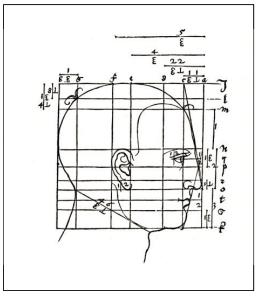
Metodologija psiholoških istraživanja

1

obrada frekvencijski 5



20. novembar 2018

IV. Obrada podataka

A. Frekvencijski nacrti

1. Univarijatni frekvencijski nacrti
2. Bivarijatni frekvencijski nacrti
3. Trivarijatni frekvencijski nacrti

2. Bivarijatni frekvencijski nacrti (BFN)

2

'svakodnevna korelacija'

- često se čuju tvrdnje koje imaju sledeći opšti oblik:
  - 'osobe koje spadaju u kategoriju X imaju osobinu Y'
  - *PRIMERI:* 'žene su loši vozači', 'bikovi su agresivni', 'crnci su muzikalni', ...
- uočimo: ovakve tvrdnje su *generalne*, *komparativne*, *korelativne* i *prediktivne*
  - *generalnost:* tvrdnja je *opšta*, o grupama osoba (o ženama, bikovima, crncima itd)
  - *komparativnost: porede* se dve grupe (žene i muškarci, bikovi i ne-bikovi, itd.)
  - *korelativnost:* tvrdi se *korelacija* između pripadnosti grupi i osobine
    - povezanost pola i umešnosti vožnje, zodijačkog znaka i ponašanja, rase i muzikalnosti ...
  - *prediktivnost:* tvrdi se da se na osnovu pripadnosti kategoriji X može *predvideti* odn. proceniti prisustvo osobine Y
    - na osnovu pola se može proceniti umešnost vožnje, na osnovu zod. znaka ličnost, itd.
- kako odlučujemo da li su takve tvrdnje *istinite*?
- čest način: pozivanje na *pojedinačne* slučajeve koji *potvrđuju* opštu tvrdnju
  - navode se *primeri* žena loših vozača, agresivnih bikova, muzikalnih crnaca itd.
- međutim: ovakvo zaključivanje je logički i statistički pogrešno!
  - naime: opšte tvrdnje ovog tipa se *ne mogu* dokazati (niti pobiti) samo navođenjem potvrđujućih (ili opovrgavajućih) *pojedinačnih* slučajeva (odn. *anegdota*)

2. Bivarijatni frekvencijski nacrti (BFN)

3

- ovakve tvrdnje se mogu tretirati kao BFN tipa 2x2
- PRIMER: 'žene su loši vozači'
  - ova tvrdnja ne znači: 'postoje žene koje su loši vozači'
    - naime, postoje loši vozači i među muškarcima
  - ova tvrdnja (bi trebalo da) znači: 'među ženama ima više loših vozača nego među muškarcima'

	loši voz.	dobri voz.
žene	a	b
muš.	c	d

- uočiti: navođenjem primera koji *potvrđuju* tvrdnju, samo se ustanovljava da *postoje* slučajevi koji spadaju u situaciju a u gornjoj tabeli
- međutim: korelacija se može ispitati *samo* ako su poznate a, b, c i d
- naime: 'potvrđujući' primeri će postojati ne samo kada korelacija postoji ($\phi > 0$), već i kada je nulta ($\phi = 0$), ili čak *obrnuta* od tvrđene ($\phi < 0$)!

$\phi > 0$	loši voz.	dobri voz.
žene	70	30
muš.	30	70

$\phi = 0$	loši voz.	dobri voz.
žene	50	50
muš.	50	50

$\phi < 0$	loši voz.	dobri voz.
žene	30	70
muš.	70	30

- takođe: 'kontraprimeri' će postojati čak i kada korelacija zaista postoji
  - npr., dokazano je da postoji korelacija pušenja i raka
  - ali, ipak će postojati i pušači koji *ne* obolevaju (situacija c)
  - korelacija ne znači da će *svaki* pušač nužno dobiti rak

	puš.	nep.
oboleli	a	b
neoboleli	c	d

2. Bivarijatni frekvencijski nacrti (BFN)

4

(2) Nacrti složeniji od tipa 2x2

nacrt tipa 3x3, primer A

	rok	ozb.	nar.	UZR.
mladi	250	100	150	500
sred.	200	80	120	400
stari	50	20	30	100
MUZ.	500	200	300	1000

apstraktni prikaz rezultata nacrta 3x3

	a_1	a_2	a_3	B
b_1	a	b	c	$fb_1 = a+b+c$
b_2	d	e	f	$fb_2 = d+e+f$
b_3	g	h	i	$fb_3 = g+h+i$
A	$fa_1 = a+d+g$	$fa_2 = b+e+h$	$fa_3 = c+f+i$	$N = a+b+ ... +h+i$

primer A, procentualni prikaz. 3. način

	rok	ozb.	nar.	UZR.
mladi	50%	20%	30%	100%
sred.	50%	20%	30%	100%
stari	50%	20%	30%	100%
MUZ.	50%	20%	30%	100%

primer B, frekvence i procenti

	rok	ozb.	nar.	UZRAST
mladi	400 (80%)	50 (10%)	50 (10%)	500 (100%)
sred.	100 (25%)	100 (25%)	200 (50%)	400 (100%)
stari	0 (0%)	50 (50%)	50 (50%)	100 (100%)
MUZ.	500	200	300	1000

- kada će postojati odn. kada *neće* postojati *korelacija* između dve varijable?
  - nepostojanje korelacije: *jednaki* odnosi frekvenci, (kao i proporcija i procenata)
  - postojanje korelacije: *nejednaki* odnosi frekvenci (kao i proporcija i procenata)
    - primer A: svi odnosi su 5:2:3, nema korelacije, isti odnos prema muzici kod svih
    - primer B: odnosi su 8:1:1, 1:1:2, 0:1:1, ima korelacije, različit odnos
- nepostojanje: jednaki *profili* podataka; postojanje: nejednaki profili podataka

2. Bivarijatni frekvencijski nacrti (BFN)

5

e. Značajnost rezultata

- dve vrste testova značajnosti u BFN: 1D i 2D
  - (1) jednodimenzionalni (1D) testovi (1D nulte hipoteze)
    - testiranje značajnosti u 1D matricama (prostim i glavnim)
    - *PRIMER:* bivarijatni nacrt tipa 2x2 ...

