6. THE CHI-SQUARE TEST

A **chi-squared test**, also referred to as **chi-square test** or χ^2 **test**, is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Also considered a chi-squared test is a test in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough.

Some examples of chi-squared tests where the chi-squared distribution is only approximately valid:

- Pearson's chi-squared test, also known as the chi-squared goodness-of-fit test or chi-squared test for independence. When the chi-squared test is mentioned without any modifiers or without other precluding context, this test is usually meant (for an exact test used in place of χ^2, see Fisher's exact test).
- Yates's correction for continuity, also known as Yates' chi-squared test.
- Cochran–Mantel–Haenszel chi-squared test.
- McNemar's test, used in certain 2 × 2 tables with pairing
- Tukey's test of additivity
- The portmanteau test in time-series analysis, testing for the presence of autocorrelation
- Likelihood-ratio tests in general statistical modelling, for testing whether there is evidence of the need to move from a simple model to a more complicated one (where the simple model is nested within the complicated one).

One case where the distribution of the test statistic is an exact chi-squared distribution is the test that the variance of a normally distributed population has a given value based on a sample variance. Such a test is uncommon in practice because values of variances to test against are seldom known exactly.

Chi-squared test for variance in a normal population

If a sample of size n is taken from a population having a normal distribution, then there is a result (see distribution of the sample variance) which allows a test to be made of whether the variance of the population has a pre-determined value. For example, a manufacturing process might have been in stable condition for a long period, allowing a value for the variance to be determined essentially without error. Suppose that a variant of the process is being tested, giving rise to a small sample of n product items whose variation is to be tested. The test statistic Tin this

instance could be set to be the sum of squares about the sample mean, divided by the nominal value for the variance (i.e. the value to be tested as holding). Then T has a chi-squared distribution with $n - 1$ degrees of freedom. For example if the sample size is 21, the acceptance region for T for a significance level of 5% is the interval 9.59 to 34.17.

In statistics, **minimum chi-square estimation** is a method of estimation of unobserved quantities based on observed data.

In certain chi-square tests, one rejects a null hypothesis about a population distribution if a specified test statistic is too large, when that statistic would have approximately a chi-square distribution if the null hypothesis is true. In minimum chi-square estimation, one finds the values of parameters that make that test statistic as small as possible.

Among the consequences of its use is that the test statistic actually does have approximately a chi-square distribution when the sample size is large. Generally, one reduces by 1 the number of degrees of freedom for each parameter estimated by this method.

Pearson's chi-squared test

Pearson's chi-squared test (χ^2) is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance. It is suitable for unpaired data from large samples. It is the most widely used of many 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 improve a distinction between the test statistic and its distribution, names similar to Pearson χ-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

Definition

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 is related to the response).

The procedure of the test includes the following steps:

1. Calculate the chi-squared test statistic, χ^2, which resembles a normalized sum of squared deviations between observed and theoretical frequencies (see below).
2. Determine the degrees of freedom, df, of that statistic, which is essentially the number of frequencies reduced by the number of parameters of the fitted distribution.
3. Compare χ^2 to the critical value from the chi-squared distribution with df degrees of freedom, which in many cases gives a good approximation of the distribution of χ^2.

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 $p = s + 1$, where s is the number of co-variates used in fitting the distribution. For instance, when checking a three-co-variate Weibull distribution, $p = 4$, and when checking a normal distribution (where the parameters are mean and standard deviation), $p = 3$. In other words, there will be $n - p$ 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 die, there would be five degrees of freedom because there are six categories/parameters (each number). The number of times the die is rolled will have absolutely no effect on the number of degrees of freedom.

6.1 CHI - SQUARE AS PROOF OF ASSOCIATION: Chi squared analysis is useful in determining the statistical significance level of association rules. We show that the chi squared statistic of a rule may be computed directly from the values of confidence, support, and lift (interest) of the rule in question. Our results facilitate pruning of rule sets obtained using standard association rule mining techniques, allow identification of statistically significant rules that may have been overlooked by the mining algorithm, and provide an analytical description of the relationship between confidence and support in terms of chi{squared and lift.

AN ASSOCIATION RULE is a rule of the form A) B (1) where A and B are item sets, that is, sets of items that appear in a database of transactions. This terminology is that of \market basket" analysis; each transaction item set represents the set of items that are purchased together in a single retail transaction. An association rule such as that in Eq. 1 is meant to represent the statement that transactions that contain the item set A are likely to also contain the item set B, at least within a particular database of transactions.

Association rules have been widely used within data mining since the development of the famous Apriori association rule mining algorithm [1], [2]. Various evaluation measures have been proposed to assess the degree to which an association rule applies to or is of interest in a given context.

