Multivariate Normal Distribution

Statistics 407, ISU

Statistical Thinking

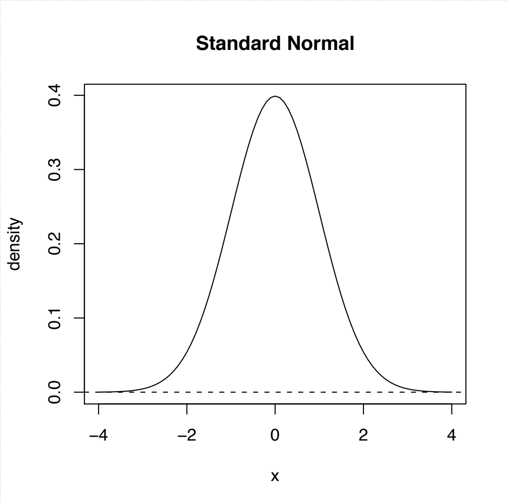
- Fundamental to statistical thinking is the idea that there exists a population which is really the quantity that we are interested in describing. The sample is some realization of the population.
- The population is often quantified numerically using a mathematical description.

Univariate Normal Distribution

- The most often used example is the univariate normal distribution which is described by the ``bell curve'' density function:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\}$$

where μ is the mean of the population and σ is the standard deviation of the population.



Properties

- The probability that a value is observed that is within 1 standard deviation of the mean is approximately 0.68, and the probability that a value is observed within 2 standard deviations of the mean is approximately 0.95.
- A standard normal distribution is defined to have mean equal to 0, and standard deviation equal to 1.

Multivariate Normal Density Function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})}{2}\}$$

$\boldsymbol{\mu}$ controls the location, center; Σ controls the shape.

Multivariate Normal Density Function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})}{2}\}$$

$\boldsymbol{\mu}$ controls the location, center; Σ controls the shape.

Population mean vector

Multivariate Normal Density Function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})}{2}\}$$

$\boldsymbol{\mu}$ controls the location, center; Σ controls the shape.

Population mean vector

Population variance-covariance matrix

Multivariate Normal Density Function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{\boxed{(\mathbf{X}-\mu)}'\Sigma^{-1}(\mathbf{X}-\mu)}{2}\}$$

μ controls the location, center; Σ controls the shape.

Population mean vector

Population variance-covariance matrix

Multivariate Normal Density Function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{\boxed{(\mathbf{X}-\boldsymbol{\mu})}'\Sigma^{-1}\boxed{(\mathbf{X}-\boldsymbol{\mu})}}{2}\}$$

$\boldsymbol{\mu}$ controls the location, center; Σ controls the shape.

Population mean vector

Population variance-covariance matrix

Multivariate Normal Density Function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{\boxed{(\mathbf{X}-\mu)}'\Sigma^{-1}\boxed{(\mathbf{X}-\mu)}}{2}\}$$

μ controls the location, center; Σ controls the shape.

Population mean vector

Population variance-covariance matrix

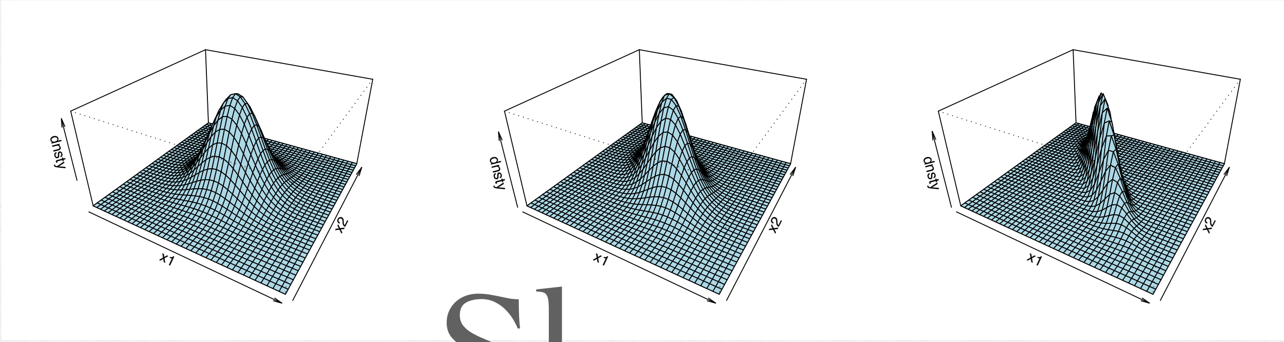
Multivariate Normal Density Function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{(\mathbf{X}-\mu)'\Sigma^{-1}(\mathbf{X}-\mu)}{2}\}$$

μ controls the location, center; Σ controls the shape.

Population mean vector

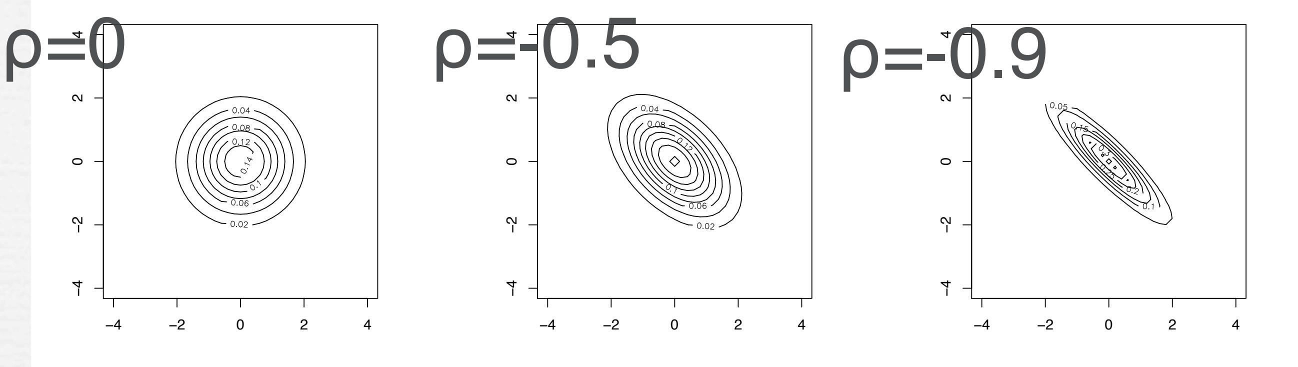
Population variance-covariance matrix



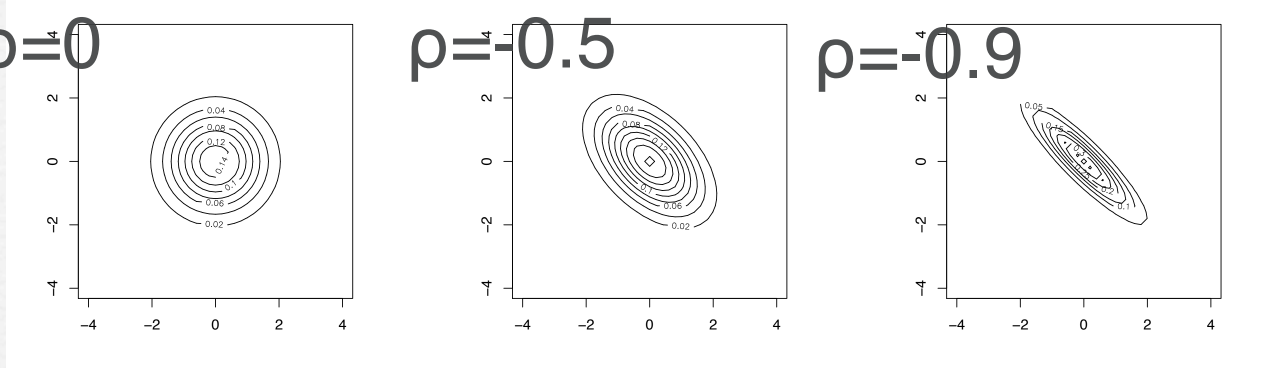
Shape

- A standard multivariate normal distribution is defined to have mean equal to the zero vector, and variance-covariance matrix equal to the identity matrix.

- The shape of a standard multivariate normal is like a nicely rounded hill. Generally, when there is correlation between measurements the hill becomes elongated in the direction of the correlation.

