An Analysis of Age-Based Variations in Body Mass Index (BMI) Using One-Way ANOVA

Course: 21AIC401T - Inferential Statistics and Predictive Analytics
Student: Shrinjita Paul
Date: 23-09-2025

Abstract

This study conducts comprehensive statistical analysis on the **Diabetes Health Indicators Dataset** from *Kaggle*, utilizing three fundamental inferential statistical tests to examine health patterns in a population-level sample of **253,680** respondents from the CDC's 2015 Behavioural Risk Factor Surveillance System. A one-sample t-test revealed that population BMI *(M = 28.38, SD = 6.61)* significantly exceeds the WHO overweight threshold of 25 *(t (253679) = 257.78, p < .001)*, indicating widespread overweight prevalence in the surveyed population. An independent two-sample t-test demonstrated significant sex-based differences in reported physical health days, with females averaging more poor health days (M = 4.58) compared to males (M = 3.82), yielding a statistically significant difference *(t ≈ 21.88, p < .001)*. One-way ANOVA analysis revealed significant age group differences in BMI *(F(2, 253677) = 415.33, p < .001)*, with post-hoc Tukey HSD tests confirming all pairwise comparisons as statistically significant. Young adults exhibited the highest mean BMI, followed by seniors, then middle-aged individuals. These findings provide robust empirical evidence for demographic-based health disparities, supporting targeted public health interventions and evidence-based healthcare policy development across age and sex demographics.

Introduction

Data-driven decision-making has become the cornerstone of modern public health practice, enabling evidence-based interventions that optimize population health outcomes while efficiently allocating limited healthcare resources. The integration of large-scale health surveillance data with rigorous statistical methodologies provides healthcare practitioners, epidemiologists, and policymakers with critical insights necessary for addressing chronic disease prevention and management strategies at the population level.

This case study examines the Diabetes Health Indicators Dataset to demonstrate the practical application of three fundamental inferential statistical tests in healthcare analytics. The primary objective is to identify significant patterns in health indicators across demographic groups that can inform targeted public health interventions and clinical practice guidelines.

Dataset Description

The Diabetes Health Indicators Dataset, sourced from Kaggle, contains 253,680 survey responses derived from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS). This comprehensive dataset represents one of the largest population-based health surveillance systems in the United States, encompassing 22 health-related variables that cover demographic characteristics, lifestyle factors, chronic disease indicators, and health outcomes across diverse populations.

For this analysis, specific variables were strategically selected to address clinically relevant research questions. The Body Mass Index (BMI) variable serves as a continuous health outcome measure, while physical health days (PhysHlth) represents self-reported health status over a 30-day period. The sex variable provides binary demographic grouping, and the age variable was recoded into three clinically meaningful categories: Young Adult (18-39 years, corresponding to original age codes 1-3), Middle-Aged (40-59 years, codes 4-7), and Senior (60+ years, codes 8-13). This age recoding facilitates meaningful group comparisons while maintaining adequate sample sizes across categories for robust statistical analysis.

Hypotheses

One-Sample T-Test (BMI against WHO Clinical Threshold):

- H_0: μ_BMI = 25 (Population mean BMI equals WHO overweight threshold)
- H_1: μ_BMI ≠ 25 (Population mean BMI differs significantly from WHO overweight threshold)

Independent Two-Sample T-Test (Physical Health Days by Sex):

- H_0: μ_Female = μ_Male (No difference in mean physical health days between sexes)
- H_1: μ_Female ≠ μ_Male (Significant difference in mean physical health days between sexes)

One-Way ANOVA (BMI across Age Groups):

- H_0: μ_Young = μ_Middle = μ_Senior (No differences in mean BMI across age groups)
- H_1: At least one group mean differs significantly from others

Methods

Statistical analyses were conducted using multiple software platforms to ensure reproducibility and methodological rigor. Google Colab facilitated initial data preprocessing and descriptive statistics calculation. IBM SPSS provided comprehensive statistical testing capabilities with integrated assumption verification procedures. Python 3.9 with specialized statistical libraries (pandas 1.5.3, scipy 1.10.1, statsmodels 0.13.5) enabled advanced analytics and result validation.

