Analysis of Participants Information for Handle and Swab Samples

SECTION 0: Required Packages and Files

The following packages must be installed and loaded into the R script prior to execution.

```
library(dplyr)
library(data.table)
library(ggplot2)
library(tidyr)
library(readr)
library(ggpubr)
```

The datasets required to do the redcap partipant analysis can be downloaded from github (https://github.com/bbi-lab)

1. redcap_SCAN_SFS_dataset.csv

SECTION 1: Creating the Swab Sample and Handle Sample Dataframes

Loading the Swab and Handle Sample Dataset

Remember to set your working directary to the appropriate path.

One dataset is loaded and read:

1. The RedCap dataset of all SCAN and SFS samples (both swab and handle).

Note that the samples are catagorized based on side_used, whether the participant used the handle or swab for nasal collection. Additionally, all the information we need to for data analysis and visualization is within this one dataset.

redcap_SCAN_SFS_dataset

```
# load the redcap SCAN/SFS dataset
dfBoth <- read_csv("redcap_SCAN_SFS_dataset.csv")
```

At the end of this section, there is one dataframe saved in the Global Environment.

1. **dfBoth**

dfBoth will be used to for data analysis and data visualization as seen in the next section.

SECTION 2: Data Analysis and Visualization for Confidence and Discomfort Between Handle and Swab

Performing Fisher's Test for Confidence and Discomfort During Nasal Swab Collection

Participant's confidence and discomfort during nasal swab collection are two variables that can be compared between handle versus swabs. To determine whether handle versus swab is significantly different, a Fisher's test can be performed.

```
# create a tables of confidence
confidence_table <- table(dfBoth$confidence, dfBoth$side_used)
confidence_prop <- prop.table(table(dfBoth$confidence, dfBoth$side_used))

# create a tables of discomfort
discomfort_table <- table(dfBoth$discomfort, dfBoth$side_used)
discomfort_prop <- prop.table(table(dfBoth$discomfort, dfBoth$side_used))

# calculate p-value for confidence of handle versus swab
fisher.test(dfBoth$confidence, dfBoth$side_used)

# calculate p-value for discomfort for handle versus swab
fisher.test(dfBoth$discomfort, dfBoth$side_used)

# binding data of confidence and discomfort together
confidence_prop_table <- data.table(confidence_prop)
colnames(confidence_prop_table) <- c("confidence","side_used","freq")
discomfort_prop_table <- data.table(discomfort_prop)
colnames(discomfort_prop_table) <- c("discomfort","side_used","freq")
nasal_swab_usability_table <- bind_rows(confidence_prop_table,discomfort_prop_table)
write_csv(nasal_swab_usability_table,"nasal_swab_usability_table.csv")
```

