[ξ] KSZI [ξ] AKTÁK

MTA KIK TTO műhelytanulmányok

2013/5

X-centage: a Hirsch-inspired indicator for distributions of percentage-valued variables and its use for measuring heterodisciplinarity

Schubert András

schuba@iif.hu

➲ http://www.mtakszi.hu/kszi_aktak/

To be published in *Scientometrics*
DOI: 10.1007/s11192-014-1281-z

X-centage: a Hirsch-inspired indicator for distributions of percentage-valued variables and its use for measuring heterodisciplinarity

András Schubert

Department of Science Policy and Scientometrics,
Library and Information Center of the Hungarian Academy of Sciences,
Budapest, Hungary

Summary

The present paper introduces two independent concepts.
X-centage is a statistical indicator characterizing distributions of percentage-valued variables in a vein similar to Hirsch's h-index.
Heterodisciplinarity is a measure of polydisciplinarity using the disciplinary categorization of references and/or citations.
The Journal Citation Reports database is used for an empirical study of using the X-centage for measuring reference heterodisciplinarity of science fields.

Introduction

The original h-index [1] and its direct generalizations outside the realm of citation distributions (e.g. [2]–[5]) are based on the equality of the value of a variable and its rank in an ordered sample. The success of the use of these indices, therefore, largely depends on a fortunate equality (at least, in order of magnitude) between the sample size and the top values of the variable.
The index introduced in this paper also hinges on the equality of the value of a variable and its position in the distribution, but in a quite different context.

Methodology

Definition and demonstration of the index

The distributions considered here have variables of percentage values. As an everyday example we may consider the alcohol content of various beverages in a given stock. In this case, the cumulative frequency distribution, F(X), has the meaning: how many percentages of the total stock contain less than or equal to X percentages of alcohol; its complement, $G(X) = 1 – F(X)$.
The definition of the X-centage value, X*, is as follows: X* is the smallest X value for which $G(X) < X$. In more formal wording (thankfully acknowledged to one of the referees of the paper): $X^* = \mathrm{argmin}_X(G(X) – X)$.
In the above example, if 90% of the stock contains at least 90% alcohol and less than 90% contains more than 90%, then X* = 90% .
If X has a constant value, X_0 (the full stock consists of beverages of the same alcohol percentage), then obviously, $X^* = X_0$. (The distribution function in this case is the step function: $F(X) = 0$ for $X < X_0$, $F(X) = 1$ for $X \geq X_0$.) In this case, of course, X* is equal to the mean value of the distribution.

X-centage and other indicators of the distribution

The behavior of the X-centage and its relation to other statistics will be demonstrated on the example of the beta distribution [6].
The beta distribution is a two-parameter distribution with the cumulative frequency distribution

$$F(X) = I_X(a,b) \text{ with } 0 \leq X \leq 1,\ a \geq 0,\ b\geq 0,$$

where $I_X(a,b)$ is the regularized incomplete Beta-function:

$$I_X(a,b) = \int_0^X t^{a-1}(1-t)^{b-1}dt / \int_0^1 t^{a-1}(1-t)^{b-1}dt \quad .$$

The beta distribution is rather flexible. With a proper choice of the parameters practically all unimodal distributions with a range of [0,1] can be well approximated. As special cases, it includes the uniform distribution (a = 1, b = 1, F(X) = X) and the power function distribution (b = 1, F(X) = X^a).

In case of uniform distribution, the defining equality is G(X) = 1 – F(X) = 1 – X = X, X* = 50%, i.e., X* is, again, equal to the mean value of the distribution.

If the distribution of X is a power function, F(X) = X^a , then the mean value of G(X) is $a/(a+1)$, and X* is the solution of the equation $X^a + X - 1 = 0$.

Figure 1 illustrates the graphical solution of the equation for a few selected values of a. This graphical procedure can be used for determining the X* value in the general case for arbitrary empirical distribution functions, as well. X* is represented on the diagram as the intersection of the G(X) curve with the straight line G(X) = X (marked on the figure by red asterisk).

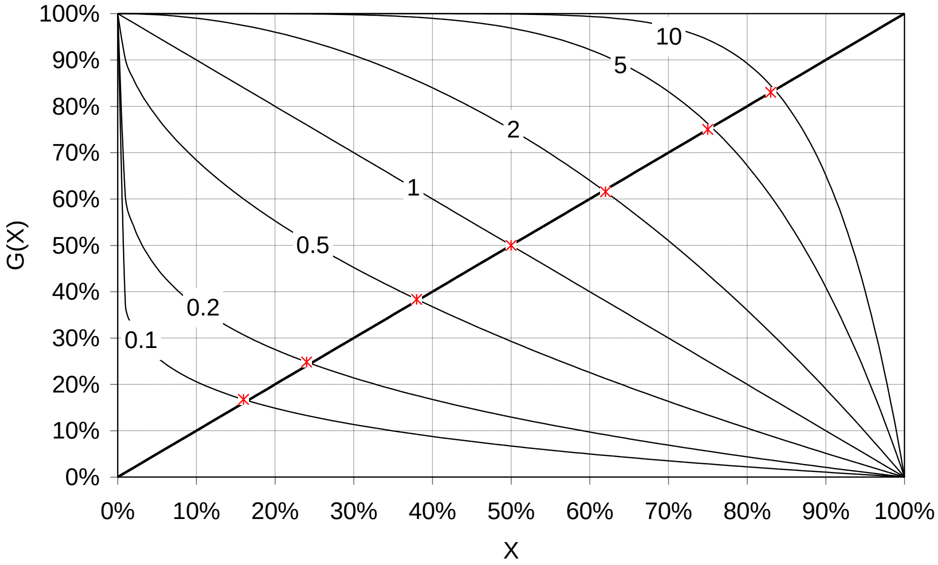


Figure 1 The graphical determination of the X* value for power functions with selected a parameters
X* is represented by the intersection of the G(X) curve with the straight line G(X) = X (marked by red asterisk).

Within the power function distribution family there is a clear monotonous relation between the X-centage and the mean value. At the same time, the difference (X*–Mean) has a strong positive correlation with the skewness of the distribution but is, apparently, uncorrelated with the standard deviation.

This relation appears to be generally valid in the two-parameter beta distribution. Figures 2 and 3 show the correlation between X* vs the mean value, as well as the standard deviation and the skewness vs (X*-Mean), respectively. The plots are based on beta distributions with integer values of a and b in the range [1,10].

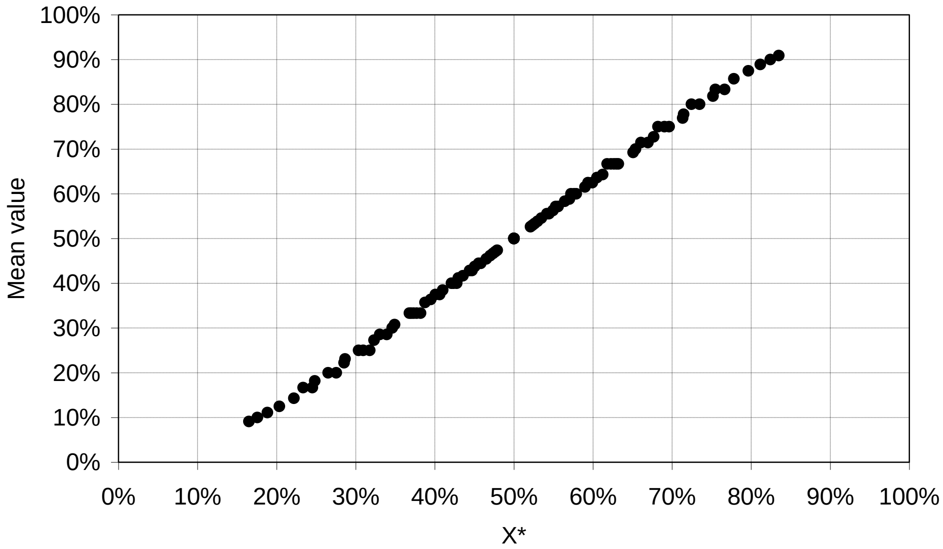


Figure 2 Plot of the mean value of beta distributions with selected *a* and *b* parameters vs X*

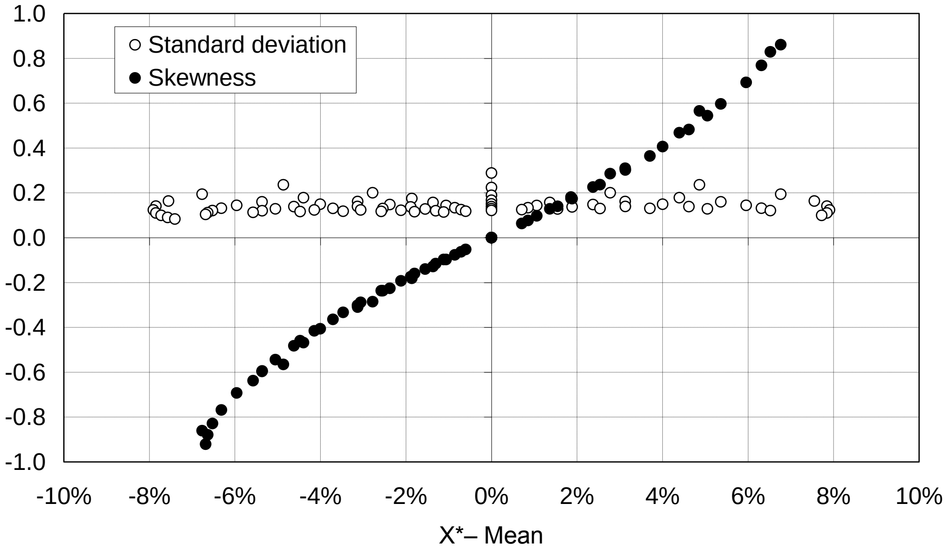


Figure 3 Plot of the standard deviation and the skewness of beta distributions with selected *a* and *b* parameters vs the difference (X*–Mean)

Based on the experiences gained with the beta distribution as model, it can be said that just as Hirsch's h-index combines the sample size and the mean value in a single measure, X-centage combines the location (mean value) and the asymmetry (skewness) of the distribution of a percentage-valued variable in a single indicator. To find the exact nature of the dependence requires further studies.

Defining the disciplinary character of journals through references and citations

It is a standard practice to study journal interdisciplinarity (as well as multidisciplinarity, pluridisciplinarity, transdisciplinarity and the like) through the citation and reference structure of the journals (see, e.g., [7–9]).

Let us consider now a set of journals classified into disciplines (categories, subject areas, etc.). Most advantageously, let these categories be mutually exclusive (like, e.g., the field categorization used in Thomson–Reuters Essential Science Indicators (ESI) [10]).

The *reference heterodisciplinarity* of a journal is defined then as the percentage share of references in the journal given to sources outside the discipline (field) of the journal itself. ("Reference multidisciplinarity" as defined in [8] for single papers.)

Likewise, the *citation heterodisciplinarity* of a journal is defined as the percentage share of citations to the journal received from sources outside the discipline (field) of the journal itself.

Low values of heterodisciplinarity indicate strong monodisciplinary character of the journal, while higher values show polydisciplinarity (whether we call it inter-, multi-, trans- or any other disciplinarities).

The somewhat unusual term heterodisciplinarity is used here to clearly distinguish the here defined indicator from the multitude of other *-disciplinarity variants. Actually our indicator is practically identical with the Citation Outside Category indicator of cross-disciplinarity as defined by Porter & Chubin [7], however, since then the term cross-disciplinarity has been used in various other contexts, as well. The relation of these related terms has been and certainly will be the target of separate studies (as initiated, e.g., in [9]).

Measuring heterodisciplinarity with the X-centage indicator

The heterodisciplinarity of a science field can be measured by the X-centage indicator as follows: if X% of the journals of the field has a heterodisciplinarity of at least X%, and less than X% has more than X%, then X% is the X-centage value of the heterodisciplinarity of the field.

Of course, both reference and citation heterodisciplinarity can be measured this way.

Results

For an empirical study, data of the "Citing Journal Package" of the Thomson–Reuters 2006 Journal Citation Reports database (JCR; SCI and SSCI editions combined) were used. All source journals covered by the database were recategorized into the ESI fields. The data of a total of 7419 journals were processed. References to all source journals receiving at least 2 references were taken into account in the study. Thus, both cited and citing journals could be uniquely assigned to one of the 22 ESI science fields. For each journal, thereby, the reference heterodisciplinarity could be calculated as defined above.

The graphical determination of the reference heterodisciplinarity of the 22 ESI fields is demonstrated in the Appendix. The X-centage values are represented by the intersection of the G(X) curve with the straight line G(X) = X (marked by a red asterisk).

