0F 2014 NNN4 Statistically-Significant Correlations 1

Statistically-Significant Correlations

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

Fall 2014
National Numeracy Network Conference
www.StatLit.org/pdf/2014-Schield-NNN4-Slides.pdf

0F 2014 NNN4 Statistically-Significant Correlations 2

Exact Solutions

For N random pairs from an uncorrelated bivariate normally-distributed distribution, the sampling distribution is not simple.

Here are three common analytic approaches:

1.Fisher transformation (using LN and Arctanh),

2.an exact solution (using a Gamma function), or

3.Student-t distribution: $t=r\text{Sqrt}[(n-2)/(1-r^2)]$; df=$n$-2

- For large n, the critical value of t (95% confidence) is 1.96.
- For small n, the critical value of t increases as n decreases.

None of these are simple or memorable.

0F 2014 NNN4 Statistically-Significant Correlations 3

Sufficient Condition

Approach: Find an equation generating a minimum correlation for statistical-significance given N.

1. Given N, find the smallest value of r where the left end of a 95% confidence interval is non-negative. Use calculator at www.vassarstats.net/rho.html or www.danielsoper.com/statcalc3/calc.aspx?id=44 For Daniels, use the results for a two-tailed test.
2. Generate correlation coefficient with simple model
3. Calculate error difference between calculated and exact using the exact as the standard. If all errors are positive, then the model is sufficient.

0F 2014 NNN4 Statistically-Significant Correlations 4

Simple Model: 2/SQRT(n)

	Minimum Correlation for Statistical Significance		
N	Exact	2/sqrt(n)	Error
400	0.10	0.10	3.0%
256	0.12	0.13	2.7%
100	0.20	0.20	2.0%
49	0.28	0.29	1.7%
25	0.40	0.40	1.3%
16	0.50	0.50	1.0%
12	0.57	0.58	0.6%
10	0.63	0.63	0.4%
7	0.75	0.76	0.4%
6	0.81	0.82	0.6%
5	0.88	0.89	1.4%
4	0.96	1.00	4.0%

All errors positive means the model is sufficient.

0F 2014 NNN4 Statistically-Significant Correlations 5

Solution

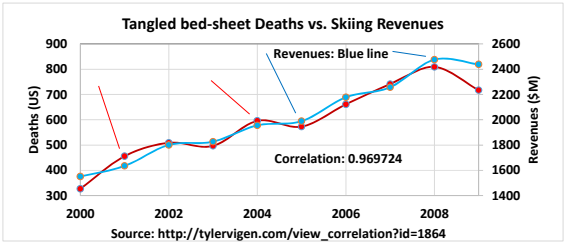
Minimum statistically-significant r = 2/Sqrt(n)
"n" is the number of pairs being correlated
Less than 5% over for n between 5 and 4,000.
Simple and memorable for two variables.

It is similar to the formula for the maximum 95% Margin of Error in samples from a binary variable:
95% ME = 1.96 Sqrt[p*(1-p)/n] < 2 Sqrt[1/(4n)]
95% ME < 1/Sqrt(n)
Simple and memorable for one binary variable.

0F 2014 NNN4 Statistically-Significant Correlations 6

Time-Series Correlations

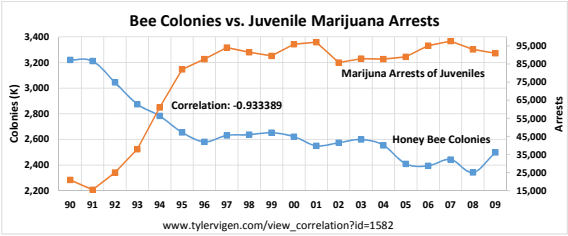
www.tylervigen.com



10 pairs; 2/Sqrt(10) = 0.63; Statistically significant

0F 2014 NNN4 Statistically-Significant Correlations 7

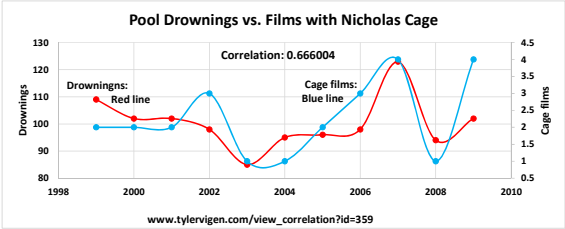
Correlation = -0.993
Bee colonies & MJ arrests



