Statistics with Excel Examples

Questions

- What is a probability density function (pdf)?
- What is a cumulative density function (cdf)?
- What is an *inverse* cdf?
- Examples of distributions:
  - Uniform, Normal, Beta, Gamma, ChiSquare
- Random Numbers
  - What is a uniform distribution?
  - The Excel worksheet function rand()
- Synthesis of distributions.
- What is an “Order Statistic”?
  - Role of Beta distribution.
  - Synthesis of order statistic distributions.

Synthesis of Distributions

- Consider a cdf, F. $P = F(x)$
  - Probability as a function of some distributed variable.
  - Examples of F: BETADIST, GAMMADIST,..
- We want a collection of x’s, x_1, x_2, x_3,…x_i,… distributed according to F.
- How to do it?
  - Generate a random number from the uniform distribution on [0,1] and plug into the inverse cdf. $x_i = F^{-1}(U_i)$
  - Examples

x=BETAINV(rand(),Alpha,Beta)

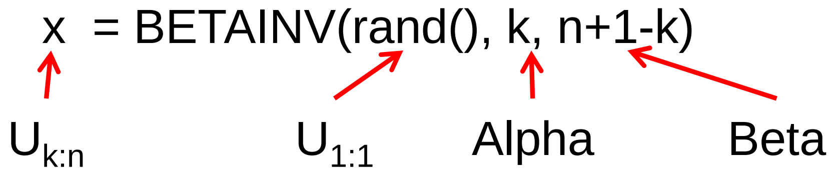
x=NORMSINV(rand()) Distributed normally, with mean 0, and sd = 1.

Order Statistics

- Sample n numbers from a distribution, F.
- Pick the kth smallest
  - k=1 is the smallest.
  - k=n is the largest.
- Do this many times.
- How is the k:n distributed?
  - It is the k:n (“k of n”) order statistic of F.
  - The 1:1 order statistic is F itself.

Synthesis of Order Statistics

- Mental furniture (just *know* it):
  - The k:n order statistic of the uniform distribution is the Beta distribution with Alpha = k, Beta = n+k-1.
- To generate numbers distributed according to the k:n order statistic of the uniform distribution:



x = BETAINV(rand(), k, n+1-k)

$U_{k:n}$ $U_{1:1}$ Alpha Beta

- And for any distribution, generate its k:n order statistic by plugging $U_{k:n}$ into the “probability” argument of its inverse cdf, for example:

$x_{k:n}$=NORMSINV($U_{k:n}$) Isn’t that nice!

Normal Distribution

- To synthesize instances distributed according to a dist’n, plug uniformly dist’d random numbers into the probability argument of the inverse cdf of the dist’n.
- Synthesis of normally distributed data
  - Mean = *m*, Variance = *V* (standard error = √V)

Inverse of standard normal dist’n. Mean = 0, variance = 1.

$$x = m + \sqrt{V} \times z \qquad z \equiv \Phi^{-1}\left(\text{Probability}\right) \text{ is called the "Probit"}$$

$$x = m + \sqrt{V} \times \Phi^{-1}\left(u\right)$$

Uniformly distributed

$$x = m + \sqrt{V} \times \text{NORMSINV}\left(\text{rand}()\right)$$

Normally distributed with mean, *m,* and standard deviation √*V.*

Synthesis of a Multi-Normal Dist’n

- For each sample, instead of generating one random number, generate one *vector* of random numbers.
- And make the numbers in each vector *correlated*.
- To do this, generalize

$$x = m + \sqrt{V} \times z$$

Square *root* of variance!

to

$$\vec{x} = \vec{m} + \overleftrightarrow{R} \times \vec{z}$$

$$\begin{bmatrix} x_1^i \\ x_2^i \\ x_3^i \\ .. \\ .. \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ .. \\ .. \end{bmatrix} + \begin{bmatrix} R_{11} & R_{12} & R_{13} & .. & .. \\ R_{21} & R_{22} & R_{23} & & \\ R_{31} & R_{32} & R_{33} & & \\ .. & & & & \\ .. & & & & \end{bmatrix} \begin{bmatrix} z_1^i \\ z_2^i \\ z_3^i \\ .. \\ .. \end{bmatrix}$$

Desired sample vector *i*.

Vector of means.

Matrix which is the “root” of the covariance matrix.

Vector of independently sampled Probits. Instances are *i* = 1, 2, 3, ..

Covariance Matrix, and Its Root

- The two dimensional covariance matrix is...

$$\overleftrightarrow{V}=\begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}$$

$$\sigma_1^2=\left\langle x_1^2\right\rangle-\left\langle x_1\right\rangle^2$$

$$\sigma_2^2=\left\langle x_2^2\right\rangle-\left\langle x_2\right\rangle^2$$

$$\sigma_1\sigma_2\rho=\left\langle x_1x_2\right\rangle-\left\langle x_1\right\rangle\left\langle x_2\right\rangle$$

- Since V is real-symmetric and +ve definite, *V* can be factorized such that $\overleftrightarrow{V}=\overleftrightarrow{R}'\overleftrightarrow{R}$

Transpose of R.

