TEMELJNE SPOZNAJE O UZORKU

doc.dr.sc. Vesna Ilakovac

Katedra za biofiziku, medicinsku statistiku i medicinsku informatiku

Medicinski fakultet Osijek

UZORAK I POPULACIJA

POPULACIJA

- osnovni skup
- skup svih jedinica promatranja (entiteta) opisanih varijablama (atributima)

UZORAK - dio jedinica populacije (osnovnog skupa)

TEORIJA UZORAKA

- ustanovljava svojstva populacije iz svojstava uzorka
- ***procjenjuje*** parametre populacije na temelju podataka dobivenih iz uzorka i ***ocjenjuje pouzdanost*** te procjene

UOBIČAJENE OZNAKE

UZORAK ↓

KARAKTERISTIČNE VELIČINE	OCJENA PARAMETRA (STATISTIKA)	PARAMETAR POPULACIJE
ARITMETIČKA SREDINA	$\bar{x}$	μ
STANDARDNA DEVIJACIJA	s	σ
PROPORCIJA	p	π

↑ POPULACIJA

PARAMETAR I STATISTIKA

POPULACIJA

1. UZORAK

2. UZORAK

⋮

n-ti UZORAK

PARAMETAR I STATISTIKA

aritmetička sredina visine populacije = 175.4

aritmetička sredina visine 1. uzorka = 172.2

aritmetička sredina visine 2. uzorka = 178.1

aritmetička sredina visine n-tog uzorka = 173.7

PARAMETAR I STATISTIKA

- **parametar:**
  - vrijednost (obično nepoznata) koja predstavlja neku karakteristiku populacije
  - unutar populacije, parametar je nepromjenljiva vrijednost koja NE VARIRA
- **statistika:**
  - veličina izračunata iz podataka izmjerenih na uzorku
  - vrijednost statistike MIJENJA SE od uzorka do uzorka

OSNOVNI POJMOVI

Koja je skupina na koju želimo generalizirati?

TEORETSKA POPULACIJA

Koja je populacija dostupna?

POPULACIJA KOJU ISTRAŽUJEMO

Na koji način možemo obuhvatiti populaciju?

OKVIR IZBORA

Tko je uključen u istraživanje?

UZORAK

UZORAK I POPULACIJA

Kvaliteta ocjene parametara ovisi o:

- **REPREZENTATIVNOSTI UZORKA**
- **ODABRANOJ VJEROJATNOSTI**

REPREZENTATIVNI UZORAK

– uzorak koji dobro opisuje populaciju

Na reprezentativnost uzorka utječu:

1. Vrsta uzorka (prema metodi odabira)
2. Veličina uzorka
3. Varijabilnost promatranog obilježja

VRSTE UZORAKA

PROBABILISTIČKI (probability samples)

– svaka jedinica promatranja u populaciji ima ***jednaku vjerojatnost izbora*** u uzorak koja je ***različita od 0***

NEPROBABILISTIČKI (non-probability samples)

– ***vjerojatnost izbora*** jedinica promatranja iz populacije je ***različita i nepoznata*** (može biti i 0)

PROBABILISTIČKI UZORCI

JEDNOSTAVNI SLUČAJNI
SUSTAVNI SLUČAJNI
SLOJEVITI (STRATIFICIRANI)
UZORAK SKUPINE
VIŠEFAZNI

JEDNOSTAVNI SLUČAJNI UZORAK
(*simple random sample*)

svojstva:

- ***svaki element*** populacije ima ***jednaku šansu*** da bude izabran
- ***svaki uzorak*** ima ***jednaku šansu*** da bude izabran

način izbora:

- lutrijska metoda
- pomoću tablice slučajnih brojeva
- pomoću programske podrške koja ima funkciju generatora slučajnih brojeva

SUSTAVNI SLUČAJNI UZORAK
(*systematic sample*)

– jedinice koje ulaze u uzorak odabiru se po nekakvom pravilu

postupak:

– numerirati jedinice populacije od 1 do N
– odrediti potrebnu veličinu uzorka (n)
– odrediti veličinu intervala k=N/n
– slučajno odabrati broj između 1 i k (početna jedinica)
– uzimati svaku k-tu jedinicu

SLOJEVITI (STRATIFICIRANI) UZORAK

– primjenjuje se u slučajevima kad je promatrano obilježje heterogeno u populaciji
– dobiva se uzimanjem jednostavnih slučajnih uzoraka iz stratuma određenih obzirom na promatrano obilježje

postupak:

– podijeliti populaciju na disjunktne skupine od
$n_1, n_2, \ldots n_s$
jedinica, pri čemu je
$n_1+n_2+\ldots+n_s = N$
– uzeti jednostavni slučajni uzorak od
$f_i = n_i/N$
jedinica iz svake skupine

UZORAK SKUPINE (CLUSTER)

- primjenjuje se u slučajevima kada treba uzeti uzorak iz populacije koja se sastoji od skupina jedinica (ulice, popisni krugovi, škole, općine, ...)

postupak:

- podijeliti populaciju na skupine jedinica
- jednostavnim slučajnim izborom odabrati skupine
- ispitati SVE jedinice unutar odabranih skupina

VIŠEFAZNI UZORAK

- uzimanje "uzorka iz uzorka"
- kombinacija više metoda odabiranja uzorka

Npr. jedan od mogućih načina dobivanja uzorka iz populacije učenika osnovnih škola u Hrvatskoj:

- podijeliti osnovne škole u stratume s obzirom na županijsku pripadnost
- iz svakog stratuma jednostavnim slučajnim izborom odabrati škole (prva faza)
- unutar odabranih škola, jednostavnim slučajnim izborom odabrati razrede (druga faza)
- unutar odabranih razreda, jednostavnim slučajnim izborom odabrati učenike (treća faza)

NEPROBABILISTIČKI UZORCI

PRIGODNI (convenience)

UZORAK KOJI SLUŽI SVRSI (purposive)

UZORAK UDJELA (quota)

PRIGODNI UZORAK (convenience sample)

- u uzorak se biraju jedinice populacije koje su "pri ruci" (npr. prolaznici, pozvani dobrovoljci, prvih 50 pacijenata u nekoj ambulanti)

UZORAK KOJI SLUŽI SVRSI (purposive sample)

- u uzorak se biraju jedinice populacije koje imaju traženo svojstvo

UZORAK UDJELA (quota sample)

- u uzorak se bira određeni broj jedinica odabranih dijelova populacije

U kojem od sljedećeg se koristi jednostavni slučajni uzorak:

a) igra “Bingo”,

b) popis stanovništva,

c) izbori za lokalnu samoupravu?

