TUNNUSPAARI ANALÜÜS

Seoste määramine

	1. Tööga rahul		*2. Pole tööga rahul*		*Kokku*	
Amet	% N1		%	N2	%	N
1. Arst	75	30	25	10	100	40
2. Med.õde	50	15	50	15	100	30
3. Abipersonal	20	6	80	24	100	30
Kokku		51		49		100

	1.Tööga rahul		*2. Pole tööga rahul*		*Kokku*	
Amet	% N1		%	N2	%	N
1. Arst	52	21	48	19	100	40
2.Med.õde	52	16	48	14	100	30
3.Abipersonal	52	16	48	14	100	30
Kokku		53		47		100

2 tunnust on omavahel seotud, kui ühe tunnuse väärtuste jaotus teise tunnuse eri väärtuste korral on erinev.

1. **Lähtumine 2 tunnuse sõltumatusest**
2. **Alamgruppide erinevuste hindamine**
3. **Tunnuste väätuspaaride võrdlemine**
4. **Vea proportsionaalne vähendamine**
5. **Korrelatsioon**

1. Lähtumine kahe tunnuse sõltumatusest

Kahte tunnust nimetatakse <u>sõltumatuteks</u>, kui ühe tunnuse konkreetse väärtuse teadmine ei anna mingit infot teise tunnuse väärtuse ennustamiseks.

$$P(X,Y) = P(X)*P(Y)$$

$$(P_{ij} = P_{i*} * P_{*j},\ i=1,\ldots,K_{T1},\ j=1,\ldots,K_{T2})$$

<u>Statistiline sõltuvus</u> leiab aset parajasti siis, kui tunnused ei ole sõltumatud.

***Lähtealus*: millised peaksid olema andmed, et puuduks seos tunnuste vahel (statistiliselt sõltumatud).**

$$\chi^2 = {}_k\Sigma\,[(f_t - f_o)^2 / f_o]\ ,\ \text{kus}$$

χ^2 - hii-ruut,
f_t - väärtuskategooria tegelik esinemissagedus,
f_o - väärtuskategooria oodatud esinemissagedus.

***Näide*. Töötajate eelistus palgamaksmisskeemile. Kas mingi skeem on eelistatud?**

Igal nädalal	2 korda kuus	1 kord kuus	N	
40	110	60	210	(f_t)
70	70	70		(f_o)

$$\chi^2 = (40-70)^2/70 + (110-70)^2/70 + (60-70)^2/70 = 37{,}1$$

Hii-ruudul põhinevad statistikud

1) *Tshuprovi seosekordaja*

$T_{AB} = \sqrt{\chi^2 / [N \sqrt{(K_A-1)(K_B-1)}]}$, kus

N - vaatluste arv,
K_A, K_B - tunnuste A ja B väärtusklasside arv.

$0 \leq T \leq 1$,

T=0 - empiiriliselt sõltumatute tunnuste korral,
T=1 - täieliku statistilise sõltuvuse korral.

T väärtus sõltub väärtusklasside arvust:

min (K_A,K_B)	2	3	5	8	10	15	50
max T	1	0,84	0,71	0,61	0,58	0,52	0,38

2) *Crameri seosekordaja*

$C_{AB} = \sqrt{\chi^2 / \{N[\min(K_A,K_B)-1]\}}$,

$C_{AB} = T_{AB} \sqrt{[\max(K_A,K_B)-1] / [\min(K_A,K_B)-1]}$

3) *Pearsoni seosekordaja*

$P_{AB} = \sqrt{\chi^2 / (\chi^2 + N)}$

P väärtus sõltub väärtusklasside arvust:

min (K_A,K_B)	2	3	5	10	20	50	100
max P	0,71	0,82	0,89	0,95	0,97	0,99	0,995

2. Alamgruppide erinevuste hindamine

Seost mõõdetakse alamgruppide proportsioonide omavahelisel võrdlemisel.

Näide.