The p-value for confidence between handle versus swab is **0.02**

The p-value for discomfort between handle verus swab is $<$ **0.01**

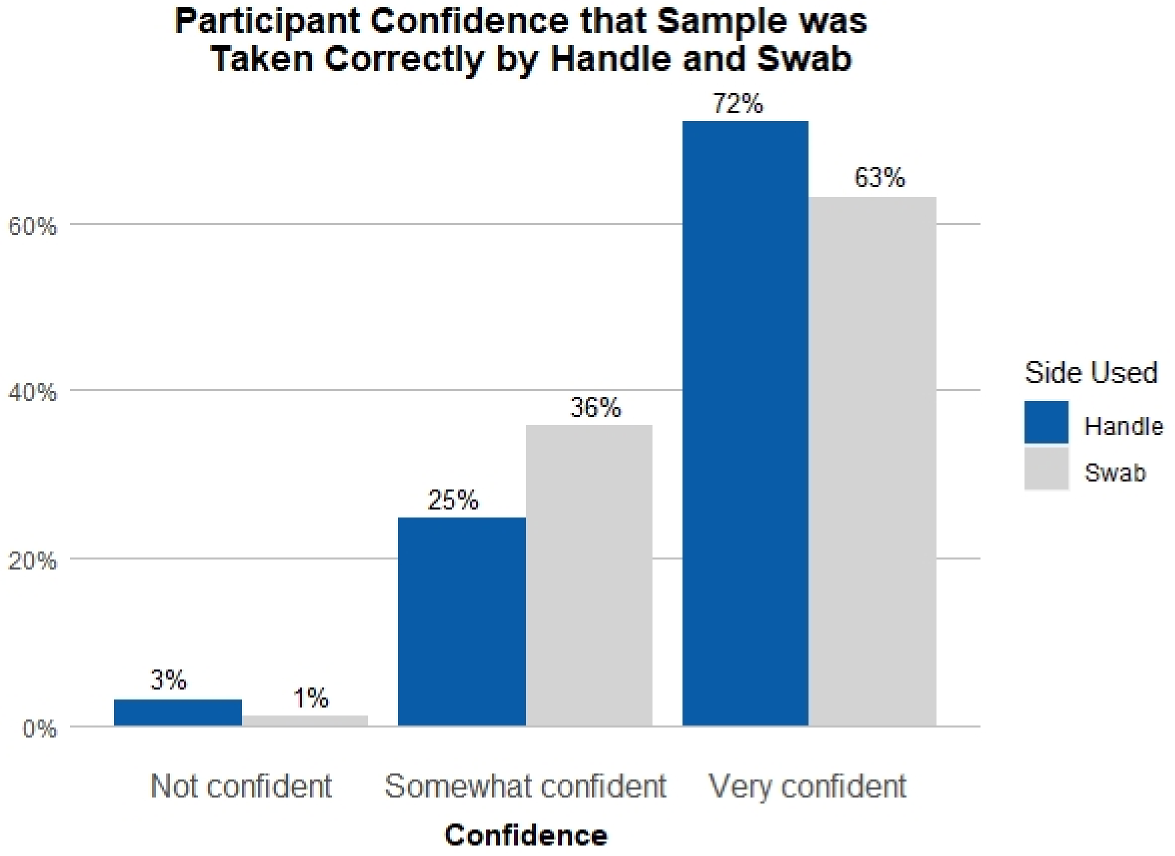
nasal_swab_usability_table is saved as a csv into the directory.

Creating Bar Plots of Confidence and Discomfort

Before the bar plots can be created, a table needs to include a relative frequency column. Once created, a bar plot showing the percentage of participants' level of confidence for either handle or swab can be generated.

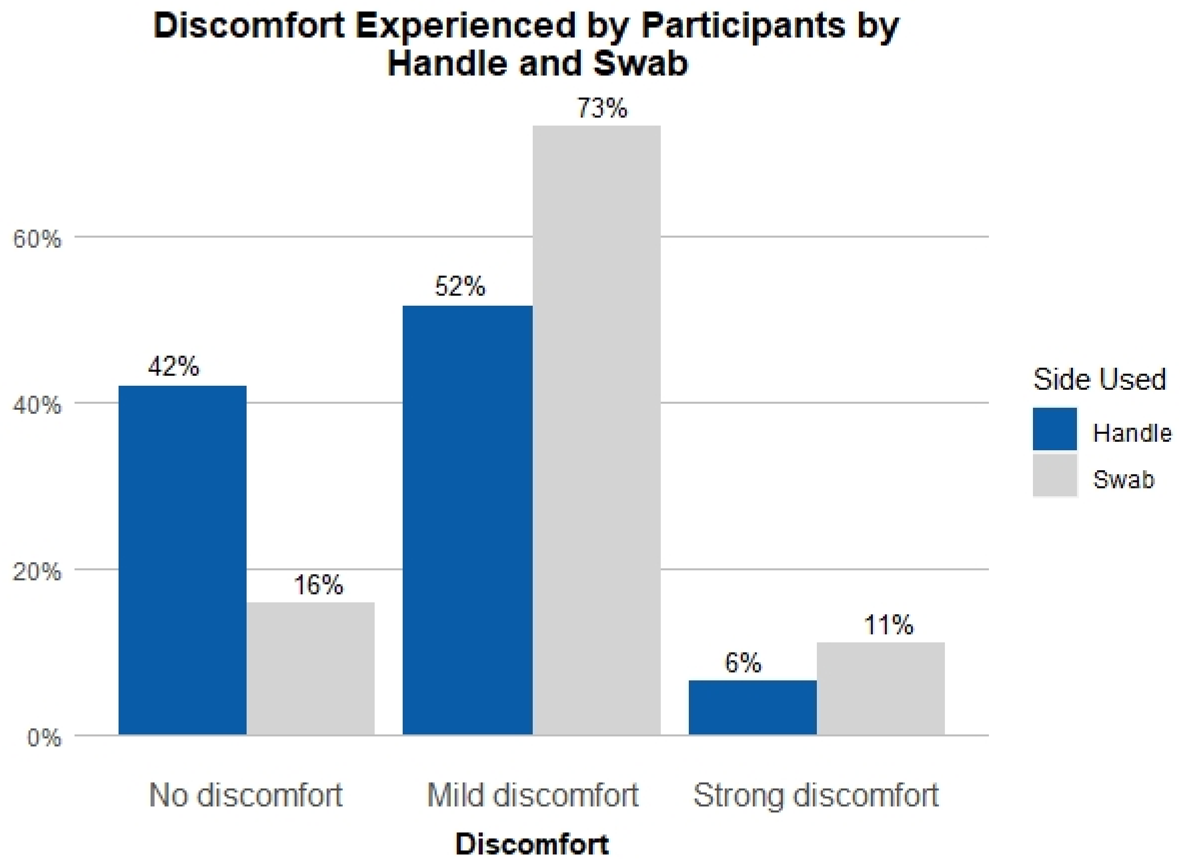
```
# creates table with relative frequency
rel_confidence_prop_table <- confidence_prop_table %>% group_by(side_used) %>%
  mutate(relfreq = freq / sum(freq))

# rename the columns in prep of bar plot generation
rel_confidence_prop_table$confidence <- recode(rel_confidence_prop_table$confidence,
                                                "not_con" = "Not confident", "some_con" =
                                                    "Somewhat confident", "v_con" = "Very
                                                  confident")
```

```
# create barplot for confidence
ggplot(data = rel_confidence_prop_table,aes(x=confidence,y=relfreq,fill=side_used))+
  geom_bar(stat="identity", position = "dodge")+
  scale_y_continuous(labels = scales::percent)+
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major.y = element_line(size = 0.5, linetype = 'solid', color = "grey"),
        panel.grid.major.x = element_blank(),
        legend.position = "right",
        plot.title = element_text(face = "bold", hjust = 0.5),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold",vjust= -1),
        axis.ticks = element_blank(),
        axis.title.y = element_blank())+
  geom_text(aes(label=scales::percent(relfreq,accuracy =1),y=relfreq),
            vjust=-0.5, size=3.5,position = position_dodge(width=1))+
  ggtitle("Participant Confidence that Sample was \n Taken Correctly by Handle and Swab")+
  scale_fill_manual(values = c("Swab" = "lightgrey",
                               "Handle" = "#095DA8"))+
  labs(x="Confidence
       ", fill = "Side Used")
```