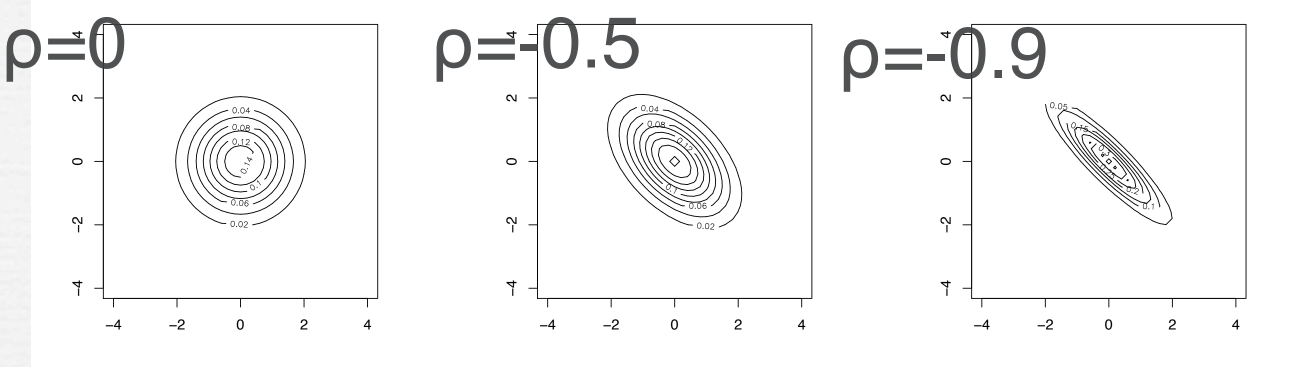


The multivariate normal density function has contours of equal density that are elliptical, a shape that you will see again when we discuss confidence regions.



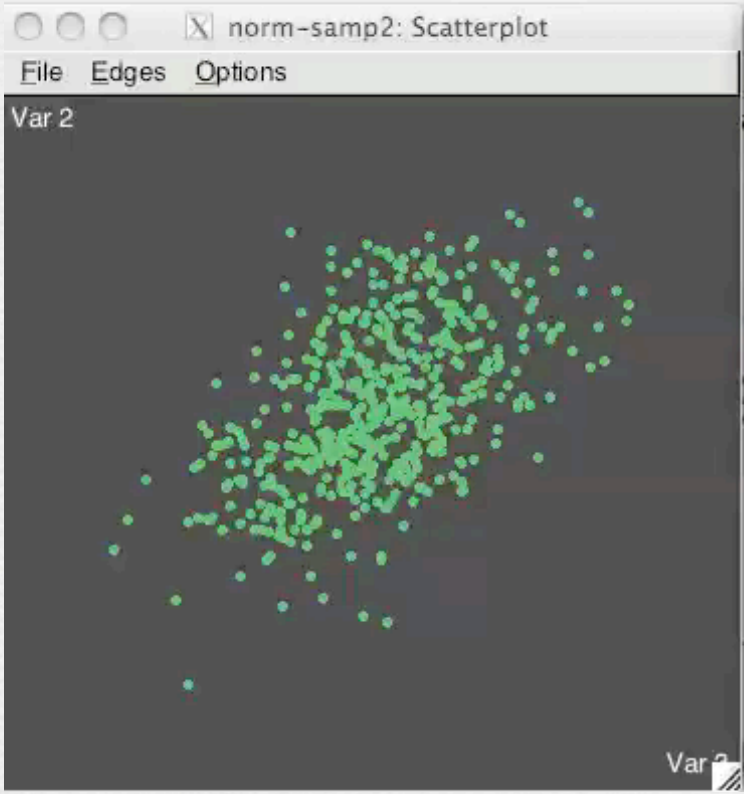
The multivariate normal density function has contours of equal density that are elliptical, a shape that you will see again when we discuss confidence regions.

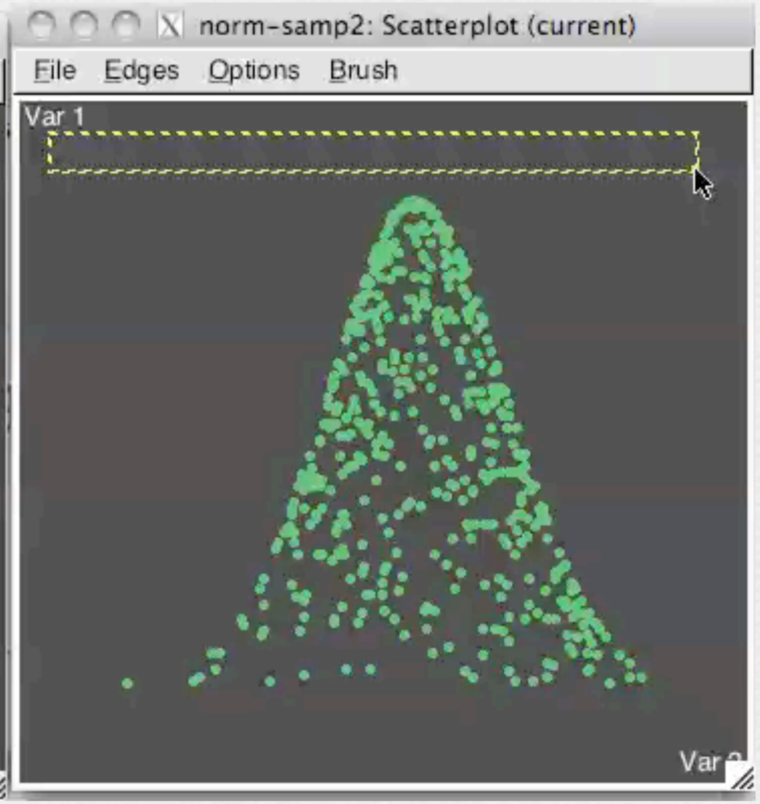
2D

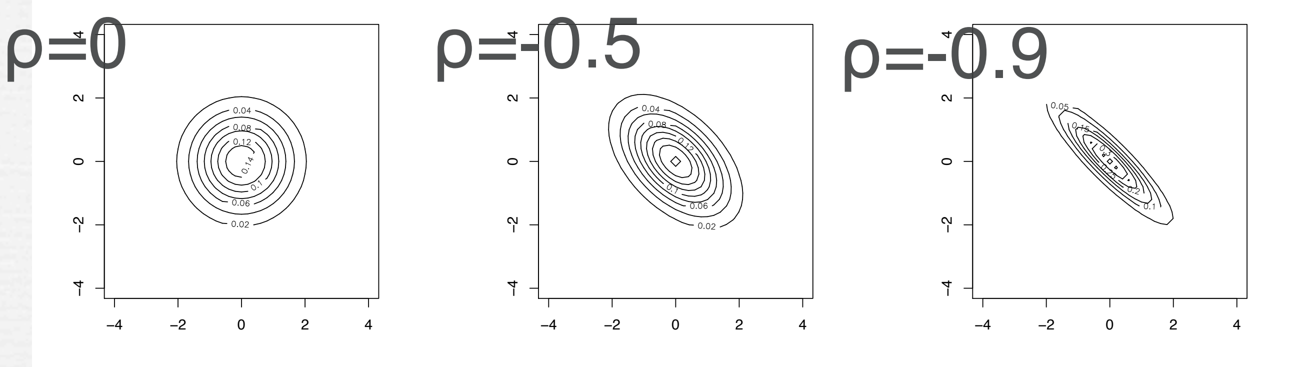


The multivariate normal density function has contours of equal density that are elliptical, a shape that you will see again when we discuss confidence regions.

