Analýza dat pro Neurovědy

RNDr. Eva Koriťáková, Ph.D.

doc. RNDr. Ladislav Dušek, Dr.

Blok 6

Jak analyzovat kategoriální a binární data II.

Osnova

1. Hodnocení diagnostických testů
2. Hledání diagnostického cut-off pomocí ROC křivek

1. Hodnocení diagnostických testů

Diagnostické testy

- Příklady: hodnocení úspěšnosti diagnostiky pomocí neuropsychologických testů, hodnocení úspěšnosti klasifikace pacientů s Alzheimerovou chorobou a kontrolních subjektů.

- Diagnostický test u dané osoby indikuje přítomnost nebo nepřítomnost sledovaného onemocnění.
- Osoba ve skutečnosti má nebo nemá sledované onemocnění.
  → **Zajímají nás diagnostické schopnosti testu.**

		Skutečnost – přítomnost nemoci	
		Ano	Ne
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

Diagnostické testy

		Skutečnost – přítomnost nemoci	
		Ano	Ne
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

- **TP** („true positive“) – kolik výsledků bylo skutečně pozitivních (tzn. kolik pacientů bylo správně diagnostikováno jako pacienti).
- **FP** („false positive“) – kolik výsledků bylo falešně pozitivních (tzn. kolik zdravých jedinců bylo chybně diagnostikováno jako pacienti).
- **FN** („false negative“) – kolik výsledků bylo falešně negativních (tzn. kolik pacientů bylo chybně diagnostikováno jako zdraví).
- **TN** („true negative“) – kolik výsledků bylo skutečně negativních (tzn. kolik zdravých lidí bylo správně diagnostikováno jako zdraví).

Senzitivita, specificita a celková správnost

		Skutečnost – přítomnost nemoci	
		Ano	Ne
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

- **Senzitivita testu**: schopnost testu rozpoznat skutečně nemocné osoby, tedy pravděpodobnost, že test bude pozitivní, když je osoba skutečně nemocná. Senzitivita testu = TP / (TP + FN)
- **Specificita testu**: schopnost testu rozpoznat osoby bez nemoci, tedy pravděpodobnost, že test bude negativní, když osoba není nemocná. Specificita testu = TN / (FP + TN)
- **Celková správnost:** (TP+TN)/(TP+FP+FN+TN)

Pozitivní a negativní prediktivní hodnota

		Skutečnost – přítomnost nemoci	
		Ano	Ne
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

- **Prediktivní hodnota pozitivního testu**: pravděpodobnost, že osoba je skutečně nemocná, když je test pozitivní.

  Prediktivní hodnota pozitivního testu = TP / (TP + FP)
- U klasifikací označována jako **přesnost** („precision“).

- **Prediktivní hodnota negativního testu**: pravděpodobnost, že osoba není nemocná, když je test negativní.

  Prediktivní hodnota negativního testu = TN / (FN + TN)

Shrnutí

		Skutečnost – přítomnost nemoci	
		Ano	Ne
Výsledek diagnostického testu	Pozitivní	TP	FP
	Negativní	FN	TN

TP + FP → **Prediktivní hodnota pozitivního testu**

FN + TN → **Prediktivní hodnota negativního testu**

TP + FN ↓ **Senzitivita testu**

FP + TN ↓ **Specificita testu**

Hodnocení diagnostických testů

- **Příklad:** Zajímá nás přesnost diagnostiky schizofrenie pomocí neuropsychologických testů. Výsledky diagnostiky jsou dány tabulkou:

Výsledek diagnostického testu	Skutečnost		
	Nemocný	Zdravý	Celkem
Nemocný	32	2	34
Zdravý	3	24	27
Celkem	35	26	61

Senzitivita testu = 32 / 35 = **91,4 %** (IS = 75,8 – 97,8)

Specificita testu = 24 / 26 = **92,3 %** (IS = 73,4 – 98,7)

Celková správnost = (32 + 24) / (32+2+3+24) = **91,8 %**

Pozitivní prediktivní hodnota testu = 32 / 34 = **94,1 %** (IS = 78,9 – 99,0)

Negativní prediktivní hodnota testu = 24 / 27 = **88,9 %** (IS = 69,7 – 97,1)

- Výpočet pomocí webových kalkulátorů:
  - https://www.medcalc.org/calc/diagnostic_test.php
  - http://vassarstats.net/clin1.html

Věrohodnostní poměr („Likelihood Ratio“)

- Věrohodnostní poměr (LR) lze definovat následovně:

$$LR = \frac{\text{(pravděpodobnost, že test dosáhne daného výsledku u nemocných pacientů)}}{\text{(pravděpodobnost, že test dosáhne daného výsledku u zdravých osob)}}$$

