Štatistická inferencia v exponenciálnej rodine založená na /-divergencii

ROBUST 2014

Vladimíra Sečkárová, Radka Sabolová, Milan Stehlík

22.1.2014

KPMS MFF UK, PRAHA
ÚTIA, PRAHA
IFAS JKU, LINZ

Kullbackova Leiblerova divergencia

Kullbackovu Leiblerovu divergencia (KLD) [S. Kullback, R.A. Leibler (1951)]:

$$KLD(p||q) = \int_Y p(y) \ln \frac{p(y)}{q(y)} dy$$

- Y je spojitá náhodná veličina
- p, q sú pravdepodobnostné hustoty

Vlastnosti KLD:

- nie je symetrická
- nezáporná
- minimum rovné nule nadobúda $\leftrightarrow p = q$ a.e.
- nie je zhora ohraničená

Kullbackova Leiblerova divergencia

Využitie:

- testovanie hypotetickej hodnoty parametru rozdelenia zrážkových úhrnov

  S.V. Weijs, N. Van De Giesen (2011): *Accounting for Observational Uncertainty in Forecast Verification: An Information-Theoretical View on Forecasts, Observations, and Truth.*

- testovanie hypotetickej hodnoty parametru + asymptotické vlastnosti testovej štatistiky

  A. Basu, A. Mandal, N. Martin, L. Pardo (2013): *Testing statistical hypotheses based on the density power divergence.*

- Bregmanova mocninná vzdialenosť (vzťah ku KLD)

  A.L. Kisslinger, W. Stummer (2013): Some Decision Procedures Based on Scaled Bregman Distance Surfaces.

Exponenciálna rodina rozdelení

$Y_1, \ldots, Y_N$ - náhodný výber z rozdelenia patriaceho do exponenciálnej rodiny

Tvar pravdepodobnostnej hustoty rozdelenia patriaceho do exponenciálnej rodiny

$$h(\mathbf{y}|\theta) = \exp\left\{-\psi(\mathbf{y}) + T(\mathbf{y})^T\gamma(\theta) - \zeta(\gamma(\theta))\right\}$$

kde $\theta \in \Theta$, $T(\mathbf{y})$ je postačujúca štatistika

Podmienky regularity:

- výberový priestor je otvorenou množinou v R^N
- množina Θ je podmnožinou R^m, $m < N$
- zobrazenie $\gamma : \Theta \to R^k$ je dvakrát spojito diferencovateľnána int(Θ)
- . . . [Pázman 1993]

Podmienky hladkosti:

- Pre $\gamma \in R^k$ označme

$$\kappa(\gamma) = \ln \int_Y \exp\left\{-\psi(\mathbf{y} + T(\mathbf{y}^T\gamma))\right\} d\mathbf{y}$$

  Predpokladáme tiež, že množina

$$\Gamma = \left\{\gamma \in R^k : \kappa(\gamma) < \infty\right\}$$

  má neprázdne vnútro

- A tiež predpokladáme, že:

$$\{t(\mathbf{y} : \mathbf{y} \in Y)\} \subseteq \left\{\frac{\partial\kappa(\gamma)}{\partial\gamma} : \gamma \in \mathrm{int}(\Gamma)\right\}$$

Definícia I-divergencie

Majme súbor náhodných veličín $Y_1, \ldots, Y_N$, $N < \infty$, ktoré

- nezávislé
- rozdělených podľa rozdelenia z exponenciálnej rodiny s pravdepodobnostnou hustotu tvaru:

$$h(y|\gamma) = \exp\{-\psi(y) + T(y)\gamma - \kappa(\gamma)\} \tag{1}$$

Ďalej budeme predpokladať:

$$\{t(\mathbf{y} : \mathbf{y} \in Y)\} \subseteq \{\mathrm{E}_\gamma[T(\mathbf{y})] : \gamma \in \mathrm{int}(\Gamma)\}$$

V regulárnej exponenciálnej rodine máme podľa [Barndorff-Nielsen (1978)]:

$$\mathrm{E}_\gamma[T(\mathbf{y})] = \frac{\partial \kappa(\gamma)}{\partial \gamma}$$

