Univerza v Ljubljani
Fakulteta za farmacijo

Urejanje in prikazovanje podatkov

Izbrana poglavja iz biomedicinske informatike 2011/2012, 2.letnik LBM2

Asist. dr. Igor Locatelli, *mag. farm.*

Ljubljana, 14. 10. 2011

Statistika

- Statistika je veda, ki kvantitativno proučuje masovne pojave v naravi in družbi ter tako z metodami, ki so njej lastne, odkriva zakonitosti teh pojavov
- Biomedicina:
  - Biostatistika ali biometrika (farmakometrika): proučevanje osnovnih dogajanj in pojavov na področju biomedicine vključujoč farmacijo
  - Zdravstvena statistika: medicinska statistika zdravstvenih služb dopolnjena s podatki demografske in vitalne statistike

Statistika

- Opisna (deskriptivna) statistika:
  - Zbiranje, urejanje in prikazovanje podatkov

- Sklepna (inferenčna, analitična) statistika:
  - Sklepanje na populacijo iz podatkov, dobljenih na majhnih skupinah (vzorcev)

Osnovni pojmi - populacija

- Populacija = statistična množica
  Skupek statističnih enot, ki ustrezajo temeljnim opredeljujočim pogojem:
  - vsebinski,
  - krajevni,
  - časovni.

Študenti 1. letnika, Fakulteta za farmacijo, Ljubljana, leto 2011/12

- Vrste populacij:
  - stvarne ali realne (opredeljene z vsemi tremi pogoji)
  - umišljene ali hipotetične (predvsem v biomedicini):
    - niso časovno, krajevno omejene, velikost populacije ni znana (vzorec)
    - populacija bolnikov s sladkorno boleznijo,
    - populacija belih laboratorijskih miši.

Osnovni pojmi - vzorec

- Del populacije, ki je izbran za proučitev določenih značilnosti populacije.
  - reprezentativnost (dobro predstavlja populacijo) :
  - naključnost: statistične enote imajo enako možnost, da so izbrane.
  - velikost vzorca: majhni (n<30), veliki vzorci
- Numerične opisne mere, izračunane za populacijo, imenujemo parametre populacije (μ), iste mere, izračunane za vzorec ($\bar{x}$), pa statistike.

Naključni izbor in verjetnostno vzorčenje

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	10480	15011	01536	02011	81647	91646	69179	14194
2	22368	46573	25595	85393	30995	89198	27982	53402
3	24130	48360	22527	97265	76393	64809	15179	24830
4	42167	93093	06243	61680	07856	16376	39440	53537
5	37570	39975	81837	16656	06121	91782	60468	81305
6	77921	06907	11008	42751	27756	53498	18602	70659
7	99562	72905	56420	69994	98872	31016	71194	18738
8	96301	91977	05463	07972	18876	20922	94595	56869
9	89579	14342	63661	10281	74553	18103	57740	84378
10	85475	36857	53342	53988	53060	59533	38867	62300
11	28918	69578	88231	33276	70997	79936	56865	05859
12	63553	40961	48235	03427	49626	69445	18663	72695
13	09429	93969	52636	92737	88974	33488	36320	17617
14	10365	61129	87529	85689	48237	52267	67689	93394
15	07119	97336	71048	08178	77233	13916	47564	81056
16	51085	12765	51821	51259	77452	16308	60756	92144
17	02368	21382	52404	60268	89368	19885	55322	44819
18	01011	54092	33362	94904	31273	04146	18594	29852
19	52162	53916	46369	58586	23216	14513	83149	98736
20	07056	97628	33787	09998	42698	06691	76988	13602

Generiranje naključnih števil z računalniškimi programi:
npr. MS Excel, funkcija RAND()

Stratificirano vzorčenje:
Naključni izbor enot znotraj vseh posameznih razredov (slojev), na katere je populacija razdeljena:
npr. po spolu, starosti itd.

Osnovni pojmi – statistične spremenljivke

- Statistična enota (npr. posamezen bolnik);
  n = število vseh enot
- Statistične spremenljivke, znaki ali variable
  (npr. starost, spol): proučevane ali neproučevane.
- Vrste statističnih spremenljivk
  - Opisne ali atributivne: spremenljivke, katerih vrednosti opisujemo z besedami
  - Številske ali numerične: spremenljivke katerih vrednosti opisujemo s številkami
- Označevanje: x ali y $\quad x = x_1, x_2, x_3, x_4 \ldots x_n$

Osnovni pojmi – opisne spremenljivke

- Razdelitev glede na število kategorij
  - Dihotomne ali binarne spremenljivke, zajemajo samo dve vrednosti oz. kategoriji;
    npr. spol (M, Ž), preživetje (živ, mrtev).
  - Politomne spremenljivke; imajo več kategorij
- Razdelitev glede na urejenost v zaporedje
  - Nominalne spremenljivke, niso urejene po logičnem zaporedju;
    npr. krvna skupina (A, B, AB, 0).
  - Ordinalne spremenljivke, so urejene v zaporedje;
    npr. stopnja bolečine (brez, blaga, zmerna, huda, zelo huda).

Osnovni pojmi – numerične spremenljivke

- Nezvezne ali diskontinuirane numerične spremenljivke
  - Zajemajo celoštevilčne vrednosti (naravna števila z nič)
  - Pridobimo jih v glavnem s štetjem

  Npr. število porodov v določenem obdobju/regiji,
  število opravljenih izpitov pri posameznem študentu

- Zvezne ali kontinuirane numerične spremenljivke
  - Zajemajo lahko vse štrvilčne vrednosti na določenem intervalu (realna števila)
  - Pridobimo jih v glavnem z merjenjem

  Npr. krvni pritisk, krvna koncentracija LDL, telesna masa, itd

- Kaj pa starost?

Urejanje statističnih podatkov

- Urejanje opisnih spremenljivk
  - Frekvenčna tabela
  - Prikaz s stolpci
- Urejanje numeričnih spremenljivk
  - Frekvenčna tabela
  - Histogram
  - Ranžirna vrsta

Urejanje opisnih spremenljivk

- Združevanje enot v skupine – kategorije:
  - Določitev števila enot v posamezni kategoriji (frekvenca)
- Spremenljivke z maloštevilčnimi vrednostmi:
  enostavna razmejitev v kategorije
  Npr: spol, zakonski stan, krvne skupine (A, B, AB, 0)?
- Spremenljivke z veliko vrednostmi in nejasnimi mejami
  Npr. barva las
- Klasifikacije
  - Mednarodna klasifikacija bolezni (MKB),
  - anatomsko-terapevtska-kemična klasifikacija zdravil (ATC)

Urejanje numeričnih spremenljivk

- Nezvezne numerične spremenljivke
  - Preštejemo frekvence posameznim celoštevilskim vrednostim – frekvenčna tabela
  - Vsak vrednost svoj razred, če je različnih vrednosti malo
- Zvezne numerične spremenljivke
  - Razvrščanje v ranžirno vrsto
  - Frekvenčna porazdelitev zveznih numeričnih spremenljivk, nato izrišemo **histogram**

Ranžirna vrsta in rang

- Ranžirna vrsta: ureditev enot po velikosti vrednosti znaka od najmanjše do največje vrednosti ali obratno
- Vsaki enoti dodamo zaporedno številko (=rang).
- Vsem enotam z isto vrednostjo pripišemo enak rang:
  - Takšen rang izračunamo tako, da seštejemo range, ki naj bi jih enote dobile, in vsoto delimo s številom vseh enot
- Primer: plazemske koncentracije kalcitriola

Primer rangiranja v ranžirno vrsto

Iz populacije bolnikov s kronično okvaro ledvic, ki se zdravijo z dializo smo izbrali velik vzorec in v plazemskih vzorcih določili koncentracijo kalcidiola. Dobili smo naslednje vrednosti (nmol/L):

59.0; 23.0; 142.0; 168.0; 22.0;
64.0; 228.0; 59.0; 32.0; 145.0;
38.0; 64.0; 164.0; 5.0; 147.0;
41.0; 112.0; 21.0; 249.0; 140.0;
133.0; 27.0; 160.0; 64.0; 63.0; 93.0

Rangiranje: primer kalcitriol

Vrednost [nmol/L]	Rang	Vrednost [nmol/L]	Rang
5.0	1	133.0	16
21.0	2	140.0	17
22.0	3	142.0	18
23.0	4	145.0	19
27.0	5	147.0	20
32.0	6	160.0	21
38.0	7	164.0	22
41.0	8	168.0	23
59.0	9.5	228.0	24
59.0	9.5	249.0	25
63.0	14		
64.0	13		
64.0	13		
64.0	13		
112.0	15		

Neparametrični statistični testi

Frekvenčna porazdelitev numeričnih spremenljivk

- Enotna porazdelitev frekvence glede na vrednosti statistične spremenljivke
- Podatke razvrstimo v razrede
- Določiti moramo število razredov (k) in širino razredov (j).

$$(k-1)\cdot j \leq (x_{MAX} - x_{MIN}) \leq k\cdot j \qquad \text{ali} \qquad (k-1)\cdot j \leq (x_{MAX} - 0) \leq k\cdot j$$

- Število razredov je običajno med 10 in 20 (vsaj 5).
- Praviloma imajo razredi enako širino.
- Razredi so lahko na začetku in na koncu lahko tudi odprti (v tem primeru nimajo sredine)

Frekvenčna porazdelitev numeričnih spremenljivk

- Meje razredov. Natančnost zapisa!

