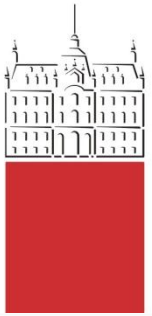
Univerza v Ljubljani
Fakulteta za farmacijo



Urejanje in prikazovanje podatkov I

Farmacevtska informatika
2011/2012, 1. letnik EMŠF

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

Ljubljana, 2. 3. 2012

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 (atributivne) spremenljivke

- Razdelitev glede na število kategorij (skupin)
  - Dihotomne ali binarne spremenljivke, zajemajo samo dve vrednosti oz. kategoriji;
    npr. spol (M ali Ž), preživetje (živ ali mrtev).
  - Politomne spremenljivke; imajo več kategorij;
    npr. genotip *CYP2C9,* barva las, opisna ocena.
- 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 opravljenih izpitov pri posameznem študentu, število porodov v določenem obdobju/regiji, ocena na izpitu (izmerjen)

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

  Npr. krvni pritisk, telesna masa, telesna višina, itd

- Kaj pa starost?

Urejanje statističnih podatkov

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

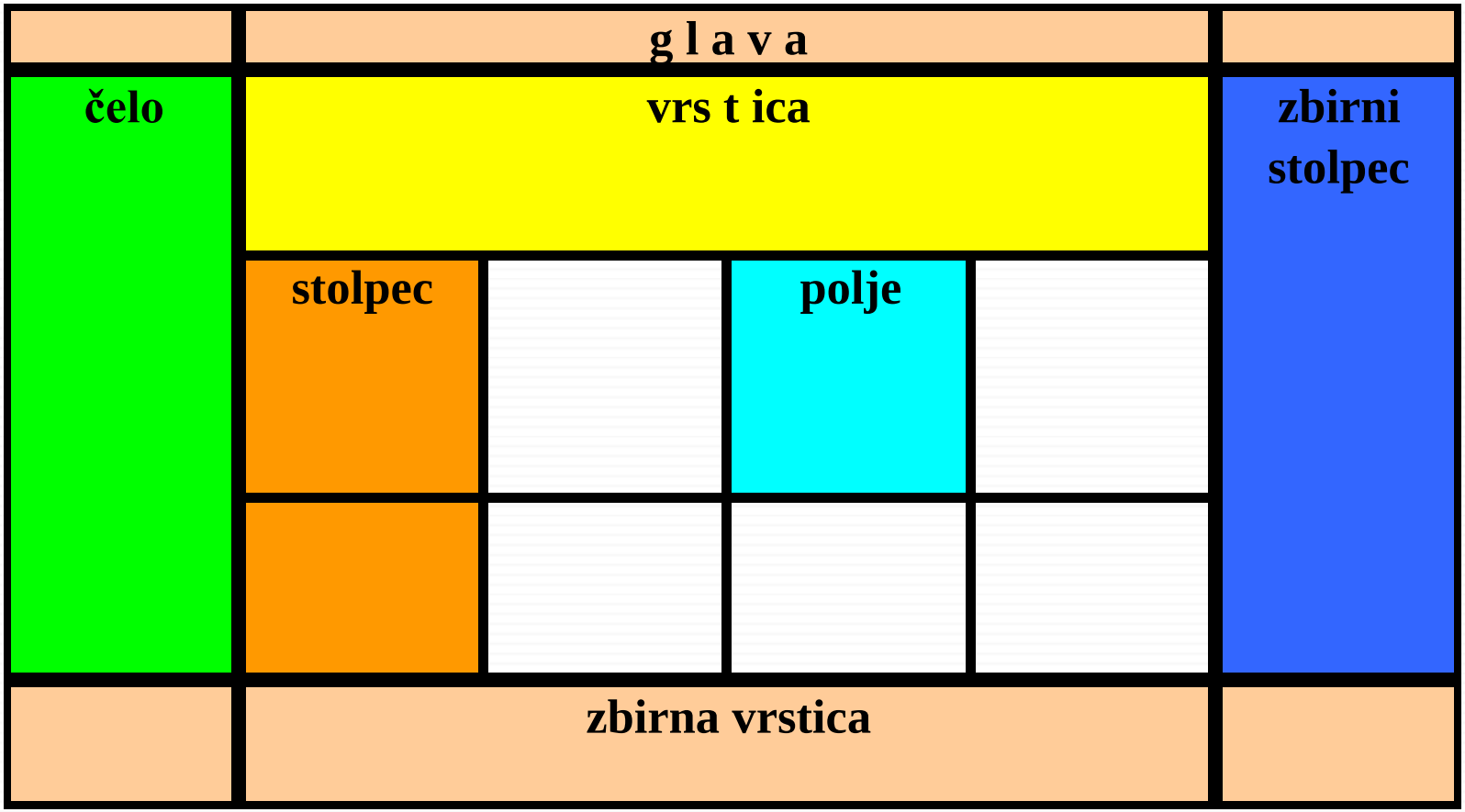
Urejanje opisnih spremenljivk

- Združevanje enot v skupine – kategorije:
  - Določitev števila enot v posamezni kategoriji (frekvenca)
- Spremenljivke z malo številnimi vrednostmi:
  enostavna razmejitev v kategorije
  Npr: spol, zakonski stan, krvne skupine (A, B, AB, 0),
  genotip *CYP2C9*.
- Spremenljivke z veliko vrednostmi in nejasnimi mejami
  Npr. barva las, barva oči
- Klasifikacije
  - Mednarodna klasifikacija bolezni (MKB),
  - anatomsko-terapevtska-kemična klasifikacija zdravil (ATC)

Prikazovanje opisnih spremenljivk

- Tabele oz. preglednice
- Grafikoni:
  - Stolpčni ali stolpičasti diagram (prikaz s stolpci)
  - Prikaz s krogi (krožni izsek)
  - Tortni prikaz (x) problem 3D prikaza

Tabele oz. preglednice



Primer frekvenčne tabele: genotip *CYP2C9*

Št. bolnika	Genotip *CYP2C9*
1	**1/*1*
2	**1/*1*
3	**1/*1*
4	**2/*3*
5	**2/*3*
6	**2/*3*
7	**2/*3*
8	**2/*3*
9	**2/*2*
10	**2/*2*
11	**2/*2*
12	**1/*3*
13	**1/*3*
14	**1/*3*
15	**1/*3*
16	**1/*3*
…	
188	**1/*3*

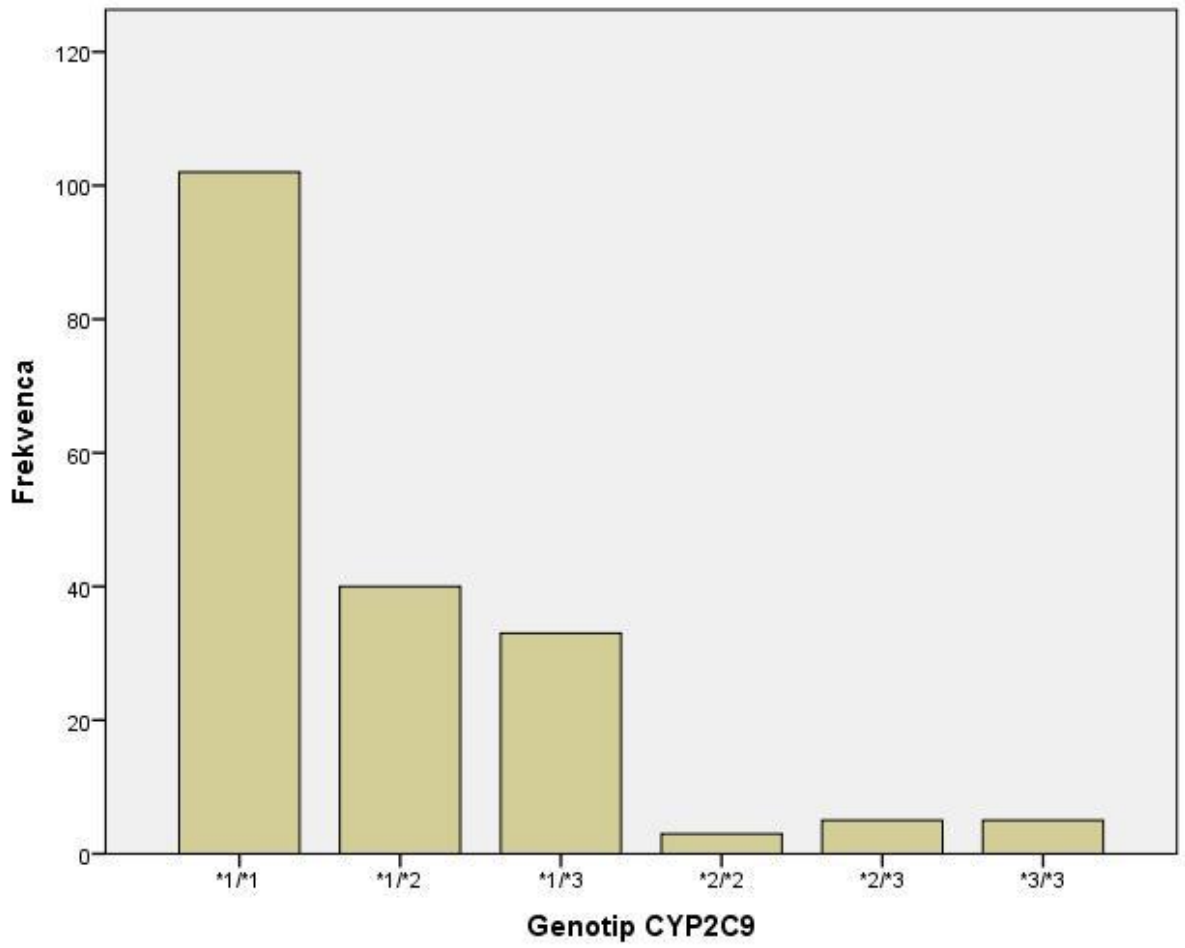
6 različnih kategorij

Genotip *CYP2C9*	Frekvenca
**1/*1*	102 (54%)
**1/*2*	40 (21,2%)
**1/*3*	33 (17,6%)
**2/*2*	3 (1,6%)
**2/*3*	5 (2,7%)
**3/*3*	5 (2,7%)
SKUPAJ	188

Stolpčni diagram (bar graph/chart)

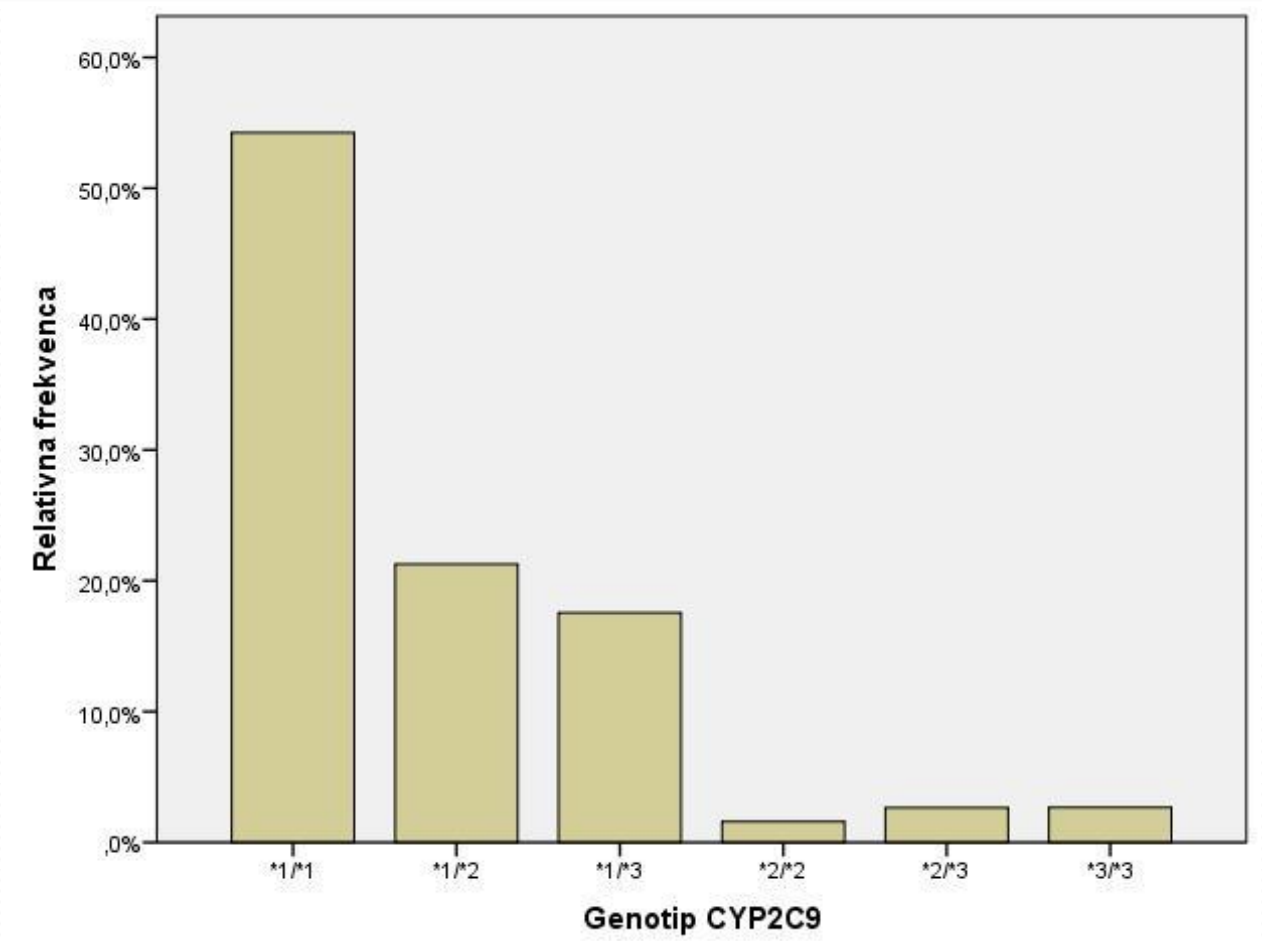
- Grafični prikaz statističnih podatkov s stolpci, katerih višina je sorazmerna s frekvenco ali velikostjo pojava
- Prikaz pojavnosti posameznih kategorij pri opisnih ali nezveznih numeričnih spremenljivkah
- Višina stolpca je lahko izražena kot absolutna (število) ali relativna (procent) frekvenca
- Stolpci običajno niso povezani (so enako oddaljeni drug od drugega)

