Bootstrap program help

Paul Barrett
Advanced Projects R&D Ltd.
Auckland
New Zealand

email: paul@pbarrett.net
Web: www.pbarrett.net

August 27th, 2010

Table of Contents

Part I

1 Introduction

This program enables a user to determine the likelihood of obtaining an agreement coefficient, as high or higher than that observed, by chance alone. The agreement indices currently available for analysis are the Gower index, the Kernel Smoothed Distance (KSD-s),and the double-scaled euclidean similarity (DSE-s) index.

It does this by creating samples of random data, with a desired number of cases per sample, corresponding to the particular measurement range of the data from within which the original coefficient was calculated. For each sample, a coefficient is calculated. The number of samples should be sufficient to provide an empirical sampling distribution of coefficients, in order that the estimated occurrence probability of the target coefficient is robust. Anywhere between 5,000 to 20,000 samples is usually sufficient to achieve this (giving you a frequency distribution comprising 5,000 or 20,000 coefficients).

Normally, I tend to resample first using 1000 or so samples to get a ball-park estimate, then I use 20,000 samples as beyond this quantity little seems to change. However, if you need precision for the estimation of likely probability, you may go as high as 50,000 samples. However, remember that the time it takes to generate the samples is a function of sample size and number of samples to be generated.

Some timings (on an I7Extreme 920XM Intel quad-core machine, 2GHz, 8Gb 1333 MHz DDR3 RAM with 1333MHz bus, running Windows-7 64-bit Ultimate, but with Aero "transparency" effects turned off). Actually, those Windows Aero effects produce a huge drag on timings, especially when a program (like this one) updates an on-screen processing status component .. it's the difference between a 12-second and 45-second completion time on a 20,000 sample job.

A single Gower coefficient **(in seconds)**
For a job with 100 cases per sample, sampling 1000 times: 1
For a job with 100 cases per sample, sampling 10000 times: 6.5
For a job with 100 cases per sample, sampling 20,000 times: 11
For a job with 1000 cases per sample, sampling 20,000 times: 17

Difference between two KSD-s coefficients
For a job with 100 cases per sample, sampling 1000 times: 1
For a job with 100 cases per sample, sampling 10000 times: 7
For a job with 100 cases per sample, sampling 20,000 times: 13.5
For a job with 1000 cases per sample, sampling 20,000 times: 20

So, even on a slower machine the bootstrapping will be relatively efficient.

This is the essence of "bootstrapping", generating samples of data from some specified statistical distribution or the sample data themselves (taking a subset of the dataset each time), and constructing a frequency distribution of coefficients found from the samples so that you can determine how likely it is to have observed your 'target' coefficient, by chance alone.

Instead of creating artificial null or alternative hypotheses, we simply determine the likelihood of a coefficient occurring if we were to calculate it from random data. We don't need to assume a sampling distribution for the target coefficient as we construct it empirically, tailored to the particular sample size, measurement range, and type of data (integer or real-valued numbers) we have at hand.

We can also take the same approach if we want to determine whether the magnitude of the difference between two coefficients could have occurred by chance alone. To do this, we specify the two coefficients we have in mind, and their respective sampling details. We then generate say 10,000 samples for each constrained-specified coefficient, taking the difference between them for each pair of samples. This "difference distribution" forms the frequency distribution against which we compare our observed difference. Again, the question we wish to answer is "what is the likelihood of observing a difference between two coefficients as high or higher than that observed, by chance alone?".

So, resampling and bootstrapping are powerful tools, enabling 'significance tests' to be undertaken without relying upon assumed sampling distributions or assumed population parameters.

Paul Barrett, *(paul@pbarrett.net)*

1.1 Data Generation

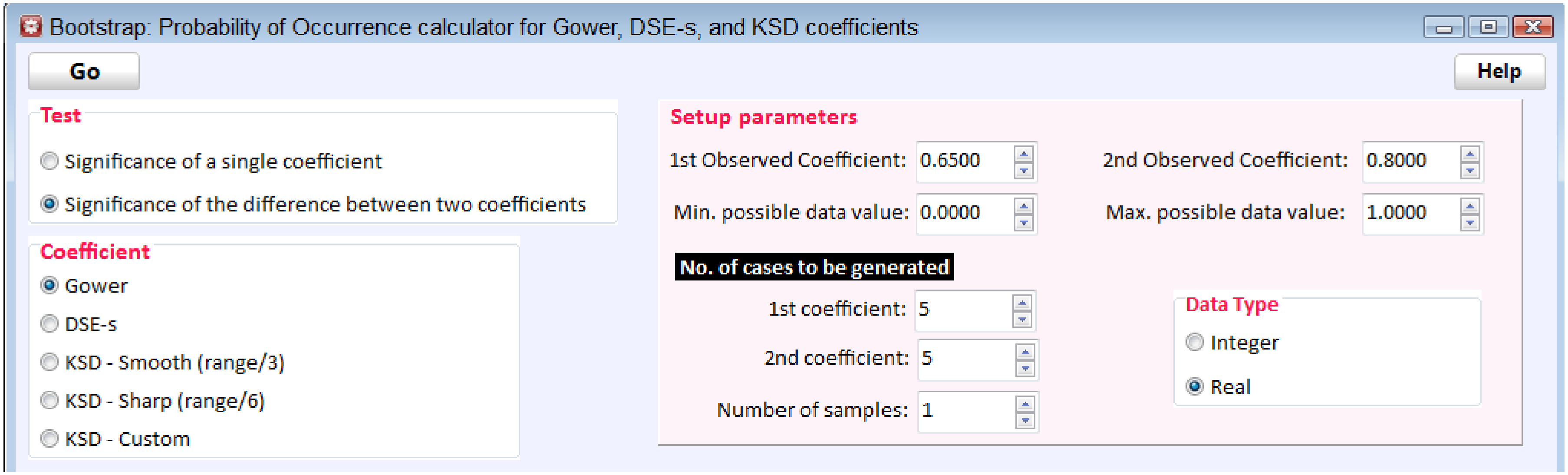
The data is generated sampling from a uniform distribution, where every value between and including the minimum and maximum possible values possesses an equal probability of selection.

Data generation can be constrained to integers (whole numbers like 1, 3, 7, 21) or reals (decimal-fraction numbers like 12.4568 or 1.02355). It matters because you are trying to generate data which is within the same range and type as your target coefficient data, and so whether the data is integer or real does matter.

When comparing the difference between two coefficients, the program assumes:

1. The coefficients are the same class (both Gower, both KSD-s etc).
2. The minimum and maximum metric range for each coefficient is equivalent.
3. The number of cases for each coefficient may be different.

