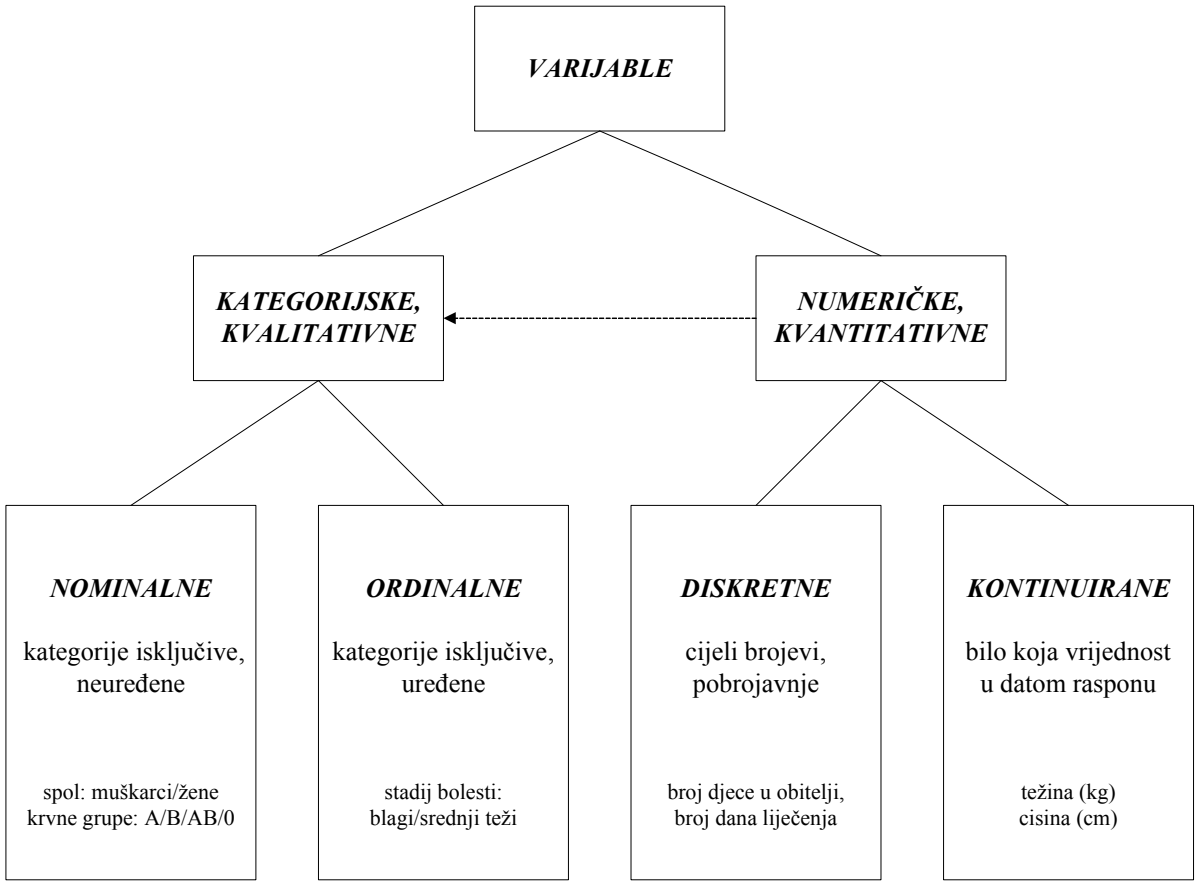
SREĐIVANJE MNOŠTVA PODATAKA

Davor Ivanković i Mirjana Kujundžić Tiljak

Liječnik u svom radu prikuplja mnoštvo podataka. Dovoljno je pogledati zdravstveni karton u ordinaciji ili povijest bolesti na bolničkom odjelu pa da se ustanovi da sve vrvi od podataka od kojih su neki liječniku praktičaru vrlo vrijedne informacije. Neki podaci, a pojavljuju se vrlo često pa se čak i ponavljaju, i ne moraju biti informativni. U jednom zdravstvenom kartonu nađe se i po nekoliko tisuća podataka.

Podaci su po svojoj prirodi različiti. Neka svojstva (obilježja, atributi, varijable) pobrojavamo, neka mjerimo. Spol primjerice biva označen riječima ili simbolom m/ž ili pak 1 za muško 2 za žensko ili binarizirano (tako će spol biti pretvoren u dvije varijable: muško 1 0 a žensko 0 1). Spol kao varijabla ne može primiti nikoju drugu vrijednost. Stanje neke bolesti u trenutku promatranja može biti pogoršano, nepromijenjeno ili poboljšano. Pritom imamo tri kategorije ali se za njih ne može reci da predstavljaju skalu sređenog intenziteta. Takva svojstva kao što su spol i spomenuti ishod bolesti zovemo *nominalno* „mjerenim“ svojstvima (lat. nomen, nominis = ime)
Intenzitet opeklina opisujemo u četiri stupnja. Stupanj predstavlja dogovorenu kategoriju intenziteta. Takovo svojstvo zovemo *ordinalno* „mjerenim“ svojstvom (lat. ordo, ordinis = red, vrsta). U nas uobičajeni sustav ocjena na ispitu (1-5) također je ordinalno mjereno svojstvo.
Uspoređujemo li intenzitet odnosno veličinu nekog svojstva s definiranim standardom, primjerice metrom, takvo svojstvo zovemo *intervalno* „mjerenim“. Intervalna skala mjerenja ima jedinicu intervala (primjerice kg, cm, mmol/L, kPa, mmHg, stupnjevi temperature i slično). Istovrijedna je na bilo kojem mjestu skale mjerenja.
Nominalno i ordinalno opisane atribute zovemo *kvalitavnima, atributivnima, nenumeričkima* odnosno *nemetričkima.*

Intervalno izmjerene atribute zovemo *kvantitativnima, numeričkima* odnosno *metričkima*. Ti pak mogu primiti *kontinuiranu* ili *diskretnu (diskontinuiranu)* vrijednost. U obitelji primjerice može biti jedno, dvoje, troje ili više djece, ali ne može 1,7 djece