- 2 druhy věrohodnostního poměru:
  1. **LR+ (LR pro pozitivní test)** – podíl pravděpodobnosti, že nemocný člověk je testem diagnostikován jako pozitivní, a pravděpodobnosti, že zdravý člověk je chybně diagnostikován jako pozitivní.
     $$LR+ = senzitivita \,/ (1 - specificita)$$
  2. **LR- (LR pro negativní test)** – podíl pravděpodobnosti, že nemocný člověk je testem chybně diagnostikován jako negativní, a pravděpodobnosti, že zdravý člověk je diagnostikován jako negativní.
     $$LR- = (1 - senzitivita) \,/\, specificita$$
- U kvalitního diagnostického testu chceme, aby LR+ bylo co nejvyšší (LR+ > 10) a LR- co nejnižší (LR- < 0,1).

Věrohodnostní poměr

- **Příklad:** Chceme zjistit věrohodnostní poměr pozitivního a negativního testu u diagnostiky schizofrenie pomocí neuropsychologických testů.

Výsledek diagnostického testu	Skutečnost		
	Nemocný	Zdravý	Celkem
Nemocný	32	2	34
Zdravý	3	24	27
Celkem	35	26	61

Senzitivita testu = 32 / 35 = **91,4 %** (IS = 75,8 – 97,8)

Specificita testu = 24 / 26 = **92,3 %** (IS = 73,4 – 98,7)

LR+ = *senzitivita* / (1-*specificita*) = 0,914 / (1-0,923) = **11,870**

LR- = (1-*senzitivita*) / *specificita* = (1-0,914)/ 0,923 = **0,093**

Úkol 1.

- **Zadání:** U 1000 žen byl proveden test, zda jejich dítě bude trpět Downovým syndromem. Výsledky jsou uvedené v tabulce. Vypočtěte senzitivitu, specificitu, pozitivní a negativní prediktivní hodnotu a věrohodnostní poměry pro diagnostický test. Zamyslete se nad tím, zda je test dobrý či nikoliv.

Výsledek diagnostického testu	Skutečnost		
	Zdravé dítě	Dítě s Downovým syndromem	Celkem
Pozitivní	122	18	140
Negativní	857	3	860
Celkem	979	21	1000

Poznámka: pro výpočet použijte libovolný z online kalkulátorů uvedených dříve. Dejte si však pozor, abyste zadali do kalkulátoru čísla správně!

Úkol 1.

- **Řešení pomocí Medcalc:**

Test	Disease: Present	n	Disease: Absent	n	Total
Positive	True Positive	a= 18	False Positive	c= 122	a + c = **140**
Negative	False Negative	b= 3	True Negative	d= 857	b + d = **860**
		a + b = **21**		c + d = **979**	

Statistic	Formula	Value	95% CI	
Sensitivity	$\frac{a}{a+b}$	85.71%	63.66% to 96.95%	←
Specificity	$\frac{d}{c+d}$	87.54 %	85.30% to 89.54%	←
Positive Likelihood Ratio	$\frac{Sensitivity}{1-Specificity}$	6.88	5.41 to 8.75	←
Negative Likelihood Ratio	$\frac{1-Sensitivity}{Specificity}$	0.16	0.06 to 0.47	←
Disease prevalence	$\frac{a+b}{a+b+c+d}$	2.10% (*)	1.30% to 3.19%	
Positive Predictive Value	$\frac{a}{a+c}$	12.86% (*)	10.39% to 15.81%	←
Negative Predictive Value	$\frac{d}{b+d}$	99.65 % (*)	99.01% to 99.88%	←
Accuracy	$\frac{a+d}{a+b+c+d}$	87.50% (*)	85.29% to 89.49%	

Závěr:
Test má sice vcelku vysokou senzitivitu a specificitu, ale celkově příliš dobrý není (ukazují na to i věrohodnostní poměry). Skvěle vychytává ženy s nemocným dítětem (vysoká negativní prediktivní hodnota), ale spousta žen je zbytečně vystrašená, že bude mít nemocné dítě, ale pak se jim narodí zdravé dítě (nízká pozitivní prediktivní hodnota).

Úkol 1.

- **Řešení pomocí Vassarstats:**

	Condition		
	Absent	Present	Totals
Test Positive	122	18	140
Test Negative	857	3	860
Totals	979	21	1000

Calculate Reset

	Estimated Value	95% Confidence Interval	
		Lower Limit	Upper Limit
Prevalence	0.021	0.013379	0.032489
Sensitivity	0.857143	0.626434	0.962357
Specificity	0.875383	0.85267	0.895089
For any particular test result, the probability that it will be:			
Positive	0.14	0.119402	0.163417
Negative	0.86	0.836583	0.880598
For any particular positive test result, the probability that it is:			
True Positive (Positive Predictive Value)	0.128571	0.080049	0.198174
False Positive	0.871429	0.801826	0.919951
For any particular negative test result, the probability that it is:			
True Negative (Negative Predictive Value)	0.996512	0.988941	0.999099
False Negative	0.003488	0.000901	0.011059
likelihood Ratios: [C] = conventional [W] = weighted by prevalence [definitions]			
Positive [C]	6.87822	5.405528	8.752136
Negative [C]	0.163194	0.057226	0.465386
Positive [W]	0.147541	0.095411	0.228154
Negative [W]	0.003501	0.001131	0.010832

