Different types of chi square tests essay

What are the different types of chi- square tests? 1. INTRODUCTION1. 1 ? 2 distribution and its properties A chi-square (? 2) distribution is a set of density curves with each curve described by its degree of freedom (df). The distribution have the following properties: – Area under the curve = 1 – All ? 2 values are positive i.

e. the curve begins from 0 (except for df= 1) increases to a peak and decreases towards 0 as its asymptote – The curve is skewed to the right, and as the degree of freedom increases, the distribution approaches that of a normal distribution Fig. 1 Graph of ? 2 distribution with differing degrees of freedom Each ? value is computed by the formula: ? 2 =? (O-E)2 E where O = observed counts from the sample Equation 1 and E= expected counts based on the hypothesized distribution1. 2 Types of ? 2 tests and their purpose For a single population, to determine if the observed distribution in the population conforms to a specific known distribution or a previously studied distribution, the ? 2 test for goodness-of-fit can be used. An example of this usage include: Mendel’s genetic model predicts that the phenotypic distribution of two phenotypes, each phenotype having a dominant and recessive allele, will follow the ratio of 9: 3: 3: 1. A study done to confirm this makes use of ? 2 test for goodness-of-fit to determine if the observed population fits into the theoretical model.

We will discuss this example in detail in the next section. To compare the distribution of two populations, the ? 2 test for homogeneity of population can be used. The data in this case can be represented in a two way table with the different populations in the rows and the distribution data based on certain categorical variable in columns. To test if the distribution of

categorical variable is the same cross several populations, the ? 2 test for homogeneity of population is used. An example of this can be to find out if the proportions of teachers with PhD teaching a specific level in high school across three different countries are the same. Data can also be represented in a two-way table even when it is taken from the same population, but divided according to two categorical variables, one in the row and one in the column.

To test if the variables have any relationships, the ? 2 test for association/independence can be used. For instance, if we want to find out if the choice of university (urban, suburban, rural) is associated with the place the student live in (urban, suburban, rural). We could tabulate the data in a 3 by 3 table and carry out a chi-square test for association/independence. 2. TEST FOR GOODNESS OF FITIn a single population with a hypothesized distribution with n outcomes, i.

e. each observation falls into one of n possible outcomes and the proportion of the outcomes, number of observations in each outcome divided by the total number of observations, is hypothesized to follow a certain predetermined distribution. We test the null hypothesis H0: The actual population proportions are equal to the hypothesized population proportion The chi-square statistics are calculated using Equation 1, X2 = ? (O-E)2/E where the expected count is calculated by multiplying the hypothesized proportions to the total number of observation. Conditions for using the ? 2 test for goodness of fit: 1. All individual expected counts must be at least 1 2. No more than 20% of the expected counts are less than 5 The X2 (chi-

square statistic) has approximately a ? 2 distribution with (n – 1) degrees of freedom.

For testing the H0 vs the alternate hypothesis, Ha: The actual population proportions are not equal to the hypothesized population proportions the P-value is P(? 2 >= X2). To evaluate the P-value using graphing calculator, the input command under Home is as follow: TIStat. chi2Cdf (X2, ? , df) The command can also be found in Apps/Home/Catalog/F3 Flash Apps/chi2Cdf(Value returned will be the P-value for ? 2 >= X2. The smaller the P-value, the less likely the sample conforms to the hypothesized distribution.

There is strong evidence against H0 when the P-value is small. The follow up analysis compared the observed counts with the expected counts. The largest component that makes up the X2 statistic is the proportion that deviates most. E.

g. 1. Chi-square test for goodness of fit used widely in the field of genetics, where geneticists test out if the phenotypic appearance is resulted from the theoretical genotypes. The hypothesized population distribution is usually derived with a Punnett square.

Two true breeding pea plants, one with yellow round seed and one with green wrinkled seed, were crossed, producing dihybrids, called the F1 generation. Yellow round seed have the genotypes YYRR, while green wrinkled seed have the genotype yyrr. Y is the genotype coding for yellow colour and is dominant over y (green), and R coding for round dominant over r (wrinkled). The F1 dihybrids are all heterozygous, having the genotype

YyRr, and appear yellow and round due to the dominant genes. Self pollination of the F1 generation produces the following distribution of F2 generation: ? YR? Yr? yR? yr ? YRYYRRYYRrYyRRYyRr yellow round ? YrYYRrYYrrYyRrYyrr yellow wrinkled ? yRYyRRYyRryyRRyyRr green round ? yryyRrYyrryyRryyrr green wrinkled Probability of each genotype is ? ? = 1/16 The phenotype ratio is therefore Yellow round : Yellow wrinkled : Green round : Green wrinkled = 9/16 : 3/16 : 3/16 : 1/16 = 9: 3: 3: 1 This is the hypothesized distribution.

