SAS® Technical Report A-106

Probability Plotting

5901

SAS® SAS Institute Inc.

Probability Plotting

by

Daniel M. Chilko

SAS® Technical Report A-106

SAS Institute Inc.
SAS Circle, Box 8000
Cary, NC 27512-8000

The correct bibliographic citation for this technical report is as follows: SAS Institute Inc., SAS® Technical Report A-106, Probability Plotting, Cary, NC: SAS Institute Inc., 1983.

SAS® Technical Report A-106, Probability Plotting

ABSTRACT

This report describes probability plotting using SAS® procedures and functions.

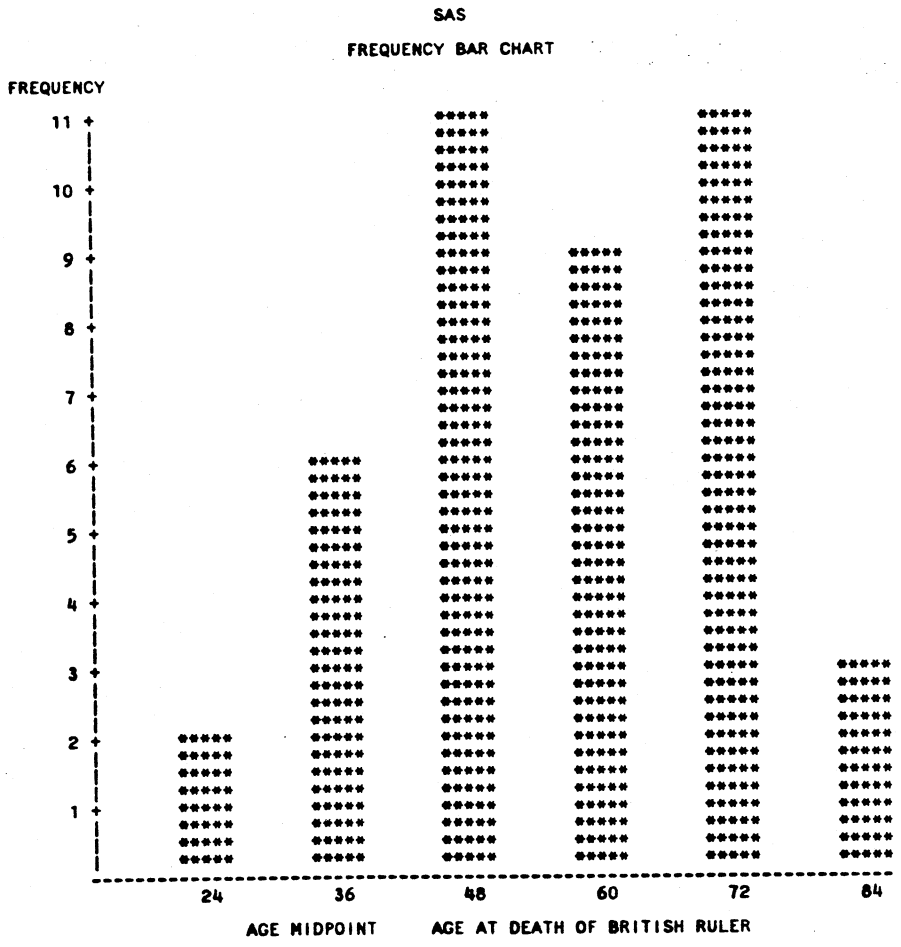
Histograms

The simplest and most frequently used graphical method for presenting sample data is a histogram or bar chart. Its use can provide answers to the following questions:

- What are the extreme values in the data?
- How many modes does the data have?
- Are the data symmetric or skewed?
- Are there outliers or gaps?

Figure 1 is a histogram produced by the SAS procedure CHART that shows the ages at death of British rulers since William the Conqueror.

