Objective 1

DANA4820

2025-03-29

# Load required libraries  
library(dplyr)

## Warning: package 'dplyr' was built under R version 4.4.2

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.2

library(binom)

## Warning: package 'binom' was built under R version 4.4.2

library(exactci)

## Warning: package 'exactci' was built under R version 4.4.2

## Loading required package: ssanv

## Loading required package: testthat

## Warning: package 'testthat' was built under R version 4.4.2

##   
## Attaching package: 'testthat'

## The following object is masked from 'package:dplyr':  
##   
## matches

##   
## Attaching package: 'exactci'

## The following object is masked from 'package:binom':  
##   
## binom.exact

library(PropCIs)  
library(epitools)

##   
## Attaching package: 'epitools'

## The following object is masked from 'package:exactci':  
##   
## binom.exact

## The following objects are masked from 'package:binom':  
##   
## binom.exact, binom.wilson

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(car)

## Warning: package 'car' was built under R version 4.4.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.4.2

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(tidyr)

## Warning: package 'tidyr' was built under R version 4.4.2

##   
## Attaching package: 'tidyr'

## The following object is masked from 'package:testthat':  
##   
## matches

# Read the dataset  
df <- read.csv('global\_health.csv')  
str(df)

## 'data.frame': 1880 obs. of 29 variables:  
## $ Country : chr "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...  
## $ Country\_Code : chr "AFG" "AFG" "AFG" "AFG" ...  
## $ Year : int 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 ...  
## $ Fertility\_Rate : num 5.83 5.7 5.56 5.41 5.26 ...  
## $ Urban\_Population\_Percent : num 24.2 24.4 24.6 24.8 25 ...  
## $ Total\_Population : num 30466479 31541209 32716210 33753499 34636207 ...  
## $ Water\_Access\_Percent : num 21.1 22 22.9 23.9 24.8 ...  
## $ Unemployment\_Rate : num 7.91 7.92 7.92 9.01 10.1 ...  
## $ Sanitary\_Expense\_Per\_GDP : num 7.9 8.81 9.53 10.11 11.82 ...  
## $ Life\_Expectancy : num 61.9 62.4 62.5 62.7 63.1 ...  
## $ Life\_Expectancy\_Female : num 63.5 64 64.3 64.6 65.1 ...  
## $ Life\_Expectancy\_Male : num 60.3 60.8 60.8 60.8 61.2 ...  
## $ Infant\_Deaths : num 70669 70906 70288 70068 69144 ...  
## $ GDP\_Per\_Capita : num 653 639 627 567 523 ...  
## $ Hospital\_Beds\_Per\_1000 : num 0.47 0.46 0.46 0.44 0.45 0.42 0.4 0.38 0.34 0.36 ...  
## $ Female\_Population : num 15067373 15594637 16172321 16682054 17115346 ...  
## $ Male\_Population : num 15399105 15946572 16543889 17071446 17520861 ...  
## $ Alcohol\_Consumption\_Per\_Capita: num 0.004 0.004 0.004 0.009 0.013 0.017 0.016 0.016 0.011 NA ...  
## $ Immunization\_Rate : num 67 64 62 64 66 66 72 72 70 66 ...  
## $ Sanitary\_Expense\_Per\_Capita : num 52.6 56.3 60.2 60.1 61.5 ...  
## $ CO2\_Exposure\_Percent : num 70.9 73.1 77.1 73.5 72.8 ...  
## $ Air\_Pollution : num 70.9 73.1 77.1 73.5 72.8 ...  
## $ Labour\_Force\_Total : num 7520865 7881567 8285362 8630724 8913938 ...  
## $ Tuberculosis\_Per\_100000 : num 189 189 189 189 189 189 189 189 183 185 ...  
## $ Suicide\_Rate\_Percent : num 3.68 3.66 3.6 3.57 3.61 3.44 3.52 3.61 NA NA ...  
## $ Obesity\_Rate\_Percent : num 10.7 11.6 10.4 11.2 12 ...  
## $ Underweight\_Rate\_Percent : num 10.15 10.79 10.17 10.52 7.88 ...  
## $ Overweight\_Rate\_Percent : num 31.6 32.7 34 35.2 36.5 ...  
## $ Safe\_Water\_Access\_Percent : num 46.7 49.5 52.2 55.1 58 ...