Koje metode odabira uzorka se koriste u ovim primjerima?

a) igra “Bingo” – jednostavni slučajni uzorak

b) popis stanovništva - ne koristi jednostavni slučajni uzorak jer SVE jedinice populacije moraju biti obuhvaćene popisom

c) izbori za lokalnu samoupravu – neprobabilistički uzorak; na izbore izlaze oni koji to žele (“dobrovoljci”)

VELIČINA UZORKA

dovoljno veliki uzorak:

- uzorak pomoću kojeg s razumnom pouzdanošću možemo prihvatiti ili odbaciti neku hipotezu i ocijeniti parametar populacije

ovisit će o:

- homogenosti populacije s obzirom na promatrano obilježje
- učestalost promatranog obilježja (obrnuto proporcionalno)

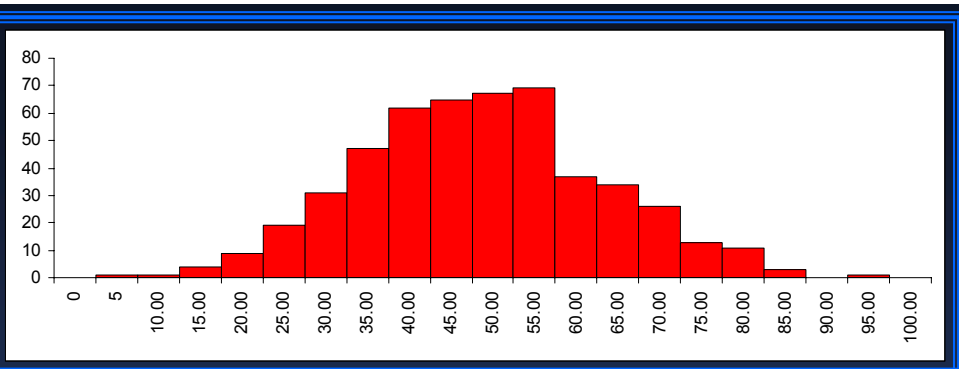
UTJECAJ VARIJABILNOSTI

- varijabilnost je često nepoznata
- poznata, a velika varijabilnost ugrožava reprezentativnost uzorka
- utjecaj varijabilnosti se smanjuje s povećanjem uzorka

STANDARDNA POGREŠKA

POKUS 1. Napraviti razdiobe aritmetičkih sredina 500 uzoraka veličine n=4, n=20 i n=50 osnovnog skupa N=101 brojeva od 0 do 100.

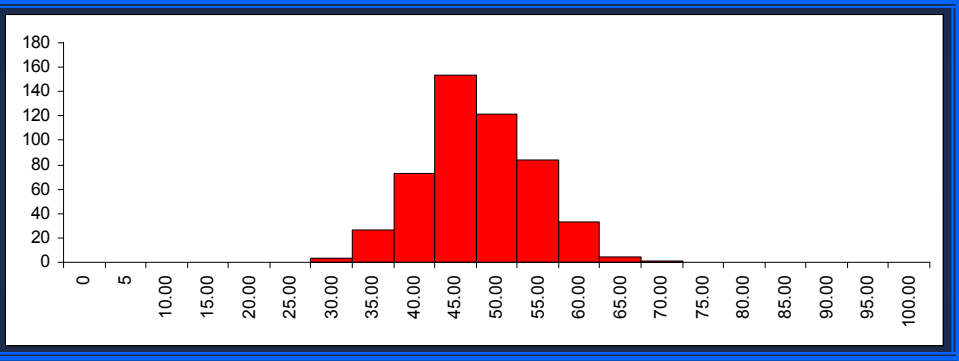
aritmetička sredina osnovnog skupa	$\mu = 50$
standardna devijacija osnovnog skupa	$\sigma = 29.15$



$n = 4$

$\bar{x} = 50.96$

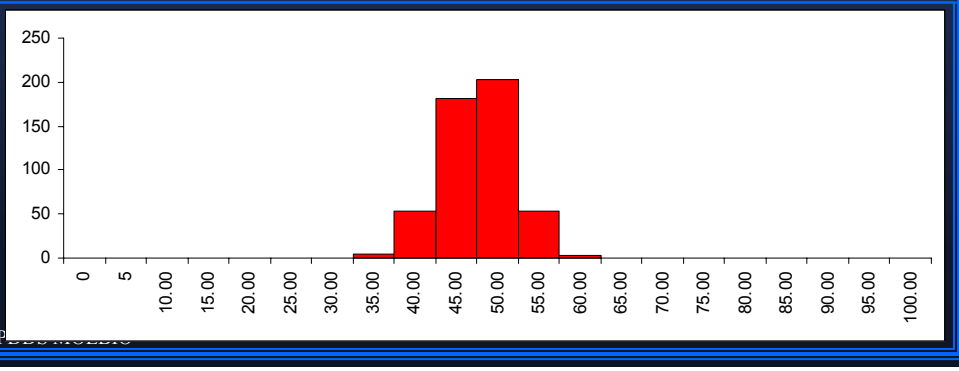
$s = 14.451$



$n = 20$

$\bar{x} = 50.10$

$s = 6.639$



$n = 50$

$\bar{x} = 50.07$

$s = 4.189$

N .... veličina osnovnog skupa

n .... veličina slučajnih uzoraka

$\binom{N}{n}$ broj svih mogućih uzoraka veličine **n** uzetih iz osnovnog skupa veličine N

1. uzorak	$x_{11}, x_{12}, \ldots, x_{1n}$	sa sredinom	$\bar{x}_1$
2. uzorak	$x_{21}, x_{22}, \ldots, x_{2n}$	sa sredinom	$\bar{x}_2$
3. uzorak	$x_{31}, x_{32}, \ldots, x_{3n}$	sa sredinom	$\bar{x}_3$
....			
k-ti uzorak	$x_{k1}, x_{k2}, \ldots, x_{kn}$	sa sredinom	$\bar{x}_k$

$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_k$ slučajna varijabla (*sampling distribucija*)

OČEKIVANJE $E(\bar{x}) = \mu$

VARIJANCA $V(\bar{x}) = \frac{\sigma^2}{n}$

standardna devijacija $s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ **STANDARDNA POGREŠKA (SE, standard error)**

- za slučajne i dovoljno velike uzorke

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$ **STANDARDNA POGREŠKA (SE, standard error)**

- standardna pogreška aritmetičke sredine (SEM, *standard error of the mean*)
- pogreška kojoj se izlažemo pri zaključivanju o populaciji na temelju uzorka

$s \uparrow \Rightarrow s_{\bar{x}} \uparrow$ ***povećava se*** s povećanjem *varijabilnosti obilježja*

$n \uparrow \Rightarrow s_{\bar{x}} \downarrow$ ***smanjuje se*** s povećanjem *veličine uzorka*

CENTRALNI GRANIČNI TEOREM

Razdioba aritmetičkih sredina uzoraka teži normalnoj razdiobi s očekivanjem μ i varijancom $\sigma^2_{\bar{x}}$ $\left[N\left(\mu, \sigma^2_{\bar{x}}\right) \right]$ kad veličina uzorka *n* teži u beskonačnost.

$\Rightarrow$ ***za dovoljno velike uzorke*** razdioba aritmetičkih sredina uzoraka bit će ***normalna***, bez obzira na razdiobu vrijednosti promatranog obilježja

za proporciju:

$$s_p = \sqrt{\frac{p \cdot q}{n}}$$

STANDARDNA POGREŠKA PROPORCIJE

$p\uparrow \Rightarrow s_{\bar{x}}\downarrow$ ***smanjuje se*** s povećanjem *homogenosti obilježja*

$n\uparrow \Rightarrow s_{\bar{x}}\downarrow$ ***smanjuje se*** s povećanjem *veličine uzorka*

STANDARDNA POGREŠKA

vs

STANDARDNA DEVIJACIJA

STANDARDNA POGREŠKA:

- procjenjuje "kvalitetu" ocjene parametra (statistike)
- velika standardna pogreška => ocjena parametra (ar. sredina, proporcija) je neprecizna

STANDARDNA DEVIJACIJA:

- opisuje varijabilnost podataka
- velika standardna devijacija => velika varijabilnost podataka

RASPON POUZDANOSTI

RASPON POUZDANOSTI

- confidence interval (CI)
- uobičajeno tumačenje:

  raspon unutar kojega se, s određenom vjerojatnošću, nalazi prava vrijednost (parametar) populacije

