An ANOVA approach for statistical comparision of human capital availability in countries with different HIV/AIDS prevalence

Wai Yu Amanda Ng - 1008217350

November 4 2022

Abstract

In past research, it has been shown that Human immunodeficiency virus infection and acquired immunodeficiency syndrome (HIV/AIDS) affects economic growth in South Africa countries *(Bell, C., 2006)*. However, global research on the relationship between HIV/AIDS prevalence and human capital availability of countries has not been established yet. In this study, we focus on the statistical comparison of human capital availability in countries with different HIV/AIDS prevalence using one-way ANOVA. The major finding of this study is countries with different HIV/AIDS prevalence generally have different human capital availability. Besides, we discuss the potential bias in the findings due to unequal sample size in human capital availability groups. Finally, we acknowledge the significance of our main finding and raise concerns on the difficulties that high HIV/AIDS prevalence countries may face in ameliorating the situation.

This report is meant for readers with experience in hypothesis testing such as two-sample t-test. Readers who have knowledge about p-test are also capable to understand the concept of two-sample t-test by reading the section "Review of Two-sample t-test".

Keywords:

Analysis of variance (ANOVA), group mean, hypothesis, F-statistic

Definitions

- An **observation** is a fact or figure we collect about a given variable. It can be expressed as a number or as a quality.
- A **mean** is a measure of the average of the sample data set.
- A **sample** is an observed sample set obtained from the population of interest.
- A **group** consists of a set of observations extracted from the sample, all of which possess one, or more, characteristics in common.
- The value of **degree of freedom** indicates the number of independent values that are free to vary as we conduct a statistics analysis to estimate parameters.
- A **test statistic** is a value calculated from a statistical test of a hypothesis. It shows how closely our observed sample data match the distribution expected under the null hypothesis of that statistical test.
- A **critical value** is a cut-off value that is used to mark the start of a region where the test statistic, obtained in hypothesis testing, is unlikely to fall in.
- The **significance level** of a statistical test is the probability that the event could have occurred by chance.

Review of two-sample T-test

Two-sample T-test is a statistical method to test whether the means of two independent groups from the population are statistically different *(Kumar, 2022)*.

Assumptions of two-sample T-test

Below are some primary assumptions made when using two-sample T-test:

- Each observation must be independent, i.e. measurements for one observation do not affect measurements for any other observation.
- The sample is obtained through random sampling from the population.
- The observations in each group come from a normal distribution population.
- Data values are continuous.
- The variances for the two independent groups are equal.

Methodology of two-sample t-test

Step 1: Formulate the null and alternate hypothesis.

Null hypothesis: There is no difference between the means of the two independent groups("group 1" and "group 2").

$$H_0 : \mu_1 = \mu_2$$

where

- μ_1 is the mean of group 1
- μ_2 is the mean of group 2

Alternative hypothesis: There is difference between the means of the two independent groups.

$$H_a : \mu_1 \neq \mu_2$$

where

- μ_1 is the mean of group 1
- μ_2 is the mean of group 2

Step 2: Calculate test statistic *t-value*

The test statistic *t-value* (T) is given by the following formula *(Ganti,2022)*.

$$T = \frac{\bar{Y_1} - \bar{Y_2}}{\sqrt{\frac{s_1^2}{n_1} - \frac{s_2^2}{n_2}}}$$

where

- $\bar{Y_1}$ and $\bar{Y_2}$ are the sample means of group 1 and group 2 respectively.
- s_1^2 and s_2^2 are the variances of group 1 and group 2 respectively.
- n_1 and n_2 are the sample size of group 1 and group 2 respectively.

The numerator of *t-value* (T) represents the difference between the two group means while the denominator represents the standard error of the difference between the two population means.

Step 3 : Calculate *degree of freedom*

The degree of freedom (*df*) is given by the following formula *(Ganti,2022)*:

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^2}{n_1^2(n_1-1)} + \frac{s_2^2}{n_2^2(n_2-1)}}$$

where

- s_1^2 and s_2^2 are the variances of group 1 and group 2 respectively.
- n_1 and n_2 are the sample size of group 1 and group 2 respectively.

Step 4: Set a *significance level*

The significance level is the fixed probability of mistaken rejection of the null hypothesis when it is true.

It is common practice to set the significance level α as 0.05. In two-sample t-test, this α indicates that there is a 5% risk of concluding that there is a difference between two group means when there is no actual difference.

