Doing Statistics with XmStat

Harald Martin Stauss, MD
Dept. of Physiology
Humboldt University - Charité Berlin
Tucholskystr. 2
10117 Berlin
Germany
e-mail: harald.stauss@rz.hu-berlin.de

October 27, 2002

Doing Statistics with XmStat: Harald Martin Stauss, 1^{st} edition, 1997

Dedicated to Margarita

Contents

1 GNU general public license **9**
1.1 Preamble 9
1.2 Terms and conditions 10

2 Introduction **17**

3 Installation **19**

4 File Commands **21**
4.1 Arrangement of data for XmStat 22
4.2 New 22
4.3 Load 22
4.4 Save 23
4.5 Close 24
4.6 Exit 24

5 Edit Commands **27**
5.1 Add Observation 28
5.2 Add Variable 28
5.3 Copy Observation 28
5.4 Copy Variable 28
5.5 Delete Observation 28
5.6 Delete Variable 29
5.7 Change Variable Name 29
5.8 Change Variable Order 29

6 Region Commands **31**
6.1 Cut 31
6.2 Copy 31
6.3 Paste 32
6.4 Transpose 32

7 Utils Commands **33**
7.1 Sort Variables 34
7.2 Dependent Variables 35
7.3 Transformations 35
7.3.1 One-variable transformations 35
7.3.2 Two-variables transformations 35

8 Statistics Commands **41**
8.1 The results window 42
8.1.1 File Save Command 43
8.1.2 File Append Command 43
8.1.3 File Print Command 43
8.1.4 File Close Command 44
8.1.5 Edit Insert Pagebreak Command 44
8.2 Data Listing 44
8.3 Descriptive Statistics 44
8.4 Regression 45
8.5 t-tests 45
8.5.1 Paired t-test 46
8.5.2 Unpaired t-test 48
8.6 ANOVA-one-way 52
8.6.1 Repeated Measures 52
8.6.2 Independent Measures 55
8.6.3 Post-hoc tests 57
8.7 ANOVA-two-way 59
8.7.1 Repeated-Repeated Design 59
8.7.2 Independent-Repeated Design 59
8.7.3 Independent-Independent Design 62
8.8 Non-Parametric Tests 65
8.8.1 Wilcoxon Test 65
8.8.2 Mann-Whitney U-Test 65

List of Figures

4.1 Commands in the File menu 21
4.2 New data set dialog 24
4.3 Main window after File New with 26 rows and 9 columns . . 25
4.4 Load Data and Save Data dialogs 25

5.1 Commands in the Edit Menu 27
5.2 The Variable Name dialog 29
5.3 The Change Variable Order dialog 30

7.1 Commands in the Utils menu 33
7.2 Sort by Variables and Select Dependent Variables dialogs 34
7.3 Spreadsheet after sorting by gender and agegroup 37
7.4 The one-variable transformations dialog 38
7.5 The two-variable transformations dialog 39

8.1 Commands in the Statistics menu 41
8.2 Data listing presented in the results window. 42
8.3 The Printing dialog 43
8.4 The paired t-test dialog 47
8.5 The unpaired t-test dialog 50
8.6 The ANOVA-one-way, Repeated Measures dialog 53
8.7 The ANOVA-one-way, Independent Measures dialog . . 56
8.8 The ANOVA-two-way, Independent-Repeated Measures dialog 60
8.9 The ANOVA-two-way, Independent-Independent Measures dialog 63

List of Tables

4.1 IQ values from women and men of different age on six successive tests, each one week apart. 23

8.1 Descriptive statistics calculated from IQ-scores of women and men on six successive weeks: 46

8.2 Paired t-tests on the IQ-scores of women and men reached on the first testing and during the following five tests.: 48

8.3 Unpaired t-tests on the IQ-scores of children versus senescents on the six successive IQ-tests: 51

8.4 1-way-ANOVA table for repeated measures 54
8.5 1-way-ANOVA table for independent measures 55
8.6 2-way-ANOVA table for independent and repeated measures . 61
8.7 2-way-ANOVA table for independent measures 64

Chapter 1

GNU general public license

Version 2, June 1991

1.1 Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software–to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

1.2 Terms and conditions

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".

   Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously

and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

   (a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.

   (b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.

   (c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

   These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

   Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise

the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

   (a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

   (b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

   (c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

   The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

   If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source

code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

   If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

   Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

Chapter 2

Introduction

This manual describes the XmStat program. XmStat is a X-Windows/Motif prgram that is designed to analyse experimental data with commonly used statistical procedures. The algorithms are based on several sources, including books on statistics [2, 4, 7, 10], computer programming [3, 8, 11], and freely available algorithms [9]. I feel that it is necessary to develop such a program since there are currently no statistical analysis programs available for Linux that meet the following criteria:

1. freely available
2. easy to use
3. powerful statistical functions
4. clearly structured print outs of the results
5. documentation available

As far as criteria one is concerned, there exist some freely available statistical analysis programs for Linux. Examples are LispStat, ViStat [13] or xldlas [12]. However, they are either difficult to use or they do not offer important statistical tests like "analysis of variance including post-hoc tests". The large statistical packages (SAS, SPSS, SYSTAT) that offer nearly unlimited statistical functions are not freely available. Thus, I have started to develope this program according to the criteria mentioned above.

In order to make this program a success, I need the feedback of people that have a need to calculated statistics on a Linux based computer system. Any suggestions to improve this program are explicitely appreciated.

Chapter 3

Installation

The statistics package XmStat is a X-windows/Motif program written for the public domain UNIX operating system Linux. XmStat is written in C++ and uses the LessTif widget set (a public domain Motif clon) and the Xbae widget set that provides the spreadsheet functionality. Therefore, the requirements to install and run XmStat are:

- IBM-compatible PC-AT with a 80386 (or better) microprocessor.
- Linux operating system
- GNU C++ compiler
- LessTif (tested with version 0.81) or
  Motif (tested with OpenMotif version 2.1.30)
  Motif 2.0 or higher is required. Motif 1.2 does not work.
- Xbae widget set (tested with version 4.8.2)

To install XmStat follow the guidline outlined below. For some of the installation steps you may need root privileges. The use of an Imakefile should (theoretically) make compilation of the program easy on various platforms.

1. Install LessTif or Motif (you have to follow the installation instructions that come with these packages)
2. Install Xbae (you have to follow the installation instructions that come with Xbae)
3. Untar the XmStat package in the directory where you put all your source codes (example: /usr/src):
   `tar -xzvf xmstat-ver.tar.gz`

4. Change to the xmstat-ver/src directory and generate the Makefile:
   ```
   cd xmstat-ver/src
   xmkmf
   ```

5. Compile the program:
   ```
   make depend; make
   ```

6. As root (su root) do:
   ```
   make install
   ```

7. Start the program from a Xterm shell:
   ```
   xmstat &
   ```

Chapter 4

File Commands

The commands in the FILE menu (see Fig. 4.1) allow to handle the data files for use with XmStat. Two file formats are supported. The ASCII format to import and export data in a format that can be read by nearly every other program. In addition, a special XmStat format is also supported. Files saved in this format include information on the variable names and other parameters.

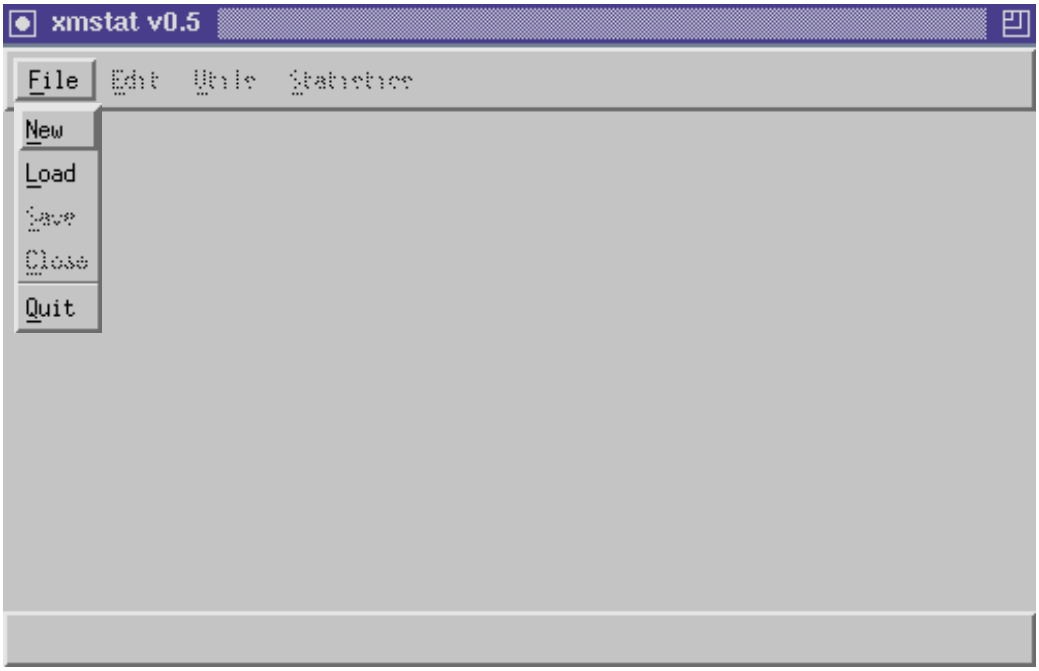


Figure 4.1: Commands in the File menu

4.1 Arrangement of data for XmStat

Arrangement of data will be explained by a simple example. Suppose you want to perform statistics on the question whether women and men of different age differ in their intelligent quotients (IQ) and whether the test persons will obtain better IQ-scores if the tests are repeated once every week for a time period of six weeks. Suppose we have 26 test persons (14 female and 12 male). The database of this study consists of 156 IQ-values (26*6). How should these values be arranged for use with XmStat? In general, data are arranged in rows and colums. Observations (IQ-values from each test person) are arranged in rows, while variables (IQ values of all persons on a specific week) are arranged in columns. The data of this example are shown in Table 4.1 and are provided in the files "iq.ascii" and "iq.stat" in the examples directory. In addition, there may be variables that do not consist of experimental data. Such variables (columns) contain information as to whether an observation (a person) belongs to a specific group (goup of women or men, agegroup).

The data in Table 4.1 are not from a real experiment. They are just brought up as an example for XmStat.

4.2 New

This menu item can be used to enter new data sets from the keyboard. Let us assume we want to enter the data from Table 4.1. These data consist of 26 observations (subjects or persons) and 9 variables (6 IQ-values and one variable for the ID-number of the person, one for the gender, and another variable for the agegroup). Thus, in the dialog that appears (see Fig. 4.2) we enter a number of 26 rows and 9 columns before we hit the OK-button. Now a spreadsheet apears with cells for 9 variables and 26 observations (see Fig. 4.3). We can use the scrollbars to scroll through the spreadsheet. It is possible to change the width of the columns by pressing the shift-key and simultaneously dragging the mouse with the 2^{nd} mouse button pressed. In order to follow the example, we could now enter the data from Table 4.1.

4.3 Load

If we don't want to enter all 156 data values manually, we can use the menu-item LOAD from the FILE menu and load the data from the file "iq.stat" or "iq.ascii" from the examples directory. As mentioned above, XmStat can read ASCII files and special XmStat files. The LOAD DATA dialog (see Fig. 4.4,

Table 4.1: IQ values from women and men of different age on six successive tests, each one week apart.

ID	gender	agegroup	week 1	week 2	week 3	week 4	week 5	week 6
1	female	mature	95	90	102	101	107	110
2	male	child	110	103	104	100	103	108
3	female	mature	84	82	91	93	97	100
4	female	teen	120	124	119	129	131	135
5	male	senescent	115	108	112	113	117	114
6	female	teen	74	76	80	86	88	92
7	female	child	93	91	97	100	105	112
8	male	mature	110	112	108	105	103	113
9	male	senescent	130	123	126	132	125	127
10	male	mature	110	112	107	108	109	110
11	female	child	93	92	97	103	110	115
12	female	senescent	85	89	92	91	97	104
13	female	child	76	79	86	91	87	98
14	male	teen	100	102	95	97	98	99
15	female	mature	98	102	106	111	113	115
16	male	teen	91	89	95	88	87	93
17	female	senescent	77	81	85	89	91	94
18	male	mature	120	105	110	122	117	116
19	male	child	113	106	105	111	107	114
20	female	senescent	90	89	96	99	103	109
21	female	teen	86	88	92	91	97	102
22	male	mature	102	99	105	95	101	97
23	female	mature	76	80	82	87	93	96
24	male	child	80	85	79	75	82	85
25	male	teen	95	97	88	92	94	91
26	female	senescent	85	87	91	93	96	95

left) allows to enter the directory that contains the data files (examples) and to select a file (iq.stat). Don't forget to select also the appropriate file type (i.e. ASCII or STAT) before you hit the OK-button.

