Law, Probability & Risk 14.3 (2015): 175 – 192

DOI: http://dx.doi.org/10.1093/lpr/mgu022

Performance of Likelihood Ratios Considering Bounds on the Probability of Observing Misleading Evidence

Lt. Co . Jose Juan Lucena Mo na, M.Sc.
Stat st cs' Department of the Cr m na st c Serv ce
Genera D rectorate of the C v Guard
Guzman e Bueno st. 110, Madr d (Spa n)
jjlucena@guardiacivil.es

Dan e Ramos Castro, PhD. Joaqu n Gonza ez Rodr guez, PhD.
ATVS B ometr c Recogn t on Group
Escue a Po tecn ca Super or, Un vers dad Autonoma de Madr d
Ca e Franc sco Tomas y Va ente 11, 28049 Madr d, Spa n
{dan e .ramos, joaqu n.gonza ez}@uam.es

Abstract

In this article we introduce a new tool, namely *Limit Tippett Plots*, in order to assess the performance of likelihood ratios in evidence evaluation including theoretical bounds on the probability of observing misleading evidence. In order to do that, we first review previous work about such bounds. Then we derive *Limit Tippett Plots*, which complements Tippett plots with information about the limits on the probability of observing misleading evidence, which are taken as a reference. Thus, a much richer way to measure performance of likelihood ratios is given. Finally, we present an experimental example in forensic automatic speaker recognition following the protocols of the Acoustics Laboratory of Guardia Civil, where it can be seen that *Limit Tippett Plots* help to detect problems in the calculation of likelihood ratios.

Keywords: forensic statistics, likelihood ratio, misleading evidence, Tippett plots, Limit Tippet plots.

1. Introduction

The statistical evaluation of the evidence by means of likelihood ratios (LR from now on) is increasingly proposed for the evaluative interpretation of results in forensic science[1]. In this context, measuring performance of LRs is critical in the process of validating statistical interpretation methods prior to its use in casework[2]. One of the effects that are associated with bad performance of LRs is *misleading evidence*, defined as evidence which has a LR in favour of the wrong proposition, *i.e.* evidence which has a LR ratio higher than one when the defence proposition (H_d) is true or smaller than one when the prosecutor proposition (H_p) is true[3].

As Aitken and Taroni stated[4] *"a change in the odds in favour of the prosecution's proposition, through the value for the evidence different from 1, is a change in the*

[1] C. Aitken, C. Berger, J. Buckleton, C. Champod, J. Curran, A. Dawid, I. Evett, P. Gill, J. Gonzalez-Rodriguez, G. Jackson, A. Kloosterman, T. Lovelock, D. Lucy, P. Margot, L. McKenn, D. Meuwly, C. Neumann, N. Daeid, A. Nordgaard, R. Puch-Solis, B. Rasmusson, M. Redmayne, P. Roberts, B. Robertson, C. Roux, M. Sjerps, F. Taroni, T. Tjin-A-Tsoi, G. Vignaux, S. Willis, G. Zadora, Expressing evaluative opinions: A position statement, Sci. Justice, Guest editorial, Vol. 51 (2011), Issue1, pp. 1-2.

[2] D. Ramos, J. Gonzalez-Rodriguez, Reliable Support: Measuring Calibration of Likelihood Ratios, Forensic Sci. Int., 230 (2013), pp. 156-159.

[3] *Ibid.* See also C. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., J. Wiley&Sons, Chichester (UK), 2004; D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, J. Forensic Sci., Available online, DOI: 10.1111/1556-4029.12233; two references from R. Royall, Statistical Evidence: A Likelihood Paradigm, Chapman&Hall, London (UK), 1997, and On the probability of observing misleading statistical evidence, Journal of the American Statistical Association, 95 (2000), pp. 760-780; and finally, I. Hacking, The Logic of Statistical Inference, Cambridge University Press, Cambridge (UK), 1965.

[4] C. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., J. Wiley&Sons, Chichester (UK), 2004.

probability of the prosecution's proposition". Therefore, the Court could be misled by making wrong decisions if misleading evidence is provided by the forensic examiner.

Misleading evidence is one of the most degrading factors in the performance of LRs, and should be somehow measured in order to evaluate its importance[5]. Moreover, the presence of strongly misleading evidence, namely LR values that support the wrong proposition with a value much greater or much smaller than 1, is even more important and degrading. Therefore, although the LR framework for evidence evaluation is logically correct and does not need to be validated, the implicit or explicit consideration of misleading evidence and strongly misleading evidence of the particular models used for LR computation is of capital importance in order to check the validity of LR procedures prior to its use in casework.

In this work we review and analyse the concept of misleading evidence from the statistics literature. In particular, we focus on theoretical work that derives bounds on the maximum proportion of misleading evidence that can be observed in a set of LRs, which follows the concept of *the probability of observing misleading evidence*[6]. Then, we apply these concepts to regular performance measures in LR-based evidence evaluation, such as Tippett plots. We propose the use of so-called *Limit Tippett Plots*, where the aforementioned bounds on the proportion of cases yielding misleading evidence in a set of LR values is explicitly represented. With *Limit Tippett Plots*, we add valuable information to regular Tippett plots: the violation of the theoretical bounds of misleading evidence is explicitly shown, revealing problems in likelihood ratio

[5] See footnote number 3, D. Ramos et al.

[6] R. Royall, Statistical Evidence: A Likelihood Paradigm, Chapman&Hall, London (UK), 1997.

computation. This represents a major improvement over Tippett plots, where there is not a baseline performance to compare with. Therefore, *Limit Tippett Plots* can be used to detect *e.g.* inadequate statistical models, bad selections of populations, etc. Moreover, a freely available software in Matlab™ has been provided by the authors in order to easily draw *Limit Tippett Plots*, which can be downloaded in http://arantxa.ii.uam.es/~dramos/software.html.

In this article, an experimental example is presented in order to illustrate the usefulness of *Limit Tippett Plots*, following the methods used by the Acoustics Laboratory of Guardia Civil in daily casework. There, *Limit Tippett Plots* are used in order to show that, if some populations are badly selected in order to compute the LR, whether for selecting wrong models, or feeding the models with inappropriate numerical values, the bounds of the probability of misleading evidence are violated. This could not be noticed by using Tippett plots, but it becomes easy to see with *Limit Tippett Plots*. In the experimental example, the protocols and databases followed by Guardia Civil are used, showing that the proposed performance representation is useful in a real operational environment.

This work is organized as follows. Firstly, we review the concept of misleading evidence according to statistical literature. Then, we present the results about theoretical bounds on the probability of misleading evidence contributed by Royall[7]. Then, we propose and describe *Limit Tippett Plots*. The aforementioned experimental example in forensic automatic speaker recognition is then presented. Finally, some conclusions are drawn.

[7] See footnote number 3, R. Royall.

2. Misleading evidence

A definition of *evidence* in statistics can be given according to the so-called *law of likelihood*, as follows[8]:

> *"If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x) / p_B(x)$, measures the strength of that evidence"*.

In forensic science, the LR paradigm exactly follows the law of likelihood, and therefore LRs express the strength of some evidence according to a pair of mutually exclusive propositions.

The next step is to define the concept of probability of misleading evidence. Using common notation in LR-based evidence evaluation, we define $P_p(A) \quad P(A|H_p)$ as the probability of A given H_p, the prosecution proposition. Conversely, we define $P_d(A) \quad P(A|H_d)$ as the probability of A given H_d, the defence proposition. Given a real value $k > 1$, we will call $P_d(LR > k)$ as the probability of having LR values greater than k, being H_d true. Alternatively, $P_p(LR < 1/k)$ is the probability of, being H_p true, having LR values smaller than $1/k$. These probabilities consider the variation in the evidence, and therefore it does not refer to the probability of the LR in a given case with a fixed observation; but the probability that, due to variation in the value of the evidence, the LR will be out of some bounds.

[8] See footnote number 3, I. Hacking.

In the above definition, the value of k determines the strength of the misleading evidence. For instance, in Royall[9] the value from which the strongly misleading evidence is considered is $k = 32$ although this is just a convention assumed by the author.

The LR with continuous data is defined as:

$$V = \frac{f(x \mid H_p)}{f(x \mid H_d)} \quad \frac{f_p(x)}{f_d(x)} \tag{1}$$

where $f(x \mid H_p) \equiv f_p(x) \equiv f_p$, when x is a value of the continuous variable X, and $f(x \mid H_d) \equiv f_d(x) \equiv f_d$, in the equation have originated the observation x of the continuous variable X. Notice that in forensic evaluation of the evidence, the value of x is usually fixed, and the LR is computed from that observation. In this research, we will consider the distribution of x with respect to the observation of misleading evidence, and therefore the probabilities $P_d(LR > k)$ and $P_p(LR < 1/k)$ consider the variation of the possible values of x. As we will see below, misleading evidence can be formally addressed by the use of LR methods to evaluate the evidence in forensic sciences.

In Royall[10] several theoretical bounds for $P_d(V > k)$ and $P_p(V < 1/k)$ are derived. These bounds constitute the motivation of this research. Thus, detecting violations of such theoretical bounds will mean that some problems may have happened in likelihood ratio

[9] See footnote number 6, R. Royall.

[10] See footnote number 3, R. Royall.

calculations. The sources of those problems may be, *e.g.*, a bad selection of models, the use of inappropriate data, etc.

3. Theoretical bounds of the probability of strong misleading evidence

3.1 The universal bound

According to[11], for any given pair of distributions $f_p(x)$ and $f_d(x)$ in (1), there is a bound of $1/k$ on the probabilities of strong misleading evidence, i.e. $P_d(V > k) < 1/k$ and $P_p(V < 1/k) < 1/k$. Therefore, for a given likelihood ratio V computed in operational conditions, if $P_d(V > k) > 1/k$ or $P_p(V < 1/k) > 1/k$, then a problem has happened in LR calculation. More details about the universal bound and its justification can be found in Royall[12]. It is important to note that this universal bound applies to any statistical distribution, no matter its kind or shape, or whether it is parametric or non-parametric.

The universal bound $1/k$ is obviously a function of k, namely the value of strong misleading evidence whose probability is being computed. Figure 1 shows such upper bound as a function of k. It is observed that the universal bound on the probability of misleading evidence will be decreasing with k. This is because if $k' < k$, the values of V representing strong misleading evidence for k are also strong misleading evidence for k', and the opposite might not be true.

[11] See footnote number 4, C. Aitken et al. See also footnote number 3, R. Royall.

[12] See footnote number 3, R. Royall.

Figure 1: Universal bound on the probability of misleading evidence as a function of $k>1$.

3.2 Normal assumption, equal variances

Here we describe a simplified scenario where the distributions of the data are assumed to be normal with equal variances. This distributional assumption founds application in many domains in forensic sciences, particularly when biometric systems are used to analyse the evidence[13]. Topics related with the theoretical behaviour of biometric systems with relation to normal distributions have attracted interest in recent years[14]. Nevertheless, this assumption is theoretically tractable, and therefore it is useful in order to illustrate how tighter bounds than the universal bounds could be obtained in some cases, as it happens in Royall[15].

[13] See footnote number 2, D. Ramos et al. See also J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition, IEEE Transactions on Audio, Speech Lang. Process., 15(7) (2007), pp. 2104-2115, and D. Ramos, Forensic evaluation of the evidence using automatic speaker recognition systems, PhD thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid (Spain), available http://atvs.ii.uam.es (accessed 21st of September, 2012).

[14] D. van Leeuwen, N. Brummer, The distribution of calibrated likelihood-ratios in speaker recognition, Interspeech 2013. Available at http://arxiv.org/abs/1304.1199. See also N. Brummer, D. Garcia-Romero, Generative Modelling for Unsupervised Score Calibration, accepted for ICASSP 2014. Available at http://arxiv.org/abs/1311.0707.

[15] R. Royall, On the probability of observing misleading statistical evidence, Journal of the American Statistical Association, 95 (2000), pp. 760-780.

Let assume a sample X_1, X_2, …, X_n of independent and identically distributed random variables with normal hypothesis-dependent distributions $f_p = N(\mu_p, \sigma)$ and $f_d = N(\mu_d, \sigma)$. From Royall[16], in this case the upper bound on the probability of strong misleading evidence is given by the so-called *bump function*:

$$P_d(V > k) = P_p\left(V < 1/k\right) < \frac{c}{2} \frac{\log k}{c} \qquad (2)$$

where

- Φ denotes the standard normal cumulative density function.
- $c = \sqrt{n}$ is the distance of the mean values, $\Delta = |\mu_1 - \mu_2|$, expressed in standard errors.

