

**StatSoft**

**Business White Paper**

## **Categorization in *STATISTICA* Frequency Tables**

Last Update: 2009

---

**U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com**

Australia: StatSoft Pacific Pty Ltd.  
Brazil: StatSoft South America  
Bulgaria: StatSoft Bulgaria Ltd.  
Czech Rep.: StatSoft Czech Rep. s.r.o  
China: StatSoft China

France: StatSoft France  
Germany: StatSoft GmbH  
Hungary: StatSoft Hungary Ltd.  
India: StatSoft India Pvt. Ltd.  
Israel: StatSoft Israel Ltd.

Italy: StatSoft Italia srl  
Japan: StatSoft Japan Inc.  
Korea: StatSoft Korea  
Netherlands: StatSoft Benelux BV  
Norway: StatSoft Norway AS

Poland: StatSoft Polska Sp. z.o.o.  
Portugal: StatSoft Ibérica Lda  
Russia: StatSoft Russia  
Spain: StatSoft Ibérica Lda

S. Africa: StatSoft S. Africa (Pty) Ltd.  
Sweden: StatSoft Scandinavia AB  
Taiwan: StatSoft Taiwan  
UK: StatSoft Ltd.

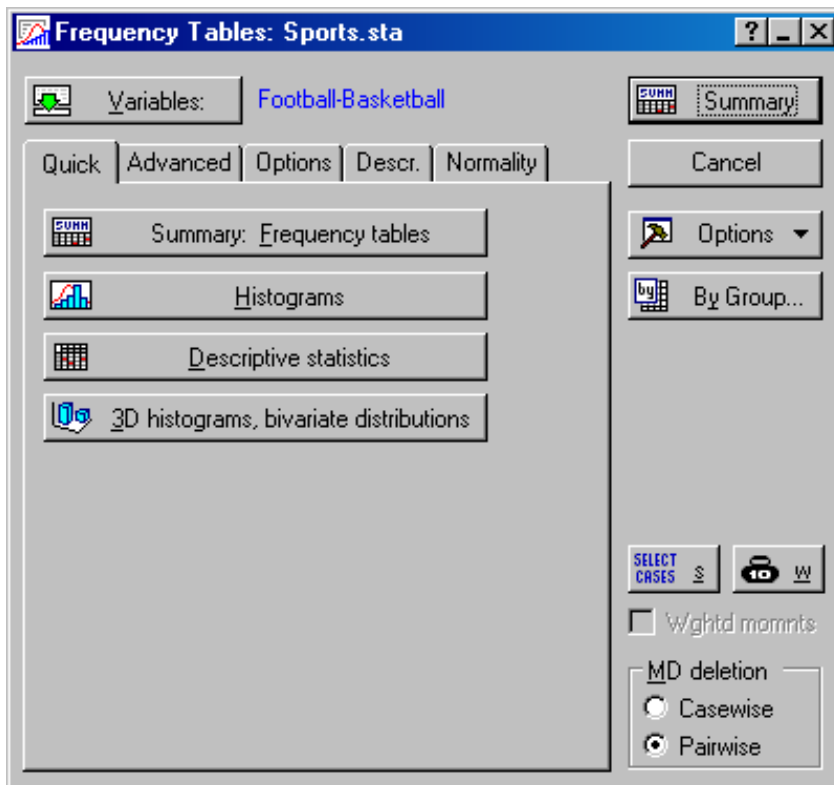
## Categorization in *STATISTICA* Frequency Tables

Frequency tables represent the simplest method for analyzing categorical data. They are often used as an exploratory procedure to review how different categories of values are distributed in a sample.

### Overview

To access *STATISTICA* frequency table options, open a data set (e.g., *Sports.sta*) and select *Frequency Tables* from *Statistics > Basic Statistics > Frequency Tables*.

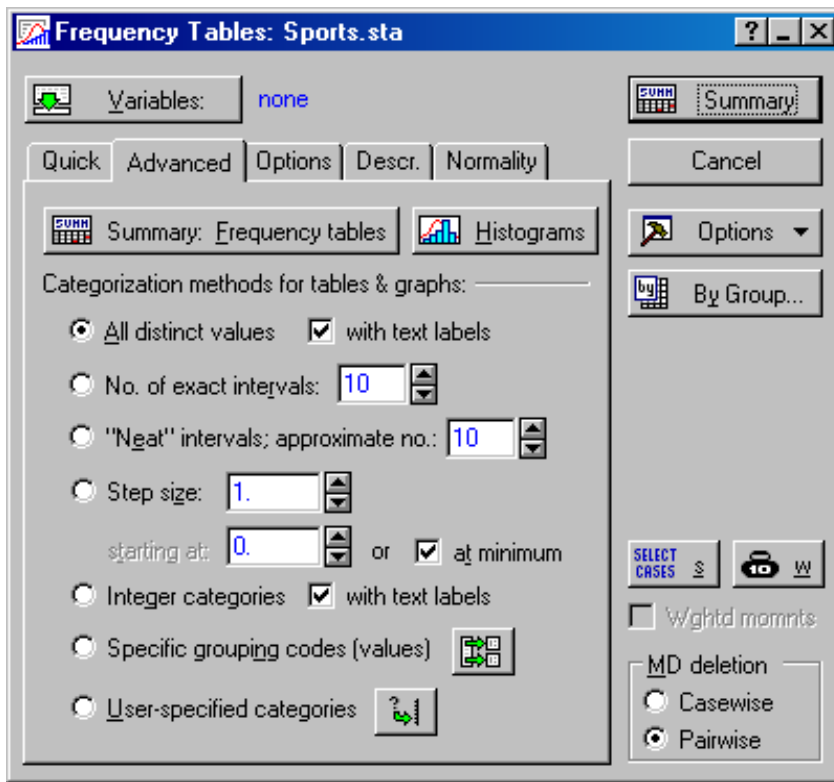
The *Quick* tab has a selection of basic options:



These include Summary: Frequency tables, Histograms, Descriptive statistics, and 3D histograms / bivariate distributions.

On other tabs of this dialog, you can enter specific codes, specify intervals, and even specify conditions that will assign specific cases to categories.

The *Advanced* tab provides methods for categorization:



Here are brief summaries of these methods:

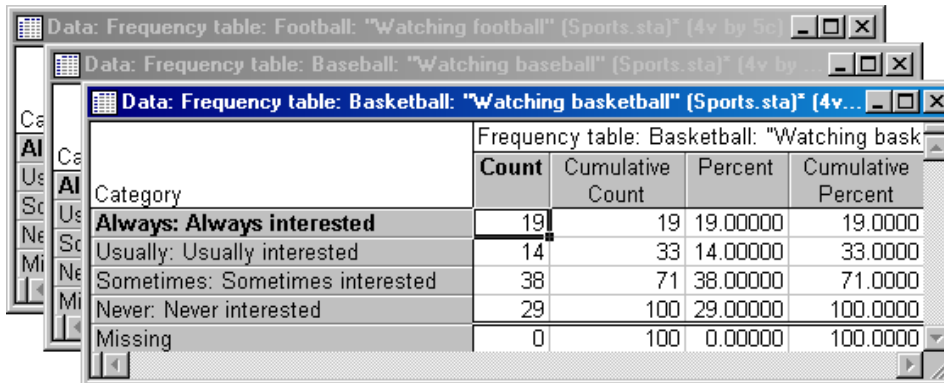
- **All distinct values** - Place each distinct value in its own category.
- **No. of exact intervals** - Divide the entire range of values into the number of intervals specified in the corresponding edit field (e.g., 10 intervals).
- **"Neat" intervals** - Round to simple values with the last digit being a 1, 2, or 5 (e.g., 10.5, 11.0, 11.5, etc.).
- **Step size** - Base the frequency tables on the user-specified step sizes (interval widths).
- **Integer categories** - Base the frequency tables on integer boundaries and step sizes, starting with the smallest integer value.
- **Specific grouping codes** - Base the frequency tables on integer categories (codes) specified by the user.
- **User-specified categories** - Base the frequency tables on a set of up to 16 logical case selection conditions specified by the user.

## Sample Applications

Consider some sample applications: (1) Quick, (2) Neat intervals, and (3) Exact binning.

### Quick

On the *Quick* tab of the *Frequency tables* dialog, you can do a quick analysis by selecting variables and then clicking the *Summary* button. This produces a cascade of frequency tables for your selected variables (one spreadsheet per variable).



The screenshot shows a window titled 'Data: Frequency table: Basketball: "Watching basketball" (Sports.sta) [4v...'. The window contains a table with the following data:

| Category                        | Count | Cumulative Count | Percent  | Cumulative Percent |
|---------------------------------|-------|------------------|----------|--------------------|
| Always: Always interested       | 19    | 19               | 19.00000 | 19.0000            |
| Usually: Usually interested     | 14    | 33               | 14.00000 | 33.0000            |
| Sometimes: Sometimes interested | 38    | 71               | 38.00000 | 71.0000            |
| Never: Never interested         | 29    | 100              | 29.00000 | 100.0000           |
| Missing                         | 0     | 100              | 0.00000  | 100.0000           |

Based a survey of spectator interest in different sports (see data set *Sports.sta*), the frequency table above shows the number, proportion, and cumulative proportion of respondents who characterized their interest in watching basketball as *Always*, *Usually*, *Sometimes*, or *Never*.

### Neat intervals

In the first application, we accepted the given defaults. Now, switch from the *Quick* tab to the *Advanced* tab, and select the "*Neat*" intervals option. As a result, for continuous variables, you get "neat" intervals with exclusive lower bin boundaries.

### Exact binning

Let's say you want to do something exact. For example, you know that the minimum value of a variable is 0 (e.g., they are measurements of response times). So you input a specific *Step size*, and a specific *Starting at* value of 0. This will produce a frequency table starting at 0 inclusive (to include the known minimum value), with upper boundaries that are exclusive. This avoids creation of the nonsensical class "<0", which we already know does not exist.

## Conclusion

In short, in some cases, inclusive lower bounds are useful and desired; and in other cases, upper inclusive bounds are more appropriate. These, and many other options for categorization (or "binning"), are all supported in *STATISTICA*. The important part is that the binning is always explicitly indicated on the graph.