AFAMRL-TR-81-132

ESTIMATING PERCENTILES OF SKEWED DATA

MEDHAT KORNA

UNIVERSITY OF DAYTON RESEARCH INSTITUTE
300 COLLEGE PARK AVENUE
DAYTON, OHIO 45469

NOVEMBER 1981

AIR FORCE AEROSPACE MEDICAL RESEARCH LABORATORY
AEROSPACE MEDICAL DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from Air Force Aerospace Medical Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with Defense Documentation Center should direct requests for copies of this report to:

Defense Documentation Center
Cameron Station
Alexandria, Virginia 22314

TECHNICAL REVIEW AND APPROVAL

AFAMRL-TR-81-132

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

CHARLES BATES, JR.
Chief
Human Engineering Division
Air Force Aerospace Medical Research Laboratory

AIR FORCE/56780/20 January 1982 — 100

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFAMRL-TR-81-132	2. GOVT ACCESSION NO. AD-A110 323	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ESTIMATING PERCENTILES OF SKEWED DATA		5. TYPE OF REPORT & PERIOD COVERED Automated Data Systems Documentation Program for Nonparametric Percentile Calculation
		6. PERFORMING ORG. REPORT NUMBER UDR-TR-81-43
7. AUTHOR(s) Medhat Korna		8. CONTRACT OR GRANT NUMBER(s) F33615-78-C-0507
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Dayton Research Institute 300 College Park Avenue Dayton, Ohio 45469		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62202F,7184-08-24
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Aerospace Medical Research Lab, Aerospace Medical Div., Air Force Systems Command, WPAFB, Ohio 45433		12. REPORT DATE November 1981
		13. NUMBER OF PAGES 37
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computer Programs
Nonparametric Procedure
Percentiles
Probability Density Function
Mean
Standard Deviation
Iterative Procedure

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report documents the experimental and theoretical approaches taken in developing the Nonparametric Percentile (Program NPPCTL) computer program. The developed computer program computes the percentile estimates of the sample data using both the Gaussian method and the nonparametric method described in this report. The report also describes the basic equations for the nonparametric method and provides a guide to the use of the computer program (NPPCTL) in addition to the source code listing.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

SUMMARY

This report documents the experimental and theoretical approaches taken in developing the Nonparametric Percentile (program NPPCTL) computer program, and illustrates the developed method. It also provides a guide to the use of the computer program in addition to the source code listing.

A method with a similar purpose has been described by Martz (1978). But this method was found to have limitations which reduced its utility. The method described in this report removes some of these limitations.

PREFACE

This work was performed under USAF Contract F33615-78-C-0507 entitled Biomechanics of Cockpit Evaluation. The Government work unit number for this contract is 71840824. Dr. Joe W. McDaniel was the initiator and monitor of this research. The UDRI technical report number for this report is UDR-TR-81-43.

The author would like to acknowledge the contributions of Dr. Joe W. McDaniel, Mr. Paul Kikta, Mr. John Moy, and Mr. Glen Potter. In addition the author would like to thank Ms. Charlene Thompson for preparing the manuscript for publication.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION 1
INTRODUCTION

Estimating percentiles is a very important statistical tool for relating an individual to a population. For example, the percentiles of anthropometric measurements are very important in designing work stations and clothing items. Since it is often impossible to design these items to fit all population personnel without modification, the usual procedure is to design for a range of values, for example in aircraft crew station design, from the 5th percentile to the 95th percentile. The most commonly used method for estimating percentiles is the Gaussian method based on the assumption that the population is normally distributed. However, nonnormally distributed parameters do exist such as age, body skinfold, strength, endurance, and reaction time.

Edmund Churchill (1981) evaluated different methods of estimating percentiles. Thirteen methods of computing percentiles from large samples were examined using 100 random samples of each of ten variables: age, weight, stature, sitting height, hip breadth, hand length, subscapular skinfold, chest, buttock, and head circumferences. The samples' values were chosen from the 1967 U.S. Air Force Flying personnel anthropometric survey. No one method was clearly superior to all others. All methods analyzed were unsatisfactory with badly skewed data such as age; however, nonparametric estimates were not studied there.

To compute the percentiles of skewed data, a "Nonparametric Method" using a nonparametric estimate of the probability density function was developed. A nonparametric procedure is a statistical procedure which is valid irrespective of the type of the probability distribution function from which the sample is obtained.

