

Priprema podataka za statističku obradu

Izv. prof. dr. sc. Gordana Brumini
Katedra za medicinsku informatiku

Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

1

Najčešća pitanja?

- Koliki ispitanika treba za doktorat, je li dovoljno 100 ispitanika?
- Kako treba upisivati podatke?
- Koje varijable koristiti?
- Jesu li ove varijable dobre/dovoljne?

Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

2

Koliko ispitanika treba tj. koliki N treba?

- Procjena veličine uzorka
- Snaga testa
- Razina značajnosti

Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

3

Razina (statističke) značajnosti

- $P < 0,05$
 - Nul-hipoteza se odbacuje, postoji statistička značajnost
 - u 95 % slučajeva statistička značajnost postoji, u 5 % slučajeva nema značajnosti
- $P \geq 0,05$
 - Nul-hipoteza se prihvaća, ne postoji statistička značajnost
 - u 95 % slučajeva statistička značajnost ne postoji, u 5 % slučajeva postoji

Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

4

Power analysis

5

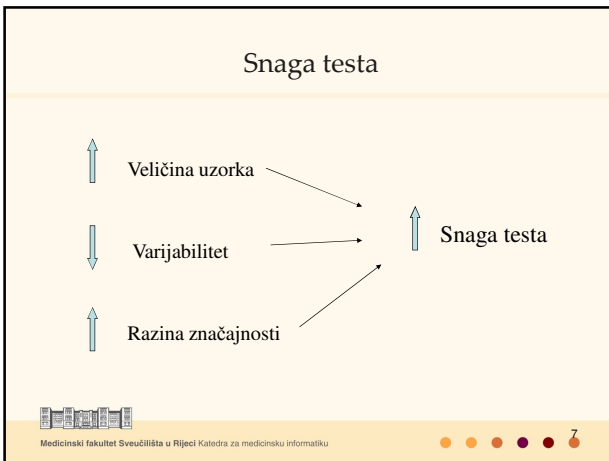
Snaga testa

Veličina je uzorka u pozitivnoj korelaciji sa snagom testa (Grafikon 1.), a općenito je prihvaćeno da snaga testa ne bi trebala biti manja od 0,8 (O'Brien i Wang, 2007) za prag značajnosti $\alpha = 0,05$. Optimalna veličina uzorka znači da uzorak treba biti dovoljno velik da zadovolji pretpostavke potrebne za izvođenje preciznih zaključaka, a opet ne previše: zbog prevencije prihvatanja lažne nulte hipoteze (false discovery rate - FDR).

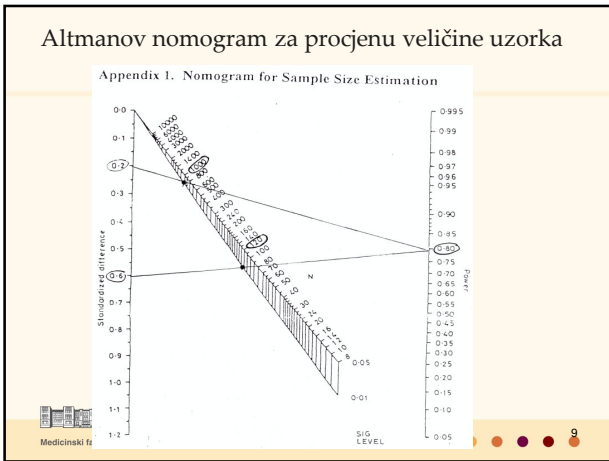
Grafikon 1. Odnos veličine uzorka i snage testa
Figure 1. Relation of sample size and power

Confalonieri i sur. (2006) ističu da bi napravljuje bilo procijeniti potreban veličinu uzorka na osnovi rezultata preliminarnih istraživanja, ali da se to rijetko čini. Veličina uzorka kvalitativnih svojstava

6

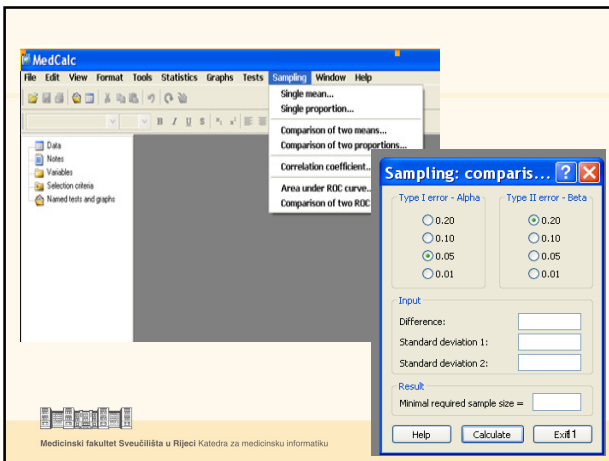


- ### Izračun veličine uzorka
- Altmanov nomogram
 - Programi
 - Psiholozi: pravilo → broj varijabli * 5 = N
- 8



Gdje se može učiniti izračun veličine uzorka?

