VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

STATISTICKÁ ANALÝZA ROC KŘIVEK

STATISTICAL ANALYSIS OF ROC CURVES

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. DAVID KUTÁLEK

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. JAROSLAV MICHÁLEK CSc.

BRNO 2010

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2009/2010

ZADÁNÍ DIPLOMOVÉ PRÁCE

student(ka): Bc. David Kutálek

který/která studuje v **magisterském navazujícím studijním programu**

obor: **Matematické inženýrství (3901T021)**

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Statistická analýza ROC křivek

v anglickém jazyce:

Statistical analysis of ROC curves

Stručná charakteristika problematiky úkolu:

V současné době se v inženýrských diagnostických postupech, v medicínské diagnostice a v celé řadě dalších oborů používají ke klasifikaci populací ROC křivky (z anglického Receiver Operating Chracteristic Curve). Značně rychle se rozvíjejí metody umožňující provádět statistickou analýzu reálných populací pomocí ROC křivek. V mnohých situacích chybí srovnání jednotlivých metod, popis jejich statistických vlastností a mnohdy není k dispozici odpovídající programátorské zázemí pro výpočet odhadů a pro provedení příslušných statistických testů.

Cíle diplomové práce:

V práci popište statistické metody pro stanovení bodového a intervalového odhadu ROC křivky v daném bodě, metody pro odhad plochy pod ROC křivkou a odhad optimální diagnostické klasifikační meze. Dále popište vybrané statistické testy pro testování hypotéz o vlastnostech dané ROC křivky a testy pro srovnání dvou ROC křivek. Popsané metody algoritmizujte a proveďte jejich srovnání. Dále proveďte jejich počítačovou implementaci v prostředí MATLAB. Funkčnost vytvořených programů demonstrujte na simulovaných i reálných datech.

Seznam odborné literatury:

X.H. Zhou, N.A. Obuchowski and D.K. McClish: Statistical methods in Diagnostic Medicine. John Wiley. 2002

Pepe, M. S.: The Statistical Evaluation of~Medical Tests for Classification and Prediction. Oxford Statistical Science Series, Oxford University Press. 2003

Vedoucí diplomové práce: doc. RNDr. Jaroslav Michálek, CSc.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2009/2010.

V Brně, dne 20.11.2009

L.S.

prof. RNDr. Josef Šlapal, CSc.	prof. RNDr. Miroslav Doupovec, CSc.
Ředitel ústavu	Děkan fakulty

Abstrakt

ROC křivka (z anglického Receiver Operating Characteristic curve) je zobrazení dvou různých distribučních funkcí F_0 a F_1, kde na osy se vynáší hodnoty $1 - F_0(c)$ a $1 - F_1(c)$. Parametr c je reálné číslo. Takto sestrojená křivka se v poslední době často využívá k posouzení kvality diskriminačního pravidla pro zařazení objektu do jedné ze dvou tříd. Jako kritérium pak slouží velikost plochy pod ROC křivkou. V reálných úlohách se pak uplatňují metody bodových a intervalových odhadů ROC křivek a testování statistických hypotéz o ROC křivkách.

Summary

The ROC (Receiver Operating Characteristic) curve is a projection of two different cumulative distribution functions F_0 and F_1. On axis are values $1 - F_0(c)$ and $1 - F_1(c)$. The c-parameter is a real number. This curve is useful to check quality of discriminant rule which classify an object to one of two classes. The criterion is a size of an area under the curve. To solve real problems we use point and interval estimation of ROC curves and statistical hypothesis tests about ROC curves.

Klíčová slova
ROC křivka, klasifikace objektu, plocha pod křivkou, bodový odhad, intervalový odhad, test statistické hypotézy.

Keywords
ROC curve, object classification, area under curve, point estimation, interval estimation, statistical hypothesis test.

KUTÁLEK, D. *Statistická analýza ROC křivek.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2010. 53 s. Vedoucí diplomové práce doc. RNDr. Jaroslav Michálek, CSc.

Prohlašuji, že jsem diplomovou práci *Statistická analýza ROC křivek.* vypracoval samostatně pod vedením doc. RNDr. Jaroslava Michálka CSc. s použitím materiálů uvedených v seznamu literatury.

David Kutálek

Děkuji doc. RNDr. Jaroslavu Michálkovi CSc. za vedení mé diplomové práce.

David Kutálek

Obsah

1 Úvod

ROC křivky (z anglického *RECEIVER OPERATING CHARACTERISTIC CURVE*) používáme při rozřazení objektů do dvou tříd, přičemž víme, že daný objekt patří právě do jedné z nich. Plocha pod křivkou pak udává kvalitu rozhodovacího kritéria.

Poprvé byly využity k vojenským účelům. Během II. světové války sloužily při analýze radarových signálů, kdy bylo třeba rozlišit vlastní vzdušné síly a nepřítele. Odtud vzniklo označení ROC. Od padesátých let pak nachází uplatnění v medicíně při vyhodnocení testování nových léků a v diagnostice.

Dnes se optimalizace klasifikace pomocí ROC křivek používá v řadě oborů. Značně rychle se rozvíjejí metody umožňující provádět statistickou analýzu reálných populací pomocí ROC křivek. V mnohých situacích chybí srovnání jednotlivých metod, popis jejich statistických vlastností a mnohdy není k dispozici odpovídající programátorské zázemí pro výpočet odhadů a pro provedení příslušných statistických testů.

Cílem této práce bude popsat statistické metody pro stanovení bodového a intervalového odhadu ROC křivky v daném bodě, metody pro odhad plochy pod ROC křivkou a odhad optimální diagnostické klasifikační meze. Dále budou popsány statistické testy pro testování hypotéz o vlastnostech dané ROC křivky a testy pro srovnání dvou ROC křivek. Výstupem této práce bude také počítačová implementace jednotlivých metod v prostředí MATLAB.

2 Základní pojmy

V této kapitole budou uvedeny základní pojmy, označení a poznatky z teorie odhadu a testování statistických hypotéz a to v souladu s [1]. Tyto budou dále využity pro popis vlastností ROC křivek a k jejich vzájemnému srovnání.

Označme ω výsledek náhodného pokusu nebo děje, tento nazýváme *elementární jev.* Množinu všech elementárních jevů značíme Ω a nazýváme ji *prostor elementárních jevů.* Mějme systém podmnožin Ω tvořící σ-algebru $\mathcal{A}$. Pak tyto podmnožiny nazýváme *náhodné jevy.* Jednotlivým množinám patřícím do $\mathcal{A}$ pak připisujeme pravděpodobnostní míru P. Trojice $(\Omega, \mathcal{A}, P)$ se nazývá *pravděpodobnostní prostor.*

Definice 2.1. Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor. Dále nechť $\mathbb{R}$ je množina reálných čísel a $\mathcal{B}$ systém jejích borelovských množin. Měřitelnou funkci $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ nazveme *náhodnou veličinou.*

Definice 2.2. Označme $Q(B)$ pravděpodobnost, že náhodná veličina X náleží do množiny B z $\mathcal{B}$ (tedy $Q(B) = P\{X \in B\},\ B \in \mathcal{B}$). Míra Q se nazývá *indukovaná míra* nebo také *rozdělení pravděpodobnosti* náhodné veličiny X.

Zvolíme-li konkrétně $B = (-\infty, x)$, dostáváme

$$Q(B) = P\{X < x\} = F(x).$$

Funkce $F(x)$ se nazývá *distribuční funkce.*

Existuje-li taková funkce $f(x)$, že

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t,$$

pak se jedná o *spojité rozdělení pravděpodobnosti* s *hustotou* f.

Definice 2.3. Měřitelné zobrazení $\boldsymbol{X} : (\Omega, \mathcal{A}) \to (\mathbb{R}^n, \mathcal{B}_n)$, kde $\mathbb{R}^n$ je n-rozměrný euklidovský prostor a $\mathcal{B}_n$ systém jeho borelovských podmnožin, nazýváme *náhodný vektor.* (Jinými slovy jde o vektor náhodných veličin $\boldsymbol{X} = (X_1, \ldots, X_n)'$ definovaných na témže pravděpodobnostním prostoru.)

Definice 2.4. Náhodné veličiny $X_1, \ldots, X_n$ se nazývají *nezávislé*, platí-li pro libovolné borelovské množiny vztah

$$P\left(\bigcap_{k=1}^{n}\{\omega : X_k(\omega) \in B_k\}\right) = \prod_{k=1}^{n} P\{\omega : X_k(\omega) \in B_k\}.$$

Poznámka. Volíme-li konkrétní borelovské množiny $B_k = (-\infty, x_k)$, pak $X_1, \ldots, X_n$ jsou nezávislé, právě tehdy, pokud sdružená distribuční funkce F je rovna součinu marginálních distribučních funkcí $F_i,\ i = 1, \ldots, n$.

$$F(x_1, \ldots, x_n) = P(X_1 < x_1, \ldots, X_n < x_n) =$$
$$= P(X_1 < x_1) \cdots P(X_n < x_n) = F_1(x_1) \cdots F_n(x_n).$$

Definice 2.5. Uspořádaná n-tice nezávislých, stejně rozdělených náhodných veličin $X_1, \dots, X_n$ se nazývá *náhodný výběr* o rozsahu n. Platí-li $X_1 \leq X_2 \leq \dots \leq X_n$ nazveme tento náhodný výběr *uspořádaný* a značíme jej $X_{(1)}, \dots, X_{(n)}$.

2.1 Teorie odhadu

Předpokládejme, že náhodný vektor $\boldsymbol{X} = (X_1, \dots, X_n)'$ má hustotu $f(\boldsymbol{x}, \boldsymbol{\theta})$ vzhledem k σ-konečné míře μ. Parametr $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$ je neznámý. Cílem je získat na základě vektoru $\boldsymbol{X}$ co nejlepší odhad vektoru $\boldsymbol{\theta}$.

Hledáme-li měřitelné zobrazení $g : (\mathbb{R}_n, \mathcal{B}_n) \to (\mathbb{R}_m, \mathcal{B}_m)$ takové, že náhodný vektor $\boldsymbol{T} = g(\boldsymbol{X})$ co možná nejlépe aproximuje hodnotu $\boldsymbol{\theta}$, pak se jedná o *bodový odhad* parametru $\boldsymbol{\theta}$. Jestliže hledáme interval nebo jinou vhodnou množinu do které s dostatečně velkou pravděpodobností $\boldsymbol{\theta}$ náleží, dostáváme *intervalový odhad.*

Definice 2.6. Řekneme, že odhad $\boldsymbol{T}$ parametru $\boldsymbol{\theta}$ je

1. *nestranný*, platí-li $E\boldsymbol{T} = \boldsymbol{\theta}$ pro každé $\boldsymbol{\theta} \in \Theta$.

2. *vychýlený*, jestliže $E\boldsymbol{T} = \boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\theta})$ a funkce $\boldsymbol{b}$ není identicky rovna nule, $\boldsymbol{b}(\boldsymbol{\theta})$ se nazývá *vychýlení* odhadu $\boldsymbol{T}$.

3. *nejlepší nestranný*, je-li rozptyl nestranného odhadu $\boldsymbol{T}$ nejmenší z rozptylů všech nestranných odhadů téhož parametru $\boldsymbol{\theta}$.

Definice 2.7. Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení Q, závislého na jednorozměrném parametru θ. Řekneme, že odhad $T_n = g_n(X_1, \dots, X_n)$ je *konsistentní*, jestliže $T_n \to \theta$ podle pravděpodobnosti při $n \to \infty$.

Věta 2.8. *Nechť střední hodnota $ET_n^2 < \infty$ pro každé přirozené n. Jestliže střední hodnota $ET_n \to \theta$ a rozptyl $varT_n \to 0$, pak T_n je konsistentní odhad parametru θ.*
Důkaz:

Pro každé $\varepsilon > 0$ platí

$$\begin{gathered} P(|T_n - \theta| > \varepsilon) = P(|T_n - ET_n + ET_n - \theta| > \varepsilon) \leq \\ \leq P\left(|T_n - ET_n| > \tfrac{\varepsilon}{2} \vee |ET_n - \theta| > \tfrac{\varepsilon}{2}\right) \leq \\ \leq P\left(|T_n - ET_n| > \tfrac{\varepsilon}{2}\right) + P\left(|ET_n - \theta| > \tfrac{\varepsilon}{2}\right). \end{gathered}$$

Jestliže $ET_n \to \theta$, pak pro $n \to \infty$: $P\left(|ET_n - \theta| > \frac{\varepsilon}{2}\right) \to 0$.
Podle Čebysevovy nerovnosti

$$P\left(|T_n - ET_n| > \frac{\varepsilon}{2}\right) \leq \frac{varT_n}{\left(\frac{\varepsilon}{2}\right)^2}.$$

Za předpokladu $varT_n \to 0$ pro $n \to \infty$ i tato pravděpodobnost konverguje k nule. Dokázali jsme tedy, že

$$P(|T_n - \theta| > \varepsilon) \to 0.$$

□

2.2 Fisherova míra informace

Nyní zavedeme *Fisherovu míru informace* o parametru θ obsaženou v náhodném vektoru $\boldsymbol{X}$.

Definice 2.9. Nechť náhodný vektor $\boldsymbol{X} = (X_1, \dots, X_n)'$ má hustotu $f(\boldsymbol{x}, \theta)$ vzhledem k nějaké σ-konečné míře μ. Předpokládejme, že platí:

- Ω je neprázdná, otevřená množina.
- Množina $M = \{\boldsymbol{x} : f(\boldsymbol{x}, \theta) > 0\}$ nezávisí na θ.
- Pro skoro všechna $\boldsymbol{x} \in M$ (vzhledem k μ) existuje konečná parciální derivace
$$f_i'(\boldsymbol{x}, \theta) = \frac{\partial f(\boldsymbol{x}, \theta)}{\partial \theta}.$$
- Pro všechna $\theta \in \Omega$ platí
$$\int_M f'(\boldsymbol{x}, \theta) d\mu(x) = 0$$
.
- Integrál
$$J_n(\theta) = \int_M \left[\frac{f'(\boldsymbol{x}, \theta)}{f(\boldsymbol{x}, \theta)}\right] f(\boldsymbol{x}, \theta) d\mu(x)$$
je konečný a kladný.

Pak se systém hustot $\{f(\boldsymbol{x}, \theta), \theta \in \Omega\}$ nazývá *regulární* a $\boldsymbol{J}_n(\theta)$ se nazývá *Fisherova míra informace.*

2.3 Princip maximální věrohodnosti

Nechť náhodný vektor $\boldsymbol{X} = (X_1, \dots, X_n)'$ má hustotu $f(\boldsymbol{x}, \theta)$, kde $\theta \in \omega$. Při pevné hodnotě $\boldsymbol{x}$ se funkce $f(\boldsymbol{x}, \theta)$ nazývá *věrohodnostní funkce.* Budeme uvažovat případ, kdy $X_1, \dots, X_n$ je náhodný výběr.

Hodnotu $\hat{\theta}$ parametru θ, maximalizující věrohodnostní funkci $f(\boldsymbol{x}, \theta)$, pak nazveme *maximálně věrohodný odhad* parametru θ.

Mějme funkci $L(\boldsymbol{x}, \theta) = \ln f(\boldsymbol{x}, \theta)$. Tato funkce se pro pevná $\boldsymbol{x}$ nazývá *logaritmická věrohodnostní funkce.*

Kasický postup hledání maxima L pomocí její derivace vede k nalezení maximálně věrohodného odhadu, který konverguje ke skutečné hodnotě parametru θ. A to s pravděpodobností blížící se k jedné. Důkaz viz. [1].

Nechť θ_0 je skutečná hodnota parametru θ. Hledáme tedy kořen $\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{X})$ *věrohodnostní rovnice*
$$\frac{\partial L(\boldsymbol{X}, \theta)}{\partial \theta} = 0$$
takový, že $|\hat{\theta}_n - \theta_0| < \epsilon$, pro každé $\epsilon > 0$.