AB	puš.	nep.	POL
muš.	90	60	150
žene	30	20	50
PUŠ.	120	80	200

    - ... sadrži 6 univarijatnih nacrta, za koje se može testirati značajnost

A/b1	puš.	nep.	
muš.	90	60	150

A/b2	puš.	nep.	
žene	30	20	50

A	puš.	nep.	
PUŠ.	120	80	200

B/a1	puš.
muš.	90
žene	30
PUŠ.	120

B/a2	nep.
muš.	60
žene	20
PUŠ.	80

B	POL
muš.	150
žene	50
PUŠ.	200

- ovakvi testovi se vrše prema ranije izloženim principima za UFN
- njima se dalje nećemo detaljnije baviti

2. Bivarijatni frekvencijski nacrti (BFN)

6

- (2) dvodimenzionalni (2D) testovi (2D nulte hipoteze)
  - 2D testovi odnose se na postojanje *korelacije* između dve varijable
- (1) Nacrti tipa 2x2
  - korelacija postoji (odn. ne postoji) ako su šanse različite (odn. iste)
    - međutim: vrlo retko će biti a/b biti identično sa c/d, tako da bude $\phi = 0$
    - skoro uvek će postojati numerički *nenulta* korelacija, tj. $\phi \neq 0$
  - ključno pitanje: da li je dobijena korelacija statistički značajna?
    - postoji u *uzorku*, ali da li postoji i u populaciji?
  - testiranje značajnosti se može podeliti na istih 5 faza kao u UFN
- *Faza I: Podaci i deskriptivne mere*
  - utvrđivanje frekvenci situacija i kategorija, i totalne frekvence
  - izračunavanje nekog pokazatelja korelacije (najčešće ϕ)
  - *PRIMER:*
    - *istraživačko pitanje:* da li su su pol i pušenje u populaciji korelirani?
    - *radna hipoteza:* pol i pušenje su korelirani

$\phi = 0.41$	puš.	nep.	POL	šanse
muš.	80	40	120	2:1
žene	20	60	80	1:3
PUŠ.	100	100	200	1:1

2. Bivarijatni frekvencijski nacrti (BFN) 7

- *Faza II: nulta hipoteza, očekivane vrednosti, devijacije*
- postavljanje statističke hipoteze
  - H0 je 2D: u populaciji nema korelacije dve varijable (pola i pušenja)
    - nenulti ϕ u uzorku je posledica slučaja
- utvrđivanje *očekivanih (teoretskih) frekvenci*
  - očekiv. frekvence: najverovatnije frekvence u uzorku, u slučaju da je H0 tačna
  - utvrđivanje očekiv. frekvenci kod BFN je nešto složeniji problem nego kod UFN
- podsetimo se: utvrđivanje očekivanih frekvenci, f', za 1D H0, kod UFN
  - UFN sa dve kategorije: f1' = f2' = N/2
    - dakle: *očekivane* frekvence pojedinih kategorija, f1' i f2', utvrđuju se na osnovu (tj. s obzirom) na *opserviranu totalnu* frekvencu, N
      - naime, dobijaju se delenjem N sa brojem kategorija, tj. 2
    - te vrednosti f1' i f2' predstavljaju najverovatniji ishod u istraživanju sa opserviranom totalnom frekvencom N, ako u populaciji važi H0
  - kod UFN sa tri kategorije važi isti princip: f1' = f2' = f3' = N/3
  - kod UFN sa četiri kategorije važi isti princip : f1' = f2' = f3' = f4' = N/4, itd.
- uočimo: očekiv. frekvence se uvek sabiraju do opservirane totalne frekvence

2. Bivarijatni frekvencijski nacrti (BFN) 8

- očekivane frekvence u BFN: slična ali složenija razmatranja
- pretpostavka: važi 2D H0: u populaciji *ne* postoji korelacija varijabli A i B
- pitanje: kolike bi, najverovatnije, tada bile frekvence situacija u matrici AB?
  - to su *očekivane* frekvence, a izračunavaju se na osnovu *opserviranih* frekvenci
- PRIMER: opservirane frekvence u matrici AB: a=80, b=40, c=20, d=60

opserv.	puš.	nep.	POL
muš.	80	40	120
žene	20	60	80
PUŠ.	100	100	200

oček.	puš.	nep.	POL
muš.	a' = ?	b' = ?	120
žene	c' = ?	d' = ?	80
PUŠ.	100	100	200

- *očekivane* frekvence, a', b', c' i d', utvrđuju se na osnovu datih *marginalnih* frekvenci i *totalne* frekvence, uz određene uslove:
- prvi uslov: očekivane frekvence se moraju sabirati do *marginalnih* frekvenci
  - slično kao što se u UFN sve očekivane frekvence f' moraju sabirati do *N*
  - dakle, u datom primeru mora biti: a'+b'=120, c'+d'=80 (i slično po kolonama)
- drugi uslov: korelacija je nula, tj. su šanse jednake: a'/b' = c'/d' = fa1/fa2
  - uočimo: ovaj odnos je *poznat*, jer su poznate marginalne frekvence fa1 i fa2
  - u datom primeru: a'/b' = c'/d' = fa1/fa2 = 100/100 = 1, tj. a'=b', i c'=d'

2. Bivarijatni frekvencijski nacrti (BFN) 9

- primenjujući ove uslove, računamo očekivane frekvence u datom primeru:

oček.	puš.	nep.	POL	šanse
muš.	a'	b'	120	1:1
žene	c'	d'	80	1:1
PUŠ.	100	100	200	1:1

a'+b'=120
a'=b'
c'+d'=80
c'=d'

oček.	puš.	nep.	POL	šanse
muš.	60	60	120	1:1
žene	40	40	80	1:1
PUŠ.	100	100	200	1:1

- kod muškaraca: koja su dva broja koja se sabiraju do 120, a jednaka su?
  - rešenje: očekivane frekvence su po 60
- kod žena: koja su dva jednaka broja koja se sabiraju do 80?
  - rešenje: očekivane frekvence su po 40
- postoje opšte *algebarske formule* za izračunavanje očekivanih frekvenci

opservirane frekvence: a, b, c, d

	a_1	a_2	B
b_1	a	b	fb_1
b_2	c	d	fb_2
A	fa_1	fa_2	N

očekivane frekvence: a', b', c', d'

	a_1	a_2	B
b_1	a' = (fa1*fb1)/N	b' = (fa2*fb1)/N	fb_1
b_2	c' = (fa1*fb2)/N	d' = (fa2*fb2)/N	fb_2
A	fa_1	fa_2	N

