The effect of milk on LDL Cholesterol level

Huong Tran

11/26/2021

Objective:

- Low-density lipoprotein (LLD) cholesterol level is the main reason causing heart disease and heart attack.
- Consuming whole-fat dairy product can have unwanted health effect of increasing LDL cholesterol level
- This project identifies the affect of several types of milk on LDL cholesterol level.

Data Resource:

- From National Health and Nutrition Examination Survey (NHANES), year 2017 - 2018.
- Population: 320,842,721
- Sample size: 16,211
- The screener response rate: 90.9%
- 9,254 completed the interview and 8,704 were examined.

About Survey Design:

The following components are used:

- Questionnaire Data: we extract information about the choice of milk.
- Laboratory Data: the measure of LDL cholesterol (mg/dL) using the standard Fredewald equation.

Note: triglyceride less than 400mg/dL.

- Demographic: sample weight, PSUs, stratum information.
- Using "SEQN" to merge data.

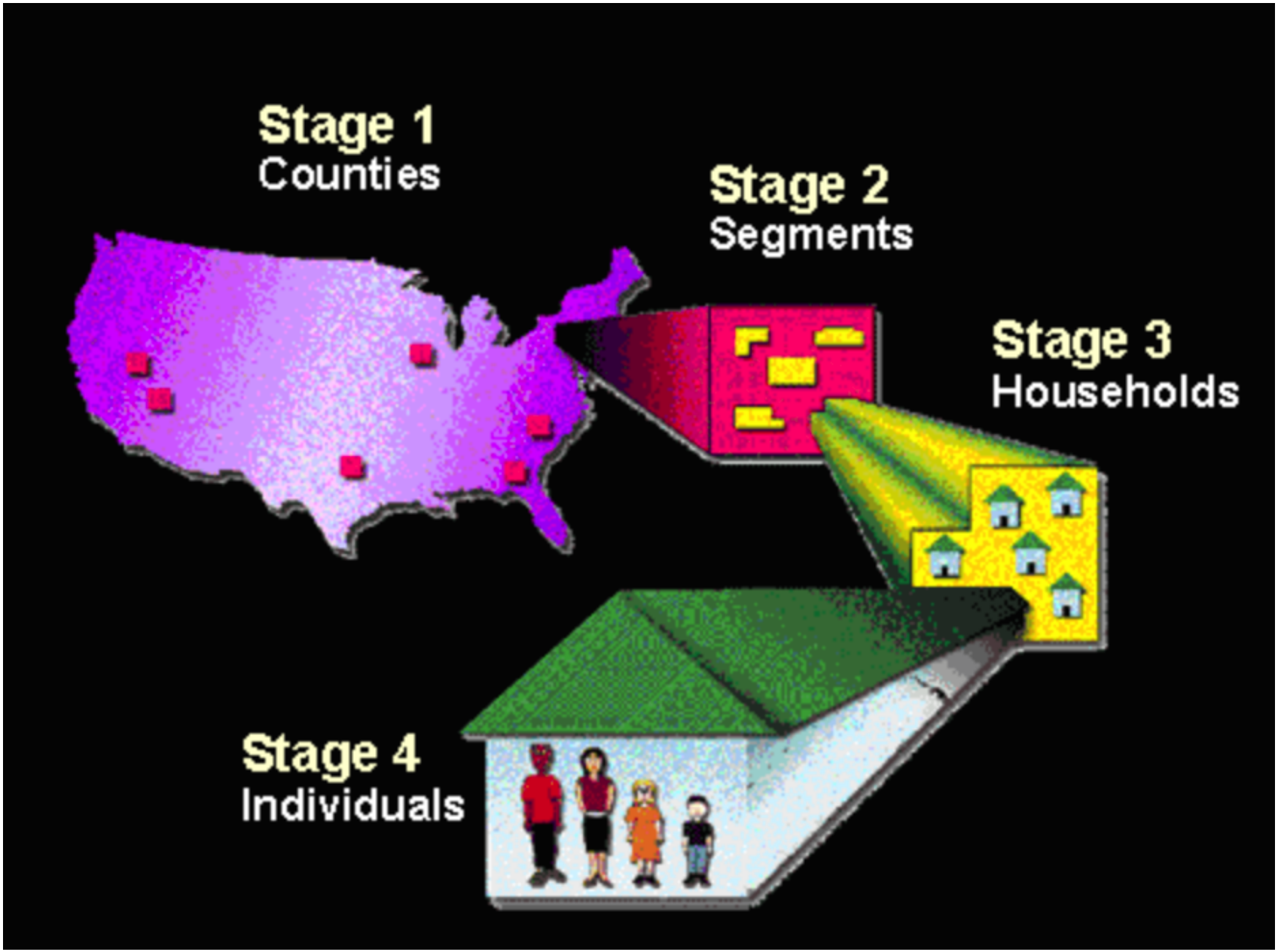
About Survey Design:

- Selection of PSUs, which are counties: these PSUs were selected from strata defined by geography, metropolitan statistical area status, and various population demographics. Two PSUs were selected from most strata.

  Note that if counties has the population of less than 5000, it will be combined with the adjacent counties to have the required number of population.

- Selection of segments within PSUs, that constitute a block/or group of blocks containing a cluster of households.
- Selection of specific households with segments, generally are city blocks.
- Selection of individual within household.

About Survey Design:



Demographics component:

- Demographics:
  - SDMVSTRA: masked variance unit pseudo-stratum.
  - SDMVPSU: massked variance unit psedo-PSU.
  - WTMEC2YR: sampling weight.

Questionnaire component:

- Questionnaire components:
  - DBQ223A: You drink whole or regular milk.
  - DBQ223B: You drink 2% fat milk.
  - DBQ223C: You drink 1% fat milk.
  - DBQ223D: You drink fat free/skim milk.
  - DBQ223E: You drink soy milk.
  - DBQ223U: You drink another type of milk.

Laboratory component:

- Laboratory Data: LBDLDL: measure of LDL cholesterol (mg/dL) using the standard Fredewald equation.

Restriction:

- We are ignoring the requirement for triglyceride less than 400mg/dL in this project.
- The cholesterol measure is valid for 12-year-older participants while the questionnaire for Diet and Behaviour are subjected for all participant.

Data Exploratory:

	SEQN	SDMVPSU	SDMVSTRA	WTMEC2YR	DBQ223A	DBQ223B	DBQ223C	DBQ223D	DBQ223E	DBQ223U	LBDLDL
1	93703	2	145	8539.731	NA	NA	12	NA	NA	NA	NA
2	93704	1	143	42566.615	NA	NA	12	NA	NA	NA	NA
3	93705	2	145	8338.420	NA	NA	NA	NA	NA	30	NA
4	93706	2	134	8723.440	NA	NA	NA	NA	NA	NA	NA
5	93707	1	138	7064.610	NA	NA	NA	NA	NA	NA	NA
6	93708	2	138	14372.489	10	NA	NA	NA	NA	NA	109
7	93709	1	136	12277.557	NA	NA	NA	NA	NA	NA	NA
8	93710	1	134	16848.020	NA	NA	NA	NA	NA	NA	NA
9	93711	2	134	12390.920	NA	NA	NA	NA	NA	30	156
10	93712	2	147	30336.654	NA	11	NA	NA	NA	NA	NA
11	93713	1	140	166841.661	NA	11	NA	NA	NA	NA	NA
12	93714	1	147	15479.581	NA	11	NA	NA	NA	NA	NA

Data Exploratory:

- Some observations give answer in at least 2 types of milk:

```
select(df, "DBQ223A", "DBQ223B", "DBQ223C", "LBDLDL")[21:23,]
```

```
##    DBQ223A DBQ223B DBQ223C LBDLDL
## 21      NA      NA      NA     NA
## 22      10      11      NA     NA
## 23      NA      11      NA     NA
```

Data Exploratory:

- Gather observations giving answer on at least 2 types of milk.
- Create new varialbe called “Category”:

```
as.matrix(unique(df$Category))
```

```
##        [,1]
##  [1,] "1% fat"
##  [2,] "another"
##  [3,] NA
##  [4,] "whole milk"
##  [5,] "2% fat"
##  [6,] "2 types"
##  [7,] "fat free"
##  [8,] "soy milk"
##  [9,] "don't know"
```

Data Exploratory:

- Delete the others columns:

```
df <- select(df, - starts_with("DBQ223"))
head(df)
```

```
##    SEQN SDMVPSU SDMVSTRA  WTMEC2YR LBDLDL   Category
## 1 93703       2      145  8539.731     NA     1% fat
## 2 93704       1      143 42566.615     NA     1% fat
## 3 93705       2      145  8338.420     NA    another
## 4 93706       2      134  8723.440     NA       <NA>
## 5 93707       1      138  7064.610     NA       <NA>
## 6 93708       2      138 14372.489    109 whole milk
```

Survey Design:

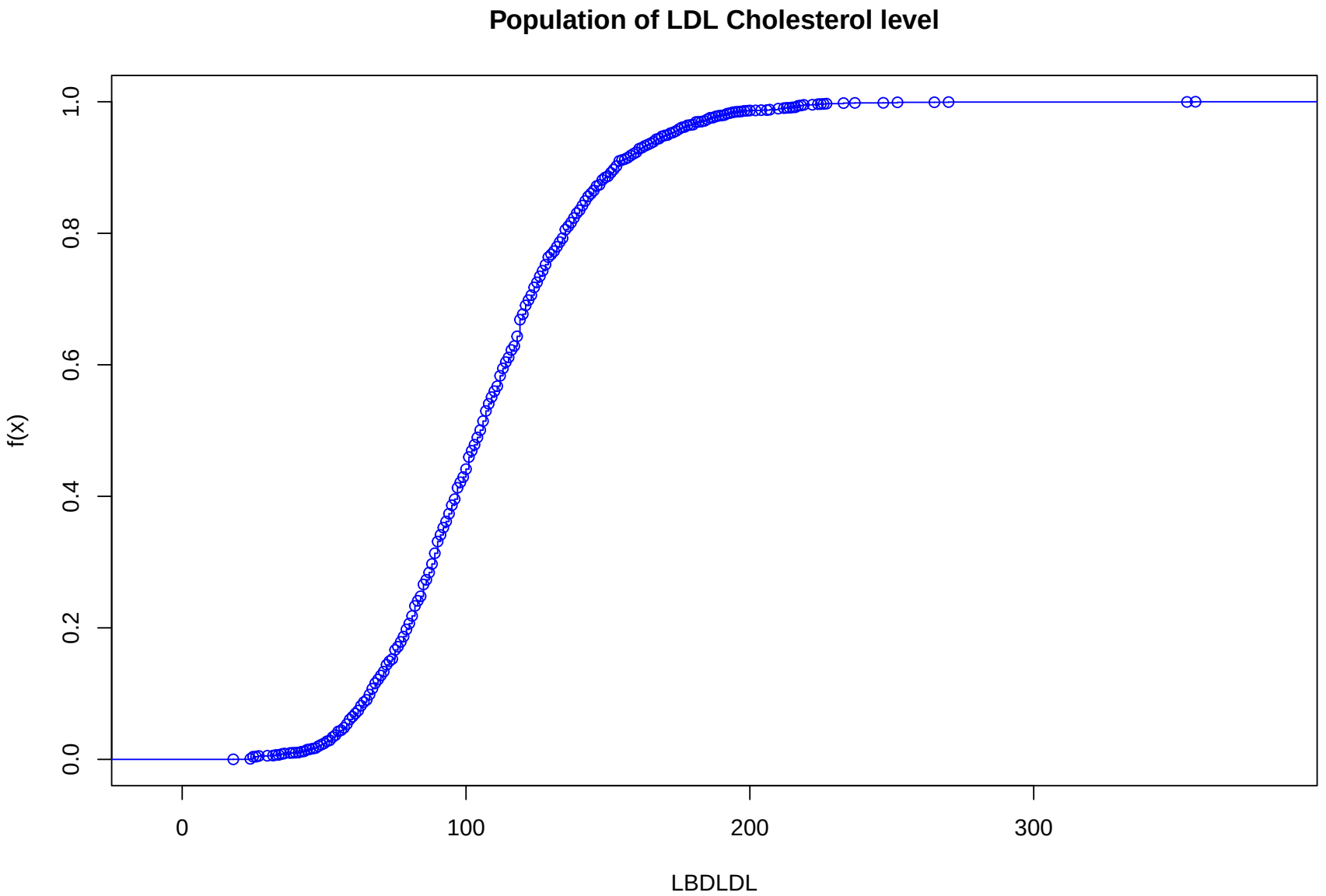
```
sv <- svydesign(id=~SDMVPSU,  strata =~SDMVSTRA,
                data=df, weights=~WTMEC2YR,nest=TRUE)

deff(df$LBDLDL, cluster = df$SDMVPSU)
```

```
##              n     clusters              rho             deff
## 2.808000e+03 2.000000e+00 4.914562e-03 7.909468e+00
```

