NUTNOST I ALTERNATIVNÍCH METOD

Ing. Pavel Kovanic, DrSc.,
Ing.Tomáš Ocelka
Zdravotní ústav se sídlem v Ostravě

ZPRACOVAT DATA: JAK?

Matematická statistika nemusí být ani jediná, ani nejlepší volba

Specifika ekologických aplikací:

- Jde o životy
- Data jsou drahá, je jich málo
- Měření jsou obtížná
- I nízké koncentrace škodlivin jsou nebezpečné (organismy je akumulují)

MATEMATICKÁ STATISTIKA

Brilantní úspěchy v oblasti hromadných dat:

- Demografie
- Státní hospodářství
- Teorie plynu
- Termodynamika
- Teorie jaderných reaktorů (a bomb!)

MATEMATICKÁ STATISTIKA

Omezení (Centrální limitní věta):

1. Náhodná proměnná x má distribuci se střední hodnotou μ a standardní odchylkou σ
2. Datové soubory jsou náhodně (nezávisle) vybrány z „populace“.

Platí-li 1. a 2., pak rozdělení μ při rostoucím počtu dat konverguje k normálnímu rozdělení nad nosičem $-\infty, +\infty$ případně $(0, +\infty)$.

NÁMITKY

- Existují distribuce, nemající μ ani σ (např. jednotlivá odlehlá data a nehomogenity).
- Reálné jevy nemají povahu hromadné náhody, mají jednotlivé reálné příčiny.
- Reálné populace nebývají homogenní.
- Reálná data nejsou nekonečná, jsou z intervalu (LB, UB), kde $-\infty < LB, UB < +\infty$.

NEURČITOST

- Základní problém: model neurčitosti
- Selhávání statistického modelu: hledání
- IPMU (International Processing and Management of Uncertainty in Knowledge-based Systems, http://ipmu.lip6.fr/
- 13 světových kongresů za 26 let
- 23 modelů neurčitosti, aplikace v 13 oborech

NEURČITOST V EKOLOGII

- WHO IPCS Harmonization Project: Characterizing and Communicating Uncertainty in Exposure Assessment(2008)
- IPCS … International Project of Chemical Safety
- http://www.who.int/ipcs/publications/methods/harmonization/exposure_assessment.pdf
- Výzkumné projekty EU:

  Heimtsa, Intarese, 2-FUN, FOKS

CENZOROVANÁ DATA

Cenzorovaná … reálná, ale neúplně určená

Necenzorovaná položka dat D: ***x = D***,

kde x je odhad skutečně změřené hodnoty.

Zdola cenzorovaná D: ***x <= D (nejvíce D)***

Shora cenzorovaná D: ***D <= x (nejméně D)***

Oboustranně cenzorovaná (intervalová) D:

DL <= x <= DU.

I cenzorovaná data obsahují informaci

ORGANICKÉ POLUTANTY

- Monitoring v ČR: ČHMÚ + ZÚ Ostrava
- Řeky: ≈ 140 organických polutantů, 106 lokalit
- Příklady cenzorování zdola:

  HxCDD, PCB207: 100% pod LOD

  1234789HpCDF, PCB206: 63 z 64 pod LOD

  PCB205, PCB209: 62 z 64 pod LOD

  Jen 44 ze 140 měřitelné nad LOD

MATEMATICKÁ GNOSTIKA

- Naše „želízko v ohni“ (ČSAV 1984)
  - Gnostická teorie neurčitých dat
  - Algoritmy založené na gnostických metodách
  - Programy realizující gnostické algoritmy
  - Aplikace gnostických programů

Pojem:

Gnostický – opak agnostického

SPECIFIKA GNOSTIKY

- Teorie a algoritmy zpracování ***jednotlivých*** dat a ***malých*** datových souborů
- Důsledné dodržování principu

  Nechť data mluví za sebe
- Maximalizace informace obsažené ve výsledcích
- Respektování přírodních zákonů

GNOSTICKÉ PROGRAMY

1) Programy gnostické marginální (jednorozměrné) analýzy
2) Programy vícerozměrné gnostické analýzy

Klíčové prostředky:

1) gnostické distribuční funkce (program GNDF)
2) program robustní vícerozměrné analýzy (GWLS)

JEDNOROZMĚRNÁ ANALÝZA

- Distribuční funkce (d.f.), výhody a použití
- EDF, ELDF, EGDF
- Homogenita datového souboru
- Meze datového souboru
- Apriorní a aposteriorní váhy dat
- Průřezová filtrace dat
- Cenzorovaná data
- Heteroskedastická data
- Robustnost d.f.
- Intervalová analýza datového souboru
- Analýza měřicích metod
- Výsledky aplikací na kontaminační i ekonomická data

DISTRIBUČNÍ FUNKCE (rozdělení pravděpodobnosti)

- Data … reálná čísla kvantifikující skutečné události
- Nosič dat …omezený interval reálných i očekávatelných dat
- Pravděpodobnost … míra očekávatelnosti dat, číslo z intervalu [0, 1]
- Distribuční funkce … izomorfismus nosiče dat a intervalu [0, 1].

NORMALITA

- Normální (jedinec, chování, jev, …):
  - obvyklý,
  - obyčejný,
  - očekávatelný,
  - v souladu s přijatým standardem.
- Normální rozdělení (Gaussovo):

Zkušenost: ***Normální rozdělení reálných jevů bývá normální (Gaussovo) spíše vzácně.***

DISTRIBUCE ELDF A EGDF

ELDF … Estimační ***lokální*** d.f.:

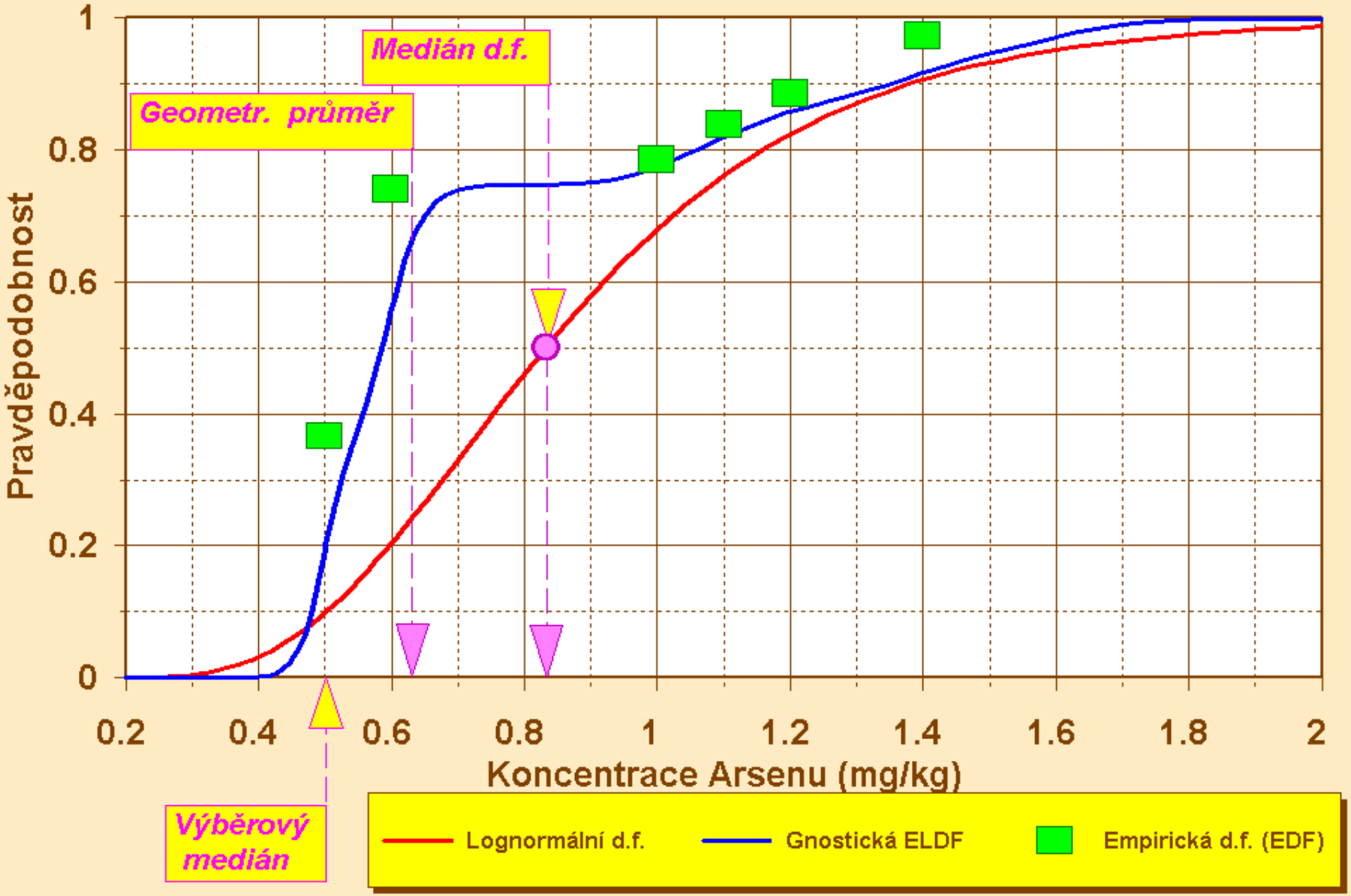
- aditivně skládá gnostická jádra,
- je pružná, může vystihnout detaily struktury datového souboru.

