

EXPLORING THE RELATIONSHIP BETWEEN TWO VARIABLES: CONDUCTING CROSS-TABULATIONS AND CHI-SQUARE ANALYSIS

Disclaimer: This draft guide is confidential and has been issued for the purpose of restricted consultation. It should not be quoted or circulated until finalised and approved for wider publication by the project funder.

The relationship between two categorical variables¹ in the Adult Social Care Survey (ASCS) or Carers Survey (PSS SACE) can be explored using **cross-tabulations** and **chi-square analysis**. Excel-based analysis tools, step-by-step guides and pre-recorded presentations are included in the MAX toolkit to help you conduct these calculations. The purpose of this summary is to provide users with an overview of cross-tabulations and chi-square and the available tools. Examples will be provided throughout and will be based on the question:

Is there a difference in the proportion of men compared to women who say they have as much control over their daily life (ASCOF 1B) as they would like?

This question will also be used in the pre-recorded webinar provided in the MAX toolkit. To provide further examples of the kinds of questions that can be explored using the cross-tabulations and chi-square tool, the step-by-step instructions (also available in the MAX toolkit) and the live webinars (to be hosted by the MAX project team in November and December 2016) will use the following two questions:

Is satisfaction with care and support services [Q1] associated with self-rated health?

Are higher ratings of satisfaction with care and support services [Q1] associated with self-rated health?

¹ A **categorical variable** is a factor where each person or case in a dataset falls into only one of two or more categories. For example, gender (male, female, missing). Categorical data can be unordered (**nominal data** e.g. gender) or ordered (**ordinal data** e.g. control over daily life). Most variables in the ASCS and PSS SACE are categorical. The exceptions are age (in years) and the calculated scores for social care-related quality of life (SCRQoL) and carer social care-related quality of life (Carer SCRQoL).

INTRODUCTION

A **cross-tabulation** summarises the frequency distribution of two variables in a dataset (e.g. gender and control over daily life) in tabular form, known as a **contingency table**. Cross-tabulation analysis is a type of **descriptive statistic** and presents all responses by groups (e.g. male/female) by control over daily life (ASCOF 1B). Cross-tabulation does not test whether the observed difference is statistically significant.

Chi-square builds on the cross-tabulation and is a form of '**hypothesis testing**' that can be used to test whether the relationship between two categorical variables is significant (i.e. there is or isn't a relationship between the variables). The statistical test generates a **p-value** which enables you to determine whether or not to reject the null hypothesis. This p value is the probability that the null hypothesis is true. If the p value is less than a set significance level (usually 5% [0.05] or 1% [0.001]), it is said that the pattern is 'statistically significant'. This means that there is a low probability that the observed pattern arose by chance.

Hypothesis testing

Hypothesis testing involves using statistical tests such as chi square to determine whether the patterns observed in a sample of data (in this instance, the sample of respondents who completed the ASCS or PSS SACE) are likely to reflect those that would be found in the entire population (in this instance, the entire population of adult social care service user or carers within a given LA). A hypothesis tests explores two hypotheses about a population: the **null hypothesis** (there is no difference between variables) and the **alternative hypothesis** (there is a difference between variables).

	Relationship	Example
Null	No difference	There is no difference in the proportion of men and women who say they have as much control over their daily life (ASCOF 1B) as they would like. In other words, the frequency counts observed in the data have arisen 'by chance'
Alternative	Difference	There is a difference in the proportion of men and women who say they have as much control over their daily life (ASCOF 1B) as they would like. In other words, the frequency counts observed in the data have arisen 'by something other than chance' (e.g. an underlying factor)

A word of caution: Hypothesis testing is not 100% certain so there is a risk of drawing an incorrect conclusion.

Type I errors: occur when the null hypothesis is rejected when, in fact, it is true (i.e. you conclude your data shows there is an effect or difference when there is none). The risk of generating a type I error can be reduced by using a lower significance value (e.g. 0.001 rather than 0.05)

Type II errors: occur when the null hypothesis is accepted when, in fact, it is false (i.e. you conclude your data shows no effect or difference when it actually does). The risk of generating a type II error can be reduced by ensuring your sample size is sufficient. Guidance notes are provided alongside the statistical tests explored in the MAX toolkit.