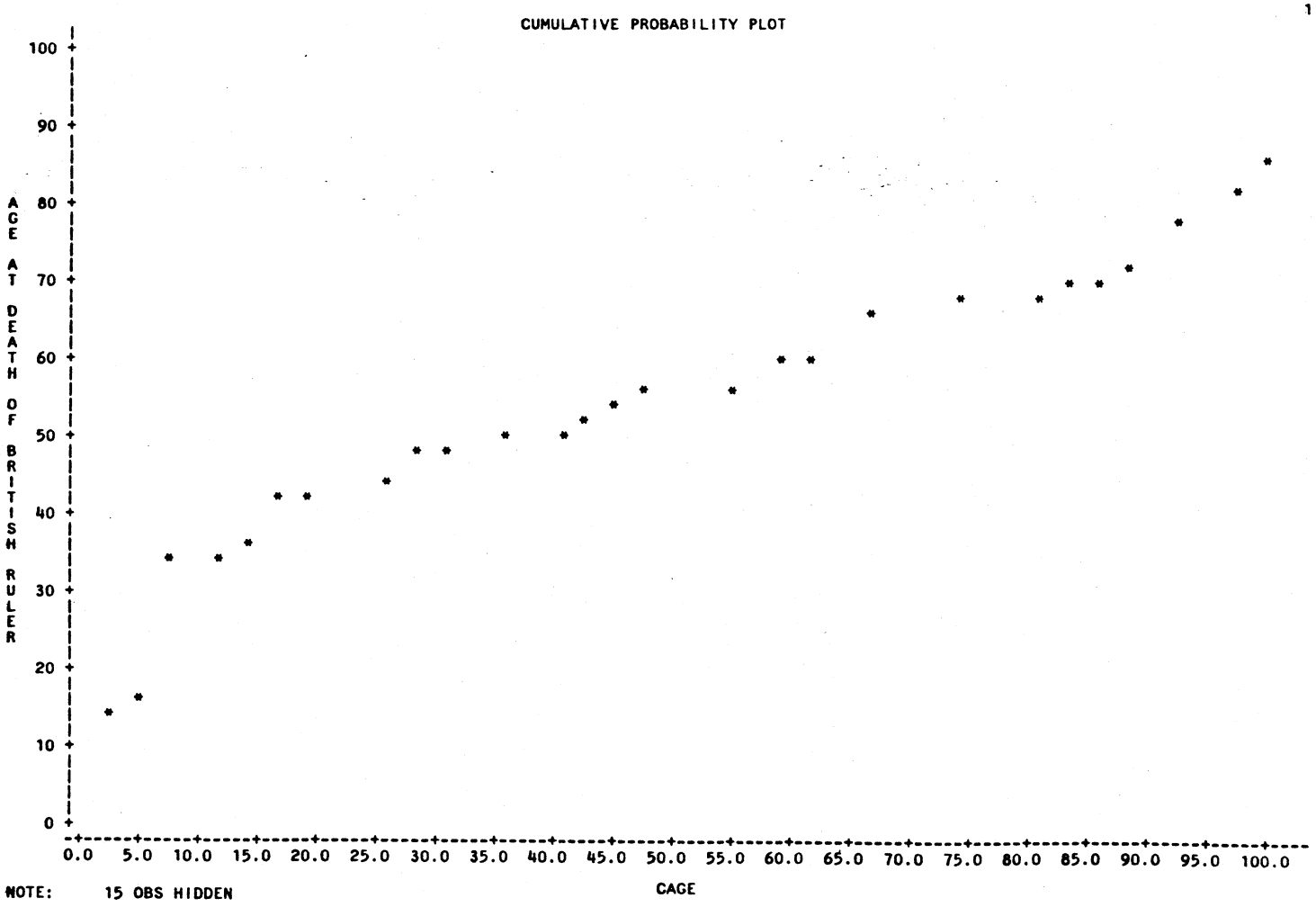
```
PROC CHART; VBAR AGE;
```



Although a histogram reveals the general shape of a distribution, it is difficult to determine from a histogram whether or not some hypothesized distribution will fit the data. Can you tell, for example, if the data shown in Figure 1 are normally distributed?

A cumulative percentage plot, made by plotting the cumulative percentage against the corresponding value of the variable of interest, can also be used to represent sample data. Using the SAS procedures RANK and PLOT, a cumulative percentage plot is produced for the ages of British rulers (Figure 2). This plot, with its characteristic S-shaped curve, is sometimes called a percentage ogive. It is difficult to work with because of its point of inflection. Rescaling the horizontal axis to form a straight line is a better alternative, since a straight line is simpler and makes interpolation easier.

```
PROC RANK DATA=RULERS PERCENT OUT=C;
    VAR AGE;
    RANKS CAGE;
PROC PLOT NOLEGEND;
    TITLE CUMULATIVE PROBABILITY PLOT;
    PLOT AGE*CAGE='*';
    FORMAT CAGE 5.1
```



Probability Plots

If the rescaled axis is based on the probability function of some hypothesized distribution, the plot is called a probability plot. A probability plot rescales the percentage axis according to some probability distribution so that, if you have chosen the proper distribution, the resulting cumulative percentage plot resembles a straight line.

This rescaling is based on the following considerations:

If F(x) is a cumulative distribution function; that is, if

$$F(x) = P(X < x),$$

and $x_1, x_2, \ldots, x_n$ are the ordered observations from a sample of size n, then

$$E(F(x_i)) = i/(n+1).$$

The rescaling of the axis is achieved by finding a set of values, $y_1, y_2, \ldots, y_n$, say, such that

$$F(y_i) = i/(n+1).$$

Plotting the pairs $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$ should result in a straight line if the x's are from a sample having a distribution with cumulative distribution function F(x). Since it is reasonable to consider a set of sample data as dependent upon a distribution function, probability plots use the variable of interest for the vertical axis and the probability distribution scaled values for the horizontal axis.

The problem of rescaling the cumulative percentage axis now becomes the problem of finding the inverse cumulative distribution function. That is,

$$y_i = F^{-1}(i/(n+1)).$$

Furthermore, the rescaling is independent of the scale and location parameters of the probability distribution, so that the rescaling problem reduces to finding the inverse of the cumulative distribution of a standardized random variable. For example, if x has a probability distribution with location parameter A and scale parameter B, a plot of the pairs (x_1,z_1), (x_2,z_2), ... (x_n,z_n) where $z_i=(y_i-A)/B$ would still result in a straight line.

