Distribucije

Distribucija

- u statistici označava raspodjelu rezultata, odnosno frekvenciju kojom se u nekom skupu rezultata, poredanih po veličini pojavljuju pojedini rezultati.

Provjera oblika distribucije

- Statistička analiza podataka počinje provjerom oblika distribucije i nastavlja se njezinom statističkom deskripcijom (određivanjem osnovnih statističkih vrijednosti- središnjih vrijednosti, varijabiliteta i sl.)
- Oblik distribucije može ukazati na to uz koji model pristaju dobiveni rezultati. To omogućuje interpretaciju rezultata, a osim toga, podatak o tome da li distribucija odstupa od određenog modela ili ne, utječe i na odabir daljnjih postupaka statističke obrade.
- Postoji veći broj matematički opisanih distribucija.

Podjela distribucija

na distribucije za:

- **kontinuirane varijable**
- **diskretne**

Ako varijabla može poprimiti bilo koju vrijednost između neke dvije specificirane vrijednosti radi se o *kontinuiranoj* varijabli (pr. težina vatrogasaca je propisana od 50 kg do 130 kg,; bilo koja vrijednost).

Ako ne može, varijabla je *diskretna* (npr. koliko je puta pala glava kod bacanja novčića: od nula do plus beskonačnosti, ali cijeli broj).

- Teorijske distribucije za diskretnu varijablu jesu : binomna i Poissonova.

- Teorijske distribucije za kontinuiranu varijablu jesu : normalna (Gaussova), Studentova t-distribucija, F-distribucija.

Binomna distribucija

- **(najjednostavnija) teorijska distribucija za alternativna obilježja.**
- **pokazuje vjerojatnost događanja međusobno isključivih događaja za svaki broj slučajeva posebno.**
- **U statistici se model binomne distribucije koristi za rezultate u dihotomnim varijablama u kojima su podaci tipa točno – netočno, muškarci – žene i sl.**

Binomna distribucija je u svezi s Bernoullijevim pokusima

- Bernoullijev pokus je slučajni pokus ovih obilježja:

➢ Pokus ima dva ishoda (uspjeh, neuspjeh)

➢ U svakom ponavljanju pokusa vjerojatnost ishoda ˝uspjeh˝ = p i ne mijenja se od pokušaja do pokušaja.

➢ Vjerojatnost ishoda ˝neuspjeh˝ $q = 1 - p$

➢ Pokušaji su neovisni.

Primjeri s novčićem

- bacanje novčića: Pismo i glava su isključivi događaji i vjerojatnost njihovog pojavljivanja je 0,5.

Primjer s dva novčića

Ako bacamo 2 novčića, postoje **3 mogućnosti ishoda** bacanja:

1. na oba pismo
2. na oba glava
3. na jednom pismo na drugom glava

- Treću mogućnost dobivamo najčešće jer – moguće su **4 kombinacije:**
  I-pismo, II-pismo
  I-pismo, II-glava
  I-glava, II-pismo
  I-glava, II-glava

- Svaka od tih kombinacija je jednako vjerojatna, pa je p od svake 25%, od 2. i 3. zajedno 50%.

- Ako p i q zamijenimo s izrazima P i G dobivamo: $(G+P)^2 = G^2+2GP+P^2$
  što znači: jedanput 2 glave+ dva puta glava-pismo+jedanput dva pisma

  $(0.5+0.5)^2=0.5^2+2*0.5*0.5+0.5^2 = 0,25 + 0,5 + 0,25$

Primjer s četiri novčića

Ako bacamo 4 komada, postoji 16 mogućih kombinacija 5 ishoda):

	I novčić	II novčić	III novčić	IV novčić	
1.	P	P	P	P	(4P)
2.	P	P	P	G	(3P, 1G)
3.	P	P	G	P	
4.	P	G	P	P	
5.	G	P	P	P	
6.	G	G	P	P	(2P,2G)
7.	G	P	P	G	
8.	P	P	G	G	
9.	P	G	G	P	
10.	P	G	P	G	
11.	G	P	G	P	
12.	P	G	G	G	(1P,3G)
13.	G	P	G	G	
14.	G	G	P	G	
15.	G	G	G	P	
16.	G	G	G	G	(4G)

