# Prevalence of STI Analysis

By: Rooney Odago

11/28/2022

## 1 Data Wrangling.

This ensures that the variables of interest have correct data type and format, ready for the analyses.

**Importing data set from STATA (.dta file)**

```
set.seed(1000)                                          # Package to help load .dta file
require(haven)                                          # load 'haven' package
library(foreign)
library(dplyr)
library(expss)
require(labelled)
stidata_unclean <- read_dta(
  "C:/Local Disk E_111920221553/STATA files/STATA TRAINING/stidata_unclean.dta"
  )
```

### Checking and dropping duplicates

```
duplct<-duplicated(stidata_unclean$idnumber)
# table(duplct)
# stidata_unclean$idnumber[duplicated(stidata_unclean$idnumber)]
stidata_unclean<- stidata_unclean[order(stidata_unclean$idnumber),]
# View(stidata_unclean[stidata_unclean$idnumber==51,])
stidata_unclean<-stidata_unclean[!(stidata_unclean$idnumber==51 & stidata_unclean$a1age ==23),]
```

### Age Category Formulation

```
stidata_unclean$AgeCat<-NA
stidata_unclean$AgeCat[stidata_unclean$a1age<=35]=1
stidata_unclean$AgeCat[stidata_unclean$a1age>35]=0
stidata_unclean$AgeCat<-factor(stidata_unclean$AgeCat,
                               levels = c(1,0),
                               labels = c("Below 35 yrs","Above 35 yrs"))
# table(stidata_unclean$AgeCat)
```

### Categorizing sex as categorical variable with factor levels

```
stidata_unclean$Sex_Cat<-NA
stidata_unclean$Sex_Cat[stidata_unclean$sex=="Female"]=1
stidata_unclean$Sex_Cat[stidata_unclean$sex=="Male"]=0
stidata_unclean$Sex_Cat<-factor(stidata_unclean$Sex_Cat,
                                levels = c(1,0),
                                labels = c("Female","Male"))
# class(stidata_unclean$Sex_Cat)
# table(stidata_unclean$Sex_Cat)
```

### Education Category

```
# table(stidata_unclean$a4levelofeducation)
# class(stidata_unclean$a4levelofeducation)
stidata_unclean$education_new<-NA
stidata_unclean$education_new[stidata_unclean$a4levelofeducation=="1 none"|
                                stidata_unclean$a4levelofeducation=="2 primary"]=1;
stidata_unclean$education_new[
  stidata_unclean$a4levelofeducation=="3 secondary"|
    stidata_unclean$a4levelofeducation=="4 tertiary"]=0;

stidata_unclean$education_new<-factor(stidata_unclean$education_new,
                                      levels = c(1,0),
                                      labels = c("Primary & Below","Secondary & Above"))
```

### Occupation

```
stidata_unclean$Occupation<-NA
stidata_unclean$Occupation[stidata_unclean$a2occupation =="1 unemployed"|
                             stidata_unclean$a2occupation == "4 student"]=1
stidata_unclean$Occupation[stidata_unclean$a2occupation =="2 informal"|
                             stidata_unclean$a2occupation =="3 formal"]=0
stidata_unclean$Occupation<-factor(stidata_unclean$Occupation,
                                   levels = c(1,0),
                                   labels = c("Unemployed", "Employed"));
```

### Religion

```
# class(stidata_unclean$a3church)
 #table(stidata_unclean$a3church)
stidata_unclean$Religion<-NA
stidata_unclean$Religion[stidata_unclean$a3church=="7 roman catholic"]=1
stidata_unclean$Religion[stidata_unclean$a3church=="2 apostolic"|
                           stidata_unclean$a3church=="3 methodist"|
                           stidata_unclean$a3church=="4 anglican"|
```

```
                            stidata_unclean$a3church=="5 pentecostal"|
                            stidata_unclean$a3church=="6 atheist 7"|
                            stidata_unclean$a3church=="8 other "]=0
stidata_unclean$Religion<-factor(stidata_unclean$Religion, levels = c(1,0),
                                 labels = c("Catholic", "Non-Catholic"))
 # table(stidata_unclean$Religion)
```