Závěr:
Test má sice vcelku vysokou senzitivitu a specificitu, ale celkově příliš dobrý není (ukazují na to i věrohodnostní poměry). Skvěle vychytává ženy s nemocným dítětem (vysoká negativní prediktivní hodnota), ale spousta žen je zbytečně vystrašená, že bude mít nemocné dítě, ale pak se jim narodí zdravé dítě (nízká pozitivní prediktivní hodnota).

Úkol 2.

- **Zadání:** Byl vytvořen algoritmus pro klasifikaci pacientů s Alzheimerovou chorobou a kontrolních subjektů. Zjistěte, jaká je úspěšnost tohoto klasifikačního algoritmu (použijte proměnné group_13_CnAd a group_klasif).

group_klasif * group_13_CnAd Crosstabulation

Count

		group_13_CnAd		
		1 CN	3 AD	Total
group_klasif	1 CN	213	8	221
	3 AD	17	189	206
Total		230	197	427

Statistic	Formula	Value	95% CI
Sensitivity	$\frac{a}{a+b}$	95.94%	92.16% to 98.23%
Specificity	$\frac{d}{c+d}$	92.61 %	88.43% to 95.64%
Positive Likelihood Ratio	$\frac{Sensitivity}{1-Specificity}$	12.98	8.21 to 20.53
Negative Likelihood Ratio	$\frac{1-Sensitivity}{Specificity}$	0.04	0.02 to 0.09
Disease prevalence	$\frac{a+b}{a+b+c+d}$	46.14% (*)	41.33% to 50.99%
Positive Predictive Value	$\frac{a}{a+c}$	91.75% (*)	87.55% to 94.62%
Negative Predictive Value	$\frac{d}{b+d}$	96.38 % (*)	93.10% to 98.13%
Accuracy	$\frac{a+d}{a+b+c+d}$	94.15% (*)	91.48% to 96.18%

Řešení pomocí Medcalc:

	Disease				
Test	**Present**	**n**	**Absent**	**n**	**Total**
Positive	True Positive	a= 189	False Positive	c= 17	a + c = **206**
Negative	False Negative	b= 8	True Negative	d= 213	b + d = **221**
Total		a + b = **197**		c + d = **230**	

Závěr: klasifikační algorimus má vysokou správnost, senzitivitu i specificitu.

Úkol 2.

- **Zadání:** Byl vytvořen algoritmus pro klasifikaci pacientů s Alzheimerovou chorobou a kontrolních subjektů. Zjistěte, jaká je úspěšnost tohoto klasifikačního algoritmu (použijte proměnné group_13_CnAd a group_klasif).

group_klasif * group_13_CnAd Crosstabulation

Count

		group_13_CnAd		
		1 CN	3 AD	Total
group_klasif	1 CN	213	8	221
	3 AD	17	189	206
Total		230	197	427

Řešení pomocí Vassarstats:

	Condition		
	Absent	Present	Totals
Test Positive	17	189	206
Test Negative	213	8	221
Totals	230	197	427

Calculate Reset

	Estimated Value	95% Confidence Interval	
		Lower Limit	Upper Limit
Prevalence	0.461358	0.413478	0.509944
Sensitivity	0.959391	0.918687	0.980996
Specificity	0.926087	0.88226	0.955041
For any particular test result, the probability that it will be:			
Positive	0.482436	0.434252	0.53094
Negative	0.517564	0.46906	0.565748
For any particular positive test result, the probability that it is:			
True Positive (Positive Predictive Value)	0.917476	0.86897	0.949738
False Positive	0.082524	0.050262	0.13103
For any particular negative test result, the probability that it is:			
True Negative (Negative Predictive Value)	0.963801	0.927309	0.983071
False Negative	0.036199	0.016929	0.072691
likelihood Ratios: [C] = conventional [W] = weighted by prevalence [definitions]			
Positive [C]	12.979994	8.207536	20.527507
Negative [C]	0.04385	0.022228	0.086504
Positive [W]	11.117647	7.038338	17.56126
Negative [W]	0.037559	0.019016	0.074183

Závěr: klasifikační algorimus má vysokou senzitivitu i specificitu. Správnost by bylo nutno dopočítat ručně.