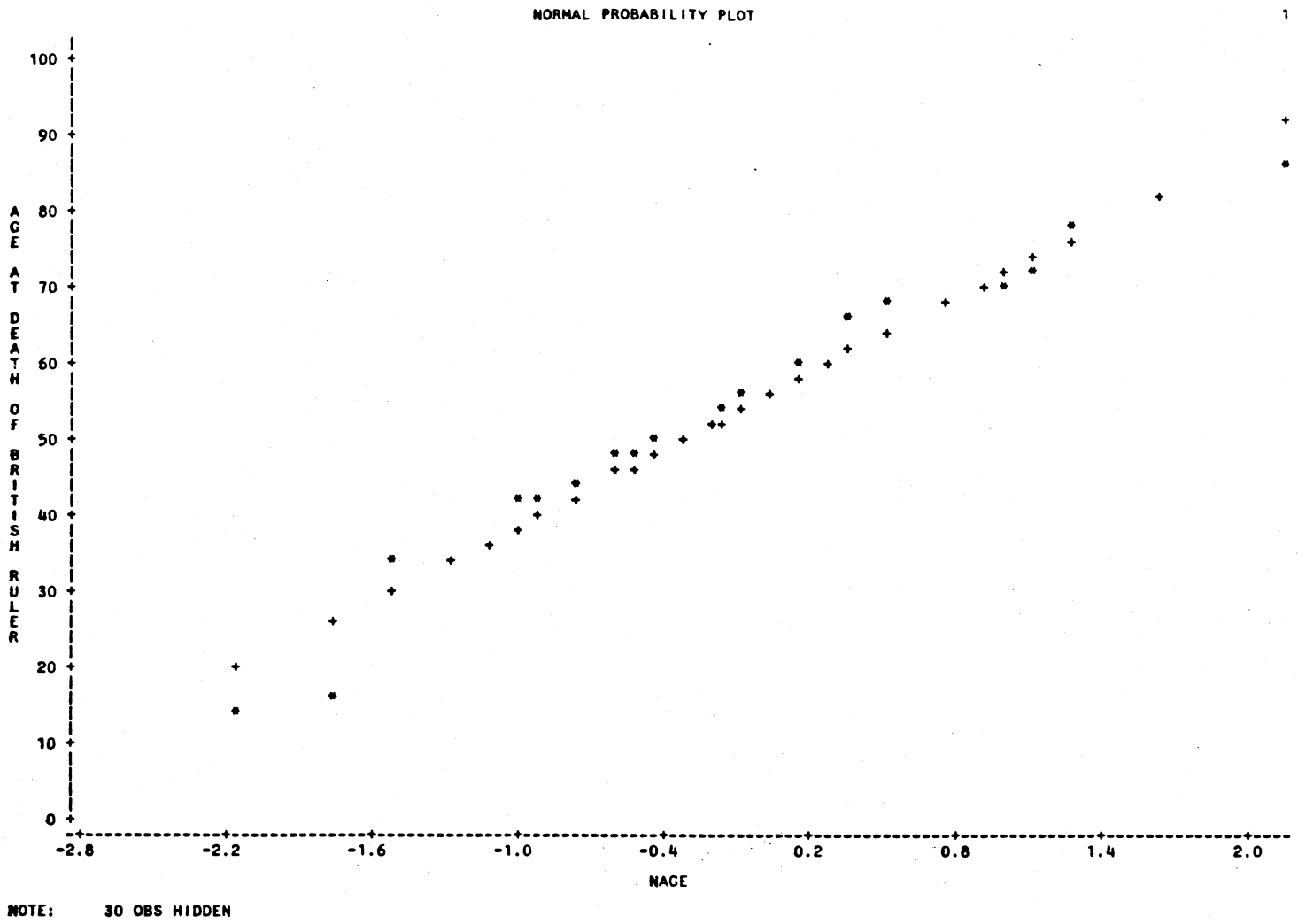
Normal Probability Plots

Many statistical tests assume that the data have a normal distribution. It is true that variables having a normal or near normal distribution occur often in nature, probably because the normal distribution is the limiting distribution of a random variable representing the sum of a series of independent random variables.

Normal probability plots can be produced using the SAS software PROC RANK and PROC PLOT, and are often used as an informal evaluation hypothesis that a sample comes from a normal distribution. If a normal distribution with the same mean and standard deviation as the sample data is plotted on the same graph, it can be used as a reference line for evaluating the distribution of the sample data.

Figure 3 shows a normal probability plot for the ages of British rulers. Since the normal probability plot of the sample data corresponds so closely to the reference line, denoted by the plotting symbol (+), this plot strongly supports the hypothesis of normality.

```
*  CALCULATE NORMAL SCORES;
PROC RANK DATA=RULERS NORMAL=BLOM OUT=R;
    VAR AGE;
    RANKS NAGE;
*  CALCULATE MEAN AND STD;
PROC MEANS NOPRINT;
    VAR AGE;
    OUTPUT OUT=M MEAN=MEAN STD=STD;
*  CREATE DATA SET WITH NORMAL SCORES AND REFERENCE POINTS;
DATA;
    SET M;
LOOP:  SET R;
    EAGE = MEAN + NAGE*STD;
    OUTPUT; GOTO LOOP;
PROC PLOT NOLEGEND;
    TITLE NORMAL PROBABILITY PLOT;
    PLOT AGE*NAGE='*' EAGE*NAGE='+' / OVERLAY;
```



Half-Normal Plots

When a random variable has a normal distribution with mean zero, the absolute value of this random variable is said to have a half-normal distribution.

In the linear regression framework

$$Y = X\,B + E,$$

where E is usually assumed to be a vector of independently and identically distributed random variables, each normally distributed with mean zero and constant variance.

In a regression analysis, the differences between observed and predicted values are called residuals. That is,

$$r = Y - \hat{Y}$$

where

$$\hat{Y} = X\,b$$

and b are the least squares estimates.

If the underlying model assumptions are true, then the r's have normal distributions, each with mean zero. They do not, in general, have constant variance; nor are they independently distributed.

