Exploring the BRFSS data

A Suarez-Pierre

Setup

Load packages

```
library(ggplot2)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.4
```

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `brfss2013`. Delete this note when before you submit your work.

```
load("brfss2013.RData")
```

Part 1: Data

This entire sample is subject to non-response bias; all the collected data comes from people who where willing to answer the phone and finish a very long survey.

It is not a great representation of the general population, but the sample size is useful to find statistical significance to even small differences.

For computing capabilities, we will refer to the data collected in the state of California only.

```
df = brfss2013
rm(brfss2013)                             # get rid of long name

df = df[20000:40000,]                     # trimming data, no CA rows deleted
df = tbl_df(df)
df = filter(df, X_state=="California")    # subsetting for CA state only

dim(df)
```

```
## [1] 11518   330
```

Due to the fact that this has been collected retrospectively we can only imply CORRELATIONS not CAUSALITY.

Part 2: Research questions

Research quesion 1: *What are the statistical characteristics of the estimaged age-gender specific maximum oxygen consumption? Does it follow a normal distribution?*

variable: `$maxvo2_` (continuous) in ml/min/kg

Research quesion 2: *Does a correlation exist between the hours slept each night and reported days with poor physical health or poor mental health in the last 30 days?*

explanatory variable: `sleptim1` in hours

response variables: `physhlth` and `menthlth` in days

```
all of these variables are continuous
```

Research quesion 3: *Does a correlation exist between the body mass index (BMI) of the individual answering the survey and having a high blood cholesterol or having a heart attack?*

explanatory variable: `_bmi5` (continuous) computed by researchers in kg/m^2

response variables: `toldhi2` and `cvdinfr4`

```
both response variables are discrete
```

Part 3: Exploratory data analysis

Research quesion 1:

What are the statistical characteristics of the estimaged age-gender specific maximum oxygen consumption? Does it follow a normal distribution?

The last two digits of this variable are implied decimal places, before doing anything this corrects to the units: `ml/min/kg`

```
df$maxvo2_ = as.numeric(df$maxvo2_)/100
```

We will now get some descritptive statistics on the computed measurement.

```
summary(df$maxvo2_)
```

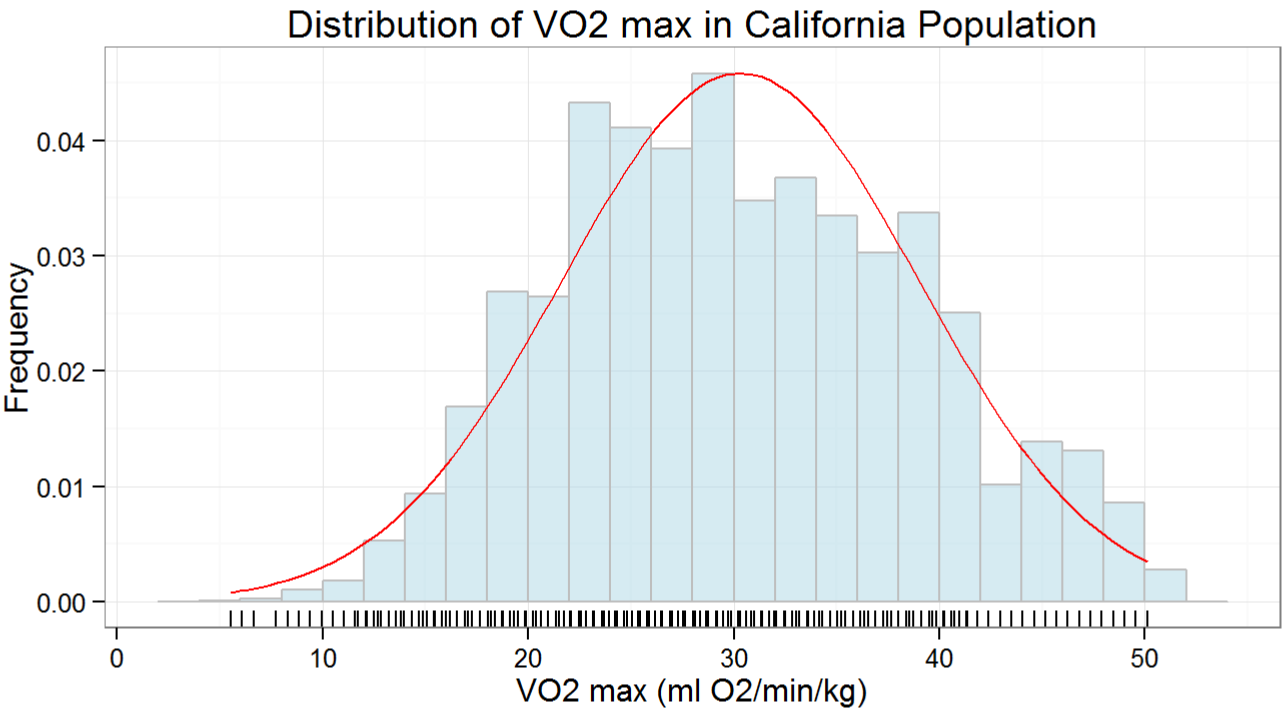
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    5.55   23.70   29.75   30.31   36.90   50.10      10
```