4P 6.25% slučajeva

3P,1G oko 25% slučajeva

2P2G oko 37.5% slučajeva

1P3G oko 25% slučajeva

4G oko 6.25% slučajeva

$$(p+q)^4 = (G+P)^4 = G^4+4G^3P + 6G^2P^2 + 4GP^3 + P^4$$

Kako to zapravo izračunavamo…

Vjerojatnost pojedinih kombinacija (P, G) izračunavamo pomoću binomne raspodjele: $\mathbf{(p+q)^n}$

S tim da je:

p-vjerojatnost da će se nešto dogoditi (npr glava)

q-vjerojatnost da se nešto neće dogoditi (ne-glava, tj.pismo)

eksponent **n** – broj faktora (u našem pr.je to br novčića)

(p+q) je uvijek 1 odnosno 100%

Binomni poučak (lat. ex binis nominibus – iz dvije oznake) je pravilo prema kojem se potencija (n=bilo koji neneg. br) nekog binoma (=matematički izraz koji se sastoji od dvije veličine povezane oznakom + ili -) razvija.

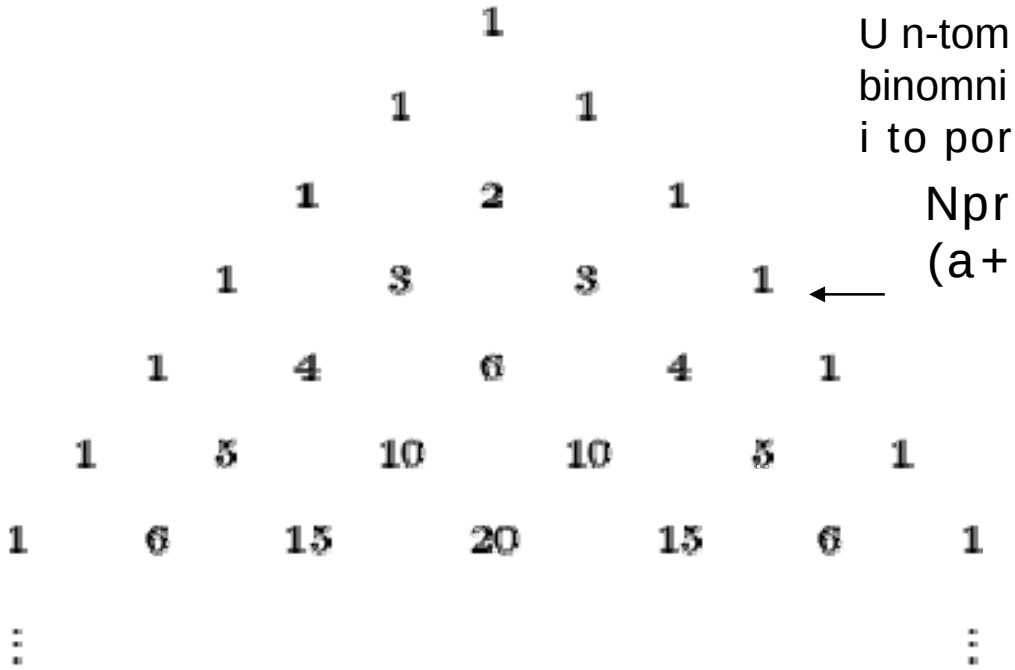
$$(a+b)^n = \sum_{k=0}^{n} \frac{n!}{(n-k)!\,k!}\, a^{n-k} b^k$$

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

$$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n$$

$$(a+b)^0 = 1$$
$$(a+b)^1 = a + b$$
$$(a+b)^2 = a^2 + 2ab + b^2$$
$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$
$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$
$$(a+b)^5 = a^5 + 5a^4b + 10\,a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$$

Parametre binomnog poučka, kombinacije, a time i očekivanu frekvenciju binomne distribucije lako dobivamo iz *Pascalovog trokuta*. Tako uz pomoć Pascalovog trokuta možemo utvrditi ove kombinacije i bez računanja.



U n-tom retku Pascalovog trokuta nalaze se binomni koeficijenti n-tog reda n=0,1,2,3..., i to poredani po razredu k=0,1,2,3...n.

Npr bacanje 3 novčića: n=3

$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$

1*3G 3*2G1P 3*1G2P 1* 3P

frekvencija kombinacija

Vidimo da je svaki element, osim rubnih, zbroj dvaju elemenata koji se nalaze s lijeve i desne strane u retku iznad.

