Analytical Studies Branch

75 Years of Excellence / Ans d'excellence

Research Paper Series

Statistics Canada / Statistique Canada

Canada

MAINFRAME SAS ENHANCEMENTS IN SUPPORT OF EXPLORATORY DATA ANALYSIS

by
Richard Johnson [1], Jane F. Gentleman [2], and Monica Tomiak [3]
No. 24

Social and Economic Studies Division
Analytical Studies Branch
Statistics Canada
1989
Revised 1992

1. Main Computer Centre Division, Informatics Branch.
2. Social and Economic Studies Division, Analytical Studies Branch and Canadian Centre for Health Statistics, Institutions and Social Statistics Branch.
3. Social and Economic Studies Division, Analytical Studies Branch.

Aussi disponible en français

Mainframe SAS Enhancements in Support of Exploratory Data Analysis

by Richard Johnson, Jane F. Gentleman, and Monica Tomiak

ABSTRACT

This document is a manual describing computer software developed for exploratory data analysis at Statistics Canada. The software comprises a collection of SAS functions and macros, with heavy emphasis on graphics. Together with some functions already available in the SAS system, these routines perform the following types of operations: evaluation of probability density functions (PDF's), cumulative distribution functions (CDF's), and inverse CDF's for nine distributions; calculation of sample quantiles; generation of random numbers; graphing of histograms with optional PDF superimposition; calculation of empirical CDF's with optional graphing and optional CDF superimposition; and graphing of Q-Q and P-P plots comparing a sample to any of the nine distributions. Detailed instructions are provided for using the software to construct Q-Q plots comparing two samples.

Key Words : Exploratory data analysis
Graphics

Received: May 5, 1989
Accepted: October 10, 1989
Second Edition Accepted : June 18, 1992

Foreword to Second Edition

This document is a manual describing computer software for exploratory data analysis, the result of a joint effort between members of the Informatics Branch and the Analytical Studies Branch at Statistics Canada. The manual is being distributed by the Informatics Branch to the computer user community within Statistics Canada, and it also appears here as a paper in the Analytical Studies Branch Research Paper Series. It is being re-published under the latter auspices because the Analytical Studies Branch Research Paper Series is intended to represent the broad array of activities being carried out within the Branch, including the production of research papers eventually destined for refereed journal publication, as well as other types of activities which support research and analysis.

The software described herein was originally developed for use by students in a data analysis class offered by Statistics Canada as part of its on-going effort to increase data analytic capability within the agency. The software is now available for general use by Statistics Canada personnel.

This Second Edition of the manual describes how the software has been updated to incorporate weights.

Following the manual are examples of graphs produced by the macros.

It is not the intention of the Informatics Branch to distribute the software outside Statistics Canada. However, those who wish may request a photocopy of the new source code from the third author.

Steven Earwaker coordinated and managed the data analysis class for which this software was produced. Louise Bergeron provided expert programming assistance.

Richard Johnson, Informatics Branch
Jane F. Gentleman, Analytical Studies Branch
Monica Tomiak, Analytical Studies Branch

Mainframe SAS Enhancements

in Support of

Exploratory Data Analysis

SAS Support Staff, Informatics Branch
and
Analytical Studies Branch
Statistics Canada

April 1992 Edition

PREFACE

The April 1992 Edition of this manual has been updated to describe modifications of the software to optionally incorporate survey weights.

This document describes SAS functions and macros developed specifically in support of the pilot presentation of Statistics Canada course 0418E entitled "The Art of Data Analysis". These facilities have been implemented on the Statistics Canada mainframe computer system. It is assumed that the reader is generally familiar with the mainframe operating environment and SAS Version 5 as configured for that environment.

Eighteen functions were written to supplement those already available in the SAS system. When viewed together with nine of the original SAS functions, they constitute a collection of routines to evaluate three types of statistical probability functions: probability density functions (PDF's), cumulative distribution functions (CDF's), and inverse cumulative distribution functions. These theoretical functions are useful tools in data analysis. In addition, macros have been developed to perform more complex operations involving both the theoretical distributions and data. In the following list, some uses for the software are described.

- Plots of PDF's can be compared to appropriately constructed histograms. A SAS macro has been developed to produce superimposed plots of this nature.

- CDF's are useful for calculating significance levels (P-values), calculating certain goodness-of-fit test statistics, calculating coordinates for P-P plots, and comparing theoretical to empirical cumulative distribution functions (ECDF's). A SAS macro has been developed to compute and plot ECDF's and optionally superimpose CDF's. Another macro has been developed to produce P-P plots.

- Inverse CDF's can be used to calculate theoretical quantiles, calculate coordinates for Q-Q plots, and generate random numbers. A SAS macro has been developed to produce Q-Q plots.

- A SAS macro for calculating sample quantiles has been developed. This is useful for analyzing the distribution of a sample of data and for calculating coordinates for Q-Q plots.

- ECDF's, histograms and quantiles can be calculated, and Q-Q plots and P-P plots can be graphed for data which are accompanied by a vector of weights. These weights correspond to the relative number of population units represented by each data value.

- SAS macros have been developed, based on available SAS random number functions, to facilitate the generation of random numbers for five distribution types.

Sections 1 and 2 of this document provide detailed descriptions of how to use the SAS functions and macros. Section 3 describes the minimal operational considerations associated with the production of graphics, and provides the user with instructions for obtaining additonal copies of this document. An Appendix contains formulas for the probability density functions.

The pilot version of "The Art of Data Analysis" was designed and presented by Dr. Jane Gentleman of Social and Economic Studies Division, Analytical Studies Branch. Coordination and logistics were managed by Stephen Earwaker of Social Survey Methods Division, Methodology Branch. For further details about the incorporation of weights in these techniques, see "Calculations of Empirical Distribution Functions, Quantiles, and Histograms for Samples of Weighted Observations" by Jane F. Gentleman. This document and the special SAS facilities described herein were created by the SAS support staff of Informatics Branch, and by members of the Analytical Studies Branch, Statistics Canada.

CONTENTS

TABLES

Section 1

PROBABILITY FUNCTIONS

The SAS system provides a variety of probability functions as documented in Chapter 6 of "SAS User's Guide: Basics, Version 5 Edition". Numerous additional functions have been written, using VS Fortran, to complement those provided by SAS.

1.1 PDF's, CDF's and Inverse CDF's for Nine Distributions

The comprehensive set of relevant functions consists of three functions for each of nine distribution types. The functions are: the probability density function (PDF); the cumulative distribution function (CDF); and the inverse of the cumulative distribution function (inverse CDF). Table 1 lists the nine distributions covered, provides the name of each function, and indicates whether the function is an original SAS function or was written at Statistics Canada.

This document describes the functions written at Statistics Canada. "SAS User's Guide: Basics" is the appropriate reference for the original SAS functions. Like SAS's own functions, the locally written functions are routines that return a value computed from one or more arguments. They are used in the context of a DATA step and, typically, are executed with each iteration of the DATA step, that is, as the DATA step processes each observation in a SAS data set. Arguments to the functions documented below are positional and mandatory. They are subjected to range validation. If invalid or missing arguments are detected, messages will be written to the SAS log and results will be set to missing.

Table 1: A Comprehensive List of Probability Functions

		Functions		
Distribution	Parameters	PDF	CDF	Inverse CDF
Normal	mu,sigsq	PDFNORM	PROBNORM [1,2]	PROBIT [1,3]
Uniform	a,b	PDFUNI	PROBUNI	UNIINV
Exponential	theta	PDFEXP	PROBEXP	EXPINV
t	df	PDFT	PROBT [1]	TINV
Chi square	df	PDFCHI	PROBCHI [1]	CHIINV
F	df1,df2	PDFF	PROBF [1]	FINV
Weibull	lo,sc,sh	PDFWEI	PROBWEI	WEIINV
Gamma	lo,sc,sh	PDFGAM	PROBGAM [1,4]	GAMINV [1,5]
Beta	a,b	PDFBETA	PROBBETA [1]	BETAINV [1]

Notes:

1. Original SAS function. (All others written at Statistics Canada.)

2. PROBNORM(x) is the CDF of a N(0,1) random variable at argument x. Use PROBNORM($(x-\mu)/\sigma$) for the CDF of a N(μ,σ^2) random variable at argument x (where μ = mu and σ^2 = sigsq).

3. PROBIT(P) is the inverse CDF of a N(0,1) random variable at argument P. Use (σ)PROBIT(P)+μ for the inverse CDF of a N(μ,σ^2) random variable at argument P (where μ = mu and σ^2 = sigsq).

4. PROBGAM(x,*sh*) is the CDF of a Gamma(0,1,sh) random variable at argument x. Use PROBGAM($(x-lo)/sc$,*sh*) for the CDF of a Gamma(*lo*,*sc*,*sh*) random variable at argument x.

5. GAMINV(P,*sh*) is the inverse CDF of a Gamma(0,1,*sh*) random variable at argument P. Use (*sc*)GAMINV(P,*sh*)+*lo* for the inverse CDF of a Gamma(*lo*,*sc*,*sh*) random variable at argument P.

1.2 Probability Density Functions

1.2.1 PDFBETA

The PDFBETA function returns the probability density at argument x of a beta distribution with parameters a and b.

General form:

```
PDFBETA(x,a,b)
```

where

x The value at which the probability density is to be evaluated. $0 < x < 1$.

a First shape parameter. $a > 0$.

b Second shape parameter. $b > 0$.

1.2.2 PDFCHI

The PDFCHI function returns the probability density at argument x of a Chi-square distribution with *df* degrees of freedom.

General form:

```
PDFCHI(x,df)
```

where

x The value at which the probability density is to be evaluated. $x > 0$.

df Number of degrees of freedom. $df \geq .5$. Argument *df* need not be an integer.

1.2.3 PDFEXP

The PDFEXP function returns the probability density at argument x of an exponential distribution with mean *theta*.

General form:

```
PDFEXP(x,theta)
```

where

x The value at which the probability density is to be evaluated. $x \geq 0$.

theta Mean *theta*. *Theta* > 0.

1.2.4 PDFF

The PDFF function returns the probability density at argument *x* of an F distribution with *df1* and *df2* degrees of freedom.

General form:

```
PDFF(x,df1,df2)
```

where

x The value at which the probability density is to be evaluated. $x > 0$.

df1 Numerator degrees of freedom. $df1 > 0$. Argument *df1* need not be an integer.

df2 Denominator degrees of freedom. $df2 > 0$. Argument *df2* need not be an integer.

1.2.5 PDFGAM

The PDFGAM function returns the probability density at argument *x* of a gamma distribution with the given *location*, *scale* and *shape* parameters.

General form:

```
PDFGAM(x,lo,sc,sh)
```

where

x The value at which the probability density is to be evaluated. $x > lo$.

lo *Location* parameter. $-\infty < lo < \infty$.

sc *Scale* parameter. $sc > 0$.

sh *Shape* parameter. $sh > 0$.

1.2.6 PDFNORM

The PDFNORM function returns the probability density at argument *x* of a Normal distribution with mean *mu* and variance *sigsq*.

General form:

```
PDFNORM(x,mu,sigsq)
```

where

x The value at which the probability density is to be evaluated. $-\infty < x < \infty$.

mu Mean (μ). $-\infty < \mu < \infty$.

sigsq Variance (σ^2). $\sigma > 0$.

1.2.7 PDFT

The PDFT function returns the probability density at argument x of a t distribution with *df* degrees of freedom.

General form:

```
PDFT(x,df)
```

where

x The value at which the probability density is to be evaluated. $-\infty < x < \infty$.

df Number of degrees of freedom. $df > 0$. Argument *df* need not be an integer.

1.2.8 PDFUNI

The PDFUNI function returns the probability density at argument x of an uniform distribution on the interval $[a,b]$.

General form:

```
PDFUNI(x,a,b)
```

where

x The value at which the probability density is to be evaluated. $a \le x \le b$.

a Lower limit of the interval. $a < b$.

b Upper limit of the interval. $a < b$.

1.2.9 PDFWEI

The PDFWEI function returns the probability density at argument x of a Weibull distribution with the given *location*, *scale* and *shape* parameters.

General form:

```
PDFWEI(x,lo,sc,sh)
```

where

x The value at which the probability density is to be evaluated. $x > lo$.

lo *Location* parameter. $-\infty < lo < \infty$.

sc *Scale* parameter. $sc > 0$.

sh *Shape* parameter. $sh > 0$.

