Statistical Analysis of Diabetes Health Indicators

STAT 344 Project

Geng Chen(leader)
Student ID: 51264893
xxxx
Student ID: xxxxx
xxxx
Student ID: xxxxx
Instructor: Wu Lang

Contributions:
Geng Chen:Group leader;Writing SRS method analysis and R Code;
Providing the LATEXTemplate; Writing part of the group project report;

November 11, 2023

Part 1

Introduction

Background:

In the 21st century, we are experiencing an unprecedented transformation in the way we live, work, and interact with our environment. While these changes have led to advances in convenience and comfort, they have also brought public health crises. One of the most significant health concerns nowadays is the relationship between diabetes and lifestyle. Diabetes, a chronic metabolic disorder characterized by elevated blood sugar levels, has become a pervasive and pressing issue that affects millions of people worldwide (CDC, 2017).

Objective:

Our project will explore factors connecting diabetes, in order to have a better understanding of the relationship between diabetes and lifestyle in modern society. Recognizing the causes and consequences of diabetes can contribute to developing effective strategies to prevent certain diseases and promote a healthier, more sustainable way of life.

Data information:

We chose the CDC Diabetes Health Indicators Dataset which is generated by the CDC (Communicable Disease Center). It was collected by an annual health-related telephone survey. The dataset contains healthcare statistics and lifestyle survey data for the general population, with information about their diabetes diagnosis. The classification target variable is the individual's health status, categorizing them as having diabetes (=2), being pre-diabetic state (=1), or healthy (=0) (TEBOUL, 2015). Moreover, there are 21 feature variables in the dataset that may be related to the diabetes diagnosis (e.g. Gender, BMI, Age, ……).

Dataset overview:

The population of the dataset is 253680. We count both having diabetes and being pre-diabetic state as diabetic patients, and we use R to get the population proportion of diabetic patients (0.158).

Sample Size Calculation

In this study, we employed Simple Random Sampling (SRS) to estimate the population mean for a continuous variable (BMI) and the population proportion for a binary variable (Smoker) from a dataset concerning diabetes health indicators.

Continuous Data (BMI):

For continuous data like BMI, we often want to estimate the mean of the population. The formula used to calculate the sample size for estimating a mean is given by:

$$n = \left(\frac{Z \times \sigma}{E}\right)^2$$

Where: - n is the sample size. - Z is the Z-score corresponding to the desired confidence level. For a 95% confidence level, the Z-score is approximately 1.96. This value comes from the standard normal distribution, representing the number of standard deviations a point must be from the mean to enclose 95% of the data. - σ is the standard deviation of the population. In your case, the standard deviation for BMI is approximately 6.61. - E is the margin of error. You've chosen a margin of error of ±2 units for BMI.

Using these values, the sample size for BMI was calculated as:

$$n = \left(\frac{1.96 \times 6.61}{2}\right)^2 \approx 42$$

This means you would need a sample size of about 42 individuals to estimate the average BMI with a margin of error of ±2 units at a 95% confidence level.

Binary Data (Smoker):

For binary data, like determining whether an individual is a smoker or not, the formula for calculating the sample size to estimate a proportion is:

$$n = \frac{Z^2 \times p \times (1-p)}{E^2}$$

Where: - p is the estimated proportion of the attribute in the population. In this case, it's the proportion of smokers, which is approximately 44.32%. - E is the margin of error, set at ±0.05 (5%).

Using these values, the sample size for the smoker variable was calculated as:

$$n = \frac{1.96^2 \times 0.4432 \times (1 - 0.4432)}{0.05^2} \approx 379$$

This suggests you need a sample size of about 379 individuals to estimate the proportion of smokers with a 5% margin of error at a 95% confidence level.

These calculations ensure that your sample is large enough to provide reliable estimates of the mean BMI and the proportion of smokers in the population with the specified precision and confidence.

Methodology

Method 1: Simple Random Sampling

In our study on the relationship between diabetes and lifestyle, we utilized Simple Random Sampling (SRS) to estimate the population mean for a continuous variable (BMI) and the population proportion for a binary variable (Smoker). SRS was chosen for its simplicity and ability to yield unbiased estimates.

For BMI, a sample size of 42 was selected, adhering to the principles of the Central Limit Theorem. This theorem suggests that with a large enough sample size, the distribution of the sample mean approximates a normal distribution, regardless of the population's actual distribution.

