## **Assumptions Assessment**

### Table 1: ANOVA and K-W Assumption Assessment

|  |  |  |  |
| --- | --- | --- | --- |
| **Assumption** | **Method used to test assumption** | **Result of method** | **Decision (Met or did not meet assumption)** |
| 1 Dependent continuous Variable | Inherent in data set | Health expenses | Met |
| 1 Independent variable or factor w/ 2+ levels | Inherent in data set | Family size (3 or 4 levels) | Met |
| Independence of Observations | Inherent in data set | Unique subjects (unique\_id) | Met |
| Normal Distribution | S-W, Q-Q plot, boxplot, histograms. | Small, normal and large family size are not normally distributed based on S-W test and graphs. | Not met (K-W methodology needed) |
| No Outliers | Box Plot | No extreme outliers | Met |
| Homoscedasticity | Levene’s Test | Test outcome p < .001 | Not Met (-K-W methodology needed) |

### Table 2: Shapiro-Wilkes Test and Q-Q Plot Interpretation for Anova and K-W H test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Family\_size Group** | **Shapiro-Wilkes test Significance (p value)** | **P-P Interpretation**  **(Normal or Not Normal)** | **Histogram Interpretation**  **(Normal or Not Normal)** | **Decision – Normally Distributed or Not Normally Distributed** |
| Small | <.001 | ~Not normal – skew | Not normal (Clear bimodal distribution) | Not Normally Distributed (Bimodal distribution) |
| Normal | <.001 | ~Not normal - skew | ~ Normal | Not Normally Distributed |
| Large | <.001 | ~Not normal - skew | Not normal | Not Normally Distributed |

### **K-W Assumptions assessment**

## Assumption Assessments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **How assumption is tested/determined** | **If test statistic - Value** | **Significance**  **(p-value)** | **Assumption met?**  **Yes or No** |
| Dependent continuous or ordinal variable | Inherent in data set (continuous expense attribute) |  |  | Yes |
| Independent variable with 3+ levels | Inherent in data set (3 or 4 family size levels) |  |  | Yes |
| Independent observations (no within or between group links) | Inherent in data set (each tuple is for a unique insuree) |  |  | Yes |
| Definition of per group data distribution similarities | K-W test for distribution similarity | K-W stat=55.987 | P < .001 | Neutral (determines the analysis path) |
| No outliers | Boxplot from K-W test |  |  | Yes |

### 

### **Chi-square Goodness of Fit Assumptions**

### Table 4: Chi-square goodness of fit assumptions validation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Assumption** | **How assumption is tested/determined** | **If test statistic - Value** | **Significance**  **(p-value)** | **Assumption met?**  **Yes or No** |
| Independence of observations | Inherent in data set |  |  | Yes |
| Expected frequencies (>=5) | Expected N = ~172.17 |  |  | Yes |