Similarly, a bar plot can be generated to display the percent range of discomfort between handle versus swab. Once again, a relative frequency column must be created.

```
# create table with relative frequency
rel_discomfort_prop_table <- discomfort_prop_table %>% group_by(side_used) %>%
  mutate(relfreq = freq / sum(freq))

# rename column names in prep of bar plot generation
rel_discomfort_prop_table$discomfort <- recode(rel_discomfort_prop_table$discomfort,
                                                "no_dis" = "No discomfort", "mild_dis" = "Mild
                                                  discomfort", "strong_dis" = "Strong
                                                  discomfort")

rel_discomfort_prop_table$discomfort <- factor(rel_discomfort_prop_table$discomfort,
                                               levels = c("No discomfort","Mild
                                                          discomfort","Strong discomfort"))

# create barplot for discomfort
ggplot(data = rel_discomfort_prop_table,aes(x=discomfort,y=relfreq,fill=side_used))+
  geom_bar(stat="identity", position = "dodge")+
  scale_y_continuous(labels = scales::percent)+
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major.y = element_line(size = 0.5, linetype = 'solid', color = "grey"),
        panel.grid.major.x = element_blank(),
        legend.position = "right",
        plot.title = element_text(face = "bold", hjust = 0.5),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold",vjust= -1),
        axis.ticks = element_blank(),
        axis.title.y = element_blank())+
  geom_text(aes(label=scales::percent(relfreq,accuracy =1),y=relfreq),
            vjust=-0.5, size=3.5,position = position_dodge(width=1))+
  ggtitle("Discomfort Experienced by Participants by \n Handle and Swab")+
  scale_fill_manual(values = c("Swab" = "lightgrey",
                               "Handle" = "#095DA8"))+
  labs(x="Discomfort
       ", fill = "Side Used")

# removing unneeded dataframes
rm(confidence_prop_table)
rm(discomfort_prop_table)
```



SECTION 3: Data Analysis and Visualization for Age Between Handle and Swab

Calculating Mean and Median Age

The mean and median age between handle and swab can be evaluated by grouping based on side used and then summarizing the mean and median for each group.

```
dfBoth %>%
  dplyr::group_by(side_used) %>%
  dplyr::summarise(mean = mean(age,na.rm=TRUE),median = median(age,na.rm=TRUE), n = n())
```

The mean and median age for handle is **59.0** and **62** respectively. The mean and median age for swab is **41.2** and **39** respectively.

