Chi-squared Tests

1. Goodness of fit test – General Procedure
2. Goodness of fit test – Fit a uniform distribution
3. Goodness of fit test – Fit a distribution with a given ratio
4. Goodness of fit test – Fit a binomial distribution (p known)
5. Goodness of fit test – Fit a binomial distribution (p unknown)
6. Goodness of fit test – Fit a Poisson distribution (λ known)
7. Goodness of fit test – Fit a Poisson distribution (λ unknown)

Chi-squared Tests

8. Goodness of fit test – Fit a Geometric distribution (p known)
9. Goodness of fit test – Fit a Geometric distribution (p unknown)
10. Goodness of fit test – Fit a Normal distribution (both μ, σ are known)
11. Goodness of fit test – Fit a Normal distribution (μ and/or σ are unknown)
12. Goodness of fit test – Fit an Exponential distribution (λ known)
13. Goodness of fit test – Fit a Exponential distribution (λ unknown)
14. Independence Test – Part I
15. Independence Test – Part II (Further examples)

Chi-squared Tests

The χ^2 significance test

χ^2 significance test is used for:

1 A χ^2 goodness-of-fit test

To know how well a particular statistical distribution models that data.

H_o : ___ distribution does provide a model for the data.

H_1 : ___ distribution does not provide a model for the data.

The χ^2 significance test

χ^2 significance test is used for:

2 A χ^2 test for independence

To know whether if two variables are independent.

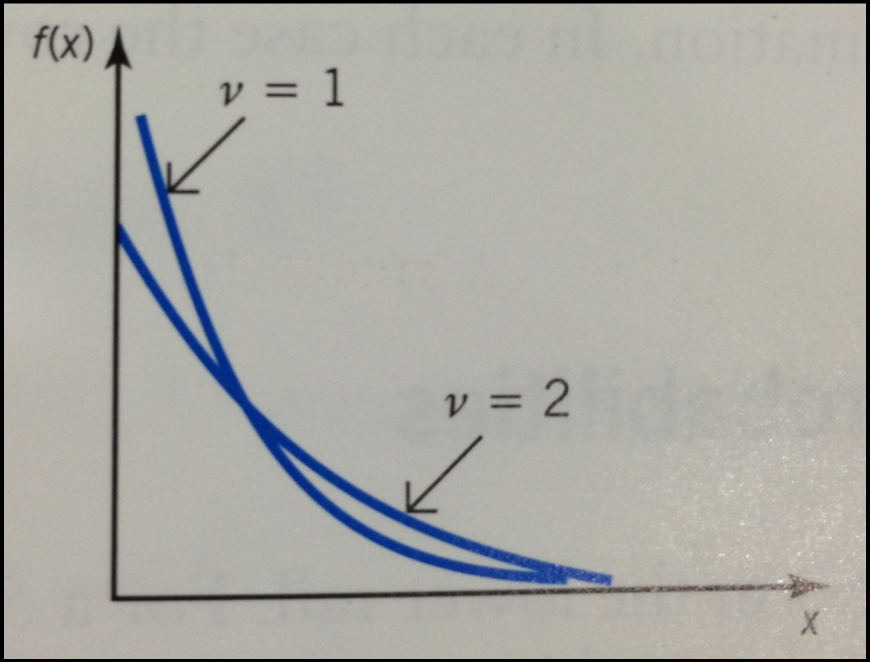
H_o : The factors are independent.

H_1 : The factors are not independent.