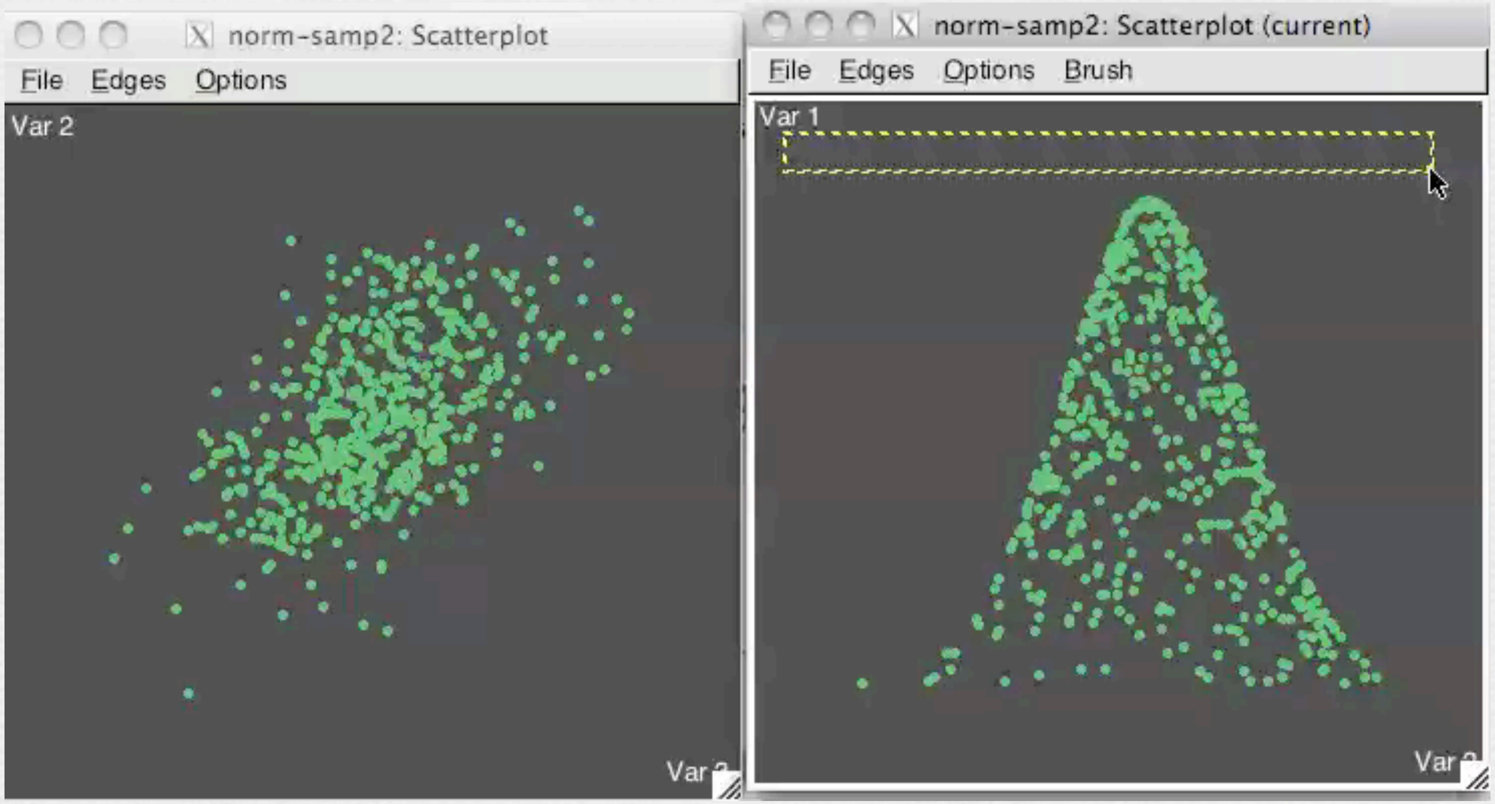
2D

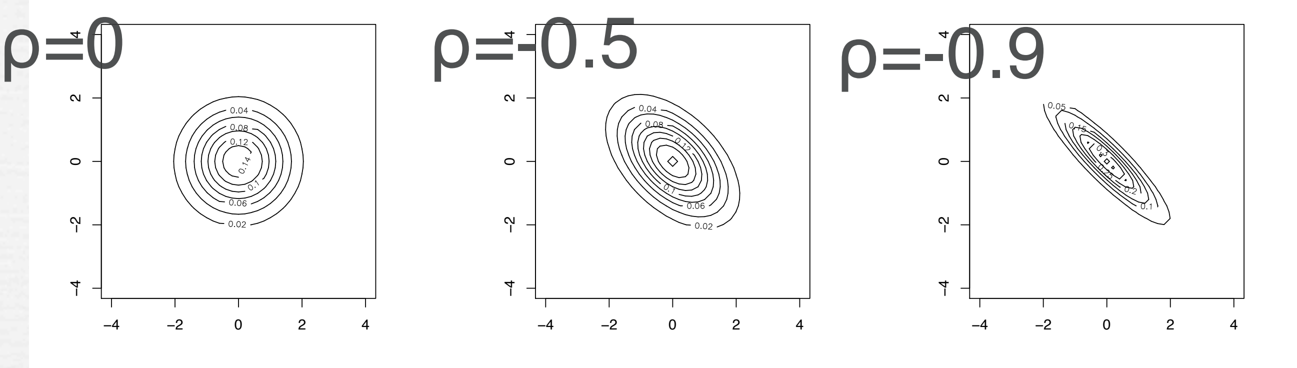






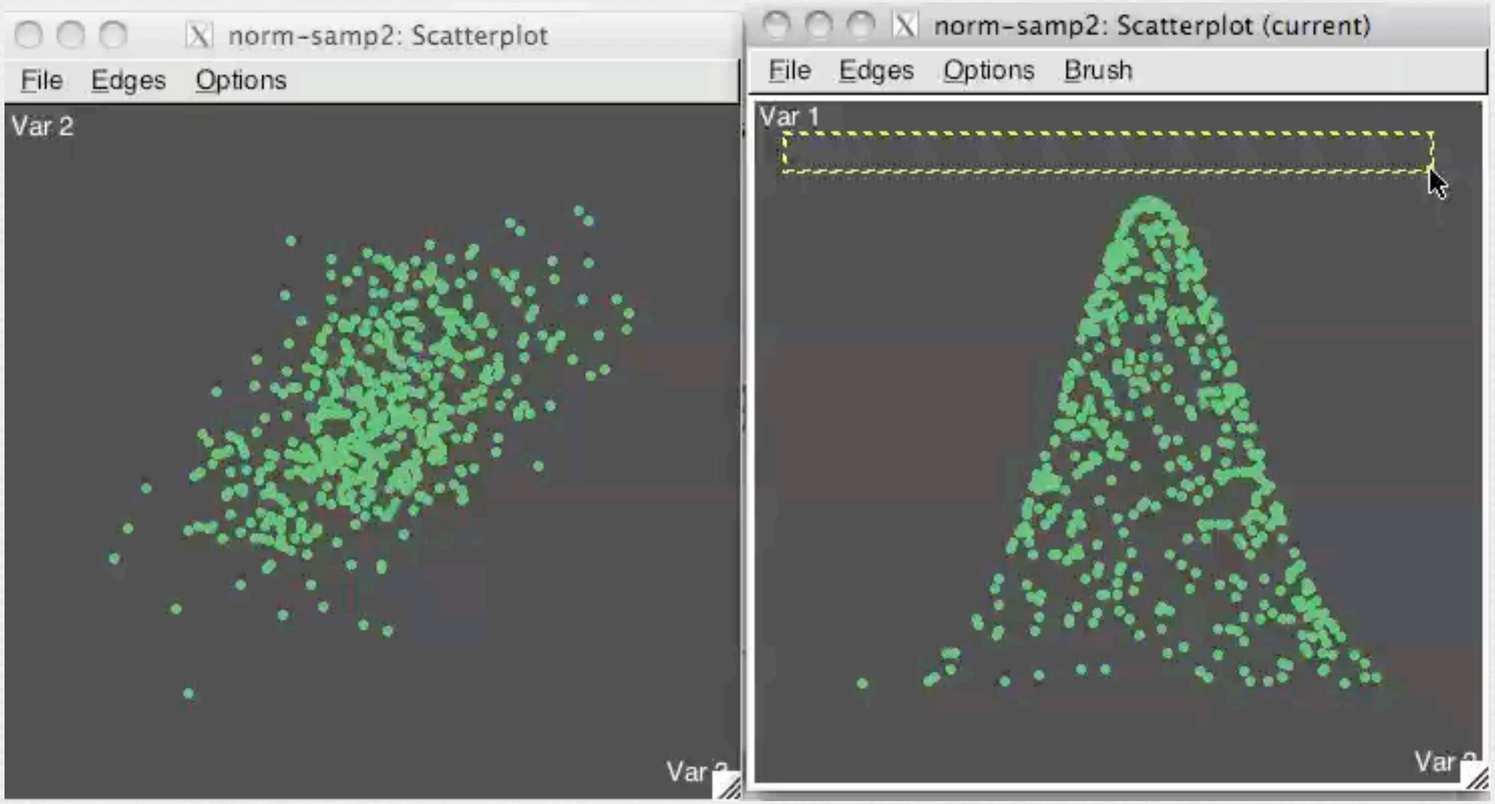
The multivariate normal density function has contours of equal density that are elliptical, a shape that you will see again when we discuss confidence regions.