Slika 1. Dijagram različitih tipova varijabli

Ovisno o tome da li su podaci kvalitativni ili kvantitativni, upotrebljavamo različite statističke metode. Premda je razlika između kategorijskih i numeričkih podataka uobičajeno jasna, u nekim situacijama može postati zamagljena. Na primjer, kad imamo varijablu s velikim brojem uređenih kategorija (npr. skala boli sa sedam kategorija) teško ju je razlikovati od diskretne numeričke varijable. Razlika između diskretnih i kontinuiranih numeričkih podataka može biti još nejasnija, premda općenito to će imati vrlo mali utjecaj na rezultate većine analiza. Dob je primjer varijable koja se često tretira kao diskretna iako je ona u potpunosti kontinuirana. Najčešće navodimo „dob zadnjeg rođendana“ pa tako žena u dobi od 30 godina može biti da je tek navršila 30 ili je već blizu 31 godine.
Numerički podaci nikada se ne smiju pohraniti u formi kategorijskih varijabli budući da se tako često gube značajne informacije. Naime, originalne numeričke podatke vrlo je jednostavno pretvoriti u kategorijske varijable, a obrnut postupak je nemoguć,

Iz originalnih podataka često se izvode novi tipovi podataka. Poboljšanje pacijenata u nekom postupku liječenja često prikazujemo u obliku *postotaka* (engl. percentages). Tako, na primjer, nakon primjene lijeka plućna funkcija pacijenta (forsirani ekspiratorni volumen u 1 sekundi, FEV1) može se povećati za 24% što je bolji prikaz stupnja poboljšanja nego apsolutna vrijednost FEV1 nakon tretmana. Nerijetko nailazimo na *omjere* ili *kvocijente* (engl. ratio or quotients) dvaju varijabli. Pri procjeni pretilosti upotrebljava se indeks tjelesne mase (engl. Body mass indeks, BMI) koji se računa kao omjer težine izražene u kg i kvadrata visine izražene u m^2. U epidemiološkim studijima učestalo se upotrebljavaju različite *stope* (engl. rates), kao na primjer broj oboljelih na 100.000 stanovnika određene regije za određeni

vremenski period. Serija odgovora na pitanja u vezi kvalitete života mogu se sumirati na određeni način te prikazati kao *skor* (engl. score) individualne procjene vlastite kvalitete života. Sve navedene *izvedene varijable* u većini analiza mogu se tretirati kao kontinuirane varijable. U slučaju kada se varijabla izvodi iz više od jedne vrijednosti izuzetno je važno zabilježiti sve originalne vrijednosti iz kojih je nova varijabla izvedena. Tako na primjer nije svejedno koja je početna vrijednost nekog obilježja bila ako je rezultat primjenjenog liječenja njegovo 10% poboljšanje.

U laboratorijskim mjerenjima nerijetka je pojava da se određenom laboratorijskom tehnikom mogu detektirati samo vrijednosti iznad određene granične vrijednosti (engl. cut-off value). Ukoliko je rezultat mjerenja razine nekog virusa „neotkrivljiv" (engl. undetectable), tj. ispod granične vrijednosti koja se registrira, to ne znači da u uzorku nije uopće prisutan virus. Opisani tip podataka naziva se *cenzorirani* (engl. cenzored). Kada iz određenih razloga ispitanici budu izgubljeni tijekom istaživanja podaci su također cenzorirani.

Postupci sređivanja podataka, ovisno o tome jesu li kvalitativni ili kvantitativni, različiti su. Tim su postupcima prilagođene i različite komercijalno dostupne programske podrške (Statistica, SAS, SPSS, S-Plus, Epi-Info i slično).

SREĐIVANJE KVANTITATIVNIH PODATAKA

Raspolažemo li s malim brojem podataka (mali uzorak) neposrednim uvidom u podatke možemo se lako orijentirati o ekstremnim vrijednostima, vidjeti koje su vrijednosti česte, da li se podaci grupiraju oko neke vrijednosti, kakvo je raspršenje podataka te kojim bi podatkom bilo najprimjerenije prezentirati čitav niz podataka.
Kada se pak radi o mnoštvu podataka ljudski mozak nema svojstvo neposrednog prepoznavanja obrasca (predloška, engl. pattern) po kojem se podaci raspoređuju. Tada primjenjujemo različite postupke sređivanja podataka.

Rezultat sređivanja podataka je *raspodjela, razdioba* odnosno *distribucija frekvencija*.

Dva su pristupa. Jedan je klasična, tradicionalna *tablica frekvencija* a drugi tzv. *stablo-list prikaz* (engl. stem and leaf), doduše informativniji ali rjeđe u uporabi.

Pri formiranju tablice frekvencija valja odrediti razrede (kategorije, klase) numeričkog obilježja i razmjestiti podatke u odgovarajući razred. Razredi se međusobno ne smiju preklapati. Nijedan podatak pritom ne smije biti izgubljen. Razredi se formiraju prema različitim kriterijima: ili tako da dobivene podskupine imaju fiziološko ili pak kliničko opravdanje ili tako da se odredi željeni broj razreda s istim *razrednim intervalom* (raspon između donje i gornje granice razreda) ili pak po međunarodnim konvencijama za dati problem kako bi se omogućila usporedivost.

Tablica 1.
Tablica frekvencija indeksa tjelesne mase izmjerenog kod 1229 odraslih muškaraca u dobi od 26 do 77 godina.

