Frequently Asked Questions

1. What do you mean by a contingency table?

Answer:
When two categorical variables are recorded we can summarize the data by counting the observed number of units that fall into each of the various intersections of category levels. The resulting counts are displayed in an array called a contingency table.
Let the attributes under consideration be A and B.

Let A be divided into r classes A_1, A2…A_r and let B be divided into s classes say B_1, B_2…B_s.
Let the observations be classified according to their classes in attributes A and B. This classification can be represented in the form of a table called contingency table as shown below.

r x s Contingency Table

B / A	B1	B2	……	Bs	M.T
A1	O11	O12	……..	O1s	(A1)
A2	O21	O22	……	O2s	(A2)
.					
.					
Ar	Or1	Or2	……..	Ors	(Ar)
M.T	(B1)	(B2)		(Bs)	N

In the above O_{ij} represents the number of n individuals falling in class A_i and B_j of the attributes A and B. (A_i) represents the number of individuals in the class A_i and (B_j) represents the number of individuals in the class Bj. $N=\sum(A_i)= \sum(B_j)$

2. Give some examples for the data which can be represented in the form of a contingency table?

Answer:
May experiments results in measurements that are qualitative or categorical rather than quantitative that is a quality or a characteristic (rather than a numerical value) is measured for each experimental unit. We can summarize this type of data by creating a list of categories or characteristics and reporting a count of the measurements that fall into each category. In some situations the researcher classifies the experimental units according to two qualitative variables to generate bivariate data

- A defective piece of furniture is classified according to the type of defect and the production shift during which it was made
- A professor is classified by his professional rank and the type of the university (public or private) at which she works
- A patient is classified according to the type of the preventive flu treatment he received and whether or not he contracted the flu during the winter

3. How do you get probabilities of an individual falling in different classes of attributes A and B under the assumption that A and B are independent?

Answer:
Suppose we have a contingency table based on two attributes A and B, then (Ai) represents the number of individuals in the class Ai and (Bj) represents the number of individuals in the class Bj.

$N=\sum (Ai) = \sum (Bj)$

The probabilities of an individual falling in different classes of attributes A and B under the null hypothesis are calculated as follows

P (A = Ai) = (Ai)/N, probability that an individual falls in class Ai

P (B = Bj) = (Bj) / N, probability that an individual falls in class Bj

P (A = Ai and B = Bj) = P (A = Ai). P (B = Bj), probability that an individual falls in classes Ai and Bj according to the attributes A and B respectively.

Under the null hypothesis = (Ai)/N. (Bj)/N

4. Give the procedure to obtain expected frequencies in a contingency table.

Answer:

Let Eij denotes the expected number of individuals with attributes A and B as Ai and Bj respectively.

Then

Eij= P (A=Ai and B = Bj).N

$$= N\frac{(A_i)}{N}\frac{(B_j)}{N} = \frac{(A_i)(B_j)}{N} \quad ; i = 1, 2\ldots r; \quad j = 1, 2\ldots s$$

5. State the test statistic used to for the test of independence.

Answer:
If we knew the expected cell counts ($E_{ij}=np_{ij}$) under the null hypothesis of independence, then we could use the chi square statistic to compare the observed and expected counts

Now under the null hypothesis $\chi^2 = \frac{\sum_i \sum_j (Oij - Eij)^2}{E_{ij}}$ is distributed as χ^2 variable where Eij's are the expected frequencies and Oij's are the observed frequencies.

6. Write a note on the Chi square test procedure for independence of attributes.

Answer:
In the first step the given data can be arranged in the form of a contingency table where in the observed frequencies for various classes or cells can be obtained.

In the second step the corresponding theoretical frequencies can be computed from the marginal totals. The test enables us to find whether there is any relationship between the two attributes say A and B. Let Oij be the observed frequencies and Eij be the corresponding expected frequencies ,i=1,2,….,r;j=1,2,..,s.

