Statistics for Computing MA4413

Lecture 17

The Chi-Squared Test

Kevin Burke

kevin.burke@ul.ie

Introduction

We have looked at tests involving the mean and the variance.

The chi-squared test is based on the **whole distribution** of values.

In particular, the test is used to compare the **observed** distribution (i.e., the data) to some **theoretical** distribution (i.e., a mathematical model).

Note: chi (pronounced “kye”) is a Greek letter: χ.

Example: Rolling a Die

We roll a die 60 times and observe the following frequencies:

Value on Die	*1*	*2*	*3*	*4*	*5*	*6*
Observed Frequency	14	9	13	5	12	7

The question that arises is: *Does this die appear to be fair?*

In other words, we wish to check whether the *observed distribution* differs significantly from the *theoretical distribution:*

Value on Die	*1*	*2*	*3*	*4*	*5*	*6*
Expected Frequency	10	10	10	10	10	10

Note: $\frac{1}{6} \times 60 = 10$.

Observed Vs Expected

In constructing the chi-squared test, we have:

- **Observed:** The number of times we *observe* a particular value in the data.

- **Expected:** The number of times we would *expect* to observe this value if the theoretical distribution fits the data.

The test is based on the fact that the observed minus expected $(o - e)$ will be small if the two are roughly in agreement.

We actually don't use $o - e$ as the *distance measure*, but rather:

$$\frac{(o-e)^2}{e}.$$

Hypotheses

For the chi-square test, the null and alternative hypotheses are:

$$H_0: \quad D = 0$$

$$H_a: \quad D > 0$$

where "D" represents the true, but unknown, scaled-squared-distance between the distributions.

In words, H_0 says that "what we observe is what we would expect", i.e., the distance is approximately zero.

The above hypotheses are usually written as:

H_0 :	the model *fits* the data
H_a :	the model *does not fit* the data

(note: "model" here means the theoretical distribution)

The Chi-Squared Test

- This test is **always one-tailed** $\Rightarrow$ **never** divide α by two.

  (being a squared distance measure, χ^2 *cannot* be negative $\Rightarrow$ upper tail only)

- Test statistic:

$$\boxed{\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}}$$

- Critical value: The $\chi^2_{\nu,\,\alpha}$ value from the **chi-squared tables**, e.g., $\chi^2_{5,\,0.05} = 11.070$, $\chi^2_{8,\,0.05} = 15.507$, $\chi^2_{11,\,0.01} = 24.725$.

- Decision: Reject H_0 if $\chi^2 > \chi^2_{\nu,\,\alpha}$, i.e., if the test statistic is greater than the critical value.

Degrees of Freedom

The degrees of freedom for the chi-squared test are:

$$\boxed{\nu = n_f - 1 - k}$$

where

- n_f is the number of observed frequencies in the frequency table.
- k is the number of *estimated parameters* for the theoretical model.

This will become clearer when we look at some examples.

Warning

The chi-squared test does **not** work well if any expected frequencies are less than 5.

We can get around this by combining cells in the frequency table (more on this later).

Goodness-of-Fit Test

The chi-squared test is called a **goodness-of-fit** test because it tells us how well the theoretical model *fits* the observed data.

"The fit is good" means that theory agrees with what we observe.

The better the fit, the more useful the mathematical model will be for describing the phenomenon under study.

Example: Rolling a Die

Value on Die	*1*	*2*	*3*	*4*	*5*	*6*	$\sum$
Observed: o_i	14	9	13	5	12	7	60
Expected: e_i	10	10	10	10	10	10	60
$o_i - e_i$	4	-1	3	-5	2	-3	
$(o_i - e_i)^2$	16	1	9	25	4	9	
$\frac{(o_i - e_i)^2}{e_i}$	1.6	0.1	0.9	2.5	0.4	0.9	6.4

$\Rightarrow \chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = 6.4$ is the test statistic.

Example: Rolling a Die

If testing at the 5% level $\Rightarrow \alpha = 0.05$ (do not divide by two here).

There are 6 frequencies ($n_f = 6$) and no estimated parameters ($k = 0$) $\Rightarrow$ degrees of freedom: $\nu = n_f - 1 - k = 6 - 1 - 0 = 5$.

Thus, the critical value is $\chi^2_{5,\,0.05} = 11.070$ (the rejection region lies above this value).

Since $6.4 < 11.07$, we accept the null hypothesis that there is no difference between what has been observed and what we expect if the die is fair.

Conclusion: the evidence suggests that the die is fair.

Example: Random Number Generator

Let's assume that we wish to generate random numbers according to the following distribution:

x_i	0	2	4	6	8
$p(x_i)$	0.2	0.2	0.5	0.05	0.05

We generate 80 numbers and we observe the following frequencies:

x_i	0	2	4	6	8
o_i	16	16	40	4	4

Does this evidence suggest that we have programmed the generator correctly?