Stolpčni diagram
Porazdelitev genotipov *CYP2C9*



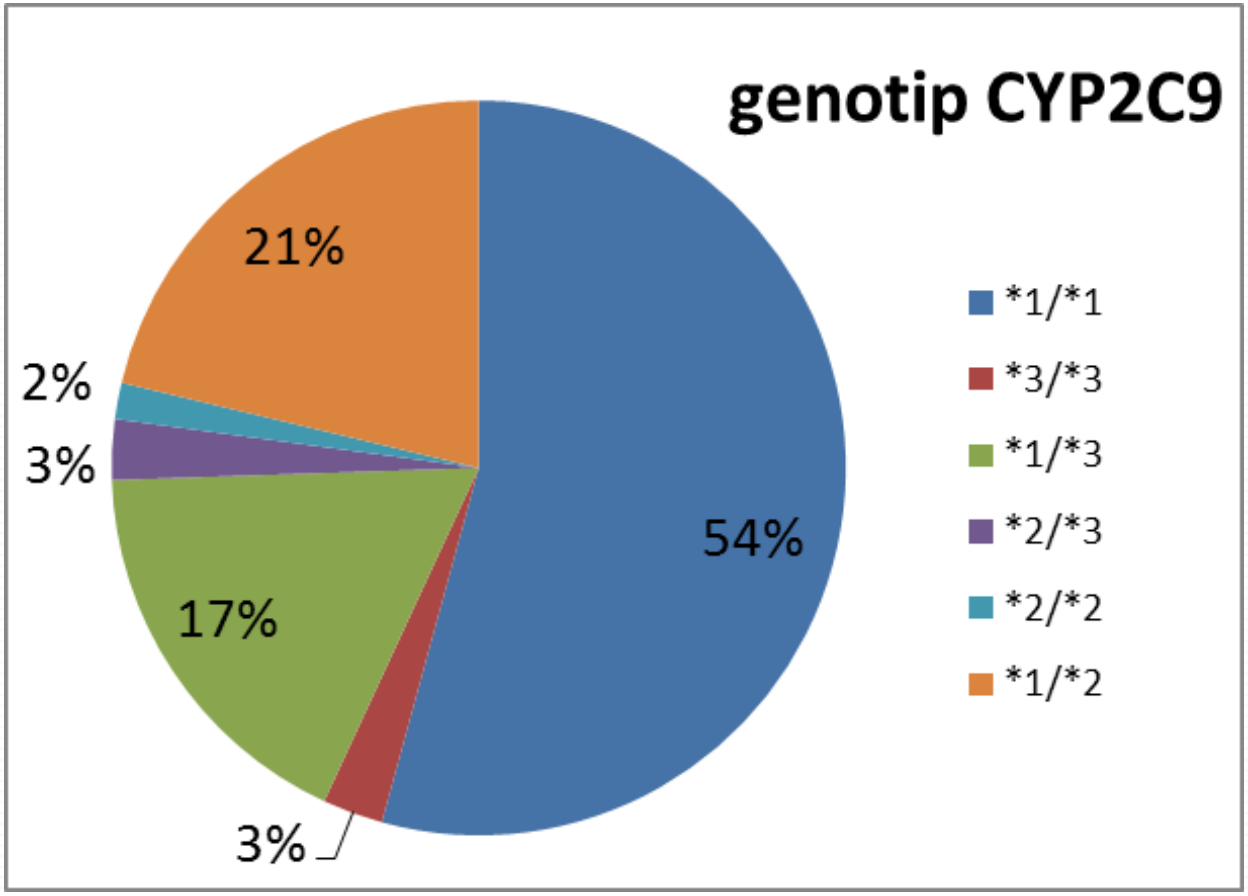
n = 188

Stolpčni diagram
Porazdelitev genotipov *CYP2C9*



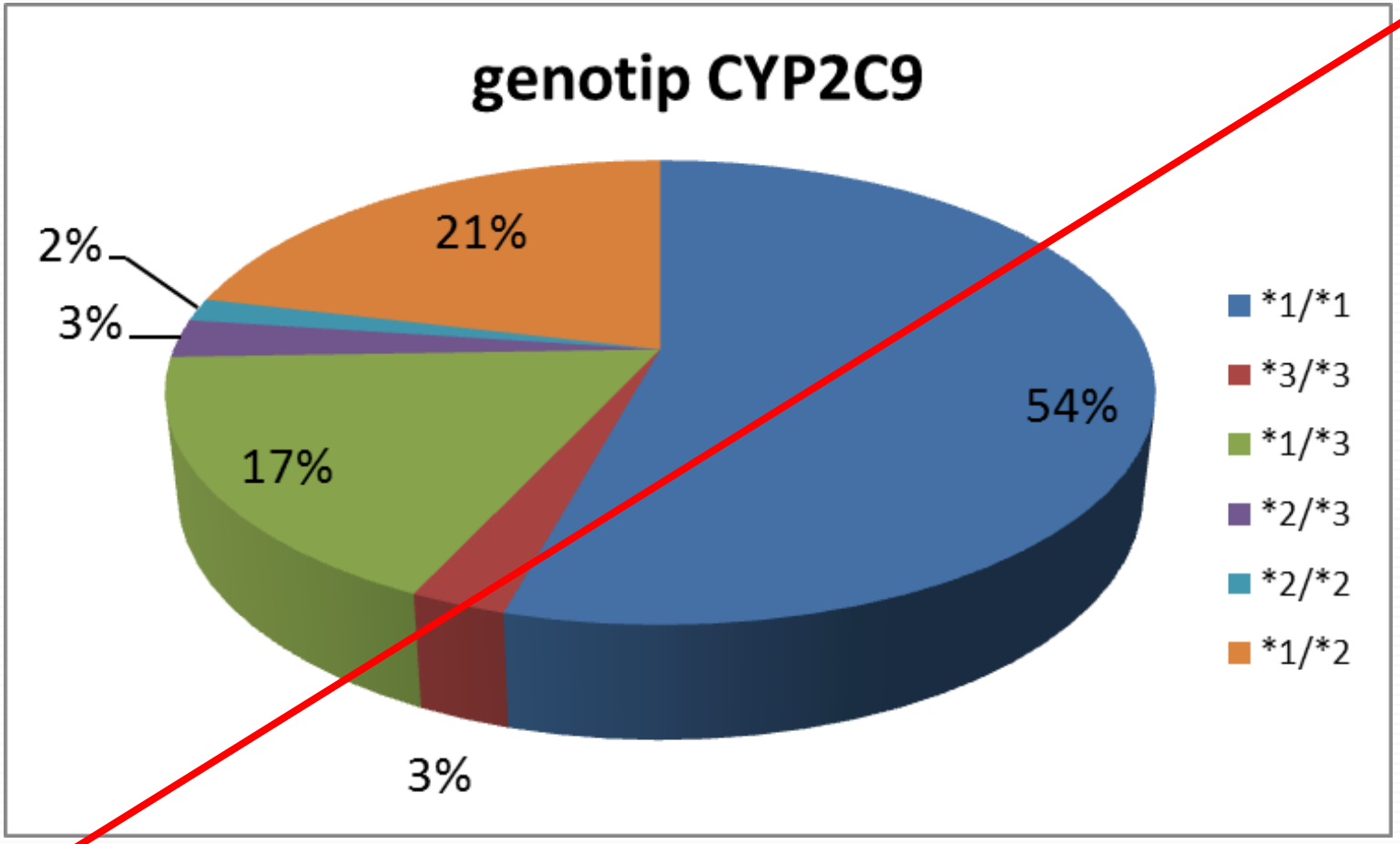
n = 188

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



n = 188

Tortni prikaz



Urejanje numeričnih spremenljivk

- Nezvezne numerične spremenljivke
  - Preštejemo frekvence posameznim celoštevilskim vrednostim – frekvenčna tabela

    Vsaka vrednost (kategorija) ima svoj razred, če je različnih vrednosti malo
- Zvezne numerične spremenljivke
  - Razvrščanje v ranžirno vrsto
  - Frekvenčna porazdelitev zveznih numeričnih spremenljivk, nato izrišemo **histogram**
  - Izračun srednjih vrednosti in mer razpršenosti (variabilnosti) podatkov

Srednje vrednosti

- Aritmetična sredina ali povprečje
- Geometrična sredina
- Modus
- Mediana

Aritmetična sredina (vzorca)

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} \quad ali \quad \mu$$

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

Lastnosti aritmetične sredine

- Če velja, da za vsak $y_i = x_i + c$, potem je $\bar{y} = \bar{x} + c$.
- Če velja, da za vsak $y_i = cx_i$, potem je $\bar{y} = c\bar{x}$.
- Če neko spremenljivko $x = (x_1, x_2, x_3, \ldots x_n)$ transformiramo v $y = (y_1, y_2, y_3, \ldots y_n)$, tako da velja za vsak $y_i = c_1x_i + c_2$, potem je $\bar{y} = c_1\bar{x} + c_2$.

Tehtana aritmetična sredina

- Aritmetična sredina vrednosti, ki so upoštevane z utežmi (w):

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i} w}$$

Geometrična sredina

- Seštejemo logaritemske vrednosti vseh enot, delimo s številom enot:

$$\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

- Geometrična sredina ($\overline{x}$) je antilogaritem od $\overline{\log x}$.

$$\overline{x} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$$

- Povezana z log-normalno porazdelitvijo.

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

Mediana (vzorca)

- Podatke moramo razvrstiti po velikosti.
- Tista vrednost spremenljivke, od katere ima polovica enot manjše ali enake, polovica enot pa večje ali enake vrednosti spremenljivke.
- Če je n liho število: mediana enaka vrednosti srednje enote oziroma
  - Mediana je m-ta največja vrednost, pri čemer je $m = (n+1)/2$.
- Če je n sodo število: mediana je povprečje vrednosti srednjega para podatkov oziroma
  - Mediana je povprečje m_1-*te* in m_2-*te* največje vrednosti, pri čemer je $m_1 = n/2$ in $m_2 = n/2 + 1$.
- 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
  - 10. centil = 1. decil
  - 90. centil = 9. decil

Določanje kvantilov

p = (per)centil (ima vrednosti od 1-100)

n= število enot

- n·p/100 ni celo število → k je navzdol zaokrožen n·p/100
  - Vrednost kvantila p je (k +1)-ta največja vrednost:
- n·p/100 je celo število → definiramo m_1 = n·p/100 in m_2 = n·p/100 +1
  - Vrednost kvantila p je povprečje med m_1-to in m_2-to največjo vrednostjo:

5 različnih načinov računanja centilov (SPSS)

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

Primer s kvantili (SPSS)

Koncentracija varfarina [mg/mL]	Mean	1,9703
	Median	2,0250
	Minimum	0,5300
	Maximum	2,9800
	Percentile 10	1,2790
	Percentile 25	1,7400
	Percentile 75	2,2650
	Percentile 90	2,5210

Percentile = centili izračunani po metodi HAVERAGE.

Primer s kvantili (Excel)

Uporabljene funkcije MEDIAN, PERCENTILE, AVERAGE, MIN, MAX in QUARTILE.

Koncentracija varfarina [mg/mL]	AVERAGE	1,9703		
	MEDIAN	2,0250		
	MIN	0,53		
	MAX	2,98	*p*	*i*
	1. decil	1,2940	0,1	19,7
	1. kvartil	1,7400	0,25	47,75
	3. kvartil	2,2625	0,75	141,25
	9. decil	2,5130	0,9	169,3

Funkcija PERCENTILE izračuna centile tako, da določi $i = (n\text{-}1)p + 1$, pri čemer je n število enot, p pa centil izražen v deležu. V primeru, da i ni celo število, je vrednost kvantila (X) določena z linearno interpolacijo oz. kot tehtano aritmetično sredino med X_i in X_{i+1}.

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 podatkov

- Vsota odklonov od povprečne vrednosti?
- 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

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

  - oznaka s2 za varianco vzorca

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

- Standardni odklon oziroma standardna deviacija
  - Kvadratni koren variance

Lastnosti variance in standardnega odklona

- Če definiramo dve spremenljivki $x = (x_1, x_2, x_3, \ldots x_n)$ in $y = (y_1, y_2, y_3, \ldots y_n)$, tako da velja za vsak $y_i = x_i + c$, potem je $s_y^2 = s_x^2$.

- Če definiramo dve spremenljivki $x = (x_1, x_2, x_3, \ldots x_n)$ in $y = (y_1, y_2, y_3, \ldots y_n)$, tako da velja za vsak $y_i = c\,x_i$, potem je $s_y^2 = c^2 s_x^2$ oziroma $s_y = c s_x$

Koeficient variacije (KV ali RSD)

$$KV(RSD) = \frac{s}{\bar{x}}$$

- Mera relativne variabilnosti: standardna deviacija utežena z aritmetično sredino, po navadi 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

Urejanje numeričnih spremenljivk
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 v plazemskih vzorcih določili koncentracijo kalcidiola. Dobili smo naslednje vrednosti (nmol/L):

59.0; 23.0; 142; 168; 22.0;
64.0; 228; 59.0; 32.0; 145;
38.0; 64.0; 164.0; 5.00; 147;
41.0; 112; 21.0; 249; 140;
133; 27.0; 160; 64.0; 63.0; 93.0

Rangiranje: primer kalcitriol

Vrednost [nmol/L]	Rang	Vrednost [nmol/L]	Rang
5.00	1	133	16
21.0	2	140	17
22.0	3	142	18
23.0	4	145	19
27.0	5	147	20
32.0	6	160	21
38.0	7	164	22
41.0	8	168	23
59.0	9.5	228	24
59.0	9.5	249	25
63.0	11		
64.0	13		
64.0	13		
64.0	13		
112	15		

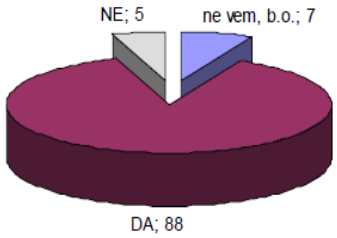
Uporaba pri neparametričnih statističnih testih

Univerza v Ljubljani
Fakulteta za farmacijo

PRIMERI

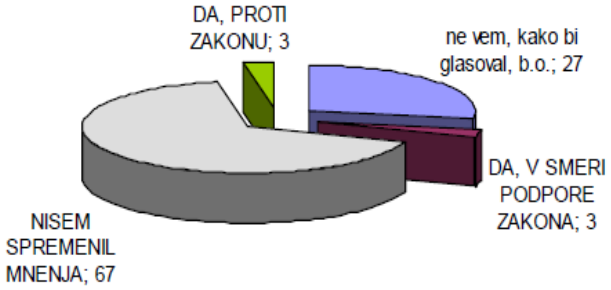
3-D prikaz

ALI MENITE, DA SO V SLOVENIJI POTREBNE REFORME?