4.4 Save

Once we have created a new spreadsheet with the File New command and entered the data that we want to analyse, we may want to save the data. We can do this with the File Save command. The dialog that appears is shown in Fig. 4.4 (right). We need to select the directory in which we want to save the data file, we need to enter a file name and we need to choose a file type. We can choose between the ASCII file format and the XmStat file

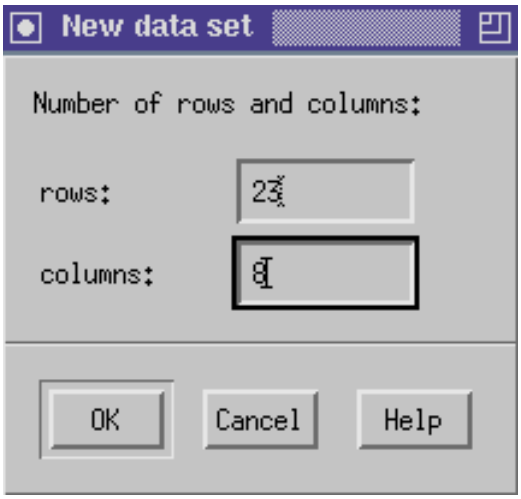


Figure 4.2: NEW DATA SET dialog

format. The ASCII file format should be chosen if we want to import the data into another program. The XmStat file format should be preferred if we want to reuse the file with the XmStat program. As mentioned earlier, the XmStat file format also saves informations on the variable names along with other information.

If the file already exists, a warning will appear that reminds you that this particular file already exists and the possibilities are offered to overwrite the file or to cancel the save process.

4.5 Close

The FILE CLOSE command closes the spreadsheet. The data are **not** saved and may be lost if you did not use the FILE SAVE command before. Thus, be careful when using this menu item.

4.6 Exit

The FILE EXIT menu item ends the program. If you didn't save the data before clicking on this menu item, your data will be lost. There is no warning that informs you that the data are not saved yet. Thus, think twice before exiting from the XmStat program.

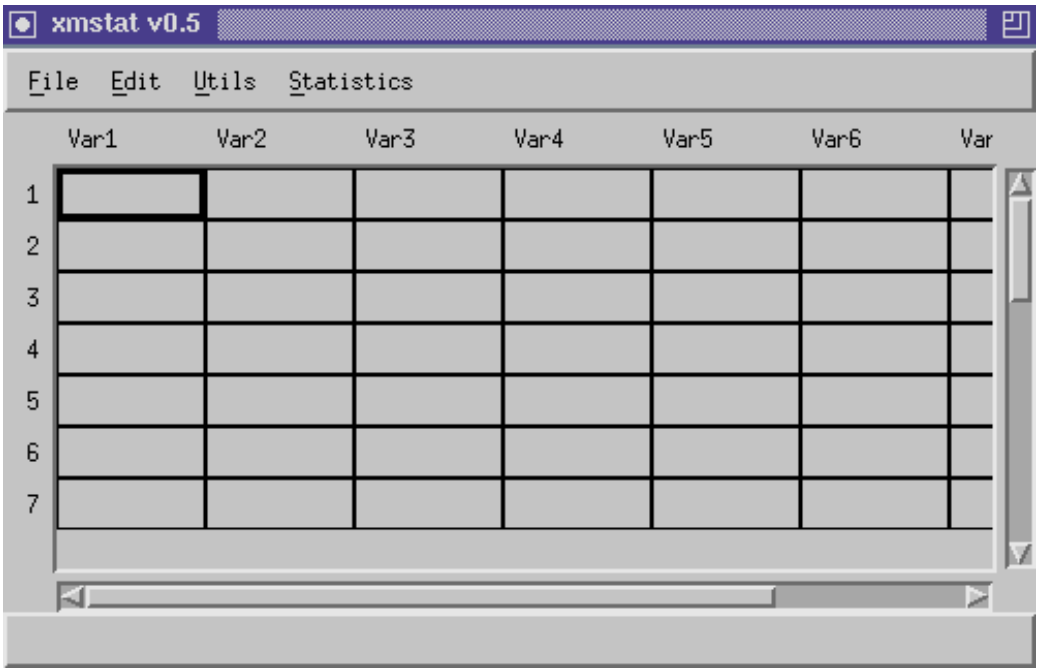


Figure 4.3: Main window after FILE NEW with 26 rows and 9 columns

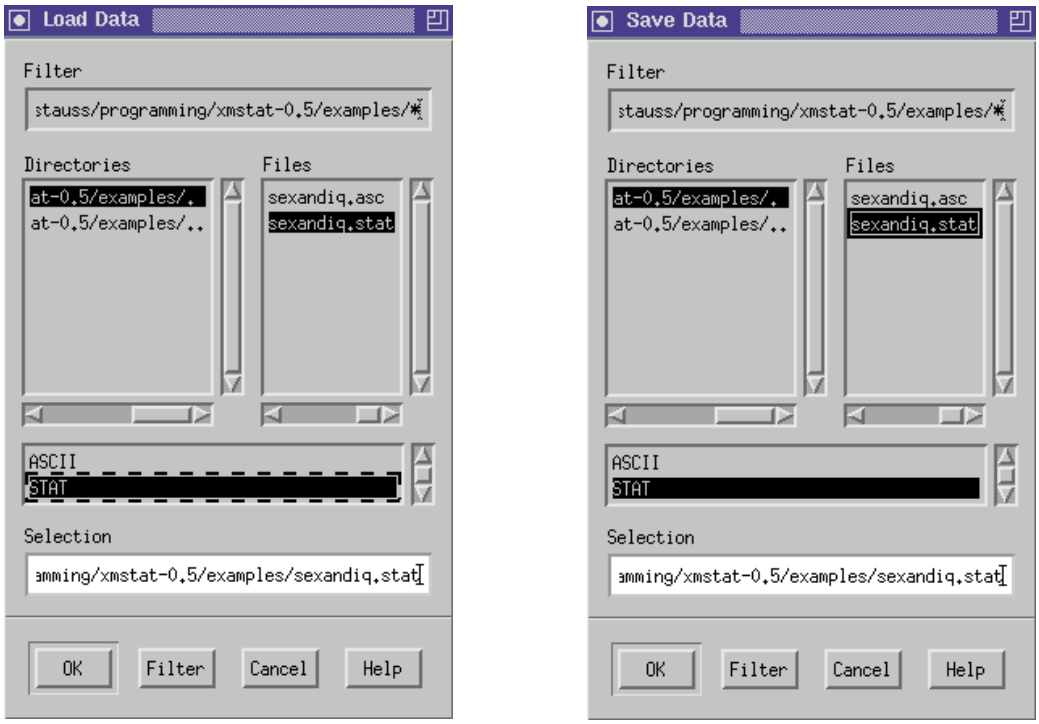


Figure 4.4: LOAD DATA and SAVE DATA dialogs

Chapter 5

Edit Commands

Once you have entered some data manually or you have loaded a data file, you may want to add additional obsersations or variables. Sometimes it is necessary to delete observations or variables that are found to be unvalid for some reasons. In addition, the names of the variables may be changed from the default variable names that consists of the prefix "VAR" and the number of the variable to more meaningful variable names. All these procedures can be reached from the EDIT menu that is shown in Fig. 5.1.

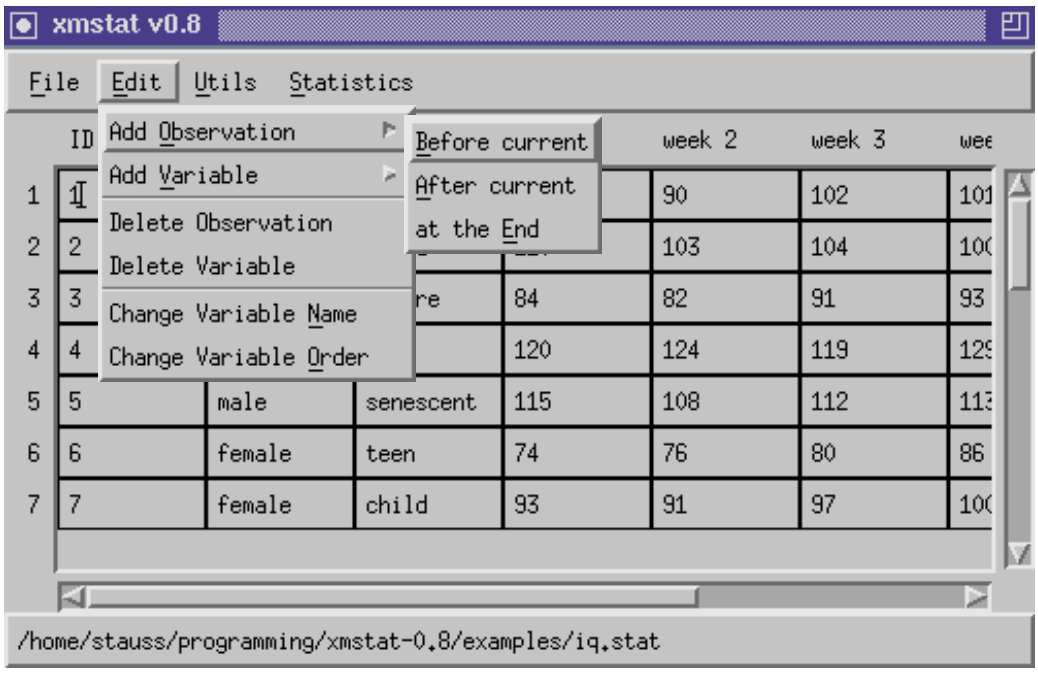


Figure 5.1: Commands in the Edit Menu

5.1 Add Observation

This menu item has a submenu (see Fig. 5.1) that allows to add new observations (p.e. new subjects for the "inteligent quotients study") before the current observation, or after the current observation, or at the end of the spreadsheet (i.e. after the last observation). Since sorting of the spreadsheet must be considered to be garbled after insertion of new observations, the sort variables that may have been defined by the UTILS SORT VARIABLE menu are unselected and the database is defined to be unsorted.

5.2 Add Variable

This menu item also has a submenu that allow to add new variables (p.e. IQ-values for a seventh week) before the current variable, or after the current variable, or at the end of the spreadsheet (i.e. after the last variable).

5.3 Copy Observation

This menu item allows to copy the current observation. A copy of the current observation is inserted to the spreadsheet at the position following the current row (observation). After copying observations, the sort variables are unselected and the database is considered to be unsorted.

5.4 Copy Variable

This menu item allows to copy the current variable. A copy of the current variable is inserted to the spreadsheet at the position following the current column (variable). This function is particularly useful, if one-variable transformations should be calculated from a variable, because variables from which one-variable transformations are performed are replaced by the transformation. Thus, it is useful to create a copy of the variable first and then perform the one-variable transformation on the copy of the variable. The variable name of the copy is created based on the original variable name and the suffix "_c".

5.5 Delete Observation

This menu item simply allows to delete the observation at which the cursor is located. At least one observation must remain in the spreadsheet. If you

want to delete all observations you may use the FILE CLOSE menu item.

5.6 Delete Variable

This menu item allows to delete the variable at which the cursor is located. At least one variable must remain. If you want to delete all variables use the FILE CLOSE menu item.

5.7 Change Variable Name

The EDIT menu also allows to change the name of the variable at which the cursor is located. A dialog box (see Fig. 5.2) will pop up that shows the current variable name in a text field. The name can be changed by entering a new name in the text field and clicking on the OK-button. Please note that it is important that all variable names differ from each other.

Figure 5.2: The VARIABLE NAME dialog

5.8 Change Variable Order

This menu item pops up a dialog box (see Fig. 5.3) with two lists. One list contains all variable names in the file and the other list ist empty. If you click on a variable name on the left list, the variable will be moved to the list on the right side. By clicking on the variable names you can create a new list of variable names (on the right side) that represents the new order of the variables in the file. The OK-button rearranges the file according to the list on the right side, the Cancel-Button leaves everything as it is. If not

all variable names are moved from left to right, the variables in the right list are preceding the variables in the left list.

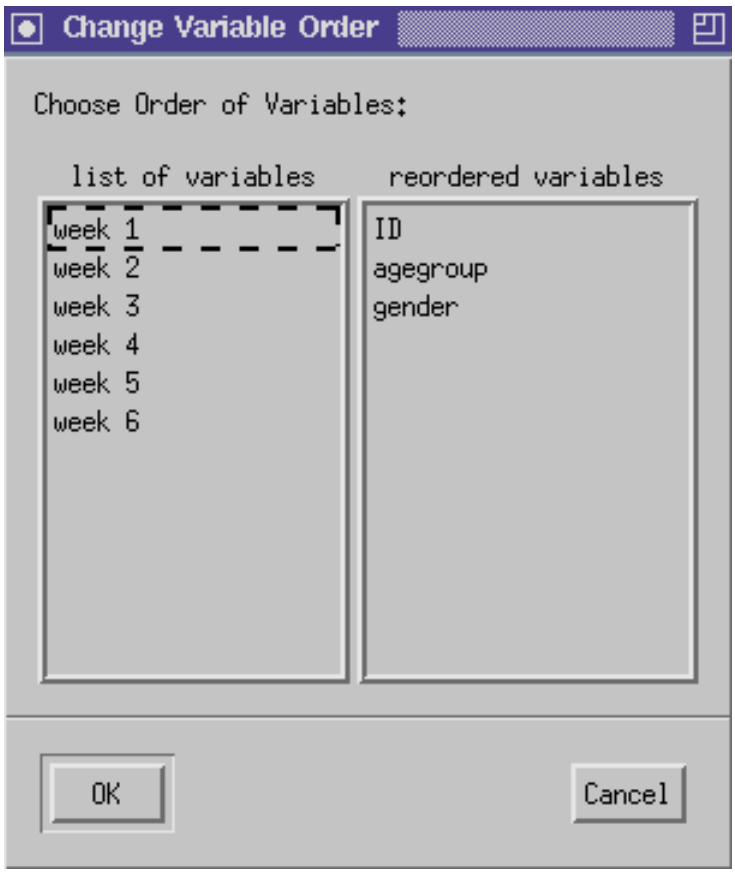


Figure 5.3: The CHANGE VARIABLE ORDER dialog

Chapter 6

Region Commands

The commands in this menu are operating on the data in a selected rectangle. A rectangular area of cells can be selected by pressing the first mouse button and dragging over the cells that should be included in the marked rectangle. The rectangle must start in a different cell than the currently selected cell, because dragging in the current cell operates on the content of this particular cell. Once a rectangular area of cells is selected, these cells can be cleared by the CUT command, copied into an internal buffer (COPY) and then pasted to a new destination. In addition, the cells in the rectangular area can be transposed, i.e. the rows and columns can be exchanged.