E.g.



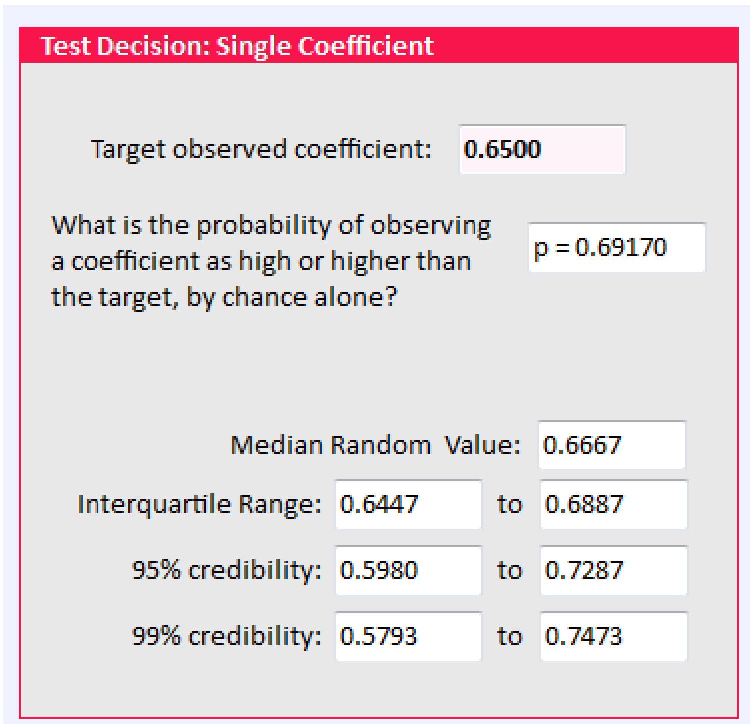
1.2 Statistical significance

In bootstrapping, "statistical significance" is interpreted as "the probability of an event/coefficient of a certain magnitude" occurring by chance alone. We might also express this as "the probability of observing a coefficient as high or higher than the one you observed", or equally as "the probability of observing a coefficient as low or lower than the one you observed".

Bottom line, you are generating a frequency distribution of coefficients calculated using purely random data, on the basis that if your target coefficient is higher than most/all of those generated, then you have evidence to indicate that your observed coefficient is unlikely to have been found by chance alone.

The fundamental logic is the same as for conventional statistical inference, EXCEPT that bootstrapping makes no assumptions about the distribution of the data or coefficients and no assumptions about hypothetical null or alternative "population" values.

The program reports its results as:



What this tells us is that a value of 0.65 occurred by chance alone in 69% of our samples ... that says our observed coefficient is not a very convincing demonstration of a systematic phenomenon.

The section on "How to use the program" explains this in much more detail, and shows the kinds of diagnostics available to you to augment the simple decision-statement.

1.3 Program Constraints

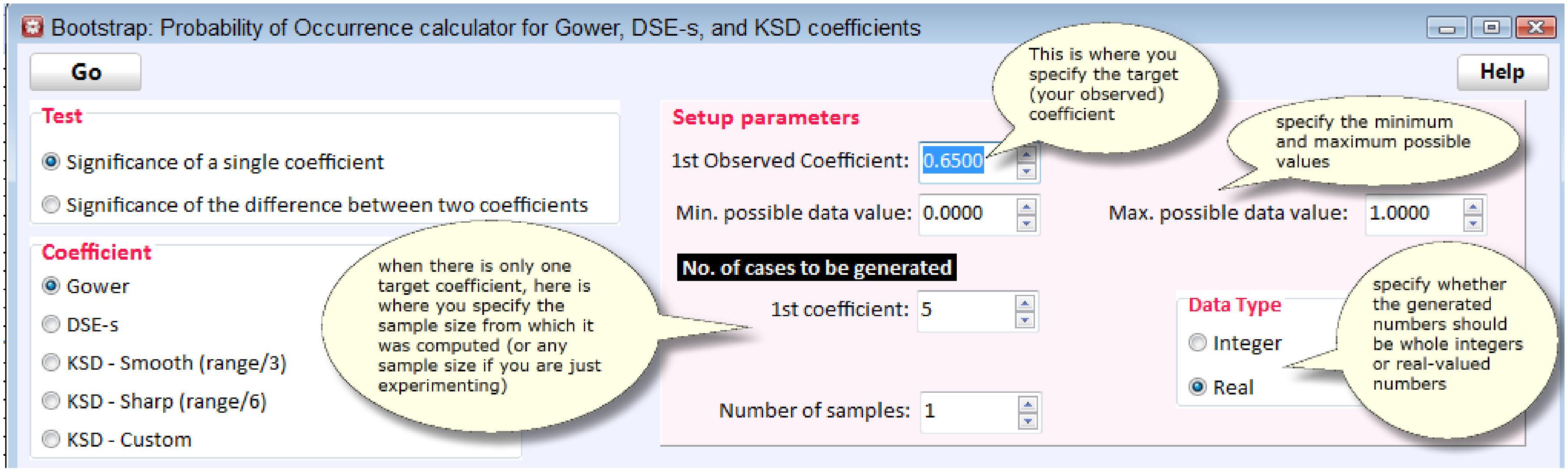
Number of cases ... between 5 and 60,000
Number of Samples ... between 1 and 60,000