Half-normal probability plots provide an informal test of normality of the residuals in a regression analysis. Half-normal plots show more sensitivity to kurtosis at the expense of not revealing skewness. They can be produced using the SAS procedures RANK and PLOT. A detailed discussion of producing half-normal probability plots is given in the SAS Technical Report A-102, "SAS Regression Applications."

Gamma Probability Plots

While the normal distribution is singularly important in statistics, the gamma probability distribution is also encountered frequently.

The general three-parameter gamma distribution is

$$f(y;\alpha,\lambda,\eta) = \frac{\lambda^{\eta}(y-\alpha)^{\eta-1}e^{-\lambda(y-\alpha)}}{\Gamma(\eta)} \quad y \geq \alpha$$

where α is the location, λ is the scale, and η is the shape parameter.

This distribution is translated into a gamma distribution with only a shape parameter whenever $x = \lambda(y - \alpha)$:

$$f(x;\eta) = \frac{e^{-x}x^{\eta-1}}{\Gamma(\eta)} \qquad x \geq 0$$

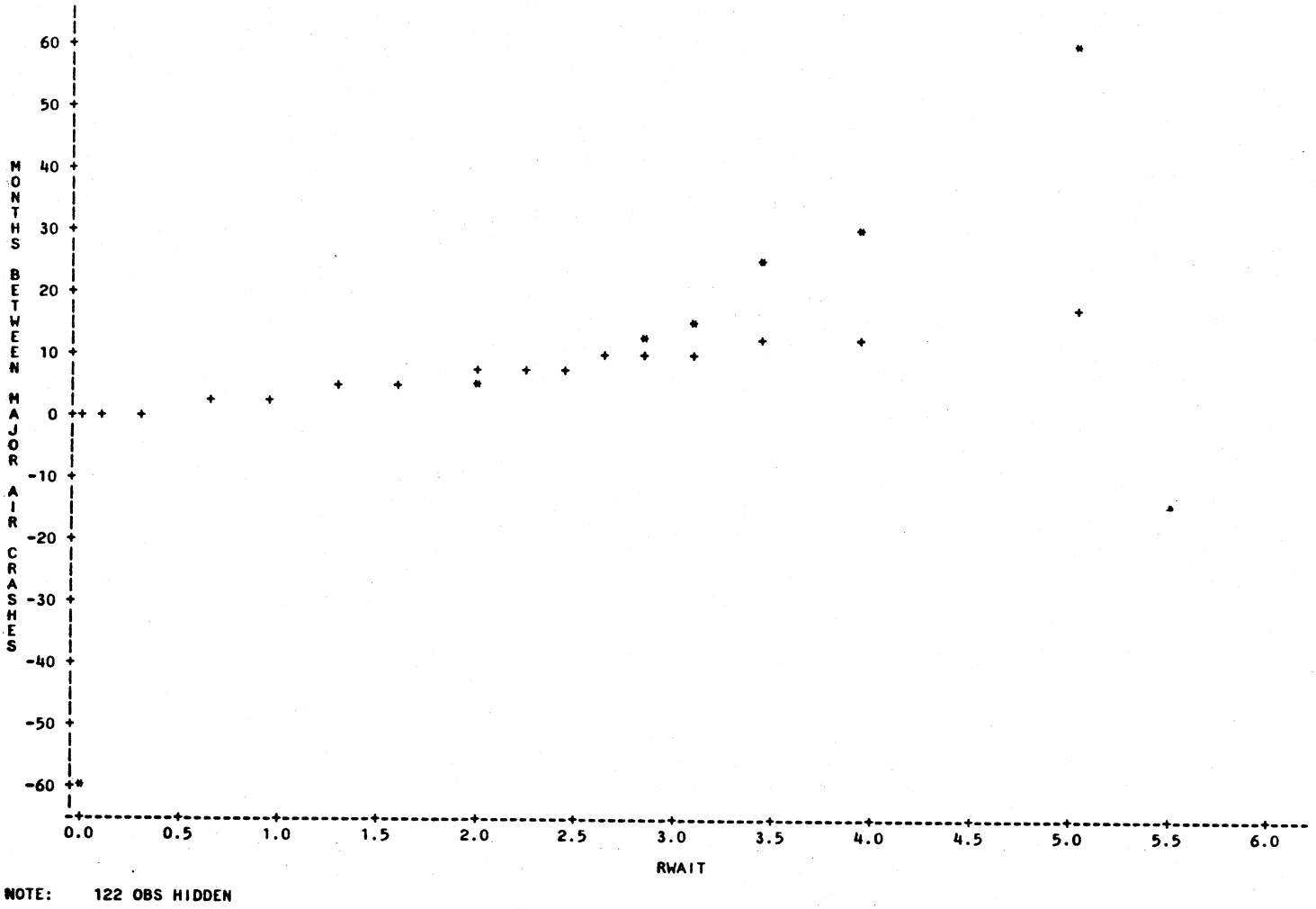
The chi-square and exponential distributions are just special cases of the gamma distribution. A chi-square random variable with degrees of freedom d is a gamma random variable with η=d/2; while an exponential distribution is a gamma distribution with η=1.

The SAS function GAMINV can be used to produce gamma probability plots. For example, if X=GAMINV(PROB,ETA), then PROB is the value of the integral from 0 to X of a gamma distribution with shape parameter ETA.