- So, since..

$$\overleftrightarrow{V}=\begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix}\begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix}=\begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix}\begin{bmatrix}1 & 0 \\ \rho & \sqrt{1-\rho^2}\end{bmatrix}\begin{bmatrix}1 & \rho \\ 0 & \sqrt{1-\rho^2}\end{bmatrix}\begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix}=\begin{bmatrix}\sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2}\end{bmatrix}\begin{bmatrix}\sigma_1 & \sigma_2\rho \\ 0 & \sigma_2\sqrt{1-\rho^2}\end{bmatrix}$$

- ..we have

$$\overleftrightarrow{R}'=\begin{bmatrix}\sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2}\end{bmatrix}$$

The upper (or lower) triangular root is the “Cholesky root”.

Covariance Matrix, and Its Root

- Other roots differing from the Cholesky root by a rotation work too..

$$\overleftrightarrow{V}=\begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}$$

$$=\begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}\begin{bmatrix}\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right) & \frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right)\\ \frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right) & \frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right)\end{bmatrix}\begin{bmatrix}\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right) & \frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right)\\ \frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right) & \frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right)\end{bmatrix}\begin{bmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{bmatrix}$$

$$=\begin{bmatrix}\sigma_1\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right) & \sigma_1\frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right)\\ \sigma_2\frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right) & \sigma_2\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right)\end{bmatrix}\begin{bmatrix}\sigma_1\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right) & \sigma_2\frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right)\\ \sigma_1\frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right) & \sigma_2\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right)\end{bmatrix}$$

- So

$$\overleftrightarrow{R}'=\begin{bmatrix}\sigma_1\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right) & \sigma_1\frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right)\\ \sigma_2\frac{1}{2}\left(\sqrt{1+\rho}-\sqrt{1-\rho}\right) & \sigma_2\frac{1}{2}\left(\sqrt{1+\rho}+\sqrt{1-\rho}\right)\end{bmatrix}$$

2-D Example

- What happens when ρ = 1, ρ = -1, ρ = 0?

$$\vec{x} = \vec{m} + \overleftrightarrow{R}'\vec{z}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} + \begin{bmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

$$x_1 = m_1 + \sigma_1 z_1$$

$$x_2 = m_2 + \sigma_2 \rho z_1 + \sigma_2 \sqrt{1-\rho^2}\, z_2$$

$$z_1 = \Phi^{-1}\left(U_1\right) = \text{NORMSINV(rand())}$$

$$z_2 = \Phi^{-1}\left(U_2\right) = \text{NORMSINV(rand())}$$

- U_1 and U_2 are independently sampled from the uniform distribution on [0,1].

Gaussian Copula

- Set $m_1 = m_2 = 0$, $\sigma_1 = \sigma_2 = 1$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

$$z_1 = \Phi^{-1}\left(U_1\right) = \text{NORMSINV(rand())}$$

$$z_2 = \Phi^{-1}\left(U_2\right) = \text{NORMSINV(rand())}$$

- x_1 and x_2 are normally distributed with mean = 0, var = 1

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \Phi\left(x_1\right) \\ \Phi\left(x_2\right) \end{bmatrix}$$

