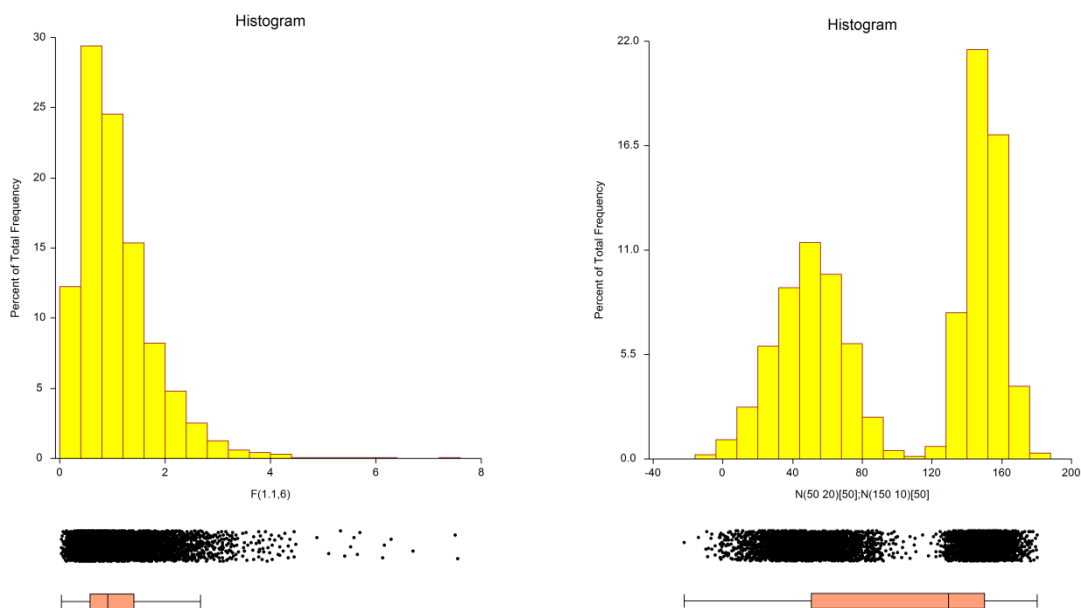


## Chapter 122

# Data Simulation

## Introduction

Because of mathematical intractability, it is often necessary to investigate the properties of a statistical procedure using *simulation* (or *Monte Carlo*) techniques. In power analysis, *simulation* refers to the process of generating several thousand random samples that follow a particular distribution, calculating the test statistic from each sample, and tabulating the distribution of these test statistics so that the significance level and power of the procedure may be investigated. This module creates a histogram of a specified distribution as well as a numerical summary of simulated data. By studying the histogram and the numerical summary, you can determine if the distribution has the characteristics you desire. The distribution formula can then be used in procedures that use simulation, such as the new t-test procedures. Below are examples of two distributions that were generated with this procedure.



## Technical Details

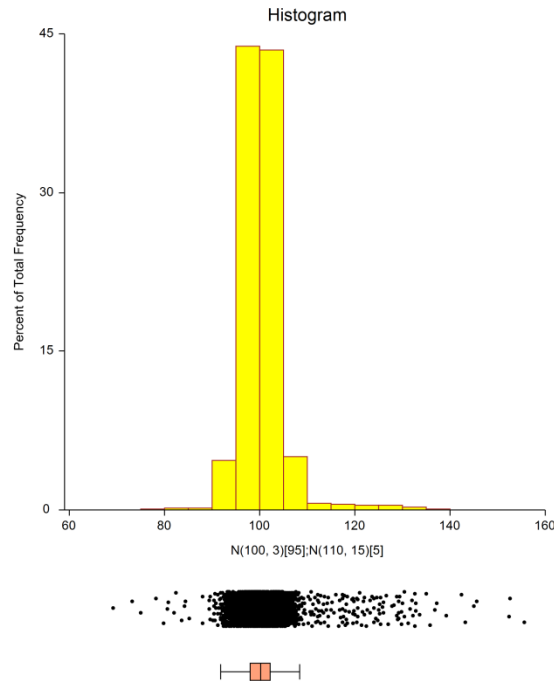
A random variable's probability distribution specifies its probability over its range of values. Examples of common continuous probability distributions are the normal and uniform distributions. Unfortunately, experimental data often do not follow these common distributions, so other distributions have been proposed. One of the easiest ways to create distributions with desired characteristics is to combine simple distributions. For example, outliers may be added to a distribution by mixing it with data from a distribution with a much larger variance. Thus, to simulate normally distributed data with 5% outliers, we could generate 95% of the sample from a normal distribution with mean 100 and standard deviation 4 and then generate 5% of the sample from a normal distribution with mean 100 and standard deviation 16. Using the standard notation

## Data Simulation

for the normal distribution, the composite distribution of the new random variable  $Y$  could be written as

$$Y \sim \delta(0 \leq X < 0.95)N(100,4) + \delta(0.95 \leq X \leq 1.00)N(100,16)$$

where  $X$  is a uniform random variable between 0 and 1,  $\delta(z)$  is 1 or 0 depending on whether  $z$  is true or false, Normal(100,4) is a normally distributed random variable with mean 100 and standard deviation 4, and Normal(100,16) is a normally distributed random variable with mean 100 and standard deviation 16. The resulting distribution is shown below. Notice how the tails extend in both directions.



The procedure for generating a random variable,  $Y$ , with the mixture distribution described above is

1. Generate a uniform random number,  $X$ .
2. If  $X$  is less than 0.95,  $Y$  is created by generating a random number from the  $Normal(100,4)$  distribution.
3. If  $X$  is greater than or equal to 0.95,  $Y$  is created by generating a random number from the  $Normal(100,16)$  distribution.

Note that only one uniform random number and one normal random number are generated for any particular random realization from the mixture distribution.

In general, the formula for a mixture random variable,  $Y$ , which is to be generated from two or more random variables defined by their distribution function  $F_i(Z_i)$  is given by

$$Y \sim \sum_{i=1}^k \delta(a_i \leq X < a_{i+1}) F_i(Z_i), \quad a_1 = 0 < a_2 < \dots < a_{k+1} = 1$$

Note that the  $a_i$ 's are chosen so that weighting requirements are met. Also note that only one uniform random number and one other random number actually need to be generated for a particular value. The  $F_i(Z_i)$ 's may be any of the distributions which are listed below.

Since the test statistics which will be simulated are used to test hypotheses about one or more means, it will be convenient to parameterize the distributions in terms of their means.

## Beta Distribution

The beta distribution is given by the density function

$$f(x) = \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} \left(\frac{x-C}{D-C}\right)^{A-1} \left(1 - \frac{x-C}{D-C}\right)^{B-1}, \quad A, B > 0, C \leq x \leq D$$

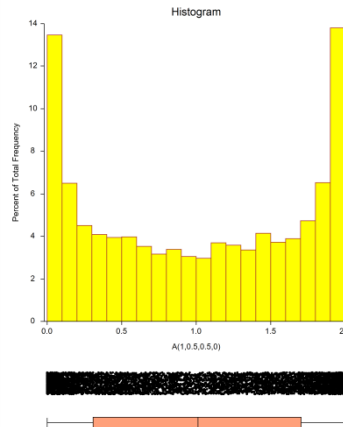
where  $A$  and  $B$  are shape parameters,  $C$  is the minimum, and  $D$  is the maximum. In statistical theory,  $C$  and  $D$  are usually zero and one, respectively, but the more general formulation used here is more convenient for simulation work. A beta random variable may be specified using either of two parameterizations: **Beta**( $A, B, C, D$ ) or **BetaMS**(*Mean, SD, C, D*). If BetaMS(.) is used, the program solves for the values of  $A$  and  $B$  from the *Mean* and *SD* using the following relationships

$$\text{Mean} = (D - C) \left[ \frac{A}{A + B} \right] + C$$

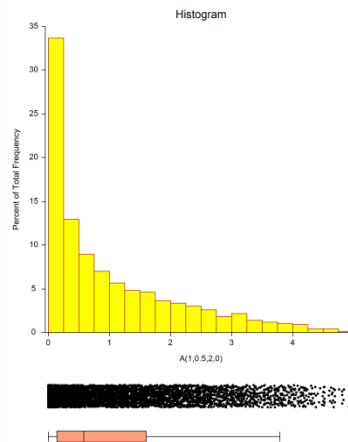
$$\text{SD} = \sqrt{\frac{(D - C)^2}{(A + B)^2} \left[ \frac{AB}{A + B + 1} \right]}$$

