

# Tabulation of multiple responses

Ben Jann  
ETH Zurich, Switzerland  
October 20, 2004

**Abstract.** Although multiple response questions are quite common in survey research, Stata's official release does not provide much possibility for an effective analysis of multiple response variables. For example, in a study on drug addiction an interview question might be, "Which substances did you consume during the last four weeks?" The respondents just list all the drugs they took, if any, e.g., an answer could be "cannabis, cocaine, heroin" or "ecstasy, cannabis" or "none", etc. Usually, the responses to such questions are stored as a set of variables and, therefore, cannot be easily tabulated. I will address this issue here and present a new module to compute one- and two-way tables of multiple responses. The module supports several types of data structure, provides significance tests, and offers various options to control the computation and display of the results. In addition, tools to create graphs of multiple response distributions are presented.

**Keywords:** st0001, mrtab, mrgraph, \_mrvmat, multiple responses, multiple testing, tabulate

## 1 Introduction

Surveys often contain questions which can have multiple responses. That is, questions are asked to which a respondent can give zero, one or more answers. For example, such a question might be: "Which of the following devices do you have in your home?" The respondent is then given a list like "1. Television, 2. Dishwasher, 3. Computer, 4. Dry cleaner . . ." and may mark any number of devices. Furthermore, the list might be open-ended, so that the respondent can also name devices that are not listed. Thus, although not necessarily so, multiple response questions often have an explorative character.

The answers to multiple response questions are, in fact, a series of answers and, thus, are usually stored as a series of variables. However, because the variables constitute some topical entity, the combined statistical distribution of all variables may be of interest rather than the separate distributions of the single variables. Moreover, depending on the storage structure, the distributions of the single variables may be completely meaningless if taken individually.

Because multiple response data is spread out over several variables, it cannot be easily tabulated. Firstly, the data structure may vary and it may be necessary to transform the data into different structures before tabulating. In particular, a storage type in which the responses are recorded in the order they have been mentioned by the respondent is quite common even if the ordering is not relevant. For most purposes, however, it is useful to transform such data to a storage mode in which the single variables indicate whether particular responses have been observed or not. Secondly,

statistical information from several variables has to be combined, which is not always just a matter of arranging separate distributions in one table. Both tasks are difficult to accomplish using standard statistical instruments that are designed for the analysis of ordinary (single response) variables. Thus, it appears that Stata's official tools are not sufficient for an effective analysis of multiple response data; additional instruments are needed.

In the following section I will briefly touch on the issue of how to store the answers of multiple response questions and then move on to the presentation of three new commands to support the analysis of multiple responses in section 3. Section 4 illustrates the capabilities and the usage of the new commands and contains some additional considerations about significance tests.

## 2 Approaches to storing multiple responses

The fact that the answers to multiple response questions are typically composed of several bits of information poses difficulties for their representation in a data set. A common way to deal with this issue is to store each part of the answer in a separate variable. Two main approaches may be distinguished: the indicator mode and the polytomous mode.

### 2.1 Indicator mode

Consider the following question which could have been part of a questionnaire on drug addiction:

**Which of the following narcotic substances did you consume during the last four weeks?**

*Check all that apply*

- Cannabis
- Cocktail (cocaine/heroin)
- Heroin (alone)
- Cocaine (alone)
- Ecstasy

Obviously, it would be most straightforward to construct a set of indicator or dummy variables in this case: One variable for each drug. Basically, the example question above is just a shortcut to five separate questions in the manner of “Did you consume cannabis during the last four weeks?”, “Did you consume a cocaine-heroin cocktail during the last four weeks?”, and so on. The data would then look like the following (1 meaning “ticked”, 0 meaning “not ticked”):

id	d1_cannabis	d2_cocktail	d3_heroin	d4_cocaine	d6_ecstasy
1	1	0	0	0	1
2	1	0	0	1	1

3	1	0	1	0	0
4	0	1	1	1	0
5	0	0	0	0	0
...					

Thus, respondent 1 ticked “Cannabis” and “Ecstasy”, respondent 2 ticked “Cannabis”, “Cocaine (alone)” and “Ecstasy”, and so on. Respondent 5 did not tick any of the boxes.

In the remainder of this paper I will use the term *indicator mode* to refer to the situation in which the data are stored as a set of indicator variables. Note that it is not crucial that the variables be dichotomous or binary. The important point is just that each item (here each drug) is represented by its own variable.

## 2.2 Polytomous mode

The indicator mode is particularly suitable if the list of response categories is fixed (like in the example above) and is not too long. However, multiple response questions are often open or half-open due to their explorative nature. For example, the question on drug addiction might also have been

**Which narcotic substances did you consume during the last four weeks?**

.....  
 .....

or maybe

**Which narcotic substances did you consume during the last four weeks?**

Check all that apply

- Cannabis
- Cocktail (cocaine/heroin)
- Heroin (alone)
- Cocaine (alone)
- Ecstasy
- Other substances:

.....  
 .....

The difference to the closed question in section 2.1 is that the respondent is also given the possibility of naming drugs other than the ones the researchers could think of when constructing the questionnaire. This means that (half-)open questions cannot be divided into separate questions for the different drugs; a last question gathering the remainder

is always needed (e.g.: “Besides the already mentioned, what *other* substances did you consume?”).

Thus, the list of possible response categories is not fixed and one would have to update the list of responses continuously while collecting the data: Each time a new drug is named, it would be added to the list and given a unique code. Because the list is not fixed, it is also not possible to set up indicator variables for all items in advance.

Furthermore, even if a complete list of possible answers is known in advance, the indicator mode could be very inefficient because of the vast number of required variables. It would be more efficient in such cases to split the multiple responses according to their “order of appearance”. Think of the answers to the above question as lists of drugs (for example, one answer is “cannabis, ecstasy”, or “cocaine, heroin, cocktail, morphine”). We would then in each case take the first token of the list and save it in a first variable, take the second token and save it in a second variable, and so on. Thus, if applying this strategy and using string variables to store the information the data could look like the following:

id	druguse1	druguse2	druguse3	druguse4
1	cannabis	ecstasy		
2	ecstasy	cocaine	cannabis	
3	heroin	cannabis		
4	cocaine	heroin	cocktail	morphine
5	LSD			
...				

A numerical representation of the same data would be, for instance

id	druguse1	druguse2	druguse3	druguse4
1	1	5	0	0
2	5	4	1	0
3	3	1	0	0
4	4	3	2	6
5	7	0	0	0
...				

with the label definitions 1 “cannabis”, 2 “cocktail”, 3 “heroin”, 4 “cocaine”, 5 “ecstasy”, 6 “morphine”, 7 “LSD”, ..., 0 “no further”.

Because the multiple responses are stored as a set of polytomous variables, I will call this the *polytomous mode*. The approach is suitable if the list of possible response categories is not clearly defined *ex ante* and/or if the list is rather long. Sometimes it is also convenient to use the polytomous mode because it reflects the way that the data have been collected. The obvious disadvantage of the approach is that the number of required variables depends on the observed maximum “length” (number of tokens) of a multiple response, which is not known in advance (unless explicitly limited in the questionnaire).<sup>1</sup> Most of the time it is possible to make reasonable assumptions about this maximum, though.

<sup>1</sup>Of course, the technical maximum length is equal to the number of distinct items covered by the question (e.g. if someone took *all* drugs) unless repeated items are allowed. However, the *reasonable* maximum is usually much smaller than the technical maximum.

Depending on the context, the ordering of the answers can be of substantial interest. For example, an instruction to order the drugs according to the frequency of consumption could have been included in the questionnaire. In this paper, however, I will not address the issue of ordered responses.