20 pairs; 2/Sqrt(20) = 0.45; Statistically-significant

0F 2014 NNN4 Statistically-Significant Correlations 8

Correlation = 0.664
Drownings & Cage films



11pairs; 2/Sqrt(11) = 0.60; Statistically-significant!

0F 2014 NNN4 Statistically-Significant Correlations 9

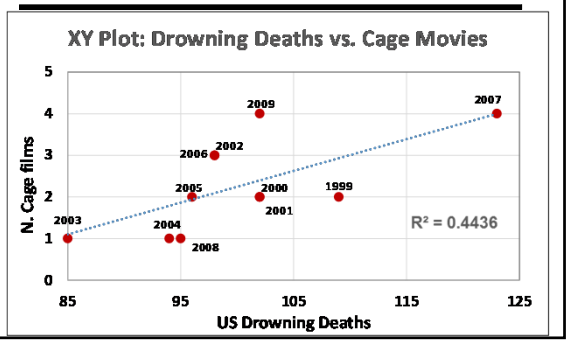
Something Seems Wrong!

1. There is nothing linear about these associations.
2. These correlations seem unbelievably high.

#1: The correlation between two time-series eliminates the common factor: time. The question is whether their mutual association is linear. To see this, an XY-plot is generated.

0F 2014 NNN4 Statistically-Significant Correlations 10

Correlation = 0.664
Drownings & Cage films



0F 2014 NNN4 Statistically-Significant Correlations 11

#2: Very High Correlations. Three Explanations

1. *Association is causal*. See Tyler Vigen's video: www.youtube.com/watch?feature=player_embedded&v=g-g0ovHjQxs
2. *Association is spurious – just random chance*. Five percent of random associations will be mistakenly classified as statistically significant.
3. *Association is cherry-picked -- after the fact*. According to Tyler, "This server has generated 24,470 correlations." Tyler just picked those with high or interesting correlations.

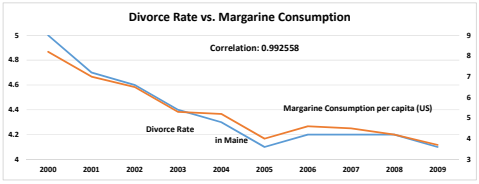
0F 2014 NNN4 Statistically-Significant Correlations 12

Conclusions

1. Use 2/Sqrt(n) as the minimum correlation for statistical significance. This criteria is sufficient, fairly accurate (within 5%) and memorable.
2. The correlation between two time-series eliminates time. Correlation determines the degree of linearity in their cross-sectional association.
3. Do not use a test for statistical significance if the data pairs were selected – after the fact via data mining – solely because of their high correlation.

0F 2014 NNN4 Statistically-Significant Correlations 13

Correlation = 0.993
Divorce & Margarine Usage



Correlation: 0.993. N=10, SS_Rho = 1/sqrt(11) =

10 pairs; 2/Sqrt(10) = 0.64. Statistically-significant

www.tylervigen.com/view_correlation?id=1703

Statistically-Significant Correlations

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

Fall 2014
National Numeracy Network Conference

www.StatLit.org/pdf/2014-Schield-NNN4-Slides.pdf

Exact Solutions

For N random pairs from an uncorrelated bivariate normally-distributed distribution, the sampling distribution is not simple.

Here are three common analytic approaches:

1. Fisher transformation (using LN and Arctanh),
2. an exact solution (using a Gamma function), or
3. Student-t distribution: $t = r\,\mathrm{Sqrt}[(n-2)/(1-r^2)]$; $df = n-2$
   - For large n, the critical value of t (95% confidence) is 1.96.
   - For small n, the critical value of t increases as n decreases.

None of these are simple or memorable.

