Tests of Independence Using Contingency Table Data

So far, we have discussed how hypothesis testing can be performed to measure the following:

- Mean
- Variance
- Correlation

In addition, hypothesis tests can be used to test the **independence** between categorical or discrete data. This requires a non-parametric test based on a chi-square, χ^2, distribution.

Consider the following contingency table of observed frequencies for analysts ratings for stocks in various sectors:

Sector	**Analyst Rating**			**Total**
	Buy	**Hold**	**Sell**	
Technology	22	3	11	**36**
Consumer Staples	12	9	10	**31**
Airline	1	3	4	**8**
Energy	6	7	5	**18**
Financial	2	2	3	**7**
Total	**43**	**24**	**33**	**100**

To test the independence between the sectors and analyst ratings for this sample, we need to calculate the chi-square test statistic using the following formula:

$$\chi^2 = \sum^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- m: Number of cells in the contingency table (excluding the totals)
- O_{ij}: **Observed** frequency in row i and column j
- E_{ij}: **Expected** frequency in row i and column j

$$E_{ij} = \frac{(\text{Total of row } i) \times (\text{Total of column } j)}{\text{Overall total}}$$

Note that the formula of E_{ij} above implies that the variables are independent. In other words, if the observations are close to the values of E_{ij} calculated using this formula (i.e., χ^2 has a small value), it is likely that the variables are independent.

There are $(r-1)(c-1)$ degrees of freedom, where r and c represent the total number of rows and columns, respectively. The hypotheses are set up as follows:

$$H_0 : \text{Independence between variables}$$
$$H_1 : \text{Dependence between variables}$$

The null hypothesis is rejected if the calculated chi-square test statistic exceeds a critical value, indicating that the variables are not independent. The chi-square test of independence has only one region of rejection, on the right side of the distribution. All else equal, the critical chi-square value will be higher as the number of degrees of freedom increases and the region of rejection narrows.

To demonstrate this process, let's work through an example. First, determine the expected frequency for the contingency table above using the E_{ij} formula. For example, the expected number of technology stocks with a buy rating is:

$$E_{11} = \frac{(\text{Total of row 1}) \times (\text{Total of column 1})}{\text{Overall total}}$$

$$= \frac{36 \times 43}{100}$$

$$= 15.48$$

The complete results for expected frequencies are shown in the table below:

Sector	Analyst Rating		
	Buy	Hold	Sell
Technology	15.48	8.64	11.88
Consumer Staples	13.33	7.44	10.23
Airline	3.44	1.92	2.64
Energy	7.74	4.32	5.94
Financial	3.01	1.68	2.31

Unsurprisingly, the cells in the table sum up to 100.

Next, compute the scaled squared standard deviation for each cell using the following formula:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

For example, for technology stocks with a buy rating, the scaled squared standard deviation is:

$$\frac{(22 - 15.48)^2}{15.48} = 2.746$$

The complete results are shown in the table below:

Sector	Analyst Rating		
	Buy	**Hold**	**Sell**
Technology	2.746	3.682	0.065
Consumer Staples	0.133	0.327	0.005
Airline	1.731	0.608	0.701
Energy	0.391	1.663	0.149
Financial	0.339	0.061	0.206

The chi-square test statistic is the sum of all the cells in the table. In this case,

$$\begin{aligned}\chi^2 &= \sum^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= 12.805\end{aligned}$$

There are (5 – 1)(3 – 1) = 8 degrees of freedom. Assuming a 5% level of significance, the critical value is 15.507. Since $\chi^2 < 15.507$, the **null hypothesis not rejected**, which means the sectors and analyst ratings are independent.

