Milestone 2

Angel Waters

2022-06-19

# Introduction

Milestone 1 was used to explore the different variables in the SWAN dataset that was subseted for the purpose of explaratory data analysis. In this report, multiple hypotheses will be tested to understand the relationships between the different groups depicted in the data subset. This dataset is used to assessed women at a crucial lifestage to properly provide health services and support for women in the 40’s and 50’s age group (Sutton-Tyrell et al. 1997).

## Data Cleaning

Additional variables were added to the Milestone 1 subset to capture support the women interviewed felt they received.

rawData <-   
 read\_csv("SWANBaselineData\_ProfessorKSubset (1).csv")

## New names:  
## Rows: 3302 Columns: 33  
## -- Column specification  
## -------------------------------------------------------- Delimiter: "," chr  
## (18): HBCHOLE0, MIGRAIN0, ANEMIA0, LISTEN0, TAKETOM0, CONFIDE0, HELPSIC0... dbl  
## (15): ...1, SWANID, AGE0, HSWRKHR0, HOSPSTA0, PULSE0, SYSBP10, DIABP10, ...  
## 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`

milestone2\_subset <- subset(rawData, select = c(  
 SWANID,  
 AGE0,  
 ANEMIA0,  
 LISTEN0,  
 TAKETOM0,  
 CONFIDE0,  
 HELPSIC0,  
 SMOKERE0,  
 PULSE0,  
 HEIGHT0,  
 WEIGHT0,  
 RACE)  
)

Data was cleaned and additional columns were formulated. Minority data was used to separate races that aren’t as frequent as others, by taking all races below 20% (one fifth of the data because there are 5 races) and assigning them as a subdivision minority. Support Scores were calculated by updating each support column to a numeric scale and adding them together. The support score scale goes from 0 support to a score of 20 which means they feel the maximum support they could feel. The average support score was also calculated.

## # A tibble: 6 x 15  
## SWANID AGE0 ANEMIA0 LISTEN0 TAKETOM0 CONFIDE0 HELPSIC0 SMOKERE0 PULSE0  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
## 1 10005 48 No 5 5 5 1 No 36  
## 2 10046 52 No 5 5 5 5 Yes 38  
## 3 10056 51 Yes 4 4 4 5 No 36  
## 4 10092 45 Yes 5 5 5 5 Yes 32  
## 5 10126 48 Yes 5 5 5 5 No 40  
## 6 10153 51 No 5 5 5 5 Yes 41  
## # ... with 6 more variables: HEIGHT0 <dbl>, WEIGHT0 <dbl>, RACE <chr>,  
## # Subdivision <chr>, SupportScore <dbl>, SupportAvg <dbl>

## tibble [3,302 x 15] (S3: tbl\_df/tbl/data.frame)  
## $ SWANID : num [1:3302] 10005 10046 10056 10092 10126 ...  
## $ AGE0 : num [1:3302] 48 52 51 45 48 51 46 47 46 47 ...  
## $ ANEMIA0 : chr [1:3302] "No" "No" "Yes" "Yes" ...  
## $ LISTEN0 : num [1:3302] 5 5 4 5 5 5 5 3 4 2 ...  
## $ TAKETOM0 : num [1:3302] 5 5 4 5 5 5 5 4 4 2 ...  
## $ CONFIDE0 : num [1:3302] 5 5 4 5 5 5 5 3 4 3 ...  
## $ HELPSIC0 : num [1:3302] 1 5 5 5 5 5 4 2 4 2 ...  
## $ SMOKERE0 : chr [1:3302] "No" "Yes" "No" "Yes" ...  
## $ PULSE0 : num [1:3302] 36 38 36 32 40 41 33 30 35 31 ...  
## $ HEIGHT0 : num [1:3302] 151 156 162 167 164 ...  
## $ WEIGHT0 : num [1:3302] 49.5 67.7 54.4 88.9 77.2 ...  
## $ RACE : chr [1:3302] "Hispanic" "Chinese/Chinese American" "Caucasian/ White Non-Hispanic" "Caucasian/ White Non-Hispanic" ...  
## $ Subdivision : chr [1:3302] "Minority" "Minority" "Majority" "Majority" ...  
## $ SupportScore: num [1:3302] 16 20 17 20 20 20 19 12 16 9 ...  
## $ SupportAvg : num [1:3302] 4 5 4.25 5 5 5 4.75 3 4 2.25 ...

