Creating Distributions Empirically

Milo Schield
Augsburg College
Editor of www.StatLit.org
US Rep: International Statistical Literacy Project

Fall 2014 NNN Conference

www.StatLit.org/pdf/
2014-Schield-Create-Distributions-Empirically-Excel-6up.pdf

Generating Distributions

In introductory statistics, students are shown how the means of random samples form a sampling distribution that -- in the limit -- forms a Normal distribution.

This is extremely useful in sampling. But the Normal distribution applies in situations that do not involve sampling – as do the Log-Normal and Exponential.

Certain kinds of random activity can be shown to generate these well-known analytic distributions.

Understanding how this happens can be very useful to managers making data-based decisions.

Generating Distributions: Normal, Log-Normal & Exponential

Mathematicians have proven:

1. The sum of an infinite number of independent random variables generates a Normal distribution
2. The product of an infinite # of random, independent, positive variables generates a Log-Normal distribution
3. A random process of assigning counts to cells in a table can generate a chi-square distribution.
4. A process having a constant chance of 'death' generates an exponential distribution.

Amazing! At the micro level, pure randomness. At the macro level, an analytic distribution emerges without any agency or intent. Micro randomness generates macro-order: spontaneous and unplanned.

Generating Distributions Empirically

Excel is used to generate these distributions empirically.

The Normal and Log-Normal are the empirical results of 10,000 independent lines of activity. Each line receives:

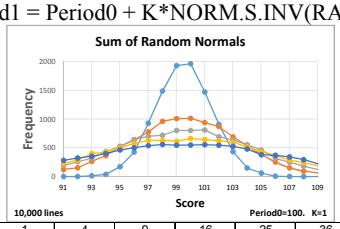
- An add by a random amount (discrete or continuous) generates a Normal distribution.
- a multiply by a positive random change (discrete or continuous) generates a Log-Normal distribution.

The Chi-square is the empirical result of randomly assign values to cells in a table (keeping the expected values equal) and then compute chi-square for each series.

The Exponential is the empirical result of 2,000 lines of activity; each line has a fixed chance of death per period.

1a Generating a Normal Dist. Sum of Random Normals

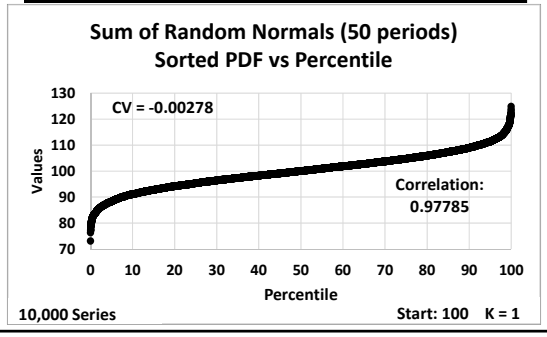
Period1 = Period0 + K*NORM.S.INV(RAND())



Period	1	4	9	16	25	36	49
Median	100.02	100.03	100.01	100.04	100.01	100.03	100.04
StdDev	1.00	2.00	3.02	3.16	3.31	3.45	3.60
PredictSD	1	2	3	4	5	6	7

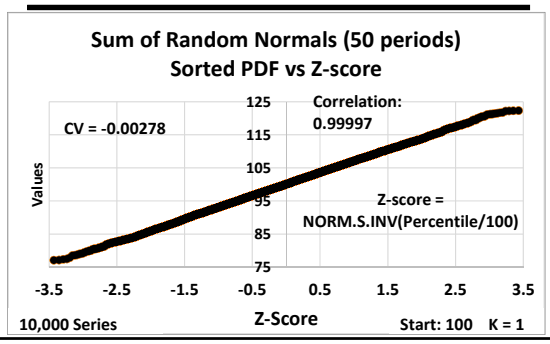
SD(T) = K*SD(1)*Sqrt(T) SD(1) = 1.

Test for Normality #1 QQ-Plot: Sorted PDF vs Percentile



0E 2014 Schield Creating Distributions Empirically 7

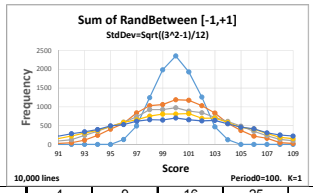
Test for Normality #2
QQ-Plot: Sorted PDF vs. Z-Score



0E 2014 Schield Creating Distributions Empirically 8

1b Generating a Normal Dist.
Sum of Random Discretes

Period1 = Period0 + K*RandBetween(-1,+1)



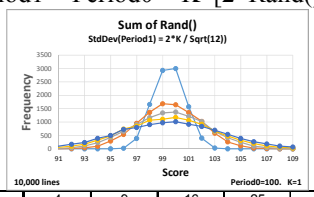
Period	1	4	9	16	25	36	49
Median	100.00	100.00	100.00	100.00	100.00	100.00	100.00
StdDev	0.815	1.64	2.45	3.27	4.10	4.90	5.72
PredictSD	0.815	1.63	2.45	3.26	4.08	4.89	5.71

SD(T) = K*SD(1)*Sqrt(T) SD(1) = Sqrt(8/12)

0E 2014 Schield Creating Distributions Empirically 9

1c Generating a Normal Dist.
Sum of Random Uniforms

Period1 = Period0 + K*[2*Rand()-1]



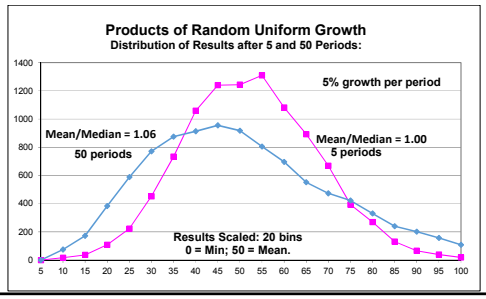
Period	1	4	9	16	25	36	49
Median	100.00	99.99	100.00	100.00	100.05	100.03	100.08
StdDev	0.576	1.17	1.75	2.32	2.87	3.43	3.98
PredictSD	0.577	1.15	1.73	2.31	2.89	3.46	4.04

SD(T) = K*SD(1)*Sqrt(T). SD(1) = 2/Sqrt(12)