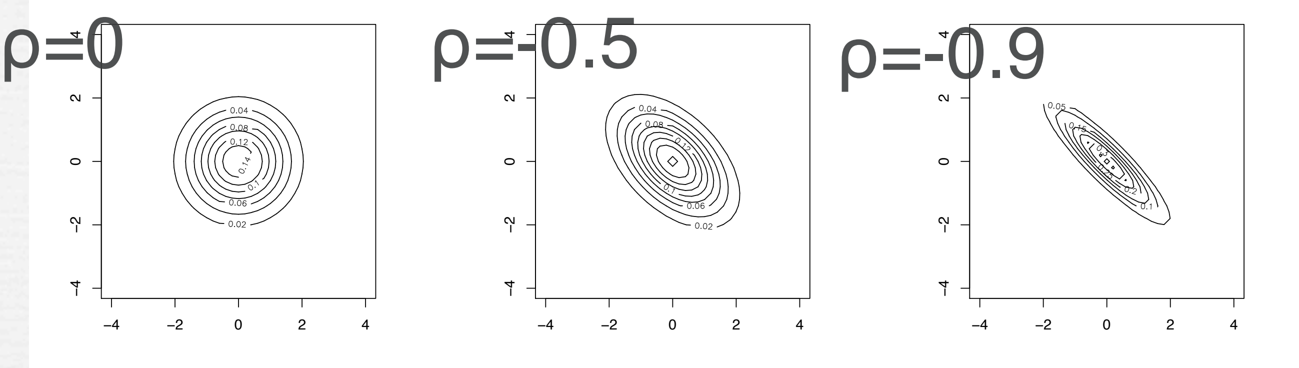




The multivariate normal density function has contours of equal density that are elliptical, a shape that you will see again when we discuss confidence regions.

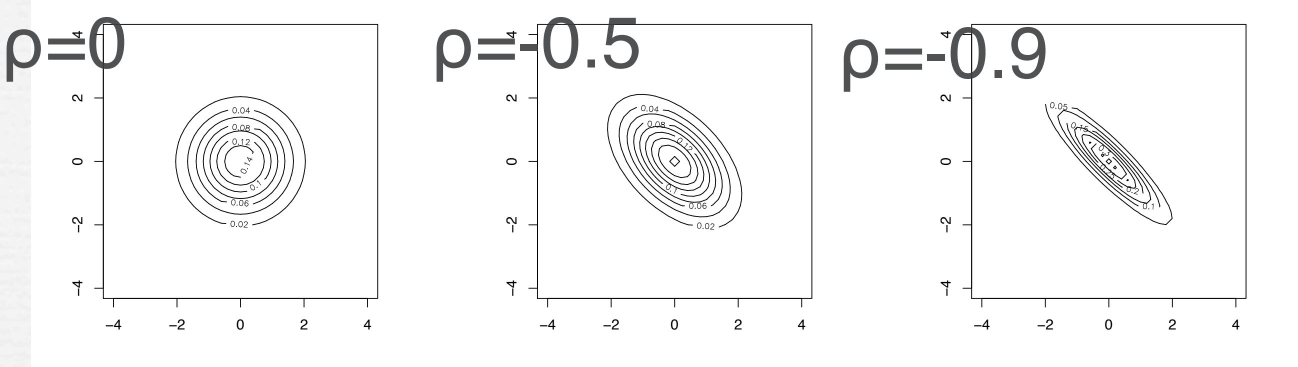
3D





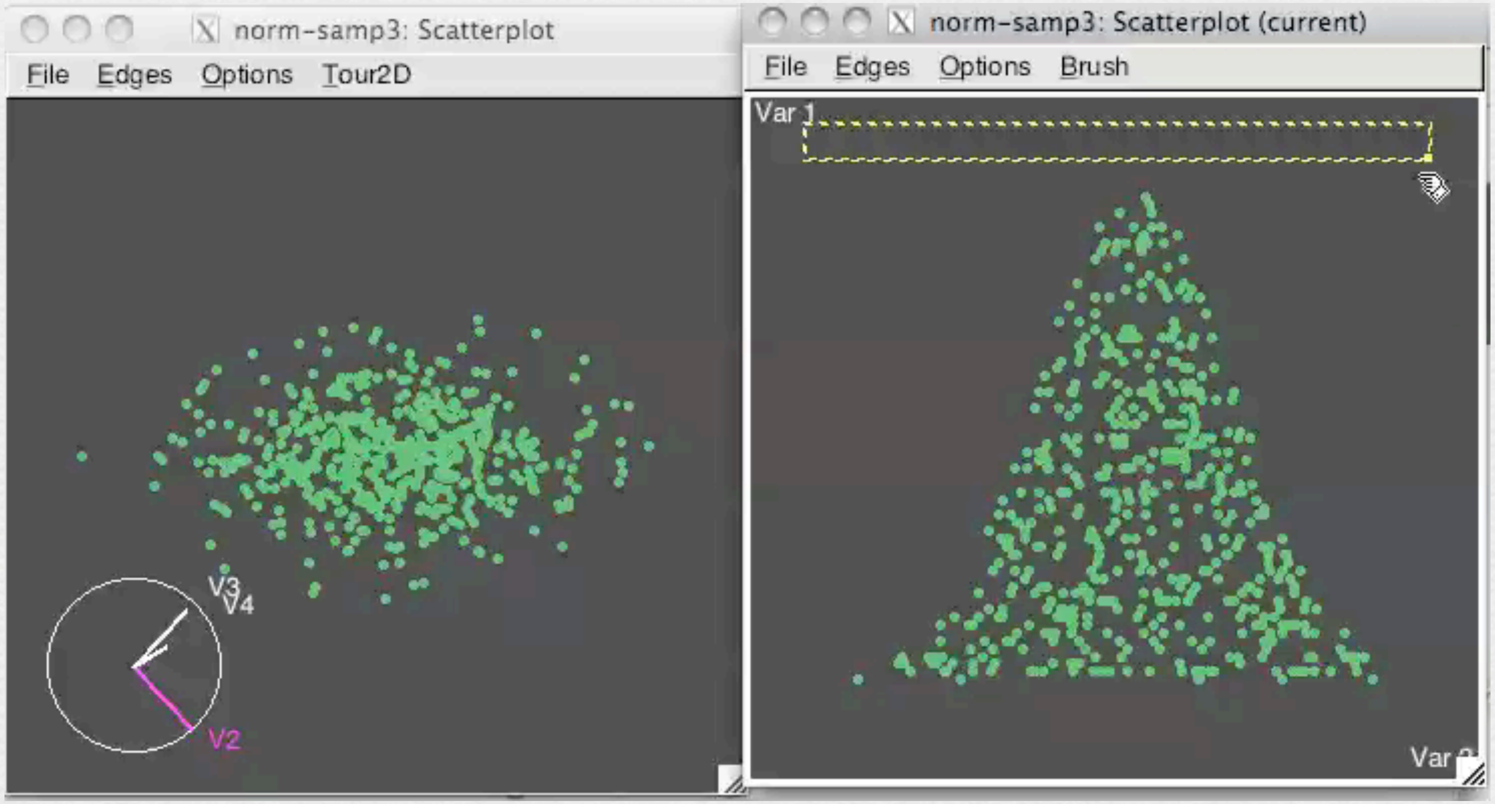
The multivariate normal density function has contours of equal density that are elliptical, a shape that you will see again when we discuss confidence regions.

3D



The multivariate normal density function has contours of equal density that are elliptical, a shape that you will see again when we discuss confidence regions.

