

Seose tugevuse mõõtmise II

$2 \times k$ ja $m \times k$ sagedustabelid
AUC, Entroopiakordaja

Meenutuseks – seose tugevus, 2×2 tabelid

	süda haige	süda terve	Kokku
Muud õlid	1 971	397 715	399 686
Oliiviõli	3 516	892 123	895 639
Kokku	5 487	1 289 838	1 295 325

$$\widehat{OR} = \frac{1\,971 \cdot 892\,123}{397\,715 / 3\,516} = 0,803 \dots$$

$$\widehat{OR} \cdot \exp\left(\pm 1,96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

$$RD = p_C - p_T$$

$$\widehat{RD} \pm 1,96 \sqrt{\widehat{p}_C(1 - \widehat{p}_C)/n_C + \widehat{p}_T(1 - \widehat{p}_T)/n_T}$$

$$RR = p_T / p_C$$

$$\widehat{RR} \cdot \exp\left(\pm 1,96 \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}\right)$$

$$RRR = (p_C - p_T) / p_C$$

$$NNT = 1 / (p_C - p_T)$$

2

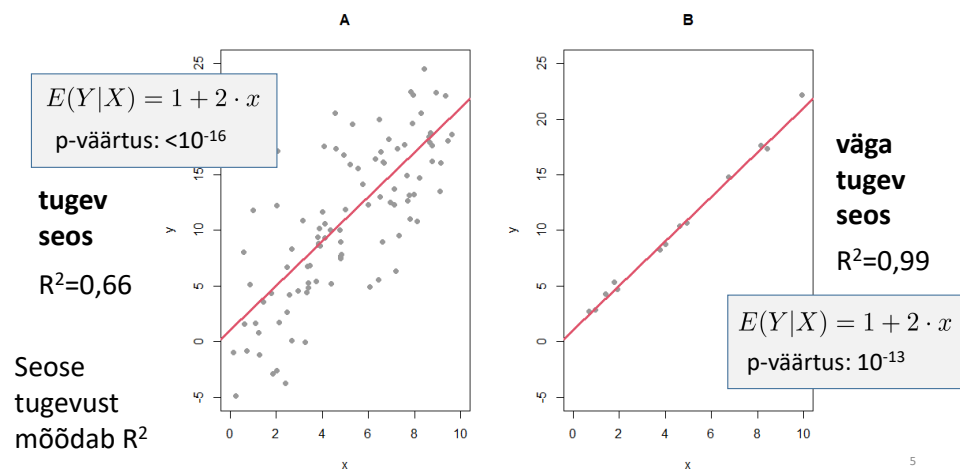
Meenutuseks – seose tugevus, $2 \times k$ tabelid

	süda haige	süda terve	Kokku
Tarbitud oliiviõli kogus päevas			
Muud õlid	1 971	397 715	399 686
0...1	2 658	638 583	641 241
1...1,5	367	106 313	106 680
1,5...	491	150 743	151 234
Kokku	5 487	1 289 838	1 295 325

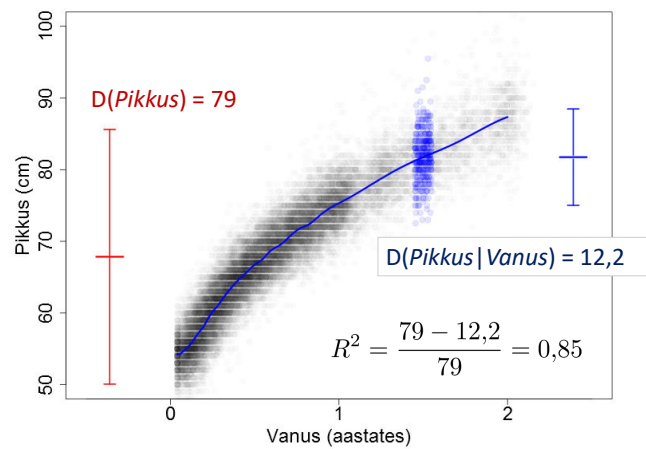
	OR	95%-usaldusintervall
ei tarbi	1	
0...1	0,84	0,79 ... 0,89
1...1,5	0,70	0,62 ... 0,78
1,5....	0,66	0,60 ... 0,73

3

Meenutuseks – seose tugevus, pidevad tunnused



5



11

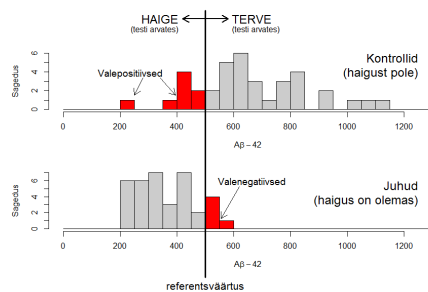
Kuidas mõõta prognoositäpsust kvalitatiivsete tunnuste puhul?