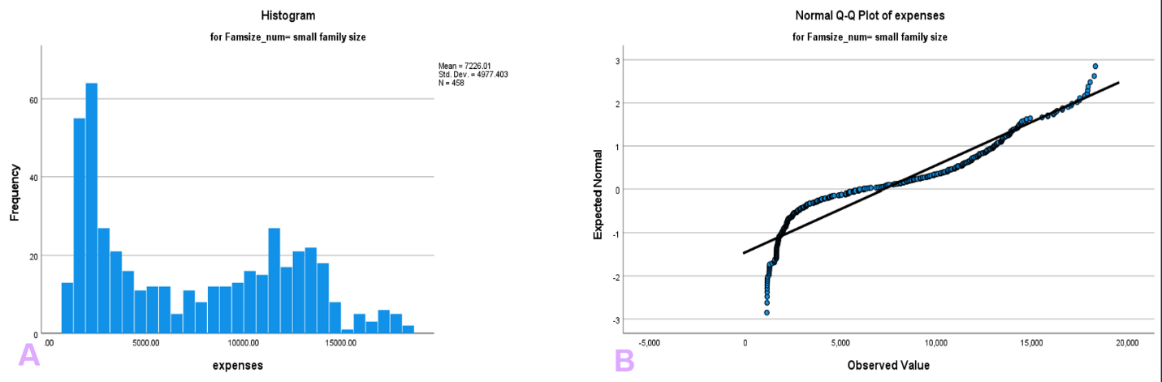
### **Chi-square Test of Association**

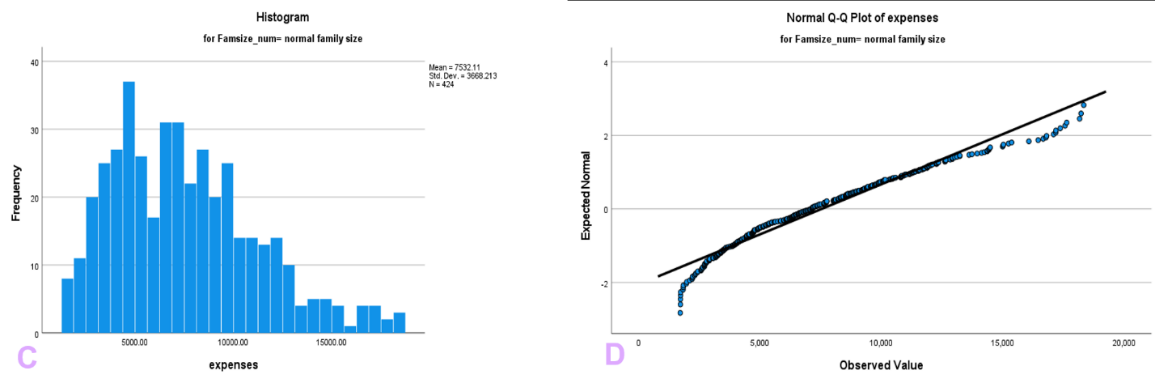
### Table 5: Chi-square Test of Association assumptions

|  |  |  |
| --- | --- | --- |
| **Assumption** | **How assumption is tested/determined** | **Assumption met?**  **Yes or No** |
| Independence of observations | Inherent in data set | Yes |
| 2 Categorical variables (pairwise family size groups) | Inherent in data set | Yes |
| Variables with 2+ factors (AA vs BA) | Inherent in data set | Yes |
| Expected frequencies >=5 (2x2 table) | Expected N from Frequency table | Yes (2x2 table) |

## **Anova Assumptions visualizations and Graphs**

Anova assumption visualizations included are the Q-Q plot, histograms, S-W test table and Levene’s test table (for purposes of simplicity).





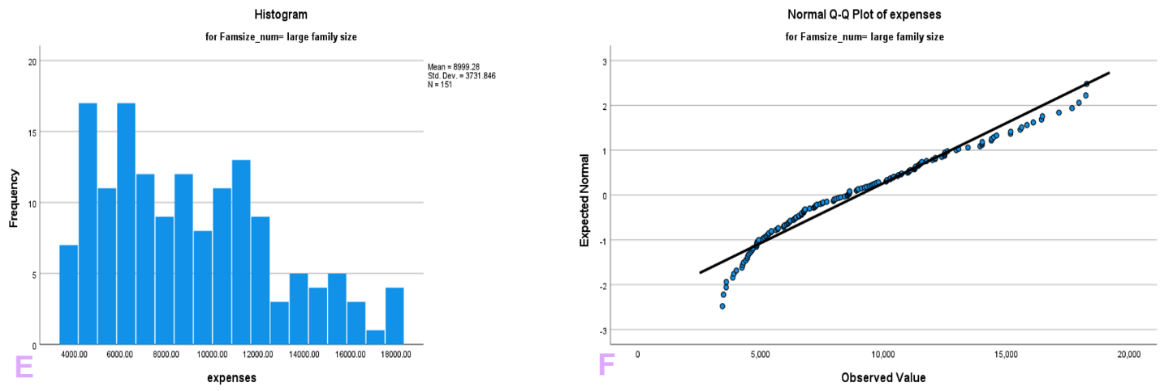


Figure 1: Histogram and QQ plots for all 3-levels of family size from original ANOVA analysis. Observe the bimodal distribution of the small family size histogram. Visit [Table 2](#_Table_2:_ANOVA) for assumption(s) assessment.