Exponential Probability Plots

The exponential distribution is often used to characterize failure or waiting time distributions. Figure 4 is an exponential probability plot produced by SAS software and shows the time in months between major airplane crashes during the period 1951-1973. The plot includes a reference line corresponding to an exponential distribution with the same mean (3.7 months) as the sample data; this helps in spotting the three outliers (24, 27, and 29 months). These outliers represent the relatively long periods of time that were free from crashes. Except for these outliers, the plot gives support to the hypothesis that the data come from an exponential distribution.

```
*RANK THE WAITING TIMES;
PROC RANK OUT=R;
    VAR WAIT;
    RANKS RWAIT;
*CALCULATE AN AVERAGE WAITING TIME;
PROC MEANS NOPRINT;
    VAR WAIT;
    OUTPUT OUT=M N=N MEAN=MEAN;
*CALCULATE EXPONENTIAL QUANTILES AND REFERENCE POINTS;
DATA;
    SET M;
LOOP:  SET R;
       RWAIT = GAMINV((RWAIT-.5)/N,1.0);
       EWAIT = MEAN*RWAIT; OUTPUT; GOTO LOOP;
PROC PLOT NOLEGEND;
    PLOT WAIT*RWAIT='*' EWAIT*RWAIT='+' / OVERLAY;
```



NOTE: 122 OBS HIDDEN

Chi-square Probability Plots

Sample variances or mean squares from a normal population have a chi-square distribution indexed by a quantity called the degrees of freedom. A chi-square probability plot can be used to provide an informal test of the homogeneity of sample variances.

In Figure 5, a chi-square probability plot of six sample variances is shown. Six types of yarns were evaluated for strength by measuring the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn. Fifty-four looms (9 of each type) were evaluated. The probability plots suggest that for five of the types, the variances are homogeneous. One type of yarn appears to have a strength that is twice as variable as the others.

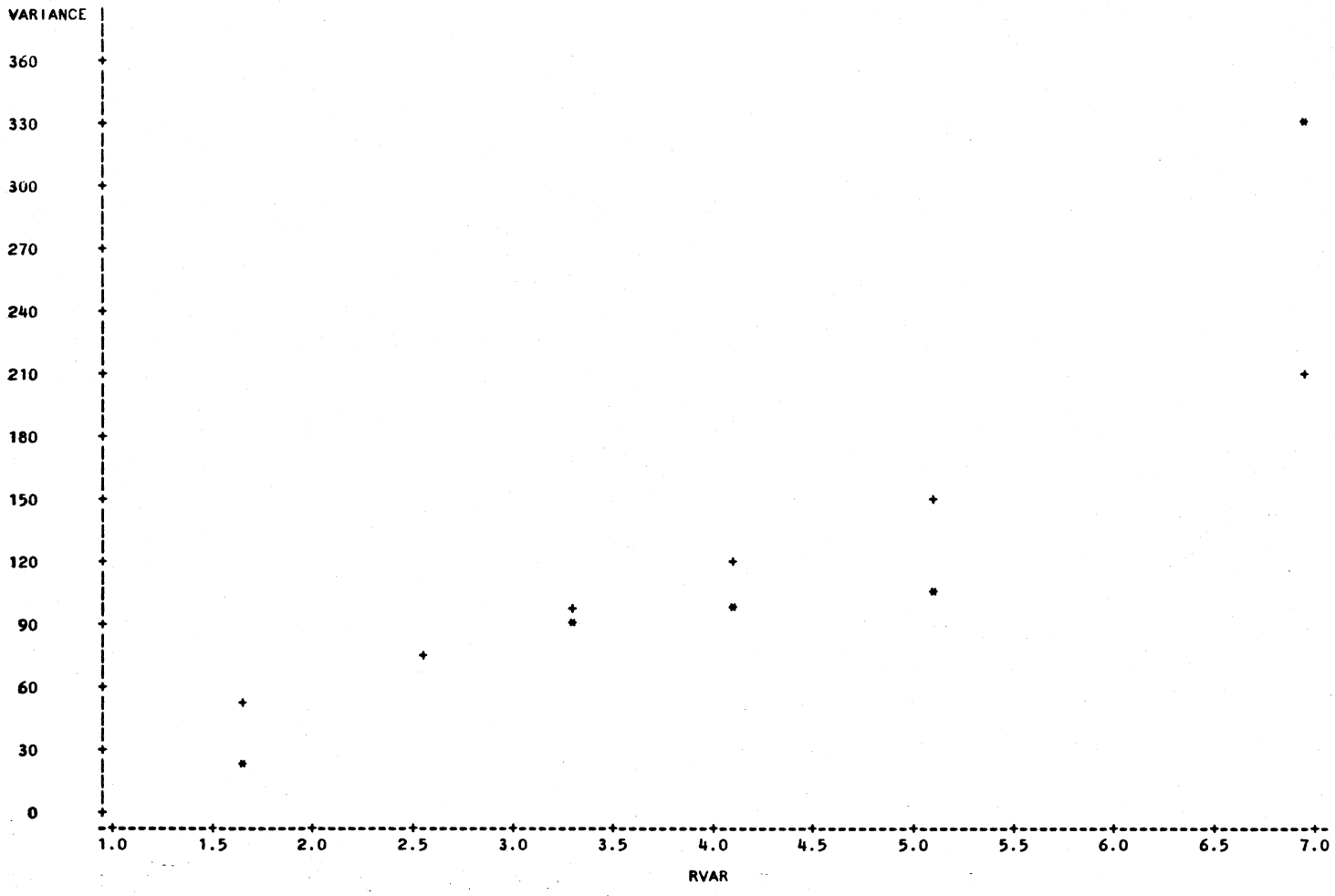
```
*CALCULATE VARIANCES, BY TYPE;
PROC MEANS N MEAN STD VAR;
    VAR BREAKS; BY TYPE;
    OUTPUT OUT=S VAR=VARIANCE;
*RANK THE VARIANCES;
PROC RANK DATA=S OUT=R;
    VAR VARIANCE;
    RANKS RVAR;
*CALCULATE AN AVERAGE VARIANCE;
PROC MEANS DATA=S;
    VAR VARIANCE; OUTPUT OUT=M N=N MEAN=AVGVAR;
*CALCULATE CHI-SQUARED QUANTILES AND REFERENCE POINTS;
DATA;
    SET M;
LOOP:  SET R;
    RVAR = GAMINV((RVAR-.5)/N,4.0);
    EVAR = AVGVAR*RVAR/4;
    OUTPUT; GOTO LOOP;
*PRODUCT CHI-SQUARED PROBABILITY PLOT;
PROC PLOT NOLEGEND;
    TITLE CHI-SQUARED PROBABILITY PLOT;
    PLOT VAR*RVAR='*' EVAR*RVAR='+' / OVERLAY;
```