For this study three subsets of the age data from the 1967 Anthropometric Survey of U.S. Air Force Flying personnel are considered. For the first subset, ten randomly selected samples of size 200 are drawn without replacement from a population of 2420 observations. Also drawn without replacement, for the second and third subsets, are ten randomly selected samples of sizes 150 and 100 respectively. The percentile estimates are computed using the Gaussian method and the nonparametric method. The average computed percentiles, and the individual computed percentiles are compared to the actual percentiles of the total population from which the data samples are drawn. The actual percentiles of the total population are computed using the well known counting procedure.

We observed that the nonparametric method outperforms the Gaussian method for skewed data, when estimating the 5th, 15th, 25th, 35th, 45th, 50th, 65th, 75th, 85th, and 95th percentiles.

This report describes the basic equations used in developing the computer program for the nonparametric method in addition to the source code listing. It also contains the examples used to illustrate the method, and explains the use of the program.

SECTION 2

THE NONPARAMETRIC ESTIMATE OF THE PROBABILITY DENSITY FUNCTION

Let X_1, X_2, ...X_n be a random sample of size n. Assume that the probability density function [f(x)], of the population from which the random sample is drawn, is unknown. Then the estimator [$f_n(x)$], of the probability density function [f(x)], may be represented by the following

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_n(x, X_i) \qquad (1)$$

where n is the sample size, X_i is the ith observation, and $K_n(x, X_i)$ is the smoothing function or the kernel. The idea of the estimator of the probability density function is the following. The empirical distribution function is a discrete distribution with mass $\frac{1}{n}$ placed at each of the observations. The formula in (1) smooths this probability out continuously, smoothing according to the choice of $K_n(x,X_i)$. Thus the choice of $K_n(x,X_i)$ is very important and to a large extent determines the properties of $f_n(x)$. The smoothing function used here is

$$K_n(x,X_i) = \frac{1}{2h} e^{-(|\frac{x-X_i}{h}|)} \qquad (2)$$

where h is a selected function of the sample size (n) such that $h \to 0$, at an appropriate rate, as $n \to \infty$. Of course the problem is to choose the function h = h(n) converging to 0 at an appropriate rate. If $h = cn^{-\alpha}$, $\alpha > 0$ the optimum choice of α is $\frac{1}{5}$. The optimum value of c is a function of the probability density function [f(x)], but since we are attempting to estimate f(x), it is unlikely that we will know enough to choose an optimum c. Nonetheless, choosing the constant $c > 0$, to be the standard deviation of the sample data, will be satisfactory. Thus

$$h = sn^{-\frac{1}{5}} \tag{3}$$

where s is the standard deviation of the random sample. Thus, the nonparametric estimator of the probability density function is

$$f_n(x) = \frac{1}{2nh} \sum_{i=1}^{n} e^{-\left(\left|\frac{x-X_i}{h}\right|\right)} \qquad -\infty<x<\infty \tag{4}$$

If the random sample is arranged in order of magnitude, then the γ_ξth percentile is the value of x such that γ_ξ percent of the observations is less than the value of x and $(100-\gamma_\xi)$ percent is greater. That is γ_ξ is the $(100)(\xi)$th percentile if

$$P[x \leq \gamma_\xi] = \xi \tag{5}$$

where $P[x \leq \gamma_\xi]$ is the probability distribution function. But

$$P[x \leq \gamma_\xi] = \int_{-\infty}^{\gamma_\xi} f_n(x)\, dx \tag{6}$$

Therefore

$$\xi = \int_{-\infty}^{\gamma_\xi} f_n(x)\, dx \tag{7}$$

$$= \int_{-\infty}^{\gamma_\xi} \frac{1}{2nh} \sum_{i=1}^{n} e^{-\left(\left|\frac{x-X_i}{h}\right|\right)} dx$$

$$= \frac{1}{2nh} \sum_{i=1}^{n} \int_{-\infty}^{\gamma_\xi} e^{-(|\frac{x-X_i}{h}|)} dx$$

$$= \frac{1}{2n^{4/5}s} \sum_{i=1}^{n} \int_{-\infty}^{\gamma_\xi} e^{-[\frac{n^{\frac{1}{5}}}{s}(|x-X_i|)]} dx \tag{8}$$

The developed program uses an iterative procedure to find γ_ξ which is the nonparametric estimate of the (100)(ξ)th percentile.

