Random databases and ϵ-entropy

Oleg Seleznjev

Umeå university, Sweden
Moscow State University, Russia

http://matstat.umu.se

Petersburg 2005

Coauthor

Bernhard Thalheim

Institute of Computer Science and Applied Mathematics,
Christian-Albrechts University, Kiel
Germany

Outline:

1. Introduction. Basic notation.
2. Probabilistic models for databases.
3. Rényi ϵ-entropy.
4. Tests in random databases.
5. Summary.

Basic notation

Database ($m \times n$-table) of m **tuples** (or records) with n **attributes** (or features), $U := \{1, \ldots, n\}$

$$R = \begin{pmatrix} t_1(1) & \cdots & t_1(n) \\ \cdots & \cdots & \cdots \\ t_m(1) & \cdots & t_m(n) \end{pmatrix}$$

Tuples $t_j(U) = (t_j(1), \ldots, t_j(n)), j = 1, \ldots, m$, are vectors with values in $D = D_1 \times \ldots \times D_n$, where D_i are **domains** $i = 1, \ldots, n$.

A set of attributes A is called a **test** in R if all tuples $t_A(i), i = 1, \ldots, m$, are different.

We say that vectors x and y in a metric space (S, d) are ϵ-**close**, $\epsilon \geq 0$, if the distance $d(x, y) \leq \epsilon$. A set of attributes A will be called a ϵ-**test** if there are no ϵ-close tuples $t_A(i), i = 1, \ldots, m$. Let $N_\epsilon(A) := \#\{\epsilon-$close tuples in $R_A\}$.

Example

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- tests : $\{2, 3\}, \ldots$
- not a test : $\{1, 4\}$
- ϵ-test, $\epsilon = 0.5$, $\{3, 5\}$ but not a 1.0-test for the Euclidian norm.

Database problems:

- Data search optimization; Tests and minimal tests.
- Database design; constraints sets complexity.

Problems:

- Probabilistic models for discrete and continuous databases;
- The distribution of the number of ϵ-coincidences $N_\epsilon(A)$
- Joining multiple tables with approximate matching.

The worst case setting approach:
Combinatorial or deterministic methods; restrictive class of models and overestimating complexity.

Average case setting approach:
Probabilistic methods; general class of models; *where* the distribution of tests concentrates (i.e., *typical* tests), and *for which* model parameters.

Probabilistic models for databases.

1. Tuples $t_j(U) \in \prod_{i \in U} D_i, j = 1, \ldots, m$, are independent random vectors;

2. $\mathcal{P}$ is a common (discrete or continuous) distribution for tuples

Examples

Uniform random database if $\mathcal{P}$ is a uniform (discrete or continuous) distribution $\mathcal{U}$ in D.

Gaussian database if $\mathcal{P}$ is a Gaussian distribution $\mathcal{G}$ in $D = \Re^n$.

(Generalized) **Bernoulli** random database if all attributes are iid random Q-variables.

For instance, the conventional Bernoulli model corresponds to a binary one for the discrete Bernoulli distribution with $D_i = \{0, 1\}$ for all attributes.

Measures of uncertainty

Shannon For a discrete distribution $\mathcal{P} = \{p(\mathbf{k}), \mathbf{k} \in D\}$,

$$h_1(\mathcal{P}) := -\sum_{\mathbf{k}} p(\mathbf{k}) \log_2 p(\mathbf{k})$$

Rényi For a discrete distribution $\mathcal{P} = \{p(\mathbf{k}), \mathbf{k} \in D\}$,

$$h_s(\mathcal{P}) := \frac{1}{1-s} \log_2(\sum_{\mathbf{k}} p(\mathbf{k})^s), \quad s \neq 1,$$

and $h_s(\mathcal{P}_A) \to h_1(\mathcal{P}_A)$ as $s \to 1$.

Rényi for a continuous random variable X, **differential entropy**. The uniform quantizer $q(X) = [NX]/N$. Then for $p_k := P\{q(X) = k/N\} = P\{k/N < X \le (k+1)/N\}$

$$h_\epsilon^R(X) := -\log_2 \sum_k p_k^2, \quad \epsilon = 1/N, s = 2$$

$$h_\epsilon^R(X) = \log_2 \frac{1}{\epsilon} - \log_2 \int_R p(x)^2 dx + \mathrm{o}(1),$$

with a straightforward generalization to the vector case R^n and the general class of entropies

$$h_\epsilon^R(X) = n \log_2 \frac{1}{\epsilon} - \log_2 \int_{R^n} p(x)^2 dx + \mathrm{o}(1).$$

Kolmogorov For a metric space (S,d) and $N_\epsilon(S,d)$ the cardinality of the minimal ϵ-net

$$H_\epsilon(S) = \log_2 N_\epsilon(S,d).$$

Kolmogorov-Shannon For random continuous variables X, Y with the mutual information

$$I(X,Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)q(y)} dxdy,$$

the **risk distortion** (or ϵ-entropy)

$$R_\epsilon(X) = \inf\{I(X,Y) : E||X-Y||^2 \leq \epsilon^2\}.$$

Posner-Rodemich For a probabilistic separable metric space (S, d, μ) and a countable ϵ-partition $\pi_\epsilon = \{A_i\}$ with diameter $d(A_i) \le \epsilon$,

$$H_\epsilon^{PR}(S, \mu) = \inf_{\pi_\epsilon} \sum \mu(A_i) \log_2(1/\mu(A_i))$$

There are ϵ, δ-variants for Kolmogorov and Posner-Rodemich when defined on $S \setminus B$ and $\mu(B) < \delta$.