6.1 Cut

This command clears the selected (marked) cells and copies the content in an internal buffer. From this buffer, the cleared cells can be copied into a new destination by the PASTE command.

6.2 Copy

Like the CUT command, this command copies the content of the selected (marked) cells in an internal buffer, but does not clear the selected cells. Again, this buffer can be copied into a new destination by the PASTE command.

6.3 Paste

The paste command allows to copy the content of the internal buffer to a certain position. At the position of the current cell, the upper left corner of the rectangle that had been transfered into the internal buffer by the CUT or COPY command will be located.

6.4 Transpose

The TRANSPOSE command allows to exchange the rows and columns in the marked rectangle. This rectangle must consist of the same number of rows and columns.

Chapter 7

Utils Commands

The commands in this menu (see Fig. 7.1) are intended to select variables that define subgroups for the calculations and to select the dependent variables from which statistics should be calculated. In addition, transformations can be performed on the values in the variables by the UTILS TRANSFORMATIONS menu item.

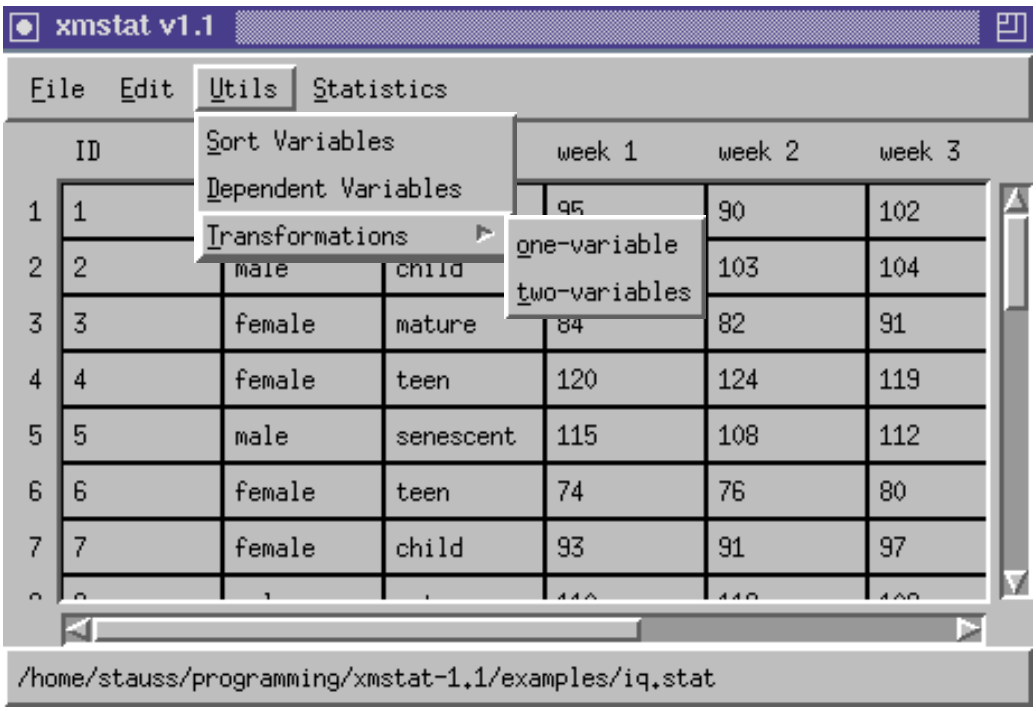


Figure 7.1: Commands in the UTILS menu

7.1 Sort Variables

Sort variables are variables that define subgroups of the database. A sort variable can be a column of the spreadsheet that defines certain characteristics of the subjects. In the database for the IQ-values from women and men (see Table 4.1) the sort variable "gender" naturally has two levels, i.e. female and male. The sort variable "agegroup" that defines subgroups based on the age of the subjects has four levels (i.e. child, teen, mature, senescent). To define sort variables just click on the variable name on the left list in the SORT BY VARIABLES dialog (see Fig. 7.2). The variable will be moved to the list of sort variables on the right side of the dialog. After clicking on the OK button, the subjects in the spreadsheet will be rearranged according to the sort variables. Fig. 7.3 shows the spreadsheet after sorting by gender and agegroup. Note how the rows of the spreadsheet (observations) are rearranged. First there are all female subjects. The subgoup of females is further sorted by the agegroups. Please also note that the order of the sort variables matter. If you calculate descriptive statistics after selecting the sort variable "gender" you get the results of the descriptive statistics separated for all female and all male subjects of the study. If more than one sort variable are selected, statistics are calculated for as many subgroups as the combination of all selected subgroups allow.

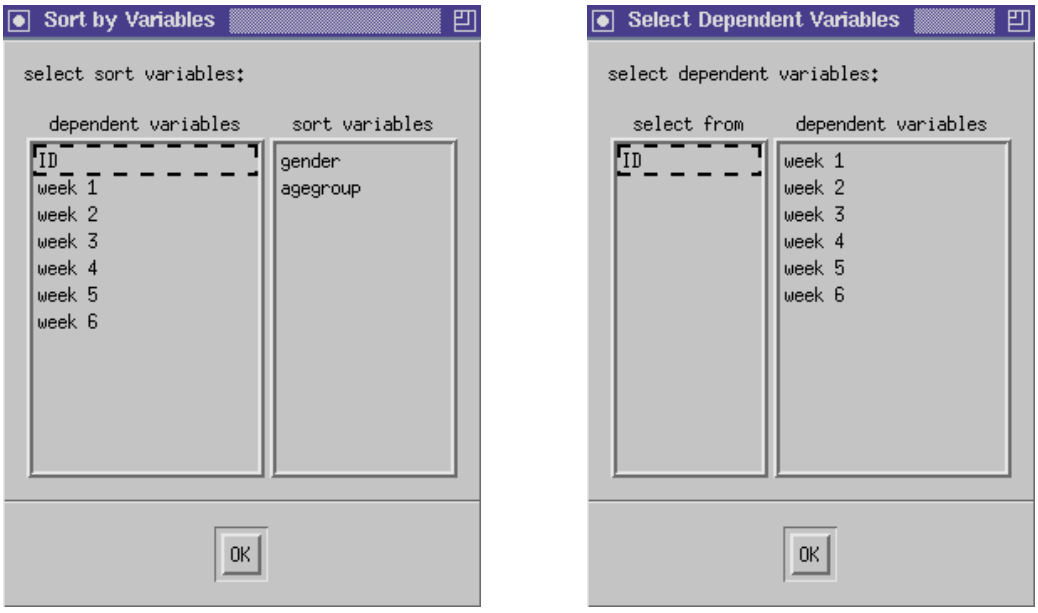


Figure 7.2: SORT BY VARIABLES and SELECT DEPENDENT VARIABLES dialogs

7.2 Dependent Variables

Dependent variables are the variables (columns) from which statistics will be calculated. You can easily select the dependent variables by clicking on the variable names on the left list of the SELECT DEPENDENT VARIABLES dialog (see Fig. 7.2). The variables will be moved to the list of dependent variables. In the database from Table 4.1 you may want to select the variables for the IQ values from the six weeks as dependent variables while you certainly don't want to calculate any statistics from the ID-numbers of the subjects.

7.3 Transformations

Transformations can be performed on a single variable or on two variables. In the case of single variable transformations, the content of the cells of this variable will be overwritten by the new (transformed) values. In the case of two variable transformations, a new variable will be created for each two-variable transformation.

7.3.1 One-variable transformations

To perform a one-variable transformation, use the UTILS TRANSFORMATIONS ONE-VARIABLE menu item. The dialog shown in Fig. 7.4 will appear.

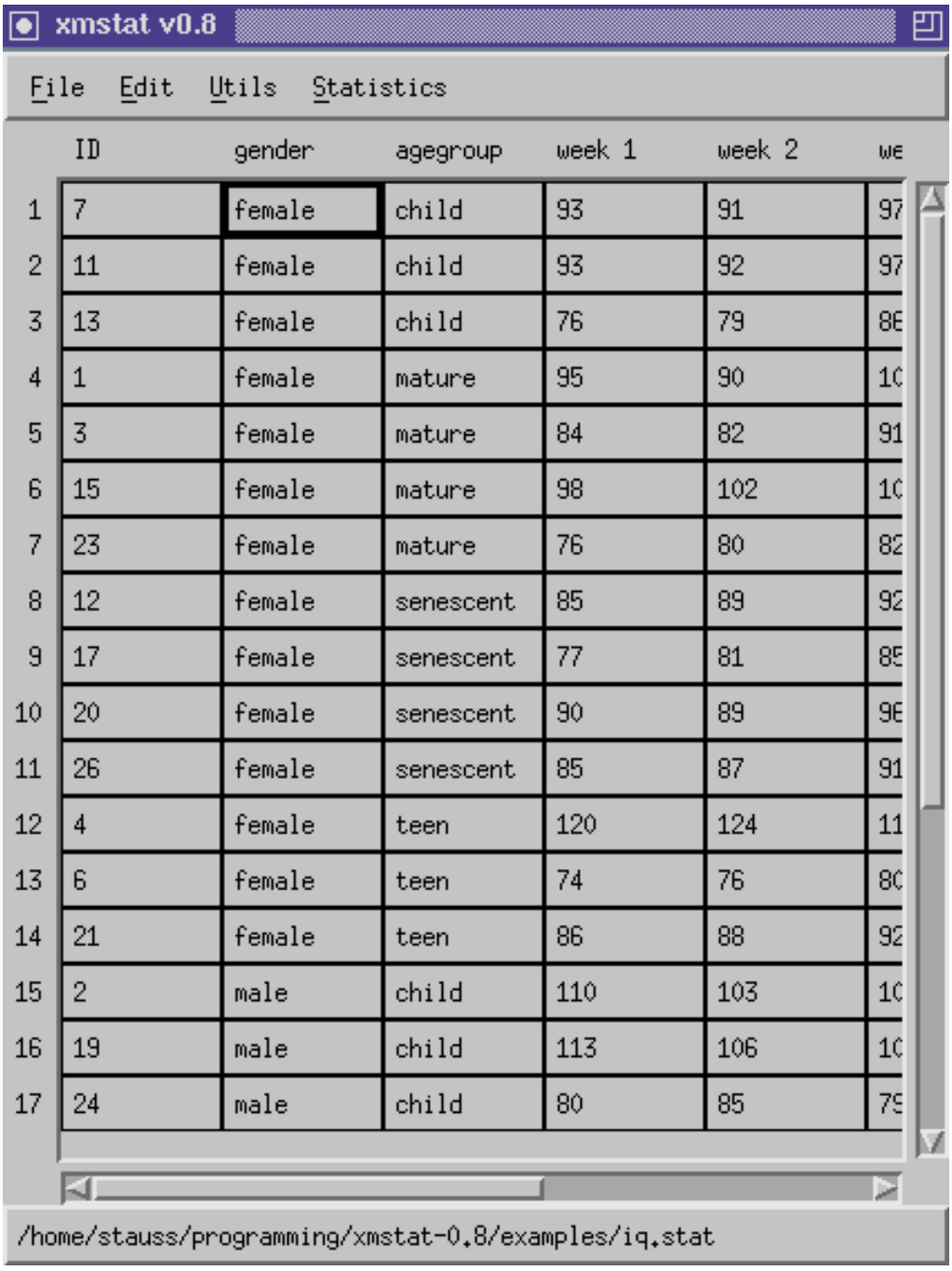
Basically, this dialog consists of a list widget, a field with radio-buttons and two text fields that allow to enter parameters for the transformations. In the list widget you can select the variables from which the transformations should be calculated. Keep in mind, that the values of the variables will be overwritten by the transformed values. Use the radio-buttons to select the transformation that you want to calculate and finally enter the values for the paramters a and b. After clicking on the OK-button, the values of the selected variables will be replaced by the transformed values.

7.3.2 Two-variables transformations

To perform two-variables transformations, use the UTILS TRANSFORMATIONS TWO-VARIABLES menu item. The dialog shown in Fig. 7.5 will appear.

This dialog allows to fill a list of result variable names via a text field widget. In addition, two lists of variables with the same number of variables as the results variables list must be filled using the add buttons. To add variables to the 1st and 2nd variable lists, first highlight the variable names in the list on the left bottom side of the dialog and press the add button under

the respective list. To delete variables from the results variable list or the lists of the first and second variable use the delete buttons. Finally, you have to select a transformation that should be applied to the lists of variables by activating the desired radio button. After pressing the OK-button the calculation will be performed and the new result variables will be added to the spreadsheet.



