Chi-Squared Tests Cheat Sheet

Goodness of fit tests are used to determine how well an observed frequency distribution fits a known distribution. For example, you might have a set of data that you suspect follows a Poisson distribution. To figure out whether the data is suitable to be modelled by a Poisson distribution, you can conduct a goodness of fit test. We will learn to carry out goodness of fit tests for discrete uniform, Poisson, binomial, and geometric distributions in addition to contingency tables.

The idea behind goodness of fit tests

Let's say that you have a dice, and you are wondering whether it is fair or not. You might decide to throw the dice 240 times and record the results. If the dice is fair, then in theory, the results should look like this:

Number on dice, n	1	2	3	4	5	6
Expected frequency	40	40	40	40	40	40

The expected results follow a discrete uniform distribution.

However, since you have taken a sample, you likely won't roll each side of the dice exactly 40 times. In practice, your results might look something like this:

Number on dice, n	1	2	3	4	5	6
Observed frequency	36	41	47	35	38	44

So, even if the dice was in fact unbiased, the scores won't necessarily exactly match the theoretical expected values. But how can we decide whether our results are 'good enough' to conclude that the dice is fair (i.e. the observed frequencies do indeed follow a discrete uniform distribution)? Well, we can set up a hypothesis test.

We take the null hypothesis to be that there is no difference between the theoretical frequency distribution and the observed frequency distribution, and that any differences can be put down to natural variations. This would mean that the dice is unbiased (i.e. the score on the dice follows a discrete uniform distribution)

We take the alternative hypothesis to be that there is in fact a difference between the theoretical frequency distribution and the observed frequency distribution that cannot be explained by natural variations. This would mean that the dice is in fact biased. (i.e. the score on the dice does not follow a discrete uniform distribution)

In order to test these hypotheses, we need a measure of how closely correlated the expected and theoretical frequencies are. In other words, we need a measure for the goodness of fit between the expected and observed frequencies. This is where the X^2 test statistic comes in:

- $X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E}$ — X^2 is sometimes called the goodness of fit statistic.

If we were to calculate X^2 for the above data, we would have:

Number on dice, n	1	2	3	4	5	6
Observed frequency (O)	36	41	47	35	38	44
Expected frequency (E)	40	40	40	40	40	40
$\frac{(O-E)^2}{E}$	$\frac{(36-40)^2}{40} = 0.4$	0.025	1.225	0.625	0.1	0.4

Each rectangle that contains an observation is known as a cell.

$\Rightarrow X^2 = 0.4 + 0.025 + \cdots + 0.4 = 2.775$

You can see that when there is a small difference between observed and expected frequencies, the value of $\frac{(O-E)^2}{E}$ is small, but when there is a bigger difference between observed and expected frequencies then the value is larger. This should convince you that the statistic X^2 is a suitable measure for the goodness of fit between the observed and expected frequencies.

Chi-squared (χ^2) family of distributions and degrees of freedom

The χ^2 family of distributions are a set of distributions that can be used as approximations for the X^2 test statistic. In other words, X^2 follows a chi-squared distribution. We use the chi-squared distributions to find critical values for our hypothesis test. You will need to calculate the number of degrees of freedom for your data before you can find a critical value. The number of degrees of freedom is what distinguishes the different members of the χ^2 family.

Degrees of freedom are a measure of how much information from the sample data has not been used up. Every time a statistic is calculated from a sample, a degree of freedom is used up.

- $Number\ of\ degrees\ of\ freedom\ (\nu) = number\ of\ columns\ (after\ any\ combining) - number\ of\ constraints$

  For the above data, $\nu = 6 - 1 = 5$

  Pronounced "nu"

  $\Rightarrow$ Any parameter that you calculate relating to the distribution counts as a constraint.

  $\Rightarrow$ We always subtract.at least 1 because the requirement that the total expected and observed frequencies must match counts as a constraint.

  $\Rightarrow$ If there are any expected values less than 5, you will need to combine columns so that there are no expected values less than 5. This is because X^2 is only approximated well by χ^2 when the expected frequencies are all above 5.

- To find critical values for your goodness of fit test, you will need to use the table of values for the percentage points of the χ^2 distribution, which is given in the formula booklet (or p192 in the textbook). Some calculators (e.g. Casio CG-50) will allow you to directly find these critical values, eliminating the need to refer to the table.

- To make clear which member of the χ^2 distribution you are talking about, you write χ^2_ν. For example, χ^2_3 is used to denote the χ^2 distribution with 3 degrees of freedom.

- The critical value of χ^2 that is exceeded with probability 5% is written as $\chi^2_\nu(5\%)$ or $\chi^2_\nu(0.05)$.