In the third step the chi square statistic of independence is given by

$\chi^2 = \frac{\sum_i \sum_j (Oij - Eij)^2}{E_{ij}}$ is computed and is compared to the critical value of no significance from the Chi square distribution, which in many cases gives a good approximation of the distribution of χ^2. The test procedure is to compute the statistic χ^2 and to compare it with the tabulated value $\chi_\alpha{}^2((r-1)(s-1))$ where r is the number of rows and s is the number of columns in the table and to reject the null hypothesis whenever the computed value exceeds the tabulated value.

7. Write a note on the degrees of freedom of Chi square test statistic for independence.

Answer:
There are r x s cells to be filled independently because of restrictions on the marginal and grand totals. Thus degrees of freedom of the above statistic are (r-1) (s-1). Therefore under the null hypothesis the above statistic is distributed as a Chi square variable with (r-1) (s-1) degrees of freedom.

The degrees of freedom for a Chi-square grid are equal to the number of rows minus one times the number of columns minus one: that is, (r-1)*(s-1). In our simple 2x2 grid, the degrees of independence are therefore (2-1)*(2-1), or 1

8. What are the assumptions for the application of chi square distribution?

Answer:

The chi-squared test, when used with the standard approximation that a chi-squared distribution is applicable, has the following assumptions:

- Simple Random sample– The sample data is a random sampling from a fixed distribution or population where each member of the population has an equal probability of selection. Variants of the test have been developed for complex samples, such as where the data is weighted.
- Sample size (whole table) – A sample with a sufficiently large size is assumed. If a chi squared test is conducted on a sample with a smaller size, then the chi squared test will yield an inaccurate inference. The researcher, by using chi squared test on small samples, might end up committing a Type II error
- Expected cell count – Adequate expected cell counts. Some require 5 or more, and others require 10 or more. A common rule is 5 or more in all cells of a 2-by-2 table, and 5 or more in 80% of cells in larger tables, but no cells with zero expected count. When this assumption is not met, Yates correction applied.
- Independence – The observations are always assumed to be independent of each other. This means chi-squared cannot be used to test correlated data (like matched pairs or panel data). In those cases you might want to turn to McNemar test

9. What action should be taken when the theoretical frequencies are less than 5?

Answer:
If any theoretical frequency is less than five it should be pooled with the preceding or succeeding frequencies so that the combined frequency is greater than five. Finally adjustment for the degrees of freedom lost in pooling should be made. Adjustments are made by reducing the total degrees of freedom by one each time when two frequencies are combined.

10. What is the Chi-square test of independence for?

Answer:
The Chi-square test is intended to test how likely that two attributes are related to each other. They are dependent on each other. It is also called a "test of independence" because it states whether the two characteristics under consideration are dependent on each other.

A Chi-square test is designed to analyze categorical data. That means that the data has been counted and divided into categories. It will not work with parametric or continuous data (such as height in inches). For example, if you want to test whether attending class influences how students perform on an exam, using test scores (from 0-100) as data

would not be appropriate for a Chi-square test. However, arranging students into the categories "Pass" and "Fail" would. Additionally, the data in a Chi-square grid should not be in the form of percentages, or anything other than frequency (count) data. Thus, by dividing a class of 54 into groups according to whether they attended class and whether they passed the exam, you might construct a data set like this and apply chi square test for independence.

	Pass	Fail
Attended	25	6
Skipped	8	15

11. Derive a test statistic for the test of independence of 2 attributes in 2X2 contingency table?

Answer:

Consider a 2 x 2 contingency table given below. To test the null hypothesis that the attributes A and B are independent

B A	B1	B2	M.T
A1	a	b	a+b
A2	c	d	c +d
M.T	a+c	b+d	a+b+ c+d=N

Under the null hypothesis the statistic Chi square is computed as follows
The expected frequency of a is E (a) = (a + b) (a + c) / N
a – E (a) = a - (a + b) (a + c) / N = (ad-bc) / N