Figure 2: The bump function is the upper bound on the probability of strong misleading evidence for normal distributions with equal variances as a function of c, the difference of the means measured in standard errors.

The bump function is a function of c and k, and therefore it can be represented as a function of both variables, as it is shown in Figure 3. From such representation, it can be seen that the bump function always decreases with k for fixed values of c. As in the case of the universal bound, this makes sense because increasing values of k decreases the cases where the value of V may exceed k. However, another effect is remarkable,

[16] See footnote number 3, R. Royall.

because for different values of c the convexity of the bound of the probability of strong misleading evidence changes significantly. This is observed in Figure 4.

Figure 3: Bump function as a function of c and k.

Figure 4: Upper bound on the probability of strong misleading evidence assuming normal distributions with equal variances. Represented as a function of k and for different values of c, the difference of the means measured in standard errors.

The maximum value of the bound on the probability of strong misleading evidence for a given k significantly varies with c, because the maximum value of the bump function is $\left(\sqrt{2\log k} \right)$ for $c = \sqrt{2\log k}$. This result allows us to plot the *maximum* upper bound on the probability of strong misleading evidence for *any* pair of normal distributions with equal variances, regardless of the difference in their means and as a function of k. With this result, it is not needed to know the means of the hypothesis-dependent distributions in order to limit the probability of strong misleading evidence. Figure 5 shows the behaviour of such maximum upper bound with respect to the ones for different values of c as a function of k. We see that the difference among such bounds is sometimes quite important. The bound also has a decreasing behaviour with k, as expected.

Figure 5: Maximum upper bound on the probability of strong misleading evidence assuming normal distributions with equal variances. Represented as a function of k and for different values of c, the difference of the means measured in standard errors.

The bound on the probability of strong misleading evidence for normal distributions with equal variances is much more restrictive than the universal bound. This is because a given distribution cannot have a looser bound than the universal bound for all distributions. Figure 6 shows the comparison between the universal bound and the normal bound under the equal-variance assumption, both for the maximum value of the normal bound, and for different values of *c*.

Figure 6: Comparison between the universal bound for the probability of misleading evidence and the bounds for normal distributions with equal variances.

3.3 Normal assumption, general case

To our knowledge, the solution for the bound on the probability of strong misleading evidence for normal distributions with different means and variances is not easily tractable using analytical methods. An example of this can be found in Alcon[17]. Also, a simulated approach from the input parameters by randomly generating samples in order to determine the probability of misleading evidence can be used in order to determine the bounds in this case. However, for the purpose and contribution of this article, this is outwith the scope of this work.

[17] M.J. Alcon, J. Amador, I. Caceres, P. Giron, C. Nieto, T. Perez, Estimation of the probability of misleading evidence in the case of normal populations with known different variances, Technical Report, Universidad Complutense de Madrid, 2007. Available at http://atvs.ii.uam.es/files/2009_TR_Alcon.pdf.

4. Application to the performance assessment of LR-based evidence evaluation methods

4.1 Empirical performance assessment of LR methods

The proposed methodology for performance assessment of LR-based evidence evaluation methods is empirical, *i.e.*, it requires the availability of a *validation* database that will be used for building simulated real cases[18]. Thus, from those simulated cases, the performance can be measured. However, in other forensic disciplines, this kind of validation database may not be available, or may be deemed non-representative, and therefore this empirical performance assessment methodology is not recommended. Anyway, data availability and representativeness for empirical performance evaluation of LR methods remains an open problem, which is outwith the scope of this work.

4.2 Tippett plots

Tippett plots[19] are a valuable tool for the assessment of LR-based methods for evidence evaluation, empirically representing the cumulative proportions of the LR in an experimental set depending on which proposition was true (H_p or H_d). Figure 7 shows an example of Tippett plots. Important performance measures which can be seen in the plots are the rates of misleading evidence at the value of $V = 1$. Moreover, any rate of misleading evidence can be seen from Tippett plots for any value $k > 1$ or $1/k < 1$.

[18] See footnote number 3, D. Ramos et al.

[19] I. Evett, J. Buckleton, Statistical analysis of STR data, Advances in Forensic Haemogenetics, Springer-Verlag, Heildeberg 6 (1996), pp. 79-86.

Figure 7: Example of Tippett plots. The rates of misleading evidence at the value V=1, represented in the title of the figure, are the proportion of cases where the value of the LR supports the wrong hypothesis, for each hypothesis H_p (right curve) and H_d (left curve).

4.3 Misleading evidence and Tippett plots: *Limit Tippett Plots*

Tippett plots can be interpreted as representing the probability of observing a $\log_{10}$-LR value greater than a given value in the x axis. Such number may be a given value of misleading evidence previously stated, say k. Therefore, we can represent the bounds derived in Section 3 for the probability of strong misleading evidence in Tippett plots. An example of that is shown in Figure 8, where it can be seen that in one of the cases the computed Tippett plots exceed the theoretical bounds.

(a) (b)

Figure 8: *Limit Tippett Plots*, *i.e.*, Tippett plots with the corresponding universal bounds on the probability of misleading evidence as a reference. In (a) both curves lay within the theoretical bounds. However, in (b) such bounds are exceeded when H_p is true.

Note that the bounds when H_d is true are essentially the same that were derived in Section 3 (see e.g. Figure 1 for the universal bound). For the bounds when H_p is true, just note that, for the universal bound:

$$P_p(V < 1/k) = P_p(\log V < \log k) = 1 \quad P_p(\log V > \log k) \tag{5}$$

This allows plotting the bound when H_p is true by simply flipping the bound when H_d is true over log $V = 0$ and obtaining the complementary.

The bounds on the probability of strong misleading evidence for normal distributions with equal variances can also be drawn in Tippett plots. It still holds that such bounds are equal for $P_d\,(V > k)$ and $P_p\,(V < 1/k)$. Figure 9 shows several examples of Tippett plots including such bounds for different values of c, and for the maximum bound according to the bump function.

For Figures 9(a), 9(b) and 9(c), Tippett plots have been generated according to the distribution of each bound which is represented (e.g., a distribution having $c = 2$ has generated the example Tippett plots in Figure 9(b)). It can be observed that, again, the bounds on the probability of strong misleading evidence are much more restrictive for the normal distribution than the universal bound, as it was argued in previous discussion. Moreover, as k approaches the value 1, this difference increases. A special mention deserves the case for $c = 1$ (Figure 9(a)), where the example Tippett plot almost exactly reaches the theoretical bound. The maximum value of the bound, as expected, is looser than for any fixed-c bound. The bounds without their corresponding Tippett plots are represented in Figure 10 for illustration.

(a) (b)

(c) (d)

Figure 9: *Limit Tippett Plots* with the corresponding bounds on the probability of misleading evidence as a reference, for the case of normal distributions with equal variances. Different values of c are represented, namely c=1 (a), c=2 (b) and c=3 (c). The maximum bound with respect to c for all k is shown in (d).

Figure 10: *B*ounds on the probability of misleading evidence for the case of normal distributions with equal variances.

Tippett plots exceeding the theoretical bounds indicates that there are problems in the LR computation process, either *e.g.* because of the selection of the statistical models, or because the use of inappropriate data. Therefore, the use of *Limit Tippett Plots* allows warning about those problems, which may not be seen in Tippett plots, and fostering the improvement of the LR computation methods themselves. Although the way in which the improvements of methods depends on many factors (databases, statistical models, type of data, etc.), detecting violations of the theoretical bounds can be applied to any method computing LR values, and therefore the usefulness of *Limit Tippett Plots* is general for any LR-based discipline.

5. Experimental example: forensic automatic speaker recognition

In this section we describe an experimental example in order to illustrate the usefulness of *Limit Tippett Plots* to detect problems in statistical LR computation methods. The example is contextualized as an experiment to measure performance of a forensic automatic speaker recognition following the procedures used in the Acoustics Laboratory of Spanish Guardia Civil (AL-GC from now on). These procedures have been recently accredited by the Spanish Accreditation Body (ENAC, www.enac.es).

It is worth noting that, although in this article we use the example of forensic automatic speaker recognition to illustrate the use of the proposed methodology, it can be applied to any forensic discipline where LRs are used as the expression of the strength of the evidence, and where an empirical performance assessment is in order. Therefore, the following experimental section can be understood without previous knowledge in forensic automatic speaker recognition. In the experimental example, we follow the procedures of data selection in forensic automatic speaker recognition in Guardia Civil, and then we assume that the validation database used in that example is representing different scenarios in casework, mainly defined by the conditions of the recordings (mainly related to the transmission channel of the speech signal). Therefore this empirical experiment can be seen as representative of future performance in casework where the conditions of the evidential recordings may fit those of the considered scenarios.

5.1 Context and Motivation

AL-GC regularly submits forensic reports to court based on a likelihood ratio methodology, for which an automatic forensic speaker recognition system is used. This system is described in Kenny et al.[20] LRs are computed by the system using procedures previously reported in Gonzalez-Rodriguez et al.[21]

In order to compute LRs, a population of recordings from speakers is needed, and therefore a database must be used to select that population. The use of an appropriate

[20] Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P, A Study of Inter-Speaker Variability in Speaker Verification, IEEE Transactions on Audio, Speech Lang. Process., 16(5) (2008), pp. 980-988.

[21] See footnote number 13, J. Gonzalez-Rodriguez, P. Rose et al.

population database is important in order to compute the LR. This case study shows how *Limit Tippett Plots* are useful in order to detect violations of the universal bounds of misleading evidence, and therefore problems in likelihood ratios due to a bad selection of the population.

The selection of a population in LR-based evidence evaluation should consider the circumstances of the case and the propositions defined for the case. In this experimental example, there are also issues concerning the conditions of the recordings used in populations (*e.g.* noise level, transmission channel, speaker emotional state, etc.). The protocols in Guardia Civil are strict about those conditions, and therefore, the cases where the available populations are not appropriate for the conditions of the evidential recordings are typically rejected by the laboratory.

5.2 Forensic Scenario

The forensic scenario in which the case study is described represents typical conditions in AL-GC casework. In particular, two recordings of speech (one questioned and the other one coming from a given suspect) are to be compared, in order to yield a LR using the automatic system. The two recordings are known to come from different sessions, and from different moments in time. Both recordings are known to come from digital wire-taps of GSM mobile conversations, obtained in accordance to operational procedures in Guardia Civil.

5.3 Validation Database

In order to generate the LR values that will be represented in *Limit Tippett Plots*, a so-called *validation database* will be used[22] simulating the conditions found in casework. The aim is to measure the performance of the method in use in the conditions in which it will be used in the case at hand. In the aforementioned forensic scenario, the *Ahumada-IV* database is used by AL-GC as a validation database. The database consists of 91 male speakers speaking in different dialects of Spanish from Spain. The identity of the speaker is known, and the utterances contain spontaneous speech presenting high variability in recording times, acoustic environment (there are indoors and outdoors recordings), dialect, noise, etc. All the recordings are acquired using the wire-tapping system used for obtaining the recordings in the described forensic scenario. For each speaker, a *long* recording of roughly 120s and about 5 *short* recordings of 20s are available. All recordings are from different sessions.

The experimental protocol to measure performance consists of computing a single LR value for each comparison between each *long* recording and each of the 20s *short* recordings. For each comparison, a LR will be generated. A comparison is denoted a *true-H_p* comparison if the *long* and the *short* recordings come from the same speaker. Conversely, a comparison will be denoted a *true-H_d* comparison if the *long* and the *short* recordings come from different speakers. Also, the LR values will be called *true-H_p LR values* or *true-H_d LR values* depending on whether they are respectively computed for a *true-H_p* or a *true-H_d* comparison. Using this protocol, a total of 442 *true-H_p* LR values and a total of 39780 *true-H_d* LR values are obtained. These LR values will be used to draw *Limit Tippett Plots*.

[22] See footnote 3, D. Ramos et al.

In order to illustrate the effect of the population selection in the performance of the LR values, two different experiments will be conducted, each one using a different population database.