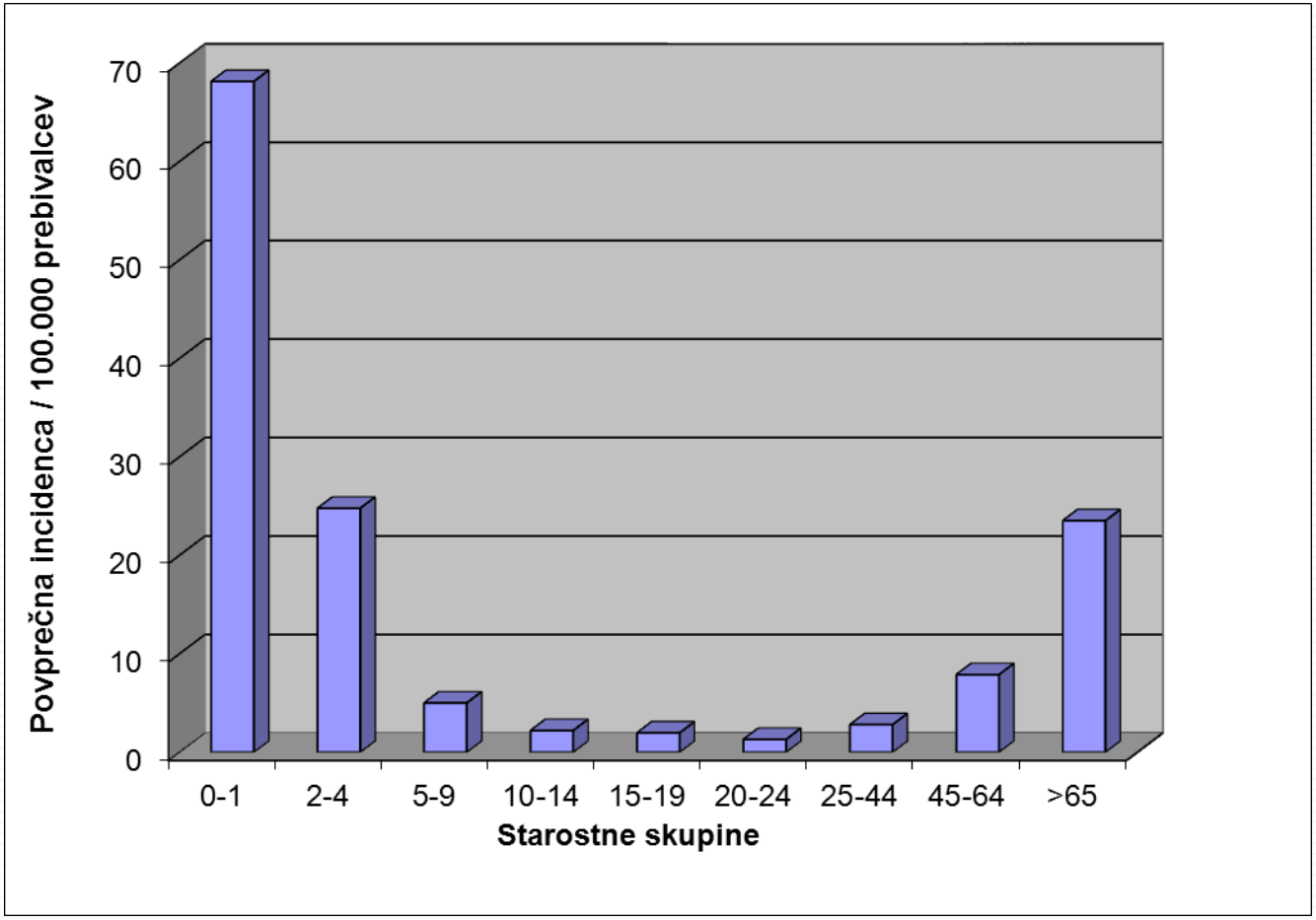
CJMMK, Politbarometer, marec 2011, N=926

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

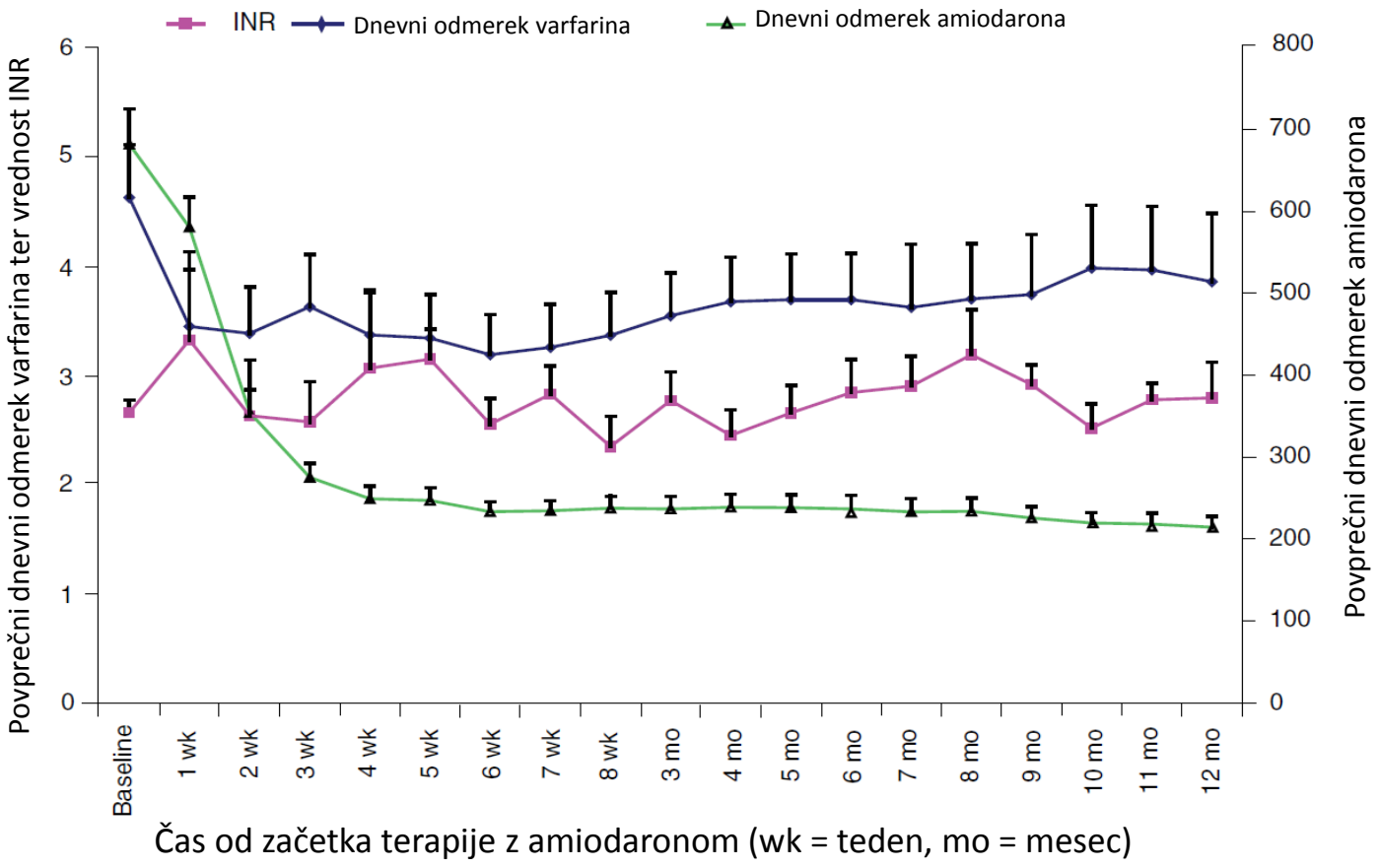


CJMMK, Politbarometer, marec 2011, N=926

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

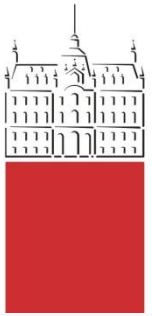


Vpliv različnih dejavnikov na učinek antikoagulantnih zdravil (Locatelli I. in Oblak E.)



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)

Univerza v Ljubljani
Fakulteta za farmacijo



Urejanje in prikazovanje podatkov II

Farmacevtska informatika
2011/2012, 1. letnik EMŠF

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

Ljubljana, 9. 3. 2012

Urejanje numeričnih spremenljivk

- Nezvezne numerične spremenljivke
  - Preštejemo frekvence posameznim celoštevilskim vrednostim – frekvenčna tabela

    Vsaka vrednost (kategorija) ima svoj razred, če je različnih vrednosti malo
- Zvezne numerične spremenljivke
  - Razvrščanje v ranžirno vrsto
  - Frekvenčna porazdelitev zveznih numeričnih spremenljivk, nato izrišemo **histogram**
  - Izračun srednjih vrednosti in mer razpršenosti (variabilnosti) podatkov

Frekvenčna porazdelitev numeričnih spremenljivk

- **Meje razredov**: najmanjša in največja možna vrednost v razredu, kam spadajo mejne vrednosti?
- **Širina razreda (j):** razlika med zgornjo in spodnjo mejo razreda
- **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, $g = f/j$

Frekvenčna porazdelitev numeričnih spremenljivk

- Histogram: 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)

Primer frekvenčne tabele: Podatki kalcitriol

j = 50 nmol/L
k= 5

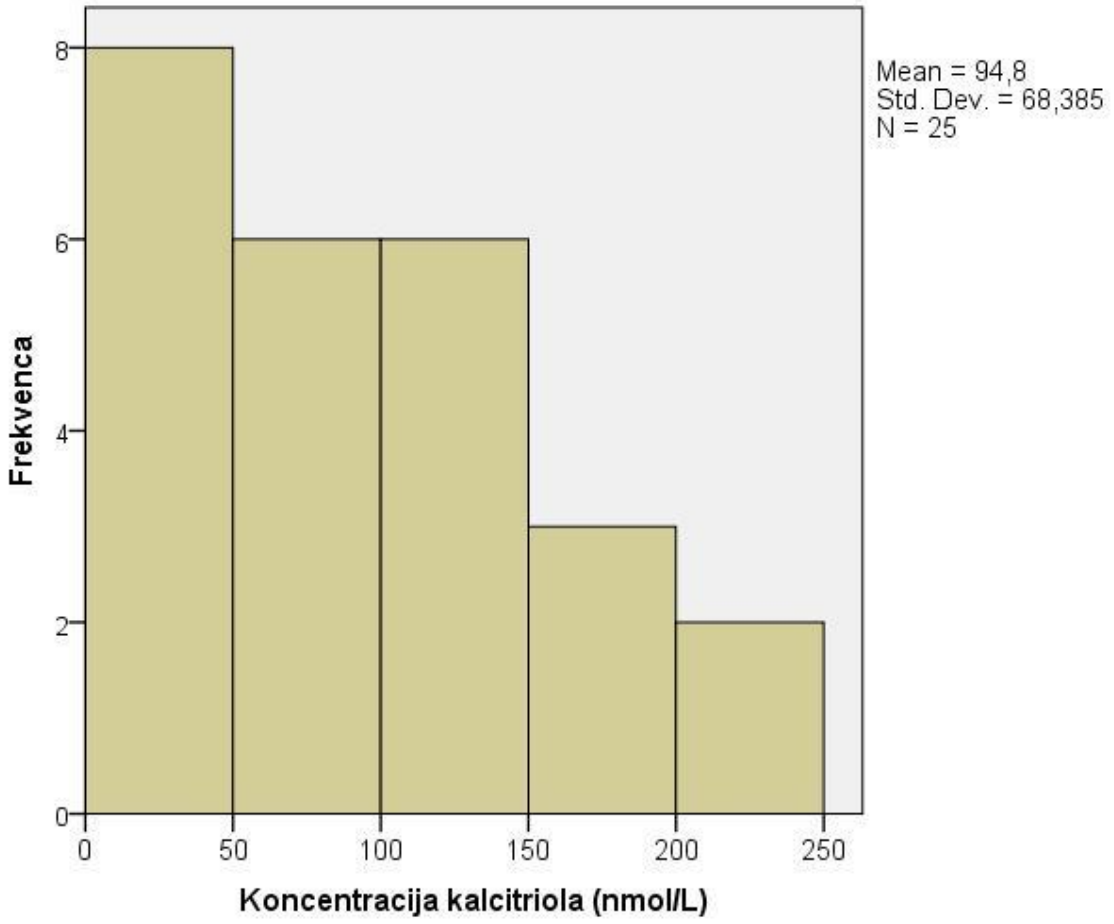
5,00 21,0 22,0 23,0 27,0 32,0 38,0 41,0
59,0 59,0 63,0 64,0 64,0 64,0 112 133
140 142 145 147 160 164 168 228 249

Št.	Meje razreda	Širina (j)	Sredina	f	f%	F	g
1.	$0 \leq x < 50$	50	25,0	8	32%	0	0.16
2.	$50 \leq x < 100$	50	75,0	6	24%	8	0.12
3.	$100 \leq x < 150$	50	125,0	6	24%	14	0.12
4.	$150 \leq x < 200$	50	175,0	3	12%	20	0.06
5.	$200 \leq x < 250$	50	225,0	2	8%	23	0.04
	skupaj			25	100%	25	