Part II

2 How to Use the Program

2.1 Significance of a single coefficient

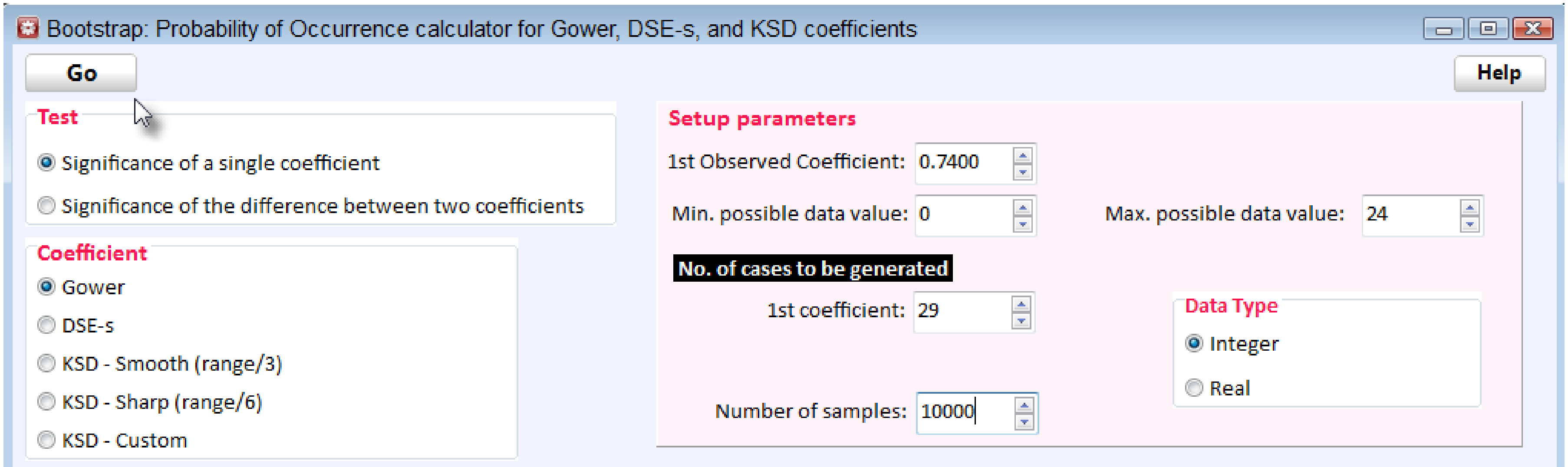
Click on the program icon to show:



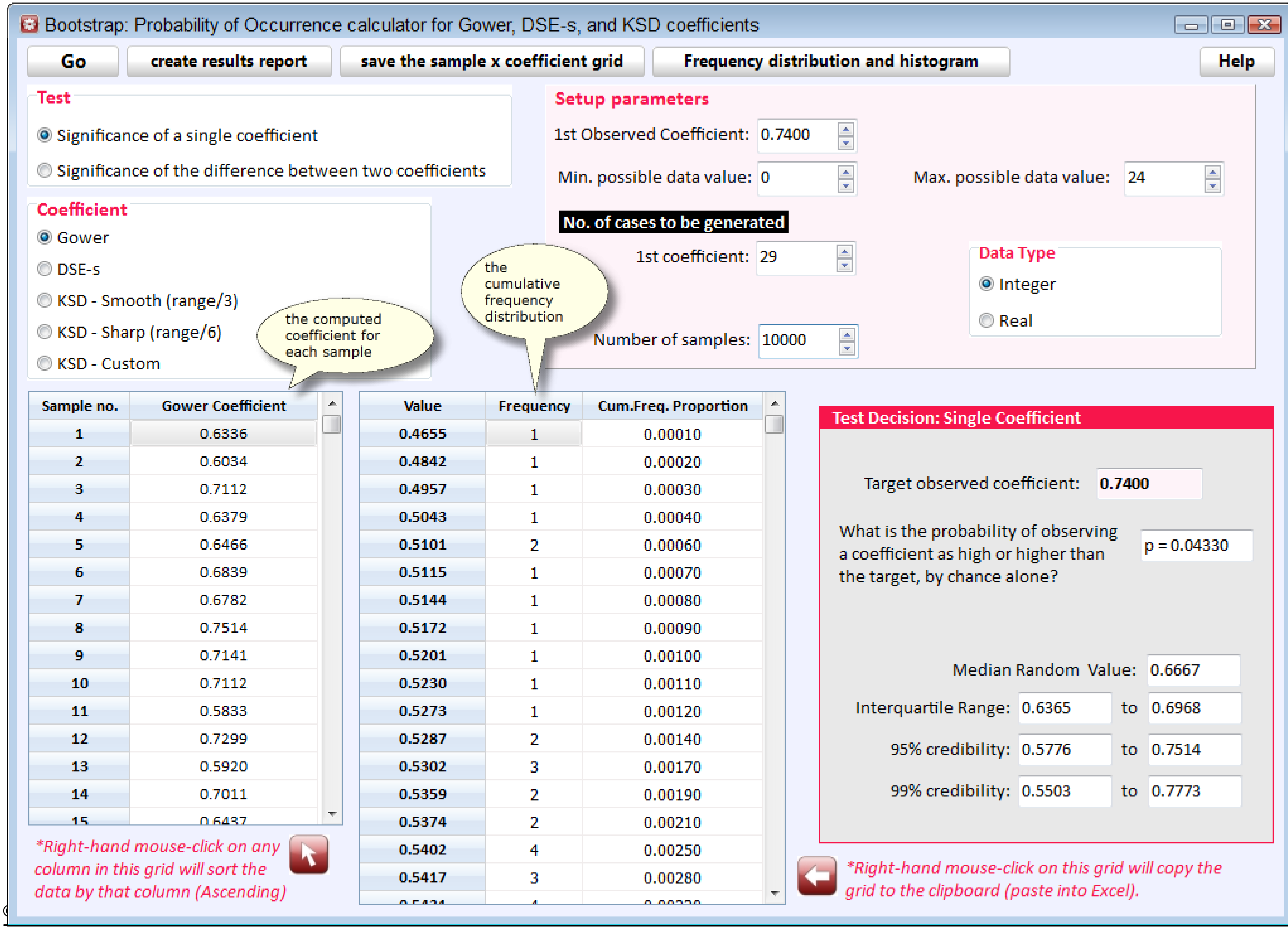
Just select the constraints you wish to use - then click on the "go" button.

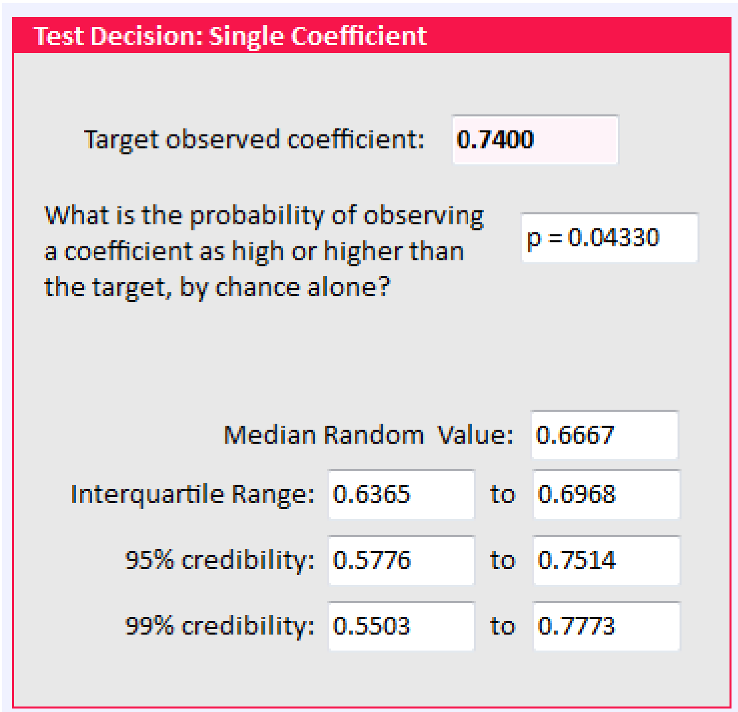
A process status bar will appear, indicating the % completion of the task, then the results will be displayed.