Tööhõive	Omab diplomit	Ei oma diplomit
Töötab	87%	38%
Töötu	13%	62%
Kokku	100%	100%

-> haridus tingib tööhõive

3) Väärtuspaaride võrdlemine

Võrreldakse omavahel kõiki objektipaare. Seose mõõduks on ühe väärtuspaaritüübi ülekaal.

Omadus A	Omadus B 1. Jah	Omadus B 2. Ei	Marginaal-summa
1. Jah	a	b	a+b
2. Ei	c	d	c+d
	a+c	b+d	N

<u>Positiivne seos</u> - kui paari üks liige (objekt) omab tunnust A ja B, samal ajal kui teine liige ei oma kumbagi.

<u>Negatiivne seos</u> - kui paari üks liige omab tunnust A ja ei oma tunnust B, kui paari teine liige omab tunnust B ja ei oma tunnust A.

Yule Q = (ad-bc)/(ad+bc)

$-1 \leq Q \leq 1$,

Q=0, kui seos puudub,
Q=1, kui täielik positiivne seos,
Q=-1, kui täielik negatiivne seos.

Näide.

Täidavad ühiskondlikke ülesandeid

	1. Jah	2. Ei	Kokku
1. Õpivad	89	33	122
2. Ei õpi	50	48	98
Kokku	139	81	220

Q = (89*48-50*33)/(89*48+50*33)=
=2622/5922=0,422

K_AxK_B andmetabelitel (K_A või K_B suurem kui 2)

Yule γ = (P-N)/(P+N)

Näide.

Suhtumine töösse

Eksamihinne	1. Kõrge	2. Keskmine	3. Madal	Kok-ku
1. Kõrge	12	6	2	20
2. Keskmine	8	10	2	20
3. Madal	0	4	16	20
Kokku	20	20	20	60

P=812, N=100; γ = (812-100)/(812+100) = 0,78.

Eksitakse 50(1- γ) % juhtudel (50(1- 0,78)=11%).

Yule V = (ad-bc)/ √ (a+b)(a+c)(b+d)(c+d).
V = (89*48-33*50)/√122*139*81*98 =
= 2622/11601= 0,226.

4. Vea proportsionaalne vähendamine

Mil määral info tunnuse B kohta aitab ennustada tunnuse A väärtust e. palju info B kohta proportsionaalselt vôimalikku viga A ennustamisel vähendab?

Lähtutakse: **A ennustamine esmalt kasutamata infot B kohta, seejärel hinnatakse palju ennustamisel viga väheneb, kui kasutame infot B kohta.**

Goodmann'i g, λ, $0 \leq g \leq 1$

Näide.
Rügemendis 63 sôjaväelast. Valime juhuslikult ühe nende seast. Kas ta on ohvitser vôi allohvitser?

Auaste

Haridus	1. Ohvitser	2. Allohvitser	Kokku
1.Erakooli	25	2	27
2.Riigikooli	5	31	36
Kokku	30	33	63

allohvitser 33/63, eksime 30/63.
Lisainfo: ta on lôpetanud riigikooli.
allohvitser 31/36, eksime 7/63,
viga vähenes 23/63 vôrra.

Teades B, ennustame A:

$g_{A/B}$ = (Σsuurimad veerusagedused - suurim reasumma)/(N- suurim reasumma)

$g_{A/B}$ = ((25+31)-36)/(63-36)= 0,74,
teades auastet, viga hariduse ennustamisel väheneb 74%.

Teades A, ennustame B:

$g_{B/A}$ = (Σsuurimad reagedused - suurim veerusumma)/(N- suurim veerusumma)

$g_{B/A}$ = ((25+31)-36)/(63-33)= 0,77,
teades haridust, viga auastme ennustamisel väheneb 77%.

<u>asümmeetriline seosekordaja</u>

5. Korrelatsioon

Funktsionaalne seos - ühe tunnuse kindlale väärtusele vastab teise tunnuse kindel väärtus
<u>Korrelatiivne seos</u> - ühe tunnuse mingile väärtusele vastab teise tunnuse mitu erinevat väärtust.