- In the first experiment, namely Ahumada-IV-Population experiment, a population that presents the same conditions as the validation database is used, which is highly desirable in forensic automatic speaker recognition[23]. This is the population used in the usual protocol followed by AL-GC in real cases for this forensic scenario where all the speech in the case comes from digital wire-tapping.
- In the second experiment, namely Ahumada-III-Population experiment, the population used to compute LR values presents different conditions that the validation database. In this case, it is expected that the models obtained from the population will be not so adequate for the data in the validation database, and therefore we expect worse performance in the Ahumada-III-Population experiment than in the Ahumada-IV-Population experiment.

5.4 Population Databases

Here we describe the databases used as reference population in both experiments presented.

- Ahumada-IV-Population: this population database comes from recordings from the Ahumada-IV database itself. From all the *long* recordings in the database, a total of 35 recordings are used as the population. In each comparison, the population

[23] See footnote 13, D. Ramos, PhD thesis.

selected did not include the recording corresponding to any of the speakers in the comparison. This population presents the same conditions as the validation database, because it is actually extracted from the same database, but using recordings from different speakers than the ones in the case. Therefore, it consists on recordings of 120s, obtained using digital wire-tapping.

- Ahumada-III-Population: this population database comes from the so-called *Ahumada-III* database, described in Ramos et al.[24] The conditions of this database are different from the conditions in Ahumada-IV. In Ahumada-III, the recordings are also originated from GSM mobile conversations, but they are recorded over magnetic tape, not digitally. This emulates the procedures used by AL-GC before 2005, year in which the digital wire-tapping system started operation in Spain. Ahumada-III is a database of real forensic cases, for which AL-GC has been allowed to use it for forensic purposes by the Spanish law. For the Ahumada-III-population experiment presented here, 69 recordings of about 120s have been used as the population. However, although the speech in Ahumada-III database comes from real forensic cases, there is a substantial difference between the conditions in the Ahumada-III and Ahumada-IV databases, because the recording technique used in both databases is also different. This suggests that the performance in the Ahumada-III-population experiment will be worse than in the Ahumada-IV-population experiment.

5.5 Results

[24] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, J.J. Lucena-Molina, Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-case database in Spanish, in Proc. of Interspeech 2008, pp. 1493-1496, 2008.

In this section, we present results that show the usefulness of *Limit Tippett Plots* as tools for measuring performance.

Figure 11 (a) shows the *Limit Tippett Plots* of the Ahumada-IV-Population experiment. As it can be seen, the Tippett plots do not violate the theoretical universal bounds for the probability of misleading evidence. According to Ramos[25], this can be partially because a proper population has been used to model between-source variability, since the conditions of the population are the same as the conditions of the cases in the validation database. Remarkably, with *Limit Tippett Plots* the adequacy of the LR values to the theoretical bounds of misleading evidence is made explicit, and consequently it adds valuable information to regular Tippett plots.

(a) (b)

Figure 11: *Limit Tippett Plots* for the Ahumada-IV-population experiment (a) and for the Ahumada-III-population experiment (b).

On the other hand, Figure 11 (b) shows the *Limit Tippett Plots* of the Ahumada-III-Population experiment. As it can be seen, in this case the Tippett plots violate the theoretical universal bound for the probability of misleading evidence when H_d is true: it can be seen that the proportion of *true-*H_d LR values greater than $10^{1.6}$ is greater than $10^{-1.6}$, and therefore the bound is exceeded. This can be attributed to a bad selection of the relevant population for LR computation, because the conditions of the Ahumada-III-population (GSM recorded over magnetic tape) are very different than the conditions of

[25] See footnote 13, D. Ramos, PhD thesis.

the validation database, Ahumada-IV (digital GSM wire-taps). Moreover, the Ahumada-IV-population experiment and the Ahumada-III-population experiment only differ in the population in use, and therefore the loss of performance can be attributed to this fact. Therefore, the violation of the universal bounds constitutes a warning about the adequacy of LR computation methods, in this case because the selection of the population was inadequate. This information can be seen in *Limit Tippett Plots* because of the explicit representation of the universal bound, but not in regular Tippett plots. In order to better illustrate the aforementioned observations, a detailed comparison of the *Limit Tippett Plots* for both experiments is shown in Figure 12, where the region of the plots where the universal bound is violated has been magnified.

Sample size effects may have an influence in the violation of the theoretical bounds in *Limit Tippett Plots*. In an experimental setup, the smaller the size of the validation database, the bigger the expected variability in the figures of performance (*e.g.* rates of misleading evidence in Tippett plots). As the database increases in size, those figures will be more robust, and their credibility will be better. This also applies to the exact point where the theoretical bounds of misleading evidence are violated. Therefore, an increase in the size of the database is always desirable. In the current example, speech databases are always very costly to be acquired. Although there are initiatives such as the NIST Speaker Recognition Evaluation campaigns[26], the protocols of any forensic laboratory need to have an available database that is as much similar to the casework as possible, and this is not always an easy task. This is why studies of the influence of data

[26] See http://www.itl.nist.gov/iad/mig//tests/sre/.

sparsity and other effects are extremely necessary in many fields[27], as well as constant data collection in forensic conditions.

Figure 12: comparison of the *Limit Tippett Plots* for the Ahumada-IV-population and the Ahumada-III-population experiments, where the region where the universal bound for misleading evidence is exceeded has been magnified.

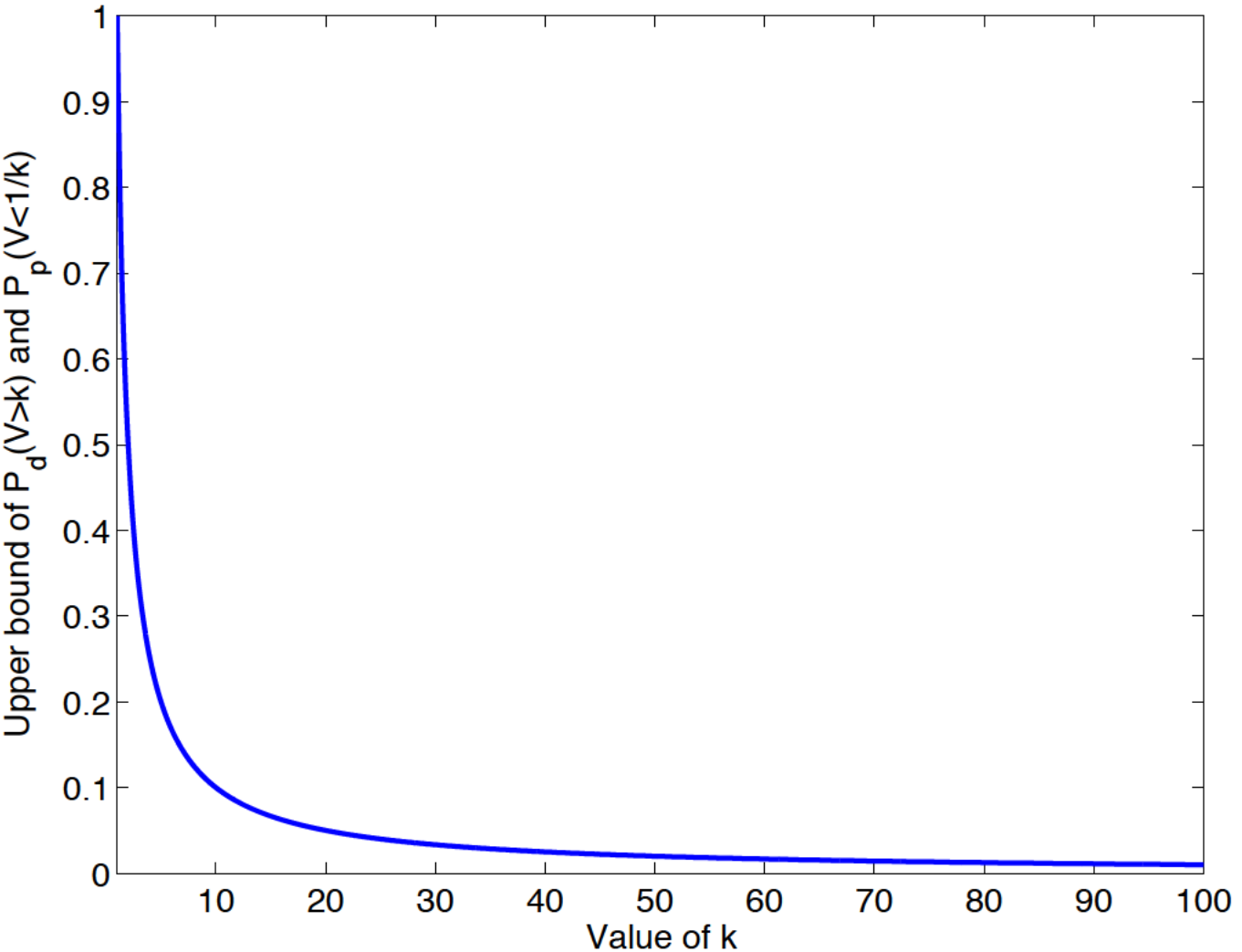
6. Conclusions

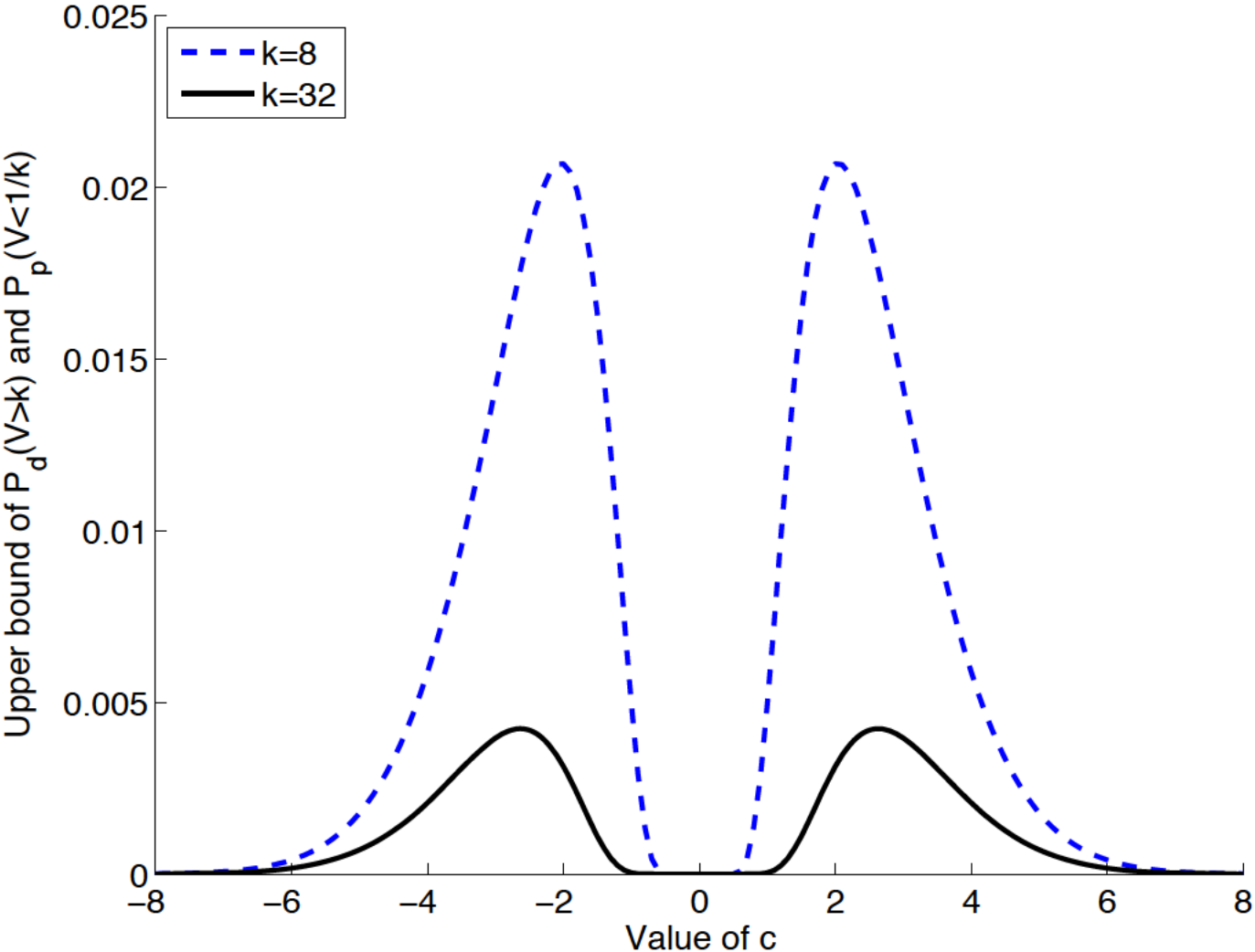
This work contributes with a novel tool to detect likelihood ratios presenting too high misleading evidence, by the proposal of so-called *Limit Tippett Plots*. Based on previous contributions[28], the behaviour of the bounds of the probability of observing strong misleading evidence in a set of LR values has been described. Three main conditions have been examined: unknown data distributions (universal bounds), and normal data distributions. Then, Tippett plots including the theoretical bounds of the probability of misleading evidence, namely *Limit Tippett Plots*, have been proposed for detecting an anomalous behaviour in LR values. This proposal is useful for examiners conducting experiments with the aim of measuring performance of LR values. A Matlab™ software has been made available by the authors, in order to easily draw *Limit Tippett Plots*. It can be downloaded in http://arantxa.ii.uam.es/~dramos/software.html.