EGDF … Estimační ***globální*** d.f.:

- poskytuje celkový pohled na datový soubor,
- gnostická jádra skládá neaditivně,
- je robustní k odlehlým datům a shlukům.

2. Srovnání distribučních funkcí A

Arsen v rybách



Vyplývá:

- Rozdělení **není** lognormální
- Rozdělení nemá tvar ani jiného „standardního“ statistického rozdělení
- Bodové statistiky (geometrický průměr, medián distribuční funkce, výběrový medián) nevypovídají o datech nic použitelného

MEZE NOSIČE DAT A PŘÍSLUŠNOSTI K SOUBORU

- Statistické pojetí:
  - Dány meze nosiče dat (apriorní předpoklad)
  - Dán typ rozdělení (apriorní předpoklad)
  - Dána významnost testu (subjektivní rozhodnutí)
- Gnostické pojetí:
  - Dána data
  - Z dat jednoznačně plyne EGDF
  - EGDF jednoznačně určí meze nosiče dat i homogenního jádra datového souboru

MEZE DAT

- Meze nosiče dat (***LB*** a ***UB***):

  Jaká je nejnižší a nejvyšší očekávatelná hodnota dat tohoto souboru?

- Meze příslušnosti (***LSB*** a ***USB***)

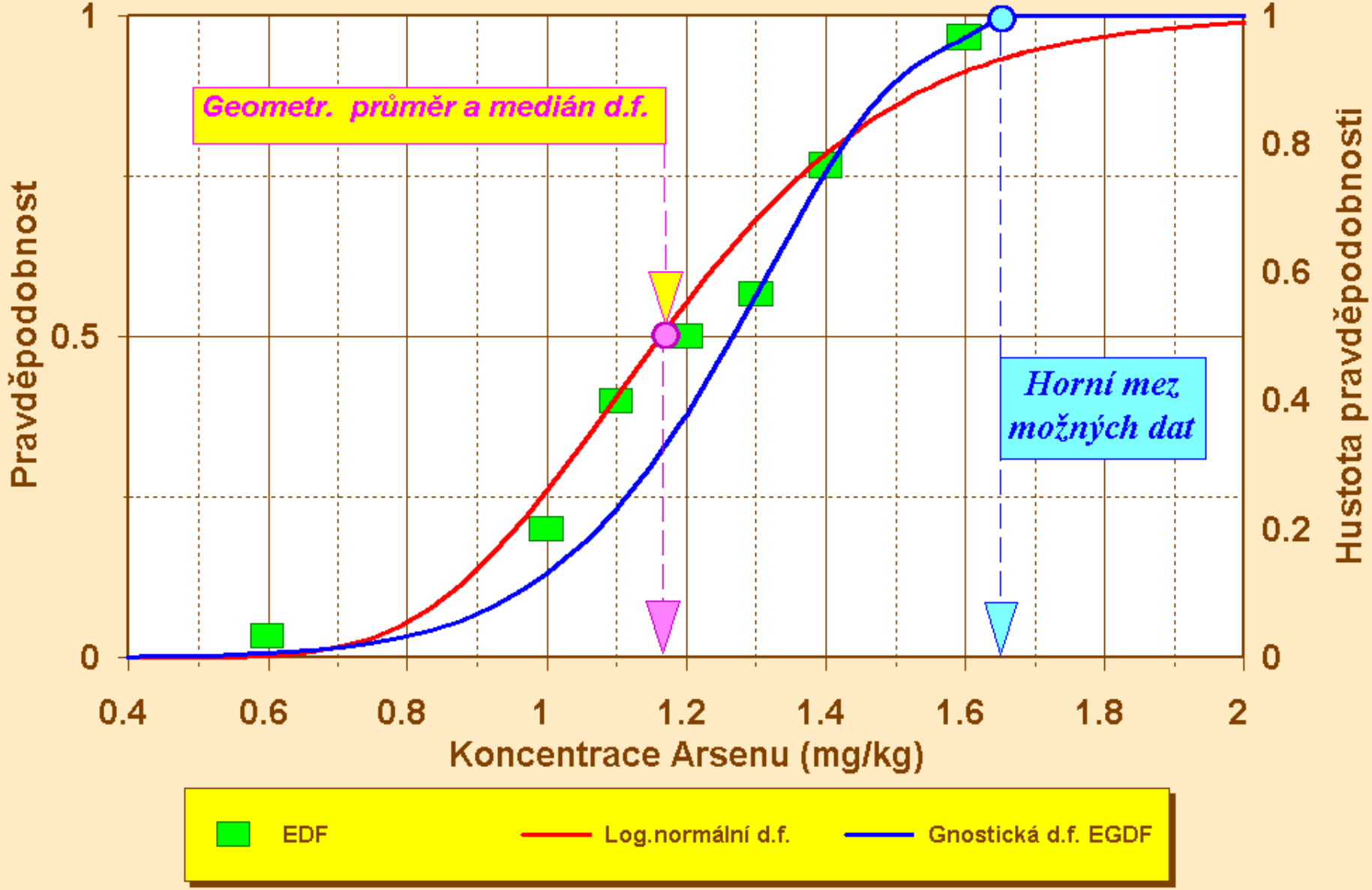
  k datovému souboru:

  Jaká je dolní a horní mez, jejíž překročení naruší homogenitu souboru?