  Mejo prvega razreda lahko postavimo nekaj pod najmanjšo vrednostjo, zadnjega razreda pa nekaj nad najvišjo vrednostjo.

- Sredina razreda: cenilka povprečne vrednosti vseh enot v razredu
- Absolutna frekvenca (f): število enot v razredu.
- Relativna frekvenca (f°, f%): strukturni delež posameznega razreda v celotni statistični masi; $f\% = 100 \cdot f/n$
- Kumulativna frekvenca (F): število enot z vrednostjo pod spodnjo mejo ustreznega razreda
- Gostota frekvence (g): je mera za število enot, ki so razporejene na enoto intervala izbranega razreda, $g = f/j$

Primer frekvenčne tabele: Podatki kalcitriol

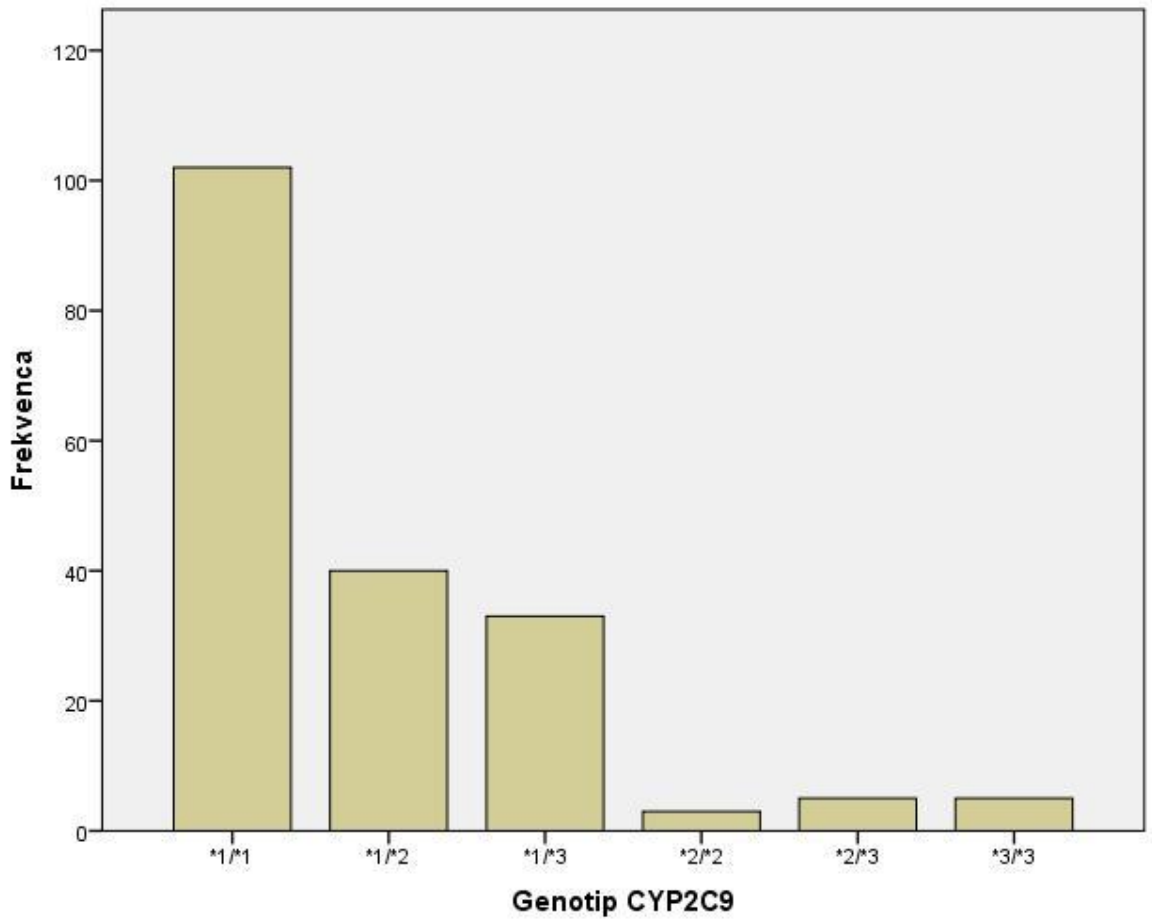
j = 50 nmol/L
k= 5

Št.	Razred	f	f%	F	g
1.	0 - 49.9	8	32%	0	0.16
2.	50.0 - 99.9	6	24%	8	0.12
3.	100.0 - 149.9	6	24%	14	0.12
4.	150.0 - 199.9	3	12%	20	0.06
5.	200.0 - 249.9	2	8%	23	0.04
	skupaj	25	100%	25	

Prikazovanje statističnih podatkov

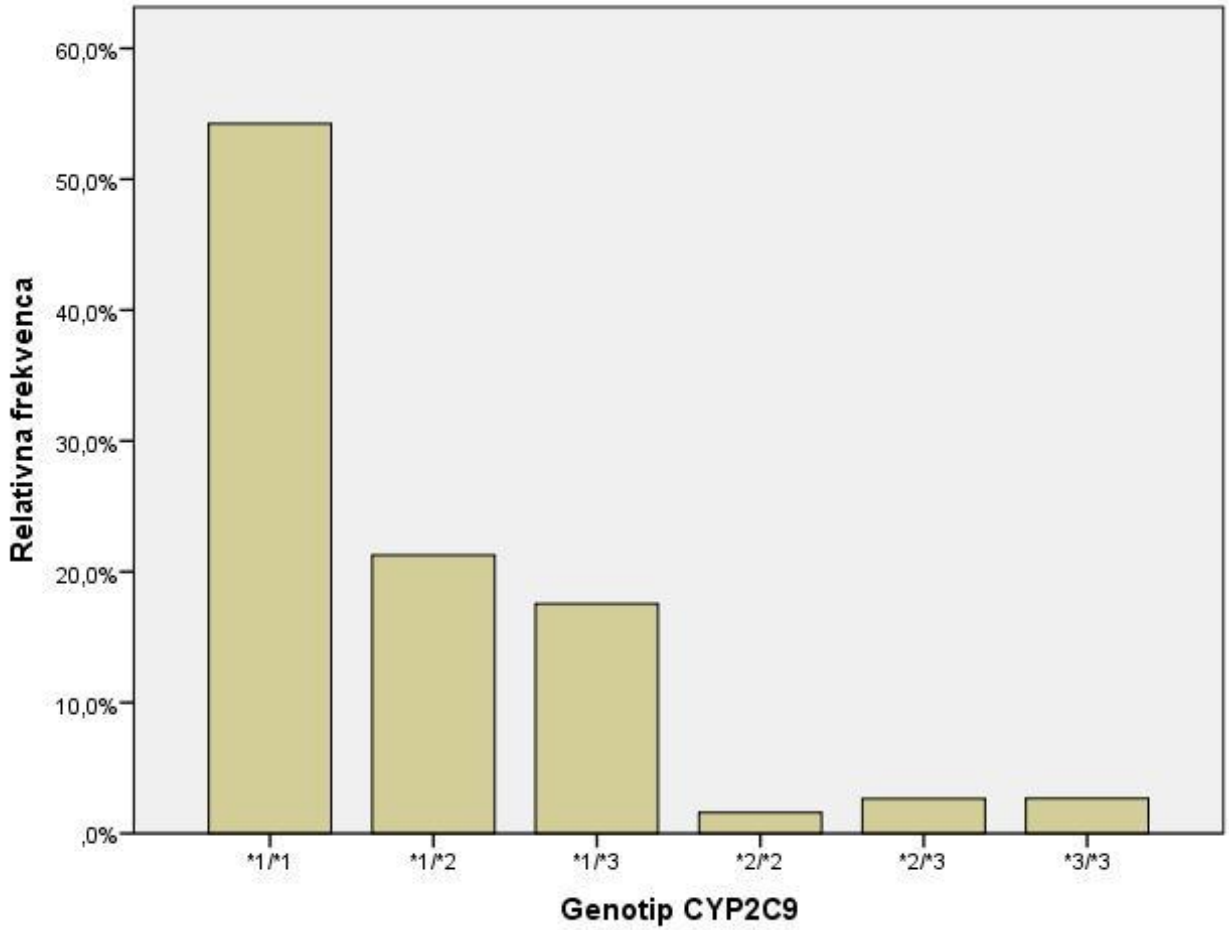
- Tabele oz. preglednice
- Grafikoni:
  - Stolpčni ali stolpičasti diagram (prikaz s stolpci)
  - Linijski ali črtni diagram
  - Prikaz s krogi (krožni izsek)
  - Histogram
  - Frekvenčni poligon
  - Histogram s številkami
  - Razsevni diagram (xy diagram)
  - Kvantilni diagram
- Tortni diagram