1.3 Cumulative Distribution Functions

1.3.1 PROBEXP

The PROBEXP function returns the probability that a random variable having the exponential distribution with mean *theta* is less than or equal to the input argument x.

General form:

```
PROBEXP(x,theta)
```

where

x The value at which the function is to be evaluated. $x \geq 0$.

theta Mean *theta*. *Theta* > 0.

1.3.2 PROBUNI

The PROBUNI function returns the probability that a random variable having the uniform distribution on the interval (a,b) is less than or equal to the input argument x.

General form:

```
PROBUNI(x,a,b)
```

where

x The value at which the function is to be evaluated. $a \leq x \leq b$.

a Lower limit of the interval. $a < b$.

b Upper limit of the interval. $a < b$.

1.3.3 PROBWEI

The PROBWEI function returns the probability that a random variable having the Weibull distribution with the given *location*, *scale* and *shape* parameters is less than or equal to the input argument *x*.

General form:

```
PROBWEI(x,lo,sc,sh)
```

where

x The value at which the function is to be evaluated. $x > lo$.

lo *Location* parameter. $-\infty < lo < \infty$.

sc *Scale* parameter. $sc > 0$.

sh *Shape* parameter. $sh > 0$.

1.4 Inverse Cumulative Distribution Functions

1.4.1 CHIINV

The CHIINV function returns the Chi-square value *x*, such that a random variable, distributed as Chi-square with *df* degrees of freedom, is less than or equal to *x* with probability *p*.

General form:

```
CHIINV(p,df)
```

where

p Probability in range [0,1].

df Number of degrees of freedom. $df \geq .5$. Argument *df* need not be an integer.

1.4.2 EXPINV

The EXPINV function returns the exponential value *x*, such that a random variable, distributed as exponential with mean *theta*, is less than or equal to *x* with probability *p*.

General form:

```
EXPINV(p,theta)
```

where

p Probability in range [0,1].

theta Mean *theta*. *Theta* > 0.

1.4.3 FINV

The FINV function returns the F value *x*, such that a random variable, distributed as F with *df1* and *df2* degrees of freedom, is less than or equal to *x* with probability *p*.

General form:

```
FINV(p,df1,df2)
```

where

p Probability in range [0,1].

df1 Numerator degrees of freedom. *df1* > 0. Argument *df1* need not be an integer.

df2 Denominator degrees of freedom. *df2* > 0. Argument *df2* need not be an integer.

1.4.4 TINV

The TINV function returns the t value *x*, such that a random variable, distributed as t with *df* degrees of freedom, is less than or equal to *x* with probability *p*.

General form:

```
TINV(p,df)
```

where

p Probability in range [0,1].

df Number of degrees of freedom. *df* > 0. Argument *df* need not be an integer.

1.4.5 UNIINV

The UNIINV function returns the uniform value *x*, such that a random variable, distributed as uniform on the interval (*a*,*b*), is less than or equal to *x* with probability *p*.

General form:

```
UNIINV(p,a,b)
```

where

p Probability in range [0,1].

a Lower limit of the interval. $a < b$.

b Upper limit of the interval. $a < b$.

1.4.6 WEIINV

The WEIINV function returns the Weibull value x, such that a random variable, distributed as Weibull with the given *location*, *scale* and *shape* parameters, is less than or equal to x with probability p.

General form:

```
WEIINV(p,lo,sc,sh)
```

where

p Probability in range [0,1].

lo *Location* parameter. $-\infty < lo < \infty$.

sc *Scale* parameter. $sc > 0$.

sh *Shape* parameter. $sh > 0$.

Section 2

MACROS

This section describes SAS macros written to perform operations on entire SAS data sets. The macros consist of macro statements, and data step programming statements and/or complete DATA and PROC steps. All of the macros were written for name-style macro calls; that is, the form of the invocation is *%macroname(parameters)* as described in Chapter 19 of "SAS User's Guide: Basics, Version 5 Edition".

Several of the macros described in this section produce plots. All plots are produced in landscape orientation on the IBM 3800-3 laser page printer. This device has been chosen because it is available to all mainframe SAS users, and also because plots sent directly to the 3800-3 are produced in the preferred landscape orientation. (By comparison, landscape orientation of graphs on a cut-sheet 3820 laser printer requires creation of a graphics catalog, followed by template replay to effect 90-degree rotation.)

Other graphics devices can be used with the macros described in this section only by modifying the GOPTIONS statement in the macro source code. For occasional needs, this can more easily be done on-line under Display Manager where the macro can be brought into the editor screen with an INCLUDE INCL(macroname) command (where *macroname* is the name of the macro but without the leading % character), modified and SUBMITted. Once a modified macro has been submitted from the editor screen, that version will take precedence over the original version in the default autocall library whenever that macro is invoked subsequently in the SAS session. In batch mode, the user must create a modified copy of the macro in a user autocall library and then specify the root of that library to the INCL parameter of the JCL procedure. The batch approach can also be used on-line and is appropriate for regular use of an alternative graphics device such as a pen plotter or graphics display terminal. Please note that some of the plotting macros described below have subordinate macros which do the actual plotting and therefore contain the GOPTIONS statement. In the macro descriptions which follow (in the narrative portion preceding the detailed specifications), the name of the macro which contains the GOPTIONS statement is given.

By default, SAS/GRAPH will automatically scale the axes of a plot by taking into consideration the ranges of values of the variables being plotted, the alignment of tick-mark values on the axes, and space required for user-specified titles and footnotes. If several data samples having differing value ranges are plotted in this way, the axes of the plots will have different limits[1] and therefore may be given different lengths even though titling may be consistent and the same graphics device is being used. Such differences in the plots may hinder comparison of the samples. However , SAS/GRAPH provides the means to force axis consistency so that

---

[1] In this discussion of axes, the term *length* will refer to the physical length (measurable in inches or centimeters), and the term *limits* will refer to the smallest and largest numerical labels on the axes.

plots of differing samples can be compared.

Thus, each of the plotting macros can operate in either of two modes regarding axis length and limits, depending on the use of available parameters to the macros. Axis length can be determined by SAS (parameter AXES=SAS or no AXES parameter specified) or can be fixed at a predetermined percentage[2] of the dimensions of the total plotting surface (AXES=FIXED). Fixed mode imposes restrictions on the use of titles and footnotes, the restrictions varying with the choice of graphics device because of differing character cell dimensions. On the 3800-3 used by the macros, titles and footnotes of default height can be used in 2-and-1 combinations: either 2 titles and one footnote, or 2 footnotes and one title. With no footnotes, a maximum of four titles of default height can be used. Heights other than the defaults may be used on a trial-and-error basis.

In all plotting macros except for the histogram macro, axis limits can be allowed to default to those of the relevant variable, or can be specified via XAXIS and YAXIS parameters. Values specified for the XAXIS and YAXIS parameters can correspond to any of the forms values can take for the ORDER option of the SAS/GRAPH AXIS statement. Values can be specified as a list (e.g., XAXIS=1 3 5 7 9), as a range (e.g., XAXIS=1 9 or XAXIS=1 TO 9), as a range with an increment (e.g., XAXIS=1 TO 9 BY 2), or as a combination of any of these forms. In the case of specification of a range with an increment, SAS will attempt to annotate the axis tick-marks with the specified incremental values.

The X-axis and Y-axis limits for plots produced by the histogram macro are determined from the data and the interval boundaries specified by the user. They cannot be specified explicitly by the user.

2.1 Histograms with Optional PDF Superimposition

The HIST macro produces a histogram in which the widths of the vertical bars are determined by the user and need not all be the same. The macro will also create an output SAS data set having one observation for each histogram interval and containing variables for lower interval boundary, upper interval boundary, frequency, and bar height. Optionally, the macro will superimpose a user-specified theoretical probability density function (PDF) between user-specified lower and upper limits. The graph will be produced on the 3800-3 laser printer. To use an alternative graphics device, it is necessary to modify the GOPTIONS statement contained in the %HIST macro. (See the general information at the beginning of this section.)

In order to scale the histogram properly so that a PDF can be superimposed, the bar heights are calculated as follows:

$$\textit{bar height} = \frac{\textit{number of data values in the interval}}{(\textit{interval width})(\textit{total number of data values})}.$$

Thus, like a PDF, the total area of the histogram is one.

---

[2] The percentage is fixed in the macro code and cannot be modified by the user without macro modification.

If the user designates a vector of weights (corresponding to the relative number of units represented by each data value), the bar heights are calculated as follows:

$$bar\ height = \frac{sum\ of\ the\ weights\ for\ data\ values\ in\ the\ interval}{(interval\ width)(sum\ of\ the\ weights\ for\ all\ data\ values)}.$$

Parameters must be provided to identify the input SAS data set and to specify interval boundaries for the histogram. Any data value which is equal to a boundary between two bars is counted as being in the higher interval. A data value equal to the lowest/highest boundary is counted as being in the leftmost/rightmost interval. If any data values fall outside the outer interval boundaries specified by the user, no bars will be graphed for them, but they will still contribute to the denominator of the bar height formula above.

The X-axis limits for the plot of the histogram are fixed to be the lowest and highest boundary values of the histogram. The user specifies which X-axis values are to be labelled using the TICKLABS= parameter described below. Y-axis values range from 0 to the maximum bar height of the histogram, or an upper limit for the Y-axis can be specified using the YUPLIM= parameter described below.

If the appropriate parameter is provided to select a PDF, then a graph of that function will be superimposed on the histogram. In this case, the user has the option of specifying lower and upper plotting limits for the PDF via another parameter, or letting the minimum and maximum sample values be used by default.

When superimposing a PDF, the user must specify the theoretical distribution's parameter values (see Appendix A). These values are typically determined from the data, e.g., using maximum likelihood estimation. In the case of weighted data, the parameter estimates should take the weights into account.

Macro parameters are in the form *keyword*=*value* and can be specified in any order.

2.1.1 %HIST Macro

General form:

```
%HIST(parameter list)
```

Parameters:

IN= the name of the input sample SAS data set for which a histogram is to be produced. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.

INVAR= the name of the variable, in the input SAS data set identified by the IN= parameter, for which a histogram is to be produced. If this parameter is not specified, the variable name X will be assumed.

WGTVAR= the name of the variable, in the input SAS data set identified by the IN= parameter, containing the weights for the data values. This parameter is optional.

BNDS= the interval boundaries for the histogram. Specify the lower bound for each bar, proceeding from left to right, and terminate the list with the upper bound of the rightmost bar. (There are no gaps between bars.) Separate the boundary values by at least one blank. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.

AXES= the method used to determine the lengths of the axes for the histogram. AXES=SAS will allow SAS to determine the lengths of the axes. AXES=FIXED will fix the axes at a predetermined percentage of the total plotting surface dimensions. If this parameter is not specified, AXES=SAS will be assumed. (See also the discussion of axis length at the beginning of this section.)

TICKLABS= a list of X-axis values to be labelled. Separate the values by at least one blank. Normally, these are the same as the interval boundaries specified using the BNDS= parameter. Only points falling between the minimum and maximum boundary values will be labelled. If this parameter is not specified, no X-axis values will be labelled.

YUPLIM= the upper limit for the Y-axis. If this parameter is not specified, the upper limit will be the maximum bar height of the histogram.

FUNC= the PDF specification which will result in a theoretical curve being superimposed on the histogram. This parameter is optional. If it is used it must be specified in the form FUNC=*function-name*(*arguments*), where *function-name* is the name of a PDF described in Section 1.2 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as X. (This is not the same X as the default INVAR variable.)

LIMITS= the minimum and maximum values of X (argument to the PDF) at which the PDF will be evaluated. (The macro will generate 200 values of X, evenly distributed between the limits. The PDF will be evaluated and plotted as 200 points connected by straight lines.) The limits must be given in the order LIMITS=*lower upper* with at least one blank separator between the limits. This parameter will be used only if FUNC= has been specified. If FUNC= is specified and LIMITS= is not specified, then the minimum and maximum sample values will be used by default. If the lower limit specified for the PDF is greater than the minimum sample value, then the latter will be used instead as the lower limit. Similarly, if the upper limit specified for the PDF is less than the maximum sample value, then the latter will be used instead as the upper limit.

OUT= the name of the output SAS data set which will contain information about the histogram. Normally, the data set will contain one observation for each interval. If any data values fall below the lowest interval boundary, the data set will contain an extra observation describing the interval between the smallest data value and the lowest interval boundary. Similarly, an extra observation is created for values above the highest interval boundary.

