ASSOCIATION OF ATTRIBUTE

The Degree of relation ship between two or more attributes is called association

Positive and negative association

Two attributes are said to positively associated if the presence of one result in the presence of the other. Eg. Education and employment

Two attributes are said to be negatively associated if the presence of one result in the absence of the other. Eg.Immunization and Infection.

Chi square- test of Independence

The Chi-Square test of Independence is used to determine if there is a significant relationship between two nominal (categorical) variables. The frequency of one nominal variable is compared with different values of the second nominal variable. The data can be displayed in an R*C contingency table, where R is the row and C is the column.

To test the hypothesis that two attributes are associated or not we used the Chi-square test for independence. To test whether there exist any dependency between the different group and the opinion. In other words we test whether the different group have the same opinion or they will have significant difference

Hypothesis:

Null hypothesis: Assumes that there is no association between the two attributes or the attributes are independent.

Alternative hypothesis: Assumes that there is an association between the two attributes.

Chi-square is defined as $\sum \frac{(O-E)^2}{E}$ where O refers to the observed frequencies and E for the expected frequencies (the ratio of the product of the row total and column total to the grand total). The df = (r-1)(c-1), r-number of rows, c=number of column.

Procedure:

First we have to calculate the expected value of the two nominal variables. We can calculate the expected value of the two nominal variables by using this formula:

- Calculate the row total R and column total C.
- Calculate the grand total N.

- Calculate the expected frequency *E* for each cell using $E = \frac{R \times C}{N}$.
- Calculate $\chi^2 = \sum \frac{(O-E)^2}{E}$.
- Calculate degrees of freedom df=*(r*-1*)(c*-1).
- Obtain the tabled value
- Make conclusion

Non-parametric test

Parametric statistical procedures *– inferential procedures that rely on testing claims regarding parameters such as the population mean μ, the population standard deviation, σ, or the population proportion, p. Many times certain requirements had to be met before we could use those procedures.*

Nonparametric statistical procedures *– inferential procedures that are not based on parameters, which require fewer requirements be satisfied to perform the tests. They do not require that the population follow a specific type of distribution.*

Nonparametric methods use techniques to test claims that are *distribution free*.

Advantages of Nonparametric Statistical Procedures

- Most of the tests have very few requirements, so it is unlikely that these tests will be used improperly.
- For some nonparametric procedures, the computations are fairly easy.
- The procedures can be used for count data or rank data, so nonparametric methods can be used on data such as rankings of a movie as excellent, good, fair, or poor.

Disadvantages of Nonparametric Statistical Procedures

- The results of the test are ***typically less powerful***. Recall that the **power of a test** refers to the probability of making a Type II error. A Type II error occurs when a researcher does not reject the null hypothesis when the alternative hypothesis is true.
- Nonparametric procedures are ***less efficient*** than parametric procedures. This means that a larger sample size is required when conducting a nonparametric procedure to have the same probability of a Type I error as the equivalent parametric procedure.

Run Test of Randomness

There are non-parametric tests or methods that are used in cases when the parametric test is not in use. One of these non-parametric tests is the run test. Run test is used for examining whether or not a set of observation constitute a random sample from an infinite population

Runs test for randomness *– used to test claims that data have been obtained or occur randomly*

Run *– sequence of similar events, items, or symbols that is followed by an event, item, or symbol that is mutually exclusive from the first event, item, or symbol*

Length *– number of events, items, or symbols in a run*

Assumption: The run test for randomness is carried out in a random model in which the observations vary around a constant mean. The observation in the random model in which the run test is carried out has a constant variance, and the observations are also probabilistically independent. The run in a run test is defined as the consecutive sequence of zeroes and ones. This test checks whether or not the number of runs are the appropriate for a randomly generated series.

Hypothesis

H_0: Sample value came from random sequence

H_1: Sample value came from non-random sequence

Procedure:

Test statistics is the number of runs "r"

For finding the number of runs, the observations are listed in their order of occurrence. Each observation is denoted by '+' sign if it is more than the previous observation by a '-' sign if it is less than the previous observation. If the observation is same as previous observation put '0'. The total number of runs up (+) and down is counted.

Accept H_0 if $r_{CL} \leq r \leq r_{Cu}$. Other wise reject H_0.

Where $r_{CL} = \mu - 1.96\,\sigma$ and $r_{CL} = \mu + 1.96\,\sigma$

$$\mu = \frac{2n-1}{3}, \quad \sigma = \sqrt{\frac{16n-29}{90}}.$$

Wald-Wolfowitz Two-sample Run test.

