The Chi-Squared Distribution

Every hypothesis test we have met so far has been about a single *parameter*: a mean, a proportion, a correlation coefficient. We now ask a more ambitious question: does a data set fit a whole *distribution*? To answer it we need a way of measuring the discrepancy between data and model, and we need to know how that measure behaves when the model is actually true. The distribution that does this job is the chi-squared distribution.

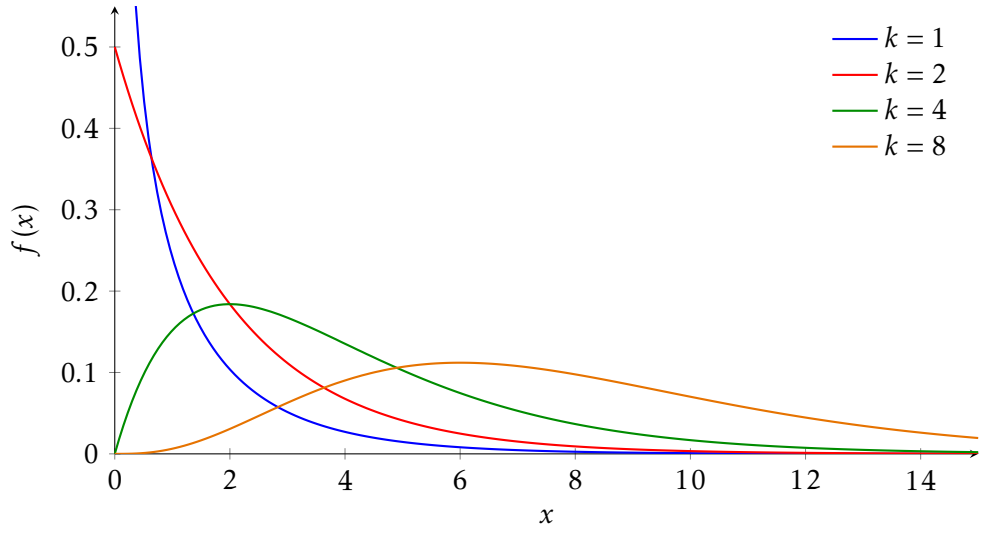
Definition. Let $Z_1, Z_2, \ldots, Z_k$ be *independent* standard normal random variables, so $Z_i \sim N(0,1)$ for each i. Then

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

has the chi-squared distribution with k degrees of freedom. We write $X \sim \chi^2_k$.

> **Remark.** χ is the Greek letter chi, and χ^2 should be read as a single symbol — there is no random variable called χ. Since X is a sum of squares, $X \geq 0$ always.

The shape of the distribution



Some observations, which you should check by simulation (e.g. square and sum columns of random normal values in a spreadsheet, or play with sliders in Geogebra):

- For $k = 1$, $X = Z^2$ and most of the mass is squashed up against 0 (squaring a number between -1 and 1 makes it smaller); the density is unbounded as $x \to 0^+$.
- For $k = 2$ the density is a decreasing exponential, $f(x) = \frac{1}{2}e^{-x/2}$.
- For $k \geq 3$ the curve is humped, with its peak at $x = k - 2$, and a long right tail.
- As k grows the shape becomes increasingly symmetric and bell-like — unsurprising, since χ^2_k is a sum of k independent identically distributed random variables (Z_i^2), so the Central Limit Theorem applies.

Fact — If $X \sim \chi^2_k$ then

$$\mathbb{E}[X] = k \qquad \text{and} \qquad \mathrm{Var}[X] = 2k.$$

Moreover, if $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ are independent, then $X + Y \sim \chi^2_{m+n}$ (a sum of m squares plus a sum of

n squares is a sum of $m+n$ squares).

Example

Given that $\mathbb{E}[Z^4] = 3$ for $Z \sim N(0,1)$, prove the fact above.

Remark (Where $\mathbb{E}[Z^4] = 3$ comes from)**.** The value $\mathbb{E}[Z^4] = 3$ comes from integration by parts on $\int z^4 \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz$. More slickly: χ^2_k is the Gamma distribution with parameters $\alpha = \frac{k}{2}$, $\lambda = \frac{1}{2}$, and the substitution $u = \frac{z^2}{2}$ in the integral $\mathbb{E}[e^{tZ^2}]$ produces the Gamma function (see the Gamma Function notes). The MGF route gives $M_{\chi^2_k}(t) = (1-2t)^{-k/2}$, from which both moments drop out.