PLATÍ: $LB <= LSB < USB <= UB$

3. Srovnání distribučních funkcí B

EGDF nadprahového arsenu v rybách

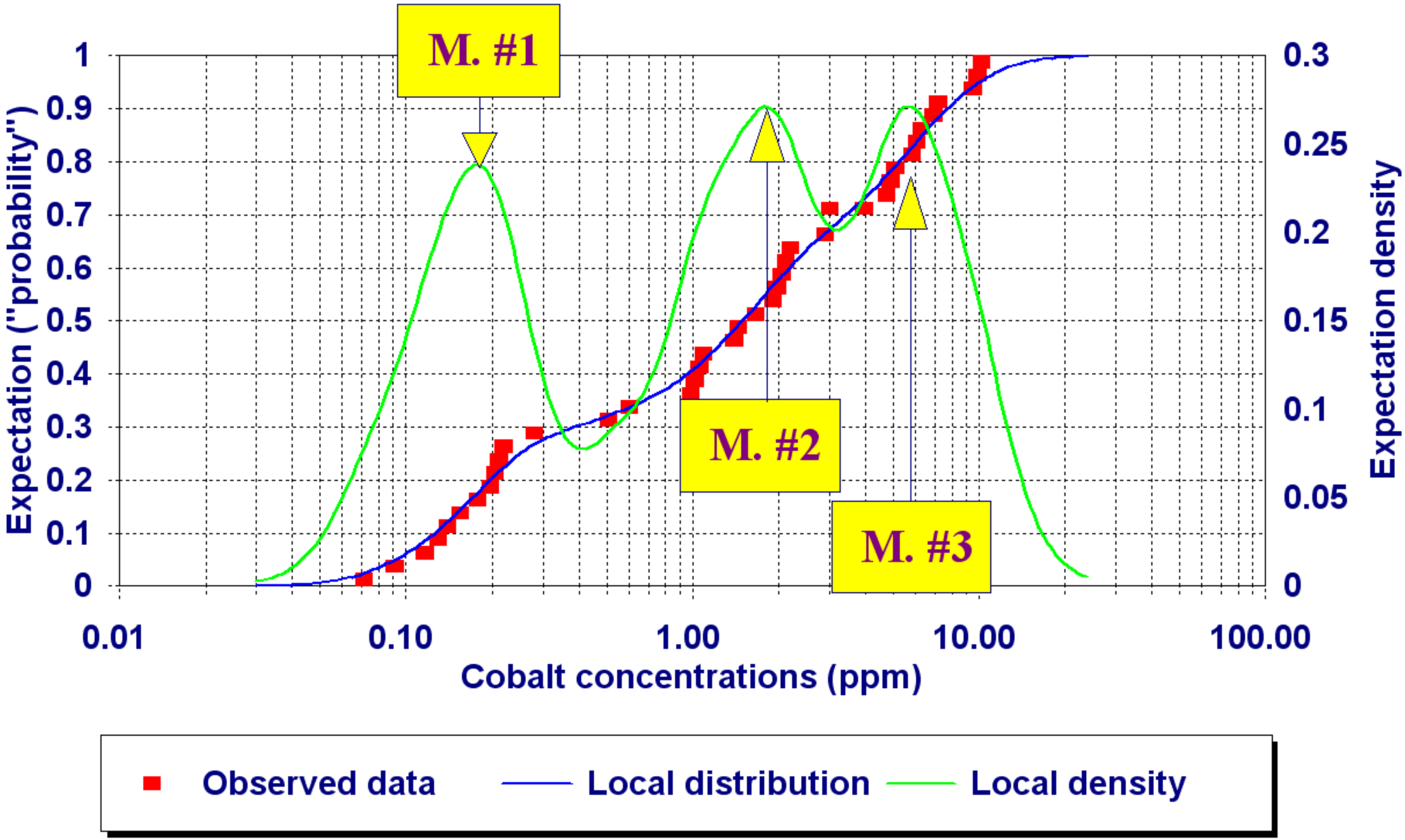


UŽITÍ ELDF

- Prezentace detailní struktury datového souboru
- Odhadování pravděpodobnosti i kvantilů (i nehomogenních) datových souborů
- Marginální shluková analýza
- Intervalová analýza:
  - klasifikační meze souboru,
  - třídění dat podle příslušnosti k podsouborům,
  - posuzování míry shody různých souborů dat