Věta 2.10. Nechť $f(x, \theta), \theta \in \Theta$ je regulární systém hustot s Fisherovou mírou informace $J(\theta)$. Nechť jsou splněny následující předpoklady:

(p1) Θ je parametrický prostor, který obsahuje takový neprázdný otevřený interval ω, že $\theta_0 \in \omega$

(p2) $\boldsymbol{X} = (X_1, \ldots, X_n)'$, kde X_i jsou nezávislé stejně rozdělené náhodné veličiny s hustotou $f(x, \theta)$ vzhledem k nějaké σ-konečné míře μ.

(p3) $M = \{x : f(x, \theta) > 0\}$ nezávisí na θ.

(p4) Nechť $\theta_1, \theta_2 \in \Omega$. Pak $f(x, \theta_1) = f(x, \theta_2)$ skoro všude právě tehdy, když $\theta_1 = \theta_2$.

(p5) Pro všechna $\theta \in \omega$ a skoro všechna $x \in M$ existuje derivace

$$f'''(x, \theta) = \frac{\partial^3 f(x, \theta)}{\partial \theta^3}$$

.

(p6) Pro všechna $\theta \in \omega$ platí:

$$\int_M f''(x, \theta) d\mu(x) = 0.$$

(p7) Existuje taková nezáporná měřitelná funkce $H(x)$, že střední hodnota $E_{\theta_0} H(X) < \infty$ a přitom pro skoro všechna $x \in M$ a pro všechna θ taková, že $|\theta - \theta_0| < \epsilon$ pro dostatečně malé $\epsilon > 0$ platí:

$$\left| \frac{\partial^3 \ln f(x, \theta)}{\partial \theta^3} \right| \leq H(x).$$

Pak platí následující tvrzení:

(i) Jestliže $n \to \infty$, pak

$$\frac{1}{\sqrt{n}} L'(\theta_0) \rightharpoonup N[0, J(\theta)]$$

(ii) Existuje-li dostatečně velké n a pro každou hodnotu $\boldsymbol{X}$ takový kořen $\hat{\theta}_n$ věrohodnostní rovnice, že $\hat{\theta}_n$ je konsistentním odhadem parametru θ_0, pak

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightharpoonup N\left[0, \frac{1}{J(\theta)}\right].$$

Na základě poznatků o maximálně věrohodných odhadech lze pak provádět asymptotické testy statistických hypotéz.

Věta 2.11. *Nechť* jsou splněny všechny předpoklady věty 2.10. Označme

$$LM(\theta_0) = \frac{[L'(\theta_0)]^2}{nJ(\theta_0)},$$

$$U_{LM} = \frac{L'(\theta_0)}{\sqrt{nJ(\theta_0)}}.$$

Pak U_{LM} má asymptoticky rozdělení N(0,1) a LM má asymptoticky χ^2 rozdělení pravděpodobnosti s jedním stupněm volnosti.

3 Teoretická konstrukce ROC křivky

V tomho oddílu budou uvedeny základní poznatky z teorie využití ROC křivek při klasifikaci objektů, vlastnosti ROC křivek a příklady jejich konstrukce pro normální a některá další rozdělení pravděpodobnosti.

3.1 Senzitivita a specificita

Uvažujme diagnostický test, který má určit zda sledovaný objekt má určitou vlastnost D. Předpokládejme, že každý objekt náleží právě do jedné ze dvou skupin π_1, π_0. Jestliže má danou vlastnost ($D = 1$), pak náleží do skupiny π_1. Naopak pokud tuto vlastnost nemá ($D = 0$), náleží do skupiny π_0. D je náhodná veličina s alternativním rozdělením pravděpodobnosti.

Označme náhodnou veličinu T, výsledek testu. Klasifikaci objektu (odhad příslušnosti k dané skupině) provedeme na základě srovnání výsledku testu T s danou mezí c. Je-li $T \geq c$ klasifikujeme objekt jako prvek skupiny π_1 a řekneme, že klasifikace je pozitivní. Je-li $T < c$ klasifikujeme objekt jako prvek skupiny π_0 a řekneme, že klasifikace je negativní. Označení mezní hodnoty c vychází z anglického výrazu(*cut-off point*).

Definice 3.12. Pro danou hodnotu klasifikační meze c definujeme *senzitivitu* testu jako podmíněnou pravděpodobnost, že výsledek testu T objektu ze skupiny π_1 je větší nebo roven c.

$$Se(c) = P(T \geq c|D = 1) = 1 - P(T < c|D = 1). \tag{3.1}$$

Dostáváme tedy pravděpodobnost správně určené pozitivní klasifikace.

Mějme nyní náhodnou veličinu T_1, výsledek testu ve skupině π_1, s distribuční funkcí F_1 a hustotou f_1. Pak získáváme ekvivalentní vyjádření senzitivity ve tvaru:

$$Se(c) = 1 - P(T < c|D = 1) = 1 - P(T_1 < c) = 1 - F_1(c). \tag{3.2}$$

Definice 3.13. Pro danou hodnotu klasifikační meze c definujeme *specificitu* testu jako podmíněnou pravděpodobnost, že výsledek testu T objektu ze skupiny π_0 je menší než c.

$$Sp(c) = P(T < c|D = 0). \tag{3.3}$$

Tedy pravděpodobnost správně určené negativní klasifikace.

Nyní označme náhodnou veličinu T_0, výsledek testu ve skupině π_0. Pak získáváme ekvivalentní vyjádření specificity pomocí distribuční funkce F_0 náhodné veličiny T_0:

$$Sp(c) = P(T < c|D = 0) = P(T_0 < c) = F_0(c). \tag{3.4}$$

Senzitivita a specificita jsou základními vlastnostmi testu. Jejich doplňky pak udávají pravděpodobnosti chybné klasifikace.

Zvolíme-li chybně negativní klasifikaci, dopouštíme se chyby prvního druhu a to s pravděpodobností $1 - Se$.

Zvolíme-li chybně pozitivní klasifikaci, dopouštíme se chyby druhého druhu a to s pravděpodobností $1 - Sp$.

3.2 ROC křivka

V následujícím textu se dostáváme k zavedení pojmu ROC křivky. Provedeme teoretickou konstrukci křivky jako funkce distribučních funkcí F_1 a F_0. Na příkladech pak ukážeme vlastnosti ROC křivek a geometrický význam senzitivity a specificity.

Definice 3.14. *ROC křivku* definujeme jako množinu bodů daných souřadnicemi

$$[1 - Sp(c), Se(c)], \quad c \in (-\infty, \infty). \tag{3.5}$$

Jestliže podle vztahů (2.2) a (2.4)

$$\begin{aligned} F_1(c) &= 1 - Se(c), \\ F_0(c) &= Sp(c), \end{aligned} \tag{3.6}$$

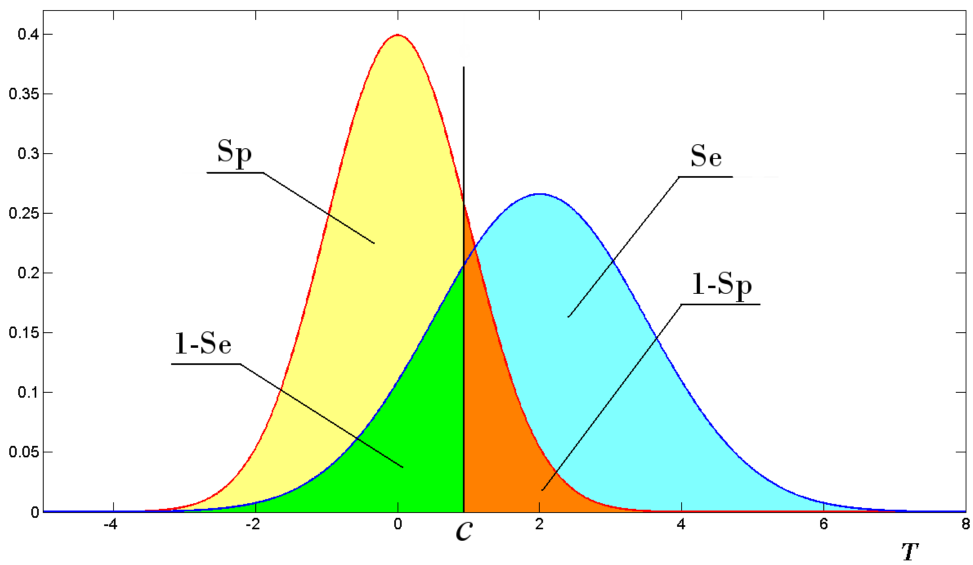
pak za předpokladu existence inverzní distribuční funkce F_0^{-1} lze ROC křivku vyjádřit vzávislosti na parametru t. Položíme

$$t = 1 - Sp(c) = 1 - F_0(c), \quad \text{pak} \quad c = F_0^{-1}(1-t), \quad Se(c) = 1 - F_1(c).$$

Dostáváme tedy ekvivalentní vztah:

$$ROC(t) = 1 - F_1(F_0^{-1}(1-t)), \quad \text{pro} \quad t \in (0,1). \tag{3.7}$$

Příklad 3.15. Mějme případ, kdy náhodná veličina T_0 má normální rozdělení pravděpodobnosti s nulovou střední hodnotou a rozptylem rovným jedné - $N(0,1)$ a T_1 má rozdělení $N(2, 1.5)$. Na obrázku 1 je pro hustoty f_0 a f_1 znázorněna senzitivita a specificita testu při klasifikaci s mezní hodnotou c.



Obrázek 1: Senzitivita a specificita pro klasifikační mez c.

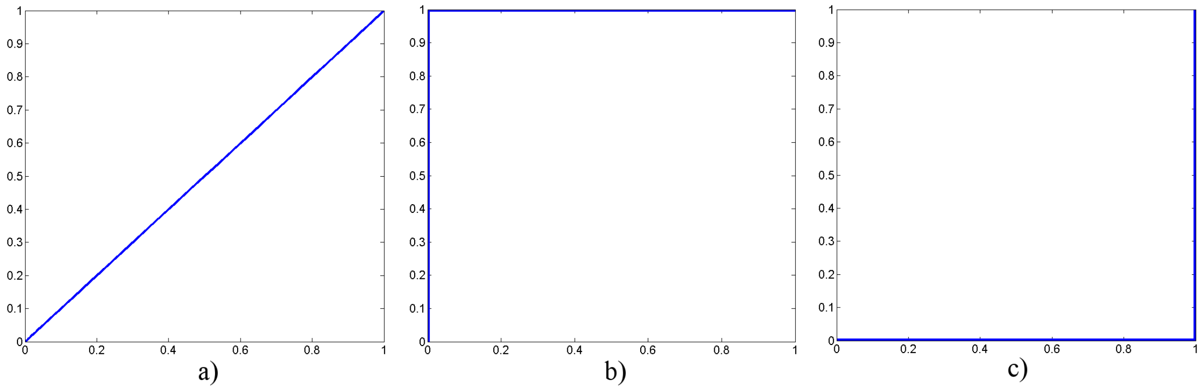
3.3 Vlastnosti a parametry ROC křivek

Z tvaru získané ROC křivky lze odhadnout rozlišovací schopnost zkoumaného testu. Jestliže křivka nejprve prudce roste a poté je téměř konstantní, poměr chybně klasifikovaných objektů bude malý. Bude-li se křivka blížit diagonále poměr chyb poroste.

Pokud se hustoty f_0 a f_1 shodují, křivka je totožná s diagonálou. Pravděpodobnosti správné a chybné klasifikace se rovnají - obrázek 2a.

V ideálním případě, kdy test je schopen správně rozřadit všechny objekty, křivka prochází bodem $[1,1]$ - obrázek 2b.

Pokud nastane případ, kdy žádný objekt nebyl zařazen správně - obrázek 2c, lze jej převrácením testovacího kriteria převést opět na ideální stav.



Obrázek 2: Extrémní případy ROC křivek.

Věta 3.16. *ROC křivka je invariantní vzhledem k monotónní rostoucí transformaci* T_0,T_1.

Důkaz:

Nechť h je monotónní rostoucí transformace taková, že

$$h : x_0 \to u, \text{ tj. } u = h(x_0),$$

$$h : x_1 \to v, \text{ tj. } v = h(x_1).$$

Označme

$$F_{0h}(x_0) = F_0(h^{-1}(x_0)) = P(T_0 \leq h^{-1}(x_0)),$$
$$F_{1h}(x_1) = F_1(h^{-1}(x_1)) = P(T_1 \leq h^{-1}(x_1)),$$

$$ROC_h(t) = 1 - 1 - F_{1h}(F_{0h}^{-1}(1-t)).$$

Dále platí

$$F_{0h}^{-1}(x_0) = h(F_0^{-1}(x_0)),$$
$$F_{1h}^{-1}(x_1) = h(F_1^{-1}(x_1)).$$

Je třeba dokázat

$$ROC(t) = ROC_h(t).$$

Pak po dosazení dostáváme

$$ROC_h(t) = 1 - F_{1h}(F_{0h}^{-1}(1-t)) = 1 - F_1(h^{-1}(F_{0h}^{-1}(1-t))) =$$

$$= 1 - F_1(h^{-1}(h(F_0^{-1}(1-t)))) = 1 - F_1(F_0^{-1}(1-t)) = ROC(t).$$

Tímto je daná vlastnost dokázána.

□

Věta 3.17. Je-li náhodná veličina T_1 stochasticky větší než náhodná veličina T_0, tedy když $F_0(c) \geq F_1(c)$ pro všechna $c \in (-\infty, \infty)$, pak ROC křivka leží nad diagonálou v jednotkovém čtverci.
Důkaz:
Je-li $F_0(x_0) \geq F_1(x_1)$, pak za předpokladu existence inverzních funkcí $F_0^{-1}(x_0) \leq F_1^{-1}(x_1)$. Potom pro $t \in (0, 1)$

$$ROC(t) = 1 - F_1(F_0^{-1}(1-t)) \geq 1 - F_1(F_1^{-1}(1-t)) = 1 - (1-t) = t.$$

Tedy $ROC(t) \geq t$ a křivka leží nad diagonálou v jednotkovém čtverci.

□

Věta 3.18. Když hustoty f_0, f_1 mají monotónní věrohodnostní poměr (tj. když existuje statistika S taková, že podíl hustot f_0/f_1 je neklesající funkcí statistiky S), pak ROC křivka je konkávní.
Důkaz:
V tomto bodě je třeba dokázat, že za daných předpokladů je ROC křivka konkávní. Tedy derivace ROC křivky je klesající funkcí. Předpokládejme existenci inverzních distribučních funkcí a potřebných derivací.
Vypočteme tedy

$$\frac{\partial ROC(t)}{\partial t} = \frac{\partial(1 - F_1(F_0^{-1}(1-t)))}{\partial t} = -\frac{\partial(F_1(F_0^{-1}(1-t)))}{\partial F_0^{-1}(1-t)} \frac{\partial F_0^{-1}(1-t)}{\partial t}.$$

Člen

$$\frac{\partial(F_1(F_0^{-1}(1-t)))}{\partial F_0^{-1}(1-t)} = f_1(F_0^{-1}(1-t)).$$

Označme $u = (1-t)$, pak dostáváme

$$\frac{\partial u}{\partial t} = -1 \text{ tedy } \partial u = -\partial t.$$

Pak platí

$$\frac{\partial F_0^{-1}(1-t)}{\partial t} = -\frac{\partial F_0^{-1}(u)}{\partial u}.$$

Nechť $w = F_0^{-1}(u) \Rightarrow u = F_0(w)$, pak

$$-\frac{\partial F_0^{-1}(u)}{\partial u} = -\frac{\partial w}{\partial F_0(w)} = -\frac{1}{\frac{\partial F_0(w)}{\partial w}} = -\frac{1}{f_0(w)} \Rightarrow$$

$$\Rightarrow -\frac{\partial F_0^{-1}(1-t)}{\partial t} = \frac{1}{f_0(F_0^{-1}(1-t))}.$$

Dosazením získáváme vztah

$$\frac{\partial ROC(t)}{\partial t} = \frac{f_1(F_0^{-1}(1-t))}{f_0(F_0^{-1}(1-t))}.$$

Pro $0 < t_1 < t_2 < 1$ platí $1 - t_1 > 1 - t_2$, protože F_0^{-1} je rostoucí funkce

$$F_0^{-1}(1-t_1) > F_0^{-1}(1-t_2).$$

Mají-li f_0 a F_1 neklesající věrohodnostní poměr, tedy pro $x_1 < x_2$

$$\frac{f_1(x_1)}{f_0(x_1)} \leq \frac{f_1(x_2)}{f_0(x_2)},$$

dostáváme

$$\frac{f_1(F_0^{-1}(1-t_1))}{f_0(F_0^{-1}(1-t_1))} \geq \frac{f_1(F_0^{-1}(1-t_2))}{f_0(F_0^{-1}(1-t_2))}.$$

Tím je tvrzení dokázáno.