0E 2014 Schield Creating Distributions Empirically 10

2a Generating a Log-Normal Dist.
Product: Random Normal Growth

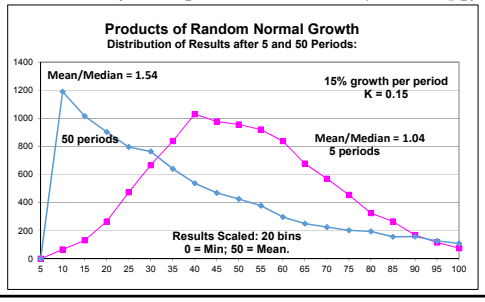
P1 = P0*{1+K*[1+NORMSINV(RAND()]}



0E 2014 Schield Creating Distributions Empirically 11

2a Generating a Log-Normal Dist.
Product: Random Normal Growth

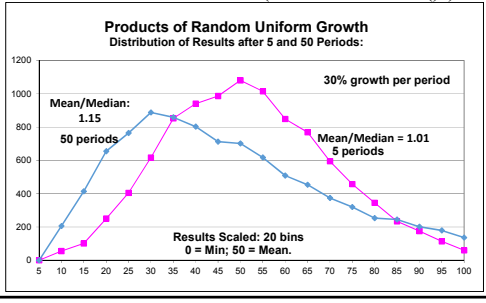
P1 = P0*{1+K*[1+NORMSINV(RAND()]}



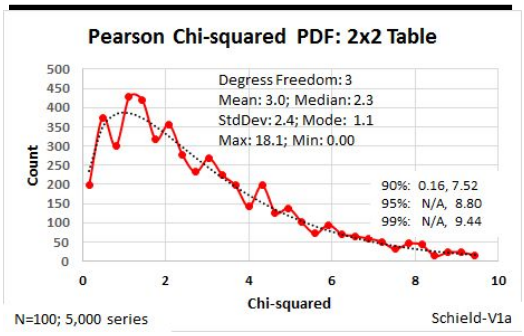
0E 2014 Schield Creating Distributions Empirically 12

2b Generating a Log-Normal Dist.
Product: Random Uniform Growth

Period1 = Period0 * (1 + K*2*Rand())

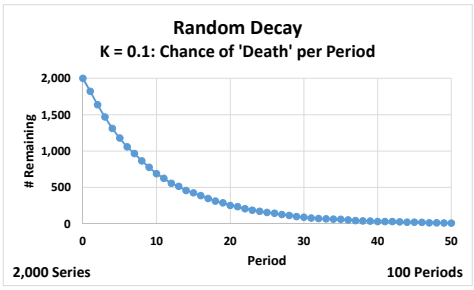


3 Generating Chi-Square Distrib.: Random Assignment to Table Cells



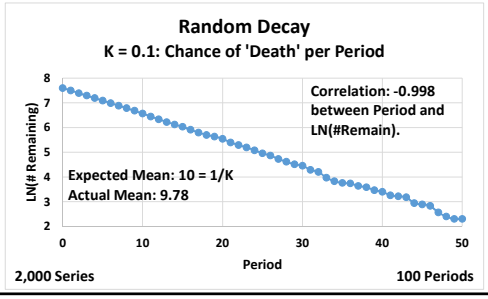
4 Generating an Exponential Dist. Random "Death"

Chance of Death: =IF(RAND() < K, 1, "")



4 Generating an Exponential Dist. Random "Death"

Exponential Test: Is LN(#Remain) vs. time linear?



Conclusion

Knowing the process that generates a distribution is helpful to anyone *who wants to understand why* a given distribution is a good fit in a particular situation.

It also helps us *understand what determines* the parameters of a distribution (or changes therein).

Why have these increased over time?

- the standard deviation of heights?
- the mean/median ratio (the skewness) for incomes?
- the economic return from higher education?
- the average lifetime of humans?

Bibliography

Aitchison and Brown (1957, 1963). *The Log Normal Distribution with Special References to its uses in Economics*. Cambridge U.

Crow, E. and Shimizu, K. (1988). *Lognormal Distributions: Theory and Applications*. New York: Marcel Dekker, Inc., 1988.

International Futures (2014) Using Lognormal Income Distributions www.du.edu/ifs/help/understand/economy/poverty/lognormal.html

Schield, M. and T. Burnham (2008). Von Mises' Frequentist Approach to Probability, 2008 ASA Proceedings of the Section on Statistical Education. [CD-ROM] P. 2187-2194. www.StatLit.org/pdf/2008SchieldBurnhamASA.pdf.

Wikipedia/Log-normal distribution, Uniform distribution (continuous) and Uniform distribution (discrete)

Using Excel to Create Distributions Empirically

All files are in www.StatLit.org/Excel/

Generating Normal Distributions with Sums of Random Variables:
* 2014-Schield-Create-Normal-Empirically-Excel-Sum-Discrete.xlsx
* 2014-Schield-Create-Normal-Empirically-Excel-Sum-Uniform.xlsx
* 2014-Schield-Create-Normal-Empirically-Excel-Sum-Normal.xlsx

Creating Log-Normals with Product of Uniform Random Variables:
* 2014-Schield-Create-LogNormal-Empirically-Excel-Product-Uniforms.xlsx

Creating Chi-Square with Random Assignment to Table Cells:
* 2014-Schield-Create-ChiSquare-Empirically-Excel-DF3.xlsx

Generating Exponentials with Fixed Chance of 'Death" per period
* 2014-Schield-Create-Exponential-Empirically-Excel.xlsx

Creating Distributions Empirically

Milo Schield
Augsburg College
Editor of www.StatLit.org
US Rep: International Statistical Literacy Project

Fall 2014 NNN Conference

www.StatLit.org/pdf/2014-Schield-NNN3-Slides.pdf

Generating Distributions

In introductory statistics, students are shown how the means of random samples form a sampling distribution that -- in the limit -- forms a Normal distribution.

This is extremely useful in sampling. But the Normal distribution applies in situations that do not involve sampling – as do the Log-Normal and Exponential.

Certain kinds of random activity can be shown to generate these well-known analytic distributions.

Understanding how this happens can be very useful to managers making data-based decisions.

Generating Distributions: Normal, Log-Normal & Exponential

Mathematicians have proven:

1. The sum of an infinite number of independent random variables generates a Normal distribution
2. The product of an infinite # of random, independent, positive variables generates a Log-Normal distribution
3. A random process of assigning counts to cells in a table can generate a chi-square distribution.
4. A process having a constant chance of 'death' generates an exponential distribution.

Amazing! At the micro level, pure randomness. At the macro level, an analytic distribution emerges without any agency or intent. Micro randomness generates macro-order: spontaneous and unplanned.

Generating Distributions Empirically

Excel is used to generate these distributions empirically.