Haussler and Opper (the volume-scaling entropy). For a probabilistic separable metric space (S, d, μ), X is a random μ-vector,

$$H_\epsilon^{HO}(S, \mu) = \mathsf{E} \log_2(1/\mu(B_\epsilon(X)))$$

The Rényi ϵ-entropy (cf. Szpankowski for discrete sequences)

Let X, Y be independent $\mathcal{P}$-distributed random vectors with values a metric space (S, d) and $B_\epsilon(x) := \{y : d(x, y) \le \epsilon\}$ be a ϵ-ball, the ϵ-*ball probability* $p_\epsilon(x) := P\{Y \in B_\epsilon(x)\}$. The generalized Rényi ϵ-entropy

$$h_{2,\epsilon}(\mathcal{P}) := -\log_2 P\{d(X, Y) \le \epsilon\} = -\log_2 p_\epsilon(X).$$

In the general case,

$$h_{s,\epsilon}(\mathcal{P}_A) := \frac{1}{1-s} \log_2 \mathsf{E} p_\epsilon(X)^{s-1}, s \neq 1,$$

the generalized Shannon ϵ-entropy as $s \to 1$,

$$h_{1,\epsilon}(\mathcal{P}) := -\mathsf{E} \log_2 p_\epsilon(X).$$

(cf. the volume-scaling entropy).

Proposition. *Let* $X = (X_A, X_B)$ *be a random* $\mathcal{P}$*-vector.*

(i) $h_{2,\epsilon}(\mathcal{P}) \geq 0$*. If* $h_{2,\epsilon}(\mathcal{P}) = 0$*, then for some* x_0 *,* $P\{X \in B_\epsilon(x_0)\} = 1$*. On the other hand, if* $P\{X \in B_\epsilon(x_0)\} = 1$*, then* $h_{2,2\epsilon}(\mathcal{P}) = 0$*;*

(ii) $h_{2,\epsilon}(\mathcal{P}_A) \leq h_{2,\epsilon}(\mathcal{P}_{A\cup B})$*;*

(iii) if $|x| = \max_{i=1,\ldots,n} |x_i|$ *and* X_A*,* X_B *are independent, then* $h_{2,\epsilon}(\mathcal{P}_{A\cup B}) = h_{2,\epsilon}(\mathcal{P}_A) + h_{2,\epsilon}(\mathcal{P}_B)$*;*

(iv) $\frac{1}{2}h_{2,\epsilon}(\mathcal{P}) \leq h_{3,\epsilon}(\mathcal{P}) \leq h_{2,\epsilon}(\mathcal{P})$*;*

(v) for every continuous distribution with compact domain D *and continuous and bounded density function* $p(x)$ *and the uniform distribution* $\mathcal{U}$ *on* D*,*

$$h_{2,\epsilon}(\mathcal{P}) \leq h_{2,\epsilon}(\mathcal{U}) + \mathrm{o}(1) \text{ as } \epsilon \to 0.$$

Discrete case, X, Y, Z are $\mathcal{P}$-iid, $\epsilon = 0$, $\mathcal{P} = \{p(\mathbf{k}) = P(X = \mathbf{k})\}$,

$$\begin{aligned} h_2(\mathcal{P}) &= -\log_2 P(X = Y) = -\log_2(\sum_{\mathbf{k}} p(\mathbf{k})^2) = -\log_2 \mathsf{E}p(X), \\ h_3(\mathcal{P}) &= -\log_2 P(X = Y, X = Z) = -1/2 \log_2(\sum_{\mathbf{k}} p(\mathbf{k})^3) \\ &= -\log_2 \mathsf{E}p(X)^2 \end{aligned}$$

Proposition. *Let $X = (X_A, X_B)$ be a random $\mathcal{P}$-vector.*

(i) $h_2(\mathcal{P}_A) \le h_2(\mathcal{P}_{A \cup B})$*;*

(ii) If X_A, X_B are independent, then $h_2(\mathcal{P}_{A \cup B}) = h_2(\mathcal{P}_A) + h_2(\mathcal{P}_B)$;

(iii) For every discrete non-uniform distributions with finite domains, $h_2(\mathcal{P}_A) < h_2(\mathcal{U}_A)$;

(iv) $\frac{3}{4} h_2(\mathcal{P}_A) < h_3(\mathcal{P}_A) \le h_2(\mathcal{P}_A)$ with the equality iff $\mathcal{P}$ is uniform.

Continuous case, $\epsilon > 0$, density function $p(x)$, the volume of $B_\epsilon(x)$ in $\Re^n$, $b_\epsilon(n) := |B_\epsilon(x)|$

Proposition. *Let $p(x), x \in D$ bounded and continuous or have a finite number of discontinuity points. Then*

$$\begin{aligned} h_\epsilon(\mathcal{P}) &= -\log_2 b_\epsilon(n) - \log_2 \int_D p(x)^2 dx + \mathrm{o}(1) \\ &= n \log_2 \frac{1}{\epsilon} - \log_2 b_1(n) - \log_2 \int_D p(x)^2 dx + \mathrm{o}(1) \text{ as } \epsilon \to 0. \end{aligned}$$

If the **differential entropy** $H_s(\mathcal{P}_A) := \frac{1}{1-s} \log_2 \int_{R^n} p(x)^s dx, s \neq 1,$

$$h_{s,\epsilon}(\mathcal{P}_A) = n \log_2 \frac{1}{\epsilon} + \log_2 b_1(n) + H_s(\mathcal{P}_A) + \mathrm{o}(1) \text{ as } \epsilon \to 0,$$