View(df)

# Clean and handle missing values  
colnames(df) <- trimws(colnames(df))  
df\_clean <- df %>%   
 filter(!is.na(Life\_Expectancy\_Female) & !is.na(Life\_Expectancy\_Male)) %>%  
 mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = TRUE), .)))  
  
num\_cols <- sapply(df\_clean, is.numeric)  
df\_clean[num\_cols] <- lapply(df\_clean[num\_cols], function(x) {  
 cap\_value <- quantile(x, 0.99, na.rm = TRUE)  
 ifelse(x > cap\_value, cap\_value, x)  
})  
  
# Keep relevant variables  
df\_clean <- df\_clean %>%  
 dplyr::select(Country, Year, Life\_Expectancy, Life\_Expectancy\_Female, Life\_Expectancy\_Male,  
 GDP\_Per\_Capita, Unemployment\_Rate, Safe\_Water\_Access\_Percent, Urban\_Population\_Percent,  
 Immunization\_Rate, Air\_Pollution, Infant\_Deaths) %>%  
 group\_by(Country) %>%  
 mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = TRUE), .))) %>%  
 ungroup() %>%  
 filter(complete.cases(.))  
  
# Categorize Life Expectancy by Gender  
df\_clean <- df\_clean %>%  
 mutate(Life\_Expectancy\_Category\_Female = case\_when(  
 Life\_Expectancy\_Female < 60 ~ 'Low',  
 Life\_Expectancy\_Female >= 60 & Life\_Expectancy\_Female < 75 ~ 'Medium',  
 Life\_Expectancy\_Female >= 75 ~ 'High'  
 ),  
 Life\_Expectancy\_Category\_Male = case\_when(  
 Life\_Expectancy\_Male < 60 ~ 'Low',  
 Life\_Expectancy\_Male >= 60 & Life\_Expectancy\_Male < 75 ~ 'Medium',  
 Life\_Expectancy\_Male >= 75 ~ 'High'  
 ))

## Objective 1: Differences in Life Expectancy Between Genders  
  
# Frequency table for Female  
lifeexp\_table\_f <- table(df\_clean$Life\_Expectancy\_Category\_Female)  
lifeexp\_table\_f

##   
## High Low Medium   
## 1002 99 739

# Frequency table for Male  
lifeexp\_table\_m <- table(df\_clean$Life\_Expectancy\_Category\_Male)  
lifeexp\_table\_m

##   
## High Low Medium   
## 438 284 1118

# Count frequencies   
lifeexp\_table <- table(Gender = rep(c("Female", "Male"), each = nrow(df\_clean)),  
 Category = c(df\_clean$Life\_Expectancy\_Category\_Female, df\_clean$Life\_Expectancy\_Category\_Male))  
lifeexp\_table

## Category  
## Gender High Low Medium  
## Female 1002 99 739  
## Male 438 284 1118

# Wald Test  
wald\_test <- function(y, n, p){  
 p\_hat <- y/n  
 se <- sqrt(p\_hat \* (1 - p\_hat) / n)  
 z <- (p\_hat - p) / se  
 pvalue <- 2 \* pnorm(abs(z), lower.tail=FALSE)  
 return(c('Z-Statistic' = z, 'P-Value' = pvalue))  
}  
wald\_test(lifeexp\_table[1,3], sum(lifeexp\_table[1,]), 0.5)