The program computes the percentiles of the sample data using both the Gaussian method and the nonparametric method. For the Gaussian method the following equation is used:

$$\xi = \int_{-\infty}^{\gamma_\xi} \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2 \left(\frac{x-\bar{X}}{\sigma}\right)^2} dx$$

Where γ_ξ is the (100)(ξ)th percentile, σ is the standard deviation, and $\bar{X}$ is the mean.

SECTION 3
THE STUDY

The design of this study is basically experimental rather than theoretical. The results reported in this report are obtained by randomly selecting samples of different sizes from skewed data (1967 USAF Survey age data).

In the 1967 Survey of USAF Flying Personnel conducted by the Air Force Aerospace Medical Research Laboratory (see Churchill, et al., 1977), 185 variables were measured and recorded for 2420 male pilots. For this study three subsets of sizes 200, 150, and 100 of the age data are considered. For each subset ten randomly selected samples are drawn without replacement from the population of 2420 observations.

The 5th, 15th, 25th, ..., 50th, 65th, ..., and 95th percentile estimates are computed using the Gaussian method and the nonparametric method. The average nonparametric percentile estimates and Gaussian estimates are computed for each of the three subsets considered in this study. The average computed percentiles from both methods are compared to the corresponding population percentiles. The population percentiles are computed using the well known counting method. The criteria used for comparing the Gaussian and nonparametric methods are as follows. The estimates of the percentiles should be close to the corresponding percentiles of the population from which the data sample is drawn. That is the estimate of the 1st percentile should be close to the population 1st percentile, the estimate of the 2nd percentile should be close to the population 2nd percentile, etc.

The total population arithmetic mean is 30.03 years, the standard deviation is 6.31 years, and the measure of skewness, using the third moment about the mean, is 0.76. The actual percentiles and the computed percentiles for the total population (2420 observations) using both the Gaussian and nonparametric methods are shown in Table 1. Also shown in Table 1 is the difference between each population percentile and each corresponding percentile estimate expressed as a percent of the actual percentile (Δ%). Table 2 shows the population percentiles for all 2420 observations, the average nonparametric percentiles estimates, and the average Gaussian estimates from the ten randomly selected samples of size 200. The population percentiles, the average nonparametric estimates, and the Gaussian estimates from the ten randomly selected samples of sizes 150 and 100 are shown in Tables 3 and 4 respectively. Also shown in Tables 2, 3, and 4 is the difference between every population percentile and the corresponding percentile estimates expressed as a percent of the actual percentile (Δ%).

Now let us consider the performance of the nonparametric method described in Section 2 of this report with that of the Gaussian method. As shown in Table 1, the nonparametric method outperforms the Gaussian method when estimating the 5th, 25th, 35th, 45th, 50th, 55th, and 95th percentiles. Using all 2420 observations it is observed from Tables 2, 3, and 4, that the nonparametric method outperforms the Gaussian method when estimating the 5th, 15th, 25th, 35th, 45th, 50th, 65th, and 95th percentiles for sizes 200, 150, and 100 respectively. It is also observed that the nonparametric method is superior to the Gaussian method at the lower half of the distribution since the data are skewed right (positive skewness).

In order to test the performance of the nonparametric method with that of the Gaussian method when dealing with different types of data, the AFAMRL unpublished strength data (weight holding in seconds) are considered. The 1st, 2.5th, 5th, 10th,...., 95th, 97.5th, and 99th percentiles are computed using the counting procedure, the Gaussian method, and the nonparametric method. The total population size is 1,066 observations, the arithmetic mean is 53.33 seconds, the standard deviation is 22.11 seconds, and the measure of skewness, using the third moment about the mean, is 0.95. Table 5 shows the population percentiles, Gaussian estimates, and nonparametric estimates for the total population (1,066 observations).