	ID	gender	agegroup	week 1	week 2	we
1	7	female	child	93	91	97
2	11	female	child	93	92	97
3	13	female	child	76	79	8E
4	1	female	mature	95	90	1C
5	3	female	mature	84	82	91
6	15	female	mature	98	102	1C
7	23	female	mature	76	80	82
8	12	female	senescent	85	89	92
9	17	female	senescent	77	81	85
10	20	female	senescent	90	89	9E
11	26	female	senescent	85	87	91
12	4	female	teen	120	124	11
13	6	female	teen	74	76	8C
14	21	female	teen	86	88	92
15	2	male	child	110	103	1C
16	19	male	child	113	106	1C
17	24	male	child	80	85	7S

Figure 7.3: Spreadsheet after sorting by gender and agegroup

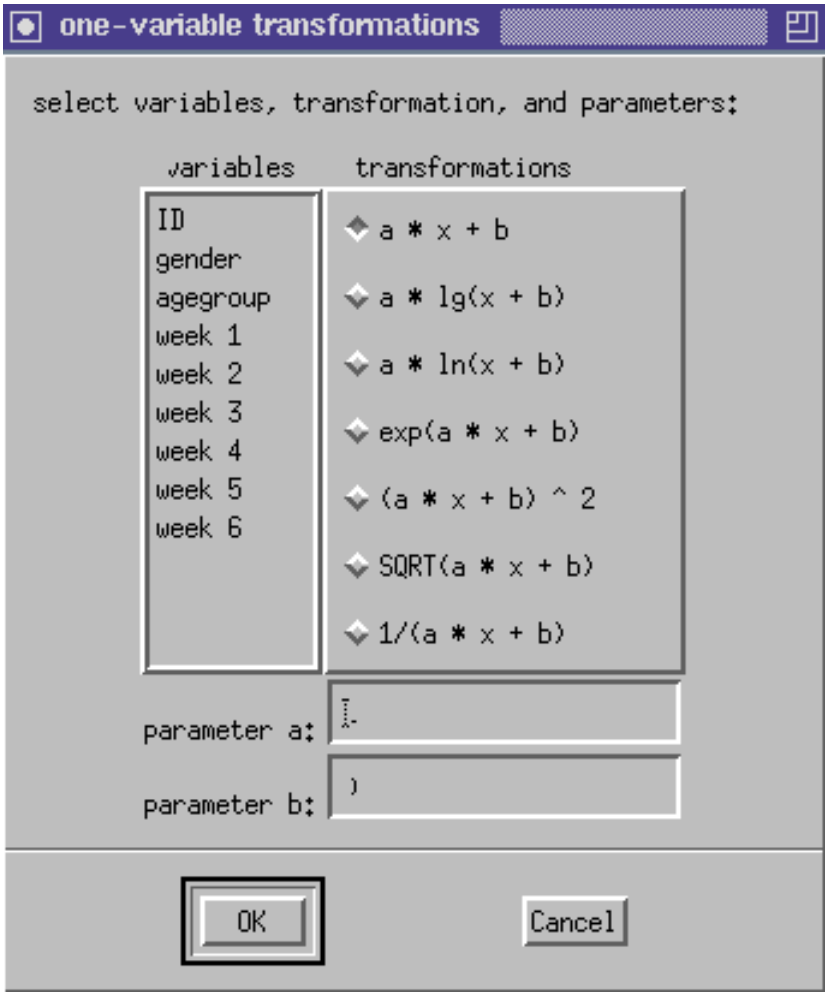


Figure 7.4: The ONE-VARIABLE TRANSFORMATIONS dialog

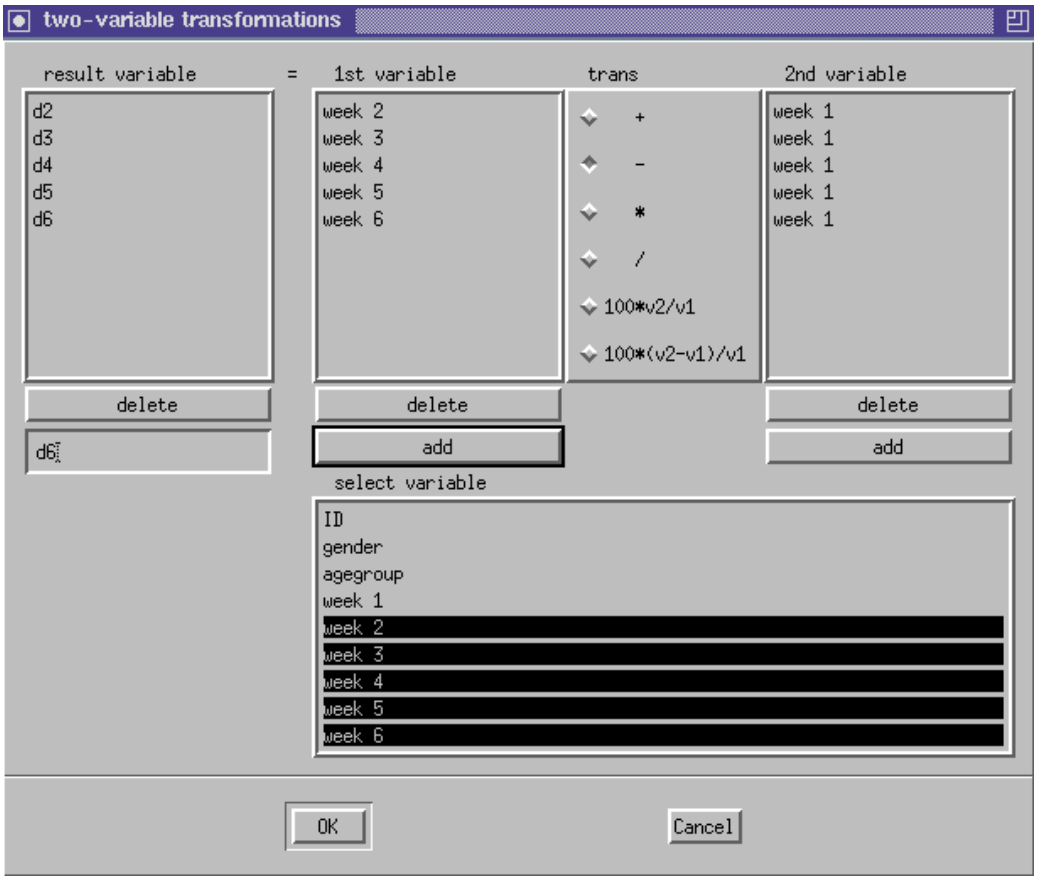


Figure 7.5: The TWO-VARIABLE TRANSFORMATIONS dialog

Chapter 8

Statistics Commands

The STATISTICS menu offers a variety of statistical analyses tools. The database can be listed, descriptive statistics, t-tests, and analyses of variances (ANOVAs), including post-hoc tests, can be calculated (Fig. 8.1).The results are presented in a text window that allows to edit the output manually (Fig. 8.2).

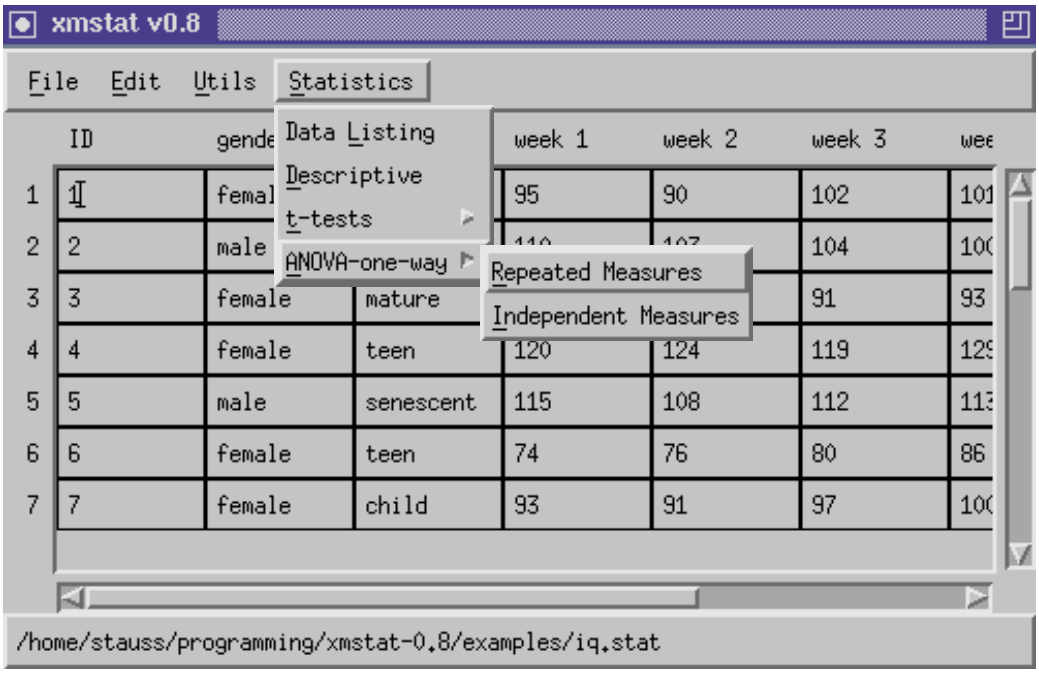
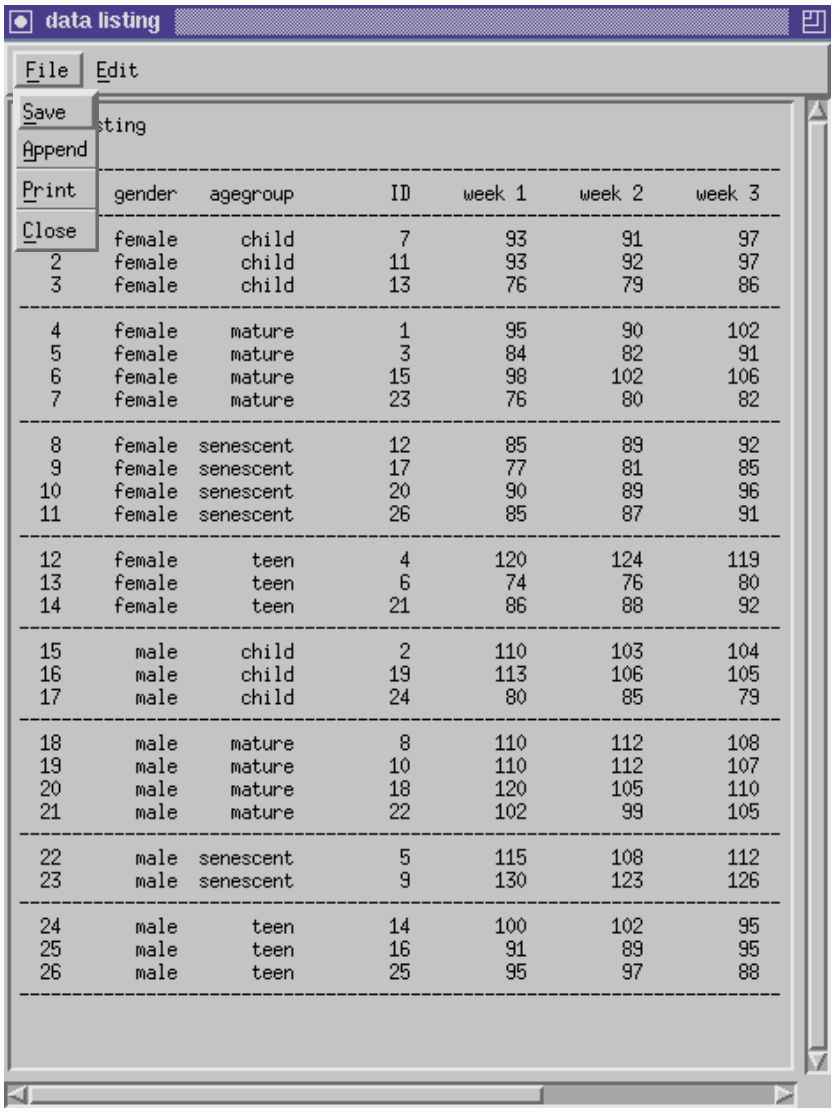


Figure 8.1: Commands in the STATISTICS menu

8.1 The results window

The results window (Fig. 8.2) presents the results that can be obtained from the STATISTICS menu (Fig. 8.1). It features a menubar that offers a FILE and an EDIT menu. The FILE menu offers four functions (Save, Append, Print, and Close). The EDIT menu allows to insert a pagebreak at the cursor position within the results window. The results window has editing capabilities as provided by the multiline ScrolledText Motif-widget.



	gender	agegroup	ID	week 1	week 2	week 3
1	female	child	7	93	91	97
2	female	child	11	93	92	97
3	female	child	13	76	79	86
4	female	mature	1	95	90	102
5	female	mature	3	84	82	91
6	female	mature	15	98	102	106
7	female	mature	23	76	80	82
8	female	senescent	12	85	89	92
9	female	senescent	17	77	81	85
10	female	senescent	20	90	89	96
11	female	senescent	26	85	87	91
12	female	teen	4	120	124	119
13	female	teen	6	74	76	80
14	female	teen	21	86	88	92
15	male	child	2	110	103	104
16	male	child	19	113	106	105
17	male	child	24	80	85	79
18	male	mature	8	110	112	108
19	male	mature	10	110	112	107
20	male	mature	18	120	105	110
21	male	mature	22	102	99	105
22	male	senescent	5	115	108	112
23	male	senescent	9	130	123	126
24	male	teen	14	100	102	95
25	male	teen	16	91	89	95
26	male	teen	25	95	97	88

Figure 8.2: Data listing presented in the results window.