3 METHODS OF CHEMICAL ANALYSES

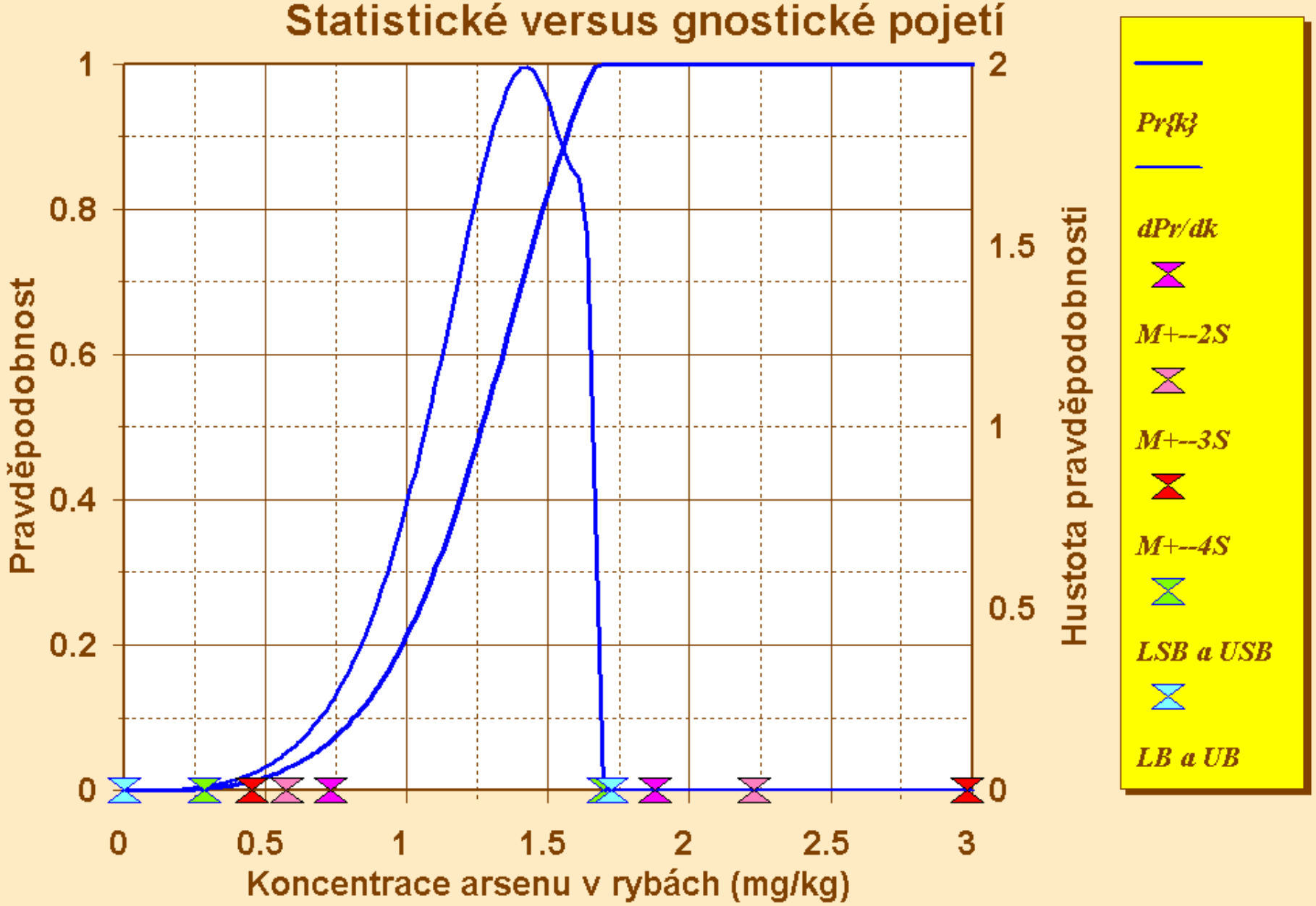
Cobalt contents in the granite samples



UŽITÍ EGDF

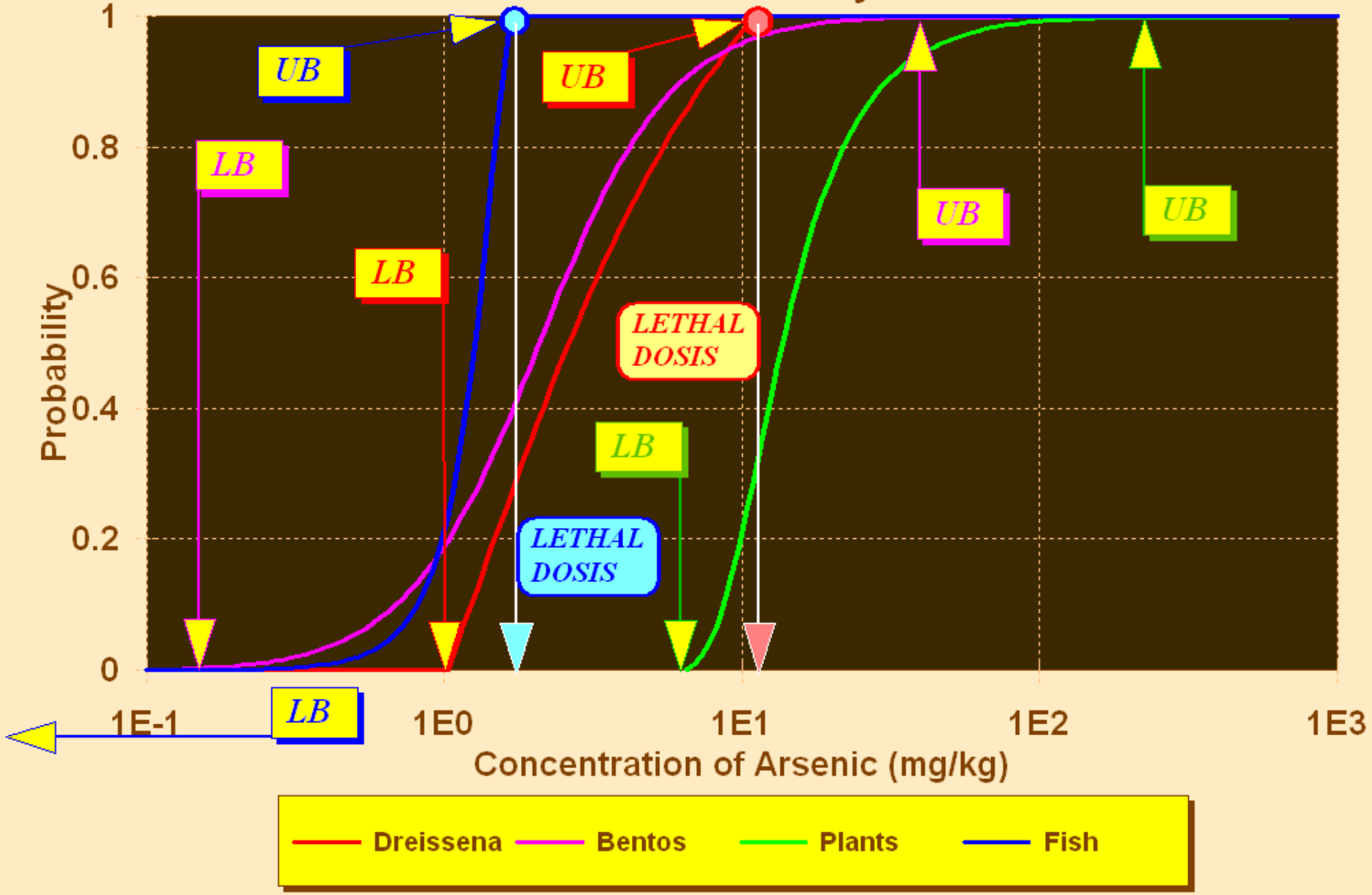
- Prezentace celkového pohledu na data
- Testování homogenity souboru
- Robustní odhadování
  - mezí nosiče dat,
  - mezí příslušnosti k datovému souboru,
  - parametrů měřítka a polohy souboru
  - pravděpodobností i kvantilů homogenních dat