RASPON POUZDANOSTI ARITMETIČKE SREDINE

$$\bar{x} - z \cdot s_{\bar{x}} \leq \mu \leq \bar{x} + z \cdot s_{\bar{x}}$$

RASPON POUZDANOSTI PROPORCIJE

$$p - z \cdot s_p \leq \Pi \leq p + z \cdot s_p$$

z - standardizirana vrijednost normalne raspodjele (ovisi o pretpostavljenoj vjerojatnosti)

RASPON POUZDANOSTI

PRIMJER. Koliki je raspon pouzdanosti ako želimo obuhvatiti μ sa:

a) 99% pouzdanosti

b) 95% pouzdanosti

c) 90% pouzdanosti

uz pretpostavku normalne razdiobe?

a) $z_{0.005}=2.756\approx2.58$ $\quad \bar{x} - 2.58 \cdot s_{\bar{x}} \leq \mu \leq \bar{x} + 2.58 \cdot s_{\bar{x}}$

b) $z_{0.025}=1.96$ $\quad \bar{x} - 1.96 \cdot s_{\bar{x}} \leq \mu \leq \bar{x} + 1.96 \cdot s_{\bar{x}}$

c) $z_{0.05}=1.65$ $\quad \bar{x} - 1.65 \cdot s_{\bar{x}} \leq \mu \leq \bar{x} + 1.65 \cdot s_{\bar{x}}$

RASPON POUZDANOSTI

PRIMJER. Od 1000 ljudi koji su cijepljeni, 200 ih je pokazalo alergične reakcije. Koliku proporciju alergičnih očekujemo u populaciji cijepljenih uz vjerojatnost od 95%?

$z_{0.025}=1.96$

$p-1.96\cdot s_p \le \Pi \le p+1.96\cdot s_p$

$p=0.20; \quad q=0.80;$

$$s_p=\sqrt{\frac{0.2\cdot 0.8}{1000}}=\sqrt{\frac{0.16}{1000}}=\sqrt{0.00016}=0.0126$$

$$0.2-1.96\cdot 0.0126 \le \Pi \le 0.2+1.96\cdot 0.0126$$

$$0.2-0.025 \le \Pi \le 0.2+0.025$$

$$0.175 \le \Pi \le 0.225$$

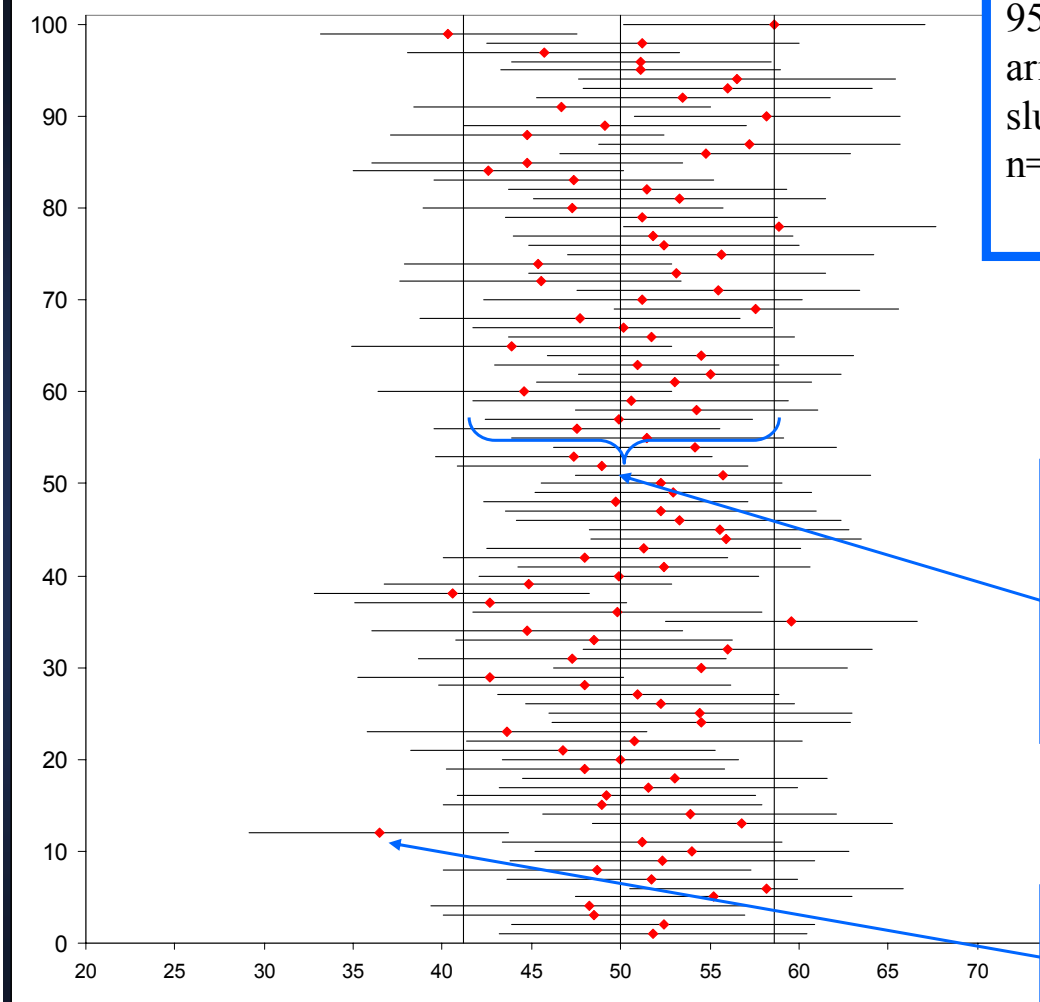
95% raspon pouzdanosti aritmetičke sredine izračunat iz nekog uzorka:

- ***uobičajeno*** se tumači kao raspon vrijednosti unutar kojeg se s 95% pouzdanosti nalazi prava vrijednost aritmetičke sredine (aritmetička sredina populacije)
- ***u stvari znači*** da očekujemo da 95% takvih intervala dobivenih iz uzoraka iste veličine dane populacije uključuje pravu vrijednost aritmetičke sredine

95% raspon pouzdanosti:

Slučajan interval čije granice se mogu izračunati iz podataka o uzorku, takav da 95 od svakih 100 takvih intervala obuhvaća pravu vrijednost parametra koji se procjenjuje.