10



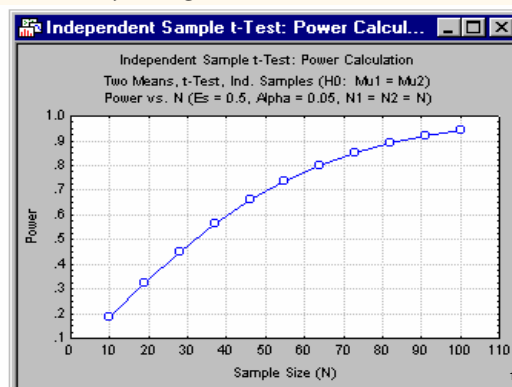
- ### Snaga testa
- Snaga testa ocjenjuje kvalitetu samog testa
 - Definiira koliki uzorak je potreban za statistički zaključak
 - Koliko je vjerojatno da se nekim statističkim testom određuje mjereno svojstvo
- 12

Analiza snage testa

- "dobar test"- snaga bi trebala biti 70-80 %
- etički neprihvatljivo, gubitak vremena i novaca provoditi istraživanje koje ima 40 % vjerojatnost da otkriva efekt koji se istražuje
- pr. mjerenje stava s upitnikom (snaga testa 80 %)
 - u 80 % postoji vjerojatnost da se tim upitnikom doista mjeri stav
 - u 20 % postoji vjerojatnost da taj upitnik ne mjeri stav



Krivulja snaga testa i veličina uzorka



Koliko ispitanika treba tj. koliki N treba?

- Procjena veličine uzorka
- Treba 2 parametra
 - snaga testa
 - razina značajnosti



Podaci

- Kvantitativni
- Omjerene i intervalne mjerne ljestvice
- Mjerljivi i imaju mjernu jedinicu
- 10 cigareta na dan
- BMI
- 45 godina
- Kvalitativni
- Nominalne i ordinalne mjerne ljestvice
- Opisni, bez mjerne jedinice
- pušači i nepušači
- opekline I. -IV. stupnja



Nominalna ljestvica

- Nazočnost prisustva na nastavi
 - da/ne
- Pušenje
 - da/ne
- Kardiovaskularna bolest
 - ima/nema
- Spol
 - muški, ženski



Ordinalna ljestvica

- Ocjenjivanje studenata 1-5 ili A - E
- Razredi u smislenom (rastućem) nizu
- Širina razreda može biti nejednaka
- Katkada se mogu tretirati kao podaci iz intervalne mjerne ljestvice
- Opekline od I do IV stupnja
- Stadiji bolesti (Ca,..)



Intervalna ljestvica

- Izražava količinu → brojna vrijednost odgovara vrijednosti obilježja
- Može imati mjernu jedinicu
- Temperatura , pH vrijednost, IQ,
- “nula” → dogovorena vrijednost (pr. 0 °C)
- $t_1 = -10\text{ °C}$, $t_2 = 20\text{ °C}$, nije 2 puta veća nego je razlika 30 °C



Omjerna ljestvica

- Kvantitativni podatak – brojnost jedinice mjerenja
- Masa , dob, koncentracija, broj studenata na I. godini
- Vrijednost “nula” → nepostojanje obilježja
- Omjer → omogućuje procjenu vrijednosti
 - npr: BMI ... dvostruko veći od ...



Pravila za prikupljanje podataka

- što više varijabli
 - za multivarijantne analize treba više zavisnih varijabli
 - izbjegavati kategorijske varijable
- kvantitativni podaci mogu se prevesti u kvalitativne
- čuvati izvorne podatke
- po završetku EU projekta revizija i nakon 5 godina, podaci iz doktorata do 10 godina



Kako upisati podatke



Programi u kojima možete upisati podatke

- MS Excel
- MedCalc
- Statistica



Kako pravilno upisati podatke u Excel

- Sve podatke upisati u jedan radni list (Sheet)
- Izbjegavati boje
- Objasniti legendu
- Što više varijabli imati
 - posebice kontinuirane podatke
- Voditi računa o ciljevima istraživanja



Primjer loše tablice

The screenshot shows an Excel spreadsheet with a table containing patient data. The table has several columns with headers like 'RUT HALURS VALGUS', 'RUT IZMEĐU I I METATARZALNE KOSTI', and 'RUT IZMEĐU II I METATARZALNE KOSTI'. The data is presented in a dense, repetitive format that is difficult to read.