- princip računanja očekivanih frekvenci:
  - za svaku očekivanu frekvencu utvrde se opservirane *marginalne* frekvence u istom *redu* i istoj *koloni*
  - te dve marginalne frekvence se *množe*, a dobijeni proizvod se *deli* sa N

	a_1	a_2	B
b_1	a'	b'	fb_1
b_2	c'	d'	fb_2
A	fa_1	fa_2	N

2. Bivarijatni frekvencijski nacrti (BFN) 10

- *PRIMER:*

opservirane frekvence: a, b, c, d

opser.	puš.	nep.	POL	šanse
muš.	80	40	120	2:1
žene	20	60	80	1:3
PUŠ.	100	100	200	1:1

očekivane frekvence: a', b', c', d'

oček.	puš.	nep.	POL	šanse
muš.	60	60	120	1:1
žene	40	40	80	1:1
PUŠ.	100	100	200	1:1

isto kao kod opser.: marginalne i tot. frekvenca
različito od opser.: šanse su jednake (obe 1:1)

- izračunajmo očekivane frekvence prema formulama:
  - a' = (120*100)/200 = 60; b' = (120*100)/200 = 60;
  - c' = (80*100)/200 = 40; d' = (80*100)/200 = 40
  - uočiti: ovi rezultati su jednaki ranije dobijenim, izvedenim logičkim razmatranjima
- sledeći korak: izračunavanje *reziduala* (d): odstupanja opserv. od očekiv. fr.
  - za svaku ćeliju, od opservirane frekvence oduzima se očekivana frekvenca
  - dobijaju se devijacije odn. *reziduali d:* d_a = a-a', d_b = b-b', d_c = c-c', d_d = d-d'

	a_1	a_2	B
b_1	da = a – a'	db = b – b'	fb_1
b_2	dc = c – c'	dd = d – d'	fb_2
A	fa_1	fa_2	N

	puš.	nep.	B
muš.	80-60 = 20	40-60 = -20	120
žene	20-40 = -20	60-40 = 20	80
A	100	100	200

- uočiti: reziduali se sabiraju do nule, i po redovima i po kolonama

2. Bivarijatni frekvencijski nacrti (BFN) 11

- *Faza III: Test-statistik*
- računa se χ^2, i to po istom principu kao kod UFN
  - kvadrirani d se dele sa odgovarajućim f', i tako dobijene vrednosti se sabiraju

reziduali

d	puš.	nep.
muš.	20	-20
žene	-20	20

očekivane frekvence

f'	puš.	nep.
muš.	60	60
žene	40	40

kvadrirani reziduali

d^2	puš.	nep.
muš.	400	400
žene	400	400

deljenje sa oč. frek.

d^2/f'	puš.	nep.
muš.	6.67	6.67
žene	10.00	10.00

zbir: χ^2 = 33.3

- koliki je broj stepeni slobode (df) za χ^2 u nacrtima tipa 2x2?
  - setimo se: u UFN nacrtima sa 4 kategorije df = 4-1 = 3
  - ali: u nacrtima 2x2, od 4 moguća stepena slobode za 4 ćelije matrice AB, 3 su izgubljena korišćenjem informacije o frekvencama u marginalnim matricama!
  - stoga je vrednost za samo *jednu* ćeliju (bilo koju) 'slobodna'
  - *PRIMER:*
    - neka je poznata samo jedna vrednost u matrici AB, i sve margin. frekvence
    - ostale tri vrednosti u matrici AB se mogu sračunati na osnovu tih informacija

	puš.	nep.	POL
muš.	80	?	120
žene	?	?	80
PUŠ.	100	100	200

	puš.	nep.	POL
muš.	80	40	120
žene	20	60	80
PUŠ.	100	100	200

zaključak: u nacrtima tipa 2x2 važi df = 1

2. Bivarijatni frekvencijski nacrti (BFN) 12

- *Faza IV: p-vrednost*
  - kao i kod UFN, na osnovu stat. teorije, kompjuter izračunava p-vrednost
  - u datom primeru važi: $\chi^2(1) = 33.3$, $p<0.05$
- *Faza V: Odluka o statističkoj značajnosti*
  - rezultat je statistički značajan, postoji korelacija između pola i pušenja
- uočiti: ako je u uzorku korelacija tačno 0:
  - tada važi a/b = c/d, tj. šanse su jednake
  - tada važi da je ϕ = 0, KŠ = 1, i KP1 = KP2
  - važi i f'=f: očekivane vrednosti *jednake su* opserviranim vrednostima
  - tada za sve reziduale važi d = f-f' = 0, pa je i χ^2 = 0

SL. 6	puš.	nep.	POL	šanse
muš.	90	60	150	3:2
žene	30	20	50	3:2
PUŠ.	120	80	200	3:2

- Alternative χ^2 – testu
  - *z-test razlika proporcija:* ista odluka o značajnosti kao i za χ^2-test
  - *Fišerov egzaktni test:* preciznije procene p-vrednosti, ali se ne može koristiti za složenije nacrte od tipa 2x2 (dok χ^2–test može)
- statistička napomena
  - u gornjim formulama za χ^2 nije uzeta u obzir tzv. korekcija za kontinuitet

2. Bivarijatni frekvencijski nacrti (BFN) 13

- alternativna formula za χ^2–test
  - χ^2 se može računati ne samo po ranije navedenoj formuli, nego još na jedan, ekvivalentan način (ako je već izračunato ϕ): $\chi^2 = N^*\phi^2$
  - iz ove alternativne formule slede neke zanimljive statističke posledice
  - gornja jednačina ima oblik A = B*C (pri čemu je A = χ^2, B = N, C = ϕ^2)
    - kada će A biti *veće?* onda kada su B i C veći, pa prema tome:
      - (a) ako je B konstantno, A će biti utoliko veće ukoliko je C veće
        - npr. ako je B=10, jednačina A = B*C glasi: A = 10*C
      - (b) ako je C konstantno (npr. C=10), A će biti utoliko veće ukoliko je B veće
  - dakle, u formuli $\chi^2 = N^*\phi^2$, χ^2 raste kada N i ϕ^2 rastu, pa zaključujemo:
    - (a) ako dva istraživ. imaju istu veličinu uzorka, tj. konstantno N, tada će:
      - ono istraživanje kod koga je korelacija jača (veće ϕ i a time i veće ϕ^2) imati veći χ^2, a time i veću šansu da bude statistički značajno (manja p-vrednost)