Of course there are other approaches to storing multiple responses besides the two that have been discussed here, e.g., composite string variables can be used or the data can be stored in long form. See Cox and Kohler (2003) for an overview and advice on how to convert the data into different structures.

### 3 Analyzing multiple responses

Because of the complex data structure even simple descriptive analyses of multiple response data like tabulating frequency distributions can be quite involved—especially if the data are stored according to the polytomous mode. A simple solution to the problem would be, of course, to tabulate the single variables on their own and count the frequencies together by hand. A more efficient approach is to transform the data to binary indicators (e.g. using `zb_qrm` by Eric Zbinden or `mr dum` by Lee Sieswerda, both available from the SSC Archive) and then use `tabstat` to tabulate the means of the indicator variables (see [R] `tabstat`). However, this approach is rather limited and still a lot of work. In addition, there are quite a few details which have to be taken into account while transforming the data and the whole process is vulnerable to mistakes.

To avoid having to figure out how to transform and analyze multiple responses over and over again, fairly general and easy-to-use commands should be available. Like there is, for example, the official `tabulate` to calculate frequency distributions of ordinary variables (see [R] `tabulate`), there should be a basic procedure to tabulate multiple responses. In the remainder of this paper I will thus present the results of some approaches to providing such general and user-friendly instruments.

#### 3.1 Tables

##### Syntax

###### *One-way tables*

```
mrtab varlist [weight] [if exp] [in range] [, poly response(numlist)
      countall include includemissing casewise title(string) width(#)
      abbrev nolabel nonames format(%fmt) integer sort[(#)] descending
      generate(prefix) nofreq ]
```

###### *Two-way tables*

```
mrtab varlist [weight] [if exp] [in range], by(varname) [ column row cell
  rcolumn rcell chi2 lrchi2 mtest[(method)] mlrchi2 wrap one-way-options
  ]
```

by ... : may be used with `mrtab`; see [R] `by`.

`fweights` and `aweights` are allowed with `mrtab`; see [U] **14.1.6 weight**.

### Description

`mrtab` tabulates multiple responses which are stored as a set of variables (*varlist*). `mrtab` can handle the two data storage modes explained above, that is, the indicator mode and the polytomous mode (see the `poly` option below). The multiple response variables have to be numeric in the first case and can be numeric or string in the latter (it is not possible to mix numeric and string variables, though). Note that the response variables, if numeric, should contain integer values only. To specify how the values of the variables be interpreted, use the `response()` option (see below).

`mrtab` will either display a one-way table of the unconditional distribution of the responses or if applying `by(varname)` a two-way table of the conditional distributions with respect to the groups defined by the values of *varname*. In the former case, counts, frequencies divided by the total of responses, as well as frequencies divided by the total of observations will be reported.<sup>2</sup> In the latter case, cell counts and optionally column, row, and/or cell percentages based on totals of responses and/or totals of observations are displayed (see the `column`, `row`, `cell`, `rcolumn`, and `rcell` options).

There are various possibilities to influence the appearance of the displayed table. For example the display formats of the cell statistics may be specified (`format()`, `integer`) and the width of the left stub of the table can be set (`width()`). It is also possible to influence the labeling (`no label`, `nonames`, `name()`) and to sort the response items in order of frequency (`sort`).

An important issue in calculating relative distributions of multiple responses is the determination of the correct denominator. Firstly, depending on the research question, one has to choose between totals of observations or totals of responses as indicated above (usually, it is proportions on the basis of observations one is interested in). Secondly and maybe less obviously, the sample of relevant observations has to be isolated depending on the nature of the data and the topic of research. The default in `mrtab` is to treat all observations containing at least one response as valid. This behavior can be changed to additionally accounting for cases with zero responses (see the `include` and `includemissing` options). Furthermore, one might want to consider cases with complete information only and neglect all cases with one or more missing values (see

<sup>2</sup>I will use the terms “observations”, “cases”, and “respondents” interchangeably in this paper to refer to the objects which data are gathered from, that is, the basic units of analysis (typically, each unit is represented by one row in the data matrix). In contrast, the term “response” refers to the single parts of an answer to a multiple response question (thus, for multiple response questions the total number of responses is typically greater than the total number of units of analysis).

the `casewise` option).

`mrtab` also provides (limited) support for significance tests in two-way tables. On the one hand, it is possible to perform a series of separate  $\chi^2$  tests for each response item and adjust the  $p$ -values to account for simultaneous testing (see the `mtest()` and `mlrchi2` options). On the other hand, overall  $\chi^2$  tests are available (`chi2` and `lrchi2`).

Irrespective of the data structure of the original variables, `mrtab` always transforms the data to binary indicator variables internally. It is possible to leave behind the generated indicators for further analysis (see the `generate()` option). The generated variables take on the values 1 or 0—with 1 indicating a response—for the sample of relevant cases (see above); non-relevant cases contain missing values (`.`) for these variables. Thus, using `mrtab` to produce indicator variables may also make sense if the data are in the indicator mode already, but are not 0/1 and/or should be harmonized to cover the same sub-sample of cases.

#### □ Technical Note

If the multiple response data are stored according to the polytomous mode and if the used variables are numeric, the information to label the rows (i.e. the response categories) is taken from the label definitions of the *first* variable. It is therefore important that the labels of the first variable are well defined. A good approach is, for example, to use just one set of label definitions which is attached to all variables used for a certain multiple response question.

#### Options

`abbrev` specifies that long response labels be abbreviated rather than wrapped.

`by(varname)` tabulates the distribution of responses against the categories of *varname* (two-way table). The `by`-variable may be string or numeric.

`casewise` specifies that observations with missing values for at least one of the response variables should be excluded listwise.

`cell` displays the relative frequency of each cell in a two-way table (base: total number of valid observations).

`chi2` requests the calculation of an overall Pearson's  $\chi^2$  statistic for the hypothesis that the distribution of response patterns is independent of the values of the `by`-variable. That is: A standard  $\chi^2$  test is applied to an expanded two-way table, where the rows represent unique combinations of responses. Note that the `chi2` option is not allowed if `aweights` are specified.

`column` displays in each cell of a two-way table the relative frequency of that cell within its column (base: column total of observations).

`countall` requests that repeated identical responses be added up (allowed only for polytomous response variables, see the `poly` option). By default, repeated identical

responses will only be counted once per observation. Notes: Significance tests may not be requested if `countall` is specified. Be careful with interpreting results that are labeled “percentage of cases”; though they reflect the mean number of responses per observation, they cannot be interpreted as proportions.

`descending` specifies that the sort order be descending. The default is to sort in ascending order. This is only relevant if the `sort` option is specified.

`format(%fmt)` specifies the display format for relative frequencies.

`generate(prefix)` creates a set of indicator variables reflecting the observed responses. The variables will be labeled and named according to the `prefix` provided. If the `name(string)` option is specified, the first eight characters of `string` are inserted into the variable labels. If the `chi2` and/or `lrchi2` options are specified, `generate` will additionally return a composite string variable, `prefixrp`, which reflects response patterns (each unique combination of responses is represented by a string of zeros and ones).

`include` specifies that observations composed of zero responses be treated as valid. Only cases with “real” missings (`.`, `.a`, `.b`, `.c`, ...) for all response variables will be excluded. Note that `include` will affect only the number of valid cases, i.e. both the absolute distribution of responses and the distribution relative to the total of responses will remain unchanged. In the case of string response variables, `include` specifies that cases with only empty strings (“”) be treated as valid.

`includemissing` is an enhancement to `include` and specifies that cases be treated as valid even if all response variables are missing. `includemissing` implies `include`. Specifying `includemissing` in connection with `casewise` has the effect that cases with missing values for at least one of the response variables will be treated as valid cases composed of zero responses.