Sufficient Condition

Approach: Find an equation generating a minimum correlation for statistical-significance given N.

1. Given N, find the smallest value of r where the left end of a 95% confidence interval is non-negative. Use calculator at www.vassarstats.net/rho.html or www.danielsoper.com/statcalc3/calc.aspx?id=44 For Daniels, use the results for a two-tailed test.
2. Generate correlation coefficient with simple model
3. Calculate error difference between calculated and exact using the exact as the standard. If all errors are positive, then the model is sufficient.

Simple Model: 2/SQRT(n)

	Minimum Correlation for Statistical Significance				
N		Exact	2/sqrt(n)		Error
400		0.10	0.10		3.0%
256		0.12	0.13		2.7%
100		0.20	0.20		2.0%
49		0.28	0.29		1.7%
25		0.40	0.40		1.3%
16		0.50	0.50		1.0%
12		0.57	0.58		0.6%
10		0.63	0.63		0.4%
7		0.75	0.76		0.4%
6		0.81	0.82		0.6%
5		0.88	0.89		1.4%
4		0.96	1.00		4.0%

All errors positive means the model is sufficient.

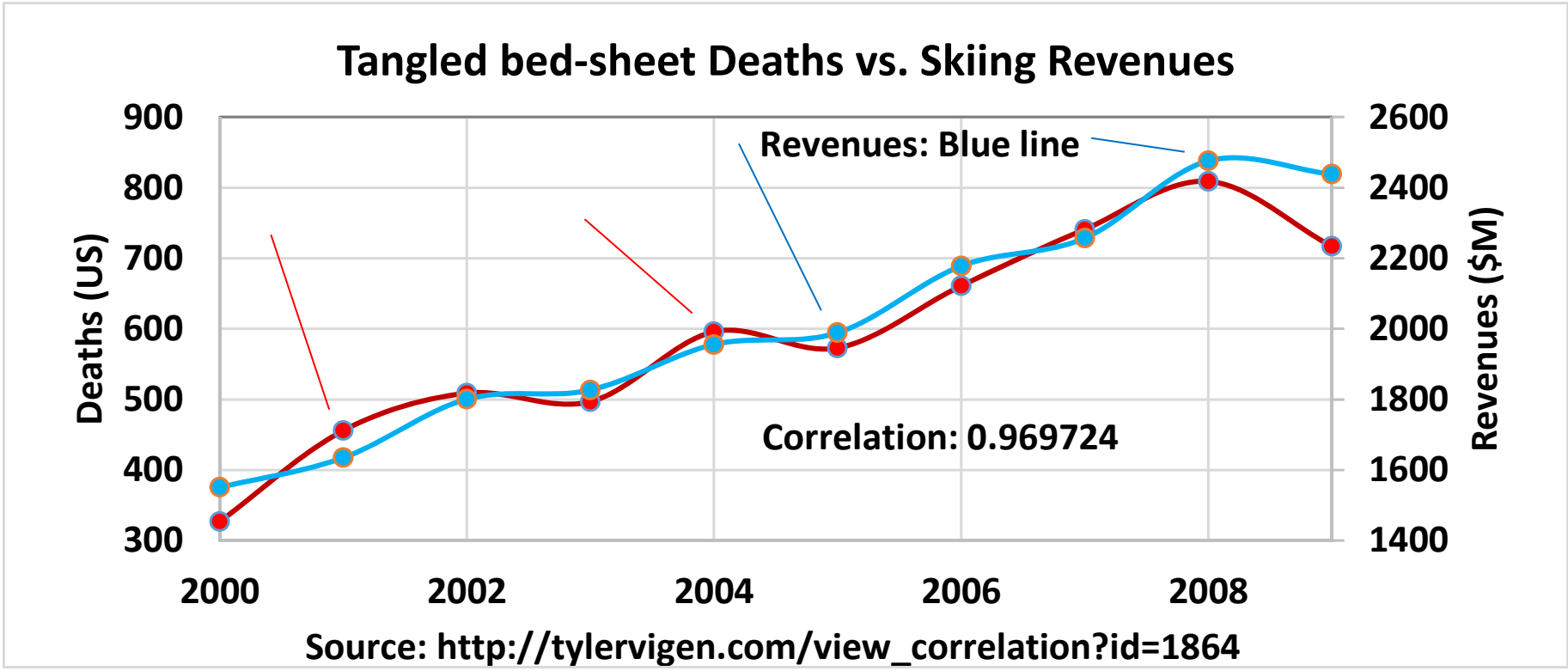
Solution

Minimum statistically-significant r = 2/Sqrt(n)
"n" is the number of pairs being correlated
Less than 5% over for n between 5 and 4,000.
Simple and memorable for two variables.

It is similar to the formula for the maximum 95% Margin of Error in samples from a binary variable:
95% ME = 1.96 Sqrt[p*(1-p)/n] < 2 Sqrt[1/(4n)]
95% ME < 1/Sqrt(n)
Simple and memorable for one binary variable.

