Frequently Asked Questions

1. Give some examples for the data which falls in categories?
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. Here are a few examples

- People can be classified into 5 income brackets
- A mouse can respond in one of three ways to a stimulus
- An M & M's candy can have one of six colors
- An industrial process manufactures items that can be classified as" acceptable", "second quality" or "defective"

2. What do you mean by Multinomial experiment?
Answer:
An experiment which satisfies the following conditions is said to be a multinomial experiment

- The experiment consists of n independent trials
- The outcome of each trial falls into one of k categories
- The probability that the outcome of a single trial falls into a particular category, say category I is pi and remains constant from trial to trial. This probability must be between 0 and 1 for each of the k categories and the sum of all k probabilities ∑pi=1
- The trials are independent
- The experimenter counts the observed number of outcomes in each category written as O_1, O,..O_k =n

3. How do you visualize Multinomial Experiment?
Answer:
We can visualize the multinomial experiment by thinking of k boxes or cells into which n balls are tossed. The n tosses are independent and on each toss the chance of hitting the ith box is the same. However the chance can vary from box to box. , it might be easier to hit box 1 than box 3 on each toss. Once all n balls have been tossed the number in each box or cell – O_1, O_2…O_k is counted.

4. What do you mean by Chi-square probability distribution?
Answer:
Consider the distribution of sample variance s^2 based on repeated random sampling from a Normal distribution with a specified mean and variance. We can show theoretically that the distribution begins at $s^2=0$, since the variance cannot be negative. With mean equal to σ^2
Its shape is non symmetric and changes with each different sample size and each different value of σ^2. We can standardize the statistic as we did for Z the standardized statistic for variance is called chi-square = (n-1) s^2 / σ^2
Is called a Chi- square variable and has a sampling distribution called the Chi-square probability distribution with n-1degrees of freedom.

5. What is the theory behind Pearson's Chi square test statistic for goodness of fit?
Answer:
A powerful test for testing the discrepancy between theory and experiment was given by Karl Pearson and is known as Chi-square test of goodness of fit.

Suppose that n=100 balls are tossed at the cells (boxes) and we know that the probability of a ball falling into the first box is p1=0.1. How many balls would we expect to fall into the first box? Intuitively you would expect to see 100(0.1)=10 balls in the first box. This would remind you of the average or expected number of success , μ=np in the Binomial experiment . In general the expected number of balls that fall into the cell i – written as Ei – can be calculated using the formula Ei=npi for any of the cells i=1,2..,k

Now suppose that we hypothesize values for each of the probabilities p_1, p_2…p_k and calculate the expected number of each category or cell . If our hypothesis is correct, the actual observed cell counts Oi should not be too different from the expected cell counts, Ei=npi.

The larger the differences the more likely it is that the hypothesis is incorrect. The Pearson's Chi-square statistic uses the differences to eliminate negative contributions and the forming a weighted average of the squared differences.

It can be shown that when n is large χ^2 has an appropriate Chi-square probability distribution in repeated sampling. If the hypothesized cell counts are correct the differences (Oi – Ei) are small and χ^2 is close.

6. Write a note on the Chi square test procedure for goodness of fit.

Answer:

In the first step the given data can be arranged in the form of a frequency distribution 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 knowledge of the population. The test enables us to find if the difference in the observed and theoretical frequencies is just due to chance or is it really the inadequacy of a theory to fit the observed data.

Let Oi be the observed frequencies and Ei be the corresponding expected frequencies, i=1, 2….n

In the third step the Pearson's statistic of goodness of fit is given by $\chi^2 = \frac{\sum_i (Oi - Ei)^2}{E_i}$ 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. A test that does not rely on this approximation is Fisher's exact test; it is substantially more accurate in obtaining a significance level, especially with few observations. The test procedure is to compute the statistic χ^2 and to compare it with the tabulated value $\chi^2_\alpha(n - p - 1)$ 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.

Answer:

Under the hypothesis that the theoretical distribution fits well the statistic is distributed as a Chi square variable with (n-p-1) degrees of freedom where n is the number of cell frequencies, p is the number parameters estimated depends on the distribution of the population under consideration

8. What are the conditions required for the application of Chi square test

Answer:

Conditions for Chi-square test are

1) The sample observations should be independent
2) The constraints on the frequencies must be linear
3) No all theoretical frequencies should be less than 5

In most of the situations conditions (1) and (2) are satisfied as random observations are taken and as only one restriction (namely) $\sum_i Oi = \sum_i Ei$ is imposed on observations

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

Answer:

If any theoretical frequency is less than 5 it should be pooled with the preceding or succeeding frequencies so that the combined frequency is greater than 5 . Finally adjustment for the degrees of freedom lost in pooling should be made. Adjustments are made by reducing the total degrees of freedom by 1 each time when 2 frequencies are combined.