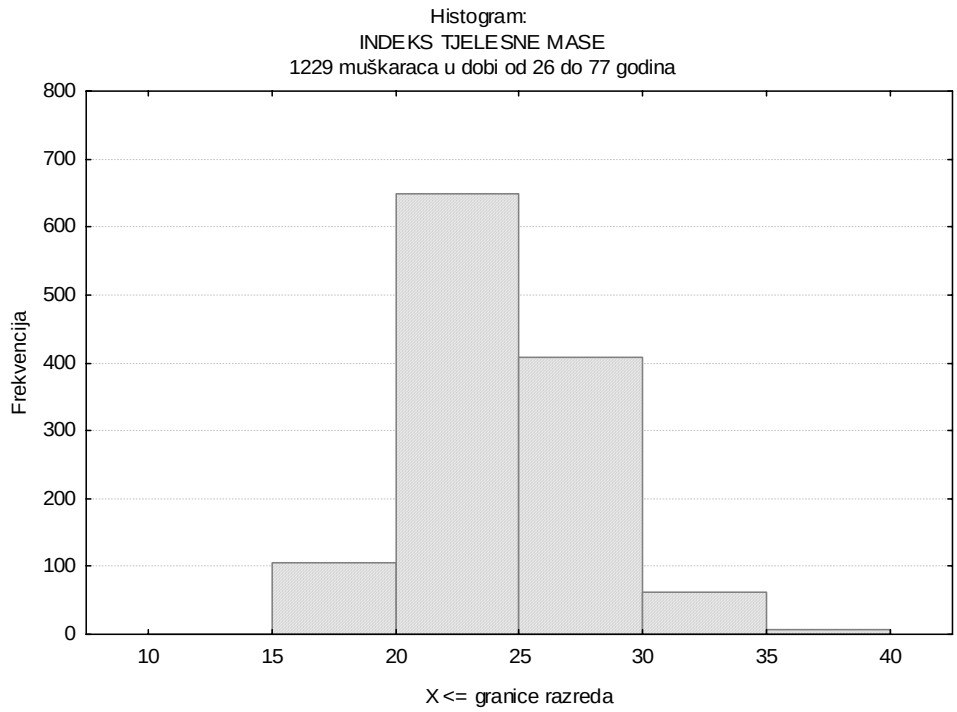
INDEKS TJELESNE MASE (BMI)	frekvencija	kumulativna frekvencija	relativna frekvencija (%)	kumulativna relativna frekvencija (%)
10,0< x ≤15,0	0	0	0,00	0,00
15,0< x ≤20,0	105	105	8,54	8,54
20,0< x ≤25,0	648	753	52,73	61,27
25,0< x ≤30,0	407	1160	33,12	94,39
30,0< x ≤35,0	61	1221	4,96	99,35
35,0< x ≤40,0	8	1229	0,65	100,00
Ukupno	1229		100,00	

Tablica 2.
Klinički reducirana tablica frekvencija indeksa tjelesne mase izmjerenog kod 1229 odraslih muškaraca u dobi od 26 do 77 godina (prema europskoj klasifikaciji stupnja pretilosti, Vrhovac – Interna medicina, str. 1445)

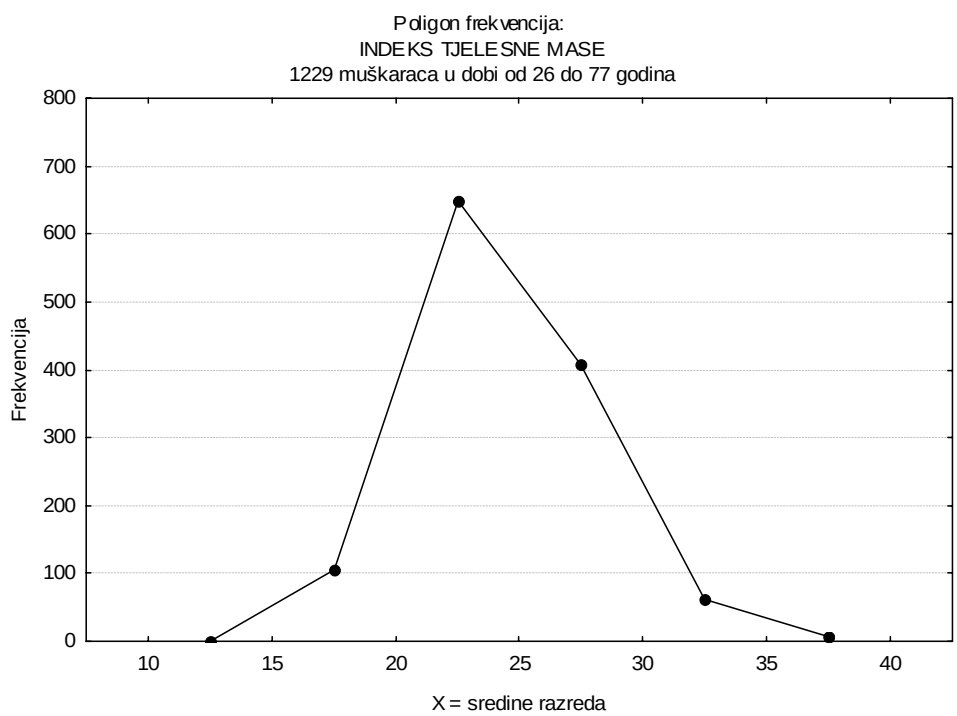
STUPNJEVI PRETILOSTI	frekvencija	kumulativna frekvencija	relativna frekvencija (%)	kumulativna relativna frekvencija (%)
Stupanj 0 (BMI ≤ 25,0)	753	753	61,27	61,27
Stupanj 1 (25,0<BMI ≤ 30,0)	407	1160	33,12	94,39
Stupanj 2 (30,0<BMI ≤ 40,0)	69	1229	5,61	100,00
Stupanj 3 (40,0<BMI)	0	1229	0,00	100,00
Ukupno	1229		100,00	

Tako sređeni podaci mogu se prikazati i grafički, kao stupčasti (histogram) i linijski dijagram, ili na neki drugi način, pregledno po razredima ili pak kumulativno. Crtež olakšava uvid u razdiobu a može poslužiti i za orijentaciono očitavanje pojedinih svojstava distribucije.

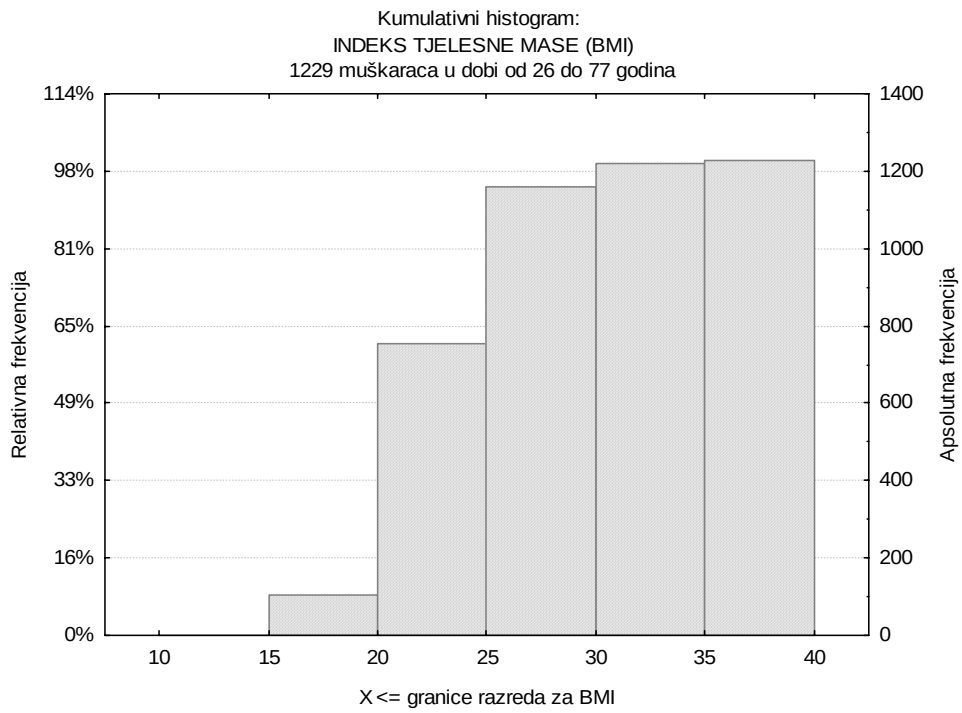
Slika 2.
Histogram indeksa tjelesne mase izmjerenog kod 1229 odraslih muškaraca u dobi od 26 do 77 godina.