The Normal and Log-Normal are the empirical results of 10,000 independent lines of activity. Each line receives:

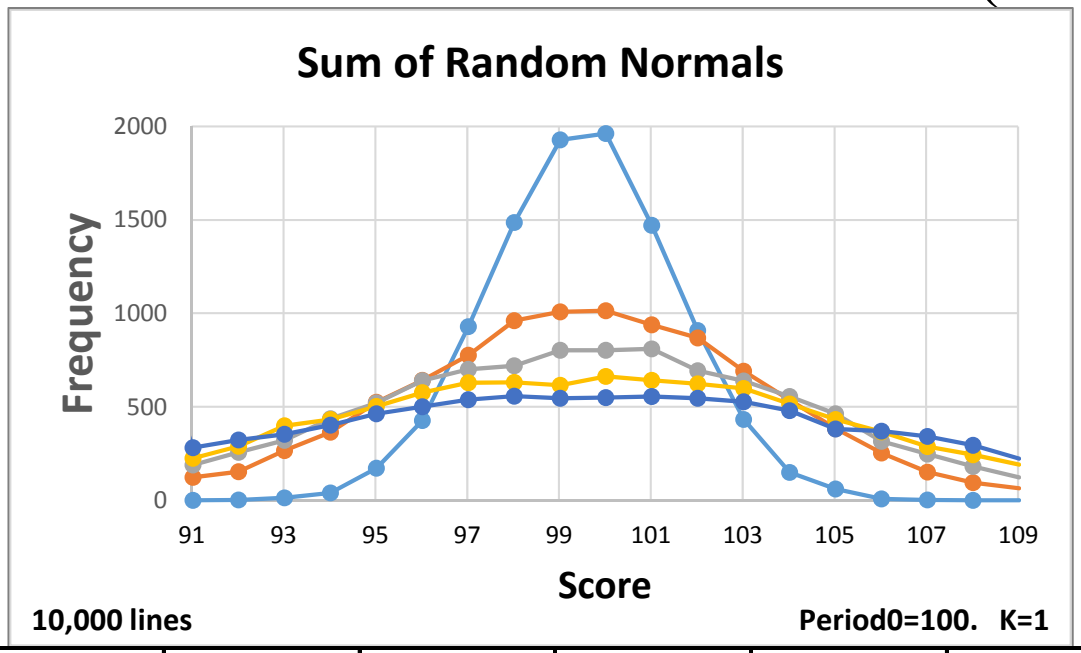
- An add by a random amount (discrete or continuous) generates a Normal distribution.
- a multiply by a positive random change (discrete or continuous) generates a Log-Normal distribution.

The Chi-square is the empirical result of randomly assign values to cells in a table (keeping the expected values equal) and then compute chi-square for each series.

The Exponential is the empirical result of 2,000 lines of activity; each line has a fixed chance of death per period.

1a Generating a Normal Dist. Sum of Random Normals

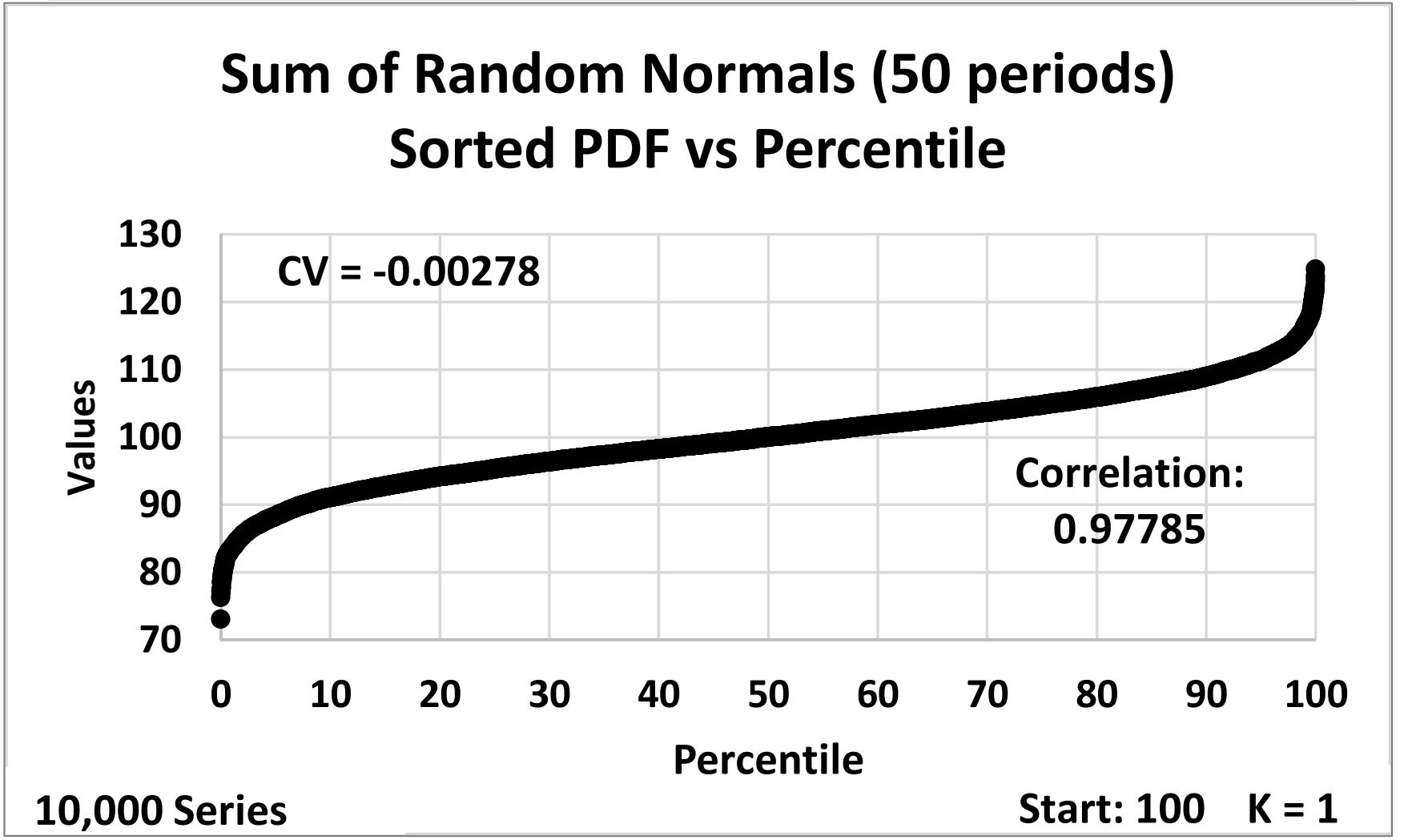
$$Period1 = Period0 + K*NORM.S.INV(RAND())$$