Which implies that our design is less precise than SRS.

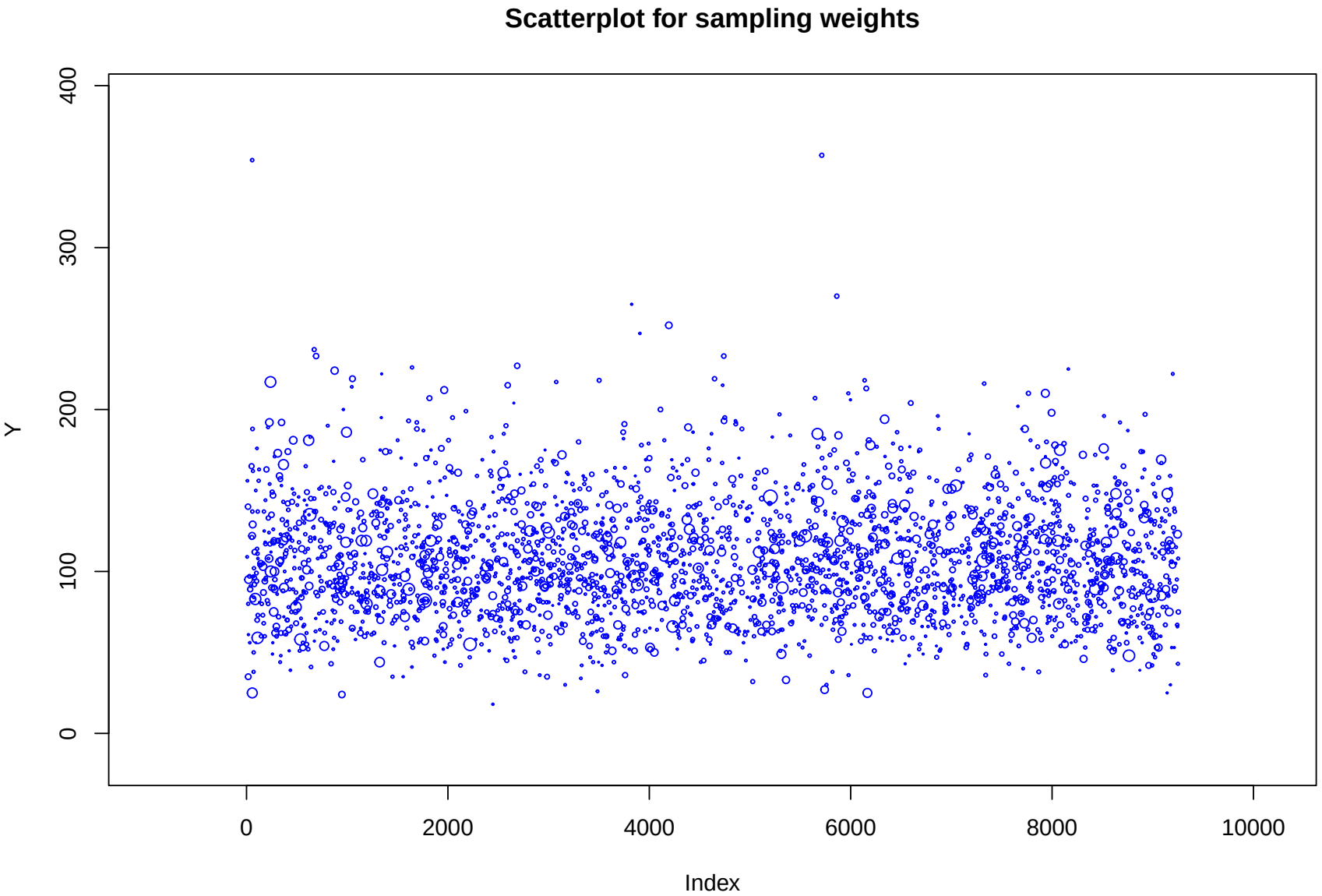
About LDL Cholesterol:

```
plot(svycdf(~LBDLDL, sv),
     main= "Population of LDL Cholesterol level", col = "blue")
```