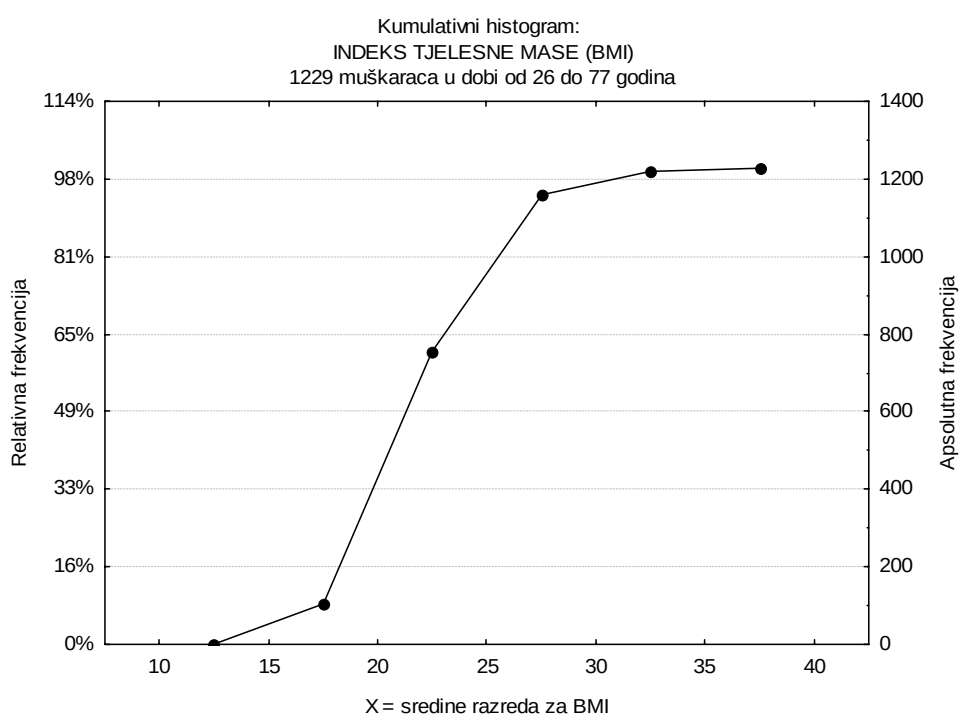
Slika 3.
Poligon frekvencija indeksa tjelesne mase izmjerenog kod 1229 odraslih muškaraca u dobi od 26 do 77 godina.



Slika 4.
Kumulativni histogram indeksa tjelesne mase izmjerenog kod 1229 odraslih muškaraca u dobi od 26 do 77 godina.



Slika 5.
Kumulativni poligon frekvencija indeksa tjelesne mase izmjerenog kod 1229 odraslih muškaraca u dobi od 26 do 77 godina.



Praktičnost tablice frekvencija je očita. Međutim, takvim se pristupom sređivanju podataka gube pojedinačne vrijednosti podataka a prepoznaje samo njihova pripadnost pojedinom razredu odnosno klasi. Želimo li sačuvati uvid u pojedinačne vrijednosti a ipak steci sliku o njihovoj raspodjeli, srediti ćemo podatke u obliku *stablo i list*. Stablo pritom čine prve znamenke izmjerene vrijednosti koje su iste nekoj klasi podataka a listove čine znamenke koje određuju vrijednost pojedinačnog podatka.

Tablica 3.
Stablo i list prikaz indeksa tjelesne mase izmjerenog kod 1229 odraslih muškaraca u dobi od 26 do 77 godina. (Programska podrška: Statistica 7.1)

Stem and Leaf Plot: INDEKS TJELESNE MASE (BMI)	one leaf = 4 cases Include condition: SPOL = muškarci	
stem°leaf (leaf unit=1,0; e.g., 6°5 = 6,5)	Class n	Percentiles
12° · · · ·	0	
13° · · · ·	0	
14° · · · ·	0	
15° · · · ·	3	
16° · · · ·	2	
17° 9 · · · ·	7	
18° 0235689 · · · ·	21	
19° 245668899· · · ·	33	
20° 00112333445566677788888999 · ·	95	
21° 001112233344455566677788899999 ·	116	
22° 00122233334445556677777888999 9 ·	109	25%
23° 000112223333444455555566666777888899999·	150	
24° 0011122222333344445555566777778888899999·	156	median
25° 000011112333344555666777888899· ·	117	
26° 0001111222333444555666777788899 ·	120	75%
27° 00112223334444556678899 · ·	90	
28° 0011122334455667899· · ·	74	
29° 0012345567899 · · ·	51	
30° 01235789 · · · ·	28	
31° 011456 · · · ·	23	
32° 02578 · · · ·	17	
33° 1 · · · ·	5	
34° · · · ·	4	
35° · · · ·	0	
36° 8 · · · ·	6	
37° 9 · · · ·	2	
38° · · · ·	0	
min = 15,5 max = 37,9 Total N:	1229	

Članovima dviju familija u kojima postoji nasljedno oboljenje deficit G6PD (bolest kod koje je čovjek uvjetno zdrav sve dok ne konzumira neke tvari kao što su mahunarka bob, kininski preparati, babiturati i drugo, kada nastupa hemoliza) izmjerena je enzimatska aktivnost G6PD (enzim glukoza-6-fosfat-dehidrogenaza). Vrijednosti su prikazane po spolu:

Muškarci:

190	187	2	173	148	160	195	9	168	6
4	163	158	171	170	9	206	14	8	165
197	230	149	149	168	175	163	169	162	176
163	153	168	19	15	197	173	187		

Žene:

115	187	170	91	124	193	120	122	184	139
190	98	162	125	161	175	179	159	184	182
158	153	192	115	163	77				

Muškarci:

Sredit ćemo podatke tradicionalno tablicom frekvencija. Pritom ćemo jednom uzeti razredni interval od 80 a drugi put od 20 internacionalnih jedinica enzimatske aktivnosti.