ϕ = 0.1	crv.	plv.	POL	šanse
muški	45	55	100	11:9
ženski	55	45	100	9:11
BOJA	100	100	200	1:1

ϕ = 0.1, ϕ^2 = 0.01, χ^2 = 200*0.01=2, p > 0.05

ϕ = 0.4	crv.	plv.	POL	šanse
muški	30	70	100	3:7
ženski	70	30	100	7:3
BOJA	100	100	200	1:1

ϕ = 0.4, ϕ^2 = 0.16, χ^2 = 200*0.16=32, p < 0.05

2. Bivarijatni frekvencijski nacrti (BFN) 14

- (b) ako dva istraživanja imaju istu jačinu korelacije ϕ, tada će:
  - ono istraživanje kod koga je veći uzorak (N) imati veći χ^2, a time i veću šansu da korelacija bude statistički značajna (manja p-vrednost)

ϕ = 0.4	crv.	plv.	POL	šanse
muški	30	70	100	3:7
ženski	70	30	100	7:3
BOJA	100	100	200	1:1

ϕ = 0.4, ϕ^2 = 0.16, χ^2 = 200*0.16=32, p < 0.05

ϕ = 0.4	crv.	plv.	POL	šanse
muški	3	7	10	3:7
ženski	7	3	10	7:3
BOJA	10	10	20	1:1

ϕ = 0.4, ϕ^2 = 0.16, χ^2 = 20*0.16=3.2, p > 0.05

  - ova dva istraživanja imaju iste šanse i istu korelaciju, ϕ = 0.4, ali se razlikuju po veličini uzorka (N1=200, N2=20), i stoga i po statističkoj značajnosti
- iz formule A = B*C se može zaključiti i sledeće:
  - neka je A konstantno; to isto A se može dobiti na (bar) dva načina:
    - B veliko a C malo; B malo a C veliko, što za formulu $\chi^2 = N^*\phi^2$ znači:
  - (c) ako dva istraživanja imaju isti χ^2, postoje dve osnovne mogućnosti:
    - (1) malo N ali veliko ϕ; (2) veliko N ali malo ϕ
  - npr.: neka je χ^2 = 5; ta se vrednost može dobiti ako je npr.:
    - (1) N = 10 i ϕ = 0.71; ili (2) N = 1000 i ϕ = 0.071
    - dakle: N i ϕ mogu se međusobno *kompenzovati* u svom doprinosu za χ^2

2. Bivarijatni frekvencijski nacrti (BFN) 15

detaljne analize dva poučna primera nacrta tipa 2x2

- primer 1: *aspirin i srčani udari*
  - koristićemo okrugle cifre, pojednostavljen prikaz nacrta
  - varijabla A (pilula): E-grupa: aspirin, K-grupa: placebo
  - varijabla B (simptom): DA: imati srčani udar, NE: nemati srčani udar

ops. frek. f	DA	NE	PILULA
placebo	200	9800	10000
aspirin	100	9900	10000
SIMPTOM	300	19700	20000

oč. frek. f'	DA	NE	PILULA
placebo	150	9850	10000
aspirin	150	9850	10000
SIMPTOM	300	19700	20000

reziduali d	DA	NE
placebo	+50	-50
aspirin	-50	+50

količnik d^2/f'	DA	NE
placebo	16.67	0.25
aspirin	16.67	0.25

$\chi^2(1) = 33.84$
$p < 0.00001$
$\phi = 0.041$

- istraživački zaključak: prisustvo srčanog udara korelira sa vrstom pilule
  - drugim rečima: srčani udari su ređi uz aspirin nego uz placebo
- stepen korelacije je vrlo mali (ϕ = 0.041), ali je kompenzovan velikim brojem ispitanika (N = 20000), koji je potreban zbog retkosti srčanog udara

2. Bivarijatni frekvencijski nacrti (BFN) 16

- računanje šansi i proporcija

	DA	NE	PILULA	šanse	proporcije DA	procenti
placebo	200	9800	10000	200/9800 = 1:49 ≈ 1:50 = 0.02	200/10000 = 0.02	2%
aspirin	100	9900	10000	100/9900 = 1:99 ≈ 1:100 = 0.01	100/10000 = 0.01	1%
SIMPTOM	300	19700	20000			

- količnik šansi: KŠ ≈ (1/50)/(1/100) = 100/50 = 2; 1/KŠ ≈ 1/2 = 0.5
  - dakle: šansa srčanog udara uz placebo je *dva puta* veća nego uz aspirin
  - obrnuto: šansa srčanog udara uz aspirin je *upola manja* nego uz placebo
- za 'DA': KP=0.02/0.01 = 2; dakle, sličan zaključak kao i na osnovu KŠ
  - kaže se: *relativni rizik* srčanog udara je aspirinom smanjen za 50% (prepolovljen)
- ali: uočimo da su udari relativno *retki* događaji, tako da su same šanse odn. proporcije, čiji se *količnik* računa za KŠ odn. KP, relativno *mali* brojevi
  - npr.: procent udara uz placebo je samo 2%, a uz aspirin je samo 1%
  - kaže se: *apsolutni rizik* udara je aspirinom smanjen za 1% (sa 2% na 1%)
- dalje: uz aspirin je 100 ljudi manje imalo srčani udar nego uz placebo
  - ali: 9800 ljudi ne bi imali udar ni bez aspirina (što se vidi iz placebo grupe)
  - dakle: od aspirina je imalo korist 100 ljudi od 10000, tj. samo 1%, a 99% nije!
- uočiti: veći broj ispravnih, ali veoma različitih prikaza istih rezultata:
- rel. rizik (manji 50%), aps. rizik (manji 1%), frek. (pomaže 100), proc. (pomaže 1%)

2. Bivarijatni frekvencijski nacrti (BFN) 17

- primer 2: *bolesti i dijagnoze*
- zamislimo neku populaciju u kojoj se javlja izvesna *bolest*
  - npr. određen broj ljudi ima HIV (AIDS, SIDA)
- *dijagnoza* bolesti se vrši različitim testovima
  - *pozitivan* test: bolest je prisutna, *negativan* test: bolest je odsutna
- testovi *nisu* savršeni, tj. dijagnoza ne mora uvek biti *tačna*
  - ishod testa ne odražava uvek potpuno ispravno prisustvo bolesti
- analizirajmo ovu problematiku kao nacrt tipa 2x2:
  - varijabla 1: *bolest*; kategorije: bolest prisutna (+B), bolest odsutna (-B)
  - varijabla 2: *dijagnoza*; kategorije: test pozitivan (+T), test negativan (-T)
  - situacije: 4 kombinacije kategorija, tipa ±B & ±T