Measuring Goodness of Fit

Example

A die is rolled 120 times, with these results:

Score	1	2	3	4	5	6
Frequency	25	17	15	23	24	16

If the die is fair we "expect" 20 of each score. The data is clearly not exactly that — but data never is. Is it *far enough* from $20, 20, \ldots, 20$ to convince us the die is biased?

We need a single number measuring how far the **observed frequencies** O_i are from the **expected frequencies** E_i.

Definition. The **goodness-of-fit statistic** is

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

summed over all the cells. The individual terms $\dfrac{(O_i - E_i)^2}{E_i}$ are called the **contributions** to the test statistic (exam questions often ask for these).

Squaring stops positive and negative discrepancies cancelling; dividing by E_i makes the measure relative (being 10 out matters far more when you expected 15 than when you expected 1500).

Why chi-squared? The two-cell case

Theorem

For two cells, with expected frequencies large enough, X^2 has approximately the χ^2_1 distribution.

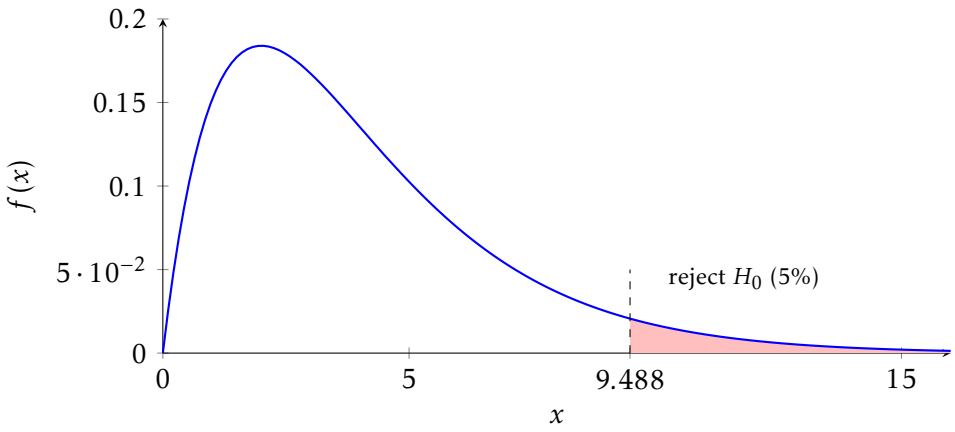
The key is that with two cells there is really only *one* free quantity, and it is approximately normal.

Remark (More cells)**.** The same idea with the multinomial distribution (and much more work) shows that for c cells, $X^2 \approx \chi^2_{c-1}$ when the cell probabilities are fully specified. Notice the degrees of freedom: c cells minus 1 constraint, because the frequencies must total n — exactly as O_2 was determined by O_1 above.

Fact (The $E \geq 5$ rule) — The argument above relies on the binomial-to-normal approximation, which fails when np is small. To keep the approximation honest, **every expected frequency must be at least 5**: where necessary, adjacent classes are combined until this holds. Combining classes is called pooling.

A right-tail test only

If the model is true, X^2 behaves like a χ^2_ν random variable. A *large* value of X^2 means observed and expected frequencies disagree badly — evidence against the model. A *small* value just means the fit is good. So the chi-squared test is always a test on the **right-hand tail only**: we reject H_0 when X^2 exceeds the critical value.



The χ^2_4 density with the 5% rejection region shaded: the critical value is 9.488.

Remark (Too good to be true?)**.** There is no significance test on the left-hand tail at A Level. But a value of X^2 deep in the left tail *should* raise an eyebrow: real data is noisy, and a suspiciously perfect fit suggests the data may have been tidied up. R. A. Fisher famously analysed Gregor Mendel's pea-breeding data and found the fit to Mendel's 3:1 ratios was far better than chance should allow — the combined chi-squared statistic was so small that data this good would arise only a few times in 100 000. Whether Mendel (or an enthusiastic assistant) trimmed the data remains controversial. For your exams: tests are right-tailed, always.

Remark. This test is in some sense backwards compared to other hypothesis tests. Rejecting H_0 gives evidence the data was *not* drawn from the specified distribution; but failing to reject does *not* give evidence that it *was* — it just means we couldn't tell the difference.

The Goodness-of-Fit Test

Tip (The procedure)

1. State H_0 and H_1 (see below for the phrasing).
2. Compute the expected frequencies E_i under H_0. **Pool** adjacent cells until every $E_i \geq 5$.
3. Compute $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.
4. Degrees of freedom: ν = (number of cells after pooling) – 1.
5. Compare with the critical value of χ^2_ν from the formula booklet at the given significance level (right tail).
6. Conclude *in context*, without over-claiming.