## Z-Statistic P-Value   
## -8.607393e+00 7.474118e-18

# Score Test  
score\_test <- prop.test(c(lifeexp\_table[1,3], lifeexp\_table[2,3]),  
 c(sum(lifeexp\_table[1,]), sum(lifeexp\_table[2,])),  
 alternative = 'two.sided')  
score\_test

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(lifeexp\_table[1, 3], lifeexp\_table[2, 3]) out of c(sum(lifeexp\_table[1, ]), sum(lifeexp\_table[2, ]))  
## X-squared = 155.32, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.2381366 -0.1738200  
## sample estimates:  
## prop 1 prop 2   
## 0.4016304 0.6076087

# Likelihood Ratio Test  
likelihood\_ratio\_test <- function(phat, pnull, y, n) {  
 LL <- c(  
 'LRT Statistic' = 2 \* log(dbinom(x = y, prob = phat, size = n) / dbinom(x = y, prob = pnull, size = n)),  
 'P-Value' = pchisq(2 \* log(dbinom(x = y, prob = phat, size = n) / dbinom(x = y, prob = pnull, size = n)), 1, lower.tail = FALSE)  
 )  
 return(LL)  
}  
likelihood\_ratio\_test(phat = 0.5, pnull = 0.5, y = lifeexp\_table[1,3], n = sum(lifeexp\_table[1,]))

## LRT Statistic P-Value   
## 0 1

# Chi-Square Test  
chi\_test <- chisq.test(lifeexp\_table)  
chi\_test

##   
## Pearson's Chi-squared test  
##   
## data: lifeexp\_table  
## X-squared = 387.61, df = 2, p-value < 2.2e-16

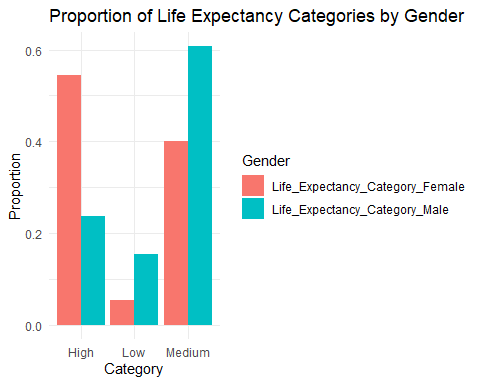
# Binomial Exact Test  
binom\_exact\_test <- binom.test(lifeexp\_table[1,3], sum(lifeexp\_table[1,]), p = 0.5, alternative = "two.sided")  
binom\_exact\_test

##   
## Exact binomial test  
##   
## data: lifeexp\_table[1, 3] and sum(lifeexp\_table[1, ])  
## number of successes = 739, number of trials = 1840, p-value < 2.2e-16  
## alternative hypothesis: true probability of success is not equal to 0.5  
## 95 percent confidence interval:  
## 0.3791312 0.4244439  
## sample estimates:  
## probability of success   
## 0.4016304

# Confidence Intervals (Wald, Wilson, Likelihood Ratio)  
wald\_ci <- binom.confint(lifeexp\_table[1,3], sum(lifeexp\_table[1,]), conf.level = 0.95, method = "asymptotic")  
wilson\_ci <- binom.confint(lifeexp\_table[1,3], sum(lifeexp\_table[1,]), conf.level = 0.95, method = "wilson")  
lrt\_ci <- binom.confint(lifeexp\_table[1,3], sum(lifeexp\_table[1,]), conf.level = 0.95, method = "lrt")  
  
list(Wald\_CI = wald\_ci, Wilson\_CI = wilson\_ci, Likelihood\_Ratio\_CI = lrt\_ci)

## $Wald\_CI  
## method x n mean lower upper  
## 1 asymptotic 739 1840 0.4016304 0.379231 0.4240299  
##   
## $Wilson\_CI  
## method x n mean lower upper  
## 1 wilson 739 1840 0.4016304 0.3794583 0.4242124  
##   
## $Likelihood\_Ratio\_CI  
## method x n mean lower upper  
## 1 lrt 739 1840 0.4016304 0.379385 0.4241535