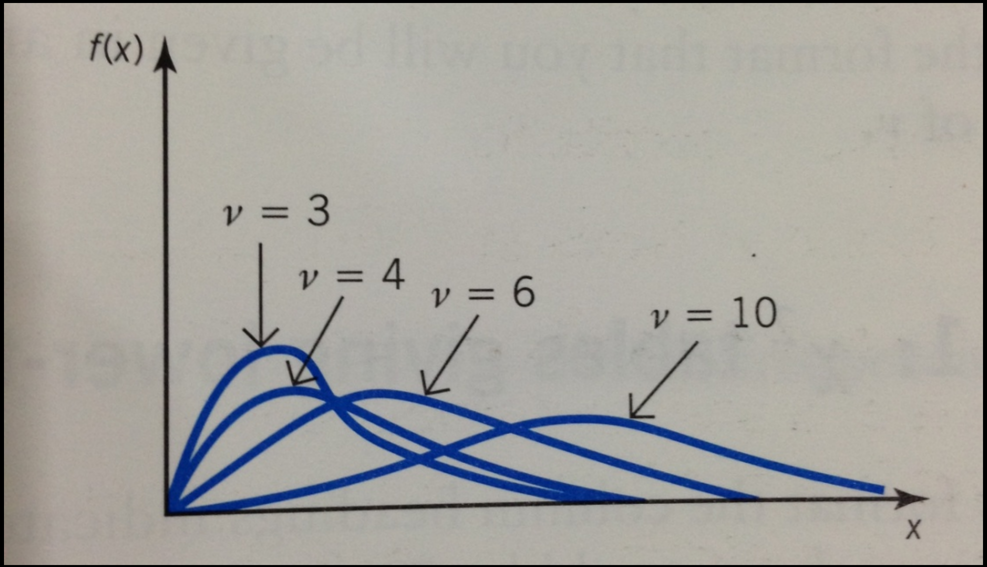
The χ^2 distribution

- Has one parameter, *v*, number of degrees of freedom.
- Mean = v ; variance = $2v$.
- χ^2_v , χ^2 with v degrees of freedom.
- As v increases, the spread of the distribution increases.
- It is reversed J-shaped for $v = 1$ and $v = 2$.
- It is positively skewed for $v > 2$.
- When v is large, the distribution is approximately normal.

The χ^2 distribution



When ν is large, the distribution is approximately normal.