2. Hledání diagnostického cut-off pomocí ROC křivek.

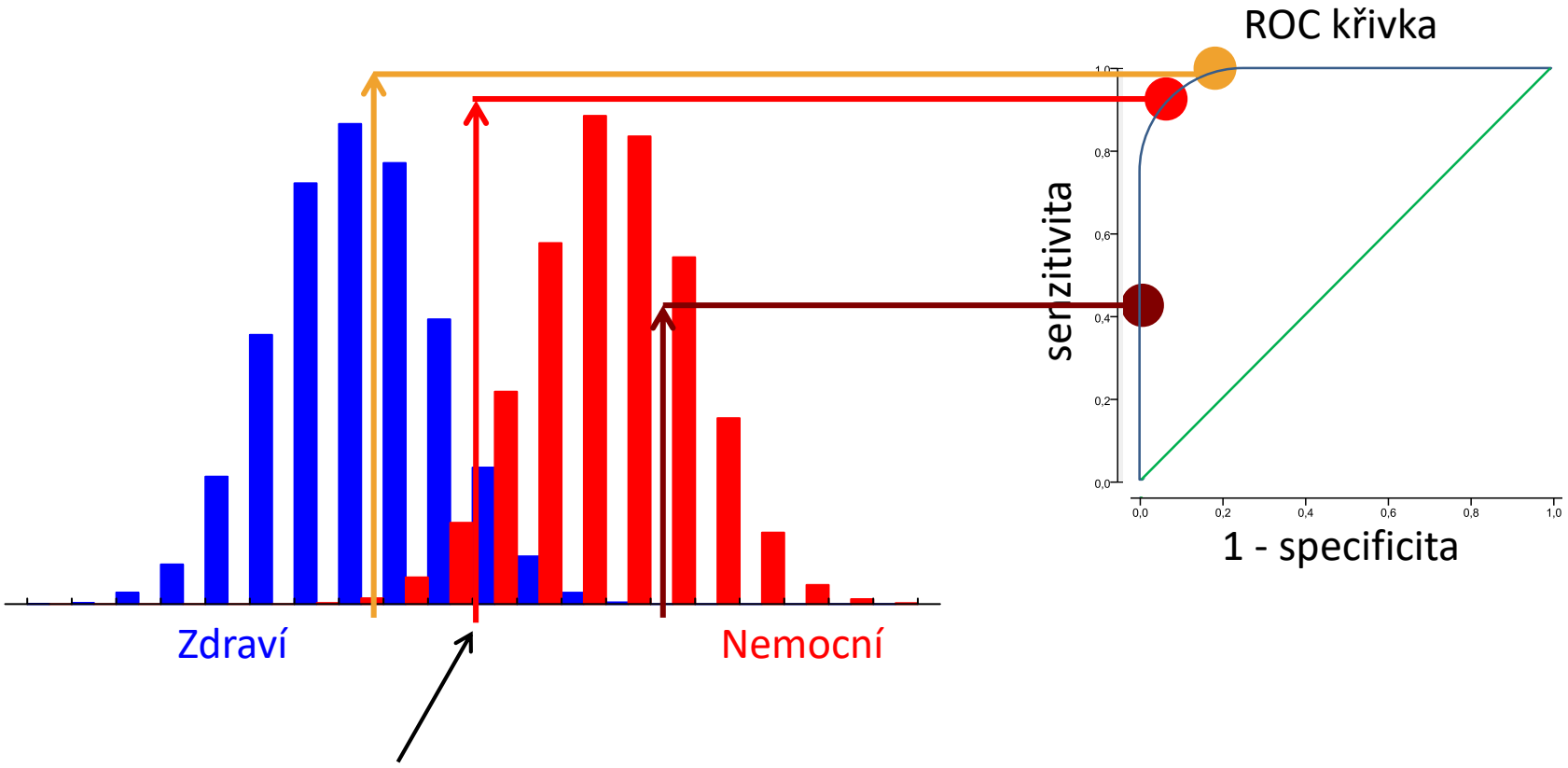
ROC analýza – motivace

- Dříve probrané ukazatele diagnostické síly testů (senzitivita, specificita apod.) **nelze použít u diagnostických testů, jejichž výstupem je spojitá (kvantitativní) proměnná** (např. koncentrace analytu v krevním séru, systolický krevní tlak).
- Pokud na základě předchozích výzkumů známe dělicí body, které odlišují normální a patologické hodnoty spojité proměnné, můžeme pomocí nich spojitou proměnnou binarizovat – tzn. vytvoření dvou kategorií „pozitivní“ / „negativní“ (např. „pod normou“ / „v normě“) – a pak použít výpočet senzitivity, specificity atd.
- Pokud dělicí body nejsou známy předem, můžeme se je snažit nalézt pomocí **ROC („Receiver Operating Characteristic“) křivky**.