Remark (Phrasing the hypotheses). Mark schemes write the hypotheses as, for example,

$$H_0 : \text{the data are consistent with the distribution } B(4, 0.5)$$
$$H_1 : \text{the data are not consistent with the distribution } B(4, 0.5)$$

Strictly this phrasing is poor — the real null assumption is that the data were drawn from a *population* with the specified distribution, and "consistency" is a property of the sample, not a hypothesis about the population. But this is how the mark schemes phrase it, so it is best to do the same.

Example (Given ratio)

A genetics model predicts that four types of flower should occur in the ratio 9 : 3 : 3 : 1. In a sample of 160 flowers the observed counts are 86, 35, 26 and 13. Test at the 5% significance level whether the data are consistent with the model.

Example (Discrete uniform)

Using the die data from earlier ($O_i = 25, 17, 15, 23, 24, 16$ in 120 rolls), test at the 5% level whether the die is fair.

Example (OCR MEI Further Statistics, June 2023 (part))

An eight-sided dice has its faces numbered $1, 2, \ldots, 8$. A student thinks that the dice may be biased. To investigate this, the student decides to roll the dice 80 times and then carry out a χ^2 goodness of fit test of a uniform distribution. The spreadsheet below shows the data for the test, where some of the values have been deliberately omitted.

	A	B	C	D
1	Score	Observed frequency	Expected frequency	Chi-squared contribution
2	1	14	10	1.6
3	2	4	10	3.6
4	3	10	10	0
5	4	15	10	
6	5	6	10	1.6
7	6	11	10	0.1
8	7	7	10	0.9
9	8		10	0.9

(i) Explain why all of the expected frequencies are equal to 10.

(ii) Determine the missing values in cells B9 and D5.

(iii) Carry out the χ^2 test at the 5% significance level.

Textbook Exercises: [CUP.S] Ch 6 §3; [S3&4] S3 Ch 5

Estimating Parameters from the Data

Sometimes the hypothesis specifies only the *family* of distribution — "the data follow a Poisson distribution" — without giving the parameter. We then estimate the parameter from the data itself (e.g. $\hat{\lambda} = \bar{x}$) before computing expected frequencies. This convenience has a price.

Fact (Degrees of freedom with estimated parameters) — Each parameter estimated from the data imposes one extra constraint on the expected frequencies (we have forced the model to match the data in one more way), so

$$\nu = (\text{cells after pooling}) - 1 - (\text{number of parameters estimated}).$$

Tip

Your conclusion must **not** mention the value of the estimated parameter, because it was never part of the null hypothesis. Write "insufficient evidence to suggest the data do not follow a Poisson distribution", *not* "…do not follow Po(2)".

Example (Poisson with estimated mean)

The number of calls arriving at a helpdesk was recorded for each of 80 one-minute intervals:

Calls per minute	0	1	2	3	4	5
Frequency	11	21	22	13	9	4

Test, at the 5% significance level, whether a Poisson distribution is a suitable model.

We met the next context in the Poisson notes: bird-watchers recording the number, N, of separate bursts of chaffinch song in 5 minute periods, with sample mean 3.55 and variance 5.6475 over 60 periods. There we judged the Poisson model informally by comparing mean and variance; the chi-squared test makes the comparison of model and data formal.

Example (OCR Further Stats, June 2024 (parts))

The complete results for the 60 periods are shown in the table.

n	0	1	2	3	4	5	6	7	8	≥ 9
Frequency	10	3	7	8	13	6	6	2	5	0

The bird-watchers carry out a χ^2 goodness of fit test at the 5% significance level.

(a) State suitable hypotheses for the test.

(b) Determine the contribution to the test statistic for $n = 3$.

(c) The total value of the test statistic, obtained by combining the cells for $n \leq 1$ and also for $n \geq 6$, is 9.202, correct to 4 significant figures. Complete the goodness of fit test.

Textbook Exercises: [CUP.S] Ch 6 §3, Ch 7 §9; [S3&4] S3 Ch 5

Contingency Tables

Definition. A contingency table splits a sample according to two attributes simultaneously, tabulating the frequency of each combination. A table with r rows and c columns is called an $r \times c$ contingency table (same convention as matrices).

Think of a contingency table as an observed sample from a *bivariate joint distribution*: each individual carries a pair of values (X, Y). Recall that X and Y are independent iff

$$\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y) \quad \text{for all } x, y.$$

We do not know the population distribution of X or Y — but, just as in a goodness-of-fit test, if the observed cell counts differ significantly from what independence would predict, we have reason to suspect that the attributes are associated.

