Subject	Statistics
Semester	04
Paper no	10
Paper Name	Testing of Hypothesis
Topic no	18
Topic name	Pearson’s Chi-Square Test for Goodness of Fit
SME	Ms Shubharekha
ID	Ms Varsha Shetty

E-learning Module
On
Pearson’s Chi-Square Test for Goodness of Fit

Learning Objectives

At the end of this session, you will be able to:

- Understand Pearson’s Chi Square test
- Explain Chi square test statistic
- Describe the conditions and type of data required to apply the test
- Explain applications of the test

Introduction

Many experiments result 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 colours

- An industrial process manufactures items that can be classified as “acceptable”, “second quality” or “defective”

These are some of the many situations in which the data set has characteristics appropriate for the multinomial experiment.

The Multinomial Experiment

- The experiment consists of **n** independent trials
- The outcome of each trial falls into one of k categories
- The trials are independent

- 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 experimenter counts the observed number of outcomes in each category written as $O_1,O_2,..,O_k = N$

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 i^{th} 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.

You have probably noticed a similarity between the multinomial experiment and the binomial experiment in fact when there are **k=2** categories, the two experiments are identical except for the notation.

Instead of p and q we write p_1 and p_2 to represent the probabilities for the two categories “success” and “failure”.

When we represented Binomial variable we made inferences about the binomial parameters using large sample methods based on **Z** statistic.

In this topic we extend this idea to make inferences about the multinomial parameters p_1,p_2…p_k using different type of statistic.

The statistic whose approximate sampling distribution was derived by a British statistician named Karl Pearson in 1900 is called the Chi-Square (or sometimes Pearson's Chi-Square) statistic

Pearson’s Chi-Square Statistic:

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 p_1=0.1

How many balls would you 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, $\mu = np$ 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

$$\mathbf{Ei=npi}$$

for any of the cells i=1,2,..k

Now suppose you hypothesize values for each of the probabilities p_1,p_2…p_k and calculate the expected number of each category or cell .

If your 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.

Although the mathematical proof is beyond the syllabus 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

The Test Procedure

1.In the first step the given data can be arranged in the form of a frequency distribution wherein the observed frequencies for various classes or cells can be obtained

2.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 given by

$$\text{⁘}^{2} \cong \frac{\underset{i}{\cong} (Oi \cong Ei)^{2}}{E_{i}}$$

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 ⁘^{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.

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.

The test procedure is to compute the statistic χ^2 and to compare it with the tabulated value

$$\chi^2_{\alpha}(n \cong p \cong 1)$$

and to reject the null hypothesis whenever the computed value exceeds the tabulated value.

Conditions or Chi-Square Test

1.The sample observations should be independent

2.The constraints on the frequencies must be linear

3.Number of theoretical frequencies should be less than 5

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.

In most of the situations conditions (1) and (2) are satisfied as random observations are taken and as only one restriction (namely)

$$\underset{i}{\cong} Oi \cong \underset{i}{\cong} Ei$$

is imposed on observations.

Type 0f Data for Chi-Square Test

What is the Chi-square test for?

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).

Applications of Chi Square Test

Pearson's chi-square 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

In contexts where it is important to make a distinction between the test statistic and its distribution, names similar to **Pearson Chi-square** 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 as 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-square 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).

What Does Chi Square Statistic Mean?

Actually, it's 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 **O**bserved frequency

Ei stands for the **E**xpected 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 sigma in front of all that is 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.

Test for Fit of A Distribution- Discrete Uniform Distribution

In this case **N** observations 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

$$Ei \cong \frac{N}{n}$$

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

Other Distributions

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 or parameters.

The number of times the dice is rolled will have absolutely no effect on the number of degrees of freedom.

The result about the number of degrees of freedom is valid when the original data was multinomial and hence the estimated parameters are efficient for minimizing the chi-square statistic.

More generally however, when maximum likelihood estimation does not coincide with minimum chi-squared estimation, the distribution will lie somewhere between a chi-square distribution with (n–p-1) and (n-1) degrees of freedom

Example 1:

Goodness of Fit

For example, to test the hypothesis that a random sample of 100 people has been drawn from a population in which men and women are equal in frequency, the observed number of men and women would be compared to the theoretical frequencies of 50 men and 50 women.

H_0: Men and women are chosen with equal probability in the sample

If there were 44 men in the sample and 56 women, then

$$\text{⁘}^{2} \cong \frac{(44 \cong 50)^{2}}{50} \cong \frac{(50 \cong 50)^{2}}{50} \cong 1.04$$

If the null hypothesis is true (that is men and women are chosen with equal probability in the sample), the test statistic will be drawn from a chi-square distribution with one degree of freedom.

Though one might expect two degrees of freedom (one each for the men and women), we must take into account that the total number of men and women is constrained (100), and thus there is only one degree of freedom (2-1)

Consultation of the chi-square distribution for 1 degree of freedom shows that the probability of observing this difference (or a more extreme difference than this) if men and women are equally numerous in the population is approximately 0.23

This probability is higher than conventional criteria for statistical significance (0.001-0.05), so normally we would not reject the null hypothesis that the number of men in the population is the same as the number of women.

That is, we would consider our sample within the range of what we would expect for a 50 by 50 male by female ratio.

This is all you need to know to calculate and understand Pearson's Chi-square test for goodness of fit.

Example 2:

According to a theory in genetics, the population of seeds of 4 types A, B, C and D are in the ratio 9:3:3:1.

In an experiment among 1600 seeds, the frequency of seeds of each of the above 4 types were 882, 313, 287, 118.

Test the results support the theory.

H_0: The experimental result supports the theory

In this case

N=1600

Oi: 882, 313, 287, 118

That is seeds are having the proportion 9:3:3:1

Probability that the seed falls in group A is

$$9/(9+3+3+1) = 9/16$$

Probability that the seed falls in group B is

$$3/(9+3+3+1) = 3/16$$

Probability that the seed falls in group C is

$$3/(9+3+3+1) = 3/16$$

Probability that the seed falls in group D is

$$1/(9+3+3+1) = 1/16$$

Therefore expected number of seeds out of 1600 seeds to fall into the Group A is

E1= Npi = (1600) 9/16 = 00

Similarly expected number of seeds out of 1600 seeds to fall into the Group B and group C are

E2 = E3 = (1600) 3/16 = 300 each

Expected number of seeds out of 1600 seeds to fall into the Group D

E4=100

$$\chi^2 = \frac{\sum_i (Oi - Ei)^2}{E_i} = 4.7266$$

From the table of Chi square probabilities

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

Since Chi Square computed is less than the table value we do not have sufficient evidence to reject the null hypothesis.

That is we conclude that the experimental results supports the theory at 5% level of significance .

It's a widely popular test because once you know the formula, it can all be done on a pocket calculator or using a software and then compared to simple charts to give you a probability value.

The Chi-square test will prove to be a handy tool for analyzing all kinds of relationships.

Once you know the basics for the applications of the test, expanding to a larger set of values is easy.