Binomna vs normalna distribucija

- Ako postoji dovoljno veliki broj takvih događaja (povećavajući br. novčića), dobili bismo konačno praktički potpuno pravilnu zvonastu ili normalnu raspodjelu.
- Ipak, razlika između binomne distribucije i normalne distribucije je u tome što binomna nastaje kombinacijom faktora kojima je pojavljivanje uvijek jednako vjerojatno, a kod normalne je situacija nešto drugačija

  (npr. kada bismo imali mnogo novčića koji nisu ispravni, tako da je svaki novčić po slučaju svinut, pa oko polovice novčića ima veću vjerojatnost da padne na glavu, a oko polovice na pismo, i takve novčiće bacamo, dobit ćemo krivulju rezultata koja će biti jednaka krivulji binomne raspodjele kada je N veliki broj).

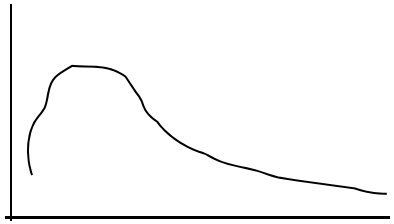
Poissonova distribucija

- je raspodjela vrlo rijetkih slučajnih događaja (kod kojih je vjerojatnost pojavljivanja vrlo mala; ako je p veoma malen, tj. ako je $p \leq 0.1$, a $n \geq 50$, tada se binomne vjerojatnosti mogu izračunati aproksimativno pomoću funkcije koju je otkrio Poisson).
- Izražava vjerojatnost broja događaja ako se ti događaji pojavljuju u fiksnom vremenskom periodu s poznatom prosječnom brzinom pojavljivanja i vremenski su nezavisne od prošlog događaja.
- Za razliku od normalne distribucije koja je potpuno definirana aritmetičkom sredinom i standardnom devijacijom, Poissonova distribucija je potpuno definirana aritmetičkom sredinom, jer je njena varijanca jednaka aritmetičkoj sredini. To znači da je ta distribucija šira što joj je aritmetička sredina veća.
- Kada je N vrlo velik, Poissonova distribucija se približava binomnoj, ali je razlika u tome što kod binomne raspodjele znamo koliko se puta neki događaj pojavio, ali i koliko se puta nije pojavio, a kod Poissonove raspodjele znamo samo koliko se puta neki događaj pojavio.

Poissonova distribucija

Npr.

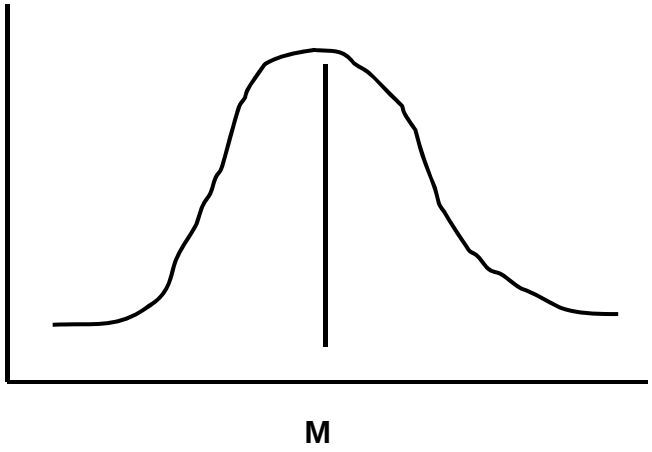
F osoba



broj nesreća na poslu zadnjih 10 god

NORMALNA (GAUSSOVA) DISTRIBUCIJA

- je najvažnija distribucija u statističkoj teoriji.



Graf normalne distribucije

- naziva se normalna ili zvonolika krivulja.

- Takav oblik distribucije rezultat je dvije tendencije ili sile koje djeluju na rezultate:

➢ **tendencija koncentriranja rezultata** koja je uvjetovana konstantnim faktorima (najčešće je to veličina pojave ili predmeta mjerenja ili opažanja)

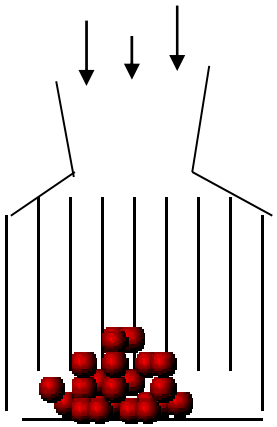
➢ **tendencija raspršivanja rezultata** koja je uvjetovana nesistematskim varijabilnim faktorima

Tendencija grupiranja i raspršenja rezultata

- Ako mnogo puta mjerimo neku pojavu
- koja je takva kakva je (to je tendencija postizanja jednakog rezultata),
- pri mjerenju radimo (svjesno ili nesvjesno) manje ili veće pogreške, pa se zato rezultati pojedinačnih mjerenja razlikuju (to je tendencija razlikovanja rezultata).