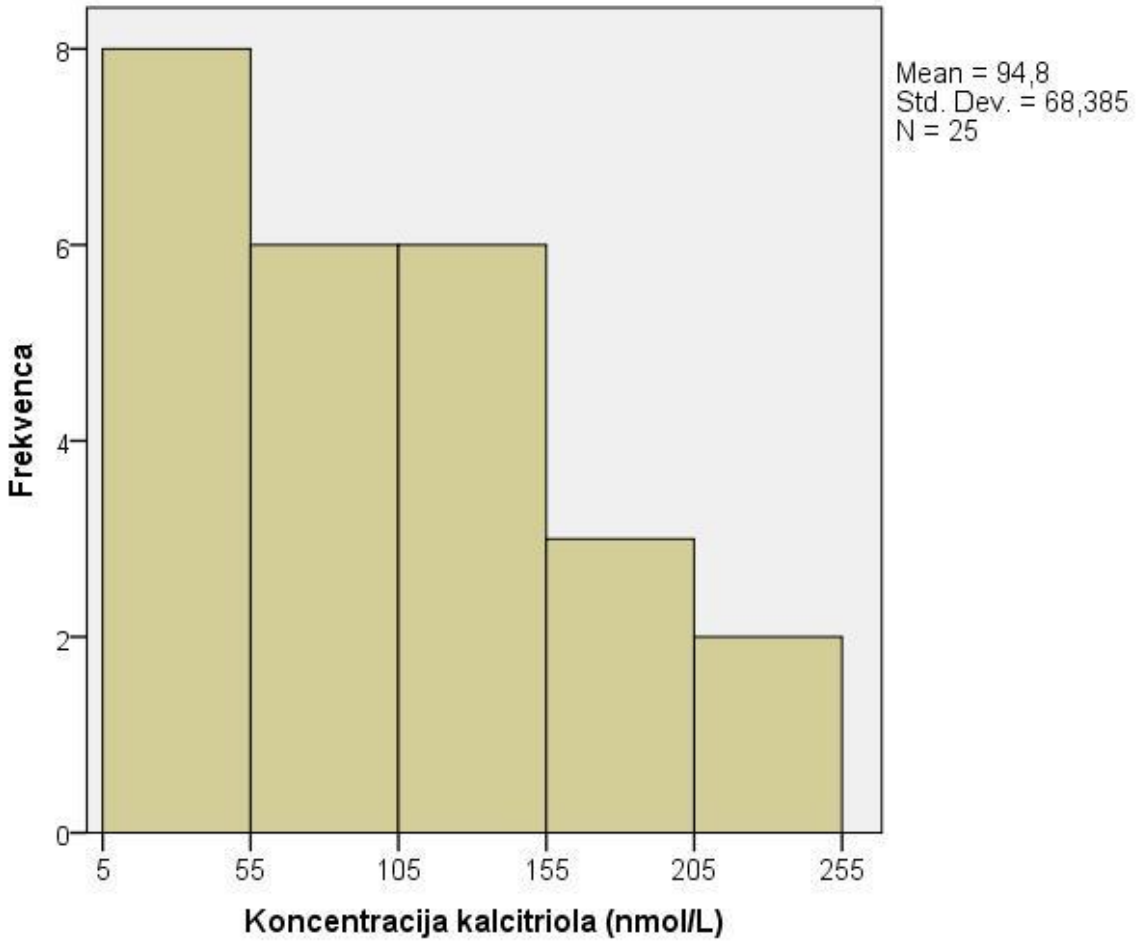
$(k-1)\cdot j \leq (x_{MAX} - x_{MIN}) \leq k\cdot j$

$(k-1)\cdot j \leq (x_{MAX} - 0) \leq k\cdot j$

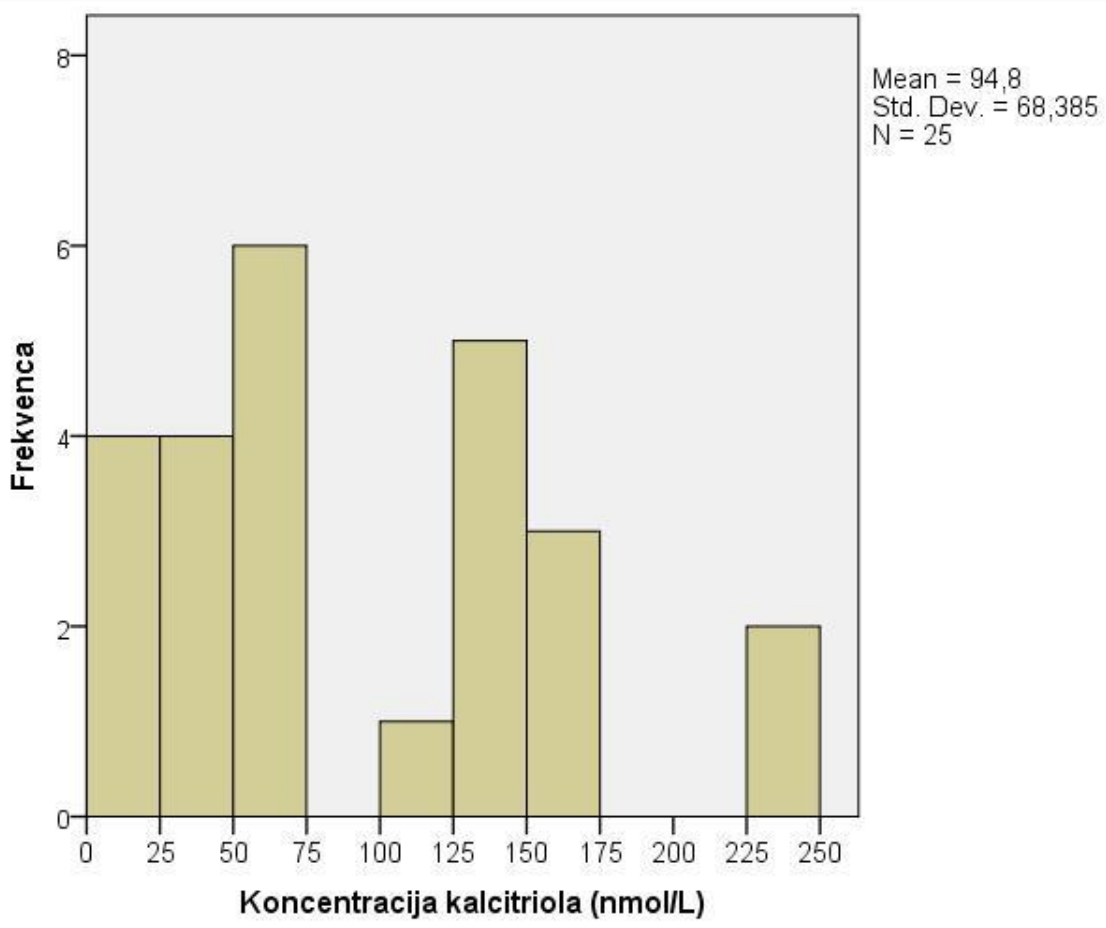
Histogram (kalcitrol; j=50; k=5)



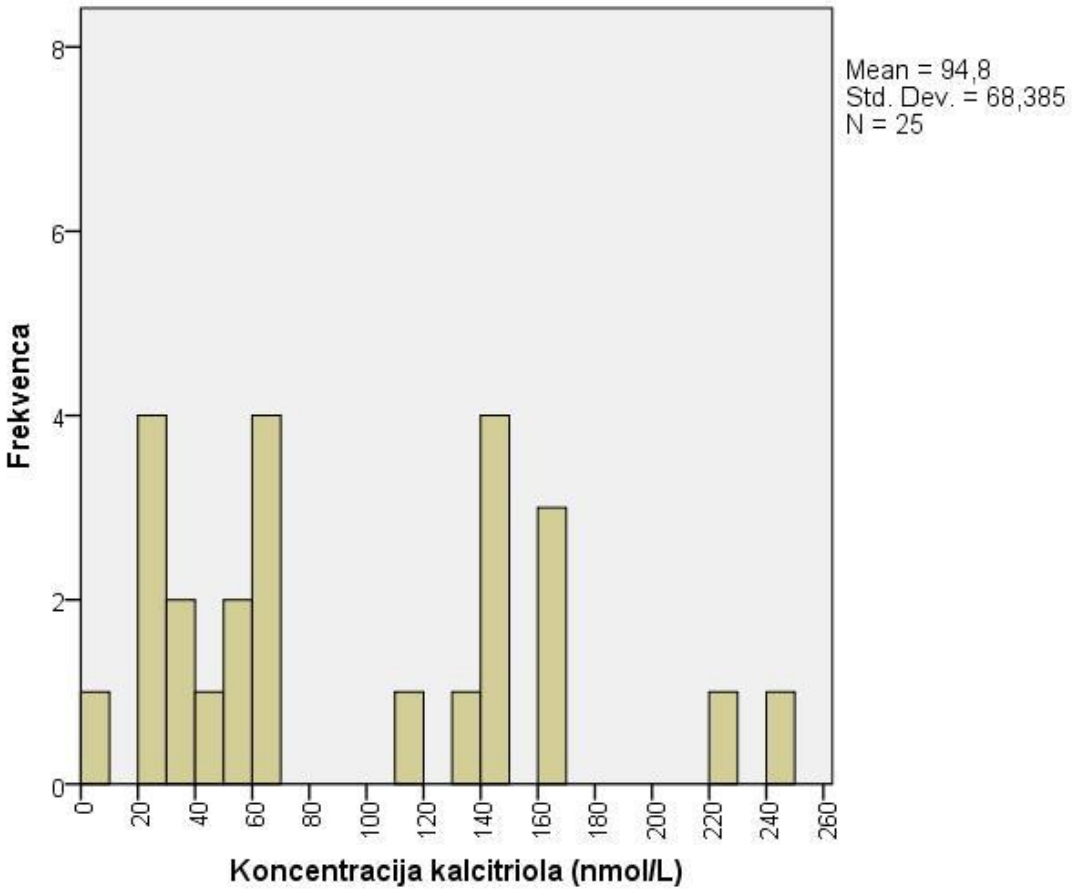
Histogram (kalcitrol; j=50; k=5; min = 5,0)



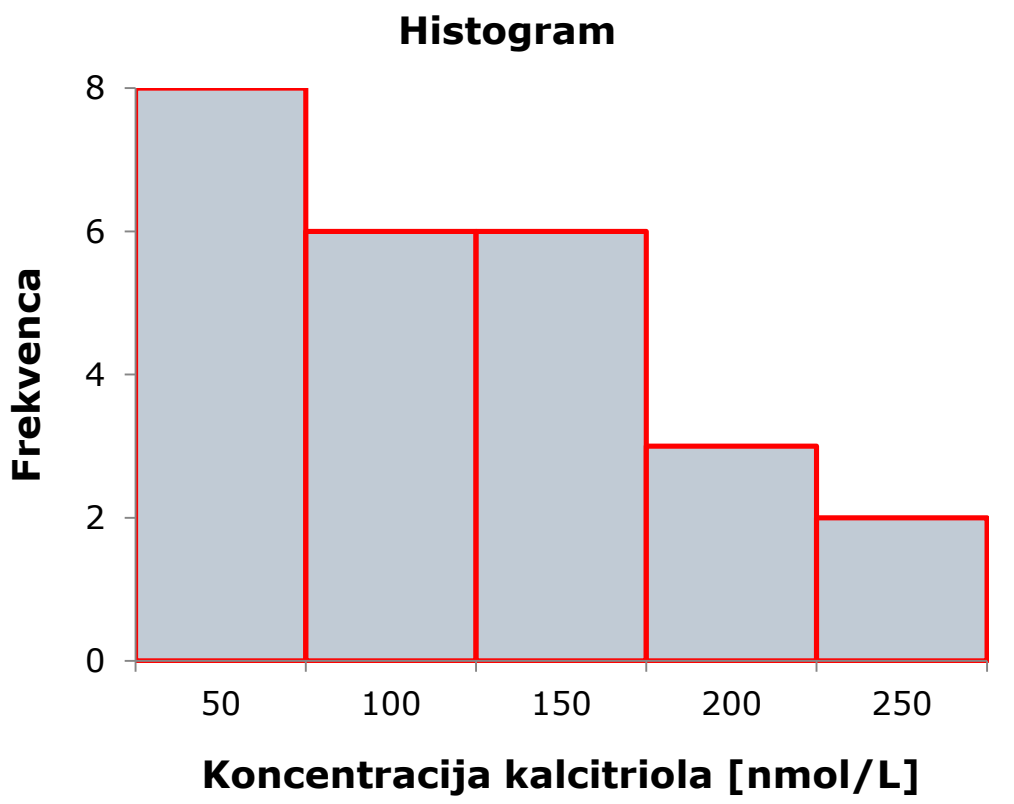
Histogram (kalcitrol; j=25; k=10)



Histogram (kalcitrol; j=10; k=25)

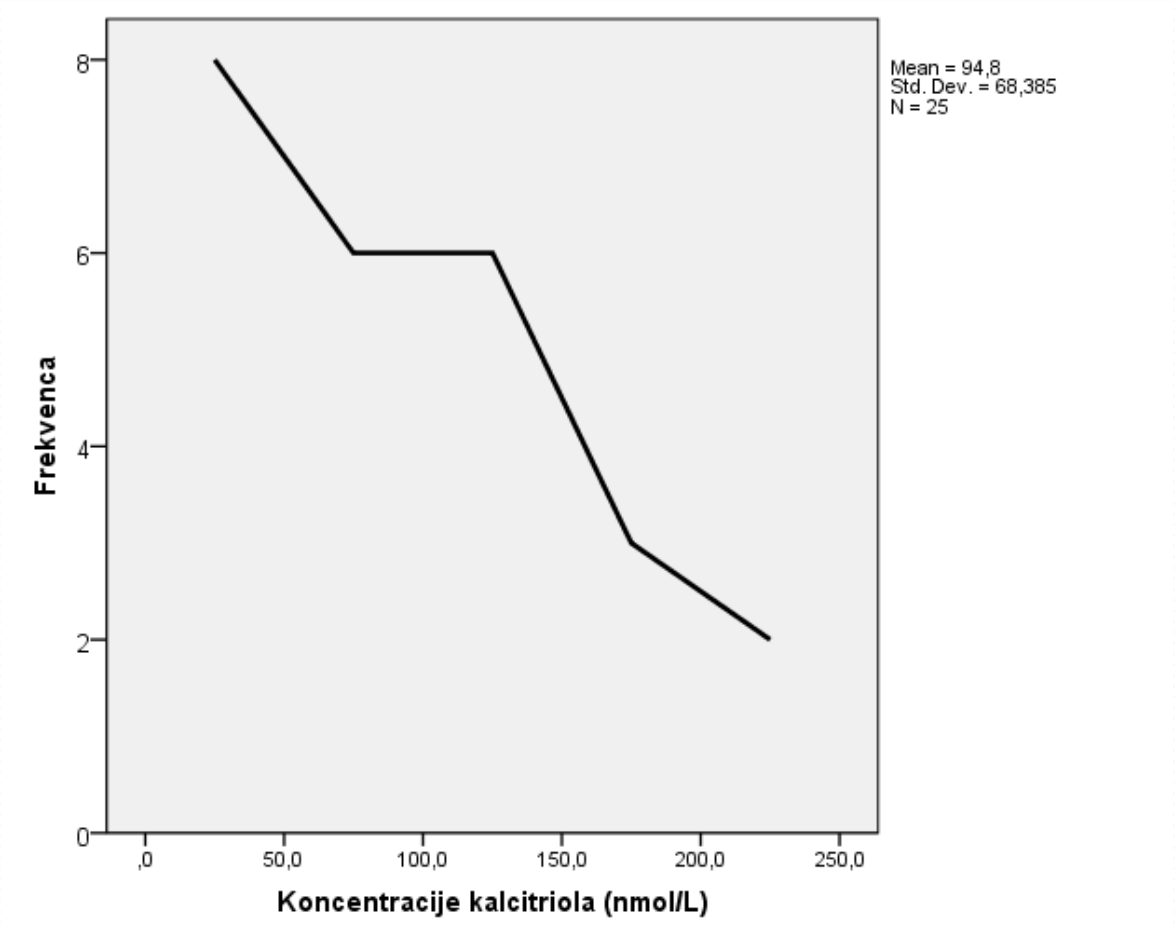


Histogram
(Excel – Add on: Data analysis)