3D



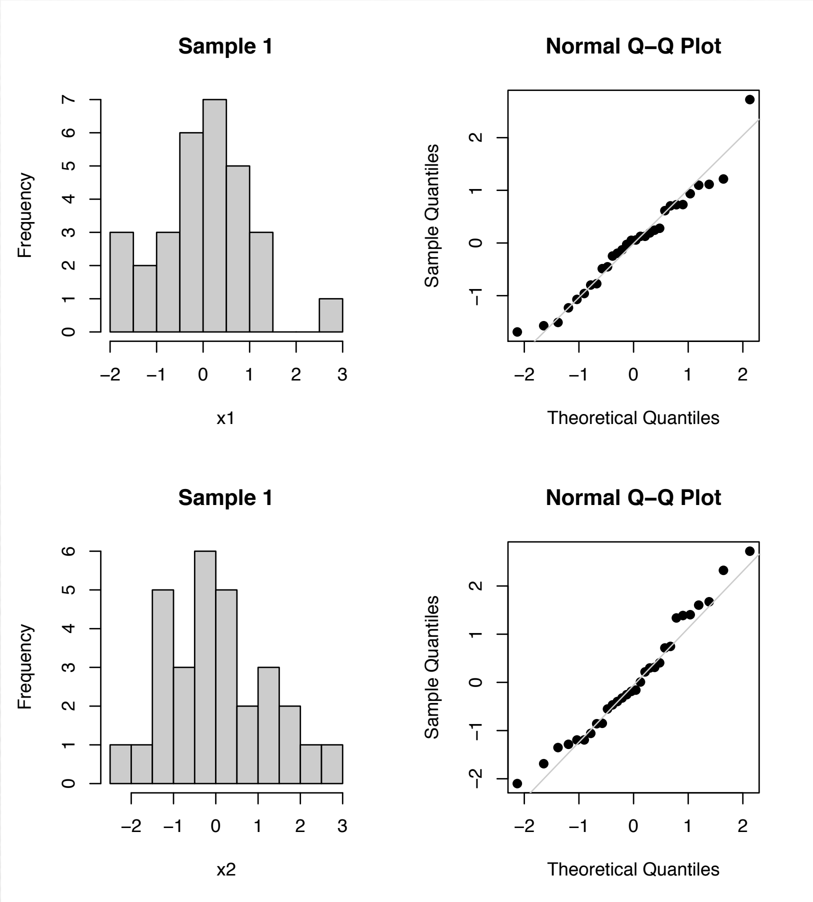
Estimating the Parameters from Data

- The population mean, $\boldsymbol{\mu}$, is estimated by the sample mean, $\bar{\mathbf{x}}$.
- The population variance-covariance, $\boldsymbol{\Sigma}$, is estimated by the sample variance-covariance matrix, $\mathbf{S}$.

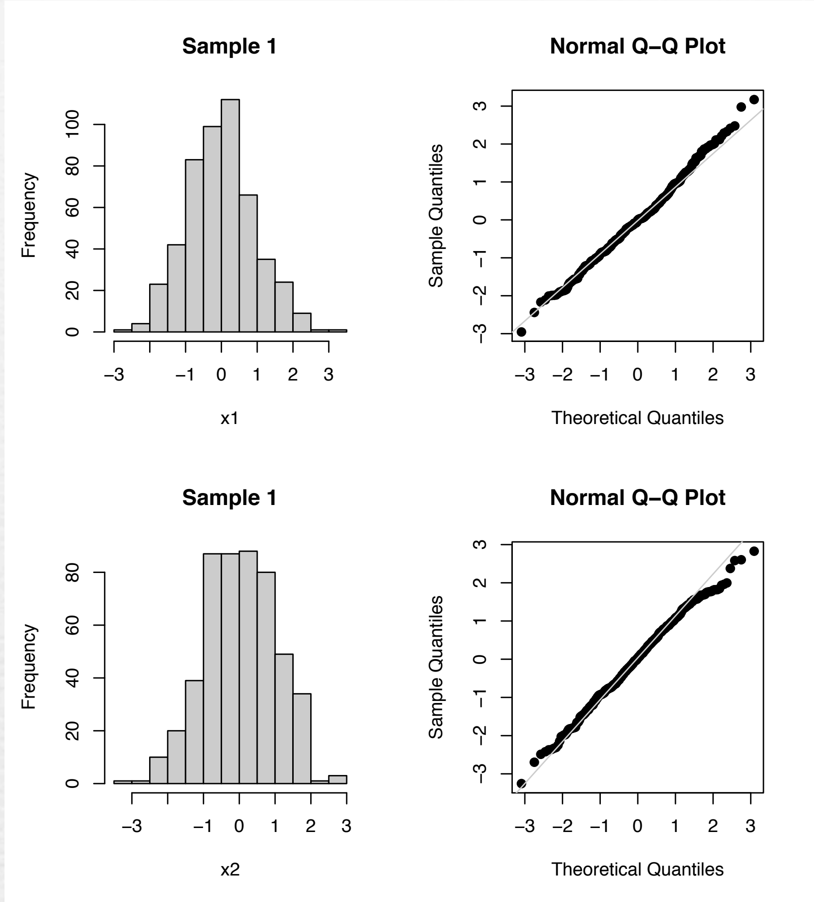
Assessing Normality

- If the sample comes from a multivariate normal population, each of the p variables in a data set will be a univariate normal.
- A normal probability plot, the sorted data values are plotted against the quantiles of a normal distribution, is used. Points should follow a straight line. It is also useful to plot the histogram along with the normal probability plot.

30 cases



500 cases



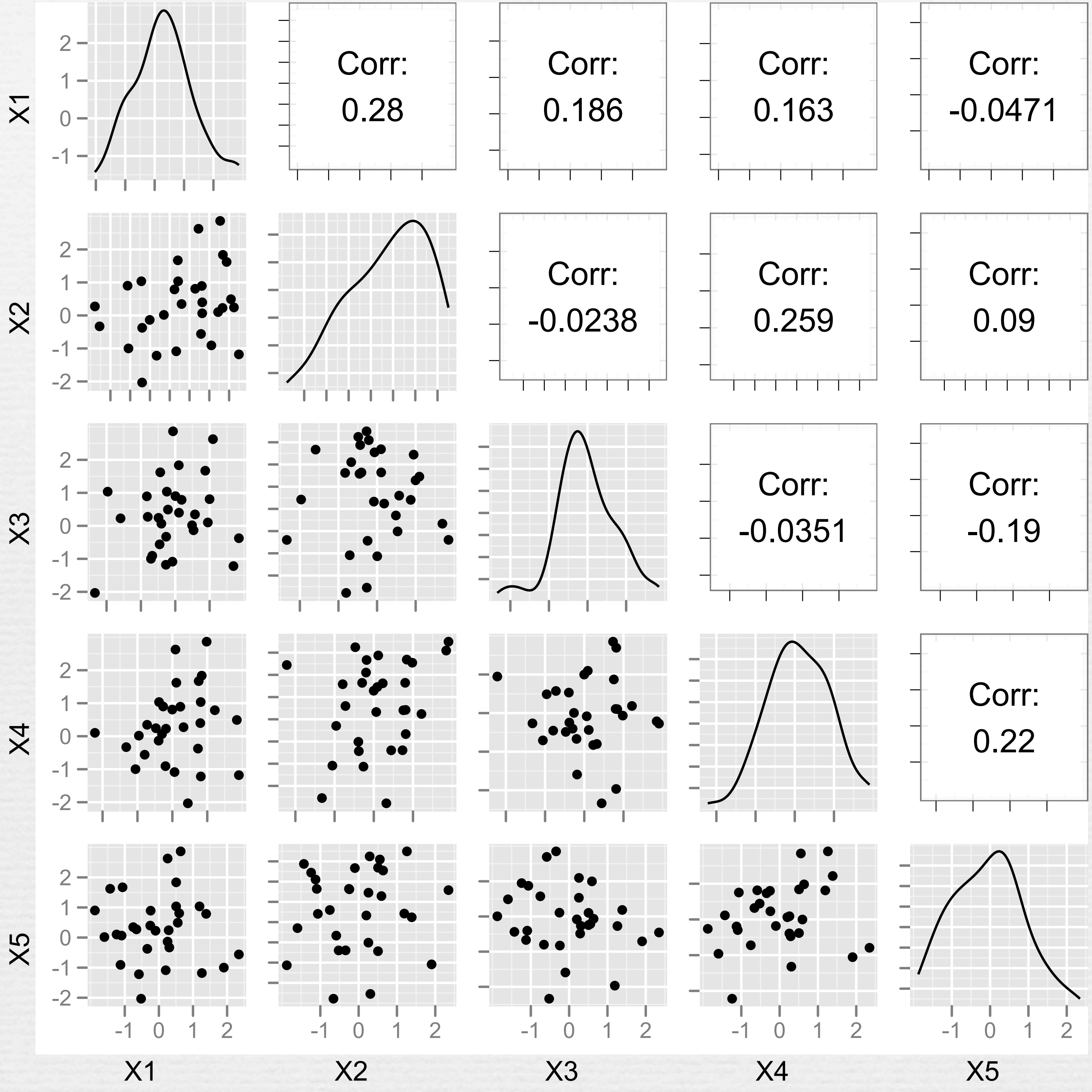
Sample size matters. Small sample size contain a lot of "artifacts". Large samples should be very regular, look bell-shaped and linear normal pp.