	-T	+T	BOL.
-B	-B&-T	-B&+T	-B
+B	+B&-T	+B&+T	+B
DIAG.	-T	+T	N

situacije

sa stanovišta uspešnosti dijagnostike postoje:
dve željene situacije: -B&-T, +B&+T
dve neželjene situacije: -B&+T, +B&-T

2. Bivarijatni frekvencijski nacrti (BFN) 18

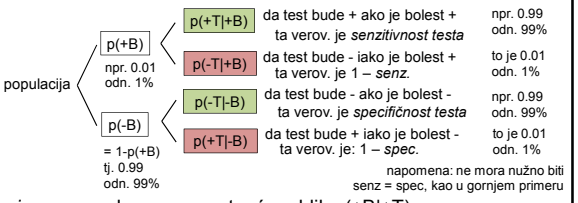
- uvedimo prikaz pomoću *verovatnoća*
- *bezuslovne verovatnoće: verovatnoće kategorija*
  - *bolest: kategorije B+ i B-*
    - *p(B+): prevalencija bolesti* (verovatnoća da osoba ima bolest)
    - *p(B-): verovatnoća da je osoba zdrava*
  - *dijagnoza: kategorije T+ i T-*
    - *p(T+): verovatnoća da je postavljena pozitivna dijagnoza*
    - *p(T-): verovatnoća da je postavljena negativna dijagnoza*
- *zajedničke verovatnoće: verovatnoće situacija*
  - p(-B&-T), p(+B&+T)
  - p(+B&-T), p(-B&+T)
- uslovne verovatnoće: dve vrste
  - 'direktne': oblika p(±T|±B): p(-T|-B), p(+T|+B), p(+T|-B), p(-T|+B)
  - 'inverzne': oblika p(±B|±T): p(-B|-T), p(+B|+T), p(+B|-T), p(-B|+T)

	-T	+T	BOL.
-B	p(-B&-T)	p(-B&+T)	p(-B)
+B	p(+B&-T)	p(+B&+T)	p(+B)
DIAG.	p(-T)	p(+T)	

videćemo: veličine *odgovarajućih* direktnih i inverznih verovatnoća (npr. p(-T|-B) i p(-B|-T)), mogu biti slične, ali i različite

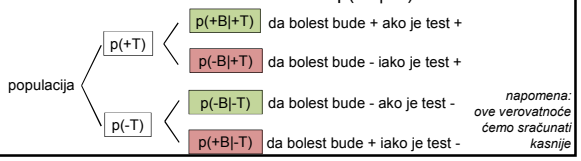
2. Bivarijatni frekvencijski nacrti (BFN) 19

- *direktne* uslovne verovatnoće: oblik p(±T|±B)



napomena: ne mora nužno biti senz = spec, kao u gornjem primeru

- *inverzne* uslovne verovatnoće: oblik p(±B|±T)

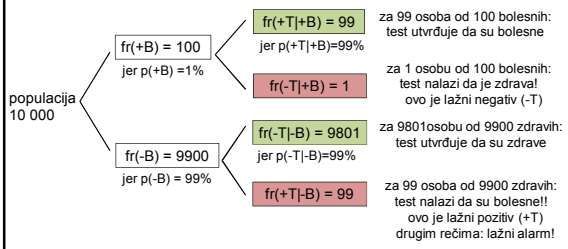


napomena: ove verovatnoće ćemo sračunati kasnije

2. Bivarijatni frekvencijski nacrti (BFN) 20

uvedimo prikaz pomoću ilustrativnih očekivanih *frekvenci*

- za ilustraciju, recimo da se vrši masovno testiranje na 10000 ljudi, da bi se utvrdilo da li su bolesni ili zdravi
- izračunajmo *očekivane* frekvence, na osnovu verovatnoća iz primera



očekivana frekvenca lažnog alarma je alarmantna!!

2. Bivarijatni frekvencijski nacrti (BFN) 21

- prikažimo ove očekivane frekvence matricama (uključujući i marginalne):

	negativan test (-T)	*pozitivan test (+T)*	*BOL.*
zdravi (-B)	9801	99	*9900*
bolesni (+B)	1	99	*100*
TEST	*9802*	*198*	10000

	negativan test (-T)	*pozitivan test (+T)*	*BOL.*
-B	tačni negativ (specifičnost)	lažni pozitiv (greška tipa 1)	*-B*
+B	lažni negativ (greška tipa 2)	tačni pozitiv (senzitivnost)	*+B*
TEST	*-T*	*+T*	N

- razmotrimo marginalne frekvence (100, 198, 9802)
- uočimo: 100 osoba od 10000 zaista je bolesno (+B)
- ali: 198 osoba od 10000, dakle skoro duplo više, imaće pozitivan test (+T)!
- zatim: od 9802 osobe sa negativnim testom (-T):
  - 9801 osoba nije bolesno, tj. p(-B|-T) = 9801/9802 ≈ 0.9999 (uporedi: p(-T|-B) = 0.99)
  - 1 osoba je bolesna, p(+B|-T) = 1/9802 ≈ 0.0001 (uporedi: p(-T|+B) = 0.01)
- ali: od 198 osoba sa pozitivnim testom (+T):
  - 99 osoba je bolesno, p(+B|+T) = 99/198 = 0.50 (uporedi: p(+T|+B) = 0.99)
  - 99 osoba je zdravo, tj. p(-B|+T) = 99/198 = 0.50 (uporedi: p(+T|-B) = 0.01)
- dakle: iako direktna uslovna verovatnoća p(+T|+B), tj. da je, ako je prisutna bolest, test pozitivan, iznosi čak 0.99 - tj. test je vrlo senzitivan (osetljiv) na prisustvo bolesti ...
- ... ipak *obrnuta* uslovna verovatnoća p(+B|+T), tj. da je, ako je test pozitivan, bolest prisutna, iznosi samo 0.50 (tj. jednaka je verovatnoći da je bolest odsutna)!!