The X-centage value and the basic statistical indicators of the 22 ESI fields are given in Table 1. The field "Multidisciplinary" containing only 17 journals and definitely not representing a specific field (discipline) is disregarded from the analysis. (Although its heterodisciplinarity, as could be expected, is extremely high.)

Table 1 X* and the basic statistical indicators of the reference heterodisciplinarity of the 22 ESI fields (in decreasing order of X*)

ESI field	X*	Mean value	X*–Mean	Standard Deviation	Skewness
Multidisciplinary	73.3%	73.12%	0.2%	23.42%	-187.82%
Pharmacology & Toxicology	72.6%	76.72%	-4.1%	15.55%	-154.65%
Molecular Biology & Genetics	63.7%	68.01%	-4.3%	14.39%	-29.10%
Biology & Biochemistry	63.5%	67.74%	-4.2%	11.89%	-26.15%
Immunology	62.5%	67.35%	-4.9%	9.87%	77.29%
Microbiology	62.3%	65.92%	-3.6%	10.46%	1.94%
Agricultural Sciences	52.6%	56.52%	-3.9%	19.67%	-2.81%
Environment/Ecology	51.4%	55.03%	-3.6%	15.99%	56.49%
Neuroscience & Behavior	50.0%	51.08%	-1.1%	11.77%	41.00%
Materials Sciences	48.4%	48.73%	-0.3%	23.14%	-5.36%
Computer Science	46.8%	47.19%	-0.4%	27.40%	17.71%
Plant & Animal Science	43.9%	42.70%	1.2%	16.21%	46.63%
Engineering	43.3%	43.16%	0.1%	24.29%	43.62%
Social Sciences, general	39.0%	34.19%	4.8%	26.81%	53.85%
Chemistry	37.6%	33.42%	4.2%	18.73%	68.81%
Physics	36.2%	33.37%	2.8%	20.08%	94.57%

Space Science	35.4%	32.25%	3.1%	29.32%	71.18%
Psychology/Psychiatry	35.3%	30.83%	4.5%	19.34%	76.07%
Clinical Medicine	31.4%	25.21%	6.2%	19.39%	101.12%
Geosciences	31.1%	26.84%	4.3%	20.33%	116.29%
Economics & Business	26.6%	19.89%	6.7%	17.71%	149.52%
Mathematics	26.5%	20.45%	6.0%	20.44%	160.88%

The X* and mean values run parallel as testified by Figure 4. Their difference, nevertheless, varies in the -4.9%–+6.7% range, and shows apparent positive correlation with the skewness and a hardly observable dependence on the standard deviation of the distribution (see Figure 5).

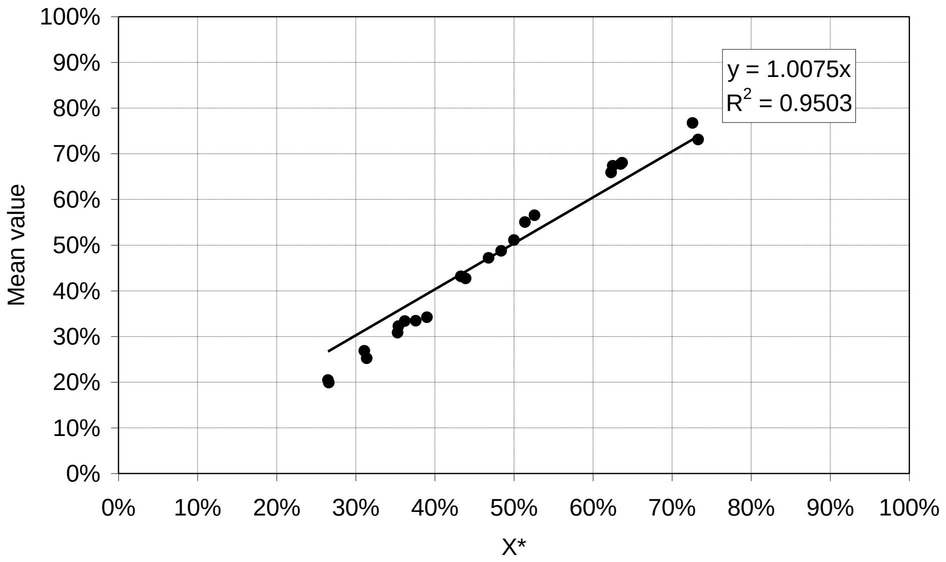


Figure 4 Plot of the mean value of the reference heterodisciplinarity of the 21 ESI fields vs X*

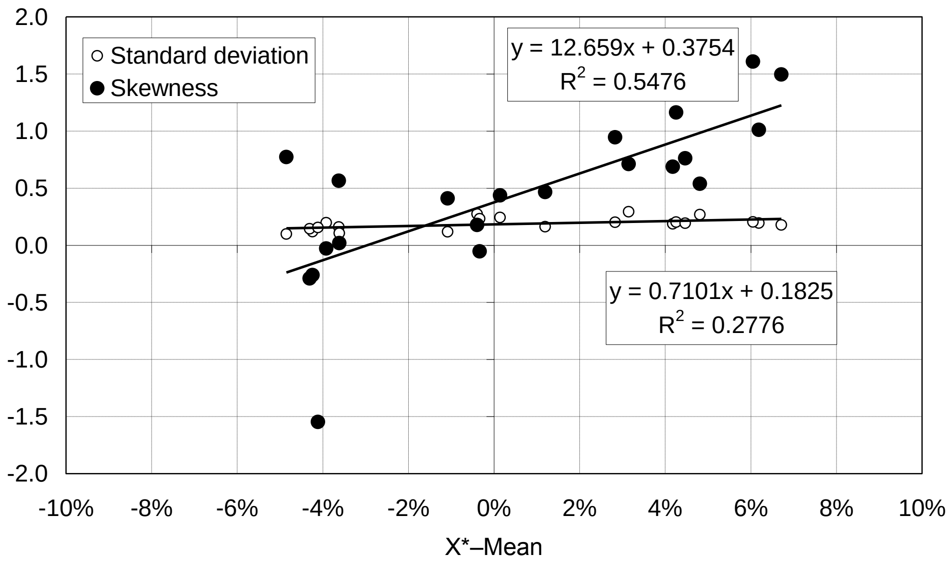


Figure 5 Plot of the standard deviation and the skewness of the reference heterodisciplinarity of the 21 ESI fields vs the difference (X*–Mean)

Apparently, the statistical behavior of the empirical samples was much like it could be expected on the basis of the model studies on the beta distributions.

Conclusion

The present paper introduces two independent concepts.
X-centage is a statistical indicator characterizing distributions of percentage-valued variables in a vein similar to Hirsch's h-index. As the h-index combines the effect of the sample size and the mean

value, X-centage reflects both the location (e.g. mean) and the asymmetry (e.g. skewness) of the distribution.

Heterodisciplinarity is a measure of polydisciplinarity of any bibliometric object (journal, field, author, etc.) using the disciplinary categorization of references and/or citations provided that the object itself is also categorized within the same system. Heterodisciplinarity is defined then as the percentage share of references/citations to/from sources outside the discipline (field) of the object itself.

It would be presumptuous to state that a new indicator would be the optimal choice to characterize a new concept. The first attempts reported here, however, suggest that using the X-centage indicator to measure reference heterodisciplinarity is a coherent procedure that may deserve further elaboration.

Both concepts can be extended far beyond the limits of the present exercise. The disciplinary categorization and characterization of journals and fields can easily be put into a wider context of aggregated knowledge flow networks or even more general economic and social frameworks, and the X-centage concept itself can lead to derivative measures similarly to the h-index.

Acknowledgment

This work was supported by the European Commission under the FP7 Science in Society Grant No. 266588 (SISOB project).

References

[1] J. E. Hirsch, An index to quantify an individual’s scientific output. Proceedings of the National Academy of Sciences of the United States of America, 102, 2005, 16569–16572.

[2] A. Korn, A. Schubert, A. Telcs, Lobby index in networks. Physica A, 388, 2009, 2221–2226.

[3] A. Schubert, A. Korn, A. Telcs, Hirsch-type indices for characterizing networks. Scientometrics, 78(2), 2009, 375–382.

[4] A. Schubert, A Hirsch-type index of co-author partnership ability. Scientometrics, 91(1), 2012, 303–308.

[5] A. Schubert, Jazz discometrics – A network approach, Journal of Informetrics, 6, 2012, 480–484.

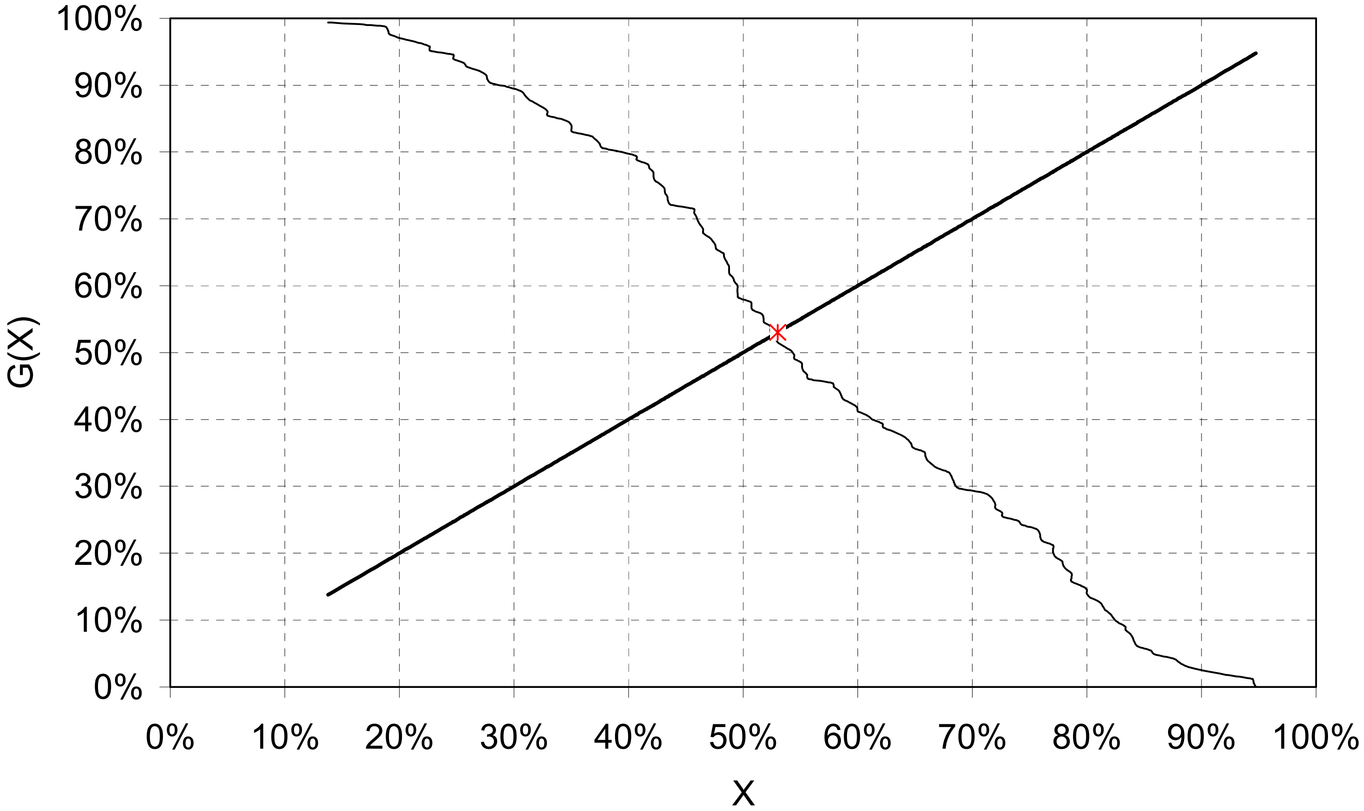
[6] http://en.wikipedia.org/wiki/Beta_distribution (last access: 14 January, 2014)

[7] A. L. Porter, D. E. Chubin, An indicator of cross-disciplinary research, Scientometrics, 8(3-4), 1985, 161–176.