Cross-tabulations and chi-square analysis can be conducted using the **SCRQoL Tool** [ASCS data] or the **Carer SCRQoL Tool** [PSS SACE data] in the MAX toolkit. **Step by step instructions** are also provided in the MAX toolkit and are accompanied by a **pre-recorded training presentation**.

CROSS-TABULATION

Use: a **cross-tabulation** can be used to summarise the frequency distribution of two variables in a dataset (e.g. gender and control over daily life) in tabular form, known as a **contingency table**. Cross-tabulations can be conducted with categorical variables and can be used to describe or explore data. An example of a contingency table based on the example question is provided below.

Control \ Gender	Male	Female	Total
Ideal state	127	232	359
No needs	121	201	322
Some needs	49	77	126
High-level needs	7	12	19
Total	304	522	826

Table 1: control over daily life (Q3a) [ASCOF 1B] by gender

CHI-SQUARE

Use: the chi-square can be used to **statistically test** whether the relationship between two categorical variables is significant (i.e. there is or isn't a significant relationship or difference between the variables). The test explores two variables at a time and applies one measurement (this is known as **univariate analysis**). Each category, however, can have a number of subcategories. Using the example question, the first category (gender) has two subcategories (male, female) and the second category (control over daily life) has four subcategories (ideal state, no needs, some needs, high needs) or fewer, if the subcategories are collapsed.

The chi-square calculation: compares the frequency counts of categories within the dataset (known as **observed data**) with the frequency counts that would have arisen if there was no relationship between the variables (i.e. the data if they arose 'by chance', known as **expected data**) by measuring the **goodness of fit** (the difference between what you would expect to find if there was no relationship between the variables and what has been observed in your population).

The greater the difference between the observed and expected data, the more likely the difference is significant.

Assumptions: the suitability and accuracy of a chi-square test are dependent on the following assumptions being fulfilled.

1. **Sufficient sample size:** the sample has to be large enough to avoid **type 2 error**. See point 2.
2. **Sufficient expected observations for all categories:** depends on the type of chi-square table being generated:
 - a. Expected frequency of 5 or more in all cells of a 2x2 table
 - b. Expected frequency of 5 or more in 80% of cells in larger tables
 - c. No cells with zero expected frequencyIf this assumption is violated, you can combine categories or conduct a Fisher's exact test. Combining categories is covered in the accompanying pre-recorded presentation. Please note that the Fisher's exact test is not covered in the MAX toolkit.

3. **Independence of observations:** each case contributes data to only one cell of the chi-square table. Chi-square analysis is therefore not suitable for data that is measured over time or data that is correlated in any way. In these instances, a McNemar's test may be more suitable. Please note that the McNemar's test is not covered in the MAX toolkit.
4. **Missing values are excluded:** incomplete data should be excluded prior to analysis. Any missing data should also be recorded in your write up.

Interpretation: the Chi square test generates a p-value. This shows whether null hypothesis, which assumes that there is no relationship between variables, can be rejected. If the p-value less is less than the **critical value**² (usually 0.05 or 0.001), the null hypothesis can be rejected and the alternative hypothesis, that there is a relationship between the variables, can be accepted.

A word of caution: chi-square analysis can show whether there is a relationship between the variables but **does not infer causality** – in other words, whether one variable has an effect on the other variable. Further statistical testing and/or research is required to establish causality.

TOOLS IN THE MAX TOOLKIT

A number of tools are provided in the MAX toolkit to help you conduct cross-tabulations and chi-square analysis. These include:

- Excel-based analysis tools for each survey;
- Step-by step instructions (based on ASCS data); and a
- Pre-recorded training presentation which provides an overview of the analysis tool and guidance on how to conduct cross-tabulations and chi-square analysis.

² To determine whether the critical value is significant, you need to know the **Degrees of Freedom (df)**. The df is the number of values in a dataset that are free to vary without influencing the result and are calculated as follows: (rows – 1) x (columns – 1)