```
sd(df$maxvo2_, na.rm=TRUE)
```

```
## [1] 8.697629
```

Next some exploratory graphs

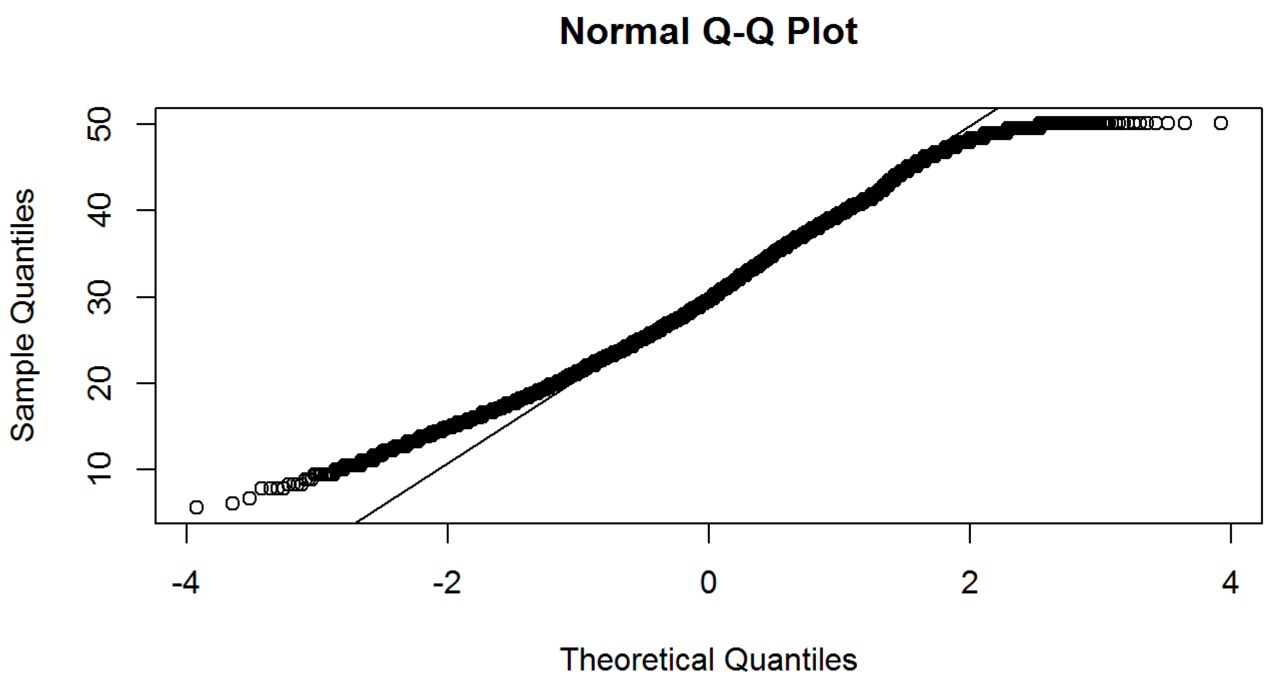
```
maxvo2 = ggplot(df, aes(x=maxvo2_)) + theme_bw() +
          geom_histogram(binwidth=2, alpha=0.5, aes(y=..density..), fill="lightblue", col="grey") +
           stat_function(fun=dnorm, args=list(mean=30.31, sd=8.7), col="red") +
            geom_rug() +
             labs(title="Distribution of VO2 max in California Population",
                  x="VO2 max (ml O2/min/kg)",
                  y="Frequency")
maxvo2
```