Calculating Significance of Age between Handle and Swab

A t-test can be used for the difference in age between handle and swab. Additionally, a Mann-Whitney-Wilcox test can be used because the distributions are skewed.

```r
# t test for difference in age
dfAge <- data.frame(side_used = dfBoth$side_used, age = dfBoth$age)
dfAge <- dfAge %>% mutate(id = row_number())
age_t_test <- dfAge %>% spread(side_used,age)
t.test(age_t_test$Handle,age_t_test$Swab)

# Mann-Whitney-Wilcox test
wilcox.test(age_t_test$Handle,age_t_test$Swab)

# remove unneeded dataframe
rm(age_t_test)
```

The p-value for the t-test $<$ **0.01**. The p-value for the Wilcox-Test is also $<$ **0.01**.

Preparing the Histogram for Age Between Handle and Swab

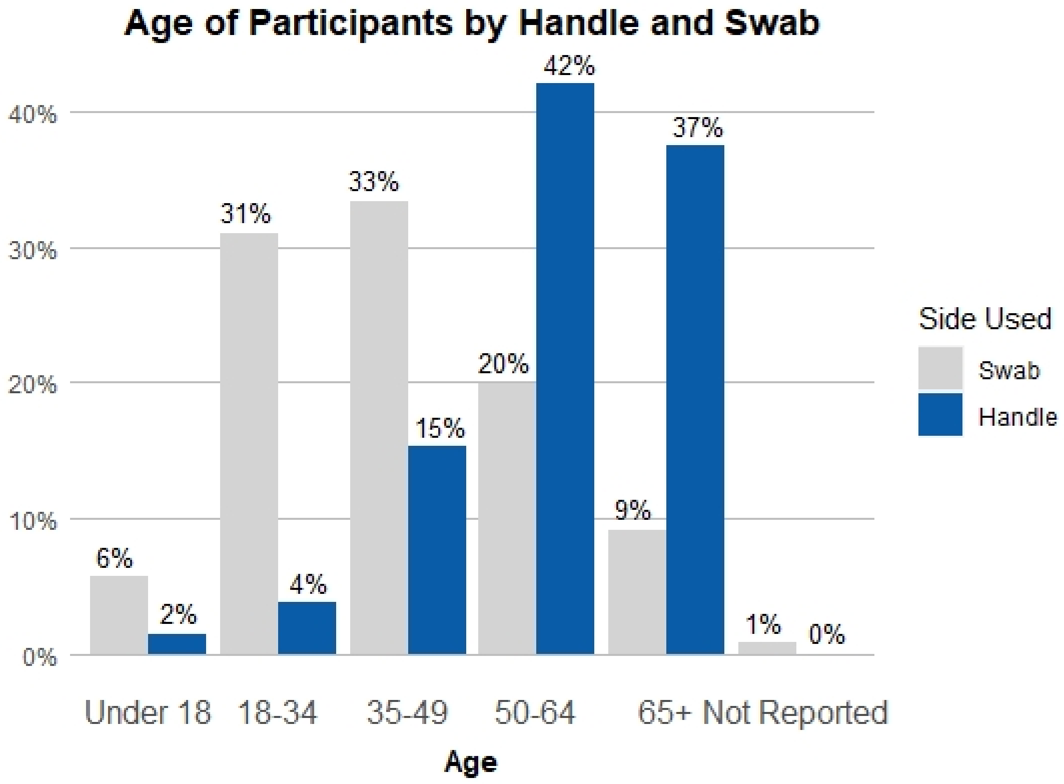
For visualization, a histogram is most appropriate to display the participants' age based on side used: handle versus swab. To prepare the histogram, age is separated into the appropriate binnings.

```r
# all the data on age is separated into age binnings
age_group <- dfBoth %>% mutate(age_group = case_when(
  age >= 65 ~ ’65+’,
  age >= 50 ~ ’50-64’,
  age >= 35 ~ ’35-49’,
  age >= 18 ~ ’18-34’,
  age < 18 ~ ’Under 18’,
  is.na(age) ~ "Not Reported"))

# create new dataframe with age binnings
age_group <- select(age_group, age_group, side_used)
names(age_group)[1] = "age"

# create new dataframe with age binnings and frequency
age_hist <- age_group %>%
  group_by(side_used,age,.drop=FALSE) %>%
  dplyr::summarise(n = n()) %>%
  complete(side_used, fill=list(count_a =0))%>%
  mutate(freq = n / sum(n))%>%
  ungroup()

# remove unneeded dataframes
rm(age_group)
rm(dfAge)

# add additional catagory (Handle: Not Reported)
age_hist <- age_hist %>% add_row(side_used = "Handle", age = "Not Reported", n=0, freq=0)

# QC-ing of age_hist, including adding a percentage column
age_hist$age <- factor(age_hist$age, levels =
                         c("Under 18","18-34","35-49","50-64", "65+","Not Reported"))
age_hist$side_used <- factor (age_hist$side_used, levels = c("Swab","Handle"))
age_hist$percent <- scales::percent(age_hist$freq)
```