Similarly E (b) = (a +b) (b + d) / N
b – E (b) = (bc – ad) / N

E (c) = (c+d) (a+c) / N
c- E (c) = (bc-ad) / N
d - E (d) = (ad-bc) / N

$$\chi^2 = \frac{(a-E(a))^2}{E(a)} + \frac{(b-E(b))^2}{E(b)} + \frac{(c-E(c))^2}{E(c)} + \frac{(d-E(d))^2}{E(d)}$$

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{1}{E(a)} + \frac{1}{E(b)} + \frac{1}{E(c)} + \frac{1}{E(d)}\right]$$

$$= \frac{(ad-bc)^2}{N^2}\left[\frac{1}{(a+c)(a+b)} + \frac{1}{(a+b)(b+d)} + \frac{1}{(c+d)(a+c)} + \frac{1}{(c+d)(b+d)}\right]$$

$$= (ad-bc)^2\left[\frac{c+d+a+b}{(a+c)(a+b)(b+d)(c+d}\right] = \left[\frac{N(ad-bc)^2}{(a+c)(a+b)(b+d)(c+d}\right]$$

Under H_0 the above statistic is distributed as Chi square variable with 1 d.f

The test procedure is to compute χ^2 and to compare it with the tabulated value $\chi_\alpha^{\ 2}(1)$ and to reject the null hypothesis if the computed value exceeds the tabulated value

12. What do you mean by Yate's correction for continuity?

Answer:

The statistic χ^2 is distributed as a chi square variable only when all the cell theoretical frequencies are greater than or equal to 5. But if any one of the frequency is less than 5, the statistic χ^2 does not follow a Chi square distribution. In such cases by applying the correction suggested by Yates, the statistic χ^2 can be regarded to follow Chi square distribution.

According to Yates in a 2X2 contingency table if any cell theoretical frequency is <5, 0.5 should be added to it and remaining frequencies should be adjusted such that the marginal frequencies remains unchanged. The χ^2 statistic then is computed as usual

13. How do you test the hypothesis for independence of attributes if any cell theoretical frequency is < 5?

Answer:
When any cell theoretical frequency is < 5, Yate's correction is applied to the test statistic

After the correction χ^2 statistic is obtained as

$$\chi^2 = \left[\frac{N\left[(a \pm \frac{1}{2})(d \pm \frac{1}{2}) - (b \mp \frac{1}{2})(c \mp \frac{1}{2}) \right]^2}{(a+c)(a+b)(b+d)(c+d} \right]$$

$$\chi^2 = \left[\frac{N\left[|ad - bc| - \frac{N}{2} \right]^2}{(a+c)(a+b)(b+d)(c+d} \right]$$

14. Derive Brandt-Snedekor's Formula for 2 x k contingency table.

Answer:

Let the observations be classified based on two attributes A and B into 2 x k classes. Let Oij represent the number of individuals falling in class Ai and Bj of the attributes A and B i = 1, 2 and j =1, 2...k

The observations may be arranged in the form of 2Xk contingency table as given below

B / A	B_1	B_2	……	B_k	M.T
A_1	O_{11}	O_{12}	……..	O_{1k}	m_1
A_2	O_{21}	O_{22}	……	O_{2k}	m_2
M.T	n_1	n_2		n_k	N

Let the null hypothesis be the attributes A and B is independent

Under H0 the expected frequencies and the Chi square statistic are calculated as follows