□

Věta 3.19. Plocha pod křivkou je rovna pravděpodobnosti $P(T_0 < T_1)$. Tedy

$$AUC = \int_0^1 ROC(t)\mathrm{d}t = P(T_0 < T_1). \tag{3.8}$$

Důkaz:

$$\int_0^1 ROC(t)\mathrm{d}t = \int_0^1 [1 - F_1(F_0^{-1}(1-t))]\mathrm{d}t = \begin{bmatrix} \textit{substituce} \\ F_0^{-1}(1-t) = x_0 \\ 1 - t = F_0(x_0) \\ t = 1 - F_0(x_0) \\ \mathrm{d}t = -f_0(x_0)\mathrm{d}x_0 \end{bmatrix} =$$

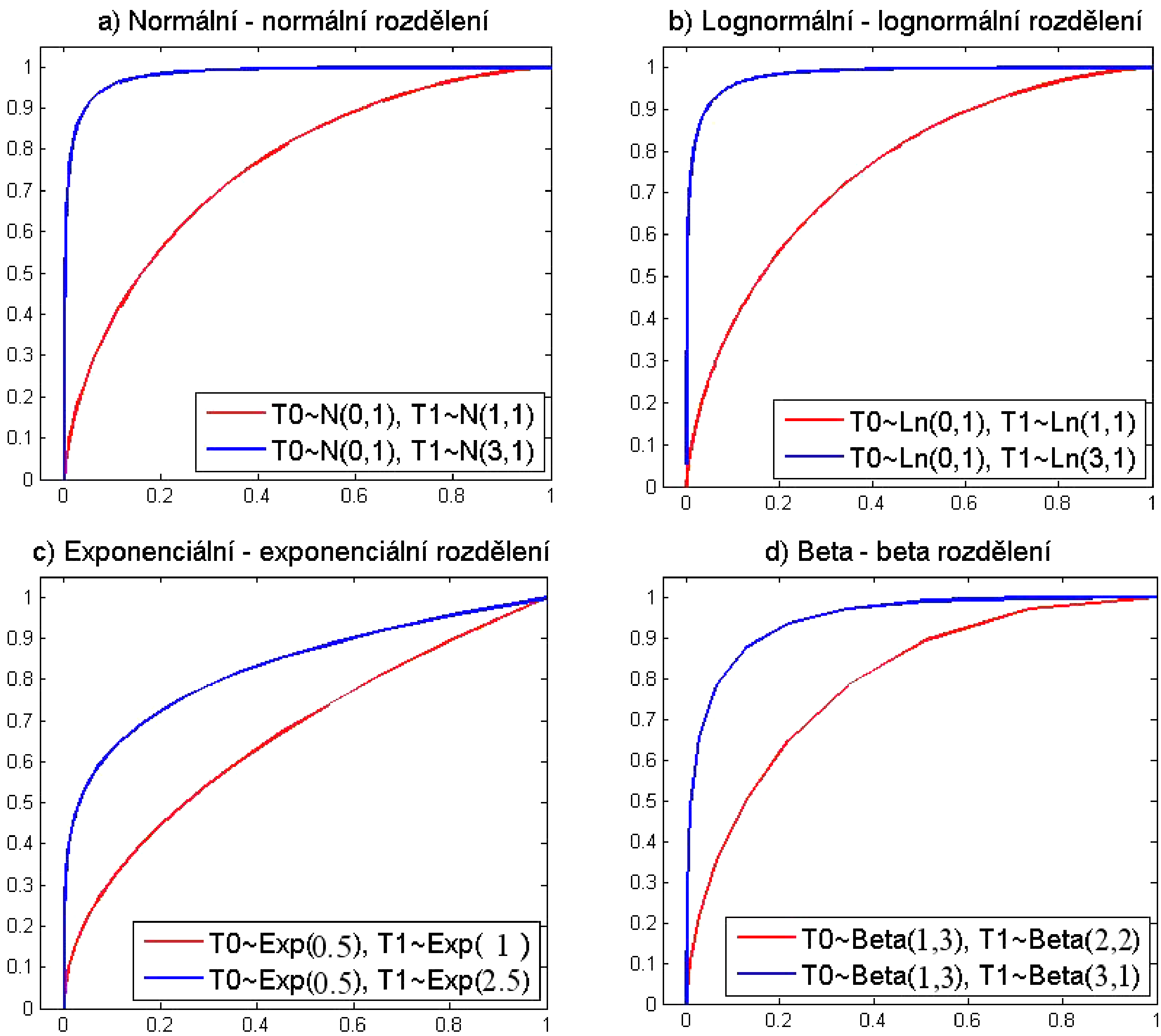
$$= \int_{-\infty}^{\infty} [1 - F_1(x_0)] f_0(x_0)\mathrm{d}x_0 = \int_{-\infty}^{\infty} \int_{x_0}^{\infty} f_1(x_1) f_0(x_0)\mathrm{d}x_0\mathrm{d}x_1 = P(T_0 < T_1).$$

□

Velikost plochy pod ROC křivkou (zkratka AUC z anglického area under curve) je jedním základních měřítek kvality diagnostického testu. Protože křivka leží v jednotkovém čtverci, může AUC obecně nabývat hodnot od 0 do 1. Splňuje-li křivka předpoklad věty 3.17, leží nad diagonálou a nabývá tedy hodnot od $0,5$ do 1.

Detailnímu popisu plochy pod ROC křivkou bude věnována kapitola 6. Tento parametr je vhodné využít také ke srovnání několika testů, více v kapitole 8.

Příklad 3.20. Na obrázku 3 jsou vykresleny příklady ROC křivek, kdy náhodné veličiny T_0 a T_1 mají a) normální, b) logaritmické normální, c) exponenciální, d) beta rozdělení pravděpodobnosti. (V popisu parametrů $T_0 \approx$T0 a $T_1 \approx$T1.)



Obrázek 3: Příklady ROC křivek.

Příklad 3.21. Mějme příklad diagnostického testu v medicíně. Bylo testováno 200 osob, přičemž 120 z nich bylo nemocných ($D = 1$) a 80 zdravých ($D = 0$). Výsledky byly zaznamenány do následující tabulky.

		Zdravotní stav	
		$D = 1$	$D = 0$
Výsledek	Nemocný	92	9
testu	Zdravý	28	71

Tabulka 1: Tabulka četností

Tedy u 92 osob ze 120 test nemoc správně diagnostikoval, 9 označil za nemocné, přestože byli zdraví, u 71 zdravých pacientů nemoc vyloučil a u 28 nemoc neodhalil, ikdyž pacient nemocný byl.

Nyní do téže tabulky zaznamenáme relativní četnosti.

		Zdravotní stav	
		$D = 1$	$D = 0$
Výsledek	Nemocný	$0,767$	$0,113$
testu	Zdravý	$0,233$	$0,887$

Tabulka 2: Tabulka relativních četností

Tento test tedy správně rozeznal nemoc u 76,7% nemocných pacientů a vyloučil u 88,7% zdravých. Získali jsme tedy odhad senzitivity a specificity použitého testu ve tvaru pozorovaných relativních četností.

4 Bodové odhady ROC křivky

V následující části budou popsány statistické metody pro stanovení hodnoty ROC křivky v daném bodě. Při neparametrickém přístupu půjde o konstrukci empirické ROC křivky, po částech lineární křivky a metodu založenou na jádrových odhadech distribučních funkcí. Dále pak zavedeme binormální model a provedeme odhad jeho parametrů.

4.1 Empirická ROC křivka

Jako první neparametrický bodový odhad ROC křivky uvedeme metodu založenou na nestranném odhadu distribuční funkce výběrovou distribuční funkcí.

Definice 4.22. Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení o dostribuční funkci $F(x)$. Nechť I_A je indikátor jevu A, tj. $I_A = 1$ jestliže jev A nastane, jinak $I_A = 0$. Pak definujeme výběrovou distribuční funkci vztahem:

$$\widehat{F}_e(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \leq x]}. \tag{4.9}$$

Označme $T_{01}, \ldots, T_{0n}$ náhodný výběr z rozdělení o distribuční funkci F_0 a $T_{11}, \ldots, T_{1m}$ náhodný výběr z rozdělení o dostribuční funkci F_1, $\widehat{F}_{e0}$, $\widehat{F}_{e1}$ odhady distribučních funcí F_0, F_1 dané vztahem 4.9. Pak parametrické zobrazení $[1-\widehat{F}_{e0}, 1-\widehat{F}_{e1}]$ nazýváme *empirická ROC křivka.*

4.2 Po částech lineární ROC křivka

Další možností založenou na neparametrickém odhadu F_0 a F_1 je konstrukce *po částech lineární ROC křivky.*

Definice 4.23. $X_{(1)}, \ldots, X_{(m)}$ je uspořádaný náhodný výběr z rozdělení o distribuční funkci F(x).Nechť středy intervalů $(X_{(i)}, X_{(i+1)})$ jsou

$$c_i = \frac{X_{(i+1)} + X_{(i)}}{2} \text{ pro } i = 1, \ldots, m-1,$$

$$c_0 = \frac{3X_{(1)} - X_{(2)}}{2},\ c_m = \frac{3X_{(m)} - X_{(m-1)}}{2}.$$

Dále nechť

$$f_l(x) = (m(c_{i+1} - c_i))^{-1}, \quad x \in \langle c_i, c_{i+1}), \quad i = 1, \ldots, m-1,$$

jinak $f(x) = 0$. Pak po částech lineární odhad distribuční funkce $F(x)$ je definován vztahem

$$\widehat{F}_l(x) = \int_{-\infty}^{x} f_l(t)dt. \tag{4.10}$$

Označme nyní $T_{0(1)}, \ldots, T_{0(n)}$ uspořádaný náhodný výběr z rozdělení o distribuční funkci F_0 a $T_{1(1)}, \ldots, T_{1(m)}$ uspořádaný náhodný výběr z rozdělení o dostribuční funkci F_1, $\widehat{F}_{l0}$, $\widehat{F}_{l1}$ odhady distribučních funcí F_0, F_1 dané vztahem 4.10. Pak parametrické zobrazení $[1-\widehat{F}_{l0}, 1-\widehat{F}_{l1}]$ nazýváme *po částech lineární ROC křivka.*

4.3 Jádrový odhad senzitivity a specificity

Výhodou této metody proti předchozím je, že získáme aproximaci ROC křivky hladkou křivkou. Pro vyjádření využijeme jádrový odhad distribuční funkce autorů Zhou, Hall, Shapiro, popsáný v [11].

Definice 4.24. Nechť funkce $k : \mathbb{R} \to \mathbb{R}$ splňuje následující podmínky:

1. nosič $\operatorname{supp}(k) = \langle -1, 1\rangle$, tedy $k(x) = 0,\ \forall x \notin \langle -1, 1\rangle$,
2. k je lipschitzovsky spojitá na $\langle -1, 1\rangle$,
   tj. $|k(x) - k(y)| \leq L|x - y|,\ L > 0, \forall x, y \in \langle -1, 1\rangle$,
3. integrál
$$\int_{-\infty}^{\infty} k(x)\mathrm{d}x = 1.$$

Pak k nazveme *jádrovou funkcí* zkráceně *jádrem.*

Významným faktorem ovlivňujícím chování jádrových odhadů je šířka *vyhlazovacího okénka* $h > 0$. Transformací
$$k_h = \frac{1}{h} k \left(\frac{x}{h}\right)$$
pak dojde ke změně nosiče jádra na interval $\langle -h, h\rangle$.

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozdělení o hustotě $f(x)$. Jádrový odhad této hustoty je pak dán vztahem:

$$\widehat{f}_k = \frac{1}{nh} \sum_{i=1}^{n} k \left(\frac{x - X_i}{h}\right)$$

s užítm jádra s dvojnásobnou váhou

$$k \left(\frac{x - X_i}{h}\right) = \frac{15}{16} \left[1 - \left(\frac{x - X_i}{h}\right)^2\right]^2, \quad x \in (X_i - h, X_i + h),$$

kde $k = 0$ jinde a šířkou vyhlazovacího okna

$$h = 0,9 \min(SD, IQR/1,34)/\sqrt[5]{n},$$

kde SD je směrodatná odchylka a IQR rozdíl 0,75-kvantilu a 0,25-kvantilu.

Odhad distribuční funkce je pak definován jako:

$$\widehat{F}_k(t) = \sum_{i=1}^{n} \int_{-\infty}^{t} \frac{1}{nh} k \left(\frac{x - X_i}{h}\right) \mathrm{d}x.$$

Při numerickém výpočtu nabývá integrál ve vztahu pro výpočet $\widehat{F}_k(t)$ následujících hodnot:

je-li $t > X_i + h$, pak je integál roven nule,
je-li $c < X_i - h$, pak je integál roven $1/n$,
je-li $X_i - h < t < X_i + h$, pak je integál roven $\frac{1}{16n}(8 - 15v + 10v^3 - 3v^5)$,
kde $v = (t - X_i)/h$.

Výsledná podoba křivky je pak dána jako $[1 - \widehat{F}_{k0}, 1 - \widehat{F}_{k1}]$, kde $\widehat{F}_{k0}, \widehat{F}_{k1}$ jsou jádrové odhady distribuních funkcí náhodných veličin T_0, T_1.

4.4 Binormální model

V této části se budeme zabývat situací, kdy obě distribuční funkce mají normální rozdělení, takzvaným binormálním modelem. Cílem bude nalézt odhad jeho parametrů.

Nechť $F_0(x)$ a $F_1(x)$ definované vztahy 3.6 jsou distribuční funkce normálního rozdělení pravděpodobnosti.

$$F_0(x) \sim N(\mu_0, \sigma_0^2), \quad F_1(x) \sim N(\mu_1, \sigma_1^2)$$

Můžeme tedy položit

$$ROC(t) = 1 - F_1(F_0^{-1}(1-t)) = 1 - \Phi\left(\frac{F_0^{-1}(1-t) - \mu_1}{\sigma_1}\right) =$$

$$1 - \Phi\left(\frac{\mu_0 + \sigma_0 \Phi^{-1}(1-t) - \mu_1}{\sigma_1}\right) = 1 - \Phi\left(\frac{\sigma_0}{\sigma_1}\Phi^{-1}(1-t) - \frac{\mu_1 - \mu_0}{\sigma_1}\right).$$

kde Φ je distribuční funkce standadizovaného normálního rozdělení.
Označme $a = \frac{\mu_1 - \mu_0}{\sigma_1}$ a $b = \frac{\sigma_0}{\sigma_1}$. Dostáváme

$$ROC(t) = 1 - \Phi(b\,\Phi^{-1}(1-t) - a), \quad t \in \langle 0, 1 \rangle\,. \tag{4.11}$$

Nyní je potřeba odhad neznámých parametrů a a b. Pokud jsou původní data skutečně binormální je možné použít pro tento účel výběrový průměr a výběrový rozptyl

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\widehat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Odhad parametrů je tedy dán vztahy:

$$\widehat{a} = \frac{\overline{X}_1 - \overline{X}_0}{S_1}, \quad \widehat{b} = \frac{S_0}{S_1}.$$

Před použitím tohoto odhadu je potřeba nejprve otestovat, zda jde o normální rozdělení pravděpodobnosti. Pokud tomu tak není provedeme vhodnou transformaci dat (za předpokladu že taková transformace existuje).

Jedna z možných metod využívá *Box-Cox transformaci* danou

$$t(y) = \begin{cases} \frac{(y^\lambda - 1)}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}.$$

Odhad parametru λ lze nalézt metodou maximální věrohodnosti nebo například připoužitím software MATLAB.