Expected frequencies under independence

Fact — With row totals R_i, column totals C_j and grand total n,

$$E_{ij} = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}} = \frac{R_i\, C_j}{n}.$$

Where does this formula come from? Derive it from the definition of independence.

Degrees of freedom

Fact — For a test of independence in an $r \times c$ contingency table,

$$\nu = (r-1)(c-1).$$

As with goodness of fit, rows or columns, as appropriate, should be combined so that each expected frequency is at least 5.

Why $(r-1)(c-1)$? Count how many cells you are genuinely free to choose.

Example (Test for independence)

150 people were asked their opinion on a proposed bypass:

	For	Against	Undecided	Total
Under 40	32	28	15	75
40 and over	18	42	15	75
Total	50	70	30	150

Test, at the 5% significance level, whether opinion is independent of age group.

Example (OCR S3, June 2012)

A study was carried out into whether patients suffering from a certain respiratory disorder would benefit from particular treatments. Each of 90 patients who agreed to take part was given one of three treatments A, B or C as shown in the table.

Treatment	A	B	C
Number in group	31	25	34

(i) It is claimed that each patient was equally likely to have been given any of the treatments. Test at the 5% significance level whether the numbers given each treatment are consistent with this claim.

(ii) After 3 months the numbers of patients showing improvement for treatments A, B and C were 14, 18 and 25 respectively. By setting up a 2×3 contingency table, test whether the outcome is dependent on the treatment. Use a 5% significance level.

(iii) If one of the treatments is abandoned, explain briefly which it should be.

Textbook Exercises: [CUP.S] Ch 6 §1; [S3&4] S3 Ch 5

Yates' Continuity Correction

The chi-squared distribution is *continuous*, but the values O_{ij} are integers, and the E_{ij} (built from integer totals) are fixed rationals — so X^2 can only take certain isolated values. As always when a discrete quantity is approximated by a continuous distribution, we should consider a continuity correction.

For large tables there are so many cells, and so many achievable values of X^2, that the approximation is fine without one. The worst case is the 2×2 table: there $\nu = 1$, so fixing the margins leaves only *one* free cell, and (check this!) all four cells have the *same* value of $|O-E|$. The achievable values of X^2 are then few and widely spaced.

Definition (Yates' correction)**.** For a 2×2 contingency table, **Yates' continuity correction** replaces the test statistic by

$$X^2 = \sum \frac{\left(|O_i - E_i| - \frac{1}{2}\right)^2}{E_i},$$

i.e. each difference is shrunk towards zero by half before squaring. This correction is used in the special case of a 2×2 table.

Remark. Only apply the correction when $|O-E| > \frac{1}{2}$. If $|O-E| \le \frac{1}{2}$ the observed values are already as close to expected as integer data can be: the fit is essentially perfect and no test is needed (blindly applying the formula would *increase* the discrepancy, which is nonsense). Some books write the corrected statistic with $\max\left(|O-E| - \frac{1}{2}, 0\right)$ for this reason.

Example (2×2 with Yates' correction)

In a trial, 80 patients were randomly assigned a drug or a placebo:

	Recovered	Did not recover	Total
Drug	26	14	40
Placebo	17	23	40
Total	43	37	80

Test, at the 5% significance level, whether recovery is independent of treatment.

Example (OCR S3, June 2007)

The students in a large university department take a trial examination some time before the proper examination. A random sample of 60 students took both examinations during a particular course. 42 students passed the trial examination, 36 passed the proper examination and 13 failed both examinations.

(i) Copy and complete the following contingency table.

		Proper		
		Pass	Fail	Total
Trial	Pass			42
	Fail		13	
	Total	36		60

(ii) Carry out a test of independence at the $\frac{1}{2}\%$ level of significance.

Remark (Fisher's exact test)**.** The chi-squared test for a 2×2 table quietly assumes the row and column totals are fixed. If they really are — pick the cell in the top-left as the variable and the hypergeometric distribution gives the *exact* probability of each possible table; summing over tables at least as extreme as the one observed is Fisher's exact test, no approximation needed. Yates' correction is precisely an attempt to nudge the chi-squared approximation towards these exact values. In real life margins are rarely all fixed: a medical trial fixes the group sizes but not the recovery counts (a *comparative trial*); a survey classifying people by two attributes fixes only the grand total (a *double dichotomy*); fixing everything is an *independence trial*. The chi-squared machinery is used for all three at this level.

Textbook Exercises: [CUP.S] Ch 6 §1; [S3&4] S3 Ch 5; [Toller] Ch 9