Let's say we observed a Gower coefficient of 0.74, in a sample of 29 cases, where the measurement range was between 0 and 24, integer numbers. We want to generate 10,000 samples of random data, constrained by those setup specifications ...

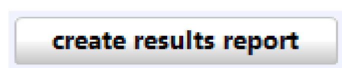


clicking on Go produces ...


Bootstrap: Probability of Occurrence calculator for Gower, DSE-s, and KSD coefficients
Go
create results report
save the sample x coefficient grid
Frequency distribution and histogram
Help
Test
Significance of a single coefficient
Significance of the difference between two coefficients
Coefficient
Gower
DSE-s
KSD - Smooth (range/3)
KSD - Sharp (range/6)
KSD - Custom
the computed coefficient for each sample
the cumulative frequency distribution
Setup parameters
1st Observed Coefficient: 0.7400
Min. possible data value: 0
Max. possible data value: 24
No. of cases to be generated
1st coefficient: 29
Number of samples: 10000
Data Type
Integer
Real
Sample no. | Gower Coefficient
1 | 0.6336
2 | 0.6034
3 | 0.7112
4 | 0.6379
5 | 0.6466
6 | 0.6839
7 | 0.6782
8 | 0.7514
9 | 0.7141
10 | 0.7112
11 | 0.5833
12 | 0.7299
13 | 0.5920
14 | 0.7011
15 | 0.6437
Value | Frequency | Cum.Freq. Proportion
0.4655 | 1 | 0.00010
0.4842 | 1 | 0.00020
0.4957 | 1 | 0.00030
0.5043 | 1 | 0.00040
0.5101 | 2 | 0.00060
0.5115 | 1 | 0.00070
0.5144 | 1 | 0.00080
0.5172 | 1 | 0.00090
0.5201 | 1 | 0.00100
0.5230 | 1 | 0.00110
0.5273 | 1 | 0.00120
0.5287 | 2 | 0.00140
0.5302 | 3 | 0.00170
0.5359 | 2 | 0.00190
0.5374 | 2 | 0.00210
0.5402 | 4 | 0.00250
0.5417 | 3 | 0.00280
Test Decision: Single Coefficient
Target observed coefficient: 0.7400
What is the probability of observing a coefficient as high or higher than the target, by chance alone?
p = 0.04330
Median Random Value: 0.6667
Interquartile Range: 0.6365 to 0.6968
95% credibility: 0.5776 to 0.7514
99% credibility: 0.5503 to 0.7773
*Right-hand mouse-click on any column in this grid will sort the data by that column (Ascending)
*Right-hand mouse-click on this grid will copy the grid to the clipboard (paste into Excel).

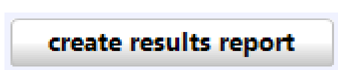


The 95% and 99% credibility intervals are those values in-between which 95% and 99% of all generated values lie. These are semi-equivalent to a confidence interval except instead of showing the values within which say 95% of all confidence intervals which contain the estimated population parameter might be observed, I'm showing the interval which actually contains 95% or 99% of expected values when using random data.

A printable summary report can be obtained by clicking on the 

button.

For example, calculating the bootstrap significance for a coefficient of 0.65 from a sample of 31 cases, whose values vary between 0 and 1 (real values) ...

and then clicking on the 

button shows:

Bootstrap Report

Print

Report Date: *Friday 27, August, 2010 - 1:57 pm*

Agreement Coefficient Bootstrap

Test = Single Gower coefficient - Real-valued data

Number of cases in each sample: 31 Number of samples: 10000

Minimum possible data value: 0 Maximum possible data value: 1

Target observed coefficient: **0.6500**

What is the probability of observing a coefficient
as high or higher than the target, by chance alone? p = 0.65720

Median random value: 0.6681

Interquartile range: from 0.6387 to 0.6960

95% Credibility (Confidence) range: from 0.5812 to 0.7463
99% Credibility (Confidence) range: from 0.5538 to 0.7674

All or some of the text in this report box may also be selected/copied into the clipboard in the usual way, and pasted into a Word document etc.

I strongly recommend you look at the video help on the program download page as the interactive nature of this program lends itself better to "show and tell"!

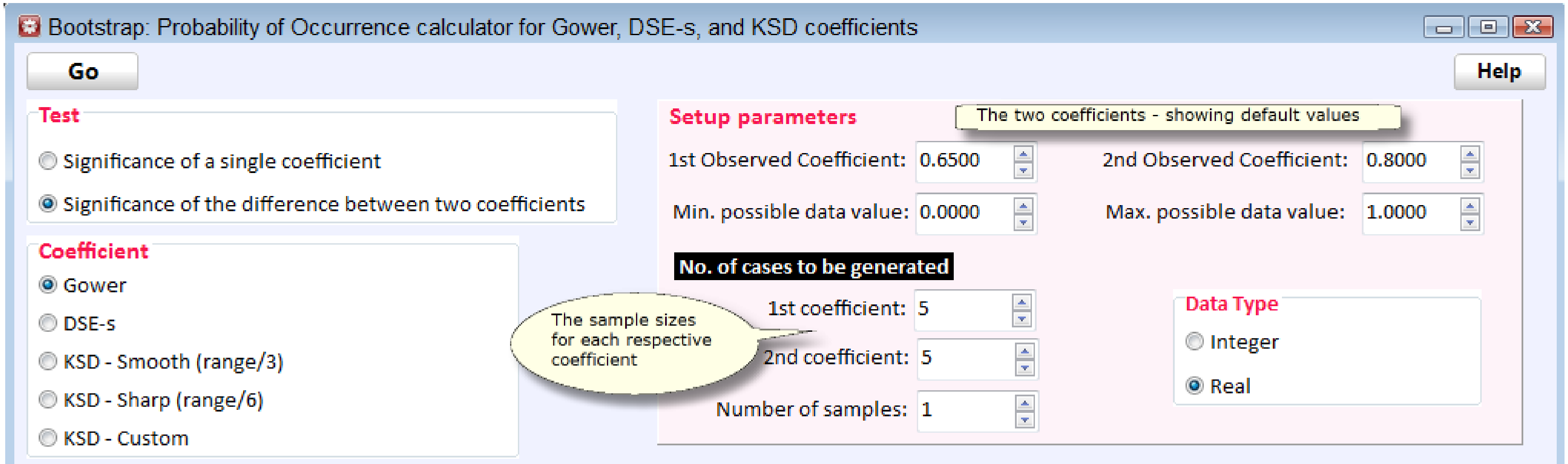
2.2 Significance of the difference between two coefficients

This where you investigate the estimated probability of occurrence of an observed difference between two coefficients.