Definícia I-divergencie

I-divergenciu potom definujeme ako

$$\begin{aligned} I_N(\boldsymbol{y}, \gamma^*) &= KLD(h(\boldsymbol{y}|\widehat{\gamma}_{\boldsymbol{y}})||h(\boldsymbol{y}|\gamma^*)) \\ &= \int_{\boldsymbol{y}} h(\boldsymbol{y}|\widehat{\gamma}_{\boldsymbol{y}}) \ln \frac{h(\boldsymbol{y}|\widehat{\gamma}_{\boldsymbol{y}})}{h(\boldsymbol{y}|\gamma^*)} d\boldsymbol{y} \\ &= \sum_{i=1}^{N} (T(y_i)\hat{\gamma}_i + \ln g(\hat{\gamma}_i)) - \sum_{i=1}^{N} (T(y_i)\gamma_i^* + \ln g(\gamma_i^*)) . \end{aligned}$$

- $\widehat{\gamma}_{\boldsymbol{y}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_N)^T$ - vektor maximálne vierohodných odhadov založených na y_i, $i = 1, \ldots, N$
- γ^* vektor obsahujúci N zvolených hodnot
- Z podmienok na hladkosť sme využili vzťah

$$\frac{\partial \kappa(\gamma)}{\partial \gamma} = T(\mathbf{y})$$

Rozklad I_N na R_N, S_N

Uvažujme maximálne vierohodný odhad $\hat{\gamma}_{MLE}$ založený na $y_1, \ldots, y_N$
Definujme R_N, S_N

$$R_N = \sum_{i=1}^{N} (\hat{\gamma}_{MLE} T(y_i) + \ln(g(\hat{\gamma}_{MLE}))) - \sum_{i=1}^{N} (\gamma_0 T(y_i) + \ln(g(\gamma_0)))$$

$$S_N = \sum_{i=1}^{N} (T(y_i)\hat{\gamma}_i + \ln(g(\hat{\gamma}_i))) - \sum_{i=1}^{N} (T(y_i)\hat{\gamma}_{MLE} + \ln(g(\hat{\gamma}_{MLE})))$$

Vidíme, že

$$I_N = R_N + S_N$$

Interpretácia častí R_N, S_N

S_N reprezentuje testovú štatistiku testu pomerom vierohodností:

$$H_0 : \gamma_1 = \ldots = \gamma_N \quad \text{vs.} \quad H_1 : \text{ľubovoľné vhodné } \gamma, \quad \gamma \in \Gamma$$

R_N representuje testovú štatistiku testu pomerom vierohodností:

$$H_0 : \gamma_1 = \ldots = \gamma_N = \gamma_0 \quad \text{vs.} \quad H_1 : \gamma \neq \gamma_0.$$

Nezávislosť R_N, S_N

- Gamma rozdelenie: R_N, S_N sú nezávislé, odvodené v [Stehlík (2003)].
- Exponenciálne rozdelenie: je špeciálny prípad Gamma rozdelenia, R_N, S_N sú nezávislé.
- Pareto rozdelenie: použijeme vzťah

$$Y \sim Pareto(x_m, \alpha) \rightarrow X = \ln\left(\frac{Y}{x_m}\right) \sim Exp(\alpha)$$

Potom máme R_N, S_N nezávislé.

Úvod

- $Y_1, \ldots, Y_N$ nezávislé, $Exp(\gamma)$, kde $\gamma = (\gamma_1, \ldots, \gamma_N)^T$
- hustota Y_i

$$h(y_i|\gamma_i) = \gamma_i e^{-\gamma_i y_i}, \quad y_i \geq 0$$

Úvod

- $Y_1, \ldots, Y_N$ nezávislé, $Exp(\boldsymbol{\gamma})$, kde $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^T$
- hustota Y_i

$$h(y_i|\gamma_i) = \gamma_i e^{-\gamma_i y_i}, \quad y_i \geq 0$$

- I-divergencia

$$I_N(\mathbf{y}, \gamma^*) = \int_{\mathbf{y}} h(\mathbf{y}|\widehat{\gamma}_{\mathbf{y}}) \ln \frac{h(\mathbf{y}|\widehat{\gamma}_{\mathbf{y}})}{h(\mathbf{y}|\gamma^*)} d\mathbf{y}$$

Úvod

- $Y_1, \ldots, Y_N$ nezávislé, $Exp(\gamma)$, kde $\gamma = (\gamma_1, \ldots, \gamma_N)^T$
- hustota Y_i

$$h(y_i|\gamma_i) = \gamma_i e^{-\gamma_i y_i}, \quad y_i \geq 0$$

- I-divergencia

$$I_N(\mathbf{y}, \gamma^*) = \int_{\mathbf{y}} h(\mathbf{y}|\widehat{\gamma}_{\mathbf{y}}) \ln \frac{h(\mathbf{y}|\widehat{\gamma}_{\mathbf{y}})}{h(\mathbf{y}|\gamma^*)} d\mathbf{y}$$