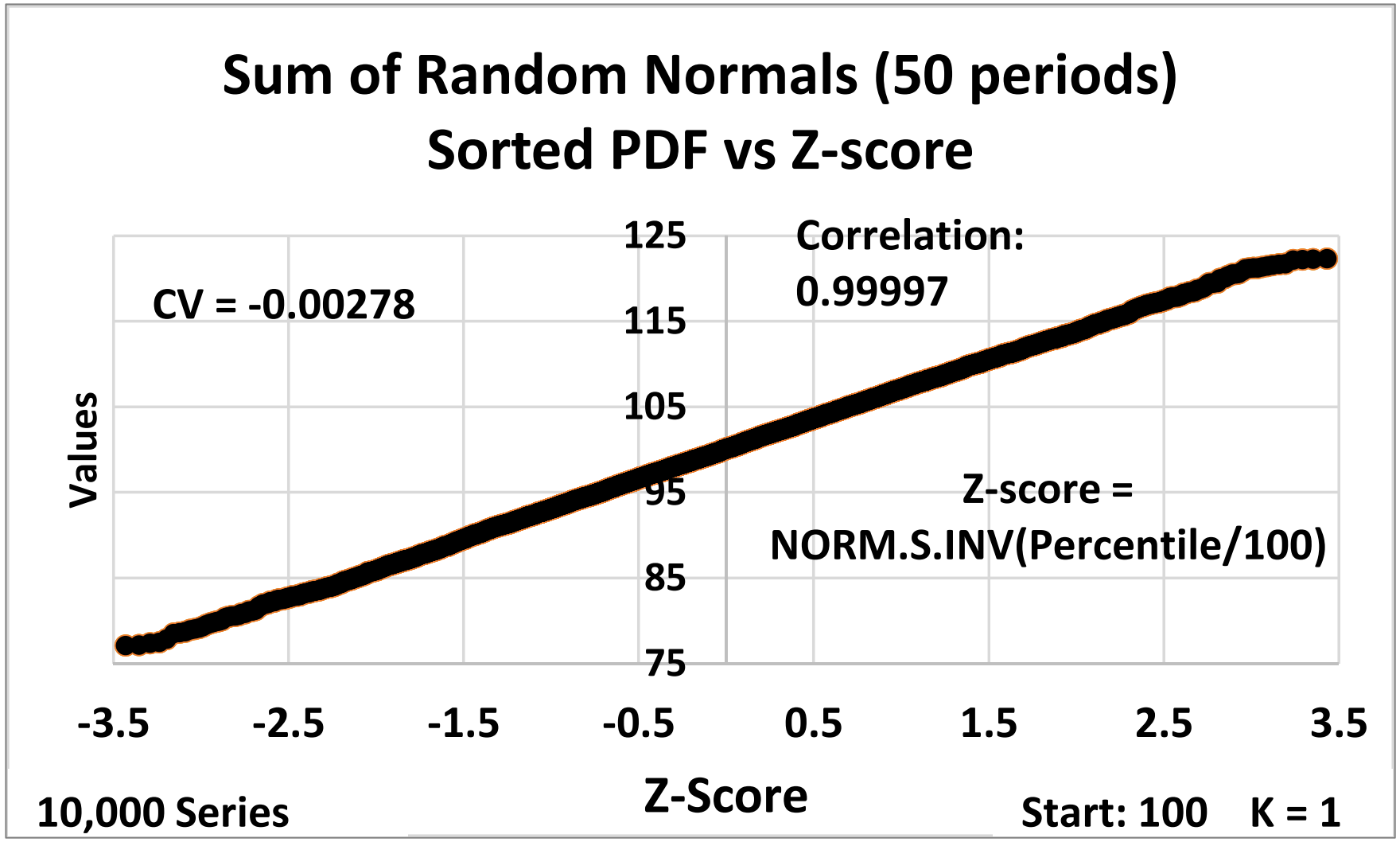
Period	1	4	9	16	25	36	49
Median	100.02	100.03	100.01	100.04	100.01	100.03	100.04
StdDev	1.00	2.00	3.02	3.16	3.31	3.45	3.60
PredictSD	1	2	3	4	5	6	7

$$SD(T) = K*SD(1)*Sqrt(T) \quad SD(1) = 1.$$

Test for Normality #1
QQ-Plot: Sorted PDF vs Percentile

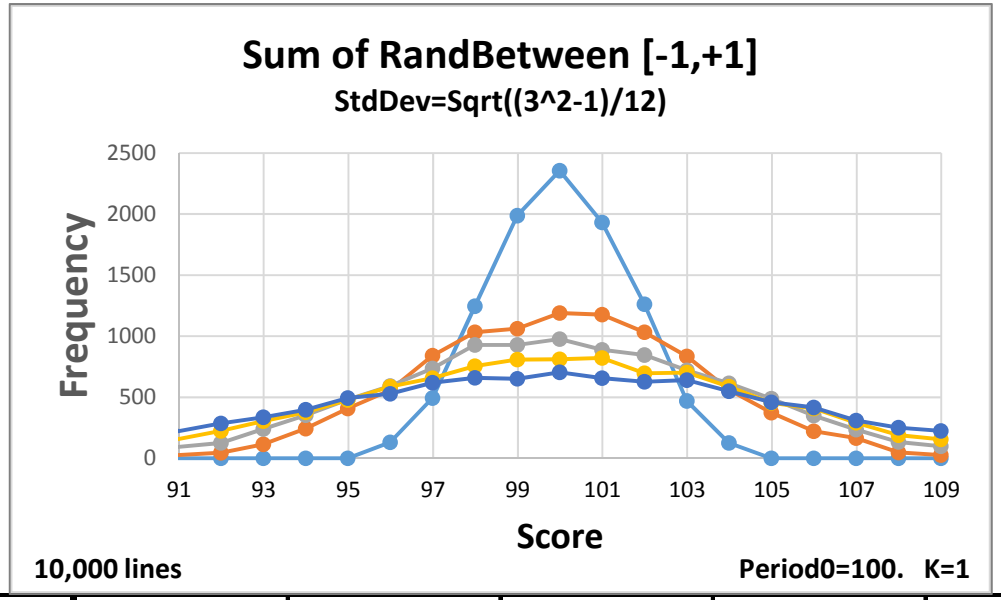


Test for Normality #2
QQ-Plot: Sorted PDF vs. Z-Score



1b Generating a Normal Dist.
Sum of Random Discretes

$$Period1 = Period0 + K*RandBetween(-1,+1)$$

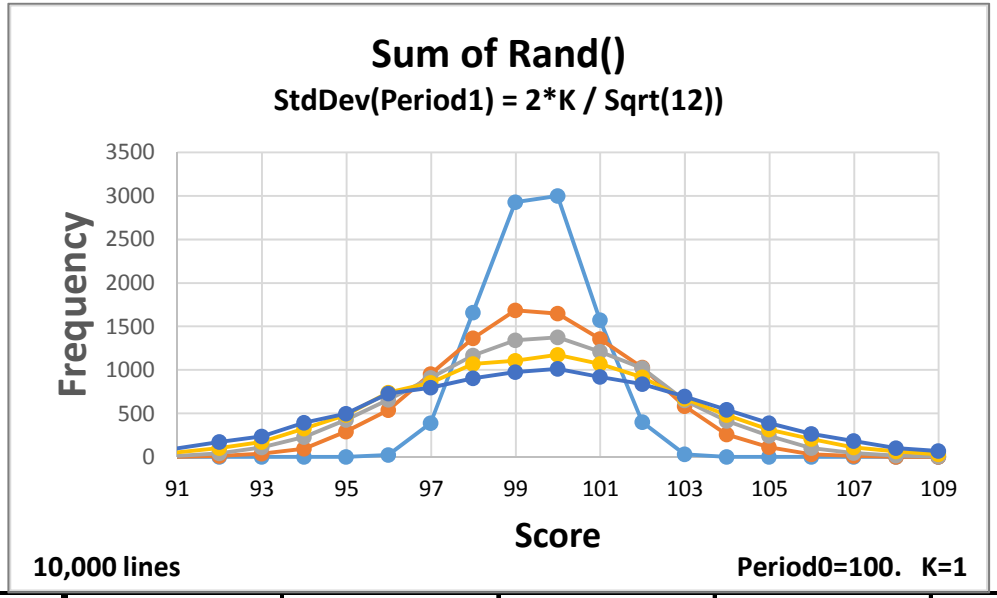


Period	1	4	9	16	25	36	49
Median	100.00	100.00	100.00	100.00	100.00	100.00	100.00
StdDev	0.815	1.64	2.45	3.27	4.10	4.90	5.72
PredictSD	0.815	1.63	2.45	3.26	4.08	4.89	5.71