The observations will contain the following variables:

LOWER the lower boundary for an interval of the histogram.

UPPER the upper boundary for an interval of the histogram.

FREQ the numerator of the bar height (see formulas above).

HEIGHT the height of the histogram bar which depicts the current interval.

INBNDS specifies whether data values in the interval fall within the bounds specified by the user. Bars will not be plotted for data values outside these bounds. INBNDS takes the value YES if data values are outside the bounds specified; otherwise INBNDS = NO.

This parameter is required in order to obtain an output data set. If it is not specified, no output data set will be produced.

The following example will produce a histogram of the variable SAMPL in SAS data set WORK.NORM using interval boundaries as specified. The histogram will be plotted as 9 bars beginning at -2.5 and ending at 3.5. It will be overlayed with a plot of the function PDFNORM for 200 values evenly distributed between -3.6 and 3.6. Arguments to PDFNORM are: the mandatory variable name X, mean 0, and variance 1. An output SAS data set named HISTINFO will be created. It will contain nine observations, one for each interval.

```
%HIST(IN=NORM,INVAR=SAMPL,
      BNDS=-2.5 -2 -1.5 -1 -.25  .25 1.5 2 2.75 3.5,
      FUNC=PDFNORM(X,0,1),LIMITS=-3.6 3.6,
      OUT=HISTINFO)
```

2.2 *ECDF's with Optional Plotting and Optional CDF Superimposition*

The ECDF macro calculates the empirical cumulative distribution function (ECDF) evaluated at each of the ordered observations of a designated variable. The ECDF is defined as follows. If $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the sample values sorted in non-descending order, then the value of the ECDF at argument $x_{(i)}$ is $G_n(x_{(i)}) = (i - .5)/n$.

If the user designates a vector of weights $w_{(1)}, w_{(2)}, \ldots, w_{(n)}$ (where $w_{(i)}$ is positive and is the weight corresponding to $x_{(i)}$), then the value of the ECDF at argument $x_{(i)}$ is

$$G_n(x_{(i)}) = \frac{\sum_{j=1}^{i} w_j - \frac{c}{2}}{\sum_{j=1}^{n} w_j},$$

where $\frac{c}{2}$ is a continuity correction. The value of c is determined by the user. Some suggested values are as follows:

c = 0 for no continuity correction, in which case, $G_n(x_{(n)}) = 1$.
c = 1, when all weights are 1, or when all weights are integers, or when there are no weights.
c = smallest weight in the sample.

The results of the calculation are placed in a new variable which is written to an output data set along with the input variable. An input data set must be specified in the parameter list. If an output data set is not specified, the input data set will be reused. It should be noted that, as a result of a sort step in the macro, the output data set will be produced in ascending order of the input variable. If the original input data set must be retained, it is necessary to specify an output data set name in the parameter list. If the user designates a vector of weights, and does not specify the type of continuity correction to be used, the smallest weight in the sample is found and is used as the default.

Optionally, a plot of the ECDF will be produced on the 3800-3 laser printer. In order to use an alternative graphics device, it is necessary to modify the GOPTIONS statement contained in the %ECDFPLT macro called by %ECDF. (See the general information at the beginning of this section.) If an ECDF plot is requested, a user-specified theoretical cumulative distribution function (CDF) can optionally be superimposed on the plot of the ECDF by specifying the appropriate parameter to select a CDF. In this case, the user has the option of specifying lower and upper plotting limits for the CDF via another parameter, or letting the minimum and maximum sample values be used by default.

When superimposing a CDF, the user must specify the theoretical distribution's parameter values (see Appendix A). These values are typically determined from the data, e.g., using maximum likelihood estimation. In the case of weighted data, the parameter estimates should take the weights into account.

Macro parameters are in the form *keyword=value* and can be specified in any order.

2.2.1 %ECDF Macro

General form:

```
%ECDF(parameter list)
```

Parameters:

IN= the name of the input SAS data set. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.

INVAR= the name of the input variable for which the ECDF is to be calculated. If this parameter is not specified, the variable name X will be assumed.

WGTVAR= the name of the variable, in the input SAS data set identified by the IN= parameter, containing the weights for the data values. This parameter is optional.

CCWT= the value of c to be used in the calculation of the continuity correction, when weights are used. If weights are used and CCWT= is not specified, the smallest weight in the sample is found and is used as the default. This parameter is optional.

NEWVAR= the name of the new variable which will contain the ECDF values. If this parameter is not specified, the name F_HAT will be used.

OUT= the name of the output SAS data set to be produced. If this parameter is not specified, a warning message will be written to the SAS log and the input data set will be reused. In this case, the sorted input variable (INVAR) and its weights (WGTVAR) will be the only variables retained from the original input data set.

PLOT= indicates whether or not an ECDF plot is to be produced on the 3800-3 laser printer. The default value is NO. Specify PLOT=YES to obtain a plot.

AXES= the method used to determine the lengths of the axes for the optional plot. AXES=SAS will allow SAS to determine the lengths of the axes. AXES=FIXED will fix the axes at a predetermined percentage of the total plotting surface dimensions. If this parameter is not specified, AXES=SAS will be assumed. If PLOT=NO is in effect, this parameter will be ignored. (See also the discussion of axis length at the beginning of this section.)

XAXIS= the X-axis limits and intermediate tick-mark values. This parameter is optional. If used, its value can correspond to any of the forms values can take for the ORDER option of the SAS/GRAPH AXIS statement. If PLOT=NO is in effect, this parameter will be ignored. (See the discussion of axis values at the beginning of Section 2 for details on the syntax of this parameter.)

YAXIS= the Y-axis limits and intermediate tick-mark values. (See XAXIS=.)

FUNC= the CDF specification which will result in a theoretical curve being superimposed on the ECDF plot produced in response to the PLOT=YES parameter. (PLOT=YES is, therefore, a prerequisite to the use of this parameter.) This parameter must be given in the form FUNC=*function-name*(*arguments*), where *function-name* is the name of a CDF described either in "SAS User's Guide: Basics, Version 5 Edition" or in Section 1.3 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as X. (This is not the same X as the default INVAR variable.)

LIMITS= the minimum and maximum values of X (argument to the CDF) at which the CDF will be evaluated. (The macro will generate 200 values of X, evenly distributed between the limits. The CDF will be evaluated and plotted as 200

points connected by straight lines.) The limits must be given in the order LIMITS= *lower upper* with at least one blank separator between the limits. This parameter will be used only if FUNC= has been specified. If FUNC= is specified and LIMITS= is not specified, then the minimum and maximum sample values will be used by default. If the lower limit specified for the CDF is greater than the minimum sample value, then the latter will be used instead as the lower limit. Similarly, if the upper limit specified for the CDF is less than the maximum sample value, then the latter will be used instead as the upper limit.

2.3 Sample Quantiles

The QUANT macro computes sample quantiles. The sample P-quantile $Q(P)$ is defined as follows. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ be the sample values sorted in non-descending order. Then $Q(P) = x_{(nP+.5)}$, where linear interpolation is used if $1 < nP+.5 < n$ and $nP+.5$ is not an integer. If $P < .5/n$, then $Q(P) = x_{(1)}$. If $P > (n-.5)/n$, then $Q(P) = x_{(n)}$. The quantity $nP+.5$ is the "index" of $Q(P)$, indicating its position among the order statistics.

If the user designates a vector of weights $w_{(1)}, w_{(2)}, \ldots, w_{(n)}$ (where $w_{(i)}$ is positive and is the weight corresponding to $x_{(i)}$), then to obtain the P-th quantile Q_p, G_n is inverted (see section 2.2). A binary search technique is used to find the interval $[G_n(x_{(i)}), G_n(x_{(i+1)}))$ (i=1,...,n-1) which contains P. Q_p is calculated as follows:

$$Q_p = x_{(1)} \text{ if } P < G_n(x_{(1)})$$

$$= x_{(i)} \text{ if } P = G_n(x_{(i)})$$

$$= x_{(n)} \text{ if } P \geq G_n(x_{(n)})$$

If Q_p falls between $x_{(i)}$'s, linear interpolation may optionally be used, in which case

$$Q_p = x_{(i)} + F$$

where :

$$F = \frac{P - G_n(x_{(i)})}{G_n(x_{(i+1)}) - G_n(x_{(i)})}(x_{(i+1)} - x_{(i)})$$

or G_n is treated as a step function, in which case

$$Q_p = x_{(i+1)}.$$

Two SAS data sets are required as input to this macro: a sample data set containing the variable from which quantiles are to be computed, and a data set containing one or more probabilities for which quantiles are to be computed. The output data set contains the input

probabilities and their corresponding quantiles and indices. Note that if no parameter is provided to name the output data set, the input data set of probabilities will be reused. In this case, all original variables will be retained and the quantiles and indices will be added.

If the user designates a vector of weights, a continuity correction may be used in calculating G_n(see section 2.2).

Macro parameters are in the form *keyword=value* and can be specified in any order.

2.3.1 %QUANT Macro

General form:

```
%QUANT(parameter list)
```

Parameters:

IN= the name of the input sample SAS data set from which quantiles are to be computed. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.

INVAR= the name of the sample variable, in the input SAS data set identified by the IN= parameter, from which quantiles are to be computed. INVAR need not be sorted. If this parameter is not specified, the variable name X will be assumed.

WGTVAR= the name of the variable, in the input SAS data set identified by the IN= parameter, containing the weights for the data values. This parameter is optional.

CCWT= the value of c to be used in the calculation of the continuity correction, when weights are used. If weights are used and CCWT= is not specified, the smallest weight in the sample is found and is used as the default. This parameter is optional.

INTERP= determines whether linear interpolation is used. If WGTVAR= is specified and INTERP=0, the ECDF is treated as a step function. If WGTVAR= is specified and INTERP is not specified, linear interpolation is used by default. This parameter is optional.

INP= the name of the input SAS data set containing one or more probabilities for which quantiles are to be computed. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate.

INPVAR= the name of the variable, in the input SAS data set identified by the INP= parameter, containing the probabilities for which quantiles are to be computed. These probabilities may be any values between zero and one and need not be sorted. If this parameter is not specified, a message will be written to the

SAS log and the macro will terminate.

OUT= the name of the output SAS data set which will contain the computed quantiles. The data set will contain one observation corresponding to each observation in the input probabilities data set. If this parameter is not specified, the input probabilities data set, specified by the INP= parameter, will be reused. In this case, all variables contained in the INP= data set will be retained.

Q= the name of the output variable which will contain the computed quantiles. If this parameter is not specified, the variable name Q will be used.

QI= the name of the output variable which will contain the computed quantile indices. If this parameter is not specified, the variable name QI will be used.

2.4 *Probability (Q-Q and P-P) Plots*

The QQ and PP macros are used to construct two types of probability plots for comparing a sample of data to a theoretical distribution: (1) Quantile-Quantile (Q-Q) plots to compare sample quantiles to theoretical quantiles, and (2) Probability-Probability (P-P) plots to compare sample cumulative proportions to theoretical cumulative probabilities. Weights may be used in constructing these plots. In addition, instructions are given below for constructing a Q-Q plot comparing two samples of data to each other.

2.4.1 Quantile-Quantile (Q-Q) Plots Comparing a Sample to a Distribution

Let $x_{(1)}, x_{(2)}, ..., x_{(n)}$ be the sample values sorted in non-descending order. Let $P_1, P_2, ..., P_k$ be a set of k probabilities selected by the user; these may be any values between zero and one and need not be sorted. Let $Q_1, Q_2, ..., Q_k$ be the corresponding sample quantiles. (The Q_i's are calculated from the entire sample of n values.)

A Q-Q plot comparing the sample to a theoretical distribution is a scatter plot of the k points $(F^{-1}(P_i),Q_i)$ (for $i = 1, ..., k$), where F^{-1} is the inverse CDF for the desired theoretical distribution (see Section 1). The user must specify the theoretical distribution's parameter values (see Appendix A). These values are typically determined from the data, e.g., using maximum likelihood estimation.

The values $k = n$ and $P_i = (i-.5)/n$ are commonly used, in which case $Q_i = x_{(i)}$.

Weights may be specified, in which case the quantiles are calculated as described in section 2.3. In that case, the paramater estimates for the theoretical distribution to which the sample is being compared should take the weights into account.

The macro for Q-Q plotting will operate in one of two modes, depending on whether or not the user provides input probabilities. In either case, a SAS data set containing an input sample is required.

