Table of Contents-Anova Analysis

Title: Hypothesis Testing using parametric and non-parametric methods to investigate the relationship between family size and healthcare expenses

Introduction

Insurance companies are faced with the task of understanding the association between health-risks and healthcare expenses in order to charge fair premiums to prospective (first-time or returning) insurees. The basic perspective of health-risks might drive one to think of scenarios like pre-existing conditions, age-related diseases, bad health habits (smoking, alcoholism etc.) amongst others; but several other factors that are socio-geo-economic like family-size, region of origin, and other behavioral aspects can drive up or down health risks hence these factors need to be considered in insurance premium determination. In an attempt to understand the effects of non-inherent risks on healthcare expenses, an insurance data set was obtained from Kaggle with descriptions of this dataset provided in a data dictionary found in the appendix. Briefly, this dataset consists of 1033 independent randomly sampled individuals that originally reported 7 attributes with the most relevant to this project being healthcare expenses and family size.

To understand the role family size plays on healthcare expenses, the number of kids reported were initially coded into three levels namely small family size (SFS), normal family size (NFS) and large family size (LFS) with details found in the data dictionary. To resolve the bimodal distribution of the initially coded SFS, this level was further subleveled into SFS_above-average-expenses (SFS_AA) and SFS_below-average-expenses (SFS_BA). Hypothesis testing was performed using parametric analysis of ANOVA and non-parametric analysis namely Kruskal-Wallis and Chi-square.

The ANOVA or Kruskal-Wallis test investigates the relationship between the categorical independent variable family size (with 3 or 4 levels) and the continuous numeric dependent variable healthcare expenses (a proxy for health insurance premium determination). The Chi-square test investigates the association between family size and health expenses in pairwise comparisons of family size levels; 3 pairwise comparisons are performed (SFS vs NFS, SFS vs LFS and NFS vs LFS) using a Bonferroni adjusted alpha level of 0.017.

Hypothesis

Four hypotheses are provided for an Anova, Kruskal-Wallis, Chi-square goodness of fit and Chi-square test of association.

Anova Hypothesis

H_0: The means for healthcare expenses are EQUAL for all three family size levels (SFS mean = NFS mean = LFS mean).

H_a: The means for healthcare expenses are NOT EQUAL for all three family size levels (SFS mean ≠ NFS mean ≠ LFS mean).

Research hypothesis: Patients with more kids or larger family sizes should spend more on health expenses compared to patients with smaller family sizes. Simply, how does family size affect health expenses by comparison of group means for the family size?

Chi-Square Goodness of Fit Hypothesis

H_0: The proportion of patients with healthcare expenses is EQUAL in each formulated category (SFS_AA % = NFS_AA % = LFS_AA % = SFS_BA % etc.); AA & BA = above average & below average health expense respectively

$H_{0\text{-}1}$: The proportion of patients with above vs below average expenses is EQUAL for each family size category (SFS AA % = SFS BA %).

H_a: The proportion of patients with health care expenses is NOT EQUAL in each formulated category (SFS_AA % ≠ NFS_AA % ≠ LFS_AA % ≠ SFS_BA % etc.)

$H_{a\text{-}1}$: The proportion of patients with above vs below average expenses is NOT EQUAL for each family size category (SFS AA % ≠ SFS BA %).

Research hypothesis: First, understand how family size and health spending habits influence the distribution of patients based on health expenses. Secondly, understand whether family size influences the health spending habits of patients by comparing the proportions for above vs below average expense for each category to expected frequencies if equal proportions was assumed.

Chi-Square test of association

H_o: The proportion of patients with above average (AA) expenses is equal for each family size category in pairwise comparisons (y = v or x = u)

H_a: The proportion of patients with above average (AA) expenses is NOT equal for each family size category in pairwise comparisons (y ≠ v or x ≠ u)

Research hypothesis: Investigate whether a larger proportion of people with relatively larger family sizes spend above average on health expenses compared to relatively small family sizes. Can family size be used as a good estimator of expense type?

		Expense_group		
		BA	AA	Total
Fam_size	SFS	x	y	x+y
	NFS	u	v	u+v
	Total	x+u	y+v	x+u+y+v

Table 1: Diagrammatic representation of Chi-square association test hypothesis.

Kruskal-Wallis Hypothesis

Path-determining Hypothesis

H_0: The distribution of expenses for Family size levels are equal or similar

H_a: The distribution of expenses for Family size levels are NOT equal or similar

Equal distributions-Ignored path

$H_{0\text{-}1}$: The medians of the family size levels are equal

H_a: The medians of the family size levels are NOT equal

Unequal distributions-Determined path

H_0: The mean ranks of the family size levels are equal

H_a: The mean ranks of the family size levels are NOT equal

Methodology

Study Design

The independent variable was derived from discrete measures of "number of children" reported by randomly sampled individuals in an insurance study. In Excel, code was written encoding "Small Family Size" to patients with zero or no kids, "Normal Family Size" to patients that had between one & two kids and "Large Family Size" for patients with greater than or equal to three kids. A data dictionary detailing these modifications of the original dataset can be accessed to gain more insight into this approach. This research focuses on understanding the interplay between the number of kids patients report when shopping for insurance and how that can affect their insurance premiums. Programmers in charge of coding pricing algorithms for insurance companies based on several parameters like region, race, kid count etc. will greatly appreciate the summarization of the kid count into these three categories which can then be used in a computer program for patient premium determination. Reiterating the research question as understanding the impact of family size (independent variable) on healthcare expenses (dependent variable) as a proxy for determining insurance premiums for patients based on family size.

