Self Employment and Health Insurance

Methods

The health insurance data is from a 1996 Medical Expenditure Panel Survey conducted in the United States, comprising information from 8,802 individuals. The dataset includes details such as age, gender, health status, whether the individual has health insurance, and self-employment status. The focus of our investigation is on the relationship between self-employment status and the possession of health insurance. Therefore, the relevant columns for our study are 'insurance' and 'selfemp.' The initial five rows of the dataset are presented below.

health	age	limit	gender	insurance	married	selfemp	family	region	ethnicity	education
yes	31	no	male	yes	yes	yes	4	south	cauc	bachelor
yes	31	no	female	yes	yes	no	4	south	cauc	highschool
yes	54	no	male	yes	yes	no	5	west	cauc	ged
yes	27	no	male	yes	no	no	5	west	cauc	highschool
yes	39	no	male	yes	yes	no	5	west	cauc	none

Figure 1. Health Insurance from a Medical Expenditure Panel Survey in the US in 1996

Since we are comparing two proportions—specifically, the proportion of individuals who are self-employed with health insurance and those who are not self-employed with health insurance—we are evaluating two independent population proportions. Within this dataset, approximately 12.17% of individuals, equivalent to 1071 people, are self-employed, while the remaining 7731 individuals are not self-employed.

To understand whether there exists sufficient evidence indicating a difference in the true population proportion of individuals with health insurance based on their self-employment status, we perform a hypothesis test comparing the two independent populations: self-employed and non-self-employed groups. We opted for this test after assuming that both independent samples are simple random samples and are independent of each other. Furthermore, the criteria for the test are met, as the number of individuals with and without health insurance in both groups are more than five, and the entire population of U.S. adults is at least twenty times the sample size. Additionally, the pooled proportion, representing the combined proportion of individuals with health insurance, is calculated based on the entire sample size, as it includes both the self-employed and non-self-employed individuals. Lastly, we used the Wald-based confidence interval method to calculate all the relevant confidence intervals, such as the confidence interval of the proportions of individuals with health insurance who are either self-employed or not and the difference in the proportions.

Results

As previously mentioned, about 12.17% of individuals, which is 1071 people, are self-employed, while the remaining 7731 individuals (87.83%) are not self-employed. As shown in Figure 2, the point estimates for the population proportion of self-employed individuals who have health insurance is

approximately 68.91% and the estimate for non-self-employed individuals who have health insurance is around 81.67%. We also estimate with 95% confidence that the true proportion of self-employed individuals who have health insurance is between 66.14% and 71.68%, whereas the 95% confidence interval for the true proportion of non-self-employed individuals who have health insurance is between 80.81% and 82.53%. Additionally, we estimate the difference in proportions (proportion of individuals with health insurance who are not self employed - proportion of individuals with health insurance who are self employed) to be approximately 12.76%. With 95% confidence, we estimate that the actual difference in proportions falls within the range of 9.86% to 15.67%.

To conduct the hypothesis test, we define the null hypothesis as there is no difference between the two proportions, and the alternative hypothesis as there is a difference between the two proportions. The mathematical representation of the null and alternative hypothesis is as such:

Let p_s be the proportion of self-employed individuals with health insurance

Let p_{ns} be the proportion of non-self-employed individuals with health insurance

Let p_c be the pooled proportion of people with health insurance from both groups

Let n_s be the number of self-employed individuals

Let n_{ns} be the number of non-self-employed individuals

$$H_0: p_s - p_{ns} = 0$$

$$H_1: p_s - p_{ns} \neq 0$$

We calculate the z score (formula below), which is about -9.808, and the p-value for the two-tailed test is approximately 1.035e-22.

$$z = \frac{(p_s' - p_{ns}') - (p_s - p_{ns})}{\sqrt{p_c * (1 - p_c) * (\frac{1}{n_s} + \frac{1}{n_{ns}})}}$$

Conclusion

With a calculated p-value, about 1.035e-22, being extremely small, at a 1% level of significance, the sample data provides ample evidence to conclude that there is a difference between the proportion of self-employed individuals with health insurance and the proportion of non-self-employed individuals with health insurance. In other words, there is substantial evidence supporting the claim that the true population proportion of individuals with health insurance differs based on self-employment status with the dataset we have. Further research can be conducted to understand the reasons behind this difference, exploring potential mitigation strategies, and examining whether insurance companies should tailor their advertising to self-employed individuals with more suitable healthcare plans.