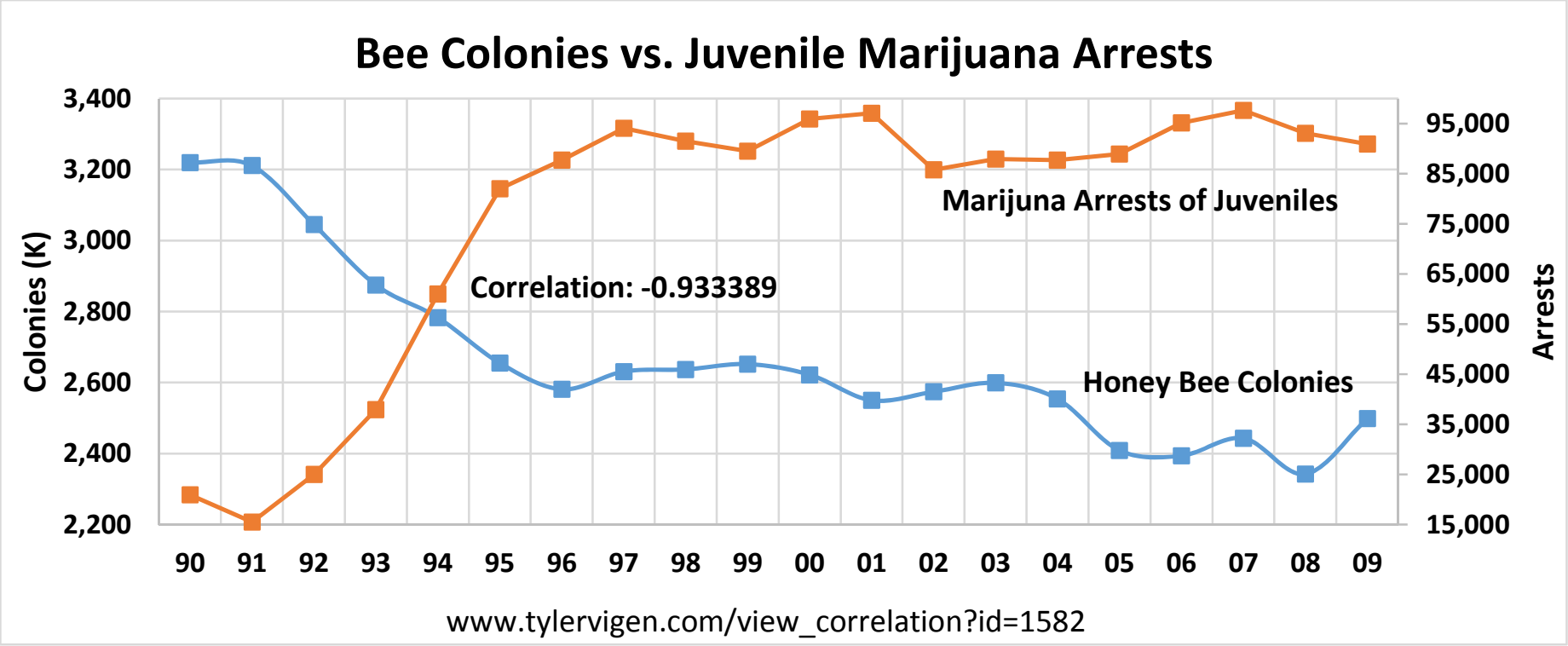
Time-Series Correlations

www.tylervigen.com