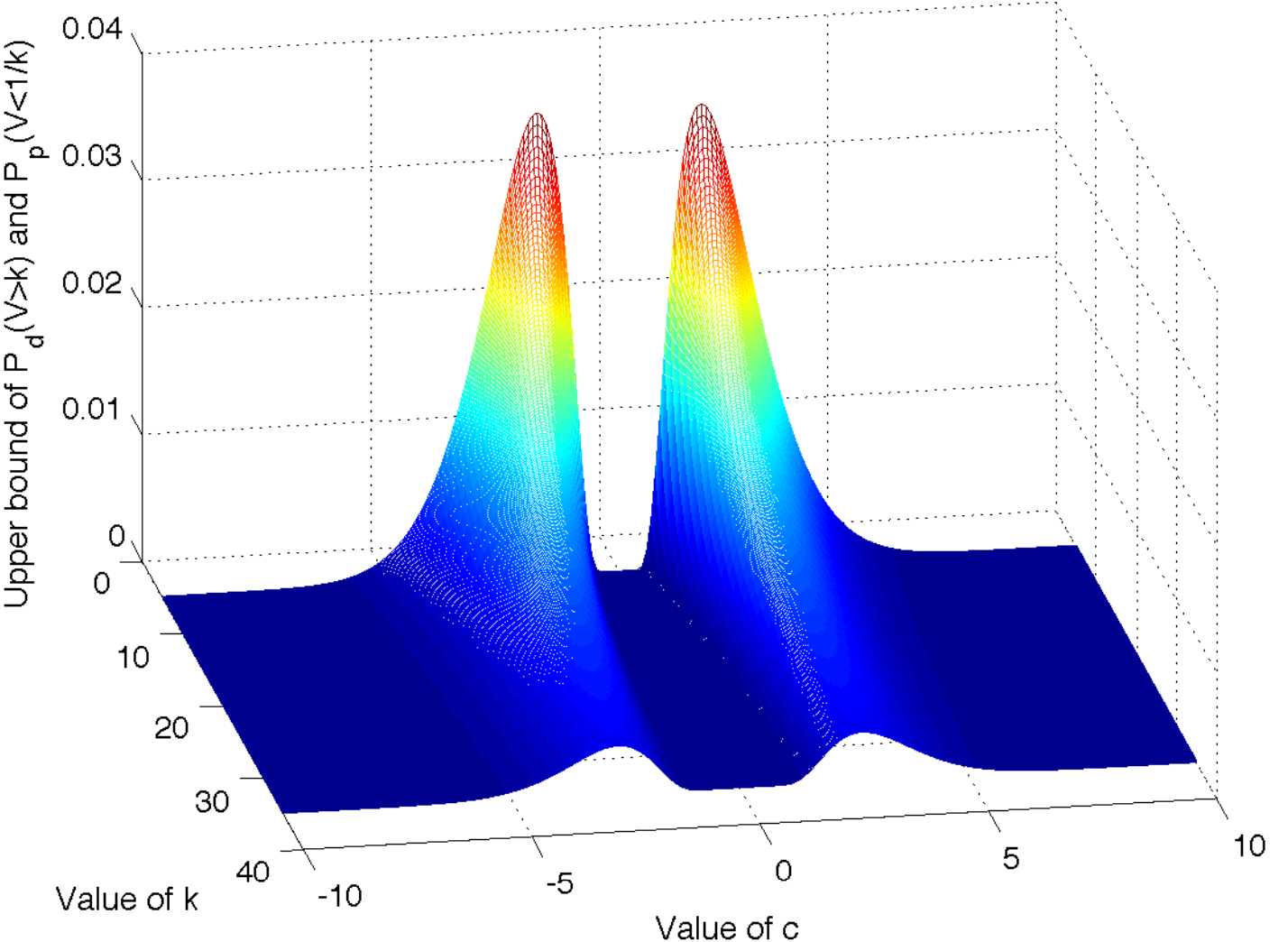
[27] R. B. Tapias, "Sistemas forenses de reconocimiento automático de locutor: determinación y análisis de sus valores más críticos", Proyecto Fin de Carrera, ETSIT UPM, Julio 2005.

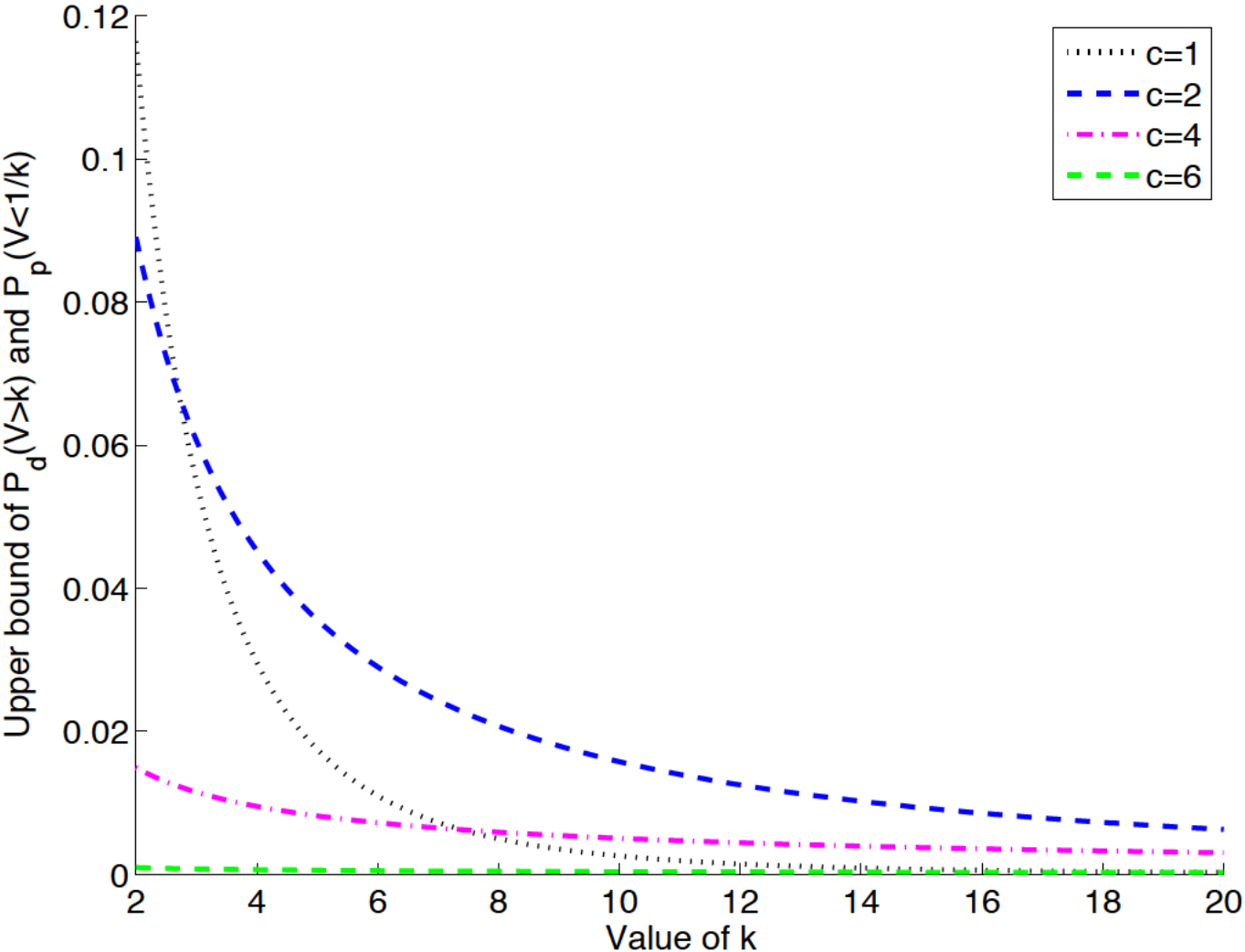
[28] See footnote number 3, C. Aitken et al. and R. Royall.

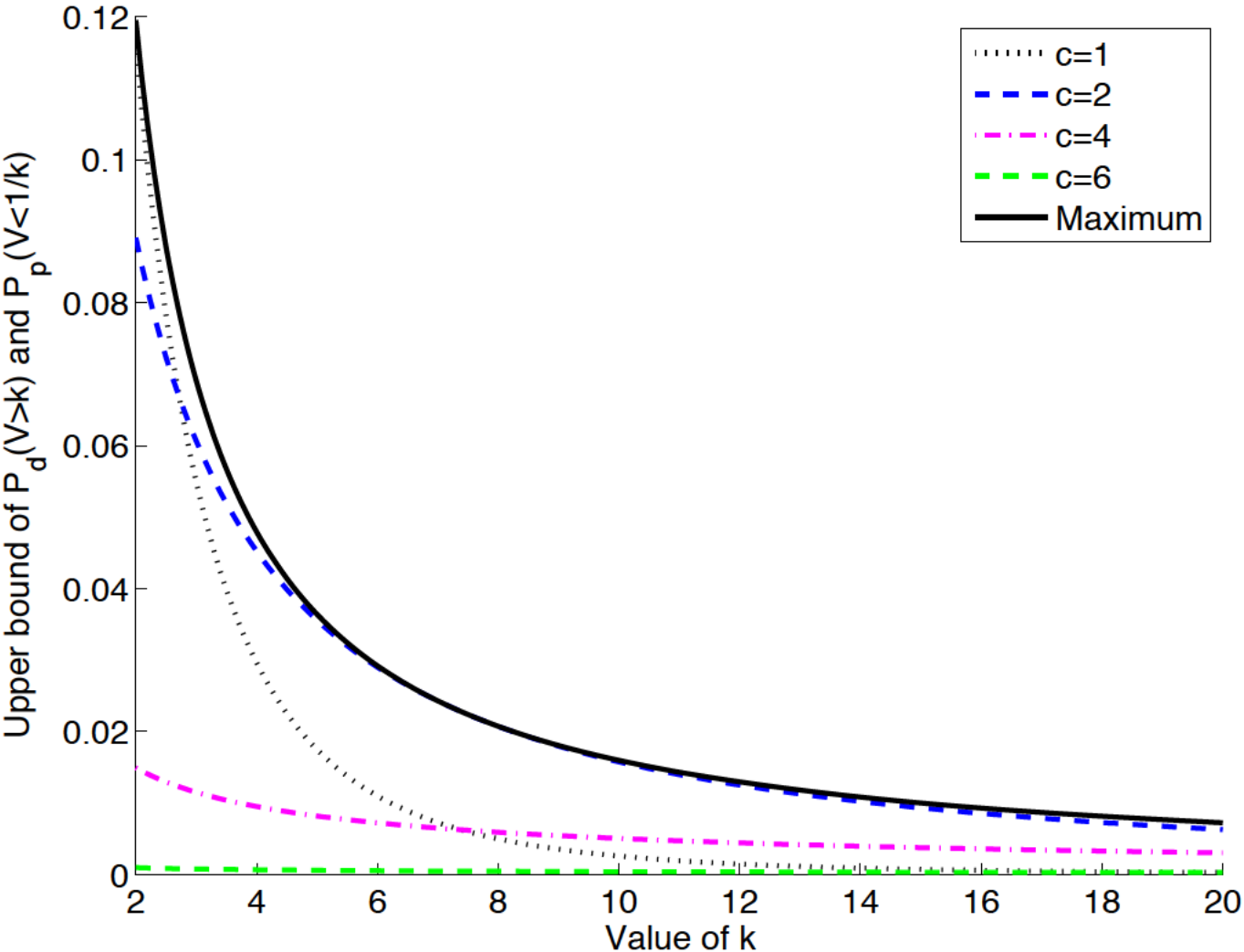
Finally, an experimental example in forensic automatic speaker recognition has been presented. There, the protocols followed by the Acoustics Laboratory of Guardia Civil are followed in order to measure performance in different scenarios, and *Limit Tippett Plots* are used in order to detect problems due to an inadequate selection of the population used to produce LR values.