SAS 1

VARIABLE	N	MEAN	STANDARD DEVIATION	VARIANCE
TYPE=1				
BREAKS	9	44.55555556	18.09772853	327.52777778
TYPE=2				
BREAKS	9	24.00000000	8.66025404	75.00000000
TYPE=3				
BREAKS	9	24.55555556	10.27267140	105.52777778
TYPE=4				
BREAKS	9	28.22222222	9.85872428	97.19444444
TYPE=5				
BREAKS	9	28.77777778	9.43103623	88.94444444
TYPE=6				
BREAKS	9	18.77777778	4.89330609	23.94444444

SAS 2

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	STD ERROR OF MEAN	SUM	VARIANCE	C.V.
VARIANCE	6	119.68981481	105.84220059	23.94444444	327.52777778	43.20989745	718.13888889	11202.571425	88.430



Uniform Probability Plots

If $x_1, x_2, \ldots, x_n$ are a random sample from an exponential distribution, then

$$y_i = \sum_{j=1}^{i} x_j / S \qquad i = 1, 2, \ldots, n-1$$

where

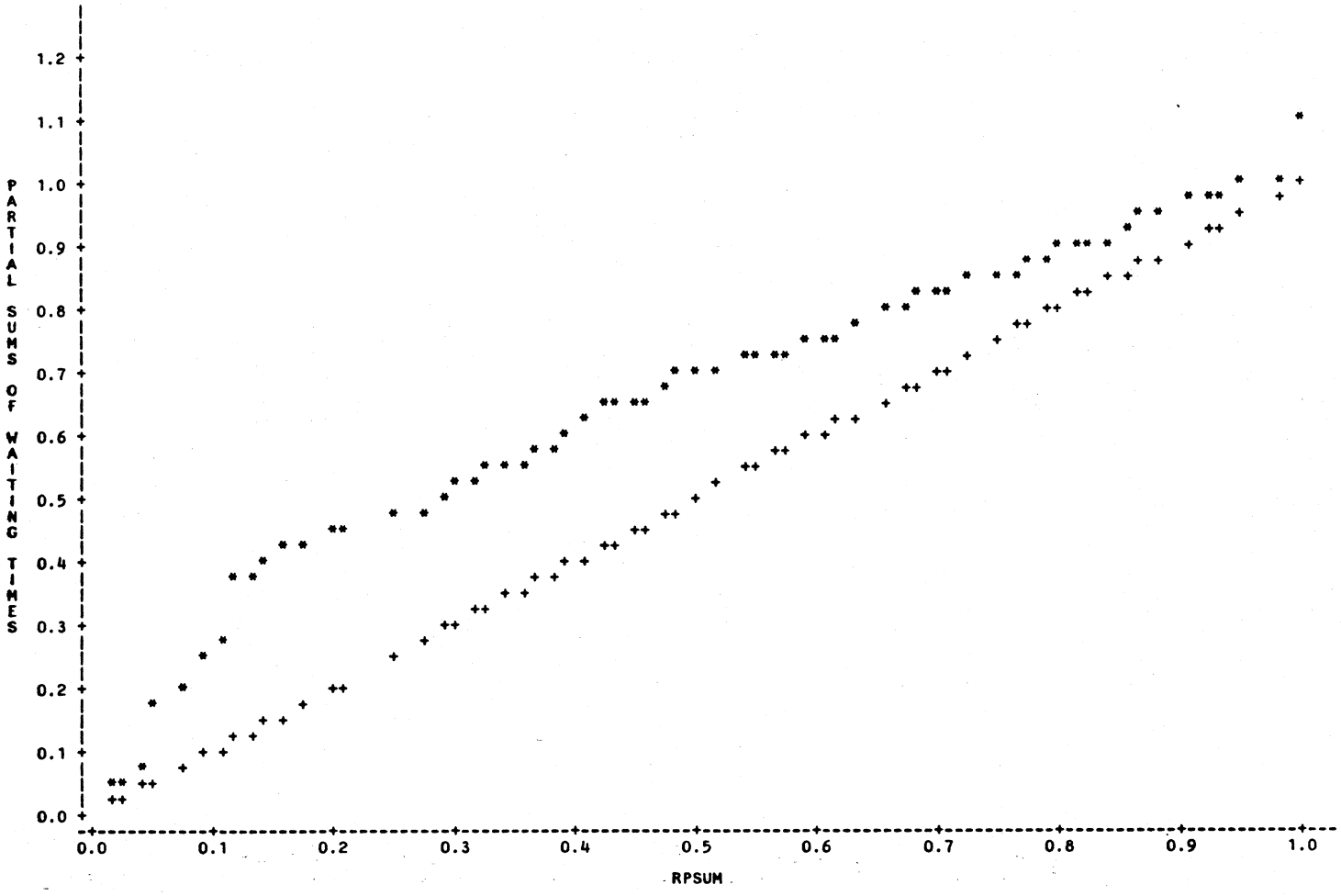
$$S = \sum_{i=1}^{n} X_i$$

have a standard uniform distribution. Therefore, a test of the hypothesis that sample data came from an exponential distribution can be converted into a test of the hypothesis that the corresponding "partial sums" came from a uniform distribution.

Uniform probability plots are easy to produce. Since the cumulative distribution function is a linear function, the inverse is also a linear function and there is no need to rescale the cumulative percentage axis. That is, a cumulative percentage plot is also a uniform probability plot. Figure 5 is a uniform probability plot for the partial sums of the waiting times between air crashes. In this form, the outliers, as departures from an exponential distribution, show up as departures from a uniform distribution.

```
*CALCULATE THE SUM OF ALL WAITING TIMES;
PROC MEANS DATA=WAITTIME NOPRINT;
    VAR WAIT;
    OUTPUT OUT=S SUM=S;
*CALCULATE PARTIAL SUMS OF WAITING TIMES;
DATA;
    SET S;
LOOP; SET WAITTIME END=EOF;
     SUM + WAIT;
     PSUM = SUM/S;
     IF NOT EOF THEN OUTPUT;
     GOTO LOOP;
KEEP PSUM;
LABEL PSUM = PARTIAL SUMS OF WAITING TIMES;
*CALCULATE CUMULATIVE PROBABILITIES;
PROC RANK FRACTION;
    VAR PSUM;
    RANKS RPSUM;
*PRODUCE UNIFORM PROBABILITY PLOT;
PLOT PLOT NOLEGEND;
    TITLE UNIFORM PROBABILITY PLOT;
    PLOT PSUM*RPSUM='*' RPSUM*RPSUM='+' / OVERLAY;
```