Stolpčni diagram
Porazdelitev genotipov *CYP2C9*



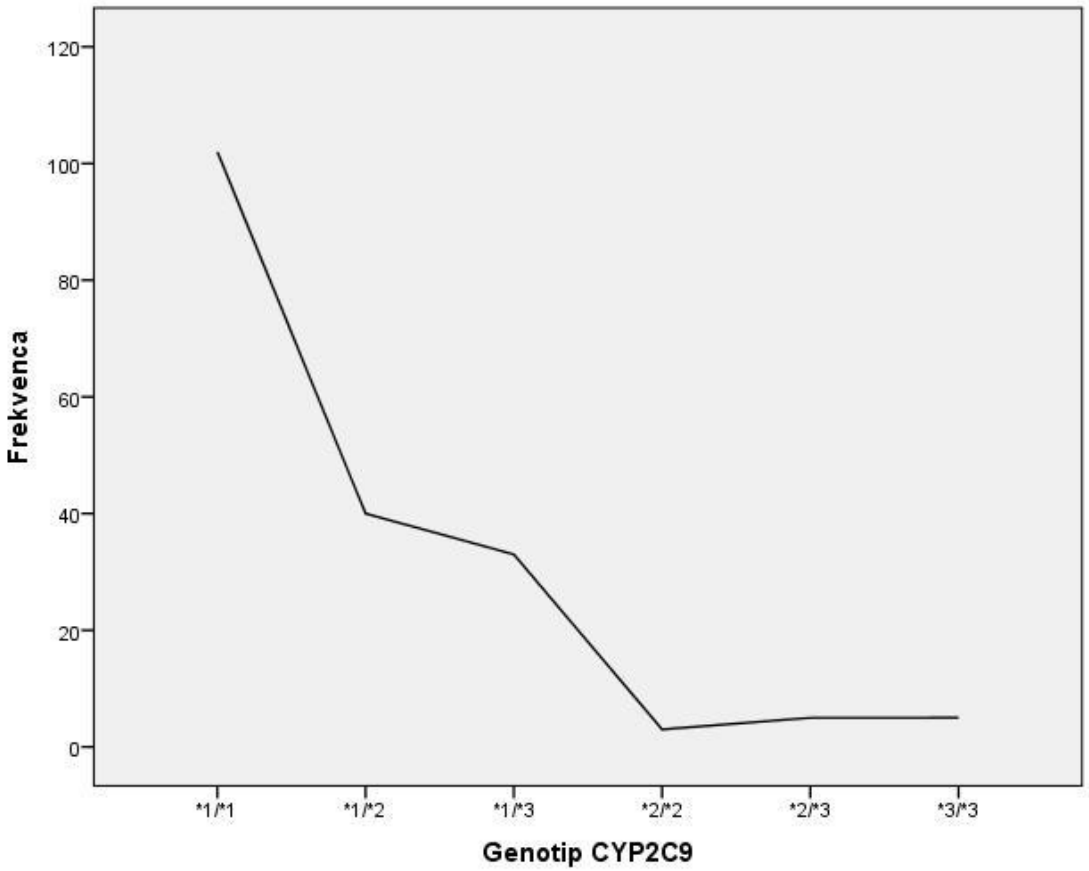
n = 188

Stolpčni diagram
Porazdelitev genotipov *CYP2C9*



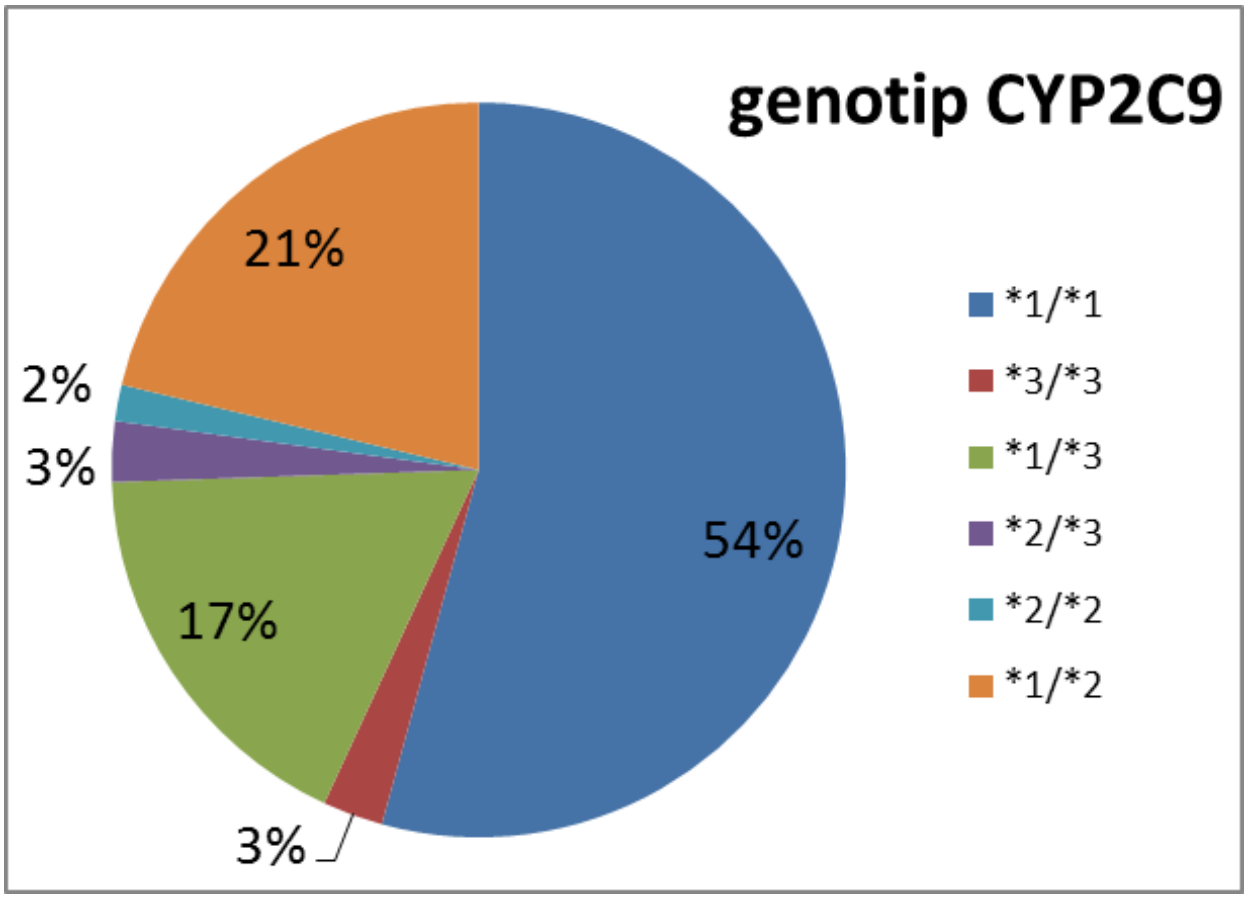
n = 188