## Marital Status

```
# table(stidata_unclean$a5maritalstatus)
# class(stidata_unclean$a5maritalstatus)
stidata_unclean$Status_Marital<-NA
stidata_unclean$Status_Marital[stidata_unclean$a5maritalstatus=="2 married"]=1
stidata_unclean$Status_Marital[stidata_unclean$a5maritalstatus=="1 single"|
                                 stidata_unclean$a5maritalstatus=="5 widowed"|
                                 stidata_unclean$a5maritalstatus=="3 co-habiting"|
                                 stidata_unclean$a5maritalstatus=="4 divorcee"]=0;
stidata_unclean$Status_Marital<-factor(stidata_unclean$Status_Marital, levels = c(1,0),
                                       labels = c("Married", "Not Married"))
# table(stidata_unclean$Status_Marital)
```

### Use of Condoms

```
#table(stidata_unclean$n11usedcondom)
#class(stidata_unclean$n11usedcondom)
stidata_unclean$Used_Condom<-NA
stidata_unclean$Used_Condom[stidata_unclean$n11usedcondom=="1 yes"]=1
stidata_unclean$Used_Condom[stidata_unclean$n11usedcondom=="2 No"]=0
stidata_unclean$Used_Condom<- factor(stidata_unclean$Used_Condom,
                                     levels = c(1,0),
                                     labels = c("Yes", "No"))
# table(stidata_unclean$Used_Condom)
```

## Sex Partner in the last One year

```
# table(stidata_unclean$sexpartner1year)
# class(stidata_unclean$sexpartner1year)
stidata_unclean$Sex_Partner_1_Yr<-NA
stidata_unclean$Sex_Partner_1_Yr[stidata_unclean$sexpartner1year==1]=1
stidata_unclean$Sex_Partner_1_Yr[stidata_unclean$sexpartner1year==0]=0
stidata_unclean$Sex_Partner_1_Yr<-factor(stidata_unclean$Sex_Partner_1_Yr,
                                         levels = c(0,1),
                                         labels = c("No","Yes"))
# table(stidata_unclean$Sex_Partner_1_Yr)
```

**CaseStatus**

*** Checking and removing inconsistency

```
# class(stidata_unclean$casestatus)
# table(stidata_unclean$casestatus)
 stidata_unclean<-stidata_unclean[order(-stidata_unclean$casestatus),]
# View(stidata_unclean[stidata_unclean$casestatus==3,])
stidata_unclean$casestatus[stidata_unclean$idnumber==1 & stidata_unclean$a1age==30]=1
stidata_unclean$casestatus[stidata_unclean$idnumber==31 & stidata_unclean$a1age==23]=1
# table(stidata_unclean$casestatus)
```

## Generating STI variable and Applying labels

```
stidata_unclean$Have_STI<-NA
stidata_unclean$Have_STI[stidata_unclean$casestatus==1]=1
stidata_unclean$Have_STI[stidata_unclean$casestatus==2]=0
stidata_unclean$Have_STI<-factor(stidata_unclean$Have_STI,
                                 levels = c(1,0),
                                 labels = c("Positive", "Negative"))
```

## Apply appropriate variable labels

```
library(labelled)
var_label(stidata_unclean$AgeCat)<-"Age Category"
var_label(stidata_unclean$Sex_Cat)<-"Sex"
var_label(stidata_unclean$education_new)<-"Education"
var_label(stidata_unclean$Occupation)<-"Employement Status"
var_label(stidata_unclean$Religion)<-"Religion"
var_label(stidata_unclean$Status_Marital)<-"Marital Status"
var_label(stidata_unclean$Used_Condom)<-"Ever Used Condom"
var_label(stidata_unclean$Sex_Partner_1_Yr)<-"Had Sex Partner in last 1Yr"
var_label(stidata_unclean$Have_STI)<-"STI Stat"
```