Creating the Age Histogram between Handle and Swab

A barplot can now be generated.

```
# histogram is generated
ggplot(age_hist, aes(x=age,y=freq,fill=side_used))+
  geom_bar(stat="identity",position = position_dodge(preserve="single"))+
  scale_y_continuous(labels = c("0%","10%","20%","30%","40%"))+
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major.y = element_line(size = 0.5, linetype = ’solid’, color = "grey"),
        panel.grid.major.x = element_blank(),
        legend.position = "right",
        plot.title = element_text(face = "bold", hjust = 0.5),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold",vjust= -1),
        axis.ticks = element_blank(),
        axis.title.y = element_blank())+
  geom_text(aes(label=scales::percent(freq,accuracy =1),y=freq),
            vjust=-0.5, size=3.5,position = position_dodge(width=1))+
  ggtitle("Age of Participants by Handle and Swab")+
  scale_fill_manual(values = c("Swab" = "lightgrey",
                               "Handle" = "#095DA8"))+
  labs(x="Age
       ", fill = "Side Used")
```



SECTION 4: Data Analysis and Visualization for Sex Between Handle and Swab

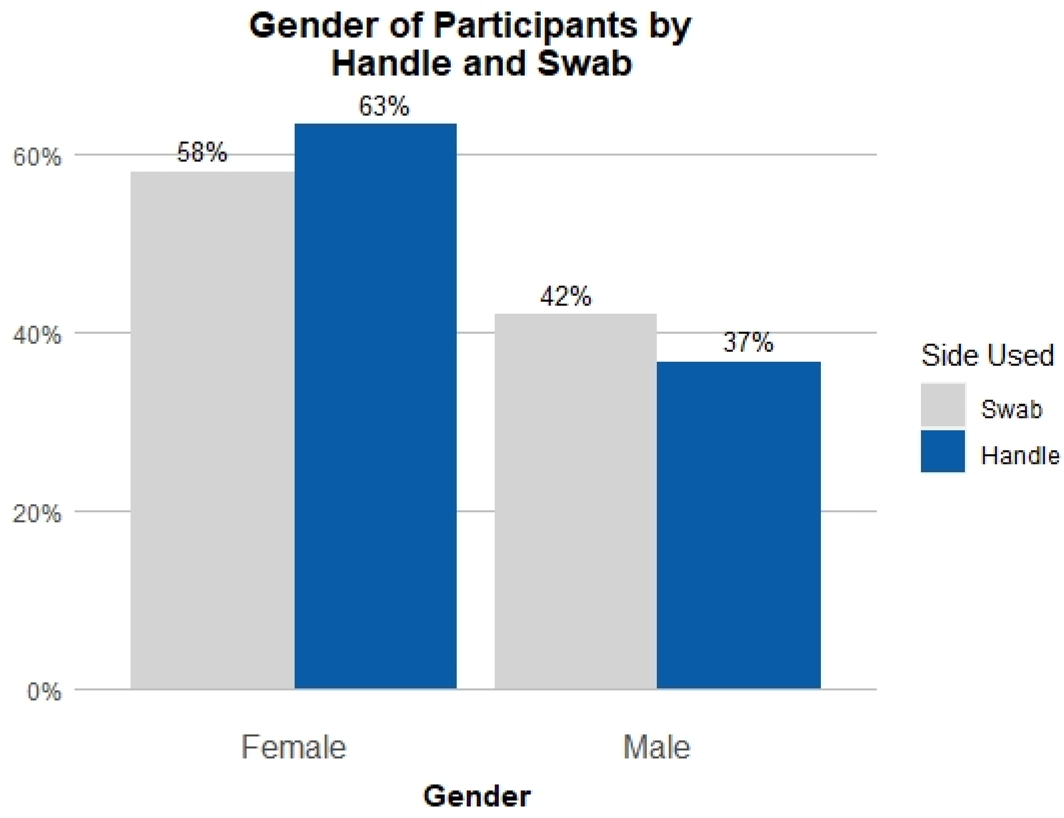
Before a visualization can be generated, participants’ sex are grouped and QC-ed (removal of NAs, addition of frequency). ## Preparing the Bar Plot for Sex Between Handle and Swab

```
# sex binning is created
sex_group <- dfBoth %>% mutate(sex_group = case_when(
  sex == "male" ~ ’Male’,
  sex == "female" ~ ’Female’,
  sex == "other" ~ ’NA’,
  sex == "dont_say" ~ ’NA’))

# NAs are removed
sex_group <- sex_group[!is.na(sex_group$sex_group), ]
sex_group <- sex_group %>% filter(!sex_group %in% "NA")

# dataframe created and modified
sex_hist <- sex_group %>%
  group_by(side_used,sex_group,.drop=FALSE) %>%
  dplyr::summarise(n = n()) %>%
  complete(side_used, fill=list(count_a =0))%>%
  mutate(freq = n / sum(n))%>%
  ungroup()

sex_hist$side_used <- factor (sex_hist$side_used, levels = c("Swab","Handle"))
```