Assessing Normality

- If the sample comes from a multivariate normal population, ALL linear combinations of the p variables in a data set will be have a normal distribution.

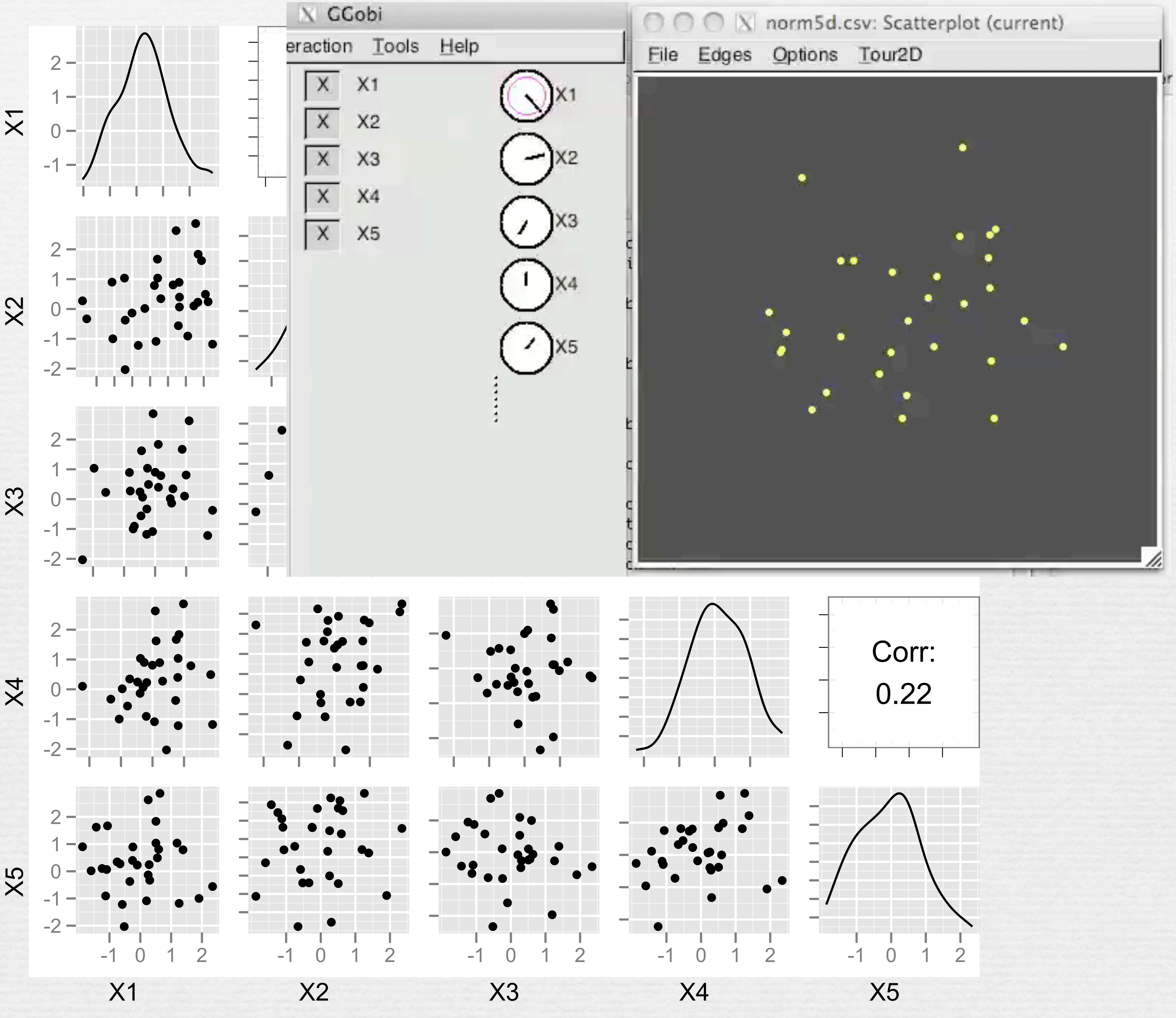
- Examine the data in a scatterplot matrix and a tour. All scatterplots should look elliptical, or spherical. (Using a 1D tour, all views should look bell-shaped.)

n=30

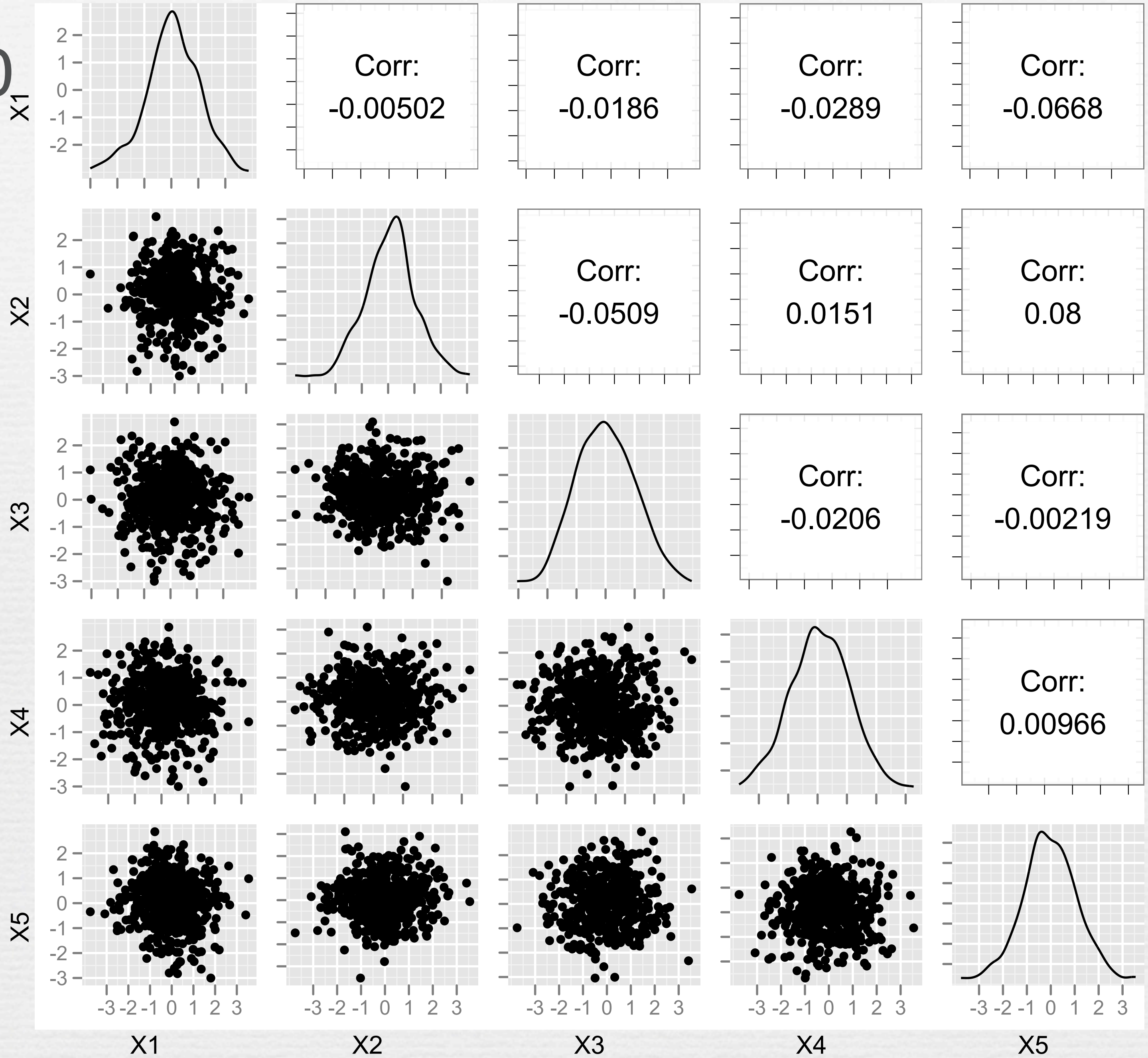


Corr: 0.28
Corr: 0.186
Corr: 0.163
Corr: -0.0471
Corr: -0.0238
Corr: 0.259
Corr: 0.09
Corr: -0.0351
Corr: -0.19
Corr: 0.22
X1
X2
X3
X4
X5