8.1.1 File Save Command

The FILE SAVE menu items allows to save the contents of the results window to an ASCII (i.e. text) file. The file name is entered via a dialog entitled SAVE RESULTS ON FILE that is similar to the dialogs shown in Fig. 4.4. If the file already exists you will be asked whether you want to overwrite the file. Alternatively, you can select a different file name.

8.1.2 File Append Command

This menu item allows to append the contents of the results window to an existing file in ASCII (i.e. text) format. The file name is entered via a dialog entitled APPEND RESULTS TO FILE that is similar to the dialogs shown in Fig. 4.4. If the file does not exist, a new file will be created. Appending results to a preexisting file is a useful feature that allows to generate an output file that holds the results of different calculations on the current database. For example, you may want to generate one output file that first contains the data listing, followed by the descriptive statistics and an ANOVA or a t-test. Once you are ready with your calculations you can load the output file into a text editor, make some final changes and print it out.

8.1.3 File Print Command

The menu item FILE PRINT allows to print the results displayed in the results window. In the dialog that appears (Fig. 8.3) you are asked to enter the command that prints ordinary text on your printer. After pressing the OK-button, the results will be printed.

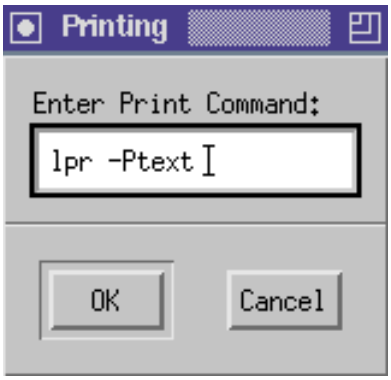


Figure 8.3: The PRINTING dialog

8.1.4 File Close Command

This menu item simply closes the results window. There is not much more to say about this.

8.1.5 Edit Insert Pagebreak Command

This menu item inserts a pagebreak at the current cursor position. The same effect can be achieved by inserting a `ctr l` character (that's pressing the `Strg` key and the `l` key simultaneously) at the location where the pagebreak should appear. In the print-out of the results, a new page will be started at the location where the pagebreak was inserted.

There are more editing features in the results window that are provided by the multiline ScrolledText Motif-widget. Within the results window you can edit the text by inserting, deleting, and copying text. For example, you can highlight a text section with the mouse and paste the highlighted text to a different location by clicking the 2rd mouse button. You can even paste the highlighted text to different programs like text editors.

8.2 Data Listing

This menu item gives a listing of all variables that are defined as sort or dependent variables. If sort variables are defined, then the list is grouped into subgroups based on the sorting variables. Between each subgroup a horizontal line is inserted. An example is given in Fig. 8.2. The sort variables were the gender and the agegroup.

8.3 Descriptive Statistics

This menu item calculates descriptive statistics on the dependent variables. Don't forget to define the dependent variables from that descriptive statistics should be calculated. If sort variables are defined, the descriptive statistics are calculated on subgroups based on the sort variables. The following descriptive statistics are calculated:

- the number of observations (n).
- the minimum (min)
- the arithmetic mean (mean)
  $\bar{x} = \frac{\sum x_i}{n}$

- the maximum (max)
- the variance (var)
  $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum {x_i}^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$
- the standard deviation (sdev)
  $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum {x_i}^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$
- the standard error of the mean (sem)
  $\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n(n-1)}} = \sqrt{\frac{\sum {x_i}^2 - \frac{\left(\sum x_i\right)^2}{n}}{n(n-1)}}$

An example of descriptive statistics is given in Table 8.1. Descriptive statistics of the IQ-scores from women and men are calculated for all IQ-tests that were performed on six successive weeks. It looks like femals start with lower IQ-scores than men. However, while women improve with successive tests, men do not show such a trend. Whether this interpretation of the descriptive statistics is correct, can only be judged by statistical tests. How these can be performed will be explained in the following chapters.

8.4 Regression

8.5 t-tests

The Student's t-test is useful to compare the mean values of two samples. However, t-tests can only be used if the samples are normaly distributed. If the samples are not normally distributed, non-parametric tests (like the U-test described by Mann and Whitney or the Wilcoxon-test) should be applied. Student's t-tests can be computed from paired and unpaired samples. Unpaired samples are samples that are totally independent. An example would be to compare the IQ-values reached by women versus the IQ-values reached by men. In contrast, paired samples are somehow dependent from each other. For example in our test dataset, each subject performed on the IQ-test on six separate time points. Thus, the IQ-values reached by the same person at different time points are paired samples. However, since there are more than two samples (six IQ-tests) the paired t-test would not be the best choise. Student's t-tests can only compare two groups. If there are more than two groups an analysis of variance (ANOVA) should be calculated.

Table 8.1: Descriptive statistics calculated from IQ-scores of women and men on six successive weeks:

gender = female

variable	n	min	mean	max	var	sdev	sem
week 1	14	74.000	88.000	120.000	143.846	11.994	3.205
week 2	14	76.000	89.286	124.000	144.220	12.009	3.210
week 3	14	80.000	94.000	119.000	105.077	10.251	2.740
week 4	14	86.000	97.429	129.000	131.341	11.460	3.063
week 5	14	87.000	101.071	131.000	137.148	11.711	3.130
week 6	14	92.000	105.500	135.000	133.962	11.574	3.093

gender = male

variable	n	min	mean	max	var	sdev	sem
week 1	12	80.000	106.333	130.000	183.879	13.560	3.914
week 2	12	85.000	103.417	123.000	106.447	10.317	2.978
week 3	12	79.000	102.833	126.000	148.879	12.202	3.522
week 4	12	75.000	103.167	132.000	239.424	15.473	4.467
week 5	12	82.000	103.583	125.000	157.356	12.544	3.621
week 6	12	85.000	105.583	127.000	154.629	12.435	3.590

8.5.1 Paired t-test

To compare two normally-distributed, paired data sets you may use the paired t-test. A t-value is calculated based on the n paired observations. From that t-value the probability that both samples have the same mean is calculated based on the t-distribution and the appropriate degrees of freedom ($df = n - 1$).

$$t = \frac{|\overline{x_i - y_i}|}{\sigma_{\overline{x_i - y_i}}} = \frac{\left|\frac{\sum (x_i - y_i)}{n}\right|}{\sqrt{\frac{\sum (x_i - y_i)^2 - \frac{\left(\sum (x_i - y_i)\right)^2}{n}}{n(n-1)}}}$$

To calculate paired t-tests first select sort variables if you want to calculate paired t-tests on subgroups of the database. You don't need to select dependent variables, since the variable pairs are selected in the PAIRED T-TEST dialog (Fig. 8.4) that appears if you select the STATISTICS T-TESTS PAIRED menu item. This dialog allows to select pairs of variables that will be compared by the paired t-test. The number of variables in list 1 and list 2

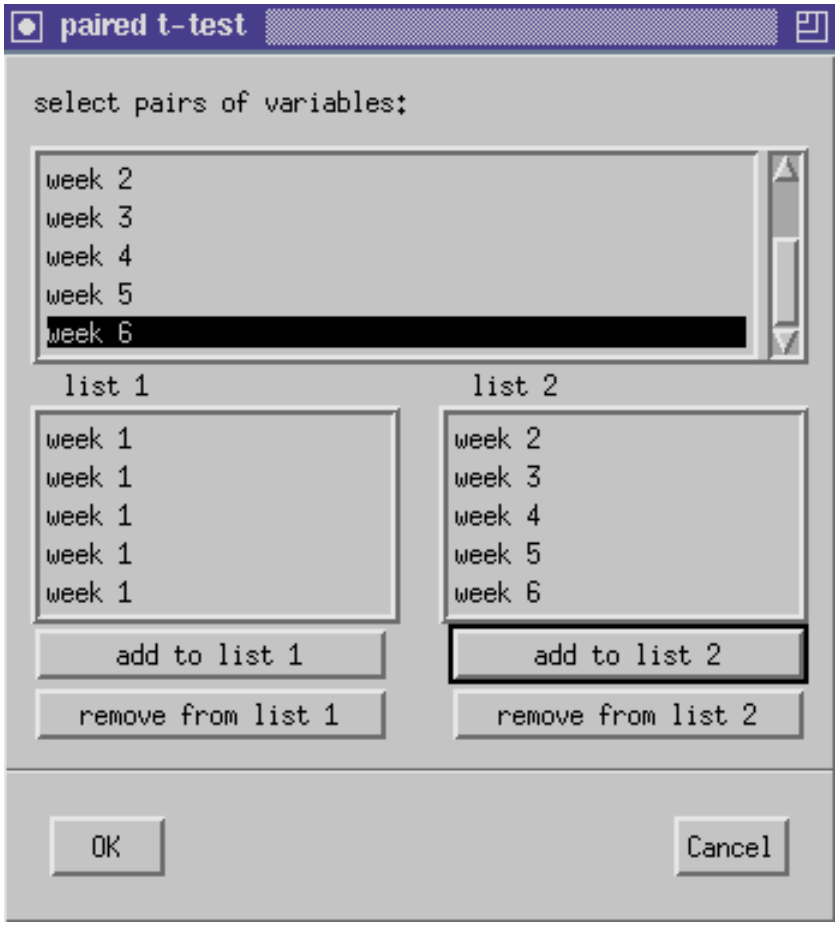


Figure 8.4: The PAIRED T-TEST dialog

must be equal since each corresponding variable in list 1 and list 2 are to be compared using the paired t-test. After both lists are defined, hit the OK-button to start the analysis.

The results window shows up and presents the results of the paired t-test. First the names of the two variables that are compared are listed, followed by the number of observations (n) and the degrees of freedom (df). In the next two columns of the results table the t-value (t) and the associated p-value (p) are given. In the last two columns the 95 % confidence interval for the difference of the means of the two groups is provided. If the confidence intervall contains 0, the two groups are likely to have the same mean. If both limits of the intervall are negative, then variable 1 is smaller than variable 2. If both limits are positive, then variable 1 is larger than variable 2.

An example is given in Table 8.2. Paired t-tests were calculated to compare the IQ-scores reached at the first test (week 1) with the IQ-scores reached at the tests at the following weeks (weeks 2-6). The calculations were performed for the subgroups of women and men by defining the sort variable "gender". In general it is not acceptable to perform multiple t-tests if more than two levels are available for a factor (in the example we have six levels of the factor "intelligent quotient") because the probability that at least one of the multiple tests would result in a Type I error increases with the number of levels. In such a scenario you should use an analysis of variance (ANOVA). Nevertheless, we calculated multiple paired t-tests in this example to demonstrate the use of the paired t-test function in XmStat. Based on the results of this paired t-test, it looks like women reached significant better IQ-scores at the 3rd, 4th, 5th, and 6th week than at the first week, while men got worse IQ-scores at the 3rd, 4th, and 5th week.

Table 8.2: Paired t-tests on the IQ-scores of women and men reached on the first testing and during the following five tests.:

gender = female

variable 1	variable 2	n	df	t	p	CI low	CI high
week 1	week 2	14	13	1.633	0.126	-2.987	0.415
week 1	week 3	14	13	8.832	0.000	-7.468	-4.532
week 1	week 4	14	13	12.050	0.000	-11.119	-7.738
week 1	week 5	14	13	23.372	0.000	-14.280	-11.863
week 1	week 6	14	13	20.908	0.000	-19.308	-15.692

gender = male

variable 1	variable 2	n	df	t	p	CI low	CI high
week 1	week 2	12	11	1.735	0.111	-0.783	6.616
week 1	week 3	12	11	2.910	0.014	0.853	6.147
week 1	week 4	12	11	3.245	0.008	1.019	5.314
week 1	week 5	12	11	3.037	0.011	0.757	4.743
week 1	week 6	12	11	0.844	0.417	-1.206	2.706

8.5.2 Unpaired t-test

Unpaired t-tests can be calculated to compare the mean values of two independent variables. An example would be to compare the IQ-scores from

children with those reached by senescents. To calculate unpaired t-tests, you must first define the dependent variables (UTILS menu) on which the unpaired t-tests should be calculated. The STATISTICS T-TESTS UNPAIRED menu item will then popup a dialog box (Fig. 8.5) with two lists. The left list contains the variables that can be used to define the two groups that should be compared. In our example we select the variable "agegroup". This variable has four different levels, i.e. child, mature, senescent, and teen that are listed on the right side. If we would like to compare the effect of sex, we would select the variable "gender" that has only two levels (female and male). Since the unpaired t-test can only compare two levels of a factor we must select the two levels of the variable that defines the two groups. With the variable "gender" this is not a problem since there are only two levels. However, with the variable "agegroup" we can choose from four agegroups. To compare the IQ-scores from children and senescents we would select the respective two levels of the factor age. If more than two levels are selected in the right list when you click on the OK-button, an error message will tell you that only two levels must be selected. Please keep in mind, that it is not acceptable to perform multiple comparisons on more than two levels by the unpaired t-test, since the probability to make a type I error would increase.