- **Cíle ROC analýzy:**
  1. Určit, zda je spojitá proměnná vhodná pro diagnostické odlišování zdravých a nemocných jedinců.
  2. Nalezení dělicího bodu („cut-off point“) na škále hodnot spojité proměnné, který nejlépe odlišuje zdravé a nemocné jedince.

ROC analýza

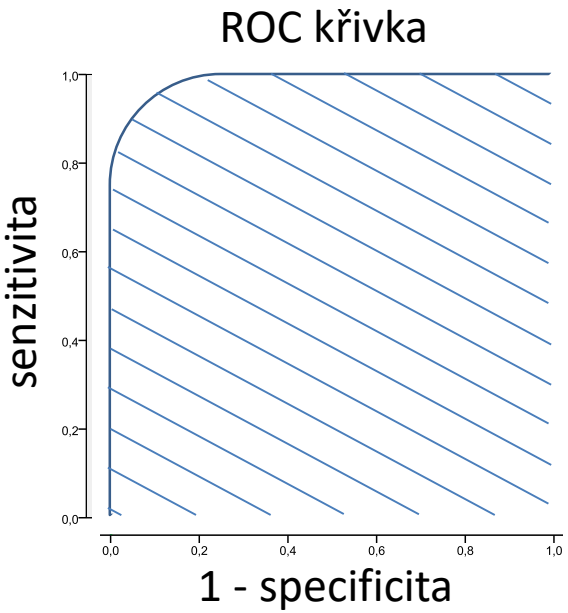
- Princip: Jakákoli hodnota spojité proměnné nějak rozlišuje zdravé a nemocné jedince, tzn. je spojena s nějakou senzitivitou a specificitou.



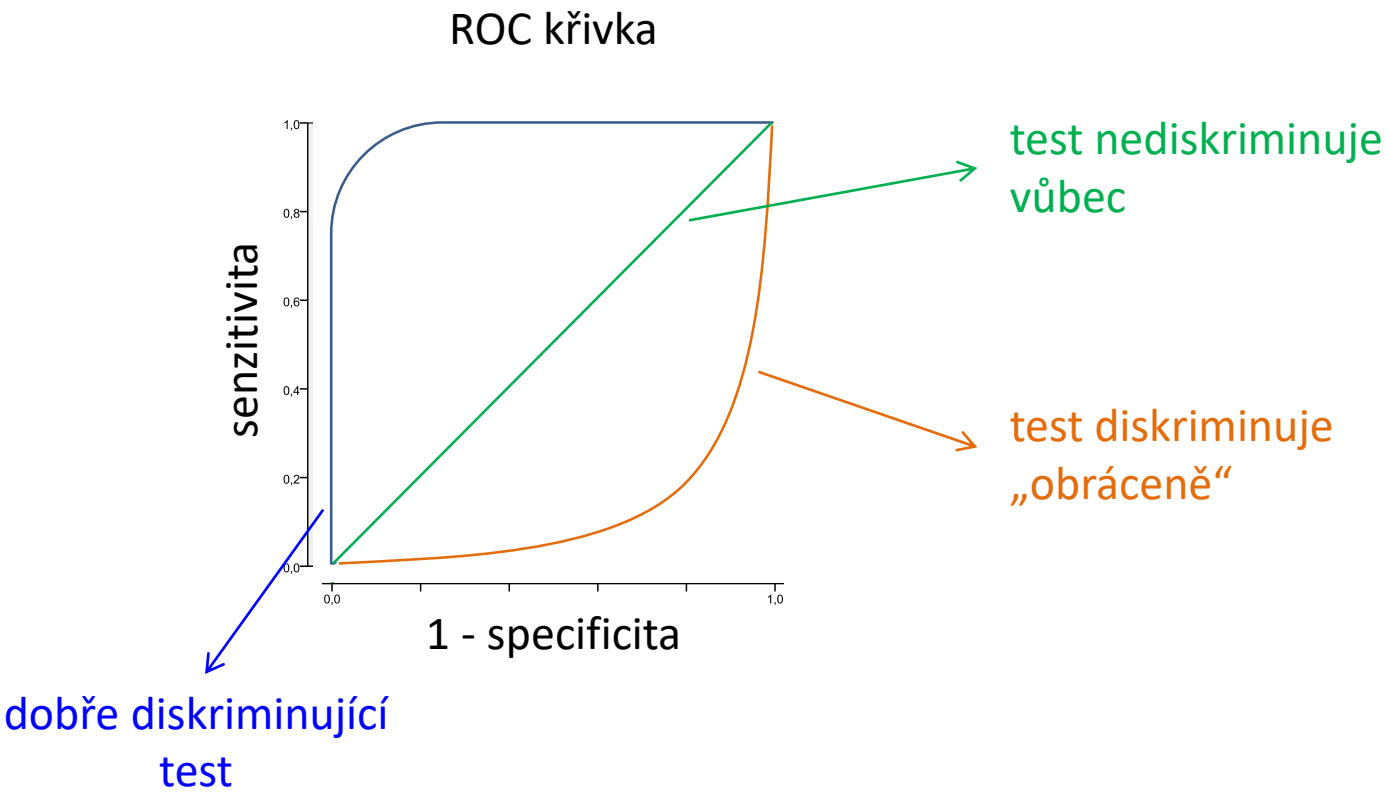
Nejlepší dělící bod („cut-off“) – nejvyšší sensitivita a specificita pro odlišení skupin – tzn. maximální součet hodnot senzitivity a specificity.