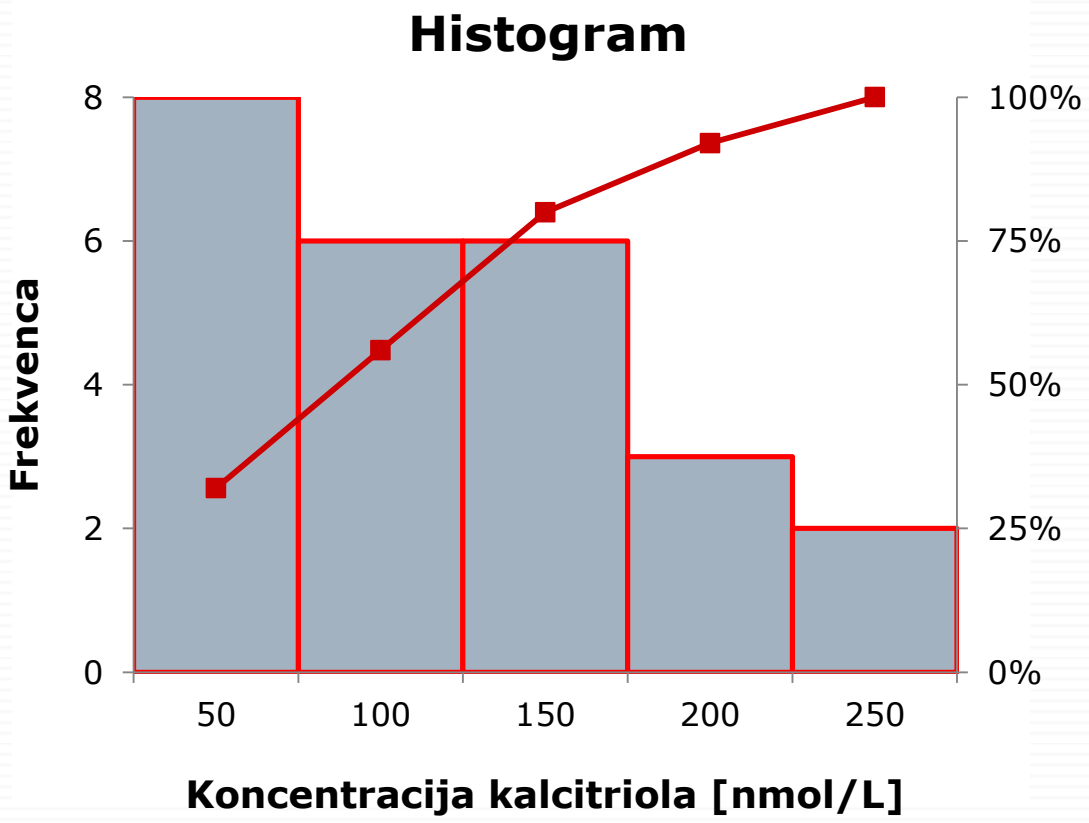


Zveznost x osi?

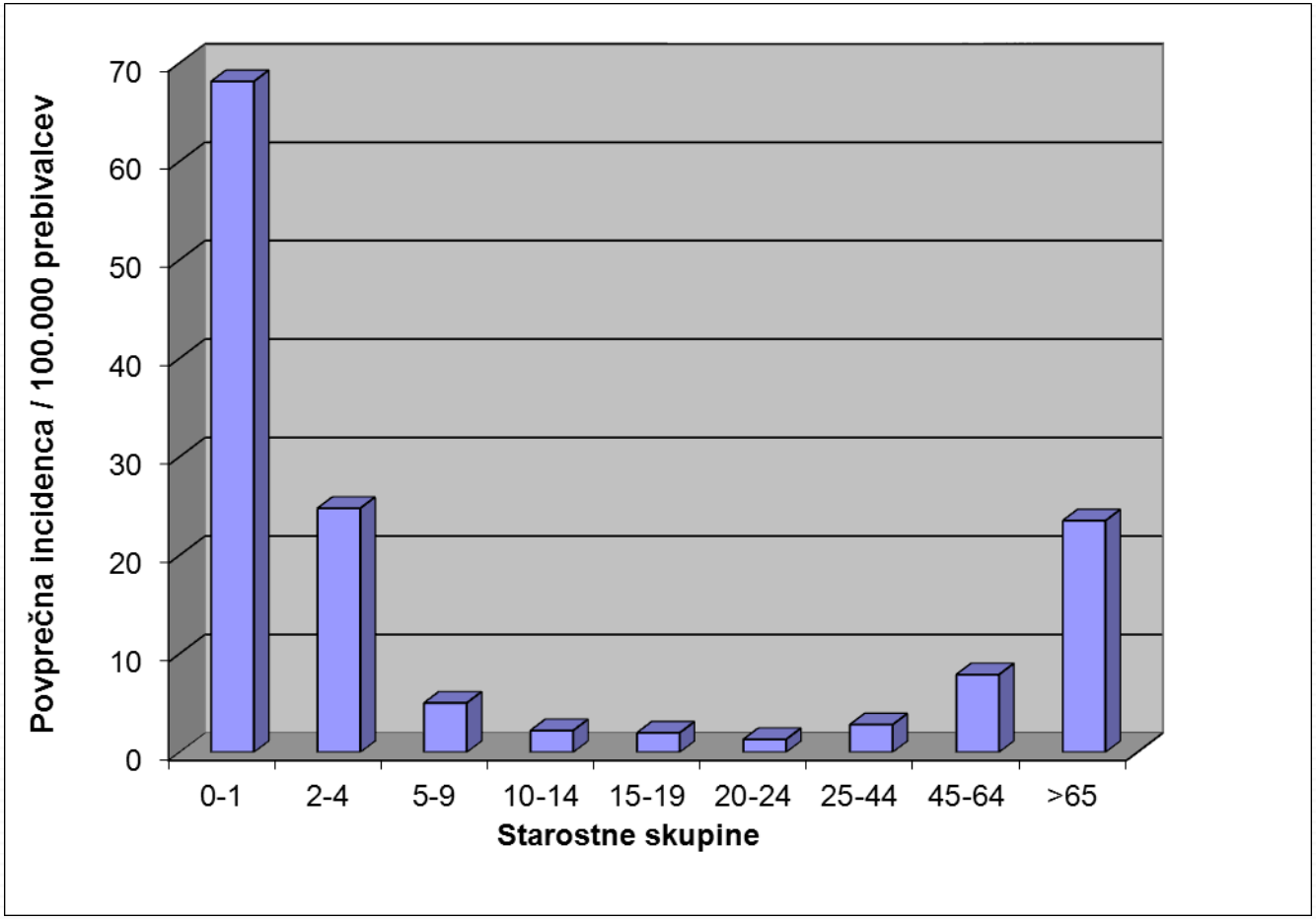
Frekvenčni poligon (kalcitriol; j=50; k=5)



Kumulativna frekvenčna porazdelitev



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



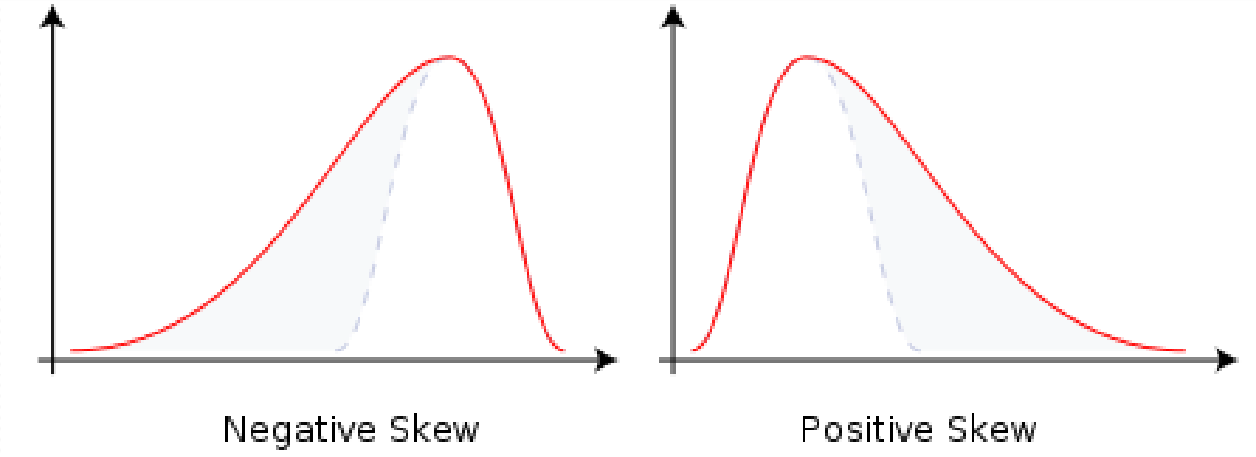
Uporaba gostote frekvence

- Širine razredov niso enake.

Starostna skupina	Incidenca (frekvenca)	Širina razreda	Gostota frekvence	Spodnja meja	Zgornja meja
0-1	68	2	34	0	2
2-4	25	3	8.3	2	5
5-9	5.0	5	1.0	5	10
10-14	2.0	5	0.40	10	15
15-19	2.0	5	0.40	15	20
20-24	1.3	5	0.26	20	25
25-44	2.8	20	0.14	25	45
45-64	7.9	20	0.39	45	65
>65	24	15	1.6	65	80

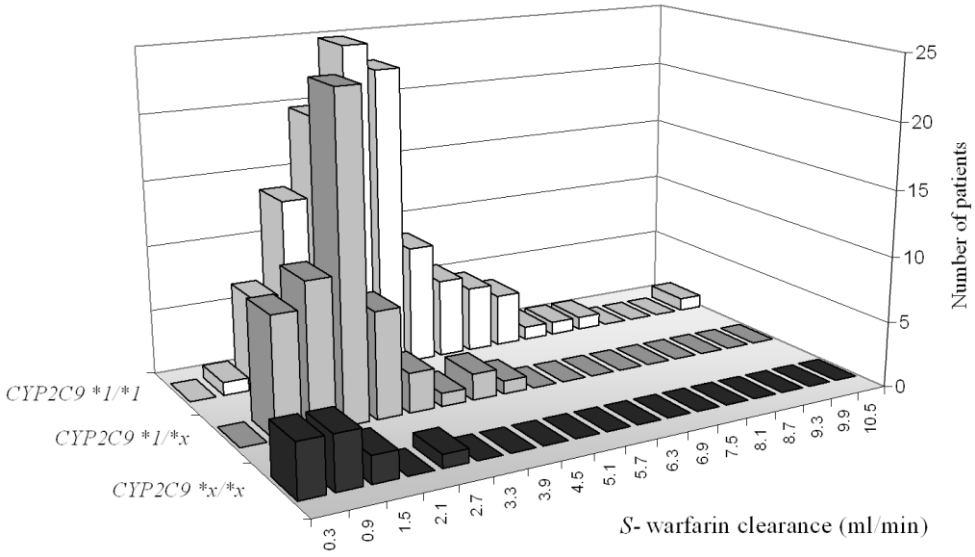
Simetričnost oz. asimetričnost frekvenčne porazdelitve

- Pozitivno asimetrična (iztegnjena v desno) – ang. positively skewed
  - večja gostota pri manjših vrednostih
- Negativno asimetrična (iztegnjena v levo) – ang. negatively skewed
  - večja gostota pri večjih vrednostih



Če je porazdelitev podatkov simetrična, potem je mediana enaka aritmetični sredini.

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



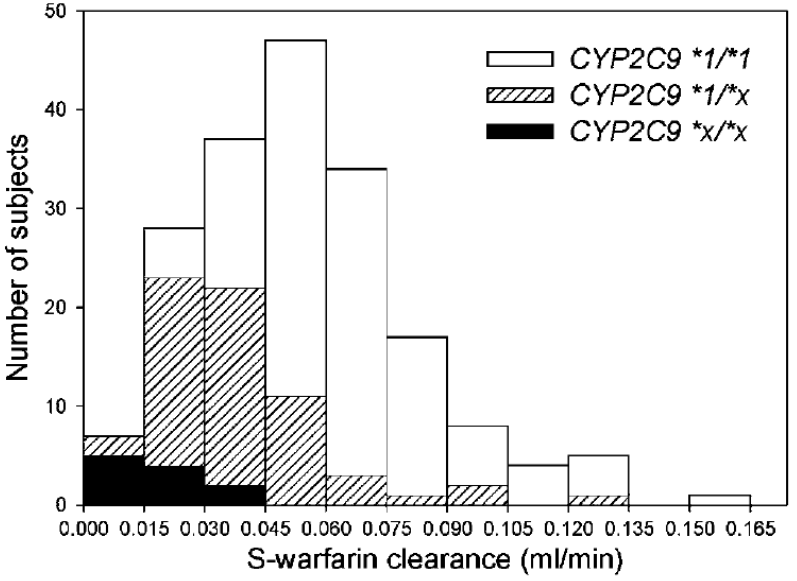
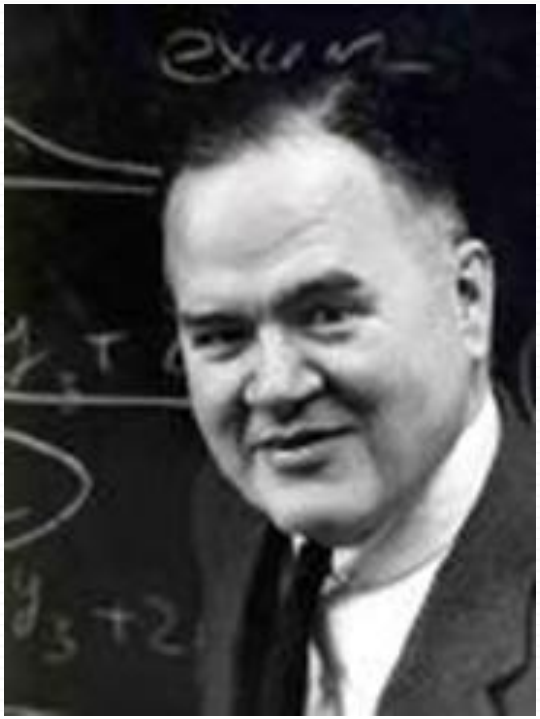


Figure 3 Frequency distribution ($N = 188$) of LBW normalized *S*-warfarin clearance according to *CYP2C9* genotype: two wild-type alleles (**1/*1*), one polymorphic allele (**1/*x*) and two polymorphic alleles (**x/*x*).