- pre exponenciálne rozdelené náhodné veličiny

$$I_N(\mathbf{y}, \gamma_0) = N \ln \sum_i y_i - \sum_i \ln y_i + \gamma_0 \sum_i y_i - N \ln(\gamma_0 \sum_i y_i) - N$$

Rozdelenie $I_N(y, \gamma_0)$

- pre $N \leq 4$, distribučná funkcia I_N je známa a je možné ju zapísať analyticky
- pre $N = 1$

$$F_1(x) = \begin{cases} \exp\{{\gamma_0}^{-1}\gamma W_0(-\exp\{-1-x\}\} - \exp\{{\gamma_0}^{-1}\gamma W_{-1}(-\exp\{-1-x\}\} \\ 0 \end{cases}$$

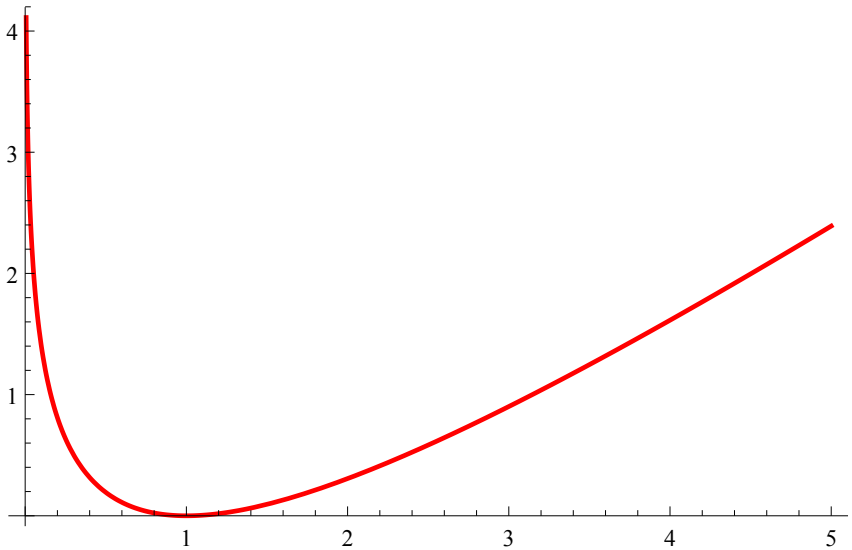
kde W_0, W_{-1} sú dve reálne vetvy Lambertovej W-funkcie