11. MEZE DATOVÉHO SOUBORU



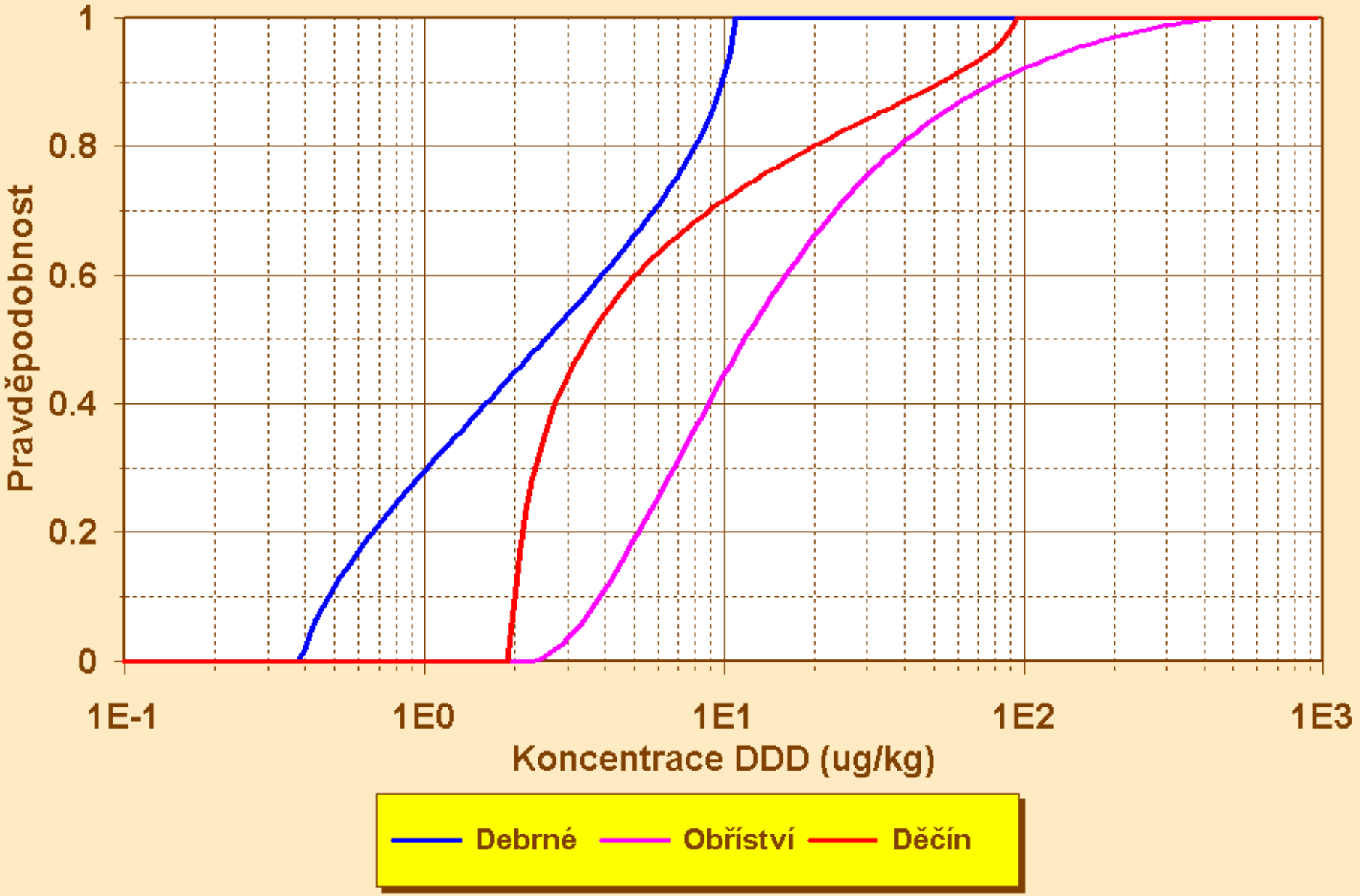
FOUR SAMPLING METHODS

Contamination of Water by Arsenic



26. SROVNÁNÍ MÍST VZORKOVÁNÍ

Kontaminace p.p.DDD v Labi

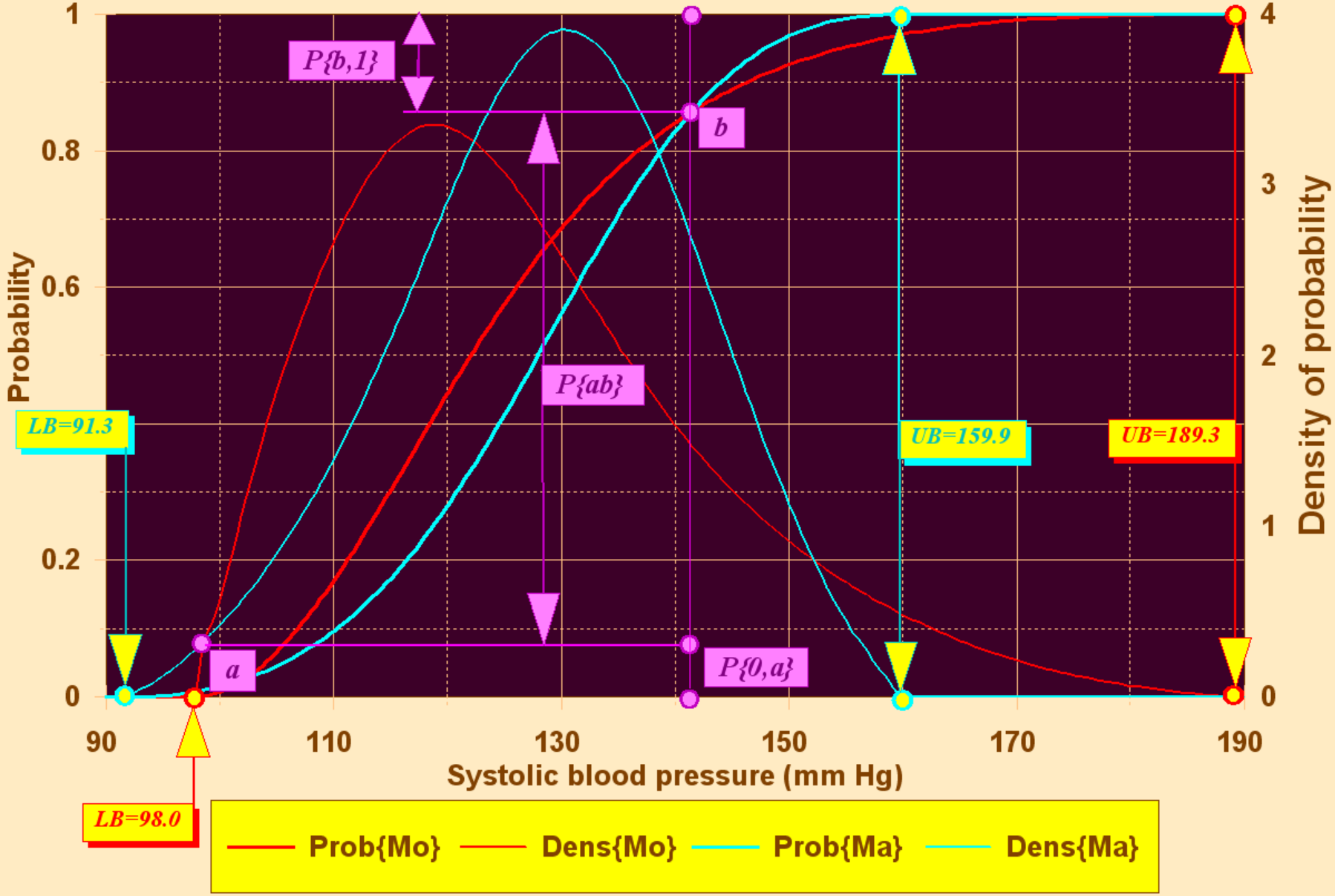


TESTOVÁNÍ HYPOTÉZ

- Data změřit
- Výsledky měření zpracovat (odhadnout „skutečné hodnoty“ a jejich vztahy)
- Interpretovat (navrhnout závěry)
- Otestovat pravděpodobnosti pravdivých i mylných závěrů

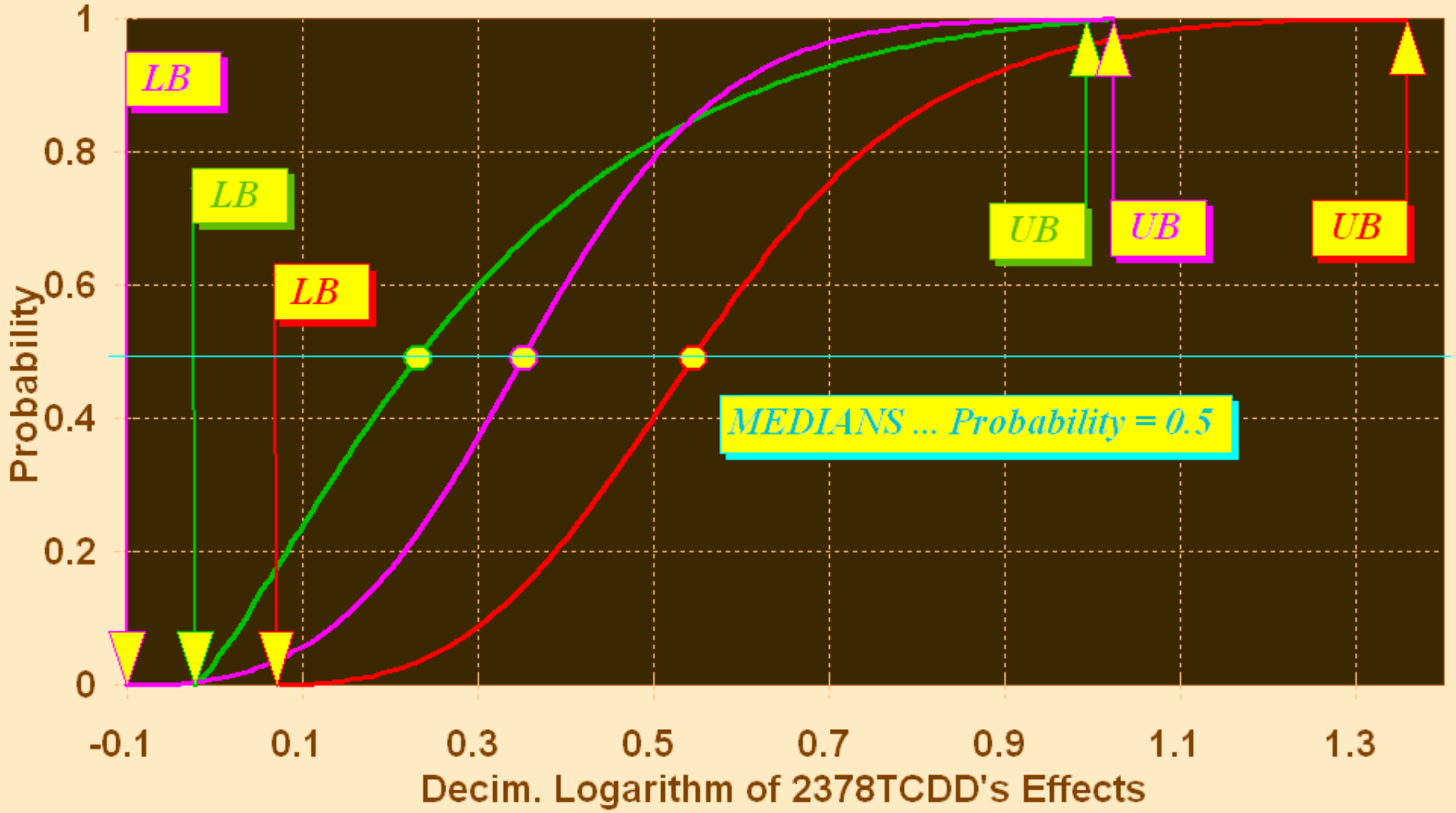
Fig.4: TEST: MEDICINE Mo AGAINST Ma

Is an alternative drug more efficient?