Example: Random Number Generator

Expected frequencies are calculated by multiplying the overall total, 80, by the theoretical probabilities:

x_i	*0*	*2*	*4*	*6*	*8*
o_i	10	11	42	11	6
$p(x_i)$	0.2	0.2	0.5	0.05	0.05
$80 \times p(x_i) = e_i$	16	16	40	4	4

All e_i values must be greater than 5 $\Rightarrow$ combine cells:

x_i	*0*	*2*	*4*	*6 or 8*
o_i	10	11	42	17
e_i	16	16	40	8

Example: Random Number Generator

x_i	0	2	4	6 or 8	$\sum$
o_i	10	11	42	17	80
e_i	16	16	40	8	80
$\frac{(o_i - e_i)^2}{e_i}$	2.250	1.562	0.100	10.125	14.037

If testing at the 5% level $\Rightarrow \alpha = 0.05$ (do not divide by two).

There are 4 frequencies (after combining cells) and no estimated parameters $\Rightarrow$ degrees of freedom: $\nu = n_f - 1 - k = 4 - 1 - 0 = 3$.

Thus, the critical value is $\chi^2_{4,\,0.05} = 7.815$ (the rejection region lies above this value).

Example: Random Number Generator

The test statistic $\chi^2 = 14.037$ is above the critical value $\chi^2_{4,\,0.05} = 7.815$ and, hence, we reject the null hypothesis at the 5% level.

$\Rightarrow$ The observed data does not match what we would expect if the theoretical distribution was true.

Conclusion: The random numbers are not being generated according to the intended distribution, i.e., there must be a mistake in our code.

Example: Poisson Road Accidents

A busy section of road was monitored over a one-year period. Each day the number of accidents was recorded and the results were:

x_i	0	1	2	3	4+	$\sum$
o_i	187	71	73	24	10	365

and the average number of accidents in the sample was $\bar{x} = 0.904$.

We wish to test the hypothesis that the above results are in line with a Poisson distribution at the 1% level:

H_0 : Poisson model *fits* the data

H_a : Poisson model *does not fit* the data

Example: Poisson Road Accidents

Since $E(X) = \lambda$ for the Poisson distribution, we can use $\bar{x} = 0.904$ as the *estimated* λ value here.

The question is as follows:

Does the data appear to follow a Poisson($\lambda = 0.904$) distribution?

We can work out the theoretical probabilities using:

$$\Pr(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{0.904^x}{x!} e^{-0.904}.$$

Example: Poisson Road Accidents

theoretical probability: p_i $\Rightarrow$ total $\times\, p_i = e_i$

$$\Pr(X = 0) = \frac{0.904^0}{0\,!} e^{-0.904} = 0.405 \quad \Rightarrow \quad 365(0.405) = 147.83$$

$$\Pr(X = 1) = \frac{0.904^1}{1\,!} e^{-0.904} = 0.366 \quad \Rightarrow \quad 365(0.366) = 133.59$$

$$\Pr(X = 2) = \frac{0.904^2}{2\,!} e^{-0.904} = 0.165 \quad \Rightarrow \quad 365(0.165) = 60.23$$

$$\Pr(X = 3) = \frac{0.904^3}{3\,!} e^{-0.904} = 0.050 \quad \Rightarrow \quad 365(0.050) = 18.25$$

$$\Pr(X \geq 4) = 1 - \Pr(X < 4) = 0.014 \quad \Rightarrow \quad 365(0.014) = 5.11$$

Example: Poisson Road Accidents

x_i	*0*	*1*	*2*	*3*	*4+*	$\sum$
o_i	187	71	73	24	10	365
e_i	147.83	133.59	60.23	18.25	5.11	365
$\frac{(o_i - e_i)^2}{e_i}$	10.38	29.32	2.71	1.81	4.68	48.9

There are 5 frequencies, $\{0, 1, 2, 3, 4+\}$, and we estimated one parameter, $\lambda = 0.904$. Thus, $\nu = n_f - 1 - k = 5 - 1 - 1 = 3$.

Thus, the critical value is $\chi^2_{3,\,0.01} = 11.345$ (for the 1% level).

Since $48.9 > 11.345$, we reject $H_0 \Rightarrow$ the Poisson model is not appropriate for this data.

Example: Exponential Battery Life

A sample of 100 laptops was selected. Each one was fully charged and the time (in hours) until the laptop reached a critical battery level was recorded.

The data was stored as a frequency table:

(recall how to construct frequency tables from Lecture1)

Class	<3	3 - 6	6 - 9	9 - 12	12 - 15	>15	$\sum$
Frequency	56	24	12	6	1	1	100

The average battery life for the sample was found to be 3.485 hours.

Example: Exponential Battery Life

Let's assume that we want to test the hypothesis that an exponential model can be used for this data.

