Statistical inference with the GSS data

Setup

Load packages

```r
library(ggplot2)
library(dplyr)
library(statsr)
```

Load data

When the data is loaded, the data file will be called `gss`.

```r
load("gss.Rdata")
```

Get an overview of the data set.

```r
summary(gss)
```

Part 1: Data

This data set contains a lot of information. We probably won't need all of the columns to answer one questions, so we will need to subset our data according the question we want to answer.

The data set is extracted from the GSS conducted by the NORC at the University of Chicago. This data set is cleaned (missing values are removed and equipped with factor variables). The responses are from a representative sample of Americans about a range of topics. There are 57061 responses (the sample size is 57061), which is less than 10% of the population in the United States. Thus it is reasonable to assume the independence of the sampling corresponding to this data set.

Part 2: Research question

Let's come up with our first question: **what is the percentage of US citizen adults (18y/o +) who are employed full (Labor force status) time in year 2008.**

Part 3: Exploratory data analysis

First, notice the min age in this data set is 18 y/o, so we don't need to worry about age column. I used the following python code to get our data cleaned.

```python
import pyreadr
import numpy as np
import pandas as pd

GSS = pyreadr.read_r("/Users/qianj/Documents/R_coursera/Coursera-Interential-Stats/gss.Rdata")
GSS
```

```
gss = GSS['gss']
us08 = gss[['year','uscitzn','wrkstat']]
us08 = us08[us08['year']==2008]
us08 = us08[us08['wrkstat']=='Working Fulltime']
us08 = us08['uscitzn']
us08.replace({'A U.S. Citizen':1},inplace=True)
us08.replace({'Not A U.S. Citizen':0},inplace=True)
us08.replace({'A U.S. Citizen Born In Puerto Rico, The U.S. Virgin Islands, Or The Northern Marianas Isl
us08.replace({'Born Outside Of The United States To Parents Who Were U.S Citizens At That Time (If Volun
us08.dropna(inplace=True)
us08
```

Let's load the cleaned data into R and take a look at the first 10 rows, where the integer 1 represents $A.U.S.Citizen$ and $0 = NotAU.S.Citizen$:

```
us08 <- read.csv("/Users/qianj/Documents/R_coursera/Coursera-Interential-Stats/gss_usct08.csv")
head(us08,10)
```

```
##        X uscitzn
## 1  51021       1
## 2  51024       0
## 3  51027       1
## 4  51028       1
## 5  51031       1
## 6  51036       0
## 7  51038       1
## 8  51041       1
## 9  51046       1
## 10 51048       1
```

It looks like the data's original response id is still in one columne, let's get rid of it by subsetting:

```
us08 = us08[c('uscitzn')]
```

Part 4: Inference

Let's first calculate **a 90% confidence interval of the proportion/percentage** of U.S. Citizens who are full time employed adults among the sample. There are a total of 147 responses.

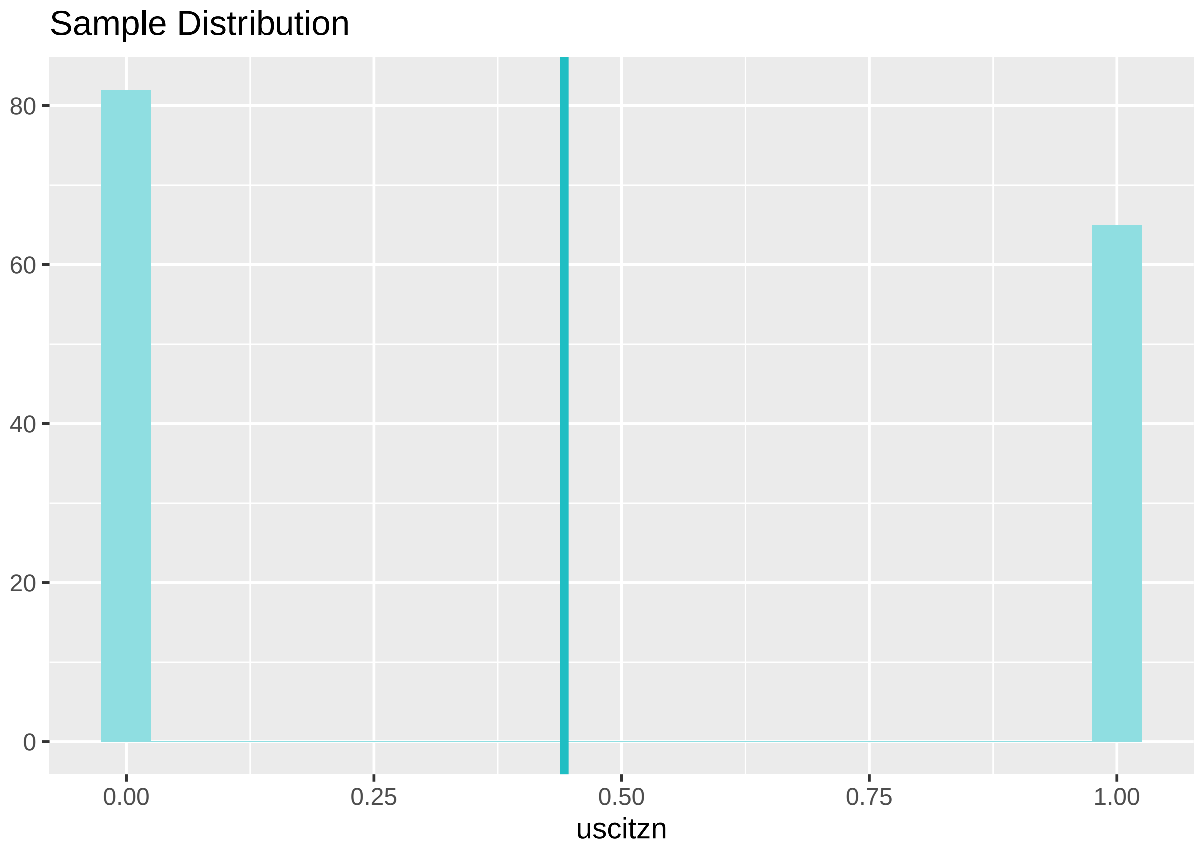
```
dim(us08)
```

```
## [1] 147   1
```