The beta density can take a number of shapes depending on the values of  $A$  and  $B$ :

1. When  $A < 1$  and  $B < 1$  the density is U-shaped.

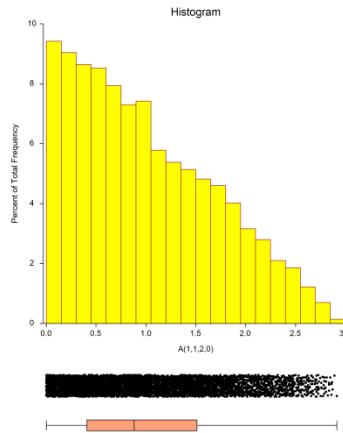


2. When  $0 < A < 1 \leq B$  the density is J-shaped.

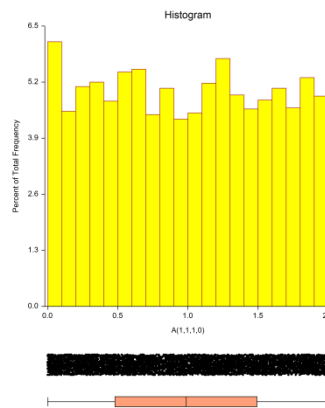


Data Simulation

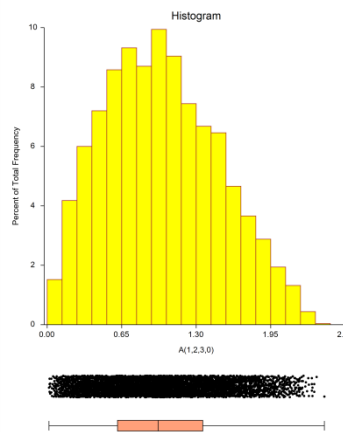
- 3. When  $A=1$  and  $B>1$  the density is bounded and decreases monotonically to 0.



- 4. When  $A=1$  and  $B=1$  the density is the uniform density.



- 5. When  $A>1$  and  $B>1$  the density is unimodal.



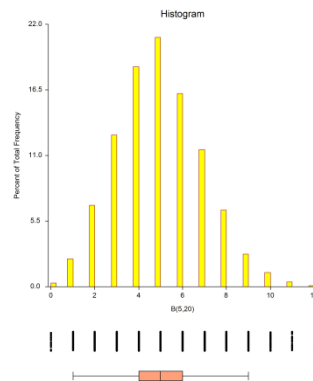
Beta random variates are generated using Cheng's rejection algorithm as given on page 438 of Devroye (1986).

## Binomial Distribution

The binomial distribution is given by the function

$$\Pr(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

A binomial random variable may be specified using either of two parameterizations: **Binomial(P, n)** or **BinomialMS(Mean, n)**. If the **BinomialMS(...)** version is used, the value of  $P$  is calculated from the *Mean* using  $P = \text{Mean}/n$ . Because of this, you must have  $0 < \text{Mean} < n$ .



Binomial random variates are generated using the inverse CDF method. That is, a uniform random variate is generated, and then the CDF of the binomial distribution is scanned to determine which value of  $X$  is associated with that probability.

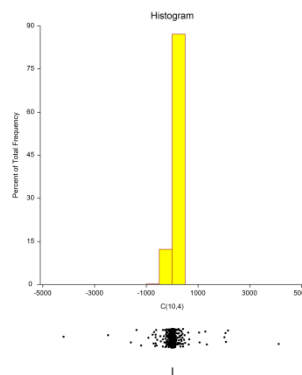
## Cauchy Distribution

The Cauchy distribution is given by the density function

$$f(x) = \left[ S\pi \left( 1 + \left\{ \frac{X - M}{S} \right\}^2 \right) \right]^{-1}, \quad S > 0$$

Although the Cauchy distribution does not possess a mean and standard deviation,  $M$  and  $S$  are treated as such. Cauchy random numbers are generated using the algorithm given in Johnson, Kotz, and Balakrishnan (1994), page 327.

In this program module, the Cauchy is specified as **Cauchy(M, S)**, where  $M$  is a location parameter (median), and  $S$  is a scale parameter.



## Constant Distribution

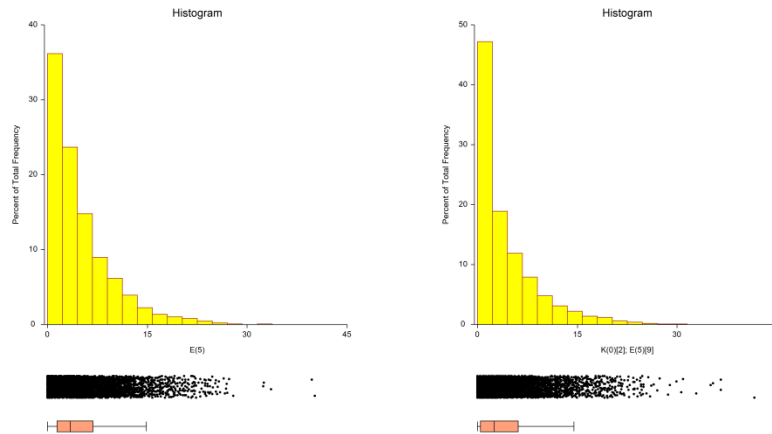
The *constant* distribution occurs when a random variable can only take a single value,  $X$ . The constant distribution is specified as *Constant*( $X$ ), where  $X$  is the value.

### Data with Many Zero Values

Sometimes data follow a specific distribution in which there is a large proportion of zeros. This can happen when data are counts or monetary amounts. Suppose you want to generate exponentially distributed data with an extra number of zeros. You could use the following simulation model:

`Constant(0)[2]; Exponential(5)[9]`

The exponential distribution alone was used to generate the histogram below on the left. The histogram below on the right was simulated by adding extra zeros to the exponential data.

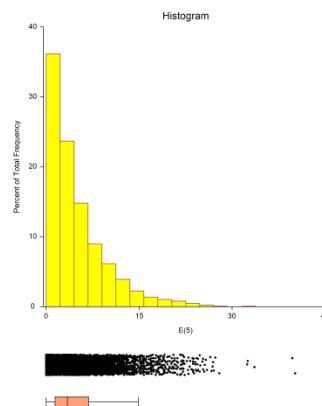


## Exponential Distribution

The exponential distribution is given by the density function

$$f(x) = \frac{1}{M} e^{-\frac{x}{M}}, \quad x > 0$$

In this program module, the exponential is specified as *Exponential*( $M$ ), where  $M$  is the mean.



Random variates from the exponential distribution are generated using the expression  $-M \ln(U)$ , where  $U$  is a uniform random variate.

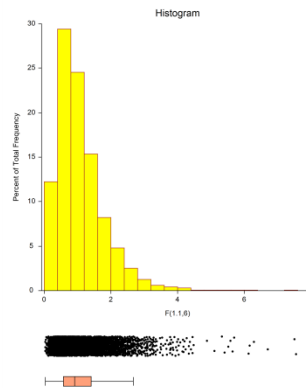
## F Distribution

Snedecor's  $F$  distribution is the distribution of the ratio of two independent chi-square variates. The degrees of freedom of the numerator chi-square variate is  $A$ , while that of the denominator chi-square is  $D$ . The  $F$  distribution is specified as  $F(M, A)$ , where  $M$  is the mean and  $A$  is the degrees of freedom of the numerator chi-square. The value of  $M$  is related to the denominator chi-square degrees of freedom using the relationship  $M=D/(D-2)$ .

$F$  variates are generated by first generating a symmetric beta variate,  $B(A/2, D/2)$ , and transforming it into an  $F$  variate using the relationship

$$F_{A,D} = \frac{BD}{A - BA}$$

Below is a histogram for data generated from an  $F$  distribution with a mean of 1.1 and  $A = 6$ .