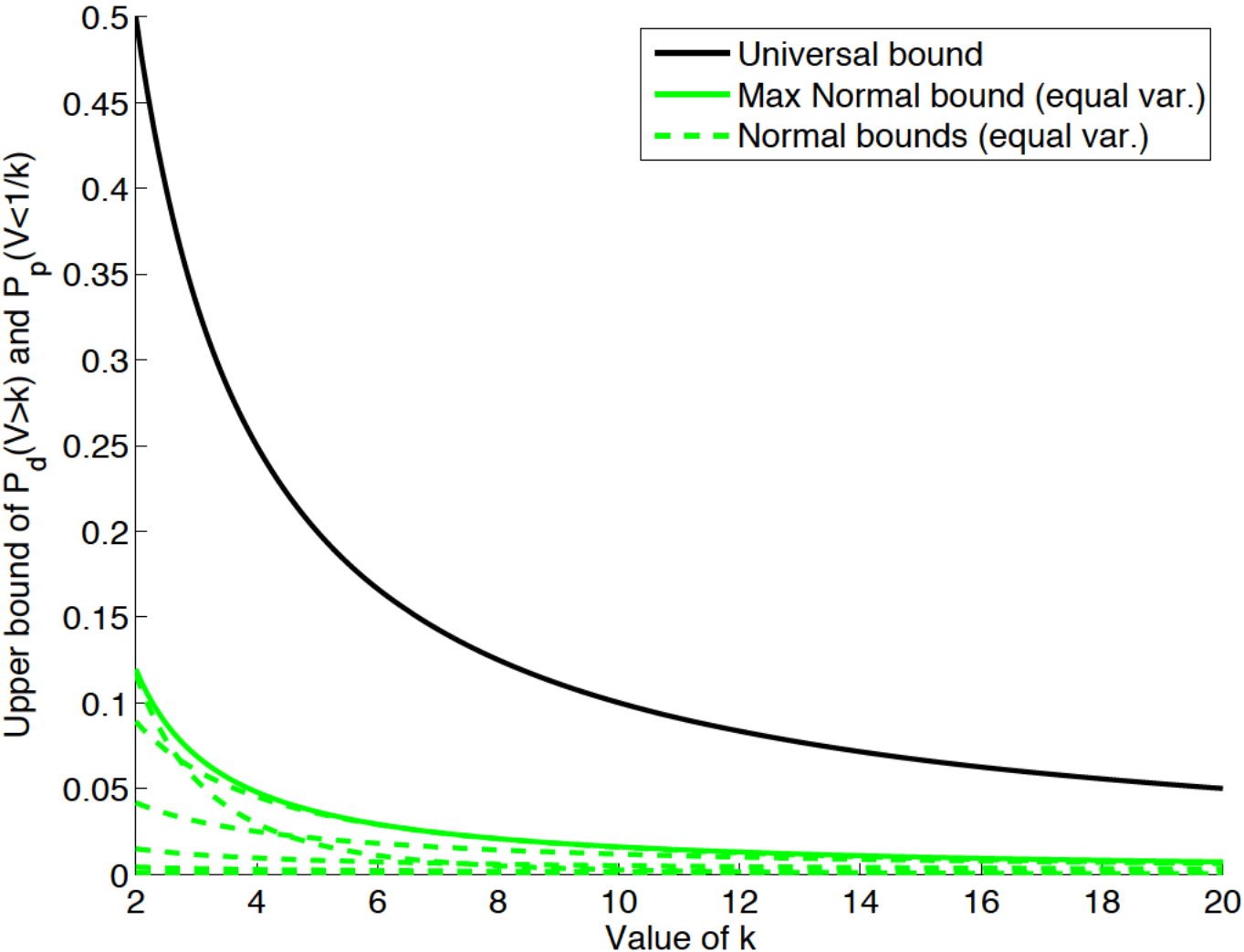

Upper bound of $P_d(V>k)$ and $P_p(V<1/k)$
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
90
100
Value of k

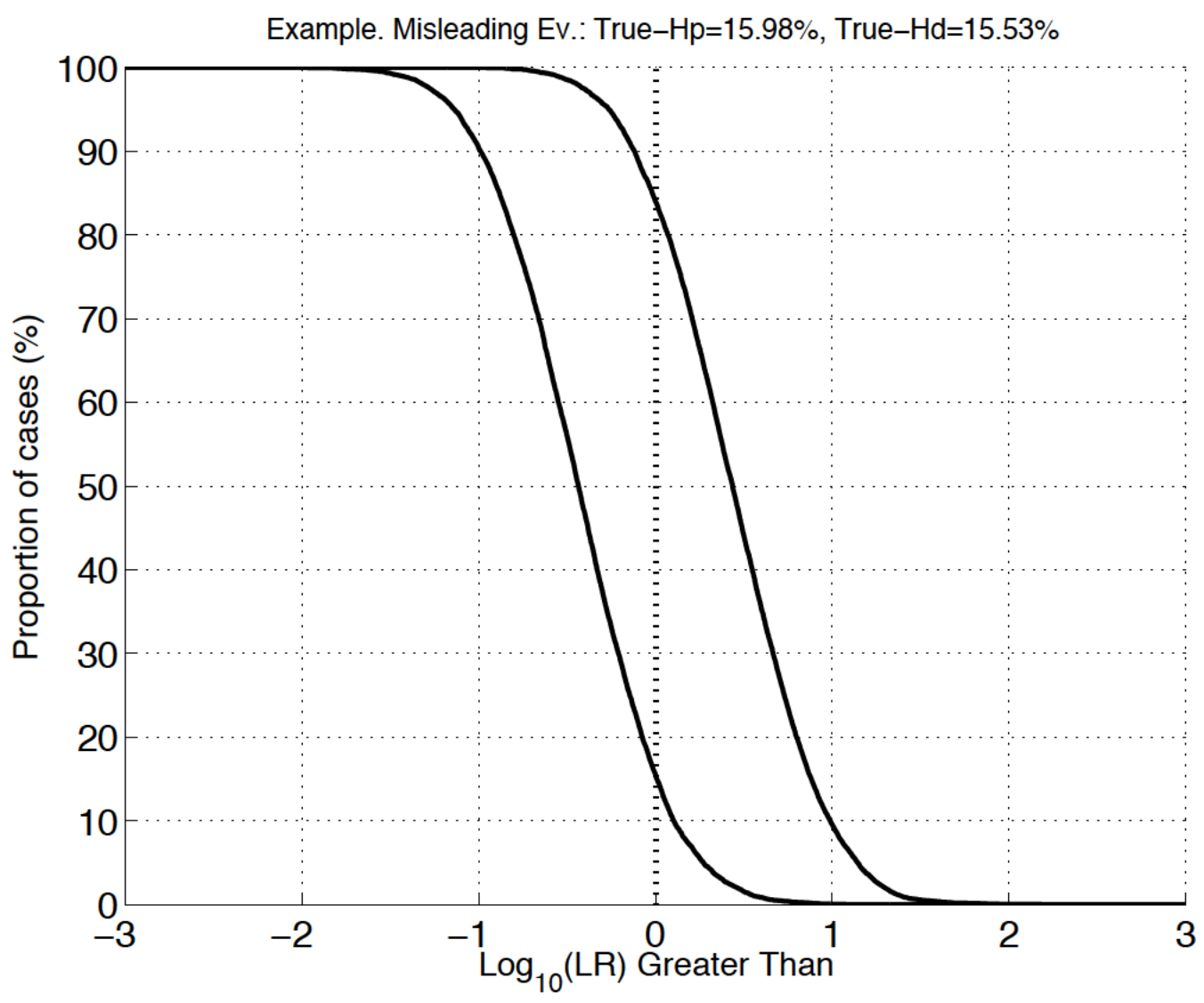

k=8
k=32
0.025
0.02
0.015
0.01
0.005
0
Upper bound of $P_d(V>k)$ and $P_p(V<1/k)$
−8
−6
−4
−2
0
2
4
6
8
Value of c

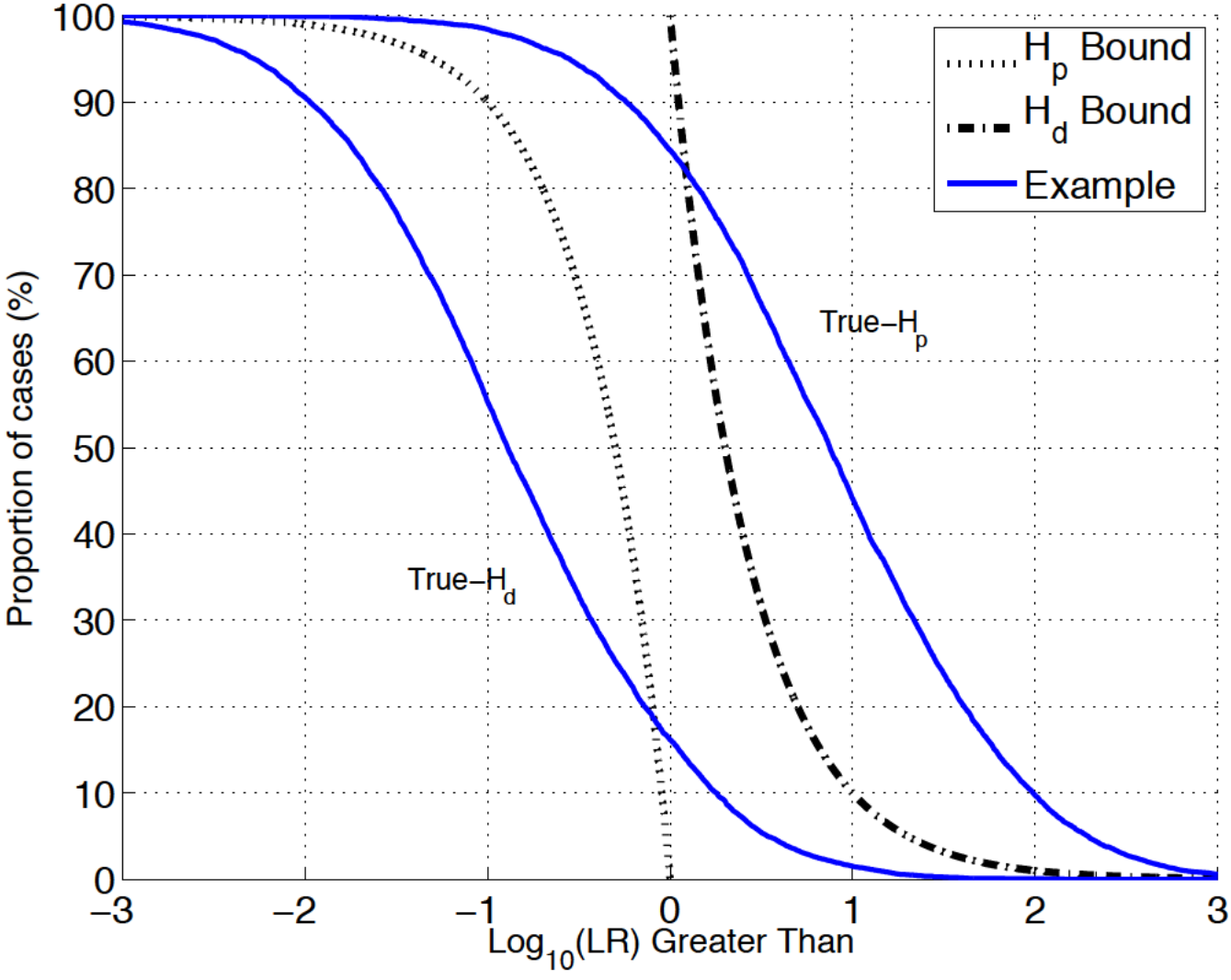

0.04
0.03
0.02
0.01
0
Upper bound of $P_d(V>k)$ and $P_p(V<1/k)$
0
10
20
30
40
Value of k
-10
-5
0
5
10
Value of c