- Lambertova W-funkcia je riešením rovnice

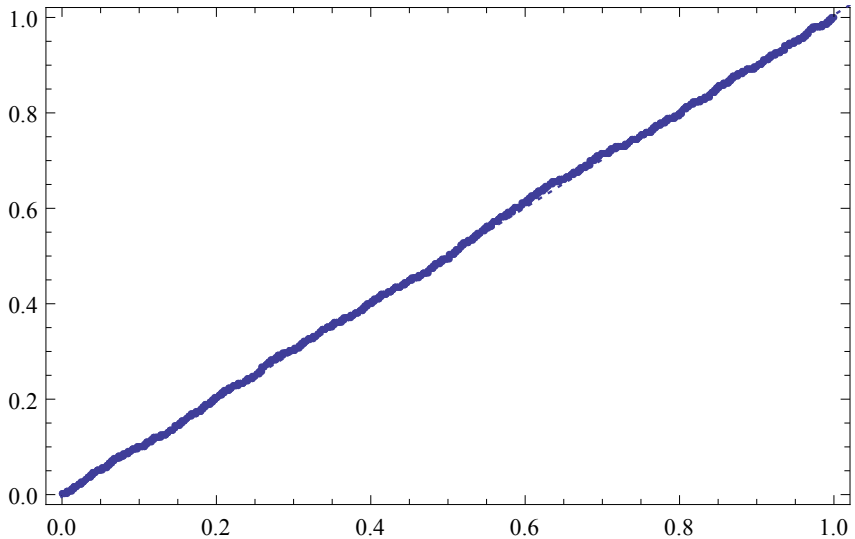
$$z = W(z)e^{W(z)}$$

pre ľubovoľné z komplexné

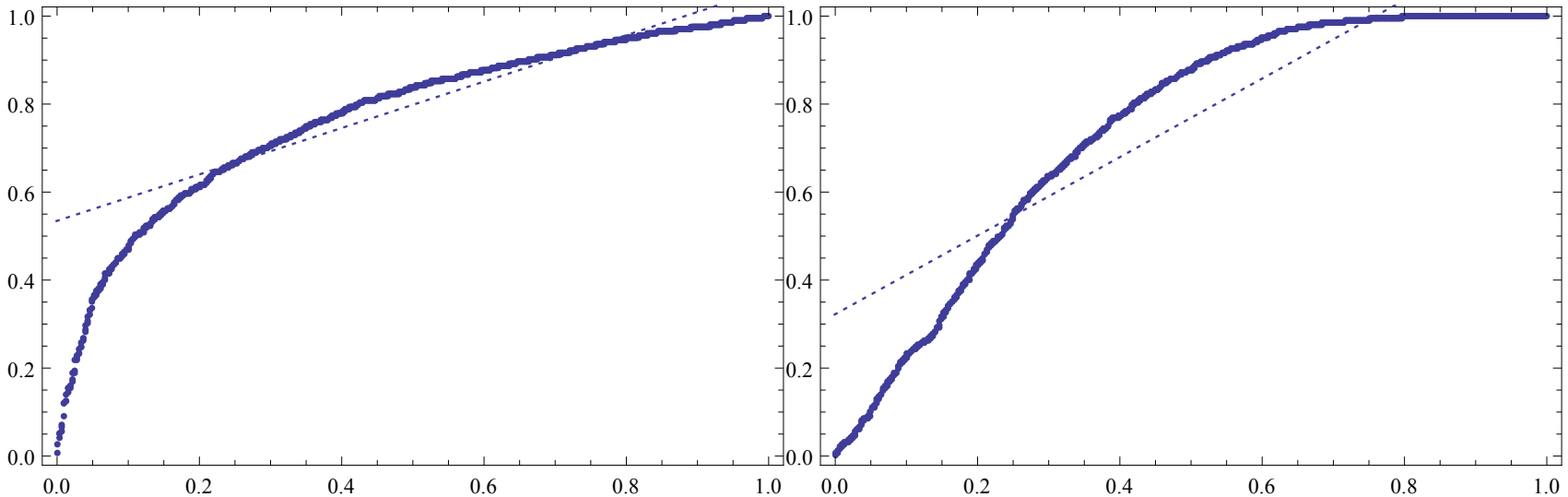
Graf I_1



QQ-plot $F_1(I_1(y, \gamma_0))$ pre $\gamma = \gamma_0 = 1$



QQ-plot $F_1(I_1(y, \gamma_0))$ pre $\gamma \neq \gamma_0$



Rozdelenie $I_N(y, \gamma_0)$ pre $N > 1$

- distribuyná funkcia I_N odvodená analyticky pre $N \leq 4$, ale komplikovaná
- využijeme dekompozíciu $I_N = R_N + S_N$, asymptotické rozdelenie R_N a približné rozdelenie S_N, kde

$$R_N = \gamma_0 \sum_i y_i - N \ln(\gamma_0 \sum_i y_i) - N + N \ln N$$

$$S_N = N \ln \sum_i y_i - N \ln N - \sum_i \ln y_i$$

Rozdelenie $I_N(y, \gamma_0)$ pre $N > 1$

- distribuyná funkcia I_N odvodená analyticky pre $N \leq 4$, ale komplikovaná
- využijeme dekompozíciu $I_N = R_N + S_N$, asymptotické rozdelenie R_N a približné rozdelenie S_N, kde