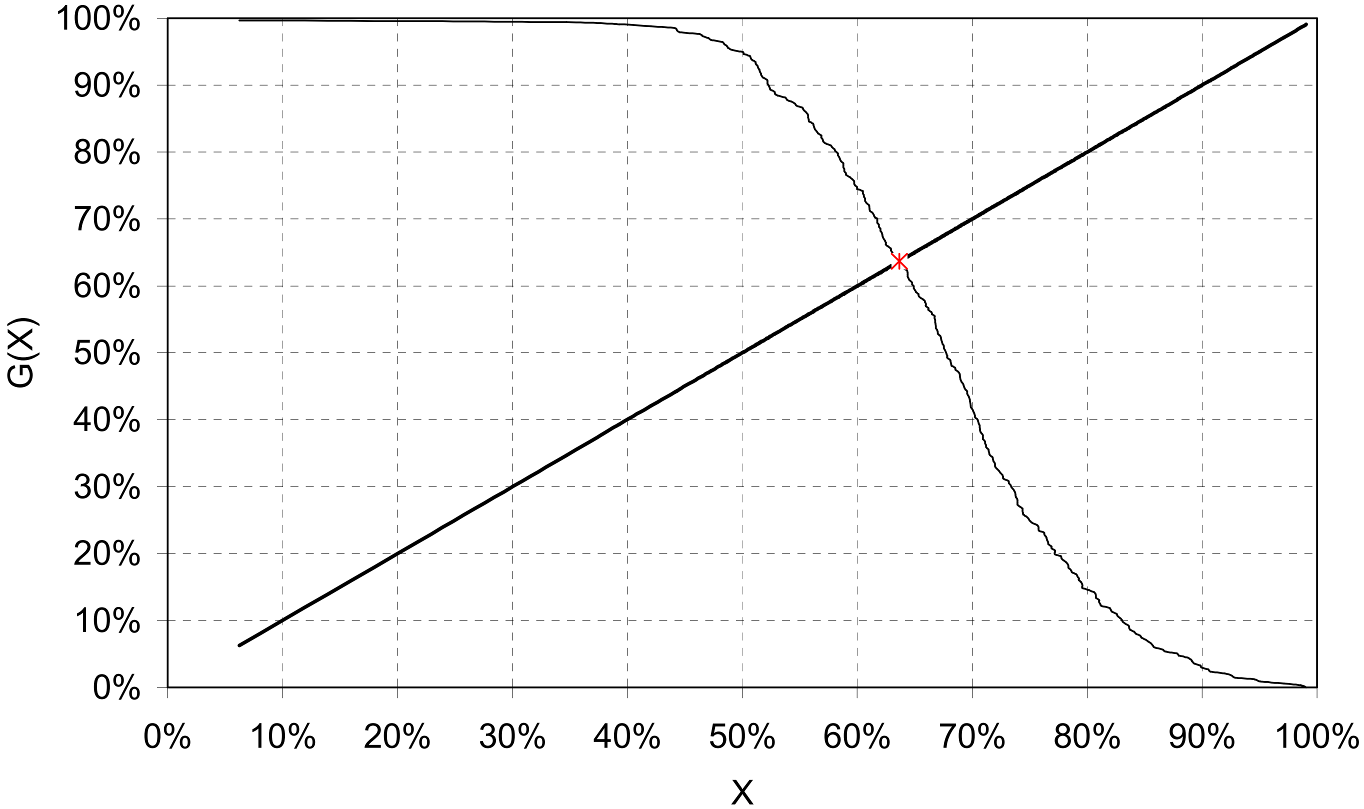
[8] W. Glänzel, A. Schubert, H.-J. Czerwon, An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis, Scientometrics, 44(3), 1999, 427–439.

[9] A. Schubert, Multi- and interdisciplinarity in medical and veterinary literature: Approaches and assertions, ISSI Newsletter, #19, 2009, 48–51.

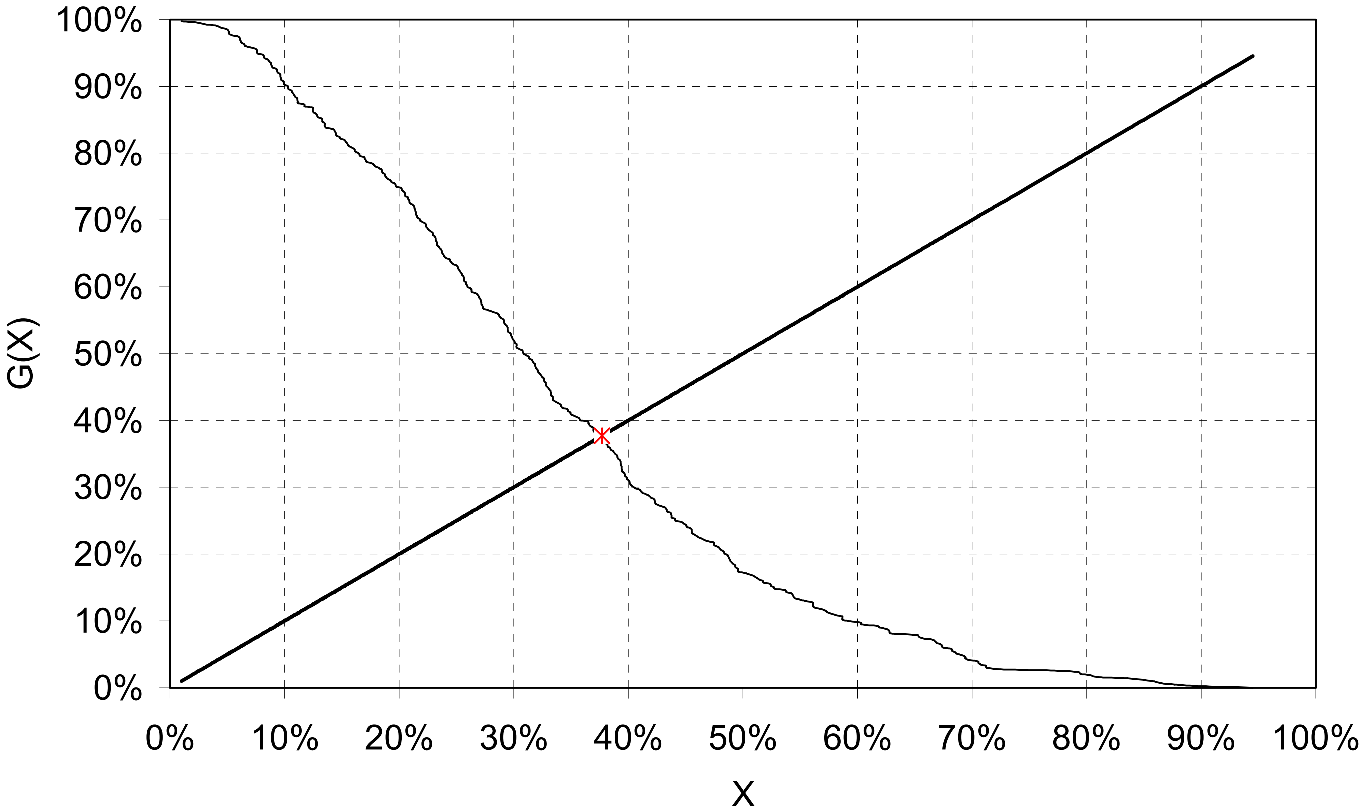
[10] http://thomsonreuters.com/essential-science-indicators/



