ESTIMATING PERCENTILES OF SKEWED DATA

MEDHAT KORNA

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

NOVEMBER 1981

Approved for public release; distribution unlimited
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/AFM/175/20 January 1982 — 100
**ESTIMATING PERCENTILES OF SKewed DATA**

**AUTHOR(s)**
Medhat Korna

**PERFORMING ORGANIZATION NAME AND ADDRESS**
University of Dayton Research Institute
300 College Park Avenue
Dayton, Ohio 45469

**CONTROLING OFFICE NAME AND ADDRESS**
Air Force Aerospace Medical Research Lab, Aerospace Medical Div.
Air Force Systems Command, WPAFB, Ohio 45433

**REPORT DATE**
November 1981

**NUMBER OF PAGES**
37

**DISTRIBUTION STATEMENT (of this Report)**
Approved for public release; distribution unlimited

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

**SUPPLEMENTARY NOTES**

**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

**ABSTRACT**
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.
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.
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

Section                                                                 Page

1  INTRODUCTION                                                        5
2  THE NONPARAMETRIC ESTIMATE OF THE PROBABILITY DENSITY FUNCTION     7
3  THE STUDY                                                           10
4  USING PROGRAM PRCNTLS                                              18
   4.1 THE PROGRAM OUTPUT                                              18
   4.2 PROGRAM INPUT                                                   20
APPENDIX A: THE PROGRAM LISTING                                      23
REFERENCES                                                           31

LIST OF ILLUSTRATIONS

1  Program PRCNTLS Sample Output.                                     19
2  Program PRCNTLS Data Flow.                                          21
3  Program PRCNTLS Sample Input.                                       22
LIST OF TABLES

Table | Description | Page
--- | --- | ---
1 | POPULATION PERCENTILES, GAUSSIAN ESTIMATES, AND NONPARAMETRIC ESTIMATES FOR THE TOTAL POPULATION (n=2420) FOR THE AGE DATA | 12
2 | POPULATION PERCENTILES, AVERAGE GAUSSIAN ESTIMATES, AND AVERAGE NONPARAMETRIC ESTIMATES FOR TEN SAMPLES OF SIZE 200 FOR THE AGE DATA | 13
3 | POPULATION PERCENTILES, AVERAGE GAUSSIAN ESTIMATES, AND AVERAGE NONPARAMETRIC ESTIMATES FOR TEN SAMPLES OF SIZE 150 FOR THE AGE DATA | 14
4 | POPULATION PERCENTILES, AVERAGE GAUSSIAN ESTIMATES, AND AVERAGE NONPARAMETRIC ESTIMATES FOR TEN SAMPLES OF SIZE 100 FOR THE AGE DATA | 15
5 | POPULATION PERCENTILES, GAUSSIAN ESTIMATES, AND NONPARAMETRIC ESTIMATES FOR THE STRENGTH DATA (WEIGHT HOLDING IN SECONDS) | 16
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 \( X_1, X_2, \ldots, X_n \) 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

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

where \( n \) is the sample size, \( X_i \) is the \( i \)th 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

\[
K_n(x, X_i) = \frac{1}{2h} e^{-\left(\frac{x-X_i}{h}\right)^2}
\]

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 \( \alpha \) is \( \frac{1}{2} \). 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
where $s$ is the standard deviation of the random sample. Thus, the nonparametric estimator of the probability density function is

$$ f_n(x) = \frac{1}{2nh} \sum_{i=1}^{n} e^{-\frac{|x-x_i|}{h}} \quad -\infty < x < \infty $$

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

$$ P[x < \gamma] = \xi $$

where $P[x < \gamma]$ is the probability distribution function. But

$$ P[x < \gamma] = \int_{-\infty}^{\gamma} f_n(x) \, dx $$

Therefore

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

$$ = \int_{-\infty}^{\gamma} \frac{1}{2nh} \sum_{i=1}^{n} e^{-\frac{|x-x_i|}{h}} \, dx $$
\[ \frac{1}{2nh} \sum_{i=1}^{n} \int_{-\infty}^{\frac{x-X_i}{h}} e^{-\frac{|x|}{h}} \, dx \]