Oblikovanje kategorijskih varijabli

Legenda: za stupanj parodontne bolesti

- 1 - max dubina džepova do 4 mm
- 2 - max dubina 4-6 mm
- 3 - Max dubina >6mm i furkacijski defekti

~~Legenda: (za broj leukocita)~~

- ~~1-normalan~~
- ~~2-povišen~~
- ~~3-snižen~~

Legenda: Candida albicans

- 0 - nema
- 1 - stupanj I
- 2 - stupanj II
- 3 - stupanj III



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

26

Oblikovanje kategorijskih varijabli

- 1- subklinička hipotireoza
- 2- hipotireoza
- 3- eutireoza
- 4- subklinička hipertireoza
- 5- hipertireoza
- može i tekstualno- paziti da se uvijek jednako piše



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

27

Kontinuirani podaci → kategorijski podaci

The screenshot shows a data table in STATISTICA software. The table has columns for '1 starost/godine', '2 spol', '3 skupina', '4 kategorija skupine', '5 kardiovaskularna bolest', '6 BMI', '7 kategorija BMI', '8 kategorija indeksa', and '9 stupanj parodontalne bolesti'. The data rows show various values for these variables, such as age (76, 40, 67, etc.), sex (Z, M), and BMI (22.86237, 26.59221, etc.).

Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

28

.. i obrnuto kada kategorijski podaci mogu postati kontinuirani

- mjerenje stava
- Likertova mjerna ljestvica od 1 do 5
 - 1 - uopće se ne slažem
 - 2 - uglavnom se ne slažem
 - 3 - niti se slažem niti se ne slažem
 - 4 - uglavnom se slažem
 - 5 - slažem se
- veliki N - dozvoljava da se ovi kategorijski podaci tretiraju kao kontinuirani podaci



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

29

Zašto su bolji kontinuirani podaci

- Parametrijski testovi
- Veća snaga testa
- Multiple analize



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

30

Dobro upisani podaci

Dobro upisani podaci

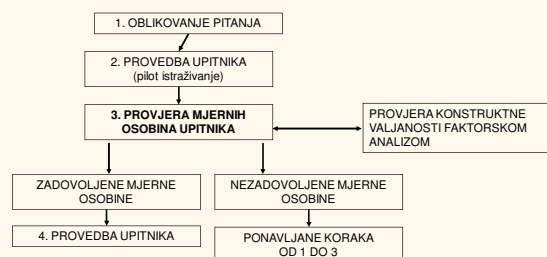
Unos podataka iz Excela u Statisticu

Podaci u MedCalc

Oblikovanje upitnika

- Definiranje ciljeva
- Odabir uzorka ispitanika
- Oblikovanje pitanja
 - otvorenog i/ili zatvorenog tipa- Likertova skala
- Provedba
- Provjera mjernih osobina

Koraci u oblikovanju upitnika



Oblikovanje pitanja

- Jasna, kratka i jednoznačna pitanja
- Izbjegavati pristrane ponuđene odgovore
 - Pušite li? Da/Ne
- Obratiti pažnju na izraze
 - sinonimi, višeznačnice, pozitivna/negativna pitanja
- Izbjegavati neugodna i hipotetska pitanja
- Poštovati privatnost i anonimnost



Primjeri pitanja

- Pušite li:
 - da
 - ne
 - svakodnevno
 - povremeno
- Računalom se koristite u svrhu:
 - posla
 - igre
 - kupovine
- Pušite li:
 - ne
 - povremeno
 - svakodnevno
- Koristite li računalao za rad:
 - da
 - ne



Primjer pitanja u upitniku 1

1. Dob (godina)
2. Spol M Ž
3. Ustanova u kojoj radite (ukoliko radite u više ustanova, upišite onu u kojoj ste zaposleni u većinskom udjelu):
 - Fakultet
 - Znanstveni institut
 - Bolnica ili druga zdravstvena ustanova



Primjer pitanja u upitniku 2

1.	Katkad sam u iskušenju da plagiram jer svi ostali (studenti, istraživači, liječnici) to čine.	1	2	3	4	5
2.	Plagiranje osiromašuje istraživački duh.	1	2	3	4	5
3.	Prekratki rokovi daju mi pravo na malo plagiranja.	1	2	3	4	5
4.	Samoplagiranje NE bi trebalo kažnjavati kao plagiranje.	1	2	3	4	5



Pisana uputa 1

- Molim Vas da budete sudionici ovog istraživanja. Istraživanje će obuhvatiti sve učenike prvog razreda srednje škole Primorsko-goranske županije. Ovim se upitnikom ispituju stavovi ljudi o pušenju.