If input probabilities are not provided, the ECDF is evaluated at each observation of the input sample, using the %ECDF macro. The resulting probabilities are used as arguments to an inverse CDF function which the user must specify as a parameter to the macro. Values returned by the inverse CDF function are used as X-coordinates for the Q-Q plot. The input sample values are used as Y-coordinates.

If input probabilities are provided, the %QUANT macro is used to obtain corresponding quantiles from the the input sample, and these are used as the Y-coordinates. The user-specified inverse CDF function operates on the same input probabilities to produce the X-coordinates.

An optional output SAS data set can be produced containing the coordinates for the Q-Q plot. The plot will be produced on the 3800-3 laser printer. In order to use an alternative graphics device, it is necessary to modify the GOPTIONS statement contained in the %PORQ macro called by %QQ. (See the general information at the beginning of this section.) Macro parameters are in the form *keyword=value* and can be specified in any order.

2.4.1.1 %QQ Macro

General Form:

```
%QQ(parameter list)
```

Parameters:

IN= the name of the input sample SAS data set. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. The input sample SAS data set need not be sorted by the user.

INVAR= the name of the variable which contains the input sample. If this parameter is not specified, the variable name X will be assumed.

WGTVAR= the name of the variable, in the input SAS data set identified by the IN= parameter, containing the weights for the data values. This parameter is optional.

CCWT= the value of c to be used in the calculation of the continuity correction, when weights are used. If weights are used and CCWT= is not specified, the smallest weight in the sample is found and is used as the default. This parameter is optional.

INTERP= determines whether linear interpolation is used, when weights are used and input probabilities are provided by the user. If WGTVAR= is specified and INTERP=0, the ECDF is treated as a step function. If WGTVAR= is specified and INTERP is not specified, linear interpolation is used by default. This parameter is optional.

INP= the name of the SAS data set containing the optional input probabilities. The macro will operate in one of two modes depending on whether or not this

parameter has been specified. (See text above for details.) If this parameter is used, the INPVAR= parameter must also be given.

INPVAR= the name of the variable containing the input probabilities. This parameter is required if the INP= parameter has been specified. If INP= has been specified and INPVAR= has not, a message will be written to the SAS log and the macro will terminate.

FUNC= the inverse CDF specification for the desired theoretical distribution. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. This function will be evaluated either at each of the default probability values derived from the ECDF calculation, or at each value of the INPVAR variable, depending on the operational mode of the macro, in order to produce theoretical quantiles to be used as X-axis coordinates for the Q-Q plot. This parameter must be given in the form FUNC=*function-name(arguments)*, where *function-name* is the name of an inverse CDF described either in "SAS User's Guide: Basics, Version 5 Edition" or in Section 1.4 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as X.

AXES= the method used to determine the lengths of the axes for the plot. AXES=SAS will allow SAS to determine the lengths of the axes. AXES=FIXED will fix the axes at a predetermined percentage of the total plotting surface dimensions. If this parameter is not specified, AXES=SAS will be assumed. (See also the discussion of axis length at the beginning of this section.)

XAXIS= the X-axis limits and intermediate tick-mark values. This parameter is optional. If used, its value can correspond to any of the forms values can take for the ORDER option of the SAS/GRAPH AXIS statement. (See the discussion of axis values at the beginning of Section 2 for details on the syntax of this parameter.)

YAXIS= the Y-axis limits and intermediate tick-mark values. (See XAXIS=.)

OUT= the name of the optional output SAS data set which will contain the X and Y coordinates for the Q-Q plot. The output data set will be produced only if this parameter has been specified.

QX= the name to be given to the output variable which will contain the theoretical quantiles used as X-axis coordinates for the Q-Q plot. If this parameter is specified without a value for the OUT= parameter, then it will be ignored. If OUT= has been specified and this parameter has not, then the default variable name QX will be used.

QY= the name to be given to the output variable which will contain the sample quantiles used as Y-axis coordinates for the Q-Q plot. If this parameter is specified without a value for the OUT= parameter, then it will be ignored. If

OUT= has been specified and this parameter has not, then the default variable name QY will be used.

2.4.2 Probability-Probability (P-P) Plots Comparing a Sample to a Distribution

Let $x_{(1)}, x_{(2)}, ..., x_{(n)}$ be the sample values sorted in non-descending order. Let $Q_1, Q_2, ..., Q_k$ be a set of k quantile values selected by the user; these need not be sorted. Let $P_1, P_2, ..., P_k$ be the corresponding values of the ECDF at the arguments Q_i. (The ECDF is calculated from the entire sample of n values.)

A P-P plot comparing the sample to a theoretical distribution is a scatter plot of the k points $(F(Q_i),P_i)$ (for $i = 1, ..., k$), where F is the CDF for the desired theoretical distribution (see Section 1). The user must specify the theoretical distribution's parameter values (see Appendix A). These values are typically determined from the data, e.g., using maximum likelihood estimation.

The values $k = n$ and $Q_i = x_{(i)}$ are commonly used, in which case $P_i = (i-.5)/n$.

Weights may be specified, in which case the ECDF is calculated as described in section 2.2. In that case, the paramater estimates for the theoretical distribution to which the sample is being compared should take the weights into account.

The macro for P-P plotting will operate in one of two modes, depending on whether or not the user provides input quantiles. In either case, a SAS data set containing an input sample is required.

If input quantiles are not provided, the ECDF is evaluated at each observation of the input sample, using the %ECDF macro. The resulting probabilities are used as the Y coordinates for the P-P plot. A CDF function which the user must specify as a parameter to the macro will also be evaluated at each observation of the input sample. Values returned by the CDF will be used as X coordinates.

If input quantiles are provided, the macro will use linear interpolation to calculate the corresponding probability values of the ECDF at each quantile argument and the results will be used as Y coordinates to the P-P plot. The user-specified CDF will be evaluated at each input quantile argument to produce the X coordinates.

An optional output SAS data set can be produced containing the coordinates for the P-P plot. The plot will be produced on the 3800-3 laser printer. In order to use an alternative graphics device, it is necessary to modify the GOPTIONS statement contained in the %PORQ macro called by %PP. (See the general information at the beginning of this section.) Macro parameters are in the form *keyword=value* and can be specified in any order.

2.4.2.1 %PP Macro

General Form:

```
%PP(parameter list)
```

Parameters:

IN= the name of the input sample SAS data set. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. The input sample SAS data set need not be sorted by the user.

INVAR= the name of the variable which contains the input sample. If this parameter is not specified, the variable name X will be assumed.

WGTVAR= the name of the variable, in the input SAS data set identified by the IN= parameter, containing the weights for the data values. This parameter is optional.

CCWT= the value of c to be used in the calculation of the continuity correction, when weights are used. If weights are used and CCWT= is not specified, the smallest weight in the sample is found and is used as the default. This parameter is optional.

INQ= the name of the SAS data set containing the optional input quantile values. The macro will operate in one of two modes depending on whether or not this parameter has been specified. (See text above for details.) If this parameter is used, the INQVAR= parameter must also be given.

INQVAR= the name of the variable containing the input quantile values. This parameter is required if the INQ= parameter has been specified. If INQ= has been specified and INQVAR= has not, a message will be written to the SAS log and the macro will terminate.

FUNC= the CDF specification for the desired theoretical distribution. This parameter is required. If it is not specified, a message will be written to the SAS log and the macro will terminate. This function will be evaluated either at each value of the input sample, or at each value of the INQVAR variable, depending on the operational mode of the macro, in order to produce theoretical probabilities to be used as X-axis coordinates for the P-P plot. This parameter must be given in the form FUNC=*function-name(arguments)*, where *function-name* is the name of a CDF described either in "SAS User's Guide: Basics, Version 5 Edition" or in Section 1.3 of this document, and *arguments* is the argument list required by the chosen function. The first element in the function's argument list, which is the name of the variable at which the function is to be evaluated, must always be coded as X.

AXES= the method used to determine the lengths of the axes for the plot. AXES=SAS will allow SAS to determine the lengths of the axes. AXES=FIXED will fix the axes at a predetermined percentage of the total

plotting surface dimensions. If this parameter is not specified, AXES=SAS will be assumed. (See also the discussion of axis length at the beginning of this section.)

XAXIS= the X-axis limits and intermediate tick-mark values. This parameter is optional. If used, its value can correspond to any of the forms values can take for the ORDER option of the SAS/GRAPH AXIS statement. (See the discussion of axis values at the beginning of Section 2 for details on the syntax of this parameter.)

YAXIS= the Y-axis limits and intermediate tick-mark values. (See XAXIS=.)

OUT= the name of the optional output SAS data set which will contain the X and Y coordinates for the P-P plot. The output data set will be produced only if this parameter has been specified.

PX= the name to be given to the output variable which will contain the theoretical probabilities used as X-axis coordinates for the P-P plot. If this parameter is specified without a value for the OUT= parameter, then it will be ignored. If OUT= has been specified and this parameter has not, then the default variable name PX will be used.

PY= the name to be given to the output variable which will contain the sample probabilities used as Y-axis coordinates for the P-P plot. If this parameter is specified without a value for the OUT= parameter, then it will be ignored. If OUT= has been specified and this parameter has not, then the default variable name PY will be used.

2.4.3 Quantile-Quantile (Q-Q) Plots Comparing Two Samples

This section provides instructions for constructing a Q-Q plot for comparing two samples, using the %QUANT macro.

Let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_m$ be two samples of data. (These need not be the same size, and they need not be sorted.) Let $P_1, P_2, ..., P_k$ be a single set of k probabilities selected by the user. Let $Q_1^x, Q_2^x, ..., Q_k^x$ be the corresponding quantiles for the x's, obtained by using the %QUANT macro, and let $Q_1^y, Q_2^y, ..., Q_k^y$ be the corresponding quantiles for the y's, obtained by using the %QUANT macro a second time.

Then a Q-Q plot comparing the distribution of the x's to that of the y's is a scatter plot of the k points (Q_i^x, Q_i^y) (for $i=1, ..., k$).

In the special case where $n=m=k$ and $P_i=(i-.5)/n$, the Q-Q plot is simply a scatter plot of the ordered y's versus the ordered x's.

2.5 *Random Variate Generators*

Several macros have been written to facilitate the creation of SAS data sets containing a random variate. Each macro will generate a different distribution: chi-square, exponential, normal, t, or uniform. Keyword parameters enable the user to specify the desired number of observations, the name of the random variate, the initial seed for the random generator function used, and parameters of the distribution. Parameters are specified in the form *keyword=value*, and can be specified in any order.

Each macro is designed to be invoked in the context of a data step for which the user supplies a DATA statement identifying the SAS data set to which the macro will output observations. For example,

```
DATA NORM01;
  %GENNORM(N=200)
RUN;
```

will produce SAS data set WORK.NORM01 containing 200 observations of normal random variate X with mean 0 and variance 1 as determined by parameter defaults.

Each of these random variate generator macros uses one of the SAS random number functions (two, in the case of %GENT). In the macro descriptions which follow, pertinent SAS random number functions are identified. The SAS random number functions are described in Chapter 6 of "SAS User's Guide: Basics, Version 5 Edition". Techniques used to generate an observation are indicated therein.

The seed parameter associated with each of these macros becomes the seed to the relevant SAS random number function(s). Your attention is directed to page 236 of "SAS User's Guide: Basics" for a detailed discussion of the initialization of a random number stream.

Note that, if data generated by these macros are to be reproducible, an initial seed having a value greater than zero must be used and that initial seed value should be recorded.

2.5.1 %GENCHI Macro

%GENCHI generates observations of a Chi-square variate. This macro uses the RANGAM function as follows:

```
ALPHA = degrees-of-freedom / 2;
variate = 2 * RANGAM(seed,ALPHA);
```

where *degrees-of-freedom*, *variate* and *seed* are macro parameters DF, RVAR and S, respectively.

General form:

```
%GENCHI(parameter list)
```

Parameters:

DF= degrees of freedom. The default value is 1. Any value specified must be an integer.

N= the number of observations to be generated. The default value is 50.

S= the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

RVAR= the name of the random variate for which values will be generated. The default value is X.

2.5.2 %GENEXP Macro

%GENEXP generates observations of an exponential variate. This macro uses the RANEXP function as follows:

```
variate = RANEXP(seed) * THETA;
```

where *variate*, *seed* and *THETA* are macro parameters RVAR, S and THETA, respectively.

General form:

```
%GENEXP(parameter list)
```

Parameters:

THETA= the mean. The default value is 1.