- Dispersiooni kasutada ei saa...
- Alustame kõige lihtsamast juhust – kui prognoositaval tunnusel on kaks võimalikku väärtust (suri/ei surnud; haige/terve; ...)

Tundlikkus (Sensitivity) ja spetsiifilisus (Specificity)

Testi **spetsiifilisus** (Specificity) on tõenäosus, et test annab tervel negatiivse tulemuse (loeb terve terveks)

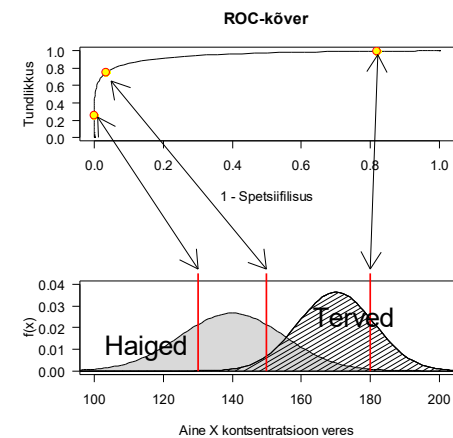
$$\text{spetsiifilisus} := P(T- | H-)$$



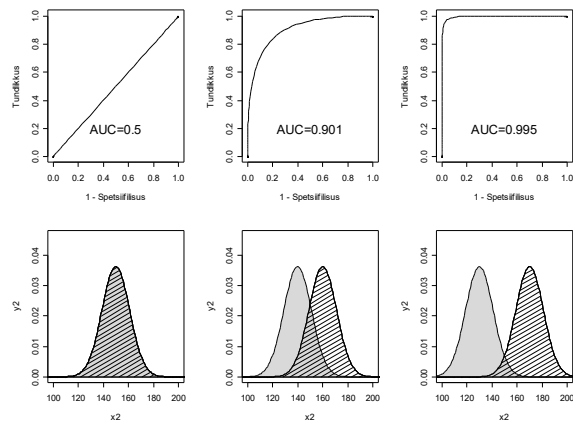
$$\text{tundlikkus} := P(T+ | H+)$$

Testi **tundlikkus** (sensitivity) on tõenäosus, et test annab testitava haiguse või seisundi olemasolul positiivse tulemuse:

13

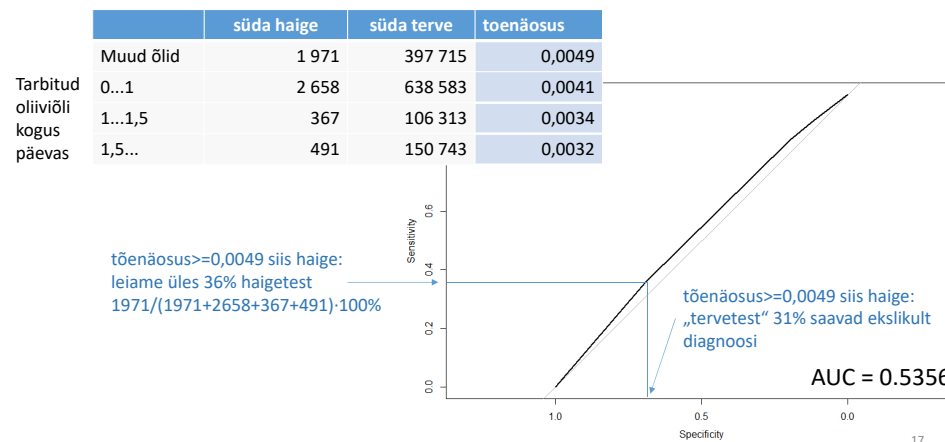


14



15

Rohkem kui kaks töötlust (2xk sagedustabelid)



17

Kuidas iseloomustada seose tugevust suuremate tabelite korral?

X \ Y	A	B	C	D
1	12	23	9	120
2	11	33	8	130
3	10	34	8	100

multiclass ROC: võta üks tulemus (näiteks A), prognoosi teda kõigi vastu -> saad AUC_A ; siis leiad AUC_B , AUC_C jne. Leiad nende keskmise

multiclass ROC: võta üks tulemus (näiteks A), üritad eristada vastuseid A ja B, saad AUC_{AB} , siis vaatad kui hästi suudad eristada A-d ja C-d, saad AUC_{AC} jne. Leiad saadud AUC väärtuste keskmise...