$$SD(T) = K*SD(1)*Sqrt(T) \quad SD(1) = Sqrt(8/12)$$

1c Generating a Normal Dist.
Sum of Random Uniforms

$$Period1 = Period0 + K*[2*Rand()-1]$$

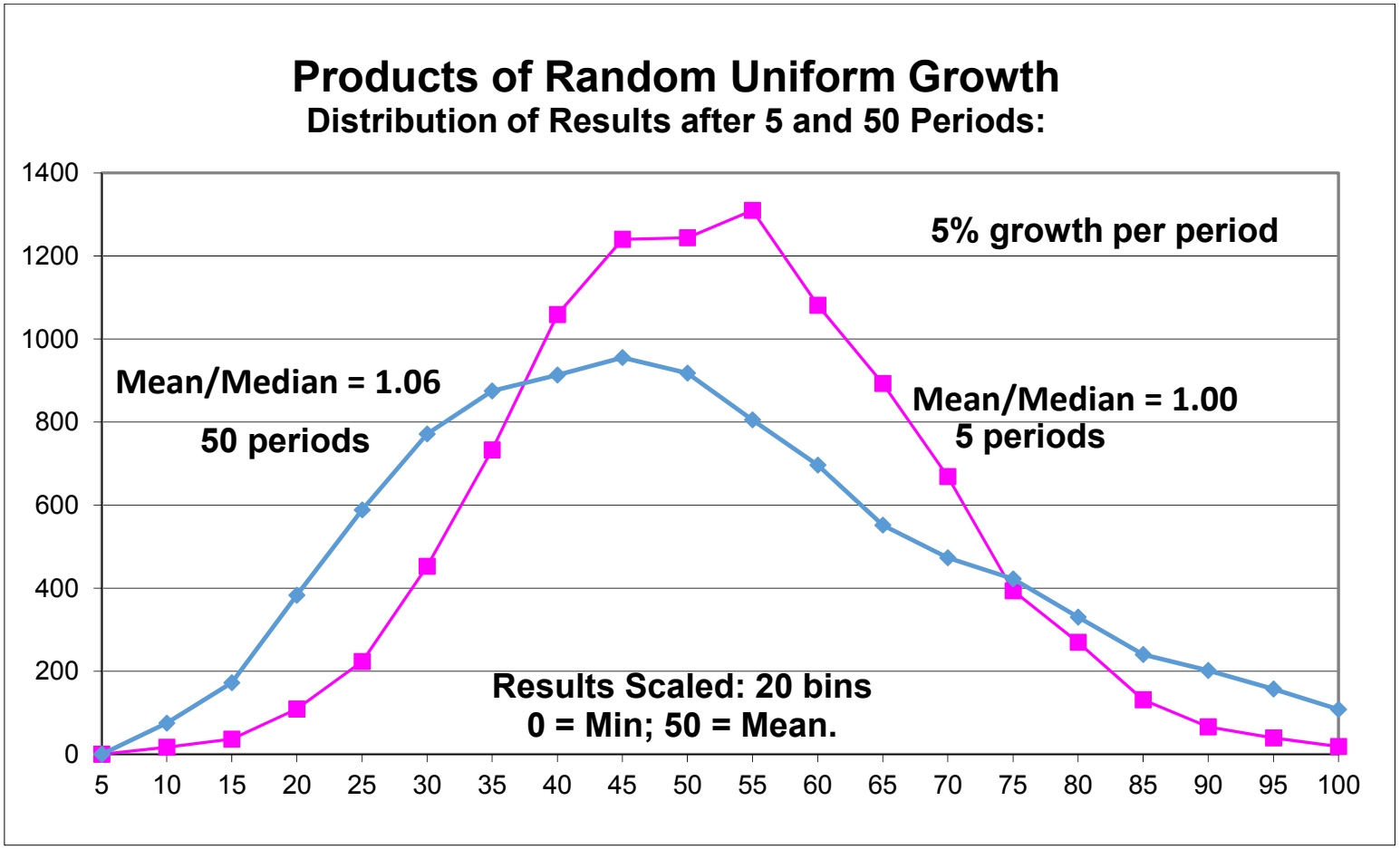


Period	1	4	9	16	25	36	49
Median	100.00	99.99	100.00	100.00	100.05	100.03	100.08
StdDev	0.576	1.17	1.75	2.32	2.87	3.43	3.98
PredictSD	0.577	1.15	1.73	2.31	2.89	3.46	4.04

