Pearson’s Standardized Residuals

Contingency tables, commonly known as cross-tabulations, show the **frequency** of observations that fall into different **combinations of categories** for two categorical variables.

Pearson's **standardized residuals** are used to **evaluate the association** between two categorical variables in a contingency table by comparing observed and expected frequencies in each cell.

Indeed, the presence of **large standardized residuals** indicates significant deviations from independence, which implies an **association** between the variables.

For instance, let's consider a bank that conducted a **survey** across three branches of its network.

The purpose of the survey is to gain an understanding of the **primary causes of customer dissatisfaction**. The bank's aim is to determine if customer dissatisfaction is dependent on the branch and, if so, to identify the **main contributors**.

The data is contained in this Excel file:

Causes_of_Dissatisfaction_in_the_Relationship	Branch	Responses
High cost of services	Branch A	23
High cost of services	Branch B	7
High cost of services	Branch C	37
Failure to comply with the conditions	Branch A	39
Failure to comply with the conditions	Branch B	13
Failure to comply with the conditions	Branch C	8
Loan Issues	Branch A	13
Loan Issues	Branch B	5
Loan Issues	Branch C	13
Staff quality	Branch A	13
Staff quality	Branch B	8
Staff quality	Branch C	8

The Crosstab node using the KNIME platform[1] has provided a 2-way frequency table below.

Cross Tabulation of Causes of Dissatisfaction in the Relationship by Branch

Frequency Expected Deviation	Branch A	Branch B	Branch C	Total
Failure to comply with the conditions	39	13	8	60
	28,2353	10,5882	21,1765	
	10,7647	2,4118	-13,1765	
High cost of services	23	7	37	67
	31,5294	11,8235	23,6471	
	-8,5294	-4,8235	13,3529	
Loan Issues	13	5	13	31
	14,5882	5,4706	10,9412	
	-1,5882	-0,4706	2,0588	
Staff quality	13	8	8	29
	13,6471	5,1176	10,2353	
	-0,6471	2,8824	-2,2353	
Total	88	33	66	187

☑ Frequency
☑ Expected
☑ Deviation
☐ Percent
☐ Row Percent
☐ Column Percent
☐ Cell Chi-Square

Max rows: 10

Max columns: 10

Statistics for Table of Causes of Dissatisfaction in the Relationship by Branch

Statistic	DF	Value	Prob
Chi-Square	6	27,4104	0,0001

The same results can be obtained by using the *chisq.test()* function in the R software[2].

```
> ctab <- xtabs( data=df, formula= Responses ~
         Causes_of_Dissatisfaction_in_the_Relationship + Branch)

> chisq.test(ctab)      # observed test statistics

     Pearson's Chi-squared test

data: ctab
X-squared = 27.41, df = 6, p-value = 0.0001213


> qchisq(0.95, df=6)      # critical value

12.59159
```

From these results, a *Chi-square* value of 27.41 is observed, which would be expected to be 12.59 in case of independence.

[1] https://www.knime.com/
[2] https://www.r-project.org/

It can be concluded that the **main causes** of customer dissatisfaction **are influenced by the specific branch**, as 27.4 > 12.6 and the *p-value* is 0.00012 (< 0.05).

In order to determine the **causes of dissatisfaction** that could have a significant impact on each branch, we can then calculate the **standardized residuals**[3]. As previously mentioned, the residuals are used to **measure the strength of the difference** between the observed and expected values using the following formula:

$$r_{ij} = \left(f_{ij}{}^{observed} - f_{ij}{}^{expected}\right) / \sqrt{f_{ij}{}^{expected}(1 - p_{i.})(1 - p_{.j})}$$

Here, $f_{ij}{}^{observed}$ and $f_{ij}{}^{expected}$ represent the observed and expected frequencies of the table cell *(i, j)*, respectively. Furthermore, $p_{i.}$ denotes the proportion in row i, calculated as $(f_{i.}{}^{observed} / f^{observed})$, while $p_{.j}$ signifies the proportion in column j, computed as $(f_{.j}{}^{expected} / f^{expected})$. Here, $f_{i.}$ and $f_{.j}$ refer to the row and column totals, respectively.

If the standardized residual r_{ij} is positive, it indicates that there are more subjects in that cell than expected, whereas if it is negative, it indicates that there are fewer.

This difference has a normal distribution with mean = 0 and standard deviation = 1, and it is considered significant if its absolute value is greater than 2 (1.96σ).