4.5 Nejlepší nestranný odhad senzitivity a specificity binormálního modelu

Tato metoda využívá Kolmogorovův nejlepší nestranný odhad distribuční funkce vyjádřený vztahy:

$$\widehat{F}_K(x) = \begin{cases} 0 & \text{pro} \quad Q(x) \le -1 \\ \frac{1}{2} - \frac{1}{2}\beta_{Q^2(x)}\left(\frac{1}{2}, \frac{m}{2} - 1\right) & \text{pro} \quad -1 < Q(x) \le 0 \\ \frac{1}{2} + \frac{1}{2}\beta_{Q^2(x)}\left(\frac{1}{2}, \frac{m}{2} - 1\right) & \text{pro} \quad 0 < Q(x) \le 1 \\ 1 & \text{pro} \quad Q(x) > 1 \end{cases},$$

kde

$$Q(x) = \frac{x - \overline{X}}{(m-1)S}\sqrt{m}$$

a funkce

$$\beta_a(p,q) = \frac{1}{\beta(p,q)} \int\limits_0^a t^{p-1}(1-t)^{q-1}dt$$

je normovaná neúplná beta funkce parametrů $a \in \langle 0, 1 \rangle$, $\quad p \ge 0, \quad q \ge 0.$
Důkaz že jde o nejlepší nestranný odhad je uveden v [6].

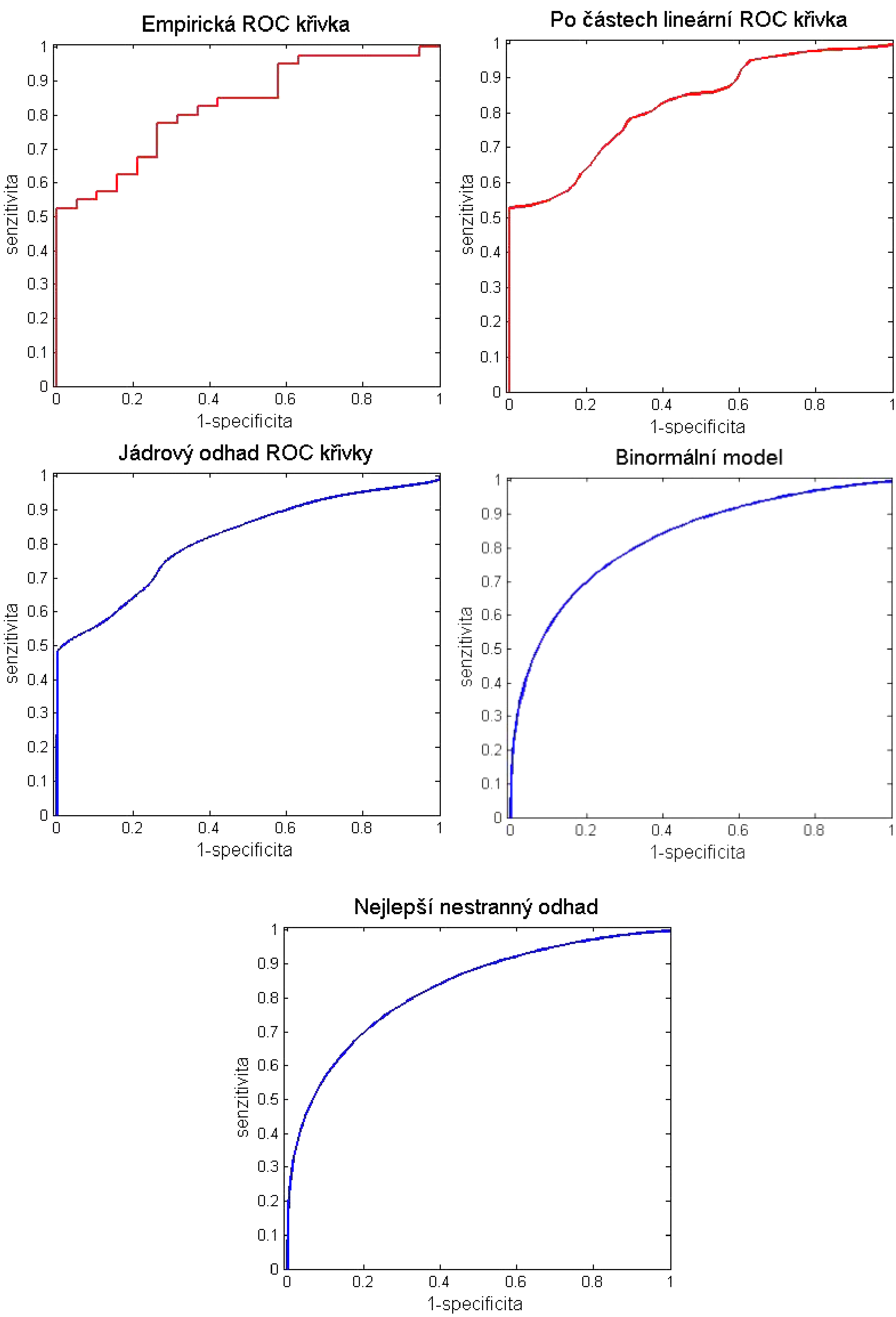
Odhah senzitivity pak opět získáme ve tvaru $(1 - \widehat{F}_{K1}(c))$ a specificitu testu odhadneme pomocí $\widehat{F}_{K0}(c)$.

Příklad 4.25. U pacientů s poraněním hlavy byla 24 hodin po úrazu měřena hodnota isoenzymu CK-BB (creatine kinase-BB). Z 59 pacientů se 19 plně zotavilo a u zbývajících 40 osob zanechal úraz trvalé následky. Na základě tohoto měření má být sestaven test, schopný predikovat podle hladiny CK-BB následný vývoj zotavení.

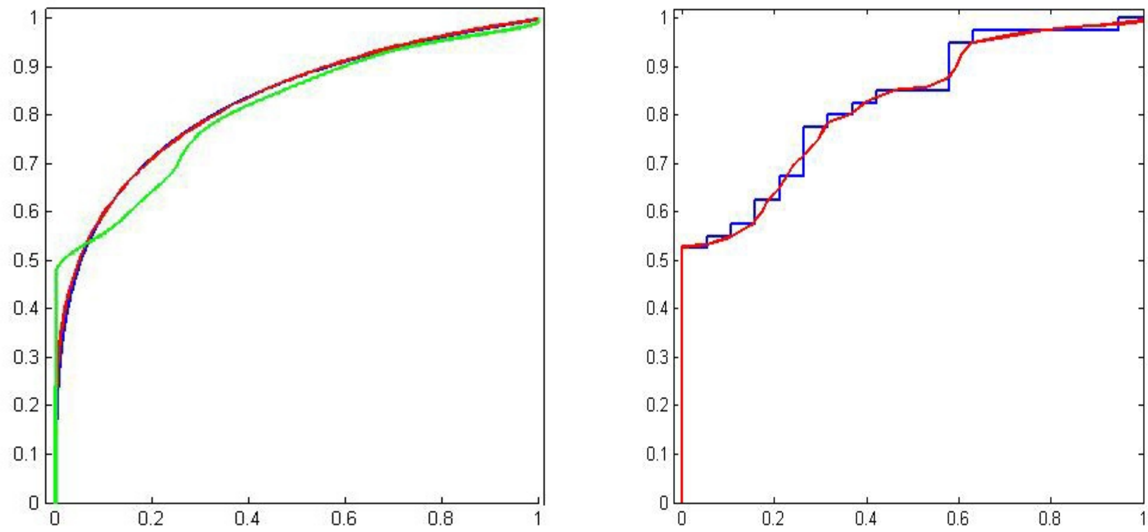
Označme T_0 hodnoty CK-BB u pacientů, kteří se plně zotavili a T_1 hladiny isoenzymu ve skupině pacientů s trvalými následky. Získané hodnoty jsou uvedeny v tabulce 3. Odhady ROC křivky jsou vykresleny na obrázku 4.

Hodnota CK-BB u pacientů					
Bez trvalých následků		S trvalými následky			
136	286	140	1087	230	183
281	23	1256	700	16	800
200	146	253	740	126	153
220	96	283	90	303	193
100	60	73	1370	543	913
17	27	230	463	60	509
126	100	576	671	80	490
253	70	156	356	323	1560
40	6	120	216	443	523
46		76	303	353	206

Tabulka 3: Hodnoty CK-BB



Obrázek 4: Odhady ROC křivky.



Obrázek 5: Srovnání jednotlivých metod odhadu ROC křivky.

Na levém grafu obrázku 5 pak vidíme porovnání jádrového odhadu (zelená), binormálního modelu (modře) a odhadu založeného na nejlepším nestranném odhadu Se a Sp (červeně). V pravé části je pak srovnání empirické (modrá) a po částech lineární ROC křivky (červená).

5 Intervalové odhady

V této části se budeme zabývat určením hranic, mezi kterými se ROC křivka s danou pravděpodobností nachází.

5.1 Pointwise confidence

Máme tedy bodový odhad binormálního mobelu ROC křivky

$$ROC(t) = 1 - \Phi(\widehat{b}\ \Phi^{-1}(1-t) - \widehat{a}),$$

dále lze určit $100(1-\alpha)\%$ interval spolehlivosti pro senzitivitu, který je dán vztahem

$$1 - \Phi\left[\widehat{b}\ \Phi^{-1}(1-t) - \widehat{a} \pm z_{1-\alpha/2}\sqrt{Var(\widehat{b}\ \Phi^{-1}(1-t) - \widehat{a})}\right], \tag{5.12}$$

zde $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$,

$$Var(\widehat{b}\ \Phi^{-1}(1-t) - \widehat{a}) = Var(\widehat{b})(\Phi^{-1}(1-t))^2 + Var(\widehat{a}) - 2(\Phi^{-1}(1-t))Cov(\widehat{a}, \widehat{b})).$$

Získáváme $100(1-\alpha)\%$ *asymptotický interval spolehlivosti.*

Rozptyly $\widehat{a}$ a $\widehat{b}$ a kovarianci $\widehat{a}, \widehat{b}$ odhadneme

$$\widehat{Var}(\widehat{a}) = \frac{n_1(\widehat{a}^2 + 2) + 2n_0\widehat{b}^2}{2n_0n_1},$$

$$\widehat{Var}(\widehat{b}) = \frac{(n_1 + n_0)\widehat{b}^2}{2n_0n_1},$$

$$\widehat{Cov}(\widehat{a}, \widehat{b}) = \frac{\widehat{ab}}{2n_0}.$$

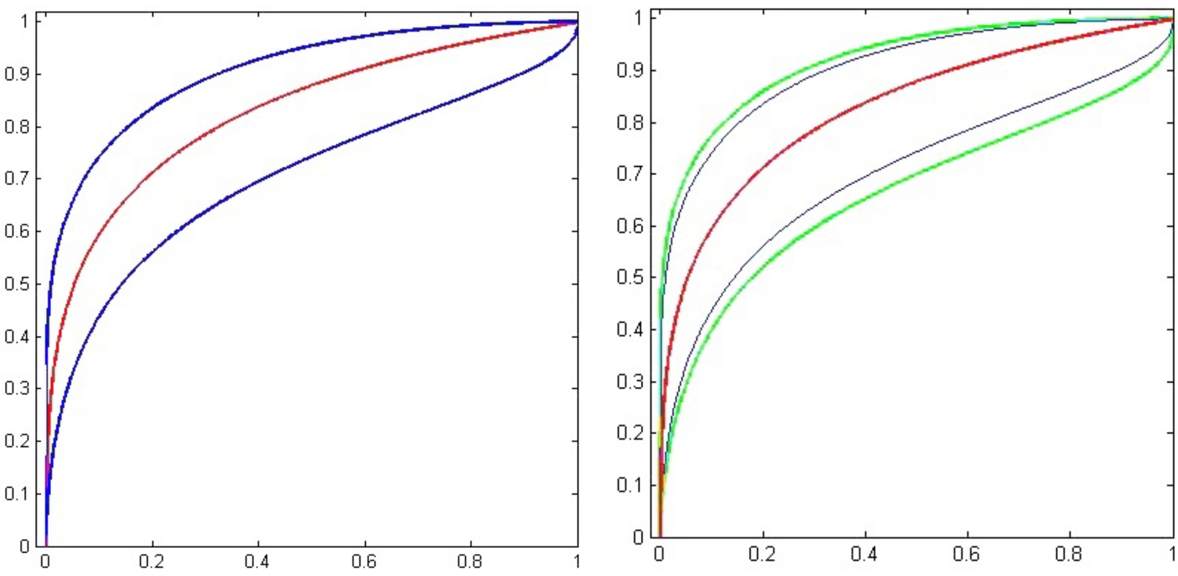
Poznámka. Alternativní konstrukcí lze získat takzvané *simultánní intervaly spolehlivosti*(simultaneous confidence bands) založené na Working-Hotelling modelu popsané v [4].

$$1 - \Phi\left[\widehat{a} - \widehat{b}\ \Phi^{-1}(1-t) \pm k_\alpha \sqrt{Var(\widehat{a} - \widehat{b}\ \Phi^{-1}(1-t))}\right],$$

kde $k = \sqrt{-2\ln(\alpha)}$.

Příklad 5.26. Odhad binormálního modelu pro data z příkladu 4.25 doplníme o asymptotický odhad 95% intervalu spolehlivosti senzitivity.

Kostrukce podle podle vztahu 5.12 je vykreslena modře na obrázku 6 vlevo a je následně srovnána s alternativní konstrukcí simultánního intervalu spolehlivosti (zeleně) v pravém grafu obrázku 6.



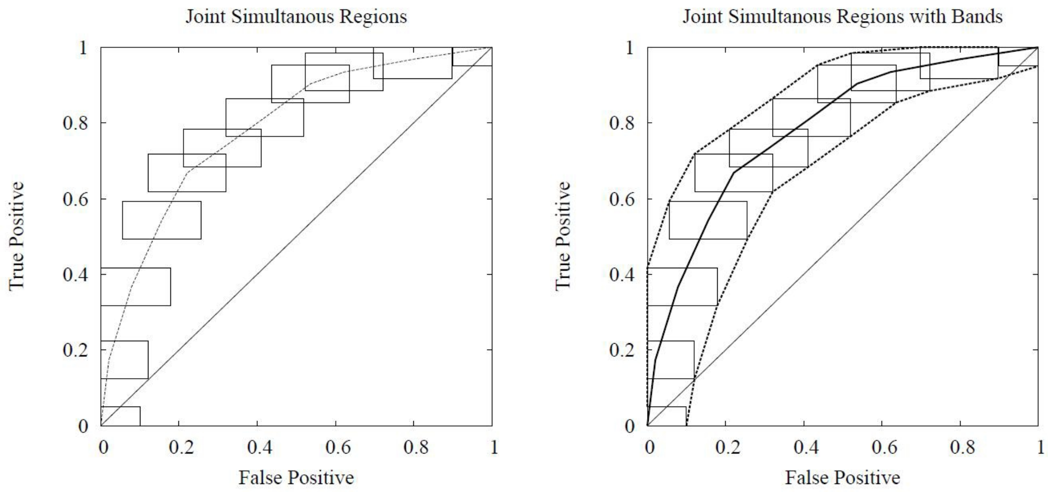
Obrázek 6: 95% intervaly spolehlivosti senzitivity testu CK-BB.

5.2 Simultánní sdružená oblast

V předchozích dvou případech byl postup založen na výpočtu intervalu spolehlivosti pro senzitivitu v daném bodě. Nyní uplatníme odlišný přístup. Pro senzitivitu a nezávisle pro specificitu určíme intervaly spolehlivosti s využitím Kolmogorovova-Smirnovova jednovýběrového testu. V bobě $[1 - F_0, 1 - F_1]$ je obdélníková oblast spolehlivosti dána

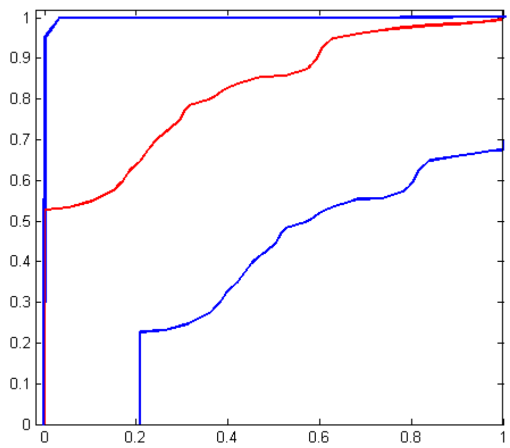
$$[1 - F_0 \pm d, 1 - F_1 \pm e],$$

kde d, e jsou příslušné kritické hodnoty K-S testu pro $(1 - \alpha)$ z tabulky 10 v příloze 2. Daný bod se pak v tomto obdélníku nachází s pravděpodobností $(1 - \alpha)^2$. Dolní hranici oblasti spohlevosti tvoří spojnice pravých dolních rohů jednotlivých obdélníků. Spojnice levých horních rohů vymezuje horní hranici.