The observed distribution of 556 pea plants grown is Yellow roundYellow wrinkledGreen roundGreen wrinkled Observed31510810132 Expected31310410435 H0: the actual distribution is equal to the hypothesized distribution. Ha: the actual distribution is different from the hypothesized distribution Conditions for using chi-square test is fulfilled and all expected counts are > 5. X2 = (315-313)2/313 + (108-104)2/104 + (101-104)2/104 + (32-35)2/35 = 0. 510 P-value = chi2Cdf (0. 510, ? , 3) = 0. 9166 The P-value is so large that we have no evidence to reject the null hypothesis, we therefore conclude that the actual population distribution of the pea plants is equal to the hypothesized ratio of 9: 3: 3: 1.

3. TEST FOR HOMOGENEITY OF POPULATION

3. 1 Setting up a two-way table The different populations are placed in the rows and the corresponding population proportions are placed in the columns. Using an example to illustrate this: we are interested to find out if the proportion of people who can memorize a string of 15 random alphabets is the same across certain age groups. What we can do is to have a simple

random sample of people in each age group as specified below, and we subject them to memorize a string of alphabets. The results can be presented in a table such as: AgeObserved CountsTotalProportion of people memorizedExpected Counts MemorizedDid not memorizeMemorizedDid not memorize <10601402000.

376. 25123. 75 10-20901102000. 4576.

25123. 75 20-30891112000. 4576. 25123. 75 30-40661342000.

3376. 25123. 75 Total3054958000. 38125305123. 75 This table is known as a two-way table that that is 4 X 2 (row X column), which shows the relationship between two categorical variables.

The age group is the explanatory variable and the success/failure of memorizing is the response variable. Fig. 2 Distribution for proportion of people who memorized across 4 age groups

3. 2 Calculating expected counts We are testing the null hypothesis that the proportions of people who can memorize the string of alphabets are the same across age groups. H0: the proportions of people who memorized are the same in each age group The alternate hypothesis states Ha: the proportions of people who memorized are not all equal in the different age groups. To use the chi square test, we need to know how to find the expected counts for each cell.

Let’s try to compute the expected count for cell 1. The probability of memorizing is (473/1657). The probability of a person <10 yrs old memorizing is the probability times the total number of subjects that are

<10 yrs old, which gives (473/1657)x(291). The generalized formula for computing expected count is thus given by: Expected count = (row total x column total)/ table total The expected counts for this example are tabulated in the table above.

Since the counts fulfill the conditions for chi square test as they are all > 5, we can proceed to use the chi-square test for homogeneity of population. The degree of freedom for this chi-square distribution is computed by df = (r-1)(c-1) where r is the number of rows and c the number of columns in the r X c table. X2 = ? (O-E)2/E = 31. 440df = (4-1)(2-1) = 3P-value = chi2Cdf (31. 44, ? , 3) = 6. 87 EE-7 Such a small P-values gives strong evidence against H0.

We therefore reject H0 and conclude that the proportions of people who can memorize the string of alphabets is significantly different across different age groups.

3. 3 Follow up analysis Since the results show that the proportions are not equal, we need to do a follow up analysis on which of the groups contributed to the biggest component of the chi-square value. By looking at the components of chi-square, i. e. (O-E)2/E value for each cell.

We find that the largest component is a result of cell 1, the observed counts of successes in memorizing is very much lower than the expected value. Note: The ? 2 test can never be one-sided because with the possibility of more than two categories, the P-value will show significance if any one of the observed proportions is different from the hypothesized distribution. This highlights the main difference between a chi-sq test for two populations and

the z-test for two population proportions. The chi-square test only returns value that it two-sided in this case. 4.

TEST FOR ASSOCIATION/INDEPENDENCEBesides population proportions, other types of categorical data can also be represented using the two way table, especially if we want to find out if one variable has an association or is independent from another variable. For instance, we are interested to find out if smoking is associated marital status. A survey was conducted on 1691 individuals in the UK on how many cigarettes they smoke in a week and what is their current marital status. The data can be represented as the following: Marital StatusSmokeTotalExpected counts No? 3 per day> 3 per dayNo? 3 per day> 3 per day Divorced1031543161120. 82115. 33624.