## SWANID AGE0 ANEMIA0 LISTEN0   
## Min. :10005 Min. :42.00 Length:3302 Min. :1.000   
## 1st Qu.:31808 1st Qu.:44.00 Class :character 1st Qu.:4.000   
## Median :54230 Median :46.00 Mode :character Median :4.000   
## Mean :54362 Mean :45.85 Mean :4.206   
## 3rd Qu.:76745 3rd Qu.:48.00 3rd Qu.:5.000   
## Max. :99992 Max. :53.00 Max. :5.000   
## NA's :5 NA's :5   
## TAKETOM0 CONFIDE0 HELPSIC0 SMOKERE0   
## Min. :1.000 Min. :1.00 Min. :1.000 Length:3302   
## 1st Qu.:4.000 1st Qu.:4.00 1st Qu.:3.000 Class :character   
## Median :5.000 Median :4.00 Median :4.000 Mode :character   
## Mean :4.174 Mean :4.19 Mean :3.746   
## 3rd Qu.:5.000 3rd Qu.:5.00 3rd Qu.:5.000   
## Max. :5.000 Max. :5.00 Max. :5.000   
## NA's :6 NA's :5 NA's :5   
## PULSE0 HEIGHT0 WEIGHT0 RACE   
## Min. :17.00 Min. :140.5 Min. : 37.60 Length:3302   
## 1st Qu.:32.00 1st Qu.:157.8 1st Qu.: 59.60 Class :character   
## Median :35.00 Median :162.4 Median : 70.60 Mode :character   
## Mean :35.19 Mean :162.4 Mean : 74.88   
## 3rd Qu.:38.00 3rd Qu.:167.0 3rd Qu.: 85.50   
## Max. :84.00 Max. :186.2 Max. :175.40   
## NA's :7 NA's :32 NA's :14   
## Subdivision SupportScore SupportAvg   
## Length:3302 Min. : 4.00 Min. :1.000   
## Class :character 1st Qu.:15.00 1st Qu.:3.750   
## Mode :character Median :17.00 Median :4.250   
## Mean :16.32 Mean :4.079   
## 3rd Qu.:19.00 3rd Qu.:4.750   
## Max. :20.00 Max. :5.000   
## NA's :6 NA's :6

## Question 1: Do women with anemia have the same pulse as women who do not have anemia?

Anemia is a blood disease which can be genetic or caused by diet and lack of specific nutrients. To understand if anemia has an effect on the pulse of women in their 40’s and 50’s, two samples of 100 were analyzed from the SWAN population, one sample set with women who have been diagnosed with anemia and one sample set with women who were not diagnosed with anemia. They were compared to each other using the Welch Two Sample t Test.

State the Null Hypothesis, Alternative Hypothesis, and Claim.

$$H\_0:\mu\_1=\mu\_2\\H\_1:\mu\_1\neq\mu\_2$$

## [1] "mu1 is equal to mu2"

## [1] "mu1 does not equal mu2"

## [1] "Women with anemia have a different average pulse than women without it"

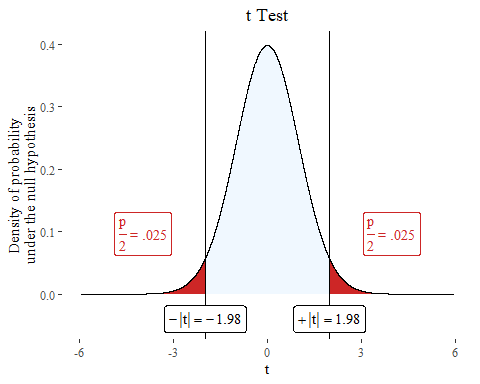
Data was subsetted for the comparison.

## SWANID AGE0 ANEMIA0 LISTEN0   
## Min. :10056 Min. :42.00 Length:1152 Min. :1.000   
## 1st Qu.:33751 1st Qu.:43.00 Class :character 1st Qu.:4.000   
## Median :57060 Median :46.00 Mode :character Median :4.000   
## Mean :55669 Mean :45.84 Mean :4.159   
## 3rd Qu.:76998 3rd Qu.:48.00 3rd Qu.:5.000   
## Max. :99809 Max. :53.00 Max. :5.000   
##   
## TAKETOM0 CONFIDE0 HELPSIC0 SMOKERE0   
## Min. :1.000 Min. :1.000 Min. :1.000 Length:1152   
## 1st Qu.:4.000 1st Qu.:4.000 1st Qu.:3.000 Class :character   
## Median :5.000 Median :4.000 Median :4.000 Mode :character   
## Mean :4.135 Mean :4.122 Mean :3.648   
## 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :5.000 Max. :5.000 Max. :5.000   
##   
## PULSE0 HEIGHT0 WEIGHT0 RACE   
## Min. :17.00 Min. :140.5 Min. : 39.00 Length:1152   
## 1st Qu.:32.00 1st Qu.:158.2 1st Qu.: 59.90 Class :character   
## Median :34.00 Median :162.7 Median : 71.00 Mode :character   
## Mean :34.95 Mean :162.7 Mean : 75.51   
## 3rd Qu.:38.00 3rd Qu.:167.0 3rd Qu.: 86.83   
## Max. :53.00 Max. :186.2 Max. :175.40   
## NA's :12 NA's :4   
## Subdivision SupportScore SupportAvg   
## Length:1152 Min. : 4.00 Min. :1.000   
## Class :character 1st Qu.:14.00 1st Qu.:3.500   
## Mode :character Median :17.00 Median :4.250   
## Mean :16.06 Mean :4.016   
## 3rd Qu.:19.00 3rd Qu.:4.750   
## Max. :20.00 Max. :5.000   
##