10. What is the Chi-square test for?

Answer:

The Chi-square test is intended to test how likely it is that an observed distribution is due to chance. It is also called a "goodness of fit" statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

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 and apply a Chi square test.

11. What are the applications of Chi square test?

Answer:

Pearson's chi-squared test (χ^2) is the best-known of several chi-squared tests (Yates, likelihood ratio, portmanteau test in time series, etc.) – statistical procedures whose results are evaluated by reference to the chi-squared distribution. Its properties were first investigated by Karl Pearson in 1900.In contexts where it is important to make a distinction between the test statistic and its distribution, names similar to Pearson Chi-squared test or statistic are used. It tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. The events considered must be mutually exclusive and have total probability 1. A common case for this is where the events each cover an outcome of a categorical variable. A simple example is the hypothesis that an ordinary six-sided die is "fair", i. e., all six outcomes are equally likely to occur.

Pearson's chi-squared test is used to assess two types of comparison: tests of goodness of fit and tests of independence.

- A test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution
- A test of independence assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality affects the response)

12. What is the meaning of Pearson’s Chi square test statistic?

Answer:

Actually, Pearson’s Chi square test statistic is a fairly simple relationship. The variables in this formula are not simply symbols, but actual concepts that we've been discussing all along. Oi stands for the Observed frequency. Ei stands for the Expected frequency. You subtract the expected count from the observed count to find the difference between the two (also called the "residual"). You calculate the square of that number to get rid of positive and negative values (because the squares of 5 and -5 are, of course, both 25). Then, you divide the result by the expected frequency to normalize bigger and smaller counts (because we don't want a formula that will give us a bigger Chi-square value just because you're working with a bigger set of data). The huge sigma sitting in front of all that is asking for the sum of every **i** for which you calculate this relationship - in other words, you calculate this for each cell in the table, then add it all together.

13. How do you test the goodness of fit of discrete uniform distribution?
Answer:
In this case Nobservations are divided among $n$cells. A simple application is to test the hypothesis that, in the general population, values would occur in each cell with equal frequency. The "theoretical frequency" for any cell (under the null hypothesis of a discrete uniform distribution) is thus calculated as

$$E_i = \frac{N}{n} ,$$

and the reduction in the degrees of freedom is p = 1, notionally because the observed frequencies O_i are constrained to sum to N.

14. How do you carry on with the test of goodness of fit of other distributions?
Answer:
When testing whether observations are random variables whose distribution belongs to a given family of distributions, the "theoretical frequencies" are calculated using a distribution from that family fitted in some standard way. The reduction in the degrees of freedom is calculated as (n-p-1), where p is the number of co-variates used in fitting the distribution. For instance, when checking a three-co-variate Weibull distribution, p=3, and when checking a normal distribution (where the parameters are mean and standard deviation), p=2. In other words, there will be (n-p-1)degrees of freedom, where $n$is the number of categories.
It should be noted that the degrees of freedom are not based on the number of observations as with a Student's t or F-distribution. For example, if testing for a fair, six-sided dice, there would be five degrees of freedom because there are six categories/parameters (each number). The number of times the dice is rolled will have absolutely no effect on the number of degrees of freedom.

15. The proportions of blood phenotypes A, B, AB and O in the population of all Caucasians in the United States are 0.41, 0.10, 0.04 and 0.45 respectively. To determine whether or not the actual population proportion s fit this set of reported probabilities a random sample of 200 Americans were selected and their blood phenotypes were recorded. The observed and expected cell counts are given as follows. Test the Goodness of fit of these blood phenotype proportions

	A	B	AB	O
Observed (Oi)	89	18	12	81
Expected (Ei)	82	20	8	90

Answer:
H_0: The experimental result support the theory that is the actual population proportion s fit this set of reported probabilities

$$\chi^2 = \frac{\sum_i (Oi - Ei)^2}{E_i} = 3.70$$

$$\chi_{\alpha}^{2}(n-p-1) = \chi_{0.05}^{2}(4-0-1) = \chi_{0.05}^{2}(3) = 7.815$$

From the table of Chi square probabilities

Since Chi Square computed is less than the table value we do not have sufficient evidence to reject H0. That is we cannot declare that the blood phenotypes for American Caucasians is different from those reported earlier. The results are non significant.