When comparing the difference between two coefficients, the program assumes:

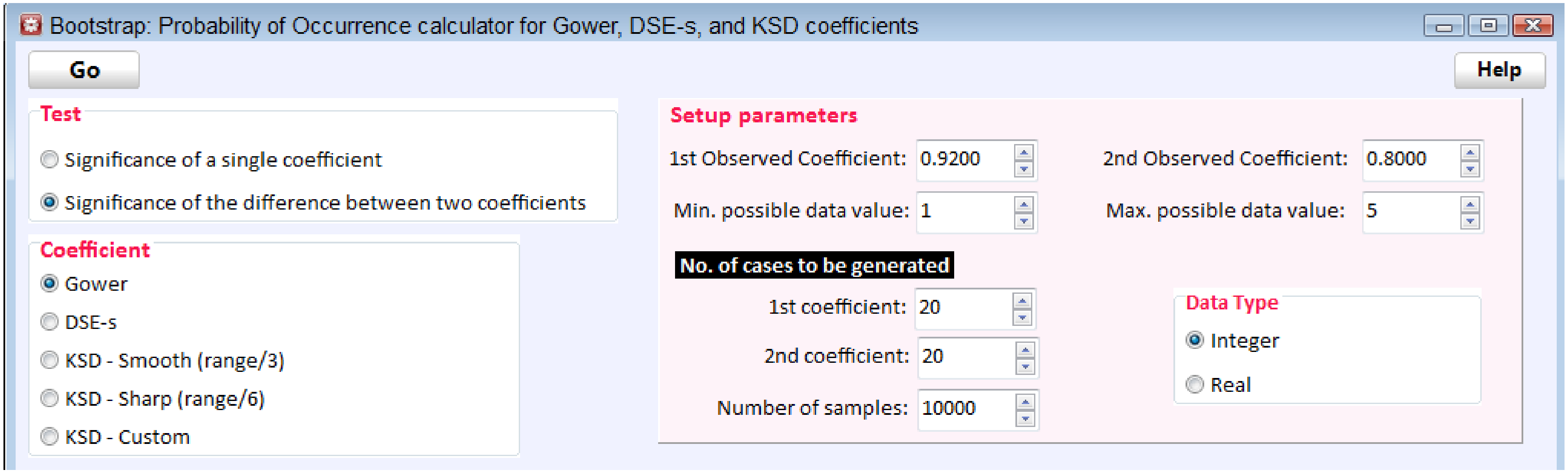
1. The coefficients are the same class (both Gower, both KSD-s etc).
2. The minimum and maximum metric range for each coefficient is equivalent.
3. The number of cases for each coefficient may be different.

When the program first opens, it defaults to examining a single coefficient .. you would click on the radiobutton: "Significance of the difference between two coefficients" .. and see ..