The R software's *chisq.test()* function can also be used to calculate standardized residuals:

```
> chisq.test(ctab)$stdres

                                         Branch

Causes_of_Dissatisfaction_in_the_Relationship   Branch A   Branch B   Branch C

        Failure to comply with the conditions  3.3785330  0.9910659 -4.3193597
        High cost of services                 -2.6061207 -1.9296653  4.2613447
        Loan Issues                           -0.6257130 -0.2427409  0.8471762
        Staff quality                         -0.2618905  1.5274419 -0.9449414
```

Based on this analysis, it can be concluded that **dissatisfaction is dependent on the following factors**:

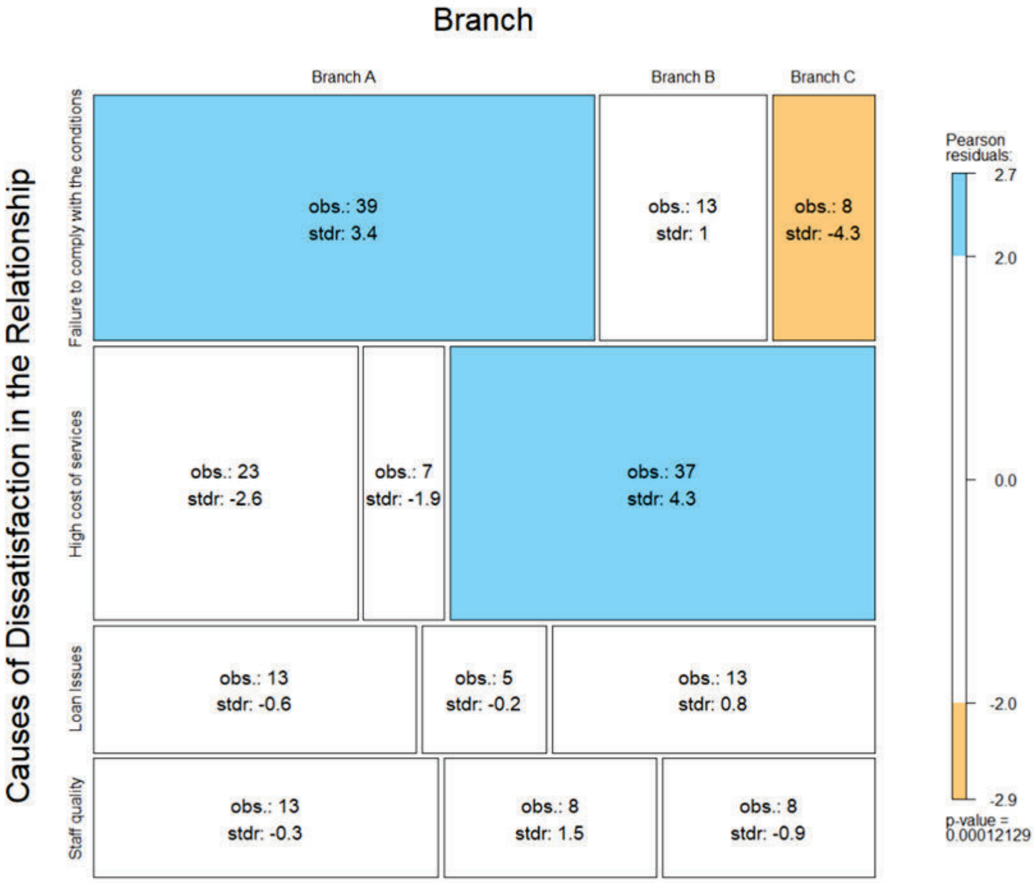
- Non-compliance conditions for Branch A

- Costs for Branch C

---

[3] For more information, please refer to section 2.4.5 of Agresti (2007).

A mosaic plot can be used to visualize a contingency table. The *mosaic()* function of R package "*vdc*" is used for this task:



The **blue** color indicates that the observed value is **higher than the expected** value if the data were random, while the **orange** color signifies that the observed value is **lower than the expected** value if the data were random.

References

Agresti, A. (2007), An Introduction to Categorical Data Analysis, 2nd Edition, New York: John Wiley & Sons.

Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. London: Sage publications.

Contacts

studio roccato
Data Science Training & Consulting

e-mail: alfredo.roccato(at)fastwebnet.it

www.alfredoroccato.it

AR 01/2014