Linijski ali črtni diagram
Porazdelitev genotipov *CYP2C9*

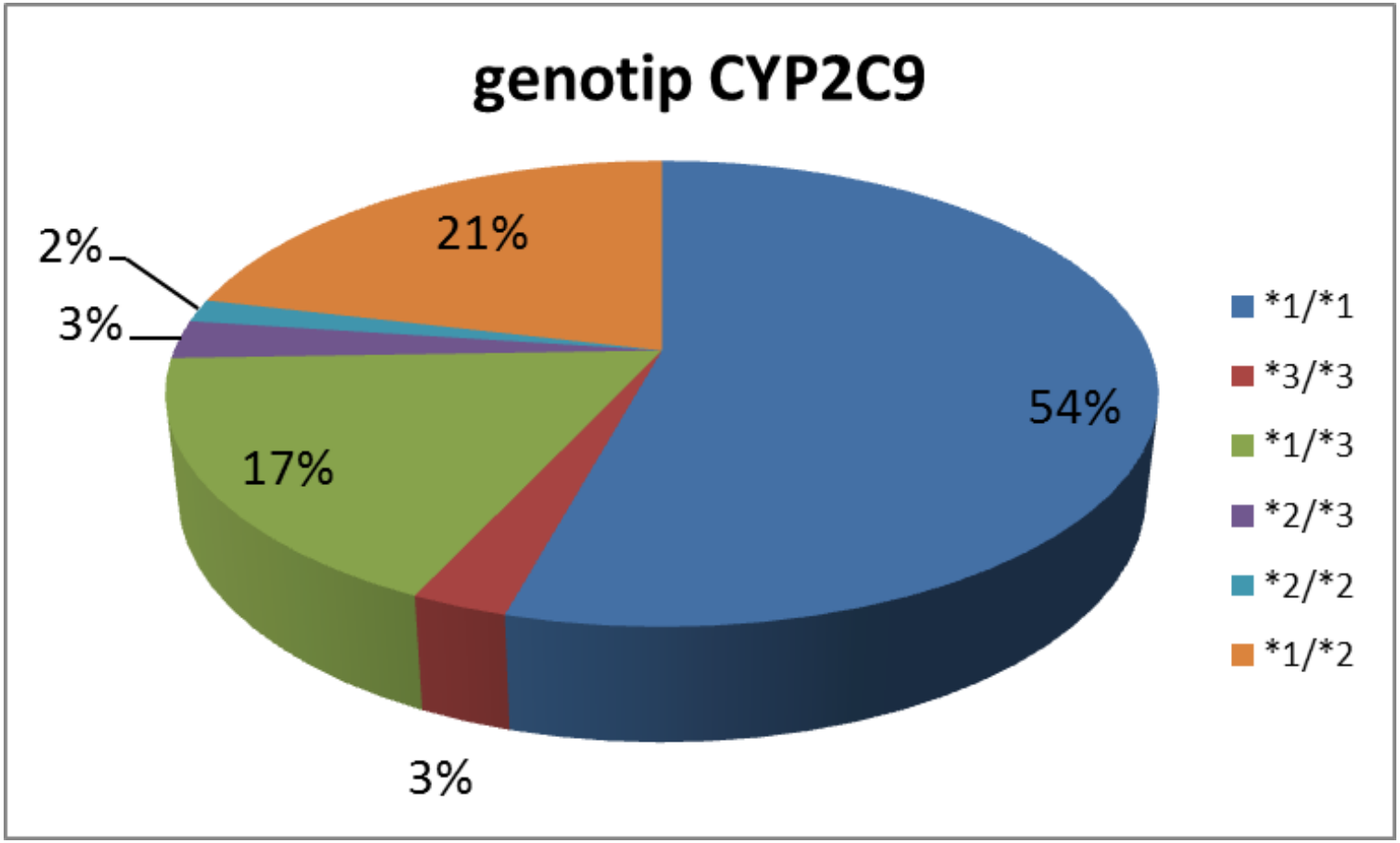


n = 188

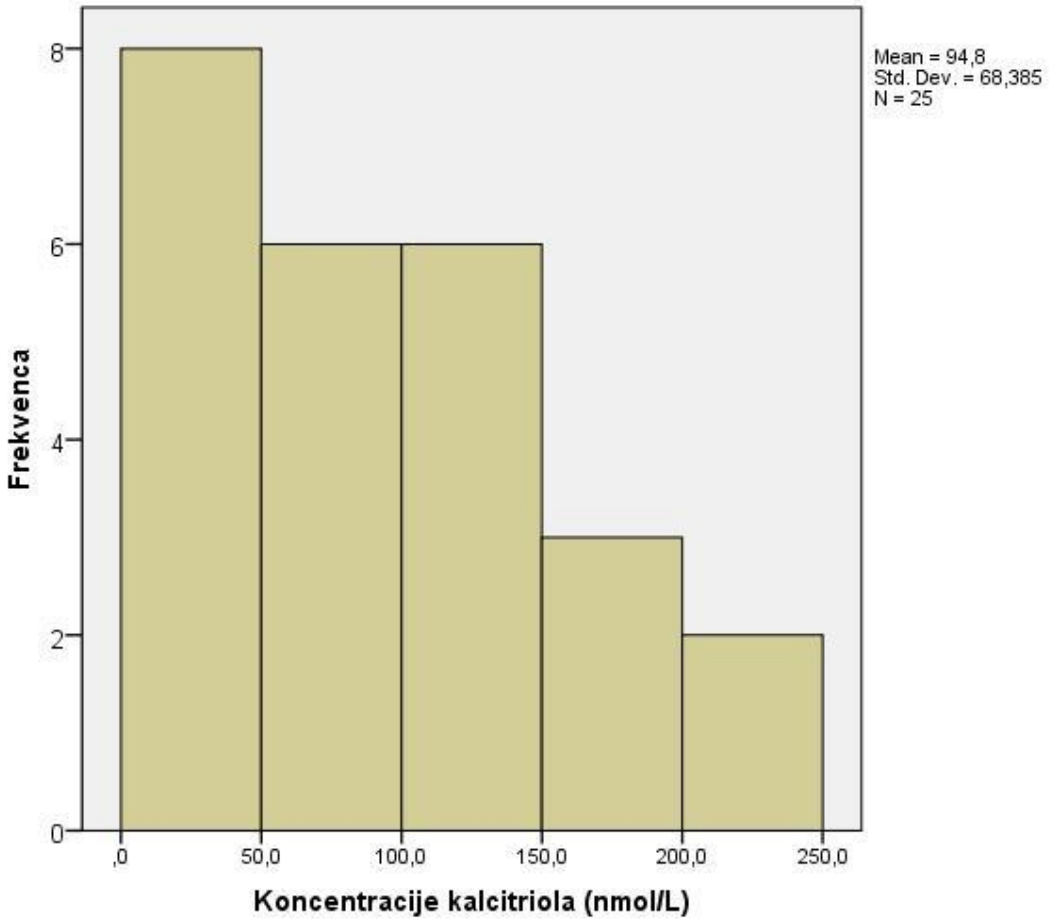
Prikaz s krogi (krožni izsek)
Porazdelitev genotipov *CYP2C9*