Nesistematski varijabilni faktori po slučaju skreću mjereni rezultat čas na jednu čas na drugu stranu, pa se ta skretanja najčešće međusobno ukidaju te zato dobivamo i najviše rezultata koji odgovaraju pravoj vrijednosti mjerene pojave, koja odgovara konstantnim faktorima.

Galtonova daska s čavličima: kuglice se sipaju kroz lijevak u kutiju s čavlićima:



- stavljanje kuglica u sredinu – tendencija grupiranja
- čavlići koji ometaju kuglice – tendencija raspršenja

Da bi se pri nekom mjerenju dobila normalna distribucija, moraju biti ispunjeni neki uvjeti:

- Ono što mjerimo moralo bi se i u prirodi normalno distribuirati (prevladava mišljenje da se gotovo sve u prirodi normalno distribuira, ali to nije točno, npr. bilirubin u krvi daje asimetričnu raspodjelu, dijametar srca daje bimodalnu raspodjelu, težina blago asimetričnu raspodjelu itd)
- Da postoji veliki broj rezultata –zakon vjerojatnosti (kod malog broja mjerenja neke pojave pa bila ona i idealno normalno distribuirana u prirodi, pukim slučajem možemo dobiti distribuciju koja nimalo ne sliči normalnoj)
- Da su sva mjerenja provedena istom metodom i u što sličnijim vanjskim prilikama (npr. mjerenje težine s odjećom/bez odjeće)
- Skupina na kojoj se vrše mjerenja morala bi biti homogena po ostalim svojstvima, a heterogena po svojstvu koje se mjeri. Npr. kod mjerenja visine da su homogeni po dobi, spolu i sl, a heterogeni po visini.

Normalna distribucija

- je matematički posve točno definirana (kompleksna formula), te je posve definirana ako joj znamo aritmetičku sredinu i sd.
- Mjesto infleksije (gdje iz konveksne prelazi u konkavnu) iznad ±1sd
- Potpuno je simetrična distribucija, zvonkolikog oblika, koja se asimptomatski približava osi apscisi.
- Svi koeficijenti asimetrije kod normalne krivulje su nula, budući da su kod simetrične distribucije M i C jednaki (npr. indeks asimetrije **α3** = [3*(M-C) / sd] ili **α3** = $m3/sd^3$).
- Vrijednost koeficijenta zaobljenosti ili kurtičnosti je kod normalne distribucije jednak 3 (**α4** = $m4/sd^4$)

Moment

- je fizikalni pojam kojeg je uveo K. Pearson.
- U statistici postoji više momenata, a definiraju se razlikom između svakog pojedinog rezultata i aritmetičke sredine svih rezultata.
- Matematički je definiran kao

$$mi = \frac{\sum x^i}{N}$$

$$mi = \frac{\sum x^i}{N}$$

gdje je:

- mi - moment prvog, drugog, trećeg ili četvrtog reda
- x^i -odstupanje svakog pojedinog rezultata od aritmetičke sredine u nekoj distribuciji rezultata podignuto na i-tu potenciju (potencija momenta prvog reda je 1, drugog 2 itd.)
- N - broj rezultata koji čini neku distribuciju.

Moment prvog reda – iznosi nula i njime je definirana aritmetička sredina

$m1 = \sum (X - M) / N$

Moment drugog reda - varijanca

$m2 = \sum (X - M)^2 / N$

Moment trećeg reda – (a)simetričnost

$$m3 = \sum (X - M)^3 / N$$

- Kada je m3= 0 distribucija je simetrična, m3>0 pozitivno asimetrična, m3<0 negativno asimetrična (slika).

- Koeficijent asimetrije α3 je omjer trećeg momenta oko sredine i sd podignute na treću potenciju

  $\alpha 3 = m3/sd^3$.

