A study on Survival to age 65 (% of cohort) based on Gender

Assignment 2

Deepthi Suresh S3991481

Last updated: 20 October, 2023

RPubs link information

https://rpubs.com/Deepthi215/1101816

Introduction

Over the years, the Life expectancy continues to increase, which can be attributed to different factors such as health care access, income, GDP etc.

The study is conducted on the Survival to the age of 65 factor across different Countries in the world, based on the Gender .Survival to age 65 refers to the percentage of a cohort of newborn infants that would survive to age 65, if subject to age specific mortality rates of the specified year.

The study also conducts an exploratory analysis on the income group of the countries and the average Survival rate, along with the trend of the rate for males and females over the years.

Problem Statement

The study aims to find if there is a statistically significant difference between average % of the cohort surviving till the age of 65 among males and females.

Descriptive statistics are also performed to identify trends between the survival rates for males and females over the years(1960-2021), as well comparison between the countries income group and the survival rate by plotting relevant visualizations.

Various statistical functions like Two-Sample t-test (hypothesis testing), QQ plots and Box plots (for identifying outliers), trend analysis and graphs (for descriptive statistics and visualizations), LeveneTest (for checking Homogeneity of variances), Welch two-sample t-test etc will be used in this study to come to a conclusion.

Data

A total of 2 datasets and one meta-data file has been used for this analysis. The dataset is collected from the World Bank website with the following links: https://data.worldbank.org/indicator/SP.DYN.TO65.MA.ZS; https://data.worldbank.org/indicator/SP.DYN.TO65.FE.ZS.

```
Survival_female <- read_excel("/Users/deepthisuresh/Library/CloudStorage/OneDrive-RMITUniversity/Applied Analytics/assignment
        2/data/female.xls",skip = 3)
Survival_male <- read_excel("/Users/deepthisuresh/Library/CloudStorage/OneDrive-RMITUniversity/Applied Analytics/assignment
        2/data/male.xls",skip = 3)
Country_meta <- read_csv("/Users/deepthisuresh/Library/CloudStorage/OneDrive-RMITUniversity/Applied Analytics/assignment
        2/data/Metadata_Country.csv")
```

The dataset is in an untidy form and requires pre-processing and merging

```
Survival_female1<- Survival_female %>% pivot_longer(names_to = "Year", values_to = "% of cohort",cols = 5:67)
Survival_male1 <- Survival_male %>% pivot_longer(names_to = "Year", values_to = "% of cohort",cols = 5:67)
Survival <- rbind(Survival_female1,Survival_male1)
head(Survival,n=2)
```

Country Name <chr>	**Country Code** <chr>	**Indicator Name** <chr>	**Ind** <ch
Aruba	ABW	Survival to age 65, female (% of cohort)	SP.D
Aruba	ABW	Survival to age 65, female (% of cohort)	SP.D

2 rows | 1-5 of 6 columns

Data Cont.

The data also required pre-processing where some of the variables had to be converted to factor variables such as country code,Indicator name (Male and Female) and country name using as.factor() and level(), to aid the merging of the country meta-dataset.The dataset was merged using Left join based on country code into “Merged” dataset

```
Survival$`Country Code` <- as.factor(Survival$`Country Code`)
Survival$`Indicator Name`<- factor(Survival$`Indicator Name`,levels= c("Survival to age 65, male (% of cohort)","Survival to age 65,
        female (% of cohort)"),labels=c("Male","Female"),ordered = TRUE)
levels(Survival$`Indicator Name`)
```

```
## [1] "Male"   "Female"
```

```
Survival$`Country Name` <- as.factor(Survival$`Country Name`)
Country_meta$`IncomeGroup` <- as.factor(Country_meta$`IncomeGroup`)
Country_meta$`Region` <- as.factor(Country_meta$`Region`)
Country_meta$`Country Code` <- as.factor(Country_meta$`Country Code`)
#Merging data with country meta data to get income group of the countries
Merged <- left_join(Survival,Country_meta, by = c ("Country Code"))
str(Merged)
```

```
## tibble [33,516 × 11] (S3: tbl_df/tbl/data.frame)
##  $ Country Name  : Factor w/ 266 levels "Afghanistan",..: 13 13 13 13 13 13 13 13 13 13 ...
##  $ Country Code  : Factor w/ 266 levels "ABW","AFE","AFG",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Indicator Name: Ord.factor w/ 2 levels "Male"<"Female": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Indicator Code: chr [1:33516] "SP.DYN.TO65.FE.ZS" "SP.DYN.TO65.FE.ZS" "SP.DYN.TO65.FE.ZS"
"SP.DYN.TO65.FE.ZS" ...
##  $ Year          : chr [1:33516] "1960" "1961" "1962" "1963" ...
##  $ % of cohort   : num [1:33516] 71.5 72.3 72.8 73.5 73.8 ...
##  $ Region        : Factor w/ 7 levels "East Asia & Pacific",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ IncomeGroup   : Factor w/ 4 levels "High income",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ SpecialNotes  : chr [1:33516] NA NA NA NA ...
##  $ TableName     : chr [1:33516] "Aruba" "Aruba" "Aruba" "Aruba" ...
##  $ ...6          : logi [1:33516] NA NA NA NA NA NA ...
```

