**Statistical Analysis of Lifestyle Risk Factors and Chronic Disease Outcomes Data Set**

**Case Study - Healthcare Applications**

|  |  |
| --- | --- |
| **Name** | KELLEN HANSEL NOSINA |
| **Roll Number** | RA2211047010137 |
| **Department** | CINTEL |
| **Subject** | Inferential Statistics and Predictive Analytics |
| **Subject Code** | 21AIC401T |

# Abstract

This study utilizes the *Statistical Analysis of Lifestyle Risk Factors and Chronic Disease Outcomes* dataset to explore key associations between health behaviors, demographic variables, and chronic disease indicators. By applying fundamental statistical methods, including t-tests and ANOVA, the analysis investigates how lifestyle-related factors influence health outcomes across different population groups. The findings demonstrate significant differences in body mass index (BMI) across demographic categories such as sex and notable variations in BMI corresponding to self-reported health status. These patterns suggest that lifestyle behaviors and perceptions of health are closely linked to measurable health risks. The results underscore the importance of data-driven public health strategies aimed at early identification and targeted interventions for populations at greater risk of chronic disease.

# Introduction

The increasing prevalence of chronic diseases, such as diabetes and cardiovascular disorders, poses a major public health challenge worldwide. Lifestyle-related factors, including diet, physical activity, and self-perceived health, play a critical role in shaping long-term health outcomes. Among the available indicators, Body Mass Index (BMI) is one of the most widely used measures for assessing weight categories and potential health risks within populations. As a simple yet effective metric, BMI provides valuable insights into overall health status when analyzed in conjunction with demographic and behavioral variables. Specifically, the analysis seeks to address the following research questions:

1. Does the average BMI of the population deviate significantly from the healthy standard?
2. Is there a significant difference in BMI between male and female participants?
3. Do lifestyle and health factors, such as smoking status and self-reported health outcomes, show a measurable relationship with BMI?

## Hypotheses

The following null (𝐻0) and alternative (𝐻1) hypotheses were formulated for statistical testing:

### 1. One-Sample t-Test

* 𝐻0:The mean BMI of the population is equal to the healthy standard of 25 (μ = 25).
* 𝐻1: The mean BMI of the is not equal to the healthy standard of 25 (𝜇 ≠ 25).

### 2. Two-Sample t-Test

* 𝐻0: There is no difference in the mean BMI between male and female participants (𝜇male = 𝜇female ).
* 𝐻1: There is a significant difference in the mean BMI between male and female participants (𝜇male ≠ 𝜇female ).

### 3. One-Way ANOVA

* 𝐻0: The mean BMI is equal across all five self-reported health groups (𝜇1 = 𝜇2 = 𝜇3 = 𝜇4 = 𝜇5).
* 𝐻1: At least one self-reported health group has a mean BMI that is significantly different from the others.

# Methods

## Data Preparation and Variable Selection

The analysis was conducted on a **synthetic healthcare dataset containing 150,000 patient records**. The dataset was designed to reflect a range of demographic, lifestyle, and clinical health indicators commonly associated with chronic illness and lifestyle-related health risks.

For this study, **Body Mass Index (BMI)** was selected as the primary continuous variable, as it serves as a widely used indicator of overweight and obesity, both of which are critical risk factors for diabetes, hypertension, and cardiovascular disease.

* **Continuous Variable:** *BMI* (Body Mass Index).
* **Grouping Variable (2 Levels):** *Gender* (Male/Female), chosen as a fundamental demographic factor to test BMI differences across sexes.
* **Grouping Variable (3+ Levels):** *Self-Reported General Health (GenHlth)*, a 5-point ordinal scale (1 = Excellent to 5 = Poor), selected to investigate whether perceived health status is associated with measurable differences in BMI.

## Statistical Approach

The analysis employed a combination of **descriptive** and **inferential statistical methods** to evaluate relationships between Body Mass Index (BMI) and selected demographic and health-related variables.