Similar to the paired t-test, a t-value will be calculated based on the n_1 and n_2 observations in the two groups and a p-value will be associated with the t-value based on the degrees of freedom (df) of the unpaired t-test. The equation for the t-value and for the degrees of freedom is dependent on whether the two samples have the same variance or not. For equal variances the equations are:

$$df = n_1 + n_2 - 2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}{n_1+n_2-2}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}$$

If the variances of the two samples that should be compared using the unpaired t-test are not the same, the degrees of freedom (df) and the t-value are calculated using different equations:

$$df = \frac{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{\sigma_2^2}{n_2}\right)^2}{n_2-1}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

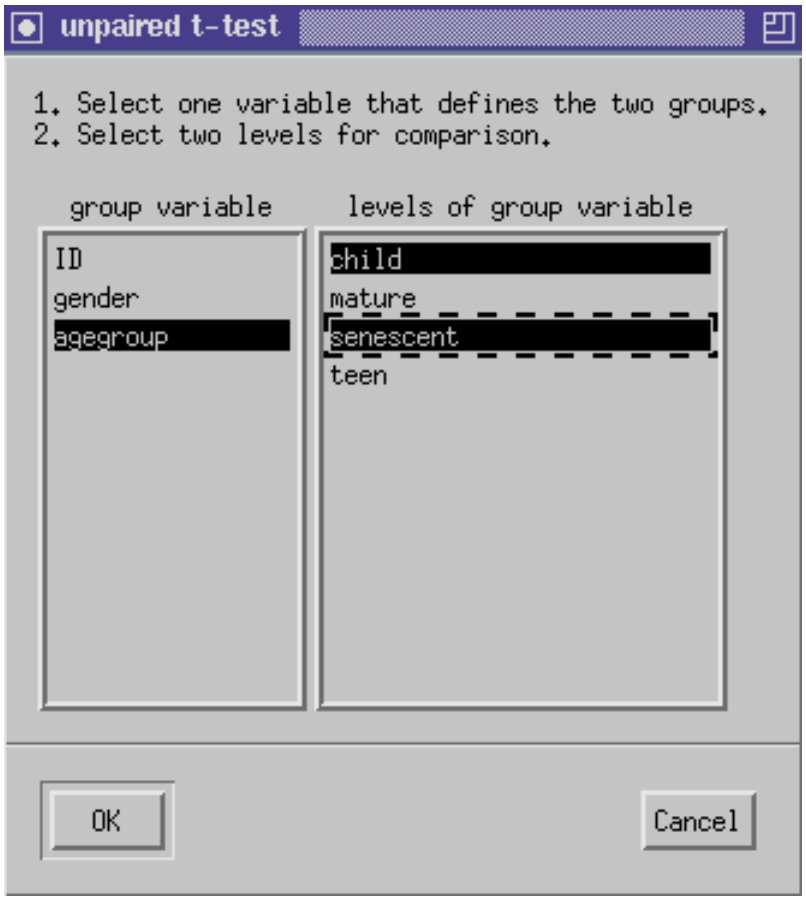


Figure 8.5: The UNPAIRED T-TEST dialog

The output of the unpaired t-test on the IQ-scores of children versus senescents on the six successive test is shown in Table 8.3. For each comparison two t-tests are calculated. This is necessary since the t-value is calculated by different equations depending on whether the samples have the same variance or not. For each comparison a F-test is calculated to find out whether the two samples have the same variance or not. The last column in the output table gives the result of this F-test. The letters "eq" (equal variances) in the last column means that the t-test was calculated based on the assumption that the two samples have the same variance, while the letters "ueq" (unequal variances) means that the t-test was based on different variances in the two samples. An asterisk (∗) followed by the letters "eq" or "ueq" indicates which t-test should be used based on the F-test of equal variances. Thus, "eq∗" means that the variances were not found to be different

(p>0.05), while "ueq∗" means that the variances are likely to be different (p≤0.05).

Table 8.3: Unpaired t-tests on the IQ-scores of children versus senescents on the six successive IQ-tests:

Independent groups Student's t-test
Group 1: agegroup = child
Group 2: agegroup = senescent

	n1	n2	df	t	p	CI low	CI high	F
week 1	6	6	10.000	-0.271	0.792	-26.163	20.496	eq*
week 1	6	6	9.136	-0.271	0.793	-26.465	20.798	ueq
week 2	6	6	10.000	-0.451	0.662	-20.808	13.808	eq*
week 2	6	6	8.556	-0.451	0.663	-21.212	14.212	ueq
week 3	6	6	10.000	-0.746	0.473	-22.590	11.257	eq*
week 3	6	6	8.663	-0.746	0.475	-22.951	11.618	ueq
week 4	6	6	10.000	-0.726	0.485	-25.100	12.767	eq*
week 4	6	6	9.226	-0.726	0.486	-25.318	12.984	ueq
week 5	6	6	10.000	-0.809	0.437	-21.895	10.228	eq*
week 5	6	6	9.807	-0.809	0.438	-21.937	10.271	ueq
week 6	6	6	10.000	-0.263	0.798	-17.389	13.722	eq*
week 6	6	6	9.964	-0.263	0.798	-17.396	13.730	ueq

The first column in the output table gives the variable for which the independent t-test was calculated. The next two columns are the number of observations in the two groups, followed by the degrees of freedom. The next two columns give the t-value and the respective p-value. Then, the confidence intervall for the difference of the mean values of the two samples are given. The last column, finally, gives the result of the F-test for equal variances. It looks like the variances in the IQ-scores from children and senescent are similar, since in all comparisons the F-test of equal variances revealed a p-value larger than 0.05, as indicated by "eq∗". Thus, the independent t-tests calculated based on equal variances (marked by "eq∗") should be used to compare the two groups. In addition, the mean values of the IQ-scores reached by children and senescents are not statistically significant different, since the p-values of the independent t-tests are all larger than 0.05.

8.6 ANOVA-one-way

If the mean values of more than two normally distributed groups of samples need to be compared, the analysis of variance is the appropriate statistical test. A F-value is calculated and based on the F-distribution a p-value is estimated that gives the probability that all mean values are identical. In other words, if the p-value is smaller than a pre-defined critical value (p.e. $p<0.05$) then at least two of the groups have different mean values. In the one-way ANOVA there is only one factor. This one factor however, usually has more than two levels. If the levels are independent from each other, than an independent measures ANOVA (completely randomized design) should be calculated. An example would be to compare the IQ-scores reach on the first test day by children, teens, matures, and senescents. The independent factor would be the age and the four levels would be the four agegroups. The dependent variable would be the IQ-scores on the first week. On the other hand, factors in a one-way ANOVA can also be dependent from each other. An example would be the IQ-scores reached on the six successive testings. Since it is possible, that a learning effect takes place when the tests are repeated, the IQ-scores reached on successive testings can not be considered to be independent. In such a case, a repeated measures ANOVA (randomized block design) should be performed.

The analysis of variance only provides information as to whether at least two mean values differ. There is, however, no information on which mean values differ from each other. Since it is often desirable to know which groups differ, post-hoc tests can be performed. In such post-hoc tests, all possible combinations of mean values ($n*(n-1)/2$) will be compared. Thus information on which mean values differ from each other is provided. There are a number of post-hoc tests described in the literature. Examples are the Scheffé test, the Tukey test, the Newman-Keuls test, or the Fisher test. These post-hoc tests differ in their "strength" or "power". In other words, it is possible that one post-hoc test finds differences between two groups while another post-hoc test does not. Therefore it is important to know some characteristics of the different post-hoc tests, that are discussed in a later section.

8.6.1 Repeated Measures

To calculate a repeated measures ANOVA use the STATISTICS ANOVA-ONE-WAY REPEATED MESURES menu item. A dialog with two lists pops up (see Fig. 8.6). In the left list, all variables that are not sort variables are listed. By clicking on the variable names, the variables are moved to the right list which represents the repeated measurements that should be compared

by the ANOVA. For example, you may want to select the six variables that hold the IQ-scores reached on the six weeks. After hitting the OK-button, the results of the repeated measurements ANOVA will be presented in the results window. If sort variables have been defined (p.e. the gender or/and the agegroup), the ANOVA will be calculated for all respective subgroups.

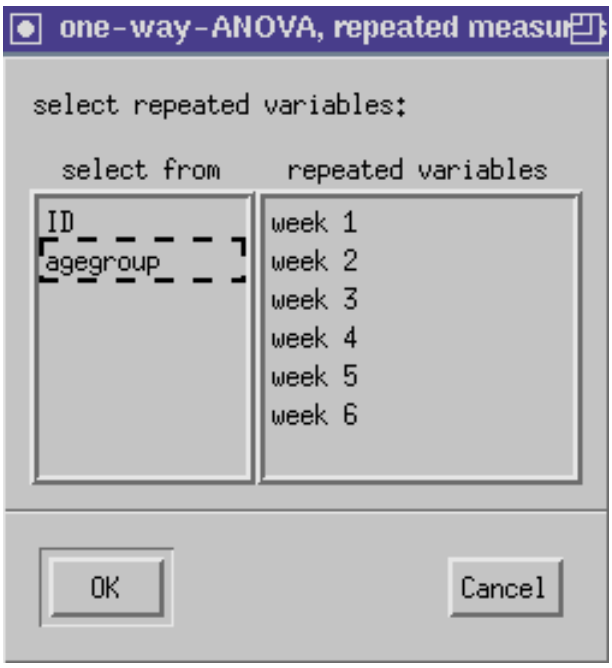


Figure 8.6: The ANOVA-ONE-WAY, REPEATED MEASURES dialog

On top of the results window, the ANOVA table (see Table 8.4) is displayed, followed by the post-hoc tests. In the ANOVA table, the total variance of all samples of all groups is separated into the variance that can be attributed to the intra-individual variance (within the six IQ-scores of each subject), inter-individual variance (between the subjects), and in the variance that is caused by the sampling variability (error). A F-value is calculated for the intra-individual variance (within) and for the inter-individual variance (between). Based on the F-distribution and the respective degees of freedom, these F-values are associated with a p-value that is given in the last column of the ANOVA table. If the p-value of the intra-individual (within) variance is small (p.e. $p<0.05$), than at least two of the levels of the repeated measurements factor differ. In our example this would mean that repeated IQ-tests reveal different results. If the p-value of the inter-individual (between) variance is small (p.e. $p<0.05$), this means that the subjects are not

very homogenious. In our example it would mean that some subjects received significant different (higher or lower) IQ-scores than some other subjects.

Source	df	sum of squares	mean squares	F	p
within	5	3234.0952	646.8190	127.0561	0.0000
between	13	10011.8105	770.1393	151.2802	0.0000
error	65	330.9028	5.0908		
total	83	13576.8086			

Table 8.4: 1-way-ANOVA table for repeated measures

Calculation of the various values in the repeated measures ANOVA table is performed according to the following scheme:

Source	df	sum of squares	mean squares	F	p
within	p-1	SSG	MSG	MSG/MSE	$F_{p-1,\ n-p-b+1}$
between	b-1	SSS	MSS	MSS/MSE	$F_{b-1,\ n-p-b+1}$
error	n-p-b+1	SSE	MSE		
total	n-1	SS (Total)			

p: number of groups (repeated measurements)
b: number of subjects (observations)
n: total number of data values (p * b)
$\bar{x}_{G_i}$: mean of all subjects for the i^{th} measurement
$\bar{x}_{S_i}$: mean of all repeated measurements for the i^{th} subject
$\bar{x}$: mean of all data values in the database

$$SSG\ (group) = \sum_{i=1}^{p} b(\bar{x}_{G_i} - \bar{x})^2$$

$$SSS\ (subject) = \sum_{i=1}^{b} p(\bar{x}_{S_i} - \bar{x})^2$$

$$SS\ (total) = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SSE\ (error) = SS - SSG - SSS$$

$$MSG = \frac{SSG}{p-1}$$

$$MSS = \frac{SSS}{b-1}$$

$$MSE = \frac{SSE}{n-b-p+1}$$

8.6.2 Independent Measures

To calculate an independent measures ANOVA, you first have to select sort variables if calculations on subgroups need to be performed. In addition, you must define dependent variables on which the independent measures ANOVA should be calculated. Once the sort and dependent variables have been defined, you can access the ONE-WAY-ANOVA, INDEPENDENT MEASURES dialog (see Fig. 8.7) via the STATISTICS ANOVA-ONE-WAY INDEPENDENT MEASURES menu item. The dialog consists of two lists. The left list shows all variables that are not sort or dependent variables. From those variables you can select the variable that defines the levels of the factor for the independent measures ANOVA. If you have selected the variable that defines the levels of the factor, the right list will contain all levels that are included in this variable. From these levels (in the right list) you can select those levels, that you want to include in the independent measures ANOVA. For example you may want to select the variable "agegroup" to define the levels of the factor for the independent measures ANOVA. This factor consists of four levels (i.e. mature, teen, child, and senescent). By default, all levels are selected. If you don't want to exclude any levels you can start calculation of the ANOVA by clicking on the OK-button. The results will be displayed in the results window.