## Gamma Distribution

The two parameter gamma distribution is given by the density function

$$f(x) = \frac{(x)^{A-1}}{B^A \Gamma(A)} e^{-\frac{x}{B}}, \quad x > 0, A > 0, B > 0$$

where  $A$  is a shape parameter and  $B$  is a scale parameter.

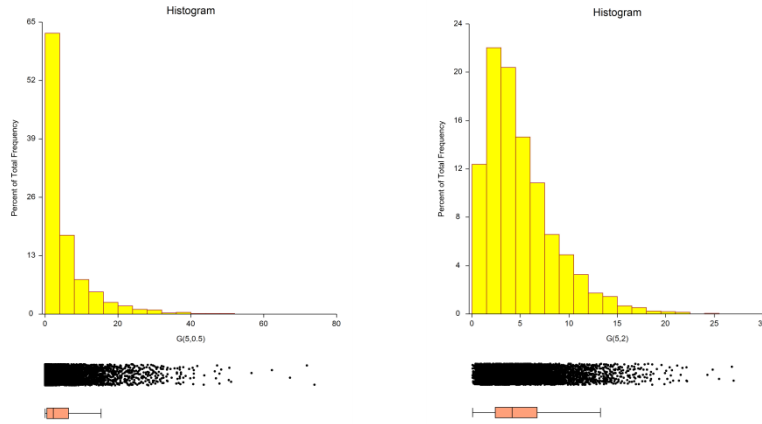
A gamma random variable may be specified using either of two parameterizations:  $Gamma(A, B)$  or  $GammaMS(Mean, SD)$ . If  $GammaMS(Mean, SD)$  is used, the values of  $A$  and  $B$  are solved for using

$$Mean = AB$$

$$SD = B\sqrt{A}$$

Gamma variates are generated using the exponential distribution when  $A = 1$ ; Best's XG algorithm given in Devroye (1986), page 410, when  $A > 1$ ; and Vaduva's algorithm given in Devroye (1986), page 415, when  $A < 1$ .

Data Simulation



### Gumbel Distribution

The two parameter Gumbel (extreme value) distribution is given by the density function

$$f(x; A, B) = \frac{1}{B} \exp\left\{-\frac{x - A}{B} - \exp\left\{-\frac{x - A}{B}\right\}\right\}$$

where  $A$  is a location parameter and  $B$  is a scale parameter.

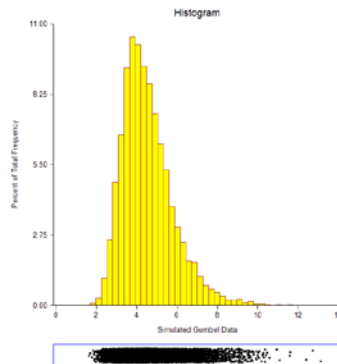
A Gumbel random variable may be specified using either of two parameterizations:  $Gumbel(A, B)$  or  $GumbelMS(Mean, SD)$ . If  $GumbelMS(Mean, SD)$  is used, the values of  $A$  and  $B$  are solved for using

$$Mean = A + 0.57722B$$

$$SD = \sqrt{1.64493B}$$

Gumbel variates may be generated using the following transformation of uniform variates

$$g_i = A - B \ln\left\{\ln\left(\frac{1}{U_i}\right)\right\}$$





## Laplace Distribution

The two parameter Laplace (or double-exponential) distribution is given by the density function

$$f(x; A, B) = \frac{1}{2B} \exp\left(-\frac{|x - A|}{B}\right)$$

where  $A$  is a location parameter and  $B$  is a scale parameter.

A Laplace random variable may be specified using either of two parameterizations:  $Laplace(A, B)$  or  $LaplaceMS(Mean, SD)$ . If  $LaplaceMS(Mean, SD)$  is used, the values of  $A$  and  $B$  are solved for using

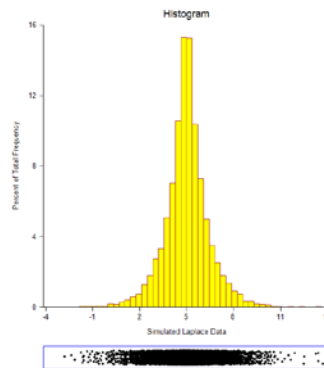
$$Mean = A$$

$$SD = B\sqrt{2}$$

Laplace variates are generated using the following transformation of uniform  $\left(-\frac{1}{2} < U < \frac{1}{2}\right)$  variates

$$x_i = A - B \operatorname{sgn}(U_i) \ln(1 - 2|U_i|)$$

Here is a histogram of Laplace data



## Logistic Distribution

The two parameter logistic distribution is given by the density function

$$f(x; A, B) = \frac{\exp\left(-\frac{x - A}{B}\right)}{B \left[1 + \exp\left(-\frac{x - A}{B}\right)\right]^2}$$

where  $A$  is a location parameter and  $B$  is a scale parameter.

A logistic random variable may be specified using either of two parameterizations:  $Logistic(A, B)$  or  $LogisticMS(Mean, SD)$ . If  $LogisticMS(Mean, SD)$  is used, the values of  $A$  and  $B$  are solved for using

$$Mean = A$$

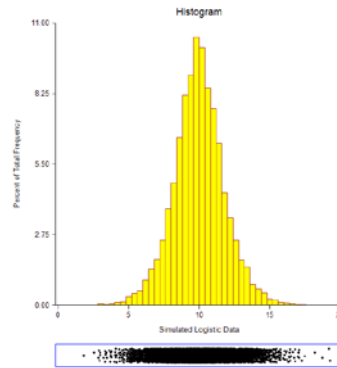
$$SD = \frac{B\pi}{\sqrt{3}}$$

Logistic variates are generated using the following transformation of uniform variates

$$x_i = A + B \ln\left(\frac{U_i}{1 - U_i}\right)$$

Here is a histogram of logistic data

## Data Simulation



## Lognormal Distribution

The two parameter lognormal distribution is given by the density function

$$f(x; A, B) = \frac{1}{xB\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln(x) - A}{B}\right)^2\right\}$$

where  $A$  is a location parameter and  $B$  is a scale parameter.

A lognormal random variable may be specified using either of two parameterizations:  $Lognormal(A, B)$  or  $LognormalMS(Mean, SD)$ . If  $LognormalMS(Mean, SD)$  is used, the values of  $A$  and  $B$  are solved for using

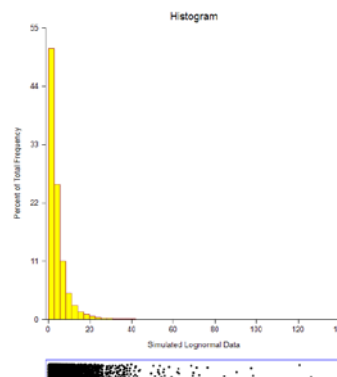
$$Mean = \exp\left\{A + \frac{B^2}{2}\right\}$$

$$SD = \exp\{2A + B^2\}[\exp\{B^2\} - 1]$$

Lognormal variates are generated the following transformation of normal variates

$$x_i = \exp(A + B z_i)$$

Here is a histogram of lognormal data



## Multinomial Distribution

The *multinomial* distribution occurs when a random variable has only a few discrete values such as 1, 2, 3, 4, and 5. The multinomial distribution is specified as ***Multinomial***( $P_1, P_2, \dots, P_k$ ), where  $P_i$  is the probability of that the integer  $i$  occurs. Note that the values start at one, not zero.

For example, suppose you want to simulate a distribution which has 50% 3's and 1's, 2's, 4's, and 5's all with equal percentages. You would enter *Multinomial*(1 1 4 1 1).

As a second example, suppose you wanted to have an equal percentage of 1's, 3's, and 7's, and none of the other percentages. You would enter *Multinomial*(1 0 1 0 0 1).

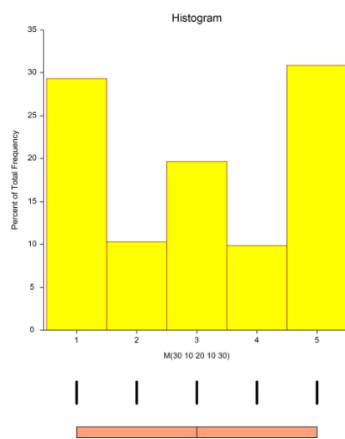
### Likert-Scale Data

Likert-scale data are common in surveys and questionnaires. To generate data from a five-point Likert-scale distribution, you could use the following simulation model:

*Multinomial*(6 1 2 1 5)

Note that the weights are relative—they do not have to sum to one. The program will make the appropriate weighting adjustments so that they do sum to one.