CENSORED DATA (2378TCDD)

44 Data of 60 Below the ST



All 60, 44 as Low-censored — All Data as Uncensored — Only 16 Data Exceeding ST

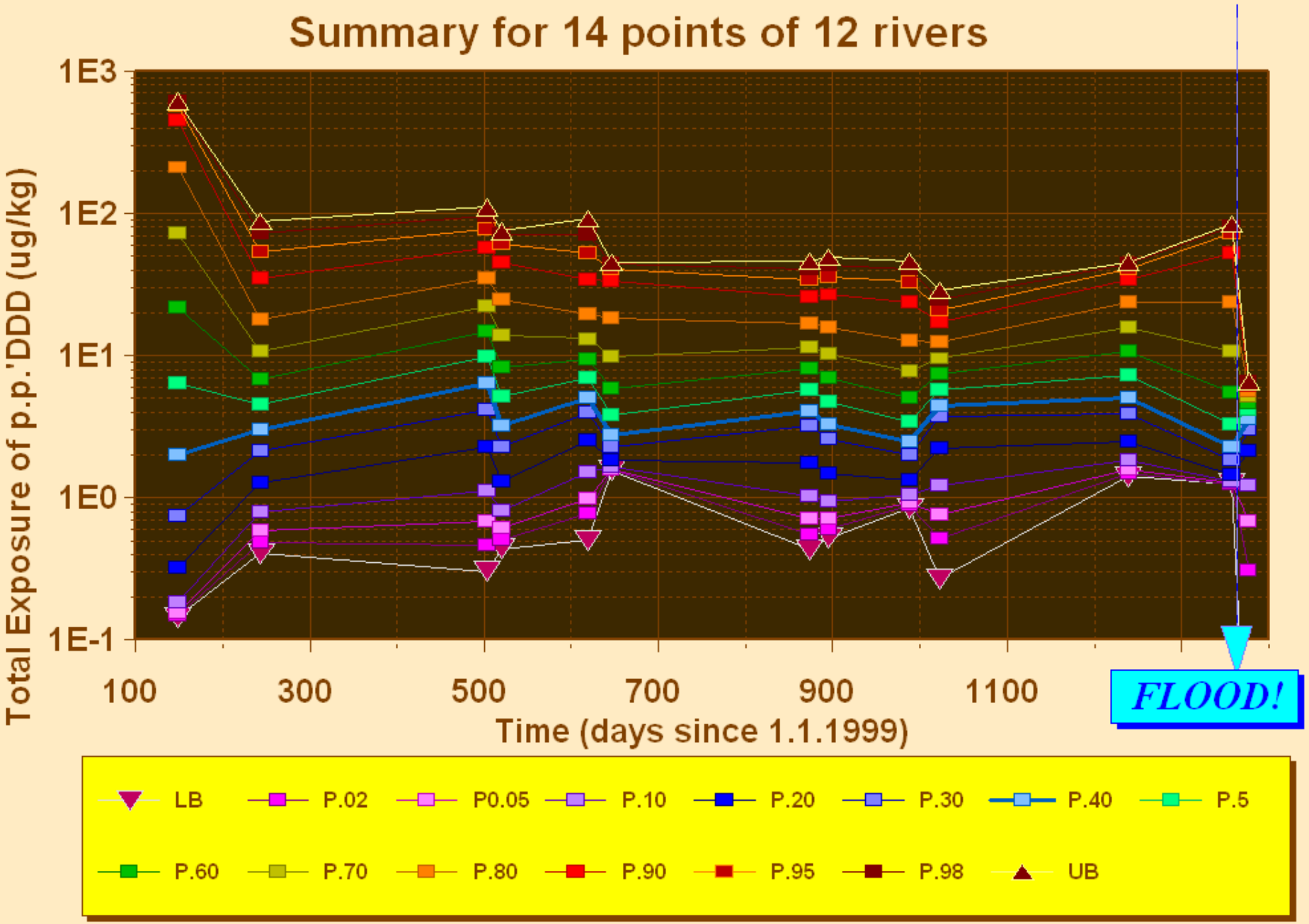
LB, UB ... the Bounds of the Data Support *ST ... the Sensitivity Threshold*

VÁHY JEDNOTLIVÝCH DAT

- ***Apriorní*** váhy … dány současně s daty, např. počet týchž dat, důvěryhodnost zdroje dat, známá přesnost měření apod. Vstupují do zpracování.
- ***Aposteriorní*** váhy … výsledek zpracování dat. Určeny algoritmem hodnotícím význam jednotlivých dat pro hledaný výsledek, jejich neurčitost a individuální příspěvek.

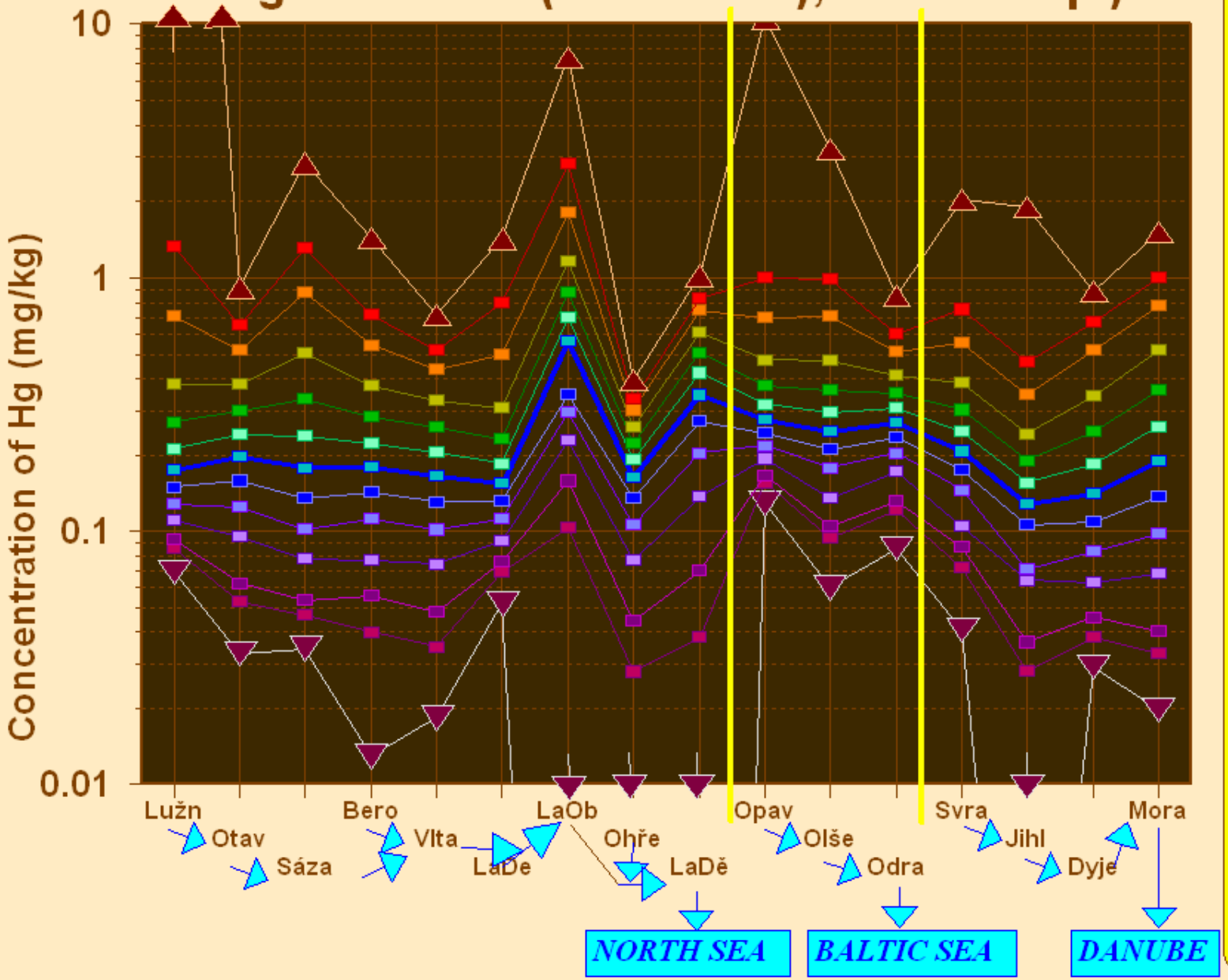
P.P.'DDD IN WATERS

Summary for 14 points of 12 rivers



CONTAMINATION OF RIVERS

Hg in Waters (1999-2002), Czech Rep.)