If we were carrying out a test at 5% significance level for the above data, our critical value would be $\chi^2_5(5\%) = 11.070$ (from tables).

The goodness of fit tests that you will be expected to carry out

You need to be able to test whether a given set of data can be modelled by a:

- Discrete uniform distribution
- Binomial distribution
- Poisson distribution
- Geometric distribution

Additionally, you need to be able to carry out goodness of fit tests to figure out whether two events are independent or not using data given in a contingency table. The process for such questions is slightly different to the others but is covered in this cheat sheet.

We have already briefly covered how a goodness of fit test is done with a discrete uniform distribution. With the help of examples, we will now look at how it is done with other distributions.

The step-by-step procedure

Regardless of which distribution you are testing for, the procedure for carrying out a goodness of fit test is largely the same:

1) Form your null and alternative hypotheses. The null hypothesis is always that the given distribution is a good model for your data while the alternative hypothesis should be that it is not a good model.

2) Work out the expected frequencies. To do this, you will need to consider the distribution you are testing for. For example, if you are testing for a binomial distribution then you will need to use the appropriate binomial distribution to figure out the expected frequencies.

3) Create a table that contains the expected frequencies and observed frequencies. Check to see if any expected frequencies are less than 5. If so, you will need to pool adjacent cells together so that there are no expected frequencies less than 5.

4) Calculate $X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E}$. Remember that this should be done **after** any columns are pooled together.

5) Calculate the number of degrees of freedom for your data using the formula $\nu = no.of\ cells - no.of\ constraints - 1$ and find the critical value using the Chi-squared percentage points table in the formula booklet. You will need the number of degrees of freedom and the significance level, which will be given to you in the question.

6) If X^2 exceeds the critical value, then you should reject the null hypothesis. Otherwise accept the null hypothesis. Remember to give your conclusion in the context of the question. Refer to the next examples for more detail.

Testing the binomial distribution as a model

Example 1: A research station is doing some work on the germination of a new variety of genetically modified wheat. They planted 120 rows containing 7 seeds in each row. The number of seeds germinating in each row was recorded. The results are as follows:

No. seeds germinating in each row	0	1	2	3	4	5	6	7
Observed number of rows	2	6	11	19	25	32	16	9

a) Write down two reasons why a binomial distribution may be a suitable model.

This question is really just asking you to state two assumptions required for the binomial distribution to be used, that are satisfied.	-Each seed either germinates or doesn't (each trial is either a success or failure). -The germination of each seed is independent of the others (independency).

b) Show that the probability of a randomly selected seed from this sample germinating is 0.6.

probability$=\frac{\text{total germinations}}{\text{total seeds planted}}$	$p = \frac{0(2) + 1(6) + 2(11) + 3(19) + 4(25) + 5(32) + 6(16) + 7(9)}{120 \times 7} = \frac{504}{840} = 0.6$

The research station used a binomial distribution with probability 0.6 of a seed germinating. The expected frequencies were calculated to 2 decimal places. The results are as follows:

No. seeds germinating in each row	0	1	2	3	4	5	6	7
Expected number of rows	0.20	2.06	s	23.22	t	31.35	15.68	3.36

c) Find the value of s and the value of t.

To find the expected values, we need to use the suggested binomial distribution. The probabilities can be found using a calculator.	Let $X = the\ number\ of\ seeds\ germinating\ in\ one\ row$. Then $X \sim B(7, 0.6)$ $P(X = 2) = 0.0774$. This is the probability that there are 2 germinated seeds in a given row. So, for 120 rows: the expected number of rows for which there are 2 germinated seeds is $120 \times 0.0774 = 9.29$ $\therefore s = 9.29$.
To find t, use the fact that the total expected frequency must be equal to the total observed frequency (120).	$t = 120 - (0.20 + 2.06 + 9.29 + 23.22 + 31.35 + 15.68 + 3.36) = 34.84$

d) Stating your hypotheses clearly, test at the 1% level of significance, whether or not the data can be modelled by a binomial distribution.

