(x = "Median life expectancy at birth of countries with major challenges in achieving SD

boxplot3 <- ggplot(data = sim3, aes(x = median, y = "")) +
              geom_boxplot(colour="black", fill="gray") +
              labs(x = "Median life expectancy at birth of countries with major challenges in achieving

hist3
```




```
boxplot3
```



Median life expectancy at birth of countries with major challenges in achieving SDG4

Education level group: **SDG Significant Challenges remain**

```
n4  <- 57
repetitions <- 10000
sim4 <- rep(NA, repetitions)
set.seed(SEED %% 200)

for (i in 1:repetitions)
{
  new_sim4 <- sample(na.omit(significant_challenges_remain$life_expectancy_at_birth) ,size = n4, replac
  sim_median4 <- median(new_sim4)
  sim4[i] <- sim_median4
}
sim4 <- tibble(median = sim4)
```

Next, we calculated a 90% confidence interval:

```
quantile(sim4$median, c(0.05, 0.95))
```

```
##      5%     95%
## 70.4697 72.8140
```

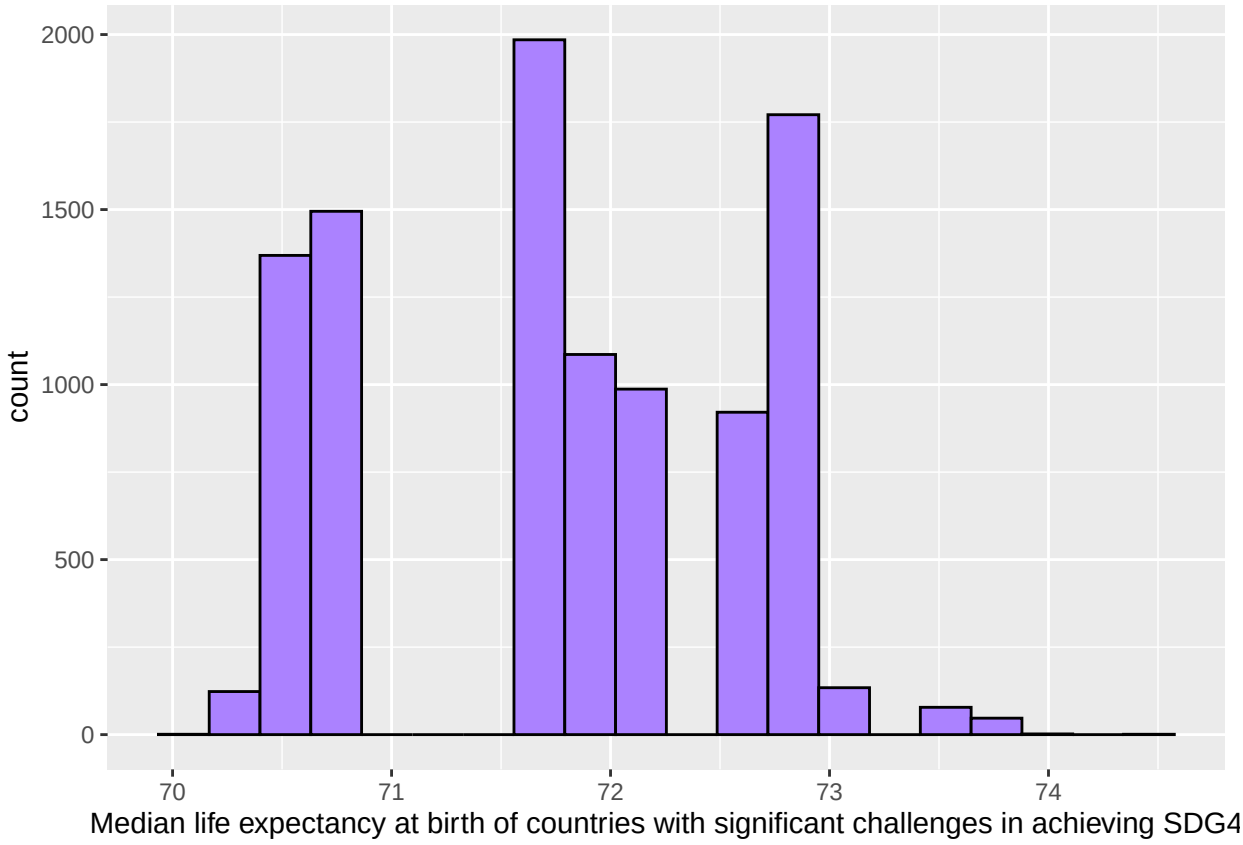
From this we know that the true median life expectancy at birth for countries who still have significant challenges regarding SDG 4 is between 68.65 and 74.13.

Visualizations

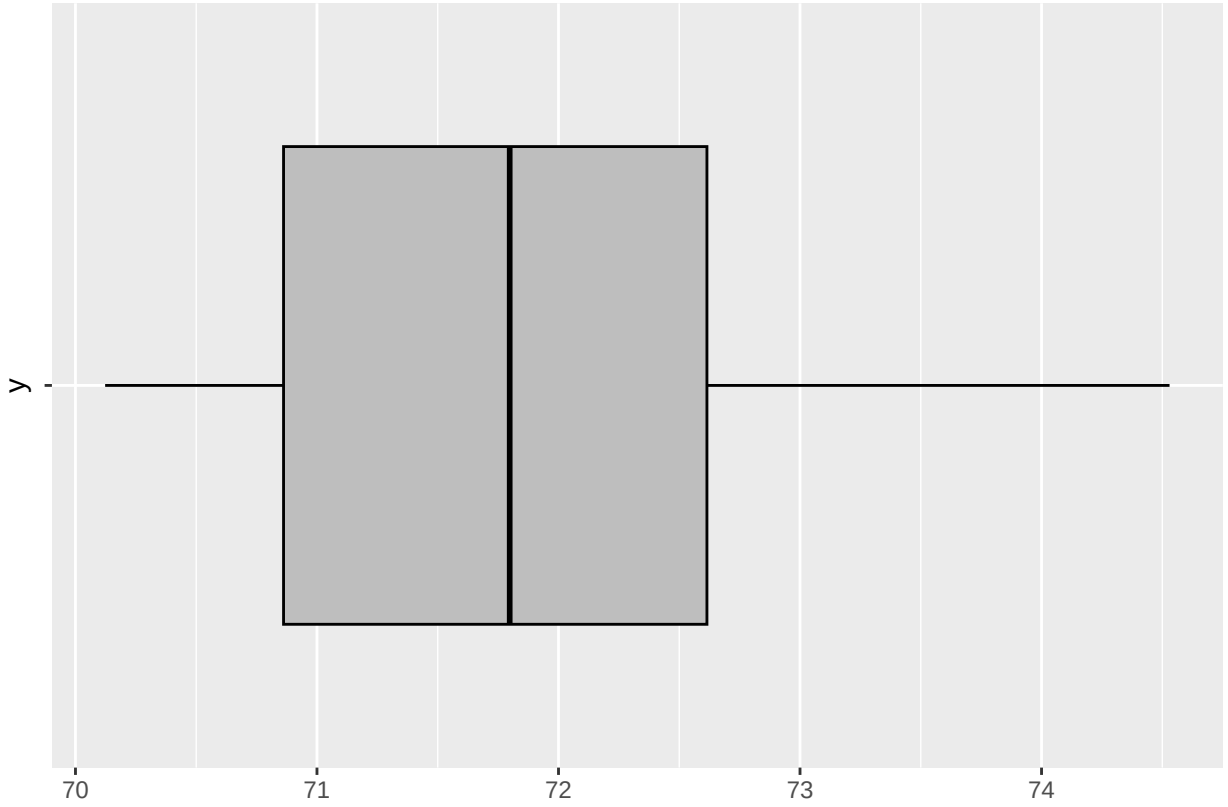
```
hist4 <- ggplot(data = sim4, aes(x = median))+
            geom_histogram(colour = "black",
                 fill = "mediumpurple1",
```

```
                bins = 20) +
            labs(x = "Median life expectancy at birth of countries with significant challenges in achie

boxplot4 <- ggplot(data = sim4, aes(x = median, y = "")) +
              geom_boxplot(colour="black", fill="gray") +
              labs(x = "Median life expectancy at birth of countries with significant challenges in achi
```