TABLE 1

POPULATION PERCENTILES, GAUSSIAN ESTIMATES, AND NONPARAMETRIC ESTIMATES FOR THE TOTAL POPULATION (n=2420) FOR THE AGE DATA

Percentile	Population Percentile	Gaussian Estimate	Nonparametric Estimate	Gaussian Δ%	Nonparametric Δ%
5.0	22.50	19.65	21.78	12.67	3.20
15.0	23.50	23.50	23.46	0.0	-0.17
25.0	24.50	25.78	24.70	-5.22	-0.81
35.0	25.50	27.60	26.06	-8.24	-2.20
45.0	27.50	29.24	27.65	-6.33	-0.55
50.0	28.50	30.03	28.59	-5.37	-0.32
55.0	29.50	30.83	29.69	-4.51	-0.64
65.0	32.50	32.46	32.18	0.12	0.98
75.0	35.50	35.34	35.78	0.45	-0.79
85.0	36.50	36.56	37.27	-0.16	-2.11
95.0	42.50	40.41	42.86	4.92	-0.85

TABLE 2

POPULATION PERCENTILES, AVERAGE GAUSSIAN ESTIMATES, AND AVERAGE NONPARAMETRIC ESTIMATES FOR TEN SAMPLES OF SIZE 200 FOR THE AGE DATA

Percentile	Population Percentile	Average Gaussian Estimates	Average Nonparametric Estimates	Gaussian Δ%	Nonparametric Δ%
5.0	22.50	20.24	21.01	10.04	6.62
15.0	23.50	24.64	23.50	- 5.06	0.0
25.0	24.50	26.29	25.13	- 7.31	-2.57
35.0	25.50	28.09	26.77	-10.16	-4.98
45.0	27.50	29.71	28.58	- 8.04	-3.93
50.0	28.50	30.49	29.58	- 6.98	-3.79
55.0	29.50	31.28	30.60	- 6.03	-3.72
65.0	32.50	32.89	32.80	- 1.20	-0.92
75.0	34.50	35.74	36.40	- 3.59	-5.51
85.0	36.50	36.95	38.21	- 1.23	-4.68
95.0	42.50	40.74	43.33	4.14	-1.95

TABLE 3

POPULATION PERCENTILES, AVERAGE GAUSSIAN ESTIMATES, AND AVERAGE NONPARAMETRIC ESTIMATES FOR TEN SAMPLES OF SIZE 150 FOR THE AGE DATA

Percentile	Population Percentile	Average Gaussian Estimates	Average Nonparametric Estimates	Average Gaussian Δ%	Average Nonparametric Δ%
5.0	22.50	20.24	20.84	10.04	7.38
15.0	23.50	23.98	23.42	- 2.04	0.34
25.0	24.50	26.28	25.26	- 7.27	-3.10
35.0	25.50	28.05	26.64	-10.0	-4.47
45.0	27.50	29.48	28.31	- 7.20	-2.94
50.0	28.50	30.33	29.37	- 6.42	-3.05
55.0	29.50	31.10	30.42	- 5.42	-3.12
65.0	32.50	32.67	32.66	- 0.52	-0.49
75.0	34.50	34.47	34.94	0.09	-1.28
85.0	36.50	36.68	37.76	- 0.49	-3.45
95.0	42.50	40.41	43.08	4.92	-1.36

TABLE 4

POPULATION PERCENTILES, AVERAGE GAUSSIAN ESTIMATES, AND AVERAGE NONPARAMETRIC ESTIMATES FOR TEN SAMPLES OF SIZE 100 FOR THE AGE DATA

Percentile	Population Percentile	Average Gaussian Estimates	Average Nonparametric Estimates	Average Gaussian Δ%	Average Nonparametric Δ%
5.0	22.50	20.20	20.52	10.22	8.80
15.0	23.50	23.99	23.29	- 2.09	0.89
25.0	24.50	26.22	24.95	- 7.02	-1.84
35.0	25.50	28.00	26.61	- 9.80	-4.35
45.0	27.50	29.61	28.47	- 7.67	-3.53
50.0	28.50	30.39	29.49	- 6.63	-3.47
55.0	29.50	31.25	30.59	- 5.92	-3.69
65.0	32.50	32.79	32.85	- 0.89	-1.08
75.0	34.50	34.57	35.18	- 0.20	-1.97
85.0	36.50	36.81	38.12	- 0.85	-4.44
95.0	42.50	40.58	43.19	4.52	-1.62

TABLE 5

POPULATION PERCENTILES, GAUSSIAN ESTIMATES, AND NONPARAMETRIC ESTIMATES FOR THE STRENGTH DATA (WEIGHT HOLDING IN SECONDS)