For the exponential distribution $E(T) = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{E(T)}$. Thus, we can use 0.287 (i.e., $\frac{1}{3.485}$) as the *estimated* λ value.

Using $\Pr(T > t) = e^{-\lambda t} = e^{-0.287 t}$, it is easy to calculate the following:

$$\Pr(T < 3) = 0.577 \qquad \Pr(9 < T < 12) = 0.044$$

$$\Pr(3 < T < 6) = 0.244 \qquad \Pr(12 < T < 15) = 0.018$$

$$\Pr(6 < T < 9) = 0.103 \qquad \Pr(T > 15) = 0.014$$

Question 1

The information from the previous slides is as follows:

Class	<3	3 - 6	6 - 9	9 - 12	12 - 15	>15	$\sum$
o_i	56	24	12	6	1	1	100
p_i	0.577	0.244	0.103	0.044	0.018	0.014	

a) State the null and alternative hypotheses.

b) Calculate the expected frequencies, e_i.

c) If necessary, combine classes if any e_i values are less than 5.

d) What is the critical value for the chi-squared test if we are testing at the 5% level?

e) Is the exponential model appropriate for this data?

Example: Defective Units

Often we wish to explore the relationship between two categorical variables in the form of a **contingency table**:

	Factory 1	Factory 2	Factory 3	$\sum$
Defective Units	8	21	14	43
Non-Defective Units	73	66	85	224
$\sum$	81	87	99	267

Note that $\Pr(D \mid F_1) = \frac{8}{81} = 0.099$

$\Pr(D \mid F_2) = \frac{21}{87} = 0.241$

$\Pr(D \mid F_3) = \frac{14}{99} = 0.141$

Example: Defective Units

The two variables of interest are “Factory” and “Defective Status”.

Based on the conditional probabilities on the previous slide, it appears that Defective Status varies with Factory, i.e., they are *dependent*

However, these results could have come about by *random chance* when in fact the variables are independent.

We wish to test the hypothesis that the variables are independent (i.e., no relationship between them).

Independence Test

In order to test whether two variables are independent, we use the chi-squared test to check if an **independence model** fits the data.

Thus we have:

H_0 : independence model fits the data

H_a : independence does not fit the data

It is more common to write these hypotheses as:

H_0 :	the variables are *independent*
H_a :	the variables are *dependent*

Marginal Probabilities

Note that the **marginal probabilities** are:

rows

$$\Pr(D) = \frac{43}{267} = 0.161$$

$$\Pr(D^c) = \frac{224}{267} = 0.839$$

columns

$$\Pr(F_1) = \frac{81}{267} = 0.303$$

$$\Pr(F_2) = \frac{87}{267} = 0.326$$

$$\Pr(F_3) = \frac{99}{267} = 0.371$$

These are the marginal distributions of Defective Status and Factory.

(so-called because they correspond to the margins of the contingency table)

Independence Model

If variables are independent, multiplying marginal probabilities gives the joint probabilities (see Lecture6).

If this is the case then the joint distribution for our data would be:

	Factory 1	Factory 2	Factory 3	
Defective Units	0.049	0.052	0.060	0.161
Non-Defective Units	0.254	0.274	0.311	0.839
	0.303	0.326	0.371	1.000

(note, for example, that $0.161 \times 0.303 = 0.049$)

This is the independence model for our data. We can use the chi-squared test to check if it fits.

Expected Frequencies

As before, $e_i = \text{total} \times p_i = 267 \times p_i$.

$\Rightarrow$ The expected frequency for $D \cap F_1$ is $267(0.049) = 13.083$.

Note that:

$$267(0.049) = 267(0.161)(0.303) = 267 \ \frac{43}{267} \ \frac{81}{267}$$

$$= \frac{43(81)}{267} = 13.045.$$

(the answers differ slightly due to rounding error)

Expected Frequencies

Clearly $\frac{43 \times 81}{267} = \frac{\text{row total} \times \text{column total}}{\text{total}}$.

(it is easy to confirm that this pattern continues for the rest of the table)

$\Rightarrow$ To calculate expected frequencies in contingency tables use:

$$\boxed{e_{ij} = \frac{r_i \times c_j}{\text{total}}}$$

where r_i and c_j are the relevant row and column totals.

In other words, we don't have to calculate the independence model probabilities as was done on slide 33.

Degrees of Freedom

The degrees of freedom are $\nu = n_f - 1 - k$. We have used the data to estimate probabilities for the independence model $\Rightarrow k \neq 0$.

How many probabilities have we estimated?

Answer: *one* row probability, and *two* column probabilities.

The remaining probabilities are given by subtraction:

Rows: $\{0.161\}$ and $\{0.839 = 1 - 0.161\}$.

Columns: $\{0.303\}$, $\{0.326\}$ and $\{0.371 = 1 - (0.303 + 0.326)\}$.