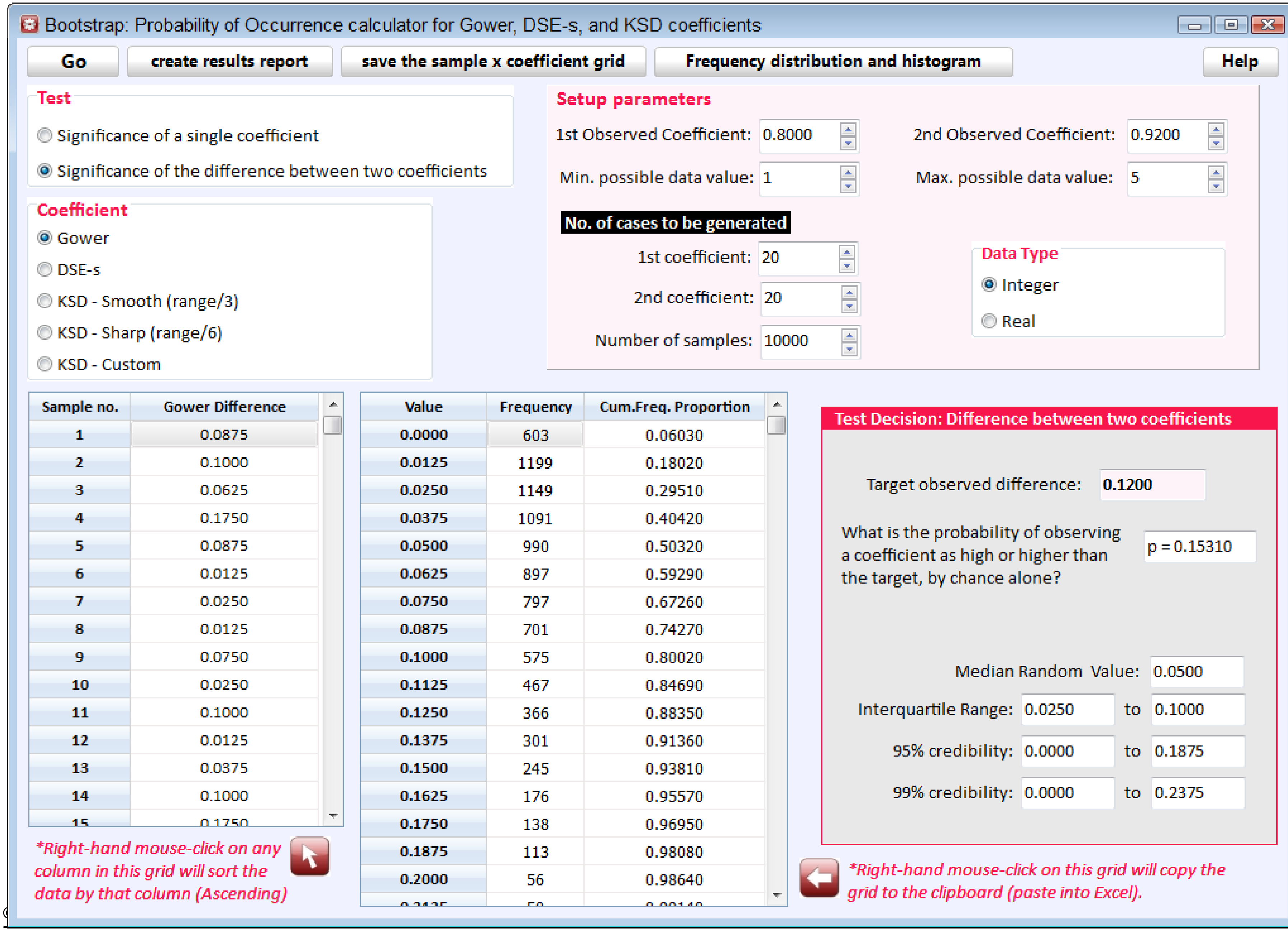


So, let's say we computed a Gower coefficient in one set of rating data, where two experienced raters had rated 20 individuals' behavior on a 1-5-point integer rating scale, with a resulting agreement coefficient of 0.92. In addition, we had another pair of raters who are fresh from training rate the same 20 individuals on the same scale, with a resulting agreement coefficient of 0.80. We want to establish whether we might expect a difference of (0.92-0.80 = 0.12) by chance alone (and so not continue with the training for our new raters), or whether the difference-value (0.12) is indicative of a systematic effect (i.e. the new raters require further training).

This is what the setup should look like ...

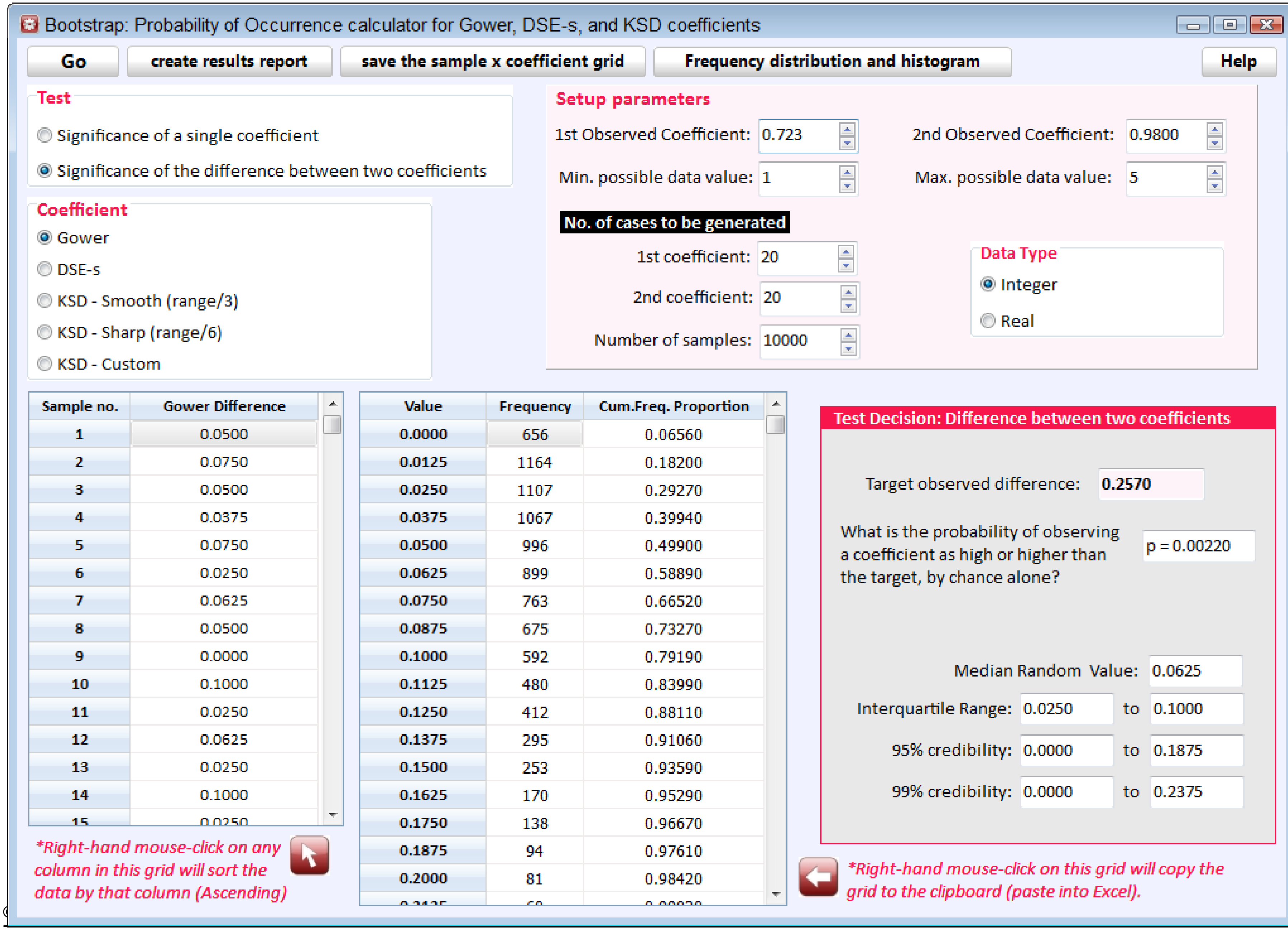


clicking on the "Go" button shows ...


Bootstrap: Probability of Occurrence calculator for Gower, DSE-s, and KSD coefficients
Go
create results report
save the sample x coefficient grid
Frequency distribution and histogram
Help
Test
Significance of a single coefficient
Significance of the difference between two coefficients
Coefficient
Gower
DSE-s
KSD - Smooth (range/3)
KSD - Sharp (range/6)
KSD - Custom
Setup parameters
1st Observed Coefficient: 0.8000
2nd Observed Coefficient: 0.9200
Min. possible data value: 1
Max. possible data value: 5
No. of cases to be generated
1st coefficient: 20
2nd coefficient: 20
Number of samples: 10000
Data Type
Integer
Real
Sample no. | Gower Difference
1 | 0.0875
2 | 0.1000
3 | 0.0625
4 | 0.1750
5 | 0.0875
6 | 0.0125
7 | 0.0250
8 | 0.0125
9 | 0.0750
10 | 0.0250
11 | 0.1000
12 | 0.0125
13 | 0.0375
14 | 0.1000
Value | Frequency | Cum.Freq. Proportion
0.0000 | 603 | 0.06030
0.0125 | 1199 | 0.18020
0.0250 | 1149 | 0.29510
0.0375 | 1091 | 0.40420
0.0500 | 990 | 0.50320
0.0625 | 897 | 0.59290
0.0750 | 797 | 0.67260
0.0875 | 701 | 0.74270
0.1000 | 575 | 0.80020
0.1125 | 467 | 0.84690
0.1250 | 366 | 0.88350
0.1375 | 301 | 0.91360
0.1500 | 245 | 0.93810
0.1625 | 176 | 0.95570
0.1750 | 138 | 0.96950
0.1875 | 113 | 0.98080
0.2000 | 56 | 0.98640
Test Decision: Difference between two coefficients
Target observed difference: 0.1200
What is the probability of observing a coefficient as high or higher than the target, by chance alone?
p = 0.15310
Median Random Value: 0.0500
Interquartile Range: 0.0250 to 0.1000
95% credibility: 0.0000 to 0.1875
99% credibility: 0.0000 to 0.2375
*Right-hand mouse-click on any column in this grid will sort the data by that column (Ascending)
*Right-hand mouse-click on this grid will copy the grid to the clipboard (paste into Excel).