Step 5: Produce *critical value*

Using the following r-code function `qt()`, we can produce a a critical value $t_{\alpha,df}$. The first parameter of the function is the significance level (α) and the second parameter is the degree of freedom (*df*).

This critical value $t_{\alpha,df}$ represents the theoretical value from the t-distribution under the assumption that null hypothesis is true *(Ganti,2022)*.

Step 6: Draw conclusion

By comparison of our test statistic, t-statistic T, and the theoretical value $t_{\alpha,df}$, we can decide to reject the null hypothesis or not *(Ganti,2022)*.

If the t-statistic T is higher than the critical value t at α level, we can reject the null hypothesis that there is no difference between the means of interested parameter of the two independent groups. Otherwise, if the t-statistic T is less than or equal to the critical value t at α level, then one cannot reject the null hypothesis.

Introduction to ANOVA

ANOVA (ANalysis Of VAriance) is a statistical test to determine if the group means are significantly different from each other by comparing the variances of individual groups *(Bevans, 2022)*.

The null hypothesis for ANOVA is all group means are equal. Hence, if the mean of any group does not match with the remaining groups, the test is regarded as statistically significant. We can then reject the null hypothesis and suggest that the groups do NOT have equal mean.

In one-way ANOVA, the sample data is categorized into several groups based on one single grouping variable, namely the **factor variable** *(Soetewey, 2020)*. If there are two factor variables involved in the classification of groups, we name the test as two-way ANOVA.

NOTE: There are several versions of the ANOVA (e.g., one-way ANOVA, two-way ANOVA, mixed ANOVA, repeated measures ANOVA, etc.). In this report, only the simplest form, one-way ANOVA, is presented and we refer to it as ANOVA in the remaining of the report.

Relationship between ANOVA and two-sample t-test

In fact, ANOVA is an extension of two-sample t-test by allowing the number of compared groups to be more than 2 groups *(Ostertagova, Ostertag, 2013)*. In particular, if we are only comparing means of two groups, ANOVA and two-sample t-test are equivalent and their underlying relationship can be given by:

$$F = t^2$$

where

- F is F-statistic / test statistic of ANOVA
- t is t-statistic / test statistic of two-sample t-test

Assumptions of ANOVA

There are three primary assumptions made when using ANOVA *(Soetewey, 2020)*:

- All observations, regardless of their associated groups, are independent of one another.
- The observations in each group come from a normal distribution population.
- The distribution of means in each group have the same variance.

Real life examples using ANOVA

1. Evaluation of academic performance of students from different schools by comparing the average GPA obtained by students of each school.
2. Comparison of average weight of individuals living in different countries or regions.
3. Comparison of effectiveness, measured by average accuracy in projecting infected cases, of three mathematical methods (nonlinear regression, SIR, and SEIR epidemic models) in tracking COVID-19 disease in nine countries affected by the SARS-CoV-2 virus *(Lucero, 2021)*.

Introduction to the study

It is shown in past research that Human immunodeficiency virus infection and acquired immunodeficiency syndrome (HIV/AIDS) affects economic growth by reducing the availability of human capital in South Africa countries *(Bell, Devarajan, Gersbach, 2006)*.

In other words, people who are infected with HIV/AIDS are believed to work less efficiently due to their physical health conditions, hence reducing the amount of work products. Thus, we can conclude that the prevalence of HIV/AIDS in a society may eventually affects the amount of resources (i.e. products of human activities) available to its society members.

Based on past results, it is reasonable to expect that populations with significantly different human capital availability have significantly different HIV/AIDS prevalence. In this study, we are interested in investigating if this hypothesis is true.

Firstly, we will classify all countries into three levels according to their Human Capital Index, which serves as a measurement of human capital availability. Then, we will compare the group means of number of people ages 15-49 who are infected with HIV/AIDS, which serves as a measurement of the populations' prevalence of HIV/AIDS. Lastly, we will preform ANOVA test to make a conclusion about our research hypothesis.

Below are the hypotheses of our research problem:

Null hypothesis:

Countries with different human capital availability DO NOT have different HIV/AIDS prevalence.

Equivalently: There is no difference between means of number of people ages 15-49 who are infected with HIV/AIDS of the 3 groups of countries (classified by Human Capital Index).