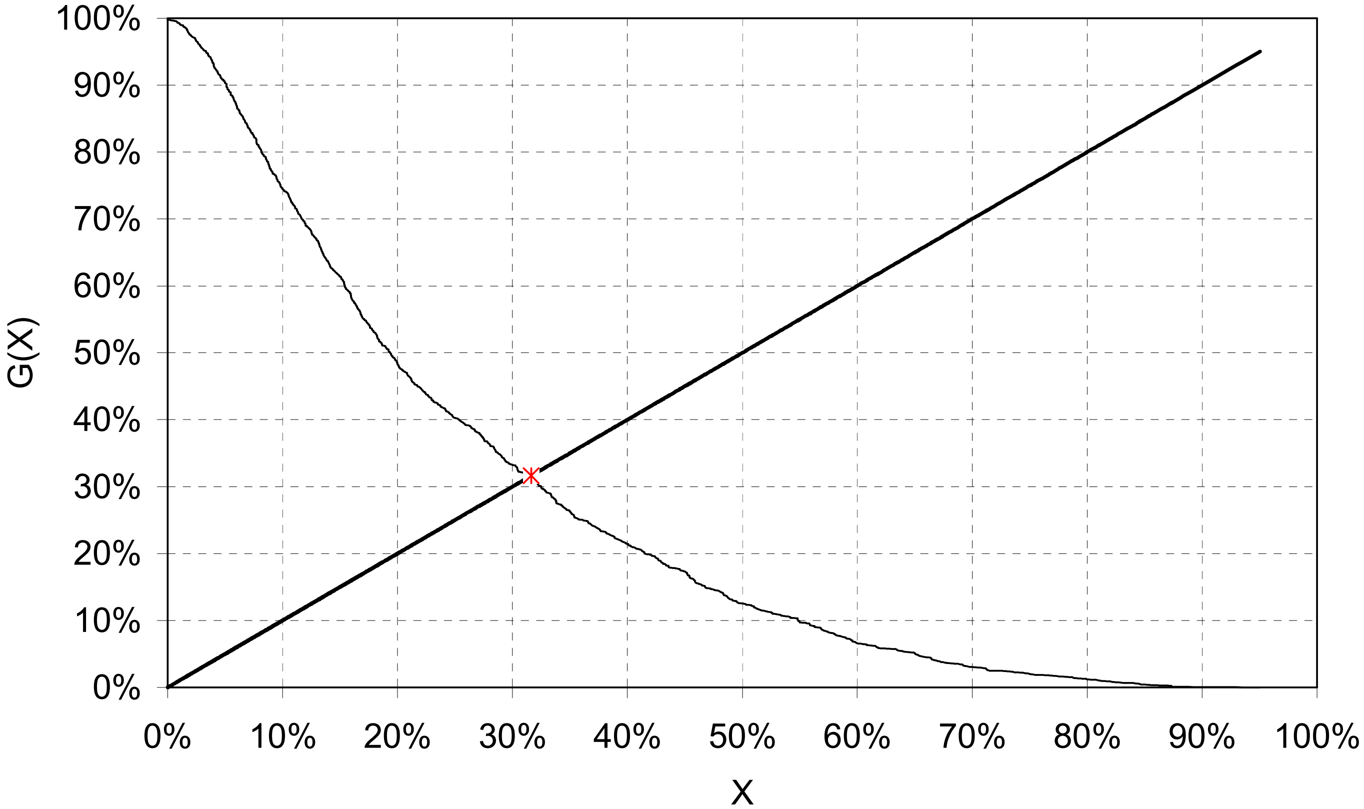
Agricultural Sciences



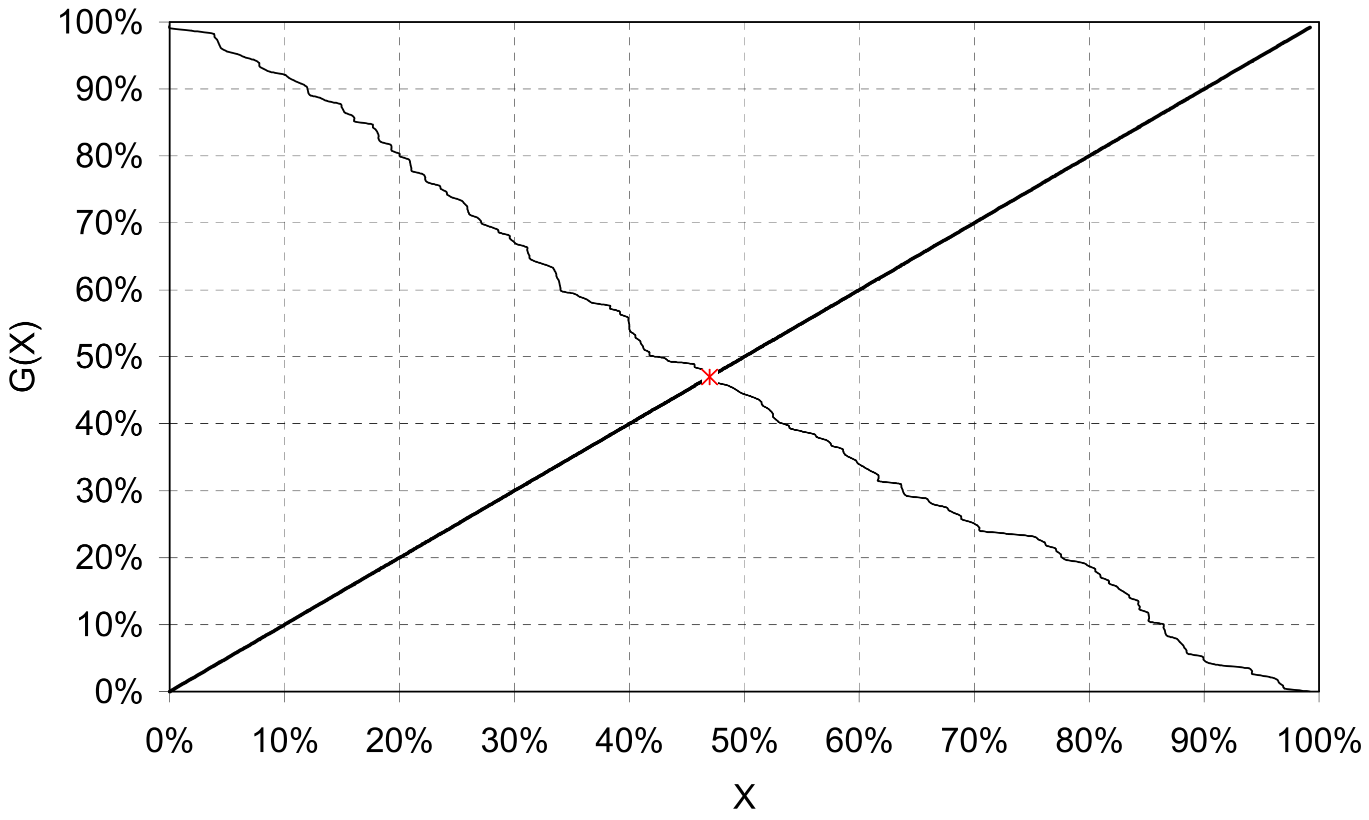
Biology & Biochemistry



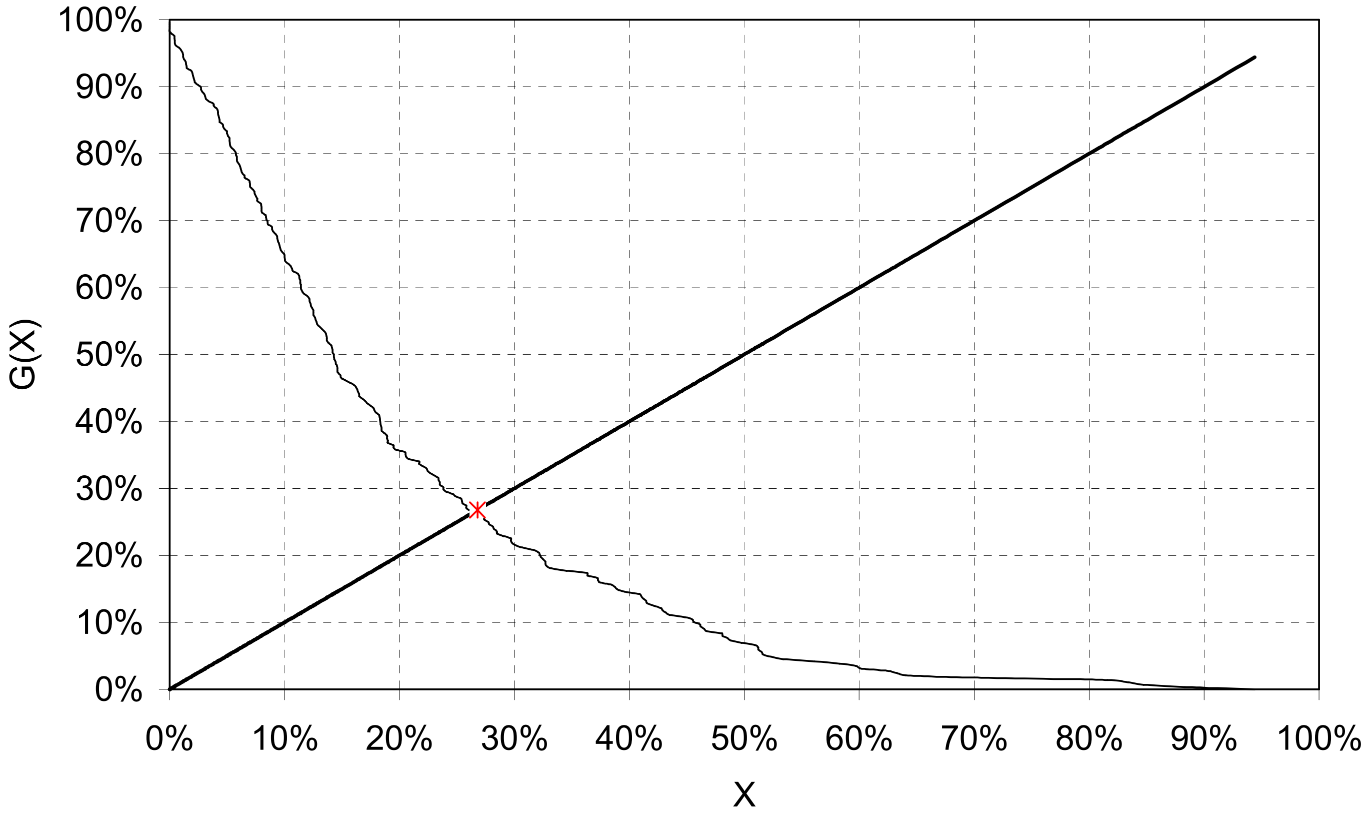
Chemistry



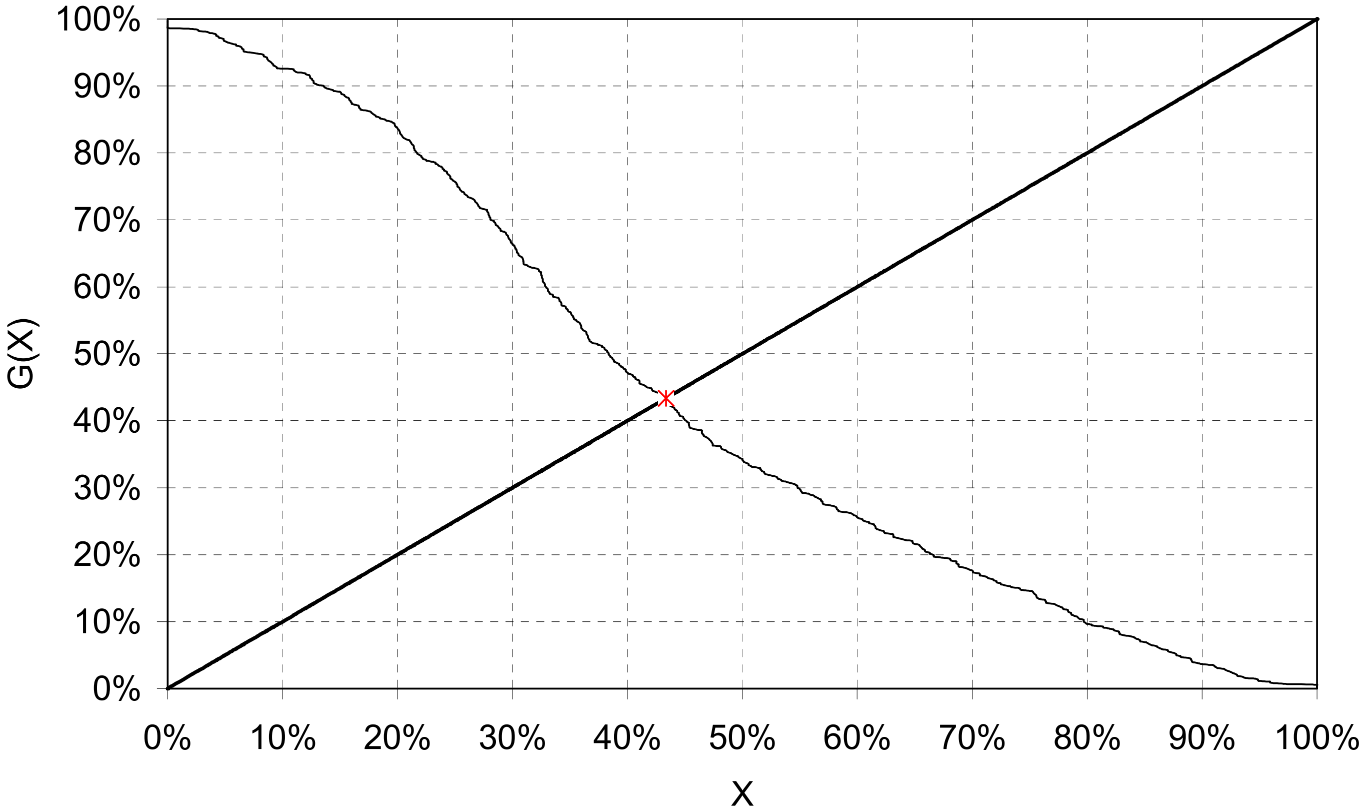
Clinical Medicine