```
  The red superimposed curve is a normal distribution with the same
  mean and standard deviation that this sample has. Nevertheless, this
  is not the most accurate way to determine if `maxvo2` has a normal
  distribution.
```

A another visual method to evaluate if this sample follow a sample distribution is a normal probability plot using the `qqnorm` function

```
qqnorm(df$maxvo2_)
qqline(df$maxvo2_)
```



The sample looks like it does not follow a normal distribution, it is seem to be ***bimodal*** as evidenced from the 'S' shape of the QQ plot.

Regardless, we can use the `Shapiro-Wilk normality test` to assess this more thoroughly. The null hypothesis (Ho) is that the data follow a normal distribution, and we will reject the Ho if the `p-value < 0.05`.

```
shapiro.test(sample(df$maxvo2_, 5000))
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  sample(df$maxvo2_, 5000)
## W = 0.98845, p-value < 2.2e-16
```

```
  Unfortunately, the shapiro.test function only allows for samples
  with sizes between 3 and 5000 observations, therefore we take a
  random sample.
```

Conclusion `p < 2.2e-16`, we can assume that this sample **does not follow a normal distribution.**

Research quesion 2:

Does a correlation exist between the hours slept each night and reported days with poor physical health or poor mental health in the last 30 days?

Turning all the values into `numeric`

```
df$sleptim1 = as.numeric(df$sleptim1)
df$physhlth = as.numeric(df$physhlth)
df$menthlth =           (df$menthlth)
```

```
               as.numeric
```

Descriptive statistics

```
stats = summarize(df, "sleptim1", mean(sleptim1, na.rm=T), sd(sleptim1, na.rm=T))
colnames(stats) = c("Variable", "mean", "sd")

temp = summarize(df, "physhlth", mean(physhlth, na.rm=T), sd(physhlth, na.rm=T))
colnames(temp) = c("Variable", "mean", "sd")
stats = rbind(stats, temp)

temp = summarize(df, "menthlth", mean(physhlth, na.rm=T), sd(menthlth, na.rm=T))
colnames(temp) = c("Variable", "mean", "sd")
stats = rbind(stats, temp)