N= the number of observations to be generated. The default value is 50.

S= the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

RVAR= the name of the random variate for which values will be generated. The default value is X.

2.5.3 %GENNORM Macro

%GENNORM generates observations of a Normal random variate. This macro uses the RANNOR function as follows:

```
variate = mu + SQRT(sigsq) * RANNOR(seed);
```

where *variate*, *mu*, *sigsq* and *seed* are macro parameters RVAR, MU, SIGSQ and S, respectively.

General form:

```
%GENNORM(parameter list)
```

Parameters:

MU= the mean. The default value is 0.

N= the number of observations to be generated. The default value is 50.

S= the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed >0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

SIGSQ= the variance of the distribution. The default value is 1.

RVAR= the name of the random variate for which values will be generated. The default value is X.

2.5.4 %GENT Macro

%GENT generates observations of a t variate. This macro uses the RANNOR and RANGAM functions as follows:

```
ALPHA = degrees-of-freedom / 2;
R1 = RANNOR(seed);                                  /* Normal(0,1) */
R2 = 2 * RANGAM(seed,ALPHA);                        /* Chi-square  */
variate = R1 / SQRT(R2 / degrees-of-freedom);       /* t           */
```

where *degrees-of-freedom*, *seed* and *variate* are macro parameters DF, S and RVAR, respectively. (Note that each function call returns a new value for *seed*, thereby ensuring the independence of the Normal and chi-square random variates.)

General form:

```
%GENT(parameter list)
```

Parameters:

DF= degrees of freedom. The default value is 1. Any value specified must be an integer.

N= the number of observations to be generated. The default value is 50.

S= the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

RVAR= the name of the random variate for which values will be generated. The default value is X.

2.5.5 %GENUNI Macro

%GENUNI generates observations of a uniform random variate on the interval (0,1). This macro uses the RANUNI function as follows:

```
variate = RANUNI(seed);
```

where *variate* and *seed* are macro parameters RVAR and S, respectively.

General form:

```
%GENUNI(parameter list)
```

Parameters:

N= the number of observations to be generated. The default value is 50.

S= the initial seed for the random number function. The default value is 0 which causes a CPU clock observation to be used as the initial seed. A reproducible series of values can be obtained by using a seed > 0. (See "SAS User's Guide: Basics, Version 5 Edition", page 236.)

RVAR= the name of the random variate for which values will be generated. The default value is X.

Section 3

USER INTERFACE

The libraries containing the macros and functions described by this document are made available automatically to users of the MCC-supported (STC2.SAS) catalogued procedures and clists. (Originally, it was necessary for a user to establish access to the libraries by means of parameters to the procedures and clists, but this requirement has been eliminated prior to distribution of this document.)

Some minor operational requirements which remain pertinent are described below.

3.1 Batch Mode

When you intend to use a macro or macro option which will generate a SAS/GRAPH plot, you must invoke the SASG3800 catalogued procedure. This procedure allocates the files necessary for graphics output to IBM laser printers via interface with GDDM[3] software, and allocates a default virtual storage region adequate for most graphics applications which use that interface. Symbolic parameters GDEST and GCOPIES can be used to route graphs to a remote 38xx printer (where applicable), and to generate multiple copies of graphs, respectively. GDEST defaults to the 3800-3 which is the target device of the graph-generating macros. If use of a remote printer is desired, it is also necessary to override the SAS/GRAPH device driver selection in the macro code. GCOPIES should be used with discretion because graphics images are retransmitted to the printer for each copy.

3.2 Interactive Mode

The default TSO logon region is adequate for most SAS sessions that do not use SAS/GRAPH. However, if graphics are to be produced, the logon region should be set to 3000K as in the following example.

```
TSO userid A(account) S(3000)
```

When you intend to use a macro or macro option which will generate a SAS/GRAPH plot, you must specify the GRAPH38 parameter when you invoke the SAS clist. This is done subsequent to issuing the START SAS command and allocating required personal data sets. The following example illustrates the command sequence.

---

[3] IBM's Graphical Data Display Manager

```
START SAS
 ...
ALLOC F(filename) DA('dsname')
 ...
SAS GRAPH38
```

Parameters GDEST and GCOPIES can be used with the SAS clist to route graphs to a remote 38xx printer (where applicable), and to generate multiple copies of graphs, respectively. GDEST defaults to the 3800-3 which is the target device of the graph-generating macros. If use of a remote printer is desired, it is also necessary to override the SAS/GRAPH device driver selection in the macro code. GCOPIES should be used with discretion because graphics images are retransmitted to the printer for each copy. (Detailed TSO help information is available for the SAS clist upon return from the START SAS command.)

3.3 *Obtaining Copies of Documentation*

In order to obtain a copy of this document, the following command sequence should be issued in TSO :

```
START TEXTAIDS
 ...
PR3820 'STC2.SAS.CPDS3820(EXPLORE)'
```

Appendix A

FORMULAS FOR PROBABILITY DENSITY FUNCTIONS

A.1 Beta Distribution

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, x^{a-1}(1-x)^{b-1}$$

$0 < x < 1$
$a > 0$
$b > 0$

A.2 Chi-square Distribution

$$f(x) = \frac{1}{2^{\frac{df}{2}}\,\Gamma\left(\frac{df}{2}\right)}\, e^{-\frac{x}{2}}\, x^{\frac{df}{2}-1}$$

$x > 0$
$df \geq .5$

A.3 Exponential Distribution

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$$

$x \geq 0$
$\theta > 0 \quad (\theta \equiv theta)$

A.4 F Distribution

$$f(x) = \frac{\Gamma\left(\frac{a+b}{2}\right)}{\Gamma\left(\frac{a}{2}\right)\Gamma\left(\frac{b}{2}\right)} \left(\frac{b}{a}\right)^{\frac{b}{2}} x^{\frac{a}{2}-1} \left(x+\frac{b}{a}\right)^{-(a+b)/2}$$

$x > 0$
$a > 0 \quad (a \equiv df1)$
$b > 0 \quad (b \equiv df2)$

A.5 Gamma Distribution

$$f(x) = \frac{\left(\frac{x-\mu}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{x-\mu}{\sigma}\right)}}{\sigma\Gamma(\alpha)}$$

$x > \mu \quad (\mu \equiv lo)$
$-\infty < \mu < \infty$
$\sigma > 0 \quad (\sigma \equiv sc)$
$\alpha > 0 \quad (\alpha \equiv sh)$

A.6 Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$-\infty < x < \infty$
$-\infty < \mu < \infty \quad (\mu \equiv mu)$
$\sigma > 0 \quad (\sigma^2 \equiv sigsq)$

A.7 t Distribution

$$f(x) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)\sqrt{\pi}\sqrt{df}} \left(1+\frac{x^2}{df}\right)^{-(df+1)/2}$$

$-\infty < x < \infty$
$df > 0$

A.8 Uniform Distribution

$$f(x) = \frac{1}{b-a}$$

$a \le x \le b$
$a < b$

A.9 Weibull Distribution

$$f(x) = \left(\frac{\alpha}{\sigma}\right)\left(\frac{x-\mu}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}}$$

$x > \mu \quad (\mu \equiv lo)$
$-\infty < \mu < \infty$
$\sigma > 0 \quad (\sigma \equiv sc)$
$\alpha > 0 \quad (\alpha \equiv sh)$

Examples of Graphs Produced Using Macros
The Data - Observations of Systolic Blood Pressure for 4,677 Canadians - is From Statistics Canada's 1978/79 Canada Health Survey

1. Histogram of systolic blood pressure data and superimposed Normal (125.5,19.4) PDF, produced by %HIST Macro using unweighted data.

   The following statement was used to generate this plot :

```
%HIST(IN=BLOODP,INVAR=SYST,OUT=OUTDAT,
      BNDS= 80 100 110 120 130 140 150 160 180 200 220 240 260,
      TICKLABS= 80 100 110 120 130 140 150 160 180 200 220 240 260,
      AXES=FIXED,YUPLIM=.025,
      FUNC=PDFNORM(X,125.523,19.3969**2),LIMITS=80 260)
```

2. Histogram of systolic blood pressure data and superimposed Normal (125.0,18.5) PDF, produced by %HIST Macro using weighted data.

   The following statement was used to generate this plot :

```
%HIST(IN=BLOODP,INVAR=SYST,WGTVAR=WGT4,OUT=OUTDAT,
      BNDS= 80 100 110 120 130 140 150 160 180 200 220 240 260,
      TICKLABS= 80 100 110 120 130 140 150 160 180 200 220 240 260,
      AXES=FIXED,YUPLIM=.025,
      FUNC=PDFNORM(X,125.0467,18.5451**2),LIMITS=80 260)
```

3. ECDF of systolic blood pressure data and superimposed Normal (125.5,19.4) CDF, produced by %ECDF Macro using unweighted data.

   The following statement was used to generate this plot :

```
%ECDF(IN=BLOODP,INVAR=SYST,OUT=WQDATA,
      PLOT=YES,AXES=FIXED,FUNC=PROBNORM((X-125.523)/19.3969))
```

4. ECDF of systolic blood pressure data and superimposed Normal (125.0,18.5) CDF, produced by %ECDF Macro using weighted data.

   The following statement was used to generate this plot :

```
%ECDF(IN=BLOODP,INVAR=SYST,WGTVAR=WGT4,OUT=WQDATA,
      PLOT=YES,AXES=FIXED,FUNC=PROBNORM((X-125.0467)/18.5451))
```

5. Q-Q Plot comparing standardized systolic blood pressure data to Normal (0,1) distribution, produced by %QQ Macro using unweighted data.

   The following statement was used to generate this plot :

```
%QQ(IN=BLOODP,INVAR=STDSYS,FUNC=PROBIT(X)
      XAXIS=-4 TO 8,YAXIS=-4 TO 8,
      OUT=COORD,AXES=FIXED)

      where STDSYS=(SYST-125.523)/19.3969
```

6. Q-Q Plot comparing standardized systolic blood pressure data to Normal (0,1) distribution, produced by %QQ Macro using unweighted data.

   The following statement was used to generate this plot :

```
%QQ(IN=BLOODP,INVAR=STDSYS,WGTVAR=WGT4,FUNC=PROBIT(X),
      XAXIS=-4 TO 8,YAXIS=-4 TO 8,
      OUT=COORD,AXES=FIXED)

      where STDSYS=(SYST-125.0467)/18.5451
```

7. P-P Plot comparing standardized systolic blood pressure data to Normal (0,1) distribution, produced by %PP Macro using unweighted data.

   The following statement was used to generate this plot :

```
%PP(IN=BLOODP,INVAR=STDSYS,
      FUNC=PROBNORM(X),AXES=FIXED,
      XAXIS=0 TO 1 BY .1,OUT=COORD)

      where STDSYS=(SYST-125.523)/19.3969
```

8. P-P Plot comparing standardized systolic blood pressure data to Normal (0,1) distribution, produced by %PP Macro using weighted data.