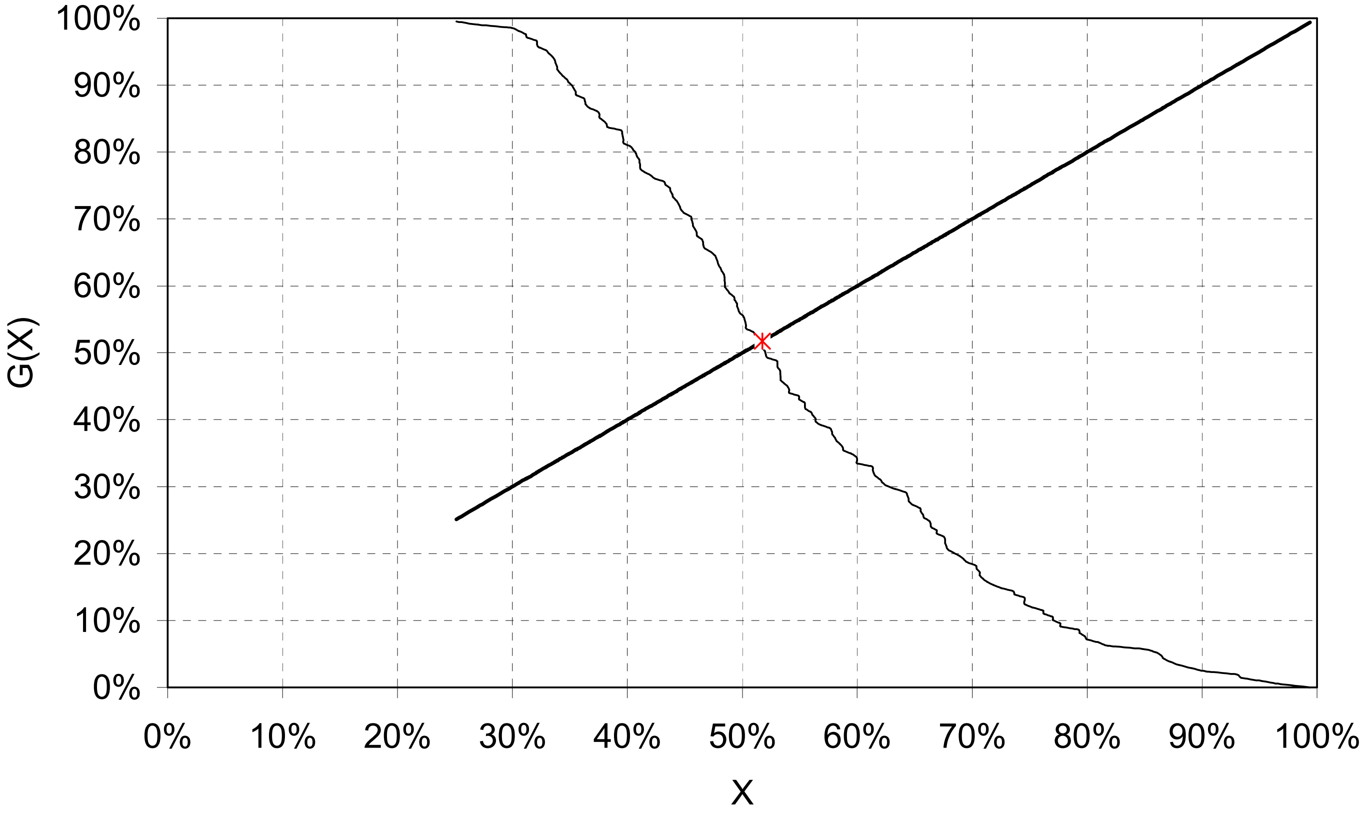
Computer Science



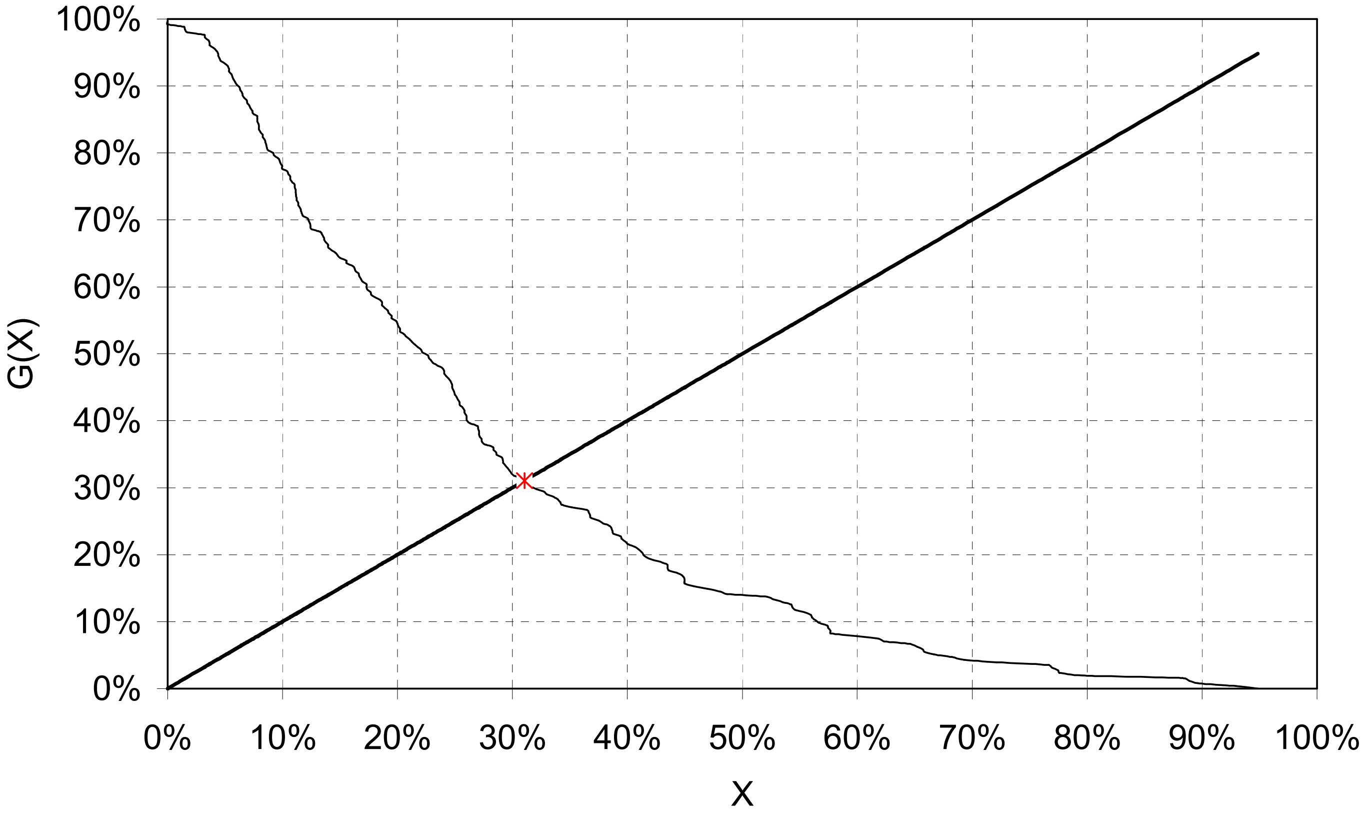
Economics & Business



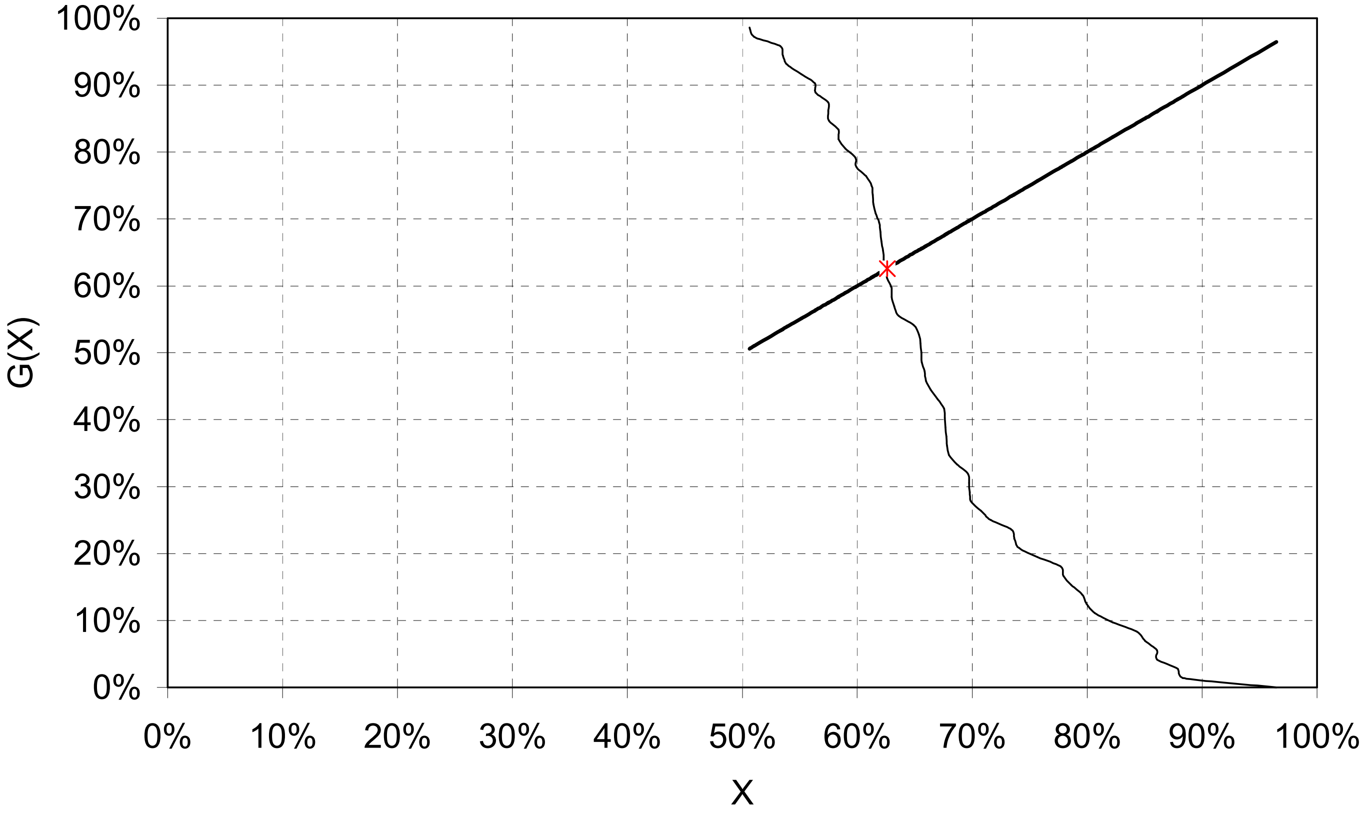
Engineering



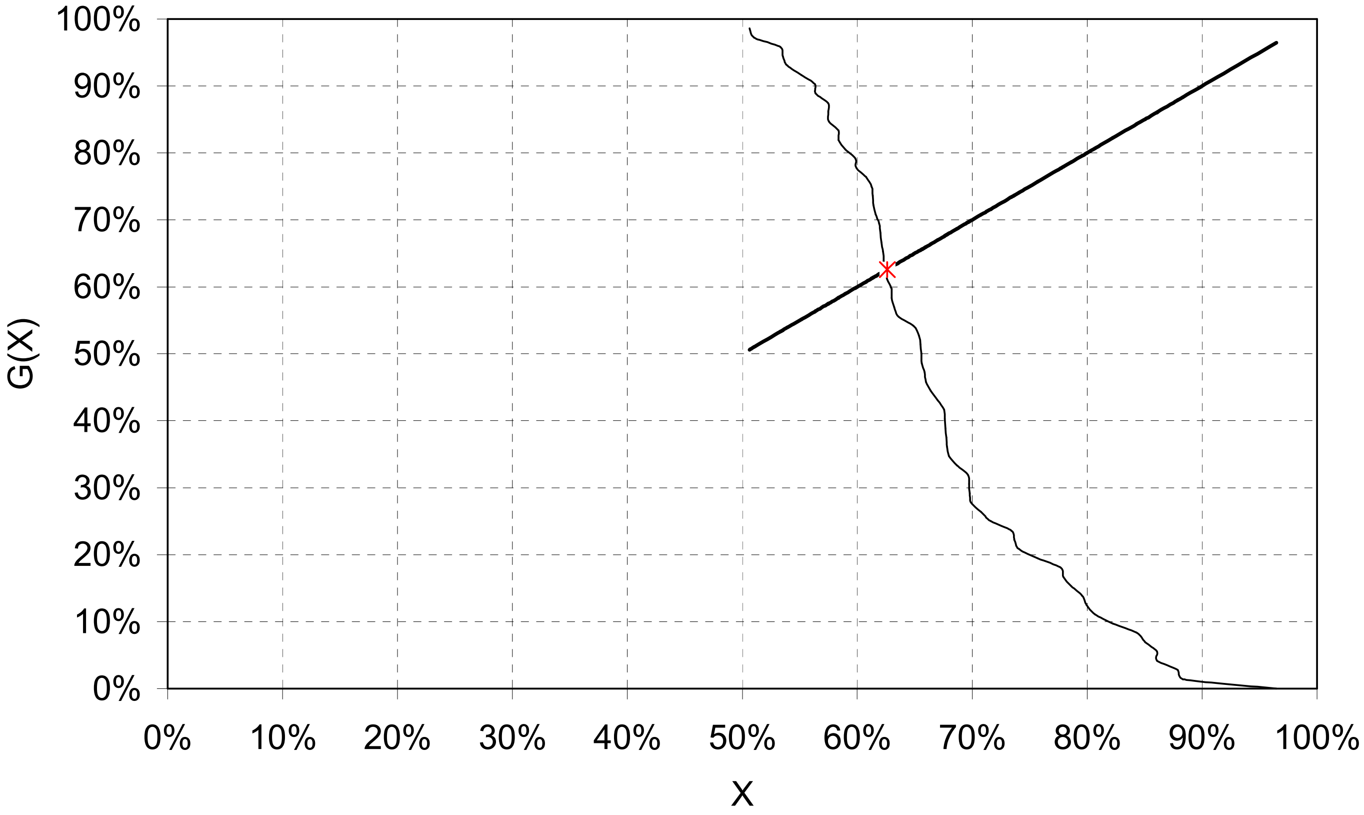
Environment/Ecology