stats
```

```
## Source: local data frame [3 x 3]
##
##   Variable     mean       sd
##      (chr)    (dbl)    (dbl)
## 1 sleptim1 7.057412 1.423875
## 2 physhlth 4.233368 8.677749
## 3 menthlth 4.233368 7.846700
```

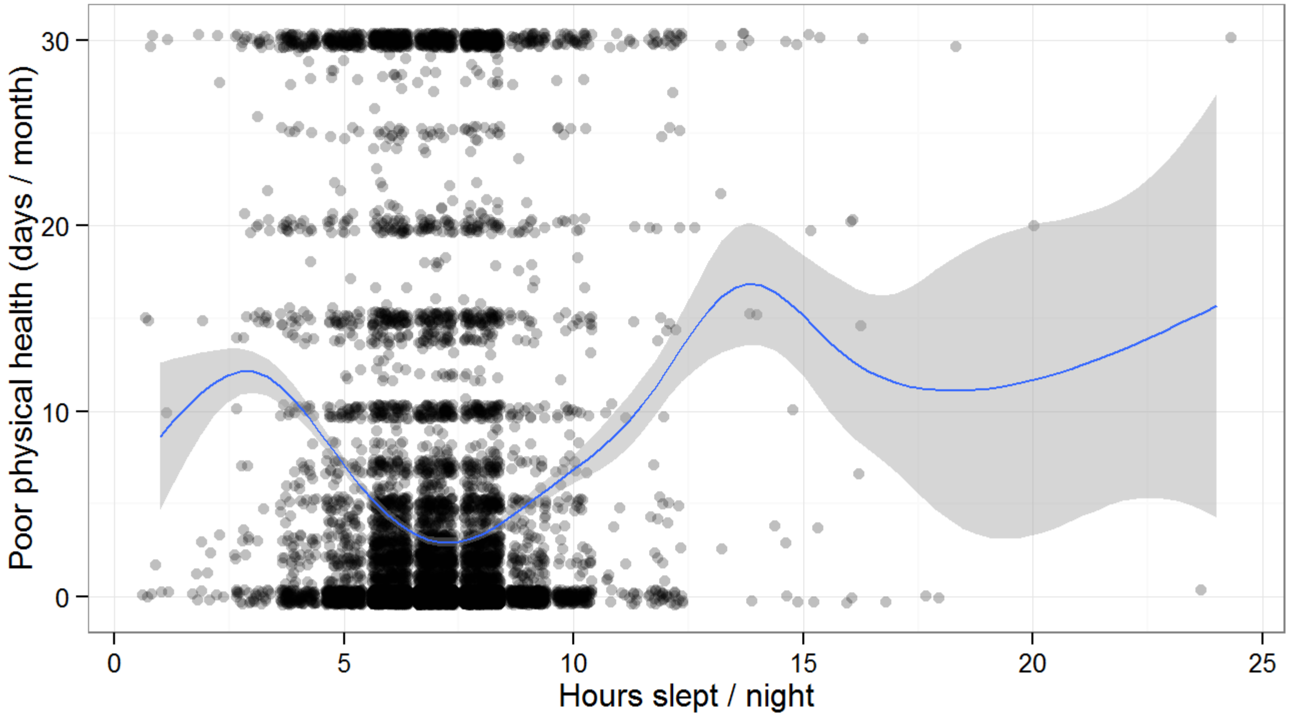
Lets test the correlation between hours slept each night and reported days with poor physical health in the last 30 days

```
ggplot(df, aes(x=sleptim1, y=physhlth)) +
 theme_bw() +
  geom_jitter(alpha=0.25) +
   geom_smooth() +
    labs(x="Hours slept / night", y="Poor physical health (days / month)")
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula:
y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.
```

```
## Warning: Removed 145 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 145 rows containing missing values (geom_point).
```