Unaprijed Vam hvala na suradnji!



Pisana uputa 2

Molim Vas da budete sudionici u istraživanju **kojega provodim u sklopu svojeg doktorskog rada**. Istraživanje će obuhvatiti sve učenike prvog razreda srednje škole Primorsko-goranske županije. Ovim se upitnikom ispituju **razmišljanja** ljudi o pušenju.

Ispitivanje je anonimno, ne zanima nas Vaše ime i prezime, nitko neće biti upoznat s Vašim odgovorima.

Unaprijed Vam hvala na suradnji!



Pouzdanost upitnika

- Određuje pogreške mjerenja u opaženom rezultatu.
- Stupanj u kojem su rezultati dosljedni, sigurni ili ponovljivi (Croanbach koeficijent $\alpha > 0,7$)
- Savršeno pouzdan test - rezultati svakog mjerenja su identični (Croanbach koeficijent $\alpha=1$)



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

43

ROC analiza

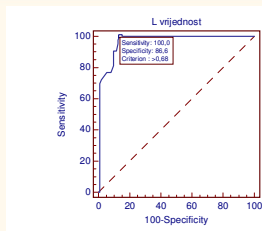
- engl. Receiver Operating Characteristic Curve
- Za određivanje valjanosti dijagnostičke metode
- Valjanost je sposobnost dijagnostičkog testa da pravilno klasificira ispitanike u bolesne i zdrave
- Određuje se temeljem
 - Osjetljivosti testa (vjerojatnost pozitivnog nalaza testa uz prisustvo bolesti)
 - Specifičnost testa



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

44

ROC krivulja



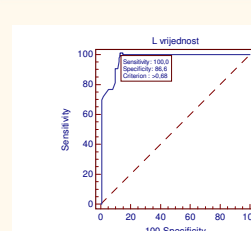
- daje podatke o:
 - osjetljivosti i specifičnosti testa
 - definira vrijednost kriterijskog prediktora (najmanji broj LP i LN)
- Kvadratić opisuje:
 - osjetljivost = 100 %, (SP/SP+LN)
 - specifičnost 86,6 % (LP)
 - kriterijski prediktor >0,68



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

45

Površina ispod ROC krivulje



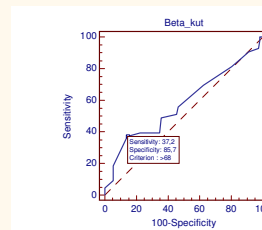
- Površina ispod ROC krivulje = 0,967; $P < 0,001$
- 0,9 - 1 = izvrstan test
- 0,8 - 0,9 = dobar test
- 0,7-0,8 = osrednji test
- 0,6 - 0,7 = slabiji test
- 0,5 - 0,6 = test bez uspjeha



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

46

ROC krivulja



- Površina ispod ROC krivulje = 0,579; $P = 0,133$

Površina ispod ROC krivulje	0,579
Standardna pogreška	0,052
95% granice pouzdanosti (CI)	0,497 do 0,657
Razina značajnosti P	0,133



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

47

Mutipla regresijska analiza

- za određivanje udjela i značajnosti doprinosa ispitivanih čimbenika na zavisnu varijablu

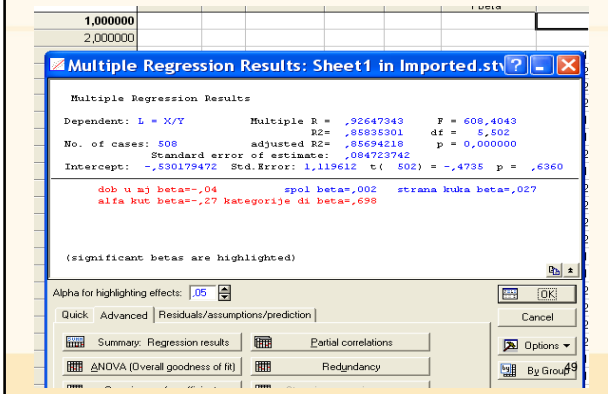
Čimbenik	β -koeficijent	P	Udio doprinosa (%)
Životna dob	-0,044	0,191	0,5
Spol	0,019	0,506	0,2
Školska sprema	0,079	0,028	1,0
Informatičko obrazovanje	0,085	0,012	1,5
Uporaba računala	0,302	<0,001	10,1



Medicinski fakultet Sveučilišta u Rijeci Katedra za medicinsku informatiku

48

Mutipla regresijska analiza u Statistici



Statistical analysis

- ..naziv odlomaka u radu u kojem pišete o statistici..
- na početku
 - Statistical analysis of data was performing by using Statistica for Windows, release 8.1 (Stasoft, INC., Tulsa, OK, USA).
- na kraju
 - All statistical values were considered significant at the P level of 0.05.