2. Bivarijatni frekvencijski nacrti (BFN) 22

- važno: ovakvi rezultati mogu se javiti kad se vrši *masovno* testiranje (10000), a velika većina populacije *nije* bolesna (9900)
  - naime, i vrlo senzitivni i specifični testovi, ipak nisu savršeni!
  - stoga, ako se testira veoma *mnogo* ljudi, onda se i veoma *mali procenti* greške mogu odnositi na relativno *veliki broj* ljudi
- međutim, ako se testira samo potencijalno bolesna populacija, situacija je drugačija
- *numerički primer 2*
  - razmotrimo uzorak od 200 osoba sa *sumnjom* na bolest, odn. konkretno:
  - pretpostavimo da je polovina uzorka bolesna
    - tj., sada su bezuslovne verovatn.:
      - p(+B) = 0.50; p(-B) = 0.50
  - pretpostavimo da test ima *istu* osetljivost i specifičnost, od po 0.99
  - po istoj logici kao ranije, očekivane frekvence biće kao u priloženoj tabeli
    - uočimo: važi ne samo da je p(+T|+B) = 0.99, kao i u prethodnoj analizi
    - već i da je i p(+B|+T) = 0.99, dok je u prethodnoj analizi bilo 0.50

očekivane frekvence

	negativan test (-T)	*pozitivan test (+T)*	*BOL.*
zdravi (-B)	99	1	*100*
bolesni (+B)	1	99	*100*
TEST	*100*	*100*	200

2. Bivarijatni frekvencijski nacrti (BFN) 23

(2) Nacrti složeniji od tipa 2x2

- analiza je vrlo slična kao za nacrte tipa 2x2

- *Faza I: Podaci i deskriptivne mere*
  - utvrđivanje frekvenci situacija i kategorija, i totalne frekvence

primer A

	rok	*ozb.*	*nar.*	*UZR.*
mladi	250	100	150	*500*
sred.	200	80	120	*400*
stari	50	20	30	*100*
MUZ.	*500*	*200*	*300*	*1000*

primer B

	rok	*ozb.*	*nar.*	*UZR.*
mladi	400	50	50	*500*
sred.	100	100	200	*400*
stari	0	50	50	*100*
MUZ.	*500*	*200*	*300*	*1000*

uočiti: marginalne frekvence u oba primera su iste

- kako se izračunava korelacija?
  - KŠ nije primeren: definisan je samo za dve šanse
  - ϕ-koeficijent nije primeren: formula važi samo za nacrte tipa 2x2
  - za nacrte veće od 2x2 primeren je tzv. *koeficijent kontingencije*, ali on ima izvesne statističke mane
  - često se koristi tzv. *Kramerov ϕ-koeficijent* (biće definisan kasnije)

2. Bivarijatni frekvencijski nacrti (BFN) 24

- *Faza II: nulta hipoteza, očekivane vrednosti, devijacije*
  - 1D nulte hipoteze: proste i marginalne matrice
  - 2D nulta hipoteza: ne postoji korelacija varijabli A i B
  - očekivane frekvence f':
    - važi ista logika i formule kao kod nacrta tipa 2x2
    - primer A
      - isti procenti u svim redovima, jednaki odnosi frekvenci (5:2:3), isti profili
      - može se pokazati: očekivane frekvence su *jednake* opserviranim
      - *nema* korelacije
    - primer B
      - različ. procenti u redovima, razl. odnosi frekv. (8:1:1, 1:1:2, 0:1:1), razl. profili
      - može se pokazati: očekivane frekvence su *različite* od opserviranih
      - *ima* korelacije
      - uočiti: kako su marginalne frekvence iste kao u primeru A (gde nema korel.), *očekivane* frekvence kod B su iste kao *opservirane* frekvence kod A
  - reziduali
    - računaju se isto kao kod nacrta tipa 2x2: d = f – f'

2. Bivarijatni frekvencijski nacrti (BFN) 25

- *Faza III:* test statistik
  - isto kao kod nacrta tipa 2x2, samo što ima više od 4 elementa
    - χ^2 se dobija kvadriranjem reziduala u svim ćelijama, deljenjem sa odgovarajućim očekivanim frekvencama, i sabiranjem svih tih vrednosti
      - χ^2 u nacrtu 3x3 ima devet sabiraka
      - primer A: $\chi^2 = 0$; primer B: $\chi^2 = 398.33$
  - broj stepeni slobode:
    - opšta formula: za nacrte tipa axb, važi da je df = (a-1)(b-1)
      - na pr., za nacrte tipa 2x2, df = (2-1)(2-1) = 1
      - u primeru, za nacrte tipa 3x3, df = (3-1)(3-1) = 4
- *Faza IV:* izračunavanje p-vrednosti
  - verovatnoća dobijene vrednosti χ^2, ili veće, ako je H0 tačna u populaciji
  - u primeru A: $\chi^2(4) = 0$, p = 1
  - u primeru B: $\chi^2(4) = 398.33$, p < 0.05
- *Faza V:* odluka o statističkoj značajnosti
  - po istim principima
    - u primeru A: nema korelacije, u primeru B: korelacija je stat. značajna

2. Bivarijatni frekvencijski nacrti (BFN) 26

- računanje koef. korelacije u nacrtima složenijim od tipa 2x2
  - podsetimo se alternativne formule za χ^2, naime $\chi^2 = N*\phi^2$
    - iz nje sledi alternativna formula za ϕ, naime $\phi = \sqrt{\chi^2/N}$
  - ova formula se ne može koristiti u nacrtima tipa axb složenijim od 2x2, ali se umesto nje može koristiti veoma slična formula, naime:
  - *Kramerov* ϕ-koeficijent, označen sa V ili $\phi_c = \sqrt{\chi^2/(N(c-1))}$
    - c je jednak *manjem* broju od brojeva a i b
      - naime: ako je a<b, onda c=a; ako je b<a onda c=b; ako je a=b, c=a=b
    - u primeru B, c=a=b=3, $\phi_c = 0.45$, tj. postoji korelacija uzrasta i muz. prefer.
    - uočimo: u nacrtu 2x2 važi a=b=2, što znači da je c=2
      - dakle, tada je formula za ϕ_c ista kao gornja formula za ϕ (jer je c-1=1)
  - prema tome: ϕ_c je *uopštenje* ϕ-koeficijenta za složenije nacrte
    - drugim rečima, za *sve* bivarijatne nacrte se može koristiti ϕ_c, a on se u slučaju nacrta 2x2 svodi na obični ϕ-koeficijent
- opisani postupak je *omnibus test*, jer se odnosi na celu matricu AB
  - ponekad je korisno testirati postojanje korelacije za neki 2D deo matrice AB
    - na pr. u matrici tipa 4x5 analiziraju se podmatrice tipa 2x2, 2x3, itd
  - ovakvi testovi se ređe koriste