About LDL Cholesterol:

```
svyplot(~LBDLDL, sv, basecol = "blue",
        main = "Scatterplot for sampling weights")
```



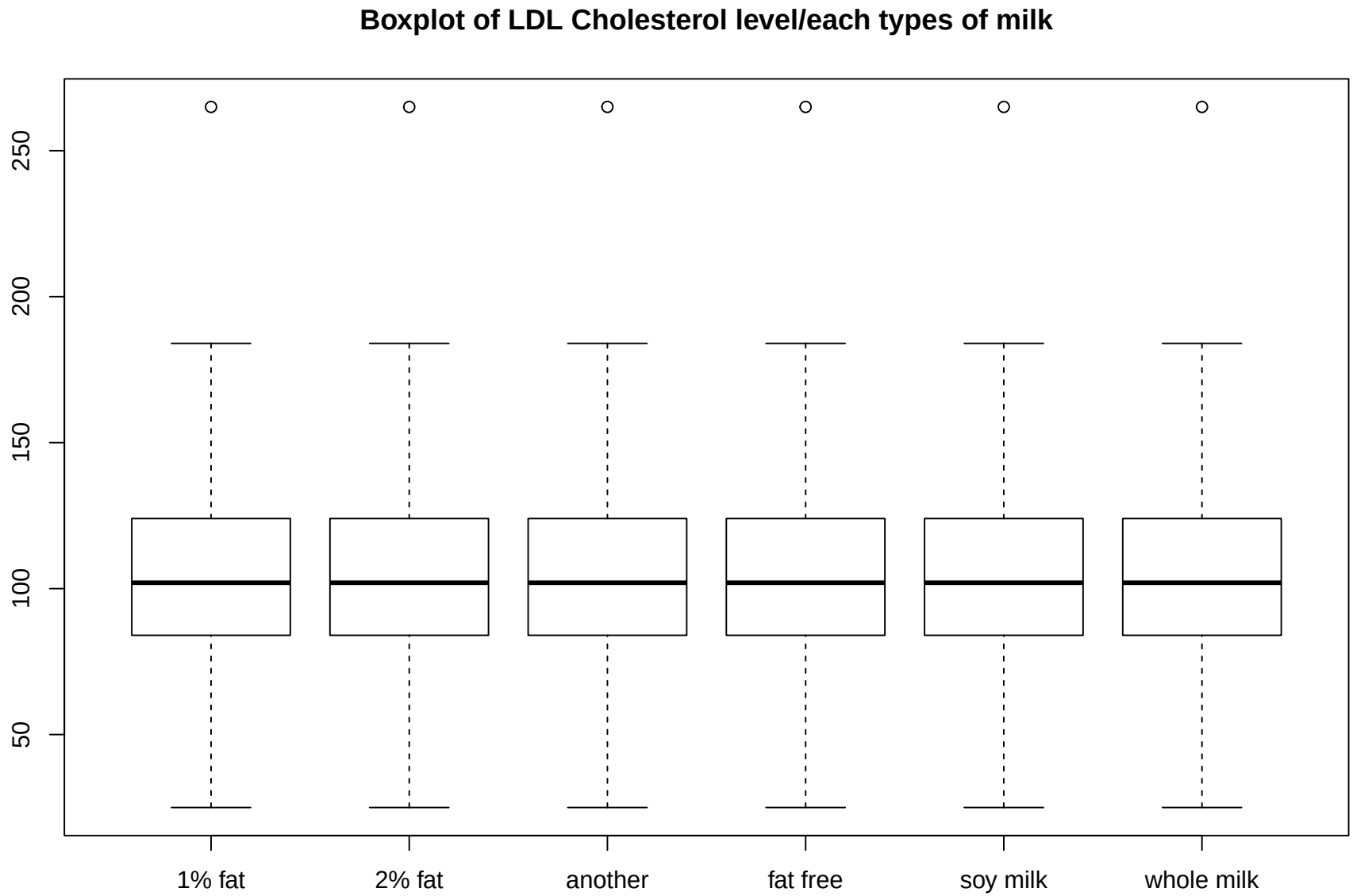
Survey Design:

- Consider the survey on sub-population for observations having their answer on a specific type of milk:

```
sub.sv <- subset(sv, is.na(df$Category) == F &
                     df$Category != "2 types" &
                     df$Category != "don't know")
```

Estimate of mean:

```
svyboxplot(LBDLDL ~ Category, design = sub.sv,
           main = "Boxplot of LDL Cholesterol level/each types
```



Estimate of mean and confidence interval:

```
mean <- svyby(~LBDLDL,~Category, design =  sub.sv,
              svymean, na.rm = T)
select(cbind(mean, confint(mean)), -"Category")
```

```
##                LBDLDL       se    2.5 %   97.5 %
## 1% fat       107.5708 2.078935 103.4961 111.6454
## 2% fat       107.8348 2.328826 103.2704 112.3993
## another      108.9764 3.470696 102.1739 115.7788
## fat free     108.1012 3.796844 100.6595 115.5428
## soy milk     107.5362 3.778740 100.1300 114.9424
## whole milk   106.5499 2.499380 101.6512 111.4486
```

Using Regression:

```
model <- svyglm(LBDLDL ~ Category, design = sub.sv )
s <- summary(model)$coefficients
print(s, digits = 2)
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           107.571        2.1 51.7432  1.8e-13
## Category2% fat          0.264        3.0  0.0868  9.3e-01
## Categoryanother         1.406        4.5  0.3132  7.6e-01
## Categoryfat free        0.530        4.8  0.1113  9.1e-01
## Categorysoy milk       -0.035        4.1 -0.0084  9.9e-01
## Categorywhole milk     -1.021        3.1 -0.3326  7.5e-01
```

Whole milk and non-fat milk:

Test the following hypothesis:

$$H_0 : \bar{y}_{10} = \bar{y}_{13}$$

```
sub.sv2 <- subset(sv, df$Category == "whole milk" |
                      df$Category == "fat free")
svyttest(LBDLDL~Category, sub.sv2)
```

```
##
##  Design-based t-test
##
## data:  LBDLDL ~ Category
## t = -0.30554, df = 14, p-value = 0.7644
## alternative hypothesis: true difference in mean is not equal
## 95 percent confidence interval:
##  -12.440791   9.338198
## sample estimates:
## difference in mean
##          -1.551297
```

Whole milk and non-fat milk:

$$p - \text{value} = 0.7644 > 0.05$$

Since the computed $p - value$ does not fall in the rejection region, we fail to reject H_0. There is insufficient evidence (at $\alpha = 0.05$) of a difference between the true mean LDL Cholesterol between people who drink whole milk and people who drink fat free milk.

Further question:

1. The requirement of Fredewald equation?
2. Considering the effect of observations who drink at least 2 types of milk at the same time.

Reference:

1. Sampling: Design and Analysis, Second Edition - Sharon Lohr.
2. https://cran.r-project.org/web/packages/survey/survey.pdf
3. https://wwwn.cdc.gov/nchs/nhanes/tutorials/default.aspx

Thank you!