```
cor(df$sleptim1, df$physhlth, use="complete.obs")
```

```
## [1] -0.04374059
```

There appears to be no clear correlation in the exploratory graph. Also the coefficient of correlation `r=-0.04` is not significant.

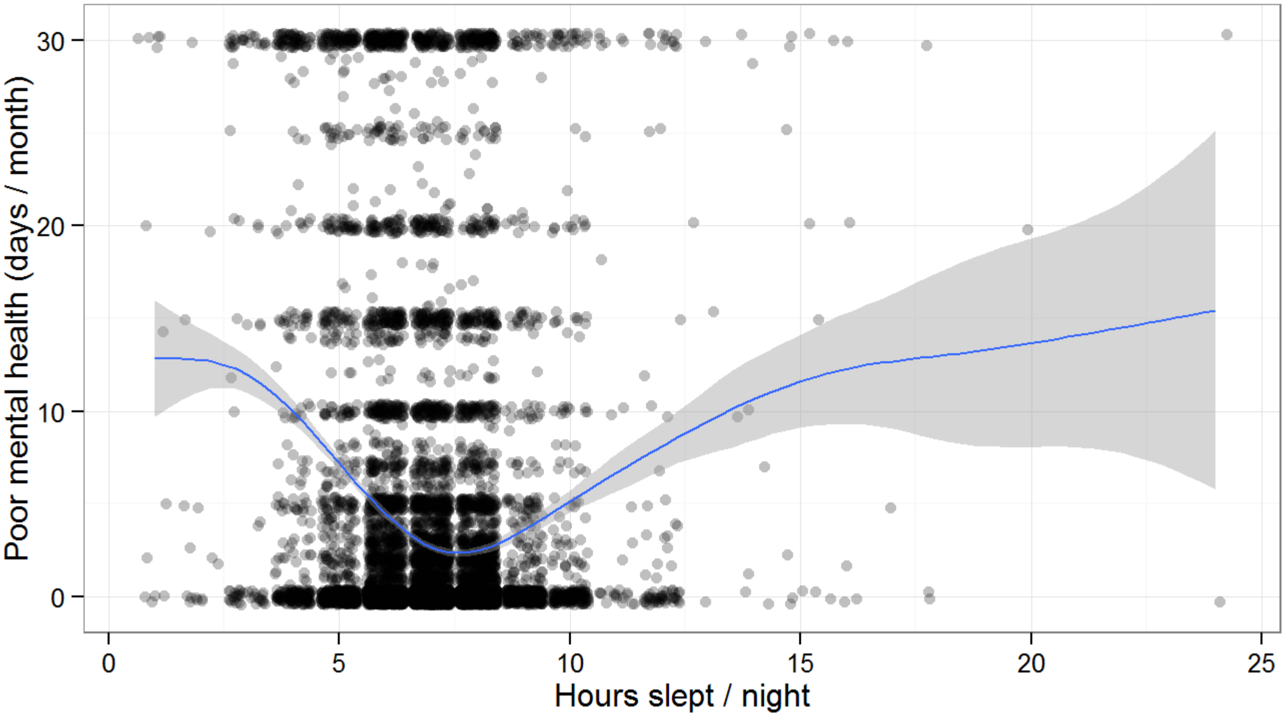
Now, lets test the correlation between hours slept each night and reported days with poor mental health in the last 30 days

```
ggplot(df, aes(x=sleptim1, y=menthlth)) +
 theme_bw() +
  geom_jitter(alpha=0.25) +
   geom_smooth() +
    labs(x="Hours slept / night", y="Poor mental health (days / month)")
```

```
## geom_smooth: method="auto" and size of largest group is >=1000, so using gam with formula:
y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.
```

```
## Warning: Removed 162 rows containing missing values (stat_smooth).
```

```
## Warning: Removed 162 rows containing missing values (geom_point).
```



```
cor(df$sleptim1, df$menthlth, use="complete.obs")
```

```
## [1] -0.1152792
```

Same thing, no visual correlation or significant coefficient can be deducted `r=-0.11`.

Conclusion: *Even though there is no linear relationship between the hours slept and both dependent variables, there is a dip in the lowess regression line plotted which leads to infer that people on both extremes of the spectrum might have a higher number of days with poor physical or mental health.*

Research quesion 3:

Does a correlation exist between the body mass index (BMI) of the individual answering the survey and having a high blood cholesterol or having a heart attack?

The last two digits of this variable are implied decimal places, before doing anything this corrects to the units: `kg/m^2`

```
df$X_bmi5 = as.numeric(df$X_bmi5)/100
```

Descriptive statistics

```
summary(df$X_bmi5)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   12.53   23.03   25.97   26.99   29.76   75.05     849
```