3. Trivarijatni frekvencijski nacrti (TFN) 27

- TFN su frekvencijski nacrti sa tri varijable
  - oznake varijabli su A, B i C, nacrt je tipa axbxc
- *PRIMERI:* nacrt 2x2x2: odnos rukosti (A), pola (B) i uzrasta (C: ml., star.)
  - nacrt 3x3x3: odnos omiljenosti vrste muzike, uzrasta i mesta stanovanja

a. Organizacija podataka

- matrica podataka tipa objekti x varij.
  - tri kolone (po jedna za svaku varijablu)
  - onoliko redova koliko ima objekata

#	RUKOST	POL	UZRAST
1.	1	1	1
2.	2	2	1
3.	1	2	2
4.	...	...	...
...	...	...	...

b. Deskriptivne statističke mere

- utvrđuju se: ukupna frekvenca, frekvence kategorija za sve tri varijable, i frekvence svih situacija (kombinacija kategorija)
- *numerički prikaz rezultata: vrši se* pomoću:
  - glavnih matrica: ABC, AB, AC, BC, A, B, C
  - prostih matrica: AB/c1, AB/c2, AC/b1, ...
- za konkretno istraživanje ne koriste se *sve* matrice za prikazivanje rezultata
  - treba prikazati *najdetaljnije* rezultate: u matrici ABC, razloženoj na proste matrice

3. Trivarijatni frekvencijski nacrti (TFN) 28

Primer: odnos rukosti (A: desn., lev.), pola (B: m, ž) i uzrasta (C: mladi, stari)

odnos pola i rukosti kod mladih (AB/c1)

AB/c1	desnoruki	levoruki	POL
muš.	396 (87.2%)	58 (12.8%)	454 (100%)
žene	561 (91.6%)	58 (9.4%)	619 (100%)
RUK.	957 (89.2%)	116 (10.8%)	1073 (100%)

odnos pola i rukosti kod starih (AB/c2)

AB/c2	desnoruki	levoruki	POL
muš.	369 (94.4%)	22 (5.6%)	391 (100%)
žene	607 (98.1%)	12 (1.9%)	619 (100%)
RUK.	976 (96.6%)	34 (3.4%)	1010 (100%)

AB: odnos pola i rukosti: zbir AB/c1 i AB/c2

AB	desnoruki	levoruki	POL
muš.	765 (90.5%)	80 (9.5%)	845 (100%)
žene	1168 (94.3%)	70 (5.7%)	1238 (100%)
RUK.	1933 (92.8%)	150 (7.2%)	2083 (100%)

AC: odnos rukosti i uzrasta: zbir AC/b1 i AC/b2

AC	desnoruki	levoruki	UZRAST
mladi	957 (89.2%)	116 (10.8%)	1073 (100%)
stari	976 (96.6%)	34 (3.4%)	1010 (100%)
RUK.	1933 (92.8%)	150 (7.2%)	2083 (100%)

odnos pola i uzrasta: zbir BC/a1 i BC/a2

BC	mladi	stari	POL
muš.	454 (53.7%)	391 (46.3%)	845 (100%)
žene	619 (50%)	619 (50%)	1238 (100%)
UZR.	1073 (51.5%)	1010 (48.5%)	2083 (100%)

1D glavne matrice

A	desnoruki	levoruki	
RUK.	1933 (92.8%)	150 (7.2%)	2083 (100%)

B	muš.	žene	
POL	845 (40.6%)	1238 (59.4%)	2083 (100%)

C	mladi	stari	
UZR.	1073 (51.5%)	1010 (48.5%)	2083 (100%)

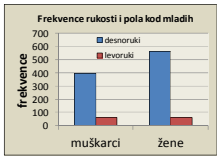
- nisu prikazani:
  - AC/b1, AC/b2
  - BC/a1, BC/a2

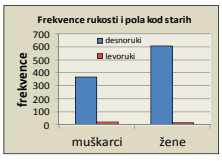
3. Trivarijatni frekvencijski nacrti (TFN) 29

grafički prikaz rezultata: pomoću već opisanih tipova grafikona za BFN

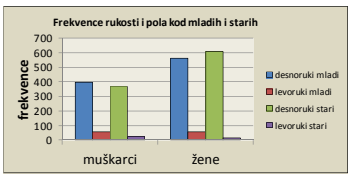
daćemo samo par primera

odvojeni 2D grafikoni za dve kategorije treće varijable (mladi i stari)





zajednički grafikon svih frekvenci



3. Trivarijatni frekvencijski nacrti (TFN) 30

d. Struktura rezultata

- počnimo od BFN sa varijablama A i B (u primeru: rukost i pol)
- proširimo ga u TFN, sa dodatnom varijablom C (u primeru: uzrast)
- odnos varijabli A i B (tj. da li su one u korelaciji) može se poreditи:
  - u marginalnoj matrici AB (odnos rukosti i pola bez obzira na uzrast)
  - u prostoj matrici AB/c1 (odnos rukosti i pola kod mladih)
  - u prostoj matrici AB/c2 (odnos rukosti i pola kod starih)
- *da li uvođenje treće varijable menja ili ne menja uvid u odnos prve dve?*
- terminologija:
  - moderatorska varijabla: treća varijabla, C (jer *moderira* odn. utiče na odnos A i B)
  - elaboracija: analiza uticaja moderatorske varijable
  - marginalna korelacija: korelacija varijabli A i B u marginalnoj matrici AB
  - parcijalne (delim.) korelacije: korelacije A i B u prostim matricama AB/c1 i AB/c2
  - u prostim matricama AB/c1 i AB/c2 varijabla C se *drži konstantnom*
    - u AB/c1, C ima vrednost c1 (mladi), a u AB/c2, C ima vrednost c2 (stari)
  - u glavnoj matrici AB varijabla C ne drži se konstantnom već *varira*
    - u toj matrici C može imati bilo vrednost c1 bilo c2
  - postoji veći broj mogućih struktura rezultata u trivarijatnim nacrtima
  - primere ćemo podeliti u dve grupe, zavisno od postojanja marginalne korelacije

3. Trivarijatni frekvencijski nacrti (TFN) 31

prva grupa primera – nema marginalne korelacije u matrici AB

AB	puš.	nep.	POL	šanse
muš.	400	200	600	2:1
žene	200	100	300	2:1
PUŠ.	600	300	900	2:1

struktura rezultata

AB/c1 (mladi)	puš.	nep.	POL	šanse
muš.	200	100	300	2:1
žene	100	50	150	2:1
PUŠ.	300	150	450	2:1