Examples

Uniform random database. Let $H(A) := \sum_{i \in A} \log_2 |D_i|$ (*information function of* A), $r = |A|$,

Discrete (ϵ=0)

$p(\mathbf{k}(A)) = 2^{-H(A)}$, Rényi entropy $a = h(\mathcal{P}_A) = H(A)$

Continuous $p(x(A)) = 2^{-H(A)}, d_{\min} = \min |D_i|$.;

$$h_\epsilon(\mathcal{P}_A) = r \log_2 \frac{1}{2\epsilon} + H(A) + \mathsf{O}(r\epsilon^2/d_{\min});$$

Bernoulli database:

Discrete(ϵ=0)
$p(\mathbf{k}(A)) = \prod_{i \in A} \mathcal{Q}(\{k(i)\})$, Rényi entropy $h(\mathcal{P}_A) = rh(Q)$;

Continuous
$p(x(A)) = \prod_{i \in A} q(x_i)$, Rényi entropy (*max*-norm, for $q(x)$) $h_\epsilon(\mathcal{P}_A) = r h_\epsilon(Q)$ and $h_\epsilon(Q) = \log_2 \frac{1}{2\epsilon} + H(Q) + \mathsf{o}(1)$;

Gaussian database:
Tuples $t_i(A)$ are iid Gaussian $N(\mu, \Sigma)$ random vectors; λ_i are eigenvalues of Σ; Rényi entropy (*max*-norm)

$$h_\epsilon(\mathcal{P}_A) = r \log_2 \frac{1}{2\epsilon} + \frac{1}{2} \sum_i \log_2(2\pi\lambda_i) + \mathsf{O}(r\epsilon^2/\lambda_{\min}),$$

– *Bernoulli* database for Gaussian tuples, $r = |A|$:

$$h_\epsilon(\mathcal{P}_A) = r(\log_2 \frac{1}{2\epsilon} + \frac{1}{2} \log_2(2\pi\sigma^2) + \mathsf{O}(\epsilon^2)) \text{ as } \epsilon \to 0.$$

Quantization ϵ-entropy

Let $X \in D \subseteq \Re^n$ be a continuous random vector and Voronoi partition $D = \cup_{i=1}^{N_\epsilon} B_\epsilon(x_i)$, $\lambda(B_\epsilon(x_i) \cap B_\epsilon(x_j)) = 0$ and $1 \leq N_\epsilon \leq \infty$. For a compact set D, assume that $N_\epsilon < \infty$. Let V_ϵ-quantizer $q(X) = x_i$, where $i = \text{argmin}_{j=1,\ldots,N_\epsilon} |X - x_j|$ and the entropy $h_\epsilon^R(X) := -\log_2 \sum_{j=1}^{N_\epsilon} p_\epsilon(x_j)^2$.

Theorem. *Let $p(x), x \in D \subseteq \Re^n$ be a continuous density function , and $q(X)$ the Voronoi V_ϵ-quantizer. Then*

(i) $h_\epsilon^R(X) = -\log_2 b_\epsilon(n) + H(\mathcal{P}) + \mathrm{o}(1)$*;*

(ii) for a compact set D,

$$h_\epsilon^R(X) \leq \log_2 N_\epsilon \text{ and } h_\epsilon(X) \leq \log_2 N_\epsilon + \mathrm{o}(1) \text{ as } \epsilon \to 0.$$

Discussion

The assertions can be directly generalized for the case of a separable metric space (S, d) with Lebesgue measure for an ϵ-ball. Independent realizations of these random functions can be archived in a database (e.g., Fourier coefficients of a realization in $L^2[0, 1]$ space or some finite dimensional realization approximations).

Example

ϵ-entropy for a Wiener measure $\mathcal{W}$. let two independent Wiener processes $W_1(t), W_2(t)$, $t \in [0,1]$, be Gaussian random vectors taking values in the Hilbert space $L^2[0,1]$. Then $X(t) = W_1(t) - W_2(t)$ is also a Wiener process with the covariance function $K(t,s) = 2\min(t,s),\ \ t,s \in [0,1]$ and the corresponding small ball probability works

$$P(||W_1 - W_2||_{L^2[0,1]} \leq \epsilon) = P\left(\int_0^1 X(t)^2 dt \leq \epsilon^2\right) \sim \frac{4\epsilon}{(2\pi)^{1/2}} \exp\{-\frac{1}{4\epsilon^2}\},$$

$$h_{2,\epsilon}(\mathcal{W}) = \frac{\epsilon^{-2}}{4} + \log_2 \frac{(2\pi)^{1/2}}{4\epsilon} + \text{o}(1) \text{ as } \epsilon \to 0.$$

If B_H is a fractional Brownian motion with Hurst constant H and $S = L^2[0,1]$, then

$$h_{2,\epsilon}(\mathcal{B}_H) \sim C_H \epsilon^{-1/H}, C_H > 0.$$

Random databases. ϵ-Test probability

Rényi entropies

(A) $a_\epsilon = a_\epsilon(m) = h_{2,\epsilon}(\mathcal{P}_{m,A_m}) \to \infty$ as $m \to \infty$.

A “relative” uncertainty in a distribution $\mathcal{P}$.