Obrázek 7: JSR [4]

Příklad 5.27. Použijeme opět data z příkladu 4.25 a pro po částech lineární odhad ROC křivky zkonstruujeme sdruženou oblast spolehlivosti.



Obrázek 8: Simultánní sdružená oblast spolehlivosti testu CK-BB.

6 Plocha pod ROC křivkou - AUC

6.1 Lichoběžníkové pravidlo

Plocha pod křivkou může být přímo odhadnuta součtem obsahů lichoběžníků daných body empirické ROC křivky.

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathbf{\Psi}(T_{1i}, T_{0j}),$$

kde $\mathbf{\Psi}$ je funkcí dvou proměnných:

$$\mathbf{\Psi}(T_{1i}, T_{0j}) = \begin{cases} 1 & T_{1i} > T_{0j} \\ \frac{1}{2} & T_{1i} = T_{0j} \\ 0 & T_{1i} > T_{0j} \end{cases}$$

6.2 Plocha a parciální plocha pod křivkou binormálního modelu

Nyní se opět zaměříme na binormální model. Parciální plochou pod ROC křivkou se rozumí plocha pod křivkou mezi dvěmi danými hodnotami Sp respektive $1 - Sp$ (dva body na vodorovné ose). Tedy

$$AUC_{(e_1 \leq 1 - Sp(c) \leq e_2)} = \int_{c_1}^{c_2} \Phi(bv - a)\phi(v)dv,$$

kde

$$c_1 = \Phi^{-1}(e_1),$$
$$c_2 = \Phi^{-1}(e_2),$$
$$\phi(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}}.$$

Je tedy zřejmé, že maximální hodnotou bude plocha obdélníka

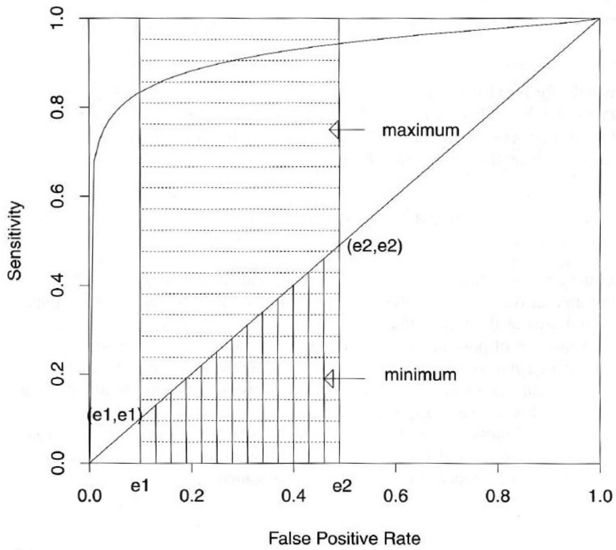
$$AUC_{(e_1 \leq 1 - Sp(c) \leq e_2)} \leq AUC_{max(e_1, e_2)} = (e_2 - e_1) \times 1.$$

Naopak minimální hodnota je plocha lichoběžníka omezeného diagonálou

$$AUC_{(e_1 \leq 1 - Sp(c) \leq e_2)} \geq AUC_{min(e_1, e_2)} = \frac{1}{2}(e_2 - e_1)(e_2 + e_1).$$

Označme náhodnou veličinu $Y = T_0 - T_1$, kde T_0 má normální rozdělení pravděpodobnosti s parametry (μ_0, σ_0^2) a T_1 má normální rozdělení pravděpodobnosti s parametry (μ_1, σ_1^2). Pak

$$Y = T_0 - T_1 \sim N(\mu_0 - \mu_1, \sigma_0^2 + \sigma_1^2) \quad , \quad U = \frac{T_0 - T_1 - (\mu_0 - \mu_1)}{\sqrt{\sigma_0^2 + \sigma_1^2}} \sim N(0, 1).$$



Obrázek 9: Maximální a minimální parciální plocha pod ROC křivkou [11].

Podle věty 3.19 vyjádříme plochu pod ROC křivkou

$$AUC = P(T_0 - T_1 \leq 0) = P\left(\frac{T_0 - T_1 - (\mu_0 - \mu_1)}{\sqrt{\sigma_0^2 + \sigma_1^2}} \leq \frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right) = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right).$$

Po dosazení $a = \frac{\mu_1 - \mu_0}{\sigma_1}$ a $b = \frac{\sigma_0}{\sigma_1}$ dostáváme

$$AUC = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right). \tag{6.13}$$

6.3 Testy hypotéz o AUC

Jak bylo již uvedeno výše, pokud je graf ROC křivky totožný s diagonálou, velikost polchy pod křivkou (přímkou) je rovna jedné polovině a zkoumaný diagnostický test má nulovou klasifikační schopnost.

Položíme tedy nulovou hypotézu

$$H_0 : \ AUC = \frac{1}{2},$$

proti alternativě

$$H_a : \ AUC \neq \frac{1}{2}.$$

Testovací statistikou bude

$$Z = \frac{\widehat{AUC} - 0.5}{\sqrt{\widehat{Var}(\widehat{AUC})}},$$

tato má přibližně(approximetely) normované normální rozdělení.

Dále je možné testovat hypotézu, že parciální AUC na daném intervalu nabývá svého maxima

$$H_0 : \ AUC_{(e_1 \leq FPR \leq e_2)} = AUC_{min},$$

proti alternativě

$$H_a : \ AUC_{(e_1 \leq FPR \leq e_2)} \neq AUC_{min}.$$

Zde bude testovací statistikou

$$Z = \frac{\widehat{AUC}_{(e_1 \leq FPR \leq e_2)} - AUC_{min}}{\sqrt{\widehat{Var}\left(\widehat{AUC}_{(e_1 \leq FPR \leq e_2)}\right)}}.$$

7 Volba optimální klasifikační meze

V této sekci se budeme zabývat problémem určení optimální meze pro klasifikaci objektu, tj. takového c, pro které jsou chyby v klasifikaci minimální.

Jak je vidět na obrázku 10 s rostoucí senzitivitou klesá specificita testu a naopak. Jestliže snižujeme chybu prvního druhu, roste chyba druhého druhu a pokud snižujeme chybu druhého druhu roste naopak chyba prvního druhu.

Mějme optimalizační úlohu ve smyslu minimalizace součtu chyby prvního a druhého druhu. V našem případě tedy

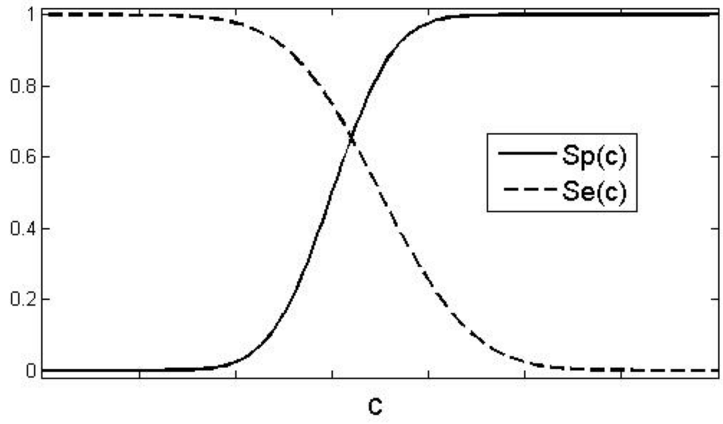
$$z(c) = 1 - Se(c) + 1 - Sp(c)$$

$$z \to \min .$$

Tuto lze převést na ekvivalentní tvar

$$z(c) = Se(c) + Sp(c) - 2$$

$$z \to \max .$$



Obrázek 10: Senzitivita a specificita

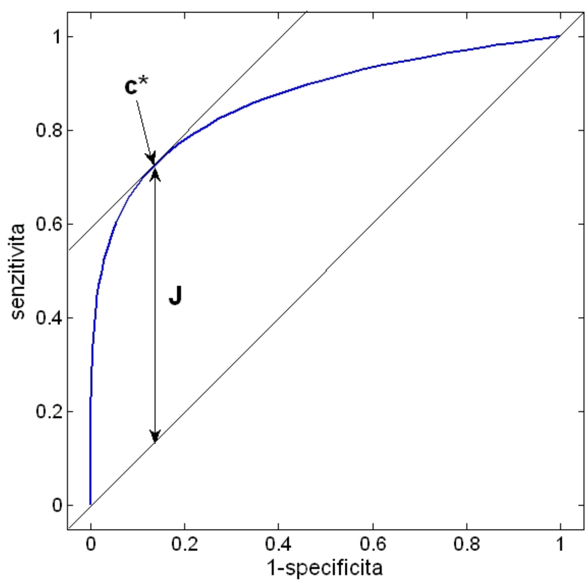
Jestliže od fukce z odečteme jedničku, pak maximum této funkce nazýváme *Youden index* (J),

$$J = \max_c \{Se(c) + Sp(c) - 1\}$$

Mějme parametrické vyjádření ROC křivky ve tvaru $[1 - Sp(c), Se(c)],\ c \in (-\infty, \infty)$. Grafem křivky $[1 - Sp(c), 1 - Sp(c)],\ c \in (-\infty, \infty)$ je diagonála v jednotkovém čtverci. Vzdálenost bodu ROC křivky od diagonály pro dané c je pak dána vztahem

$$\sqrt{(Se(c) - 1 + Sp(c))^2} = Se(c) + Sp(c) + 1$$

Graficky je tedy možné Youden index interpretovat jako největší vertikální vzdálenost mezi křivkou a diagonálou.



Obrázek 11: Youden index, optimální klasifikační mez

Poznámka. Řešení výše uvedeného problému je tedy bod křivky, ve kterém je směrnice tečny rovna jedné. V praxi se ale setkáváme s případy, kdy chyby prvního a druhého druhu nemají stejnou váhu. Pak řešíme úlohu

$$z(c) = k_1(1 - Se(c)) + k_2(1 - Sp(c))$$

$$z \to \min,$$

kde, k_1, k_2 jsou váhy jednotlivých chyb. Řešením tohoto problému je pak bod, ve kterém je směrnice tečny rovna k_2/k_1.

Příklad 7.28. Rozšíříme příklad 4.25 o určení optimální klasifikační meze. Pro výpočet odhadu Youden indexu využijeme metody bodových odhadů senzitivity a specificity z kapitoly 4. Výsledné hodnoty jsou uvedeny v tabulce 4

Metoda odhadu ROC	J	c
Empirická ROC křivka	0.53	286
Po částech lineární ROC křivka	0.53	286
Jádrový odhad ROC křivky	0.486	304.3
Odhad binormálního modelu ROC křivky	0.514	201.3
Nejlepší nestranný odhad Se a Sp	0.513	207.2

Tabulka 4: Optimální klasifikační mez CK-BB

Velký rozdíl mezi výsledky neparametrických a parametrických metod je v tomto případě zaviněn špatnou schopností prvních tří metod aproximovat ROC křivku v počáteční části, kdy křivka rychle roste.

8 Srovnání dvou ROC křivek

Označíme-li míru přesnosti daného diagnostického testu ϑ, pak tuto míru můžeme použít jako kritérium při srovnání dvou testů. Tedy testujeme nulovou hypotézu

$$H_0: \ \vartheta_1 = \vartheta_2,$$

proti

$$H_a: \ \vartheta_1 \neq \vartheta_2.$$

Opět použijeme statistiku

$$Z = \frac{\widehat{\vartheta}_1 - \widehat{\vartheta}_2}{\sqrt{Var(\widehat{\vartheta}_1 - \widehat{\vartheta}_2)}}.$$

8.1 Testy odlišnosti

Pro přímé srovnání dvou ROC křivek uvažujeme následující tvrzení:

1. Dvě křivky jsou shodné.

   Binormální model je plně popsán parametry a a b. Mají-li se dvě ROC křivky shodovat, musí se i jejich parametry rovnat. Položíme tedy nulovou hypotézu

   $$H_0: \ a_1 = a_2 \ \text{ a } \ b_1 = b_2,$$

   proti alternativě

   $$H_a: \ a_1 \neq a_2 \ \text{ nebo } \ b_1 \neq b_2.$$

   Pro tento test využijeme statistiku [Metz, Wang, Kronman]

   $$X^2 = \frac{\widehat{a}_{12} Var(\widehat{b}_{12}) + \widehat{b}_{12}^2 Var(\widehat{a}_{12}) - 2\widehat{a}_{12}\widehat{b}_{12} Cov(\widehat{a}_{12}, \widehat{b}_{12})}{Var(\widehat{a}_{12}) Var(\widehat{b}_{12}) - Cov(\widehat{a}_{12}, \widehat{b}_{12})^2},$$

   kde $a_{12} = a_1 - a_2$ a $b_{12} = b_1 - b_2$ jsou rozdíly parametrů srovnávaných křivek. Pro výpočet jednotlivých rozptylů a kovariancí lze využít vztahy uvedené výše v části 4.1.
   Tato statistika má asymptoticky chi-kvadrát rozdělení pravděpodobnosti s dvěma stupni volnosti.

2. Dvě křivky se shodují v partikulárním bodě.

   Opačný přístup je postaven na srovnání křivek v jednotlivých bodech. Pro tento účel zavádíme difrenci $D(Z_e)$ takto:

   $$D(Z_e) = (b_1 Z_e - a_1) - (b_2 Z_e - a_2) = b_{12} Z_e - a_{12}.$$

   Tato odpovídá rozdílu hodnot v bodě, kdy $1 - Sp(c) = e$. Jako nulovou hypotézu pak položíme

   $$H_0: \ D(Z_e) = 0, \ \text{ proti } \ H_a: \ D(Z_e) \neq 0.$$

   Testovací statistika

   $$Z = \frac{\widehat{D}(Z_e)}{\sqrt{Var[\widehat{D}(Z_e)]}}$$

   pak má normované normální rozdělení pravděpodobnosti.

8.2 Test ekvivalence

Narozdíl od předešlých testů, nyní posoudíme možnost výskytu statisticky významného rozdílu mezi dvěma diagnostickými testy, proti hypotéze, že tyto testy jsou ekvivalentní.

Pak je tedy testována nulová hypotéza

$$H_0: \ (\vartheta_1 - \vartheta_2) \leq \Delta_L \quad \text{nebo} (\vartheta_1 - \vartheta_2) \geq \Delta_U,$$

proti alternativní hypotéze (ekvivalence)

$$H_a: \ \Delta_L < (\vartheta_1 - \vartheta_2) < \Delta_U,$$

kde Δ_L je stanovená dolní mez a Δ_U horní mez.

Reálně jde o úlohu skládající se ze dvou testů

$$Z_1 = \frac{(\widehat{\vartheta}_1 - \widehat{\vartheta}_2) - \Delta_L}{\sqrt{Var(\widehat{\vartheta}_1 - \widehat{\vartheta}_2)}} \text{ a } Z_2 = \frac{\Delta_U - (\widehat{\vartheta}_1 - \widehat{\vartheta}_2)}{\sqrt{Var(\widehat{\vartheta}_1 - \widehat{\vartheta}_2)}}.$$

Nulovou hypotézu zamítáme, jestliže obě statistiky Z_1 i Z_2 jsou větší než příslušné kritické hodnoty na hladině α.

Pokud je první test alespoň tak dobrý jako druhý ($\vartheta_1 \geq \vartheta_2$), pak dostáváme nulovou hypotézu

$$H_0: \ (\vartheta_1 - \vartheta_2) \geq \Delta_M$$

a alternativu

$$H_a: \ (\vartheta_1 - \vartheta_2) < \Delta_M,$$

kde Δ_M je nejmenší možný rozdíl přesností, který ještě neznamená ekvivalenci.