The developed program uses an iterative procedure to find \( Y_\xi \) which is the nonparametric estimate of the \((100)(\xi)\)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:

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

Where \( Y_\xi \) is the \((100)(\xi)\)th percentile, \( \sigma \) is the standard deviation, and \( X \) 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>
<thead>
<tr>
<th>Population</th>
<th>Gaussian Estimate</th>
<th>Nonparametric Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>22.50</td>
<td>19.65</td>
</tr>
<tr>
<td>15.0</td>
<td>23.50</td>
<td>23.46</td>
</tr>
<tr>
<td>25.0</td>
<td>24.50</td>
<td>24.70</td>
</tr>
<tr>
<td>35.0</td>
<td>25.50</td>
<td>26.06</td>
</tr>
<tr>
<td>45.0</td>
<td>27.50</td>
<td>27.63</td>
</tr>
<tr>
<td>55.0</td>
<td>28.50</td>
<td>30.03</td>
</tr>
<tr>
<td>65.0</td>
<td>29.50</td>
<td>30.83</td>
</tr>
<tr>
<td>75.0</td>
<td>31.00</td>
<td>32.46</td>
</tr>
<tr>
<td>85.0</td>
<td>32.50</td>
<td>35.34</td>
</tr>
<tr>
<td>95.0</td>
<td>34.50</td>
<td>36.56</td>
</tr>
</tbody>
</table>

12
<table>
<thead>
<tr>
<th>Percentile</th>
<th>Population Percentile</th>
<th>Average Gaussian Estimates</th>
<th>Average Nonparametric Estimates</th>
<th>Gaussian Δ%</th>
<th>Nonparametric Δ%</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>22.50</td>
<td>20.24</td>
<td>21.01</td>
<td>10.04</td>
<td>6.62</td>
</tr>
<tr>
<td>15.0</td>
<td>23.50</td>
<td>24.64</td>
<td>23.50</td>
<td>-5.06</td>
<td>0.0</td>
</tr>
<tr>
<td>25.0</td>
<td>24.50</td>
<td>26.29</td>
<td>25.13</td>
<td>-7.31</td>
<td>-2.57</td>
</tr>
<tr>
<td>35.0</td>
<td>25.50</td>
<td>28.09</td>
<td>26.77</td>
<td>-10.16</td>
<td>-4.98</td>
</tr>
<tr>
<td>45.0</td>
<td>27.50</td>
<td>29.71</td>
<td>28.58</td>
<td>-8.04</td>
<td>-3.93</td>
</tr>
<tr>
<td>55.0</td>
<td>28.50</td>
<td>30.49</td>
<td>29.58</td>
<td>-6.98</td>
<td>-3.79</td>
</tr>
<tr>
<td>65.0</td>
<td>29.50</td>
<td>31.28</td>
<td>30.60</td>
<td>-6.03</td>
<td>-3.72</td>
</tr>
<tr>
<td>75.0</td>
<td>32.50</td>
<td>32.89</td>
<td>32.80</td>
<td>-1.20</td>
<td>-0.92</td>
</tr>
<tr>
<td>85.0</td>
<td>34.50</td>
<td>35.74</td>
<td>36.40</td>
<td>-3.59</td>
<td>-5.51</td>
</tr>
<tr>
<td>95.0</td>
<td>42.50</td>
<td>40.74</td>
<td>43.33</td>
<td>4.14</td>
<td>-1.95</td>
</tr>
</tbody>
</table>
### Table 3

Population percentiles, average Gaussian estimates, and average nonparametric estimates for ten samples of size 150 for the age data