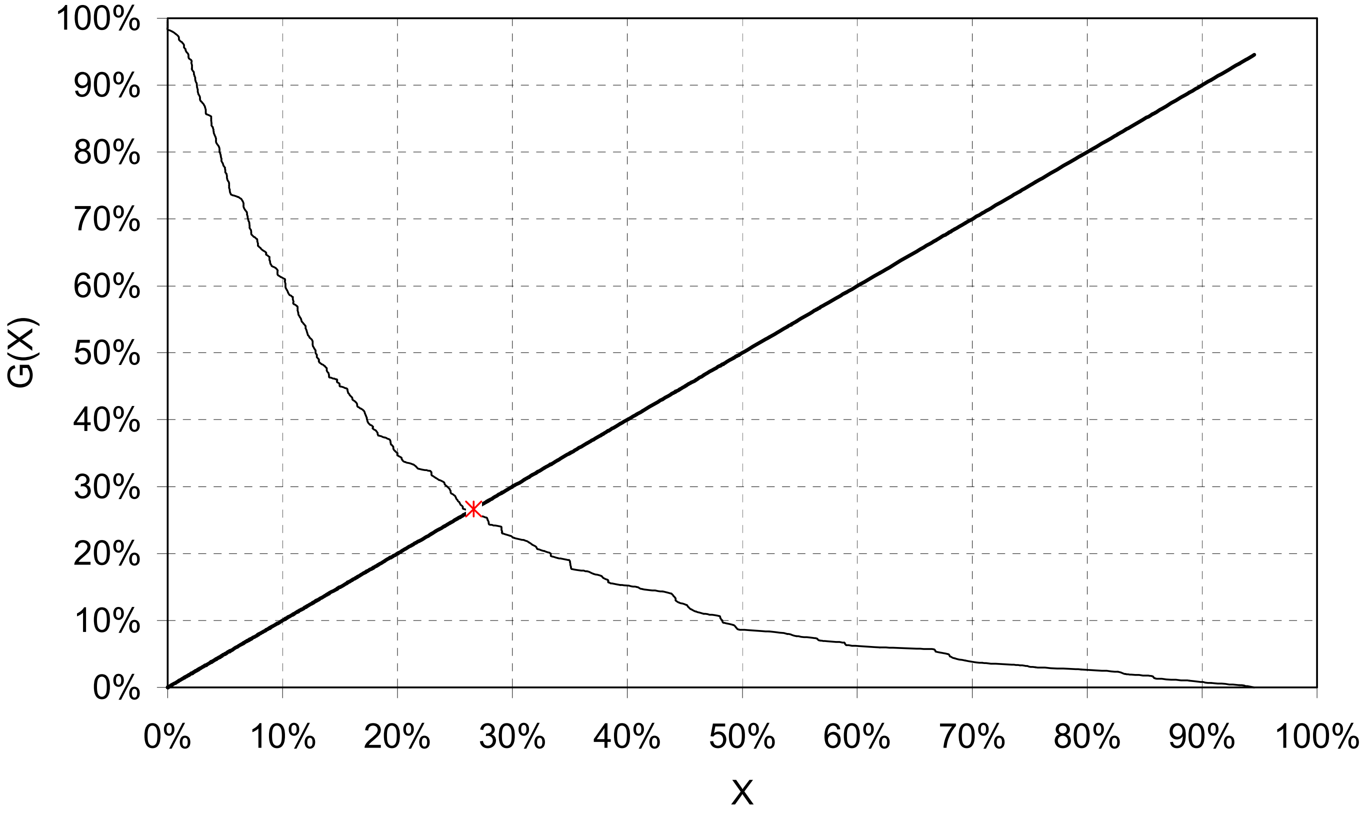
Geosciences



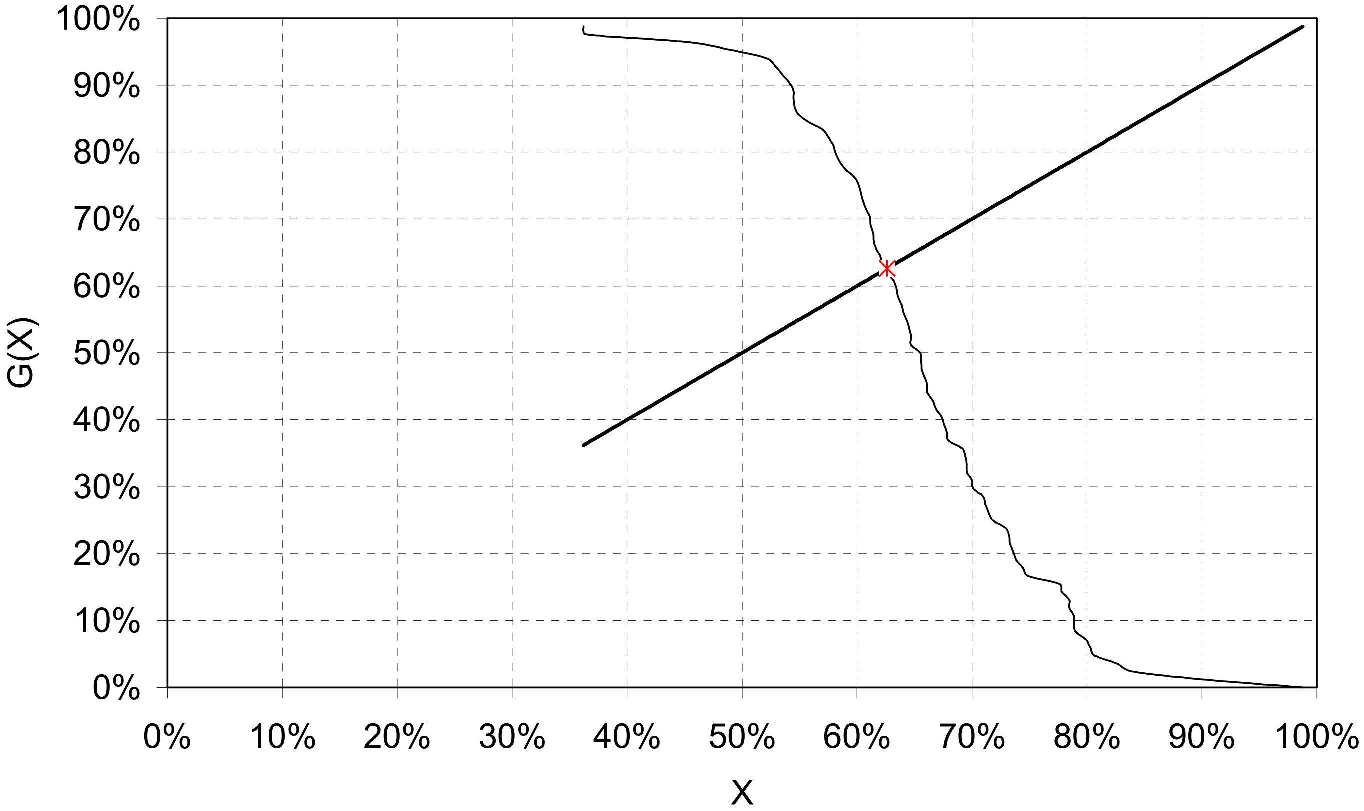
Immunology



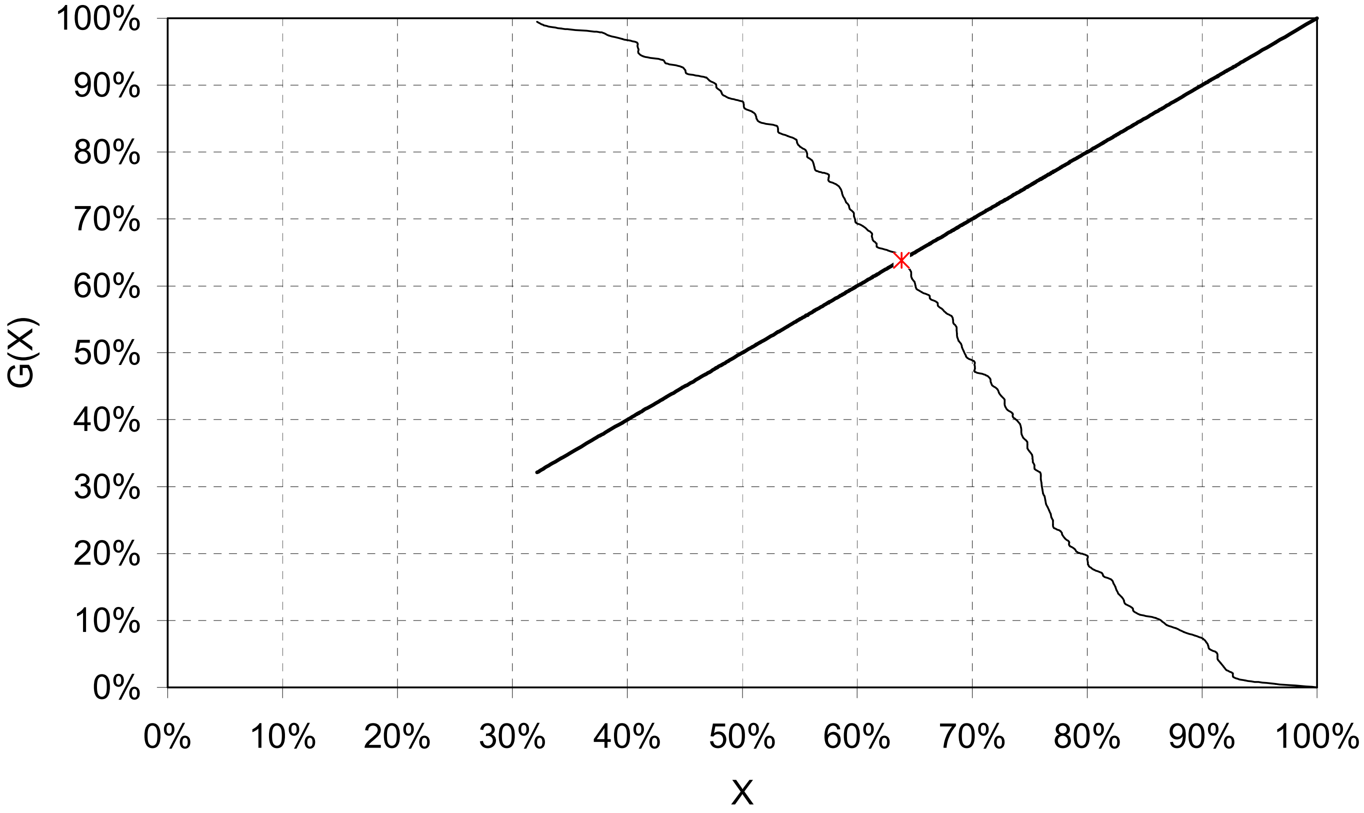
Materials Sciences



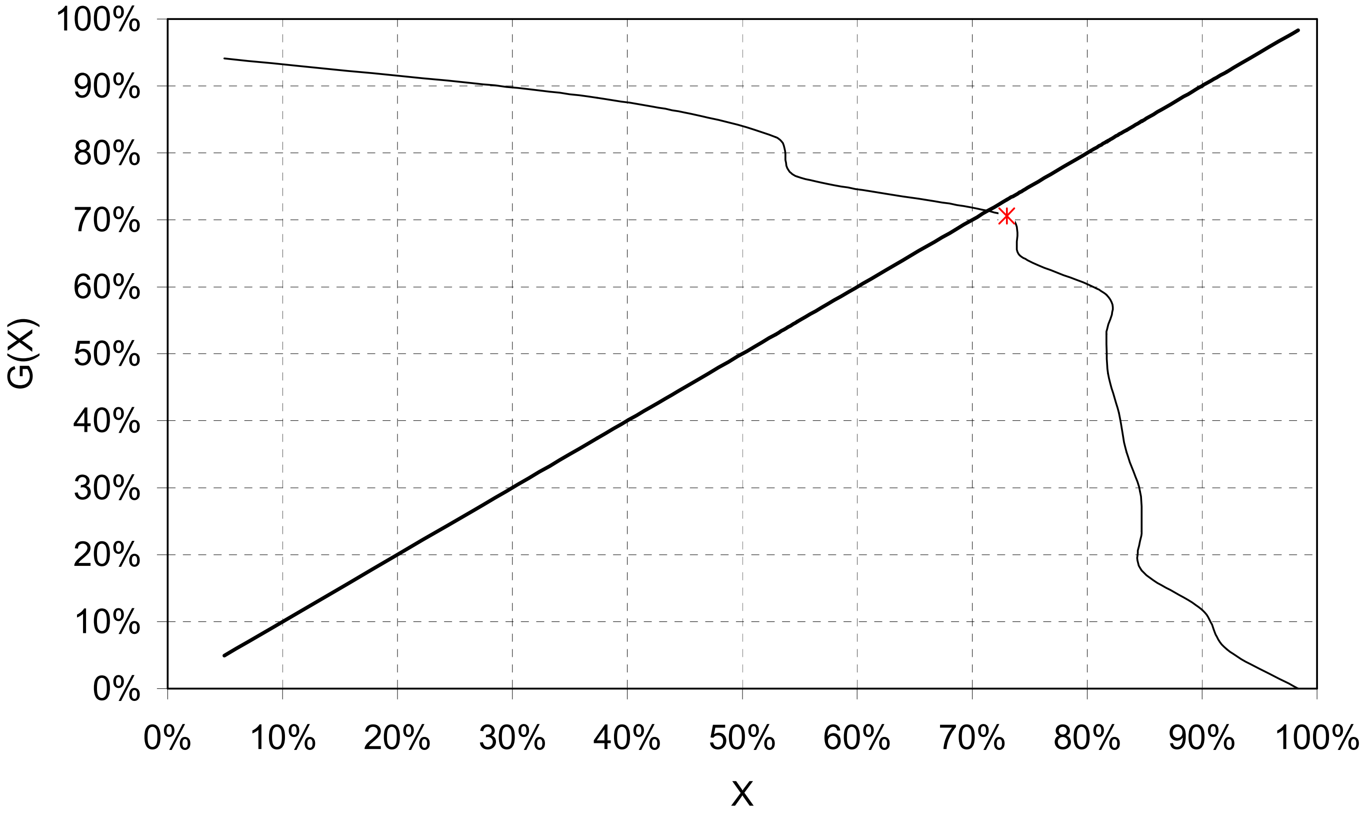
Mathematics