(B) $$\delta_\epsilon := \delta_\epsilon(\mathcal{P}) := 4\, h_{3,\epsilon}(\mathcal{P})/h_{2,\epsilon}(\mathcal{P}) - 3 > 0.$$

(i) (B) is valid e.g. for Uniform and Gaussian databases.

(ii) For a discrete distribution $\mathcal{P}$, ϵ=0, $0 < \delta(\mathcal{P}) \le 1$ with the equality only for uniform distribution.

Let the mean number of ϵ-close tuples, $M = m(m-1)/2$,

$$\lambda_\epsilon = \lambda_m(\epsilon, A) := \mathsf{E} N_\epsilon(A) = MP(|t_1(A) - t_2(A)| \le \epsilon) = M2^{-a_\epsilon}.$$

Theorem. *Let $R_m, m \ge 1$, be a sequence of random tables and (A), (B) hold.*

(i) For all $m \ge 1$ and $\lambda_\epsilon > 0$,

$$|P\{R_m \models_\epsilon A\} - e^{-\lambda_\epsilon}| \le d_{TV}(\mathcal{L}(N_{A,\epsilon}), Po(\lambda_\epsilon)) \le 8 \cdot 2^{-\delta_\epsilon a_\epsilon/2} \lambda_\epsilon^{1/2}.$$

(ii) Let λ_0 be a positive constant. Then

$$P\{R_m \models_\epsilon A\} \to \begin{cases} 0, & \text{if } \lambda_{m,\epsilon} \to \infty, \\ e^{-\lambda_0}, & \text{if } \lambda_{m,\epsilon} \to \lambda_0, \\ 1, & \text{if } \lambda_{m,\epsilon} \to 0, \end{cases} \quad \text{as } m \to \infty.$$

Discussion

The most likely ϵ-test candidates are amongst sets with maximal ϵ-entropies. Let $a_\epsilon(r) \geq 2\log_2 m + c_m$ and $c_m \to +\infty$. Then

$$P\{R_m \models_\epsilon A\} = 1 - \mathrm{o}(1) \text{ as } m \to \infty.$$

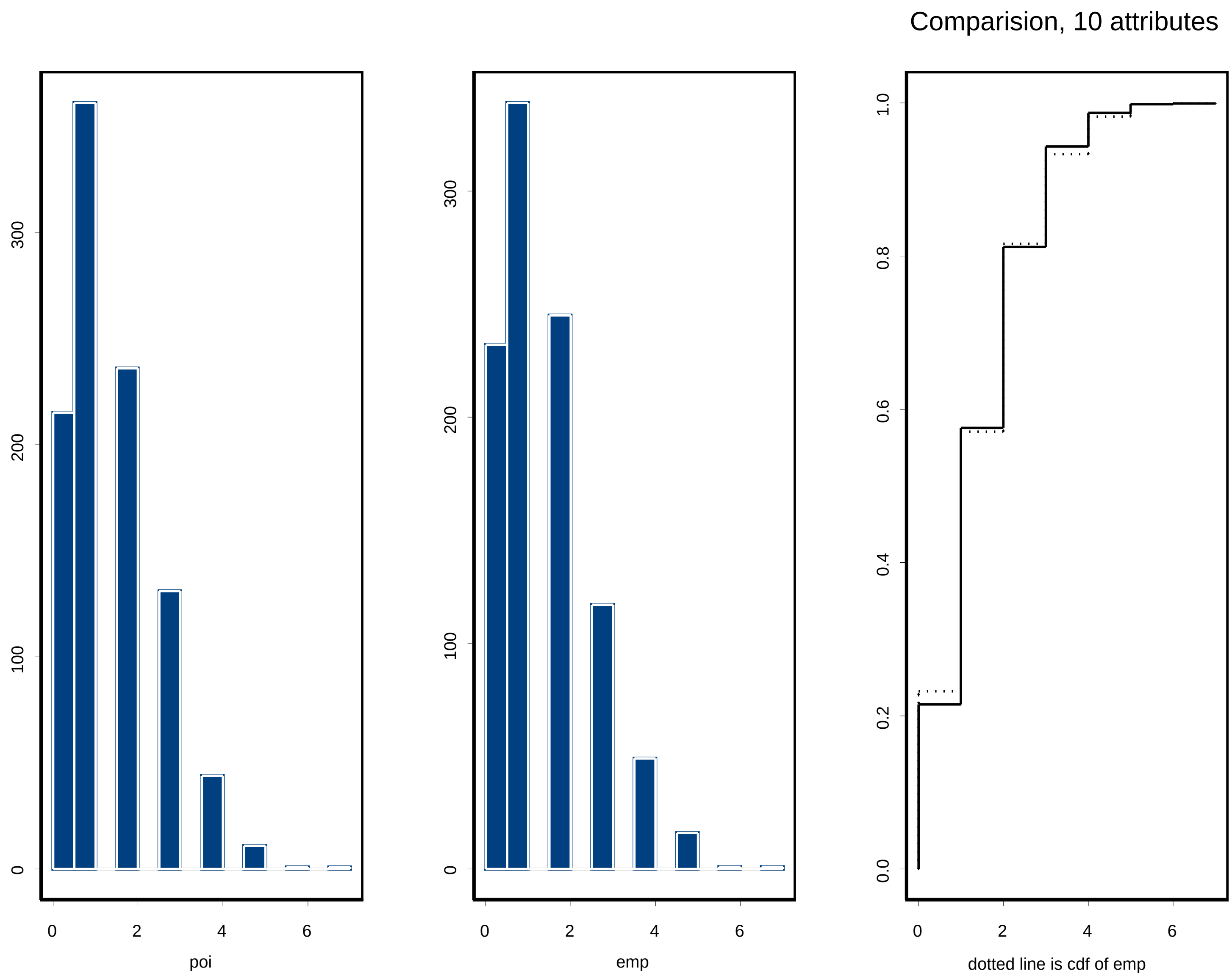
These entropies characterize **typical** ϵ-tests in a random database.