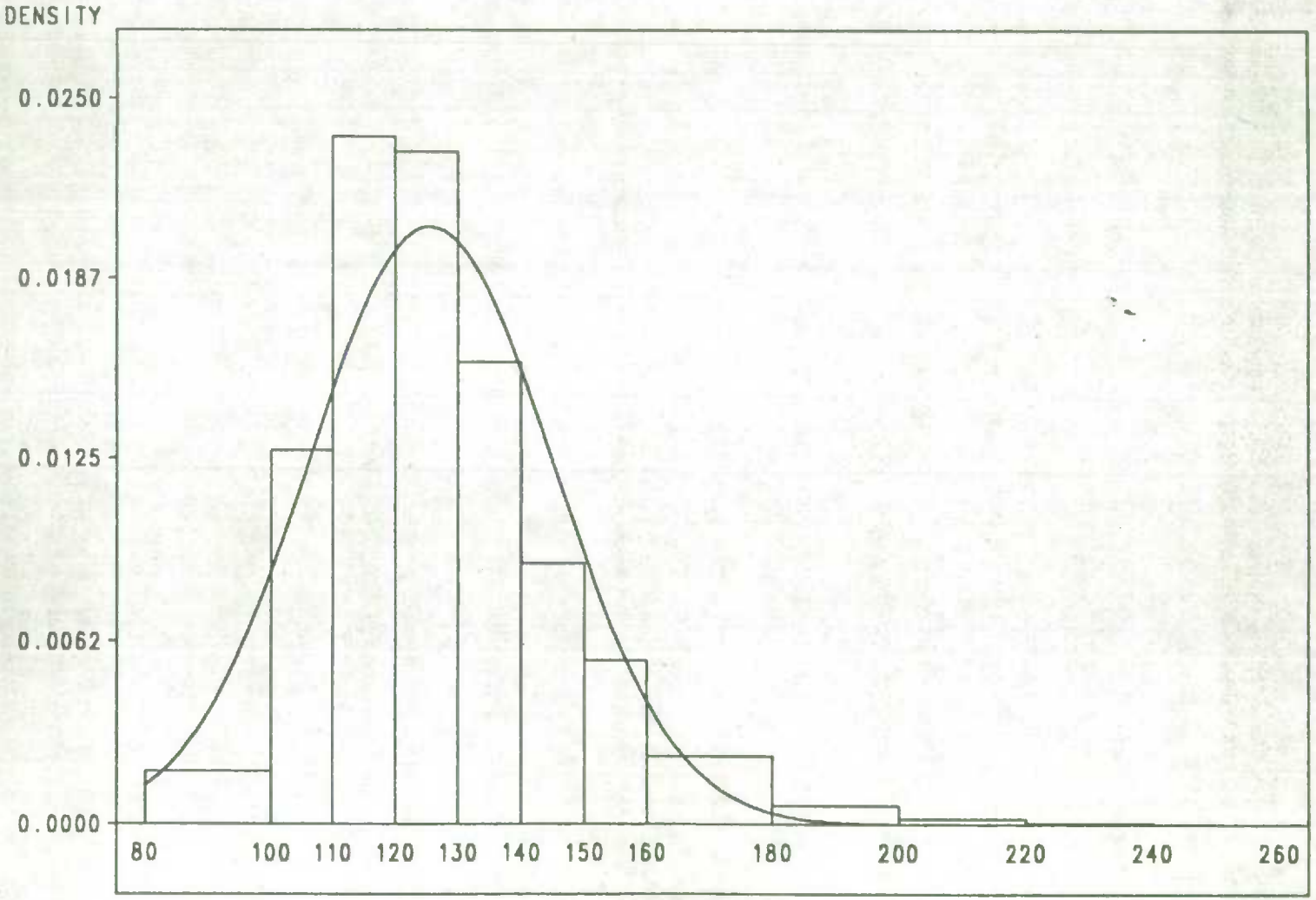
   The following statement was used to generate this plot :

```
%PP(IN=BLOODP,INVAR=STDSYS,WGTVAR=WGT4,
      FUNC=PROBNORM(X),AXES=FIXED,
      XAXIS=0 TO 1 BY .1,OUT=COORD)

      where STDSYS=(SYST-125.0467)/18.5451
```