ROBUSTNOST D.F.

Robustnost d.f. …snížená citlivost ke „špatným“ datům, zdůraznění vlivu „dobrých“ dat.

EGDF a ELDF … robustní k ***odlehlým*** (periferním) datům a shlukům dat

QGDF a QLDF … robustní k ***vnitřním*** datům souborů dat

POLOZÁVĚRY

Analýza dat má ***dva*** hlavní cíle:

I. Zjistit, co říkají data o tom, ***co*** se stalo.

II. Zjistit, ***proč*** a ***jak*** se to stalo.

Postupy:

1) Kvalitní a promyšlené měření
2) Pokročilé metody zpracování dat
3) Odborná interpretace výsledků analýzy
4) Vyvození závěrů a **ZPĚT** K BODU 1).

STUDIE I: FAKTORY ZDRAVÍ

Přímé či nepřímé souvislosti se zdravím:

- Škodliviny (ekologie)
- Kouření
- Alkohol
- Obezita

POPs ŠKODÍ ZDRAVÍ

OZP…objektivní zdravotní potíže

(Prob{OZP>1}|POPs v 1Q) = 0.50

(Prob{OZP>1}|POPs v 4Q) = 0.87

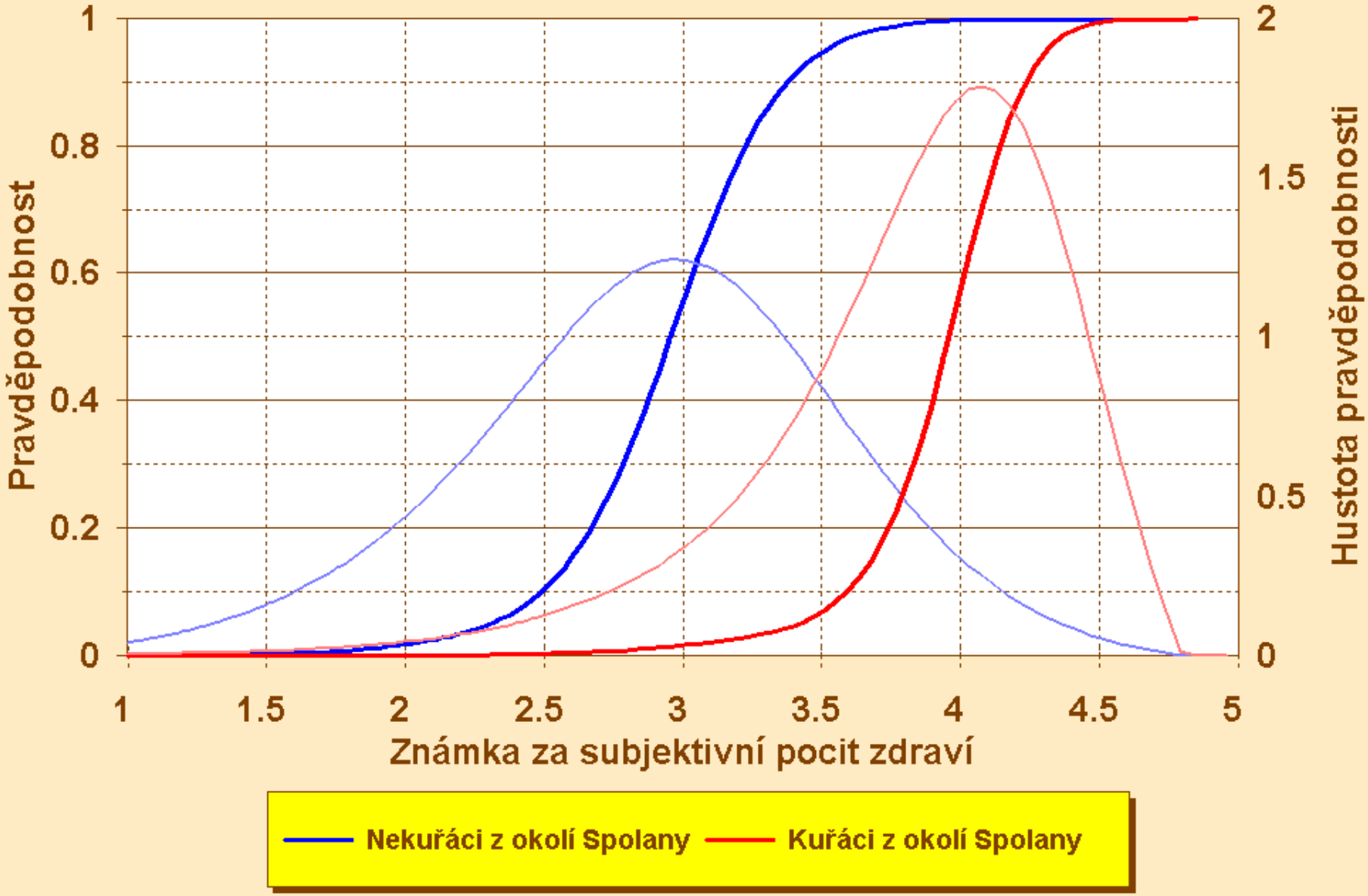
(POPs v 1Q) ➡ (UB = 8.1)

(POPs v 4Q) ➡ (UB = 982)

POŠKOZOVÁNÍ ZDRAVÍ PROKÁZÁNO A KVANTIFIKOVÁNO

Obr.14: SOUVISLOST ZDRAVÍ S KOUŘENÍM

Nekuřáci versus kuřáci (ok. Spolany)