$$SD(T) = K*SD(1)*Sqrt(T). \quad SD(1) = 2/Sqrt(12)$$

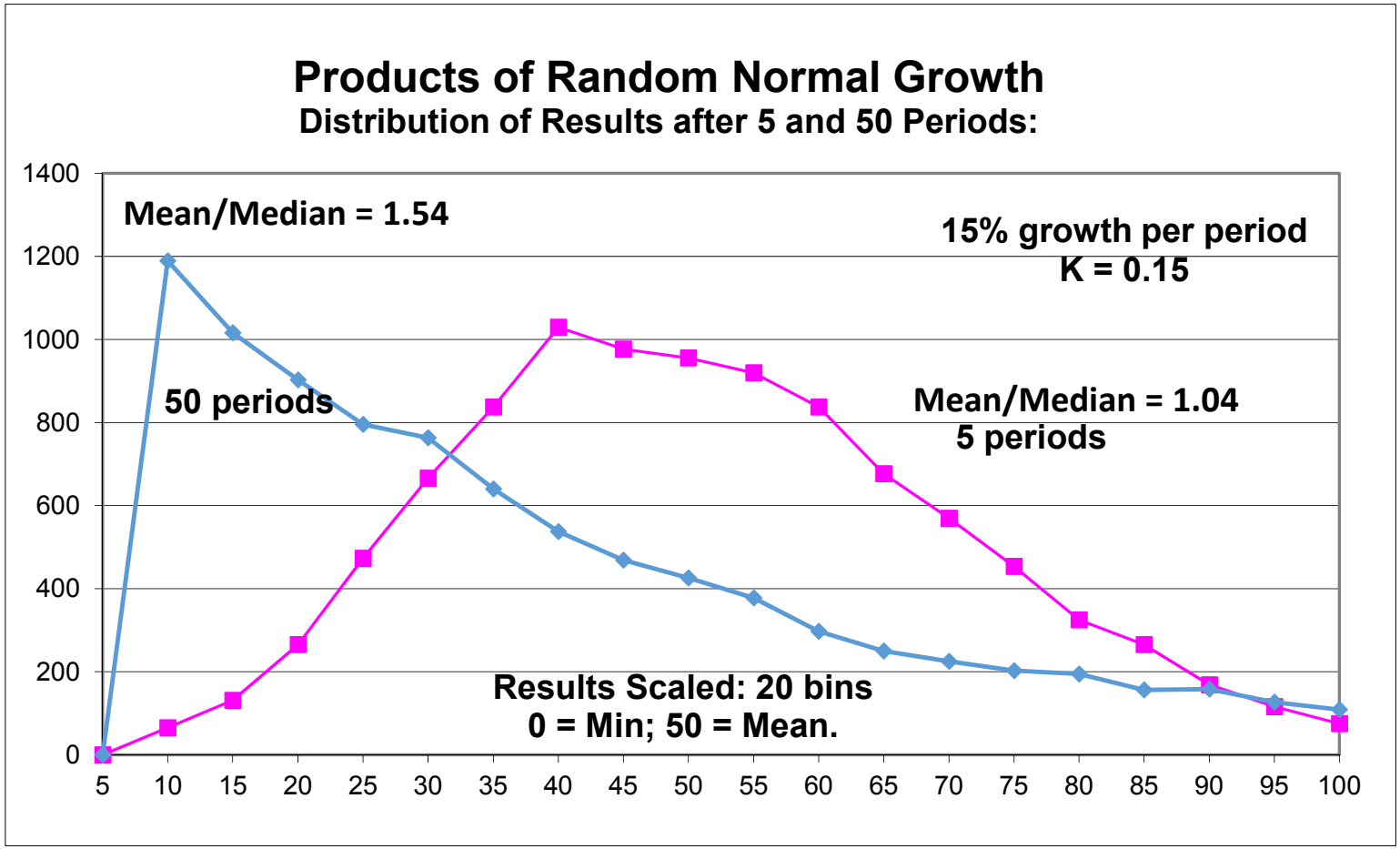
2a Generating a Log-Normal Dist.
Product: Random Normal Growth

$$P1 = P0*\{1+K*[1+NORMSINV(RAND()]\}$$



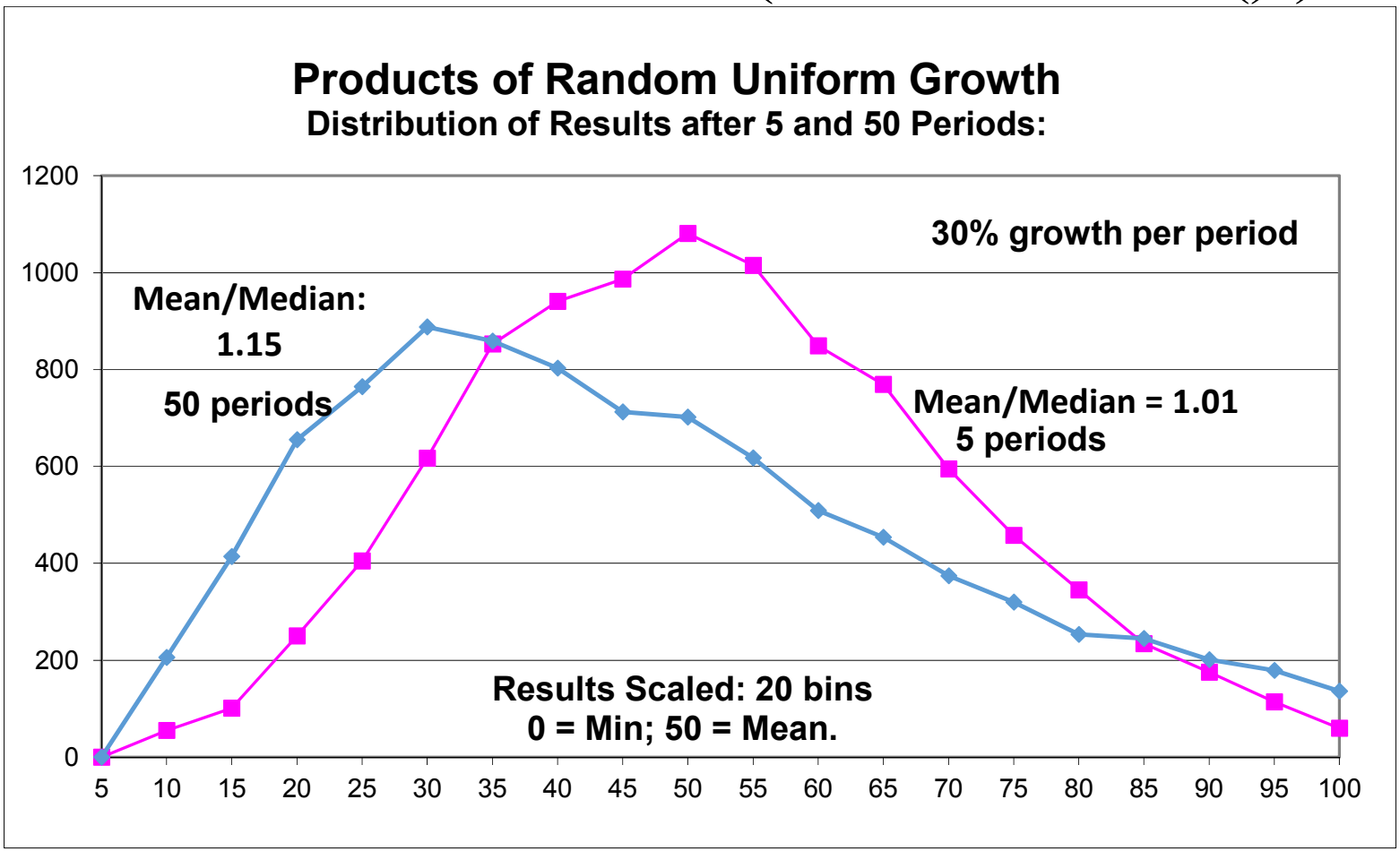
2a Generating a Log-Normal Dist.
Product: Random Normal Growth