The above expression generated the following histogram.



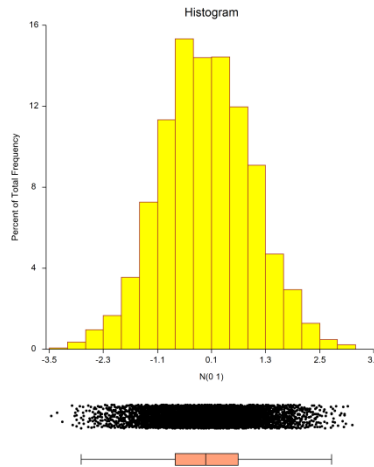
## Normal Distribution

The normal distribution is given by the density function

$$f(x) = \phi\left(\frac{x - \mu}{\sigma}\right), \quad -\infty \leq x \leq \infty$$

where  $\phi(z)$  is the usual standard normal density. The normal distribution is specified as *Normal*( $M, S$ ), where  $M$  is the mean and  $S$  is the standard deviation.

The normal distribution is generated using the Marsaglia and Bray algorithm as given in Devroye (1986), page 390.



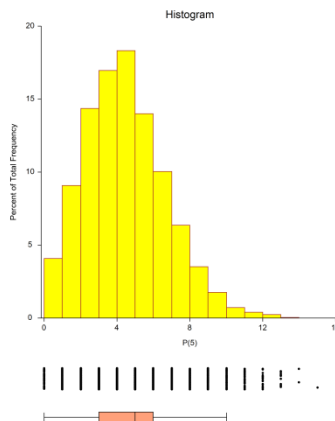
## Poisson Distribution

The Poisson distribution is given by the function

$$\Pr(X = x) = \frac{e^{-M} M^x}{x!}, \quad x = 0, 1, 2, \dots, M > 0$$

In this program module, the Poisson is specified as *Poisson*( $M$ ), where  $M$  is the mean.

Poisson random variates are generated using the inverse CDF method. That is, a uniform random variate is generated and then the CDF of the Poisson distribution is scanned to determine which value of  $X$  is associated with that probability.



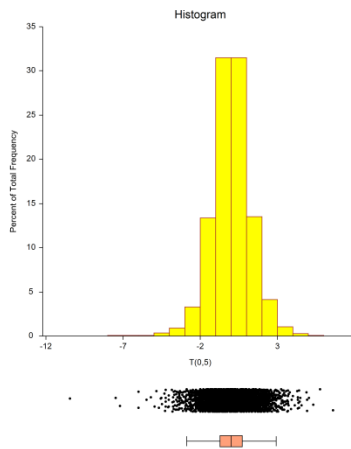
## Student's T Distribution

Student's  $T$  distribution is the distribution of the ratio of a unit normal variate and the square root of an independent chi-square variate. The degrees of freedom of the chi-square variate are the degrees of freedom of the  $T$  distribution. The  $T$  is specified as  $T(M, A)$ , where  $M$  is the mean and  $A$  is the degrees of freedom. The central  $T$  distribution generated in this program has a mean of zero, so, to obtain a mean of  $M$ ,  $M$  is added to every data value.

$T$  variates are generated by first generating a symmetric beta variate,  $B(A/2, A/2)$ , with mean equal to 0.5. This beta variate is then transformed into a  $T$  variate using the relationship

$$T = \sqrt{A} \frac{X - 0.5}{\sqrt{X(1-X)}}$$

Here is a histogram for data generated from a  $T$  distribution with mean 0 and 5 degrees of freedom.



## Tukey's Lambda Distribution

Hoaglin (1985) presents a discussion of a distribution developed by John Tukey for allowing the detailed specification of skewness and kurtosis in a simulation study. This distribution is extended in the work of Karian and Dudewicz (2000). Tukey's idea was to reshape the normal distribution using functions that change the skewness and/or kurtosis. This is accomplished by multiplying a normal random variable by a skewness function and/or a kurtosis function. The general form of the transformation is

$$Y = G_g(z)H_h(z)z, \quad X = A + BY$$

where  $z$  has the standard normal density. The skewness function Tukey proposed is

$$G_g(z) = \frac{e^{gz} - 1}{gz}$$

The range of  $g$  is typically -1 to 1. The value of  $G_0(z) \equiv 1$ . The kurtosis function Tukey proposed is

$$H_h(z) = e^{hz^2/2}$$

The range of  $h$  is also -1 to 1.

Hence, if both  $g$  and  $h$  are set to zero, the variable  $X$  follows the normal distribution with mean  $A$  and standard deviation  $B$ . As  $g$  is increased toward 1, the distribution is increasingly skewed to the right. As  $g$  is decreased towards -1, the distribution is increasingly skewed to the left. As  $h$  is increased toward 1, the data are stretched out

## Data Simulation

so that more extreme values are probable. As  $h$  is decreased toward  $-1$ , the data are concentrated around the center—resulting in a beta-type distribution.

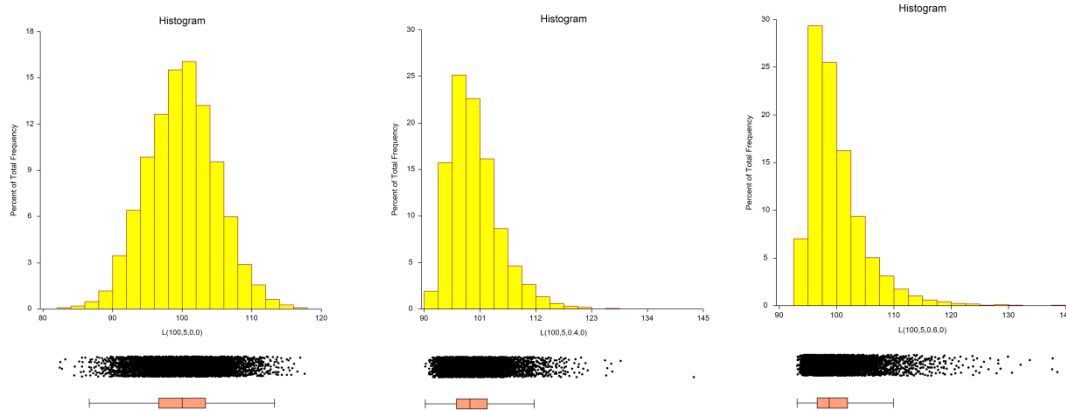
The mean of this distribution is given by

$$M = A + B \left( \frac{e^{g^2/2(1-h)} - 1}{g\sqrt{1-h}} \right), \quad 0 \leq h < 1$$

which may be easily solved for  $A$ . The value  $B$  is a scale factor (when  $g=h=0$ ,  $B$  is the standard deviation).

Tukey's lambda is specified in the program as **TukeyGH**( $M$ ,  $SD$ ,  $g$ ,  $h$ ) where  $M$  is the mean,  $SD$  is the standard deviation ( $B = SD/\text{Sqrt}(\text{Var}(Y))$ ); see Hoaglin (1985) for  $\text{Var}(Y)$  formula),  $g$  is the amount of skewness, and  $h$  is the kurtosis. The formula for  $\text{Var}(Y)$  requires that  $0 \leq h < 0.5$ .

Random variates are generated from this distribution by generating a random normal variate, applying the skewness and kurtosis modifications, and scaling to get the desired mean and standard deviation. Here are some examples as  $g$  is varied from 0 to 0.4 to 0.6. Notice how the amount of skewness is gradually increased. Similar results are achieved when  $h$  is varied from 0 to 0.5.