n=30

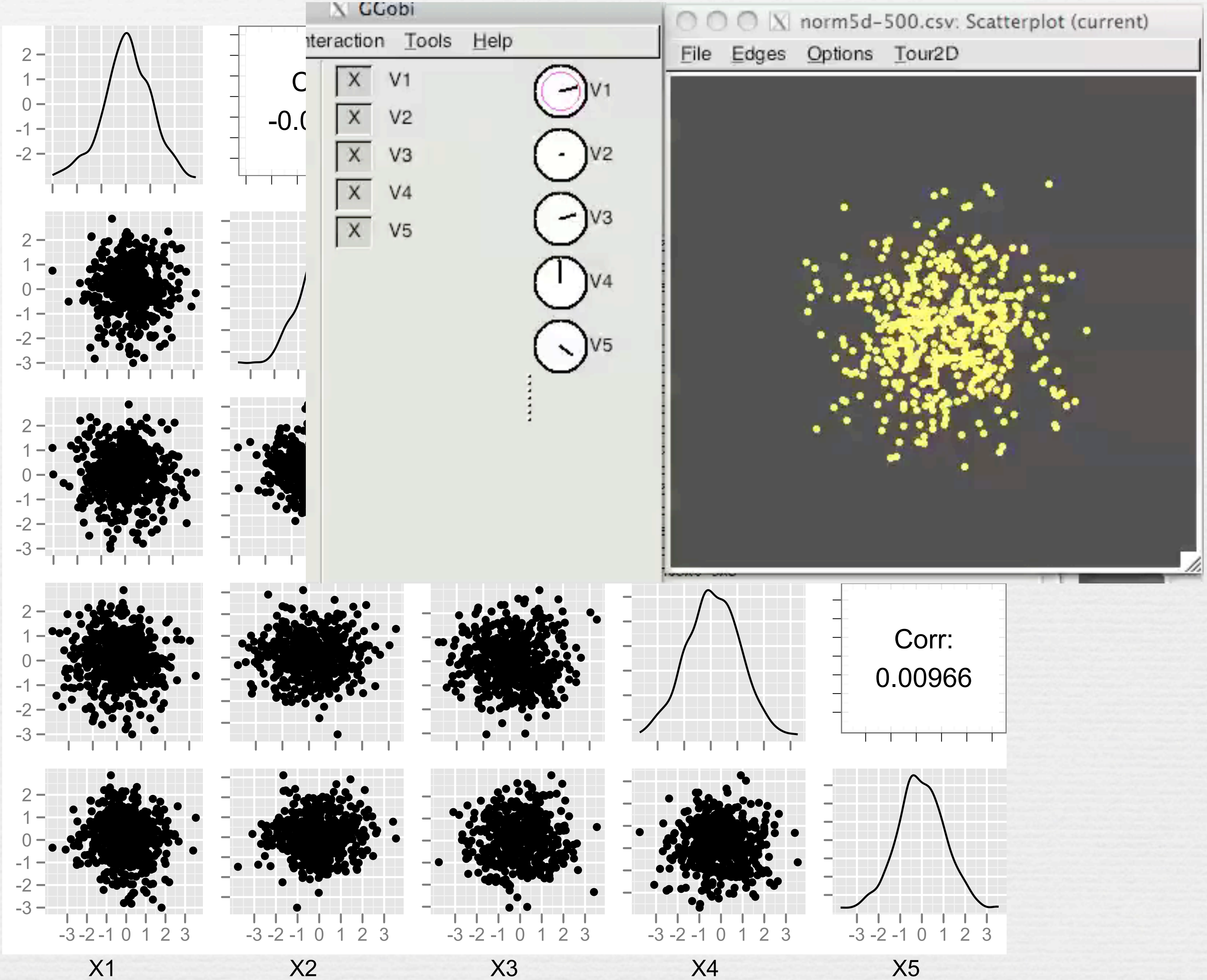

GGobi
eraction Tools Help
X1
X2
X3
X4
X5
norm5d.csv: Scatterplot (current)
File Edges Options Tour2D
Corr:
0.22
X1
X2
X3
X4
X5

n=500



Corr: -0.00502
Corr: -0.0186
Corr: -0.0289
Corr: -0.0668
Corr: -0.0509
Corr: 0.0151
Corr: 0.08
Corr: -0.0206
Corr: -0.00219
Corr: 0.00966
X1
X2
X3
X4
X5

n=500



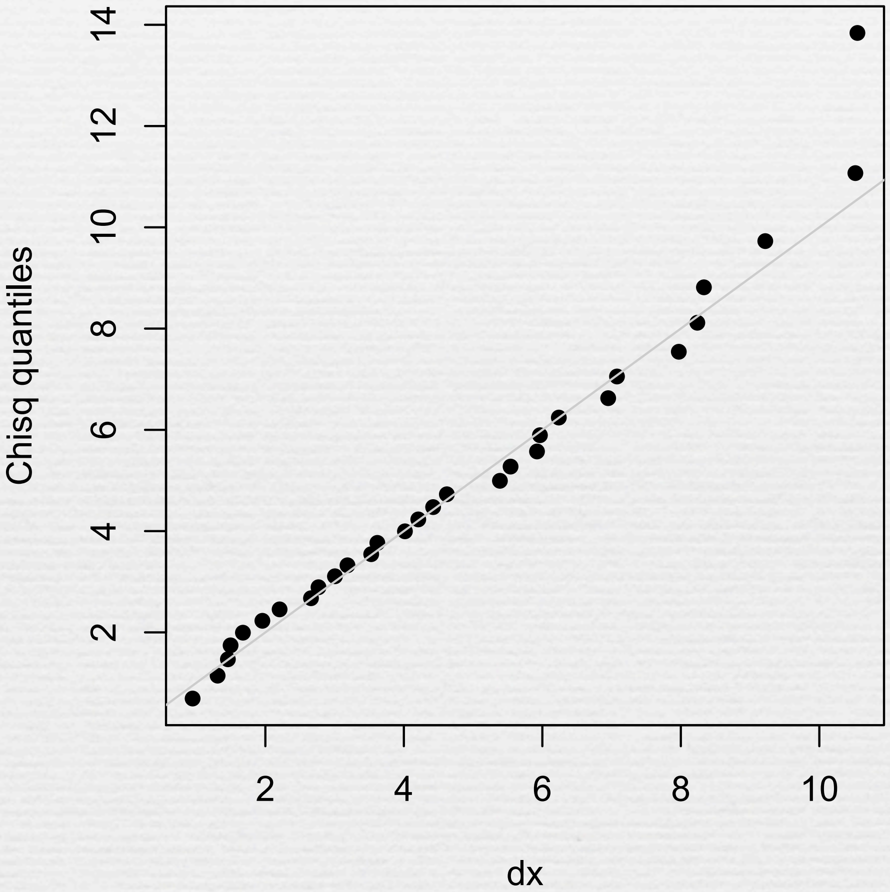
GGobi
Interaction Tools Help
V1
V2
V3
V4
V5
norm5d-500.csv: Scatterplot (current)
File Edges Options Tour2D
Corr:
0.00966
X1
X2
X3
X4
X5