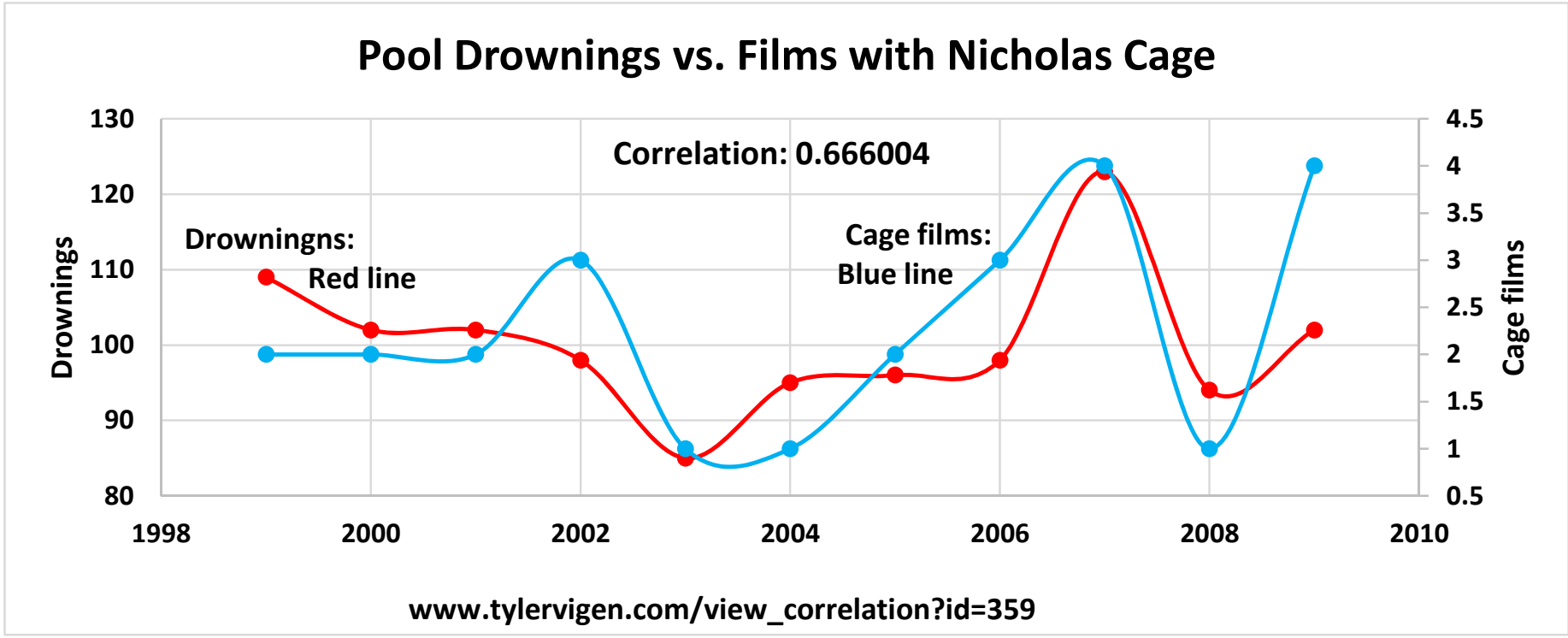
10 pairs; 2/Sqrt(10) = 0.63; Statistically significant