For each subgroup as many independent measures ANOVAs are calculated as you have selected dependent variables. In the ANOVA table (see Table 8.5) the total variance in all levels of the factor is separated into the inter-individual variance (between) and the variance caused by the sampling variability (error). For the inter-individual variance a F-value and a corresponding p-value is calculated. A small p-value (p.e. $p<0.05$) means that at least two levels differ from each other. In our example it would mean that the IQ-scores reached by subjects belonging to different agegroups differ. Or in other words, IQ-scores would be dependent on the age of the subjects. However, since in our example $p=0.8412$ (see Table 8.5) there are no effects of age on the IQ-scores. This makes sense, since IQ-scores are already age-adjusted.

week 1	df	sum of squares	mean squares	F	p
between	3	143.1667	47.7222	0.2764	0.8412
error	10	1726.8333	172.6833		
total	13	1870.0000			

Table 8.5: 1-way-ANOVA table for independent measures

Calculation of the various values in the independent measures ANOVA

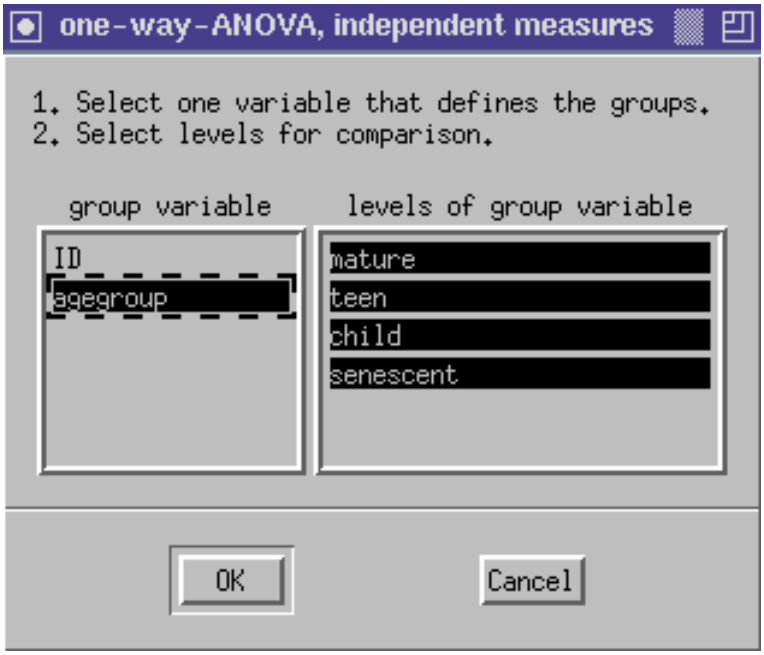


Figure 8.7: The ANOVA-one-way, Independent Measures dialog

table is performed according to the following scheme:

Source	df	sum of squares	mean squares	F	p
between	p-1	SSG	MSG	MSG/MSE	$F_{p-1,\ n-p}$
error	n-p	SSE	MSE		
total	n-1	SS (Total)			

p: number of groups (independent measurements)
n_i: number of observations (subjects) in the i^{th} group
n: total number of data values ($\sum_{i=1}^{p} n_i$)
$\bar{x}_i$: mean of all subjects in the i^{th} group
x_{ij}: j^{th} data value in the i^{th} group
$\bar{x}$: mean of all data values in the database

$$SSG\ (group) = \sum_{i=1}^{p} n_i(\bar{x}_i - \bar{x})^2$$

$$SSE\ (error) = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$SS\ (total) = SSG\ +\ SSE$$

$$MSG = \frac{SSG}{p-1}$$

$$MSE = \frac{SSE}{n-p}$$

8.6.3 Post-hoc tests

An ANOVA is usually computed if the mean values of more than two normally distributed groups need to be compared. However, the ANOVA only answers the question as to whether at least two groups differ from each other. To investigate which of the groups differ from each other, post-hoc tests can be calculated. Currently, the Scheffé, Tukey, Newman-Keuls, and Fisher post-hoc tests are supported in XmStat.

Scheffé test

The Scheffé test is valid only, if the ANOVA table reveals a significant p-value. A $\hat{F_S}$ is calculated and based on ν_1 and ν_2 degrees of freedom a p-value is obtained from the F-distribution. $\hat{F_S}$, ν_1, and ν_2 are calculated according to:

ν_1:	p-1
ν_2:	n-p
p:	number of groups
n_i:	number of data values in the i^{th} group
n:	total number of data values in the database ($\sum_{i=1}^{p} n_i$)
$\bar{x_i}$:	mean value of the i^{th} group
MSE:	MSE from ANOVA table

$$\hat{F_S} = \frac{(\bar{x_a} - \bar{x_b})^2}{MSE\left(\frac{1}{n_a} + \frac{1}{n_b}\right)(k-1)}$$

Fisher test

The Fisher test is also called the "least significant difference method (LSD)" and is based on the t-test. Like the Scheffé test, this test is only valid if the p-value in the ANOVA table is significant. If the ANOVA is significant, then each mean is compared with each other mean using a t-test. Since homogeneity of variance is typically assumed for Fisher's LSD procedure, the estimate of variance is based on all the data, not just on the data for the two groups being compared. In order to make the relationship between Fisher's LSD and other methods of computing pairwise comparisons clear, the formula for the studentized t (t_s) rather than the usual formula for t is used:

p:	number of groups
n_i:	number of data values in the i^{th} group
n:	total number of data values in the database ($\sum_{i=1}^{p} n_i$)
$\bar{x_i}$:	mean value of the i^{th} group
MSE:	MSE from ANOVA table
n_h:	harmonic mean of the sample sizes of the two groups $\frac{2}{\frac{1}{n_a}+\frac{1}{n_b}}$

$$t_s = \frac{\bar{x_a} - \bar{x_b}}{\sqrt{\frac{MSE}{n_h}}}$$

Based on t_s, a number of degrees of freedom according to MSE ($df = n - p$), and two mean values, a p-value is obtained from the studentized range distribution. The Fisher test may be appropriate if less than six groups are compared.

Newman-Keuls test

For the Newman-Keuls test, the studentized t_s-value is calculated in the same way as for the Fisher test. However, the p-value is obtained from the studentized range distribution based on the same degrees of freedom as in the Fisher procedure, but on a number of mean values that equals the difference of the ranks of the mean values of the two groups that are compared plus one. To this end, the groups in the ANOVA are rank-ordered by the mean values of the groups from smallest to largest. Then, the smallest mean is compared to the largest mean using the studentized t_s. If the test is not significant, then no pairwise tests are significant and no more testing is done. If the difference between the largest mean and the smallest mean is significant, then the difference between the smallest mean and the second largest mean as well as the difference between the largest mean and the second smallest mean are tested. This procedure is repeated until all groups are compared. The difference to the Fisher test, however, is that the p-value is obtained from the studentized range distribution based on a number of mean values that correspondes to the number of rankes spaned by the two groups that are compared. The basic idea is that when a comparison that spans k means is significant, comparisons that span k-1 means within the original span of k means are performed. In contrast to the Scheffé and Fisher tests, the Newman-Keuls test is also valid, if the p-value calculated in the ANOVA table is not significant. The Newman-Keuls test should be used if more than seven groups are compared.

Tukey test

The Tukey test is performed in a similar way as the Fisher test. However, the p-value is obtained from t_s based on the same degrees of freedam as in the Fisher test, but based on a number of mean values that is equal to the number of groups in the ANOVA. Thus, the Tukey test is much more strict than the Fisher test, i.e. may be not significant, whereas the Fisher test is significant. The Tukey test is also valid if the p-value obtained in the ANOVA table is not significant.

8.7 ANOVA-two-way

In the case of the two-way ANOVA there are two factors with several levels each that impacts the dependent variable. Now the levels of these factors can be independent or dependent. The effect of independent factors is usually tested in different subjects, whereas the effect of dependent factors is often tested in the same individual subjects. Thus, in XmStat, independent factors are arranged as columns (variables) containing the information on the level of the independent factor for each observation. Thus, there is one variable for each independent factor. In contrast, dependent variables have one column for each level of the dependent variable, containing the values of the dependent variable for that level of the dependent variable. Thus, for dependent factors there are as many variables (columns) as there are levels in the factor.

8.7.1 Repeated-Repeated Design

This is not implemented yet.

8.7.2 Independent-Repeated Design

An example of of this design is the question, if the changes in the IQ-scores of repeated IQ-tests depend on the age. The age is the independent factor, since each subjects belongs only to one specific agegroup. In addition, the results of the repeated IQ-tests is the second, repeated (or dependent) factor, since each subject performed repeated tests. Thus, the results of later IQ-tests is related to the results of the former IQ-tests.

To actually calculate this test, we need to open the dialog box that can be reached by the STATISTICS ANOVA-TWO-WAY INDEPENDENT-REPEATED MESURES menu item (Fig. 8.8). In our example, we would select the agegroup as the independent factor. This automatically selects all levels of this

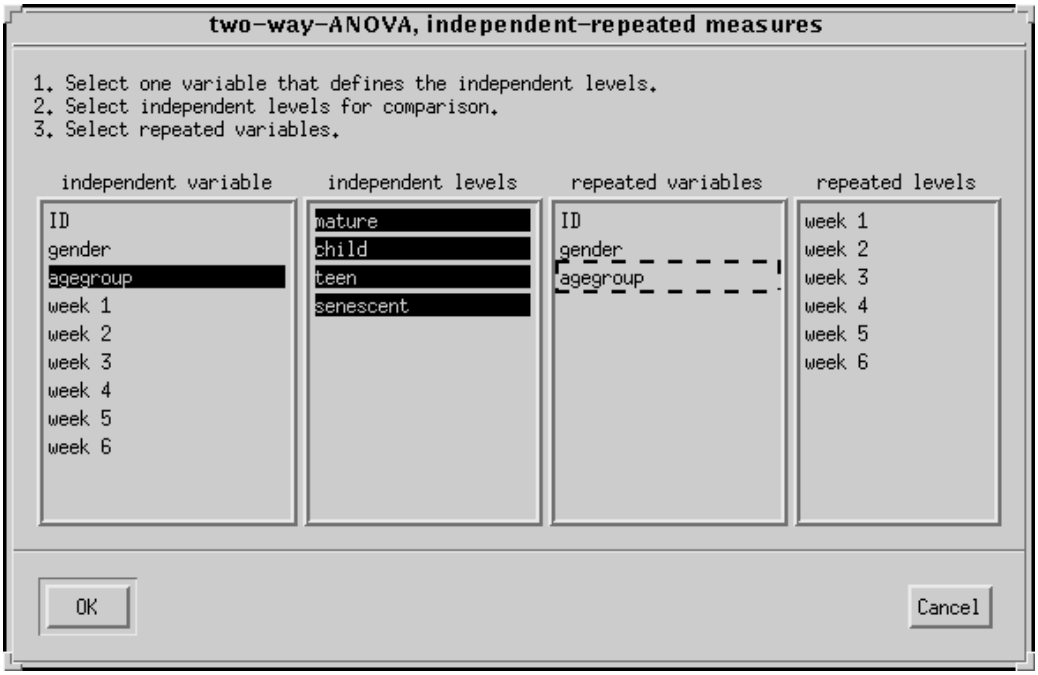


Figure 8.8: The ANOVA-two-way, Independent-Repeated Measures dialog

factor (mature, child, teen, and senescent). Now we need to select the repeated variables, e.g. week 1, week 2, week 3, week 4, week 5, and week 6. Then we can start the calculation by clicking on the OK button. This opens the results window shown in table 8.6.

This table provides P values for the between factor (agegroup), the within factor (time point of IQ-test) and the interaction of the within and between factors. The interaction is significant, if it depends on the agegroup, if the results of the IQ-test depends on the time when the test was performed. In our example, the factor agegroup is not significant ($P>0.05$). Thus, there is no difference in the IQ in different agegroups. This is the anticipated result, since the IQ-tests are usually standardized for age. However, the within factor is significant ($P<0.05$), indicating, that the results of the IQ-tests changes when the tests are repeated on successive weeks. Specifically, the result of the IQ-test becomes better, the more often the test is repeated. This may reflect a "learning effect". Finally, the interaction is not significant ($P>0.05$) indicating, that the age does not modify the behaviour of the test results over time. Specifically, in all agegroups, the test results become better when the tests are repeated.