`integer` specifies the display of frequencies as integers even if `aweights` are applied.

`lrchi2` requests the calculation of an overall likelihood-ratio  $\chi^2$  statistic (as an alternative to `chi2`). Note that the `lrchi2` option is not allowed if `aweights` are specified and that the statistic will not be calculated if there are empty cells.

`mtest[method]` requests the calculation of separate Pearson  $\chi^2$  statistics for each response category. That is, a test is carried out for each response category to establish whether the probability of observing the response depends on the values of the by-variable. Note that the `mtest` option is not allowed if `aweights` are specified. Multiple-test adjustments may be requested by specifying the method in parentheses. Currently available methods are `bonferroni`, `holm`, `sidak`, and `noadjust`. See the online help for `_mtest` for further information.

`mlrchi2` requests `mtest` to use the likelihood-ratio  $\chi^2$  statistics instead of Pearson’s  $\chi^2$ .

`nofreq` suppresses printing the frequencies (i.e., the whole frequency table will be suppressed unless `cell`, `column`, `row`, `rcell` or `rcolumn` is specified for two-way tables).

`nolabel` suppresses the printing of labels.



`nonames` suppresses the printing of variable names or category values in the left stub of the table, i.e. only the labels will be printed. This option has no effect if the responses are recorded by string variables and it is not allowed if the response variables are unlabeled or the `noLabel` option is specified.

`poly` specifies that the responses are stored according to the polytomous mode. If `poly` is not specified, `mrtab` assumes that the responses are stored according to the indicator mode. However, string response variables imply `poly`.

`rCell` displays the relative frequency of each cell in a two-way table (base: total number of responses).

`rColumn` displays in each cell of a two-way table the relative frequency of that cell within its column (base: column total of responses).

`response(numlist)` specifies the (range of) response values. If the data are stored according to the indicator mode, `response()` specifies the values which indicate a response to the item. `response()` defaults to 1 in this case. Note that the indicator variables do not necessarily have to be dichotomous since a list or range of values may be specified. If the data are stored according to the polytomous mode, `response()` specifies the list or range of responses that are to be tabulated. The default is to tabulate every value observed for the response variables (except for missing values). In the case of string variables, the `response()` option is obsolete.

`row` displays in each cell of a two-way table the relative frequency of that cell within its row (base: row total of responses; this is equal to the row total of observations unless `countall` is specified).

`title(string)` may be used to label the multiple response set. `string` will be printed at the head of the table.

`sort[(#)]` displays the table rows in ascending order of frequency. In the case of a two-way table the sorting will correspond to the row totals unless a reference column is specified in parentheses. That is, `sort(1)` will sort in order of the frequencies in the first column (first by-group), `sort(2)` in order of the frequencies in the second column, and so on. Specify the `descending` option to sort in descending order.

`width(#)` specifies the maximum width (number of characters) used to display the labels of the responses. Labels that are too wide are wrapped (or abbreviated if the `abbrev` option is specified). The default width is 30. The minimum width is 11.

`wrap` requests that no action is taken on wide two-way tables to make them readable. Unless `wrap` is specified, wide tables are broken into pieces to enhance readability.

### Saved Results

`mrtab` saves in `r()`:

Scalars			
<code>r(N)</code>	number of valid cases	<code>r(p)</code>	$p$ -value of the overall Pearson's $\chi^2$
<code>r(N_miss)</code>	number of missing cases	<code>r(chi2_lr)</code>	overall likelihood-ratio $\chi^2$ if <code>lrchi2</code> is specified
<code>r(r)</code>	number of response categories	<code>r(p_lr)</code>	$p$ -value of the overall likelihood-ratio $\chi^2$
<code>r(c)</code>	number of by-groups if <code>by()</code> is specified	<code>r(df)</code>	degrees of freedom of the overall $\chi^2$ tests
<code>r(chi2)</code>	overall Pearson's $\chi^2$ if <code>chi2</code> is specified		
Macros			
<code>r(list)</code>	list of the labels of the responses if available	<code>r(bylist)</code>	list of the labels of the by-groups if available
<code>r(mode)</code>	either <code>indicator</code> or <code>poly</code> depending on the mode of the multiple response variables	<code>r(bytype)</code>	either <code>numeric</code> or <code>string</code> depending on the storage type of the by-variable
<code>r(type)</code>	either <code>numeric</code> or <code>string</code> depending on the storage type of the multiple response variables		
Matrices			
<code>r(responses)</code>	frequencies of responses	<code>r(mchi2)</code>	Pearson's $\chi^2$ and (adjusted) $p$ -values of the separate tests if <code>mtest</code> is specified
<code>r(cases)</code>	cases in by-groups if <code>by()</code> is specified	<code>r(mchi2_lr)</code>	likelihood-ratio $\chi^2$ and (adjusted) $p$ -values of the separate tests if <code>mtest</code> and <code>mlrchi2</code> are specified

## 3.2 Graphs

### Syntax

```
mrgraph { bar | hbar | dot | tab } varlist [weight] [if exp] [in range] [, poly
  response(numlist) countall include includemissing casewise sort[(#)]
  descending by(varname[, by_subopts]) stat(statname) rtotal ctotal
  nopercent nolabel addval[(string)] width(#) height(#)
  oversubopts(over_subopts) graph_options ]
```

where `by_subopts` is `inboard`  
 or `outboard` [ `over_subopts` ]  
 or `separate` [ `suboptions` ]

and `statname` is { `freq` | `column` | `row` | `cell` | `rcolumn` | `rcell` }

`fweights` and `aweights` are allowed with `mrgraph`; see [U] 14.1.6 `weight`.

## Description

`mrgraph` is a utility to produce graphs of multiple response variables. The syntaxes `mrgraph bar`, `mrgraph hbar` and `mrgraph dot` are specified to indicate use of `graph bar`, `graph hbar` and `graph dot` respectively (see [G] `graph bar` and [G] `graph dot`). The size (height or length) of the bars or the position of the dots in the graph corresponds to the frequencies (or, optionally, proportions) of the responses. Thus, `mrgraph` works pretty much like Nick Cox's `catplot` (Cox 2004), only that the frequencies of the categories of a multiple response question are plotted instead of those of an ordinary categorical variable. Furthermore, `mrgraph tab` produces table plots in the manner of `tabplot` (Cox 2004). `mrgraph` is implemented as a wrapper for `mrtab` followed by `_mrvsmat` (see below) and `graph`.

## Options

`addval` [*string*] specifies that labels *and* values (or variable names in the case of the indicator mode) are used to mark the responses in the graph. The values and labels will be separated by *string* if specified or by a blank otherwise (use quotes, if the desired delimiter is supposed to have leading and/or trailing blanks, i.e. `addval(" : ")`). If the `addval` option is not specified, then labels are used exclusively. If no labels are available, however, values are used and `addval` will have no effect. Furthermore, `addval` will have no effect if the response variables are string.

`by`(*varname* [, *by\_subopts*]) draws the conditional distributions of responses for the categories of *varname*. The by-variable may be string or numeric. The *by\_subopts* control the grouping of the results in the graph if the graph type is `mrgraph bar`, `mrgraph hbar`, or `mrgraph dot`. Possible specifications are:

### inboard

The categories of the by-variable are grouped within the categories of the multiple response variables. This is the default.

### outboard [*over\_subopts* ]

The categories of the multiple response variables are grouped within the categories of the by-variable. The separation of the by-groups is implemented as an additional `over` statement in the internal `graph` call. Thus, *over\_subopts* may be specified. See [G] `graph bar` and [G] `graph dot`.