```
sd(df$X_bmi5, na.rm=TRUE)
```

```
## [1] 5.725637
```

```
table(df$toldhi2) # High cholesterol?
```

```
##
##   Yes   No
## 3782 5684
```

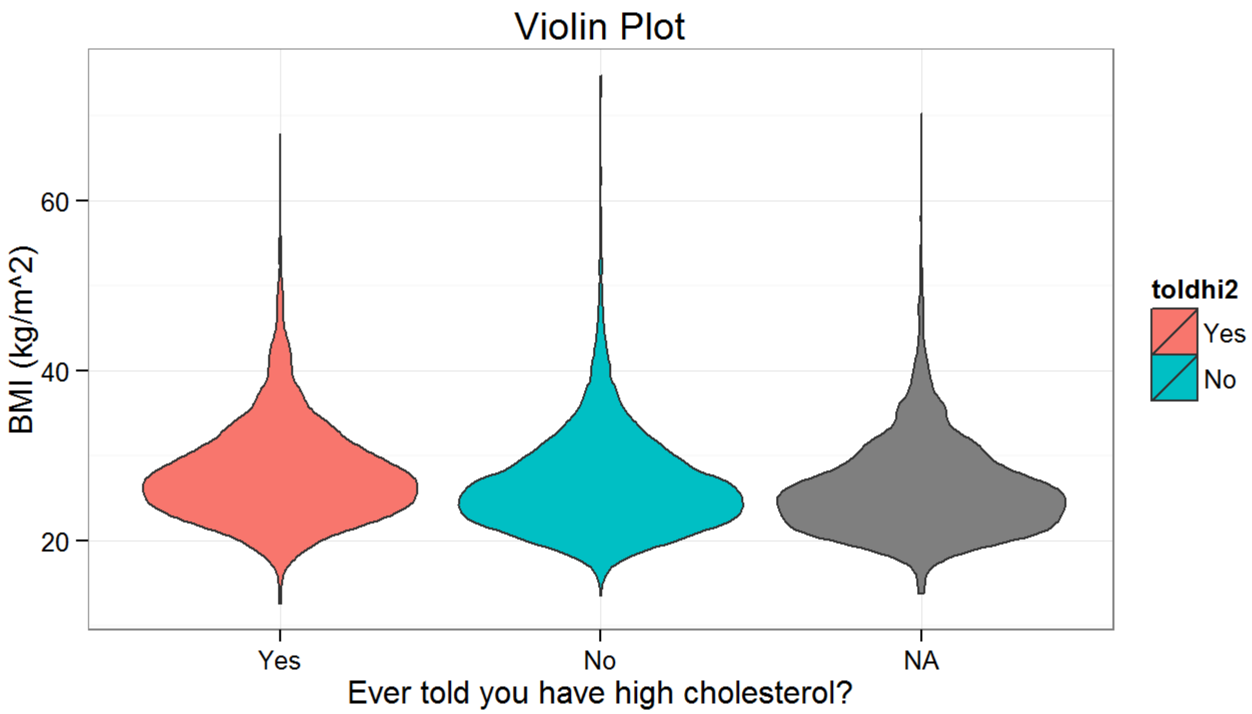
```
table(df$cvdinfr4) #Ever had a heart attack?
```

```
##
##    Yes    No
##    483 11009
```

Exploratory graph for high cholesterol

```
ggplot(df, aes(x=toldhi2, y=X_bmi5, fill=toldhi2)) +
 theme_bw() +
  geom_violin() +
   labs(title="Violin Plot",
        x="Ever told you have high cholesterol?",
        y="BMI (kg/m^2)")
```

```
## Warning: Removed 849 rows containing non-finite values (stat_ydensity).
```



Statistical analysis for high cholesterol

```
temp = filter(df, toldhi2=="Yes" | toldhi2=="No") %>%
            select(X_bmi5, toldhi2)
chisq.test(temp$X_bmi5, temp$toldhi2)
```

