drawnorm — Draw sample from multivariate normal distribution

Syntax

```
drawnorm newvarlist [ , options ]
```

Options

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>clear</td>
</tr>
<tr>
<td>double</td>
</tr>
<tr>
<td>n(#)</td>
</tr>
<tr>
<td>sd(vector)</td>
</tr>
<tr>
<td>corr(matrix</td>
</tr>
<tr>
<td>vector)</td>
</tr>
<tr>
<td>cov(matrix</td>
</tr>
<tr>
<td>vector)</td>
</tr>
<tr>
<td>cstorage(full)</td>
</tr>
<tr>
<td>cstorage(lower)</td>
</tr>
<tr>
<td>cstorage(upper)</td>
</tr>
<tr>
<td>forcepsd</td>
</tr>
<tr>
<td>means(vector)</td>
</tr>
<tr>
<td>seed(#)</td>
</tr>
</tbody>
</table>

Menu

Data > Create or change data > Other variable-creation commands > Draw sample from normal distribution

Description

drawnorm draws a sample from a multivariate normal distribution with desired means and covariance matrix. The default is orthogonal data with mean 0 and variance 1. The covariance matrix may be singular. The values generated are a function of the current random-number seed or the number specified with set seed(); see [R] set seed.

Options

clear specifies that the dataset in memory be replaced, even though the current dataset has not been saved on disk.
double specifies that the new variables be stored as Stata doubles, meaning 8-byte reals. If double is not specified, variables are stored as floats, meaning 4-byte reals. See [D] data types.

\texttt{n(\#)} specifies the number of observations to be generated. The default is the current number of observations. If \texttt{n(\#)} is not specified or is the same as the current number of observations, \texttt{drawnorm} adds the new variables to the existing dataset; otherwise, \texttt{drawnorm} replaces the data in memory.

\texttt{sds(vector)} specifies the standard deviations of the generated variables. \texttt{sds()} may not be specified with \texttt{cov()}.

\texttt{corr(matrix | vector)} specifies the correlation matrix. If neither \texttt{corr()} nor \texttt{cov()} is specified, the default is orthogonal data.

\texttt{cov(matrix | vector)} specifies the covariance matrix. If neither \texttt{cov()} nor \texttt{corr()} is specified, the default is orthogonal data.

\texttt{cstorage(full | lower | upper)} specifies the storage mode for the correlation or covariance structure in \texttt{corr()} or \texttt{cov()}. The following storage modes are supported:

- \texttt{full} specifies that the correlation or covariance structure is stored (recorded) as a symmetric $k\times k$ matrix.
- \texttt{lower} specifies that the correlation or covariance structure is recorded as a lower triangular matrix. With $k$ variables, the matrix should have $k(k+1)/2$ elements in the following order:

\[
C_{11} \ C_{21} \ C_{22} \ C_{31} \ C_{32} \ C_{33} \ldots \ C_{k1} \ C_{k2} \ldots \ C_{kk}
\]

- \texttt{upper} specifies that the correlation or covariance structure is recorded as an upper triangular matrix. With $k$ variables, the matrix should have $k(k+1)/2$ elements in the following order:

\[
C_{11} \ C_{12} \ C_{13} \ldots C_{1k} \ C_{22} \ C_{23} \ldots C_{2k} \ldots C_{(k-1)k} \ C_{kk}
\]

Specifying \texttt{cstorage(full)} is optional if the matrix is square. \texttt{cstorage(lower)} or \texttt{cstorage(upper)} is required for the vectorized storage methods. See \texttt{Example 2: Storage modes for correlation and covariance matrices}.

\texttt{forcepsd} modifies the matrix $C$ to be positive semidefinite (psd), and so be a proper covariance matrix. If $C$ is not positive semidefinite, it will have negative eigenvalues. By setting negative eigenvalues to 0 and reconstructing, we obtain the least-squares positive-semidefinite approximation to $C$. This approximation is a singular covariance matrix.

\texttt{means(vector)} specifies the means of the generated variables. The default is \texttt{means(0)}.