Testovací statistika

$$Z_{NI} = \frac{\widehat{\vartheta}_1 + \Delta_M - \widehat{\vartheta}_2}{\sqrt{Var(\widehat{\vartheta}_1 - \widehat{\vartheta}_2)}}$$

má asymptoticky normované normální rozdělení pravděpodobnosti.

9 Ordinální data

V této části se budeme zabývat případem, kdy výsledek zkoumaného dagnostického testu může nabývat pouze konečného počtu uspořádaných hodnot (například 1 = velmi špatný, 2 = špatný, 3 = dobrý, 4 = velmi dobrý). Vysoké hodnoty pak značí pozitivní klasifikaci, nízké naopak negativní.

9.1 Empirická ROC křivka

Nechť výsledek testu T nabývá hodnot $1, \dots, K$. Pro každou ordinální hodnotu testu T, definujeme senzitivitu jako

$$\widehat{Se}(i) = P(T \geq i | D = 1) = \frac{1}{n_1} \sum_{j=i}^{K} s_j$$

a hodnotu $1 - Sp(c)$

$$1 - \widehat{Sp}(i) = P(T \geq i | D = 0) = \frac{1}{n_0} \sum_{j=i}^{K} r_j,$$

kde jednotlivé parametry jsou dány tabulkou

Reálný stav (D)	Výsledek testu (T) 1	...	K	Celkem
$D = 1$	s_1	...	s_K	n_1
$D = 0$	r_1	...	r_K	n_0
Celkem	m_1	...	m_K	N

Tabulka 5: Test s ordináními daty

Tedy s_j je počet jedinců s pozitivním sledovaným znakem a výsledkem testu $T = j$. Naopak r_j je počet jedinců se stejným výsledkem testu, ale negativním sledovaným znakem.

Empirická ROC křivka pro ordinální data je pak dána vykreslením párů $[1 - \widehat{Sp}(i), \widehat{Se}(i)]$, pro $i = 1 \dots K$, spojených lomenou čarou.

9.2 Parametricý model aproximace hladkou křivkou

Empirická křivka odhadnutá z $2 \times K$ hodnot je poměrně nepřesná a dává pouze hrubou představu o vlastnostech příslušného testu. Proto se snažíme najít vhodnou aproximaci.

Předpokládejme že, data ordinálního typu vznikla z dat původně spojitých. Obecně je tedy výsledek u pozitivních jedincú náhodná veličina T_1^* s distribuční funkcí F_1. Ve skupině negativních pak T_0^* s distribuční funkcí F_0. Je-li c klasifikační mez, ROC křivku pak získáme v parametrickém tvaru

$$[1 - F_0(c), 1 - F_1(c)], \quad -\infty < c < \infty$$

$$\begin{array}{lcl} T_i^* \leq \tilde{c}_1 & \rightarrow & T_i = 1 \\ \tilde{c}_{j-1} < T_i^* \leq \tilde{c}_j & \rightarrow & T_i = j, \quad j = 2, 3, \ldots, K-1 \\ T_i^* > \tilde{c}_{K-1} & \rightarrow & T_i = K \end{array}$$

Pro ordinální data předpokládáme $K-1$ neznámých klasifikačních mezí $\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_{K-1}$ takových, že pro $i = 0, 1$

Většinou předpokládáme, že F_1 i F_0 jsou distribuční funkce normálního rozdělení. Tedy že T_i^* jsou náhodné veličiny s normálním rozdělením pravděpodobnosi nebo existuje monotónní transformace dat na toto rozdělení. Pak

$$T_1^* \sim N(\mu_1, \sigma_1^0), \quad T_0^* \sim N(\mu_0, \sigma_0^2).$$

Dále pak postupujeme jako při odhadu binormálního modelu spojitých náhodných veličin. Více například v [8]

10 Simulační studie

V tomto úseku budou na simulovaných datech srovnány jednotlivé výše popsané metody.

10.1 Bodové odhady ROC křivky

Z dat vygenerovaných v programu Matlab provedeme konstrukci a následné srovnání empirické ROC křivky (EM), po částech lineární ROC křivky (PL), odhadu založeného na jádrových odhadech distribučních funkcí (JO), binormálního modelu (B) a ROC křivky pro nejlepší nestranný odhad senzitivity a specificity (K).

Pro binormální model s parametry $F_0 \sim N(0,1)$, $F_1 \sim N(1,1)$, byly vygenerovány výběry o celkovém rozsahu ($n = n_0 + n_1 = 20, 40, 100, 200, 800$) a to v poměru $n_0 = n_1$, $n_0 = 3n_1$ a $n_1 = 3n_0$. Pro každý stav bylo spuštěno 500 simulací. Pro srovnání teoretické křivky s jejím odhadem byla měřena vzdálenost bodu $[1 - F_0(c_i), 1 - F_1(c_i)]$ od jeho odhadu $[1 - \widehat{F_0}(c_i), 1 - \widehat{F}(c_i)]$

$$v_i = \sqrt{(\widehat{F_0}(c_i) - F_0(c_i))^2 + (\widehat{F_1}(c_i) - F_1(c_i))^2}$$

pro všechna generovaná c_i, $i = 1 \ldots, n$. Z těchto hodnot pro každý bodový odhad ROC křivky vypočteme směrodatnou chybu $RMSE$

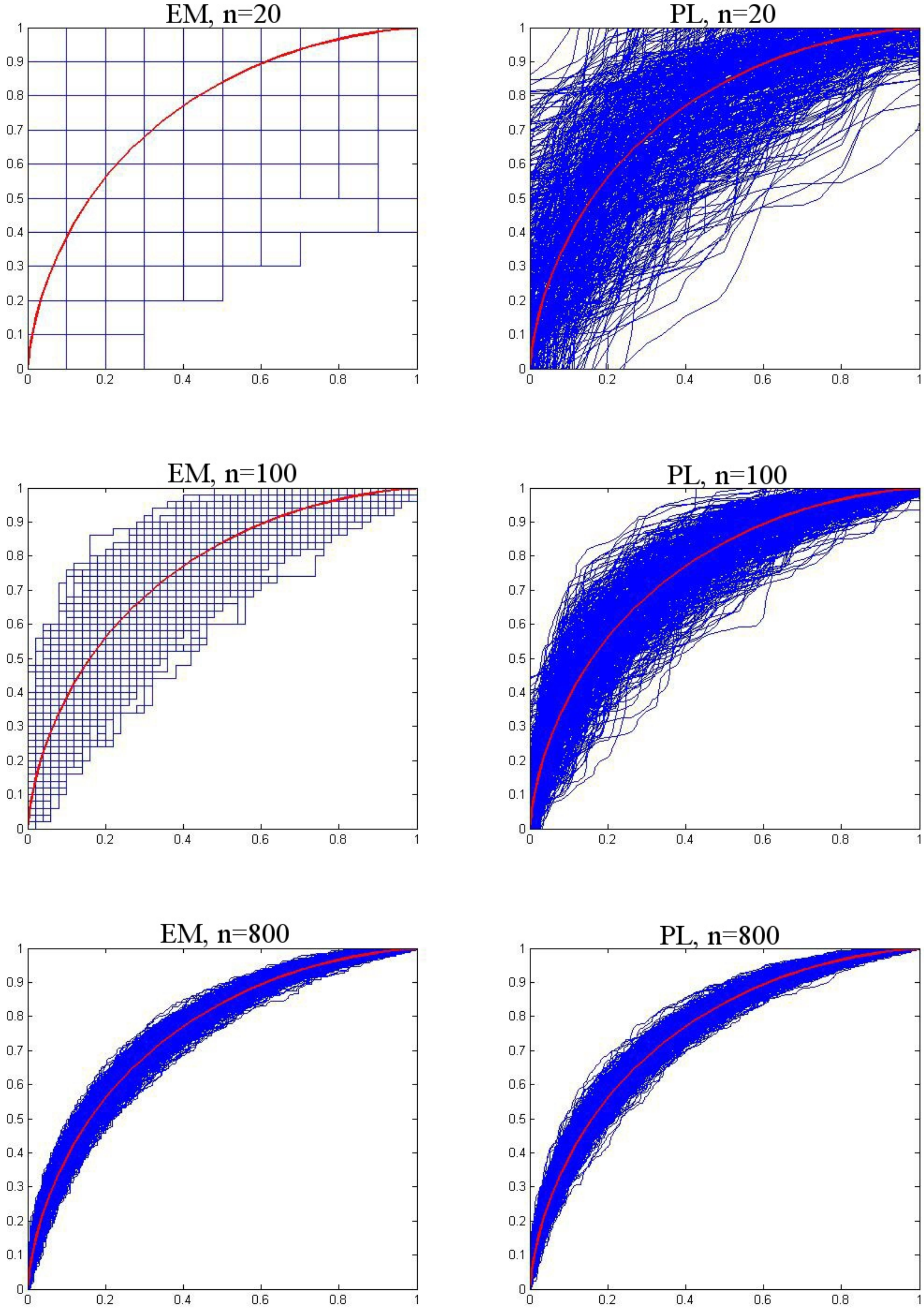
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} v_i^2}.$$

Takto pro každou jednotlivou metodu bodového odhadu ROC křivky vznikne soubor 500 hodnot RMSE. V tabulce 6 jsou pak uvedeny výběrové průměry a směrodatné odchylky RMSE.

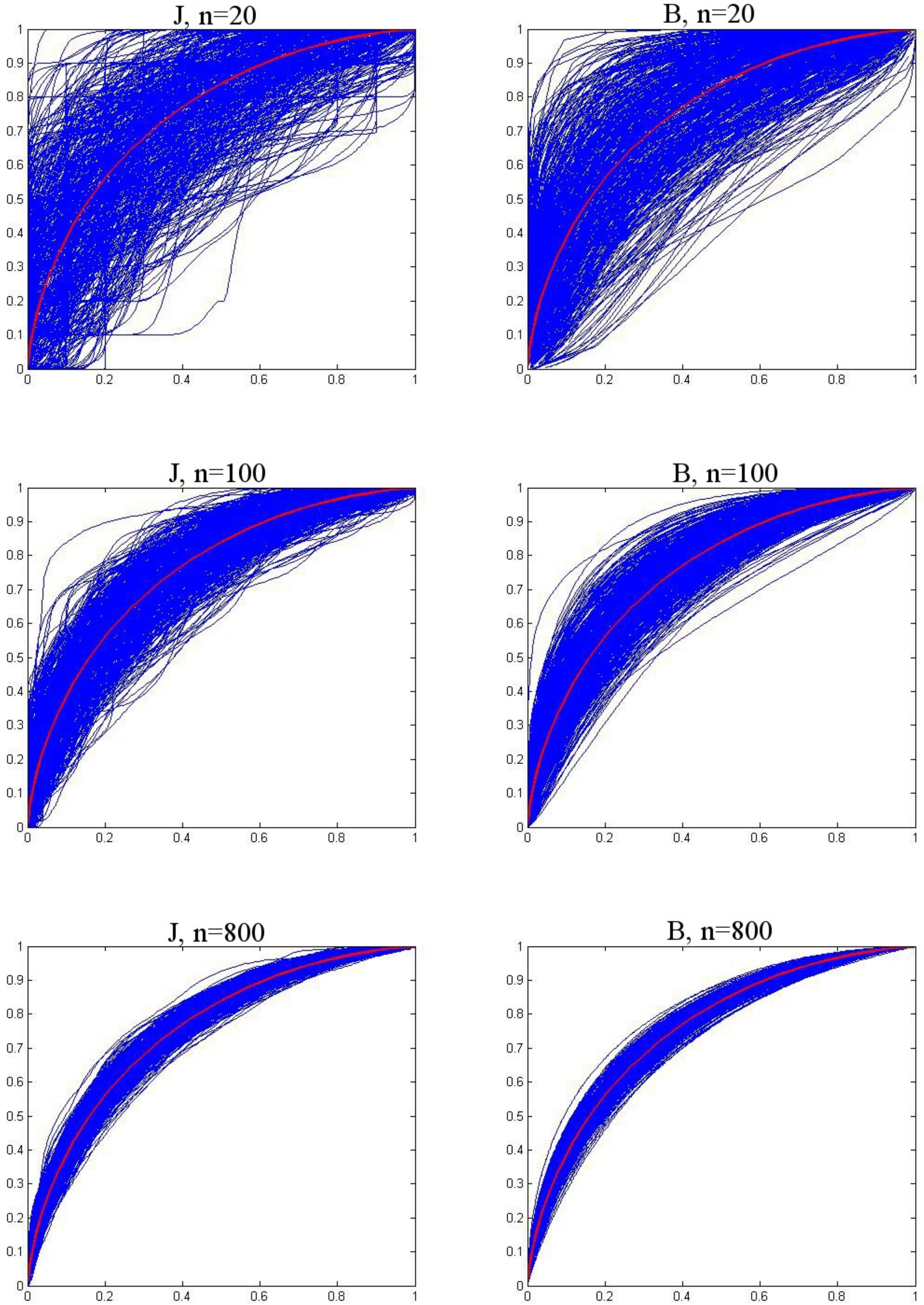
Průběh simulací pro $n = 20, 100, 800$ E, PL, JO a B proti teoretické ROC křivce (červeně) je vykreslen na obrázcích 12 a 13. Hodnoty B a K byly v tomto případě téměř identické, proto simulační studie ROC křivek založená na nejlepším nestranném odhadu senzitivity a specificity není zobrazena.

Metoda		$n = 20$		$n = 40$		$n = 100$		$n = 200$		$n = 800$	
		$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std
Empirická ROC křivka	$n_0 = n_1$	0.1686	0.0525	0.1164	0.0334	0.0730	0.0219	0.0522	0.0156	0.0259	0.0080
	$3n_0 = n_1$	0.1893	0.0555	0.1325	0.0390	0.0809	0.0255	0.0592	0.0200	0.0291	0.0091
	$n_0 = 3n_1$	0.1936	0.0557	0.1327	0.0402	0.0843	0.0261	0.0593	0.0177	0.0292	0.0087
Po částech lineární ROC křivka	$n_0 = n_1$	0.1548	0.0495	0.1105	0.0321	0.0712	0.0216	0.0516	0.0156	0.0258	0.0080
	$3n_0 = n_1$	0.1692	0.0540	0.1232	0.0378	0.0775	0.0256	0.0577	0.0200	0.0289	0.0090
	$n_0 = 3n_1$	0.1722	0.0557	0.1226	0.0403	0.0810	0.0261	0.0578	0.0177	0.0291	0.0087
Jádrový odhad ROC křivky	$n_0 = n_1$	0.1475	0.0539	0.1041	0.0346	0.0667	0.0229	0.0485	0.0164	0.0244	0.0083
	$3n_0 = n_1$	0.1632	0.0586	0.1182	0.0410	0.0730	0.0270	0.0544	0.0210	0.0274	0.0095
	$n_0 = 3n_1$	0.1664	0.0600	0.1178	0.0435	0.0770	0.0274	0.0546	0.0186	0.0276	0.0090
Odhad binormálního modelu ROC křivky	$n_0 = n_1$	0.1318	0.0579	0.0877	0.0367	0.0556	0.0238	0.0403	0.0172	0.0198	0.0087
	$3n_0 = n_1$	0.1451	0.0629	0.1020	0.0434	0.0603	0.0279	0.0444	0.0210	0.0222	0.0104
	$n_0 = 3n_1$	0.1465	0.0656	0.1020	0.0459	0.0638	0.0293	0.0459	0.0194	0.0223	0.0100
Nejlepší nestranný odhad Se a Sp ROC křivky	$n_0 = n_1$	0.1327	0.0578	0.0880	0.0367	0.0557	0.0238	0.0404	0.0172	0.0198	0.0087
	$3n_0 = n_1$	0.1481	0.0625	0.1028	0.0435	0.0605	0.0279	0.0444	0.0210	0.0222	0.0104
	$n_0 = 3n_1$	0.1495	0.0651	0.1028	0.0460	0.0639	0.0293	0.0460	0.0195	0.0223	0.0100