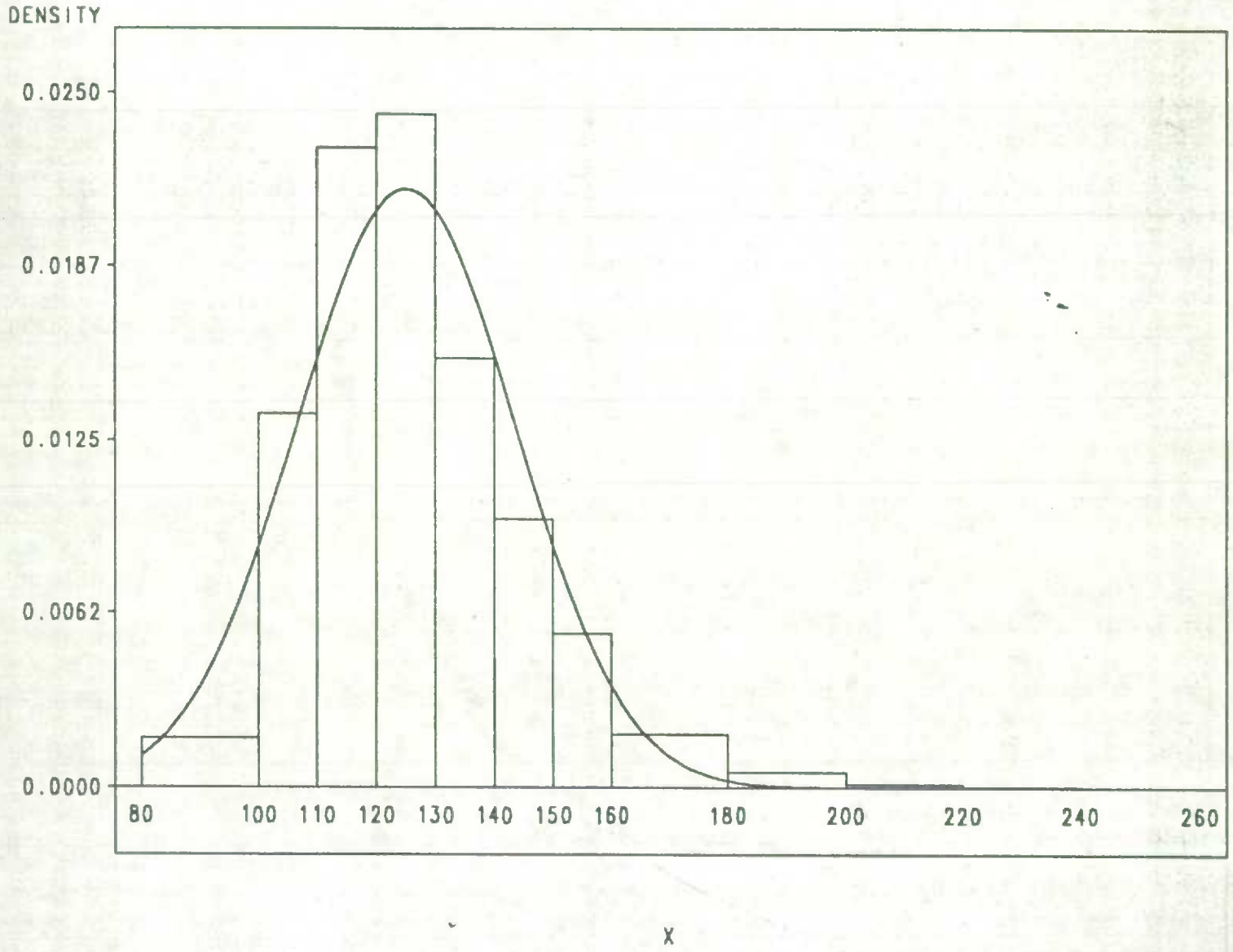
HISTOGRAM – SYSTOLIC BLOOD PRESSURE

MALES AND FEMALES – UNWEIGHTED



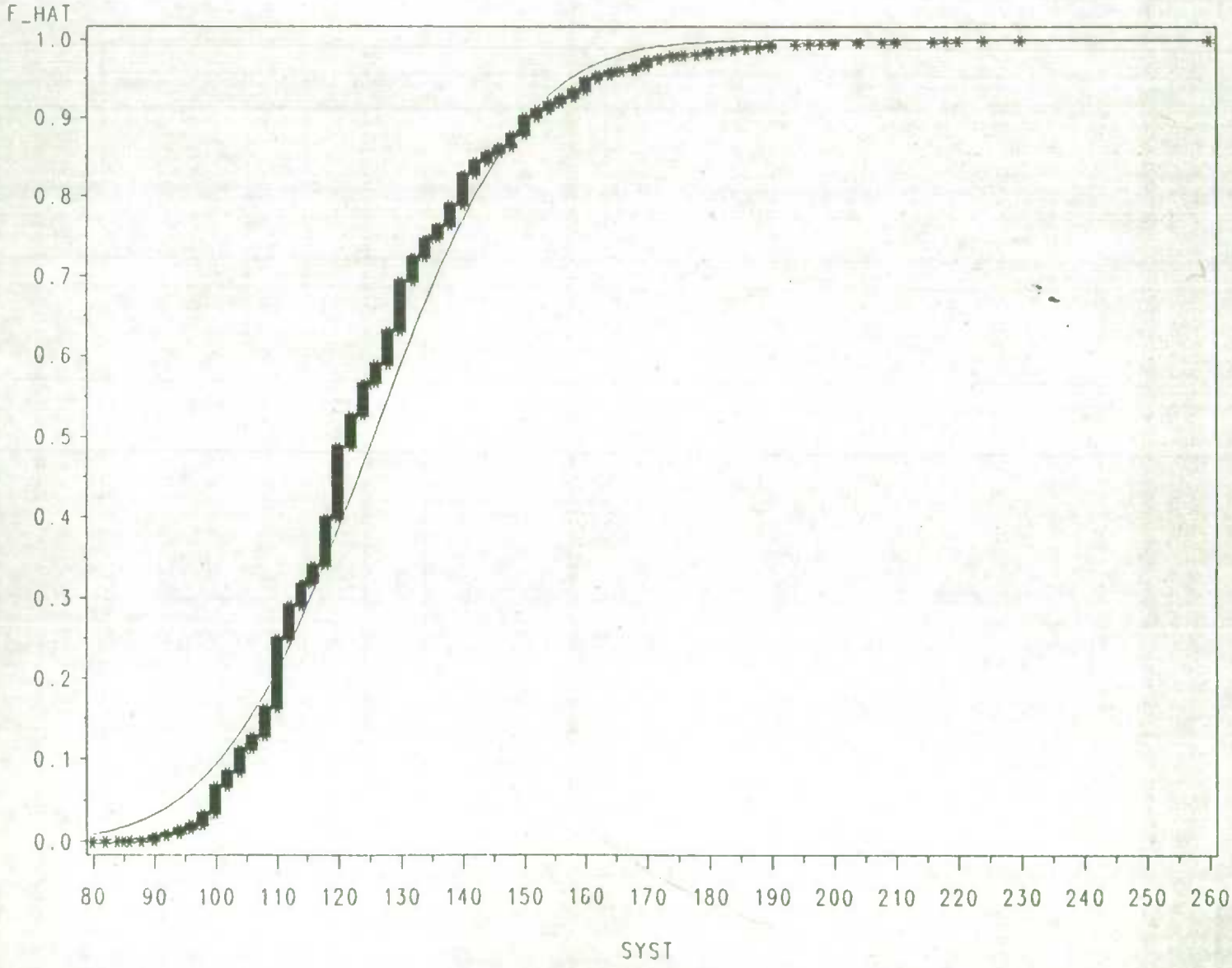
HISTOGRAM – SYSTOLIC BLOOD PRESSURE

MALES AND FEMALES - WEIGHTED



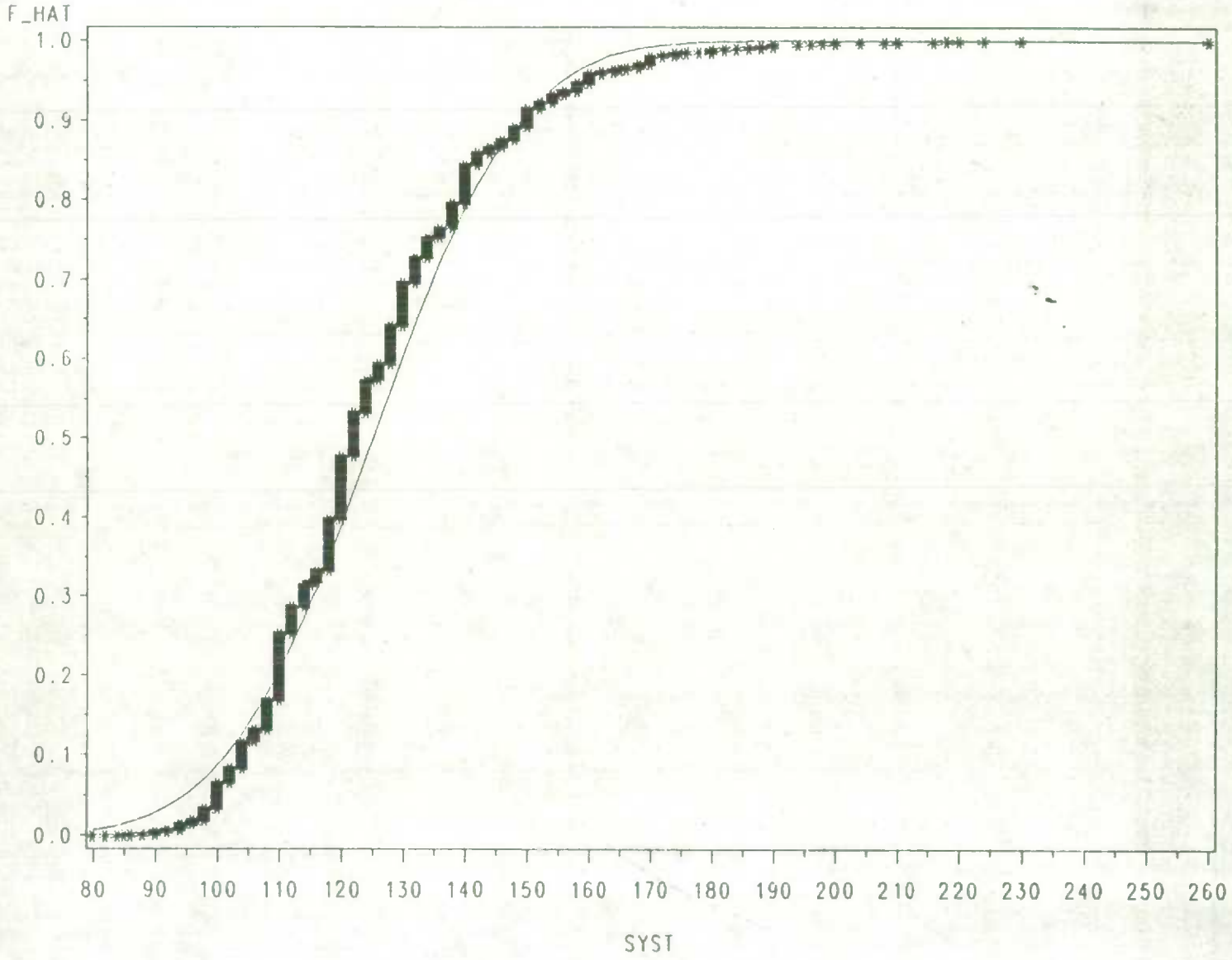
ECDF PLOT – SYSTOLIC BLOOD PRESSURE

MALES AND FEMALES - UNWEIGHTED



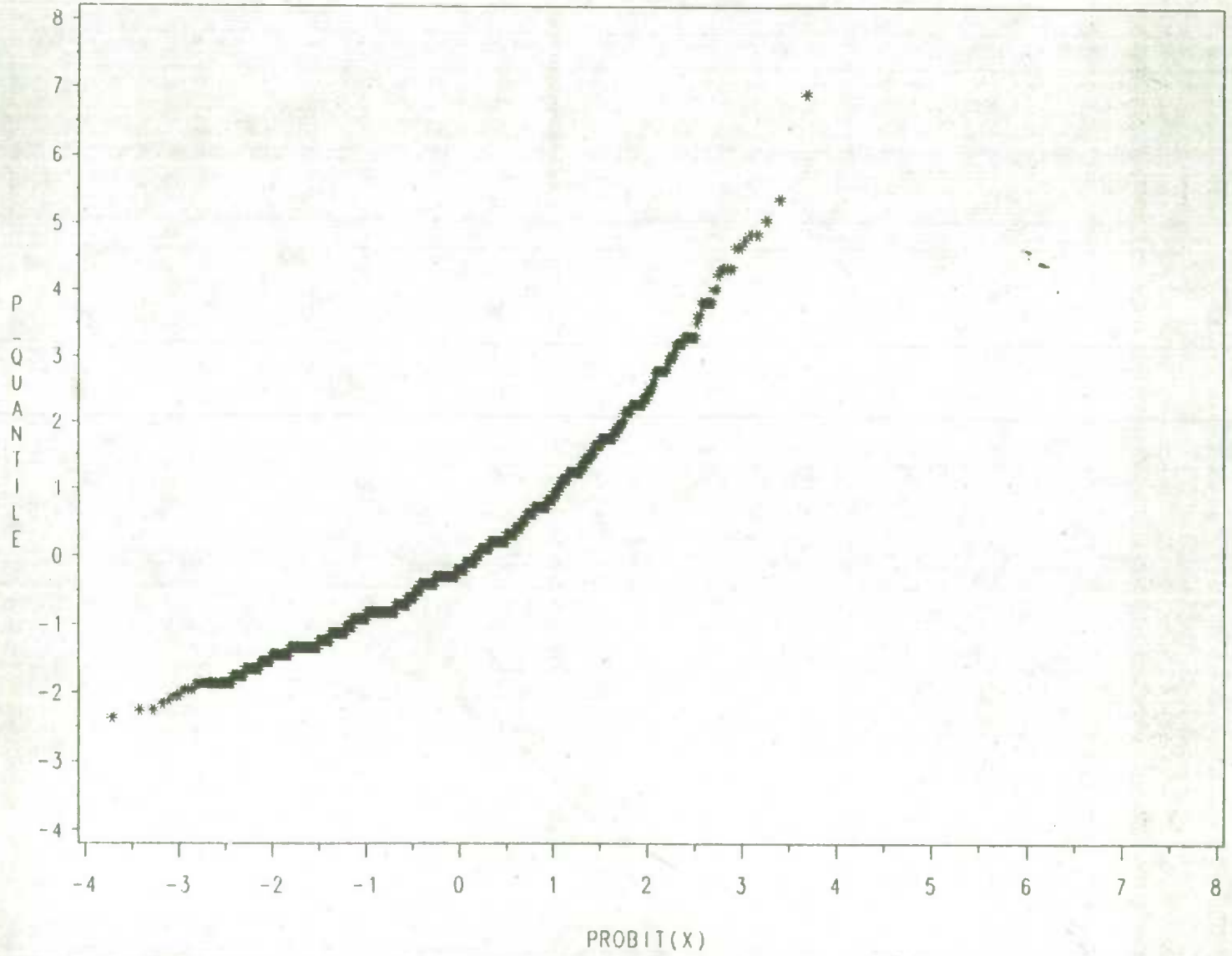
ECDF PLOT – SYSTOLIC BLOOD PRESSURE

MALES AND FEMALES - WEIGHTED



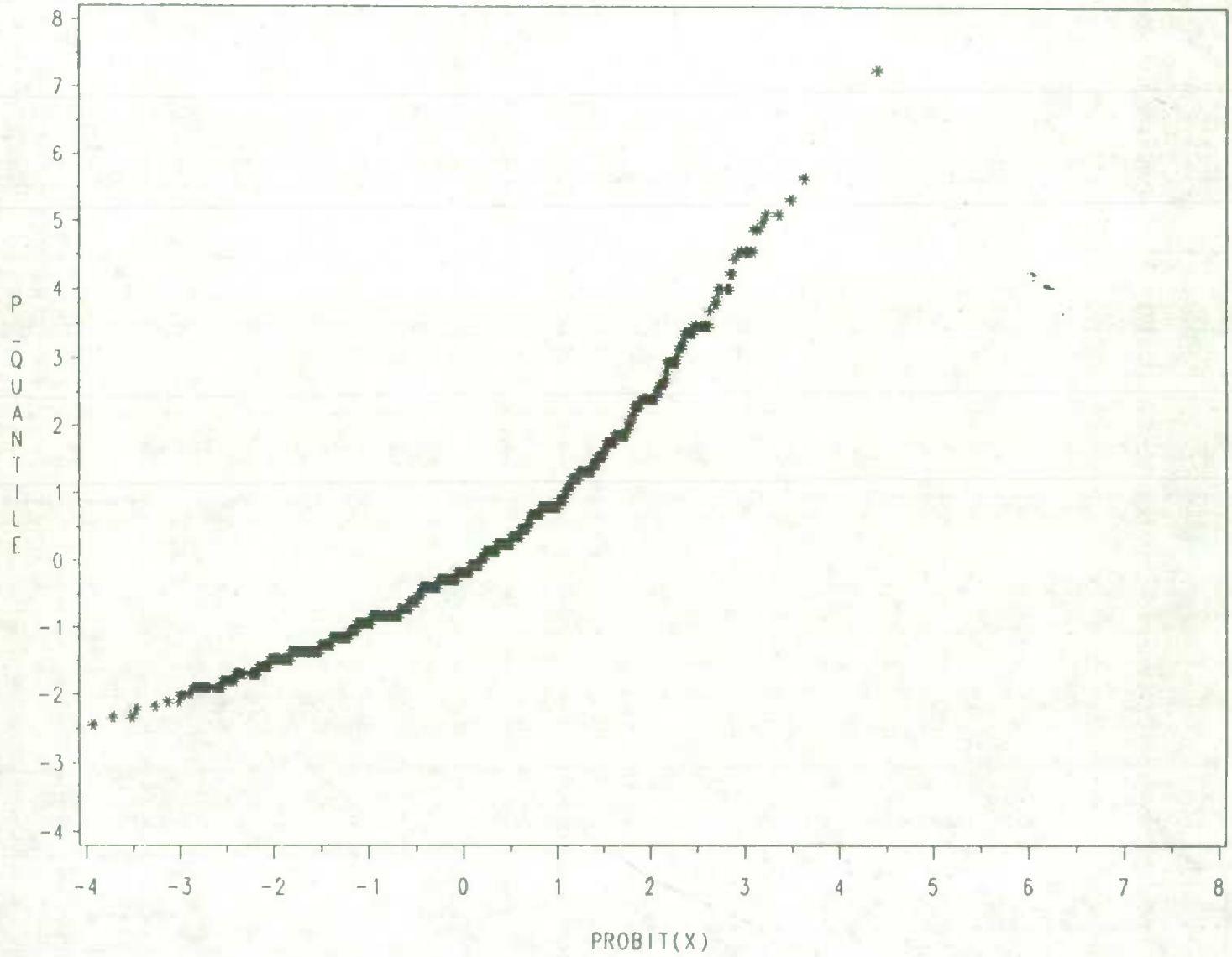
QQ PLOT – STANDARDIZED SYSTOLIC VS N(0,1)

NO PROBABILITIES SPECIFIED - UNWEIGHTED



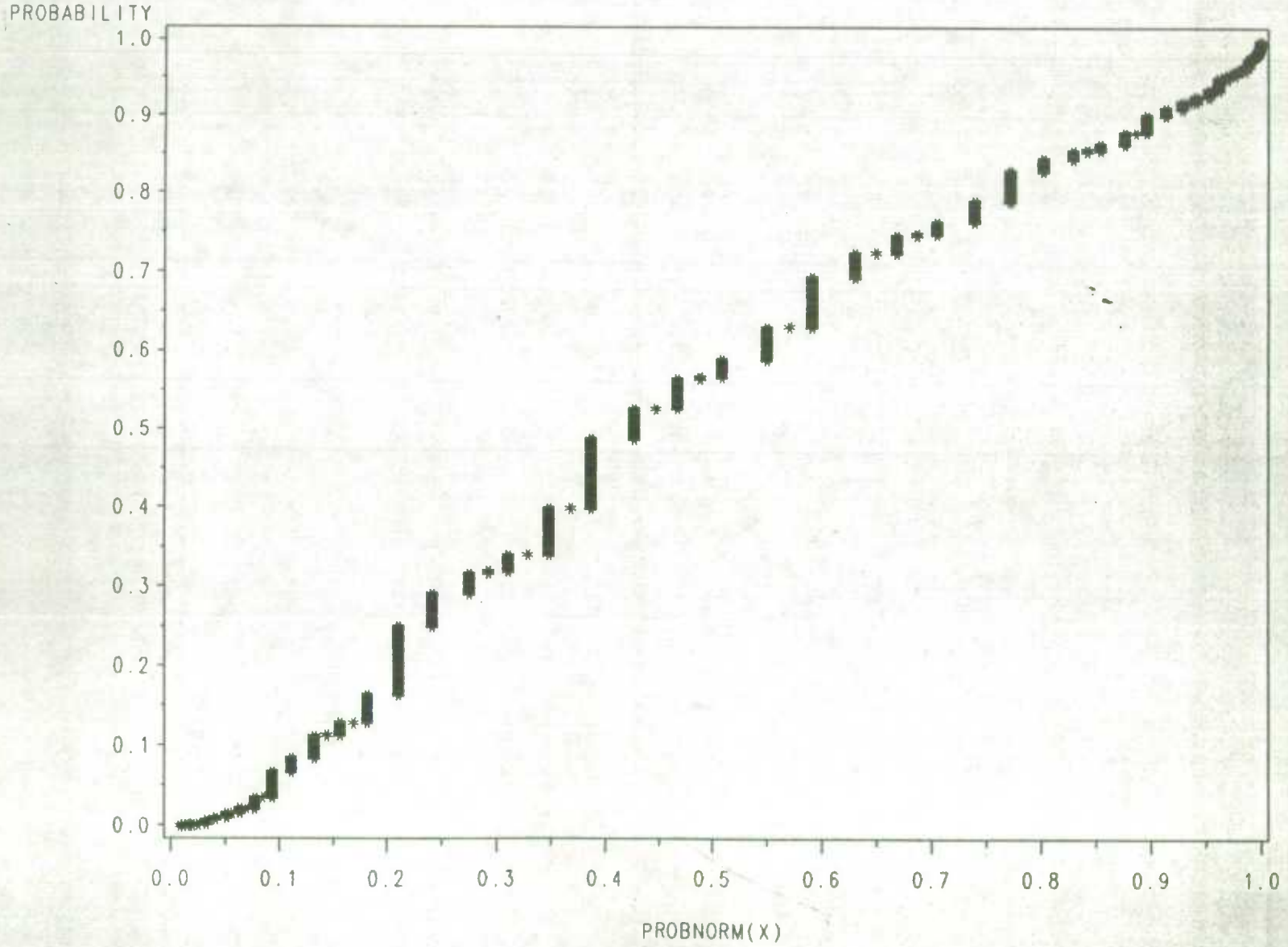
QQ PLOT – STANDARDIZED SYSTOLIC VS N(0,1)