$$P1 = P0*\{1+K*[1+NORMSINV(RAND()]\}$$



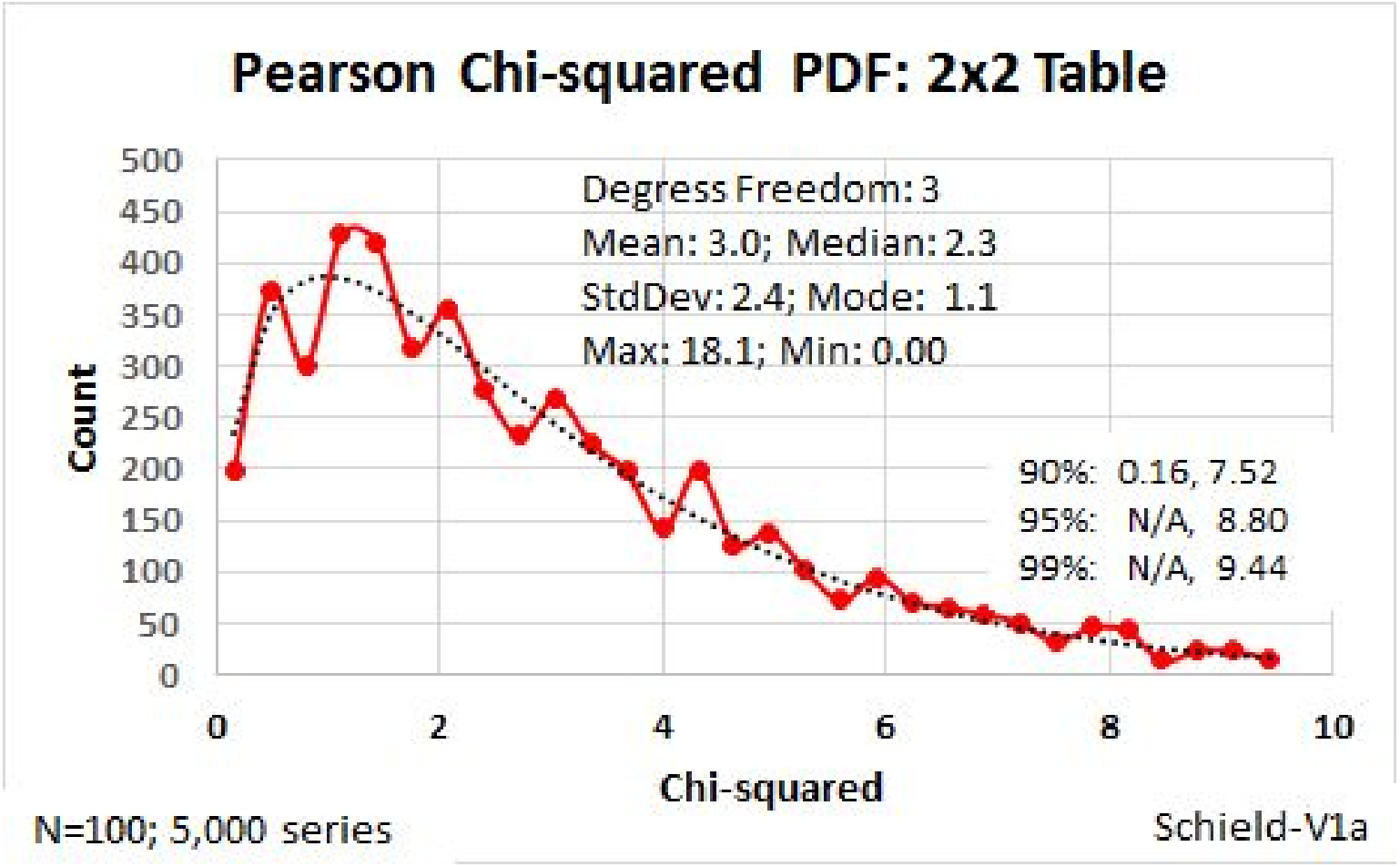
2b Generating a Log-Normal Dist.
Product: Random Uniform Growth

$$Period1 = Period0 * (1 + K*2*Rand())$$



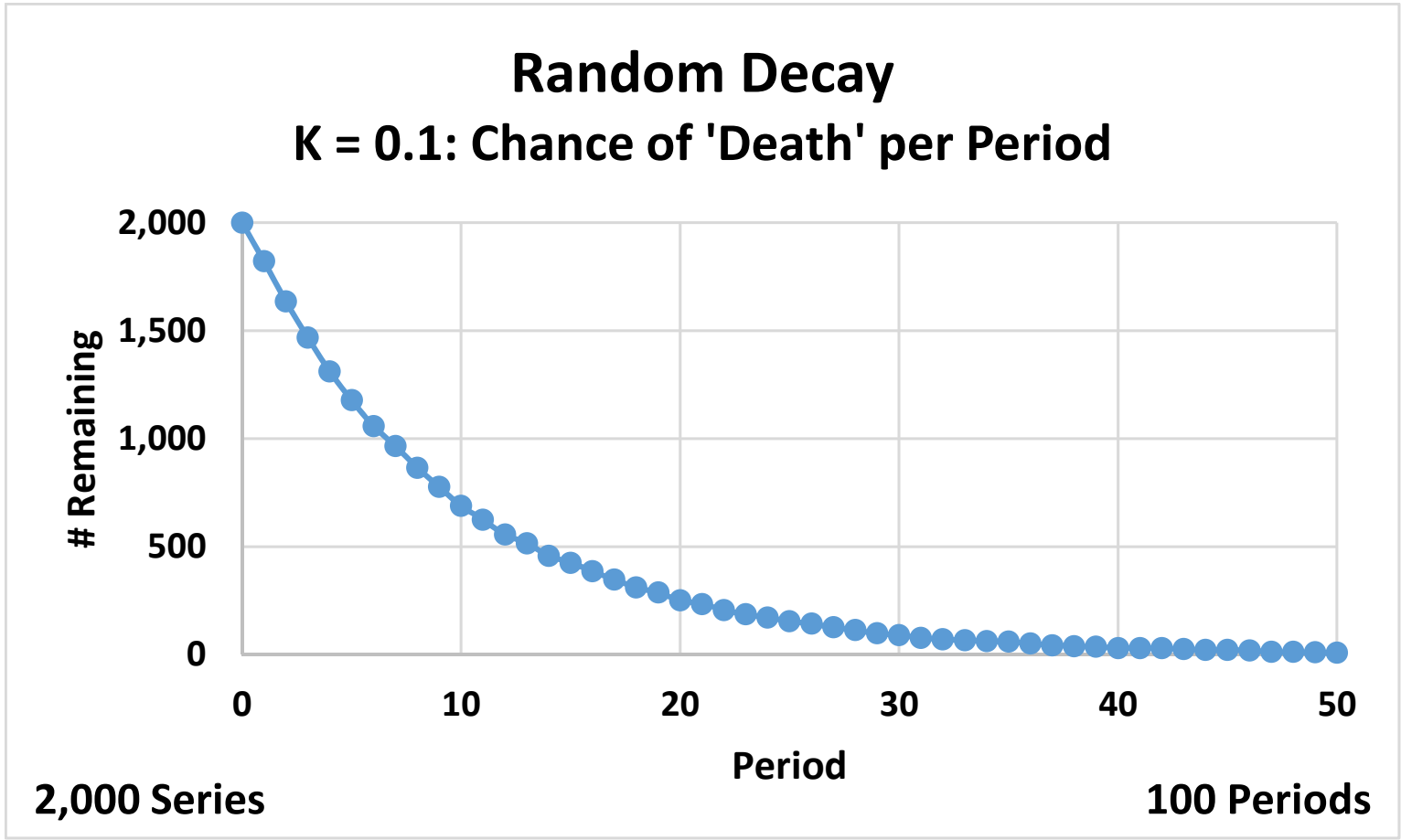
3 Generate Chi-Square Distribution: Random Assignment to Table Cells