- takodjer i raspon poželjnih vrijednosti parametra populacije (prihvatljiva nul-hipoteza)



95% rasponi pouzdanosti aritmetičkih sredina 100 slučajnih uzoraka veličine n=50

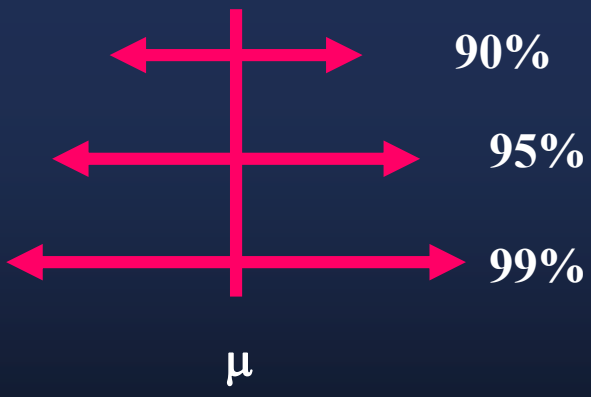
raspon unutar kojeg očekujemo 95% aritmetičkih sredina uzoraka

nije "pogrešan", samo nema sreće

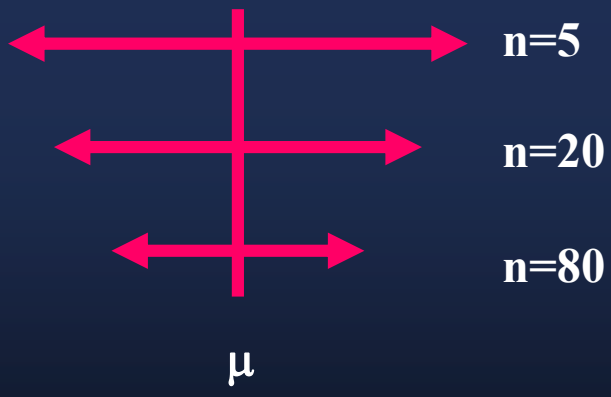
Širina raspona pouzdanosti ovisi o:

- **pretpostavljenoj vjerojatnosti**
- **varijabilnosti promatranog obilježja**
- **veličini uzorka**

širi se s povećanjem pouzdanosti



sužava se s povećanjem uzorka



POTREBNA VELIČINA UZORKA
za procjenu aritmetičke sredine

Ovisit će o:

- **pogrešci procjene koju ćemo tolerirati**
- **stupnju pouzdanosti**
- **pretpostavljenoj varijabilnosti**

$$E = z \cdot s_{\bar{x}} = z \cdot \frac{s}{\sqrt{n}}$$

POGREŠKA PROCJENE

$$n = \left(\frac{z \cdot s}{E} \right)^2$$

POTREBNA VELIČINA UZORKA

PRIMJER. Koliko ispitanika treba izabrati u uzorak kako bi se procijenila prosječna starost stanovnika nekog sela u 95% rasponu pouzdanosti od 2 godine? Pretpostavlja se kako je standardna devijacija populacije 8 godina.

Koliki uzorak treba biti ako toleriramo pogrešku od najviše ± 10 mjeseci?

$z_{0.025}=1.96$ $E=1$ $\sigma = 8$

$$n = \left(\frac{z_{0.025} \cdot 8}{1}\right)^2 = \left(\frac{1.96 \cdot 8}{1}\right)^2 = 15.68^2 = 245.86 \approx 246$$

$z_{0.025}=1.96$ $E=10/12=0.83$ $\sigma = 8$

$$n = \left(\frac{z_{0.025} \cdot 8}{0.83}\right)^2 = \left(\frac{1.96 \cdot 8}{0.83}\right)^2 = 18.89^2 = 356.83 \approx 357$$

P

POTREBNA VELIČINA UZORKA za procjenu proporcije

Ovisit će o:

- pogrešci procjene koju ćemo tolerirati
- stupnju pouzdanosti
- pretpostavljenoj proporciji

$$E = z \cdot s_p = z \cdot \sqrt{\frac{p \cdot q}{n}}$$

POGREŠKA PROCJENE

$$n = \left(\frac{z}{E}\right)^2 \cdot p \cdot q$$

POTREBNA VELIČINA UZORKA

PRIMJER.Studija provedena na Fakultetu javnog zdravstva na Harvardu utvrdila je da 19% studenata nikada ne piju alkohol. Koliki uzorak vam je potreban za procjenu proporcije studenata koji ne piju alkohol na vašem fakultetu unutar raspona od 10% uz pouzdanost od 95%, vodeći se rezultatima harvardske studije?

$z_{0.025}=1.96$ $E=0.1/2=0.05$ $p = 0.19$

$$n = \left(\frac{z}{E}\right)^2 \cdot p \cdot q = \left(\frac{1.96}{0.05}\right)^2 \cdot 0.19 \cdot 0.81 = 39.3^2 \cdot 0.19 \cdot 0.81 = 236.49 \approx 237$$

p = 0.5 **- koristimo kada nemamo prethodnih saznanja o pretpostavljenoj proporciji**

$z_{0.025}=1.96$ $E=0.1/2=0.05$ $p = 0.5$

$$n = \left(\frac{z}{E}\right)^2 \cdot p \cdot q = \left(\frac{1.96}{0.05}\right)^2 \cdot 0.5 \cdot 0.5 = 39.3^2 \cdot 0.25 = 384.15 \approx 385$$

ZADATAK : **Farmaceutska tvrtka predlaže novi lijek za ublažavanje simptoma PMS-a. U prvim kliničkim istraživanjima lijek se pokazao učinkovit kod 7 od 10 žena.**

a) **izračunajte pogrešku procjene proporcije populacije uz pouzdanost od 95%**

b) **konstruirajte 95% raspon pouzdanosti za proporciju populacije**

c) **izračunajte pogrešku procjene i konstruirajte 95% raspon pouzdanosti proporcije populacije za istu proporciju dobivenu iz uzorka od 100 ispitanica.**

ZADATAK :

n=10 **p=7/10=0.7** **q=1-0.7=0.3**

a) **izračunajte pogrešku procjene proporcije populacije uz pouzdanost od 95%**

$$E = z \cdot s_p = z \cdot \sqrt{\frac{p \cdot q}{n}} = 1.96 \cdot \sqrt{\frac{0.7 \cdot 0.3}{10}} = 0.284$$

b) **konstruirajte 95% raspon pouzdanosti za proporciju populacije**

$$0.416 \leq \Pi \leq 0.984$$

ZADATAK :

c) izračunajte pogrešku procjene i konstruirajte 95% raspon pouzdanosti proporcije populacije za istu proporciju dobivenu iz uzorka od 100 ispitanica.

n=100 p=0.7 q=1-0.7=0.3

$$E = z \cdot s_p = z \cdot \sqrt{\frac{p \cdot q}{n}} = 1.96 \cdot \sqrt{\frac{0.7 \cdot 0.3}{100}} = 0.09$$

$$0.610 \leq \Pi \leq 0.790$$

ANALIZA KVALITATIVNIH PODATAKA

TABLICA KONTINGENCIJE

- tablica koja u retcima i stupcima sadrži frekvencije atributivnih obilježja
- predstavlja empirijsku razdiobu frekvencija obilježja mjerenih nominalnom ili ordinalnom ljestvicom mjerenja

TABLICA S "JEDNIM ULAZOM" (1×k)

- opažanja su klasificirana samo po jednom obilježju

PRIMJER.