Assessing Normality

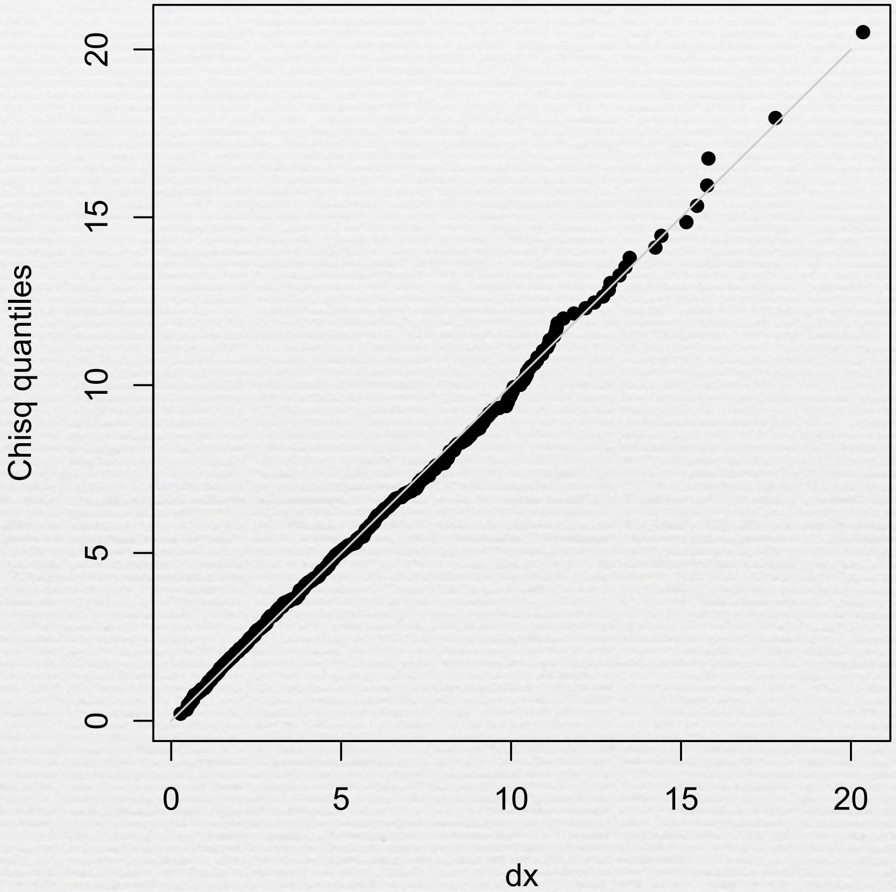
- If the sample comes from a multivariate normal population, the Mahalanobis distances between all points and the mean vector have a χ_p^2 distribution.

- Construct a quantile-quantile plot, sorted distances against quantiles from the χ_p^2 distribution. Points should follow a straight line.

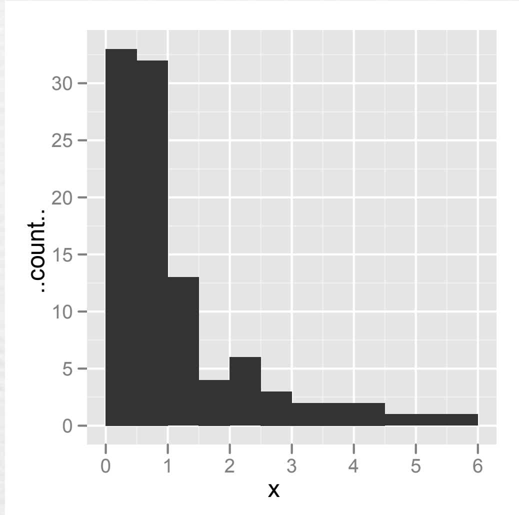
n=30


Chisq quantiles
dx
2
4
6
8
10
12
14

n=500


Chisq quantiles
dx
0
5
10
15
20

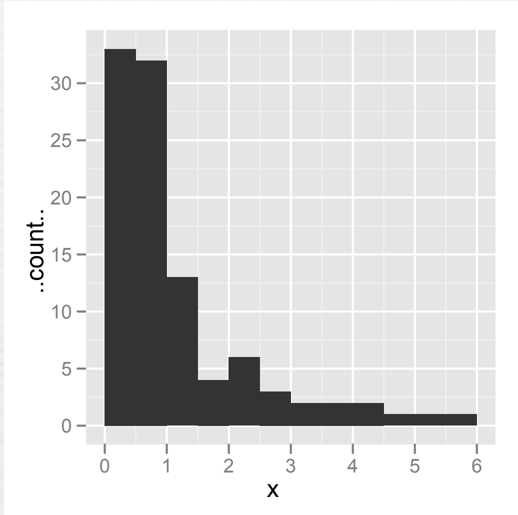
Transformation to Normality

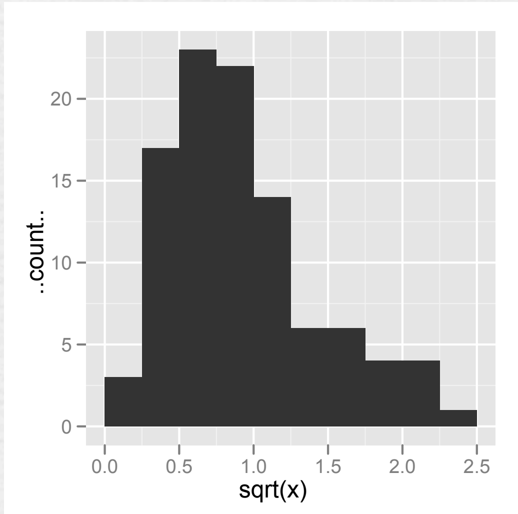


- Transform individual variables - hope it fixes multivariate normality.
- Power transformations, x^a:
  - Right-skewed data: Use $a<1$
  - Left-skewed data: Use $a>1$
  - $a=0$ is a log

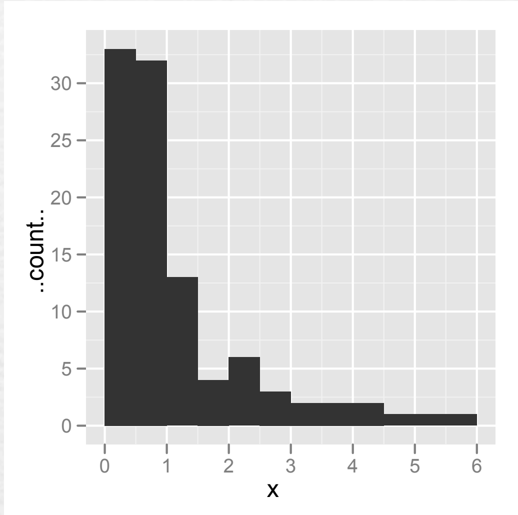
Transformation to Normality

- Transform individual variables - hope it fixes multivariate normality.
- Power transformations, x^a:
  - Right-skewed data: Use $a<1$
  - Left-skewed data: Use $a>1$
  - $a=0$ is a log

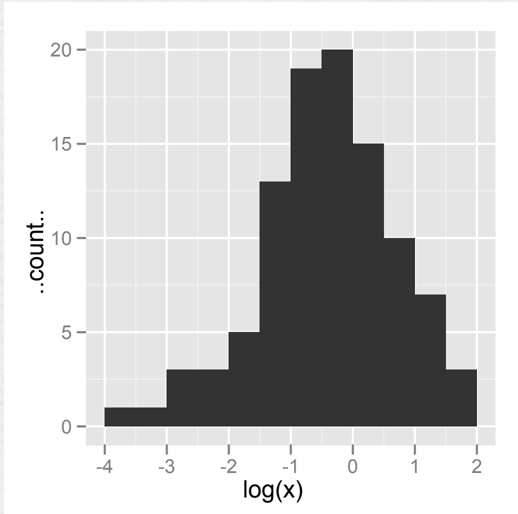




Transformation to Normality



- Transform individual variables - hope it fixes multivariate normality.
- Power transformations, x^a:
  - Right-skewed data: Use $a<1$
  - Left-skewed data: Use $a>1$
  - $a=0$ is a log



Transformation to Normality

- Transform individual variables - hope it fixes multivariate normality.
- Power transformations, x^a:
  - Right-skewed data: Use $a<1$
  - Left-skewed data: Use $a>1$
  - $a=0$ is a log