Sufficient conditions

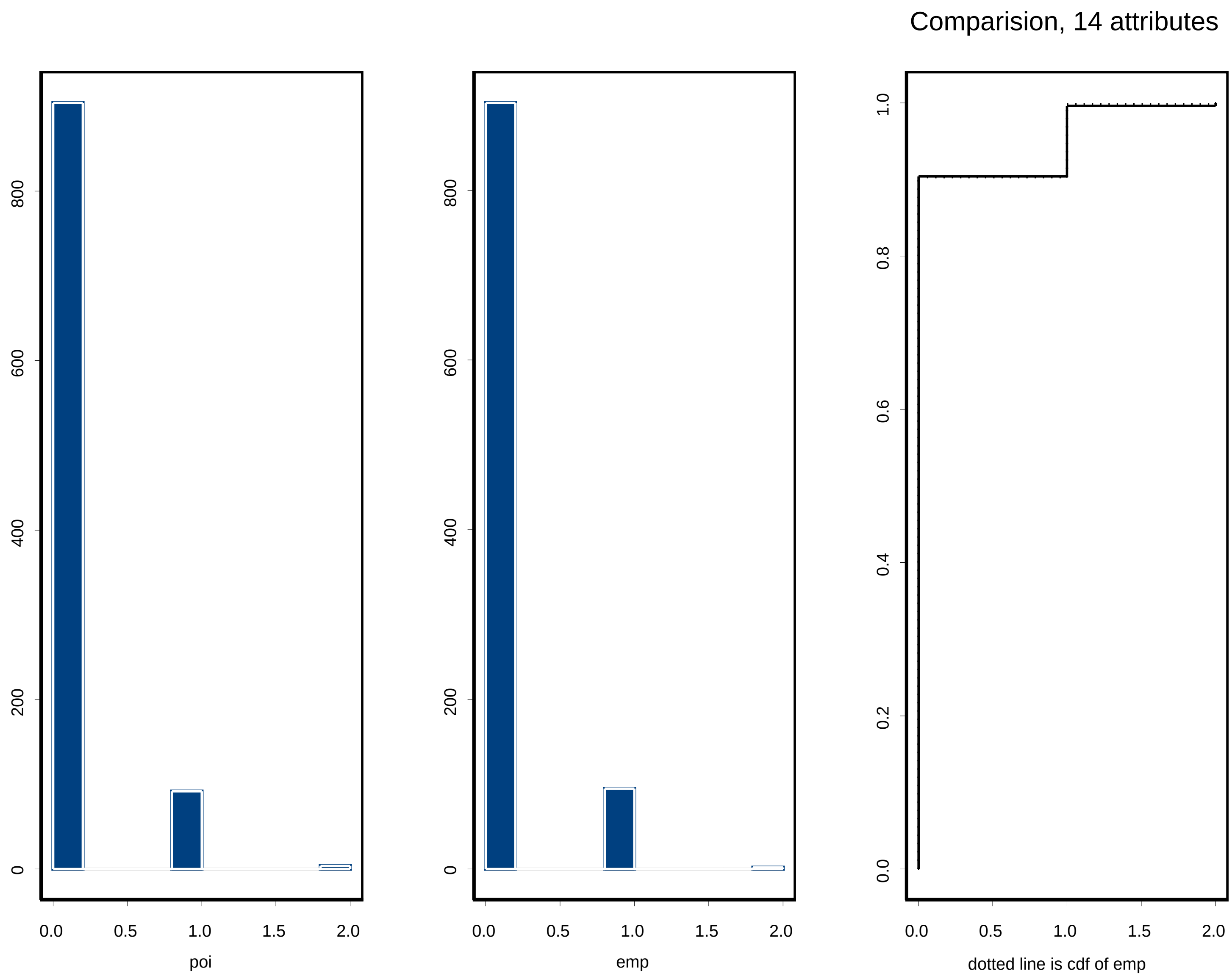
(A) $\Longleftarrow p_{\epsilon,\max} := \max_{x\in D} p_\epsilon(x) \to 0$ as $m \to \infty$.

(B) $\Longleftarrow p_{\epsilon,\min} > p^2_{\epsilon,\max}$.