Data Cont.

The Merged dataset has some unused columns such as Indicator code,special notes,table names,which has been removed by sub setting. Indicator Name is renamed as Gender for readability.

```
#subsetting data to remove unwanted columns
Merged1 <- Merged %>% select ("Country Name","Country Code","Region","IncomeGroup" ,"Indicator Name","Year","% of cohort")
Merged1 <- Merged1 %>% rename(Gender =`Indicator Name`)
head(Merged1, n=3)
```

Country Name <fct>	**Country Code** <fct>	**Region** <fct>	**IncomeGr...** <fct>	**Gen...** <ord>
Aruba	ABW	Latin America & Caribbean	High income	Female
Aruba	ABW	Latin America & Caribbean	High income	Female
Aruba	ABW	Latin America & Caribbean	High income	Female

3 rows

Data Variables

The final “Merged1” dataset has the following key variables:

Country Name - Name of the country

Country code - abbreviation of the country

Income group - The country have been grouped into high income,low income, upper and lower middle income categories

Gender - Male and Female data

Year - data pertaining to 1960-2022

% of cohort - % of infants that survives to the age of 65

Descriptive Statistics - Handling Missing Values

The Merged data needs to be checked and corrected for missing values and outliers. 2022 data and data for non-classified countries is unavailable, so we can exclude them by using complete.cases(). Region and Income groups are missing for aggregates and hence can be excluded (we are only considering countries and not aggregates for this study).

```
colSums(is.na(Merged1))
```

```
## Country Name Country Code       Region  IncomeGroup       Gender         Year 
##            0            0         6174         6300            0            0 
##  % of cohort 
##          656
```

```
Merged1 <- Merged1[complete.cases(Merged1), ]
colSums(is.na(Merged1))
```

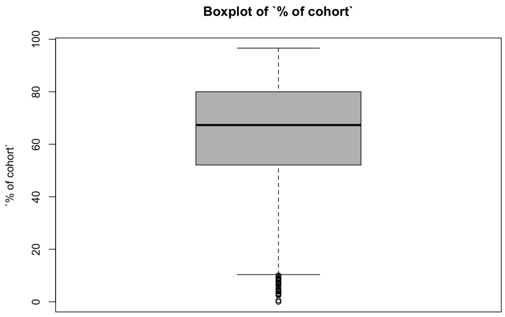
```
## Country Name Country Code       Region  IncomeGroup       Gender         Year 
##            0            0            0            0            0            0 
##  % of cohort 
##            0
```

Descriptive Statistics Cont.- handling Outliers

Identifying and correcting outliers by using the box plot and z-scores function. the length() function shows that there are 48 outliers

```
Merged1$`% of cohort` %>%  boxplot(main="Boxplot of `% of cohort`", ylab="`% of cohort`", col = "grey")
z.scores <- Merged1$`% of cohort` %>%  scores(type = "z")
length (which(abs(z.scores) >3 ))
```