Continuous Variable (BMI): The sample mean ($\bar{x}$) and standard error (SE) of the sample mean were calculated. The SE is crucial as it represents the variability of the sample mean and is given by:

$$SE(\bar{x}) = \sqrt{(1 - \frac{n}{N}) \cdot \frac{s^2}{n}}$$

Here, s is the sample standard deviation, n is the sample size (42), and N is the population size. The Finite Population Correction (FPC) is applied due to the ratio of the sample size to the population. The 95% confidence interval for the mean BMI is then:

$$CI = \bar{x} \pm z \cdot SE(\bar{x})$$

with z as the z-score for 95% confidence.

Binary Variable (Smoker): With a sample size of 379 for the binary variable, we determined the sample proportion ($\hat{p}$) and its standard error. The SE for the proportion is calculated using:

$$SE(\hat{p}) = \sqrt{(1 - \frac{n}{N}) \cdot \frac{\hat{p}(1 - \hat{p})}{n}}$$

The 95% confidence interval for the population proportion of smokers is:

$$CI = \hat{p} \pm z \cdot SE(\hat{p})$$

```
[1] 28.52381

$se_bmi [1] 0.8299979

$prop_smoker [1] 0.4248021

$se_smoker [1] 0.02537218

$BMI_CI [1] 26.89704 30.15058

$Smoker_CI [1] 0.3750736 0.4745307
```

This method's primary advantage is its straightforward and unbiased nature, but it may lack efficiency compared to methods like stratified sampling, especially in diverse populations. Additionally, it requires a complete list of the population, which might not always be available. The primary advantage of SRS is its straightforward implementation and the assurance of unbiased estimates, provided that the samples are selected randomly. However, SRS may not be as efficient as other methods, such as stratified sampling, when there are clear subgroups within the population. Moreover, SRS necessitates a complete list of the population, which may not always be feasible or practical.

Data Analysis

Your data analysis section goes here…

Discussion

Your results section goes here…

Conclusion

Your conclusion goes here…

Part 2

The paper "The Emperor's New Tests" by Michael D. Perlman and Lang Wu discusses a significant issue in statistical hypothesis testing, particularly focusing on the usage and criticism of Likelihood Ratio Tests (LRTs). The paper primarily challenges the prevailing notion in statistical circles that traditional likelihood ratio tests (LRTs) are inferior in multi-parameter hypothesis testing. The authors argue that the criticism of LRTs being biased and less effective compared to newer tests is flawed. They contend that these newer tests, often hailed as superior, can lead to scientifically unacceptable conclusions. Perlman and Wu emphasize that the goal of hypothesis testing is not just to adhere to specific statistical criteria like minimizing bias or maximizing power. Instead, it's about using the data to assess the evidence within the context of the proposed hypotheses. In practical terms, this means that despite some criticisms, LRTs remain a reasonable and sensible choice for evaluating statistical hypotheses, especially in non-Bayesian parametric contexts, where they align well with the actual needs of scientific investigation.

Appendix for part 1

R Code for Calculations:

```r
# Load the dataset
data <- read.csv('diabetes_012_health_indicators_BRFSS2015.csv')
set.seed(0)
# Calculate sample mean and standard error for BMI
n_bmi <- 42
N <- nrow(data)
sample_bmi <- sample(data$BMI, n_bmi)
mean_bmi <- mean(sample_bmi)
std_dev_bmi <- sd(sample_bmi)
se_bmi <- sqrt((1 - n_bmi/N) * (std_dev_bmi^2 / n_bmi))

# Calculate sample proportion and standard error for Smoker
n_smoker <- 379
sample_smoker <- sample(data$Smoker, n_smoker)
prop_smoker <- mean(sample_smoker)
se_smoker <- sqrt((1 - n_smoker/N) * (prop_smoker * (1 - prop_smoker) / n_smoker))

# Confidence intervals
z_score <- qnorm(0.975)
ci_bmi <- c(mean_bmi - z_score * se_bmi, mean_bmi + z_score * se_bmi)
ci_smoker <- c(prop_smoker - z_score * se_smoker, prop_smoker + z_score * se_smoker)

# Output the results
list(mean_bmi = mean_bmi, se_bmi = se_bmi)
```

```
## $mean_bmi
## [1] 28.52381
##
## $se_bmi
## [1] 0.8299979
```

```r
list(prop_smoker = prop_smoker, se_smoker = se_smoker)
```

```
## $prop_smoker
## [1] 0.4248021
##
## $se_smoker
## [1] 0.02537218
```

```r
list(BMI_CI = ci_bmi, Smoker_CI = ci_smoker)
```

```
## $BMI_CI
## [1] 26.89704 30.15058
##
## $Smoker_CI
## [1] 0.3750736 0.4745307
```

CDC. U.S. Diabetes Surveillance System. Atlanta, GA: US Department of Health and Human Services, CDC; 2017. https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html

Teboul, Alex. "Diabetes Health Indicators Dataset." Kaggle, 8 Nov. 2021, www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/.