UNIFORM PROBABILITY PLOT



NOTE: 22 OBS HIDDEN

```
* CREATE A SAS DATA SET WITH NAME AND AGE OF BRITISH RULER;
DATA RULERS;
   INPUT NAME : & $16. AGE @@;
   LABEL AGE = AGE AT DEATH OF BRITISH RULER;
CARDS;
WILLIAM I  60   WILLIAM II  43   HENRY I  67
STEPHEN  50     HENRY II  56     RICHARD I  42
JOHN     50     HENRY III   65   EDWARD I  68
EDWARD II  43   EDWARD III  65   RICHARD II  34
HENRY IV   47   HENRY V     34   HENRY VI  49
EDWARD IV   41  EDWARD V  13     RICHARD III  35
HENRY VII  53   HENRY VIII  56   EDWARD VI  16
MARY I  43      ELIZABETH  69    JAMES I  59
CHARLES I  48   CROMWELL I  59   CROMWELL II  86
CHARLES II  55  JAMES II  68     WILLIAM III  51
MARY II  33     ANNE  49         GEORGE I  67
GEORGE II  77   GEORGE III  81   GEORGE IV  67
WILLIAM IV  71  VICTORIA  81     EDWARD VII  68
GEORGE V  70    EDWARD VIII  77  GEORGE VI  56
PROC PRINT;
* PRODUCE A HISTOGRAM OF AGES;
PROC CHART;
   VBAR AGE;
* CALCULATE CUMULATIVE PERCENTS;
PROC RANK DATA=RULERS PERCENT OUT=C;
   VAR AGE;
   RANKS CAGE;
* PRODUCT CUMULATIVE PROBABILITY PLOT OF AGE;
PROC PLOT NOLEGEND;
   TITLE CUMULATIVE PROBABILITY PLOT;
   PLOT AGE*CAGE='*';
   FORMAT CAGE 5.1;
* CALCULATE NORMAL SCORES;
PROC RANK DATA=RULERS NORMAL=BLOM OUT=R;
   VAR AGE; RANKS NAGE;
* CALCULATE MEAN, STD, AND NOBS;
PROC MEANS NOPRINT;
   VAR AGE;
   OUTPUT OUT=M MEAN=MEAN STD=STD;
* MERGE THE MEAN WITH THE SCORES, CREATING
  THE REFERENCE POINTS;
DATA;
    SET M;
LOOP: SET R;
    EAGE = MEAN + NAGE*STD;
    OUTPUT; GOTO LOOP;
* PRODUCE NORMAL PROBABILITY PLOT;
PROC PLOT NOLEGEND;
   TITLE NORMAL PROBABILITY PLOT;
    PLOT AGE*NAGE='*' EAGE*NAGE='+' /OVERLAY;
```

```
* CREATE A SAS DATASET OF DATES OF MAJOR AIR CRASHES;
DATA EVENTS;
    INPUT EVENT : DATE7. @@;
CARDS;
1DEC51 1DEC52 1MAR53 1JUN53 1NOV55
1JUN56 1JUN56 1AUG57 1FEB58 1FEB60
1MAR60 1JUL60 1DEC60 1FEB61 1SEP61
1SEP61 1NOV61 1MAR62 1MAR62 1MAR62
1JUN62 1JUN62 1NOV62 1FEB63 1JUN63
1NOV63 1DEC63 1FEB64 1MAR64 1MAY64
1FEB65 1MAY65 1JAN66 1FEB66 1MAR66
1APR66 1SEP66 1DEC66 1MAR67 1APR67
1JUN67 1JUN67 1JUL67 1OCT67 1NOV67
1DEC67 1APR68 1MAY68 1SEP68 1MAR69
1MAR69 1JUN69 1SEP69 1NOV69 1DEC69
1FEB70 1JUL70 1JUL70 1AUG70 1OCT70
1NOV75 1DEC70 1MAY71 1JUL71 1AUG71
1SEP71 1MAR72 1AUG72 1OCT72 1DEC72
1DEC72 1JAN73 1APR73 1JUN73 1JUL73
1JUL73 1JUL73 1AUG73
PROC SORT;
   BY EVENT;
* CALCULATE TIME IN MONTHS (WAIT) BETWEEN CRASHES;
DATA WAITTIME;