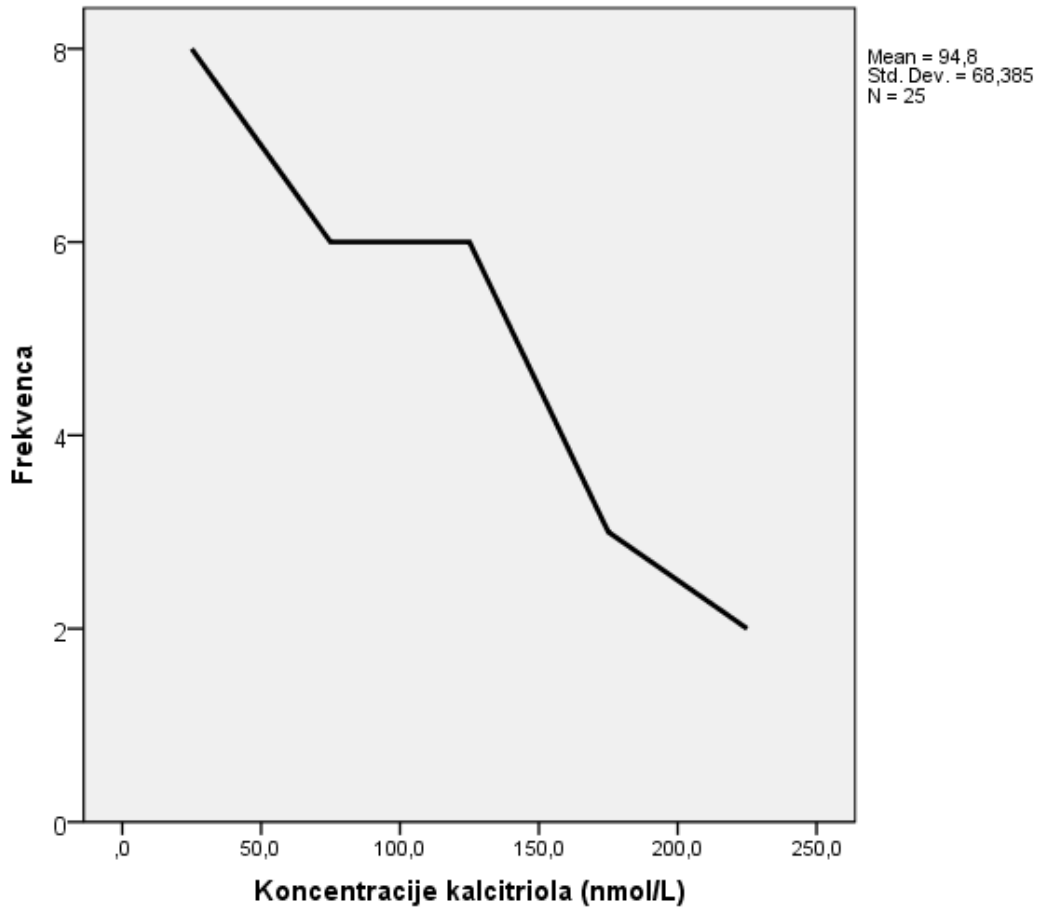
Tortni prikaz



Histogram (kalcitrol)



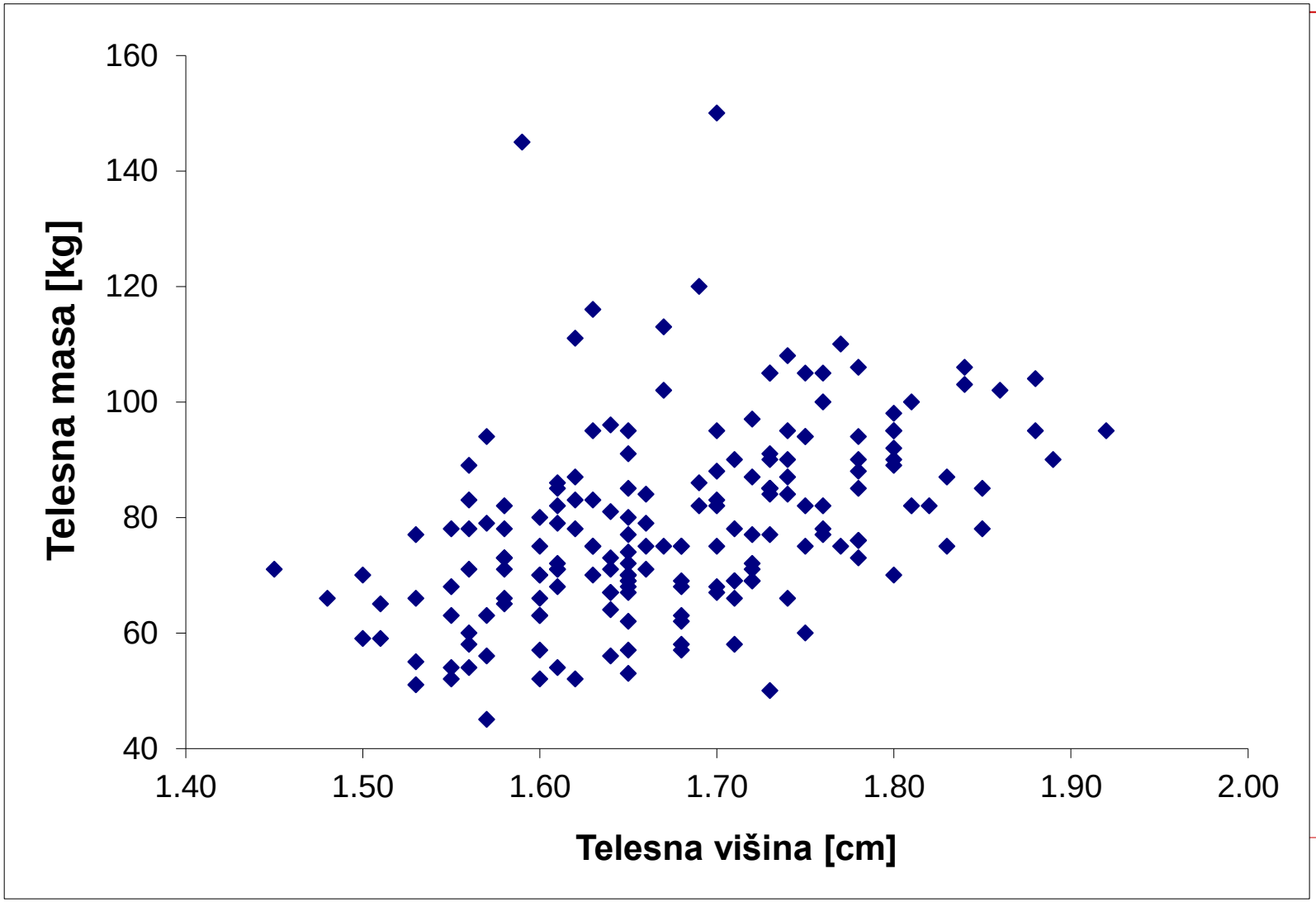
Frekvenčni poligon (kalcitriol)



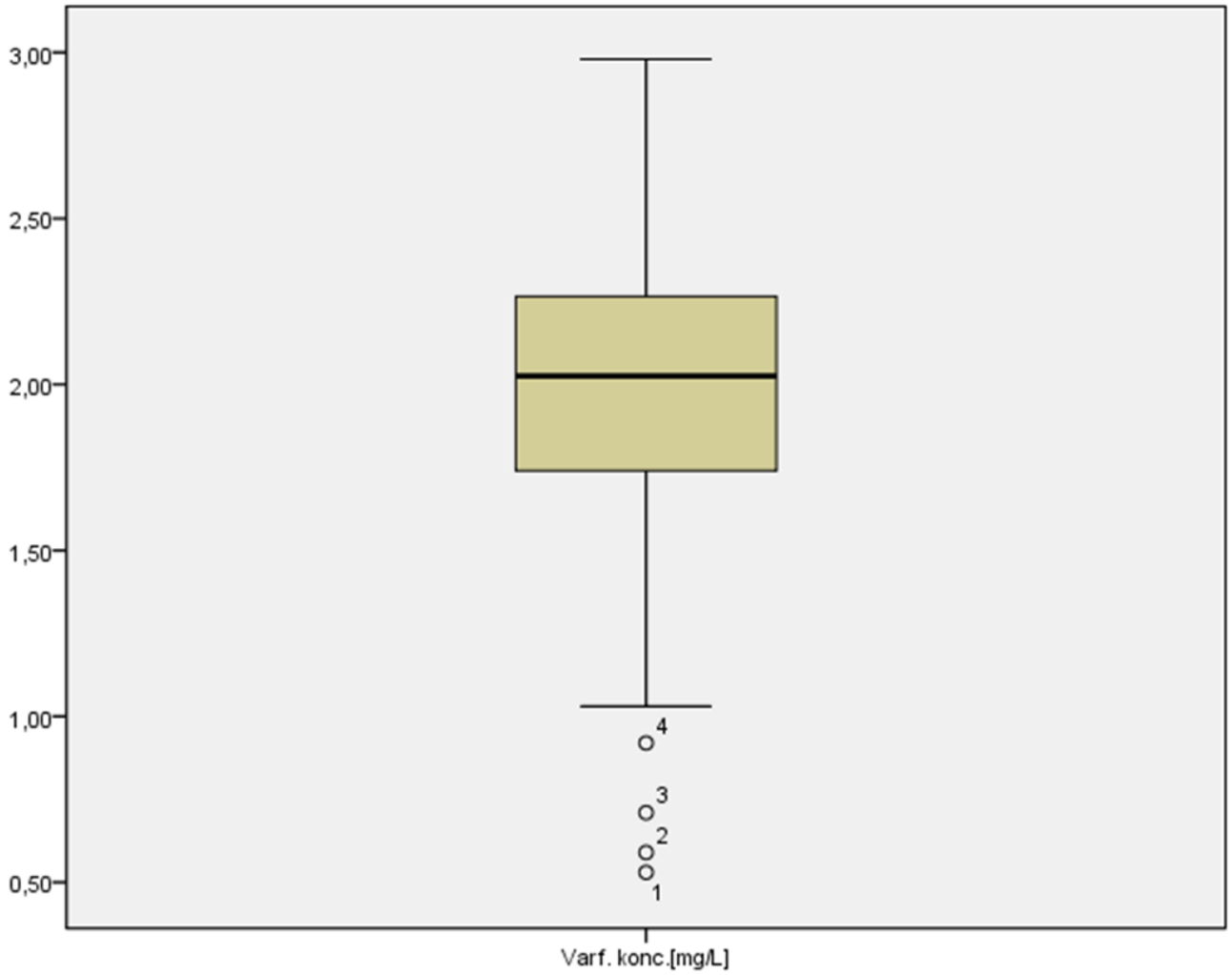
Histogram s številkami
Koncentracije varfarina (mg/L)

```
Frequency    Stem &  Leaf

    4,00 Extremes    (=<,92)
    2,00       10 .  33
    7,00       11 .  0226677
    6,00       12 .  145578
    5,00       13 .  01334
    3,00       14 .  399
    6,00       15 .  115579
    9,00       16 .  133467899
   15,00       17 .  002344555777899
   14,00       18 .  00013344467799
   20,00       19 .  00111222333455566679
   18,00       20 .  002344566777789999
   20,00       21 .  00011113334445678889
   19,00       22 .  0111234555667788999
    9,00       23 .  345666899
    9,00       24 .  001145788
   10,00       25 .  0112334888
    4,00       26 .  0148
    4,00       27 .  0236
    2,00       28 .  19
    2,00       29 .  38

Stem width:       ,10
Each leaf:       1 case(s)
```