Degrees of Freedom

The number of estimates is $k = (n_r - 1) + (n_c - 1)$.

(i.e., one less than the number of rows $+$ one less than the number of columns)

Thus, the degrees of freedom are $\nu = n_f - 1 - [(n_r - 1) + (n_c - 1)]$.

Since, the number of observed frequencies is $n_f = n_r \times n_c$, it is straightforward to show that the above can be written as:

$$\boxed{\nu = (n_r - 1) \times (n_c - 1)}$$

Example: Defective Units

Observed	Factory 1	Factory 2	Factory 3	$\sum$
Defective	8	21	14	43
Non-Defective	73	66	85	224
$\sum$	81	87	99	267

Expected	Factory 1	Factory 2	Factory 3	$\sum$
Defective	$\frac{43(81)}{267} = 13.04$	$\frac{43(87)}{267} = 14.01$	$\frac{43(99)}{267} = 15.94$	43
Non-Defective	$\frac{224(81)}{267} = 67.96$	$\frac{224(87)}{267} = 72.99$	$\frac{224(99)}{267} = 83.07$	224
$\sum$	81	87	99	267

Example: Defective Units

$\frac{(o_i - e_i)^2}{e_i}$	Factory 1	Factory 2	Factory 3
Defective	1.95	3.49	0.24
Non-Defective	0.37	0.67	0.05

Adding these numbers gives: $\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = 6.76$.

Degrees of freedom: $\nu = (n_r - 1) \times (n_c - 1) = (2 - 1)(3 - 1) = 2$
$\Rightarrow$ critical value is $\chi^2_{2,\,0.05} = 5.991$ (for the 5% level).

Since $6.76 > 5.991$ we reject the null hypothesis that there is no relationship between defects and factory.

Conclusion: The proportion of defects varies across factories.

Example: Defective Units

We must now describe the nature of the dependence.

Recall that we calculated:

$$\Pr(D \mid F_1) = 0.099, \quad \Pr(D \mid F_2) = 0.241, \quad \Pr(D \mid F_3) = 0.141.$$

$\Rightarrow$ Factory 1 appears to be the best (lowest defective rate), followed by Factory 3 and then Factory 2.

Example: Defective Units

We can also describe the nature of the dependence by inspecting the values of the **raw difference scores**, $o_i - e_i$:

$o_i - e_i$	Factory 1	Factory 2	Factory 3
Defective	-5.04	6.99	-1.94
Non-Defective	5.04	-6.99	1.94

Factory 1 has less defects than we would expect - so does Factory 3 (but the difference is not as large).

Factory 2 has more defects than we expect.

Question 1

The different grades given by 3 different examiners marking an exam were as follows:

	Examiner 1	Examiner 2	Examiner 3	$\sum$
Grade A	9	6	10	25
Grade B	11	20	26	57
Grade C	25	19	74	118
$\sum$	45	45	110	200

a) Show that there is a significant relationship between the examiner and the grade awarded (at the 5% level).

b) Using the raw difference scores, comment on nature of the dependence.

R Code: Goodness-of-fit

To run the chi-squared test in R, the `chisq.test` function is used:

```
oi = c(10, 11, 42, 17)
pi = c(0.2,0.2,0.5,0.1)
chisq.test(oi, p=pi)
```

Compare this with slide 14.

Note: since we have combined the frequencies for $X = 6$ and $X = 8$, we also combine the theoretical probabilities $0.05 + 0.05 = 0.1$.

R Code: Goodness-of-fit (Estimated Parameters)

For the road accident data we have:

```
oi = c(187, 71, 73, 24, 10)
pi = c(0.405,0.366,0.165,0.050,0.014)
chisq.test(oi, p=pi)
```

Compare this with slide 19.

Note: R has no way of knowing that we estimated a λ parameter to calculate the theoretical probabilities $\Rightarrow$ the degrees of freedom are **wrong** and, hence, the p-value is wrong.

R Code: Goodness-of-fit (Estimated Parameters)

The following code can be used to calculate the correct p-value in the case where parameters have been estimated:

```
oi = c(187, 71, 73, 24, 10)
pi = c(0.405,0.366,0.165,0.050,0.014)
chi = chisq.test(oi, p=pi)
k = 1    # the number of estimated parameters
pchisq(chi$stat, df=(chi$par-k), lower=F)[[1]]
```

Note: In this particular case, adjusting the degrees of freedom doesn't alter our conclusion (the p-value is tiny either way).

R Code: Independence Test

For the independence test, we provide the contingency table as a `matrix`.

We do not provide theoretical probabilities in this case; they are automatically calculated.

```
oi = matrix(c(8,21,14,73,66,85), nrow=2,
                                 ncol=3, byrow=T)

oi    # look at the structure of oi here

chisq.test(oi)
```

Compare this with slide 33.