$$R_N = \gamma_0 \sum_i y_i - N \ln(\gamma_0 \sum_i y_i) - N + N \ln N$$

$$S_N = N \ln \sum_i y_i - N \ln N - \sum_i \ln y_i$$

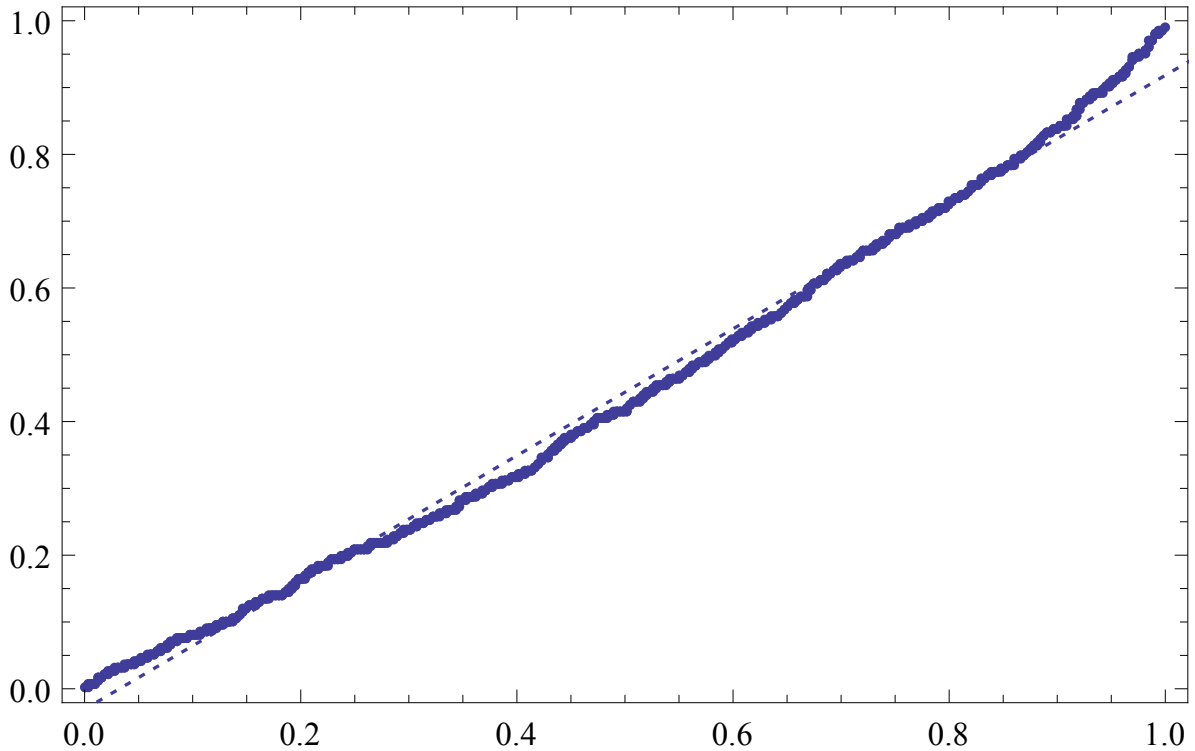
- asymptoticky

$$R_N \sim \chi^2_1$$