Razsevni diagram



Kvantilni diagram



Srednje vrednosti

- Aritmetična sredina ali povprečje
- Modus
- Mediana

Aritmetična sredina

- Najpogosteje uporabljena srednja vrednost.
- Seštejemo vrednosti vseh enot in delimo s številom enot.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Nanjo vplivajo posamezne vrednosti vsake statistične enote.
- Vsota vseh odklonov od aritmetične sredine je enaka nič.
- Povezana z normalno oz. Gaussovo porazdelitvijo.

Mediana

- Tista vrednost spremenljivke, od katere ima polovica enot manjše ali enake, polovica enot pa večje ali enake vrednosti spremenljivke.
- Podatke moramo razvrstiti po velikosti.
- Če je n liho število: mediana enaka vrednosti srednje enote
- Če je n sodo število: mediana je povprečje vrednosti srednjega para podatkov
- Ni povezana z nobeno teoretično porazdelitvijo, pri popolnoma normalni porazdelitve je enaka aritmetični sredini
- Uporabno v primerih, ko je statistična spremenljivka porazdeljena nesimetrično

Kvantili

- Kvartili: podatke razdelimo na četrtine (3 kvartili)
  - 2. Kvartil = Mediana
- Decili: podatke razvrstimo na desetine (9 decilov)
  - 5. decil = Mediana
- Centili (percentili): podatke razvrstimo na stotine
  - 50. centil = Mediana
  - 25. centil = 1. kvartil
  - 75. centil = 3. kvartil

Določanje kvantilov

- Mediana

Liho število enot (n): m-ta največja vrednost: $m = (n+1)/2$

Sodo število enot (n): povprečje med m_1-to in m_2-to vrednostjo
$m_1 = n/2$, $m_2 = n/2 + 1$

- Centili

p = (per)centil (1-100);

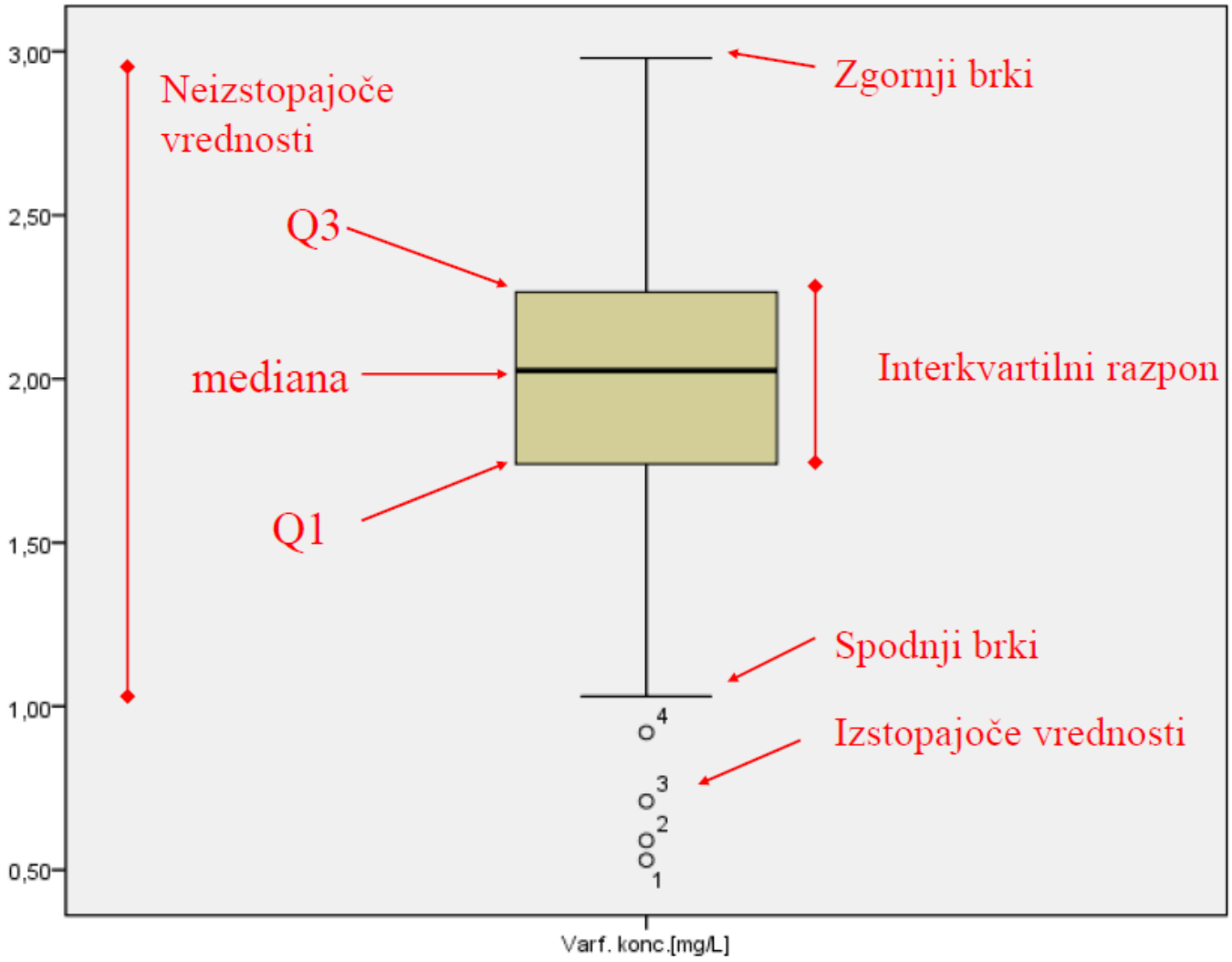
$n \cdot p/100$ ni celo število → (k +1)-ta največja vrednost:
k je navzdol zaokrožen $n \cdot p/100$

$n \cdot p/100$ je celo število → povprečje med m_1-to in m_2-to vrednostjo:
$m_1 = n \cdot p/100$, $m_2 = n \cdot p/100 + 1$

5 različnih načinov računanja centilov (SPSS)

w is the sum of the weights, p is the specified percentile divided by 100, and Xi is the value of the ith case (cases are assumed to be ranked in ascending order).	
HAVERAGE	Weighted average at X(w + 1)p. The percentile value is the weighted average of Xi and Xi + 1, where i is the integer part of (w + 1)p. This is the default if PERCENTILES is specified without a keyword.
WAVERAGE	Weighted average at Xwp. The percentile value is the weighted average of Xi and X(i + 1), where i is the integer portion of wp.
ROUND	Observation closest to wp. The percentile value is Xi or Xi + 1, depending upon whether i or i + 1 is "closer" to wp.
EMPIRICAL	Empirical distribution function. The percentile value is Xi, where i is equal to wp rounded up to the next integer.
AEMPIRICAL	Empirical distribution with averaging. This is equivalent to EMPIRICAL, except when i=wp, in which case the percentile value is the average of Xi and Xi + 1.

