Chi-square

Agenda

- Chi-square test
  - Goodness of Fit
  - Independence of Attributes

Tests for Categorical Data

Tests for categorical data

- The collected data is at times best represented by categories
- These categories are summarized by their frequency of occurrence. It may be of interest whether this frequency is equal to the expectation/claim
- It may also be of interest to know whether the categories are statistically independent

Chi-square tests

- These interests are tested by non-parametric ways
- Tests based on the chi-square distribution are used
- The chi-square tests are used to test:
    - The goodness of fit
    - The independence of two attributes
- Chi-square tests are also used to test for population variance

Chi-square distribution

- Chi-square distribution is sum of squared of standard normals

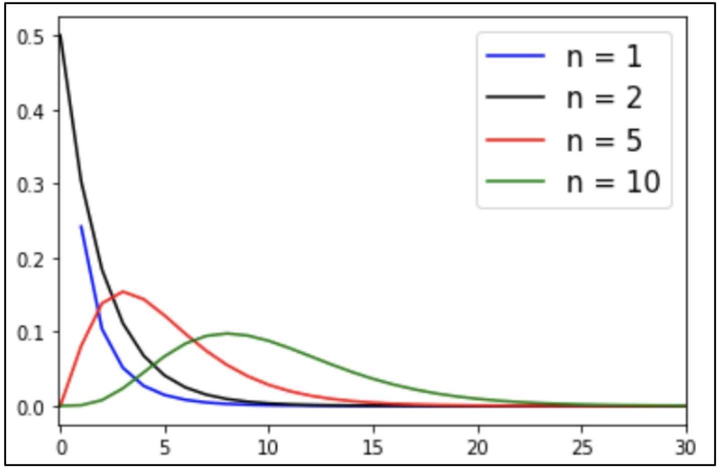
  Let $X_1, X_2, \ldots, X_n$ be n standard normal variates,

  then $Y = X^2_1 + X^2_2 + \ldots + X^2_n$,

  Y follows χ^2 distribution with *n* degrees of freedom.

- The mean of the distribution is *n* and its variance is *2n*

- The distribution is positively skewed



χ^2 Test for Goodness of Fit

Chi-square test for goodness of fit

At an emporium, the manager is interested in knowing age group which visits the mall during the day. He defines categories as - children (age < 13), teenagers (13 ≤ age < 20), adults (20 ≤ age < 55) and senior citizens (55 ≤ age). Moreover, he wishes to plan his inventory of goods accordingly.

He claims that out of all the people who visited 5% are children, 38% are teenagers, 2% are senior citizens are remaining are adults.

Can the owner verify the managers claim?

Chi-square test for goodness of fit

- The hypothesis to test whether the data fits the a specified distribution

H_0: There is no difference between observed frequencies and expected frequencies

ainst

H_1: There is difference between observed frequencies and expected frequencies

- Failing to reject H_0, implies that there is no difference between observed frequencies and expected frequencies

Chi-square test for goodness of fit

- The test statistic is given by

observed frequency

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{e_i} - N$$

total number of observations

estimated frequency

- Under H_0, the test statistics follows χ^2 distribution with k-p-1 d.f

where *k*: number of class frequencies

p: number of parameter estimated for fitting

Ref: Test statistic for goodness of fit (A.1)

Chi-square test for goodness of fit

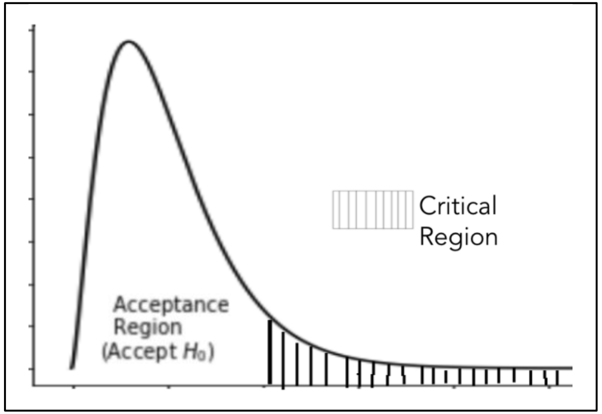
Decision Rule:

Reject H_0 if $\chi^2_{calc} \geq \chi^2_{k-p-1,\alpha}$

or

Reject H_0 if p-value $\leq \alpha$

Where, α is the level of significance (l.o.s.)