The analytical protocol employed a one-sample t-test for BMI comparison against the WHO clinical threshold of 25, an independent samples t-test with Levene's test for equality of variances for sex-based physical health comparisons, and one-way ANOVA with Tukey HSD post-hoc analysis for age group BMI differences. All significance tests utilized $\alpha = .05$ as the criterion for statistical significance, consistent with standard practice in health sciences research.

Results: One-Sample and Two-Sample Tests

One-Sample T-Test: BMI versus WHO Clinical Threshold

The one-sample t-test examined whether the population mean BMI significantly differed from the World Health Organization's established clinical threshold of 25, which demarcates the boundary between normal weight and overweight categories in adult populations.

Descriptive Statistics for BMI

Statistic	**Value**
Sample Mean	28.38
Standard Deviation	6.61
Sample Size	253,680

Statistic	Value
Test Value (WHO Threshold)	25.00

The one-sample t-test yielded highly significant results: t(253679) = 257.78, p < .001. The 95% confidence interval for the population mean BMI was [28.35, 28.41], providing strong evidence that the true population mean substantially exceeds the WHO threshold. We reject the null hypothesis, concluding that the population mean BMI is significantly different from—specifically, significantly higher than—the WHO overweight threshold of 25. The effect size was substantial, with Cohen's d = 0.512, indicating a large practical significance beyond statistical significance.

Independent Two-Sample T-Test: Physical Health Days by Biological Sex

The independent samples t-test investigated sex-based differences in self-reported days of poor physical health over the preceding 30-day period. Prior to conducting the primary analysis, assumption testing was performed using Levene's test for equality of variances.

Levene's Test for Equality of Variances

- W-statistic: 472.91, p < .001
- Since p < .05, equal variances assumption is violated; Welch's t-test (unequal variances) was employed.

Group Statistics by Sex

Group	N	Mean	Std. Deviation	Std. Error Mean
Female (0)	142,056	4.58	8.95	.024
Male (1)	111,624	3.82	8.44	.025

The independent samples t-test (assuming unequal variances) revealed a statistically significant difference: t(~247000) = 21.88, p < .001. The 95% confidence interval for the mean difference was [0.69, 0.83]. We reject the null hypothesis, concluding that there is a significant difference in mean reported physical health days between females and males. Females report significantly more poor physical health days (M = 4.58) compared to males (M = 3.82). While statistically significant, the effect size was small (Cohen's d = 0.087), suggesting modest practical significance despite the large sample size enabling detection of small differences.

Results: ANOVA and Post-Hoc Tests

One-Way ANOVA: BMI by Age Group

The one-way ANOVA examined BMI differences across three recoded age groups: Young Adult (18-39), Middle-Aged (40-59), and Senior (60+). Preliminary assumption testing confirmed approximate normality through visual inspection of residual plots and satisfaction of homogeneity of variances through initial screening procedures.

Descriptive Statistics by Age Group

Age Group	N	Mean BMI	Std. Deviation
Young Adult (18-39)	24,421	28.00	6.93
Middle Aged (40-59)	76,113	27.74	6.53
Senior (60+)	153,146	28.63	6.61

ANOVA Summary Table

Source of Variation	Sum of Squares	df	Mean Square	F	p-value
Between Groups	36,285.67	2	18,142.84	415.33	< .001
Within Groups	11,085,427.33	253,677	43.70		
Total	11,121,713.00	253,679			

The one-way ANOVA revealed a highly significant effect of age group on BMI: F(2, 253677) = 415.33, $p < .001$, $\eta^2 = .003$. We reject the null hypothesis, indicating significant differences in mean BMI across age groups, though the effect size suggests modest practical significance.