Percentile	Population Percentile	Gaussian Estimates	Nonparametric Estimates
1.0	10.00	1.90	6.24
2.5	15.00	9.99	12.43
5.0	20.00	16.96	18.09
10.0	27.00	24.98	25.40
15.0	32.00	30.42	30.42
20.0	35.00	34.71	34.40
25.0	38.00	38.43	37.83
30.0	42.00	41.74	41.01
35.0	45.00	44.82	43.96
40.0	47.00	47.74	46.67
45.0	50.00	50.55	49.27
50.0	52.00	53.33	51.82
55.0	54.00	56.12	54.41
60.0	56.00	58.93	57.06
65.0	59.00	61.85	59.87
70.0	62.00	64.92	62.93
75.0	65.00	68.24	66.43
80.0	69.00	71.95	70.55
85.0	74.00	76.24	75.77
90.0	81.00	81.68	82.68
95.0	90.00	89.71	93.00
97.5	101.00	96.67	103.50
99.0	113.00	104.77	117.52

As with the age data, the nonparametric method is superior to the Gaussian method especially at the lower end of the distribution.

In summary, based on the comparison shown in this report, the nonparametric method is superior to the Gaussian method at the lower half of the distribution since the data are skewed right (positive skewness). The criteria used for comparing the two methods are as follows. The estimates of the percentile should be close to the corresponding percentiles of the population from which the data sample is drawn.

During this study different sample sizes of the age data and other anthropometric dimensions were considered and the results were examined. For small samples ($n \leq 100$), neither of the two methods was superior to the other. But for samples greater than 100 the nonparametric method is superior to the Gaussian method for skewed data. The degree of performance of the nonparametric method was proportional to the amount of skewness.

Finally, when there is substantial reason to believe that the sample was drawn from a skewed population (that is, where the third moment about the mean is ≥ 0.6), the nonparametric method provides a better estimate of population percentiles. More effort is needed to examine the possibilities of using the method for nonskewed data (e.g. normally distributed data), and negatively skewed data.

SECTION 4
USING PROGRAM PRCNTLS

Program PRCNTLS is written in CDC EXTENDED FORTRAN IV and can be run on most large mainframe machines with minimal modifications. On a CDC 175, 47K octal words of memory were required for execution. The program is designed to compute the nonparametric percentile estimates, Gaussian percentile estimates, and the true population percentiles (optional). The nonparametric percentile estimates are computed using the method described in Section 2 of this report. The Gaussian estimates are computed using the following:

$$\xi = \int_{-\infty}^{\gamma_\xi} \frac{1}{\sigma\sqrt{2\pi}} \, e^{-1/2\left(\frac{x-\bar{x}}{\sigma}\right)^2} dx$$

Where γ_ξ is the (100)(ξ)th percentile, γ_ξ is the standard deviation, and $\bar{x}$ is the mean.

The population percentiles are computed using the counting procedure. The data are arranged in order of magnitude, and then are grouped into convenient class intervals. Then, the number of observations below each upper class limit are counted, divided by the total number of observations, and multiplied by 100 to determine the percentile rank.

4.1 THE PROGRAM OUTPUT

Program PRCNTLS writes to UNIT 6 and contains the following (see Figure 1):

(1) the variable name,
(2) the survey name,
(3) the arithmetic mean for that variable,
(4) the standard deviation,
(5) the sample size,
(6) the Gaussian percentile estimates,
(7) the nonparametric percentile estimates, and

ESTIMATED PERCENTILES OF SKEWED DATA

1 AGE (IN YEARS)

2 1967 FLYING PERSONNEL

3 MEAN.......... 31.00
4 STANDARD DEV.. 6.33
5 SAMPLE SIZE... 200

PERCENTILE	6 GAUSSIAN ESTIMATE	7 NON-PARAMETRIC ESTIMATE	8 COUNTING METHOD
1.0	16.28	17.70	22.50
2.5	18.59	19.71	22.50
5.0	20.59	21.23	23.50
10.0	22.86	22.75	23.50
15.0	24.44	23.73	24.50
20.0	25.67	24.55	24.50
25.0	26.73	25.36	25.50
30.0	27.68	26.21	25.50
35.0	28.56	27.15	26.50
40.0	29.39	28.17	27.50
45.0	30.20	29.27	29.50
50.0	31.00	30.39	30.50
55.0	31.79	31.50	31.50
60.0	32.60	32.55	32.50
65.0	33.43	33.54	33.50
70.0	34.31	34.54	34.50
75.0	35.26	35.65	35.50
80.0	36.32	36.95	35.50
85.0	37.55	38.69	37.50
90.0	39.11	41.03	41.50
95.0	41.40	43.69	43.50
97.5	43.40	45.76	43.50
99.0	45.71	48.24	45.50

Figure 1. Program PRCNTLS Sample Output.