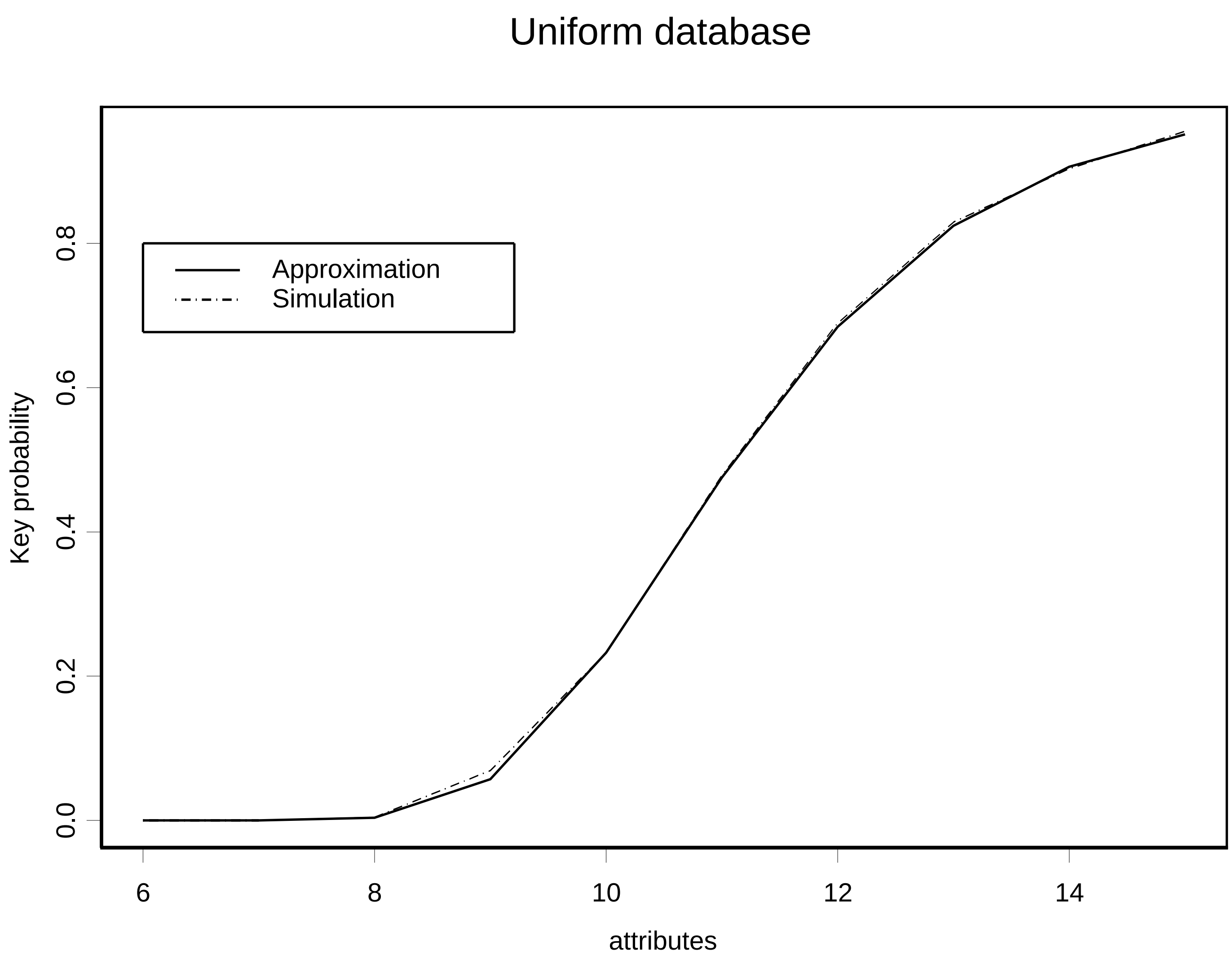
The **test property** for a set of attributes is determined by the ϵ-entropy $h_{2,\epsilon}(\mathcal{P}_{m,A_m})$.



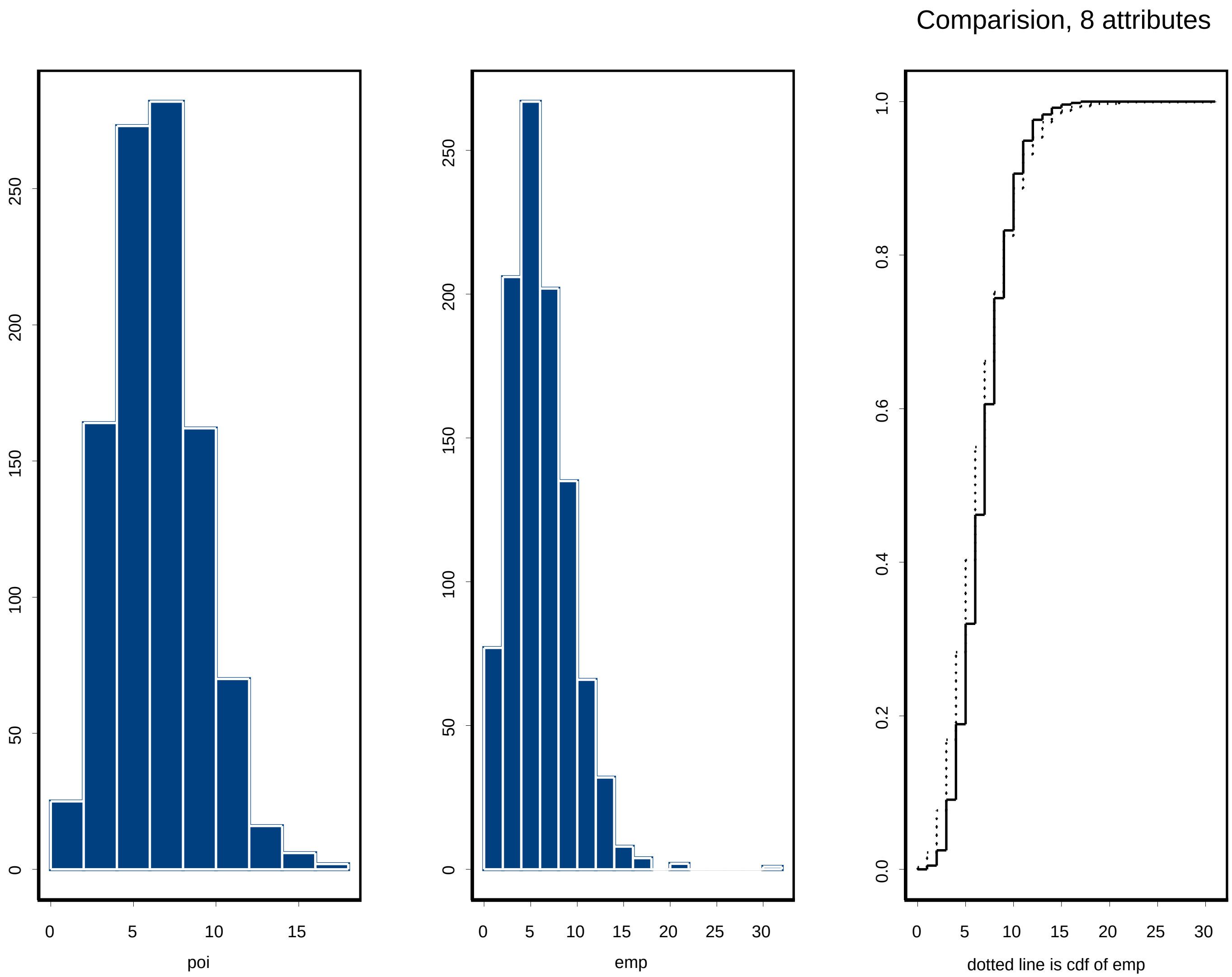
Poisson approximation for standard continuous uniform database, 10 U(0,1)-attributes, $m = 50$,$\epsilon = 0.3$.