  Koef asimetrije poprima vrijednosti od najčešće +/-2

Moment četvrtog reda-kurtičnost ili zaobljenost

$$m4 = \sum (X - M)^4 / N$$

- Koef zaobljenosti $\alpha 4 = m4/sd^4$

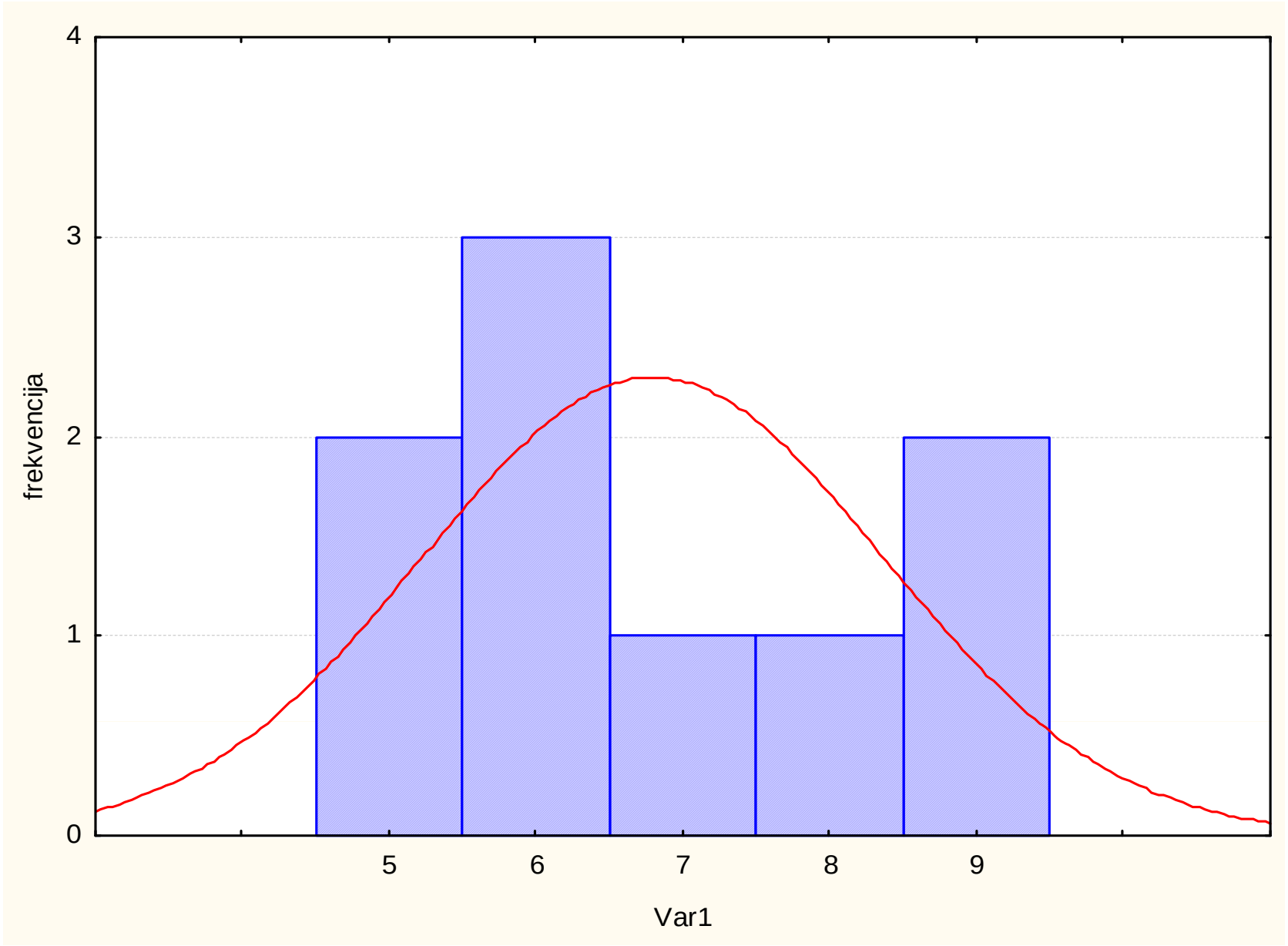
- Koef zaobljenosti normalne distribucije je 3. Takva distribucija je mezokurtična.
- Ako je veći od 3, distribucija je leptokurtična (šiljastija višeg i užeg vrha), ako je manji od 3 platokurtična (plosnatija).