Since the data is now recorded in 1 and 0, the mean can represent the percentage/proportion in our context. It satisfies the conditions for CLT. A 90% confidence interval for the proportion of US citizen adult working full time in 2008 among all adult working full time in 2008 is $(0.3741, 0.5102)$:

```
inference(uscitzn,data=us08,statistic='mean',type='ci',method='theoretical',conf_level = 0.9)
```

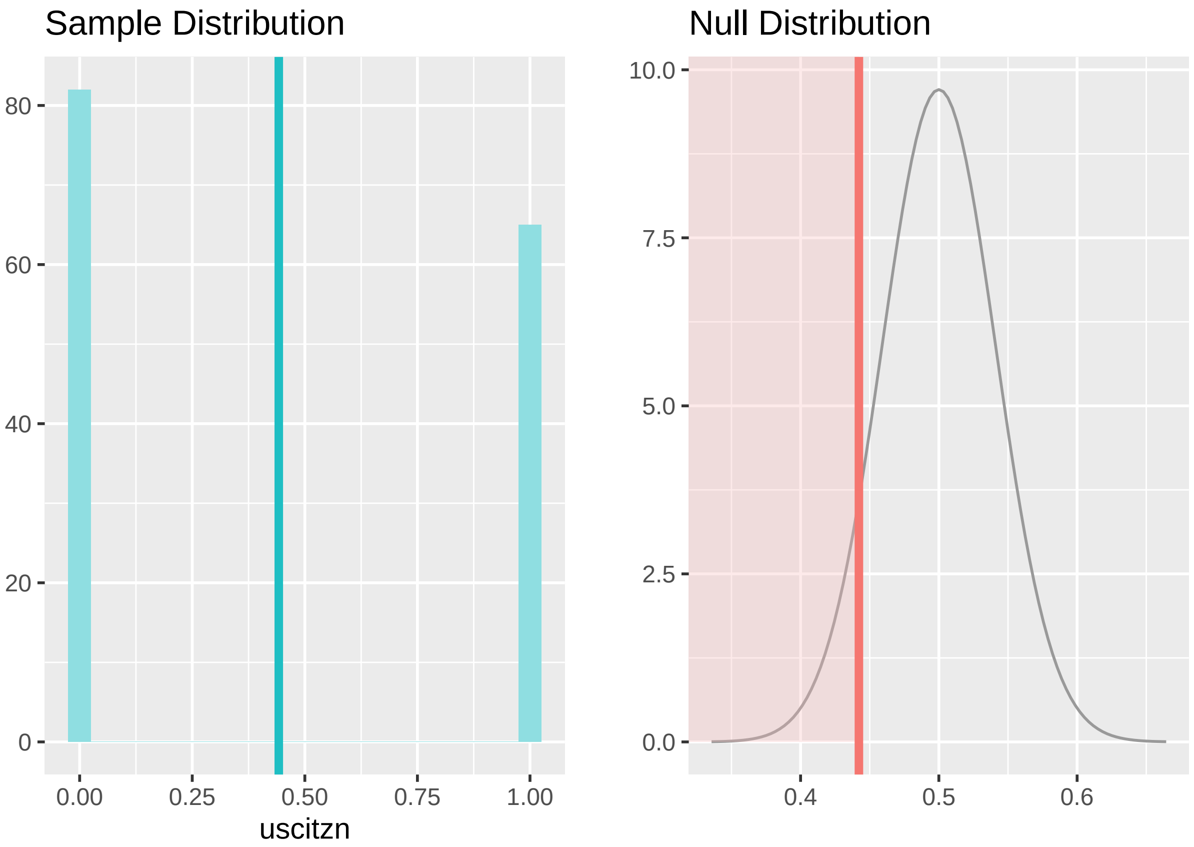
```
## Single numerical variable
## n = 147, y-bar = 0.4422, s = 0.4983
## 90% CI: (0.3741 , 0.5102)
```



Next, let's claim that at most 50 of adults who works full time in 2008 are U.S. Citizens. In this case, our null hypothesis is $H_0 = \underline{\mu = 0.5}$ and the alternative is $H_A = \underline{\mu < 0.5}$. Let's consider a standard $\alpha = 0.05$ significance level.

```
inference(uscitzn,data=us08,statistic='mean',type='ht',alternative='less',method='theoretical',null=0.5
```

```
## Single numerical variable
## n = 147, y-bar = 0.4422, s = 0.4983
## H0: mu = 0.5
## HA: mu < 0.5
## t = -1.4068, df = 146
## p_value = 0.0808
```



We get $p-$value $= 0.0808 > \alpha$, thus we don't have enough evidence to suggest out null hypothesis is false; we accept our null hypothesis.