### separate [*suboptions* ]

For each category of the by-variable a separate plot is drawn within a single graph. This conforms to the default behavior of the `by` option in Stata's `graph` commands (which, however, is not the default in `mrgraph`). See [G] *by-option* for details on the *suboptions*.

`casewise` specifies that observations with missing values for at least one of the response variables should be excluded listwise.

**countall** requests that repeated identical responses be added up. See section 3.1 for details on this option.

**ctotal** specifies that column totals be reported.

**descending** specifies that the sort order be descending. The default is to sort in ascending order. This is only relevant if the **sort** option is specified.

**height(#)** controls the amount of available graph space taken up by bars in the table plot. This option is only relevant if the graph type is **mrgraph tab**. The default is 0.8.

**include** specifies that observations composed of zero responses be treated as valid. See section 3.1 for details on this option.

**includemissing** is an enhancement to **include** and specifies that cases be treated as valid even if all response variables are missing. See section 3.1 for details on this option.

**no label** specifies that labels be ignored.

**nopercent** specifies that relative frequencies be formatted as proportions (e.g. “.271”) instead of percentages (e.g. “27.1”).

**oversubopts**(*over\_subopts*) may be used to pass through suboptions to the **over** option which is applied by **mrgraph** in the internal call of the **graph** command. This is only relevant for the graph types **mrgraph bar**, **mrgraph hbar**, and **mrgraph dot**. For further explanations on the **over** option and its suboptions see [G] **graph bar** and [G] **graph dot**. Do not use the **sort** suboption; use **mrgraph**’s own **sort** option instead (see below).

**poly** specifies that the responses are stored in polytomous mode. See section 3.1 for details on this option.

**response(numlist)** specifies the (range of) response values. See section 3.1 for details on this option.

**rtotal** specifies that row totals be reported.

**sort[(#)]** draws the categories in ascending order of frequency. If the **by()** option is specified, the sorting will correspond to the totals over all groups unless a reference group is specified in parentheses. That is, **sort(1)** will sort in order of the frequencies in the first by-group, **sort(2)** in order of the frequencies in the second by-group, and so on. Specify the **descending** option to sort in descending order.

**stat(statname)** determines the statistic which the graph be based on. *statname* is either **freq**, if raw frequencies be used, or **column** (base: column total of observations), **row** (base: row total), **cell** (base: grand total of valid observations), **rcolumn** (base: column total of responses), or **rcell** (base: grand total of responses), if relative frequencies be used. **stat(freq)** is the default.

**width(#)** specifies the maximum width (number of chars) used to draw the labels of

the responses. Labels that are too wide are wrapped. Note that the single words in the labels will not be broken. If no `width` is specified, labels are not wrapped.

### 3.3 From tables to datasets

#### Syntax

```
_mrsvmat [ , stat(statname) rtotal ctotal nopercent nolabel clear ]
```

where *statname* is { `freq` | `column` | `row` | `cell` | `r`column | `r`cell }

#### Description

`_mrsvmat` is a low-level utility which may be used after `mrtab` to prepare a data matrix for creating graphs of the tabulated distribution. Do not use `_mrsvmat` unless you are confident that `mrgraph` does not meet your needs. It is important to understand, that `_mrsvmat` will destroy the data in memory. Therefore, it should always be used in connection with `preserve` (see [R] `preserve`). `_mrsvmat` will replace the data in memory with a data matrix created from the results left behind by `mrtab`. Each row of the new data will represent one row of the table displayed by `mrtab`, that is, each row will represent one response category. Additionally, a row holding column totals will be added if the `ctotal` option is specified. `_mrsvmat` will generate the following variables:

Variable	Description
R	Values of the response categories or names of the variables representing the response items depending on the structure of the original data.
L	Labels of the response categories. Variable L will be suppressed if either no labels are found or the <code>nolabel</code> option is specified.
C1, C2, ...	Frequencies of responses. In the case of two-way tables, each variable represents one column of the multiple response table. In the case of a one-way table, just one variable, C1, will be created. Depending on the users' choice, the data cells will either contain raw counts or frequencies relative to a specified base (see the <code>stat()</code> option).
T	Row totals if option <code>rtotal</code> is specified, suppressed otherwise.

#### Options

`clear` allows `_mrsvmat` to clear the data in memory without asking for confirmation.

`ctotal` specifies that column totals be saved.

`nolabel` specifies that labels be ignored.

`nopercent` specifies that relative frequencies be formatted as proportions (e.g. “.271”)

instead of percentages (e.g. “27.1”).

`rtotal` specifies that row totals be saved.

`stat(statname)` determines the statistic to be saved. *statname* is either `freq` for raw frequencies or `column` (base: column total of observations), `row` (base: row total), `cell` (base: grand total of valid observations), `rcolumn` (base: column total of responses), or `rcell` (base: grand total of responses) for relative frequencies. `stat(freq)` is the default.

## 4 Remarks and examples

In order to illustrate the usage of the multiple response commands I will present analyses of data collected by Braun et al. (2001). The respondents in this study were drug addicts in three major cities in Switzerland in 1997.

### 4.1 One-way tables

#### The indicator mode

The respondents in the study by Braun et al. (2001) were asked to indicate their sources of income in the last three months. A list of possible sources was provided in the questionnaire. Braun et al. recorded the data with a set of 0/1-variables, one for each income source:

```
. use drugs.dta
(1997 Survey Data on Swiss Drug Addicts)
. describe inco1-inco7
```

variable name	storage type	display format	value label	variable label
inco1	byte	%8.0g	yesno	private support (partner, family, friends)
inco2	byte	%8.0g	yesno	public support (unemployment insurance, social benefits)
inco3	byte	%8.0g	yesno	drug dealing
inco4	byte	%8.0g	yesno	housebreaking, theft, robbery
inco5	byte	%8.0g	yesno	prostitution
inco6	byte	%8.0g	yesno	"mischeln"/begging
inco7	byte	%8.0g	yesno	legal occupation

Note that “mischeln” is a Swiss German word and describes a form of begging in which the beggar actively approaches people for money (i.e. walks up to them and asks them for money).

In the first step of the data analysis one would probably like to tabulate the sample distribution of the income sources. However, because the information is stored across several variables, the `tabulate` command is not very convenient: each of the indicator variables would have to be tabulated separately. Because the variables are 0/1 in the

present case, we could, however, use the `tabstat` command to tabulate the sums and the proportions of the responses:

```
. tabstat inco1-inco7, s(sum mean) c(s)
```

variable	sum	mean
inco1	226	.2325103
inco2	607	.6244856
inco3	293	.3014403
inco4	50	.0514403
inco5	82	.0843621
inco6	151	.1553498
inco7	352	.3621399

We see, for example, that 36% of the respondents ticked the 7th income source (legal occupation). Unfortunately, the `tabstat`'s output is rather meagre. A more informative output can, however, be obtained by using the new `mrtab` command:

```
. mrtab inco1-inco7, sort title(Sources of income)
```

	Sources of income	Frequency	Percent of responses	Percent of cases
inco4	housebreaking, theft, robbery	50	2.84	5.19
inco5	prostitution	82	4.66	8.52
inco6	"mischeln"/begging	151	8.57	15.68
inco1	private support (partner, family, friends)	226	12.83	23.47
inco3	drug dealing	293	16.64	30.43
inco7	legal occupation	352	19.99	36.55
inco2	public support (unemployment insurance, social benefits)	607	34.47	63.03
	Total	1761	100.00	182.87
Valid cases:	963			
Missing cases:	9			

The different response categories are nicely labeled and it is possible, for instance, to sort the categories in order of frequency. Therefore, it is immediately evident from the table above that "public support" is the source of income which has been named the most often, whereas "housebreaking, theft, robbery" is the least frequent source of income.