Creating the Bar Plot for Sex between Handle and Swab

With sex_hist created, a bar plot can be now generated.

```
# a bar plot can be generated
ggplot(sex_hist, aes(x=sex_group,y=freq,fill=side_used))+
  geom_bar(stat="identity",position = position_dodge(preserve="single"))+
  scale_y_continuous(labels = scales::percent)+
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major.y = element_line(size = 0.5, linetype = ’solid’, color = "grey"),
        panel.grid.major.x = element_blank(),
        legend.position = "right",
        plot.title = element_text(face = "bold", hjust = 0.5),
        axis.text.x = element_text(size = 12),
        axis.title.x = element_text(face = "bold",vjust= -1),
        axis.ticks = element_blank(),
        axis.title.y = element_blank())+
  geom_text(aes(label=scales::percent(freq,accuracy =1),y=freq), vjust=-0.5,
            size=3.5,position = position_dodge(width=1))+
  ggtitle("Gender of Participants by \n Handle and Swab")+
  scale_fill_manual(values = c("Swab" = "lightgrey",
                               "Handle" = "#095DA8"))+
  labs(x="Gender", fill = "Side Used")

rm(sex_group)
```



Calculating Significance of Sex Between Handle and Swab

A chi-square test with a Yates' continuity correction can be used to determine whether sex is significant between handle and swab.

```
# create a 2x2 dataframe in prep of chi-square test
dfSex <- data.frame(side_used = sex_group$side_used, sex = sex_group$sex_group)

# chi-square test
chisq.test(dfSex$sex,dfSex$side_used)

# display table
table(dfSex)
rm(dfSex)
```

The p-value for sex between handle and sex is **0.28**. There is no significant difference in handle use between female and male.

SECTION 5: Data Analysis and Visualization for Income Between Handle and Swab

Similarly to the last section ~ Before a visualization can be generated, participants' income are grouped and QC-ed (removal of NAs, addition of frequency). ## Preparing the Bar Plot for Income Between Handle and Swab

```
# new dataframe created
dfIncome <- dfBoth

# combine income_levels and income together into new column
dfIncome$income_both <- coalesce(dfIncome$income_levels,dfIncome$income)

# select for certain columns
dfIncome <- subset(dfIncome, select =
                     c("side_used","collection_id","study","income","income_levels","income_both"))

# income binnings are created
income_group <- dfIncome %>% mutate(income_group = case_when(
  income_both == "100k_125k" ~ '100-125k',
  income_both == "125k_150k" ~ '125-150k',
  income_both == "25k_50k" ~ '25-50k',
  income_both == "50k_75k" ~ '50-75k',
  income_both == "75k_100k" ~ '75-100k',
  income_both == "dont_know" ~ "NA",
  income_both == "dont_say" ~ "NA",
  income_both == NA ~ "NA",
  income_both == "less_25k" ~ '<25k',
  income_both == "more_150k" ~ '>150k',))
income_group$income_group <- income_group$income_group %>% replace_na("NA")
rm(dfIncome)
```