Test for goodness of fit

Question:

At an emporium, the manager is interested in knowing age group which visits the mall during the day. He defines categories as - children, teenagers, adults and senior citizens. He plans to have his inventory of goods accordingly. He claims that out of all the people who visited 5% are children, 38% are teenagers, 2% are senior citizens are remaining are adults.

From a sample of 180 people it was seen that 25 were children, 50 were teenagers, 90 were adults and 15 were senior citizens

Test the manager's claim at 95% confidence level.

Test for goodness of fit

Solution:

We can tabulate the given data as follows:

	Manager Claimed Frequency	The frequency expected from 180 customers (e_i)	The frequency observed from 180 customers (O_i)
Children	5%	0.05 x 180 = 9	25
Teenagers	38%	0.38 x 180 = 68.4 ≅ 68	50
Adults	55%	0.55 x 180 = 99	90
Senior Citizens	2%	0.02 x 180 = 3.6 ≅ 4	15

Test for goodness of fit

Solution:

To test, H_0: The managers claim is correct   Against H_1: The managers claim is false

The test statistic is $$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2}{e_i} - N$$

$$= [\frac{25^2}{9} + \frac{50^2}{68} + \frac{90^2}{99} + \frac{15^2}{4}] - 180$$

$$= 64.27$$

Test for goodness of fit

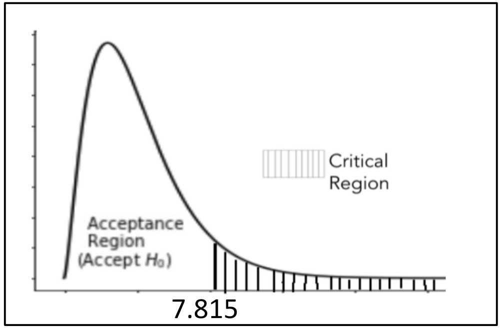
Solution:

Here there are 4 class frequencies, i.e k = 4. Since no parameter was calculated p = 0

From the statistical table for χ^2 distribution, $\chi^2_{k-p-1,\alpha} = \chi^2_{3,0.05} = 7.815$

The test statistic $\chi^2_{calc} = 64.27$

Since $\chi^2_{calc} > \chi^2_{k-p-1,\alpha}$, reject H_0.



The managers claim is false, his claim is different than what was observed from the data

Test for goodness of fit

Python solution:

```
# given observed values
observed_value = [25, 50, 90, 15]

# expected count
exp_count = [0.05, 0.38, 0.55, 0.02]

# calculate the expected values for each category
# expected_value = (np.array(exp_count) * 180)
expected_value = [9, 68, 99, 4]

# use the 'chisquare()' to perform the goodness of fit test
# the function returns the test statistic value and corresponding p-value
# pass the observed values to the parameter, 'f_obs'
# pass the expected values to the parameter, 'f_exp'
stat, p_value = chisquare(f_obs = observed_value, f_exp = expected_value)

print('Test statistic:', stat)
print('p-value:', p_value)
```

```
Test statistic: 64.2773321449792
p-value: 7.160266387019384e-14
```

As p-value < 0.05, we reject H_0.

- If the expected frequencies $e_i \geq 5$ and the total frequencies are large (≥ 50) the test can be used

- If $e_i < 5$, the class is merged with the neighbouring class for observed and expected frequencies until the it becomes ≥ 5

- It is not applicable for testing the goodness of fit of a straight line or any curve (exponential curve, second degree curve)

χ^2 Test for Independence of Attributes

Chi-square test for independence of attributes

- The hypothesis to test independence of attributes

  H_0: The attributes are independent against H_1: The attributes are dependent

- Failing to reject H_0, implies that the attributes are independent

- Decision rule: Reject H_0 at α l.o.s if $\chi^2_{(r-1)(s-1)} \geq \chi^2_{(r-1)(s-1);\alpha}$

  or

  Reject H_0 if p-value $\leq \alpha$

  Where, α is the level of significance (l.o.s.)

Chi-square test for independence of attributes

- The test statistic is given by

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{e_{ij}} - N$$

observed frequency in i^{th} row and j^{th} column

total number of observations

estimated frequency in i^{th} row and j^{th} column

- Under H_0, the test statistics follows χ^2 distribution with (r-1)(c-1) d.f

where *r* levels for a category and *c* levels for another category

Ref: Test statistic for independence of attributes (A.2)

Test for independence of attributes

Question:

A study was conducted to test the effect of the malaria parasite - plasmodium falciparum - on heterozygous and homozygous humans. The vaccine was given to a cohort of 252 humans. Test whether the heterozygous humans are better protected than homozygous.