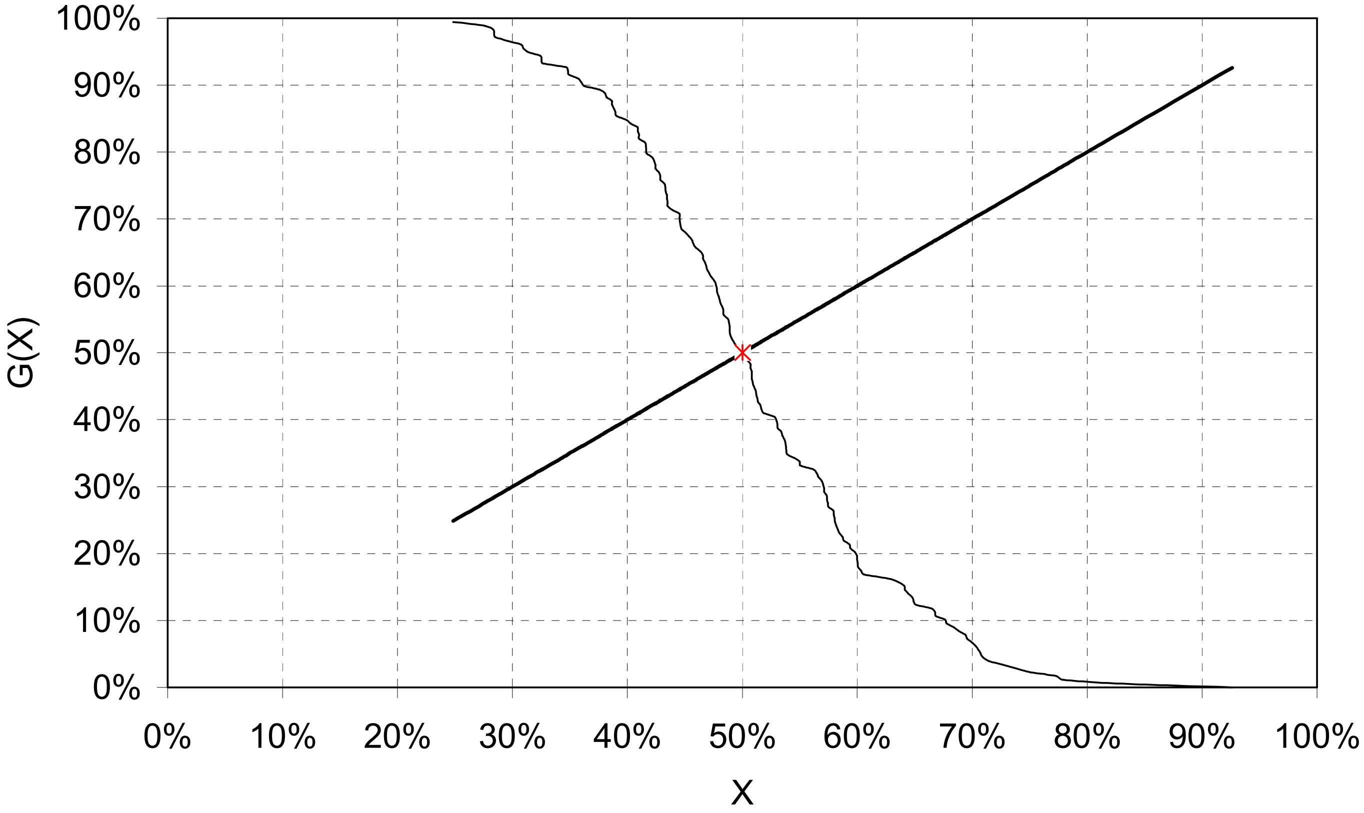
Microbiology



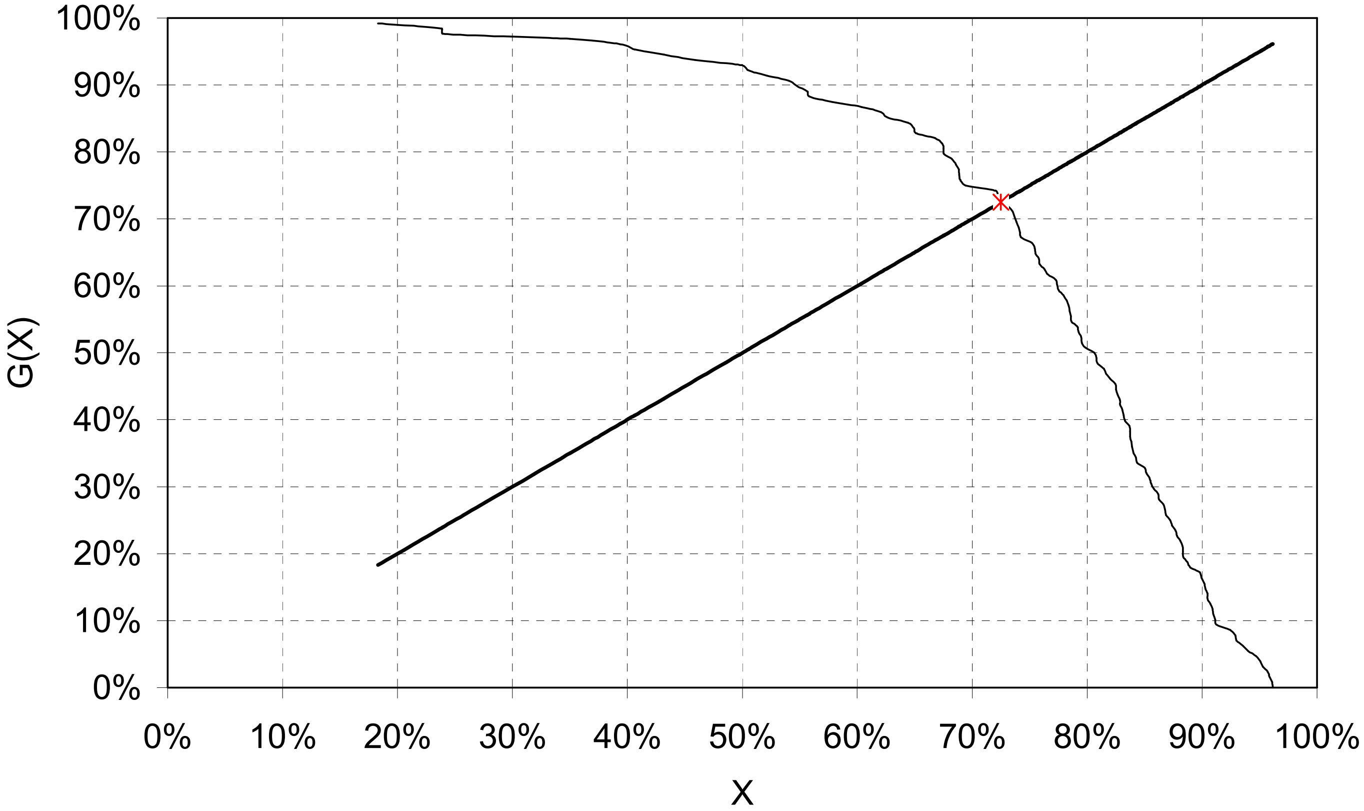
Molecular Biology & Genetics



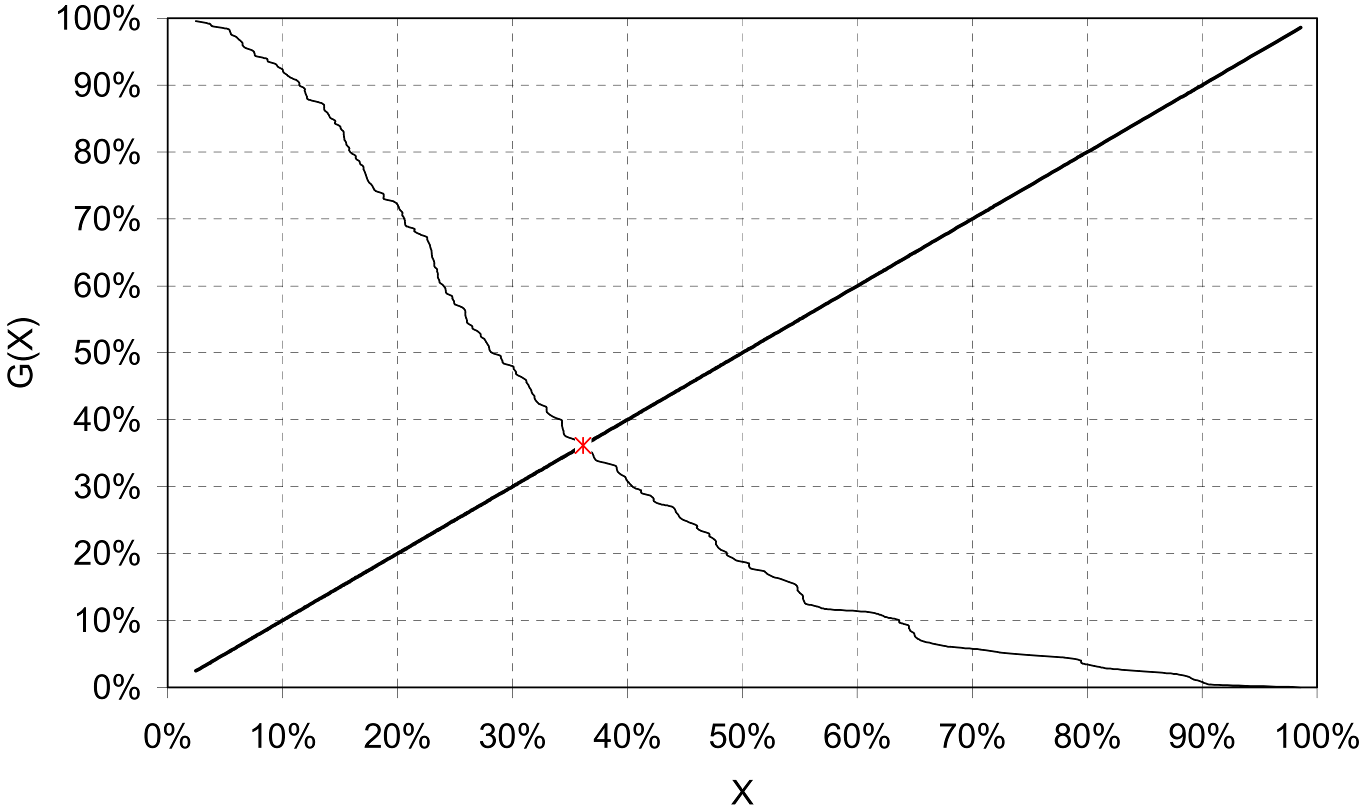
Multidisciplinary



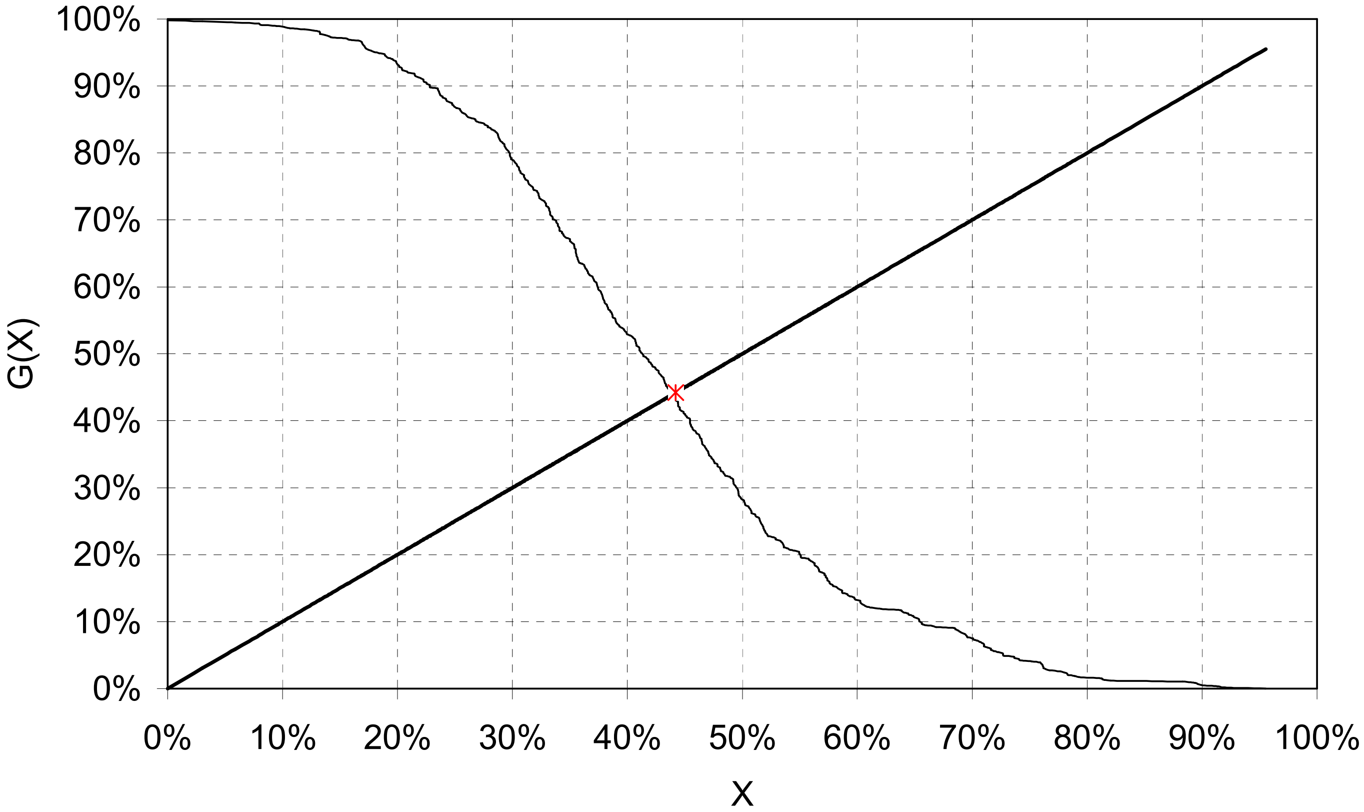
Neuroscience & Behavior