We observed a difference as large as 0.12 in 15.3% of all of the generated samples. That suggests that the difference we observed looks more like "random error" or a chance effect rather than something systematic. That is, a difference of 0.12 is not really a "rare" event when comparing random dataset coefficient values.

If we had observed a coefficient of 0.723 for the trainee raters, we would have produced the following result:


Bootstrap: Probability of Occurrence calculator for Gower, DSE-s, and KSD coefficients
Go
create results report
save the sample x coefficient grid
Frequency distribution and histogram
Help
Test
Significance of a single coefficient
Significance of the difference between two coefficients
Coefficient
Gower
DSE-s
KSD - Smooth (range/3)
KSD - Sharp (range/6)
KSD - Custom
Setup parameters
1st Observed Coefficient: 0.723
2nd Observed Coefficient: 0.9800
Min. possible data value: 1
Max. possible data value: 5
No. of cases to be generated
1st coefficient: 20
2nd coefficient: 20
Number of samples: 10000
Data Type
Integer
Real
Sample no. Gower Difference
1 0.0500
2 0.0750
3 0.0500
4 0.0375
5 0.0750
6 0.0250
7 0.0625
8 0.0500
9 0.0000
10 0.1000
11 0.0250
12 0.0625
13 0.0250
14 0.1000
Value Frequency Cum.Freq. Proportion
0.0000 656 0.06560
0.0125 1164 0.18200
0.0250 1107 0.29270
0.0375 1067 0.39940
0.0500 996 0.49900
0.0625 899 0.58890
0.0750 763 0.66520
0.0875 675 0.73270
0.1000 592 0.79190
0.1125 480 0.83990
0.1250 412 0.88110
0.1375 295 0.91060
0.1500 253 0.93590
0.1625 170 0.95290
0.1750 138 0.96670
0.1875 94 0.97610
0.2000 81 0.98420
Test Decision: Difference between two coefficients
Target observed difference: 0.2570
What is the probability of observing a coefficient as high or higher than the target, by chance alone?
p = 0.00220
Median Random Value: 0.0625
Interquartile Range: 0.0250 to 0.1000
95% credibility: 0.0000 to 0.1875
99% credibility: 0.0000 to 0.2375
*Right-hand mouse-click on any column in this grid will sort the data by that column (Ascending)
*Right-hand mouse-click on this grid will copy the grid to the clipboard (paste into Excel).

Now the probability of observing a difference as high or higher than 0.257 is 0.0020. We would conclude that the size of such a difference is a very rare occurrence by chance alone, hence it is indicative of what looks to be a systematic (non-chance) difference.

I strongly suggest you look at the video help on the program download page as the interactive nature of this program lends itself better to "show and tell"!

Part III

3 The coefficients

3.1 Double-scaled euclidean similarity (DSE-s)

This index computes the squared discrepancy between two Vector's values, then divides this squared discrepancy by the maximum possible squared discrepancy for that case/variable. Summing and taking the square root of these “scaled” discrepancies across cases/variables yields a scaled Euclidean distance. But, the metric of this scaled and cumulatively summed variable discrepancy distance varies between 0 and some value greater than 1.0. In order to scale this coefficient into a unit (0 to 1) metric, a further scaling operation takes place. That is, the initially scaled Euclidean distance is divided by the square root of the number of variables comprising the distance computation. This second scaling now produces a coefficient which always varies between 0 (no distance between variables) to 1 (maximum possible distance between variables given the designated maximum and minimum values for each variable).

This dual scaling ensures that the coefficient is comparable between studies and samples where different variable magnitudes might otherwise distort a conventional Euclidean distance. Further, because the initial scaling of distance is linear (rather than the non-linear operation used within the more common solution of converting data to normalized z-scores prior to any distance calculation), the linear distance relations between magnitudes on the variables remains unchanged. Finally, in order to complete the process, the double-scaled distance is expressed as a similarity index by subtracting it from 1, thus yielding the double-scaled Euclidean similarity (DSE-S) measure, where 0 for this coefficient indicates maximum possible disagreement, and 1 indicates that all cases/variables possess identical rating magnitudes.

The formula for this coefficient is:

$$dse_s = 1 - \left[\frac{\sqrt{\sum_{i=1}^{n} \left(\frac{(case_{1i} - case_{2i})^2}{md_i} \right)}}{\sqrt{n}} \right] = 1 - \sqrt{\left[\frac{\sum_{i=1}^{n} \left(\frac{(case_{1i} - case_{2i})^2}{md_i} \right)}{n} \right]}$$

where

$n =$ the number of cases

$md_i =$ (maximum-minimum possible value for case i)^2

$case_{1i} =$ the value for case i of n from the first vector

$case_{2i} =$ the value for case i of n from the second vector

The similarity to the Gower coefficient is obvious, but these will produce different coefficient sizes and distribution densities as the Gower is based upon absolute value discrepancy while the double-scaled Euclidean is based upon squared discrepancies.

The DSE-S coefficient computes a scaled similarity coefficient, utilizing scaled discrepancies. It varies between 0 and +1, where +1 is equal to identity between the two vectors being compared.

Paul Barrett, *(paul@pbarrett.net)*

3.2 Gower agreement

Relative to the maximum possible absolute (*unsigned*) discrepancy between the two pairs of observations, the **gower** *discrepancy* coefficient indicates the % average absolute discrepancy between all pairs of observations. When expressed as a similarity coefficient (by subtracting it from 1), it indicates the % average similarity between all pairs of observations.