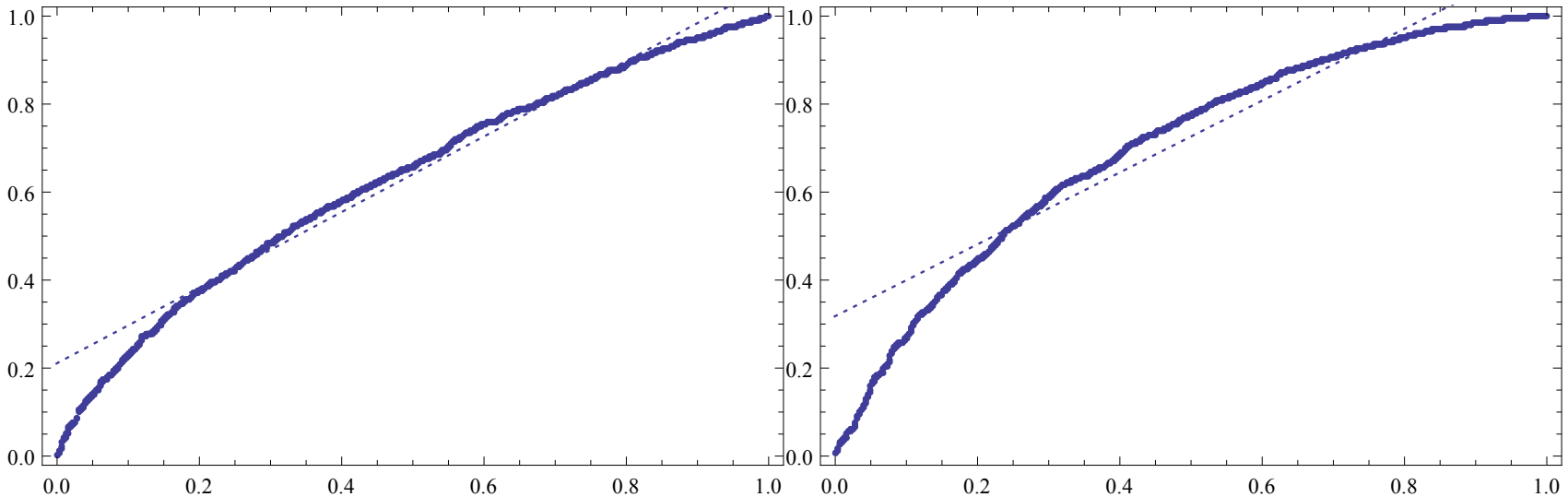
- približne

$$S_N \sim \frac{1}{2}\left(1 + \frac{1 + \frac{1}{N}}{6}\right)\chi^2_{N-1}$$

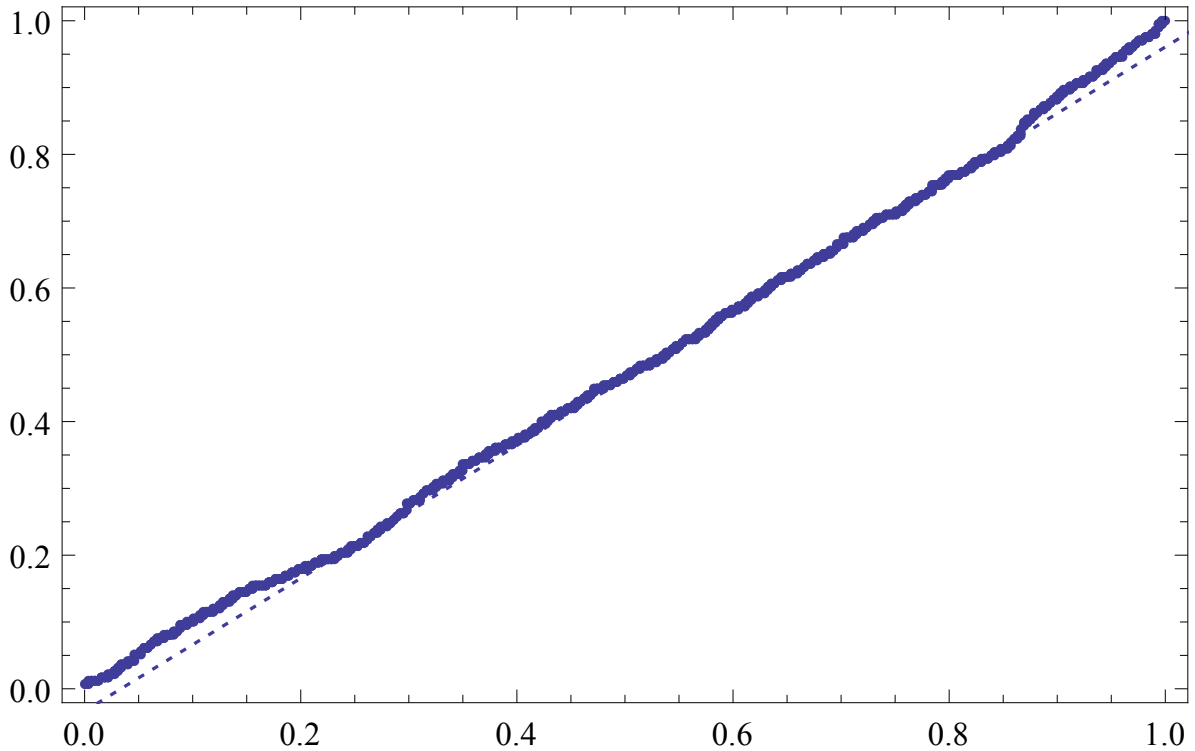
QQ-plot pre $\hat{F}_{10}(I_{10}(y, \gamma_0)), \gamma = \gamma_0 = 1$