Primjer

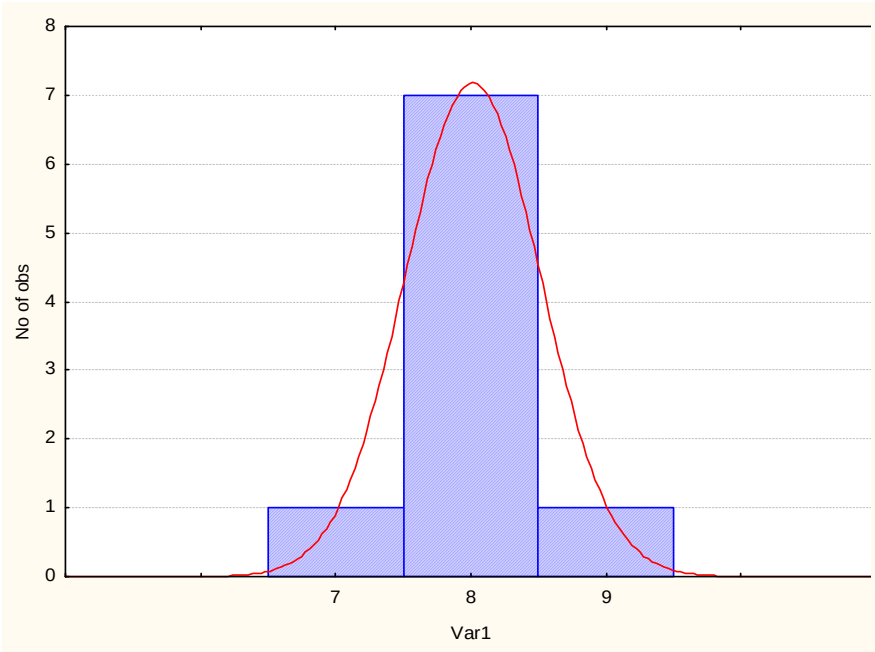
	1 v	2 m1 (v-6,78)	3 m2	4 m3	5 m4
1	5,00	-1,78	3,17	-5,64	10,04
2	5,00	-1,78	3,17	-5,64	10,04
3	6,00	-0,78	0,61	-0,47	0,37
4	6,00	-0,78	0,61	-0,47	0,37
5	6,00	-0,78	0,61	-0,47	0,37
6	7,00	0,22	0,05	0,01	0,00
7	8,00	1,22	1,49	1,82	2,22
8	9,00	2,22	4,93	10,94	24,29
9	9,00	2,22	4,93	10,94	24,29
SUM case 1		-0,02	19,5556	11,005432	71,983899