In addition to percentages on the basis of observations (i.e. cases) `mrtab` also reports percentages with respect to the overall sum of responses. For example, 35% of all responses are "public support". Furthermore, `mrtab` also provides a row reporting the totals over all response categories. Note that the total printed in the "percentage of cases" column reflects the average number of responses per subject (multiplied by 100). Thus, the mean number of sources of income is approximately 1.8 in this study.

#### □ Technical Note

Note that in the above example the percentages reported by `mrtab` slightly differ from the means calculated by `tabstat`. The reason for the difference is that `mrtab` excluded 9 observations for which no sources of income were recorded. In such cases it is often not clear a priori whether the respondent refused to answer the question (and thus should be excluded from analysis) or whether the correct response is simply “none” (in which case the observation should be accounted for). Because the determination of the correct denominator depends on the context (that is, e.g., the specific construction of the question, the topic, how the data have been recorded, etc.) `mrtab` provides specific options to determine the sample of valid observations (`include`, `includemissing`, `casewise`). In the context of the present example it, in fact, seems reasonable to include the observations with zero responses because the given list of seven different income types is probably not exhaustive. We, thus, should have specified:

```
. mrtab inco1-inco7, include sort title(Sources of income)
```

	Sources of income	Frequency	Percent of responses	Percent of cases
inco4	housebreaking, theft, robbery	50	2.84	5.14
inco5	prostitution	82	4.66	8.44
inco6	"mischeln"/begging	151	8.57	15.53
inco1	private support (partner, family, friends)	226	12.83	23.25
inco3	drug dealing	293	16.64	30.14
inco7	legal occupation	352	19.99	36.21
inco2	public support (unemployment insurance, social benefits)	607	34.47	62.45
	Total	1761	100.00	181.17
Valid cases:	972			
Missing cases:	0			

#### □ Technical Note

Although the variables in the indicator mode are interpreted in a dichotomous manner, they do not necessarily have to be 0/1 nor dichotomous in general. Basically, all kinds of variables are allowed as long as they are integer. By default, `mrtab` interprets the value “1” as a response and all the other values (e.g. “0”) as “no response”. However, the default response value may be changed via the `response(numlist)` option. If specified, any value of `numlist` will be considered as indicating a response.

The respondents in the drug study were also asked about their experiences with crime. (“Have you been involved in the following offences ... during the last 12 months: hit someone; use a weapon against someone; sexual harassment, rape; robbery; blackmail?”) For each of the different types of crimes a variable like the following has been recorded:

```
. codebook crime1
```

crime1	hit someone



```

type: numeric (byte)
label: crime
range: [0,3]
unique values: 4
units: 1
missing .: 65/972

tabulation: Freq.  Numeric  Label
              716      0      no
              62       1  yes, as committer
              97       2  yes, as victim
              32       3  yes, both
              65       .

```

As indicated by the value labels, it is probably sensible to distinguish between criminal experiences in the role of a committer and criminal experiences in the role of the victim. Thus, with the help of the `response()` option, we derive the following results:

```

. mrtab crime1-crime5, include response(1 3) title(Crime (as committer))
  nonames

```

Crime (as committer)	Frequency	Percent of responses	Percent of cases
hit someone	94	54.02	10.36
use a weapon against someone	20	11.49	2.21
sexual harassment, rape	1	0.57	0.11
robbery (including drug theft)	51	29.31	5.62
blackmail	8	4.60	0.88
Total	174	100.00	19.18

```

Valid cases:      907
Missing cases:    65

```

```

. mrtab crime1-crime5, include response(2 3) title(Crime (as victim)) nonames

```

Crime (as victim)	Frequency	Percent of responses	Percent of cases
hit someone	129	41.08	14.22
use a weapon against someone	27	8.60	2.98
sexual harassment, rape	31	9.87	3.42
robbery (including drug theft)	99	31.53	10.92
blackmail	28	8.92	3.09
Total	314	100.00	34.62

```

Valid cases:      907
Missing cases:    65

```

The first part of the output reports the frequencies of criminal experiences as a committer (or committer *and* victim), the second the frequencies of experiences as a victim (or victim *and* committer). Apparently, the respondents reported that they had been a victim of a crime considerably more often than they had committed a crime themselves. Thus, either the sample is selective, the respondents did not report all the crimes they committed, or there are a few subjects committing many crimes repeatedly. Note that 65 respondents with missing information for this question were excluded.

### The polytomous mode

Sometimes, if the list of possible response categories is open, for instance, it is convenient to store the answers to multiple response questions as a set of polytomous variables. For example, the answers to the question on income sources of drug addicts in Switzerland could have been recorded in polytomous mode rather than using a set of indicator variables. The maximum number of possible responses per observation is seven in this specific case because the list of response categories is limited to seven different income sources. However, the realized maximum of responses is only six. Thus, six polytomous response variables are required. The following output shows one of them.<sup>3</sup>

```
. codebook pinco1
```

---

```
pinco1 (unlabeled)
```

---

```

      type: numeric (byte)
      label: inco
      range: [-1,7]
unique values: 8
      units: 1
      missing .: 0/972
      tabulation: Freq.  Numeric  Label
                   9         -1   no further sources
                   98         1   private support (partner,
                                family, friends)
                   373        2   public support (unemployment
                                insurance, social benefits)
                   138        3   drug dealing
                   17         4   housebreaking, theft, robbery
                   38         5   prostitution
                   74         6   "mischeln"/begging
                   225        7   legal occupation
```

Multiple responses, which are stored in such a way, cannot be tabulated using `tabstat` because the information on the single response categories is spread out over several variables. The data would have to be transformed to indicator mode beforehand, which is quite a complicated task. However, polytomous multiple response variables can be tabulated by `mrtab` if the `poly` option is specified:

```
. mrtab pinco1-pinco6, poly response(1/7) include sort abbrev
```

	Frequency	Percent of responses	Percent of cases
4 housebreaking, theft, robbery	50	2.84	5.14
5 prostitution	82	4.66	8.44
6 "mischeln"/begging	151	8.57	15.53
1 private support (partner, fami	226	12.83	23.25
3 drug dealing	293	16.64	30.14
7 legal occupation	352	19.99	36.21
2 public support (unemployment i	607	34.47	62.45

<sup>3</sup>The data have been generated by saving each respondent's answers in random order. The shuffling of the answers was done by the `sortlistby2` command, which is available from the SSC Archive (type `ssc describe sortlistby`).

```

Total |      1761      100.00      181.17
Valid cases:      972
Missing cases:      0

```

Furthermore, string variables can also be tabulated by `mrtab` as is shown in the following example. Note that it is not necessary to specify the `poly` option in the case of string variables.

```

. codebook sinco1
-----
sinco1                                     (unlabeled)
-----
type: string (str56)
unique values: 7                          missing "": 9/972
tabulation: Freq. Value
              9 ""
              74 "mischeln"/begging"
             138 "drug dealing"
              17 "housebreaking, theft, robbery"
             225 "legal occupation"
              98 "private support (partner, family,
                 friends)"
              38 "prostitution"
             373 "public support (unemployment insurance,
                 social benefits)"

warning: variable has embedded blanks
. mrtab sinco1-sinco6, include sort abbrev
-----
                                     Frequency   Percent of   Percent
                                     responses     of cases
-----
housebreaking, theft, robbery         50           2.84         5.14
prostitution                          82           4.66         8.44
"mischeln"/begging                   151          8.57        15.53
private support (partner, fami       226          12.83        23.25
drug dealing                         293          16.64        30.14
legal occupation                     352          19.99        36.21
public support (unemployment i       607          34.47        62.45
-----
Total                                 1761         100.00       181.17

Valid cases:      972
Missing cases:      0

```

#### □ Technical Note

In the case of “half-open” multiple response questions a mixed storage design is sometimes applied (half-open means that some categories are spelled out and can be ticked, but there are also some empty lines which can be filled with further answers). That is, the first part of the answers is stored in indicator mode (the tickable items) while the rest is held in polytomous mode. It may be convenient to use `mrtab`’s `generate()` option in such a case to transform the polytomous part to indicator mode.