<u>sümmeetriline seosekordaja</u>

1) Lineaarne (Pearson'i) **paariskorrelatsioonikordaja**

$$r = \Sigma [(X_i-M_X)(Y_i-M_Y)]/ \sqrt{N*\sigma_X*\sigma_Y},$$

$$-1 \leq r \leq 1.$$

$r = [N*\Sigma XY - (\Sigma X)(\Sigma Y)] / \sqrt{[N*\Sigma X^2 - (\Sigma X)^2]*[N*\Sigma Y^2 - (\Sigma Y)^2]}$,
kus summeeritakse üle objektide, i=1,…,N.

Näide.

X (Tööpinge)	**43**	**55**	**67**	**38**	**49**	**70**	**80**	**62**	**73**	**83**
Y (Tööga rahulolu)	**4**	**5**	**6**	**4**	**5**	**7**	**9**	**5**	**6**	**9**

$N=10$, $\Sigma X=620$, $\Sigma Y= 60$, $\Sigma X^2= 40570$, $\Sigma Y^2= 390$, $\Sigma XY= 3951$.

$r = (10*3591-620*60)/\sqrt{(10*40570-620^2)(10*390-60^2)}$
$= 2310/2528 = 0{,}91$.

2) Järjekorrelatsioonikordajad

- Spearman'i järjekorrelatsioonikordaja

$r_S = 1- (6*\Sigma d^2)/(N^3-N)$, kus

d - järjenumbrite hälve.

Näide.

Traktor	Tähtsuse järgi	Kasutamise sageduse järgi	d	d^2
A	10	8	2	4
B	7	9	-2	4
C	4	4	0	0
D	1	1	0	0
E	3	5	-2	4
F	2	2	0	0
G	9	10	-1	1
H	5	6	-1	1
I	8	7	1	1
J	6	3	3	9
			Σ	24

- Kendalli järjekorrelatsioonikordaja

$\tau = 2*\Sigma Z_i/[0,5(N-1)N] - 1$, kus

Z_i - teise tunnuse järjenumbrite arv, alates (i+1)st, mille väärtus on suurem tema (teise tunnuse) i-ndast järjenumbrist, i=1,2,...,N.

Näide.

Traktor	Tähtsuse järgi	Kasutamise sageduse järgi	Zi
D	1	1	9
F	2	2	8
E	3	5	5
C	4	4	5
H	5	6	4
J	6	3	4
B	7	9	1
I	8	7	2
G	9	10	0
A	10	8	0
		Σ	38

$\tau = 2*38/45 - 1 = 0,68.$

- Informatsiooniline (Linfoot'i) korrelatsiooni-kordaja

$L = 1 - 2^{-2*I}$, kus

I = H(A) + H(B) - H(A,B), kus I tähistab tunnuste A ja B vastastikust informatsiooni (kui palju infot annab ühe tunnuse väärtuse teadmine teise tunnuse suhtes).

$H(A) = -\ {}_i\Sigma p_i \log_2 p_i$ $\quad H(B) = -\ {}_j\Sigma p_j \log_2 p_j$

$H(A,B) = -\ {}_i\Sigma\ {}_j\Sigma p_{ij} \log_2 p_{ij}$, kus

$p_i = N_{i*}/N$, $\quad p_j = N_{j*}/N$, $\quad p_{ij} = N_{ij}/N$, kus

$i=1,\ldots,K_A$, $\quad j=1,\ldots,K_B$

$0 \leq L \leq 1$

L = I = 0, kui H(A,B) = H(A) + H(B)

I -> max, kui H(A) = H(B) = H(A,B)
($I \leq \min(H(A), H(B))$)

min (K_A,K_B)	2	3	4	5	10	20
max I	1	1,58	2	2,32	3,32	4,32
max L	0,867	0,94	0,968	0,98	0,995	0,999