NB! Keskmist on võimalik leida mitut moodi!

Kuni multiclass AUC/multiclass ROC-iga selgus saabub, kasutame hoopis entroopiakordajat...

Mis on (Shanoni) entroopia?

Entroopia

juhusliku suuruse Shannoni entroopia

$$H(Y) = E\left(\log_2\left(\frac{1}{P(Y)}\right)\right) = \sum_i p_i \log_2\left(\frac{1}{p_i}\right) = -\sum_i p_i \log_2(p_i)$$

y_i	Kull	Kiri
$P(Y=y_i)$	0,5	0,5
	0	1

$$H(Y) = \sum_i p_i \log_2\left(\frac{1}{p_i}\right) = 0.5 \cdot \log_2(2) + 0.5 \cdot \log_2(2) = 1$$

Sellise juhusliku suuruse väärtuse edastamiseks läheb keskmiselt tarvis ühte bitti

Entroopia

juhusliku suuruse Shannoni entroopia

$$H(Y) = E \left(\log_2 \left(\frac{1}{P(Y)} \right) \right) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = - \sum_i^J p_i \log_2 (p_i)$$

y_i	A
$P(Y=y_i)$	1

$$H(Y) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = 1 \cdot \log_2(1) = 0$$

Sellise juhusliku suuruse ühe realiseerimise edastamiseks pole tarvis ühtegi bitti (teame niigi, mis tulemuseks oli)

Entroopia

juhusliku suuruse Shannoni entroopia

$$H(Y) = E \left(\log_2 \left(\frac{1}{P(Y)} \right) \right) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = - \sum_i^J p_i \log_2 (p_i)$$

x_i	A	B	C	D
$P(X=x_i)$	0,25	0,25	0,25	0,25
	00	01	10	11

$$H(Y) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = 0,25 \cdot \log_2(4) + \dots + 0,25 \cdot \log_2(4) = 2$$

Sellise juhusliku suuruse ühe realiseerimise edastamiseks vajame keskmiselt 2 bitti

Entroopia

juhusliku suuruse Shannoni entroopia

$$H(Y) = E \left(\log_2 \left(\frac{1}{P(Y)} \right) \right) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = - \sum_i^J p_i \log_2 (p_i)$$

x_i	A	B	C
$P(X=x_i)$	0,5	0,25	0,25
	0	10	11

$$E(pikkus) = 1 \cdot 0,5 + 2 \cdot 0,25 + 2 \cdot 0,25 = 1,5$$

$$\begin{aligned} H(Y) &= \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = 0,5 \cdot \log_2(2) + 0,25 \cdot \log_2(4) + 0,25 \cdot \log_2(4) \\ &= 0,5 \cdot 1 + 0,25 \cdot 2 + 0,25 \cdot 2 \\ &= 1,5 \end{aligned}$$

Sellise juhusliku suuruse ühe realiseerimise edastamiseks vajame keskmiselt 1,5 bitti

Entroopia

juhusliku suuruse Shannoni entroopia

$$H(Y) = E \left(\log_2 \left(\frac{1}{P(Y)} \right) \right) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = - \sum_i^J p_i \log_2 (p_i)$$

x_i	A	B	C
$P(X=x_i)$	0,5	0,5	0

$$\begin{aligned} H(Y) &= \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = 0,5 \cdot \log_2(2) + 0,5 \cdot \log_2(2) + 0 \cdot \log_2(0) \\ &= 1 \end{aligned}$$
$$\lim_{\varepsilon \rightarrow 0+} \varepsilon \log(\varepsilon) = 0$$

Tõenäosusega 0 esineva väärtuse C lisamine jaotusele ei muuda juhusliku suuruse entroopiat.

Entroopia

juhusliku suuruse Shannoni entroopia

$$H(Y) = E \left(\log_2 \left(\frac{1}{P(Y)} \right) \right) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = - \sum_i^J p_i \log_2 (p_i)$$

Märkus:

Mitte alati ei mõõdetata informatsioonisisaldust bittides. Informatsiooni võib mõõta ka baitides, kümnendnumbrites jne. Eri valdkondes võib olla kasutusel erinevad mõõtühikud informatsiooni mõõtmiseks!

$$H(Y) = - \sum_i^J p_i \log_{10} (p_i)$$

mõõdame informatsiooni vajalike kümnendnumbrite arvu kasutades

$$H(Y) = - \sum_i^J p_i \log_{\exp(1,38 \cdot 10^{-23})} (p_i)$$

Füüsikute entroopia

$$H(Y) = - \sum_i^J p_i \ln (p_i)$$

statistikute ja tõenäosusteoretikute lemmik

Seosekordajate defineerimise seisukohast pole tähtis, millist mõõtühikut kasutatakse informatsiooni mõõtmiseks!