```
## Warning in chisq.test(temp$X_bmi5, temp$toldhi2): Chi-squared approximation
## may be incorrect
```

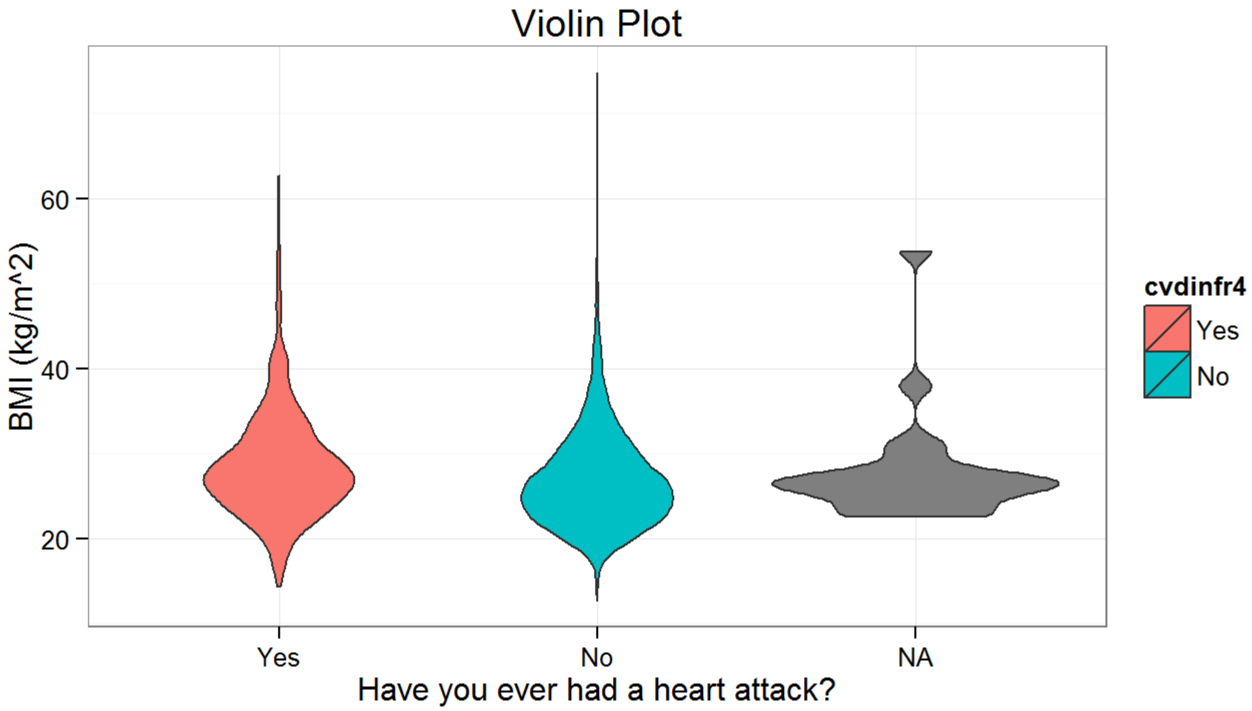
```
##
##  Pearson's Chi-squared test
##
## data:  temp$X_bmi5 and temp$toldhi2
## X-squared = 1555, df = 1368, p-value = 0.0002935
```

`p=2.9e-4` therefore we can reject the Ho that states that there is no correlation between having a higher BMI and suffering hypercholesterolemia.

Exploratory graph for heart attack

```
ggplot(df, aes(x=cvdinfr4, y=X_bmi5, fill=cvdinfr4)) +
 theme_bw() +
  geom_violin() +
   labs(title="Violin Plot",
        x="Have you ever had a heart attack?",
        y="BMI (kg/m^2)")
```

```
## Warning: Removed 849 rows containing non-finite values (stat_ydensity).
```



Statistical analysis for heart attack

```
temp = filter(df, cvdinfr4=="Yes" | cvdinfr4=="No") %>%
            select(X_bmi5, cvdinfr4)
chisq.test(temp$X_bmi5, temp$cvdinfr4)
```

```
## Warning in chisq.test(temp$X_bmi5, temp$cvdinfr4): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  temp$X_bmi5 and temp$cvdinfr4
## X-squared = 1864.4, df = 1449, p-value = 6.473e-13
```

`p=6.47e-13` therefore we can reject the Ho that states that there is no correlation between having a higher BMI and having a heart attack.

Conclusion: ***Having a higher BMI correlates statistically with having hypercholesterolemia or having a heart attack.***