	GODINA STUDIJA						
	I	II	III	IV	V	VI	UKUPNO
BROJ STUDENATA	64	48	32	28	18	15	205

TABLICA S "DVA ULAZA" (r×k)

- opažanja klasificirana po više atributa
- opažanja iz više uzoraka klasificirana po kategorijama jednog atributa

2×2 ....najjednostavnija tablica s "dva ulaza"

obilježje A	obilježje B		UKUPNO
	DA	NE	
DA	n_{11}	n_{12}	n_{1y}
NE	n_{21}	n_{22}	n_{2y}
UKUPNO	n_{x1}	n_{x2}	n_{xy}

- može se promatrati kao:
  - jedan uzorak (sa n_{xy} ispitanika)
  - dva uzorka (sa n_{1y}, n_{2y} ispitanika)

χ^2 TEST

- ocjena slaganja s poznatom razdiobom
- ocjena razlike razdiobe kategoričkog svojstva u nezavisnim uzorcima
- ocjena razlike dihotomnog svojstva u zavisnim uzorcima

χ^2 TEST ZA OCJENU SLAGANJA S POZNATOM RAZDIOBOM

- uz unaprijed poznatu razdiobu očekivanih frekvencija, test statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

gdje je: O_i .....opažena frekvencija
E_i .....očekivana frekvencija
k .....broj kategorija

ima χ^2 razdiobu s **df= k-1-m** stupnjeva slobode

k ... broj kategorija

m ... broj parametara u modelu koje treba procijeniti
m=1 za Poissonovu i binomnu raspodjelu
m=2 za normalnu raspodjelu

- **uz**

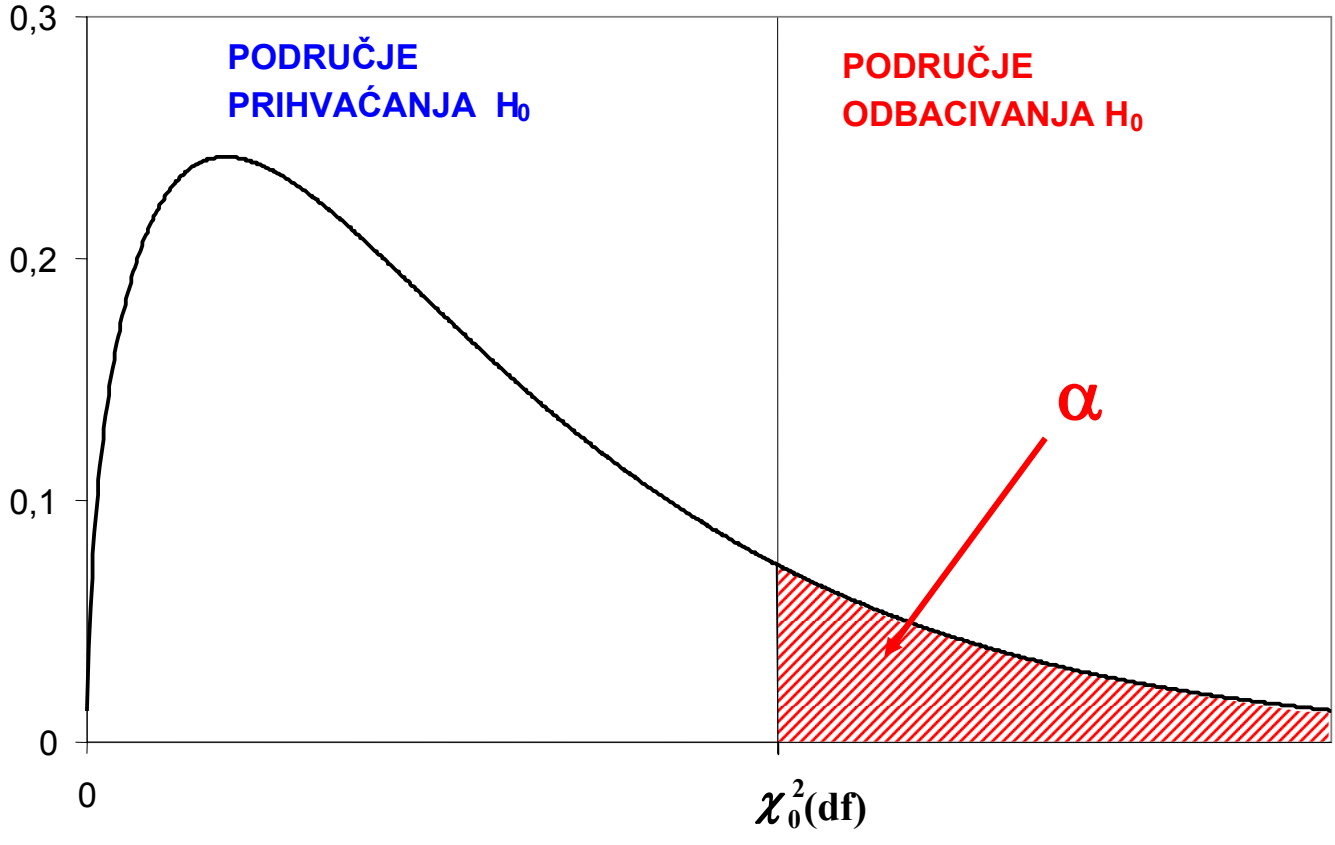
  H_0 ... nema razlike u razdiobi O_i i E_i

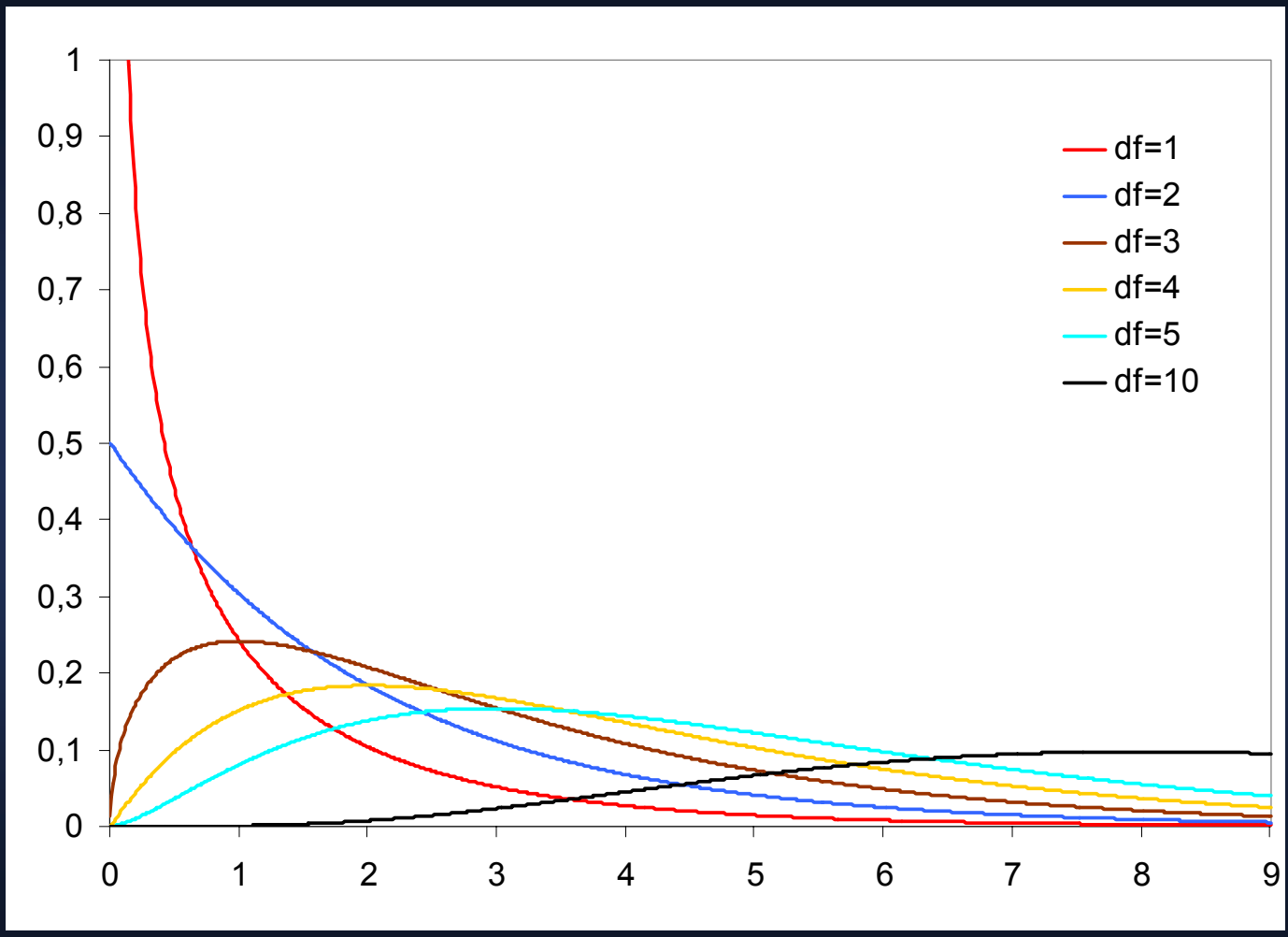
  granični χ^2 za dani α i df

za

$$\chi^2 > \chi^2_{(1-\alpha)} \Rightarrow P(\chi^2) < P(\chi^2_{(1-\alpha)})$$ **ODBACI H_0**

$$\chi^2 < \chi^2_{(1-\alpha)} \Rightarrow P(\chi^2) > P(\chi^2_{(1-\alpha)})$$ **PRIHVATI H_0**