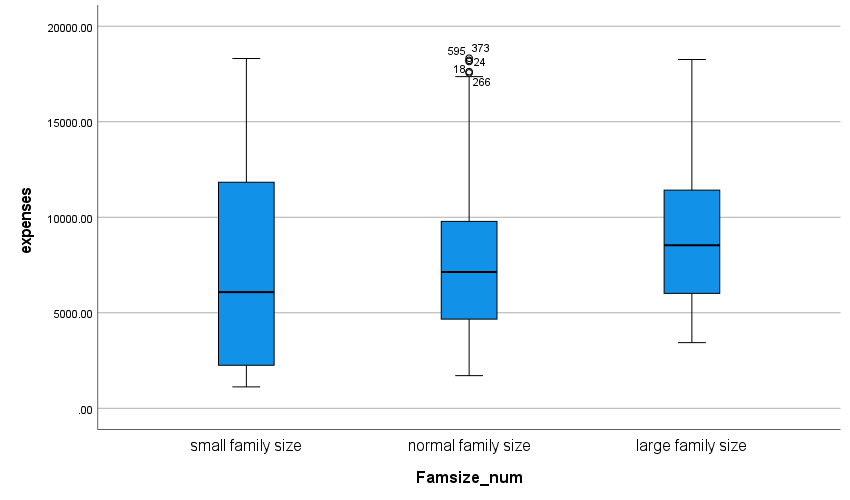


Figure 2: Boxplot for 3-levels of family size. There are no extreme outliers present so assumption met.

### Table 6: Shapiro-Wilk (S-W) test for normality of data. The null assumes normality of the data so REJECT the null (p < .001 for all 3-levels). Non-normality for all levels.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | | |
|  | Famsize\_num | Kolmogorov-Smirnova | | | Shapiro-Wilk (S-W) | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| expenses | small family size | .158 | 458 | <.001 | .895 | 458 | <.001 |
| normal family size | .070 | 424 | <.001 | .958 | 424 | <.001 |
| large family size | .100 | 151 | <.001 | .951 | 151 | <.001 |
| 1. Lilliefors Significance Correction | | | | | | | |

### Table 7: Levene’s test for homogeneity of variance between 3-levels of family size. The null hypothesis assumes variance(s) homogeneity so REJECT the null hypothesis (Levene-stat = 54.889, p <.001).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tests of Homogeneity of Variances** | | | | | |
|  | | Levene Statistic | df1 | df2 | Sig. |
| expenses | Based on Mean | 65.678 | 2 | 1030 | <.001 |
| Based on Median | 54.889 | 2 | 1030 | <.001 |
| Based on Median and with adjusted df | 54.889 | 2 | 1019.970 | <.001 |
| Based on trimmed mean | 64.545 | 2 | 1030 | <.001 |

## **Kruskal-Wallis Assumptions visualizations and Graphs**

Table 9: K-W test for assumption of equal expense distributions for family size levels

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypothesis Test Summary** | | | | |
|  | Null Hypothesis | Test | Sig.a,b | Decision |
| 1 | The distribution of health expenses is the same across categories of Famsize\_num. | Independent-Samples Kruskal-Wallis Test | <.001 | Reject the null hypothesis. |
| a. The significance level is .017. | | | | |
| b. Asymptotic significance is displayed. | | | | |

|  |  |
| --- | --- |
| **Independent-Samples Kruskal-Wallis Test Summary** | |
| Total N | 1033 |
| Test Statistic | 25.849a |
| Degree Of Freedom | 2 |
| Asymptotic Sig.(2-sided test) | <.001 |
| a. The test statistic is adjusted for ties. | |