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

where

- μ_1 is the means of the number of people ages 15-49 who are infected with HIV/AIDS of the group of countries with 'low' human capital availability.
- μ_2 is the means of the number of people ages 15-49 who are infected with HIV/AIDS of the group of countries with 'middle' human capital availability.
- μ_3 is the means of the number of people ages 15-49 who are infected with HIV/AIDS of the group of countries with 'high' human capital availability.

Alternative hypothesis (Research hypothesis):

Countries with different human capital availability have different HIV/AIDS prevalence.

Equivalently: There is difference between the means of the number of people ages 15-49 who are infected with HIV/AIDS of the 3 groups of countries (classified by Human Capital Index).

$$H_a : \mu_1 \neq \mu_2 \quad \text{or} \quad \mu_2 \neq \mu_3 \quad \text{or} \quad \mu_1 \neq \mu_3$$

where

- μ_1 is the means of the number of people ages 15-49 who are infected with HIV/AIDS of the group of countries with 'low' human capital availability.
- μ_2 is the means of the number of people ages 15-49 who are infected with HIV/AIDS of the group of countries with 'middle' human capital availability.
- μ_3 is the means of the number of people ages 15-49 who are infected with HIV/AIDS of the group of countries with 'high' human capital availability.

Data

1. The sample data set

1.1 Important variables

```
# Load data set
data <- read_csv("data.csv",
                 col_types = cols(
                   `Adults (ages 15+) living with HIV [SH.DYN.AIDS] - 2020 [YR2020]` =
                     col_integer(),
                   `Human capital index (HCI) (scale 0-1) [HD.HCI.OVRL] - 2020 [YR2020]` =
                     col_number()))
```

The sample data set "data.csv" is extracted from World Bank's "Health Nutrition and Population Statistics" report by choosing the following variables:

- "Country name"
- "Adults (ages 15+) living with HIV [SH.DYN.AIDS] - 2020 [YR2020]"
- "Human capital index (HCI) (scale 0-1) [HD.HCI.OVRL] - 2020 [YR2020]"

The variable "Adults (ages 15+) living with HIV [SH.DYN.AIDS] - 2020 [YR2020]" measures the number of people ages 15-49 who are infected with HIV/AIDS of a country in year 2020.

The variable "Human capital index (HCI) (scale 0-1) [HD.HCI.OVRL] - 2020 [YR2020]" indicates the Human Capital Index of a country in year 2020. It is a continuous variable which ranges from 0 to 1, which calculates the amount of resources available for the next generation.

1.2 Cleaning Process

```
# Cleaning process
data_1 <- data%>%
  rename(HIV = `Adults (ages 15+) living with HIV [SH.DYN.AIDS] - 2020 [YR2020]`)%>%
  rename(HCI = `Human capital index (HCI) (scale 0-1) [HD.HCI.OVRL] - 2020 [YR2020]`)%>%
  filter(!(is.na(HIV)))%>%
  filter(!(is.na(HCI)))%>%
  select(`Country Name`, HIV, HCI)
```

For easier coding, we first rename the variable names from "Adults (ages 15+) living with HIV [SH.DYN.AIDS] - 2020 [YR2020]" to "HIV" and from "Human capital index (HCI) (scale 0-1) [HD.HCI.OVRL] - 2020 [YR2020]" to "HCI" which represents Human Capital Index in 2020.

NOTE: *The variable "Country name" is selected only for neater display of the sample data set, we will not use it in ANOVA test.*

We clean the data set by filtering away all observations with "NA" indicated in the variables "HIV" or "HCI" (or both) to make sure the sample data does not contain observations with missing value.

1.3 Grouping of observations

```
# Grouping
data_1 <- data_1%>%
  mutate(HCI_level = case_when((0.7 <= HCI & HCI<= 1) ~ 'High',
                               (0.5 <= HCI & HCI<0.7) ~ 'Middle',
                                (HCI < 0.5) ~ 'Low'))
```

By creating a new variable "HCI_level", we can group the observations (i.e. countries) according to their Human Capital Index ("HCI"). The variable "HCI_level" consists of three levels: "High", "Middle" and "Low".

- If the HCI of a country is between 1(inclusive) and 0.7 (inclusive), we classify this country under "High".
- If the HCI of a country is between 0.7(exclusive) and 0.5(inclusive), we classify this country under "Middle".
- Otherwise, the country is classified under "Low".