NO PROBABILITIES SPECIFIED - WEIGHTED



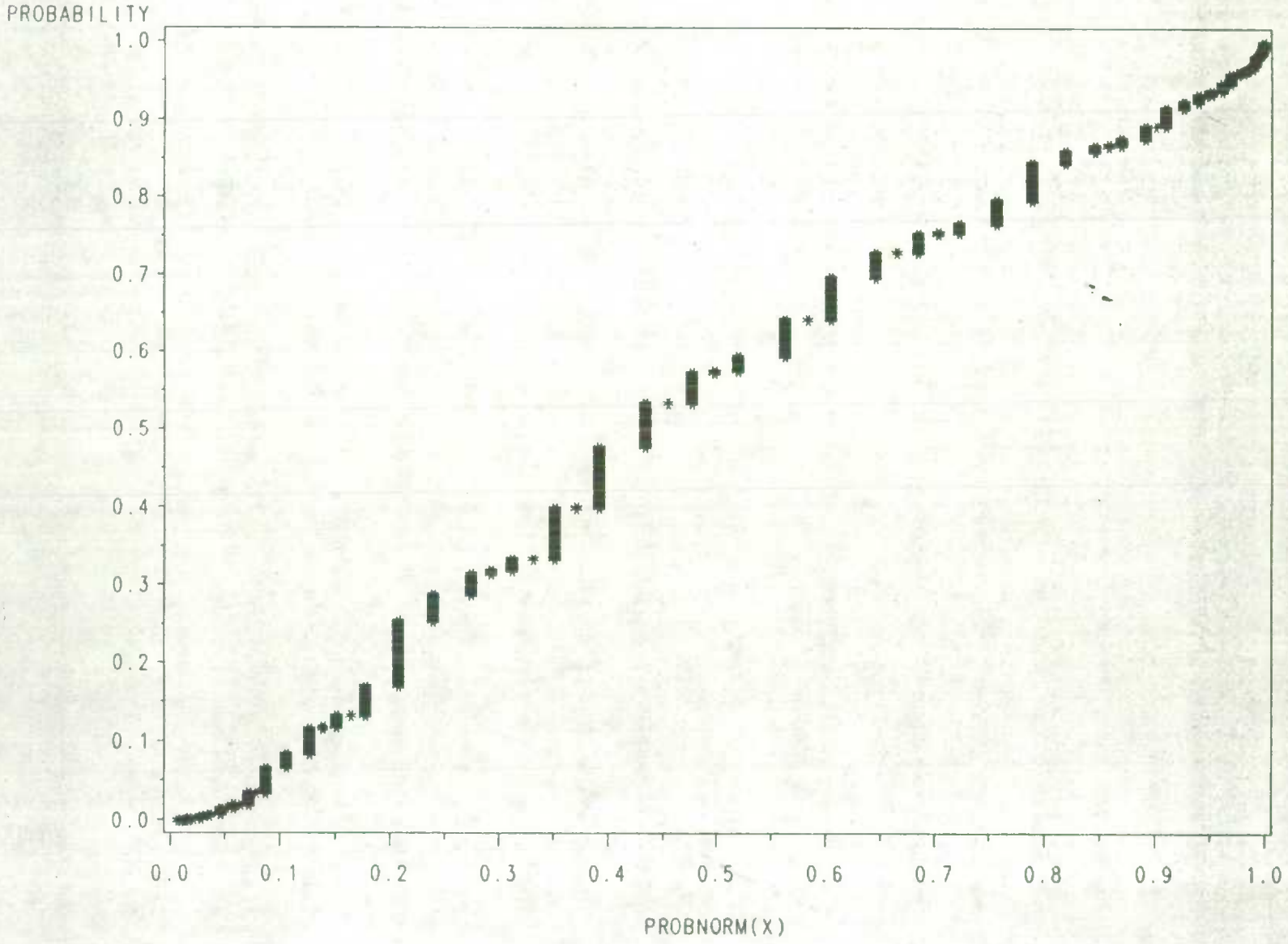
P–P PLOT – STANDARDIZED SYSTOLIC VS N(0,1)

QUANTILES NOT SPECIFIED - UNWEIGHTED



P-P PLOT - STANDARDIZED SYSTOLIC VS N(0,1)

QUANTILES NOT SPECIFIED - WEIGHTED



ANALYTICAL STUDIES BRANCH
RESEARCH PAPER SERIES

No.

1. *Behavioural Response in the Context of Socio-Economic Microanalytic Simulation, Lars Osberg*

2. *Unemployment and Training, Garnett Picot*

3. *Homemaker Pensions and Lifetime Redistribution, Michael Wolfson*

4. *Modelling the Lifetime Employment Patterns of Canadians, Garnett Picot*

5. *Job Loss and Labour Market Adjustment in the Canadian Economy, Garnett Picot and Ted Wannell*

6. *A System of Health Statistics: Toward a New Conceptual Framework for Integrating Health Data, Michael C. Wolfson*

7. *A Prototype Micro-Macro Link for the Canadian Household Sector, Hans J. Adler and Michael C. Wolfson*

8. *Notes on Corporate Concentration and Canada's Income Tax, Michael C. Wolfson*

9. *The Expanding Middle: Some Canadian Evidence on the Deskilling Debate, John Myles*

10. *The Rise of the Conglomerate Economy, Jorge Niosi*

11. *Energy Analysis of canadian External Trade: 1971 and 1976, K.E. Hamilton*

12. *Net and Gross Rates of Land Concentration, Ray D. Bollman and Philip Ehrensaft*

13. *Cause-Deleted Life Tables for Canada (1972 to 1981): An Approach Towards Analyzing Epidemiologic Transition, Dhruva Nagnur and Michael Nagrodski*

14. *The Distribution of the Frequency of Occurence of Nucleotide Subsequences, Based on Their Overlap Capability, Jane F. Gentleman and Ronald C. Mullin*

15. *Immigration and the Ethnolinguistic Character of Canada and Quebec, Réjean Lachapelle*

16. *Integration of Canadian Farm and Off-Farm Markets and the Off-Farm Work of Women, Men and Children, Ray D. Bollman and Pamela Smith*

17. *Wages and Jobs in the 1980s: Changing Youth Wages and the Declining Middle, J. Myles, G. Picot and T. Wannell*

18. *A Profile of Farmers with Computers, Ray D. Bollman*

19. *Mortality Risk Distributions: A Life Table Analysis, Geoff Rowe*

20. *Industrial Classification in the Canadian Census of Manufactures: Automated Verification Using Product Data, John S. Crysdale*

21. *Consumption, Income and Retirement, A.L. Robb and J.B. Burbridge*

22. *Job Turnover in Canada's Manufacturing Sector, John R. Baldwin and Paul K. Gorecki*

23. *Series on The Dynamics of the Competitive Process, John R. Baldwin and Paul K. Gorecki*

    A. *Firm Entry and Exit Within the Canadian Manufacturing Sector.*
    B. *Intra-Industry Mobility in the Canadian Manufacturing Sector.*
    C. *Measuring Entry and Exit in Canadian Manufacturing: Methodology.*
    D. *The Contribution of the Competitive Process to Productivity Growth: The Role of Firm and Plant Turnover.*
    E. *Mergers and the Competitive Process.*
    F. *(in preparation)*
    G. *Concentration Statistics as Predictors of the Intensity of Competition.*
    H. *The Relationship Between Mobility and Concentration for the Canadian Manufacturing Sector.*

24. *Mainframe SAS Enhancements in Support of Exploratory Data Analysis, Richard Johnson and Jane F. Gentleman*

25. *Dimensions of Labour Market Change in Canada: Intersectoral Shifts, Job and Worker Turnover, John R. Baldwin and Paul K. Gorecki*

26. *The Persistent Gap: Exploring the Earnings Differential Between Recent Male and Female Postsecondary Graduates, Ted Wannell*

27. *Estimating Agricultural Soil Erosion Losses From Census of Agriculture Crop Coverage Data, Douglas F. Trant*

28. *Good Jobs/Bad Jobs and the Declining Middle: 1967-1986, Garnett Picot, John Myles, Ted Wannell*

29. *Longitudinal Career Data for Selected Cohorts of Men and Women in the Public Service, 1978-1987, Garnett Picot and Ted Wannell*

30. *Earnings and Death - Effects Over a Quarter Century, Michael Wolfson, Geoff Rowe, Jane F. Gentleman adn Monica Tomiak*

31. *Firm Response to Price Uncertainty: Tripartite Stabilization and the Western Canadian Cattle Industry, Theodore M. Horbulyk*

32. *Smoothing Procedures for Simulated Longitudinal Microdata, Jane F. Gentleman, Dale Robertson and Monica Tomiak*

33. *Patterns of Canadian Foreign Direct Investment Abroad, Paul K. Gorecki*

34. *POHEM - A New Approach to the Estimation of Health Status Adjusted Life Expectancy, Michael C. Wolfson*

35. *Canadian Jobs and Firm Size: Do Smaller Firms Pay Less?, René Morissette*

36. *Distinguishing Characteristics of Foreign High Technology Acquisitions in Canada's Manufacturing Sector, John R. Baldwin and Paul K. Gorecki*

37. *Industry Efficiency and Plant Turnover in the Canadian Manufacturing Sector, John R. Baldwin*

38. *When the Baby Boom Grows Old: Impacts on Canada's Public Sector, Brian B. Murphy and Michael C. Wolfson*

39. *Trends in the distribution of Employment by Employer Size: Recent Canadian Evidence, Ted Wannell*

40. *Small Communities in Atlantic Canada: Their Industrial Structure and Labour Market conditions in the Early 1980s, Garnett Picot and John Heath*

41. *The Distribution of Federal/Provincial Taxes and Transfers in rural Canada, Brian B. Murphy*

42. *Foreign Multinational Enterprises and Merger Activity in Canada, John Baldwin and Richard Caves*

43. *Repeat Users of the Unemployment Insurance Program, Miles Corak*

44. *POHEM -- A Framework for Understanding and Modelling the Health of Human Population, Michael C. Wolfson*

45. *A Review of Models of Population Health Expectancy: A Micro-Simulation Perspective, Michael C. Wolfson and Kenneth G. Manton*

46. *Career Earnings and Death: A Longitudinal Analysis of Older Canadian Men, Michael C. Wolfson, Geoff Rowe, Jane Gentleman and Monica Tomiak*

47. *Longitudinal Patterns in the Duration of Unemployment Insurance Claims in Canada, Miles Corak*

48. *The Dynamics of Firm Turnover and the Competitive Process, John Baldwin*

49. *Development of Longitudinal Panel Data from Business Registers: Canadian Experience, John Baldwin, Richard Dupuy and William Penner*

50. *The Calculation of Health-Adjusted Life Expectancy for a Multi-Attribute Utility Function: A First Attempt, J.-M. Berthelot, R. Roberge and M.C. Wolfson*

51. *Testing The Robustness of Entry Barriers, J. R. Baldwin, M. Rafiquzzaman*

52. *Canada's Multinationals: Their Characteristics and Determinants, Paul K. Gorecki*

For further information, contact the Chairperson, Publications Review Committee, Analytical Studies Branch, R.H. Coats Bldg., 24th Floor, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, (613) 951-8213.