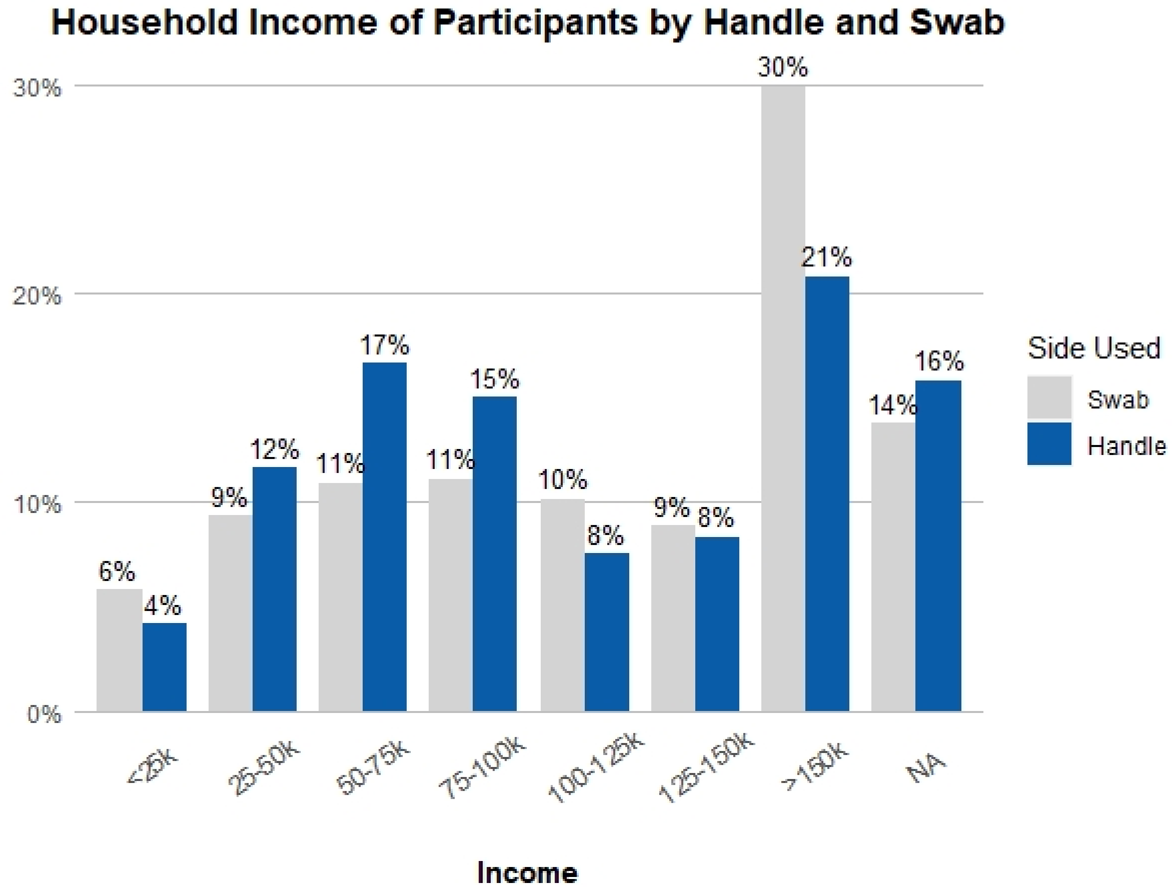
At the end of this mini session, the participants' income are merged together and catagorized into income_group. A dataframe called income_hist is created and details information on frequency and count for each income_group based on handle versus swab.

```
# dataframe created with frequency and count
income_hist <- income_group %>%
  group_by(side_used,income_group,.drop=FALSE) %>%
  dplyr::summarise(n = n()) %>%
  complete(side_used, fill=list(count_a =0))%>%
  mutate(freq = n / sum(n))%>%
  ungroup()
sum(income_hist$freq)

# QC-ing of the dataframe (factor)
income_hist$income_group <- factor(income_hist$income_group, levels =                               c("
income_hist$side_used <- factor (income_hist$side_used, levels = c("Swab","Handle"))
```