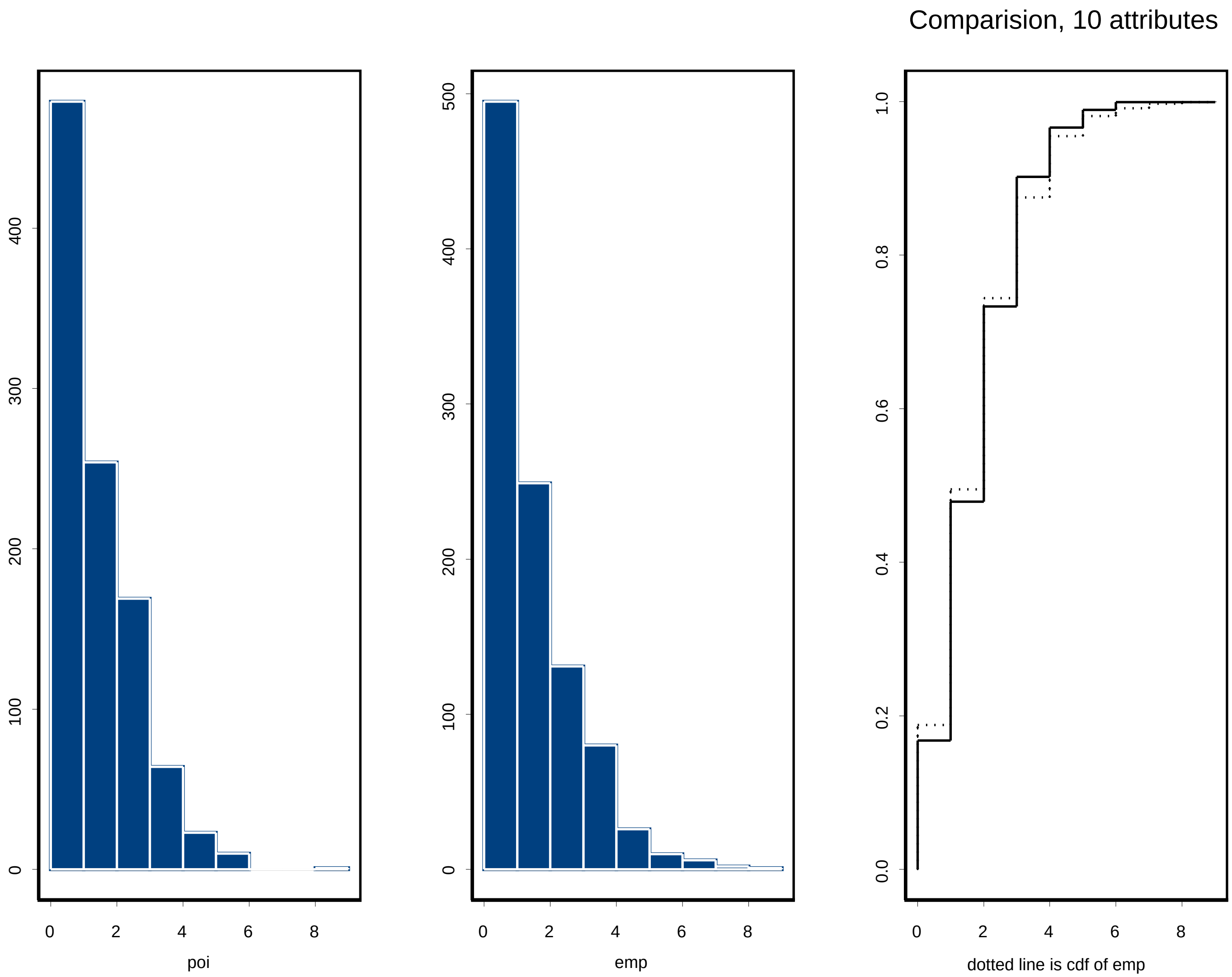
Poisson approximation for standard continuous uniform database, 14 U(0,1)-attributes, $m = 50$,$\epsilon = 0.3$.



