

STP 226

ELEMENTARY STATISTICS

NOTES

PART 1V INFERENCE STATISTICS

CHAPTER 12

CHI-SQUARE PROCEDURES

12.1 The Chi-Square Distribution

A variable has a chi-square distribution if the shape of its distribution follows the shape of a right-skewed curve called the chi-square (χ^2) curve.

Basic Properties of χ^2 -Curves

1. The total area under a χ^2 -curve is equal to 1.
2. A χ^2 -curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
3. A χ^2 -curve is right skewed.
4. As the number of degrees of freedom becomes larger, χ^2 -curves look increasingly like normal curves.

Using the χ^2 -Table

The table (Table V) gives the area to the right of a χ^2 –value.

χ^2_{α} - has an area of α to its right for a specified degree of freedom.

12.2 Chi-Square Goodness-of-fit Test

Chi-square goodness-of-fit test –procedure used to perform a hypothesis test about the distribution of a qualitative (categorical) variable or a discrete quantitative variable having only finitely many possible values.

Expected frequency $E = np$ where n is the sample size and p is the relative frequency.

For a sample of size $n = 500$

Type of violent crime	Relative Frequency (p)	Expected Frequency (E=np)
Murder	0.012	$500 \cdot 0.012 = 6.0$
Forcible rape	0.054	$500 \cdot 0.054 = 27.0$
Robbery	0.323	$500 \cdot 0.323 = 161.5$
Agg. Assault	0.61	$500 \cdot 0.611 = 305.5$

Type of violent crime	Observed Frequency	Expected Frequency	Difference	Square of distance	Chi-square subtotal
x	O	E	O - E	$(O - E)^2$	$(O - E)^2/E$
Murder	9	6.0	3.0	9.00	1.500
Forcible rape	26	27.0	-1.0	1.00	0.037
Robbery	144	161.5	-17.5	306.25	1.896
Agg. assault	321	305.5	15.5	240.25	0.786
SUMS	500	500.0	0		4.219

Distribution of the χ^2 –statistic for a chi-square goodness-of-fit Test

For a chi-square goodness-of-fit test, the test statistic, $\chi^2 = \sum(O - E)^2/E$ has approximately a chi-square distribution if the null hypothesis is true. The number of degrees of freedom is one less than the number of possible values for the variable under consideration.

The Chi-Square Goodness-Of-Fit Test

ASSUMPTIONS

1. All expected frequencies are 1 or greater.
2. At most 20% of the expected frequencies are less than 5.

STEPS

1. The null and alternative hypotheses are

H_0 : The variable under consideration has the specified distribution.

H_a : The variable under consideration does not have the specified distribution.
2. Calculate the expected frequency for each possible value of the variable under consideration using the formula $E = np$, where n is the sample size and p is the relative frequency (or probability) given for the value in the null hypothesis.
3. Check whether the expected frequencies satisfy Assumptions 1 and 2. If they do not, this procedure should not be used.
4. Decide on the significance level, α .

5. The critical value is χ^2_{α} with $df = k - 1$, where k is the number of possible values for the variable under consideration. Use Table V to find the critical value.
6. Compute the value of the test statistic $\chi^2 = \Sigma(O - E)^2/E$, where O and E denote observed and expected frequencies, respectively.
7. If the value of the test statistic falls in the rejection, reject H_0 ; otherwise, do not reject H_0 .
8. State the conclusion in words.

The p-value approach will only compare the p-value of the test statistic with the α level (the p-value of the test)

12.3 Contingency Tables; Association

Chi-square independence test – Chi-square procedure

Contingency Tables

Univariate data – data values of one variable of a population

Bivariate data – data values of two variables of a population

Contingency table (two-way table) – frequency distribution for bivariate data

Cell – the intersection of a row and a column in a contingency table

Eg. of a two-way table

Class Level					
Political Party	Freshmen	Sophomore	Junior	Senior	Total
Democratic	1	4	5	3	13
Republican	4	8	4	2	18
Other	1	3	3	2	9
Total	6	15	12	7	40

Association is when knowing the value of one variable imparts information about the value of the other variable where the two variables are categorical or quantitative with only finitely many possible values.

Class Level					
Political Party	Freshmen	Sophomore	Junior	Senior	Total
Democratic	0.167	0.267	0.417	0.429	0.325
Republican	0.667	0.533	0.333	0.286	0.450
Other	0.167	0.200	0.250	0.286	0.225
Total	1.000	1.000	1.000	1.000	1.000

Conditional distribution – the distribution of political party for each of the freshmen, sophomores, juniors or seniors. Freshmen are divided into three classes (class levels) (16.7% Dem., 66.7% Rep., 16.7% Other). This is called the conditional distribution of political party for freshmen. The same applies to the students in the other groups.

Marginal Distribution – the total column gives the unconditional distribution of political party. (32.5% Dem., 45% Rep., 22.5% Other)

Segmented bar graph can be used to view/understand the concept of association

First 4 segmented bars give the conditional distribution for political party for freshmen, sophomores, juniors, and seniors.

The last segmented bar gives the marginal distribution of political party.

There is an **association** between two variables of a population if the conditional distributions of one variable given the other are not identical.

Statistically dependent – there is an association between two variables

Statistically independent – there is no association between two variables

12.4 Chi-square Independence Test

Distribution of the χ^2 –statistic for a chi-square independence test

For a chi-square independence test, the test statistic $\chi^2 = \sum \frac{(O - E)^2}{E}$

Has approximately a chi-square distribution if the null hypothesis of nonassociation is true. The number of degrees of freedom is $(r - 1)(c - 1)$, where r and c are the number of possible values for the two variables under consideration.

The Chi-square Independence Test

ASSUMPTIONS

1. All expected frequencies are 1 or greater.
2. At most 20% of the expected frequencies are less than 5.

STEPS

1. The null and alternate hypothesis are

H_0 : The two variables under consideration are not associated.

H_a : The two variables under consideration are associated.

2. Calculate the expected frequencies using the formula $E = \frac{R \cdot C}{n}$
where R = row total, C = column total, and n = sample size. Place each expected frequency below its corresponding observed frequency in the contingency table.
3. Check whether the expected frequencies satisfy assumptions 1 & 2.
If they do not, this procedure should not be used.
4. Decide on the significance level, α .
5. The critical value is χ^2_{α} with $df = (r - 1)(c - 1)$, where r and c are the number of possible values for the two variables under consideration. Use Table V to find the critical value.
6. Compute the value of the test statistic
where O and E represent observed and expected frequencies, respectively.

7. If the value of the test statistic falls in the rejection region, reject H_0 ; otherwise, do not reject H_0 .
8. State the conclusion in words.