Here, the variable "HCI_level" is an ordinal variable because there is a clear ordering of its elements.

1.4 Re-ordering of variables for better visualization

```
# Re-ordering
data_1$HCI_level <- ordered(data_1$HCI_level,
                            levels = c('High', 'Middle', 'Low'))
```

In the original sample data set, the observations are not automatically arranged in the correct order according to their human capital availability level (constructed by the variable "HCI_level"). For tidier visualization of the data set, we re-order them so that the observations will follow the order: high human capital availability level, middle human capital availability level, low human capital availability level.

2. Data summary

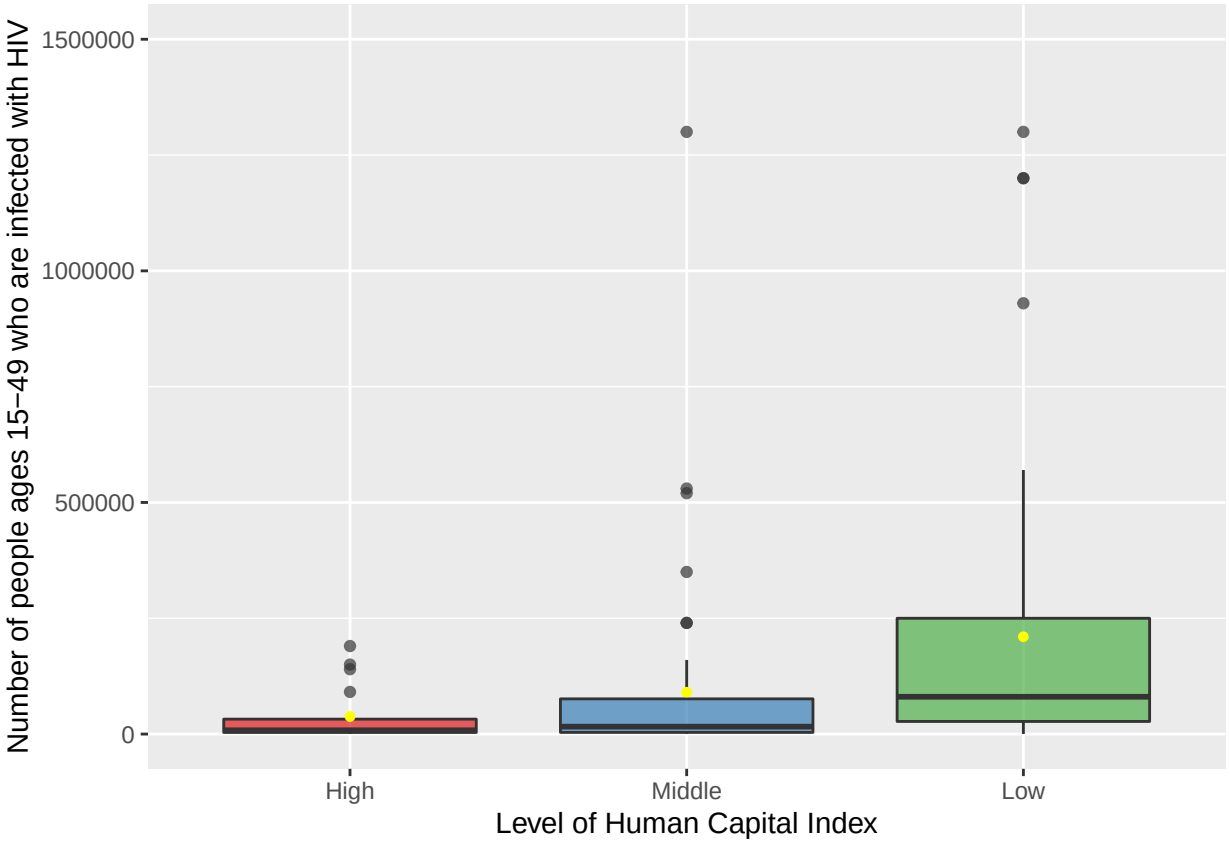
2.1 Table Summary

Group	Number of observations	Mean*
Countries with high human capital availability	20	37811.00
Countries with middle human capital availability	49	90208.37
Countries with low human capital availability	53	394401.89

*Mean represents the mean of number of people ages 15-49 who are infected with HIV/AIDS of the group.

Among all three groups, countries with low human capital availability has the highest group mean while countries with high human capital availability has the lowest.