Correlation = -0.993
Bee colonies & MJ arrests



20 pairs; 2/Sqrt(20) = 0.45; Statistically-significant

Correlation = 0.664
Drownings & Cage films



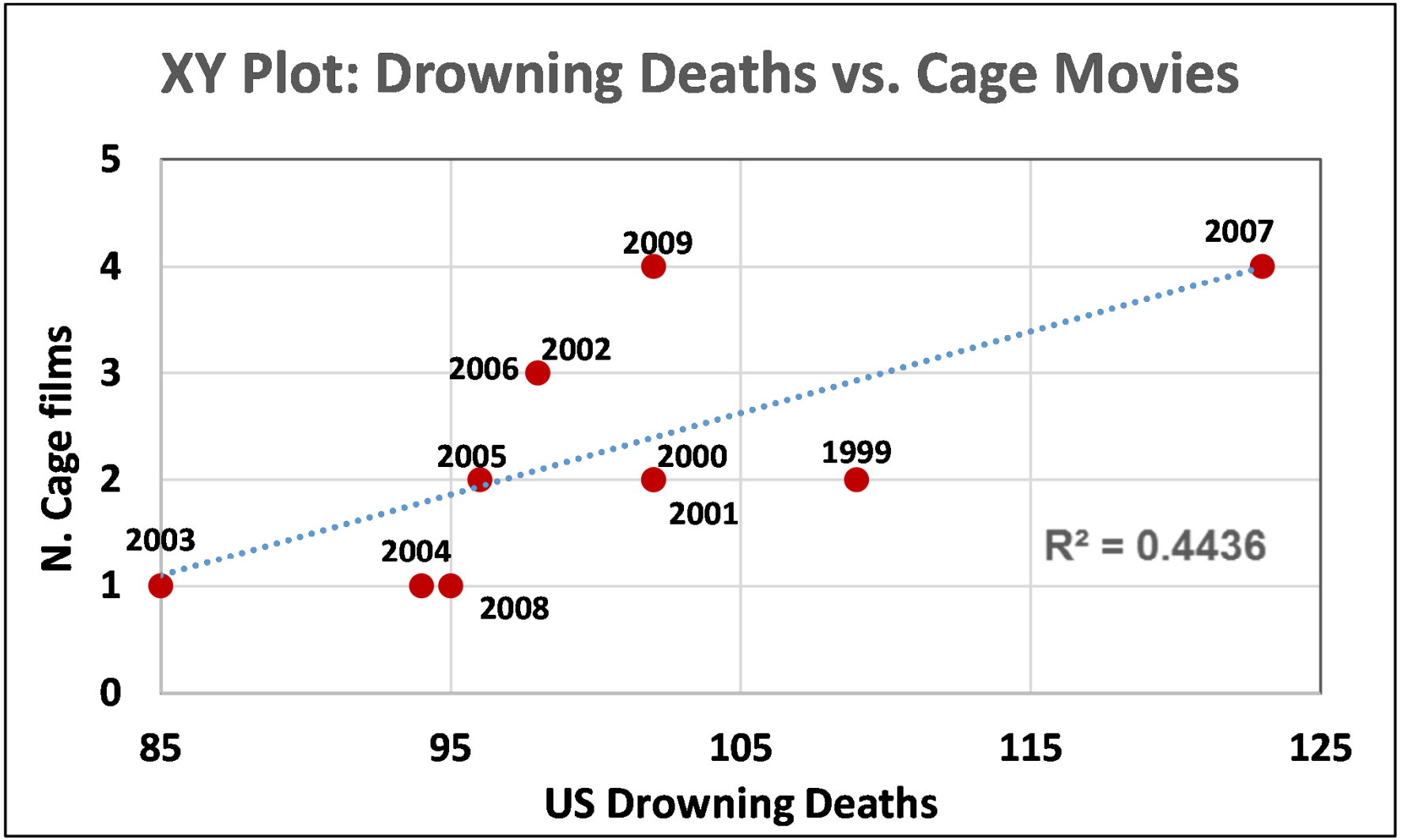
11pairs; 2/Sqrt(11) = 0.60; Statistically-significant!

Something Seems Wrong!

1. There is nothing linear about these associations.
2. These correlations seem unbelievably high.

#1: The correlation between two time-series eliminates the common factor: time. The question is whether their mutual association is linear. To see this, an XY-plot is generated.