945 Married6695687812609. 3676. 8303125. 81 Separated469146951. 78066.

5286810. 6907 Single2696693428321. 1940. 496766. 3134 Widowed1821425221165. 84820.

910734. 2413 Total126916026216911269160262 The expected counts are obtained in the same way: E = (row total X column total)/ table total A comparison between the amounts of smoke taken can be represented using a stacked bar chart as shown: Marital StatusSmoke No? 3 per day> 3 per day Divorced0. 6397520. 0931680. 267081 Married0.

8238920. 0689660. 107143 Separated0. 6666670. 1304350.

202899 Single0. 6285050. 1542060. 21729 Widowed0.

8235290. 0633480. 113122The conditions for chi-square test is satisfied, all expected counts > 5. The null hypothesis in this case is H0: there is relationship between marital status and whether the person smokes.

Versus the alternate hypothesis that Ha: there is a relationship between marital status and the occurrence of smoking. The chi-square statistic is computed using the same formula as follow: X2 = ? (O-E)2/E = 76. 794df = (5-1)(2-1) = 4 P-value = chi2Cdf (76. 794, ? , 4) = 8. 331EE-16 Such a small P-value provides strong evidence against H0.

We therefore reject H0 and conclude that there is a relationship between marital status and smoking tendencies. The size and nature of this association is shown in Fig. 3 above. Notes: The last two types of chi-square tests can be differentiated by examining the design of the study. For the test for homogeneity of populations, the data comes from samples taken from two or more populations, and each individual is grouped based on one categorical variable. In the test for association/independence, the data is taken from one population, but grouped according to two categorical variables.

The aim of the test will also differ, based on these different sampling designs, as can be seen from the null hypothesis. The chi-square test is a non-parametric test, so the data itself does not need to conform to normality. However, one important condition is that the deviations of the data conform to normality. This condition will be satisfied if the sample is a simple random sample. After thoughts Follow up analysis in this case only identifies the component(s) of chi-square that is the largest, i. e.

most likely to have created the non-conformity to the hypothesized distribution. However, the textbook fails to mention how to test if that component is the only proportion that is causing the misfit into the hypothesized distribution. A more complete analysis would be to take out the component having the largest deviation, and subject the rest of the results to another round of testing. This should be continued until the remaining populations are homogeneous or until the last two populations are proven to be non-homogeneous. We can then say with certainty which proportions are the ones that are significantly different from and which proportions actually follow the hypothesized distribution.

There is an ambiguity regarding the comparison between chi-square test for homogeneity of population between two populations and the two proportions z-test. Since z-test requires the sample to be normal, while the chi-square test does not. Yet the tests return the same value for a two-sided hypothesis. It would be logical to think that the normality requirements which place more constraints on the data would give a more accurate P-value. For the test for homogeneity of populations, it would seem that the nature of the sampling design is an experimental design instead of simply data collection, where the categorical data classify an individual in each population is the response variable to a certain treatment.

It was not made clear if it indeed need to be an experiment. APPENDIXThe formula for the probability density function of the chi-square distribution is where ? is the degree of freedom and ? is the gamma function defined by for values of z that are positive integers, Using this property, we can sub in increasing values of x to get the chi-square distribution for the various

degrees of freedom as shown in Fig. 1. REFERENCESEngineering statistics handbook. Chi-square distribution.

available [online] http://www. itl. nist. gov/div898/handbook/eda/section3/eda3666. htm Date retrieved: 21/3/08 Wikipedia. Chi-square distribution.

available [online] http://en. wikipedia. org/wiki/Chi-square_distributionDate retrieved: 21/3/08 Wikipedia. Gamma function. available [online] http://en. wikipedia.

org/wiki/Gamma_function Date retrieved: 21/3/08 Engineering statistics handbook. Underlying assumptions. available [online] http://www. itl. nist. gov/div898/handbook/eda/section2/eda21.

htm Date retrieved: 23/3/08 Stats4Schools. Large Datasets on Smoking. available [online] http://www. stats4schools.

gov. uk/large_datasets/smoking/default. asp Date retrieved: 21/3/08 Neil A. Campbell and Jane B.

Reece. Biology. 7th Edition. Pearson publishing. 2005. Daniel S.

Yates et. al. The Practice of Statistics. 2nd Edition. W.

H. Freeman and Company. 2003.