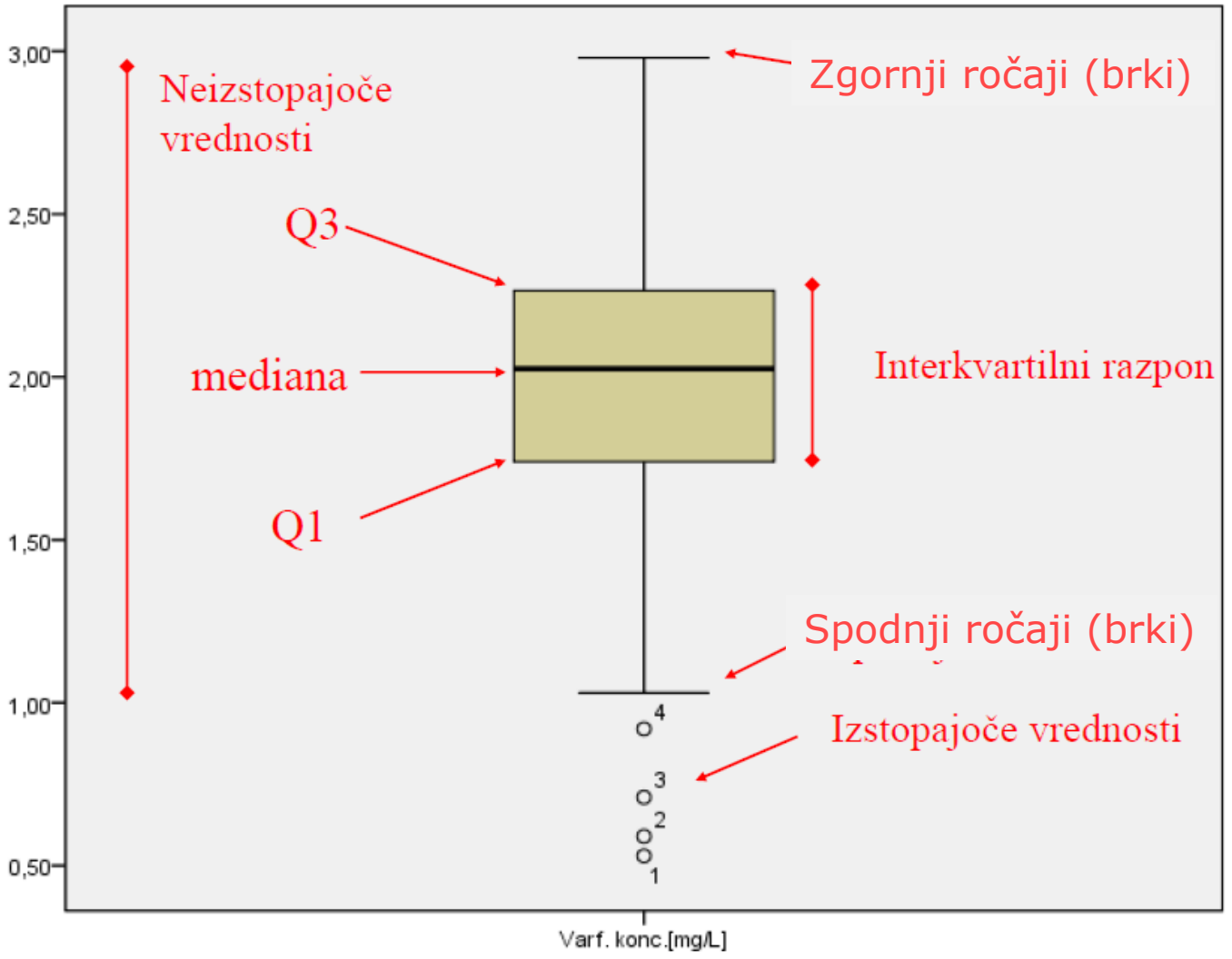
John Wilder Tukey

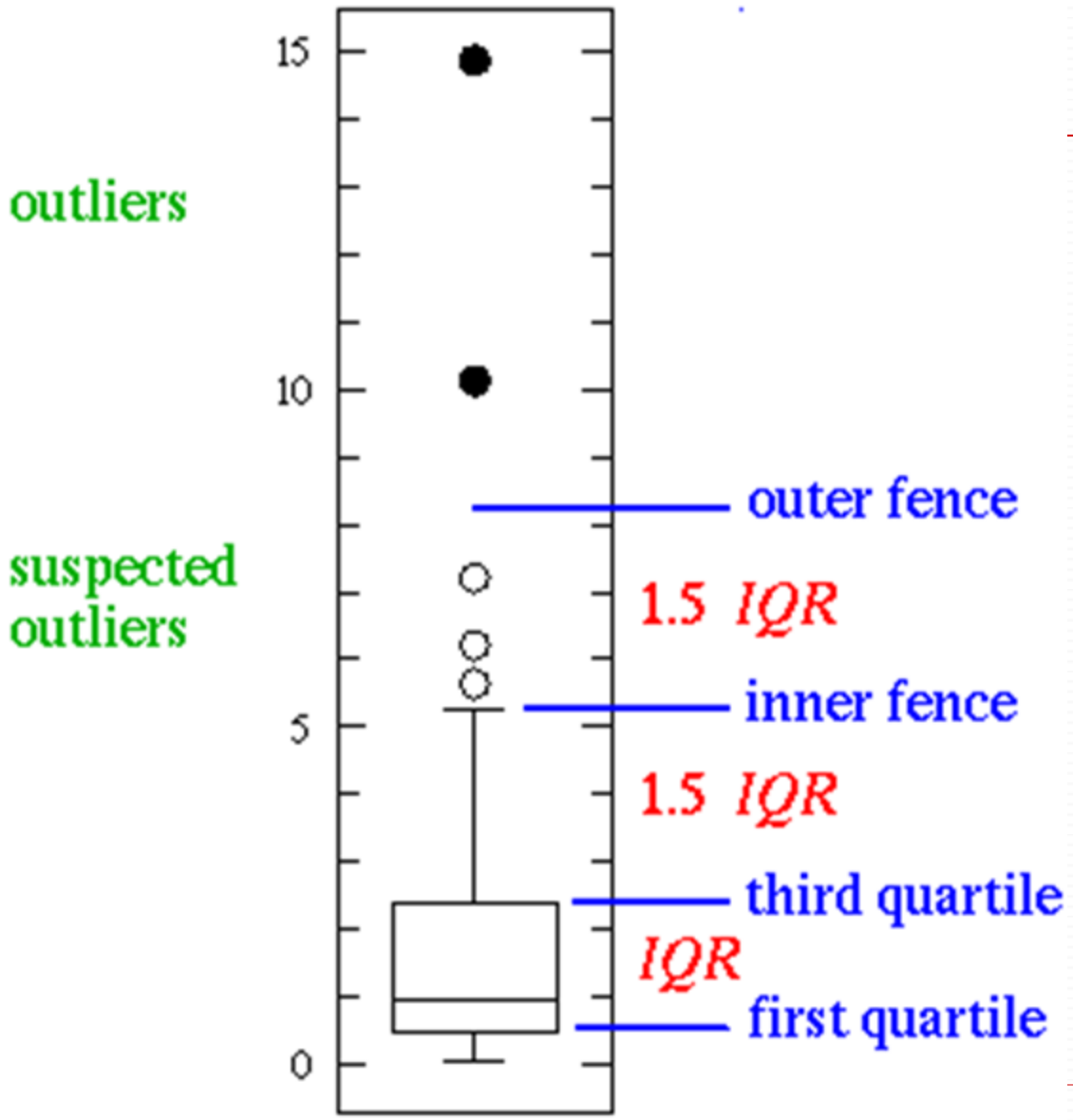


- Ameriški statistik
- 1915-2000

- Exploratory Data Analysis
  - box and whisker plot (box plot)
  - Stem-and-Leaf plot (stem plot)

Kvantilni diagram - (box and whisker plot)




outliers
suspected outliers
15
10
5
0
outer fence
1.5 IQR
inner fence
1.5 IQR
third quartile
IQR
first quartile

Izstopajoče vrednosti
Extremne vrednosti

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

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

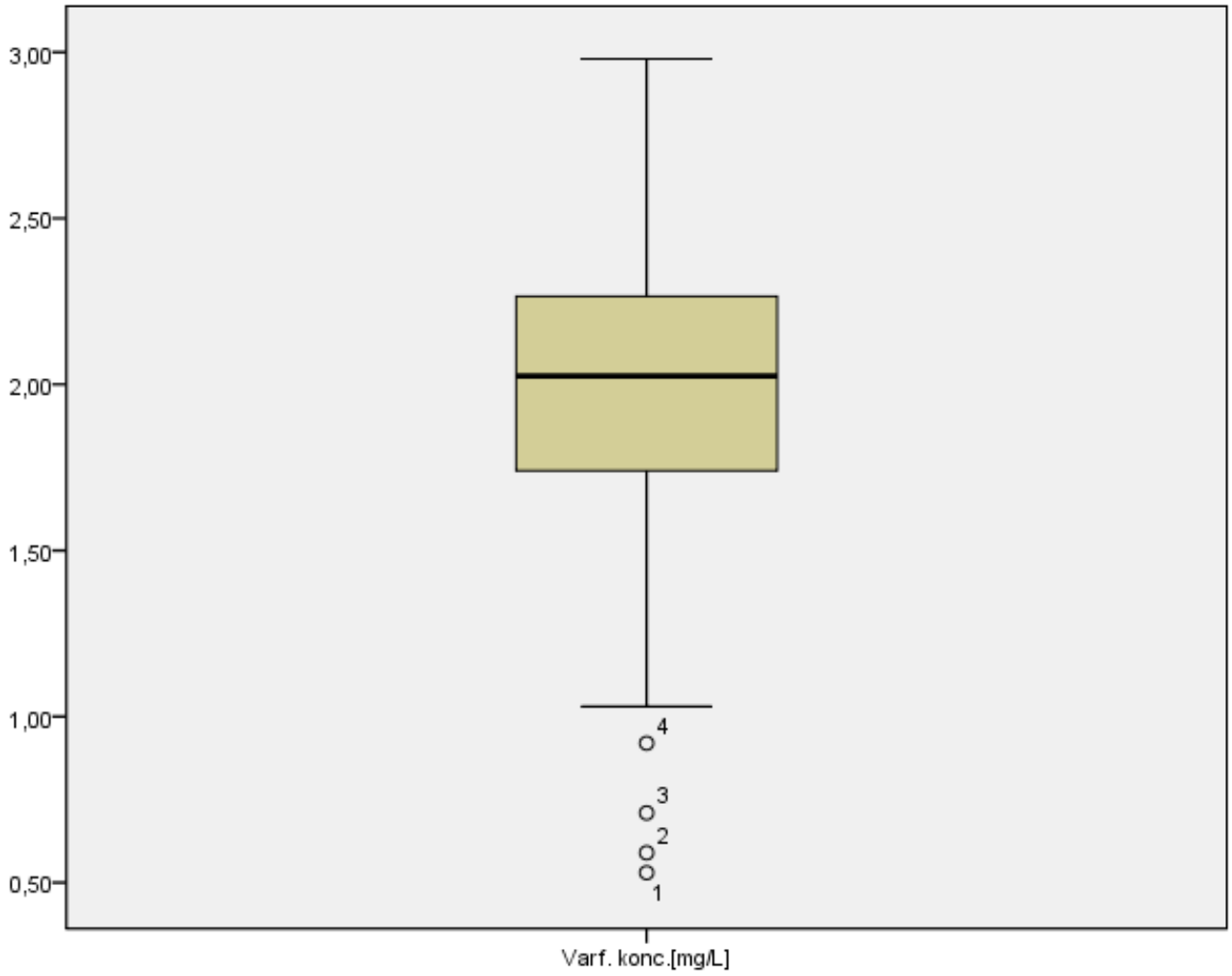
$$X > X_{Q3} + 3 \cdot \left(X_{Q3} - X_{Q1}\right)$$

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

Kvantilni diagram - postopek

- Določitev mediane, 1. kvartila (X_{Q1}) in 3. kvartila (X_{Q3}).
- Izris škatle.
- Izračun
  - Medkvartilni razmik: $IQR = X_{Q3} - X_{Q1}$
  - Zgornja notranja ograja: $X_{Q3} + 1{,}5 \cdot IQR$
  - Zgornja zunanja ograja: $X_{Q3} + 3 \cdot IQR$
  - Spodnja notranja ograja: $X_{Q1} - 1{,}5 \cdot IQR$
  - Spodnja zunanja ograja: $X_{Q1} - 3 \cdot IQR$
- Določitev najvišje in najnižje neizstopajoče vrednosti.
- Izris ročajev.
- Izris izstopajočih vrednosti (točke).
- Izris ekstremno izstopajočih vrednosti (točke).