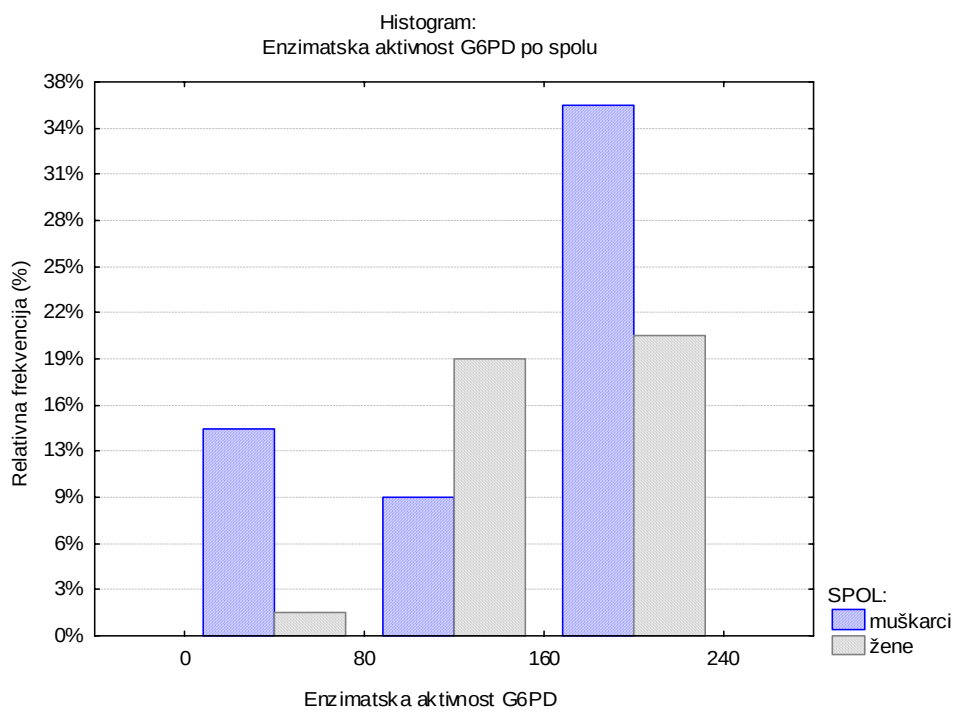
Tablica 4.
Tablica frekvencija enzimatske aktivnosti G6PD izmjerene kod 64 ispitnika (38 muškaraca i 26 žena) s razrednim intervalima od 80 internacionalnih jedinica

Enzimatska aktivnost G6PD (i.u.)	SVI		MUŠKARCI		ŽENE	
	frekvencija	relativna frekvencija (%)	frekvencija	relativna frekvencija (%)	frekvencija	relativna frekvencija (%)
0 ≤ 80	10	15,6	9	23,7	1	3,8
80 ≤ 160	17	26,6	5	13,2	12	46,2
160 ≤ 240	37	57,8	24	63,2	13	50,0
Ukupno	64	100,0	38	100,0	26	100.0

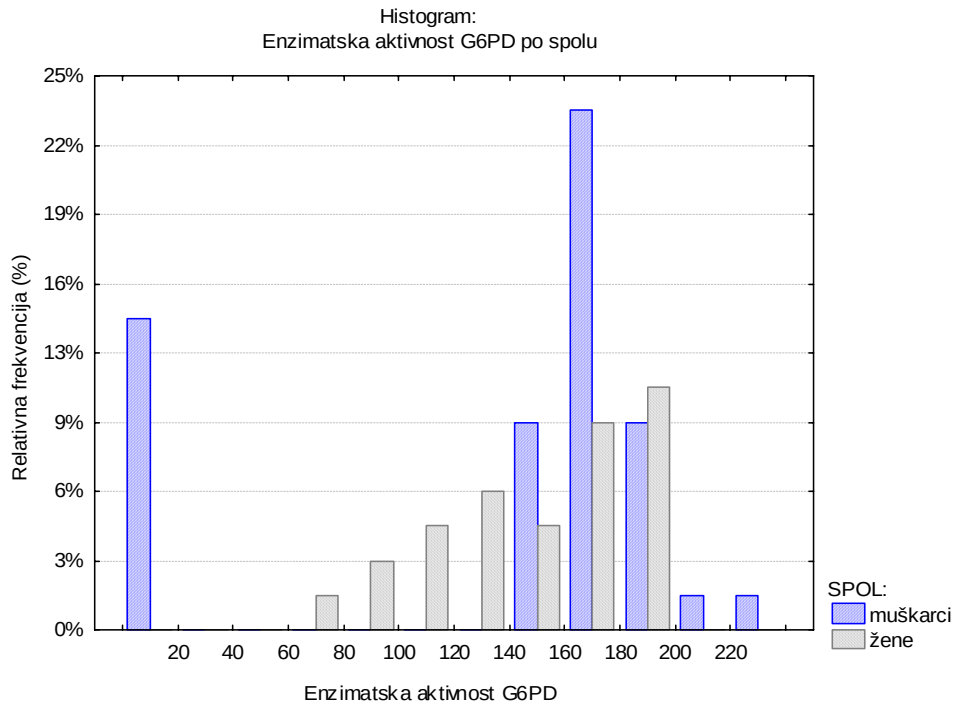
Tablica 5.
Tablica frekvencija enzimatske aktivnosti G6PD izmjerene kod 64 ispitnika (38 muškaraca i 26 žena) s razrednim intervalima od 20 internacionalnih jedinica