## Uniform Distribution

The uniform distribution is given by the density function

$$f(x) = \frac{1}{B - A}, \quad A \leq x \leq B$$

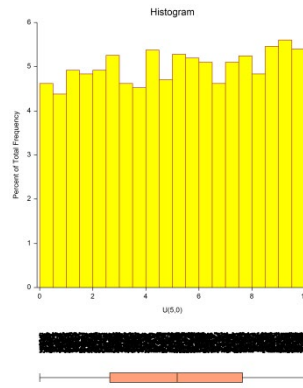
The uniform is specified as either **Uniform**( $A$ ,  $B$ ) or **UniformMS**(Mean,  $SD$ ). If **UniformMS**(Mean,  $SD$ ) is used, the program calculates  $A$  and  $B$  using the relationships

$$\text{Mean} = \frac{A + B}{2}$$

$$\text{SD} = \sqrt{\frac{B - A}{12}}$$

Following is a histogram of thousand's of uniform random variates.

## Data Simulation



Uniform random numbers are generated using Makoto Matsumoto's Mersenne Twister uniform random number generator which has a cycle length greater than  $1.0E+6000$  (that's a one followed by 6000 zeros).

## Weibull Distribution

The Weibull distribution is indexed by a shape parameter,  $B$ , and a scale parameter,  $C$ . The Weibull density function is written as

$$f(x|B, C) = \frac{B}{C} \left(\frac{x}{C}\right)^{(B-1)} e^{-\left(\frac{x}{C}\right)^B}, \quad B > 0, C > 0, x > 0.$$

A Weibull random variable may be specified using either of two parameterizations: **Weibull**( $A, B$ ) or **WeibullMS**(*Mean, SD*). If **WeibullMS** (*Mean, SD*) is used, the values of  $A$  and  $B$  are found for using

$$\text{Mean} = C \Gamma\left(1 + \frac{1}{B}\right)$$

$$\text{SD} = C \sqrt{\Gamma\left(1 + \frac{2}{B}\right) - \Gamma^2\left(1 + \frac{1}{B}\right)}$$

### Shape Parameter - B

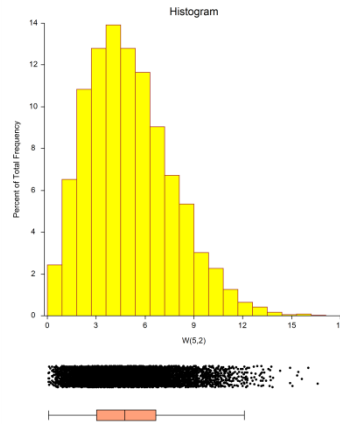
The shape parameter controls the overall shape of the density function. Typically, this value ranges between 0.5 and 8.0. One of the reasons for the popularity of the Weibull distribution is that it includes other useful distributions as special cases or close approximations. For example, if

- B = 1     The Weibull distribution is identical to the exponential distribution.
- B = 2     The Weibull distribution is identical to the Rayleigh distribution.
- B = 2.5   The Weibull distribution approximates the lognormal distribution.
- B = 3.6   The Weibull distribution approximates the normal distribution.

### Scale Parameter - C

The scale parameter only changes the scale of the density function along the  $x$  axis. Some authors use  $1/C$  instead of  $C$  as the scale parameter. Although this is arbitrary, we prefer dividing by the scale parameter since that is how one usually scales a set of numbers.

## Data Simulation




---

## Combining Distributions

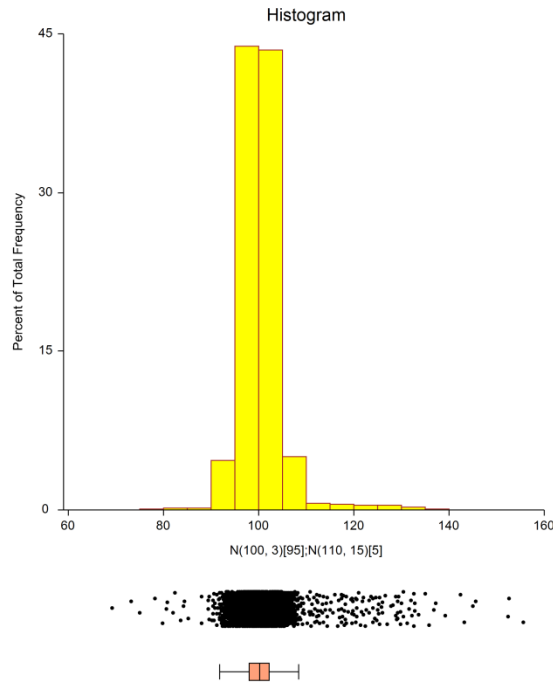
A random variable's probability distribution specifies its probability over its range of values. Examples of common continuous probability distributions are the normal and uniform distributions. Unfortunately, experimental data often do not follow these common distributions, so other distributions have been proposed. One of the easiest ways to create distributions with desired characteristics is to combine simple distributions. For example, outliers may be added to a distribution by mixing it with data from a distribution with a much larger variance. Thus, to simulate normally distributed data with 5% outliers, we could generate 95% of the sample from a normal distribution with mean 100 and standard deviation 4 and then generate 5% of the sample from a normal distribution with mean 100 and standard deviation 16. Using the standard notation for the normal distribution, the composite distribution of the new random variable  $Y$  could be written as

$$Y \sim \delta(0 \leq X < 0.95)Normal(100,4) + \delta(0.95 \leq X \leq 1.00)Normal(100,16)$$

where  $X$  is a uniform random variable between 0 and 1,  $\delta(z)$  is 1 or 0 depending on whether  $z$  is true or false,  $Normal(100,4)$  is a normally distributed random variable with mean 100 and standard deviation 4, and  $Normal(100,16)$  is a normally distributed random variable with mean 100 and standard deviation 16. The resulting distribution is shown below. Notice how the tails extend in both directions.



Data Simulation



The procedure for generating a random variable,  $Y$ , with the mixture distribution described above is

1. Generate a uniform random number,  $X$ .
2. If  $X$  is less than 0.95,  $Y$  is created by generating a random number from the  $N(100,4)$  distribution.
3. If  $X$  is greater than or equal to 0.95,  $Y$  is created by generating a random number from the  $N(100,16)$  distribution.

Note that only one uniform random number and one normal random number are generated for any particular random realization from the mixture distribution.

In general, the formula for a mixture random variable,  $Y$ , which is to be generated from two or more random variables defined by their distribution function  $F_i(Z_i)$  is given by

$$Y \sim \sum_{i=1}^k \delta(a_i \leq X < a_{i+1}) F_i(Z_i), \quad a_1 = 0 < a_2 < \dots < a_{K+1} = 1$$

Note that the  $a_i$ 's are chosen so that weighting requirements are met. Also note that only one uniform random number and one other random number actually need to be generated for a particular value. The  $F_i(Z_i)$ 's may be any of the distributions which are listed below.

Since the test statistics which will be simulated are used to test hypotheses about one or more means, it will be convenient to parameterize the distributions in terms of their means.

## Creating New Distributions using Expressions

The set of probability distributions discussed above provides a basic set of useful distributions. However, you may want to mimic reality more closely by combining these basic distributions. For example, paired data is often analyzed by forming the differences of the two original variables. If the original data are normally distributed, then the differences are also normally distributed. Suppose, however, that the original data are exponential. The difference of two exponentials is not a common distribution.

## Data Simulation

## Expression Syntax

The basic syntax is

$$C1 D1 operator1 C2 D2 operator2 C3 D3 operator3 \dots$$

where C1, C2, C3, etc. are coefficients (numbers), D1, D2, D3, etc. are probability distributions, and *operator* is one of the four symbols: +, -, \*, /. Parentheses are only permitted in the specification of distributions.

Examples of valid expressions include

Normal(4, 5) – Normal(4, 5)

2 Exponential(3) – 4 Exponential(4) + 2 Exponential(5)

Normal(4, 2)/ Exponential(4) - Constant(5)

**Notes about the Coefficients: C1, C2, C3**

The coefficients may be positive or negative decimal numbers such as 2.3, 5, or -3.2. If no coefficient is specified, the coefficient is assumed to be one.

**Notes about the Distributions: D1, D2, D3**

The distributions may be any of the distributions listed above such as normal, exponential, or beta. The expressions are evaluated by generating random values from each of the distributions specified and then combining them according to the operators.

**Notes about the operators: +, -, \*, /**

All multiplications and divisions are performed first, followed by any additions and subtractions.