   SET;
   RETAIN EVENT1;
   WAIT = MONTH(EVENT) - MONTH(EVENT1)
          + 12*(YEAR(EVENT)- YEAR(EVENT1));
   IF WAIT ¬= . THEN OUTPUT;
   KEEP WAIT EVENT;
   LABEL WAIT = MONTHS BETWEEN MAJOR AIR CRASHES;
   EVENT1 = EVENT;
PROC PRINT;
   VAR EVENT WAIT;
    FORMAT EVENT DATE7.;
* PRODUCE A HISTOGRAM OF WAITS;
PROC CHART;
   VBAR WAIT;
* RANK THE WAITING TIMES;
PROC RANK OUT=R;
   VAR WAIT;
   RANKS RWAIT;
* CALCULATE AN AVERAGE WAITING TIME;
PROC MEANS NOPRINT;
   VAR WAIT;
   OUTPUT OUT=M N=N MEAN=MEAN;
* CALCULATE EXPONENTIAL QUANTILES AND
  REFERENCE POINTS;
DATA;
    SET M;
LOOP: SET R;
```

```
    RWAIT = GAMINV((RWAIT-.5)/N,1.0);
    EWAIT = MEAN*RWAIT;
    OUTPUT; GO TO LOOP;
* PRODUCE EXPONENTIAL PROBABILITY PLOT;
PROC PLOT NOLEGEND;
    TITLE EXPONENTIAL PROBABILITY PLOT;
    PLOT WAIT*RWAIT='*' EWAIT*RWAIT='+' /OVERLAY;
* CALCULATE THE SUM OF ALL WAITING TIMES;
PROC MEANS DATA=WAITTIME NOPRINT;
    VAR WAIT;
    OUTPUT OUT=S SUM=S;
* CALCULATE PARTIAL SUMS OF WAITING TIMES;
DATA;
    SET S;
LOOP: SET WAITTIME END=EOF;
    SUM + WAIT;
    PSUM = SUM/S;
    IF NOT EOF THEN OUTPUT;
    GOTO LOOP;
KEEP PSUM;
LABEL PSUM = PARTIAL SUMS OF WAITING TIMES;
* CALCULATE CUMULATIVE PROBABILITIES;
PROC RANK FRACTION;
    VAR PSUM;
    RANKS RPSUM;
* PRODUCE UNIFORM PROBABILITY PLOT;
PROC PLOT NOLEGEND;
    TITLE UNIFORM PROBABILITY PLOT;
    PLOT PSUM*RPSUM='*' RPSUM*RPSUM='+' /OVERLAY;
```

```
* CREATE A SAS DATASET OF WARP BREAKS;
DATA WARP;
    INPUT TYPE @;
    N = 9;
LOOP: IF N = 0 THEN RETURN;
    INPUT BREAKS @;
    OUTPUT;
    N = N - 1; GOTO LOOP;
DROP N;
CARDS;
1 26 30 54 25 70 52 51 26 67
2 18 21 29 17 12 18 35 30 36
3 36 21 24 18 10 43 28 15 26
4 27 14 29 19 29 31 41 20 44
5 42 26 19 16 39 28 21 39 29
6 20 21 24 17 13 15 15 16 28
* CALCALUTE VARIANCES, BY TYPE;
PROC MEANS N MEAN STD VAR;
   BY TYPE; VAR BREAKS;
   OUTPUT OUT=S VAR=VARIANCE;
* RANK THE VARIANCES;
PROC RANK DATA=S OUT=R;
   VAR VARIANCE;
   RANKS RVAR;
* CALCULATE AN AVERAGE VARIANCE;
PROC MEANS DATA=S;
   VAR VARIANCE;
   OUTPUT OUT=M N=N MEAN=AVGVAR;
* CALCULATE CHI-SQUARED QUANTILES AND
  REFERENCE LINE;
DATA;
    SET M;
LOOP: SET R;
    RVAR = GAMINV((RVAR-.5)/N,4);
    EVAR = AVGVAR*RVAR/4;
    OUTPUT; GOTO LOOP;
* PRODUCE CHI-SQUARED PROBABILITY PLOT;
PROC PLOT NOLEGEND;
    TITLE CHI-SQUARED PROBABILITY PLOT;
    PLOT VAR*RVAR='*' EVAR*RVAR='+' /OVERLAY;
```

Index