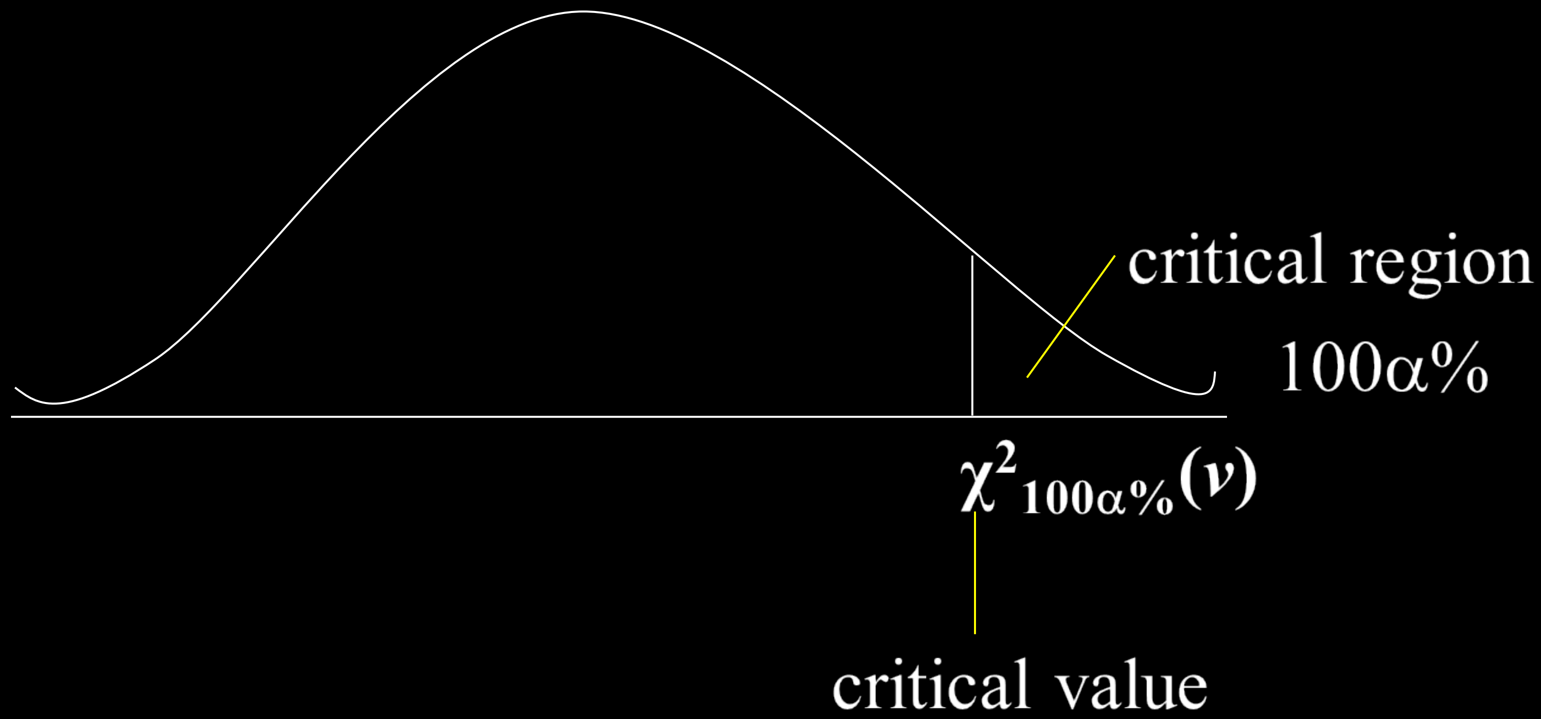
Degrees of freedom, *v*

Number of degreed of freedom

= number of classes - number of constraints/restrictions

Critical values and levels of significance

- χ^2 test is conducted as a one-tailed (upper tail) test.



- If χ^2 lies in the critical region $\Rightarrow$ reject H_o.

The test statistic χ^2

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

O_i=observed frequency
E_i=expected frequency

- χ^2 distribution can be used as an approximation for the distribution of χ^2 (If none of the $E_i < 5$)

$$\chi^2 \sim \chi^2(v)$$

- The smaller the χ^2, the better the fit.

- If $E_i < 5$ for any class, combine adjacent classes to form a class that is sufficiently large.

The χ^2 Goodness-of-fit test

- Random numbers, x, consist of lists of ten digits.

  0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

  $\Rightarrow P(X = x) = 0.1$

  This is the discrete uniform distribution.

Use calculator to generate 100 digits, and arrange the digits in a frequency table:

Digit	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
frequency	***4***	***10***	***7***	***16***	***12***	***8***	***10***	***11***	***12***	***10***	***Total 100***

The χ^2 Goodness-of-fit test

- A χ^2 goodness fir test is used to test whether the numbers generated on the calculator are random enough.
- H_0: the digits are random

  H_1: the digits are not random

- If the digits are random,

  $\Rightarrow$Expected frequency for each digit is 100 x 0.1 = 10

Digit	0	1	2	3	4	5	6	7	8	9	
Frequency (O)	*4*	*10*	*7*	*16*	*12*	*8*	*10*	*11*	*12*	*10*	*Total 100*
Expected frequency (E)	*10*	*10*	*10*	*10*	*10*	*10*	*10*	*10*	*10*	*10*	*Total 100*

The χ^2 Goodness-of-fit test

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- **If** $\chi^2 = 0$ (exact agreement between the observed & expected frequencies)
- χ^2 low (good fit)
- χ^2 high (poor fit)

$$\chi^2 = \frac{(4-10)^2}{10} + \frac{(10-10)^2}{10} + \cdots + \frac{(10-10)^2}{10} = 9.4$$

The χ^2 Goodness-of-fit test

v = number of classes – number of restrictions

= 10 – 1

one restriction $\sum f = 100$

Let $\alpha = 0.05$

$$\chi^2_{5\%}(9) = 16.919$$

Reject H_o if $\chi^2 > 16.919$

$\Rightarrow$ Fail to reject H_o.