Kvantilni diagram (box and whisker plot)



Simetričnost!
Različne vrste.

Izstopajoče vrednosti
Extremne vrednosti

$$X > X_{Q3} + 1.5 \cdot \left(X_{Q3} - X_{Q1}\right)$$

$$X < X_{Q1} - 1.5 \cdot \left(X_{Q3} - X_{Q1}\right)$$

$$X > X_{Q3} + 3 \cdot \left(X_{Q3} - X_{Q1}\right)$$

$$X < X_{Q1} - 3 \cdot \left(X_{Q3} - X_{Q1}\right)$$

Modus

- Modus je najpogostejša vrednost neke spremenljivke
- Ugotavljamo le za razmeroma veliko populacijo, pri manjših pa ga ni mogoče uporabiti
- Opis porazdelitve populacije:
  - Unimodalna
  - Bimodalna
  - Polimodalna

Mere razpršenosti (variabilnosti)

- Koliko posamezni podatki odstopajo od srednje vrednosti
- Dejavniki, ki vplivajo na variabilnost:
  - Napake pri meritvah: npr. zaradi aparature, delovnih razmer, netočnosti metode
  - Intraindividualni razlogi: variabilnost pri osebkih npr. emocionalno stanje, dnevni ritem, menstruacijski ciklus
  - Interindividualni razlogi: variabilnost med osebki npr. Genetski dejavniki, spol, starost, prehrana, zdravstveno stanje

Mere razpršenosti (variabilnosti)

- Standardni odklon ali deviacija
- Varianca
- Koeficient variacije (relativni standardni odklon)
- Variacijski razpon ali razmik
- Kvartilni razpon
- Decilni razpon

Varianca in standardni odklon

- Varianca (σ^2 oz. s^2) je povprečje kvadratov odklonov posameznih vrednosti od aritmetične sredine
  - oznaka σ^2 za populacijski parameter
  - oznaka s^2 za varianco vzorca

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

- Standardni odklon oziroma deviacija
  - Kvadratni koren variance

Koeficient variacije (KV ali RSD)

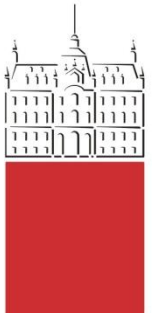
$$KV(RSD) = \frac{\sigma}{\mu}$$

- Mera relativne variabilnosti: standardna deviacija utežena z aritmetično sredino, ponavadi podano kot odstotek (pomnoženo s 100%)
- Ko želimo primerjati variabilnost različnih spremenljivk, ki so med seboj v vsebinski zvezi Npr. višine pri odraslih in otrocih

Variacijski, kvartilni in decilni razpon

- Variacijski razmik: $(x_{max}-x_{min})$
- Decilni razmik: D_9-D_1
- Kvartilni razmik: Q_3-Q_1

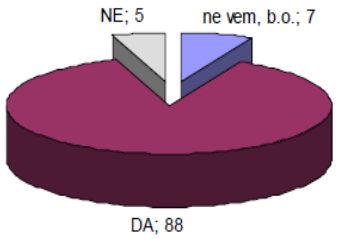
Ko so podatki nesimetrično porazdeljeni, uporabljamo za srednjo vrednost mediano, kvantile pa kot mero razpršenosti; varianca oz. standardni odklon bi bili v tem primeru neustrezna mera razpršenosti



PRIMERI

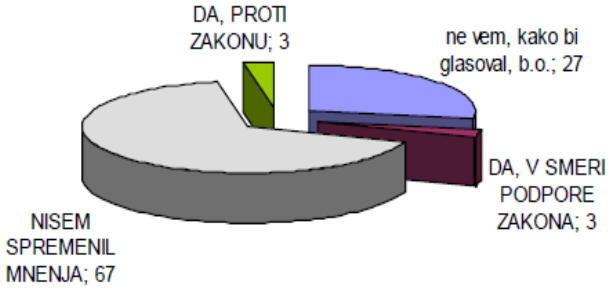
3-D prikaz

ALI MENITE, DA SO V SLOVENIJI POTREBNE REFORME?



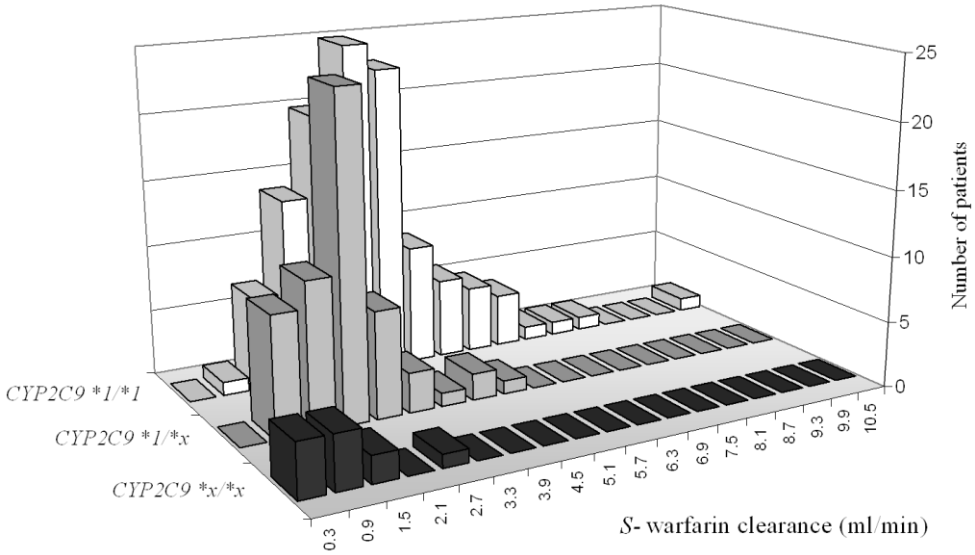
CJMMK, Politbarometer, marec 2011, N=926

ALI STE V ZADNJEM MESECU SPREMENILI SVOJE MNENJE O ZAKONU O MALEM DELU?



CJMMK, Politbarometer, marec 2011, N=926

The distribution of S-warfarin clearance according to *CYP2C9* polymorphism (Locatelli et al.)