ROC analýza – plocha pod ROC křivkou

- Plocha pod ROC křivkou = „Area Under the Curve“ (AUC).
- Nabývá hodnot od 0 do 1.
- Slouží k vyjádření diagnostické síly (efektivity) testu.
- Čím větší hodnota AUC, tím lepší diagnostický test je (hodnota AUC nad 0,75 většinou poukazuje na uspokojivou diskriminační schopnost testu).

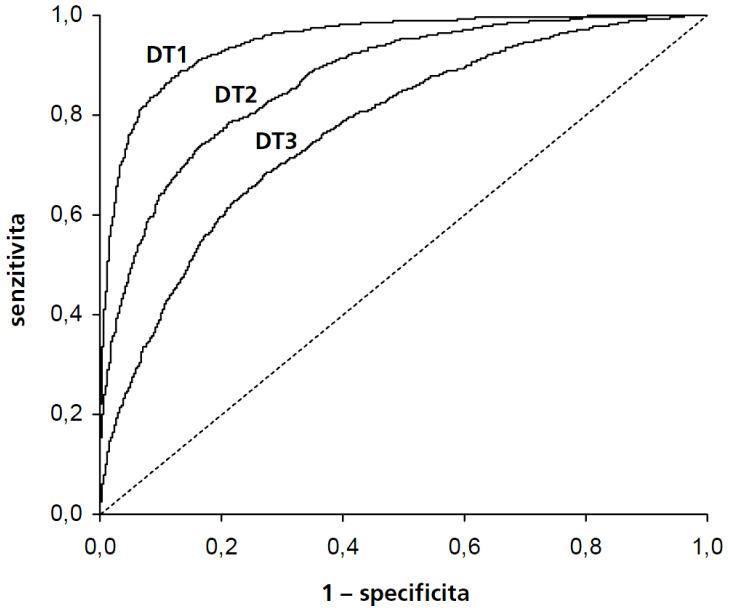


ROC analýza – srovnání diagnostické síly různých testů



ROC analýza – srovnání diagnostické síly různých testů

- Lze srovnat i velmi rozdílné testy (např. testy založené na různých proměnných).

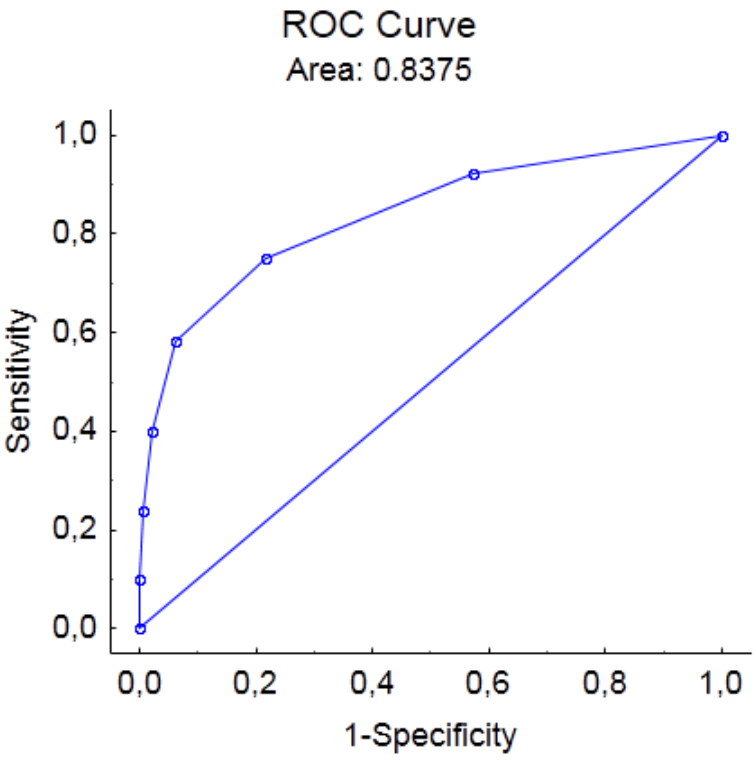


Diagnostický test	AUC	
DT1	0,949	→ nejlepší
DT2	0,872	
DT3	0,770	→ nejhorší

Zdroj: Dušek, Pavlík, Jarkovský, Koptíková, Analýza dat v Neurologii, Cesk Slov Neurol N 2011; 74/ 107(4)