Članak u CMJ

- Lang T. Twenty statistical errors even YOU can find in biomedical research articles. CMJ.2004;45(4):361-370.

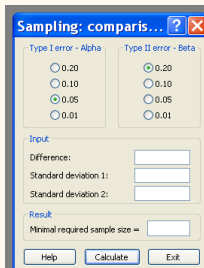
Hvala na pozornosti

Dva su načina na koji se može živjeti:
 Jedan je - kao da ništa nije čudo.
 Drugi je - kao da je sve čudo. (Einstein)



Pogreške testiranja hipoteze

- vjerojatnost pogreške I. vrste, α
 - odbacujemo istinitu nul-hipotezu (ako je $P < \alpha$)
- vjerojatnost pogreške II. vrste, β
 - prihvaćamo neistinitu nul-hipotezu, zaključujemo da nema efekta kada on stvarno postoji
- $1 - \beta$ = snaga testa



ROC analiza - izračun vrijednosti kriterijskog prediktora

Kriterij	Osjetljivost/%	95% CI	Specifičnost/%	95% CI	+LP	-LN
>=0,34	100,00	91,7 - 100,0	0,00	0,0 - 3,3	1,00	
>0,34	100,00	91,7 - 100,0	0,89	0,1 - 4,9	1,01	0,00
>0,35	100,00	91,7 - 100,0	1,79	0,3 - 6,3	1,02	0,00
>0,38	100,00	91,7 - 100,0	2,88	0,6 - 7,6	1,03	0,00
>0,43	100,00	91,7 - 100,0	3,57	1,0 - 9,9	1,04	0,00
>0,44	100,00	91,7 - 100,0	4,46	1,5 - 10,1	1,05	0,00
>0,45	100,00	91,7 - 100,0	6,25	2,6 - 12,5	1,07	0,00
>0,47	100,00	91,7 - 100,0	11,61	6,3 - 19,0	1,13	0,00
>0,48	100,00	91,7 - 100,0	13,39	7,7 - 21,1	1,15	0,00
>0,5	100,00	91,7 - 100,0	16,07	9,8 - 24,2	1,19	0,00
>0,51	100,00	91,7 - 100,0	17,86	11,3 - 26,2	1,22	0,00
>0,52	100,00	91,7 - 100,0	23,21	15,8 - 32,1	1,30	0,00
>0,53	100,00	91,7 - 100,0	27,68	19,6 - 36,9	1,38	0,00
>0,54	100,00	91,7 - 100,0	29,46	21,2 - 38,8	1,42	0,00
>0,55	100,00	91,7 - 100,0	37,50	28,5 - 47,1	1,60	0,00
>0,56	100,00	91,7 - 100,0	41,96	32,7 - 51,7	1,72	0,00
>0,57	100,00	91,7 - 100,0	52,88	43,0 - 62,2	2,11	0,00
>0,58	100,00	91,7 - 100,0	55,36	45,7 - 64,8	2,24	0,00
>0,59	100,00	91,7 - 100,0	58,04	48,3 - 67,3	2,38	0,00
>0,6	100,00	91,7 - 100,0	62,50	52,9 - 71,5	2,67	0,00
>0,61	100,00	91,7 - 100,0	64,29	54,7 - 73,1	2,80	0,00
>0,63	100,00	91,7 - 100,0	67,86	58,4 - 76,4	3,11	0,00
>0,64	100,00	91,7 - 100,0	69,64	60,2 - 79,0	3,29	0,00
>0,65	100,00	91,7 - 100,0	75,00	65,9 - 82,7	4,00	0,00
>0,66	100,00	91,7 - 100,0	83,04	74,8 - 89,5	5,89	0,00
>0,67	100,00	91,7 - 100,0	83,93	75,8 - 90,2	6,22	0,00
>0,68	100,00	91,7 - 100,0	86,61	78,9 - 93,3	7,47	0,00
>0,69	97,67	87,7 - 99,6	87,50	79,9 - 93,0	7,81	0,027