### Significance tests

The test for equality of proportions in matched samples proposed by Cochran (1950) may be applied to evaluate the significance of the differences among the proportions of the single response categories. If  $c$  is the number of response categories,  $T_j$  the number of responses in the  $j$ th category,  $\bar{T}$  the mean number of responses per category, and  $u_i$  the number of responses in the  $i$ th observation, the test statistic proposed by Cochran is then defined as

$$Q = \frac{c(c-1) \sum_j (T_j - \bar{T})^2}{c \sum_i u_i - \sum_i u_i^2}$$

For large samples,  $Q$  is  $\chi^2$ -distributed with  $(c-1)$  degrees of freedom. Note that in the case of only two response categories the Cochran test is equal to the McNemar  $\chi^2$  test implemented in `mcc` (see [ST] `epitab`).

In our example on income sources, the differences in the proportions of the various responses are highly significant (this is not surprising since the proportions range from 5% to 62%!):

```
. mrtab pinco1-pinco6, poly response(1/7) nofreq include generate(_inco)
. cochran _inco*
Test for equality of proportions in matched samples (Cochran's Q):
Number of obs      =      972
Cochran's chi2(6)  =  1123.529
Prob > chi2        =    0.0000
. drop _inco*
```

Note that the `cochran` command is not part of the official release of Stata. However, it can be obtained from the SSC Archive (type `ssc install cochran`).

### □ Technical Note

An alternative approach to testing for differences in the proportions would be to transform the data to long format (see [R] `reshape`) and then estimate, for example, a logistic regression including dummy variables for the different response categories and correcting the standard errors for the clustering on observations:

```
. preserve
. mrtab pinco1-pinco6, poly response(1/7) nofreq include generate(_inco)
. reshape long _inco, i(id) j(R)
(output omitted)
. xi: logit _inco i.R, cluster(id) nolog
i.R          _IR_1-7          (naturally coded; _IR_1 omitted)
Logit estimates                Number of obs   =      6804
                                Wald chi2(6)    =      895.23
                                Prob > chi2    =      0.0000
Log pseudo-likelihood = -3299.692          Pseudo R2   =      0.1519
                                (standard errors adjusted for clustering on id)
```

<code>_inco</code>	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
<code>_IR_2</code>	1.702822	.1020924	16.68	0.000	1.502725	1.902919
<code>_IR_3</code>	.3537421	.1016399	3.48	0.001	.1545315	.5529527
<code>_IR_4</code>	-1.720332	.1598888	-10.76	0.000	-2.033708	-1.406955
<code>_IR_5</code>	-1.190312	.137136	-8.68	0.000	-1.459093	-.9215301
<code>_IR_6</code>	-.4990527	.1091339	-4.57	0.000	-.7129512	-.2851541
<code>_IR_7</code>	.6281023	.1018616	6.17	0.000	.4284573	.8277473
<code>_cons</code>	-1.194191	.0759684	-15.72	0.000	-1.343086	-1.045295

```
. restore
```

The null hypothesis of equal proportions is rejected if the overall Wald  $\chi^2$  of the model is significant (which is obviously the case here). Note that the choice of the logistic regression model is not crucial here. We might as well use Probit or even linear regression. The only really important thing is to take account of the clustering (an issue which is sometimes referred to as “matched samples”). We might also think of applying panel models, e.g. `xtlogit` or `xtreg` (see [XT] `xt`; the tests based on panel models will usually be more efficient).

## 4.2 Two-way tables

In most applications not only the marginal distribution of the responses is of interest, but also the conditional distributions with regard to the values of some other variable. For example, it might be interesting to examine the differences among the drug scenes in the three cities covered in the study by Braun et al. (2001). Again considering the income sources of the drug addicts we could use `mrtab` to produce the following two-way table:

```
. mrtab inco1-inco7, include sort title(Sources of income) nonames
      width(28) by(city) column
```

Sources of income	City in which the interview was conducted			Total
	Basel	Bern	Zurich	
housebreaking, theft, robbery	13 3.74	23 7.99	14 4.17	50 5.14
prostitution	20 5.75	38 13.19	24 7.14	82 8.44
"mischeln"/begging	43 12.36	62 21.53	46 13.69	151 15.53
private support (partner,	92	71	63	226

*Tabulation of multiple responses*

family, friends)	26.44	24.65	18.75	23.25
drug dealing	102 29.31	101 35.07	90 26.79	293 30.14
legal occupation	132 37.93	103 35.76	117 34.82	352 36.21
public support (unemployment insurance, social benefits)	205 58.91	172 59.72	230 68.45	607 62.45
Total	607 174.43	570 197.92	584 173.81	1761 181.17
Cases	348	288	336	972

Valid cases: 972  
Missing cases: 0

It is striking that the drug scene in Bern is quite different from the scenes in the two other cities. The Bern scene seems to be much more criminal: the measured rates of theft/robbery, prostitution, and drug dealing are clearly highest. In addition, the proportion of beggars is highest in the Bern sample. Another substantial difference in the table is that Zurich has a relatively high rate of drug addicts who live off public support, whereas the proportion of respondents receiving private support is relatively low. Possibly, there is a tradeoff between public and private support.

**Significance tests**

In order to assess whether the observed differences between the cities are significant or not, we can, for example, conduct an overall Pearson or likelihood-ratio  $\chi^2$  test which is based on an expanded table of the frequencies of response “patterns” by cities. Applied to the above table, the results of the tests are:

```
. mrtab inco1-inco7, include by(city) nofreq chi2 lrchi2 generate(_inco)
Overall Test(s) of Significance:
      Pearson chi2(150) = 210.0806   Pr = 0.001
likelihood-ratio chi2(150) = .
```

The Pearson  $\chi^2$  test indicates that the differences are indeed significant (assuming a 5% significance level). But what about the likelihood-ratio test? Apparently it could not be evaluated because of the occurrence of empty cells. The problem is that the test is based on a variable identifying response “patterns”. In the example above the patterns variable looks as follows (first six cases only):

```
. list _incorp in 1/6, clean
      _incorp
1. 0000010
2. 0100000
3. 0000010
4. 0100000
5. 0000001
```

```
6. 1100000
. drop _inco*
```

That is, a response pattern is a composite string of zeros and ones with ones indicating a response. For instance, respondent no. 1's source of income is a legal occupation, respondent no. 6's sources are theft and prostitution, etc. As the number of response categories increases, the number of unique response patterns grows exponentially. In the present case of seven categories there are  $2^7 = 128$  unique patterns. This means that the table of the patterns variable against the three cities potentially has  $128 \cdot 3 = 384$  cells. Quite a few observations are needed to fill all these cells. The overall test, thus, is not very powerful (many degrees of freedom) and is only suited if there are just very few categories.