Enzimatska aktivnost G6PD (i.u.)	SVI		MUŠKARCI		ŽENE	
	frekvencija	relativna frekvencija (%)	frekvencija	relativna frekvencija (%)	frekvencija	relativna frekvencija (%)
0 ≤ 20	9	14,1	9	23,7	0	0,0
20 ≤ 40	0	0,0	0	0,0	0	0,0
40 ≤ 60	0	0,0	0	0,0	0	0,0
60 ≤ 80	1	1,6	0	0,0	1	3,8
80 ≤ 100	2	3,1	0	0,0	2	7,7
100 ≤ 120	2	3,1	0	0,0	2	7,7
120 ≤ 140	5	7,8	0	0,0	5	19,2
140 ≤ 160	8	12,5	5	13,2	3	11,5
160 ≤ 180	22	34,4	16	42,1	6	23,1
180 ≤ 200	13	20,3	6	15,8	7	26,9
200 ≤ 220	1	1,6	1	2,6	0	0,0
220 ≤ 240	1	1,6	1	2,6	0	0,0
Ukupno	64	100,0	38	100,0	26	100.0

Slika 6.
Histogram enzimatske aktivnosti G6PD s razrednim intervalima od 80 internacionalnih jedinica prema spolu ispitanika



Slika 7.
Histogram enzimatske aktivnosti G6PD s razrednim intervalima od 20 internacionalnih jedinica prema spolu ispitanika



Uz veliki razredni interval i zanemarujući spol dobili smo jednu relativno simetričnu raspodjelu. Lako bismo upali u zamku da govorimo o jednoj homogenoj skupini s nešto češćim nizom vrijednostima enzimatske aktivnosti. Međutim uz mali razredni interval, još uvijek zanemarujući spol, skupina se jasno razdvaja na dvije: jednu sa vrlo niskim vrijednostima i jednu sa višim vrijednostima, asimetričnu sa pomakom u lijevo. Kada pak sažmemo podatke uzimajući u obzir spol uz veliki razredni interval dobijemo dvije slične distribucije s tim da je distribucija za žene pomaknuta u lijevo. Uz mali razredni interval dobijemo posve različite distribucije za muškarce i žene. Distribucija za žene kontinuirana je sa vrijednostima od 77 naviše, sa pomakom u lijevo. Distribucija za muškarce ima dva vrha s tim da od vrijednosti 19 do 148 nema nijednog podatka. Ovdje se dakle ne radi o jednoj već o dvije jasno razdvojene distribucije što sugerira da je promatrano svojstvo spolno vezano monogeno determinirano. Muškarci s niskom enzimatskom aktivnošću su osobe sa genskim defektom. Kako se ovdje radi o selekcioniranom familijama za očekivati je da se među ženama nađe znatan broj onih koje su prenosioci patološkog svojstva. Pažljivom inspekcijom distribucije za žene uočava se lijeva asimetričnost. Žene sa vrijednostima ispod 160 možemo smatrati prenosiocima iako su sve zajedno klinički uvjetno zdrave. U uzorku nađemo i četiri muškarca s vrijednostima ispod 160 ali znatno iznad 20. Ako je vrijednost 160 arbitrarna granica normale onda ovih četiri muškarca imaju predstavljati laboratorijsku grešku mjerenja.

Sredimo li podatke u maniri stablo i list dvojbe o prirodi pojave koju promatramo znatno su manje.

Tehnicki termin za ovakve distribucije nastale sređivanjem prikupljenih podataka je EMPIRIJSKA DISTRIBUCIJA.

Razdiobe mogu biti simetrične, asimetrične bilo prema višim bilo prema nižim vrijednostima obilježja, sa jednim ili vise vrhova. Primjer simetrične razdiobe: antropometrijski podaci (visina, tjelesna težina, slučajne greške mjerenja). Lijevo asimetrične, s većinom podataka s višim vrijednostima (kvocijent inteligencije, hemoglobin u žena). Desno asimetrične, s većinom podataka prema nižim vrijednostima (glukoza u krvi natašte a pogotovo nakon opterećenja glukozom, kolesterol u serumu, krvni tlak sve to u populacijama koje žive u obilju). Raspodjele koje imaju vise od jednog vrha kao i asimetrične raspodjele ukazuju na postojanje jednog ili vise snažnih čimbenika koji čine podatke jednoznačno heterogenima. Ako na neko svojstvo utječe vise čimbenika od kojih nijedan nije jak već imaju kumulativni učinak u oba smjera podjednako podaci se raspodjeljuju pravilno, simetrično i zvonoliko. Primjerice visina normalne i zdrave djece u zdravoj okolini određena je poligeno i multifaktorijalno.

SREĐIVANJE KVALITATIVNIH OBILJEŽJA

Sređivanje kvalitativnih podataka jednostavnije je utoliko što su kategorije po kojima podatke prebrojavamo unaprijed određene kao isključivo moguće ili arbitražom. Podaci se razvrstaju u tablice koje mogu biti vrlo jednostavne, s jednim ulazom ili složenije sa dva, tri i vise ulaza.

Takve tablice zovemo *tablicama kontingencije*.

Primjer tablice s jednim ulazom:

Tablica 6.
Udio uzroka smrti u Hrvatskoj 2003. godine – 10 vodećih uzroka smrti (Izvor: Hrvatsko zdravstveno-statistički ljetopis za 2003. godinu)

Uzroci smrti - dijagnoza	Apsolutna frekvencija	Relativna frekvencija
Ishemične bolesti srca	10436	19,85
Cerebrovaskularne bolesti	8360	15,90
Insuficijencija srca	3809	7,24
Zloćudna novotvorina dušnika, dušnica i pluća	2640	5,02
Zlućudne novotvorine debelog crijeva	1622	3,09
Kronične bolesti jetre, fibroza i ciroza	1246	2,37
Bronhitis, emfizem, astma	1171	2,23
Dijabetes melitus	1069	2,03
Ateroskleroza	995	1,89
Zloćudna novotvorina želuca	995	1,89
Komplikacije i nedovoljno definirani opisi srčane bolesti	988	1,88
Ostalo	19244	36,60
Ukupno	52575	100,0

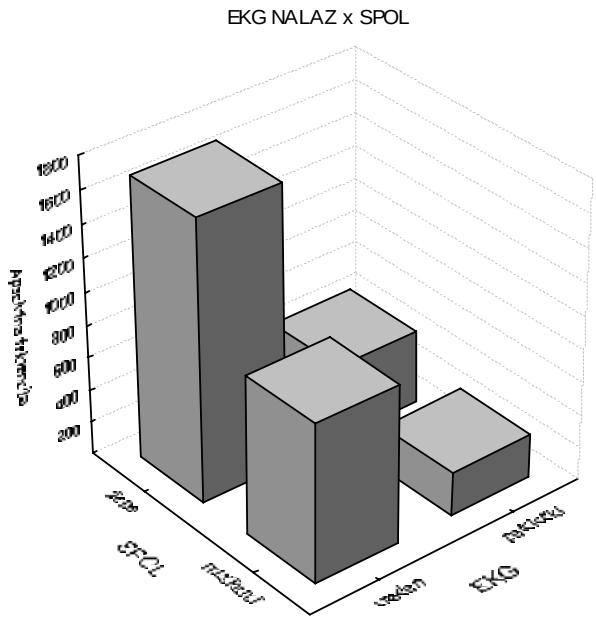
Tablice s više ulaza služe zornom prikazu relacija među odabranim obilježjima ili kao podloga za ispitivanje ima li među tim varijablama povezanosti ili pak razlikuju li se uzorci po prisustvu nekog svojstva.

