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 12, 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):

The formula for calculating the sample size to estimate a mean is:

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

Where:

- n is the sample size.

- Z is the Z-score corresponding to the desired confidence level. For a 95% confidence level, this Z-score is approximately 1.96. This number originates from the standard normal distribution, indicating the number of standard deviations a point needs to be from the mean to encompass 95% of the data under the bell curve.
- σ is the standard deviation of the population. In this study, the standard deviation for BMI, as derived from preliminary data analysis, is approximately 6.61.
- E is the margin of error, which in this context is chosen to be ± 2 units for BMI.

Substituting these values into the formula, the calculated sample size for BMI is:

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

Thus, approximately 42 individuals are required to estimate the average BMI within a ± 2 unit margin of error at a 95% confidence level. This sample size is crucial to ensure the representativeness and reliability of the statistical estimates, adhering to the principles of inferential statistics.

Binary Data (Smoker):

For binary data, such as determining the smoking status (Smoker/Non-Smoker), the sample size calculation aims to estimate a population proportion with a given level of precision and confidence.

The formula for this calculation is:

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

Where:

- p represents the estimated proportion of the attribute in the population. In this scenario, the estimated proportion of smokers is approximately 44.32%.
- E is the margin of error, which has been set at ± 0.05 (5%) for this study.

By applying these values:

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

This calculation indicates that a sample size of roughly 379 individuals is required to estimate the proportion of smokers with a 5% margin of error at a 95% confidence level.

The determination of this sample size is fundamental for achieving statistical significance and ensuring that the estimated proportion accurately reflects the broader population's characteristics. It is a critical step in the research methodology that underpins the validity and credibility of the study's findings.

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.

- **Estimate (Sample Mean)**: The calculated sample mean for Body Mass Index (BMI) is 28.52. This figure provides an estimated average BMI for the population under study.
- **Standard Error (SE)**: The standard error of the mean BMI is 0.83. Standard error measures the dispersion or variability of the sample mean relative to the true population mean. It is a pivotal element in inferential statistics as it quantifies the precision of the sample mean as an estimator of the population mean.
- **Confidence Interval (CI)**: The 95% confidence interval for the mean BMI is [26.90, 30.15]. This interval suggests that, under repeated sampling, we can be 95% confident that the true population mean BMI falls within this range. The width of this interval is influenced by the standard error and the chosen confidence level, which in this case, reflects a standard normal distribution's 95th percentile.

Interpretation of BMI Results: The average BMI of 28.52, along with a confidence interval ranging from 26.90 to 30.15, indicates a tendency towards higher BMI values within the population. Given that a BMI over 25 is typically classified as overweight, these results may suggest a prevalence of overweight or obesity issues in the studied population. The relative narrowness of the confidence interval denotes a high precision in the estimate, bolstering the reliability of these results.

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:

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

- **Estimate (Sample Proportion)**: The estimated proportion of smokers in the population is 0.425. This proportion is a crucial indicator of the prevalence of smoking habits within the studied group.
- **Standard Error (SE)**: The standard error of the sample proportion is approximately 0.025. Similar to the continuous case, this measure indicates the precision of the sample proportion in estimating the true population proportion.
- **Confidence Interval (CI)**: The 95% confidence interval for the proportion of smokers is [0.375, 0.475]. This interval provides a range within which the true population proportion of smokers is expected to fall with a 95% level of confidence.

Interpretation of Smoker Results: The estimated proportion of smokers (42.5%) is significant, indicating that nearly half of the population engages in smoking. The confidence interval, which is relatively narrow, underlines the precision of this estimate and its potential public health implications, especially when considering smoking's known health risks.

Assumptions and Methodological Considerations:

- **Random Sampling**: Both analyses assume that the samples were randomly drawn from the population, ensuring representativeness and the applicability of inferential statistics.
- **Central Limit Theorem**: The application of the Central Limit Theorem is assumed, particularly for the calculation of confidence intervals, implying that the sampling distribution of the mean (or proportion) approaches normality as the sample size increases.
- **Finite Population Correction**: This was applied in the standard error calculations, accounting for the sample size's proportion relative to the overall population size.

Advantages and Disadvantages:

- **Simple Random Sampling (SRS)**:
  - **Advantages**: SRS ensures each member of the population has an equal chance of being selected, providing an unbiased sample. It is straightforward and easy to understand, making it a widely accepted method in statistical practice.
  - **Disadvantages**: SRS may not be the most efficient method, especially in heterogeneous populations. It may require larger sample sizes to achieve the same level of precision as more complex sampling techniques like stratified sampling. Additionally, SRS presupposes the availability of a complete population list, which can be impractical or unfeasible in some scenarios.

Conclusion:

The statistical analysis of the BMI and smoking data provides valuable insights into the population's health status, with implications for public health policy and interventions. The use of SRS, while methodologically straightforward, has yielded precise and presumably unbiased estimates of these key health indicators. However, the choice of sampling method should always be considered in light of the research objectives and the population's characteristics to ensure the most efficient and accurate data collection approach.

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/.