So, a Gower similarity coefficient of say 0.90 indicates that relative to the maximum possible absolute (*unsigned*) discrepancy between them, the observations agree to within 90% of each other's values.

If you change the value of that maximum possible discrepancy, then the Gower coefficient will change to reflect this, as the discrepancies between pairs of observations are divided (scaled) by that maximum possible discrepancy value.

E.g. if two observations differ by 5, and the measurement range of each observation is 10, then the relative discrepancy is 0.5. However, if the measurement range for each observation was say 100, then the relative discrepancy would be just 0.1.

But that's the whole point of the Gower, it tells you how discrepant (or similar) observations are, RELATIVE to how discrepant they could have been. A 5-point difference in a 10-point maximum measurement range is not very good. A 5-point difference between observations within a 100-point measurement range is pretty accurate.

The equation for the gower similarity index is:

$$Gower_{similarity} = 1 - \left[\frac{\sum_{i=1}^{n} \left(\frac{|obs_{1i} - obs_{2i}|}{range} \right)}{n} \right]$$

$n =$ the number of cases

$range =$ the maximum possible discrepancy between the two variables

$obs_{1i} =$ the observed value for case i of n in the first set of observations

$obs_{2i} =$ the observed value for case i of n in the second set of observations

The Gower coefficient computes a scaled similarity coefficient, utilizing scaled discrepancies. It varies between 0 and +1, where +1 is equal to identity between the two vectors being compared.

For example, look at the following scores on two tests, with the minimum and maximum possible scores for each test between 0 and 50 ..

	Var1	Var2
1	30	5
2	30	5
3	30	5
4	30	5
5	30	5
6	30	5
7	30	5
8	30	5
9	30	5
10	30	5

The discrepancy between each pair of observations is 25, which is exactly half the maximum possible discrepancy range of 50. So each paired scaled discrepancy is exactly 0.5, with a resulting Gower and DSE-s of **0.5**

That really is the logic of a discrepancy-based coefficient in a nutshell. A value of 0.5 tells you that the similarity between observations is 50% of maximum (which = identity).

Another way of expressing this is that the average discrepancy between observations is one half of the maximum possible discrepancy.

However, we have to be careful here. For the Gower, " the maximum possible discrepancy" is the range of the data (the difference between the maximum possible and minimum possible values). The Gower is the average of the *absolute* discrepancies, divided through by the range, subtracted from 1 to provide the measure of similarity.

There is a lot more information about this coefficient and the consequences of its practical application (and comparisons with other coefficients) in:

Barrett, P.T. (2010) *Test reliability and validity: The inappropriate use of the Pearson and other variance ratio coefficients for indexing reliability and validity*. Technical Whitepaper #9, available at: www.pbarrett.net, from the Whitepapers link.

The coefficient itself was first presented to the research community in 1971 ...
Gower, J. C., 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857-874.

Paul Barrett, *(paul@pbarrett.net)*

3.3 Kernel Smoothed Distance (KSD-s)

This coefficient is based upon a very simple idea that a distance function should be ***shaped*** in such a way that if the simple arithmetic unsigned difference between a person's attribute value and a target value is computed to be within a certain range, then the computed distance should reflect a very small distance, almost regardless of the actual distance. But, as that distance grows larger, then the computed distance should be accelerated in size. In short, an "**inertial**" effect was aimed for – translated into a distance metric. The coefficient itself is scaled as a measure of similarity, varying between 0 (maximal dissimilarity) to 100 (identity).

$$KSD = \frac{\sum_{i=1}^{n}\left[\frac{1}{s\sqrt{2\pi}} e^{-\left[\frac{(case_{1i}-case_{2i})^2}{2s^2}\right]}\right]\cdot\left(100\cdot\left(s\cdot\sqrt{2\pi}\right)\right)}{n}$$

where

$$s = \frac{range_i}{smoother}$$

$smoother$ = the "smoothing" parameter

$range_i$ = (maximum-minimum value) for case/variable i

$case_{1i}$ = the rating value for case i of n from the first vector

$case_{2i}$ = the rating value for case i of n from the second vector

n = the number of cases

The key to using this coefficient is selecting an appropriate value of the smoother constant which produces the desired inertial effect. The selection choice is application-specific; the function in fact needs to be calibrated for each specific application, taking into account the costs and benefits of a sharper or smoother distance/discrepancy function. Person-target profiling applications are the most readily understood in this regard, where profile similarity can be adjusted to reflect only very close matches, with even "nearly similar" is reduced to "no-match" with even tiny departures of a person's profile from a target. Likewise, if a broad screen is required, then the smoothing function can be more gradual - providing a kind of "plateau" effect around the target value, before the discrepancy between person and target is accelerated by the non-linear function.

In essence, this coefficient needs to be "calibrated", to match the "by eye" judgment of the user. That is, when plotting two profiles, or when deciding whether two values are to be adjudged similar to one another given the range of the rating scale being used, it is the user who has to decide when two values are to be adjusted "similar", and not the "mathematics . The KSD coefficient is sensitive only to magnitude discrepancy (not monotonicity). It also takes into account the range of ratings or score values from which the person-target, or rater reliability ratings are drawn.

For example, consider the comparison of two sets of scores from two Raters ...

Rater 1	Rater 2
3	4
4	3
3	4
4	3
3	4
4	3
3	4
4	3
4	4
4	4

If we apply the coefficient formula to the data in the table, with a KSD smoother value of 5,

Rater 1	Rater 2	KSD
3	4	45.783
4	3	45.783
3	4	45.783
4	3	45.783
3	4	45.783
4	3	45.783
3	4	45.783

4	3	45.783
4	4	100
4	4	100

the average of these 10 rater pairs is 56.63% similarity, or using a 0-1 coefficient scaling, 0.566.

Incidentally, the Pearson correlation for these data is – 0.67, ICC model 2 = – 0.80, and model 3 = – 0.67, while the Gower is +0.80, and DSE-S = +0.78.

Such data will produce highly negative Pearson and Intraclass correlations. Assuming a rating scale range of **1-5**, the KSD coefficient with smoother factor of 5 is: **0.57**. If we assume the scores are drawn from a range **1-20**, then the KSD coefficient is **0.97**. Clearly, the design of the coefficient introduces the element of relativity; a score discrepancy of 1 looks reasonably important when the range is just 1 to 5, but trivial when the range is 1-20.

In order to look at the effect of various smoother values over the measurement range, a interactive utility program is available from my website which allows a user to select various values and see the achieved smoothing/plateau effect. This file also contains further help information about the coefficient.

Paul Barrett, *(paul@pbarrett.net)*

Index

- D -

- G -

- K -