Tabulka 6: Simulační studie pro $F_0 \sim N(0,1)$ a $F_1 \sim N(1,1)$

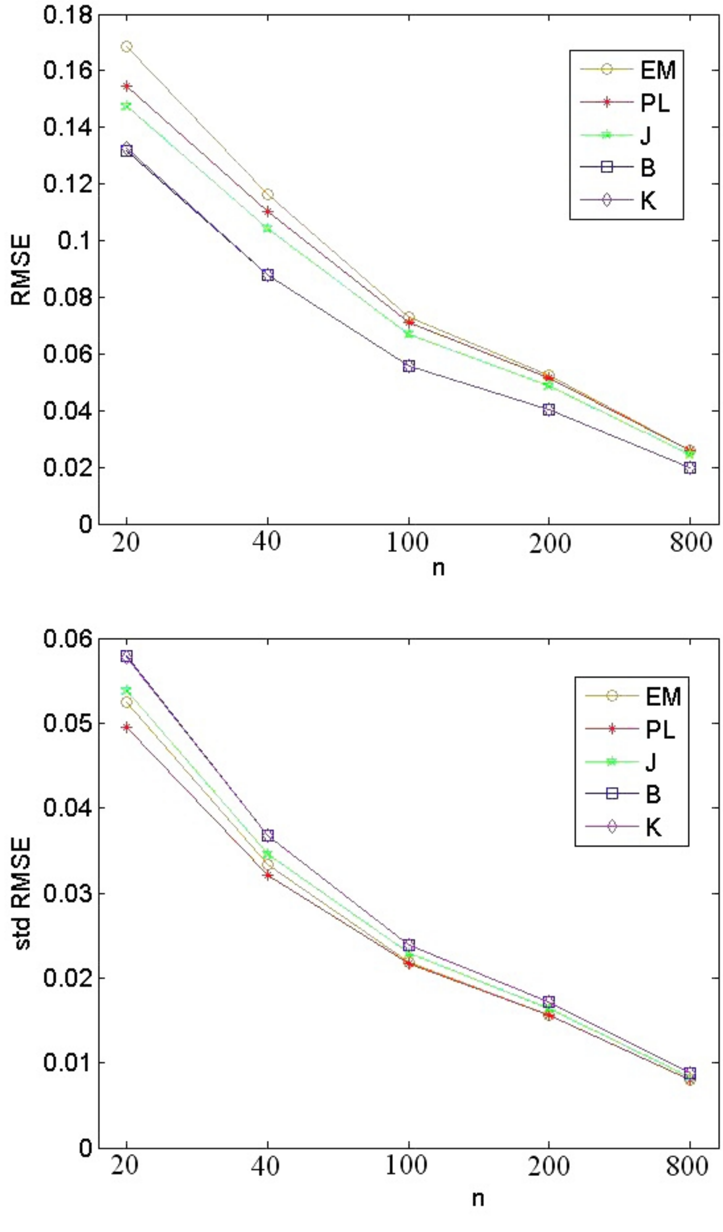


Obrázek 12: Simulace: empirické a pocástech lineární ROC křivky.



Obrázek 13: Simulace: jádrové odhady a odhady binormálního modelu ROC křivky.

Na obázku 14 vidíme srovnání metod konstrukce bodového odhadů ROC křivky pomocí průměrné hodnoty RMSE (graf v horní části) a směrodatných odchylek RMSE (dolní graf).



Obrázek 14: Simulace: srovnání bodových odhadů ROC křivky.

Dále byly stejným postupem provedeny simulace pro binormální model s parametry $F_0 \sim N(0,1)$, $F_1 \sim N(3,1)$ a model kdy F_0 a F_1 měly exponenciální rozdělení pravděpodobnosti s parametry $F_0 \sim exp(0.5)$, $F_1 \sim exp(1)$. Výsledky jsou zaznamemány v tabulkách 11 a 12 v příloze 3.

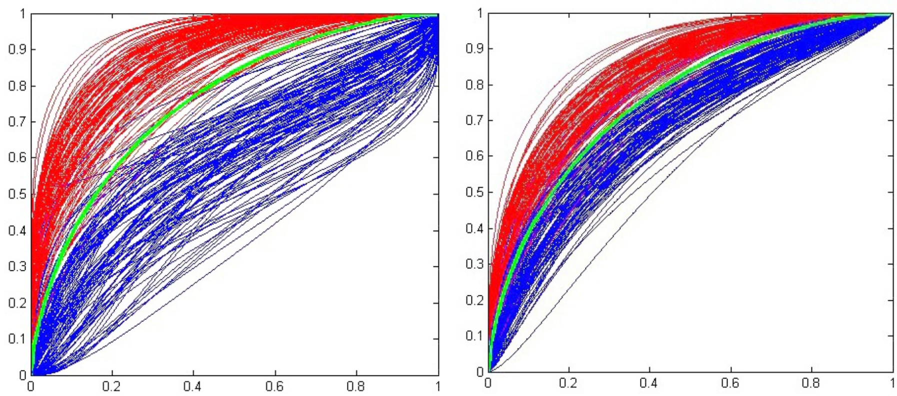
Ve všech případech z průběhů simulací pozorujeme, že pro malé rozsahy odhady ROC křivky pokrývají většinu polchy jednotkového čtverce. Nejvyšší přesnost vykazuje odhad binormálního modelu. Nejlepší nestranný odhad Se a Sp dává podobné výdledky, ale výpočetní náročnost této metody je vyšší. Jádrový odhad v popsaném tvaru nedokáže s dostatečnou přesností aproximovat ROC křivku v části úvodního rychlého růstu. To může způsobovat problém v následném hodnocení ROC křivky nebo při odhadu optimální klasifikační meze. Empirická a po částech lineární ROC křivka, hlavně pro malá n, dává pouze hrubou představu o tvaru křivky a je vhodné ji doplnit některým dalším odhadem.

10.2 Intervalové odhady

V této části bude provedena simulační studie metod intervalových odhadů ROC křivek popsaných v kapitole 5. Pro binormální model s parametry $F_0 \sim N(0,1)$, $F_1 \sim N(1,1)$, byly vygenerovány výběry o rozsahu $n = n_0 = n_1 = 10, 20, 50, 100, 400$. Pro každý stav bylo spuštěno 100 simulací. U jednotlivých metod pak byl sledován počet případů, ve kterých teoretická křivka zasahovala mimo odhadnuté hranice. Do tabulky 7 byly zaznamenány pozorované spolehlivosti (AI - asymptotické intervaly spolehlivosti, SI - simultánní intervaly spolehlivosti, JSR - simultánní sdružené oblasti spolehlivosti).

	n				
Metoda	10	20	50	100	400
AI	81%	76%	82%	90%	88%
SI	86%	81%	86%	94%	93%
JSR	100%	100%	100%	100%	100%

Tabulka 7: Pozorovaná spolehivost intervalových odhadů

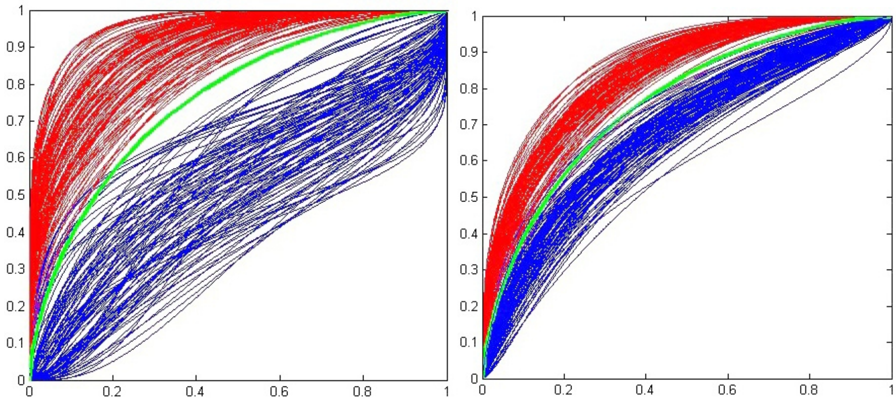


Obrázek 15: Průběh simulací AI pro n=20 vlevo a n=100 vpravo.

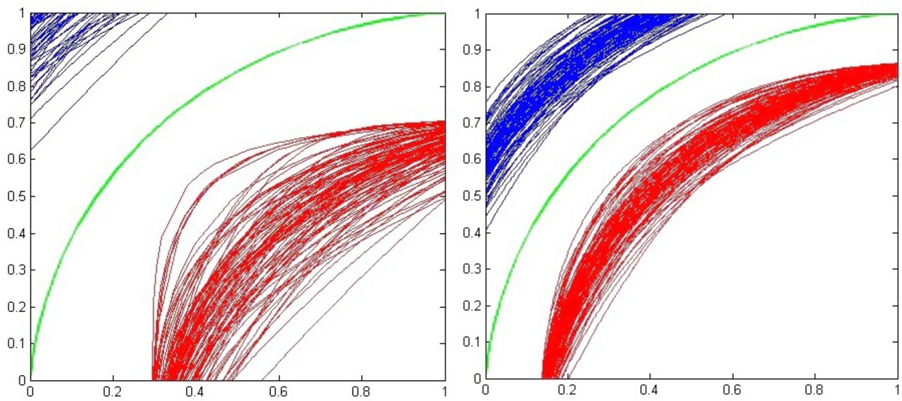
Při použití prvních dvou metod, nebyla ani v jednom případě dosažena spolehlivost 95%. Naopak u třetí metody teoretická křivka ležela vždy v odhadnuté oblasti viz. obrázek 17. V tomto případě jsou ale hranice nejširší.

10.3 Youden index a optimální c

V této části bude pomocí metod bodových odhadů senzitivity a specificity odhadnut youden index a příslušná hodnota optimální klasifikační meze. Průběh simulací je shodný jako při simulacích bodových odhadů ROC křivek pro binormální model s parametry $F_0 \sim N(0,1)$, $F_1 \sim N(1,1)$. Teoretická hodnota youden indexu $J = 0{,}383$ a optimální klasifikační mez $c = 0{,}5$.



Obrázek 16: Průběh simulací SI pro n=20 vlevo a n=100 vpravo.



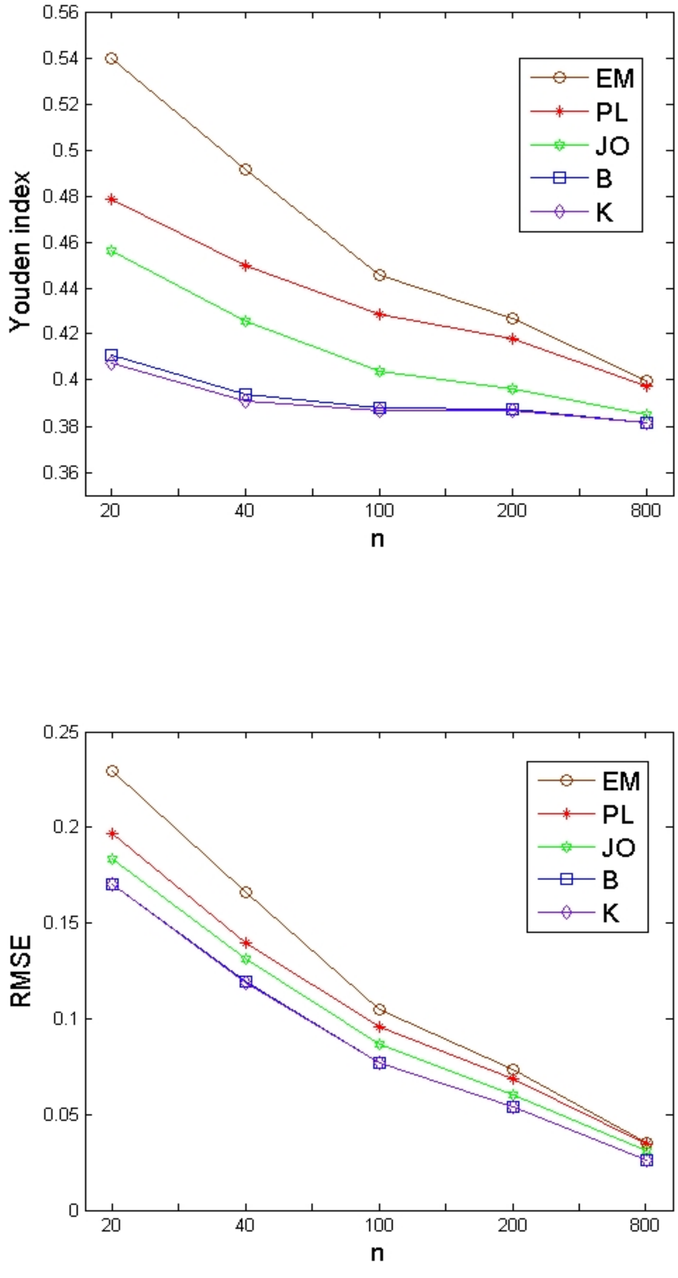
Obrázek 17: Průběh simulací JSR pro n=20 vlevo a n=100 vpravo.

	$n = 20$		$n = 40$		$n = 100$		$n = 200$		$n = 800$	
Metoda	J	RMSE	J	RMSE	J	RMSE	J	RMSE	J	RMSE
EM	0,54	0,230	0,49	0,169	0,45	0,105	0,43	0,074	0,40	0,0353
PL	0,48	0,196	0,45	0,140	0,43	0,0958	0,42	0,069	0,40	0,034
JO	0,46	0,184	0,43	0,131	0,40	0,087	0,40	0,060	0,39	0,031
B	0,41	0,171	0,39	0,1192	0,39	0,077	0,39	0,054	0,38	0,027
K	0,41	0,170	0,39	0,119	0,39	0,0769	0,39	0,054	0,38	0,026

Tabulka 8: Youden index pro $F_0 \sim N(0,1)$ a $F_1 \sim N(1,1)$

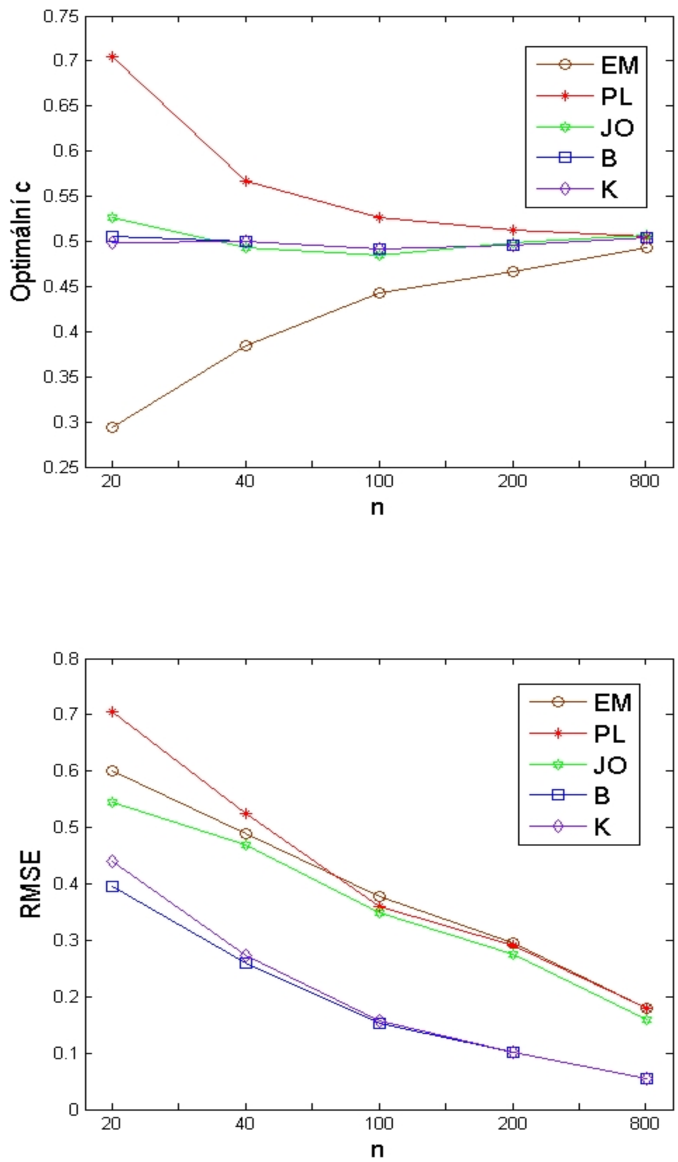
	$n = 20$		$n = 40$		$n = 100$		$n = 200$		$n = 800$	
Metoda	c	RMSE	c	RMSE	c	RMSE	c	RMSE	c	RMSE
EM	0,29	0,600	0,38	0,489	0,44	0,377	0,46	0,296	0,49	0,172
PL	0,7	0,705	0,57	0,524	0,52	0,359	0,51	0,2916	0,51	0,178
JO	0,53	0,545	0,49	0,469	0,48	0,347	0,50	0,276	0,51	0,160
B	0,51	0,395	0,50	0,259	0,49	0,153	0,49	0,100	0,50	0,054
K	0,50	0,439	0,50	0,273	0,49	0,157	0,49	0,102	0,50	0,054

Tabulka 9: Optimální c a RMSE



Obrázek 18: Odhad Youden indexu a jeho RMSE

I v případě hledání odhadu youden indexu se ukazuje jako nejvhodnější metoda odhad senzitivity a specificity jako distribuční funkce normálního rozdělení pravděpodobnosti. Všechny odhady s rostoucím rozsahem výběru klesají k teoretické hodnotě 0,383. Bereme-li hodnotu J jako měřítko kvality testu, pak je tento odhad nadhodnocený.