X=RND() If(X>.75,4,If(X<.25,1,If(X>.5,3,2)))



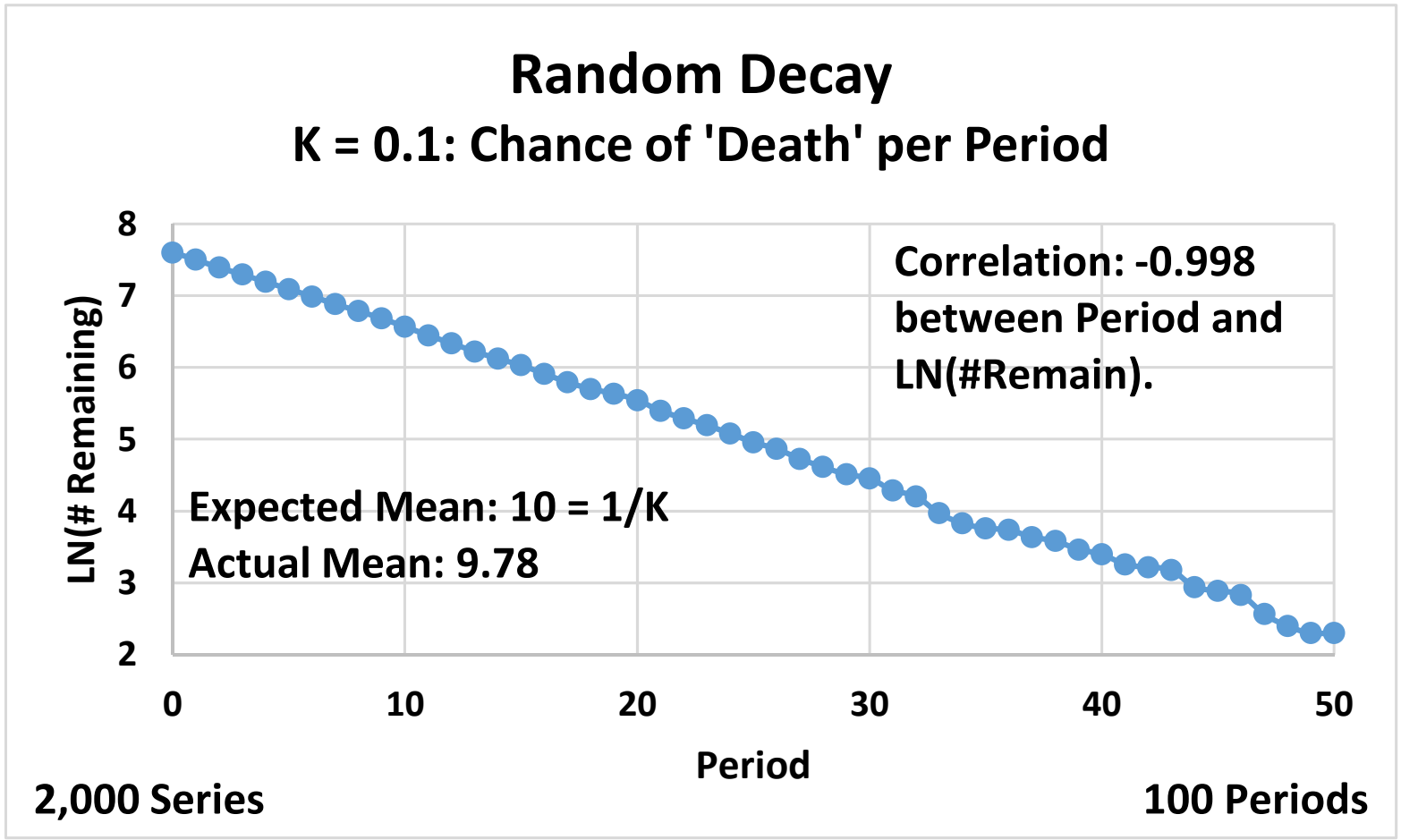
4 Generating an Exponential Dist.
Random "Death"

Chance of Death: =IF(RAND() < K, 1, "")



4 Generating an Exponential Dist. Random "Death"

Exponential Test: Is LN(#Remain) vs. time linear?



Conclusion

Knowing the process that generates a distribution is helpful to anyone *who wants to understand why* a given distribution is a good fit in a particular situation.

It also helps us *understand what determines* the parameters of a distribution (or changes therein).

Why have these increased over time?

- the standard deviation of heights?
- the mean/median ratio (the skewness) for incomes?
- the economic return from higher education?
- the average lifetime of humans?

Bibliography

Aitchison and Brown (1957, 1963). *The Log Normal Distribution with Special References to its uses in Economics*. Cambridge U.

Crow, E. and Shimizu, K. (1988). *Lognormal Distributions: Theory and Applications*. New York: Marcel Dekker, Inc., 1988.

International Futures (2014) Using Lognormal Income Distributions www.du.edu/ifs/help/understand/economy/poverty/lognormal.html

Schield, M. and T. Burnham (2008). Von Mises' Frequentist Approach to Probability, 2008 ASA Proceedings of the Section on Statistical Education. [CD-ROM] P. 2187-2194. www.StatLit.org/pdf/2008SchieldBurnhamASA.pdf.

Wikipedia/Log-normal distribution, Uniform distribution (continuous) and Uniform distribution (discrete)

Using Excel to Create Distributions Empirically

All of these files are in www.StatLit.org/Excel/

Generating Normal Distributions with Sums of Random Variables:
* 2014-Schield-Create-Normal-Empirically-Excel-Sum-Discrete.xlsx
* 2014-Schield-Create-Normal-Empirically-Excel-Sum-Uniform.xlsx
* 2014-Schield-Create-Normal-Empirically-Excel-Sum-Normal.xlsx

Creating Log-Normals with Product of Uniform Random Variables:
* 2014-Schield-Create-LogNormal-Empirically-Excel-Product-Uniforms.xlsx

Creating Chi-Square with Random Assignment to Table Cells:
* 2014-Schield-Create-ChiSquare-Empirically-Excel-DF3.xlsx

Generating Exponentials with Fixed Chance of 'Death" per period
* 2014-Schield-Create-Exponential-Empirically-Excel.xlsx