(8) optionally, the actual population percentiles using the counting method.

Population percentiles by the Counting Method are included to show the user of the program how well the two percentile estimation techniques fared on his data.

4.2 PROGRAM INPUT

The input to program PRCNTLS is read from Unit 5 and consists of the following:

- the variable name,
- the survey name,
- the sample size,
- the counting method indicator (1 if the percentiles by the counting method are desired; 0 if not),
- the data format, and
- the data itself.

As many sets of input as desired may be run together, ending with either a blank card or an end-of-file (EOF). The general data deck layout is shown in Figure 2. The data format is as follows:

- The variable name and survey name,

  columns 1-30 the variable name (3A10)
  columns 41-70 the survey name (3A10)

- The sample size and counting method indicator,

  columns 1-5 the sample size (I5)
  columns 7 the counting method indicator (I2)

- The data format,

  columns 1-80 the data format enclosed in parenthesis (8A10)

- The data as specified in the data format.

Figure 3 is the input example that produced the output of Figure 1.

.
.
.

4) The data

3) Data format

2) Sample size and counting method indicator

1) Variable name and survey name

Figure 2. Program PRCNTLS Data Flow.

```
AGE (IN YEARS)                                  1967 FLYING PERSONNEL

  200    1

  (20F4.1)
  425 425 415 315 305 435 415 445 345 365 375 245 225 245 275 245 255 225 235 235
  275 235 365 235 365 295 315 255 275 435 485 305 305 315 295 385 405 405 295 265
  245 285 345 285 265 295 315 315 425 245 285 265 315 345 355 325 325 305 275 275
  295 275 335 325 235 245 235 225 225 235 255 235 235 235 235 325 325 285 295 345
  355 255 265 355 335 415 335 325 355 245 265 275 325 395 245 385 325 355 255 345
  415 415 345 375 365 425 425 435 365 315 245 235 235 265 255 275 275 275 235 245
  295 235 235 255 275 375 365 225 325 265 265 295 385 265 265 275 255 275 335 465
  245 475 265 295 375 365 255 265 325 265 255 325 265 255 375 315 335 315 255 365
  265 335 355 315 225 225 265 245 245 225 235 245 255 275 275 275 255 345 275 265
  335 365 365 355 335 325 265 345 275 295 355 425 325 295 375 355 355 345 335 305
```

Figure 3. Program PRCNTLS Sample Input.