\texttt{seed(\#)} specifies the initial value of the random-number seed used by the \texttt{runiform()} function. The default is the current random-number seed. Specifying \texttt{seed(\#)} is the same as typing \texttt{set seed \#} before issuing the \texttt{drawnorm} command.
Example 1

Suppose that we want to draw a sample of 1,000 observations from a normal distribution $N(M, V)$, where $M$ is the mean matrix and $V$ is the covariance matrix:

```
. matrix M = 5, -6, 0.5
. matrix V = (9, 5, 2 \ 5, 4, 1 \ 2, 1, 1)
. matrix list M
  M[1,3]
    c1 c2 c3
   r1  5 -6 .5
. matrix list V
  symmetric V[3,3]
    c1 c2 c3
   r1 9
   r2 5 4
   r3 2 1 1
. drawnorm x y z, n(1000) cov(V) means(M)
(obs 1000)
. summarize
  Variable |     Obs  Mean Std. Dev.     Min     Max
-------------|--------|--------|--------|--------|--------|
    x | 1000 5.001715 3.00608 -4.572042 13.66046
    y | 1000 -5.980279 2.004755 -12.08166 -.0963039
    z | 1000 .5271135 1.011095 -2.636946 4.102734
. correlate, cov
(obs=1000)
|     x     y     z
-------------|--------|--------|--------|
x | 9.03652
y | 5.04462 4.01904
z | 2.10142 1.08773 1.02231
```

Technical note

The values generated by `drawnorm` are a function of the current random-number seed. To reproduce the same dataset each time `drawnorm` is run with the same setup, specify the same seed number in the `seed()` option.

Example 2: Storage modes for correlation and covariance matrices

The three storage modes for specifying the correlation or covariance matrix in `corr2data` and `drawnorm` can be illustrated with a correlation structure, $C$, of 4 variables. In full storage mode, this structure can be entered as a $4 \times 4$ Stata matrix:

```
. matrix C = ( 1.0000, 0.3232, 0.1112, 0.0066  \ ///
            0.3232, 1.0000, 0.6608, -0.1572  \ ///
            0.1112, 0.6608, 1.0000, -0.1480  \ ///
            0.0066, -0.1572, -0.1480, 1.0000 )
```
Elements within a row are separated by commas, and rows are separated by a backslash, \.
We use the input continuation operator /// for convenient multiline input; see [P] comments. In this
storage mode, we probably want to set the row and column names to the variable names:

. matrix rownames C = price trunk headroom rep78
. matrix colnames C = price trunk headroom rep78

This correlation structure can be entered more conveniently in one of the two vectorized storage
modes. In these modes, we enter the lower triangle or the upper triangle of C in rowwise order; these
two storage modes differ only in the order in which the \( k(k+1)/2 \) matrix elements are recorded.
The lower storage mode for C comprises a vector with \( 4(4+1)/2 = 10 \) elements, that is, a \( 1 \times 10 \)
or \( 10 \times 1 \) Stata matrix, with one row or column,

. matrix C = ( 1.0000, ///
           0.3232, 1.0000, ///
           0.1112, 0.6608, 1.0000, ///
           0.0066, -0.1572, -0.1480, 1.0000)

or more compactly as

. matrix C = ( 1, 0.3232, 1, 0.1112, 0.6608, 1, 0.0066, -0.1572, -0.1480, 1 )

C may also be entered in upper storage mode as a vector with \( 4(4+1)/2 = 10 \) elements, that is,
a \( 1 \times 10 \) or \( 10 \times 1 \) Stata matrix,

. matrix C = ( 1.0000, 0.3232, 0.1112, 0.0066, ///
           1.0000, 0.6608, -0.1572, ///
           1.0000, -0.1480, ///
           1.0000 )

or more compactly as

. matrix C = ( 1, 0.3232, 0.1112, 0.0066, 1, 0.6608, -0.1572, 1, -0.1480, 1 )

Methods and formulas

Results are asymptotic. The more observations generated, the closer the correlation matrix of the
dataset is to the desired correlation structure.

Let \( V = A' A \) be the desired covariance matrix and \( M \) be the desired mean matrix. We first
generate \( X \), such that \( X \sim N(0, I) \). Let \( Y = A' X + M \), then \( Y \sim N(M, V) \).

References


Also see

[D] corr2data — Create dataset with specified correlation structure
[R] set seed — Specify initial value of random-number seed