## 2. Descriptive Statistics

### 2.1 Influence of Socio-demographic factors on STI Prevalence

```
library(gtsummary)
library(dplyr)
Vr<-stidata_unclean %>% select(AgeCat,Sex_Cat,education_new,Occupation,Religion,
                               Status_Marital,Used_Condom,Sex_Partner_1_Yr,Have_STI)

Vr %>%
  tbl_summary(by = Have_STI,missing=  'no',percent = "row") %>%
  add_overall() %>%
  modify_header(label~"**Factor**") %>%
  modify_spanning_header(c("stat_1","stat_2")~"**STI Case Status**") %>%
  bold_labels()%>%
  modify_caption("***Tabl 1: Demographic factors against STI prevalence***")
```

Table 1: ***Tabl 1: Demographic factors against STI prevalence***

| **Factor** | **Overall**, N = 226 | **Positive**, N = 113 | **Negative**, N = 113 |
|---|---|---|---|
| **Age Category** | | | |
| Below 35 yrs | 198 (100%) | 101 (51%) | 97 (49%) |
| Above 35 yrs | 28 (100%) | 12 (43%) | 16 (57%) |
| **Sex** | | | |
| Female | 107 (100%) | 53 (50%) | 54 (50%) |
| Male | 117 (100%) | 59 (50%) | 58 (50%) |
| **Education** | | | |
| Primary & Below | 27 (100%) | 15 (56%) | 12 (44%) |
| Secondary & Above | 199 (100%) | 98 (49%) | 101 (51%) |
| **Employement Status** | | | |
| Unemployed | 90 (100%) | 52 (58%) | 38 (42%) |
| Employed | 136 (100%) | 61 (45%) | 75 (55%) |
| **Religion** | | | |
| Catholic | 41 (100%) | 22 (54%) | 19 (46%) |
| Non-Catholic | 117 (100%) | 60 (51%) | 57 (49%) |
| **Marital Status** | | | |
| Married | 149 (100%) | 74 (50%) | 75 (50%) |
| Not Married | 77 (100%) | 39 (51%) | 38 (49%) |
| **Ever Used Condom** | 49 (100%) | 22 (45%) | 27 (55%) |
| **Had Sex Partner in last 1Yr** | 52 (100%) | 42 (81%) | 10 (19%) |

**Interpretation**

The table 1 above shows a summary of a total of 226 people who participated in a study to determine the whether demographic factors influence prevalence of Sexually Transmitted Infections. We find that 51%,that is 101 of study participants above 35 years of age tested Positive while 49%, that is 97 of the same group tested Negative. The same interpretation follows for the rest of the factors except for Had Sex Partner in the last 1Yr which reveals that 81% of participants who had a sex partner in the last one year tested Positive while only 10% of the same cohort showed Negative results.

### 2.1 Influence of Socio-demographic factors on STI Prevalence

```
library(gtsummary)
library(dplyr)
Vr<-stidata_unclean %>% select(AgeCat,Sex_Cat,education_new,Occupation,Religion,
                               Status_Marital,Used_Condom,Sex_Partner_1_Yr,Have_STI)

Vr %>%
  tbl_summary(by = Have_STI,missing = 'no',percent = "row") %>%
  add_overall() %>%
  modify_header(label~"**Factor**") %>%
  add_p() %>%
bold_labels()%>%
modify_caption("***Tabl 1: Test between Demographic factors against STI prevalence***")
```

