© 1986 Gary Perlman

Section 1.1 Capabilities and Requirements

|STAT is a small statistical package I have developed on the UNIX operating system (Ritchie & Thompson, 1974) at the University of California San Diego and at the Wang Institute of Graduate Studies. Over twenty programs allow the manipulation and analysis of data and are complemented by this documentation and manual entries for each program. The package has been distributed to hundreds of UNIX sites and the portability of the package, written in C (Kernighan & Ritchie, 1979), was demonstrated when it was ported from UNIX to MSDOS at Cornell University on an IBM PC using the Lattice C compiler. This handbook is designed to be a tutorial introduction and reference for the most popular parts of release 5.3 of |STAT (January, 1987) and updates through February, 1987. Full reference information on the programs is found in the online manual entries and in the online options help available with most of the programs.

Dataset Sizes

|STAT programs have mostly been run on small datasets, the kind obtained in controlled psychological experiments, not the large sets obtained in surveys or physical experiments. The programs' performances on datasets with more than about 10,000 points is not known, and the programs should not be used for them.

System Requirements

The programs run on almost any version of UNIX. They are compatible with UNIX systems dating back to Version 6 UNIX (circa 1975). On MSDOS, the programs run on versions 2.X through 3.X. MSDOS versions earlier than 2.0 may not support the pipes often used with |STAT programs, and MSDOS version 4.0 formats are not compatible. Space requirements for MSDOS are about 1 megabyte of disk space, and at least 96 kilobytes of main memory. Hard disk storage is preferred, but not mandatory.

Availability and Distribution

Please take care to follow all the instructions below.

- Please indicate the items that you would like to order.
- Orders must be prepaid. Purchase orders are not acceptable.
- Make your check/(postal)money order payable to G. Perlman.
- Checks must be in US funds drawn on a US bank.
- Please include a delivery address label to speed service.
  International orders: please indicate your country name.

UNIX C Source Version of |STAT: $20/30
C Language Source Code & Online Manual Entries
1/2 inch 9 track mag tape, 1600 bpi tar format ($20)
1/4 inch cartridge tape, tar format ($30)
DOS Executable Version of |STAT: $15
Preformatted Manuals & Executables (without Source Code) ($15)
2S/2D (360K) DOS 5.25" floppy diskettes
HD (1.2M) DOS 5.25" or 3.5" floppies (by special request)
DOS Turbo C Source Code Version for |STAT: $10
Turbo C Language Source Code, Project Files, Online Manual
HD (1.2M) DOS 5.25"inch or 3.5" floppy diskette
Handbook (highly recommended for new users): $10
Examples, Ref. Materials, CALC & DM Manuals, Manual Entries
Typeset Manual (over 100 8.5 x 11 inch pages)
Prices include cost of media and airmail delivery worldwide.

© 1986 Gary Perlman

Section 1.2 Design Philosophy

|STAT programs promote a particular style of data analysis. The package is interactive and programmable. Data analysis is typically not a single action but an iterative process in which a goal of understanding some data is approached. Many tools are used to provide several analyses of data, and based on the feedback provided by one analysis, new analyses are suggested.

The design philosophy of |STAT is easy to summarize. |STAT consists of several separate programs that can be used apart or together. The programs are called and combined at the command level, and common analyses can be saved in files using UNIX shell scripts or MSDOS batch files.

Understanding the design philosophy behind |STAT programs makes it easier to use them. |STAT programs are designed to be tools, used with each other, and with standard UNIX and MSDOS tools. This is possible because the programs make few assumptions about file formats used by other programs. Most of the programs read their inputs from the standard input (what is typed at the keyboard, unless redirected from a file), and all write to the standard output (what appears on the screen, unless saved to a file or sent to another program). The data formats are readable by people, with fields (columns) on lines separated by white space (blank spaces or tabs). Data are line-oriented, so they can be operated on by many programs. An example of a filter program on UNIX and MSDOS that can be used with the |STAT programs is the **sort** utility, which puts lines in numerical or alphabetical order. The following command sorts the lines in the file **input** and saves the result in the file **sorted**.

```
sort < input > sorted
```

The **<** symbol causes **sort** to read from **input** and the **>** causes **sort** to write to the file **sorted**. Because **sort** exists on UNIX and MSDOS, it is not necessary to duplicate its function in |STAT, which does not duplicate existing tools. (In all following examples, **this font will be used** to show text (e.g., commands and program names) that would be seen by people using the programs.