```
hist4
```

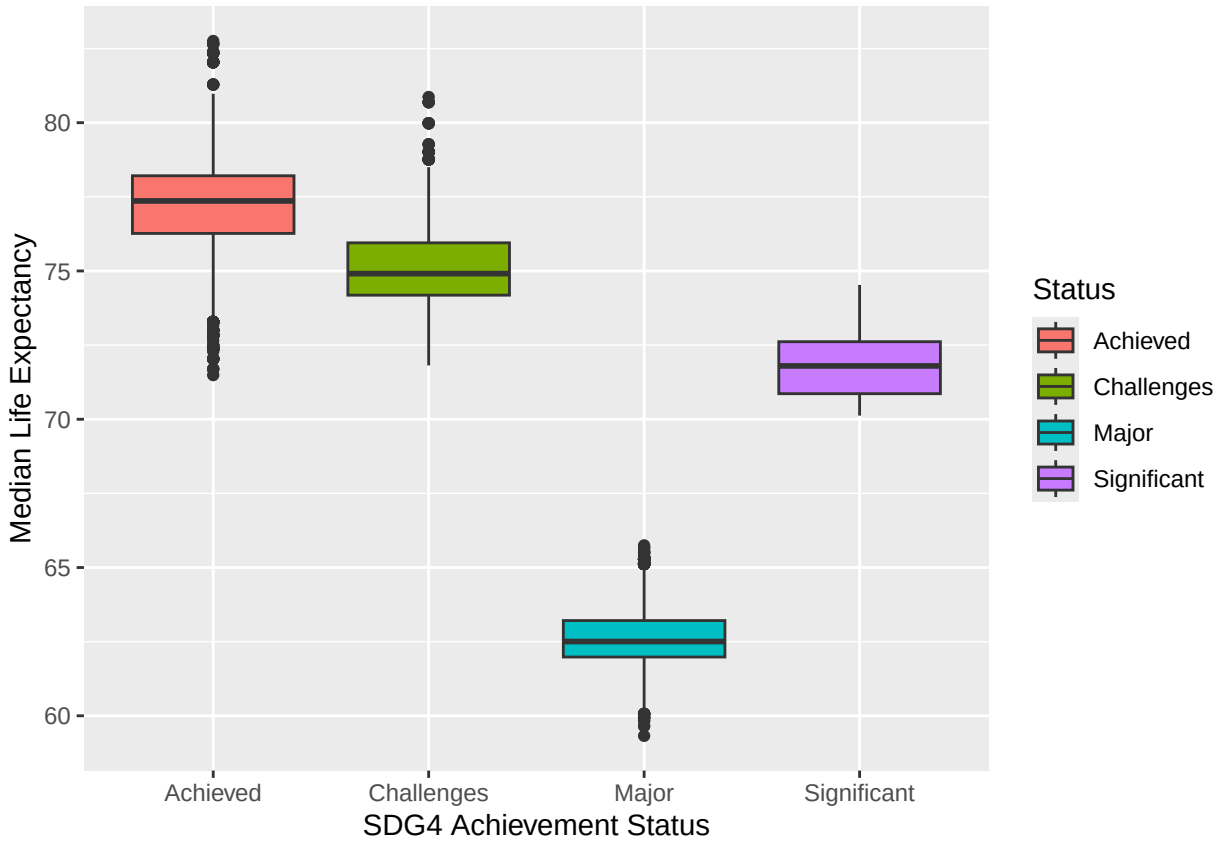



```
boxplot4
```



Median life expectancy at birth of countries with significant challenges in achieving SDG4

```
combined_data <- rbind(mutate(sim1, Status = "Achieved"),
                       mutate(sim2, Status = "Challenges"),
                       mutate(sim3, Status = "Major"),
                       mutate(sim4, Status = "Significant"))

ggplot(combined_data, aes(x = Status, y = median, fill =Status)) +
  geom_boxplot() +
  labs(x = "SDG4 Achievement Status",
       y = "Median Life Expectancy")
```