Wald-Wolfowitz Two-sample Run test is used to examine whether two random samples came from population having same distribution. This test can detect differences in average or spread or any other important aspect between the two population.

Hypothesis

H_0: Sample value came from the population having same distribution

H_1: Sample value came from the population having diffrent distribution

Procedure:

Let $\boldsymbol{r}$ denote the number of runs. To obtain $\boldsymbol{r}$, list the $n_1 + n_2$ observations from two samples in ascending order of magnitude. Denote observations from one sample as x's and other by y's. Count the number of runs. In case x and y observation are same value, place the observation x(y) first if run of x(y) observation is counting.

Accept H_0 if $r \geq r_C$. Other wise reject H_0.

Where $r_C = M - 1.96\, S$ with $M = 1 + \frac{2n_2n_1}{n_1+n_2}$ and $S = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$.

Sign test

The sign test can be used to test hypotheses about the central tendency of non normal distributions. In order to measure the central tendency the mean would not be an appropriate measure for the center, because distribution might be skewed or have outliers. Instead we will use the median as a measure for the center of the distribution. Remember: The median of a distribution is the value, so that the probability to fall below equals 0.5, or 50% of the measurements in the population fall below. The median of a sample, is the value so that 50% of the sample data fall below the median.

One-sample sign test

The one sample sign test is used to test the null hypothesis that the median of a distribution is equal to some value.

Hypothesis

H_0: $M = M_0$

H_1: $M \neq M_0$ or $M > M_0$ or $M < M_0$

Procedure:

The observations in a sample of size n are $x_1, x_2, \ldots, x_n$ the null hypothesis is that the population median is equal to some value M_0. Replace the observation with + sign if the value of the observation is greater than M_0 and with – if the value is less than M_0. Suppose that $\boldsymbol{t}$ be the number of + and $\boldsymbol{u}$ the number of − sign. Values of x which are exactly equal to M are ignored. The test statistics is $\boldsymbol{t}$.

Accept H_0 if $r_L \leq t \leq r_u$ if H_1. $M \neq M_0$

Accept H_0 if $r_{CL} \leq t$ if H_1. $M < M_0$

Accept H_0 if $t \leq r_{Cu}$ if H_1. $M > M_0$

Where r_{CL} , r_{CU} , r_L and r_u are obtained from the binomial table.

Inferences about the Differences between Two Medians (paired sample):

The two sample sign test is used to test the null hypothesis that the median of two distribution are equal.

Hypothesis

H_0: $M_1 = M_2$

H_1: $M_1 \neq M_2$ or $M_1 > M_1$ or $M_1 < M_2$

Procedure:

The observations in a sample of size n are (x_1,y_1), (x_2,y_2). . . , (x_n,y_n) . Determine the difference d= y-x for each data pair. Assign a + sign for a positive difference, a – sign for a negative difference, and a 0 for no difference. . Suppose that $\boldsymbol{t}$ be the number of + and $\boldsymbol{u}$ the number of − sign. The test statistics is $\boldsymbol{t}$.

Accept H_0 if $r_L \leq t \leq r_u$ if H_1. $M_1 \neq M_2$

Accept H_0 if $r_{CL} \leq t$ if H_1. $M_1 < M_2$

Accept H_0 if $t \leq r_{Cu}$ if H_1. $M_1 > M_2$

Where $\mathbf{r_{CL}}$, $\mathbf{r_{CU}}$, $\mathbf{r_L}$ and $\mathbf{r_u}$ are obtained from the binomial table.

Wilcoxon Rank Sum Test

Suppose measures of central tendency of two populations shall be compared but it is not save to assume that the populations are normally distributed. If the sample sizes are not large it is not appropriate to use the t-test. In this case we will use the Wilcoxon Rank Sum Test. Instead of using the original measurements, all sample data is ranked from smallest to largest and according to its position in the combined data set the rank is assigned.

Procedure:

The difference (d = y-x) between the members of each pair (x, y) is found out. Rank-this d, from smallest to largest, the differences (between the paired observations) for each pair without regard to the sign of the difference (i.e., rank order the absolute differences). Ignore all zero differences (i.e., pairs with equal scores). Affix the original signs to the ranks. All pairs with equal absolute differences (ties) get the same rank. Sum all positive ranks (*T*+) and all negative ranks (*T*-) and determine the total number of pairs (N). The Wilcoxon statistic, *T*, is the smaller of *T*+ or *T*-. This value is compared with tabled critical values of the Wilcoxon statistic.