

OAKLAND UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

STA5229: STAT METHOD IN SAMPLE SURVEY

Milk and LDL Cholesterol level

Professor: Dorin Drignei

Student: Huong Tran

November 18, 2021

Contents

1	Abstract:	2
1.1	Objective:	2
1.2	Data overviews:	2
1.3	Methods:	3
2	Data exploratory:	3
3	Comparison between the types of milk:	3
4	Further question:	3

1 Abstract:

1.1 Objective:

Consuming whole-fat dairy products might cause the increase in low-density lipoprotein (LDL) cholesterol level, which cause heart disease and heart attack. This project identifies the affect of several types of milk on LDL cholesterol level. The comparison in LDL cholesterol level on milk-consumers will be derived.

1.2 Data overviews:

The data used to produce insights of the difference between several types of milk in this report is extracted from National Health and Nutrition Examination Survey (NHANES) 2017 - 2018. The following components are used:

- Questionnaire Data: from component Diet and Behaviour, we extract information about the choice of milk. The following variables are used:
 - 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 Data: In this project, we are interested in the measure of LDL cholesterol (mg/dL) using the standard Fredewald equation, which is stored in the variable LBDLDL. Please note that, we are ignoring the requirement for triglyceride less than 400mg/dL in this project.

Note that, the cholesterol measure is valid for 12-year-old participants while the questionnaire for Diet and Behaviour are subjected for all participant. Therefore, weights of sample should be taken into account carefully.

- Demographics: Because NHANES uses complex surveys designs, sample weights should be taken into account carefully:
 1. Selection of PSUs, which are counties. 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.
 2. Selection of segments within PSUs, that constitute a block/or group of blocks containing a cluster of households.
 3. Selection of specific households with segments.

4. Selection of individual within household.

Data for cholesterol level were collected in the MEC exam, therefore variable WT-MEC2YR (Full sample 2 year MEC exam weight) must be used.

1.3 Methods:

To observe the effect of various types of milk on consumers, we stratify the population based on the type of milk they are drinking.

2 Data exploratory:

First, we download data from NHANES website:

```
milk <- read.xport("dataset/DBQ_J.xpt")
choles <- read.xport("dataset/TRIGLY_J.xpt")
w <- read.xport("dataset/DEMO_J.xpt")

# save as an R data frame
saveRDS(milk, file="dataset/DEMO_I.rds")
saveRDS(choles, file="dataset/TRIGLY_J.rds")
saveRDS(w, file="dataset/DEMO_J.rds")
```

In NHANES data, the unique identifier for each sample person is known as the sequence number (SEQN). Using SEQN, we are able to merge the three dataframes:

```
df.tem <- merge(w, milk, by.x = "SEQN", by.y = "SEQN")
df.join <- merge(df.tem, choles, by.x = "SEQN", by.y = "SEQN")
myvars <- c("SEQN", "DBQ223A", "DBQ223B", "DBQ223C", "DBQ223D",
            "DBQ223E", "DBQ223U", "LBDLDL", "WTMEC2YR")
df <- df.join[myvars]
```

Some statistics of LDL Cholesterol:

```
summary(df$LBDLDL)
```

The result is as follow:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
18.0	82.0	103.0	106.9	128.0	357.0	228

3 Comparison between the types of milk:

4 Further question:

References