Kvantilni diagram – koncentracija varfarina (mg/L)



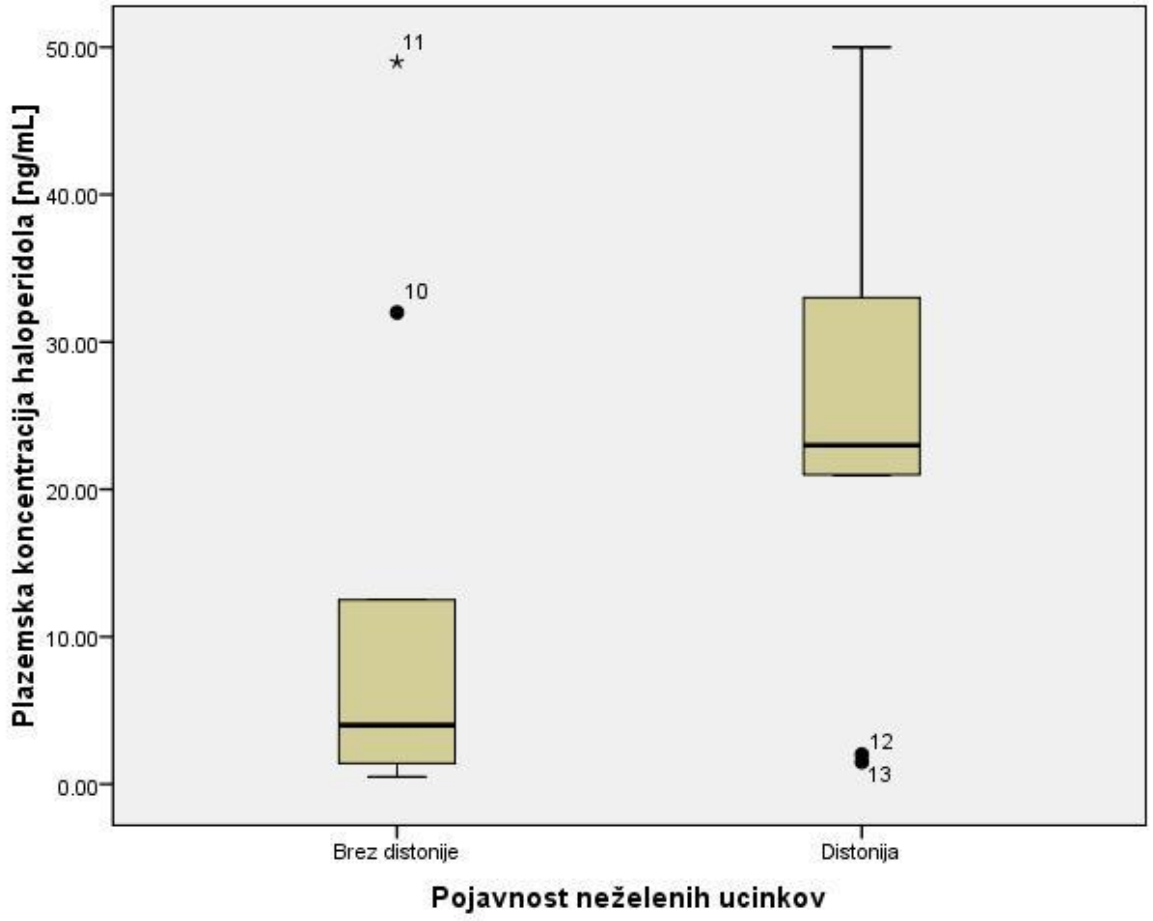
n = 188

Me	2,025
X_{Q1}	1,74
X_{Q3}	2,265
IQR	0,525
ograje	3,0525
	3,84
	0,9525
	0,165

Kvantilni diagram - lastnosti

- Horizontalno ali vertikalno postavljen
- Princip 5 točkovnega opisa podatkov (ang. five-number summary)
- Ročaja se včasih zaključita
  - z minimalno oz. maksimalno vrednostjo
  - s prvim oz. devetim decilom
- Opisuje porazdelitev podatkov (simetričnost/asimetričnost)
- Opisuje vpliv dejavnikov na proučevano spremenljivko

Plazemska koncentracija haloperidola in pojav distonije



Plazemska koncentracija haloperidola [ng/mL]	
Bolniki brez distonije	Bolniki z distonijo
0,5	1,5
0,5	2
1,4	**21**
1,43	21
2,54	21,5
4	**23**
4,63	25
12,5	33
12,5	**33**
32	33
44	45

n = 11

Histogram s številkami
Stem-and-Leaf plot

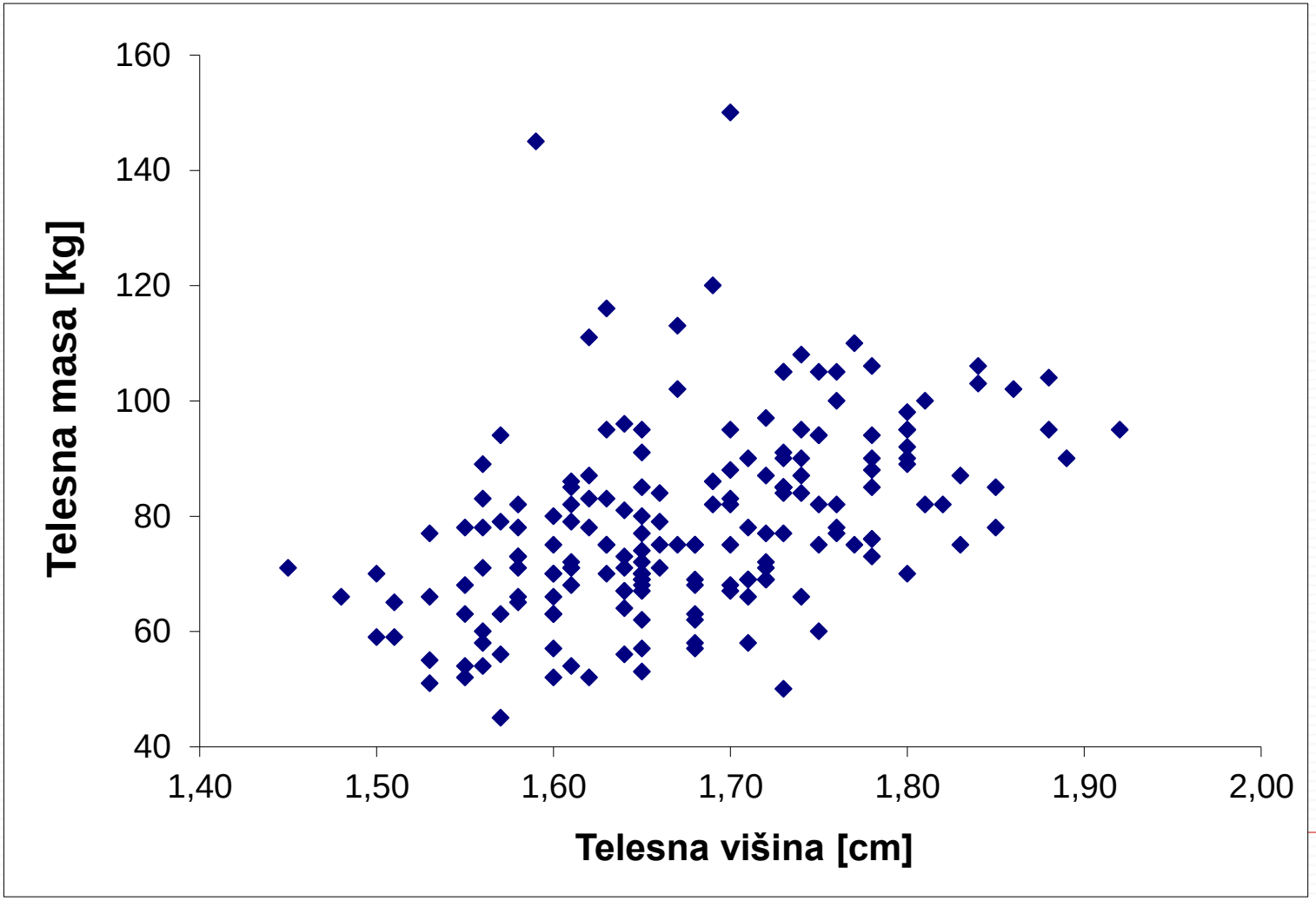
```
Frequency    Stem &  Leaf

    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)
```

0.53	1
0.59	2
0.71	3
0.92	4
1.03	5
1.03	6
1.10	7
1.12	8
1.12	9
1.16	10
1.16	11
1.17	12
1.17	13
1.21	14
1.24	15
1.25	16
1.25	17
1.27	18
1.28	19
1.30	20
1.31	21
1.33	22
1.33	23
1.34	24
1.43	25
1.49	26
1.49	27

Razsevni diagram

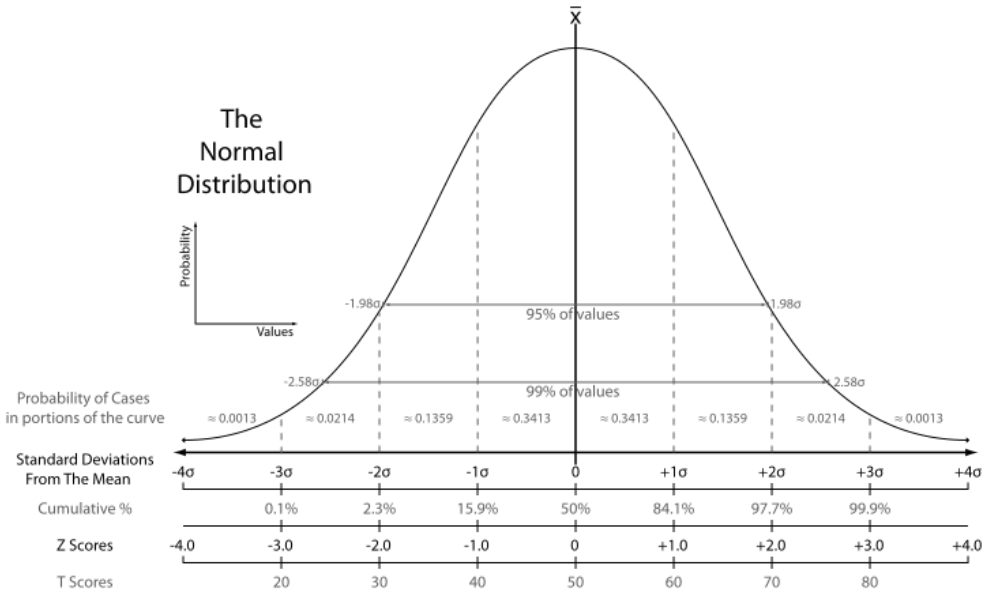


Vrste porazdelitev oz. distribucij

- Empirične (izkustvena)
- Teoretične: porazdelitve, ki jih dobimo s teoretičnim razglabljanjem ter s pomočjo matematičnih postopkov

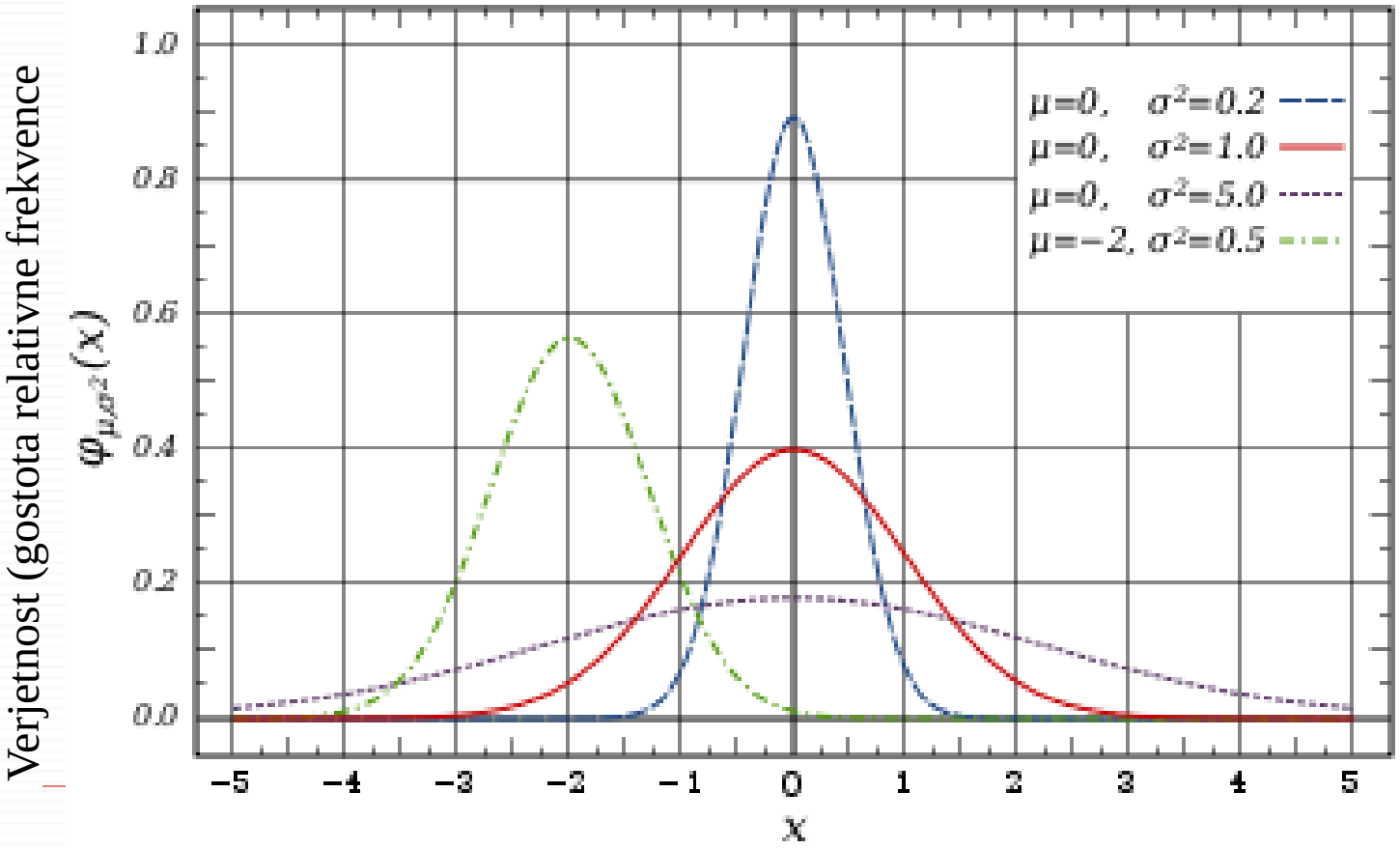
Normalna ali Gaussova porazdelitev

- Unimodalna
- Simetrična
- Zvonasta
- Max gostota pri aritmetični sredini
- Zvezno padajoča gostota
- N(μ,σ)