2.2 Visualization (boxplot)



- This is a boxplot representing the distribution of the number of people ages 15-49 who are infected with HIV/AIDS in each group level.
- The yellow dots in each boxes represent the values of group mean.

NOTE: Three observations are removed from the boxplot for better visualization. All three observations are from the group of 'Countries with low human capital availability' and their "HIV" value are all above 1500 thousands (the maximum limit set in this boxplot).

Methods

To test the null hypothesis H_0, we will use ANOVA test. Below are 2 parameters which will be used in ANOVA test.

1. Test statistic of ANOVA

The test statistic for ANOVA is denoted as F. For an independent variable with k groups, the F statistic evaluates whether the group means are significantly different. Below is a table explaining the components of F *(Bevans, 2022)*:

	Sum of Squares	df (Degree of freedom)	Mean square
Group	SSR (regression sum of squares)	df_r (model degrees of freedom)	MSR (regression mean square)
Error	SSE (error sum of squares)	df_e (error degrees of freedom)	MSE (mean square error)

where

- k = the total number of groups (levels of the factor variable)
- n = the total number of observations
- $df_r = k - 1$
- $df_e = n - k$
- $MSR = \frac{SSR}{df_r}$
- $MSE = \frac{SSE}{df_e}$

F statistic itself can be computed with the following formula:

$$F = \frac{MSR}{MSE}$$

2. P-value

P-value is the probability of getting a result at least as extreme as the one that was actually observed, given that the null hypothesis is true. In ANOVA, we will measure the P-value of F-statistic which represents the probability of how likely our F-statistic (calculated from the sample data set) will occur under the condition that there is no difference between the means of number of people ages 15-49 who are infected with HIV of the 3 groups of countries (classified by Human Capital Index).

Assumptions

We assume that all observations (i.e. countries) are independent of each other, the distributions of means* in each group are normal and all groups have the same variance.

*Mean represents the mean of the number of people ages 15-49 who are infected with HIV/AIDS of the group.

Results

Using the r code function `aov()`, we can obtain the F statistic and its components easily.

```
# Compute the analysis of variance
data.aov <- aov(HIV ~ HCI_level, data_1)
# Summary of the analysis
summary(data.aov)
```

	Sum of Squares	df (Degree of freedom)	Mean square	F statistic	p-value
Group	3097 billion	2	1548 billion	3.235	0.0428
Error	56950 billion	119	478.6 billion		

Here, we have the F statistic as 3.235 and the P-value of F statistic as 0.0428.

Since the p-value is lower than the significant level ($\alpha = 0.05$), there is strong evidence against the null hypothesis that all group means of number of people ages 15-49 who are infected with HIV are equal. Therefore, we can reject the null hypothesis and conclude that countries with different human capital availability have different HIV/AIDS prevalence.

Conclusions

We started by stating the null hypothesis as countries with different prevalence of HIV/AIDS have the same human capital availability.

We first divided countries into three groups according to their human capital availability measured by Human Capital Index. Then, we compared the groups' prevalence of HIV/AIDS by comparing the group means of number of people ages 15-49 who are infected with HIV/AIDS. Next, we perform ANOVA test and obtain a P-value of 0.0428 which is statistically significant. Thereby, we can conclude that means of number of people ages 15-49 who are infected with HIV among countries with different level of Human Capital Index are the NOT same.

As a result, we rejected the null hypothesis and can accept our research hypothesis that countries with different prevalence of HIV/AIDS have different human capital availability.

Weaknesses

In our sample data set, the number of observations(i.e. number of countries) in each group(i.e. level of Human Capital Index) are not the same. In particular, the group "Countries with high HCI" consists only 20 countries, which is significantly lower than that of the group "Countries with low HCI" with size of 53. Ideally, all three groups should have the same sample size. This is because unequal sample size will lead to:

1. Reduced statistical power *(Zach, 2021)*

- Statistical power refers to the probability that a test will detect some effect when there actually is one. When we use ANOVA to compare groups, the statistical power of the test is highest when each group has an equal sample size.

2. Reduced Robustness to unequal variance *(Zach, 2021)*

- One of the assumption made is the variance between each group is equal. In general, ANOVA is considered to be robust against violations of the equal variances assumption, but only if each group has the same sample size. Therefore, unequal sample size will lower ANOVA's power in opposing unequal variances. Hence, the effect of unequal variances in the sample data set may yield to lower accuracy of the result of ANOVA.