E_{1j}= E (O_{1j}) = m_1 n_j /N

E_{2j}= E (O_{2j}) = m_2 n_j /N

$$\chi^2 = \frac{\sum_{i=1}^{2}\sum_{j=1}^{k}(Oij - Eij)^2}{E_{ij}}$$

$$\chi^2 = \frac{\sum_{j=1}^{k}(O_{1j} - E_{1j})^2}{E_{1j}} + \frac{\sum_{j=1}^{k}(O_{2j} - E_{2j})^2}{E_{2j}}$$

$$= \frac{\sum_{j=1}^{k}(O_{1j} - \frac{m_1 n_j}{N})^2}{\frac{m_1 n_j}{N}} + \frac{\sum_{j=1}^{k}(O_{2j} - \frac{m_2 n_j}{N})^2}{\frac{m_2 n_j}{N}}$$

$$= \frac{N}{m_1}\sum_{j=1}^{k} n_j(\frac{O_{1j}}{n_j} - \frac{m_1}{N})^2 + \frac{N}{m_2}\sum_{j=1}^{k} n_j(\frac{O_{2j}}{n_j} - \frac{m_2}{N})^2$$

$$= \frac{1}{p}\sum_{j=1}^{k} n_j(p_j - p)^2 + \frac{1}{q}\sum_{j=1}^{k} n_j(q_j - q)^2$$

Where $p_j = \frac{O_{1j}}{n_j}; q_j = \frac{O_{2j}}{n_j}; p = \frac{m_1}{N}; q = \frac{m_2}{N}$

It may be noted that $p_j + q_j = 1; p + q = 1$

Hence $(q_j - q) = -(p_j - p) \Rightarrow (q_j - q)^2 = (p_j - p)^2$

$$\chi^2 = \frac{1}{p}\sum_{j=1}^{k} n_j(p_j - p)^2 + \frac{1}{q}\sum_{j=1}^{k} n_j(p_j - p)^2$$

$$= (\frac{1}{p} + \frac{1}{q})\sum_{j=1}^{k} n_j(p_j - p)^2 = \frac{1}{pq}\sum_{j=1}^{k} n_j(p_j - p)^2$$

$$= \frac{1}{pq}\sum_{j=1}^{k} n_j(\frac{O_{1j}}{n_j} - \frac{m_1}{N})^2 = \frac{1}{pq}(\sum_{j=1}^{k}\frac{{O_{1j}}^2}{n_j} + \frac{{m_1}^2}{N} - \frac{2{m_1}^2}{N})$$

$$= \frac{N^2}{m_1 m_2}(\sum_{j=1}^{k}\frac{{O_{1j}}^2}{n_j} - \frac{{m_1}^2}{N})$$

Under H0 the statistic hi square is distributed as a Chi square variable with (2-1)(K-1) = (k-1) d.f

The test procedure in this case is to reject the null hypothesis whenever the computed value exceeds the tabulated value $\chi_\alpha^2(k-1)$

15. What is the Chi-Square test NOT for?

Answer:

This is also an important question to tackle, of course. Using a statistical test without having a good idea of what it can and cannot do means that you may misuse the test, but also that you won't have a clear grasp of what your results really mean. Even if you don't understand the detailed mathematics underlying the test, it is not difficult to have a good comprehension of where it is or isn't appropriate to use.

First of all, the Chi-square test is only meant to test the probability of independence of a

distribution of data. It will NOT tell you any details about the relationship between them. If you want to calculate how much more likely it is that a woman will be a Democrat than a man, the Chi-square test is not going to be very helpful. However, once you have determined the probability that the two variables *are* related (using the Chi-square test), you can use other methods to explore their interaction in more detail. For a fairly simple way of discussing the relationship between variables, (probably odds ratio) Some further considerations are necessary when selecting or organizing your data to run a Chi-square test. The variables you consider must be mutually exclusive; participation in one category should not entail or allow participation in another. In other words, the data from all of your cells should add up to the total count, and no item should be counted twice.

You should also never exclude some part of your data set. If your study examined males and females registered as Republican, Democrat, *and* Independent, then excluding one category from the grid might conceal critical data about the distribution of your data.

It is also important that you have enough data to perform a viable Chi-square test. If the estimated data in any given cell is below 5, then there is not enough data to perform a Chi-square test. In a case like this, you should research some other techniques for smaller data sets: for example, there is a correction for the Chi-square test to use with small data sets, called the Yates correction. There are also tests written specifically for smaller data sets, like the Fisher's Exact Test.