### 1. Descriptive Statistics

1. The distribution of BMI was summarized using measures of central tendency (mean, median) and variability (variance, standard deviation).
2. Frequency counts were used to describe categorical variables such as Gender and General Health (GenHlth).

### 2. One-Sample t-Test

1. Objective: To determine whether the mean BMI of the entire dataset significantly differs from the standard healthy threshold of 25.
2. Rationale: BMI ≥ 25 indicates overweight; hence testing deviation from this standard provides insight into overall population health status.

### 3. Two-Sample t-Test (Independent, Unequal Variances)

1. Objective: To test whether there is a significant difference in mean BMI between male and female participants.
2. Rationale: Gender-based differences in body composition and health risk profiles are well established, making this a critical comparison.

### 4. One-Way ANOVA

1. Objective: To evaluate whether mean BMI differs significantly across the five General Health (GenHlth) categories (Excellent, Very Good, Good, Fair, Poor).
2. Rationale: Self-reported health status is often correlated with measurable indicators like BMI; ANOVA helps assess if these differences are statistically significant across multiple groups.

# Results

## One-Sample t-Test

A one-sample t-test was performed to evaluate whether the mean BMI of the synthetic healthcare population (𝑛 = 150,000) differs significantly from the healthy reference standard of 25.0.

The analysis revealed that the **sample mean BMI was approximately 28.1**, placing the average participant in the **overweight category**.

The test results (Table 1) indicated a very large t-statistic and a **p-value < 0.001**. Since 𝑛 < 0.05, the null hypothesis was rejected. This confirms that the mean BMI of the dataset population is statistically significantly greater than the healthy reference standard of 25.

### *Results of the One-Sample t-Test for BMI*

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Sample Size (n) | 150,000 |
| Sample Mean BMI | 28.1 |
| Test Value |  |
| t-statistic |
| p-value |

## Two-Sample Independent t-Test

An independent samples t-test was conducted to compare the mean BMI between **female participants (coded as 0)** and **male participants (coded as 1)**.

The analysis showed that the **mean BMI for males was approximately 28.7**, while the **mean BMI for females was approximately 28.1**. Although the numerical difference between the two groups was modest, the test produced a **p-value < 0.001**, which is well below the alpha threshold of 0.05.

### *Results of the Two-Sample t-Test for BMI by Sex*

|  |  |  |  |
| --- | --- | --- | --- |
| **Group** | **n** | **Mean BMI** | **Std. Deviation** |
| Female (Sex=0) | 72,000 | 28.11 | 5.5 |
| Male (Sex=1) | 78,000 | 28.70 | 5.8 |
| *t-statistic = >10 , p-value* < *0.001* | | |  |

## One-Way ANOVA

A one-way Analysis of Variance (ANOVA) was performed to examine the relationship between **selfreported general health status (GenHlth: 1 = Excellent to 5 = Poor)** and **Body Mass Index (BMI)**.The results, summarized in **Table 3**, were statistically significant:

0.001F(4,149995)≈323,460.22,p<0.001

This strong outcome leads us to reject the null hypothesis, confirming that there are **significant differences in mean BMI among the five General Health groups**. Specifically, participants reporting poorer general health (GenHlth = 5) had a markedly higher mean BMI compared to those reporting excellent health (GenHlth = 1).

### *ANOVA Summary Table for BMI by General Health Group*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Source of Variation** | **Sum of Sq.** | **df** | **Mean Sq.** | **F-statistic** | **p-value** | **F crit** |
| Between Groups | 3,368,777 | 4 | 842,194.2 | 323,460.2 | 0.001 | 2.37 |
| Within Groups | 390,542 | 149,995 | 2.60 |  |  |  |
| Total | 3,759,319 | 149,999 |  |  |  |  |

## (Post-hoc Analysis – Tukey’s HSD)

To further explore the significant ANOVA findings, a **Tukey’s Honest Significant Difference (HSD) post-hoc test** was conducted to compare BMI means between all pairs of General Health groups (GenHlth 1–5).