There is sufficient evidence at 5% significance level to conclude that the digits are random digits.

Test 1 :
Goodness-of-fit test for a uniform distribution

The table shows the number of employees absent for just one day during a particular period of time.

Day of the week	**Mon**	**Tues**	**Wed**	**Thurs**	**Fri**	
Number of absentees	***121***	***87***	***87***	***91***	***114***	***Total 500***

(a) Find the frequencies expected according to the hypothesis that the number of absentees is independent of the day of the week.

(b) Test at the 5% level whether the difference in the observed and expected data are significant.

Test 2 : Goodness-of-fit test for a distribution in a given ratio

In genetic work it is predicted that the children with both parents of blood group AB will fall into blood groups AB, A and B in the ratio 2:1:1. Of a random sample of 100 such children 55 were blood group AB, 27 blood group A and 18 blood group B. Test at the 10% significance level whether the observed results agree with the theoretical prediction.

<u>Test 3(i): Goodness-of-fit test for a Binomial distribution, *p* known.</u>

A game contains 20 pieces, each of which has probability 0.08 of being defective.

(a) Suggest a suitable distribution to model the number of defective pieces in a game.

Let X represent the number of defective pieces in a game.

(b) Copy and complete the following probability distribution.

x	0	1	2	3	4	5	6 or more
$P(X = x)$	*0.1887*			*0.1414*	*0.0523*	*0.0145*	

Test 3(i): Goodness-of-fit test for a Binomial distribution, *p* known.

A random sample of 1000 games was checked for defective pieces and the following table produced to summarise the number of defective pieces in each game.

Number of defective pieces	**0**	**1**	**2**	**3**	**4**	**5**	**6 or more**
Number of games	*194*	*344*	*266*	*137*	*46*	*10*	*3*

(c) Use a χ^2 test to test, at the 5% level, whether or not the observed results are consistent with those exoected under the model specified in (a).

Test 3(ii): Goodness-of-fit test for a Binomial distribution, *p* unknown.

Samples of size 5 are selected regularly from a production line and tested. During one week 500 samples are taken and the number of defective items in each sample is recorded.

Number of defectives, *x*	**0**	**1**	**2**	**3**	**4**	**5**
Frequency, f	*170*	*180*	*120*	*20*	*8*	*2*

(a) Find the frequencies of the number of defectives per sample given by the binomial distribution having the same mean and total as the observed distribution.

(b) Test whether the observed distribution follows a binomial pattern.

Test 4:
Goodness-of-fit test for a Geometric distribution

Below is the start of a list of digits obtained by using the random number generator on a calculator. If such a list is random, then each digit is equally likely to appear in any position in the list. As a result odd and even digits are also equally likely to appear in any position on the list.

6 3 3 0 4 5 3 2 1 4 1 3 2 1 2 7 2 2 5 1 1 1 4 1 2 5 3….

The list is repeated below with a line placed after each odd digit. This has the effect of breaking the sequence up into 'run', consisting of even digits, until the last digit, which is odd. The length of the runs below are 2, 1, 3, 1, 2, 2, etc.

6 3 | 3 | 0 4 5 3 | 2 1 | 4 1 | 3 | 2 1 | 2 7 | 2 2 5 | 1 | 1 | 1 | 4 1 | 2 5 | 3 ….

Test 4: Goodness-of-fit test for a Geometric distribution

If the digits are random, then the length, *Y*, of a run will have a geometric distribution. Considering even number as a 'failure' and an odd number as a 'success'. In this case,

p = q = 0.5, Y~Geo(0.5). Table below gives the frequency distribution of run length for 50 runs.

Length of run	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**>7**
Frequency	***21***	***18***	***5***	***2***	***3***	***0***	***1***	***0***

Carry out a χ^2 goodness of fit test at the 5% significance level to determine whether or not the above distribution can be reasonably modelled by the Geometric (0.5).