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Population Percentile</th>
<th>Average Gaussian Estimates</th>
<th>Average Nonparametric Estimates</th>
<th>Average Gaussian $\Delta_%$</th>
<th>Average Nonparametric $\Delta_%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>22.50</td>
<td>20.24</td>
<td>20.84</td>
<td>10.04</td>
<td>7.38</td>
</tr>
<tr>
<td>15.0</td>
<td>23.50</td>
<td>23.98</td>
<td>23.42</td>
<td>-2.04</td>
<td>0.34</td>
</tr>
<tr>
<td>25.0</td>
<td>24.50</td>
<td>26.28</td>
<td>25.26</td>
<td>-7.27</td>
<td>-3.10</td>
</tr>
<tr>
<td>35.0</td>
<td>25.50</td>
<td>28.05</td>
<td>26.64</td>
<td>-10.0</td>
<td>-4.47</td>
</tr>
<tr>
<td>45.0</td>
<td>27.50</td>
<td>29.48</td>
<td>28.31</td>
<td>-7.20</td>
<td>-2.94</td>
</tr>
<tr>
<td>50.0</td>
<td>28.50</td>
<td>30.33</td>
<td>29.37</td>
<td>-6.42</td>
<td>-3.05</td>
</tr>
<tr>
<td>55.0</td>
<td>29.50</td>
<td>31.10</td>
<td>30.42</td>
<td>-5.42</td>
<td>-3.12</td>
</tr>
<tr>
<td>65.0</td>
<td>32.50</td>
<td>32.67</td>
<td>32.66</td>
<td>-0.52</td>
<td>-0.49</td>
</tr>
<tr>
<td>75.0</td>
<td>34.50</td>
<td>34.47</td>
<td>34.94</td>
<td>0.09</td>
<td>-1.28</td>
</tr>
<tr>
<td>85.0</td>
<td>36.50</td>
<td>36.68</td>
<td>37.76</td>
<td>-0.49</td>
<td>-3.45</td>
</tr>
<tr>
<td>95.0</td>
<td>42.50</td>
<td>40.41</td>
<td>43.08</td>
<td>4.92</td>
<td>-1.36</td>
</tr>
</tbody>
</table>
### Table 4

Population Percentiles, Average Gaussian Estimates, and Average Nonparametric Estimates for Ten Samples of Size 100 for the Age Data

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Population Percentile</th>
<th>Average Gaussian Estimates</th>
<th>Average Nonparametric Estimates</th>
<th>Average Gaussian Δ%</th>
<th>Average Nonparametric Δ%</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0</td>
<td>22.50</td>
<td>20.20</td>
<td>20.52</td>
<td>10.22</td>
<td>8.80</td>
</tr>
<tr>
<td>15.0</td>
<td>23.50</td>
<td>23.99</td>
<td>23.29</td>
<td>-2.09</td>
<td>0.89</td>
</tr>
<tr>
<td>25.0</td>
<td>24.50</td>
<td>26.22</td>
<td>24.95</td>
<td>-7.02</td>
<td>-1.84</td>
</tr>
<tr>
<td>35.0</td>
<td>25.50</td>
<td>28.00</td>
<td>26.61</td>
<td>-9.80</td>
<td>-4.35</td>
</tr>
<tr>
<td>45.0</td>
<td>27.50</td>
<td>29.61</td>
<td>28.47</td>
<td>-7.67</td>
<td>-3.53</td>
</tr>
<tr>
<td>50.0</td>
<td>28.50</td>
<td>30.39</td>
<td>29.49</td>
<td>-6.63</td>
<td>-3.47</td>
</tr>
<tr>
<td>55.0</td>
<td>29.50</td>
<td>31.25</td>
<td>30.59</td>
<td>-5.92</td>
<td>-3.69</td>
</tr>
<tr>
<td>65.0</td>
<td>32.50</td>
<td>32.79</td>
<td>32.85</td>
<td>-0.89</td>
<td>-1.08</td>
</tr>
<tr>
<td>75.0</td>
<td>34.50</td>
<td>34.57</td>
<td>35.18</td>
<td>-0.20</td>
<td>-1.97</td>
</tr>
<tr>
<td>85.0</td>
<td>36.50</td>
<td>36.81</td>
<td>38.12</td>
<td>-0.85</td>
<td>-4.44</td>
</tr>
<tr>
<td>95.0</td>
<td>42.50</td>
<td>40.58</td>
<td>43.19</td>
<td>4.52</td>
<td>-1.62</td>
</tr>
</tbody>
</table>
TABLE 5