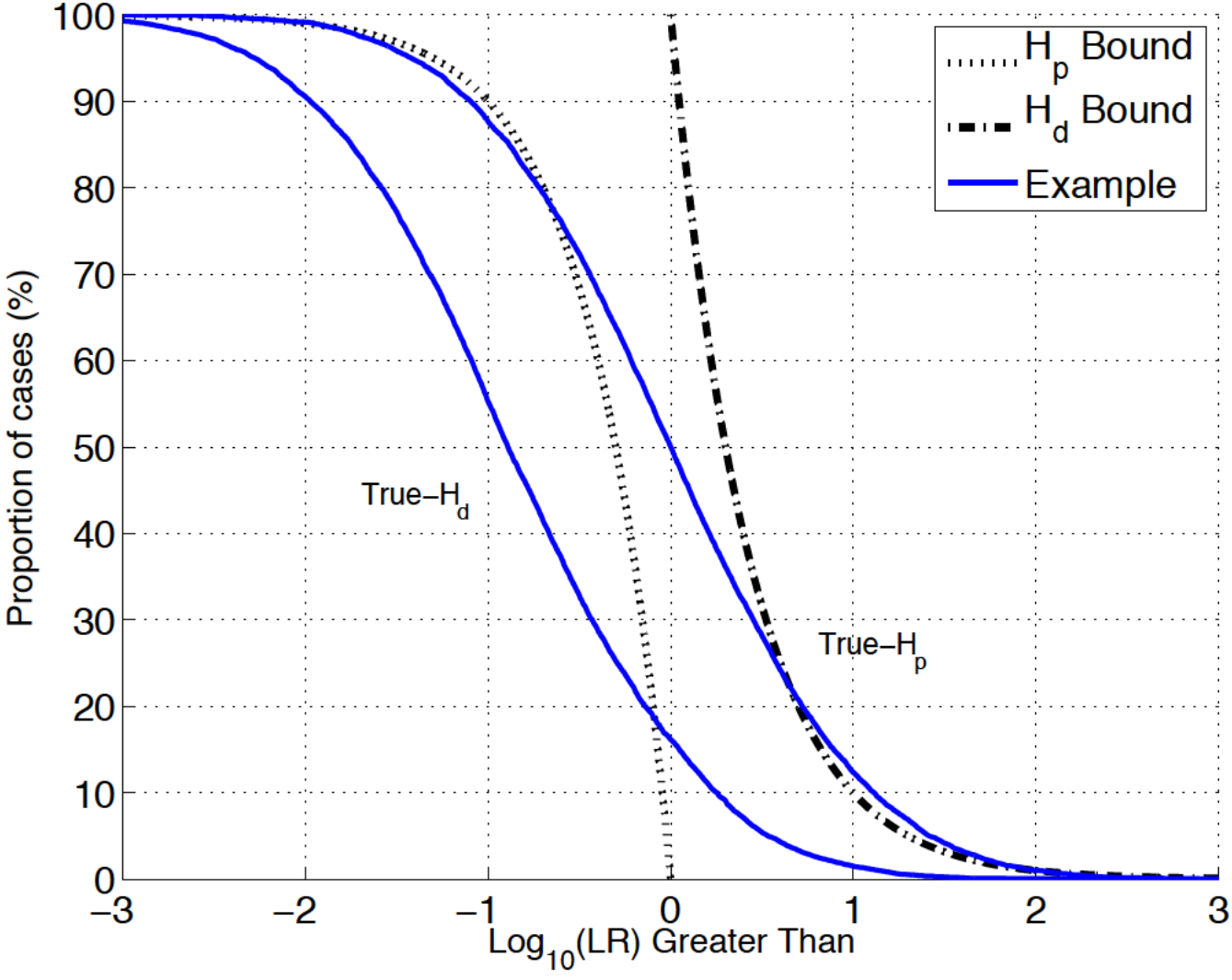

0.12
0.1
0.08
0.06
0.04
0.02
0
2
4
6
8
10
12
14
16
18
20
Upper bound of $P_d(V>k)$ and $P_p(V<1/k)$
Value of k
c=1
c=2
c=4
c=6

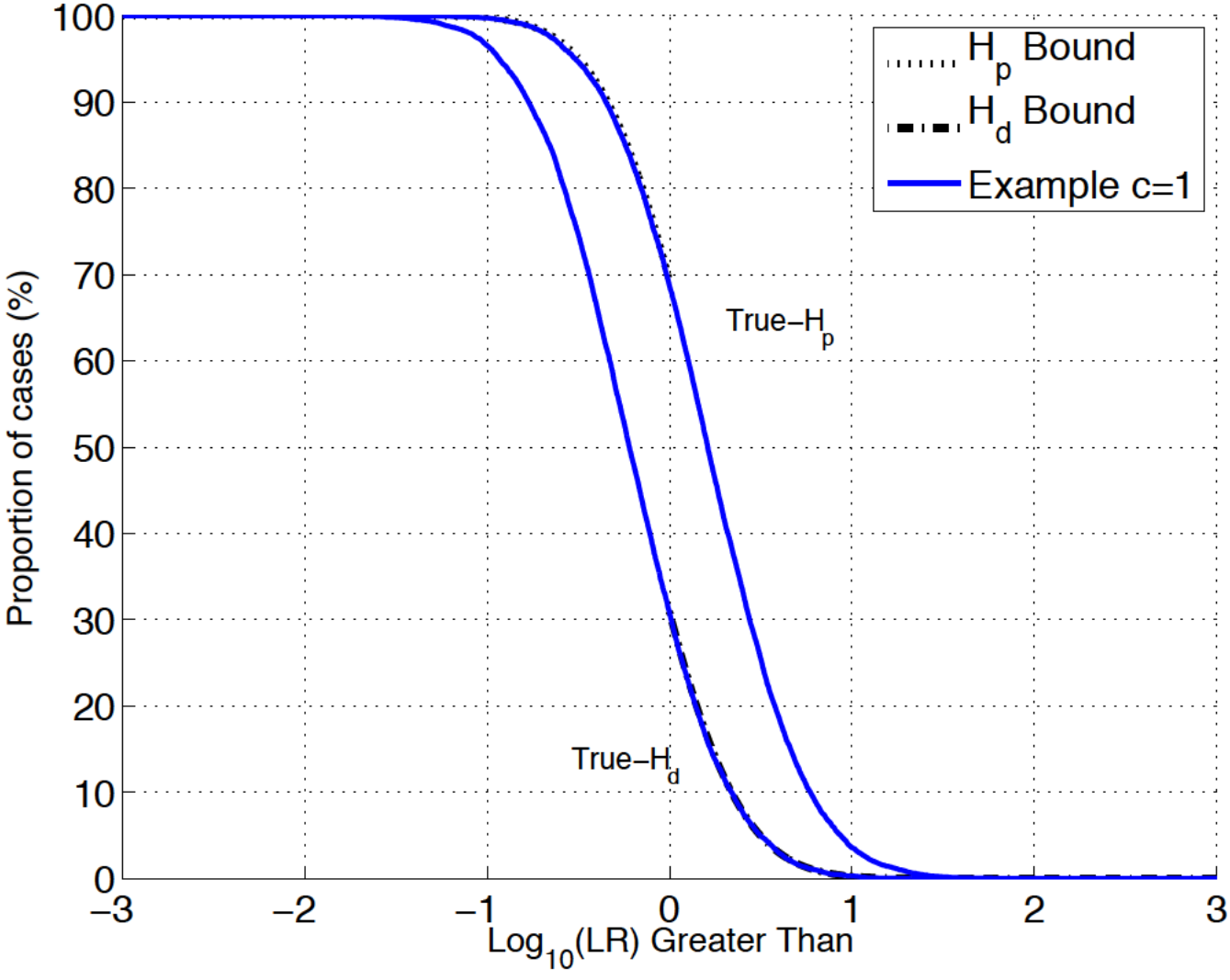

0.12
0.1
0.08
0.06
0.04
0.02
0
Upper bound of $P_d(V>k)$ and $P_p(V<1/k)$
2
4
6
8
10
12
14
16
18
20
Value of k
c=1
c=2
c=4
c=6
Maximum

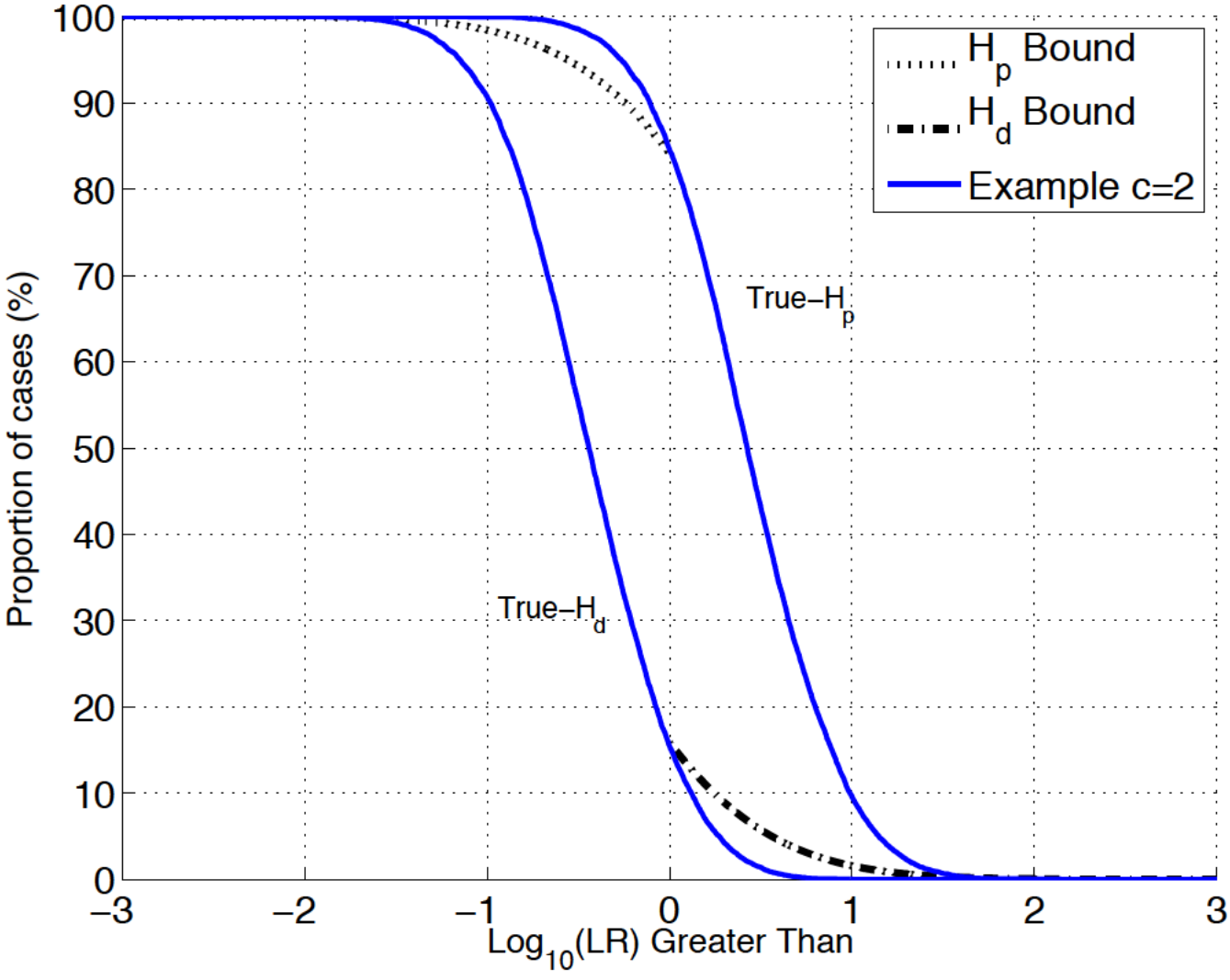

Universal bound
Max Normal bound (equal var.)
Normal bounds (equal var.)
Upper bound of $P_d(V>k)$ and $P_p(V<1/k)$
Value of k
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
2
4
6
8
10
12
14
16
18
20