Tablica 7. Tablica kontingencija nalaza EKG-a prema spolu ispitanika

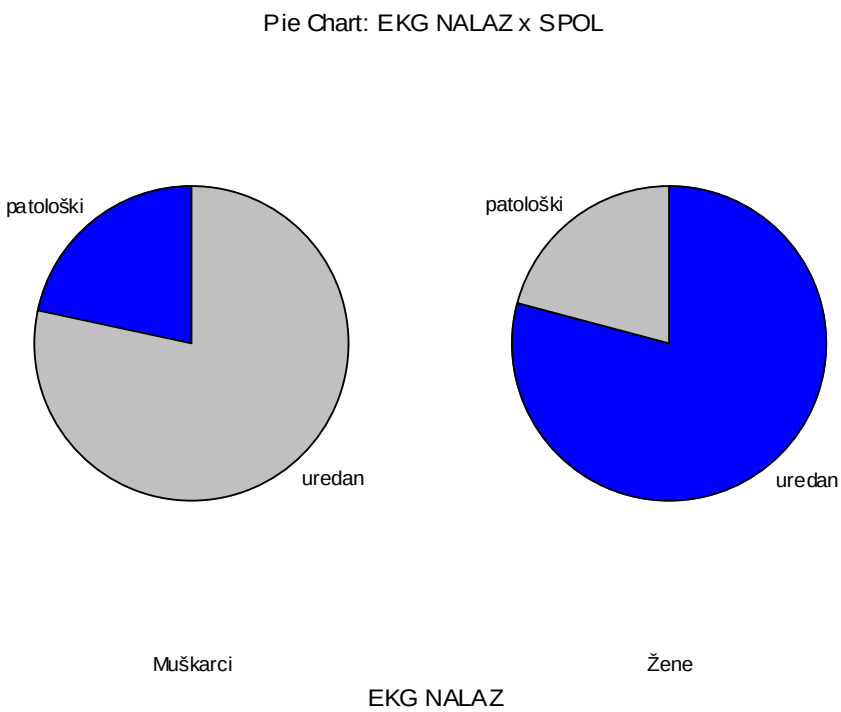
		EKG NALAZ		Ukupno
		uredan	patološki	
SPOL	muškarci	962	267	1229
	žene	1700	451	2151
Ukupno		2662	718	3380

Tablica 7, 2×2 tablica kontingencije, prikazuje učestalost urednih i patoloških nalaza EKG po spolu ispitanika. Može poslužiti ispitivanju pretpostavke o postojanju povezanosti jednog spola s učestalijim patološkim nalazima EKG-a odnosno za ispitivanje valjanosti pretpostavke o postojanju razlika među spolovima s obzirom na pojavu patoloških EKG nalaza.
Isto se može prikazati i grafički – trodimenzionalno (slika 8) ili pomoću „pita“ (engl. pie-charts, slika 9):

Slika 8.
Trodimenzionalni prikaz povezanosti nalaza EKG-a i spola ispitanika



Slika 9. Grafički prikaz („pite“) povezanosti nalaza EKG-a i spola ispitanika

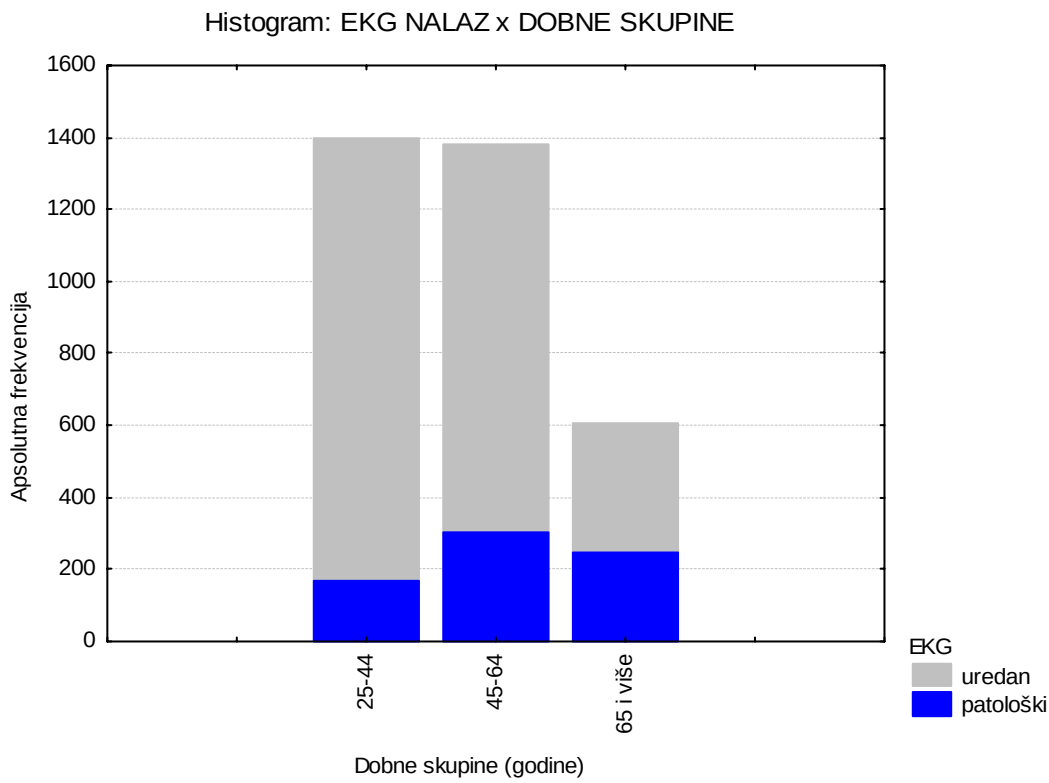


Tablica 8, 3×2 tablica kontingencije, prikazuje povezanost dobi i nalaza EKG. Dob je pritom kategorizirana u 3 dobne skupine. Podloga je ispitivanju povezanosti između ovih dvaju obilježja, tj. imaju li stariji ispitanici češće patološki nalaz EKG-a od mlađih.