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

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

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

Where \( \gamma_\xi \) is the \((100)\)th percentile, \( \gamma_{\xi} \) 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

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Gaussian Estimate</th>
<th>Non-Parametric Estimate</th>
<th>Counting Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>16.28</td>
<td>17.70</td>
<td>22.50</td>
</tr>
<tr>
<td>2.5</td>
<td>18.59</td>
<td>19.71</td>
<td>22.50</td>
</tr>
<tr>
<td>5.0</td>
<td>20.59</td>
<td>21.23</td>
<td>23.50</td>
</tr>
<tr>
<td>10.0</td>
<td>22.86</td>
<td>22.75</td>
<td>23.50</td>
</tr>
<tr>
<td>15.0</td>
<td>24.44</td>
<td>23.73</td>
<td>24.50</td>
</tr>
<tr>
<td>20.0</td>
<td>25.67</td>
<td>24.55</td>
<td>24.50</td>
</tr>
<tr>
<td>25.0</td>
<td>26.73</td>
<td>25.36</td>
<td>25.50</td>
</tr>
<tr>
<td>30.0</td>
<td>27.68</td>
<td>26.21</td>
<td>25.50</td>
</tr>
<tr>
<td>35.0</td>
<td>29.56</td>
<td>27.15</td>
<td>26.50</td>
</tr>
<tr>
<td>40.0</td>
<td>29.39</td>
<td>26.17</td>
<td>27.50</td>
</tr>
<tr>
<td>45.0</td>
<td>30.20</td>
<td>29.27</td>
<td>29.50</td>
</tr>
<tr>
<td>50.0</td>
<td>31.00</td>
<td>30.39</td>
<td>30.50</td>
</tr>
<tr>
<td>55.0</td>
<td>31.79</td>
<td>31.50</td>
<td>31.50</td>
</tr>
<tr>
<td>60.0</td>
<td>32.60</td>
<td>32.55</td>
<td>32.50</td>
</tr>
<tr>
<td>65.0</td>
<td>33.43</td>
<td>33.54</td>
<td>33.50</td>
</tr>
<tr>
<td>70.0</td>
<td>34.31</td>
<td>34.54</td>
<td>34.50</td>
</tr>
<tr>
<td>75.0</td>
<td>35.26</td>
<td>35.65</td>
<td>35.50</td>
</tr>
<tr>
<td>80.0</td>
<td>36.32</td>
<td>36.95</td>
<td>35.50</td>
</tr>
<tr>
<td>85.0</td>
<td>37.55</td>
<td>38.69</td>
<td>37.50</td>
</tr>
<tr>
<td>90.0</td>
<td>39.11</td>
<td>41.03</td>
<td>41.50</td>
</tr>
<tr>
<td>95.0</td>
<td>41.40</td>
<td>43.69</td>
<td>43.50</td>
</tr>
<tr>
<td>97.5</td>
<td>41.40</td>
<td>45.76</td>
<td>43.50</td>
</tr>
<tr>
<td>99.0</td>
<td>45.71</td>
<td>48.24</td>
<td>45.50</td>
</tr>
</tbody>
</table>