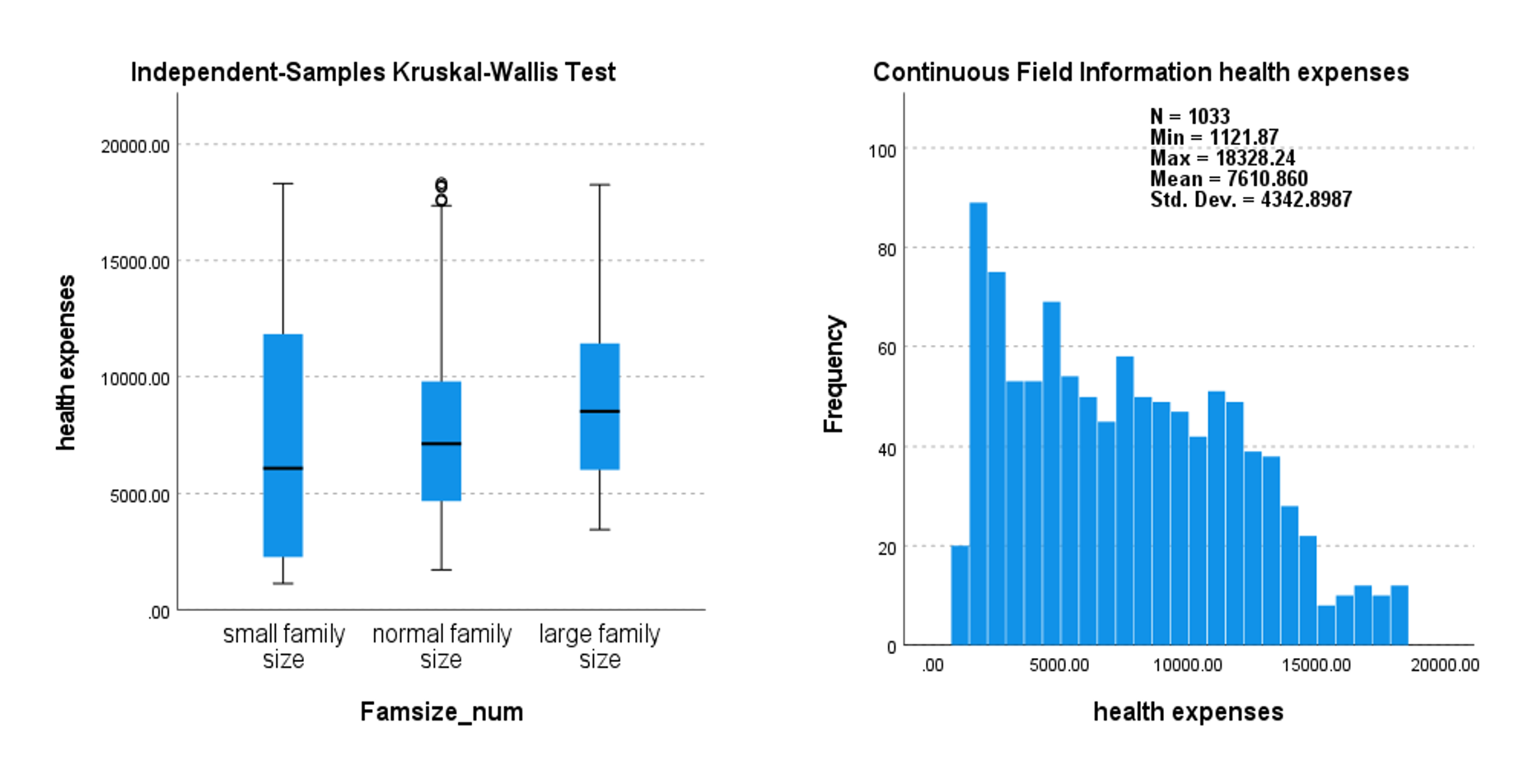


Figure 3: Boxplot of family size levels (left) based on health expenses. Indicative of unequal distributions for all 3-levels. Right is a histogram of the expense attribute just to get a gauge of normality.