A better and probably more informative strategy is to test each row of the table separately. This can be done using standard  $\chi^2$  or likelihood-ratio tests for two-way tables (in fact, each row of a two-way multiple response table is nothing else than a dense version of a  $(2 \times k)$ -table of a specific response indicator against the by-variable). Use `mrtab`'s `mtest` option to add a column to the multiple response table reporting the results of the row-specific tests:

```
. mrtab inco1-inco7, include sort title(Sources of income) nonames
width(19) by(city) column mtest(bonferroni)
```

Key					
<i>frequency of responses</i>					
<i>column percent of cases</i>					
Sources of income	City in which the interview was conducted			Total	chi2/p*
	Basel	Bern	Zurich		
housebreaking, theft, robbery	13 3.74	23 7.99	14 4.17	50 5.14	6.840 0.229
prostitution	20 5.75	38 13.19	24 7.14	82 8.44	12.427 0.014
"mischeln"/begging	43 12.36	62 21.53	46 13.69	151 15.53	11.433 0.023
private support (partner, family, friends)	92 26.44	71 24.65	63 18.75	226 23.25	6.111 0.330
drug dealing	102 29.31	101 35.07	90 26.79	293 30.14	5.232 0.512
legal occupation	132 37.93	103 35.76	117 34.82	352 36.21	0.751 1.000
public support (unemployment)	205 58.91	172 59.72	230 68.45	607 62.45	7.938 0.132

insurance, social benefits)				
Total	607	570	584	1761
	174.43	197.92	173.81	181.17
Cases	348	288	336	972

\* Pearson  $\chi^2(2)$  / Bonferroni adjusted  $p$ -values

Valid cases: 972  
Missing cases: 0

Note that in the example the  $p$ -values have been adjusted for the fact that a series of tests is performed (see, e.g. Wright 1992). There are several adjustment methods available (`bonferroni`, `holm`, `sidak`; see the online help for `mtest` for formulas and further details). The results above indicate that the differences among the three cities concerning prostitution and begging are significant.

#### □ Technical Note

An alternative and more flexible strategy would again be to reshape the data and apply regression models. Firstly, dummy variables for the different response categories need to be introduced to the model to account for the different levels of prevalence. To evaluate the overall significance of the differences among the response distributions in the three cities, we then include dummy variables for the cities, as well as for all interactions between cities and response categories, and perform a joint test for all these parameters:

```
. preserve
. mrtab inco1-inco7, include nofreq generate(_inco)
. reshape long _inco, i(id) j(R)
. (output omitted)
. xi: logit _inco i.R*i.city, cluster(id)
. (output omitted)
. unab I : _Icity* _IRX*
. test 'I'
( 1) _Icity_2 = 0
( 2) _Icity_3 = 0
( 3) _IRXcit_2_2 = 0
( 4) _IRXcit_2_3 = 0
( 5) _IRXcit_3_2 = 0
( 6) _IRXcit_3_3 = 0
( 7) _IRXcit_4_2 = 0
( 8) _IRXcit_4_3 = 0
( 9) _IRXcit_5_2 = 0
(10) _IRXcit_5_3 = 0
(11) _IRXcit_6_2 = 0
(12) _IRXcit_6_3 = 0
(13) _IRXcit_7_2 = 0
(14) _IRXcit_7_3 = 0
      chi2( 14) = 43.12
      Prob > chi2 = 0.0001
```



```
. restore
```

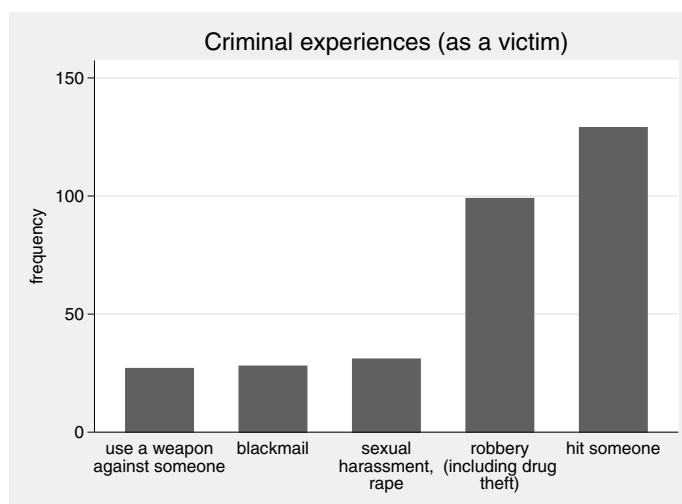
Again, the choice of a specific model is not so critical as long as the specification is correct and the clustering is accounted for. Very similar results are obtained, for example, using `xtlogit` (see [ST] `xtlogit`):

```
. quietly xi: xtlogit _inco i.R*i.city, re i(id)
. estimates store A
. quietly xi: xtlogit _inco i.R, re i(id)
. lrtest A
(log-likelihoods of null models cannot be compared)
likelihood-ratio test          LR chi2(14) =    49.23
(Assumption: . nested in A)   Prob > chi2 =    0.0000
```

### 4.3 Graphs

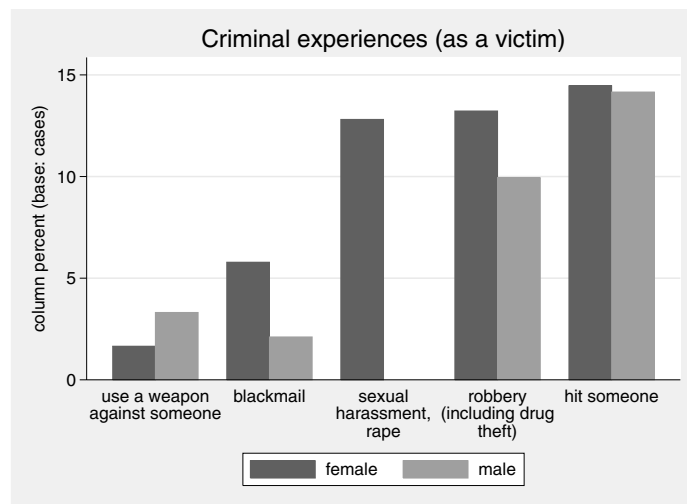
A convenient way to illustrate multiple response distributions are bar charts whereby the sizes of the bars represent the frequencies of the response categories. For example, the frequencies of criminal experiences as a victim (see section 4.1) could be plotted as follows:

```
. mrgraph bar crime1-crime5, include response(2 3) sort width(15)
      title(Criminal experiences (as a victim)) ylabel(,angle(0))
```



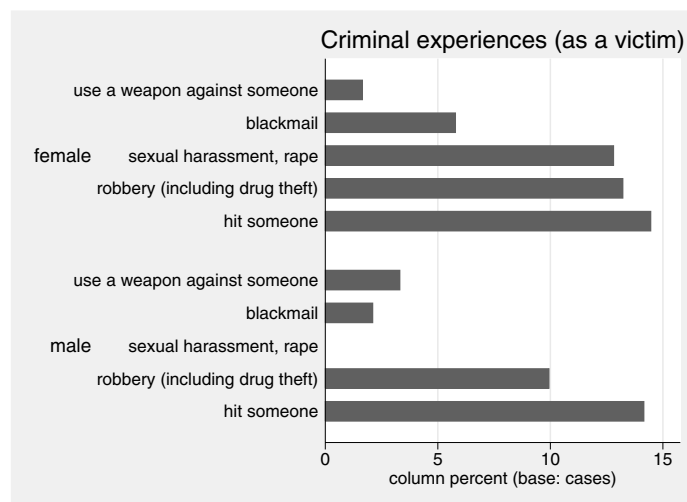
If the frequencies of two-way multiple response tables be plotted, e.g. criminal experiences by sex, the grouping of the bars becomes relevant. Thus, the `by()` option of the `mrgraph` command comes in different flavors. The default is to group the categories of the `by`-variable within the response categories as is illustrated in the following example:

```
. mrgraph bar crime1-crime5, include response(2 3) sort width(15) by(sex)
  stat(column) title(Criminal experiences (as a victim))
  ylabel(,angle(0)) legend(bmargin(t+1))
```