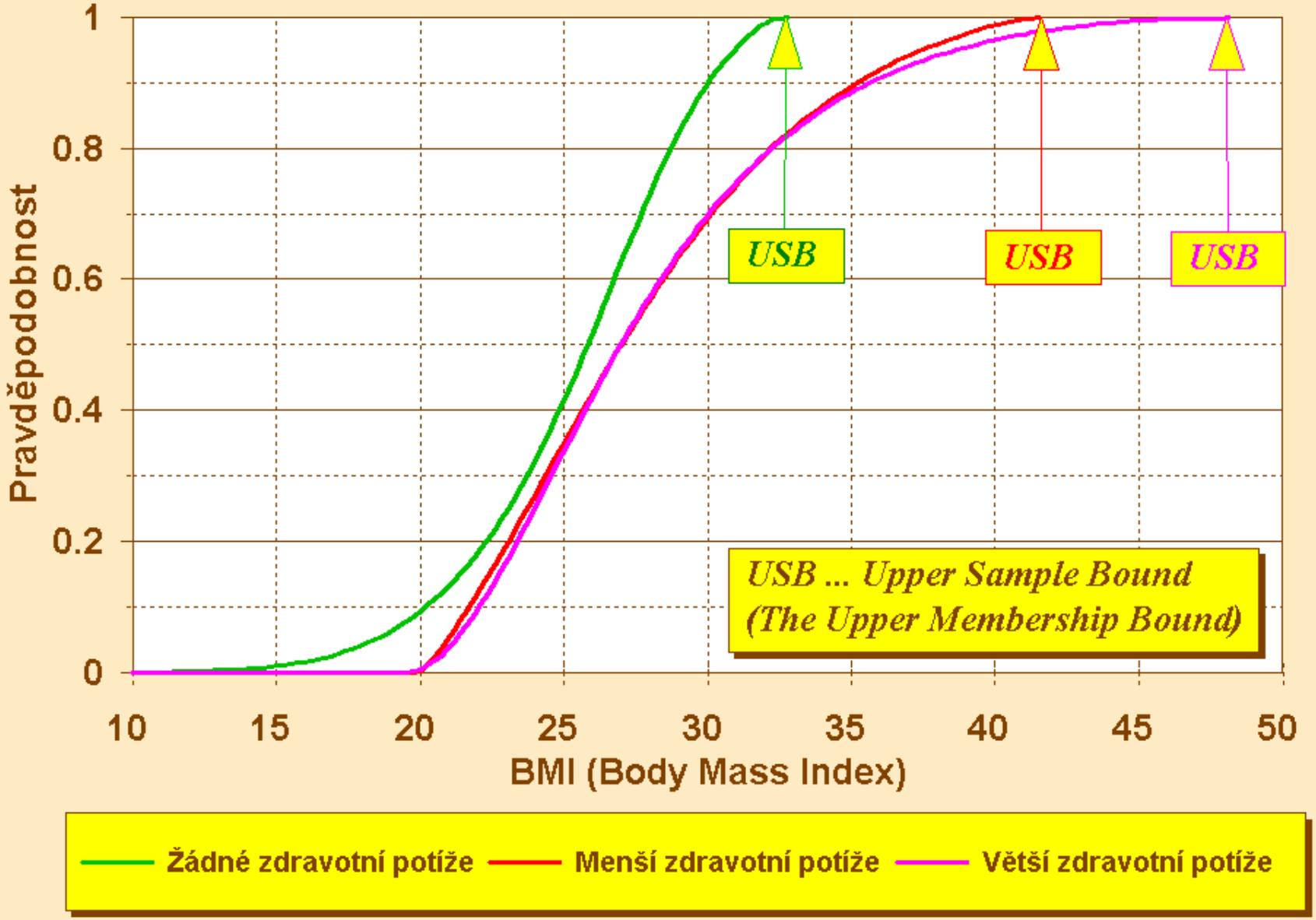
Souvislosti s kouřením

- Sami probandi dokládají: kouření ničí naše zdraví
- Kouření souvisí s rizikem vyšší akumulace POPs
- Kouření nesnižuje BMI, ačkoliv

  existuje významná souvislost zvyšování BMI se zhoršováním zdraví

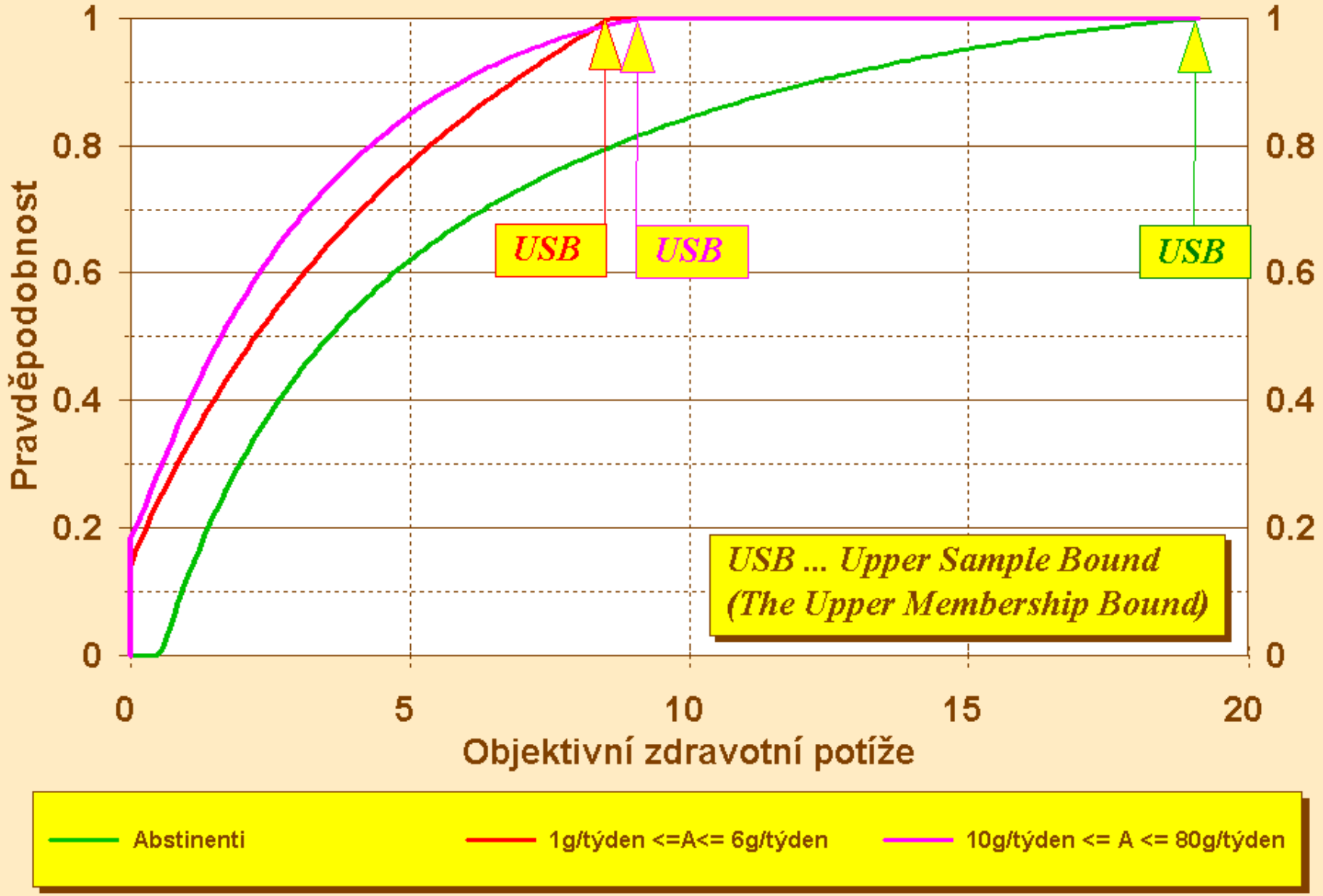
Obr.17: BMI A OBJEKTIVNÍ NEMOCNOST

Závislost BMI na objektivním zdraví



Obr.19: ALKOHOL A OBJEKTIVNÍ NEMOCNOST

Souvislost alkoholu s nemocností



Závěry o alkoholu

- Probandy členíme do tří shluků podle konzumace alkoholu
- Nikdo z probandů nebyl alkoholik
- Pití alkoholu se u probandů projevuje na zdraví příznivě
- Existují příznivé i nepříznivé souvislosti mezi pitím alkoholu a akumulací POPs

Vícerozměrná analýza

- Robustní korelační matice i pro statistiku:

  regresní analýza, hlavní komponenty, faktorová i diskriminační analýza
- Robustní MD regresní modely:

  explicitní,

  implicitní,

  v pravděpodobnostech
- Srovnatelnost vícerozměrných modelů
- Robustní uspořádání v MD prostoru
- Robustní vícerozměrná shluková analýza

SROVNÁNÍ METOD VÍCEROZMĚRNÉ ANALÝZY

Nmet=0 ... Klasická (nerobustní) statistická metoda *OLS* .
Nmet=1... Gnostická verze metody *IWLS*.
Nmet=2... Robustní statistická metoda Huberova.
Nmet=3... Robustní statistická metoda Hampelova.
Nmet=4... Robustní statistická metoda zvaná Bisquare

Nmet	*R-square*	*STDfitY*	*MeanW*	*MErr*	*MAErr*	*MsqErr*
0	0.3195	26.65	1.000	8.12e-14	15.92	0.320
1	0.9583	7.111	0.504	0.657	5.356	0.073
2	0.2005	19.82	0.959	1.383	12.27	0.198
3	0.3970	17.26	0.973	0.638	12.19	0.197
4	0.5740	12.72	0.915	0.522	10.02	0.152

NABÍDKA

- Dejte nám data: ukážeme, co z nich lze dostat
- Dáme vám gnostické programy
- Přihlašte se na naši letní školu
- Naučíme, pomůžeme při implementaci
- Detaily: příspěvek Ing.Pavlisky

tomas.ocelka@zuova.cz

kovanic@tiscali.cz

lubomir.pavliska@tiscali.cz