- c_1 and c_2 are uniform on [0,1], so simulate any marginal distn’s

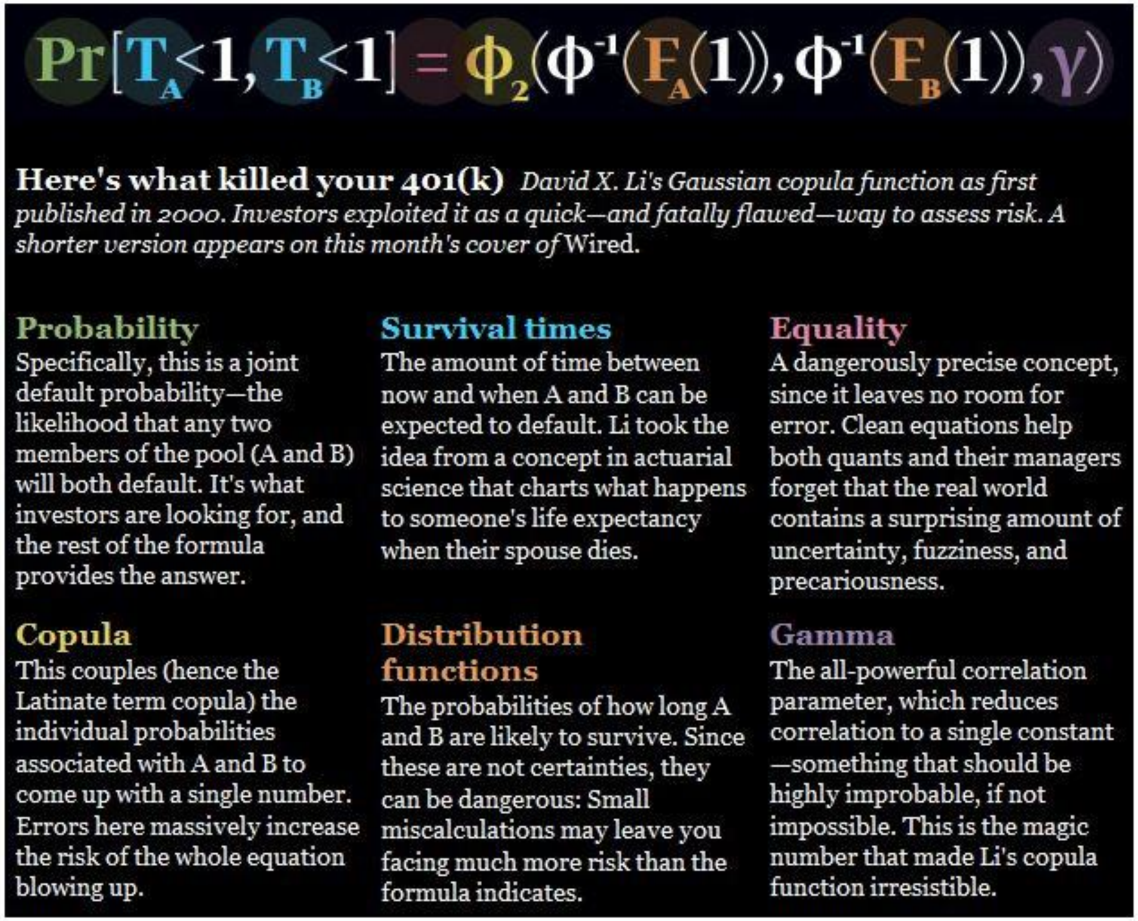
$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} A^{-1}\left[\Phi\left(x_1\right)\right] \\ B^{-1}\left[\Phi\left(x_2\right)\right] \end{bmatrix}$$

A, B are arbitrary cdfs.

inverse!

Gaussian Copula

- Infamously implicated in financial disaster:



$$\Pr[T_A<1, T_B<1] = \phi_2(\phi^{-1}(F_A(1)), \phi^{-1}(F_B(1)), \gamma)$$

Here's what killed your 401(k) *David X. Li's Gaussian copula function as first published in 2000. Investors exploited it as a quick—and fatally flawed—way to assess risk. A shorter version appears on this month's cover of* Wired.

Probability
Specifically, this is a joint default probability—the likelihood that any two members of the pool (A and B) will both default. It's what investors are looking for, and the rest of the formula provides the answer.

Copula
This couples (hence the Latinate term copula) the individual probabilities associated with A and B to come up with a single number. Errors here massively increase the risk of the whole equation blowing up.

Survival times
The amount of time between now and when A and B can be expected to default. Li took the idea from a concept in actuarial science that charts what happens to someone's life expectancy when their spouse dies.

Distribution functions
The probabilities of how long A and B are likely to survive. Since these are not certainties, they can be dangerous: Small miscalculations may leave you facing much more risk than the formula indicates.

Equality
A dangerously precise concept, since it leaves no room for error. Clean equations help both quants and their managers forget that the real world contains a surprising amount of uncertainty, fuzziness, and precariousness.

Gamma
The all-powerful correlation parameter, which reduces correlation to a single constant—something that should be highly improbable, if not impossible. This is the magic number that made Li's copula function irresistible.

Recipe for Disaster:
The Formula that Killed Wall Street.
Wired Mag. February 2009

Extension to N Dimensions

- Gaussian copulas are easily extended to N dimensions.

$$\overleftrightarrow{V}=\begin{bmatrix} V_{11} & V_{12} & V_{13} & .. \\ V_{21} & V_{22} & V_{23} & .. \\ V_{31} & V_{32} & V_{33} & .. \\ .. & .. & .. & .. \end{bmatrix} \qquad V_{ij}=V_{ji}=\left\langle x_i x_j \right\rangle - \left\langle x_i \right\rangle \left\langle x_j \right\rangle$$

- If all marginal distributions have m = 0 and σ = 1, then

$$\overleftrightarrow{V}=\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & .. \\ \rho_{21} & 1 & \rho_{23} & .. \\ \rho_{31} & \rho_{32} & 1 & .. \\ .. & .. & .. & .. \end{bmatrix}$$

ρ_{ij} = ρ_{ji} is the correlation coefficient between variables *i* and *j*.

- Calculation of the Cholesky root of V.
  - Analytically messy for N > 2.
  - But algorithms are easily available.