ROC analýza

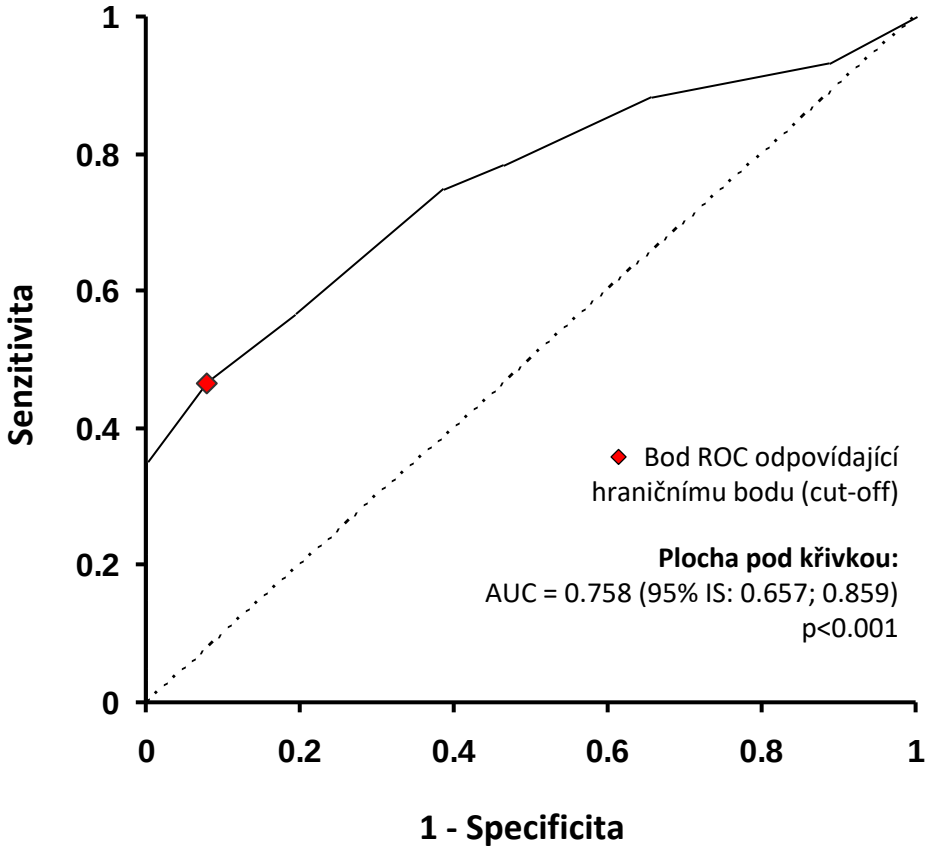
Příklad: Zjistěte, zda je MMSE skóre vhodné na diagnostiku mírné kognitivní poruchy (MCI). Najděte dělící bod (cut-off), který nejlépe odlišuje pacienty s MCI od kontrolních subjektů.



MMSE skóre	Sensitivity	1-Specificity	Specificity	Sensitivity + Specificity
22	0,000	0,000	1,000	1,000
23,5	0,002	0,000	1,000	1,002
24,5	0,101	0,000	1,000	1,101
25,5	0,239	0,004	0,996	1,235
26,5	0,399	0,022	0,978	1,377
27,5	0,581	0,061	0,939	1,520
28,5	0,749	0,217	0,783	1,531
29,5	0,924	0,574	0,426	1,350
31	1,000	1,000	0,000	1,000

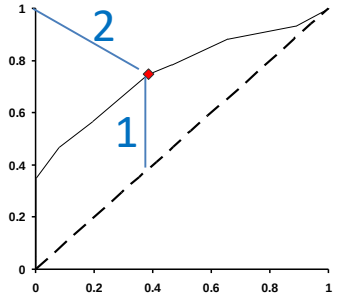
Hledání cut-off – doplnění

Příklad:



Sens	Spec	Sens+Spec
1.000	0.000	1.000
0.933	0.115	1.049
0.883	0.346	1.229
0.783	0.538	1.322
0.750	0.615	1.365
0.567	0.808	1.374
0.467	0.923	1.390
0.350	1.000	1.350
0.217	1.000	1.217
0.150	1.000	1.150
0.050	1.000	1.050
0.033	1.000	1.033
0.000	1.000	1.000

Hledání cut-off – kritéria



Kritérium	Vzoreček	Reference
1. Youdenova J statistika [1] – maximalizace vzdálenosti od diagonály	$\max(se + sp)$	• W. J. Youden (1950) "Index for rating diagnostic tests". Cancer, 3, 32–35. • R-kový balík pROC • http://www.medicalbiostatistics.com/roccurve.pdf
2. Nejbližší bod levému hornímu rohu grafu	$\min((1-se)^2+(1-sp)^2)$	• R-kový balík pROC • http://www.medicalbiostatistics.com/roccurve.pdf
3. Maximalizace součinu senzitivity a specificity	$\max(se * sp)$	• R-kový balík OptimalCutpoints • dr. Budíková používá maximalizaci geometrického průměru sens a spec