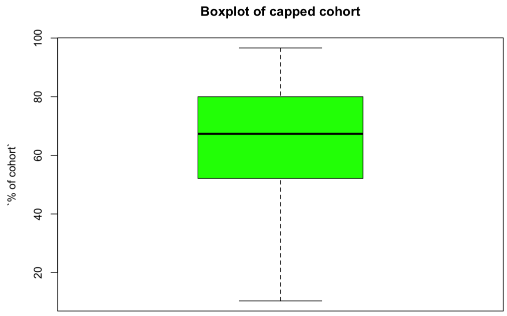
```
## [1] 48
```



Descriptive Statistics Cont.

In order to handle the outliers a capping function is created, that caps the values outside the limits. The new capped variable is then added to the Merged1 dataset for further analysis.

```
test <- Merged1 %>% select(`% of cohort`)
cap <- function(x) {
  quantiles <- quantile(x, c(0.05, 0.25, 0.75, 0.95))
  x[x < quantiles[2] - 1.5 * IQR(x)] <- quantiles[1]
  x[x > quantiles[3] + 1.5 * IQR(x)] <- quantiles[4]
  (x)
}
test <- sapply(test, FUN = cap)
Merged1 <- Merged1 %>% mutate(.,cappedcohort = sapply(test, FUN = cap))
Merged1$`cappedcohort` %>%  boxplot(main="Boxplot of capped cohort", ylab="`% of cohort`", col = "green")
```

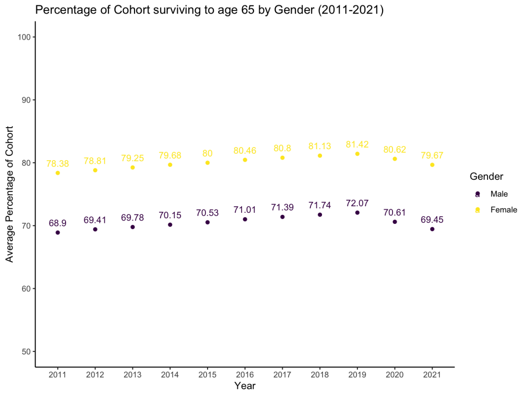


Descriptive Statistics Cont.

The point graph shows the mean survival rate over the last 10 years for male and females. The yellow points represents the female and the black points represents the male. It can be seen that,over the years females have a higher average survival rate when compared to males.

```
filtered_year <- Merged1 %>%
  group_by(Gender,Year)%>%
  filter(Year >= 2011 & Year <= 2021) %>%
  summarise(n = n(),mean_cohort = round(mean(cappedcohort, na.rm = TRUE),2))

ggplot(data = filtered_year,
       aes(x = Year, y = mean_cohort, color = Gender)) +
  geom_point() +labs(title = "Percentage of Cohort surviving to age 65 by Gender (2011–2021)",x = "Year",y = "Average Percentage of
        Cohort") +coord_cartesian(ylim = c(50, 100))+ geom_text(aes(label = mean_cohort),vjust = -1, size = 3.5) + theme_classic()
```

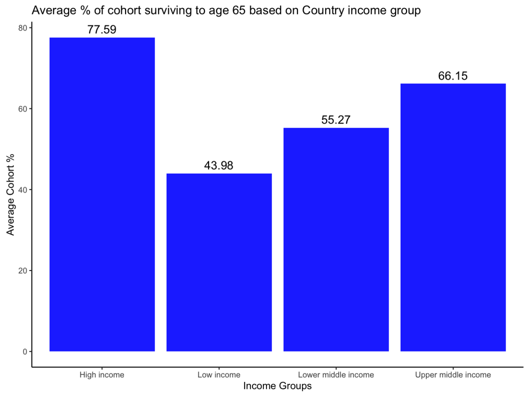


Descriptive Statistics Cont.