Next Steps

In ANOVA test, a significant p-value indicates that some of the group means are different, but we don't know exactly which pair(s) of groups are different. In this study, since the ANOVA test is significant, we can use Tukey multiple pairwise-comparisons to determine if the mean difference between specific pairs of group are statistically significant *(Soetewey, 2020)*.

Discussion

We concluded that countries with different prevalence of HIV/AIDS have different human capital availability. Broadly speaking, we may further deduce that there is a relationship between prevalence of HIV/AIDS and human capital availability in a country. Since HIV/AIDS is a chronic, potentially life-threatening, sexually transmitted infectious disease, it is important to educate the public, especially the next generation about how the prevention in order to lower its prevalence. However, high prevalence of HIV/AIDS often contributes to low human capital availability (this has been proven in the research conducted in South Africa but the general global trend is yet to be proven in future research). In addition, low human capital availability implies that there are limited resources such as education and health care for the next generation and, as a result, a vicious cycle is formed. Future studies should aim at discovering solutions to generate more human capital and resources amid the HIV/AIDS situation in order to break the cycle.

Bibliography

Bell, C., Devarajan, S.,; Gersbach, H. (2006). The long-run economic costs of AIDS: A model with an application to South Africa. The World Bank Economic Review, 20(1), 55–89. https://doi.org/10.1093/wber/lhj006

Bevans, R. (2022, July 9). ANOVA in R: A step-by-step guide. Scribbr. Retrieved October 17, 2022, from https://www.scribbr.com/statistics/anova-in-r/

Ganti, A. (2022, September 20). Degrees of freedom in statistics explained: Formula and example. Investopedia. Retrieved October 17, 2022, from https://www.investopedia.com/terms/d/degrees-of-freedom.asp

Kumar, A. (2022, February 16). One-way ANOVA test: Concepts, Formula & Examples. Data Analytics. Retrieved October 17, 2022, from https://vitalflux.com/one-way-anova-test-concepts-formula-examples/#Real-world_examples_of_One-way_ANOVA_test

Kumar, A. (2022). Two-sample T-test. JMP. Retrieved October 17, 2022, from https://www.jmp.com/en_ca/statistics-knowledge-portal/t-test/two-sample-t-test.html

Lucero, J., Sánchez, S., Hernández, D., Villa, G., & Robles, A. (2021, March 26). ANOVA to compare three methods to track covid-19 in nine countries. Revista mexicana de ingeniería biomédica. Retrieved October 29, 2022, from https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0188-95322021000100103

Ostertagova, E., ; Ostertag, O. (2013, November). Methodology and application of one-way ANOVA - researchgate. Retrieved October 17, 2022, from https://www.researchgate.net/publication/259291691_Methodology_and_Application_of_One-way_ANOVA

Soetewey , A. (2020, October 12). ANOVA in R. Stats and R. Retrieved October 17, 2022, from https://statsandr.com/blog/anova-in-r/#introduction

Two-sample T-test. JMP. (n.d.). Retrieved October 17, 2022, from https://www.jmp.com/en_ca/statistics-knowledge-portal/t-test/two-sample-t-test.html

Zach. (2021, September 18). How to perform an ANOVA with unequal sample sizes. Statology. Retrieved October 17, 2022, from https://www.statology.org/anova-unequal-sample-size/

Appendix

A1: Ethics Statement

The open data set "Health Nutrition and Population Statistics" contributed by World Bank is used in the report. Open data is ethical because the data is usually published alongside with publications to verify findings. They are easily accessible to others and can be reused wherever possible. Hence, open data creates opportunities for new research and enhances the research process's transparency and replication.

A2: Supplementary Materials

Below is the r-code to produce the boxplot in *section 2.2 Visualization (boxplot)*.

```
# Box plot
plot <- ggplot(data_1, aes(x=HCI_level, y=HIV, fill=HCI_level)) +
    geom_boxplot(alpha=0.7) +
  xlab('Level of Human Capital Index')+
   ylab('Number of people ages 15-49 who are infected with HIV')+
    stat_summary(fun=mean, geom="point", shape=20, size=2, color="yellow", fill="red") +
    theme(legend.position="none") +
    scale_fill_brewer(palette="Set1")

plot + ylim(0, 1500000)
```