[1] Youdenova J statistika je definována jako: $J = se + sp - 1$; při hledání maxima lze ale člen (-1) zanedbat

Hledání cut-off – vážená kritéria (dle R balíku pROC)

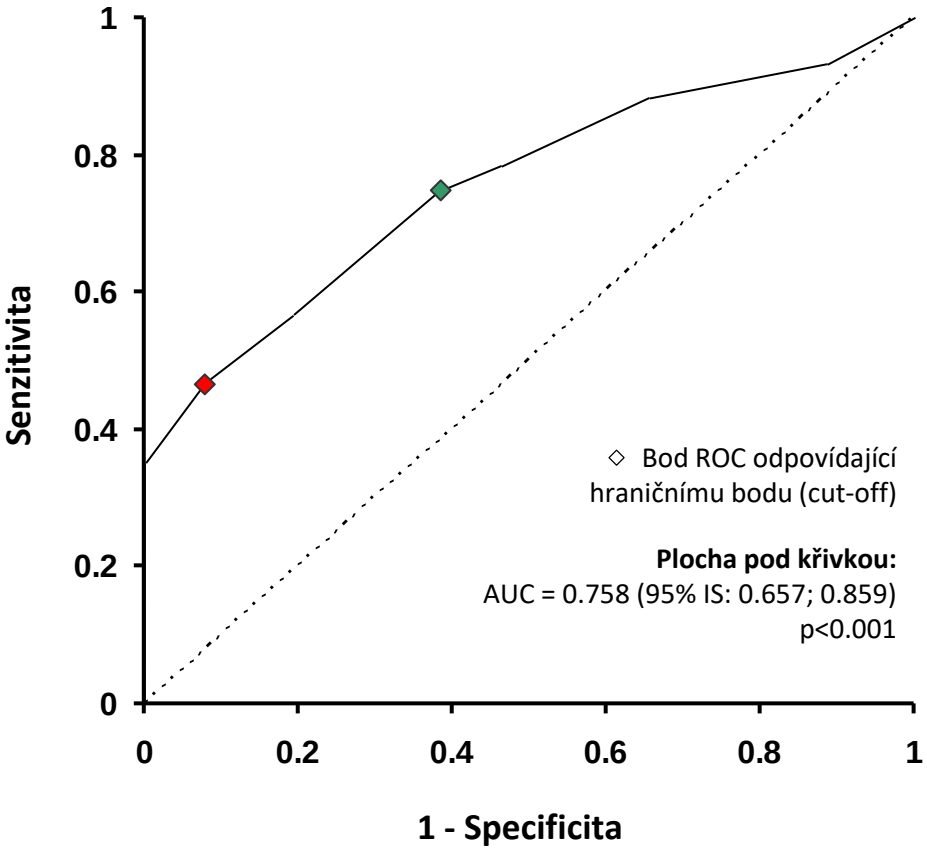
Kritérium	Vzoreček
Youdenova J statistika [1] – maximalizace vzdálenosti od diagonály	$\max(se + r * sp)$
Nejbližší bod levému hornímu rohu grafu	$\min((1 - se)^2 + r * (1 - sp)^2)$

kde: $r = \frac{1 - prevalence}{cost * prevalence}$ $prevalence = \frac{n_{cases}}{n_{cases} + n_{controls}}$

$cost$ – penalizace falešně negativních výsledků

defaultně: $prevalence = 0{,}5$ a $cost = 1$

Příklad - pokračování



Sens	Spec	Sens+ Spec	closest. topleft	Sens* Spec
1.000	0.000	1.000	1.000	0.000
0.933	0.115	1.049	0.787	0.108
0.883	0.346	1.229	0.441	0.306
0.783	0.538	1.322	0.260	0.422
0.750	0.615	1.365	0.210	0.462
0.567	0.808	1.374	0.225	0.458
0.467	0.923	1.390	0.290	0.431
0.350	1.000	1.350	0.423	0.350
0.217	1.000	1.217	0.614	0.217
0.150	1.000	1.150	0.723	0.150
0.050	1.000	1.050	0.903	0.050
0.033	1.000	1.033	0.934	0.033
0.000	1.000	1.000	1.000	0.000

Poděkování…

Příprava výukových materiálů předmětu „DSAN01 Analýza dat pro Neurovědy “ byla finančně podporována prostředky projektu FRVŠ č. 942/2013 „Inovace materiálů pro interaktivní výuku a samostudium předmětu Analýza dat pro Neurovědy“