Creating Bar Plot of Income Between Handle and Swab

Once income_hist is created, a box plot can be generated.

```
ggplot(income_hist, aes(x=income_group,y=freq,fill=side_used, width = 0.8))+
  geom_bar(stat="identity",position = position_dodge(preserve="single"))+
  scale_y_continuous(labels = scales::percent)+
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major.y = element_line(size = 0.5, linetype = 'solid', color = "grey"),
        panel.grid.major.x = element_blank(),
        legend.position = "right",
        plot.title = element_text(face = "bold", hjust = 0.5),
        axis.text.x = element_text(size = 10, angle = 35, vjust=1),
        axis.title.x = element_text(face = "bold",vjust= -1),
        axis.ticks = element_blank(),
        axis.title.y = element_blank())+
  geom_text(aes(label=scales::percent(freq,accuracy =1),y=freq), vjust=-0.5, size=3.5,
            position = position_dodge(width=.8))+
  ggtitle("Household Income of Participants by Handle and Swab")+
  scale_fill_manual(values = c("Swab" = "lightgrey",
                               "Handle" = "#095DA8"))+
  labs(x="Income", fill = "Side Used")
```

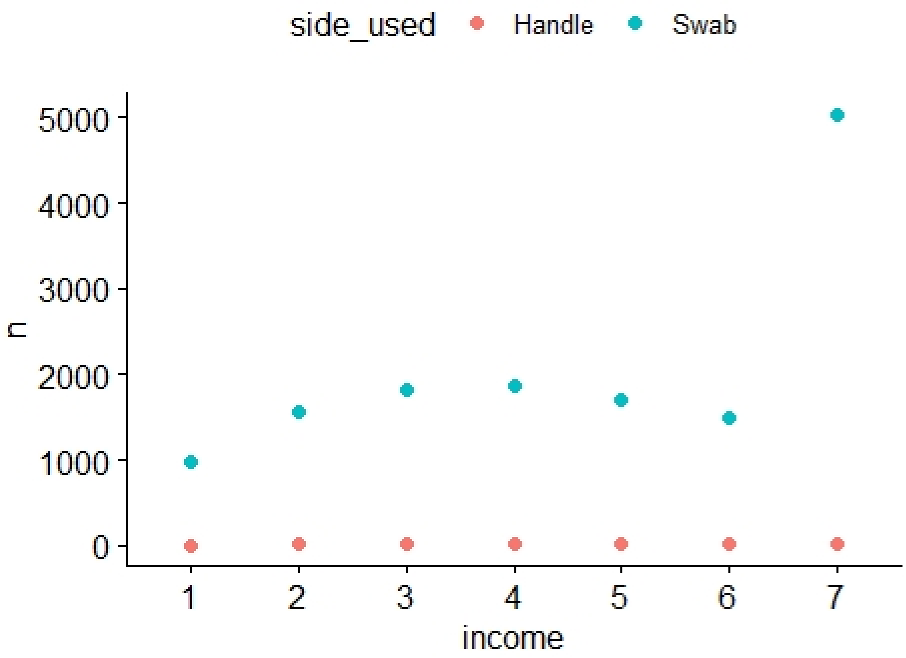


Calculating Correlation of Income Between Handle and Swab

To determine if there is a correlation of income between handle and swab, a scatter plot can be generated.

```
# create new dataframe and convert income binning into numerics
income_cor <- data.table(side_used = income_group$side_used,
                         income = recode(income_group$income_group, "<25k" = 1 ,"25-50k" = 2, "50-75k"


# create another dataframe that includes frequency and count
```

```
income_cor_hist <- income_cor%>%
  group_by(side_used,income,.drop=FALSE) %>%
  dplyr::summarise(n = n()) %>%
  complete(side_used, fill=list(count_a =0))%>%
  mutate(freq = n / sum(n))%>%
  ungroup()

# QC the dataframe
income_cor_hist[c(8,16), ]<-NA
income_cor_hist <- na.omit(income_cor_hist)

# create scatterplot to visually see the correlation
ggscatter(income_cor_hist, x = "income", y = "n", color = "side_used",
          cor.coef = TRUE, cor.method = "pearson")

# insert side_used column for each dataframe
incorhisthandle <- income_cor_hist %>% filter(side_used == "Handle")
incorhistswab <- income_cor_hist %>% filter(side_used == "Swab")

# calculate correlation between income and handle versus swab
cor(as.numeric(incorhisthandle$income),incorhisthandle$n)
cor(as.numeric(incorhistswab$income),incorhistswab$n)
```



Based on the scatter plot, there is no noticable pattern in terms of income and side-used for nasal collection. Additionally, there is no correlation between income and number of participants who used the handle for nasal collection (p-value = 0.45).