Start by stating your hypotheses.	H_0: A binomial distribution is a suitable model for these data. H_0: A binomial distribution is not a suitable model for these data.
Create a table showing the expected and observed values and use it to find X^2. You will need to pool the first three cells and the last two cells, so that all expected frequencies are above 5.	No. seeds germinated: 0-2 | 3 | 4 | 5 | 6-7 Observed number of rows: 19 | 19 | 25 | 32 | 25 Expected number of rows: 11.55 | 23.22 | 34.84 | 31.35 | 19.04 $\frac{(O-E)^2}{E}$: 4.805 | 0.767 | 2.779 | 0.013 | 1.866 $\therefore X^2 = 4.805 + 0.767 + \cdots + 1.866 = 10.23$
Find degrees of freedom.	There are 5 cells after pooling. We also calculated p in part (a) so that counts as a constraint. $\therefore \nu = 5 - 1 - 1 = 3$
State the critical value.	$\nu = 3, significance\ level = 1\%$ $critical\ value = \chi^2_3(1\%) = 11.345$
Compare X^2 with the critical value and write a conclusion in context.	$10.23 < 11.345$ ∴ result is insignifcant, accept H_0. Evidence suggests that a binomial distribution is a suitable model for the given data.

Testing the Poisson distribution as a model

Example 2: The number of accidents on a particular stretch of motorway was recorded each day for 200 consecutive days. The results are summarised in the following table:

No. accidents	0	1	2	3	4	5
Frequency	47	57	46	35	9	6

a) Show that the mean number of accidents per day for these data is 1.6.

mean $=\frac{\text{total accidents}}{\text{total days}}$	$mean = \frac{0(47) + 1(57) + 2(46) + 3(35) + 4(9) + 5(6)}{200} = \frac{320}{200} = 1.6$

A motorway supervisor believes that the number of accidents per day on this stretch of motorway can be modelled by a Poisson distribution. She uses the mean found in part (a) to calculate the expected frequencies for this model. Some of her results are given in the following table.

No. accidents	0	1	2	3	4	5 or more
Expected frequency	40.38	64.61	r	s	11.03	t

b) Complete the table, giving your answers to 2 decimal places.

To find the expected values, we need to use the suggested Poisson distribution. The number of accidents follow the distribution Po(1.6). The probabilities can be found using a calculator.	Let $X = the\ number\ of\ accidents\ in\ one\ day$. Then $X \sim Po(1.6)$ $P(X = 2) = 0.2584$. This is the probability that there are 2 accidents in a given day. $P(X = 3) = 0.1378$. This is the probability that there are 3 accidents in a given day. So, for 200 days: the expected number of days for which there are 2 accidents is $200 \times 0.2584 = 51.69$ the expected number of days for which there are 3 accidents is $200 \times 0.1378 = 27.57$ $\therefore r = 51.69$ and $s = 27.57$.
To find t, use the fact that the total expected frequency must be equal to the total observed frequency (200).	$t = 200 - (40.38 + 64.61 + 51.69 + 27.57 + 11.03) = 4.72$.

c) Stating your hypotheses clearly, use a 10% level of significance to test the motorway supervisor's belief.

Start by stating your hypotheses.	H_0: Po(1.6) is a suitable model for these data. H_0: Po(1.6) is not a suitable model for these data.
Create a table showing the expected and observed values and use it to find X^2. Remember that the last cell had an expected frequency less than 5 so we combine the final two cells together before calculating X^2.	No. accidents: 0 | 1 | 2 | 3 | 4 or more Observed frequency: 47 | 57 | 46 | 35 | 9+6=15 Expected frequency: 40.38 | 64.61 | 51.69 | 27.57 | 11.03+4.72=15.75 $\frac{(O-E)^2}{E}$: 0.164 | 0.896 | 0.626 | 2.002 | 0.035 $\therefore X^2 = 0.164 + 0.896 + \cdots + 0.035 = 3.72$
Find degrees of freedom.	There are 5 cells. We also calculated the mean in part (a) so that counts as a constraint. $\therefore \nu = 5 - 1 - 1 = 3$
State the critical value.	$\nu = 3, significance\ level = 10\%$ $critical\ value = \chi^2_3(10\%) = 6.251$
Compare X^2 with the critical value and write a conclusion in context.	$3.72 < 6.251$ ∴ result is insignifcant, accept H_0. Evidence suggests that a poisson distribution with mean 1.6 is a suitable model for the number of accidents per day on the motorway.

Using contingency tables

Another type of question you can be asked is one where you are given a contingency table (sometimes called a two-way table) and you need to conduct a goodness of fit test to figure out whether two variables are independent. The process for such questions is slightly different.

1) Form your null and alternative hypotheses. The null hypothesis is always that the two variables are independent (no association). The alternative hypothesis is always that the two variables are not independent (there is an association).

2) Work out the expected frequency for each cell using the formula $expected\ frequency = \frac{row\ total \times column\ total}{grand\ total}$.

3) Find $X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E}$.

4) Work out the degrees of freedom using the formula $\nu = (rows - 1)(columns - 1)$.