Table 2: ***Tabl 1: Test between Demographic factors against STI prevalence***

| **Factor** | **Overall**, N = 226 | **Positive**, N = 113 | **Negative**, N = 113 | **p-value** |
|---|---|---|---|---|
| **Age Category** | | | | 0.4 |
| Below 35 yrs | 198 (100%) | 101 (51%) | 97 (49%) | |
| Above 35 yrs | 28 (100%) | 12 (43%) | 16 (57%) | |
| **Sex** | | | | 0.9 |
| Female | 107 (100%) | 53 (50%) | 54 (50%) | |
| Male | 117 (100%) | 59 (50%) | 58 (50%) | |
| **Education** | | | | 0.5 |
| Primary & Below | 27 (100%) | 15 (56%) | 12 (44%) | |
| Secondary & Above | 199 (100%) | 98 (49%) | 101 (51%) | |
| **Employement Status** | | | | 0.057 |
| Unemployed | 90 (100%) | 52 (58%) | 38 (42%) | |
| Employed | 136 (100%) | 61 (45%) | 75 (55%) | |
| **Religion** | | | | 0.8 |
| Catholic | 41 (100%) | 22 (54%) | 19 (46%) | |
| Non-Catholic | 117 (100%) | 60 (51%) | 57 (49%) | |
| **Marital Status** | | | | 0.9 |
| Married | 149 (100%) | 74 (50%) | 75 (50%) | |
| Not Married | 77 (100%) | 39 (51%) | 38 (49%) | |
| **Ever Used Condom** | 49 (100%) | 22 (45%) | 27 (55%) | 0.4 |
| **Had Sex Partner in last 1Yr** | 52 (100%) | 42 (81%) | 10 (19%) | <0.001 |

## 2.2 Hypothesis Testing

$H_o$ :$\beta_j = 0$ (Demographic factors above do not have an influence on STI prevalence)
$H_1$ :$\beta_j \neq 0$ (Demographic factors above have an influence on STI prevalence)

### 2 .2 .2 Interpretation

Let $\alpha = 0.05$, the critical value.
From the Table 2 above the factor **Had Sex Partner in Last 1Yr** had a p-value of "$< 0.001$" which is less than critical value, therefore statistically significant in determining the STI prevalence. Hence, we conclude that there was no enough information to reject $H_o$ for all the risk factors except for *Had Sex Partner in Last 1Yr* at 95% Level of signifcance. We therefore fit a generalized linear model to assist in determining the level of influence the factor had on STI prevalence.

### 2.3 Multivariate Regression Analysis of Risk Factors for STI(Using Backward method)

By picking only the statistically significant variable, we have:

```
library(equatiomatic)
Model1<-glm(Have_STI~Sex_Partner_1_Yr,
            family = binomial(link = logit), data = stidata_unclean)
```

```
equatiomatic::extract_eq(Model1)
```

$$\log\left[\frac{P(\text{Have\_STI} = \text{Negative})}{1 - P(\text{Have\_STI} = \text{Negative})}\right] = \alpha + \beta_1(\text{Sex\_Partner\_1\_Yr}_{\text{Yes}}) \tag{1}$$

```
equatiomatic ::extract_eq(Model1,use_coefs=TRUE)
```

$$\log\left[\frac{P(\widehat{\text{Have\_STI}} = \text{Negative})}{1 - P(\widehat{\text{Have\_STI}} = \text{Negative})}\right] = 0.37 - 1.81(\text{Sex\_Partner\_1\_Yr}_{\text{Yes}}) \tag{2}$$

```
tbl_regression(Model1, exponentiate = T, intercept = T)
```

| **Characteristic** | **OR** | **95% CI** | **p-value** |
|---|---|---|---|
| (Intercept) | 1.45 | 1.07, 1.97 | 0.016 |
| Had Sex Partner in last 1Yr | | | |
| No | — | — | |
| Yes | 0.16 | 0.07, 0.34 | <0.001 |

### Conclusion

Since the reference group for whether a participant Had Sex Partner in the last 1 Year is *No*, the Odds Ratio of 0.16 reveals that people who have never had sex partner in the last 1 year are 0.16 less likely to contract STI.