Križanjem dviju vrsta biljki dobivena je u slijedećoj generaciji ova razdioba opaženih genotipova:

genotip	opažene frekvencije
Aa	53
AA	23
aa	24

Odgovara li ova razdioba očekivanoj razdiobi 2:1:1 uz α=0.01 ?

genotip	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Aa	53	50	3	9	0.18
AA	23	25	-2	4	0.16
aa	24	25	-1	1	0.04
					0.38

$\chi^2 = 0.38$ $k = 3;\ m = 0$ $df = 3 - 1 - 0 = 2$

χ^2 RAZDIOBA

df \ p	0.99	0.98	0.95	0.9	0.8	0.7	0.5
1	6.635	5.412	3.841	2.706	1.642	1.074	0.455
2	9.210	7.824	5.991	4.605	3.219	2.408	1.386
3	11.345	9.837	7.815	6.251	4.642	3.665	2.366
4	13.277	11.668	9.488	7.779	5.989	4.878	3.357

za df=2: $\chi^2_{(1-\alpha)} = \chi^2_{(0.99)} = 9.210$

$\chi^2 < \chi^2_{(0.99)} \Rightarrow P(\chi^2) > P(\chi^2_{(0.99)})$ **PRIHVATI H_0**

Radioaktivna tvar je promatrana tijekom 2608 jednakih vremenskih intervala. Za svaki interval registriran je broj čestica koje padnu u brojač. U tablici su dani brojevi intervala Oi u koje padne točno x čestica. Slažu li se podatci s Poissonovom raspodjelom na razini značajnosti od 0.05

x	O i
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	16

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}, \quad P(0) = e^{-\mu}$$

vrijedi:

$$p(x) = \frac{\mu}{x} p(x-1)$$

x	Oi	x*Oi	μ/x	p(x)	Ei=n*p(x)	Oi-Ei	$(Oi-Ei)^2$	$(Oi-Ei)^2/Ei$
0	57	0	-	0.0209	54.51	2.49	6.2001	0.1137
1	203	203	3.87	0.0809	210.99	-7.99	63.8401	0.3026
2	383	766	1.94	0.1569	409.2	-26.2	686.4400	1.6775
3	525	1575	1.29	0.2024	527.86	-2.86	8.1796	0.0155
4	532	2128	0.97	0.1963	511.95	20.05	402.0025	0.7852
5	408	2040	0.77	0.1512	394.33	13.67	186.8689	0.4739
6	273	1638	0.65	0.0983	256.37	16.63	276.5569	1.0787
7	139	973	0.55	0.0541	141.09	-2.09	4.3681	0.0310
8	45	360	0.48	0.0260	67.81	-22.8	520.2961	7.6729
9	27	243	0.43	0.0112	29.21	-2.21	4.8841	0.1672
10	16	160	0.39	0.0044	11.48	4.52	20.4304	1.7797
n=	2608	10086		1.0026	2614.8		$\chi 2=$	14.0979

$$\mu = \frac{\sum_{i=0}^{10} x_i f_i}{\sum_{i=0}^{10} f_i} = \frac{1}{n}\sum_{i=0}^{10} x_i f_i = 3.87$$

k=11
m=1 (za Poissonovu raspodjelu)
df=11-1-1=9
α=0.05

$$\chi^2 = 14.0979$$

χ^2 DISTRIBUCIJA

df \ p	0,99	0,98	0,95	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,02
1	6,635	5,412	3,841	2,706	1,642	1,074	0,455	0,148	0,064	0,016	0,004	0,001
2	9,210	7,824	5,991	4,605	3,219	2,408	1,386	0,713	0,446	0,211	0,103	0,040
3	11,345	9,837	7,815	6,251	4,642	3,665	2,366	1,424	1,005	0,584	0,352	0,185
4	13,277	11,668	9,488	7,779	5,989	4,878	3,357	2,195	1,649	1,064	0,711	0,429
5	15,086	13,388	11,070	9,236	7,289	6,064	4,351	3,000	2,343	1,610	1,145	0,752
6	16,812	15,033	12,592	10,645	8,558	7,231	5,348	3,828	3,070	2,204	1,635	1,134
7	18,475	16,622	14,067	12,017	9,803	8,383	6,346	4,671	3,822	2,833	2,167	1,564
8	20,090	18,168	15,507	13,362	11,030	9,524	7,344	5,527	4,594	3,490	2,733	2,032
9	21,666	19,679	16,919	14,684	12,242	10,656	8,343	6,393	5,380	4,168	3,325	2,532
10	23,209	21,161	18,307	15,987	13,442	11,781	9,342	7,267	6,179	4,865	3,940	3,059

za df=9: $\chi^2_{(1-\alpha)} = \chi^2_{(0.95)} = 16.919$

$$\chi^2 < \chi^2_{(0.95)} \Rightarrow P(\chi^2) > P(\chi^2_{(0.95)})$$ **PRIHVATI H_0**

χ^2 TEST ZA NEZAVISNE UZORKE

postupak:

- formirati tablicu kontingencije (r x k)
- na osnovu postavljene hipoteze izračunati očekivane frekvencije
- test statistika dana je sa:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

gdje je: r .....broj redaka
k .....broj stupaca

ima χ^2 razdiobu s
df= (r-1)·(k-1) stupnjeva slobode

VAŽNE NAPOMENE

a) u tablicu smijemo unijeti SAMO APSOLUTNE FREKVENCIJE

b) uzorci moraju biti nezavisni

c) u 2x2 tablici:

- ako je N<40:
  - NITI JEDNA očekivana frekvencija ne smije biti <5
  - preporuča se uvesti Yatesovu korekciju (umanjiti svaku razliku O-E prije kvadriranja)

d) ako je u r x k tablici E < 5 u više od 20% polja, NE MOŽEMO KORISTITI χ^2 TEST

rješenje:

- spajanje susjednih razreda (frekvencija susjednih polja)
- Fisherov egzaktni test

Pri istraživanju djelovanja nekog cjepiva, opažena je sljedeća učestalost oboljenja kod određene grupe ljudi:

	cijepljeni	necijepljeni	ukupno
oboljeli	3	10	13
nisu oboljeli	144	117	261
ukupno	147	127	274

Postoji li povezanost između učestalosti bolesti i cijepljenja (je li učestalost bolesti jednaka kod cijepljenih i necijepljenih) uz $\alpha = 0.01$?