Correlation = 0.664
Drownings & Cage films



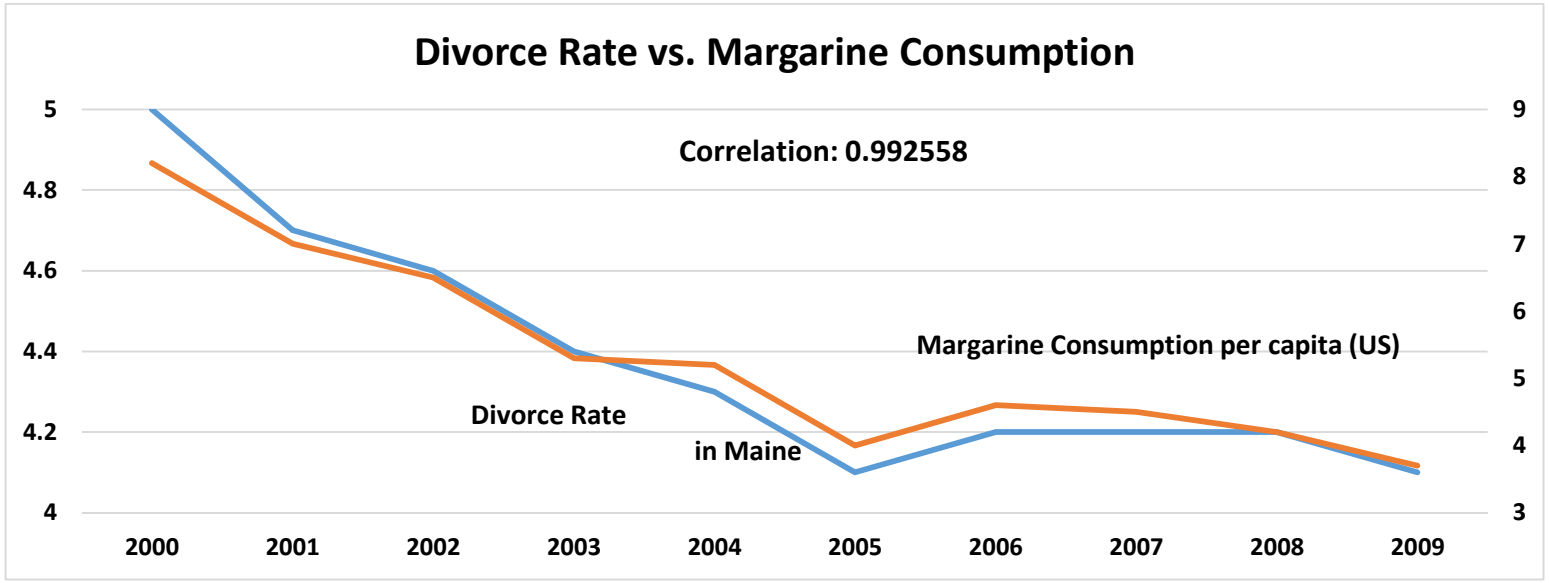
#2: Very High Correlations. Three Explanations

1. *Association is causal*. See Tyler Vigen's video: www.youtube.com/watch?feature=player_embedded&v=g-g0ovHjQxs
2. *Association is spurious – just random chance*. Five percent of random associations will be mistakenly classified as statistically significant.
3. *Association is cherry-picked -- after the fact*. According to Tyler, "This server has generated 24,470 correlations." Tyler just picked those with high or interesting correlations.

Conclusions

1. Use 2/Sqrt(n) as the minimum correlation for statistical significance. This criteria is sufficient, fairly accurate (within 5%) and memorable.
2. The correlation between two time-series eliminates time. Correlation determines the degree of linearity in their cross-sectional association.
3. Do not use a test for statistical significance if the data pairs were selected – after the fact via data mining – solely because of their high correlation.

Correlation = 0.993
Divorce & Margarine Usage



Correlation: 0.993. N=10, SS_Rho = 1/sqrt(11) =

10 pairs; 2/Sqrt(10) = 0.64. Statistically-significant

www.tylervigen.com/view_correlation?id=1703