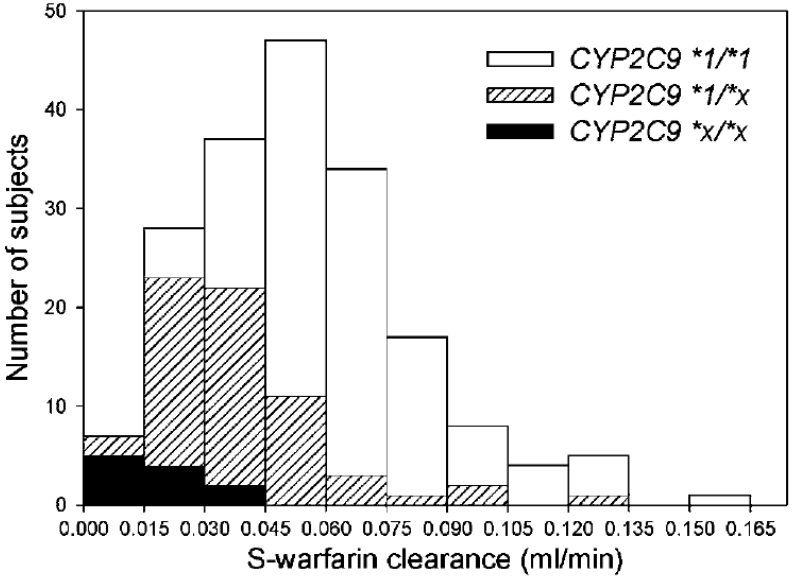
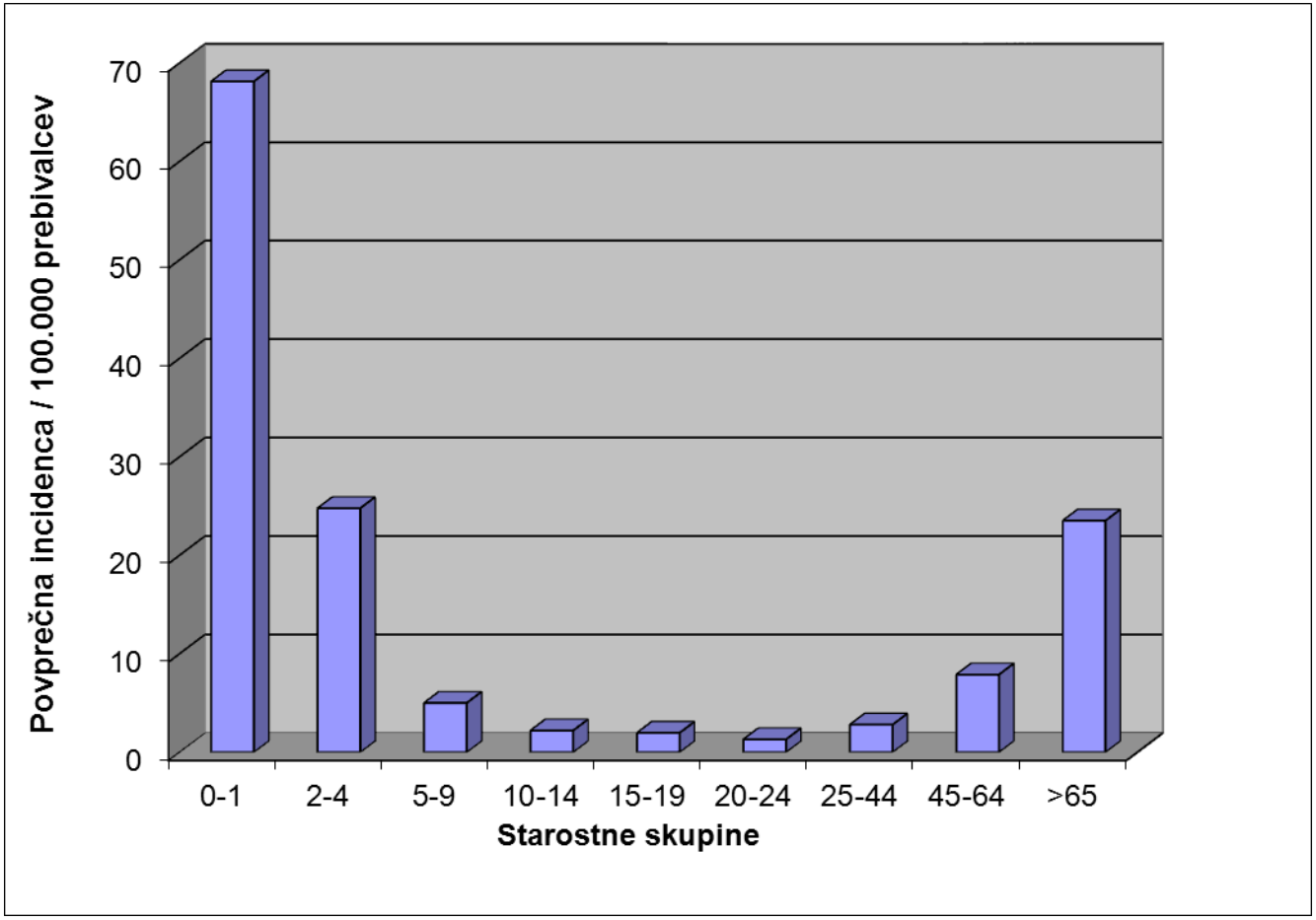
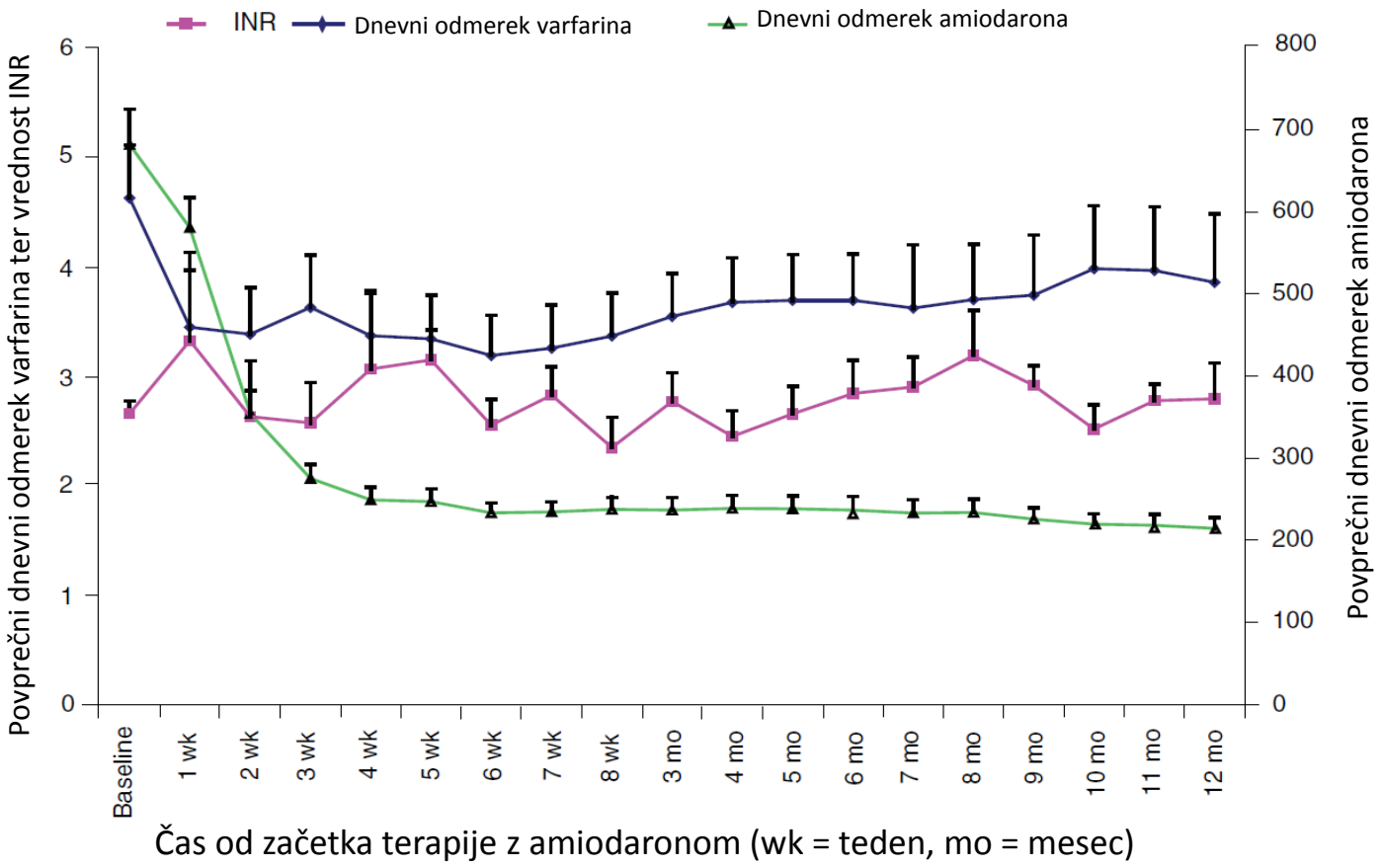


Figure 3 Frequency distribution ($N = 188$) of LBW normalized *S*-warfarin clearance according to *CYP2C9* genotype: two wild-type alleles (**1/*1*), one polymorphic allele (**1/*x*) and two polymorphic alleles (**x/*x*).

Incidenca invazivnih pnevmokoknih bolezni glede na starostne skupine (2004-2010)



Vpliv različnih dejavnikov na učinek antikoagulantnih zdravil (Locatelli I. in Oblak E.)



Slika 1. Vpliv uvajanja terapije z amiodaronom ob že vzpostavljeni terapiji z varfarinom. Prikazane so povprečne vrednosti, odkloni predstavljajo standardno napako (n=70). (Povzeto po 10)

Kakšen graf je prikazan spodaj?