Chi-Squared Contingency Table Tests

Steps

1. State the H_0 (null hypothesis) and H_1 (alternative hypothesis). The former is that there is no association / they are independent.
2. Calculate the expected frequencies (for if they were independent), using (row total x col. total) / total.
   - Note: if any expected frequencies < 5, group appropriately.
3. Calculate the test statistic using:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

If O and E are close $\Rightarrow X^2$ will be small. Otherwise, X^2 will be large.

4. Obtain any critical values from the chi-squared table.

Degrees of freedom

- For an $m \times n$ table, there are $(m-1)(n-1)$ degrees of freedom.

Yates' continuity correction

- In the case of a 2×2 table (1 degree of freedom)

$$X^2 = \sum \frac{(|O-E| - 0.5)^2}{E}$$ is a better approximation.

Worked Example

Observed (binge drinking)	Never	Occassional	Frequent
Trouble w/ police	71	154	398
No trouble w/ police	4992	2808	2737

H_0: Binge drinking and trouble with police are independent.
H_1: There is an association between binge drinking and trouble with the police.

- Degrees of freedom : $(2-1)(3-1) = 2$

- Expected (binge drinking)

	Never	Occasional	Frequent
Trouble w/ police	282.6	168.4	175.0
No trouble w/ police	4780.4	2796.6	2960.0

using $\frac{\text{row total} \times \text{col total}}{\text{total}}$

try to have many d.p.s as it can make a difference

$\Rightarrow X^2 = 469.6$

meaning the p-value is $< 2.2 \times 10^{-16}$ $\Rightarrow$ very strong evidence against H_0.

calculated using software

very large value $\Rightarrow$ very far from expected

Yates' correction for 2 × 2 contingency tables

Given the appropriate conditions $X^2 = \Sigma\frac{(O-E)^2}{E}$ can be approximated by a χ^2 distribution. In the case of a 2 × 2 contingency table the approximation can be improved by using $\Sigma\frac{(|O-E|-0.5)^2}{E}$ instead of $\Sigma\frac{(O-E)^2}{E}$. This is known as Yates' correction.

The underlying reason for this is that the Os are discrete but the χ^2 distribution is continuous. Hence this is often called Yates' continuity correction.

For a 2 × 2 table, $\Sigma\frac{(|O-E|-0.5)^2}{E}$ should be calculated.
This is known as Yates' correction.

$|x|$ means the numerical value of x. Thus $|6| = 6$ and $|-3| = 3$.

Worked example 7.5

A university requires all entrants to a science course to study a non-science subject for one year. In the first year of the scheme entrants were given the choice of studying French or Russian. The number of students of each sex choosing each language is shown in the following table:

	French	Russian
Male	39	16
Female	21	14

Use a χ^2 test at the 5% significance level to test whether choice of language is independent of gender.

Solution

H_0 Subject chosen is independent of gender
H_1 Subject chosen is not independent of gender

	O	E	$O-E$	$\lvert O-E\rvert-0.5$	$\frac{(\lvert O-E\rvert-0.5)^2}{E}$
Male/French	39	36.67	2.33	1.83	0.091
Male/Russian	16	18.33	−2.33	1.83	0.183
Female/French	21	23.33	−2.33	1.83	0.144
Female/Russian	14	11.67	2.33	1.83	0.287

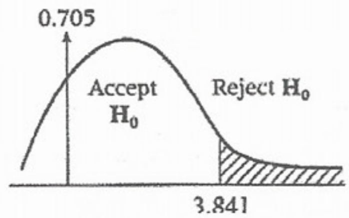
Be careful to find the modulus of $O-E$ (i.e. $|O-E|$) **before** subtracting 0.5.

Note: Rounding the Es to 2 d.p. may lead to small errors in the calculated value. In this case, a more accurate value is 0.707. Such small differences are of little importance.

$$\Sigma\frac{(|O-E|-0.5)^2}{E} = 0.705$$

There are $(2-1) \times (2-1) = 1$ degrees of freedom. Critical value for 5% significance level is 3.841.

Accept that choice of subject is independent of gender.



Learning objectives

After studying this chapter, you should be able to:

- analyse contingency tables using the χ^2 distribution
- recognise the conditions under which the analysis is valid
- combine classes in a contingency table to ensure the expected values are sufficiently large
- apply Yates' correction when analysing 2×2 contingency tables.

We use a chi-squared test to test whether 2 variables are independent (the null hypothesis) or whether there is an association between them. We do this by looking at "goodness of fit" and comparing the observed values with the expected values. If the difference is small, we would conclude they are independent.

Step 1: State the Hypotheses H_0: the variables are independent/no association (include context of question).

Step 2: Calculate the expected frequencies $\qquad \dfrac{\text{row total} \times \text{column total}}{\text{TOTAL}}$

***if any expected frequencies ≤ 5 must combine classes

Step 3: Calculate the test statistic (formula book) $\qquad X^2 = \sum \dfrac{(O - E)^2}{E}$

$X^2 = \Sigma \dfrac{(O - E)^2}{E}$ may be approximated by the χ^2 distribution provided:

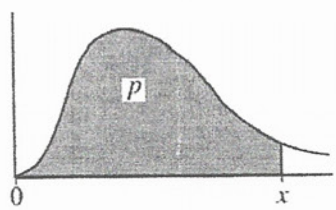
(i) the Os are frequencies,

(ii) the Es are reasonably large, say >5.

Step 4: Obtain critical value from the chi-squared table

TABLE 6 PERCENTAGE POINTS OF THE χ^2 DISTRIBUTION

The table gives the values of x satisfying $P(X \leqslant x) = p$, where X is a random variable having the χ^2 distribution with ν degrees of freedom.



p / ν	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995	p / ν
1	0.00004	0.0002	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750	5

An $m \times n$ contingency table has $(m - 1)(n - 1)$ degrees of freedom.

Example

The results of a recent police survey of traffic travelling on motorways produced information about the genders of drivers and the speeds, S miles per hour, of their vehicles, as tabulated below.

	$S \leqslant 70$	$70 < S \leqslant 90$	$S > 90$	Total
Male	17	40	70	127
Female	30	25	18	73
Total	47	65	88	200

Investigate, at the 1% level of significance, the claim that there is no association between the gender of the driver and the speed of the car. *(11 marks)*

H_0: gender and car speed are independent.
H_1: gender and car speed have some association.

Expected values:

	$S \le 70$	$70 < S \le 90$	$S > 90$
Male	29.845	41.275	55.88
Female	17.155	23.725	32.12

$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
−12.845	164.994025	5.52836...
−1.275	1.625625	0.03938 5...
14.12	199.3744	3.567902649
12.845	164.994025	9.617838823
1.275	1.625625	0.068519442
−14.12	199.3744	6.2071731 0

≈ 25.029

Degrees of freedom $= (2-1)(3-1) = 3$

As $25.029 > 11.345$, there is sufficient evidence to reject H_0 in favour of H_1 at the 1% level of significance.