# Proportional comparison barplot  
prop\_df <- df\_clean %>%  
 pivot\_longer(cols = c(Life\_Expectancy\_Category\_Female, Life\_Expectancy\_Category\_Male),   
 names\_to = "Gender", values\_to = "Life\_Category") %>%  
 group\_by(Gender, Life\_Category) %>%  
 summarise(Count = n(), .groups = 'drop') %>%  
 group\_by(Gender) %>%  
 mutate(Proportion = Count / sum(Count))  
  
ggplot(prop\_df, aes(x = Life\_Category, y = Proportion, fill = Gender)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Proportion of Life Expectancy Categories by Gender", x = "Category", y = "Proportion") +  
 theme\_minimal()



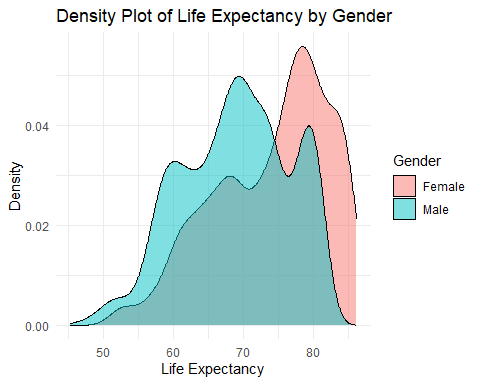
# Summary Stats Comparison  
summary\_stats <- df\_clean %>%  
 summarise(  
 Mean\_Female = mean(Life\_Expectancy\_Female),  
 Mean\_Male = mean(Life\_Expectancy\_Male),  
 Median\_Female = median(Life\_Expectancy\_Female),  
 Median\_Male = median(Life\_Expectancy\_Male),  
 SD\_Female = sd(Life\_Expectancy\_Female),  
 SD\_Male = sd(Life\_Expectancy\_Male)  
 )  
summary\_stats

## # A tibble: 1 × 6  
## Mean\_Female Mean\_Male Median\_Female Median\_Male SD\_Female SD\_Male  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 74.1 68.9 76.1 69.3 8.10 7.71

# Paired T-Test between Male and Female Life Expectancy  
t\_test <- t.test(df\_clean$Life\_Expectancy\_Female, df\_clean$Life\_Expectancy\_Male, paired = TRUE)  
t\_test

##   
## Paired t-test  
##   
## data: df\_clean$Life\_Expectancy\_Female and df\_clean$Life\_Expectancy\_Male  
## t = 111.89, df = 1839, p-value < 2.2e-16  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## 5.111561 5.293961  
## sample estimates:  
## mean difference   
## 5.202761

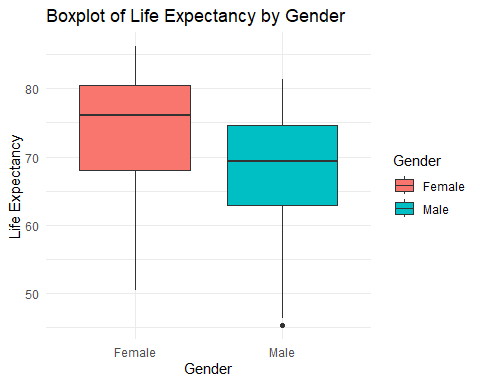
# Density plot for Male and Female  
long\_df <- df\_clean %>%  
 dplyr::select(Country, Life\_Expectancy\_Female, Life\_Expectancy\_Male) %>%  
 pivot\_longer(cols = starts\_with("Life\_Expectancy"), names\_to = "Gender", values\_to = "Life\_Expectancy")  
  
long\_df$Gender <- ifelse(grepl("Female", long\_df$Gender), "Female", "Male")  
  
ggplot(long\_df, aes(x = Life\_Expectancy, fill = Gender)) +  
 geom\_density(alpha = 0.5) +  
 labs(title = "Density Plot of Life Expectancy by Gender", x = "Life Expectancy", y = "Density") +  
 theme\_minimal()