Note that if only addition and subtraction are used in the expression, the mean of the resulting distribution is found by applying the same operations to the individual distribution means. If the expression involves multiplication or division, the mean of the resulting distribution is usually difficult to calculate directly.

**Creating New Distributions using Mixtures**

Mixture distributions are formed by sampling a fixed percentage of the data from each of several distributions. For example, you may model outliers by obtaining 95% of your data from a normal distribution with a standard deviation of 5 and 5% of your data from a distribution with a standard deviation of 50.

**Mixture Syntax**

The basic syntax of a mixture is

$$D1[W1]; D2[W2]; \dots; Dk[Wk]$$

where the D's represent distributions and the W's represent weights. Note that the weights must be positive numbers. Also note that semi-colons are used to separate the components of the mixture.

Examples of valid mixture distributions include

Normal(4, 5)[19]; Normal(4, 50)[1]      95% of the distribution is Normal(4, 5), and the other 5% is Normal(4, 50).

Weibull(4, 3)[7]; K(0)[3]      70% of the distribution is Weibull(4, 3), and the other 30% is made up of zeros.

Normal(4, 2)-Normal(4,3)[2]; Exponential(4)\*Exponential(2)[8]      20% of the distribution is Normal(4, 2) - Normal(4,3), and the other 80% is Exponential(4)\*Exponential(2).

**Notes about the Distributions**

The distributions D1, D2, D3, etc. may be any valid distributional expression.

## Data Simulation

### Notes about the Weights

The weights  $w_1$ ,  $w_2$ ,  $w_3$ , etc. need not sum to one (or to one hundred). The program uses these weights to calculate new, internal weights that do sum to one. For example, if you enter weights of 1, 2, and 1, the internal weights will be 0.25, 0.50, and 0.25.

When a weight is not specified, it is assumed to have the value of '1.' Thus

```
Normal(4, 5)[19]; Normal(4,50)[1]
```

is equivalent to

```
Normal(4, 5)[19]; Normal(4,50)
```

---

## Special Functions

A set of special functions is available to modify the generator number after all other operations are completed. These special functions are applied in the order they are given next.

### Square Root (Absolute Value)

This function is activated by placing a  $\wedge$  in the expression. When active, the square root of the absolute value of the number is used.

### Logarithm (Absolute Value)

This function is activated by placing a  $\sim$  in the expression. When active, the logarithm (base e) of the absolute value of the number is used.

### Exponential

This function is activated by placing an  $\&$  in the expression. When active, the number is exponentiated to the base e. If the current number  $x$  is greater than 70,  $\exp(70)$  is used rather than  $\exp(x)$ .

### Absolute Value

This function is activated by placing a  $|$  in the expression. When active, the absolute value of the number is used.

### Integer

This function is activated by placing a  $\#$  in the expression. When active, the number is rounded to the nearest integer.

---

## Procedure Options

This section describes the options that are specific to this procedure. These are located on the Data tab. To find out more about using the other tabs such as Labels or Plot Setup, turn to the chapter entitled Procedure Templates.

---

### Data Tab

The Data tab contains the parameters used to specify a probability distribution.

---

### Data Simulation

#### Probability Distribution to be Simulated

Enter the components of the probability distribution to be simulated. One or more components may be entered from among the continuous and discrete distributions listed below the data-entry box.

The  $W$  parameter gives the relative weight of that component. For example, if you entered

Poisson(5)[1];Constant(0)[2], about 33% of the random numbers would follow the Poisson(5) distribution, and 67% would be 0. When only one component is used, the value of  $W$  may be omitted. For example, to generate data from the normal distribution with mean of five and standard deviation of one, you would enter Normal(5, 1), not Normal(5, 1)[1].

---

### Data Simulation – For Summary and Histogram

#### Number of Simulated Values

This is the number of values generated from the probability distribution for display in the histogram. We recommend a value of about 5000.

Note that the histogram, box plot, and dot plot row limits must be set higher than this amount or the corresponding plot will not be displayed. These limits are modified by selecting Edit, Options, and Limits from the spreadsheet menu.

---

### Storage Tab

The Storage tab is used to specify the columns that will hold the stored values.

---

### Storage of Simulated Values to Spreadsheet

#### Store Values in Column

This is the variable in which the simulated values will be stored. Any data already in this column will be replaced.

#### Numbers of Values Stored

This is the number of generated values that are stored in the current database.

## Reports Tab

The following options control the format of the reports.

---

### Select Report

#### Numerical Summary

This option controls the display of this report.

---

### Report Options

#### Precision

This allows you to specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also note that all calculations are performed in double precision regardless of which option you select here. This is for reporting purposes only.

---

### Report Options – Percentile Options

#### Percentile Type

This option specifies which of five different methods is used to calculate the percentiles.

RECOMMENDED: **Ave  $X_{p(n+1)}$**  since it gives the common value of the median.

In the explanations below,  $p$  refers to the fractional value of the percentile (for example, for the 75th percentile  $p = .75$ ),  $Z_p$  refers to the value of the percentile,  $X[i]$  refers to the  $i$ th data value after the values have been sorted,  $n$  refers to the total sample size, and  $g$  refers to the fractional part of a number (for example, if  $np = 23.42$ , then  $g = .42$ ). The options are

- **Ave  $X_{p(n+1)}$**

This is the most commonly used option. The 100pth percentile is computed as

$$Z_p = (1-g)X[k_1] + gX[k_2]$$

where  $k_1$  equals the integer part of  $p(n+1)$ ,  $k_2=k_1+1$ ,  $g$  is the fractional part of  $p(n+1)$ , and  $X[k]$  is the  $k$ th observation when the data are sorted from lowest to highest.

- **Ave  $X_{p(n)}$**

The 100pth percentile is computed as

$$Z_p = (1-g)X[k_1] + gX[k_2]$$

where  $k_1$  equals the integer part of  $np$ ,  $k_2=k_1+1$ ,  $g$  is the fractional part of  $np$ , and  $X[k]$  is the  $k$ th observation when the data are sorted from lowest to highest.

- **Closest to  $np$**

The 100pth percentile is computed as

$$Z_p = X[k_1]$$

where  $k_1$  equals the integer that is closest to  $np$  and  $X[k]$  is the  $k$ th observation when the data are sorted from lowest to highest.

## Data Simulation

- **EDF**

The 100pth percentile is computed as

$$Z_p = X[k_1]$$

where  $k_1$  equals the integer part of  $np$  if  $np$  is exactly an integer or the integer part of  $np+1$  if  $np$  is not exactly an integer.  $X[k]$  is the  $k$ th observation when the data are sorted from lowest to highest. Note that EDF stands for empirical distribution function.

- **EDF w/Ave**

The 100pth percentile is computed as

$$Z_p = (X[k_1] + X[k_2])/2$$

where  $k_1$  and  $k_2$  are defined as follows: If  $np$  is an integer,  $k_1=k_2=np$ . If  $np$  is not exactly an integer,  $k_1$  equals the integer part of  $np$  and  $k_2 = k_1+1$ .  $X[k]$  is the  $k$ th observation when the data are sorted from lowest to highest. Note that EDF stands for empirical distribution function.

### Smallest Percentile

This option lets you assign a different value to the smallest percentile value shown on the percentile report. The default value is 1.0. You can select any value between 0 and 100, including decimal numbers.

### Largest Percentile

This option lets you assign a different value to the largest percentile value shown on the percentile report. The default value is 1.0. You can select any value between 0 and 100, including decimal numbers.

---

## Report Options – Report Decimal Places

### Means – Values

Specify the number of decimal places used when displaying this item.

GENERAL: Display the entire number without special formatting.

---

## Plots Tab

This panel sets the options used to define the appearance of the histogram.

---

### Select Plot

#### Histogram

This option controls the display of the histogram. Click the plot format button to change the plot settings.

## Example 1 – Generating Normal Data

In this example, 5000 values will be generated from the standard normal (mean zero, variance one) distribution. These values will be displayed in a histogram and summarized numerically.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Data Simulation window.