5) Find the critical value using ν and the significance level. Compare this to your value for X^2 and decide whether to accept or reject the null hypothesis.

Example 3: A researcher investigates the results of candidates who took their driving test at one of three driving test centres. A random sample of 620 candidates gave the following results

		A	B	C	Total
Results	Pass	99	110	68	277
	Fail	108	116	119	343
Total		207	226	187	620

Test, at the 5% level of significance, whether there is an association between the results of candidates' driving tests and the driving test centre. State your hypotheses and show your working clearly. You should state your expected frequencies correct to 2 decimal places.

Start by stating your hypotheses.	H_0: There is no association between the driving test results and the driving test centre (they are independent). H_1: There is an association between the driving test results and the driving test centre (they are not independent)
Find the expected frequencies for each cell of the table using the formula $expected\ frequency = \frac{row\ total \times column\ total}{grand\ total}$	For example, expected frequencies for Pass/A will be $\frac{277 \times 207}{620} = 92.48$. Doing this for all cells gives us: EXPECTED FREQUENCIES | A | B | C | Total Results – Pass: 92.48 | 100.97 | 83.55 | 277 Fail: 114.52 | 125.03 | 103.45 | 343 Total: 207 | 226 | 187 | 620
Now calculate $X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E}$.	$X^2 = \frac{(99 - 92.48)^2}{92.48} + \frac{(110 - 100.97)^2}{100.97} + \cdots + \frac{(119 - 103.45)^2}{103.45} = 7.52$
Find degrees of freedom.	$\nu = (rows - 1)(columns - 1) = (2 - 1)(3 - 1) = 2$
Find critical value, compare it to our X^2 value and write a conclusion.	$critical\ value = \chi^2_2(5\%) = 5.991$ $7.52 > 5.991$ ∴ result is significant. Evidence suggests that there is in fact an association between the driving test centre and the results of candidates.

Testing the geometric distribution as a model

Example 4: Each day after work Katie flags down a taxi to take her home. She records how many taxis she needs to flag down before one stops, over the course of 100 days.

Attempts	1	2	3	4	5	Total
Frequency	76	17	4	2	1	100

Katie thinks she can model the number of attempts each day using a geometric random variable $X \sim Geo(p)$.

a) Using the observed frequencies, find an estimate for p (to 3 d.p.).

p$=\frac{\text{total successes}}{\text{total attempts}}$	$p = \frac{100}{1(76) + 2(17) + 3(4) + 4(2) + 5(1)} = 0.741$

b) Conduct a goodness of fit test at the 2.5% significance level, and say whether a geometric random variable is a good model for the data.

Start by stating your hypotheses.	H_0: A geometric distribution is a suitable model. H_1: A geometric distribution is not a suitable model.
Next, we work out the expected frequencies. We use the geometric distribution with $p = 0.741$ to do so. The formula $P(X = i) = p(1 - p)^{i-1}$ is used. Remember that the last expected value will be given by $100 - (sum\ of\ the\ other\ expected\ values)$ since you need to make sure the total expected frequency matches up with the total observed frequency (100).	$Expected\ frequency = 100 \times P(X = i)$, since there are 100 days. $\therefore E_1 = 100 \times (0.741)(1 - 0.741)^0 = 74.1$ $E_2 = 100 \times (0.741)(1 - 0.741)^1 = 19.2$ $E_3 = 100 \times (0.741)(1 - 0.741)^2 = 4.98$ $E_4 = 100 \times (0.741)(1 - 0.741)^3 = 1.29$ $E_{\geq 5} = 100 - (74.1 + 19.2 + 4.98 + 1.29) = 0.42$
Now we put the expected and observed values in a table. Notice that there were some expected values less than 5. We need to pool the final three columns to make sure that all expected frequencies are above 5.	Attempts: 1 | 2 | ≥3 | *Total* Observed frequency: 76 | 17 | 7 | 100 Expected frequency: 74.1 | 19.2 | 6.71 | 100
We use our table to find X^2.	Observed frequency: 76 | 17 | 7 Expected frequency: 74.1 | 19.2 | 6.71 $\frac{(O-E)^2}{E}$: 0.049 | 0.252 | 0.013 $\therefore X^2 = 0.049 + 0.252 + 0.013 = 0.313$
Find the number of degrees of freedom. We subtract an extra 1 because we estimated p in part a.	$\nu = 3 - 1 - 1 = 1$
Find critical value, compare it to our X^2 value and write a conclusion.	$critical\ value = \chi^2_1(2.5\%) = 5.024$ $0.313 < 5.024$ ∴ result is significant. Evidence suggests that there is in fact an association between the driving test centre and the results of candidates.