Test 5(i): Goodness-of-fit test for a Poisson distribution, λ known.

An analysis if the number of goals scored by the local football team gave the following results:

Goals per match (x)	0	1	2	3	4	5	6	>7	
Number of matches	*14*	*18*	*29*	*18*	*10*	*7*	*3*	*1*	*Total 100*

Carry out a χ^2 goodness of fit test at the 10% significance level to determine whether or not the above distribution can be reasonably modelled by a Poisson distribution with parameter 2.

Test 5(ii): Goodness-of-fit test for a Poisson distribution, λ unknown.

Can the data of Test 4(i) be modelled by a Poisson distribution having the same as the observed data?

Test at the 10% level.

Test 6(i): Goodness-of-fit test for a Normal distribution, Mean and variance known.

For a period of six months 100 similar hamsters were given a new type of feedstuff. The gains in mass are recorded in the table below:

Gain in mass (g), x	**Observed frequency, f**
$-\infty < x \le -10$	3
$-10 < x \le -5$	6
$-5 < x \le 0$	9
$0 < x \le 5$	15
$5 < x \le 10$	24
$10 < x \le 15$	16
$15 < x \le 20$	14
$20 < x \le 25$	8
$25 < x \le 30$	3
$30 < x \le 35$	2

Test 6(i): Goodness-of-fit test for a Normal distribution, Mean and variance known.

It is thought that these data follow a normal distribution, with mean 10 and variance 100. Use the χ^2 distribution at the 5% level of significance to test this hypothesis.

Describe briefly how you would modify this test if the mean and variance were known.

Test 6(ii):
Goodness-of-fit test for a Normal distribution, Mean and variance unknown.

The lengths (in mm) of 50 leaves that had fallen from an oak tree were measured. The results are summarised as below.

Length (mm)	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99	100 – 109
Frequency	*3*	*9*	*15*	*9*	*6*	*4*	*3*	*1*

By carrying out a goodness of fit test at the 5% significance level test whether the length of an oak leaf can be modelled by a normal distribution.

Test 7:
Goodness-of-fit test for a Exponential distribution

The table below shows the time intervals, in seconds, between successive white cars in free flowing traffic on an open road. Can these times be modelled by a exponential distribution?

Time	0 -	20 -	40 -	60 -	90 -	120 - 180
Frequency	*41*	*19*	*16*	*13*	*9*	*2*

The χ^2 significance Test for independence

Sometimes when data are classified to two factors, displayed in a table, known as contingency table, we can use a χ^2 test to investigate whether the two factors are independent or whether there is an association between them.

H_0: Two factors are independent

H_1: There is an association between them

Example 1:

Two schools enter their pupils for a particular public examination and the results obtained are shown below.

	Credit	Pass	Fail
School A	*51*	*10*	*19*
School B	*39*	*10*	*21*

By using an approximate χ^2 statistics, asses at the 5% level whether or not there is a significant difference between the two schools with respect to the proportions of pupils in the three grades. State your null and alternative hypothesis.

Finding the number of degrees of freedom, *v*, in an *h* by k contingency table

$$v = (h - 1)(k - 1)$$

Example 2:

A university sociology department believes that students with a good grade in A level General Studies tend to do well on sociology degree courses. To check this it collected information on a random sample of 100 students who had just graduated and had also taken General Studies at A level. The student's performance in General Studies was divided into two categories, those with grade A or B and 'other'. Their degree classes were recorded as Class I, Class II, Class III and Fail. The data are given in the table below.

Example 2:

	Class I	Class II	Class III	Fail	Total
Grade A	*11*	*22*	*6*	*1*	*40*
Others	*4*	*28*	*24*	*4*	*60*
Total	*15*	*50*	*30*	*5*	*100*

Use these data to test, at the 1% significance level, the hypothesis that degree class is independent of General Studies A level performance. State your conclusion clearly.