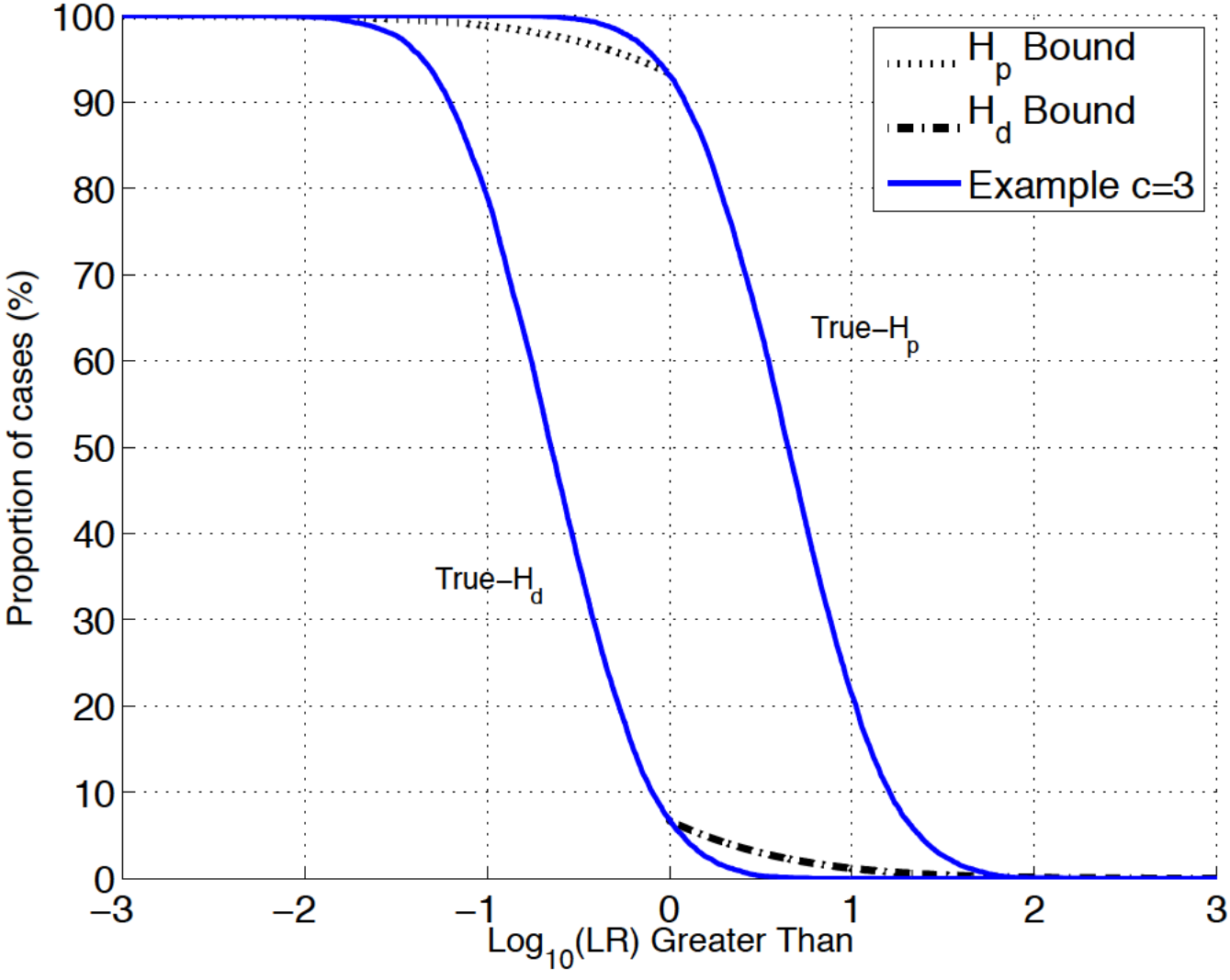

Example. Misleading Ev.: True−Hp=15.98%, True−Hd=15.53%
Proportion of cases (%)
100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Log10(LR) Greater Than

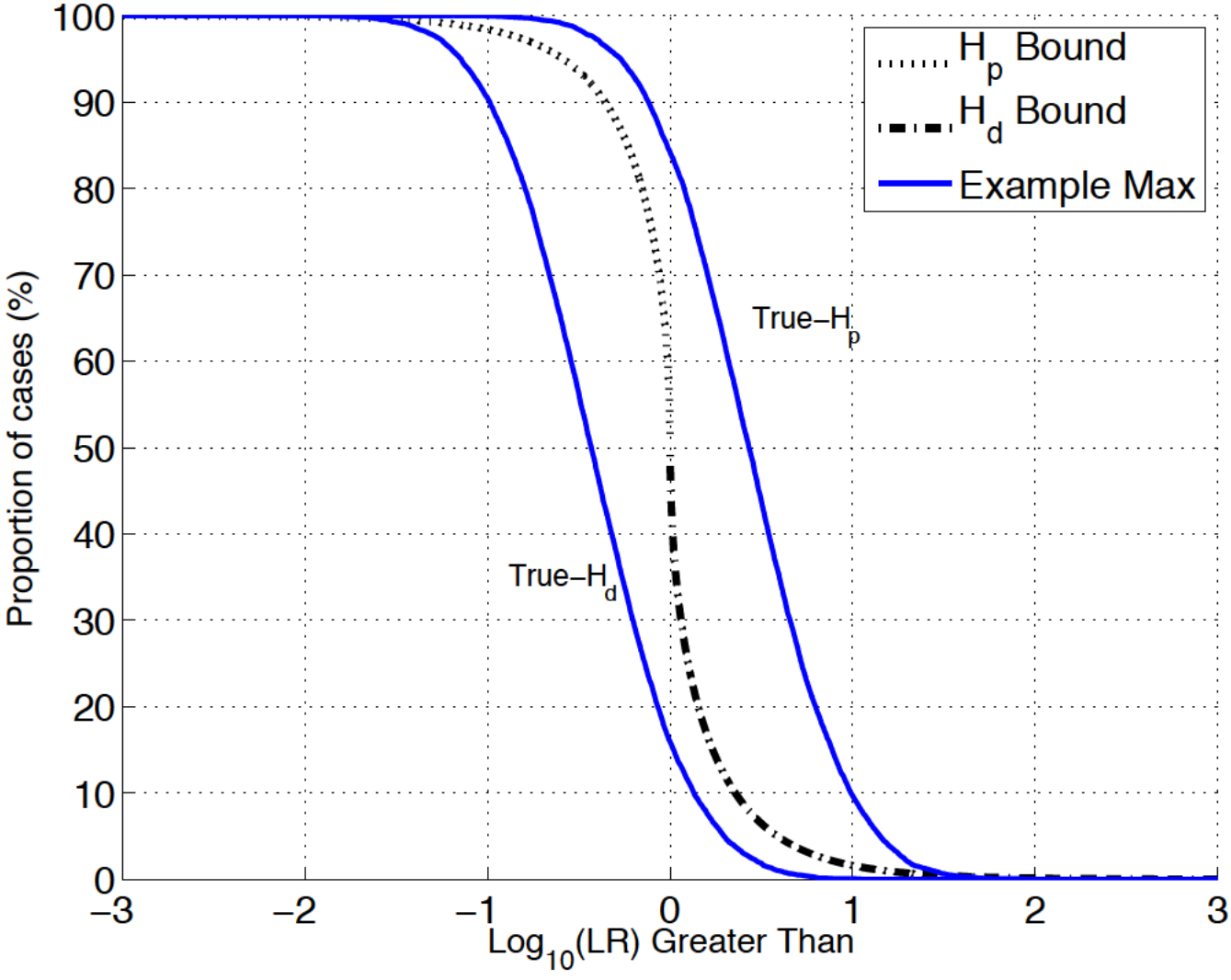

100
90
80
70
60
50
40
30
20
10
0
Proportion of cases (%)
−3
−2
−1
0
1
2
3
Log10(LR) Greater Than
Hp Bound
Hd Bound
Example
True−Hp
True−Hd

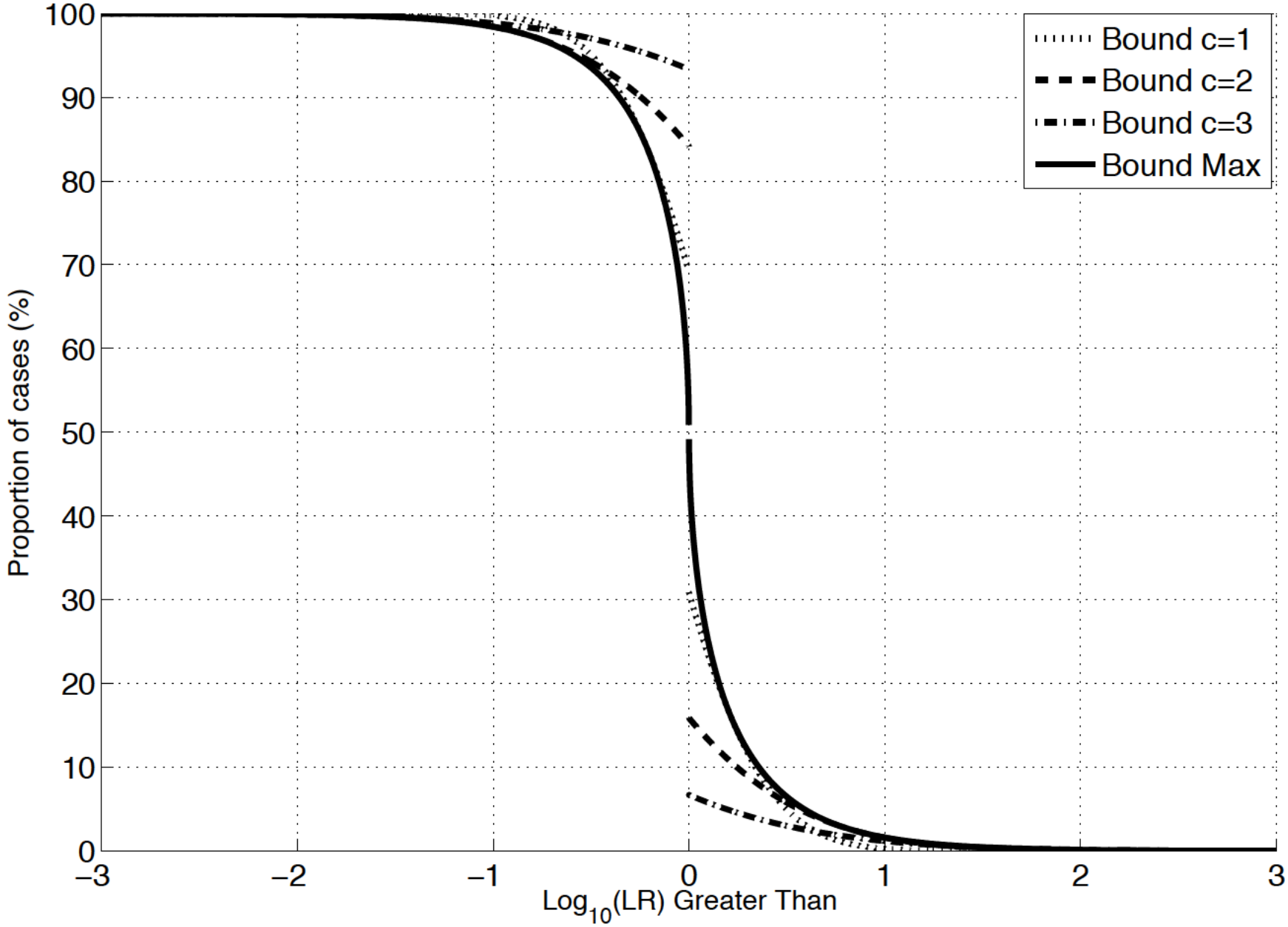

100
90
80
70
60
50
40
30
20
10
0
Proportion of cases (%)
−3
−2
−1
0
1
2
3
Log10(LR) Greater Than
Hp Bound
Hd Bound
Example
True–Hd
True–Hp