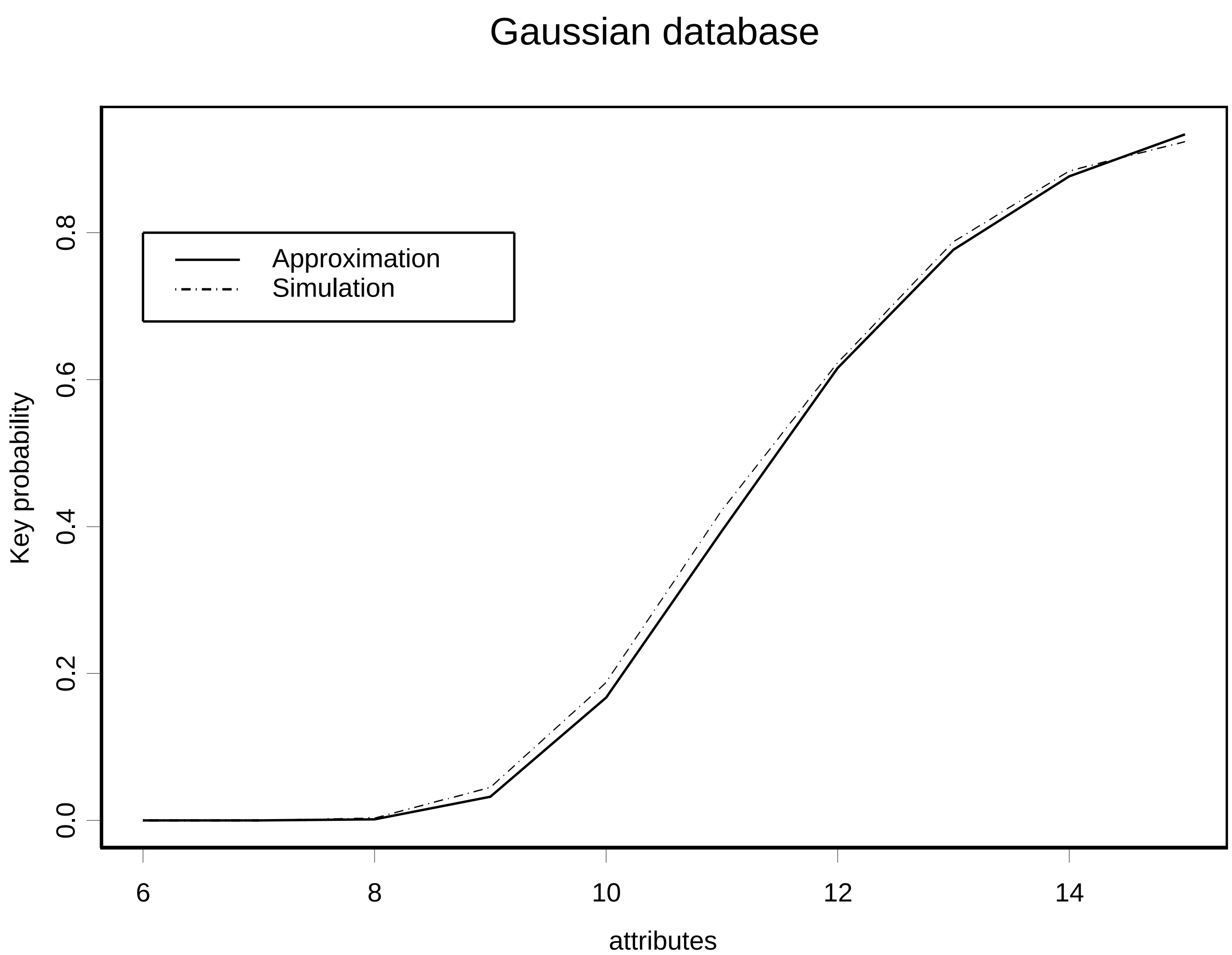
Poisson approximation for standard continuous uniform database, U(0,1)-attributes, $m = 50$, $\epsilon = 0.3$. Empirical distribution (*simulation*), $N_{sim} = 1000$.



Poisson approximation for standard continuous Gaussian database, 8 N(0,1)-attributes, $m = 50$, $\epsilon = 1.0$.



Poisson approximation for standard continuous Gaussian database, 10 N(0,1)-attributes, $m = 50$, $\epsilon = 1.0$.



Poisson approximation for standard continuous Gaussian database, N(0,1)-attributes, $m = 50$, $\epsilon = 1.0$. Empirical distribution (*simulation*), $N_{sim} = 1000$.

Summary

Instead of

Worst case setting and exhaustive search

Stochastic modelling and statistical inference

References

- J. Demetrovics, G.O.H. Katona, D. Miklós, O. Seleznjev, and B. Thalheim. Asymptotic properties of keys and functional dependencies in random databases. *Theor. Computer Science,* 190:151–166, 1998.
- O. Seleznjev, and B. Thalheim. Average case analysis in database problems. *Methodology and Computing in Applied Probability,* 5:395-418, 2003.
- O. Seleznjev, and B. Thalheim. Random databases with approximate record matching and epsilon-entropy. Univ. Umeå Research Report, Dep. Math. and Math. Stat., 2004:4, http://www.matstat.umu.se/personal/Oleg/personal/EntropyDB2005.pdf, 2004.

Thanks!