ANOVA (two-way) for independent and repeated measures

Source	df	Sum of Square	Mean Square	F	P
Between	25	22880.00464			
agegroup	3	872.30347	290.76782	0.291	0.831655
error	22	22007.70117	1000.35004		
Within	130	4370.83516			
repeated	5	1791.56934	358.31387	16.243	0.000000
interaction	15	152.67256	10.17817	0.461	0.954770
error	110	2426.59326	22.05994		

Table 8.6: 2-way-ANOVA table for independent and repeated measures

Calculation of the values in the 2-way-ANOVA table for independent and repeated measures is rather complicate. Especially helpful in implementing this function was the book by Brandt [1] and the Internet page by David M. Lane [4]. The details are provided in the following schematic:

Source	df	Sum of Square	Mean Square	F	P
Between	n-1	SSB + SSBE			
agegroup	i-1	SSB	MSB	MSB/MSBE	$F_{i-1,\ n-i}$
error	n-i	SSBE	MSBE		
Within	n*(j-1)	SSW+SSBW+SSBWE			
repeated	(j-1)	SSW	MSW	MSW/MSBWE	$F_{j-1,\ (n-i)(j-1)}$
interaction	(i-1)*(j-1)	SSBW	MSBW	MSBW/MSBWE	$F_{(i-1)(j-1),\ (n-i)(j-1)}$
error	(n-i)*(j-1)	SSBWE	MSBWE		

n: number of subjects
i: number of levels of the independent factor
j: number of levels of the repeated factor
$\bar{x}$: mean of all values
ni: number of values in an independent subgroup
$\bar{xi}$: mean of all values in an independent subgroup
nsi: number of subjects in an independent subgroup
$\bar{xsi}$: mean of all repeated values of only one subject
$\bar{xir}$: mean of all values of a repeated variable for only one independent subgroup
nj: number of values in a repeated variable
$\bar{xj}$: mean of all values in a repeated variable

$$SSB\ (between) = \sum_{k=1}^{i} ni_k(\bar{xi}_k - \bar{x})^2$$

$$SSBE\ (between\ error) = \sum_{k=1}^{i}\sum_{l=1}^{nsi_k} j(\bar{xsi}_{l,k} - \bar{xi}_k)^2$$

$$SSW\ (within) = \sum_{m=1}^{j} nj_m(\bar{xj}_m - \bar{x})^2$$

$$SSBW\ (interaction) = \sum_{k=1}^{i}\sum_{m=1}^{j} nsi_k(\bar{xir}_{m,k} + \bar{x} - \bar{xi}_k - \bar{xj}_m)^2$$

$$SSBWE\ (interaction\ error) = \sum_{k=1}^{i}\sum_{l=1}^{nsi_k}\sum_{m=1}^{j} (x_{m,l,k} - \bar{x})^2 - SSB - SSBE - SSW - SSBW$$

$$MSB = \frac{SSB}{i-1}$$

$$MSBE = \frac{SSBE}{n-i}$$

$$MSW = \frac{SSW}{j-1}$$

$$MSBW = \frac{SSBW}{(i-1)*(j-1)}$$

$$MSBWE = \frac{SSBWE}{(n-i)*(j-1)}$$

8.7.3 Independent-Independent Design

A two-way ANOVA with two independent factors would be appropriate, if the question is how the IQ-scores depend on the gender, the age, and the interaction of both parameters. Both factors (gender and age) are independent, since each subject belongs only two one gender and one agegroup. In addition, this design allows to investigate the interaction of both factors, i.e. the question, as to whether it depends on the gender, if there are effects of age on the IQ-scores. Alternatively, the question may be, if it depends on the age, if there is a dependency of the IQ-scores on gender.

Fig. 8.9 shows the dialog box for the two-way ANOVA with two independent measures. Before opening this dialog box (STATISTICS, ANOVA-TWO-WAY, INDEPENDENT-INDEPENDENT MEASURES), dependent variables must be selected, on which the ANOVA will be calculated. Then, the two independent factors and the levels of the two factors must be selected. Finally, a method must be selected for the calculation of the sums of squares in the case of an unbalanced design. An unbalanced design is a design, in which the number of subjects in each subgroup (e.g. gender-age combination) is not the same. The dataset iq.stat, provided in the example directory reveals an unbalanced design, since there are 4 mature females, but only 2 senescent male subjects. Finally, we can start the calculation by clicking on the OK BUTTON. This opens the results window shown in Table 8.7.

The ANOVA-table provides three P-values, one for each independent variable and one for the interaction of both variables. In this example, there is a

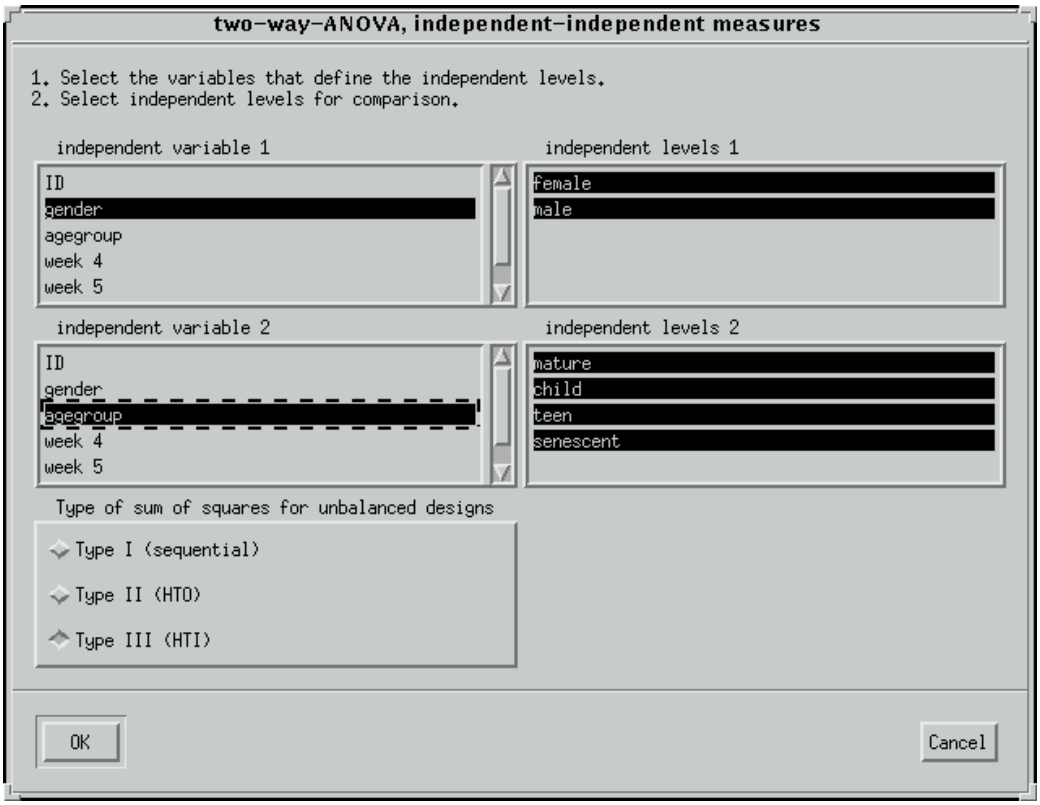


Figure 8.9: The ANOVA-TWO-WAY, INDEPENDENT-INDEPENDENT MEASURES dialog

clear effect of the gender (males have higher IQ-scores), while the agegroup does not matter (IQ-scores are adjusted for age). The interaction is also not significant ($P>0.05$). If it were significant, it would mean that it depends on the gender, if there is a change in the IQ-scores with age.

Calculation of the values in the 2-way-ANOVA table for independent measures is straight forward, if there is a balanced design (same number of observations in each subgroup). In the case of an unbalanced design, calculation of the sums of squares becomes difficult. In this software, the algorithm described by Donal B. Macnaughton [5, 6] was used. This algorithm provides three different types of sums of squares:

- Type I: sequential sums of squares. The sums of squares depend on the order in which the two factors are defined. This is typically not desired.
- Type II: Higher-level Terms are Omitted (HTO). With this method,

ANOVA (two-way) for independent measures
(Type III (HTI) for unbalanced designes)

week 1	df	sum of squares	mean squares	F	P
gender	1	2245.68774	2245.68774	14.921513	0.001140
agegroup	3	336.91891	112.30630	0.746221	0.538530
Interaction	3	991.09424	330.36475	2.195115	0.123870
Error	18	2709.00000	150.50000		
Total	25	6064.46094			

Table 8.7: 2-way-ANOVA table for independent measures

the sums of squares do not add up to the same value as SST.

- Type III: Higher-level Terms are Included (HTI). With this method, the sums of squares do not add up to the same value as SST. This is typically the best choice.

Type III, marginal sums of squares are corrected for as many other factors in the model as possible. They also provide estimates which are not a function of the frequency of observations in any group, i.e. for unbalanced data structures, where we have unequal number of observations in each group, the group(s) with more observations do not per se have more importance than group(s) with fewer observations. For purely nested designs, some polynomial regressions, and some models involving balanced data fitted in the right order, we can sometimes need Type I, sequential sums of squares.

Source	df	Sum of Square	Mean Square	F	P
ind. var. 1	i-1	SSV1	SSV1/(i-1)	(SSV1/(i-1))/MSE	$F_{i-1,\ dfe}$
ind. var. 2	j-1	SSV2	SSV2/(j-1)	(SSV2/(j-1))/MSE	$F_{j-1,\ dfe}$
interaction	(i-1)*(j-1)	SSI	SSI/((i-1)*(j-1))	(SSI/((i-1)*(j-1)))/MSE	$F_{(i-1)(j-1),\ dfe}$
error	dfe	SSE	MSE		
total	n-1	SST			

i	number of levels for factor 1
j	number of levels for factor 2
n	total number of observations in ANOVA
dfe	n-1-(i-1)-(j-1)-(i-1)*(j-1)
SSV1	sum of squares for factor 1
SSV2	sum of squares for factor 2
SSI	sum of squares for interaction
SSE	sum of squares for error term
SST	total sum of squares
MSE	SSE/dfe

8.8 Non-Parametric Tests

These tests differ from the t-test or the ANOVA in that they do not require that the data were sampled from a normal distribution. As with the t-test and the ANOVA there are tests available for comparing only two groups or two repeated measures versus more than two groups or repeated measures. Furthermore, appropriate tests for paired and unpaired samples must be selected.

8.8.1 Wilcoxon Test

This test can be used if in each subject, two repeated measurements were done. The test will then determine if the first measurement differ from the second. In addition, the test does not require normal-distribution of the parameters measured or the populations from which the data are sampled.

8.8.2 Mann-Whitney U-Test

This test allows to compared data sampled from two independent groups. Again, no normal-distribution of the parameters measured or the population from which the data are sampled is required.

An example (based on the example data set iq.stat) would be if we are interested in the question if there are gender differences in the intelligent quotients (IQ) in children, teens, mature, and senescent test persons at the first time, the IQ-test is presented to the candidates. We would select the variable *agegroup* as a sort variable and the variable *week1* as the dependent variable. Then we would select the *Mann-Whitney U-test* from the *Non-Parametric* submenu of the *Statistics* menu. Now we select the variable *gender* as the variable that defines the two groups. Of course, this variable has only two levels (male and felmale) that are already selected as the levels that should be analysed. The results window that appears after hitting OK reveals that a significant gender difference exists in mature and senescent candidates.

The Mann-Whitney U-test is based on a rank-sum test. Basically, the data from both groups are combined and sorted in ascending order. The smallest value will be assigned a rank of 1, the second smallest a rank of 2 etc. Equal observations are assigned the mean of the ranks occupied by those observations. Then the ranks of the samples of the group with the smaller number of observations will be added. If both groups have the same sample size, it does not matter which group is used to calculate the sum of ranks. This sum is referred to as the rank-sum (R). Now the mean and the variance

of the (normal) distribution of the R-values is calculated according to (n_1 and n_2 are the number of observations in the smaller (n_1) and larger (n_2) groups):

$$\mu_R = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma_R^2 = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

The R-value is then transformed to a standard normal distribution (with mean 0 and variance 1) by:

$$z = \frac{R - \mu_R}{\sigma_R}$$

Finally, the P-value that corresponds to the z-value is looked up in the normal distribution function.

Bibliography

[1] Siegmund Brandt. *Data Analysis - Statistical and Computational Methods for Scientists and Engineers.* Springer, New York, Berlin, Heidelberg, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo, 1997. [ISBN 0-387-98498-4].

[2] H. T. Hayslett, Jr. *Statistics made simple.* Doubleday, New York, London, Toronto, Sydney, Auckland, 1 edition, 1968. ISBN 0-385-02355-3.

[3] Dan Heller and Paula M. Ferguson. *Motif programming manual.* O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2 edition, 1994. ISBN 1-56592-016-3.

[4] David M. Lane. HyperStat online. *Internet*, 2000. [http://davidmlane.com/hyperstat/index.html].

[5] Donald B. Macnaughton. Computing numerator sums of squares in unbalanced analysis of variance: two-way case. *http://www.matstat.com.*

[6] Donald B. Macnaughton. Which sums of squares are best in unbalanced analysis of variance? *http://www.matstat.com*, 1998.

[7] James T. McClave, Frank H. Dietrich, and Terry Sincich. *Statistics.* Prentice-Hall International, Inc., Upper Saddle River, NJ 07458, USA, 7th edition, 1997. ISBN 0-13-492950-0.

[8] Willian H. Press, William T. Vetterling, Saul A. Teukolsky, and Brian P. Flannery. *Numerical recipes in C.* Cambridge University Press, Cambridge, New York, Melbourne, 2nd edition, 1992. ISBN 0-521-43108-5 ISBN 0-521-43720-2 ISBN 0-521-43714-8 ISBN 0-521-43724-5 ISBN 0-521-43715-6.

[9] Royal Statistical Society. StatLib - applied statistics algorithms. *Applied Statistics*, 1968. [http://lib.stat.cmu.edu/].

[10] Lothar Sachs. *Statistische Methoden.* Springer Verlag, Berlin, Heidelberg, New York, 5 edition, 1982. ISBN 3-540-11762-8 ISBN 0-387-11762-8.

[11] Herbert Schildt. *C++ from the ground up.* Osborn McGraw-Hill, Berkeley, CA, USA, 1 edition, 1994. ISBN 0-07-881969-5.

[12] Thor Sigvaldason. xldlas - a program for statistics. *Linux Journal*, 34:66–68, February 1997. ftp://sunsite.unc.edu/pub/Linux/X11/xapps/math/.

[13] Forrest W. Young. *ViSta the visual statistics system.* Young, Forrest W., Chapel Hill, NC, USA, 1996. http://forrest.psych.unc.edu ftp://ftp.psych.unc.edu/pub/forrest/vista.