### 1 Open the Data Simulation window.

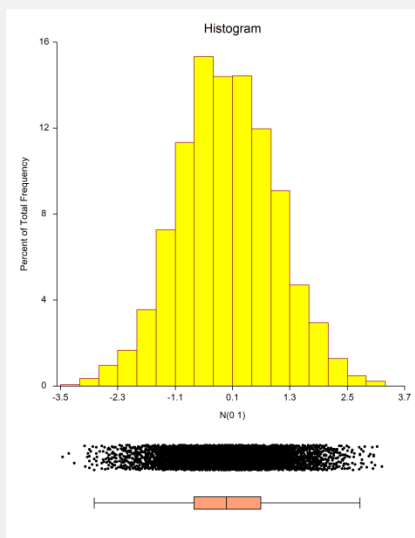
- Using the Tools menu or the Procedure Navigator, find and select the **Data Simulation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 2 Specify the variables.

- On the Data Simulator window, select the **Data** tab.
- Enter **Normal(0 1)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

### 3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.



#### Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	8.627304E-03	Minimum	-3.453244
Standard Deviation	0.9910903	1st Percentile	-2.289698
Skewness	-6.010016E-02	5th Percentile	-1.65369
Kurtosis	2.847444	10th Percentile	-1.298245
Coefficient of Variation	114.8783	25th Percentile	-0.6598607
Count	5000	Median	2.659537E-02
		75th Percentile	0.6922649
		90th Percentile	1.276735
		95th Percentile	1.620931
		99th Percentile	2.274097
		Maximum	3.645533

This report shows the histogram and a numerical summary of the 5000 simulated normal values. It is interesting to check how well the simulation did. Theoretically, the mean should be zero, the standard deviation one, the skewness zero, and the kurtosis three. Of course, your results will vary from these because these are based on generated random numbers.

## Example 2 – Generating Data from a Contaminated Normal

In this example, we will generate data from a contaminated normal. This will be accomplished by generating 95% of the data from a  $N(100,3)$  distribution and 5% from a  $N(110,15)$  distribution.

You may follow along here by making the appropriate entries or load the completed template **Example 2** by clicking on Open Example Template from the File menu of the Data Simulation window.

### 1 Open the Data Simulation window.

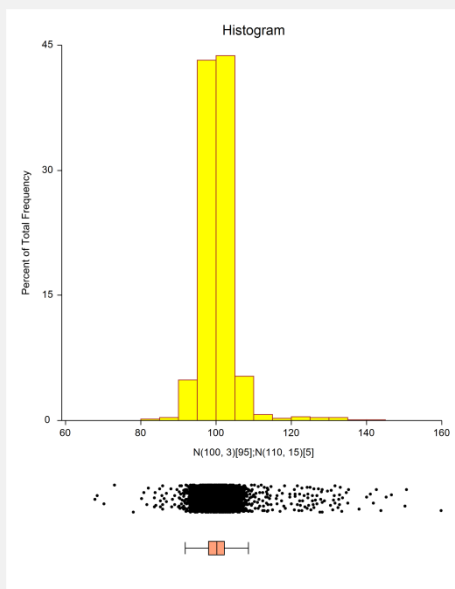
- Using the Tools menu or the Procedure Navigator, find and select the **Data Simulation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **Normal(100 3)[95];Normal(110 15)[5]** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

### 3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.



#### Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	100.3549	Minimum	71.11177
Standard Deviation	4.716588	1st Percentile	92.40327
Skewness	2.67692	5th Percentile	94.82752
Kurtosis	23.11323	10th Percentile	96.01687
Coefficient of Variation	4.699908E-02	25th Percentile	97.88011
Count	5000	Median	100.0328
		75th Percentile	102.18
		90th Percentile	104.266
		95th Percentile	105.9543
		99th Percentile	120.4605
		Maximum	145.2731

This report shows the data from the contaminated normal. The mean is close to 100, but the standard deviation, skewness, and kurtosis have non-normal values. Note that there are now some very large outliers.



## Example 3 – Likert-Scale Data

In this example, we will generate data following a discrete distribution on a Likert scale. The distribution of the Likert scale will be 30% 1's, 10% 2's, 20% 3's, 10% 4's, and 30% 5's.

You may follow along here by making the appropriate entries or load the completed template **Example 3** by clicking on Open Example Template from the File menu of the Data Simulation window.

### 1 Open the Data Simulation window.

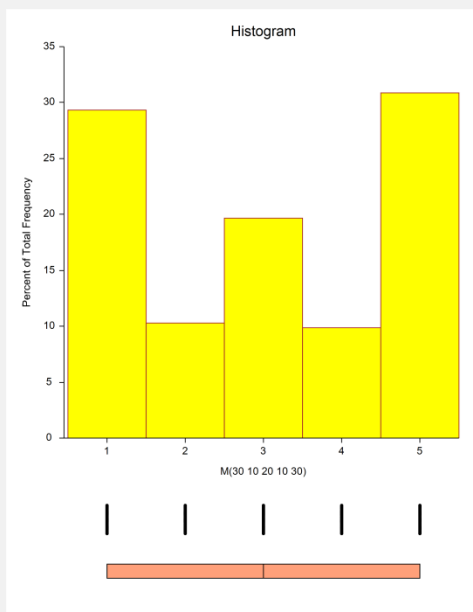
- Using the Tools menu or the Procedure Navigator, find and select the **Data Simulation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **Multinomial(30 10 20 10 30)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

### 3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.



#### Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	2.9858	Minimum	1
Standard Deviation	1.622506	1st Percentile	1
Skewness	1.169226E-02	5th Percentile	1
Kurtosis	1.436589	10th Percentile	1
Coefficient of Variation	0.5434074	25th Percentile	1
Count	5000	Median	3
		75th Percentile	5
		90th Percentile	5
		95th Percentile	5
		99th Percentile	5
		Maximum	5

This report shows the data from a Likert scale.

## Example 4 – Bimodal Data

In this example, we will generate data that have a bimodal distribution. We will accomplish this by combining data from two normal distributions, one with a mean of 10 and the other with a mean of 30. The standard deviation will be set at 4.

You may follow along here by making the appropriate entries or load the completed template **Example 4** by clicking on Open Example Template from the File menu of the Data Simulation window.

### 1 Open the Data Simulation window.

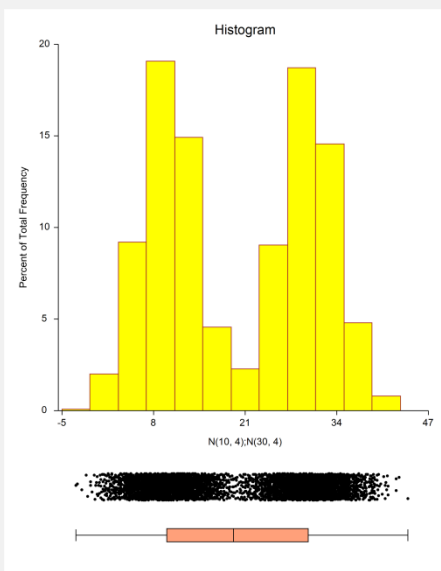
- Using the Tools menu or the Procedure Navigator, find and select the **Data Simulation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 2 Specify the variables.

- On the Data Simulator window, select the **Data** tab.
- Enter **Normal(10 4);Normal(30 4)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

### 3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.



#### Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	19.82392	Minimum	-7.067288
Standard Deviation	10.81724	1st Percentile	1.819361
Skewness	3.503937E-02	5th Percentile	4.949873
Kurtosis	1.500108	10th Percentile	6.473907
Coefficient of Variation	0.545666	25th Percentile	9.863808
Count	5000	Median	17.73235
		75th Percentile	29.94195
		90th Percentile	33.4243
		95th Percentile	35.11861
		99th Percentile	38.04005
		Maximum	43.71141

This report shows the results for the simulated bimodal data.

## Example 5 – Gamma Data with Extra Zeros

In this example, we will generate data that have a gamma distribution, except that we will force there to be about 30% zeros. The gamma distribution will have a shape parameter of 5 and a scale parameter of 10.

You may follow along here by making the appropriate entries or load the completed template **Example 5** by clicking on Open Example Template from the File menu of the Data Simulation window.