APPENDIX A

THE PROGRAM LISTING

```
      PROGRAM PRCNTLS
     # (INPUT,OUTPUT,TAPE5=INPUT,TAPE6=OUTPUT)
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C *                                                                   *
C *       THIS PROGRAM COMPUTES THE NONPARAMETRIC PERCENTILE          *
C *     ESTIMATES, THE GAUSSIAN ESTIMATES, AND THE POPULATION         *
C *     PERCENTILES USING THE COUNTING METHOD (OPTIONAL).             *
C *                                                                   *
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
C *                                                                   *
C *  INPUT                                                            *
C *   1. VARIABLE NAME, SURVEY NAME             (3A10,10X,3A10)       *
C *        UP TO 30 CHARACTERS EACH                                   *
C *   2. SAMPLE SIZE, COUNT METHOD INDICATOR     (I5,I2)              *
C *        1 FOR THE COUNT METHOD INDICATOR IF THE COUNT METHOD       *
C *        PERCENCTILES ARE DESIRED; BLANK OR ZERO IF  NOT.           *
C *   3. DATA FORMAT                            (8A10)                *
C *   4. DATA                                   (AS PER FORMAT)       *
C *   5. REPITIONS OF NUMBERS 1-4 AS DESIRED                          *
C *                                                                   *
C * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
      DIMENSION D(23),GAMMA(23),P(23),ITER(23),PCNT(23),XX(2420)
      DIMENSION SURVEY(3),VRNAME(3)
      REAL NHMEAN(23),ESMEAN(23)
C
      DATA D/-2.326,-1.96,-1.645,-1.282,-1.036,-.842,-.674,
     *       -.524,-.385,-.253,-.126,
     *       0.0,.126,.253,.385,.524,.674,.842,1.036,1.282,
     *       1.645,1.96,2.326/
      DATA P/ 1.,2.5, 5.,10.,15.,20.,25.,30.,35.,40.,45.,
     *       50.,55.,60.,65.,70.,75.,80.,85.,90.,95.,97.5,99./
      DATA PCNT/23*1.E20/,BLANK/10H          /
      DATA NP/23/
   10 CONTINUE
C                      READ VARIABLE & SURVEY NAMES
      READ(5,300) VRNAME,SURVEY
      IF(VRNAME(1).EQ.BLANK.OR.EOF(5).GT.0)   STOP
C                      READ SAMPLE SIZE
      READ(5,301) NS,ICNT
C                      READ IN SAMPLE
      CALL RDDAT(XX,XBAR,XSD,NS)
C                      COUNT METHOD PERCENTILES
      IF(ICNT.GT.0) CALL CNTPRCN(XX,NS,PCNT)
C                      CALCULATE GAUSSIAN ESTIMATES
      DO 40 I=1,NP
      GAMMA(I)=P(I)/100.
      ESMEAN(I)=XBAR+XSD*D(I)
C                      CALCULATE NON-PARAMETRIC ESTIMATES
      CALL NONPAR(ESMEAN(I),GAMMA(I),XX,NHMEAN(I),NS,ITER(I),XSD)
   40 CONTINUE
```

```
C                          WRITE RESULTS
      WRITE(6,202) VRNAME,SURVEY
      WRITE(6,200) XBAR,XSD,NS
      IF(ICNT.GT.0) GO TO 50
      WRITE(6,204)
      WRITE(6,401)(P(J),ESMEAN(J),NWMEAN(J),J=1,NP)
      GO TO 10
C
   50 WRITE(6,206)
      WRITE(6,402)(P(J),ESMEAN(J),NWMEAN(J),PCNT(J),J=1,NP)
      GO TO 10
C
  200 FORMAT(//52X,*MEAN..........*,F8.2,
     *         /52X,*STANDARD DEV..*,F8.2,
     *         /52X,*SAMPLE SIZE...*,I8//)
  202 FORMAT(1H1,5(/),45X,36HESTIMATED PERCENTILES OF SKEWED DATA ,
     # ///,21X,3A10,30X,3A10  )
  204 FORMAT(57X,8HGAUSSIAN,9X,14HNONPARAMETRIC,/,37X,10HPERCENTILE,
     # 10X,8HESTIMATE,12X,6HESTIMATE,/  )
  206 FORMAT(47X,8HGAUSSIAN,9X,14HNON-PARAMETRIC,9X,8HCOUNTING,/,27X,
     # 10HPERCENTILE,10X,8HESTIMATE,12X,8HESTIMATE,13X,6HMETHOD,/  )
  300 FORMAT( 3A10,10X,3A10 )
  301 FORMAT(I5,I2)
  401 FORMAT(39X,F5.1,12X,F8.2,12X,F8.2  )
  402 FORMAT(29X,F5.1,12X,F8.2,12X,F8.2,12X,F8.2  )
C
      END
```

```
      SUBROUTINE RDDAT(XX,XBAR,XSD,NS)
      DIMENSION XX(2420),FMT(8)
C
      XBAR=0.
      XSD=0.
C                             READ INPUT FORMAT
      READ(5,100) FMT
C                             READ SAMPLE
      READ(5,FMT) (XX(I),I=1,NS)
C                             CALCULATE MEAN & STD DEV
      DO 20 I=1,NS
      XBAR=XBAR+XX(I)
      XSD=XSD+XX(I)*XX(I)
   20 CONTINUE
C
      XBAR=XBAR/NS
      XSD=XSD/NS
      XSD=XSD-XBAR**2
      XSD=SQRT(XSD)
C
  100 FORMAT(8A10)
      RETURN
      END
```

```
      SUBROUTINE NONPAR(START,ALPHA,X,END,N.INDEX,SD)
      DIMENSION X(2420)
C                        SET UP INITIAL CONDITIONS
      INDEX=0
      TOP=START
      BOTTOM=START
      END=START
      XN=N
      H=SD/XN**.2
      VALUE=XN*(2*ALPHA-1)
      DIFF=.00001*XN
      T=SD/10
C                        CALCULATE PROB OF .LE. END
    5 CONTINUE
      INDEX=INDEX+1
      SUM=0.
      DO 10 I=1,N
      XX=(END-X(I))/H
      IF(XX.LT.0.) GO TO 7
      SUM=SUM+1.-EXP(-XX)
      GO TO 10
    7 SUM=SUM-1.+EXP(XX)
   10 CONTINUE
C                        HOW CLOSE ARE WE ?
      DIST=VALUE-SUM
      IF(INDEX.GT.50.OR.ABS(DIST).LE.DIFF) RETURN
C
      IF(DIST.LT.0.) GO TO 20
      IF(END.NE.TOP) GO TO 15
C                        SHIFT INTERVAL RIGHT
      BOTTOM=TOP
      TOP=TOP+T
      END=TOP
      GO TO 5
C                        TAKE RIGHT HALF OF INTERVAL
   15 BOTTOM=END
      END=(BOTTOM+TOP)/2.
      GO TO 5
C
   20 CONTINUE
      IF(BOTTOM.NE.END) GO TO 25
C                        SHIFT INTERVAL LEFT
      TOP=BOTTOM
      BOTTOM=BOTTOM-T
      END=BOTTOM
      GO TO 5
C                        TAKE LEFT HALF OF INTERVAL
   25 TOP=END
      END=(BOTTOM+TOP)/2.
      GO TO 5
```

```
C                                                                       00016000
      END                                                               00016100
```

```
      SUBROUTINE SORT(X,N)
      DIMENSION X(1)
C                            THIS IS A SIMPLE SORT
      DO 100 I=2,N
      IM=I-1
      XX=X(I)
        DO 50 J=1,IM
        IF(X(J).LT.XX) GO TO 50
        CALL SHIFT(X,J,I)
        GO TO 100
C
   50   CONTINUE
  100 CONTINUE
C
      RETURN
      END
```

```
      SUBROUTINE SHIFT(X,J,I)
      DIMENSION X(1)
C
      INT=I-J
      XX=X(I)
      DO 10 K=1,INT
      X(I-K+1)=X(I-K)
   10 CONTINUE
      X(J)=XX
C
      RETURN
      END
```

```
      SUBROUTINE CNTPRCN(XX,NS,PCNT)
C
      DIMENSION XX(2420),GAMMA(23),P(23),PCNT(23)
      DATA P/ 1.,2.5, 5.,10.,15.,20.,25.,30.,35.,40.,45.,
     *        50.,55.,60.,65.,70.,75.,80.,85.,90.,95.,97.5,99./
C
      N=23
      CALL SORT (XX,NS)
      DO 100 I=1,N
      GAMMA(I)=P(I)/100.
      M=GAMMA(I)*NS+.5
      IF(M.GT.0) PCNT(I)=XX(M)
  100 CONTINUE
      RETURN
      END
```