```
sim1_quantile <- quantile(sim1$median, c(0.05, 0.95))
sim2_quantile <- quantile(sim2$median, c(0.05, 0.95))
sim3_quantile <- quantile(sim3$median, c(0.05, 0.95))
sim4_quantile <- quantile(sim4$median, c(0.05, 0.95))

SDG4_status_quantiles <- data.frame("Achieved" = sim1_quantile,
                                    "Challenges_remain" = sim2_quantile,
                                    "Major_challenges" = sim3_quantile,
                                    "Significant_challenges" = sim4_quantile)

SDG4_status_quantiles
```

```
##     Achieved Challenges_remain Major_challenges Significant_challenges
## 5%  74.13025          73.67645         61.20495                 70.4697
## 95% 78.71290          77.06710         64.36360                 72.8140
```

Analysis.

From our confidence intervals we obtained the following ranges for true life expectancy at birth median:

SDG Achieved = 74.1 - 78.7

SDG Challenges Remain = 73.7 - 77.1

SDG Major Challenges Remain = 61.2 - 64.4

SDG Significant Challenges Remain = 70.5 - 72.8

In relation to the historgrams for each of the SDG4 status categories, the shape of the bootstrapping distributions appears to be following a normal distribution for countries which have

challenges remaining and have major challenges remaining towards SDG4, on the contrary, the shape of the bootstrapping distribution for countries which achieved SDG4 and countries which have significant challenges in achieving SDG4 appears to be left skewed and multimodal respectively. The values for SDG4 achieved, challenges remain and major challenges categories appears to be concentrated around the range 73-78, 73.5-76, 61.5-63.5 respectively, whereas, in the case of significant challenges remain category there appears to be 3 separate clusters of values. Additionally, in terms of centre they are, around 78, 75, 62 and 71.5 respectively.

Regarding the box-plots of each SDG4 status category, the median life expectancy of countries that have achieved SDG goal 4 is around 77, the next is by countries that have challenges remaining with a median life expectancy of around 75, which is followed by countries that have significant challenges at around 72 and lastly countries with major challenges at 62.5. The countries that have achieved SDG goal 4 appear to have the greatest number of extreme outliers when compared to the other 3 categories, conversely, the countries with significant challenges have no extreme outliers. Additionally, the interquartile range of countries that have achieved, have challenges remaining and have significant challenges remaining towards SDG4 are are about the same, whereas, countries with major challenges in acheiving SDG4 has the smallest IQR indicating that the spread of values is the lowest compared to the aforementioned categories.

Furthermore, it is noteworthy that the median life expectancy of countries that have achieved SDG goal 4 is comparatively higher than those that have not. Also, it is interesting that the median life expectancy of countries with significant challenges in achieving SDG4 is significantly higher than that of countries with major challenges in achieving SDG4 which introduces the possibility of confounding factors which can skew the results of this study such as the availability of healthcare to the general public, lifestyle choices and income disparity. One limitation of this study is the way SDG4 status is classified, this due to the fact that some countries that have significant challenges in achieving SDG4 has a higher goal 4 score than countries that have achieved this goal.

As we can observe, the estimated median values for all for groups lie between 60.74 and 78.72, which means we have quite a wide range of values. We found that the groups where SDG 4 was already achieved and where there were still a few challenges had very similar ranges, with the group for SDG Achieved having a slightly higher end of the confidence interval.

We also observe a big difference when going from the two groups with higher advance in SDG 4 to the group where major challenges remain, as the lower end of the range of values drops from 73.6 to 60.7, more than 10 points. This was quite surprising for us, however, the most surprising finding was that the group where significant challenges remain, has a higher range of values for the median life expectancy at birth than the group where major challenges remain. The difference between the groups is almost 10 points on the lower end and 6 points on the higher end.

The lowest range value is found in the SDG Major Challenges Remain group (60.74), on the other hand, countries in the group SDG Achieved had the highest possible value for the true life expentancy at birth (78.72). Furthermore, we see the largest range of possible for values for SDG Achieved (5.11), which means that the estimates from the sampling distribution were more spread apart, so there was a larger variability for the test statistics in this group. However, most of the ranges for the other groups also lie around 4, except for the group where significant challenges remain, where the range is only of 2.5 points, indicating the lowest variability in values.