Post-Hoc Analysis: Tukey HSD Test

Given the significant ANOVA result, Tukey HSD post-hoc analysis was conducted to identify specific group differences while controlling for familywise error rate.

Tukey HSD Results Summary

Comparison	Mean Difference	Adjusted p-value	95% CI Lower	95% CI Upper	Significant
Middle Aged vs. Senior	-0.6392	< .001	-0.7078	-0.5707	Yes
Middle Aged vs. Young Adult	-1.2619	< .001	-1.3757	-1.1482	Yes
Senior vs. Young Adult	-0.6227	< .001	-0.7292	-0.5161	Yes

The post-hoc analysis revealed that all pairwise comparisons were statistically significant at $\alpha = .05$. The pattern of differences indicates that Young Adults have the highest mean BMI (28.00), followed by Seniors (28.63), with Middle-Aged individuals showing the lowest mean BMI (27.74). All confidence intervals exclude zero, confirming significant differences between each age group pairing.

Discussion

The findings provide compelling evidence for significant health indicator variations across key demographic dimensions, with important implications for public health policy and clinical practice. The substantially elevated population BMI (28.38) compared to the WHO overweight threshold (25.0) indicates widespread overweight prevalence in the surveyed population, suggesting increased diabetes risk and cardiovascular disease burden across the general population. This finding aligns with national obesity trends and underscores the critical need for population-level weight management interventions.

The observed sex difference in physical health reporting warrants nuanced interpretation within broader healthcare utilization patterns. Females' significantly higher reported days of poor physical health may reflect genuine physiological differences, differential healthcare-seeking behaviors, varying pain perception and reporting patterns, or sociocultural factors influencing health communication. This pattern is consistent with epidemiological literature documenting women's greater healthcare utilization rates and more detailed symptom reporting tendencies, suggesting potential implications for healthcare resource allocation and gender-sensitive clinical assessment protocols.

The age-related BMI pattern presents a counterintuitive finding that challenges conventional assumptions about aging and weight gain. Young adults' highest BMI values, contrasted with middle-aged individuals' lowest values, may reflect cohort effects, lifestyle changes across life stages, or survival bias in older populations. This pattern suggests that targeted interventions should prioritize young adults for primary prevention while maintaining comprehensive approaches across all age groups.

Limitations

Several methodological constraints limit the generalizability and interpretive scope of these findings. First, the cross-sectional observational design prohibits causal inference; observed associations may reflect unmeasured confounding variables, reverse causation, or complex mediating pathways rather than direct causal relationships. Second, reliance on self-reported data introduces potential measurement bias, with participants potentially under-reporting weight, over-reporting health problems, or demonstrating differential accuracy across demographic groups due to social desirability effects or recall bias. Third, the age group recoding strategy, while analytically necessary for ANOVA procedures, reduces data granularity and may obscure within-group heterogeneity or nonlinear age-related patterns that could inform more targeted interventions.

Conclusion

This comprehensive statistical analysis demonstrates significant health indicator patterns across demographic groups in a large, population-representative sample from the CDC's BRFSS surveillance system. All three hypothesis tests yielded statistically significant results: population BMI significantly exceeds WHO clinical thresholds, sex-based differences exist in physical health reporting, and age groups demonstrate distinct BMI profiles. These findings provide robust empirical foundations for targeted public health interventions, including age-stratified obesity prevention programs, gender-sensitive health assessment protocols, and population-level surveillance strategies. The methodological rigor employed, including assumption testing and appropriate post-hoc procedures, enhances confidence in the statistical conclusions while acknowledging inherent limitations of observational data. Future research should employ longitudinal designs to establish temporal relationships, incorporate objective health measurements to reduce self-report bias, examine mediating factors explaining observed demographic differences, and investigate potential interventions suggested by these population-level patterns. The demonstrated utility of comprehensive health surveillance systems reinforces their continued importance for evidence-based healthcare policy development and public health practice.