The bar chart represents the average survival rate across countries under different income group.It can be noticed that high income countries have the highest survival rate of 77.59%, whereas the lowest survival rate of 43.98% is in the low income countries.

```r
cohort <- Merged1 %>%
    group_by(IncomeGroup) %>% # group by categorical variable
    summarise(n = n(),mean = round(mean(cappedcohort, na.rm = TRUE),2)) %>% # count number of observations in each group
    mutate(prop = round(n / sum(n),2))


ggplot(data = cohort, # specify data
       aes(x =IncomeGroup , y = mean)) + # specify x and y variables
  geom_col(fill="blue", alpha=0.9) + # specify type of plot (geom)+
  labs(x = "Income Groups", y = "Average Cohort %", # add axis labels
  title = "Average % of cohort surviving to age 65 based on Country income group") +geom_text(aes(label = mean), # add frequencies to bars
            vjust = -0.5,size = 4.5) +theme_classic()
```



Hypothesis Testing

H0:The average percentage of cohorts that can survive till 65 among males and females are equal

H1:The average percentage of cohorts that can survive till 65 among males and females are not equal

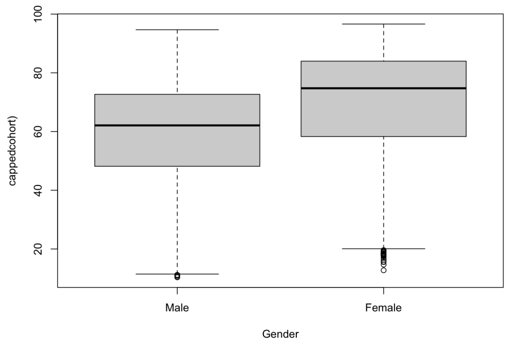
```
Merged1 %>% group_by(Gender) %>% summarise(Min = min(cappedcohort,na.rm = TRUE),
                                             Q1 = quantile(cappedcohort,probs = .25,na.rm = TRUE),
                                             Median = median(cappedcohort, na.rm = TRUE),
                                             Q3 = quantile(cappedcohort,probs = .75,na.rm = TRUE),
                                             Max = max(cappedcohort,na.rm = TRUE),
                                             Mean = mean(cappedcohort, na.rm = TRUE),
                                             SD = sd(cappedcohort, na.rm = TRUE),
                                             n = n(),
                                             Missing = sum(is.na(cappedcohort)))
```

Gen...	Min	Q1	Median	Q3	Max	Mean	SD
<ord>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Male	10.35209	48.15872	62.09503	72.67790	94.68853	60.09081	16.63218
Female	12.73489	58.30608	74.74541	83.98292	96.61929	70.12068	17.39383

2 rows

Hypothesis Testing Cont.

```
Merged1 %>% boxplot(cappedcohort ~ Gender, data = ., ylab = "cappedcohort)")
```



The box plot shows that females have a higher average survival rate than males.The two-sample t-test will help us consider whether this difference is statistically significant.

Hypothesis Testing- Assumptions

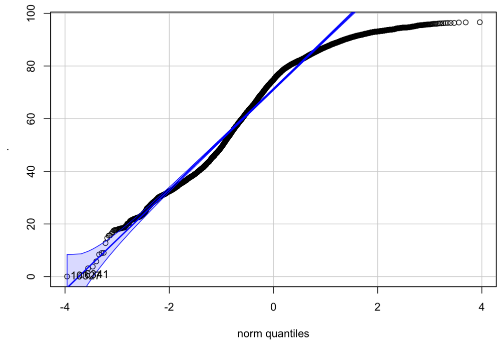
$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

But before conducting the two-sample t-test we need to check the assumptions of normality and variance homogeneity through QQ plots and other statistical tests

```
#checking QQ plot for Gender- female
Merged1_female <- Merged1 %>% filter(Gender == "Female")
Merged1_female$`% of cohort` %>% qqPlot(dist="norm")
```

```
## [1] 10327  6341
```