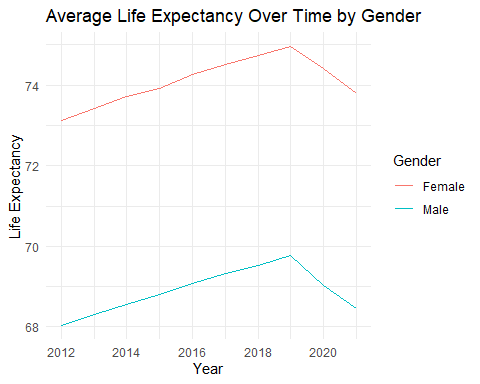
# ANOVA to compare Life Expectancy by Gender using long format  
anova\_df <- df\_clean %>%  
 dplyr::select(Country, Life\_Expectancy\_Female, Life\_Expectancy\_Male) %>%  
 pivot\_longer(cols = c(Life\_Expectancy\_Female, Life\_Expectancy\_Male),  
 names\_to = "Gender", values\_to = "Life\_Expectancy",  
 names\_transform = list(Gender = ~ifelse(. == "Life\_Expectancy\_Female", "Female", "Male")))  
  
# Run one-way ANOVA  
anova\_model <- aov(Life\_Expectancy ~ Gender, data = anova\_df)  
summary(anova\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Gender 1 24903 24903 398.3 <2e-16 \*\*\*  
## Residuals 3678 229952 63   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 1. Side-by-Side Boxplots of Life Expectancy by Gender  
ggplot(long\_df, aes(x = Gender, y = Life\_Expectancy, fill = Gender)) +  
 geom\_boxplot() +  
 labs(title = "Boxplot of Life Expectancy by Gender", x = "Gender", y = "Life Expectancy") +  
 theme\_minimal()



# 2. Line Plot of Life Expectancy Over Time by Gender  
df\_long\_time <- df\_clean %>%  
 dplyr::select(Year, Country, Life\_Expectancy\_Female, Life\_Expectancy\_Male) %>%  
 pivot\_longer(cols = starts\_with("Life\_Expectancy\_"), names\_to = "Gender", values\_to = "Life\_Expectancy") %>%  
 mutate(Gender = ifelse(Gender == "Life\_Expectancy\_Female", "Female", "Male"))  
  
ggplot(df\_long\_time, aes(x = Year, y = Life\_Expectancy, color = Gender)) +  
 geom\_line(stat = "summary", fun = mean) +  
 labs(title = "Average Life Expectancy Over Time by Gender", x = "Year", y = "Life Expectancy") +  
 theme\_minimal()



# 3. Heatmap of Life Expectancy by Country and Gender  
gap\_df <- df\_clean %>%  
 mutate(Gender\_Gap = Life\_Expectancy\_Female - Life\_Expectancy\_Male)  
  
# 🔍 Improved Gender Gap Plot (Top 20 countries with largest gap)  
top\_gap\_df <- gap\_df %>%  
 mutate(Gender\_Gap = Life\_Expectancy\_Female - Life\_Expectancy\_Male) %>%  
 group\_by(Country) %>%  
 summarise(Avg\_Gap = mean(Gender\_Gap, na.rm = TRUE)) %>%  
 arrange(desc(Avg\_Gap)) %>%  
 slice(1:20)  
  
ggplot(top\_gap\_df, aes(x = reorder(Country, Avg\_Gap), y = Avg\_Gap)) +  
 geom\_col(fill = "salmon") +  
 coord\_flip() +  
 labs(title = "Top 20 Countries with Highest Gender Gap in Life Expectancy",  
 subtitle = "Positive values mean females live longer than males",  
 x = "Country", y = "Life Expectancy Difference (Female - Male)") +  
 theme\_minimal(base\_size = 12)