m1=-0,0022 m2=2,17 m3=1,22 m4=7,99

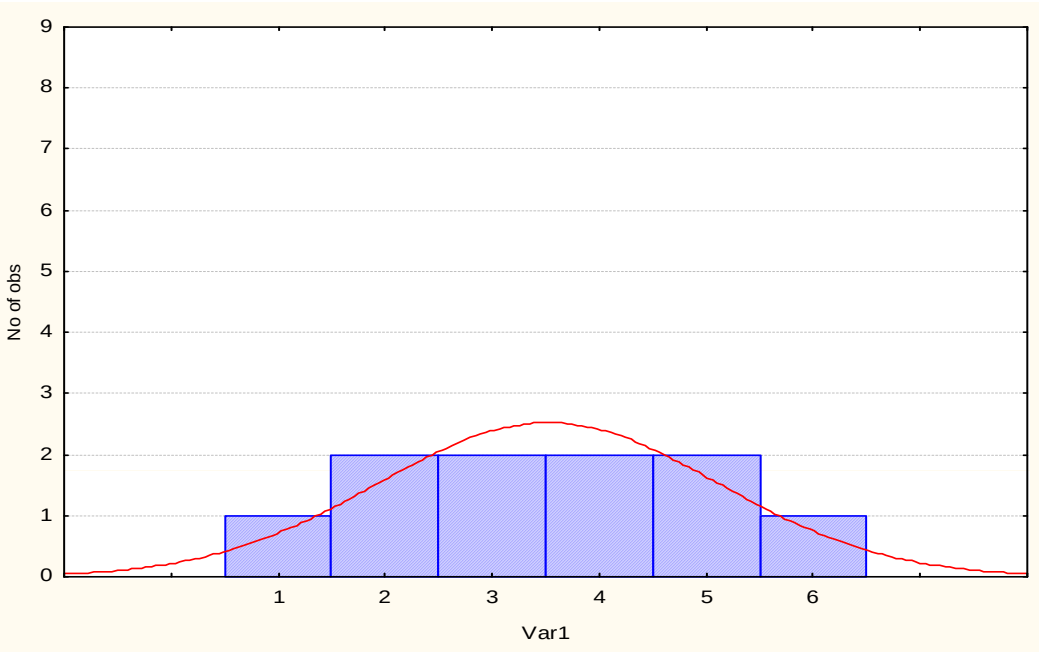
α3=0,39 α4=1,71

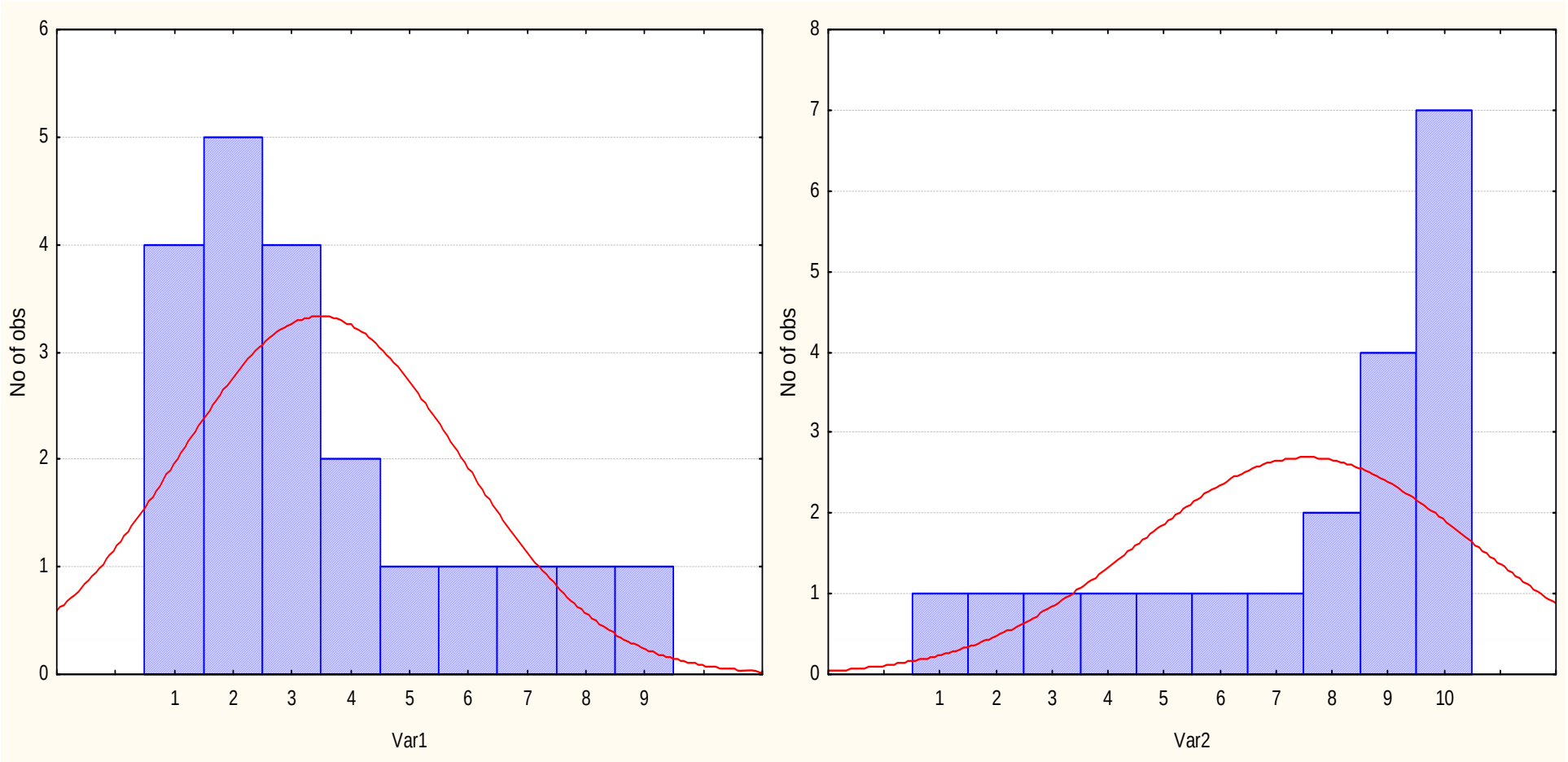

4
3
2
1
0
frekvencija
5
6
7
8
9
Var1

Leptokurtična



Platokurtična





Pozitivno asimetrična **Negativno asimetrična**

Još o normalnoj distribuciji…

- Normalna distribucija je jedan od osnovnih pojmova statističkog rezoniranja jer je osnova za razumijevanje glavnih statističkih pojmova vjerojatnosti.