H_0 .... učestalost je ista kod cijepljenih i necijepljenih

iz H_0 => proporcije oboljelih trebaju biti jednake u obje skupine

zajednička proporcija oboljelih: $zpo = \frac{13}{274} = 0.0474$

zajednička proporcija zdravih: $zpz = \frac{261}{274} = 0.9526$

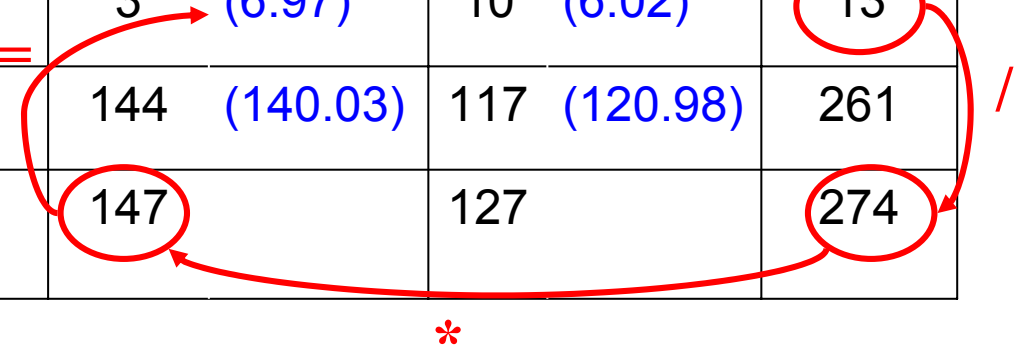
E oboljelih: u grupi cijepljenih..... 147*0.0474= 6.97

u grupi necijepljenih.. 127*0.0474= 6.02

E zdravih: u grupi cijepljenih..... 147*0.9526= 140.03

u grupi necijepljenih.. 127*0.9526= 120.98

	cijepljeni	necijepljeni	ukupno
oboljeli	3 (6.97)	10 (6.02)	13
nisu oboljeli	144 (140.03)	117 (120.98)	261
ukupno	147	127	274



Oi	Ei	Oi - Ei	$(Oi-Ei)^2$	$(Oi-Ei)^2/Ei$
3	6.97	-3.97	15.7609	2.26
10	6.02	3.98	15.8404	2.63
144	140.03	3.97	15.7609	0.11
117	120.98	-3.98	15.8404	0.13
				χ^2= 5.13

Yates-ova korekcija:

Oi	Ei	$(Oi - Ei)_{corr}$	$(Oi-Ei)_{corr}^2$	$(Oi-Ei)_{corr}^2/Ei$
3	6.97	-3.47	12.0409	1.73
10	6.02	3.48	12.1104	2.01
144	140.03	3.47	12.0409	0.09
117	120.98	-3.48	12.1104	0.10
				χ^2= 3.93

$\chi^2 = 3.93$

za $\alpha = 0.01$, df=1: $\chi_0^2 = 6.635$

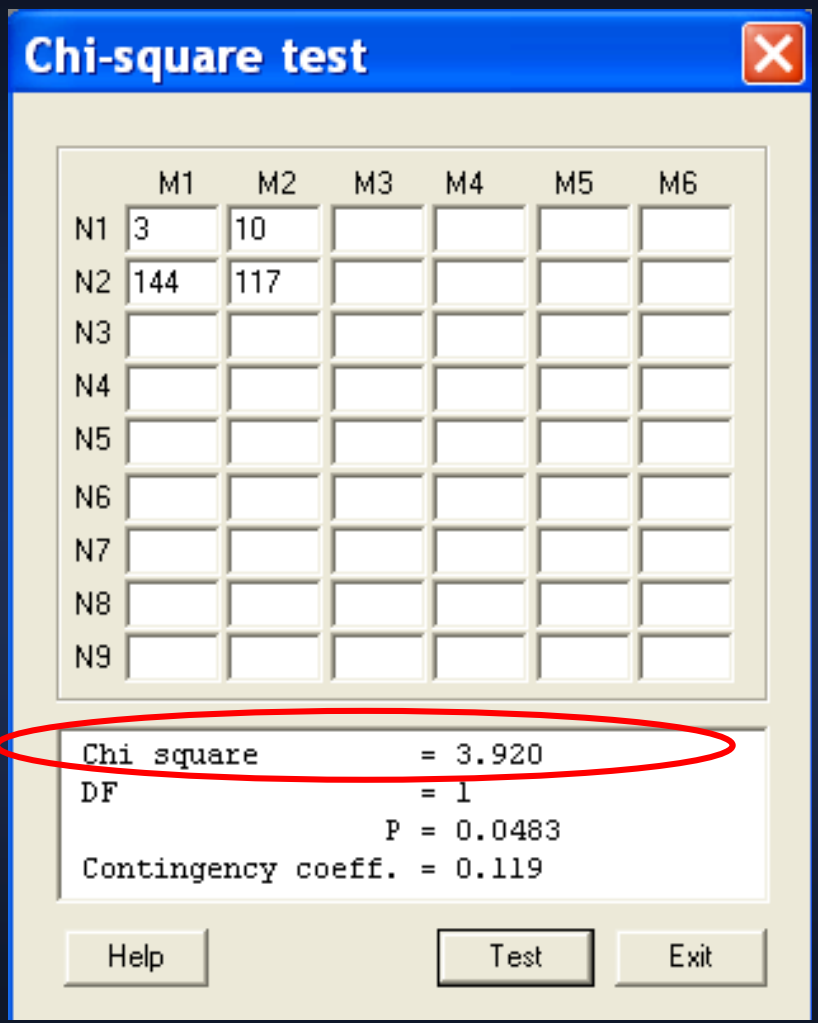
$\chi^2 < \chi_0^2 \Rightarrow P > 0.01$

$\Rightarrow$ ne postoji povezanost između učestalosti bolesti i cijepljenja

χ^2 TEST - MedCalc:

Tests-> Chi-square test

- za tablice 2x2, MedCalc primjenjuje Yatesovu korekciju



ZADATAK :

Ispitivana je ćud (benignost/malignost) tumora mozga prema lokalizaciji. Od 100 bolesnika s benignim tumorom, tumor je bio kod 21 lociran na frontalnom, kod 28 na temporalnom, a kod ostalih na drugim režnjevima mozga. Od 50 bolesnika s malignim tumorom kod 19 se radilo o tumoru frontalnog, kod 2 temporalnog a kod 29 o tumoru ostalih režnjeva mozga. Ocijenite postoji li povezanost malignosti s lokalizacijom tumora na mozgu na razini značajnosti od 0.05.