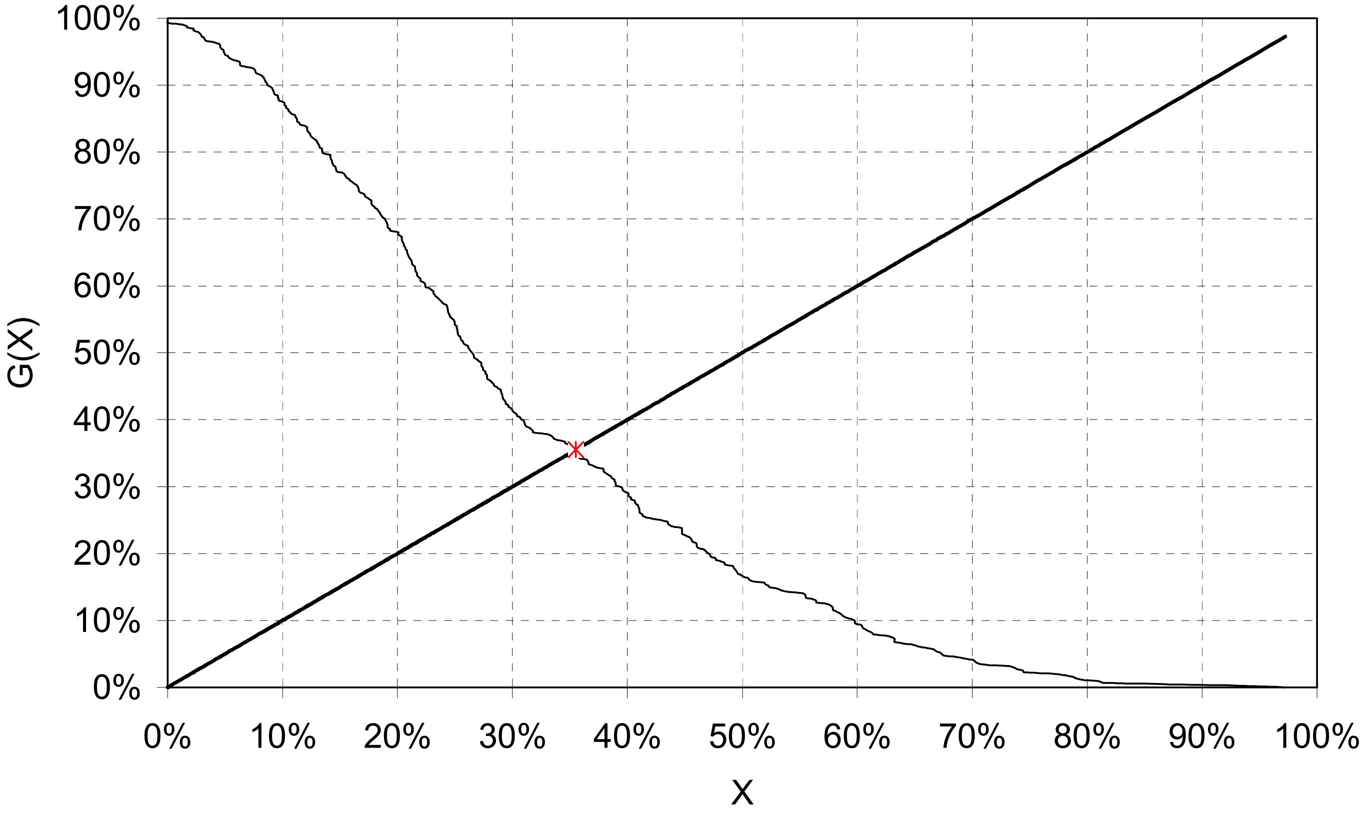
Pharmacology & Toxicology



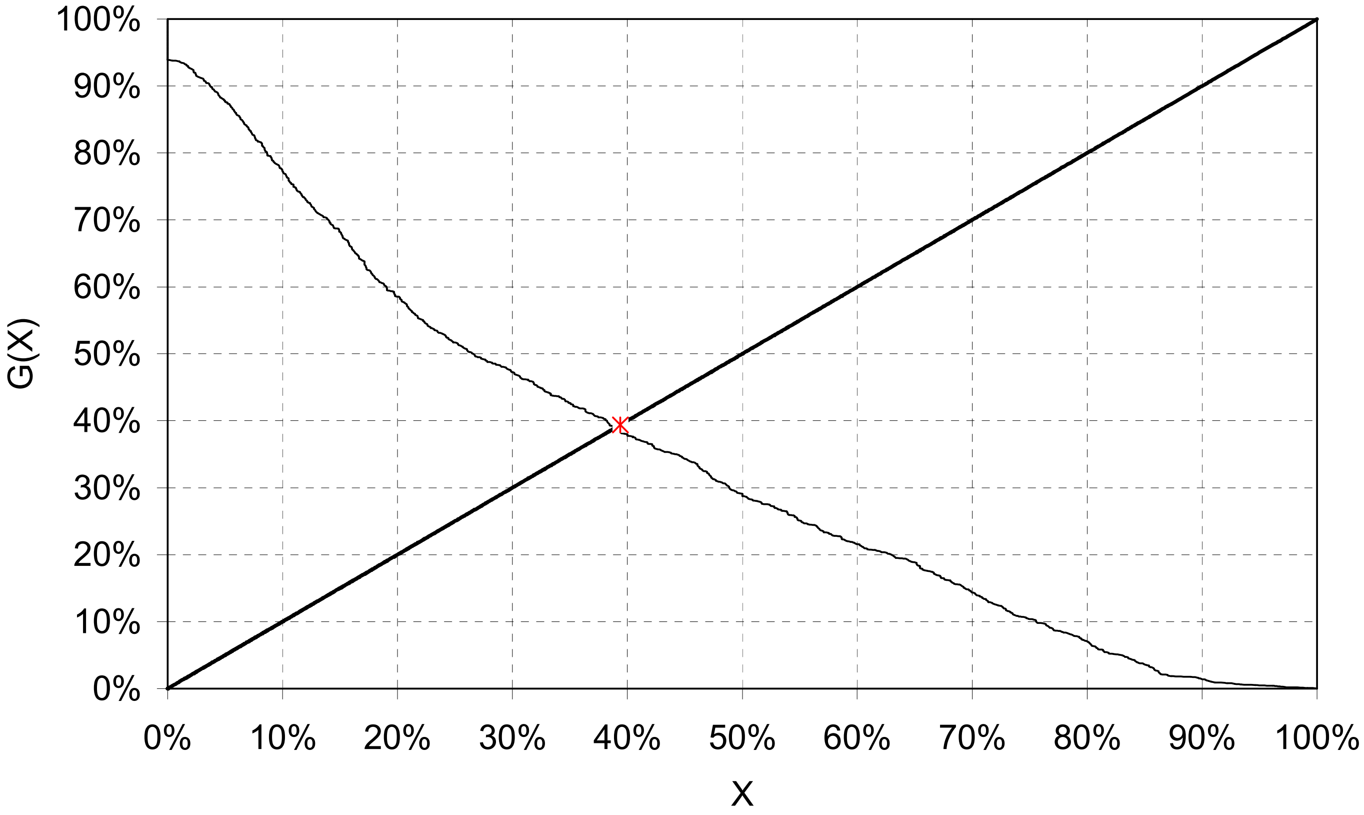
Physics



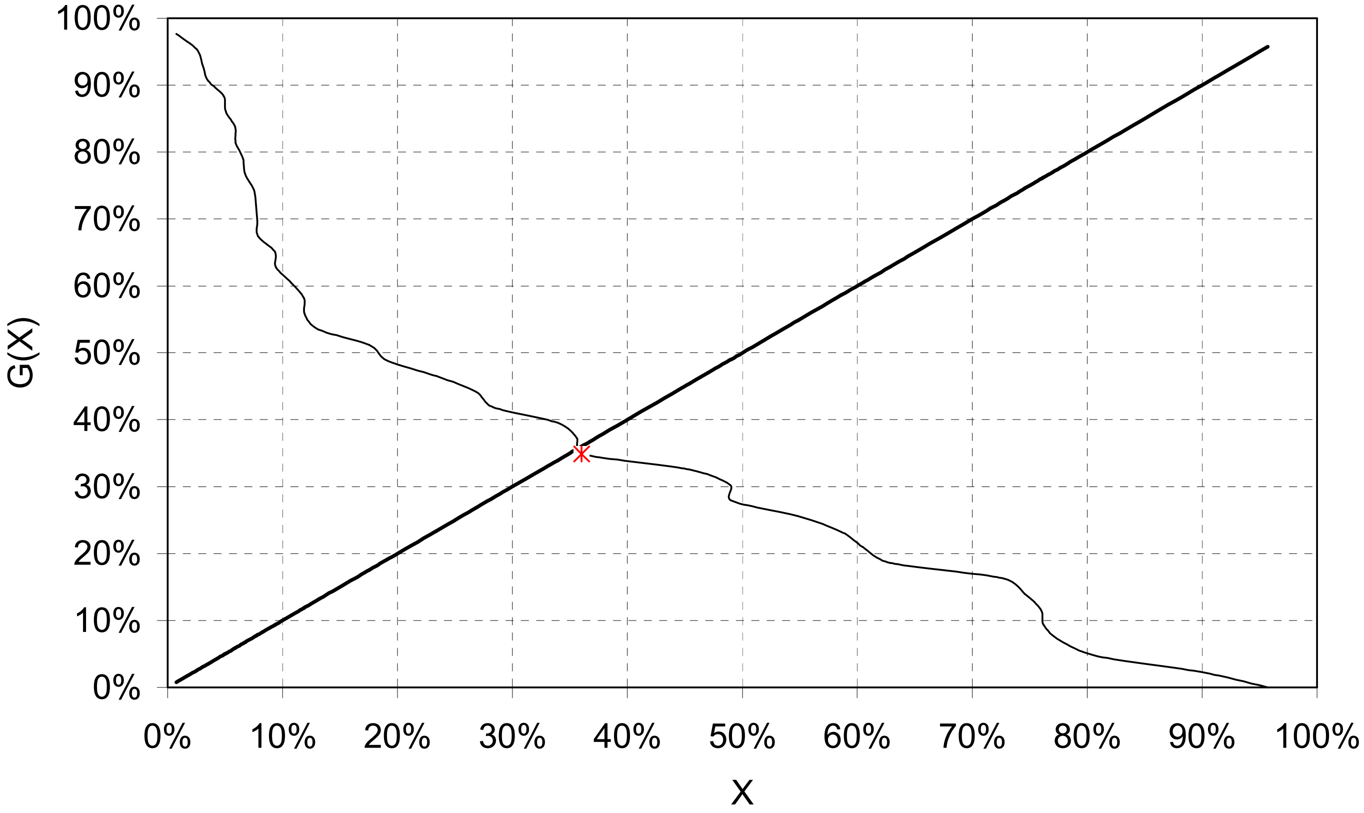
Plant & Animal Science



Psychology/Psychiatry



Social Sciences, general



Space Science