It would be valuable for the analyst to know which cells have observations that deviate significantly from their expectations, assuming the variables are independent. This deviation is captured by the *standardized residual* (a.k.a. *Pearson residual*).

$$\text{Standardized residual} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

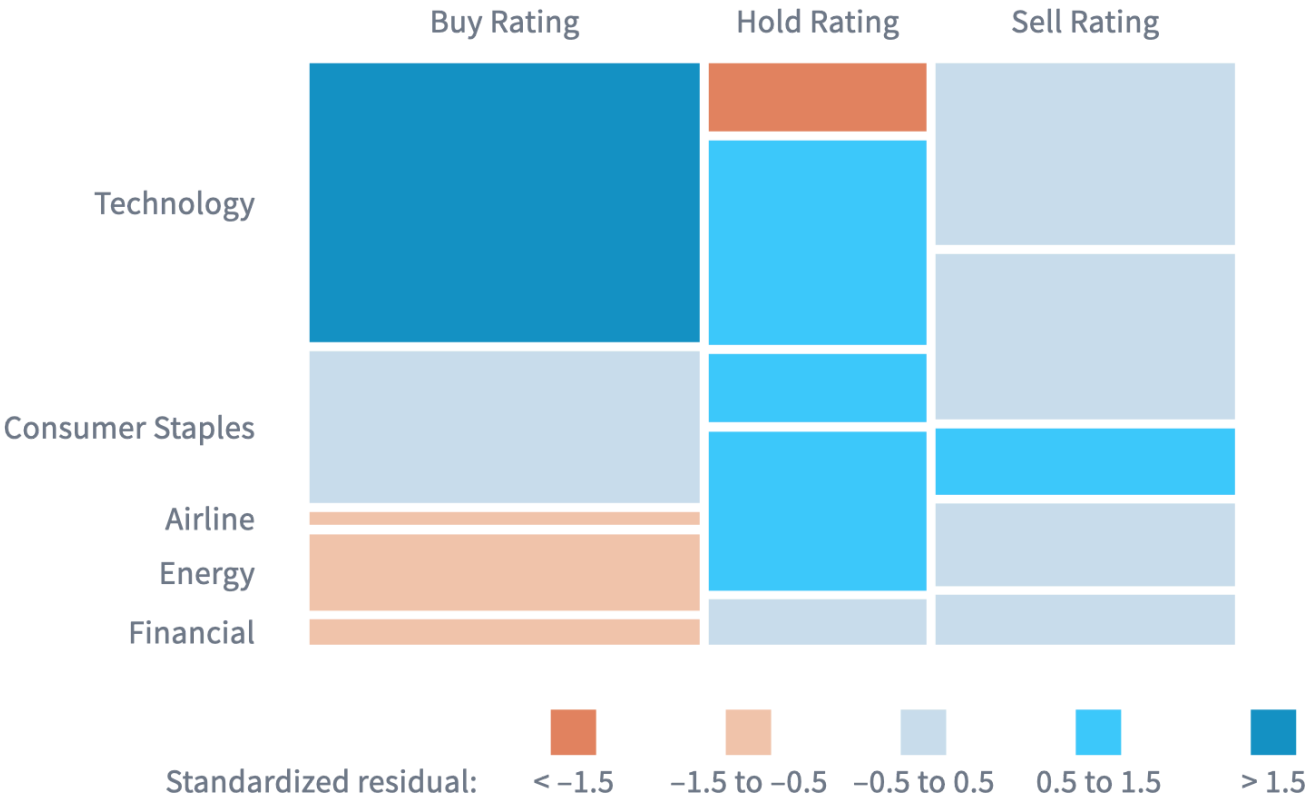
For example, the standardized residual for technology stocks with a buy rating is calculated as follows:

$$\frac{(22 - 15.48)}{\sqrt{15.48}} = 1.657$$

The standardized residuals for each cell in our example are shown in the table below:

Sector	Analyst Rating		
	Buy	**Hold**	**Sell**
Technology	1.657	–1.919	–0.255
Consumer Staples	–0.364	0.572	–0.072
Airline	–1.316	0.779	0.837
Energy	–0.625	1.289	–0.386
Financial	–0.582	0.247	0.454

Finally, the standardized residuals can be represented visually in a *mosaic*, such as the one shown below:



In this mosaic, the width of the rectangles represents the proportion based on the analyst ratings, while the height represents the proportion based on the sectors. This visualization helps use to determine that the largest sources of dependence are:

- Technology stocks with a buy rating (observed frequency greatly exceeds expected frequency).

- Technology stocks with a hold rating (observed frequency is significantly lower than expected).