OPAŽENE FREKVENCIJE	Frontalni	Temporalni	Ostali	Ukupno
Benigni	21	28	51	100
Maligni	19	2	29	50
Ukupno	40	30	80	150

100*40/ 150 100*30/150 100*80/150

OČEKIVANE FREKVENCIJE	Frontalni	Temporalni	Ostali	Ukupno
Benigni	26.67	20.00	53.33	100.00
Maligni	13.33	10.00	26.67	50.00
Ukupno	40.00	30	80.00	150.00

50*40/150 50*30/150 50*80/150

Excel:

=A2-B2

=C2^2
ili
=POWER(C2;2)

=D2/B2

	A	B	C	D	E
1	Oi	Ei	Oi-Ei	$(Oi-Ei)^2$	$(Oi-Ei)^2/Ei$
2	21	26.67	-5.67	32.1489	1.2054
3	19	13.33	5.67	32.1489	2.4118
4	28	20.00	8.00	64.0000	3.2000
5	2	10.00	-8.00	64.0000	6.4000
6	51	53.33	-2.33	5.4289	0.1018
7	29	26.67	2.33	5.4289	0.2036
8	150	150		$\chi 2=$	13.5226
				P=	0.001

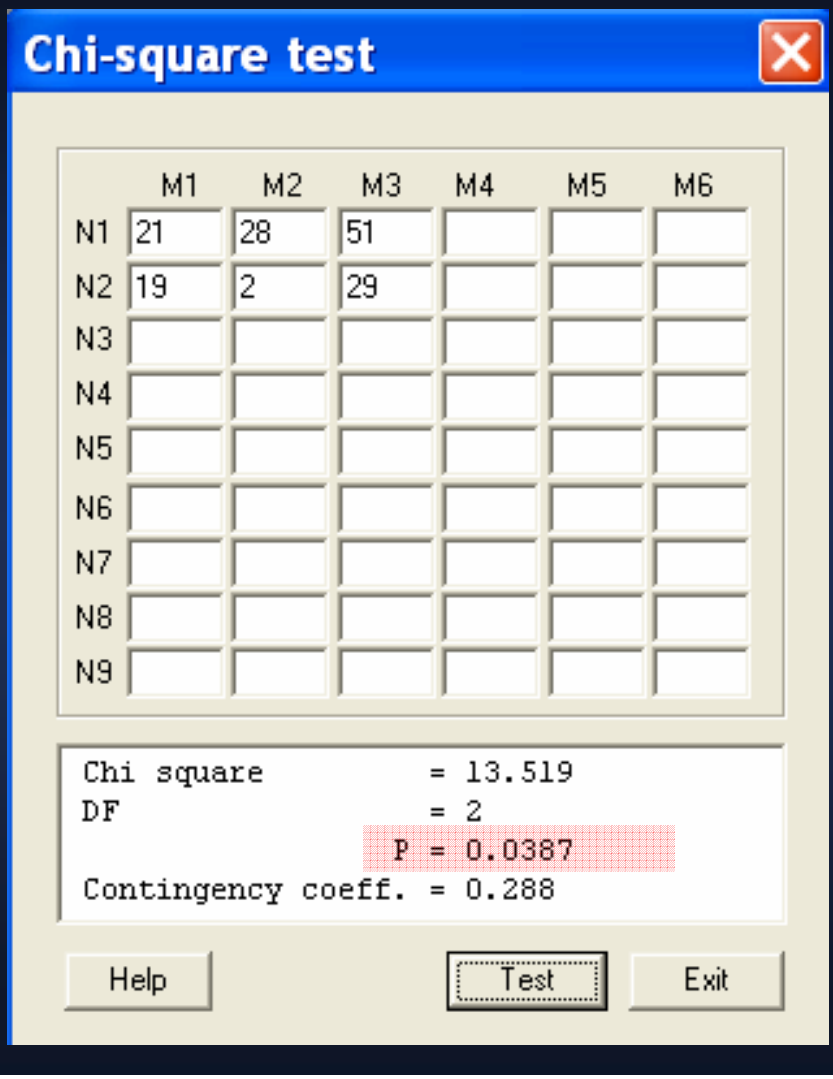
Σ

df = (3-1)(2-1)=2

=CHIDIST(E8;2)

χ^2 TEST - MedCalc: (ver 4.1)

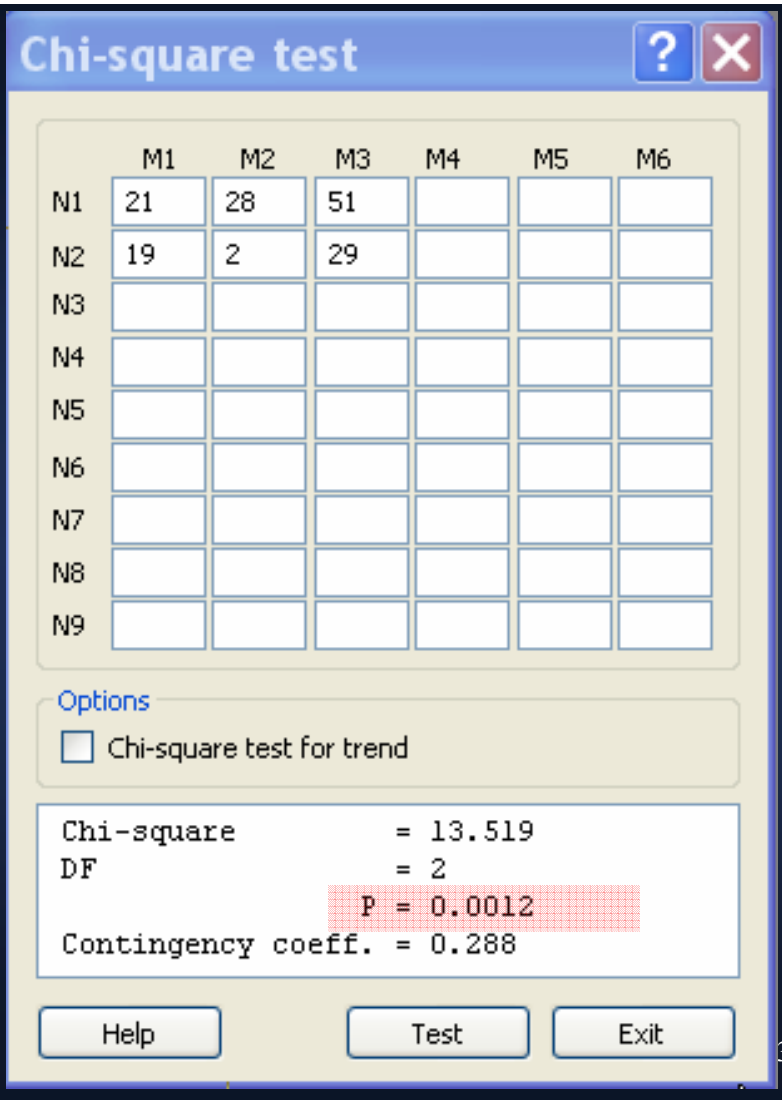
Tests-> Chi-square test



PDDS MOLBIO

χ^2 TEST - MedCalc: (ver 8.0)

Tests-> Chi-square test



χ^2 TEST ZA ZAVISNE UZORKE (McNemarov test)

- testiranje značajnosti razlike (ili vjerojatnosti povezanosti) između podataka dobivenih na uzorcima parova

$$\chi^2 = \frac{(|b-c| - 1)^2}{b+c}$$

b, c .... frekvencije parova koji se **ne slažu** po prisutnosti obilježja

OBILJEŽJE A		UZORAK I	
		DA	NE
UZORAK II	DA	a	b
	NE	c	d

Iz skupine od 150 bolesnika formirano je 75 parova. Jedan član para liječen je novom, a drugi standardnom terapijom.

		standardna terapija		
		poboljšano	ne poboljšano	ukupno
nova terapija	poboljšano	40	25	65
	ne poboljšano	10	0	10
	ukupno	50	25	75

Postoji li razlika u učinkovitosti standardne i nove terapije uz $\alpha = 0.05$?

$$\chi^2 = \frac{\left(|25-10|-1\right)^2}{25+10} = \frac{14^2}{35} = 5.60$$

$$df = 1, \quad \alpha = 0.05$$

$$\chi_0^2 = 3.84$$

$$\chi^2 > \chi_0^2 \Rightarrow p < 0.05$$