	Infected with plasmodium falciparum	Not infected with plasmodium falciparum
Heterozygous humans	93	51
Homozygous humans	68	40

Test for independence of attributes

Solution:

Let X: The zygote type - Homozygous or Heterozygous
Y: Whether infected or not with malaria parasite

Here X and Y are two attributes.

To test, H_0: The attributes are independent Against H_1: The attributes are dependent

Here there are 2 rows and 2 columns.

Let us computed the expected frequency.

Test for independence of attributes

Solution:

In order to compute the expected frequency, first compute the total of the each column and row.

	Infected with plasmodium falciparum	Not infected with plasmodium falciparum	Total
Heterozygous humans	93	51	144
Homozygous humans	68	40	108
Total	161	91	252

Test for independence of attributes

Solution:

The expected frequencies are computed as

$$e_{ij} = \frac{\text{Total}_{\text{row}} \times \text{Total}_{\text{column}}}{N}$$

$$e_{11} = \frac{144 \times 161}{252} = 92$$

	Infected with plasmodium falciparum	Not infected with plasmodium falciparum	Total
Heterozygous humans	$\frac{144 \times 161}{252} = 92$		144
Homozygous humans			108
Total	161	91	252

Test for independence of attributes

Solution:

Similarly compute the expected frequencies for other classes

	Infected with plasmodium falciparum	Not infected with plasmodium falciparum	Total
Heterozygous humans	$\frac{144 \times 161}{252} = 92$	$\frac{144 \times 91}{252} = 52$	144
Homozygous humans	$\frac{108 \times 161}{252} = 69$	$\frac{108 \times 91}{252} = 39$	108
Total	161	91	252

Test for independence of attributes

Solution:

The observed frequency (O_{ij})

	Infected with plasmodium falciparum	Not infected with plasmodium falciparum	Total
Heterozygous humans	93	51	144
Homozygous humans	68	40	108
Total	161	91	252

The expected frequency (e_{ij})

	Infected with plasmodium falciparum	Not infected with plasmodium falciparum	Total
Heterozygous humans	92	52	144
Homozygous humans	69	39	108
Total	161	91	252

Test for independence of attributes

Solution:

The test statistic is computed as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{e_{ij}} - N$$

$$= \frac{93^2}{92} + \frac{51^2}{52} + \frac{68^2}{69} + \frac{40^2}{39} - 252$$

$$= 0.070$$

Test for independence of attributes

Solution:

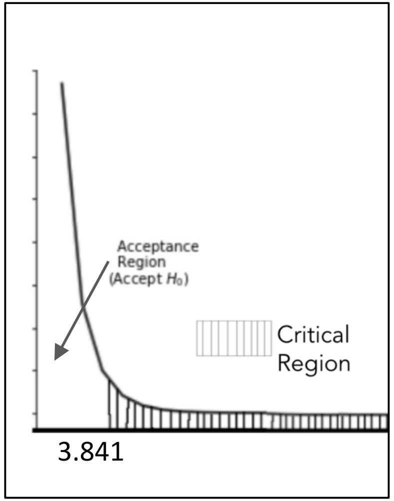
Here there are 2 levels of one attribute and 2 levels of another.
Thus the degrees of freedom are (r-1)(c-1) = (2-1)(2-1) = 1

From the statistical table for χ^2 distribution, $\chi^2_{(r-1)(s-1),\alpha} = \chi^2_{1,0.05} = 3.841$

The test statistic $\chi^2_{calc} = 0.070$

Since $\chi^2_{calc} < \chi^2_{k-p-1,\alpha}$, we fail to reject H_0.

The attributes are independent.



Test for independence of attributes

Python solution:

```
# use the 'chi2_contingency()' to check the independence of variables
# pass the observed values to the parameter, 'observed'
# 'correction = False' will not apply the Yates' correction
test_stat, p, dof, expected_value = chi2_contingency(observed = observed_value, correction = False)

# print the output
print("Test statistic:", test_stat)
print("p-value:", p)
```

```
Test statistic: 0.07023411371237459
p-value: 0.790996215494177
```

As p-value > 0.05, we fail to reject H_0.

Independence of attributes

Question:

A psychologist wants study whether the happiness quotient of children in the house is related to the family income. He collects data of 1300 children is there enough evidence to claim that they are related.

	Low income	Moderate income	High income
Happy	245	354	243
Unsatisfied	98	220	140

Tests based on Chi-squared distribution for categorical data are one tailed tests.