Entroopia - omadused

juhusliku suuruse Shannoni entroopia

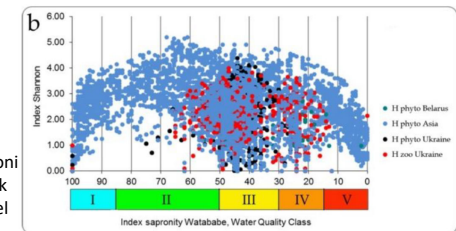
$$H(Y) = E \left(\log_2 \left(\frac{1}{P(Y)} \right) \right) = \sum_i^J p_i \log_2 \left(\frac{1}{p_i} \right) = - \sum_i^J p_i \log_2 (p_i)$$

$$0 \leq H(Y) \leq \log_2(J)$$

Kui leidub selline väärtus y_j , et $P(Y=y_j)=1$

Kui iga $j=1 \dots J$ korral $P(Y=y_j)=1/J$

Shannoni entroopia (siin kui Shannoni indeks H) kasutamine elurikkuse ehk looduse mitmekesisuse kirjeldamisel



Tinglik entroopia

$$H(Y) - H(Y|X)$$

Näitab kui palju informatsiooni Y-tunnuse kohta on olemas X-tunnuses

Ühisjaotus:

X \ Y	A	B	C
1	0,25	0,125	0,125
2	0,25	0,25	0

Y-i marginaaljaotus: $H(Y) = -0,5 \log_2(0,5) - 0,375 \log_2(0,375) - 0,125 \log_2(0,125) = 1,4056$

y_i	A	B	C
$P(Y=y_i)$	0,5	0,375	0,125

Y-i tinglikud jaotused

y_i	A	B	C
$P(Y=y_i X=1)$	0,5	0,25	0,25
$P(Y=y_i X=2)$	0,5	0,5	0

$$H(Y|X=1) = 1,5$$

$$H(Y|X=2) = 1$$

$$H(Y|X) = H(Y|X=1)P(X=1) + H(Y|X=2)P(X=2) = 1,5 \cdot 0,5 + 1 \cdot 0,5 = 1,25$$

Nominaaltunnuste sõltuvuskordaja

Theil'i U ; *Uncertainty coefficient*

$$U(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)}$$

Näitab „säätatud“ bittide osakaalu kui peame edastama juhusliku suuruse Y-tunnuse väärtuseid, aga teame eelnevalt X-tunnuse väärtuseid.

X \ Y	A	B	C
1	0,25	0,125	0,125
2	0,25	0,25	0

$$H(Y) = 1,4056$$

$$H(Y|X) = 1,25$$

$$U(Y|X) = \frac{1,4056 - 1,25}{1,4056} \approx 0,11$$

Teades X-tunnuse väärtuseid suudaksime kokku hoida iga 9. biti Y-tunnuse väärtuseid edastades (andmete edastamiskiirus tõuseks 11%)

Nominaaltunnuste sõltuvuskordaja

Theil'i U ; *Uncertainty coefficient*

$$U(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)}$$

Näitab „säätetud“ bittide osakaalu kui peame edastama juhusliku suuruse Y -tunnuse väärtuseid, aga teame eelnevalt X -tunnuse väärtuseid.

$X \setminus Y$	A	B	C
1	0,5	0	0
2	0	0,25	0,25

$$H(Y) = 1,5$$

$$H(Y|X) = 0,5$$

$$U(Y|X) = \frac{1,5 - 0,5}{1,5} \approx 0,666 \dots$$

$$U(X|Y) = ?$$

Teades X -tunnuse väärtuseid suudaksime andmeid edastada 67% vähemate bittide abil (andmete edastamiskiirus tõuseks 67%)

Nominaaltunnuste sõltuvuskordaja

Theil'i U ; *Uncertainty coefficient*

$$0 \leq U(Y|X) \leq 1$$

Kui X ja Y on sõltumatud

Kui Y on X -i järgi täpselt leitav

Ei sõltu valitud informatsiooni mõõtühikust

Hinnatav läbilõikelise uuringu korral – teiste uuringuplaanide korral pole juhuslike suuruste ühisjaotust võimalik hinnata!

$$U(Y|X) \neq U(X|Y)$$

$X \setminus Y$	A	B	C	D
1	0,25	0,25	0	0
2	0	0	0,25	0,25

$$U(Y|X) = 0,5$$

$$U(X|Y) = 1$$