AB/c2 (stari)	puš.	nep.	POL	šanse
muš.	200	100	300	2:1
žene	100	50	150	2:1
PUŠ.	300	150	450	2:1

1. validacija (replikacija)
(nema korelacije ni u AB/c1 ni u AB/c2)

AB/c1 (mladi)	puš.	nep.	POL	šanse
muš.	40	80	120	1:2
žene	40	80	120	1:2
PUŠ.	80	160	240	1:2

AB/c2 (stari)	puš.	nep.	POL	šanse
muš.	360	120	480	3:1
žene	160	20	180	8:1
PUŠ.	520	140	660	26:7

2. delimična validacija (nema korelacije u AB/c1, ali je ima u AB/c2)

AB/c1 (mladi)	puš.	nep.	POL	šanse
muš.	300	100	400	3:1
žene	50	50	100	1:1
PUŠ.	350	150	500	7:3

AB/c2 (stari)	puš.	nep.	POL	šanse
muš.	100	100	200	1:1
žene	150	50	200	3:1
PUŠ.	250	150	400	5:3

3. invalidacija
(ima korelacije i u AB/c1, i u AB/c2)
uočiti: ovo je moguć rezultat!

3. Trivarijatni frekvencijski nacrti (TFN) 32

druga grupa primera – ima marginalne korelacije u matrici AB

AB	puš.	nep.	POL	šanse
muš.	180	120	300	3:2
žene	120	180	300	2:3
PUŠ.	300	300	600	1:1

struktura rezultata

AB/c1 (mladi)	puš.	nep.	POL	šanse
muš.	90	60	150	3:2
žene	60	90	150	2:3
PUŠ.	150	150	300	1:1

AB/c2 (stari)	puš.	nep.	POL	šanse
muš.	90	60	150	3:2
žene	60	90	150	2:3
PUŠ.	150	150	300	1:1

4. validacija (replikacija)
(ima korelacije i u AB/c1 i u AB/c2)

AB/c1 (mladi)	puš.	nep.	POL	šanse
muš.	80	20	100	4:1
žene	20	80	100	1:4
PUŠ.	100	100	200	1:1

AB/c2 (stari)	puš.	nep.	POL	šanse
muš.	100	100	200	1:1
žene	100	100	200	1:1
PUŠ.	200	200	400	1:1

5. delimična validacija (specifikacija)
(ima korelacije u AB/c1, ali je nema u AB/c2)
specifikovan je izvor marginalne korelacije

AB/c1 (mladi)	puš.	nep.	POL	šanse
muš.	160	80	240	2:1
žene	40	20	60	2:1
PUŠ.	200	100	300	2:1

AB/c2 (stari)	puš.	nep.	POL	šanse
muš.	20	40	60	1:2
žene	80	160	240	1:2
PUŠ.	100	200	300	1:2

6. invalidacija (prividna marginalna korelacija)
(nema korelacije ni u AB/c1, ni u AB/c2)
uočiti: moguć rezultat!

3. Trivarijatni frekvencijski nacrti (TFN) 33

dodatni primeri prividne marg. korelacije ('Simpsonov paradoks')

- PRIMER 1: korelacija matematičke sposobnosti i broja cipele
  - test matematike primeren srednjem uzrastu (4-5 razred osnovne škole)
  - varijabla A: uspeh đaka na testu (položio, pao)
  - varijabla B: broj cipele đaka (mali, veliki)

AB: svi	vel.	mali	USPEH	šanse
položili	180	120	300	3:2
pali	120	180	300	2:3
BROJ CIP.	300	300	600	1:1

  - moderatorska varijabla C: uzrast: mlađi (treći razred), stariji (sedmi razred)

AB: mlađi	vel.	mali	USPEH	šanse
položili	20	40	60	1:2
pali	80	160	240	1:2
BROJ CIP.	100	200	300	1:2

AB: stariji	vel.	mali	USPEH	šanse
položili	160	80	240	2:1
pali	40	20	60	2:1
BROJ CIP.	200	100	300	2:1

  - korelacija A i B postoji u celom uzorku, ali ne postoji ni kod mlađih ni kod starijih!
- PRIMER 2: polna diskriminacija prilikom zapošljavanja?
  - varijabla A: pol (muški, ženski)
  - varijabla B: zapošljavanje nastavnika na univerzitetu (primljeni, odbijeni)

3. Trivarijatni frekvencijski nacrti (TFN) 34

AB: svi	prim.	odb.	POL	šanse
muš.	40	60	100	2:3
žene	40	100	140	2:5
ZAPOŠLJ.	80	160	240	1:2
šanse	1:1	3:5	5:7	

- postoji polna diskriminacija!
  - ne apsolutno nego procentualno
- muš.: primljeno 40%, odbijeno 60%
- žene: primljeno 29%, odbijeno 71%
- uvedimo moderatorsku varijablu C: fakultet

AB: filološki	prim.	odb.	POL	šanse
muš.	10	30	40	1:3
žene	30	90	120	1:3
ZAPOŠLJ.	40	120	160	1:3
šanse	1:3	1:3	1:3	

AB: mašinski	prim.	odb.	POL	šanse
muš.	30	30	60	1:1
žene	10	10	20	1:1
ZAPOŠLJ.	40	40	80	1:1
šanse	3:1	3:1	3:1	

ni na filološkom ni na mašinskom fakultetu ne postoji polna diskriminacija!?!

- jednak broj kandidata (po 40) prima se na filološki (FF) i na mašinski (MF)
- međutim, 2x više kandidata se prijavljuje na FF (160) nego na MF (80)
- 3x strožiji je odbir kandidata (odnos prim:odb) na FF (1:3) nego na MF (1:1)
- na FF se prijavljuje 3x više žena nego muškaraca (1:3), na MF 3x više muš. nego ž. (3:1)

BC	prim.	odb.	FAK.	šanse
filološki	40	120	160	1:3
mašinski	40	40	80	1:1
ZAPOŠLJ.	80	160	240	1:2
šanse	1:1	3:1	2:1	

AC	muš.	žen.	FAK.	šanse
filološki	40	120	160	1:3
mašinski	60	20	80	3:1
POL	100	140	240	5:7
šanse	2:3	5:1	2:1	

- usled ovakvog sklopa biva odbijen veći procent žena (71%) nego muškaraca (60%)