**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.
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
295 275 335 325 235 245 235 225 225 235 235 235 235 235 235 235 325 285 295 345
355 255 265 355 335 415 335 325 355 245 265 275 325 395 245 385 325 355 255 345
415 415 343 375 365 425 425 435 365 315 245 235 235 265 255 275 275 275 235 245
295 235 235 255 275 375 365 225 325 265 265 295 385 265 265 275 255 275 315 465
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)
  *
  * THIS PROGRAM COMPUTES THE NONPARAMETRIC PERCENTILE
  * ESTIMATES, THE GAUSSIAN ESTIMATES, AND THE POPULATION
  * PERCENTILES USING THE COUNTING METHOD (OPTIONAL).
  *
  * INPUT
  * 1. VARIABLE NAME, SURVEY NAME (3A0,10X,3Al0)
  * UP TO 30 CHARACTERS EACH
  * 2. SAMPLE SIZE, COUNT METHOD INDICATOR (I5,I2)
  * 1 FOR THE COUNT METHOD INDICATOR IF THE COUNT METHOD
  * PERCENTILES ARE DESIRED; BLANK OR ZERO IF NOT.
  * 3. DATA FORMAT (8A10)
  * 4. DATA (AS PER FORMAT)
  * 5. REPETIONS OF NUMBERS 1-4 AS DESIRED
  *
  DIMENSION 0(23),GAMMA(23),P(23),ITER(23),PCNT(23),XX(2420)
  REAL NWMEAN(23),ESMEAN(23)
  DATA D/-2.326,-1.96,-1.645,-1.282,-1.036,-0.8142,-.674,
  * -0.524,-.385,-.253,-.126,
  * 0.0,+.126,.253,.385,.524,.674,.8142,1.036,1.282,
  * 1.645,1.96,2.326/.
  DATA P/0.,.25,.5,10.,15.,20.,25.,30.,35.,40.,45.,
  * 50.,55.,60.,65.,70.,75.,80.,85.,90.,95.,97.5,99/,.
  DATA PCNT/231*.EZO/,BLANK/tOH
  DATA NP/23/
  CONTINUE
  READ VARIABLE & SURVEY NAMES
  READ(5,300) VRNAME,SURVEY
  IF(VRNAME.EQ.BLANK.OR.EOF(5).GT.0) STOP
  READ SAMPLE SIZE
  READ(5,301) NS,ICNT
  READ (5,301) NS,ICNT
  CALL RDATA(XX,XBAR,XSD,NS)
  CALL COUNT METHOD PERCENTILES
  IF(ICNT.GT.0) CALL CNTPRCN(XX,NS,PCNT)
  CALL CALCULATE GAUSSIAN ESTIMATES
  DO 40 I=1,NP
    GAMMA(I)=P(I)/100.
    ESMEAN(I)=XBAR*XSD*D(I)
    CALL CALCULATE NON-PARAMETRIC ESTIMATES
    CALL NONPAR(ESMEAN(I),GAMMA(I),XX,NWMEAN(I),NS,ITER(I),XSD)
  CONTINUE
 23
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 DEVIATION.*,F8.2,
* /52X,*SAMPLE SIZE....*,I8/*)
202 FORMAT(1H1,5(/),49X,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( 3A2.0,I10X,3AL0
301 FORMAT(15,22
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 RDOAT(XX,XBAR,XSD,NS)
DIMENSION XX(2420),FMT(8)
C
XBAR=0.
XSD=0.
C
READ(5,100) FMT
C
READ(5,FMT) (XX(I),I=1,NS)
C
DO 20 I=1,NS
XBAR=XX(I)+XBAR
XSD=XSD+XX(I)**2
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, SO)
DIMENSION X(2420)

C INDEX=0
TOP=START
BOTTOM=START
XN=START

C SET UP INITIAL CONDITIONS
M*SO/XN**.2
VALUE=XN*(2*ALPHA-1)
DIFF=.00001*XN
T=SO/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.SUM.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 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 PRINT INDEX, START, END, X, SO, VALUE, DIFF, T
C STOP
C END
SUBROUTINE SORT(X,N)
DIMENSION X(N)

C THIS IS A SIMPLE SORT

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

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
N=GAMMA(I) * NS+.5
IF(N.GT.0) PCNT(I)=XX(M)
100 CONTINUE
RETURN
END
REFERENCES