Obrázek 19: Odhad optimální klasifikační meze a RMSE

Odhad optimální klasifikační meze na základě maximalizace youden indexu se u metod JO, B a K ukázal jako přesný i u malých rozsahů.

11 Závěr

V úvodních částech byly uvedeny základní poznatky z teorie odhadu a testování statitických hypotéz. Dále byla zavedena ROC křivka jako funkce senzitivity a specificity zkoumaného testu a na teoretických příkladech byly demonstrovány její základní vlastnosti a parametry. Zde vidíme první možnost posouzení kvality testu z tvaru ROC křivky.

V části 4 byly popsány metody bodových odhadů ROC křivky. Z neparametrických metod to byl odhad empirické a po částech lineární ROC křivky. Ze simulačních studií v kapitole 10 vyplývá, že tyto metody poskytují pouze hrubý odhad, často poměrně vzdálený od teoretické křivky. Dalším neparametrickým odhadem ROC křivky byla metoda založená na jádrových odhadech distribučních funkcí. Výhodou této metody je schopnost aproximovat ROC křivku hladkou křivkou, nevýhodou je pak nepříznivý vliv hraničního efektu.

Parametrická metoda odhadu binormálního modelu a metoda nejlepšího nestranného odhadu senzitivity a specificity binormálního modelu pak na základě srovnání simulačních studií vycházejí jako nejpřesnější. Nevýhodou metody nejlepšího nestranné odhadu Se a Sp je její výpočetní náročnost.

Dále pak byly popsány metody intervalových odhadů ROC křivek. Také tyto byly srovnány pomocí simulovaných dat. Pozorovaná spolehlivost byla nejvyšší u sdržené simultánní oblasti, ale hranice této oblasti jsou podstatně širší než u ostatních metod.

V následujících kapitolách je pak popsána problematika výpočtu plochy pod ROC křivkou, která udává kvalitu testovacího kritéria, volba optimální klasifikační meze. Kapitola 8 se zabývá testy statistických hypotéz o ROC křivkách, sloužících k vzájemnému srovnání testů.

Téma statistické analýzy ROC křivek je poměrně rozsáhlé. Tato práce přináší popis základních používaných metod a jejich srovnání. Dále je možné se zabývat hledáním a úpravou jednotlivých metod pro konkrétní praktické úlohy, nebo naopak pokračovat v popisu nových metod na teoretické úrovni.

Reference

[1] ANDĚL, J. *Matematická statistika.* SNTL/ALFA Praha, 1978.

[2] GREINER, M. - PFEIFFER, D. - SMITH, R.D.: Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. In *Preventive Veterinary Medicine 45.* p. 23-41, 2000

[3] KUTÁLEK, D: *Užití ROC křivek ke klasifikaci objektů.* [Bakalářská práce] Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2008.

[4] MACSKASSY, A. - PROVOST F. Confidence Bands for ROC Curves: Methods and an Empirical Study. Proceedings of the First Workshop on ROC Analysis in AI. August 2004.

[5] MICHÁLEK, J. - VESELÝ, V. A Comparison of the ROC curve estimators by simulations.

[6] MICHÁLEK, J. - VESELÝ, V. The ROC and ODC curve estimators in binomial model based on the best unbiased estimator of CDF. XXIII International Colloquium on the Acqusition Process Managemant. Universitz of Defence Brno 2005.

[7] PAVLÍK, J. *Aplikovaná statistika.* 1. vyd. Vysoká škola chemicko-technologická v Praze, Praha 2005, ISBN 80-7080-569-2.

[8] PEPE, M.S.: *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press, 2004

[9] SEDLAČÍK, M.: *Využití ROC křivek při konstrukci klasifkačních a regresních stromů* [Disertační práce.] Brno: Masarykova univerzita, Přírodovědecká fakulta, 2006.

[10] SCHISTERMAN E.F. - PERKINS N.J. - LIU A., BONDELL H. Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Pooled Blood Samples, 2005.

[11] ZHOU X.H., OBUCHOVSKI N.A., McCLISH D.K. *Statistical methods in Diagnostic Medicine.* John Wiley. 2002

[12] *Receiver operating characteristic* [online], poslední revize 12.6.2010 [cit. 2010-14-06]. Dostupné z `<http://en.wikipedia.org/wiki/Receiver_operating_characteristic>`

12 Seznam použitých zkratek a symbolů

AI	asymptotický interval spolehlivosti
AUC	area under curve
B	odhad binormálního modelu ROC křivky
EM	empirická ROC křivka
J	Youden index
JO	jádrový odhad ROC křivky
JSR	simultánní sdružena oblast spolehlivosti
K	nejlepší nestranný odhad senzitivity a specificity binormálního modelu podle Kolmogorova
PL	po částech lineární ROC křivka
ROC	receiver operating characteristic
Se	senzitivita
SI	simultánní interval spolehlivosti
Sp	specificita

$\boldsymbol{a}$	sloupcový vektor reálných čísel
$\boldsymbol{a}'$	transponovaný vektor
$\boldsymbol{A}'$	transponovaná matice
$\boldsymbol{A}^{-1}$	inverzní matice
$\|\boldsymbol{A}\|$	determinant matice
$\mathcal{B}$	systém borelovských množin
$\mathrm{E}X$	střední hodnota náhodné veličiny X
$varX$	rozptyl náhodné veličiny X
F^{-1}	inverzní distribuční funkce
$\mathbb{R}^m$	reálný m-rozměrný prostor
$P(a\|b)$	podmíněná pravděpodobnost jevu a za podmínky b
Θ	parametrický prostor
Ω	prostor elementárních jevů

13 Seznam příloh

1. CD s implementací jednotlivých algoritmů v jazyce MATLAB.
2. Tabulka kritických hodnot pro jednovýběrový K-S test.
3. Tabulky výsledků simulačních studií.

14 přílohy

n	α 0.2	0.1	0.05	0.02	0.01	n	α 0.2	0.1	0.05	0.02	0.01
1	0.900	0.950	0.975	0.990	0.995	31	0.187	0.214	0.238	0.266	0.285
2	0.684	0.776	0.842	0.900	0.929	32	0.184	0.211	0.234	0.262	0.281
3	0.565	0.636	0.708	0.785	0.829	33	0.182	0.208	0.231	0.258	0.277
4	0.493	0.565	0.624	0.689	0.734	34	0.179	0.205	0.227	0.254	0.273
5	0.447	0.509	0.563	0.627	0.669	35	0.177	0.202	0.224	0.251	0.269
6	0.410	0.468	0.519	0.577	0.617	36	0.174	0.199	0.221	0.247	0.265
7	0.381	0.436	0.483	0.538	0.576	37	0.172	0.196	0.218	0.244	0.262
8	0.358	0.410	0.454	0.507	0.542	38	0.170	0.194	0.215	0.241	0.258
9	0.339	0.387	0.430	0.480	0.513	39	0.168	0.191	0.213	0.238	0.255
10	0.323	0.369	0.409	0.457	0.4899	40	0.165	0.189	0.210	0.235	0.252
11	0.308	0.352	0.391	0.438	0.468	41	0.163	0.187	0.208	0.232	0.249
12	0.296	0.338	0.375	0.419	0.449	42	0.162	0.185	0.205	0.229	0.246
13	0.285	0.325	0.361	0.404	0.432	43	0.160	0.183	0.203	0.227	0.243
14	0.275	0.314	0.349	0.390	0.418	44	0.158	0.181	0.201	0.224	0.241
15	0.266	0.304	0.338	0.377	0.404	45	0.156	0.179	0.198	0.222	0.238
16	0.258	0.295	0.327	0.366	0.392	46	0.155	0.177	0.196	0.219	0.235
17	0.250	0.286	0.318	0.355	0.381	47	0.153	0.175	0.194	0.217	0.233
18	0.244	0.279	0.309	0.346	0.371	48	0.151	0.173	0.192	0.214	0.231
19	0.237	0.271	0.301	0.337	0.361	49	0.150	0.171	0.190	0.213	0.228
20	0.232	0.265	0.294	0.329	0.352	50	0.148	0.170	0.188	0.211	0.226
21	0.226	0.259	0.287	0.321	0.344	51	0.147	0.168	0.187	0.209	0.224
22	0.221	0.253	0.281	0.314	0.337	52	0.146	0.166	0.185	0.207	0.222
23	0.216	0.247	0.275	0.307	0.330	53	0.144	0.165	0.183	0.205	0.220
24	0.212	0.242	0.269	0.301	0.323	54	0.143	0.163	0.181	0.203	0.218
25	0.208	0.238	0.264	0.295	0.317	55	0.142	0.162	0.180	0.201	0.216
26	0.204	0.233	0.259	0.290	0.311	56	0.140	0.160	0.178	0.199	0.214
27	0.200	0.229	0.254	0.284	0.305	57	0.139	0.159	0.177	0.198	0.212
28	0.197	0.225	0.250	0.279	0.300	58	0.138	0.158	0.175	0.196	0.210
29	0.193	0.221	0.246	0.275	0.295	59	0.137	0.156	0.174	0.194	0.208
30	0.190	0.218	0.242	0.270	0.290	$n \geq 60$	$\frac{1.07298}{\sqrt{n}}$	$\frac{1.22387}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$

Tabulka 10: Kritické hodnoty pro jednovýběrový Kolmogorovův-Smirnovův test

		$n = 20$		$n = 40$		$n = 100$		$n = 200$		$n = 800$	
Metoda		$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std
Empirická ROC křivka	$n_0 = n_1$	0.1346	0.0419	0.0935	0.0279	0.0581	0.0172	0.0411	0.0115	0.0204	0.0059
	$3n_0 = n_1$	0.1371	0.0437	0.0947	0.0275	0.0601	0.0176	0.0424	0.0124	0.0208	0.0058
	$n_0 = 3n_1$	0.1410	0.0439	0.0974	0.0285	0.0604	0.0172	0.0422	0.0115	0.0209	0.0056
Po částech	$n_0 = n_1$	0.1302	0.0402	0.0909	0.0269	0.0575	0.0170	0.0409	0.0115	0.0204	0.0059
lineární	$3n_0 = n_1$	0.1328	0.0403	0.0935	0.0267	0.0596	0.0173	0.0420	0.0121	0.0207	0.0058
ROC křivka	$n_0 = 3n_1$	0.1330	0.0406	0.0949	0.0275	0.0595	0.0173	0.0420	0.0116	0.0208	0.0057
Jádrový	$n_0 = n_1$	0.1179	0.0432	0.0828	0.0281	0.0530	0.0179	0.0380	0.0120	0.0193	0.0062
odhad	$3n_0 = n_1$	0.1156	0.0395	0.0839	0.0273	0.0548	0.0181	0.0391	0.0127	0.0195	0.0061
ROC křivky	$n_0 = 3n_1$	0.1183	0.0421	0.0860	0.0282	0.0547	0.0177	0.0392	0.0122	0.0196	0.0059
Odhad	$n_0 = n_1$	0.1028	0.0467	0.0711	0.0307	0.0439	0.0189	0.0313	0.0130	0.0157	0.0065
binormálního modelu	$3n_0 = n_1$	0.1037	0.0425	0.0720	0.0301	0.0453	0.0194	0.0330	0.0138	0.0157	0.0068
ROC křivky	$n_0 = 3n_1$	0.1064	0.0443	0.0743	0.0310	0.0453	0.0188	0.0317	0.0131	0.0157	0.0064
Nejlepší nestranný	$n_0 = n_1$	0.1032	0.0466	0.0712	0.0307	0.0439	0.0189	0.0314	0.0130	0.0157	0.0065
odhad Se a Sp	$3n_0 = n_1$	0.1046	0.0421	0.0722	0.0300	0.0454	0.0194	0.0330	0.0138	0.0157	0.0068
ROC křivky	$n_0 = 3n_1$	0.1072	0.0440	0.0745	0.0310	0.0453	0.0188	0.0317	0.0131	0.0157	0.0064

Tabulka 11: Simulační studie pro $F_0 \sim N(0, 1)$ a $F_1 \sim N(1, 1)$

		$n = 20$		$n = 40$		$n = 100$		$n = 200$		$n = 800$	
Metoda		$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std	$\overline{\text{RMSE}}$	std
Empirická ROC křivka	$n_0 = n_1$	0.1707	0.0485	0.1189	0.0355	0.0733	0.0221	0.0525	0.0163	0.0256	0.0075
	$3n_0 = n_1$	0.2014	0.0676	0.1405	0.0470	0.0866	0.0294	0.0621	0.0214	0.0307	0.0097
	$n_0 = 3n_1$	0.1867	0.0561	0.1273	0.0386	0.0807	0.0247	0.0567	0.0168	0.0285	0.0089
Po částech	$n_0 = n_1$	0.1572	0.0460	0.1123	0.0341	0.0713	0.0218	0.0518	0.0162	0.0255	0.0075
lineární	$3n_0 = n_1$	0.1757	0.0657	0.1283	0.0459	0.0830	0.0291	0.0608	0.0215	0.0305	0.0097
ROC křivka	$n_0 = 3n_1$	0.1680	0.0554	0.1187	0.0380	0.0776	0.0245	0.0554	0.0168	0.0283	0.0089
Jádrový	$n_0 = n_1$	0.1476	0.0503	0.1043	0.0357	0.0665	0.0224	0.0490	0.0166	0.0251	0.0075
odhad	$3n_0 = n_1$	0.1688	0.0701	0.1206	0.0488	0.0770	0.0289	0.0569	0.0214	0.0297	0.0095
ROC křivky	$n_0 = 3n_1$	0.1612	0.0614	0.1132	0.0404	0.0736	0.0257	0.0526	0.0175	0.0275	0.0090
Odhad	$n_0 = n_1$	0.1376	0.0529	0.0926	0.0393	0.0570	0.0252	0.0410	0.0184	0.0214	0.0082
binormálního modelu	$3n_0 = n_1$	0.1658	0.0778	0.1120	0.0531	0.0695	0.0321	0.0488	0.0228	0.0256	0.0105
ROC křivky	$n_0 = 3n_1$	0.1515	0.0657	0.0989	0.0451	0.0636	0.0274	0.0450	0.0191	0.0233	0.0098
Nejlepší nestranný	$n_0 = n_1$	0.1382	0.0529	0.0928	0.0394	0.0570	0.0252	0.0410	0.0184	0.0213	0.0082
odhad Se a Sp	$3n_0 = n_1$	0.1687	0.0780	0.1127	0.0532	0.0696	0.0321	0.0488	0.0229	0.0256	0.0105
ROC křivky	$n_0 = 3n_1$	0.1535	0.0660	0.0992	0.0453	0.0636	0.0275	0.0450	0.0192	0.0233	0.0098

Tabulka 12: Simulační studie pro $F_0 \sim Exp(0.5)$ a $F_1 \sim Exp(1)$