Tablica 8. Tablica kontingencija nalaza EKG-a prema dobi ispitanika

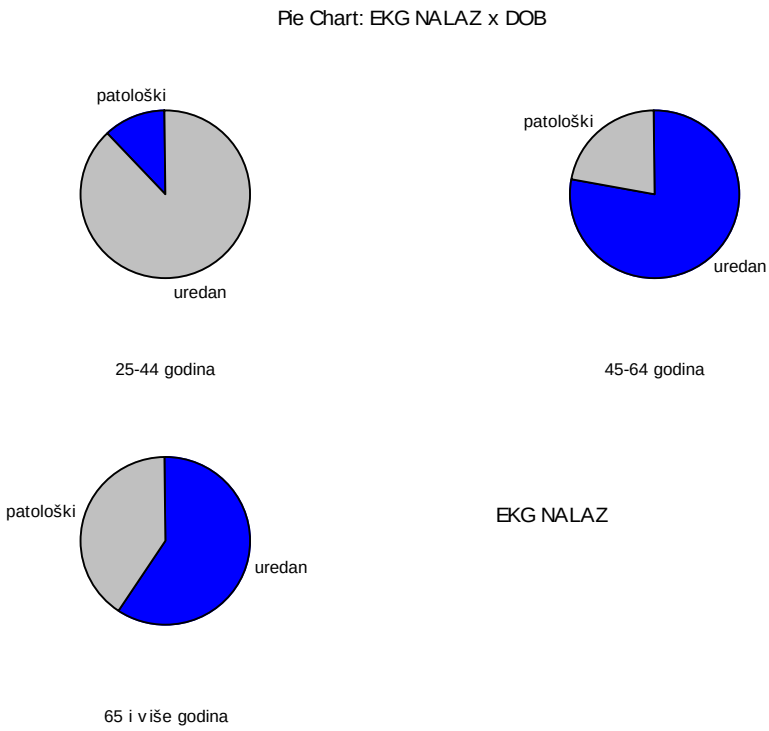
		EKG NALAZ		Ukupno
		uredan	patološki	
DOB	25-44 godine	1228	168	1396
	45-64 godine	1074	305	1379
	65 i više godina	360	245	605
Ukupno		2662	718	3380

Slika 10. Histogram povezanosti EKG nalaza i dobi ispitanika



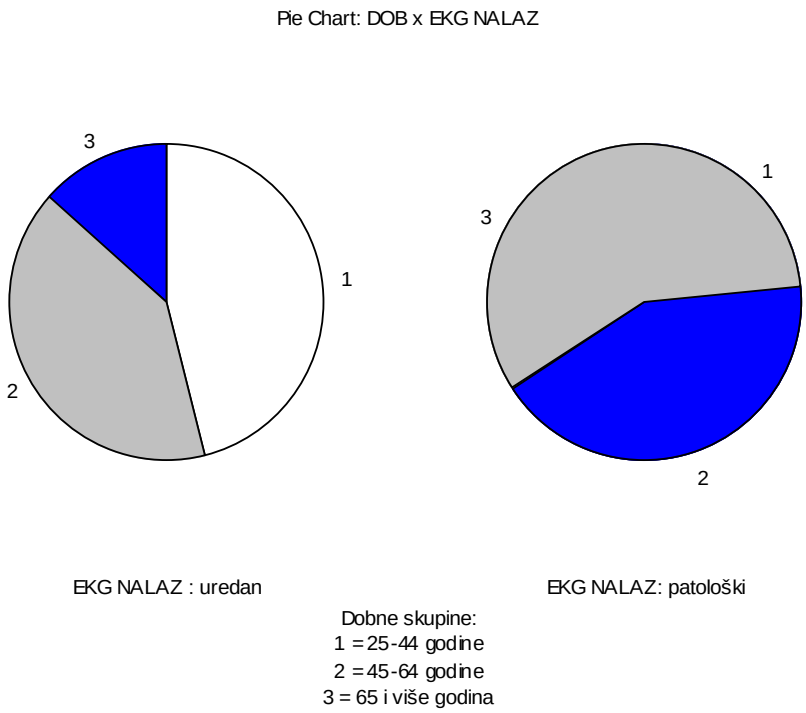
Histogram na slici 10 prikazuje udio različitiha EKG nalaz u svakoj od opisane tri dobne skupine ispitanika. To možemo prikazati i pomoću pita što prikazuje slika 11. Vidimo da udio patoloških nalaza EKG raste u starijim dobnim skupinama.

Slika 11. Grafički prikaz povezanosti EKG nalaza i dobi ispitanika



Grafički prikaz povezanosti dobi i EKG nalaza možemo obrnuti. Na slici 12. vidimo koliki su udjeli različitih dobnih skupina skupinama ispitanika s urednim i patološkim EKG nalazima. Udio najmlađih ispitanika (25-44 godine) veći je u skupini ispitanika s urednim nalazom EKG-a, dok je udio najstarijih bolesnika (65 i više godina) veći u skupini ispitanika čiji je EKG nalaz bio patološki.

Slika 12. Grafički prikaz povezanosti dobi ispitanika i EKG nalaza



Literatura:

1. *Ivanković D, i sur. Osnove statističke analize za medicinare. Zagreb: Medicinski fakultet Sveučilišta u Zagrebu, 1989.*
2. *Petrie A, Sabin C. Medical Statistics at a Glance (2nd Ed). Oxford: Blackwell Science Ltd, 2005.*
3. *Glantz. SA. Primer of Biostatistics (4th Ed). New York: McGraww-Hill: 1997.*
4. *Altman DG. Practical Statistics for Medical Research. London. Chapman & Hall, 1991.*
5. *Bland M. An Introduction to Medical Statistics (3rd Ed). Oxford: Oxford University Press, 2005.*
6. *Armitage P, Berry P. Statistical Methods in Medical Research. Oxford: Blackwell Science Ltd, 1994.*