## SWANID AGE0 ANEMIA0 LISTEN0   
## Min. :10005 Min. :42.00 Length:2126 Min. :1.000   
## 1st Qu.:30444 1st Qu.:44.00 Class :character 1st Qu.:4.000   
## Median :52970 Median :46.00 Mode :character Median :4.000   
## Mean :53631 Mean :45.85 Mean :4.232   
## 3rd Qu.:76745 3rd Qu.:48.00 3rd Qu.:5.000   
## Max. :99992 Max. :53.00 Max. :5.000   
##   
## TAKETOM0 CONFIDE0 HELPSIC0 SMOKERE0   
## Min. :1.000 Min. :1.000 Min. :1.000 Length:2126   
## 1st Qu.:4.000 1st Qu.:4.000 1st Qu.:3.000 Class :character   
## Median :5.000 Median :4.000 Median :4.000 Mode :character   
## Mean :4.196 Mean :4.228 Mean :3.801   
## 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :5.000 Max. :5.000 Max. :5.000   
## NA's :1   
## PULSE0 HEIGHT0 WEIGHT0 RACE   
## Min. :19.00 Min. :141.0 Min. : 37.60 Length:2126   
## 1st Qu.:32.00 1st Qu.:157.3 1st Qu.: 59.50 Class :character   
## Median :35.00 Median :162.1 Median : 70.40 Mode :character   
## Mean :35.31 Mean :162.2 Mean : 74.54   
## 3rd Qu.:38.00 3rd Qu.:167.0 3rd Qu.: 85.00   
## Max. :84.00 Max. :184.0 Max. :172.10   
## NA's :20 NA's :10   
## Subdivision SupportScore SupportAvg   
## Length:2126 Min. : 4.00 Min. :1.000   
## Class :character 1st Qu.:15.00 1st Qu.:3.750   
## Mode :character Median :17.00 Median :4.250   
## Mean :16.46 Mean :4.115   
## 3rd Qu.:19.00 3rd Qu.:4.750   
## Max. :20.00 Max. :5.000   
## NA's :1 NA's :1

## # A tibble: 6 x 15  
## SWANID AGE0 ANEMIA0 LISTEN0 TAKETOM0 CONFIDE0 HELPSIC0 SMOKERE0 PULSE0  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
## 1 77803 45 Yes 2 2 3 4 Yes 37  
## 2 53815 43 Yes 5 5 5 4 No 41  
## 3 86330 48 Yes 4 4 4 4 Yes 32  
## 4 82127 48 Yes 4 4 4 2 Yes 38  
## 5 48532 42 Yes 4 5 4 3 Yes 25  
## 6 30144 48 Yes 5 5 4 5 No 30  
## # ... with 6 more variables: HEIGHT0 <dbl>, WEIGHT0 <dbl>, RACE <chr>,  
## # Subdivision <chr>, SupportScore <dbl>, SupportAvg <dbl>

## # A tibble: 6 x 15  
## SWANID AGE0 ANEMIA0 LISTEN0 TAKETOM0 CONFIDE0 HELPSIC0 SMOKERE0 PULSE0  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <chr> <dbl>  
## 1 28625 43 No 4 4 4 1 No 34  
## 2 35238 42 No 4 4 4 4 Yes 38  
## 3 92035 48 No 2 2 2 4 Yes 33  
## 4 67693 51 No 5 5 5 5 No 42  
## 5 41659 45 No 5 5 4 5 No 38  
## 6 40956 44 No 5 5 5 4 No 32  
## # ... with 6 more variables: HEIGHT0 <dbl>, WEIGHT0 <dbl>, RACE <chr>,  
## # Subdivision <chr>, SupportScore <dbl>, SupportAvg <dbl>

Critical values were calculated for a two tailed test with an alpha of 0.05. The critical value was calculated to be -1.98 to 1.98. which can be seen in the plot below. 