User efficiency is supported over program efficiency. That does not mean the programs are slow, but ease-of-use is not sacrificed to save computer time. Input formats are simple and readable by people. There is extensive checking to protect against invalid analyses. Output formats of analysis programs are designed to be easy to understand. Data manipulation programs are designed to produce uncluttered output that is ready for input to other programs.

On UNIX and MSDOS, a *filter* is a program that reads from the standard input, also called **stdin** (the keyboard, unless redirected from a file) and writes to the standard output, also called **stdout** (the screen, unless redirected to a file). Most |STAT programs are filters. They are small programs that can be used alone, or with other programs. |STAT users typically keep their data in a *master data file*. With data manipulation programs, extractions from the master data file are transformed into a format suitable for input to an analysis program. The original data do not change, but copies are made for transformations and analysis. Thus, an analysis consists of an extraction of data, optional transformations, and some analysis. Pictorially, this can be shown as:

```
data | extract | transform | format | analysis | results
```

where a copy a subset of the data has been extracted, transformed, reformatted, and analyzed by chaining several programs. Data manipulation functions, sometimes built into analysis programs in other packages, are distinct programs in |STAT. The use of pipelines, signaled with the pipe symbol, **|**, is the reason for the name |STAT.

© 1986 Gary Perlman

Section 1.3 Table of |STAT Programs

|STAT programs are divided into two categories. There are programs for data manipulation: data generation, transformation, formatting, extraction, and validation. And there are programs for data analysis: summary statistics, inferential statistics, and data plots. The data manipulation programs can be used for tasks outside of statistics.

Data Manipulation Programs

`abut`	join data files beside each other
`colex`	column extraction/formatting
`dm`	conditional data extraction/transformation
`dsort`	multiple key data sorting filter
`linex`	line extraction
`maketrix`	create matrix format file from free-format input
`perm`	permute line order randomly, numerically, alphabetically
`probdist`	probability distribution functions
`ranksort`	convert data to ranks
`repeat`	repeat strings or lines in files
`reverse`	reverse lines, columns, or characters
`series`	generate an additive series of numbers
`transpose`	transpose matrix format input
`validata`	verify data file consistency

Data Analysis Programs

`anova`	multi-factor analysis of variance
`calc`	interactive algebraic modeling calculator
`contab`	contingency tables and chi-square
`desc`	descriptions, histograms, frequency tables
`dprime`	signal detection d' and beta calculations
`features`	display features of items
`oneway`	one-way anova/t-test with error-bar plots
`pair`	paired data statistics, regression, scatterplots
`rankind`	rank order analysis for independent conditions
`rankrel`	rank order analysis for related conditions
`regress`	multiple linear regression and correlation
`stats`	simple summary statistics
`ts`	time series analysis and plots

© 1986 Gary Perlman

Section 1.4 Table of UNIX and MSDOS Utilities

The UNIX and MSDOS environments are similar, at least as far as |STAT is concerned, but many command names differ. The following table shows the pairing of UNIX names with their MSDOS equivalents.

UNIX	MSDOS	Purpose
`cat`	`type`	print files to stdout
`cd,pwd`	`cd`	change/print working directory
`cp`	`copy`	copy files
`diff`	`comp`	compare and list file differences
`echo`	`echo`	print text to standard output
`grep`	`find`	search for pattern in files
`ls`	`dir`	list files in directory
`mkdir`	`mkdir`	create a new directory
`more`	`more`	paginate text on screen
`mv`	`rename`	move/rename files
`print`	`print`	print files on printer
`rm`	`del,erase`	remove/delete files
`rmdir`	`rmdir`	remove an empty directory
`sort`	`sort`	sort lines in files
`shell-script`	`batch-file`	programming language
`$1,$2`	`%1,%2`	variables
`/dev/tty`	`con`	terminal keyboard/screen
`/dev/null`	`nul`	empty file, infinite sink

© 1986 Gary Perlman

CHAPTER 1

Introduction

1.1 **Capabilities and Requirements** 1-1
Dataset Sizes
System Requirements
Availability and Distribution
1.2 **Design Philosophy** 1-2
1.3 **Table of |STAT Programs** 1-3
Data Manipulation Programs
Data Analysis Programs
1.4 **Table of UNIX and MSDOS Utilities** 1-4

The purpose, environment, and philosophy of the |STAT programs are introduced.

© 1986 Gary Perlman