REFERENCES

Churchill, Edmund, The Smoothing of Percentiles: An Evaluation of Several Methods. A report to AFAMRL/HEG, Wright-Patterson Air Force Base, Ohio (1981).

Churchill, Edmund, T. Churchill, and P. E. Kikta, The AMRL Anthropometric Data Bank Library: Volume I-V. AMRL-TR-77-1, Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio (October 1977).

Kikta, P. E. and E. Churchill, Manual of Computer Programs for the Editing of Anthropometric Survey Data. AMRL-TR-78-38, Air Force Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio (July 1978).

Martz, H. F., "Estimating Percentiles of Non-normal Anthropometric Populations," Los Alamos Scientific Laboratory, Los Alamos, New Mexico (1978).

Parzen, E., "On Estimation of a Probability Density Function and its Mode," Annals of Math. Stat., 33, pp. 1065-76 (1962).

Rosenblatt, M., "Remarks on Some Nonparametric Estimates of a Density Function," Annals of Math. Stat., 27, pp. 832-7 (1956).

Wegman, E. J., "Nonparametric Probability Density Estimating: I. A Summary of Available Methods," Technometrics, 14, pp. 533-46 (1972a).

Wegman, E. J., "Nonparametric Probability Density Estimation: II. A Comparison of Density Estimation Methods," Journal of Statistical Computation and Simulation, 1, pp. 225-45 (1972b).

*U.S.Government Printing Office: 1982 — 559-007/6018