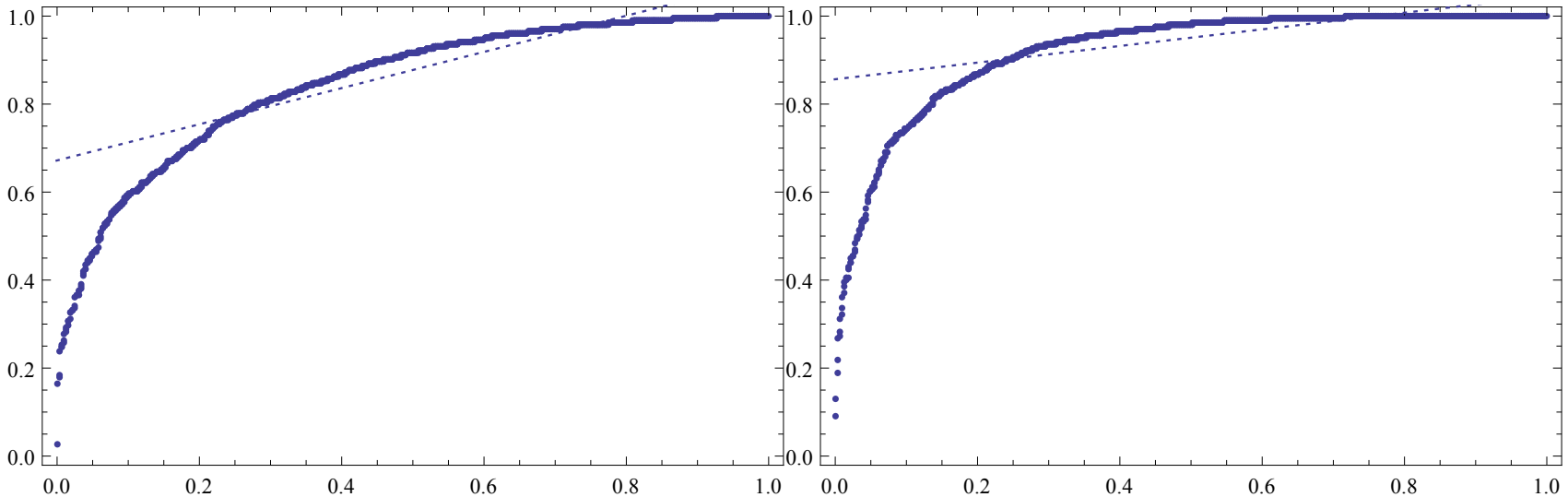
QQ-plot pre $\hat{F}_{10}(I_{10}(y, \gamma_0)), \gamma \neq \gamma_0$



QQ-plot pre $\hat{F}_{50}(I_{50}(y, \gamma_0)), \gamma = \gamma_0 = 1$



QQ-plot pre $\hat{F}_{50}(I_{50}(y, \gamma_0)), \gamma \neq \gamma_0$



Aproximácia hustoty MLE

- saddlepoint aproximácia hustoty postačujúcej štatistiky

$$q_T(t|\gamma) = (2\pi)^{-N/2} \prod_i \gamma_i \exp\{t_i\gamma_i + 1\}$$

- saddlepoint aproximácia hustoty MLE

$$q_T(\hat{\gamma}|\gamma) = (2\pi)^{-N/2} \prod_i \frac{\gamma_i}{\hat{\gamma}_i} \exp\left\{1 - \frac{\gamma_i}{\hat{\gamma}_i}\right\}$$

Referencie

S. Kullback, R.A. Leibler (1951): On Information and Sufficiency, *Ann. Math. Statist.* 22(1) 79–86.

S.V. Weijs, N. Van De Giesen (2011): Accounting for Observational Uncertainty in Forecast Verification: An Information-Theoretical View on Forecasts, Observations, and Truth, *Mon. Weather Rev.* 139 2156–2162.

A. Basu, A. Mandal, N. Martin, L. Pardo (2013): Testing statistical hypotheses based on the density power divergence, Ann Inst Stat Math 65 319–348.

A.L. Kisslinger, W. Stummer (2013): Some Decision Procedures Based on Scaled Bregman Distance Surfaces, in F. Nielsen, F. Barbaresco (Eds.): GSI 2013, LNCS 8085 479-–486. Springer-Verlag

Referencie

O.E. Barndorff-Nielsen (1978): Information and exponential families in statistical theory. John Wiley & Sons, New York 111–115.

M. Stehlík (2003): Distributions of exact tests in the exponential family, Metrika 57 145–164.

M.S. Bartlett, D.G. Kendall (1946): The statistical analysis of variance-heterogeneity and the logarithmic transformation, Supplement to the Journal of the Royal Statistical Society 8(1) 128–138.

Ďakujeme za pozornosť.