The advantage of this illustration is that the differences between the by-groups can be seen immediately for each response category. However, it might sometimes be preferable to separate the conditional distributions and thus to group the response categories within by-categories:

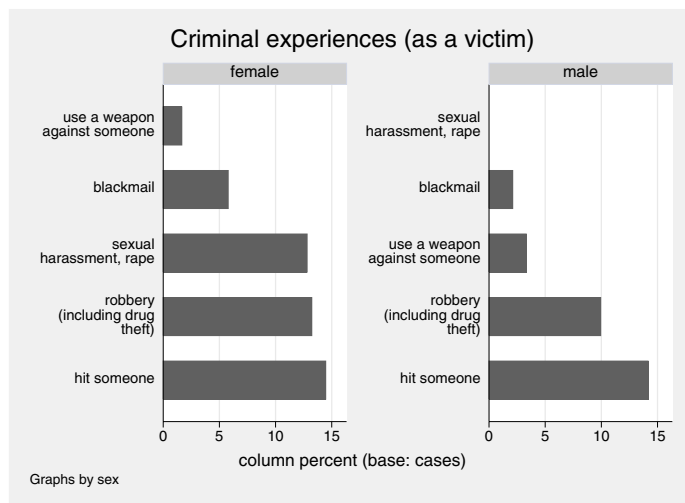
```
. mrgraph hbar crime1-crime5, include response(2 3) sort by(sex, outboard)
  stat(column) title(Criminal experiences (as a victim)) ylabel(,angle(0))
```



It is also possible to completely separate the conditional distributions and display them as several plots within the same graph. A nice feature of this procedure is that it

is possible to sort the separate distributions individually and thus illustrate differences in the ordering of the frequencies:<sup>4</sup>

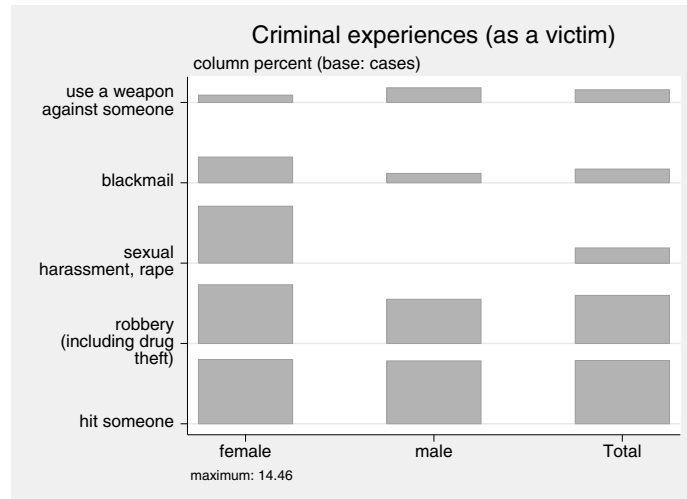
```
. mrgraph hbar crime1-crime5, include response(2 3) sort(1) width(16)
    stat(column) by(sex, separate title(Criminal experiences (as a victim)))
```



Last but not least, a very straightforward method to illustrate a two-way multiple response table is to construct a “table plot” as proposed by Cox (2004) for ordinary two-way tables:

```
. mrgraph tab crime1-crime5, include response(2 3) sort width(16)
    stat(column) by(sex) rtotal title(Criminal experiences (as a victim))
```

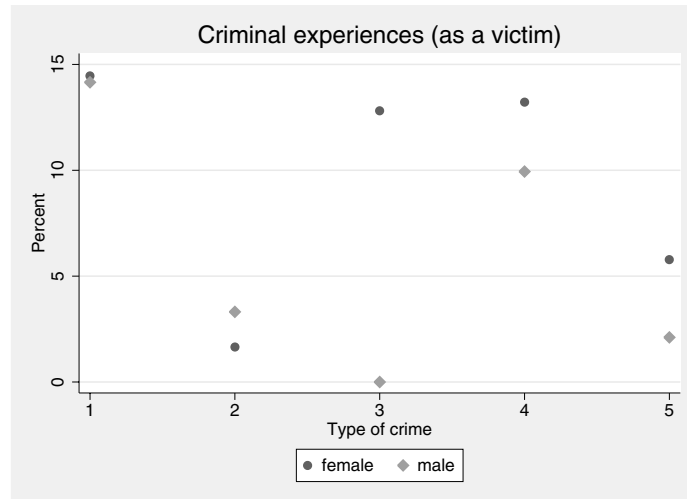
<sup>4</sup>Note that it is important to specify `sort(1)` here to achieve the individual sorting. Specifying the `sort` option without argument would sort in order of the unconditional distribution.



#### □ Technical Note

The `mrgraph` command is implemented as a wrapper for `mrtab` followed by the low-level utility `_mrsvmat`. The purpose of `_mrsvmat` is to pick up the results calculated by `mrtab` and compose a data set representing the results (note that the data in memory will be lost). It can be helpful to use the `_mrsvmat` command manually to produce graphs which are not supported by `mrgraph`. The following piece of output provides an example:

```
. mrtab crime1-crime5, include response(2 3) by(sex) nofreq
. preserve
. _mrsvmat, stat(column) cttotal rttotal clear
. list, clean string(20)
      R   L                C1      C2      T
1.  crime1  hit someone    14.46281  14.15663  14.23841
2.  crime2  use a weapon against..  1.652893  3.313253  2.869757
3.  crime3  sexual harassment, r..  12.80992      0  3.421633
4.  crime4  robbery (including d..  13.22314  9.939759  10.81678
5.  crime5  blackmail       5.785124  2.108434  3.090508
6.      T  Total           47.93388  29.51807  34.43708
. generate x=_n
. scatter C1 C2 x in 1/5, xtitle(Type of crime) ytitle(Percent)
      title(Criminal experiences (as a victim))
```



Note that the `_mrvsmat` utility can also be useful for exporting the results of `mrtab`. For example, use Roger Newson's `listtex` command after having applied `_mrvsmat` in order to create a  $\text{\LaTeX}$  table (see Newson 2003).

## 5 References

- Braun, N., B. Nydegger Lory, R. Berger, and C. Zahner. 2001. *Illegale Märkte für Heroin und Kokain*. Bern: Haupt.
- Cochran, W. G. 1950. The Comparison of Percentages in Matched Samples. *Biometrika* 37(3/4): 256–266.
- Cox, N. J. 2004. Speaking Stata: Graphing categorical and compositional data. *The Stata Journal* 4(2): 190–215.
- Cox, N. J. and U. Kohler. 2003. Stata FAQ on “Dealing with multiple responses”. <http://www.stata.com/support/faqs/data/multresp.html>.
- Newson, R. 2003. Confidence intervals and p-values for delivery to the end user. *The Stata Journal* 3(3): 245–269.
- Wright, S. P. 1992. Adjusted  $P$ -Values for Simultaneous Inference. *Biometrics* 48(4): 1005–1013.

## 6 Acknowledgements

The `mrtab` command is coauthored by Hilde Schaeper (HIS Hannover, Germany). I am greatly indebted to her for starting the project and taking me aboard. I would like to thank her and her colleagues at HIS for many important suggestions and for debugging.

Furthermore, I would like to thank Elisabeth Coutts (ETH Zurich), Jens Lauritsen and Christian Færgemann (Odense Universitetshospital), Sandro Leidi (University of Reading), and Eric Zbinden (University of Geneva) for helpful comments and suggestions.

**About the Author**

Ben Jann (jann@soz.gess.ethz.ch) is research assistant at the Department of Sociology of the Swiss Federal Institute of Technology Zurich (ETH) and a Ph.D. candidate at the University of Bern in Switzerland.