The t statistic was then calculate to compare against the critical values. If the t was located in the red regions of the t Test graph, it would result in a reject the Null Hypothesis, otherwise it would fail to reject.

## t   
## -0.2571732

Making the decision based on the critical value and t statistic, do not reject the null hypothesis because the t statistic is not in the critical region and is -1.98 < t < 1.98.

## [1] "Do not reject Null Hypothesis"

Summary of results.

## There is not enough evidence to support the claim: Women with anemia have a different average pulse than women without it

Because the data resulted in a fail to reject the Null Hypothesis, there is not enough evidence to support the claim that there is a difference in pulse between patients with previously diagnosed anemia and patients who were not diagnosed with anemia.

## Question 2: Is the proportion of women who smoke at age 45 the same as all women who smoke in the SWAN dataset?

The mean age in years of the SWAN dataset is slightly over 45 years old, Smokers vs non-smokers is relatively even in terms of proportions (review Milestone 1 for that analysis). To understand if 45 year olds are distributed the same as the remainder of the population, proportion of smokers from both groups were analyzed to understand the relationship.

Data was subsetted for the purpose of this analysis to include a sample of 45 year olds from the SWAN dataset.

smokers <- milestone2\_subset %>% filter(SMOKERE0=="Yes") %>% nrow()  
Total <- filter(milestone2\_subset, !is.na(SMOKERE0)) %>% nrow()  
pop\_prop <- smokers/Total  
fortyfivers <- filter(milestone2\_subset, AGE0==45)

State the Null Hypothesis, Alternative Hypothesis, and Claim.

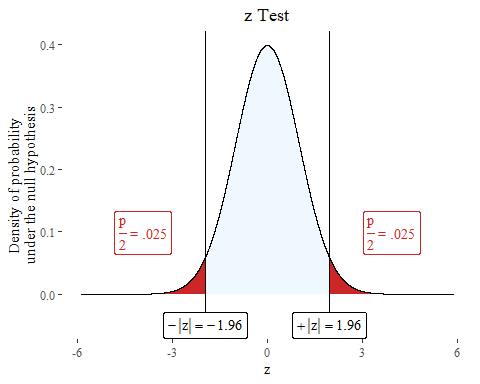
## Null: p = 43 %

## Alternative: p neq 43 %

## [1] "The proportion of smokers at age 45 is equal to the proportion of smokers in the SWAN dataset"

Proportions were calculated for the one sample Z-test for a proportion.

p <- pop\_prop  
q <- 1-p  
smoker\_45 <- fortyfivers %>% filter(SMOKERE0=="Yes") %>% nrow()  
n <- filter(fortyfivers, !is.na(SMOKERE0)) %>% nrow()  
phat <- smoker\_45/n

Critical values were calculated for a two tailed test with an alpha of 0.05. The critical value was calculated to be -1.96 to 1.96. which can be seen in the plot below. 

The z statistic was then calculate to compare against the critical values. If the z was located in the red regions of the z Test graph, it would result in a reject the Null Hypothesis, otherwise it would fail to reject.

## [1] 1.111843

Making the decision based on the critical value and z statistic, do not reject the null hypothesis because the z statistic is not in the critical region and is -1.96 < z < 1.96.

decision <- if(abs(cv)>abs(z)){  
 "Do not reject Null Hypothesis"  
}else{  
 "Reject Null Hypothesis"  
}  
decision

## [1] "Do not reject Null Hypothesis"

Summary of results:

## There is enough evidence to support the claim: The proportion of smokers at age 45 is equal to the proportion of smokers in the SWAN dataset

The claim aligned with the Null Hypothesis in this instance. The summary for this analysis that there was enough evidence to support the claim that there is no statistical difference between the proportion of smokers at age 45 to those in the SWAN dataset.

There was an additional question that tried to identify if there was a difference in support between minorities and between majority racial subdivisions. After further analysis, the data was determined to be skewed and not normally distributed, therefore that analysis is not included in this report.

**Bibliography**

Sutton-Tyrrell, Kim, Selzer, Faith, Sowers, MaryFran, R. (Mary Frances Roy), Neer, Robert, Powell, Lynda, Gold, Ellen B., … McKinlay, Sonja. Study of Women’s Health Across the Nation (SWAN): Baseline Dataset, [United States]. (1997). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-05-15. <https://doi.org/10.3886/ICPSR28762.v5\>

Waters, A. (2022). Milestone 1.