Contingency tables

The data given in the table below is from the 2014 study on the relationship between volume and type of alcohol consumed during pregnancy in relation to miscarriage.

	Miscarriage		
Average number of alcoholic drinks per week	Yes	No	
4+ drinks per week	11	21	32
1-3 drinks per week	66	337	403
No alcohol intake	95	531	626
	172	889	1061

Is this data consistent with miscarriage being independent of average number of alcoholic drinks consumed? To answer this question we need to determine what we would expect the data to look like if these two variables were independent. If miscarriage and average alcoholic drinks consumed are independent, then

If A and B are independent $P(A \cap B) = P(A)P(B)$

$$\mathbb{P}(\text{'Miscarried – Yes' AND 'No alcohol intake}) = P(\text{'Miscarried-Yes'})\,P(\text{'No Alcohol intake'})$$

We don't know $\mathbb{P}$('Miscarried – Yes') or $\mathbb{P}$('No alcohol intake), but we could estimate these from the data by

$$\widehat{\mathbb{P}}(\text{'Miscarried – Yes'}) = \frac{172}{1061} \simeq 0.162 \qquad \widehat{\mathbb{P}}(\text{'No alcohol intake}) = \frac{626}{1061} \simeq 0.59$$

So assuming these two variables are independent, the expected number of women in the study who miscarried and had no alcohol intake is

$$\text{no. of women in study} \times \widehat{\mathbb{P}}(\text{'Miscarried-Yes'}) \times \widehat{\mathbb{P}}(\text{'No Alcohol'}) = 1061 \times \frac{172}{1061} \times \frac{626}{1061} \simeq 101.48$$

In this way we can create another table containing the expected number of women in each category assuming independence between miscarriage and alcohol consumption.

	Miscarriage	
Average number of alcoholic drinks per week	Yes	No
4+ drinks per week	$\frac{32 \times 172}{1061} \simeq 5.19$	$\frac{32 \times 889}{1061} \simeq 26.81$
1-3 drinks per week	$\frac{403 \times 172}{1061} \simeq 65.33$	$\frac{403 \times 889}{1061} \simeq 337.67$
No alcohol intake	101.48	$\frac{626 \times 889}{1061} \simeq 524.52$

If the tables of the observed and expected counts differ greatly, then this would be evidence against miscarriage and alcohol consumption being independent. We measure the difference between the two tables using the statistic

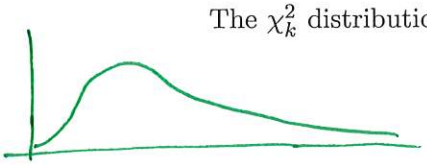
$$X^2 = \sum_i \frac{(e_i - o_i)^2}{e_i} \qquad = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

where e_i and o_i are respectively the expected and observed counts for the i-th cell of the table.

In this case, the test statistic evaluates to ...

$$X^2 = \frac{(11-5.19)^2}{5.19} + \frac{(21-26.81)^2}{26.81} + \frac{(66-65.33)^2}{65.33}$$
$$+ \frac{(337-337.67)^2}{337.67} + \frac{(95-101.48)^2}{101.48} + \frac{(531-524.52)^2}{524.52}$$
$$= 8.4298$$

Assuming that the two variables are independent, X^2 has (approximately) a chi-squared distribution with $(r-1) \times (c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the table. This distribution is often denoted by $\chi^2_{(r-1)(c-1)}$. For this approximation to be reasonable we generally need the expected count for all cells to be at least one and in 80% of cells the expected count is least 5. The χ^2_k distribution has probability density function



$$f(x;k) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}, \quad x > 0,$$

Arises as sum of squared $N(0,1)$

$\mathbb{E}\chi^2_k = k$ $\quad \mathrm{Var}(\chi^2_k) = 2k$

where Γ is the gamma function. As with the normal and t-distributions, the cumulative distribution function of the χ^2_k distribution has been tabulated so we can determine the p-value for the test by looking up the appropriate table:

$$p - \text{value} = \mathbb{P}\left(\chi^2_{(r-1)(c-1)} \geqslant X^2\right).$$

Therefore, the p-value for testing $\quad r = 3, \; c = 2 \quad$ degrees of freedom $= 2$.

- H_0: 'Miscarriage' and 'Alcohol consumption' are independent, against
- H_1: There is some association between 'Miscarriage' and 'Alcohol consumption'

is

using tables $\quad \mathbb{P}(\chi^2_2 \geqslant 7.378) = 0.025 \quad$ and $\quad \mathbb{P}(\chi^2_2 \geqslant 9.210) = 0.01$

p-value $= \mathbb{P}(\chi^2_2 \geqslant 8.4298)$ is between 0.01 and 0.025

Hence, we conclude ...

MATLAB `chi2cdf(8.4298, 2, 'upper') = 0.0148`

R `pchisq(8.4298, 2, lower.tail = FALSE).`

There is moderate evidence against the null hypothesis, suggesting an association between miscarriage and alcohol consumption.

In our example only the number of women in the study was fixed. The proportion of women in each category in the study is reflective of the population from which the sample was taken. An alternative way for a study like this to be conducted is to fix the number of women in each of the groups of alcohol consumption. Denote the probability of miscarriage in the three alcohol consumption groups ('No alcohol', '1-3 drinks per week', '4+ drinks per week') by p_0, p_1 and p_2, respectively. We would then be interested in testing

- H_0: $p_0 = p_1 = p_2$, against
- H_1: at least one of the p_i is different.

Computation of the p-value in this instance proceeds in exactly the same way as for the test of independence. It is only the interpretation of the result which is different. This is sometimes called a test for *homogeneity*.

Exercise: In a study of public attitudes to green roof systems[1], 450 people filled in questionnaires. 66.9% of the respondents replied that they might be interested in installing a green roof on their house. A table recording the respondents age group and interest is given below. Is there a difference in attitudes towards green roof systems across age groups?

Age group	Under 18	18–25	26–40	Over 40	Total
Interested	136	42	48	75	301
Not interested	83	18	31	17	149
Total	219	60	79	92	450

The null and alternative hypotheses are:

H_0: "Interest" and "Age" are independent (no association).

H_1: some association between "interest" and "age"

[1]Fernandez-Cañero, R., Emilsson, T., Fernandez-Barba, C. & Herrera Machuca, M.A. (2013) Green roof systems: A study of public attitudes and preferences in southern Spain. Journal of Environmental Management, 128, 106-115.

The p-value is ...

Expected counts

	<18	18-25	26-40	40+
Interested	$\frac{219 \times 301}{450}$	$\frac{60 \times 301}{450}$	$\frac{79 \times 301}{450}$	$\frac{92 \times 301}{450}$
	$= 146.49$	$= 40.13$	$= 52.84$	$= 61.54$
Not Interested	$\frac{219 \times 149}{450}$	$\frac{60 \times 149}{450}$	$\frac{79 \times 149}{450}$	$\frac{92 \times 149}{450}$
	$= 72.51$	$= 19.87$	$= 26.16$	$= 30.46$

$$X^2 = \sum \frac{(e_i - o_i)^2}{e_i} = \frac{(146.49 - 136)^2}{146.49} + \frac{(40.13 - 42)^2}{40.13} +$$

$$+ \cdots + \frac{(30.46 - 17)^2}{30.46} = 12.7625$$

degrees of freedom $= (r-1) \times (c-1) = (2-1)\times(4-1) = 3$

p-value $= \mathbb{P}(\chi^2_3 \geq 12.7625) = 0.0052$

Hence, we conclude ...

There is strong evidence against the null hypothesis, suggesting an association between 'interest' in green roof systems and age.