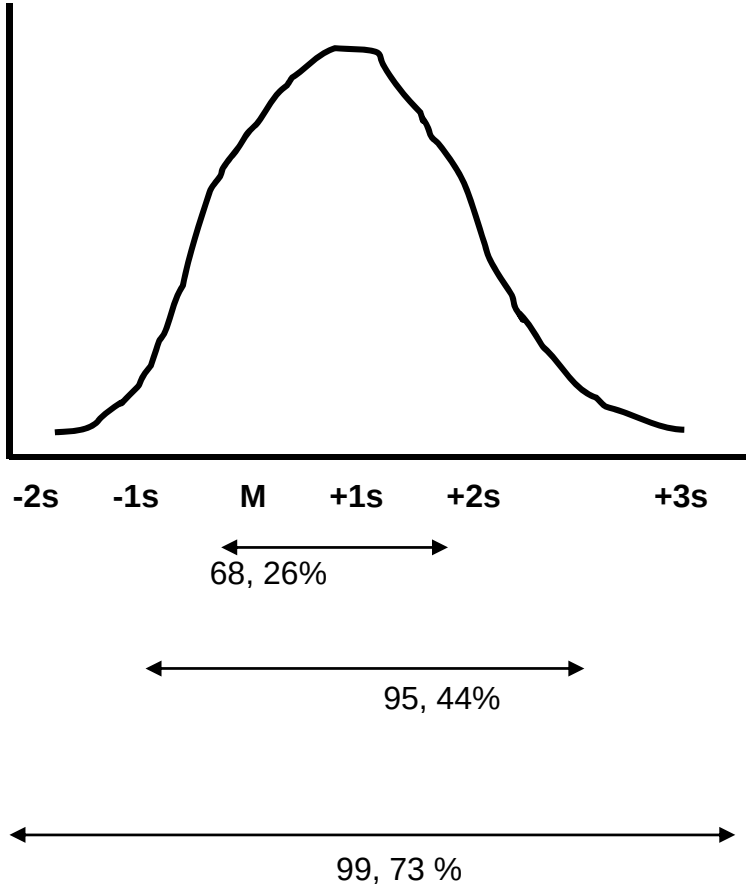
- Ukupna površina normalne distribucije se bilježi sa 1,0 ili 100 %.

Ako aritmetičkoj sredini dodamo lijevo i desno po jednu standardnu devijaciju, obuhvatili smo površinu koja čini oko 68% cijele površine krivulje, odnosno 68,26% svih rezultata.

S dvije s.d oko aritmetičke sredine, obuhvaćamo oko 95, 44% svih rezultata,

a s tri standardne devijacije gotovo sve rezultata, tj. 99,73% rezultata.

Doslovno se ne mogu obuhvatiti svi rezultati i s nekoliko s.d., jer se krivulja normalne distribucije asimptomatski približava apscisi , pa se teoretski spajaju u beskonačnost.



Zašto je to važno?

- Ako je neki rezultat točno na +1s, onda je lako izračunati koliko je udaljen od drugih rezultata:

- postoji oko 16 % rezultata koji su bolji od njega
- oko 34 % rezultata do aritmetičke sredine
- ili oko 84 % rezultata koji su slabiji od njega…

Primjer 1.

Na jednom testu iz statistike prosječno osvojeni broj bodova bio je 20, a standardna devijacija bila je 3. Odrediti koliki broj bodova se može očekivati kod najslabijeg studenta u grupi, ukoliko se rezultati ove grupe približno raspoređuju prema normalnoj distribuciji.

Primjer 2.

Odrediti površinu ispod normalne krivulje

a) lijevo od $x=0,33$

b) lijevo od $x=1,25$

c) desno od $x=1,25$

d) između $x=-0,54$ i $x=0,57$.

Primjer 3.

Broj kupaca u jednom supermarketu ima približno normalan raspored sa srednjom vrijednošću M=180 i SD=9. Odrediti vjerojatnost da će u toku dana broj kupaca biti

veći od 200

manji od 155.

Primjer 4.

Broj mušterija subotom u jednom kozmetičkom salonu ima približno normalan raspored sa srednjom vrijednošću M=28 i SD=4. Odrediti vjerojatnost da će do sljedećeg petka broj mušterija biti

a. veći od 35

b. manji od 22

c. veći od 22 a manji od 35.