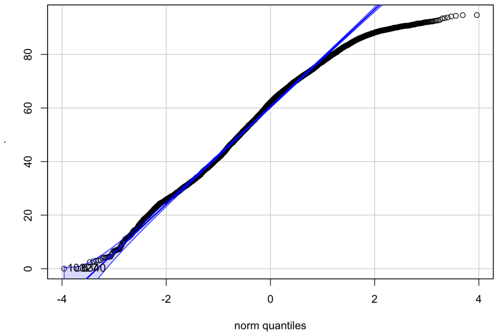


Hypothesis Testing - Assumptions

For both the groups, the data seems to be skewed, however according to CLT, since the sample size in both the groups is more than 30 therefore sampling distribution will be approximately a normal distribution.

```
#checking QQ plot for Gender - Male
Merged1_male <- Merged1 %>% filter(Gender == "Male")
Merged1_male$`% of cohort` %>% qqPlot(dist="norm")
```

```
## [1] 10327  6340
```



Hypothesis Testing- Homogeneity of variances

```
leveneTest( cappedcohort ~ Gender, data = Merged1)
```

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	1	14.71006	0.0001256627
	26782	NA	NA

2 rows

Th p-value for the LeveneTest of equal variance for % of cohort and gender was p = 0.0001, which is less that 0.05, and hence we can reject the null hypothesis that it is not safe to assume equal variance

Hypothesis Testing Cont.

Here we perform the Welch two-sample t-test. The level of confidence used for hypothesis testing is 0.95

```
t.test(
  cappedcohort ~ Gender,
  data = Merged1,
  var.equal = FALSE,
  alternative = "two.sided"
)
```

```
##
##  Welch Two Sample t-test
##
## data:  cappedcohort by Gender
## t = -48.23, df = 26728, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Male and group Female is not equal to 0
## 95 percent confidence interval:
##  -10.43749  -9.62226
## sample estimates:
##   mean in group Male mean in group Female
##             60.09081             70.12068
```

Our decision should be to reject $H0: \mu1 = \mu1$ as the $p < .05$ and the 95% CI of the estimated population difference [-10.437, -9.622], which did not capture $H0: \mu1 - \mu1 = 0$. The results of the two-sample t-test were therefore statistically significant. This meant that the mean survival rates for males and females was significantly different.

Hypothesis Testing Assumptions and Interpretation.

A two-sample t-test was used to test for a significant difference between the mean survival rate of males and females. The QQ plot showed non-normality for both males as well as females, however according to the central limit theorem, the t-test can be applied as the sample size is greater than 30 in each group (according to CLT).

The LeveneTest of homogeneity of variance indicated that equal variance could not be assumed.

The results of the two-sample t-test assuming unequal variance found a statistically significant difference between the mean survival rates of males and females, t(df=26728)=−48.23, p=<2.2e-16, 95% CI for the difference in means [-10.438 -9.622]. The results of the investigation suggest that males and females have significantly different average survival rates.

Discussion

-The results of the two-sample t-test assuming unequal variance found a statistically significant difference between the mean survival rates of males and females

-The average survival rate is the highest(77.59%) for high income countries and lowest (43.98%)for low income countries.

- A strength was the number of data points available, the larger the sample size better will be our generalizations of the population

-A limitation was that there were missing values and outliers which could have lead to loss of good information or trends

-Further investigations can be conducted on country wise difference in the survival rate based on economics factors such as income,GDP, health care investments etc.

In conclusion, there was a significant difference between the female and male mean survival rate, females is seen to have a higher survival rate over the years. Further analysis is required to find the reasons behind this.

References

Data. (2022). Retrieved from World Bank, accessed on 20 Oct 2023 https://data.worldbank.org/indicator/SP.DYN.TO65.FE.ZS

Data. (2022). Retrieved from World bank, accessed on 20 Oct 2023 https://data.worldbank.org/indicator/SP.DYN.TO65.MA.ZS

Laleh Tafakori,Applied Analytics. (2023).RMIT University,Melbourne Retrieved from Canvas: https://astral-theory-157510.appspot.com/secured/MATH1324_Module_07.html#Testing_the_Assumption_of_Normality

Wickham H, Bryan J (2023). _readxl: Read Excel Files_. R package version 1.4.3, https://CRAN.R-project.org/package=readxl.

Wickham H, Vaughan D, Girlich M (2023). _tidyr: Tidy Messy Data_. R package version 1.3.0, https://CRAN.R-project.org/package=tidyr.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016