Confidence and support are the most commonly used measures in part because of their central role in the Apriori algorithm. In it is shown that the best rules according to many other measures can be found among those that lie along an upper\confidence support border."

Despite the abundance of alternative evaluation measures, chi-squared analysis, a classical technique in statistics for determining the closeness of two probability distributions, continues to be one of the most widely used tools for statistical significance testing in many scientific circles, e.g. bioinformatics.

It was suggested that chi-squared analysis be used to assess the statistical significance level of the dependence between antecedent and consequent in association rules. It also describes a mining algorithm that uses chi-square significance levels to prune the search for item sets during mining. It has been acknowledged that the chi-square statistic does not by itself measure the strength of the dependence of the antecedent and consequent of a rule, and that an

additional measure (such as confidence) is needed for this purpose; it uses the interest (also known as lift) associated with individual cells in the contingency table of the antecedent and consequent.

Using chi-squared analysis to prune the set of mined rules was proposed where lift is again used as a measure of dependence for individual cells of the contingency tables.

The statistical significance of an association rule may be gauged through chi-square analysis. In the approach to this problem, the presence or absence of each item that appears in a rule is viewed as a random variable.

This requires one dimension for each item, leading to high{dimensional contingency tables. Here, we adopt an alternate approach. For a rule A) B, we aggregate the items of the antecedent A and, separately, the items of the consequent B.

In other words, we view the boolean product over each of these item sets as a single binary{valued random variable. This allows us to deal with two{dimensional contingency tables regardless of the number of items that appear in a rule. One advantage of using lower{dimensional tables is that it becomes easier to achieve the minimum cell counts required for validity of chi{squared analysis.

6.2 USE OF CHI-SQUARE: Two key questions in many types of research are whether two variables are correlated, and if so, the strength (or significance) of that relationship. Is there a significant correlation, for example, between gender or ethnicity and political affiliation? The chi-square test is a widely used method for measuring if a significant relationship exists between two nominal or categorical variables, such as gender and political affiliation.

Begin with a hypothesis before you start your data analysis. A common hypothesis in much research is that there is no correlation between the two variables of interest. The chi (rhymes with "my") square test measures the level of deviance from a given hypothesis. The larger the chi-square statistic, the less well the hypothesis fits the data. For example, suppose we are looking at a set of data that asked 125 registered voters (65 women and 60 men) their political party affiliation

(Democratic or Republican). Suppose we know from previous research that 55 percent of voters identified themselves as Democrats. Our working hypothesis is that this 55 percent will be evenly distributed between men and women.

Calculate the expected values based on your hypothesized model of political affiliation by gender. Based on 125 voters, we expect that 55 percent (69 voters) will identify themselves as Democrats. By gender, we expect that 36 women and 33 men will express a preference for the Democratic Party, leaving 29 women and 27 men favoring the Republican Party. Organize your data in a 2-by-2 matrix (two rows and two columns). Let party affiliation be the column variables and gender be the row variables.

3

Compare the actual values from your data with the expected values you estimated in Step 2. For this example, let's say that among the 65 women, 44 percent identified themselves as Democrats and 21 as Republicans, while 36 men claimed a Democratic affiliation and 24 preferred the Republican Party.

4

Calculate the chi-square statistic, which is the sum of the squared differences between the observed and expected values (also known as the residuals), divided by the expected values. You will need this for the four possible combinations of gender and political affiliation specified in your model. If you're using a computer, many statistical and spreadsheet programs can calculate the chi-square statistic for you. In our example, the sum of squared differentials divided by expected values is 4.59.

5

Determine whether the chi-square statistic you calculated in Step 4 is statistically significant. To do this, you need to know two things: the degrees of freedom and the significance level. Degrees of freedom is the number of rows in your table minus one, times the number of columns minus one. Significance level refers to the probability that the observed correlation could have occurred by chance alone. Many researchers prefer a .05 significance level, meaning there is only a 5 percent likelihood that the observed relationship is pure chance. In our example, we have only 1 degree of freedom. Using your statistics book (usually in the appendix), look up the chi-square value that corresponds to the significance level and degrees of freedom. For our example, the chi-square value for 1 degree of freedom and .05 significance level is 3.84. Our value of 4.59 is greater, meaning there is a

statistically significant relationship between gender and political affiliation, with women being significantly more likely to identify themselves as Democrats.

6.3 CHI-SQUARE GOODNESS OF FIT TEST AS: With the goodness-of-fit test one is interested in determining whether a given distribution of data follows an expected pattern. For the test of association, however, one is interested in learning whether two (or more) categorical variables are related. Typically one will find two categorical variables depicted in a contingency table (a cross-tabulation of the frequencies for various combinations of the variables).