Methodology Rationale

Anova is chosen as the parametric method of analysis despite some violations of assumptions in order to compare the results to the Kruska-Wallis results.

A Kruskal-Wallis non-parametric test is chosen as an equivalent to the parametric non-repeated measures Anova.

The Chi-square test is chosen to probe into the association between family sizes based on health expenses.

Results

Anova

A one-way omnibus Anova was performed in a comparison of health expense means between 3-levels of family size. Assumptions of normality and homoscedasticity (variance homogeneity) were violated albeit other assumptions being fulfilled; but given that all 3-levels of family size were diagnosed with these violations, an Anova was still performed as a secondary analysis to crosscheck with findings from non-parametric tests of Kruskal-Wallis and Chi-square test of association.

Health expense means between family size levels differed statistically significantly (Welch-$F_{(2,1030)\ (dfb,\ dfw)} = 11.655$, $p < .001$, N =1033). The effect size (Adjusted Welch Omega squared) was calculated as 0.0202 which is considered a small effect size and indicates that mean health expense differences between family size levels are not of actionable significance when determining insurance premium. In summary our results revealed a statistically significant difference (of small effect) in mean health expenses between our family size levels.

Investigating pairwise comparison between all 3-levels invoked the use of post-hoc analysis specifically Games-Howell due to violations of the homoscedasticity assumption. Post-hoc analysis revealed that the large family size differed significantly from small and normal family sizes while the small versus normal family size showed no statistically significant difference.

Kruskal-Wallis

The K-W test was performed to determine differences in health expense for 4 levels of family size (a proxy for number of children reported by prospective insurees). Mean ranks were statistically significantly different between family size groups; $\chi^2_{(2)} = 560$, $p < .001$ (mean ranks: SFS_BA=182, SFS_BA =829, normal=521, large=621).

Subsequently, pairwise comparisons were performed using Dunn's procedure with a Bonferroni correction for multiple comparisons. The post hoc analysis revealed statistically significant differences between all levels analyzed ($p < .001$).

Chi-square Goodness of Fit

The minimum expected frequency for our equal hypothesis goodness of fit test was 172.2. Results showed that insuree proportions within each group differed significantly from the expected proportions ($\chi^2_{(5)}$ = 175.37, p < .001, N=1033). Rejection of the null hypothesis is tenable. [1]

Chi-square Association Tests

A chi-sq test of association was conducted to detect association between family size and healthcare expenses. For all pairwise comparisons, statistically significant association (alpha = 0.017) between family size and health expenses, was only observed between NFS vs LFS ($\chi^2_{(1)}$ = 5.881, p = .015[2]).

Assessment of the strength of association between family size and expense, using the Phi statistic uncovered a moderate (|Phi| > .100) association between family size and expense type for the NFS vs LFS pair (ϕ = -0.101, p = .015). A statistically significant moderate association exists between family size and expense type, in the NFS vs LFS comparison (Akoglu, 2018).

[1] This goodness of fit test in real-life hypothesizes that family size and healthcare expense (in this dichotomous construct), will affect the distribution for the proportions of people that spend on health care. There is something unique about the groups to drive a deviation from the expected equal proportions or unequal proportions.

[2] Based on the Bonferroni corrected alpha level of .0167.

Conclusions

Statistically significant difference between mean ranks of all groups used in Scheffe's post-hoc analysis. The fascinating finding from Kruskal-Wallis is the SFS_AA mean rank being higher than all the groups compared, interpreted as SFS can be divided into a health-centric group and non-health centric group. The health-centric subgroup in the SFS level spends significantly higher than subjects in the LFS group, which contradicts the initial research hypothesis that spending should increase as family size increases.

The Chi-square test of association was statistically significant for the NFS vs LFS pairwise comparison. The null hypothesis is rejected inferring association between family size and expense type for this pair. However, the association for this pairwise comparison is considered a moderate association and so insurance pricing algorithms can apply a weighted approach when figuring out premium price differences for this pair. The Chi-square approach is validated by the differences in expense between those who spend above average vs spend below average (with a difference of $3,333,339.81 based on sum and average difference of $7409.6).

Conclusively, a statistically significant difference in health expenses is observed between all levels of family size based on the Kruskal-Wallis test/Scheffe post-hoc analysis and with the bimodal distribution of the small family size, special emphasis should be placed on expense differences between the normal vs large family size as indicated by the Chi-square tests.