Standardizirana normalna porazdelitev

$$\rho_{(x)} = \frac{1}{\sigma * \sqrt{2\Pi}} * e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

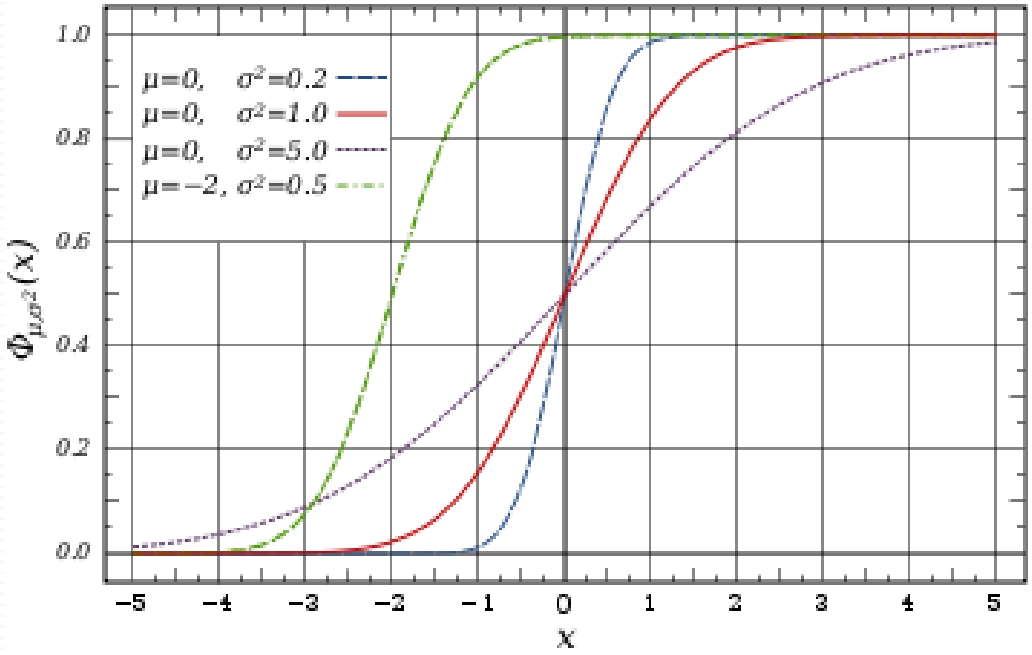


Vir: http://en.wikipedia.org/wiki/Normal_distribution

Kumulativna relativna frekvenca

- F(x): Kumulativna relativna frekvenca:
  - površina pod krivuljo poljubne normalne porazdelitve za vrednost spremenljivke od -∞ do +∞

$$\int_{-\infty}^{x} \rho_{(x)} dx = F_{(X)}$$

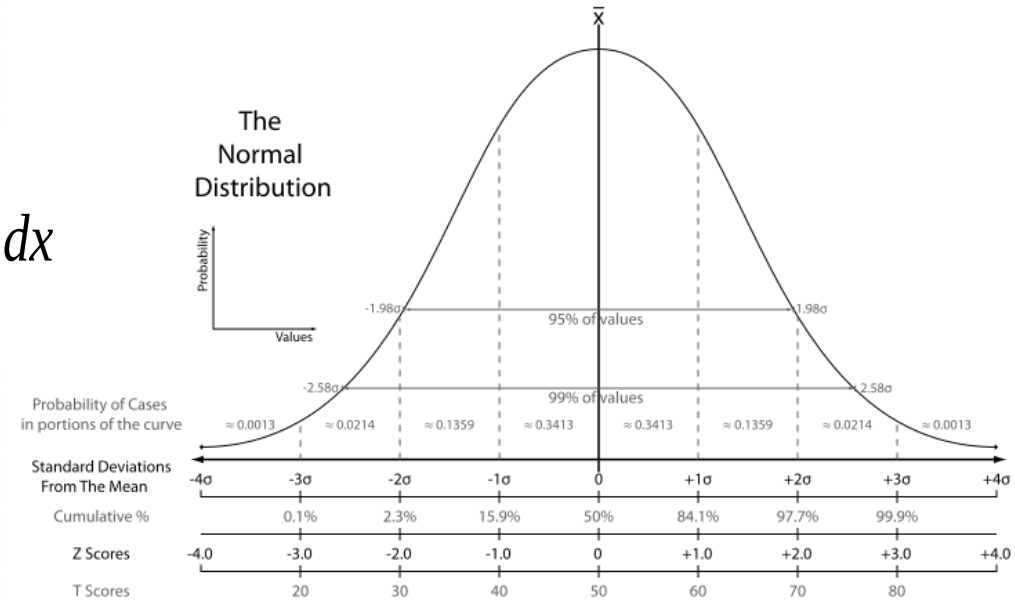


Normalna porazdelitev - lastnosti

$$F_{(\infty)} = \int_{-\infty}^{\infty} \rho_{(x)} dx = 1$$

$$F_{(x=\mu)} = \int_{-\infty}^{\mu} \rho_{(x)} dx = 0{,}5 = \int_{\mu}^{\infty} \rho_{(x)} dx$$

$$F_{(-x)} = 1 - F_{(x)}$$



$$F_{(x1<x<x2)}{}^{0} = \int_{x1}^{x2} \rho_{(x)} * dx = \int_{-\infty}^{x2} \rho_{(x)} * dx - \int_{-\infty}^{x1} \rho_{(x)} * dx = F_{(X2)}{}^{0} - F_{(X1)}{}^{0}$$

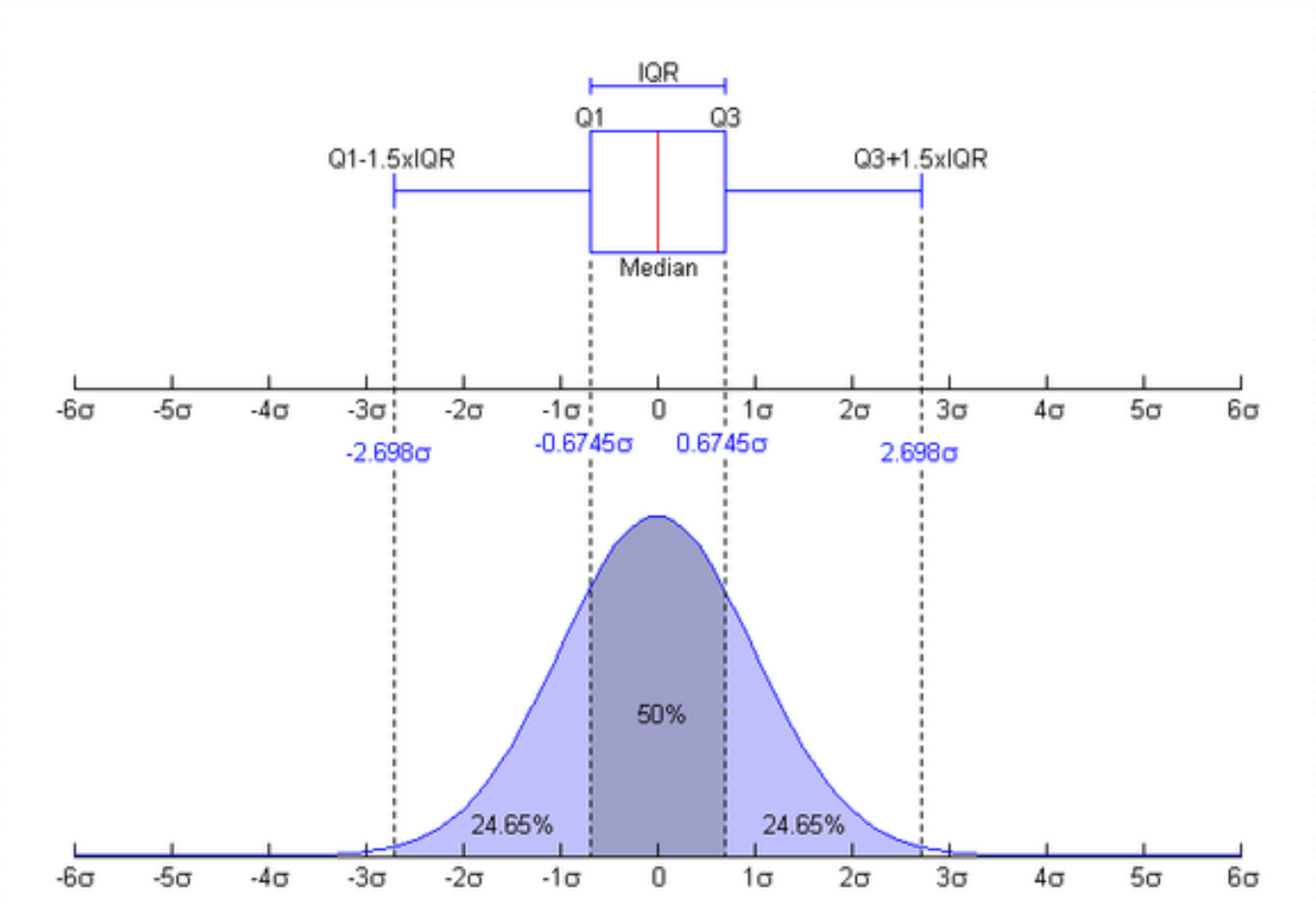
Normalna porazdelitev – standardni odklon

$$F_{(-1*\sigma 1<x<1*\sigma)}{}^{0} = 0{,}6825$$

$$F_{(-2*\sigma<x<2*\sigma)}{}^{0} = 0{,}9545$$

$$F_{(-3*\sigma<x<3*\sigma)}{}^{0} = 0{,}9973$$

Normalna porazdelitev - kvantili



Standardizirani odklon

- $x = \mu + z^* \sigma$

$$z = \frac{x - \mu}{\sigma}$$

- Nakažemo mesto enote v populaciji
- Velja:
  - $x = \mu \Rightarrow z = 0$
  - $x > \mu \Rightarrow z > 0$
  - $X < \mu \Rightarrow z < 0$

Standardizirana normalna porazdelitev

- N(0,1) oz. $\mu_z=0$; $\sigma_z=1$

$$z = \frac{x - \mu}{\sigma}$$

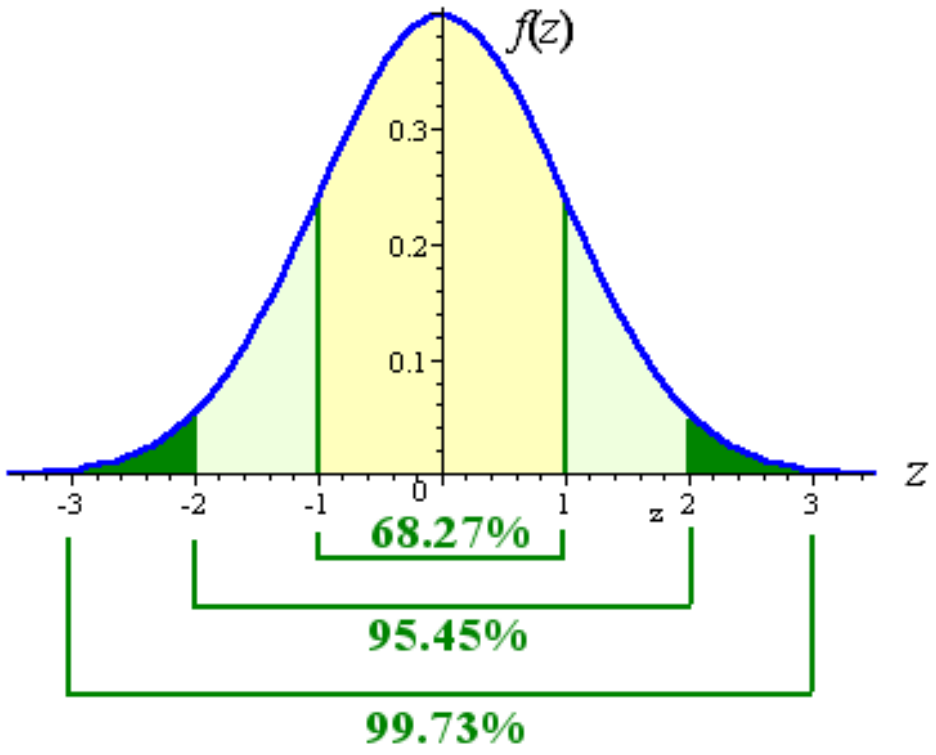
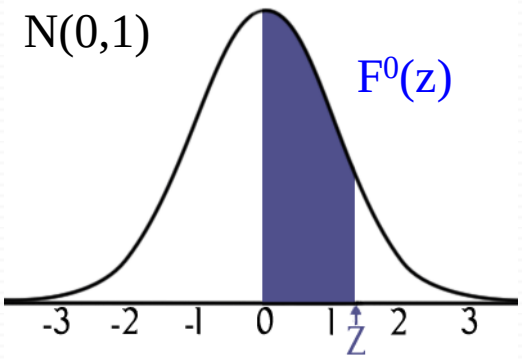


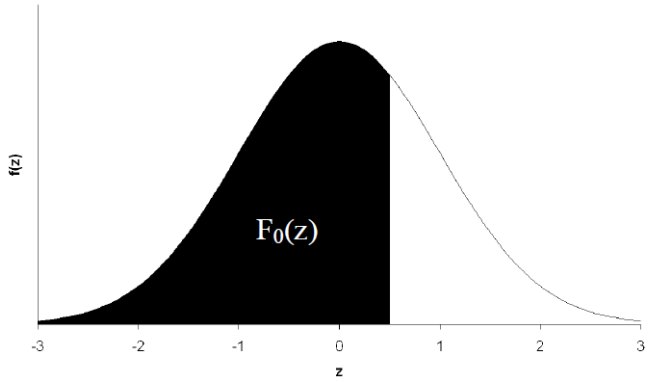
Tabela standardizirane normalne porazdelitve $F_{(z)}$



$F^0(z)$: Kumulativna frekvenca pod krivuljo standardizirne normalne porazdelitve = površina pod krivuljo standardizirane normalne distribucije med aritmetično sredino (0) in dano vrednostjo *z*.

z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Tabela standardizirane normalne porazdelitve $F_{(z)}$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830

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