The results are presented in **Table 4**. Most pairwise comparisons were statistically significant (p < 0.001), with participants in poorer self-reported health groups consistently showing higher BMI values compared to those reporting better health. The only non-significant difference was observed between **Fair (4)** and **Poor (5)** health groups.

### *Summary of Tukey HSD Post-Hoc Test for BMI*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Group 1** | **Group 2** | **Mean Diff.** | **p-adj** | **95% C.I.** | **Significant?** |
| 1 (Excellent) | 2 (Very Good) | 1.88 | 0.001 | (1.77, 1.98) | Yes |
| 1 (Excellent) | 3 (Good) | 3.74 | 0.001 | (3.64, 3.85) | Yes |
| 1 (Excellent) | 4 (Fair) | 4.92 | 0.001 | (4.79, 5.05) | Yes |
| 1 (Excellent) | 5 (Poor) | 4.80 | 0.001 | (4.62, 4.98) | Yes |
| 2 (Very Good) | 3 (Good) | 1.87 | 0.001 | (1.78, 1.95) | Yes |
| 2 (Very Good) | 4 (Fair) | 3.04 | 0.001 | (2.93, 3.16) | Yes |
| 2 (Very Good) | 5 (Poor) | 2.93 | 0.001 | (2.76, 3.10) | Yes |
| 3 (Good) | 4 (Fair) | 1.18 | 0.001 | (1.06, 1.29) | Yes |
| 3 (Good) | 5 (Poor) | 1.06 | 0.001 | (0.89, 1.23) | Yes |
| **4 (Fair)** | **5 (Poor)** | **-0.12** | **0.437** | **(-0.30, 0.07)** | **No** |

## Discussion

The results of this analysis provide important insights into the health characteristics of the synthetic healthcare population. The finding that the mean BMI (~28.1) is significantly higher than the healthy reference value of 25 confirms that, on average, the population is in the **overweight category**. This carries substantial public health implications, as elevated BMI is a well-established risk factor for conditions such as **diabetes, hypertension, and cardiovascular disease**.

The observed difference in BMI between males and females, while statistically significant, was relatively small in absolute terms. This suggests that although gender is a factor, other health behaviors and lifestyle variables, such as **smoking, exercise frequency, and sleep patterns**, may play a more influential role in determining BMI.

The strongest finding was the clear, graded relationship between **self-reported general health status (GenHlth)** and BMI. Participants reporting excellent health had the lowest average BMI, while those in poorer health categories had progressively higher BMI values. The only non-significant difference was observed between the “Fair” and “Poor” health groups, which may indicate that once BMI reaches a high threshold (falling within the obese category), perceived health status plateaus at a consistently low level. This relationship is likely **bidirectional**: higher BMI contributes to poorer health outcomes, while declining health can encourage sedentary behaviors and further weight gain.

## Limitations

This study has several limitations that should be acknowledged:

1. **Cross-Sectional Nature:** The dataset captures patient data at a single point in time, limiting the ability to establish causal relationships.
2. **Self-Reported Measures:** Variables such as *GenHlth* reflect personal perception and may be influenced by individual bias or recall errors.
3. **Uncontrolled Confounders:** The analysis did not explicitly control for factors such as dietary habits, genetic predispositions, or medication use, all of which may influence BMI.
4. **Software Constraints:** While Excel was useful for initial descriptive and inferential analyses, more advanced statistical techniques (e.g., Tukey’s HSD) required SPSS or Python for robust implementation.

## Conclusion

This statistical analysis of a large-scale synthetic healthcare dataset demonstrates that the population under study is, on average, **overweight**, highlighting an important public health concern. The findings confirm that BMI is influenced by demographic factors such as gender and, most notably, by **selfperceived general health status**, which showed a strong, consistent relationship with measured BMI.

These results reinforce BMI’s role as a **critical health indicator** and provide evidence supporting targeted public health strategies. Moreover, the analysis lays the groundwork for applying **predictive modeling and machine learning approaches** to better understand and forecast the impact of lifestyle and demographic variables on chronic disease risk.