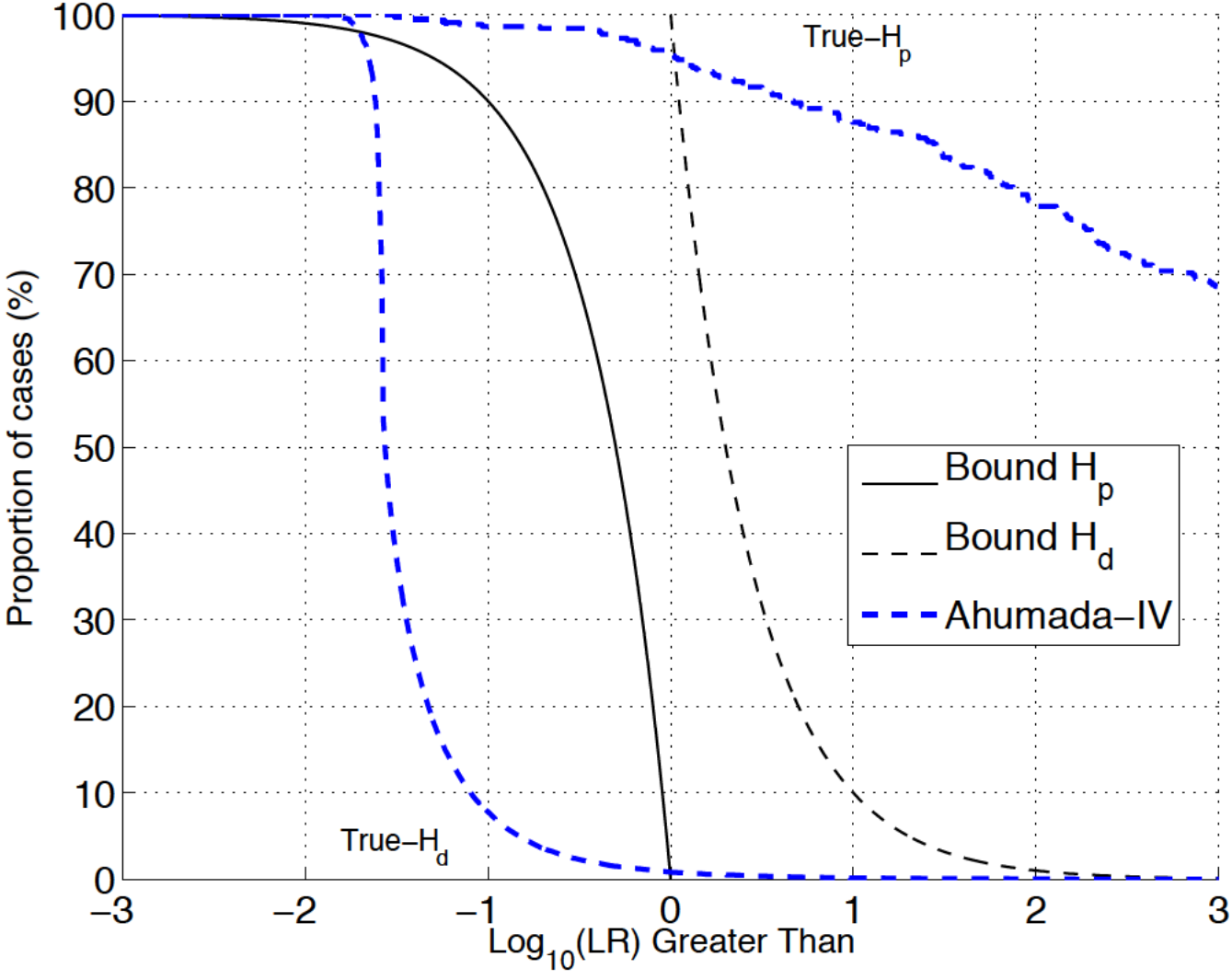

100
90
80
70
60
50
40
30
20
10
0
Proportion of cases (%)
−3
−2
−1
0
1
2
3
Log10(LR) Greater Than
Hp Bound
Hd Bound
Example c=1
True−Hp
True−Hd

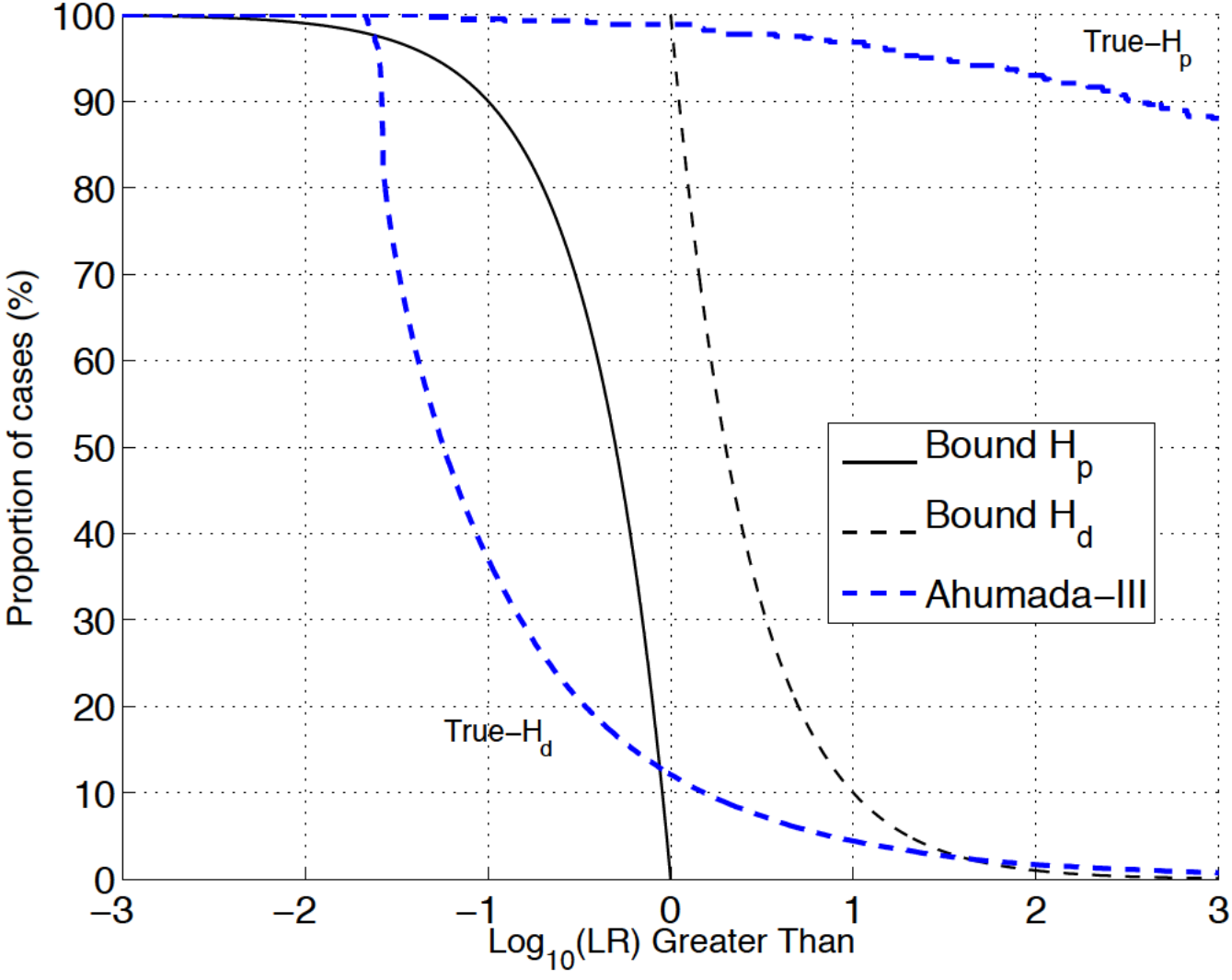

100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Proportion of cases (%)
Log10(LR) Greater Than
Hp Bound
Hd Bound
Example c=2
True−Hp
True−Hd

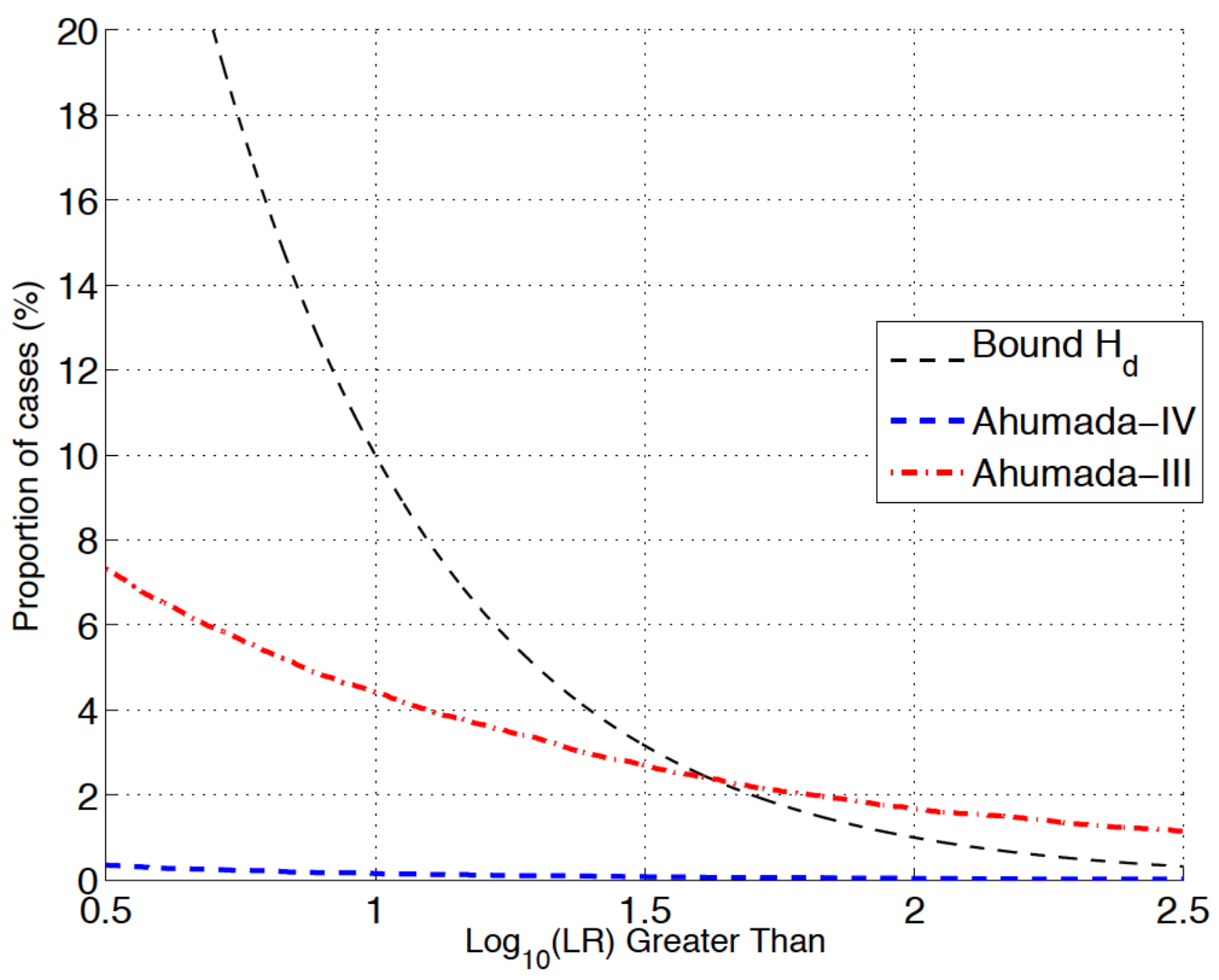

100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Proportion of cases (%)
Log10(LR) Greater Than
Hp Bound
Hd Bound
Example c=3
True−Hp
True−Hd


100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Proportion of cases (%)
Log10(LR) Greater Than
Hp Bound
Hd Bound
Example Max
True−Hp
True−Hd


Bound c=1
Bound c=2
Bound c=3
Bound Max
Proportion of cases (%)
100
90
80
70
60
50
40
30
20
10
0
−3
−2
−1
0
1
2
3
Log10(LR) Greater Than


100
90
80
70
60
50
40
30
20
10
0
Proportion of cases (%)
−3
−2
−1
0
1
2
3
Log10(LR) Greater Than
True−Hp
True−Hd
Bound Hp
Bound Hd
Ahumada−IV


100
90
80
70
60
50
40
30
20
10
0
Proportion of cases (%)
−3
−2
−1
0
1
2
3
Log10(LR) Greater Than
True−Hp
True−Hd
Bound Hp
Bound Hd
Ahumada−III


20
18
16
14
12
10
8
6
4
2
0
Proportion of cases (%)
0.5
1
1.5
2
2.5
Log10(LR) Greater Than
Bound Hd
Ahumada–IV
Ahumada–III