### 1 Open the Data Simulation window.

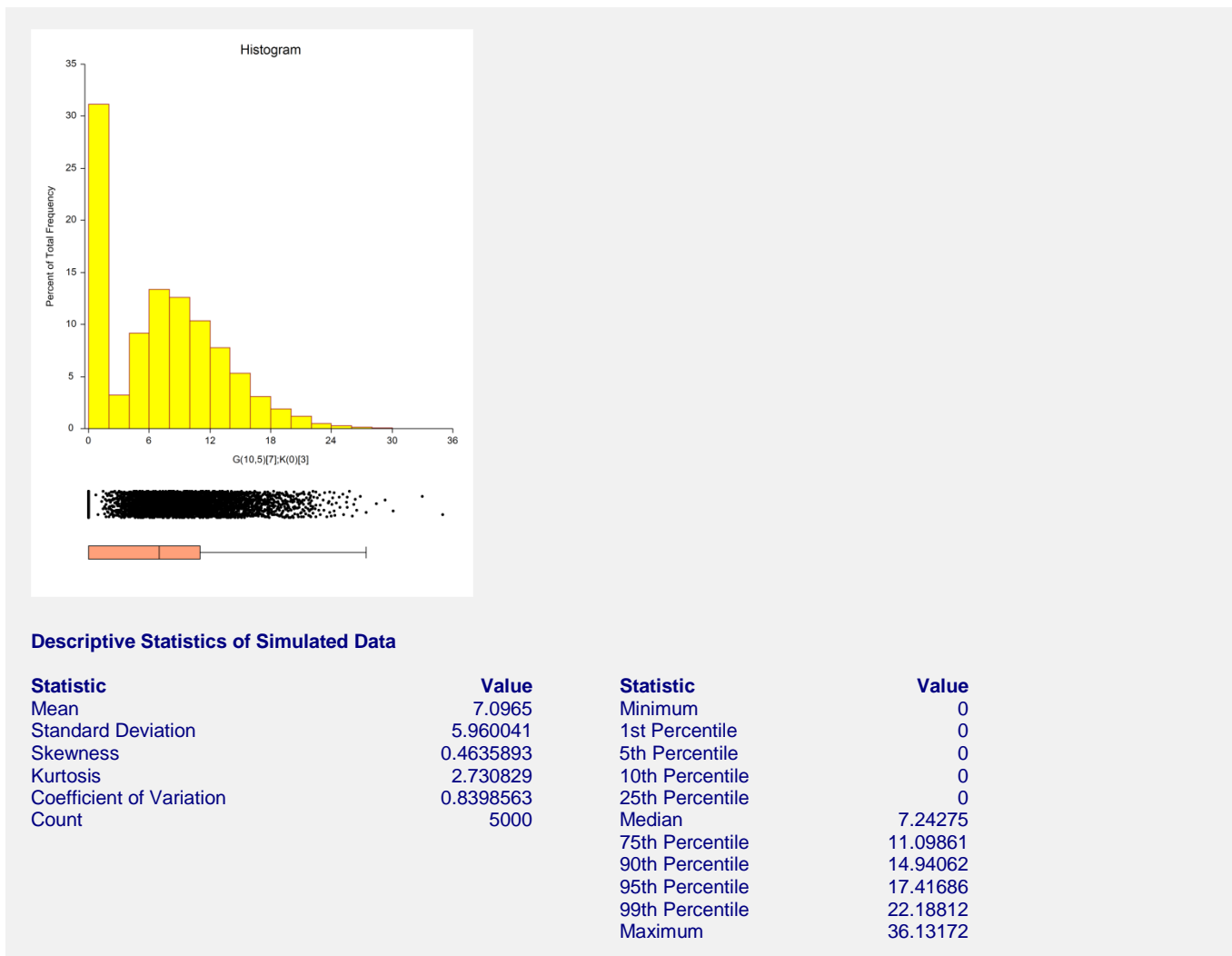
- Using the Tools menu or the Procedure Navigator, find and select the **Data Simulation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **Gamma(10 5)[7];Constant(0)[3]** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

### 3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.



This report shows the results for the simulated bimodal data.

## Example 6 – Mixture of Two Poisson Distributions

In this example, we will generate data that have a mixture of two Poisson distributions. 60% of the data will be from a Poisson distribution with a mean of 10 and 40% from a Poisson distribution with a mean of 20.

You may follow along here by making the appropriate entries or load the completed template **Example 6** by clicking on Open Example Template from the File menu of the Data Simulation window.

### 1 Open the Data Simulation window.

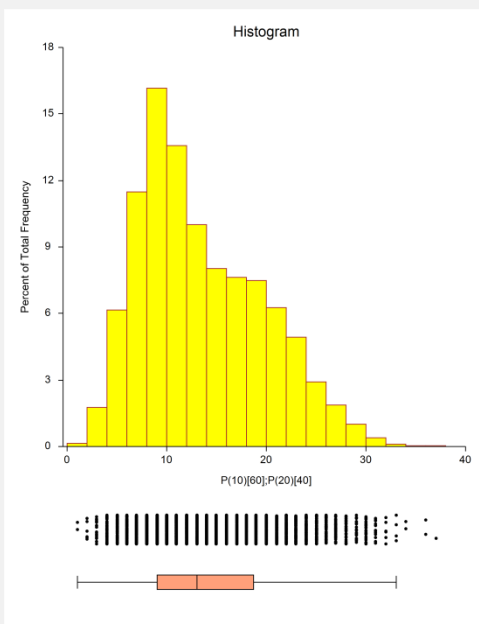
- Using the Tools menu or the Procedure Navigator, find and select the **Data Simulation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 2 Specify the variables.

- On the Data Simulator window, select the **Data** tab.
- Enter **Poisson(10)[60];Poisson(20)[40]** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

### 3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.



#### Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	13.8662	Minimum	0
Standard Deviation	6.088589	1st Percentile	4
Skewness	0.5726398	5th Percentile	6
Kurtosis	2.580317	10th Percentile	7
Coefficient of Variation	0.4390957	25th Percentile	9
Count	5000	Median	13
		75th Percentile	18
		90th Percentile	23
		95th Percentile	25
		99th Percentile	29
		Maximum	36

This report shows the results for the simulated mixture-Poisson data.

## Example 7 – Difference of Two Identically Distributed Exponentials

In this example, we will demonstrate that the difference of two identically distributed exponential random variables follows a symmetric distribution. This is particularly interesting because the exponential distribution is skewed. In fact, the difference between any two identically distributed random variables follows a symmetric distribution.

You may follow along here by making the appropriate entries or load the completed template **Example 7** by clicking on Open Example Template from the File menu of the Data Simulation window.

### 1 Open the Data Simulation window.

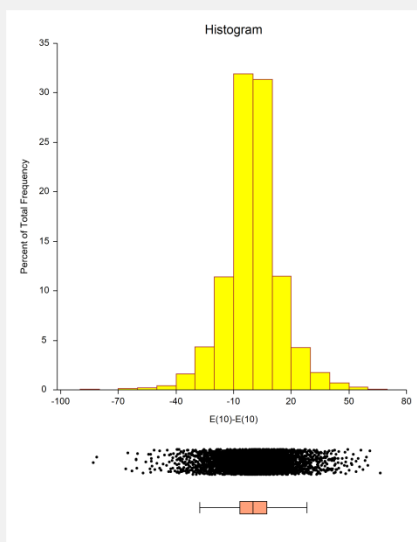
- Using the Tools menu or the Procedure Navigator, find and select the **Data Simulation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 2 Specify the variables.

- On the Data Simulator window, select the **Data tab**.
- Enter **Exponential(10) - Exponential(10)** in the **Probability Distribution to be Simulated** box.
- Leave all other options at their default values.

### 3 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.



#### Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	-0.3431585	Minimum	-99.37727
Standard Deviation	14.20288	1st Percentile	-41.02162
Skewness	-7.391038E-02	5th Percentile	-23.19279
Kurtosis	6.115685	10th Percentile	-16.70725
Coefficient of Variation	-41.38868	25th Percentile	-7.131404
Count	5000	Median	-0.1380393
		75th Percentile	6.575472
		90th Percentile	15.78555
		95th Percentile	22.12444
		99th Percentile	38.82129
		Maximum	72.36995

This report shows demonstrates that the distribution of the difference is symmetric.